Reading In-Between the Lines: An Analysis of Dissenter

Erik Rye
CMAND
rye@cmand.org

Jeremy Blackburn
Binghamton University
jblackbu@binghamton.edu

Robert Beverly
Naval Postgraduate School
rbeverly@nps.edu

ABSTRACT

Efforts by content creators and social networks to enforce legal and policy-based norms, e.g. blocking hate speech and users, has driven the rise of unrestricted communication platforms. One such recent effort is Dissenter, a browser and web application that provides a conversational overlay for any web page. These conversations hide in plain sight – users of Dissenter can see and participate in this conversation, whereas visitors using other browsers are oblivious to their existence. Further, the website and content owners have no power over the conversation as it resides in an overlay outside their control.

In this work, we obtain a history of Dissenter comments, users, and the websites being discussed, from the initial release of Dissenter in Feb. 2019 through Apr. 2020 (14 months). Our corpus consists of approximately 1.68M comments made by 101k users commenting on 588k distinct URLs. We first analyze macro characteristics of the network, including the user-base, comment distribution, and growth. We then use toxicity dictionaries, Perspective API, and a Natural Language Processing model to understand the nature of the comments and measure the propensity of particular websites and content to elicit hateful and offensive Dissenter comments. Using curated rankings of media bias, we examine the conditional probability of hateful comments given left and right-leaning content. Finally, we study Dissenter as a social network, and identify a core group of users with high comment toxicity.

CCS CONCEPTS

• **Networks → Social media networks**; • **Information systems → Social networks**.

KEYWORDS

Dissenter, Social Networks, Toxicity

ACM Reference Format:
Erik Rye, Jeremy Blackburn, and Robert Beverly. 2020. Reading In-Between the Lines: An Analysis of Dissenter. In *ACM Internet Measurement Conference (IMC '20), October 27–29, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3419394.3423615

1 INTRODUCTION

Virtual communities and discussion permeate modern society, and have fundamentally changed the way information is disseminated and consumed. Platforms that support such information exchange face not only technical challenges, but also legal and policy-based content concerns. Recently, major platforms have established and enforced policies to restrict hate speech [39] and users that engage in it. Participants of these communities have therefore moved to new platforms that either passively tolerate or actively support this content. For example [28, 42] examine speech and political characteristics of Gab, while [40] characterizes hate speech in Twitter.

Dissenter began as a web browser plugin, but, after being banned from the major browsers' respective plugin and extension stores [5], morphed into a self-contained, full-fledged browser based on Brave [7, 9]. Characterizing itself as the "free speech web browser," Dissenter provides, for any URL, a tightly integrated discussion forum specific to that URL. Only Dissenter users see this discussion forum. Thus, while e.g. a news site may have its own discussion forum for a particular article, Dissenter provides a parallel universe where its community of users are free to discuss (presumably a dissenting opinion) without restriction. Notably, the website and content owners have no power over this discussion forum as it resides in an overlay outside their control.

Similar forms of web annotation and augmentation have been created in the past, e.g. Google Sidewiki [1] (now defunct) and Hypothesis [6]. These efforts, however, were launched in an era pre-dating restrictions on social media content and not aimed at freedom of speech or providing a platform for fringe groups to discuss particular websites and content. In our work, the first to attempt to measure and characterize Dissenter, we obtain a history of Dissenter comments, users, and the websites being discussed from the initial release of Dissenter in Feb. 2019 through Apr. 2020 (14 months). We find more than 101k Dissenter users contributing more than 1.68M comments on 588k unique URLs.

Given recent debate surrounding censorship and the role of social media platforms in society – with the United States President signing executive orders to prevent censorship – our work is especially timely [38]. Toward a deeper understanding of Dissenter as an emergent platform, we make the following contributions:

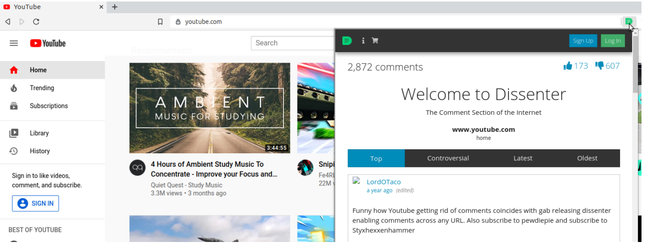
- Characterization of the Dissenter user base, including the intersection with Gab and Reddit
- Analysis of the Dissenter social network, including influential users and the set of users within individual comment threads.
- Classification of the toxicity of Dissenter comments and correlation of classes with both the political bias of the Uniform Resource Locators (URLs) being commented on, as well as its content.

2 DISSENTER

In this section, we describe Dissenter, its relationship to its parent application, Gab, and discuss major changes Dissenter has undergone since its launch in early 2019. We then define Dissenter-specific terminology.

IMC '20, October 27–29, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8138-3/20/10…$15.00
https://doi.org/10.1145/3419394.3423615



Funny how Youtube getting rid of comments coincides with gab releasing dissenter enabling comments across any URL. Also subscribe to pewdiepie and subscribe to Styxhexxenhammer

Figure 1: Dissenter browser view of YouTube

2.1 History

A description of the Dissenter plugin, browser, and comment overlay system, necessarily begins with Gab [36]. Gab was founded by Andrew Torba in 2016 as an alternative social network to more mainstream platforms; Gab counts among its users figures banned from Twitter such as provocateur Milo Yiannopoulos, and reality television star Tila Tequila. While Gab's stated purpose is to "champion free speech, individual liberty and the free flow of information online" [36], studies [27, 42] suggest Gab is primarily a fringe social network that contains hate speech and extremist content. Gab received significant attention after anti-Semitic hate speech was discovered on the Gab account of the individual responsible for the 2018 Tree of Life synagogue shooting in Pittsburgh [4, 44].

Dissenter, billed as the "Comment Section of the Internet", was released by Gab in February 2019 as a reaction to the disabling of "comment sections" of some websites by content providers, including YouTube. A full 25% of Dissenter users we examine in this study refer to "censorship" in their profile's biography, suggesting that perceived censorship is a common motivator for much of Dissenter's user base. In order to restore the ability for Internet users to comment on web content, Dissenter acts as a comment aggregation platform, receiving comments pertaining to URLs and displaying them to other users of its service. In this manner, Dissenter acts as a kind of overlay, displaying this "hidden" content only to users of Dissenter, while visitors that do not use the application remain unaware of its existence.

Dissenter initially took the form of a Firefox and Chrome browser extension which, when toggled, allowed users to post and view comments for a given URL. In April 2019, only two months after launching, the Dissenter extension was removed by both the Mozilla and Chrome extension stores. Both stores cited a terms of service violation, claiming that the extension was used to post hate speech [5]. Dissenter then morphed into a standalone browser by forking the Brave web browser [7, 9]. Figure 1 shows an example of the conversation overlay visible when viewing the YouTube home page using the Dissenter browser.

While providing a standalone browser extricated the Dissenter plugin from oversight by corporations that might otherwise attempt to crack down on the speech of its users, it requires users to switch their default web browser.

To augment the Dissenter browser, and provide a second method to access Dissenter comments, Gab deployed a news aggregation site called Gab Trends in October 2019 [8]. Gab Trends presents titles and short summaries of news articles from around the web, and includes Dissenter comment threads for each article. The comment thread visible via the Dissenter browser and Gab Trends is identical. Registered Dissenter users can participate in the Dissenter discussion by using the Trends web portal. Further, the Trends home page allows submission of new URLs. Upon submitting the new URL, the user is directed to a web page containing all of the Dissenter comments that have previously been made about this URL; if the URL is new to the Dissenter and Gab Trends system, this page contains no comments, but allows new users that navigate to it to make comments about this URL.

The popularity of Gab Trends is disputed. Gab itself claims 3M monthly views [2], while independent sources estimate 67k unique views per month [3].

2.2 Terminology

Analogous to other social networks, Dissenter users have a *home page*. A home page lists their username (a unique handle e.g. "`@a`"), display name (which may differ from the username, e.g. "Andrew Torba"), a biographical statement, and a profile image. Importantly, home pages list all of the URLs the user has commented upon.

Each URL that has received a Dissenter comment or been entered into Dissenter has a *comment page*. A comment page is analogous to a home page for a particular URL. Each comment page contains a title, which generally corresponds to the title content in the HTML of the page the URL points to, and a brief description of the content, which is typically generated by the first paragraph at the underlying URL. Some exceptions to title and description content exist, particularly when the content being commented upon is from another social network. For instance, both YouTube and Twitter content is typically embedded in the comment page by the Dissenter system, leading to ambiguous or altogether absent titles and descriptions. In order to make up for this lack of content data from Dissenter itself, we handle YouTube separately, which we describe in §3.3. Each comment page contains all of the Dissenter comments that have been made pertaining to the URL, and all of the replies to those comments.

While uncovering the operation of Dissenter, we find several undocumented identifiers within their HTML and JavaScript. We use these unique 12 byte Dissenter identifiers to prevent duplication and ensure uniqueness of users and content. Each user has a unique 24 hexadecimal digit *author-id*. Similarly, each distinct URL in Dissenter has a 12 byte *commenturl-id* identifier. Finally, every comment and reply is also assigned a 12 byte *comment-id*.

We discovered that these identifiers are not entirely random or a hash, but rather contain some structure. Analyzing the identifier, we find that all three encode state about their creation time. The first 4 bytes of the *author-*, *commenturl-*, and *comment-ids* are a Unix timestamp in seconds that describes the creation of entity; for example, an account created on February 28, 2019 at 16:23:53 UTC, will have an *author-id* beginning with `5c780b19`. Similarly, a *commenturl-id* encodes the first time a URL appears in Dissenter. While there appears to be additional structure in the remaining 16 hexadecimal digits, we are unable to determine its meaning as of this writing. In order to verify these findings, we created our own Dissenter accounts and posted innocuous content.

Finally, when a user posts a comment or reply in Dissenter, they have the option to label it as Not Safe For Work (NSFW). By default, these posts are invisible both to unauthenticated and authenticated

Dissenter users; in order to view this content, a logged-in user must explicitly "opt-in" via the Dissenter settings page. Because NSFW posts are hidden from all but authenticated users that have opted-in, this effectively creates hidden content within a shadow overlay. Similarly, an "offensive" label also exists for Dissenter comments, although unlike NSFW, it is not tagged by the user creating the content. As with NSFW content, an authenticated user must opt-in to viewing "offensive" comments.

3 METHODOLOGY

The primary interface to Dissenter is their browser or website. While Dissenter has an API, it is neither documented nor intended for public use. Hence, our methodology required basic reverse engineering of their platform and the application of several techniques to completely and programmatically gather Dissenter-internal data including users, user meta-data (including the social network), comments, and the URLs for each comment thread. Using this methodology, we effectively mirror the Dissenter database.

This section first details our data collection campaign and shows the steps taken to verify its accuracy and completeness. We then augment the Dissenter data by gathering the content of selected URLs commented upon for additional context. Finally, we describe our comment content classification techniques to enrich our understanding of comment content and user behavior. Ethical considerations of our work are provided in §8.

3.1 Gab-Based Username Harvesting

While Dissenter home pages provide important data on users and links to comments, there is no publicly available central database of usernames. Thus, a central component of our methodology is to harvest Dissenter usernames. Note that registering for Dissenter requires an active Gab account. Therefore, Dissenter users are necessarily Gab users, and we leverage this fact in order to enumerate Dissenter users before beginning to crawl other Dissenter content.

Initially, we attempted to gather Gab usernames via a combination of mining Pushshift.io [16] and crawling the most popular Gab account's ("@a", belonging to Gab founder Andrew Torba) followers, which is automatically followed by new users on the platform when their account is created. However, this methodology failed to uncover users that had not posted on Gab or had manually ceased following @a, and our results suggested a period of time before the @a handle was automatically followed by new users.

As discussed in §2.2, each Dissenter user account is associated with a unique identifier. In a similar vein, Gab accounts also have a unique user identifier. Unlike Dissenter's *author-id*s, however, Gab user IDs do not encode creation time, but are instead a counter beginning at 1, the user ID associated with "@e", belonging to former Gab Chief Technology Officer (CTO) Ekrem Büyükkaya. Having created a test account for which the Gab ID is known, we query the Gab API endpoint https://gab.com/api/v1/accounts/<GabID> for IDs between 1 and our account's ID to retrieve JSON-encoded information pertaining to that user. Gab's API helpfully returns an error when an ID is not associated with a user account, and in this manner, we are able to exhaustively enumerate Gab's user base. Among the information present in the user data JSON is the account creation date and time, which largely confirms the hypothesis that the Gab ID is a monotone increasing counter. Some exceptions to monotonicity exist in which Gab assigned unallocated, lower-valued ID numbers to new user accounts; whether these ID values became free after a user deleted their account, or whether gaps were deliberately placed between consecutively allocated IDs earlier in its history is unclear. Figure 2 shows when the account associated with each Gab ID number was created; apart from two distinct time periods in mid-2017 and mid-2019, Gab IDs are generally assigned sequentially and are strictly increasing.

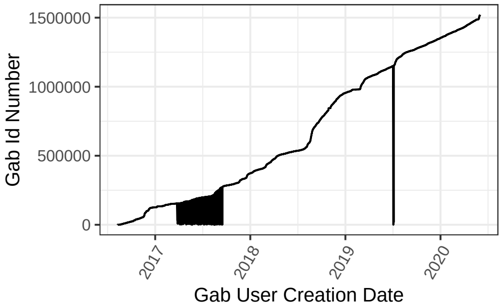


Figure 2: Gab User IDs Assigned to New Accounts Over Time

This enumeration process reveals 1.3M distinct accounts, a significant number more than discovered in prior work in 2018 [42], which discovered 336k users. In addition to the two years separating [42] and our study, the difference is explained by a large number of discovered Gab users that have not posted any messages, do not follow any other Gab users, and are similarly not followed by any other users. While many or most of these accounts may belong to inactive users, we find several thousand "friendless" and "silent" Gab users that are otherwise active on Dissenter. Finally, because our Gab user discovery methodology enumerates all Gab users, our Dissenter results in §4 do not suffer from questions of completeness or sensitivity that a user harvesting strategy based on spidering a high-degree user's followers might.

Next, we determine which Gab users are also Dissenter users. For each Gab username discovered, we send an HTTP request to the URL of the corresponding Dissenter home page, if it exists (https://dissenter.com/user/<Gabusername>). Based on the HTTP response sizes, we are able to identify Dissenter accounts, which are at least 10 kB; responses for non-existent users are ∼150 bytes. Of the 1.3M Gab usernames we enumerated via its API, 101k also have Dissenter accounts, representing approximately 8% of all Gab users.

3.2 Dissenter Comment Harvesting

With the usernames discovered in §3.1, we crawled Dissenter for URLs that users comment on, the comments they made about those pages, as well as replies to other users' comments. Our crawler first visits the home page of each Dissenter user to capture their meta-data, including username, display name, *author-id*, and biography. Then the crawler gathers the set of URLs the user has commented on.

We then iterate over the set of commented-upon URLs. For each URL, we visit its comment page in Dissenter and collect the *commenturl-id*, the number of comments and number of up- and down-votes the URL has received, as well as the title and brief description. As noted in §2.2, the title and description may be ambiguous or empty, depending on the underlying content the commented-upon URL describes and the ability of the Dissenter system to parse this data. Within each comment page, we iterate over the comments and replies. For each comment, we record the *author-id*, *comment-id*, and the comment text. Comment text appears to have no character limit; the longest comment we find is >90k characters, consisting of the word "ha" repeated 45k times in response to a YouTube video discussing Facebook's political bias. Replies can be made in response to both comments as well as to replies themselves; there also appears to be no practical limit on the depth at which replies might be made, e.g., a reply to a reply to a reply is valid. In addition to the data we collect for comments, we also note the *comment-id* of the content to which the reply is replying. Although the HTTP response headers indicate that a rate-limit of 10 requests per minute is employed by the servers hosting Dissenter content, this counter is *per-URL*; because we do not need to request the same URL twice in our crawl, we are unimpeded by this rate-limit.

Each comment is available from the URL https://dissenter.com/comment/<CID>/, where `<CID>` is the *comment-id*. We call this page the *comment-page*, as its purpose is to display a single comment as well as any replies. For reasons that are unclear, comment pages contain a JavaScript element with an unused (commented-out) JavaScript variable called "`commentAuthor`". `commentAuthor` defines an array with user data. While much of this embedded user data is identical to what is available to us via the user's home page, it also contains otherwise undiscoverable meta-data including the user's language setting, permissions, and view-filter preferences. We save these additional hidden meta-data as part of our per-user characterization.

To obtain the NSFW and "offensive" content described in §2.2, we re-spider Dissenter using the HTTP cookies of an authenticated account we created with NSFW and "offensive" content enabled separately, so that we are able to discern between content labeled NSFW by the submitter and comments marked as "offensive". These comments have no specific flag or other identifier present in the document body to indicate their presence; therefore, we infer NSFW and "offensive" comments as those found when authenticated with these flags enabled that were not previously discovered. In order to ensure we do not erroneously mislabel content as NSFW or "offensive" because of crawler errors, we monitor request timeouts and re-request missed pages, and ensure that subsequent runs only consider comments made during the initial spider's time frame. Additionally, we manually confirm 100 comments classified as NSFW or "offensive" by our subsequent crawls by attempting to view these comments both while authenticated with the NSFW and "offensive" view preferences enabled and while not authenticated. Our NSFW and "offensive" comment results are discussed in §4.3.1.

In total, we obtain 1.68M comments and replies on 588k distinct URLs made by >101k users via our methodology. Macro characteristics of these data are discussed in detail in §4.

3.3 YouTube Crawling

Typically, we rely on the title and description provided by the Dissenter application in the comment page in order to gain valuable context about URLs being commented upon. However, Dissenter's own methodology for URL content appears unable to handle the most popular source of commented-upon URLs: YouTube videos. These pages generally appear with the title "/watch" and a null description, although the video itself is embedded in the page. Therefore, because YouTube content in particular represents a sizable percentage of our data (128k URLs) and because we seek to understand the content that generates comments, we also gather the content of the underlying web page for YouTube Dissenter comments. Because YouTube pages require JavaScript to render properly, we use Selenium [13] to automate content retrieval. The data we seek (e.g., video title, uploader name) resides in large blocks of JavaScript, which may explain its absence from Dissenter. For each URL, we classify the content as one of three distinct types – "video", pages that contain a single YouTube video, "user", a home page for a particular YouTube user, and "channel", which is a collection of videos under a single banner.

3.4 Social Network Crawling

Finally, we return to Gab in order to gain context about the social network that Dissenter users inhabit. Social relationships on Gab are directional; much like in Twitter, a user may become a follower of another user, and may accumulate followers themselves. While Dissenter users are able to "follow" other Dissenter users, none of the Dissenter browser, plugin, user home page, web application, or hidden meta-data reveal followers, or allow even an authenticated user to view his or her followers and following users. Presumably this is because the social network aspect of Dissenter is a subset of Gab, and is an as-yet unimplemented part of the Dissenter experience. Therefore, we use Gab followers as a proxy, and because Gab users are a strict superset of Dissenter users, any two Dissenter users can follow each other on Gab.

We use the Gab API in order to obtain these relationships for further analysis in §4. Using the Gab API, we gather the followers and followed users of each Dissenter user. We note that Gab exposes its rate-limiting in the HTTP response headers by including the number of remaining requests, as well as the time at which the request limit will be refreshed. To minimize impact on the service, we issue at most one request per second, and monitor the number of remaining requests. If necessary, we wait until the number of available requests has been refreshed before continuing to issue new requests for Gab friends. Note that results from querying the Gab API for the social network are paginated, thus we can ensure that we gather the complete network graph.

Finally, by removing non-Dissenter users from the followers and those followed obtained by querying Gab, we construct a Dissenter-specific social network graph.

3.5 Classification

To gain a more complete understanding of Dissenter, we must understand the content and context of the comments and replies. In particular, we are interested in assessing the degree of toxicity and offensiveness. While significant prior work exists on automatically labeling hate and toxic speech, current approaches yield accuracies between 70-80% [19] and it remains an open research problem. For example, there are indications that the models encode racial bias [35], while some models can be deceived [26].

To underscore the difficulty of the problem, consider an innocuous comment about the country Pakistan. This comment could be construed as hateful as it contains the substring "paki," a false positive. However, not performing stemming and fuzzy matching can yield false negatives, for instance if the hate word is succeeded with a "z" when using slang. Words themselves are ambiguous and must be taken in full context. For example, the term "skank" can be used as a hate term or in reference to a style of dance.

Because our work is focused on characterizing Dissenter rather than improving the state-of-the-art in hate speech detection, we therefore explored multiple approaches to label comments in order to bound our estimates of its toxicity.

- ***Dictionary*** We utilize the modified Hatebase [11] dictionary of toxic terms used by the authors of [25] and [42]. This dictionary contains 1,027 hate words. We tokenize each Dissenter comment and reply, perform stemming, and then count the number of tokens that match a term in the dictionary. Our per-comment hate dictionary score is then the ratio of hate words over the number of tokens in the comment. While this metric is simple, it misses important context in the comments. For instance, the ambiguous terms "queen" and "pig" appear in the dictionary. However, by using the same dictionary as these prior works, we can draw direct meaningful comparisons.
- ***Perspective*** Next, we leverage the Google Perspective API. The Perspective API provides several models that provide scores for different aspects of toxicity. Perspective allows us to effectively outsource comment scoring; however, as with the other methods, it has limitations. The API is trained primarily on Wikipedia data, and thus there are some questions about its portability.
- ***NLP*** Finally, we employ Natural Language Processing models to build a three-class (hate, offensive, or neither) comment classifier. To train our classifier, we use labeled data from [19] which contains 1,194 hate, 16,025 offensive, and 20,499 neither labels of Twitter tweets gathered via crowd-sourcing. Because of the imbalanced complexion of data, we use ADASYN to oversample [23].

  We experiment with neural networks, decision trees, and support vector machines (SVMs) using 1 and 2-grams of cleaned and stemmed word tokens. Using grid search to tune the hyperparameters, we achieve the highest accuracy using SVMs. With 5-fold cross-validation, we achieve an F1 score of 0.87 on the Twitter training dataset. Using this SVM model, we compute the probability of each of the three possible classes for all Dissenter comments and replies.

To better understand the differences of each classifier as applied to Dissenter data, we evaluate all comments and replies with the three approaches and compare the resulting scores in Figure 3. We see that the classifiers largely agree – the distribution of Perspective toxicity for comments that score low with Hatebase or the NLP classifier is significantly skewed toward less toxicity, while those with high scores are skewed toward higher Perspective toxicity.

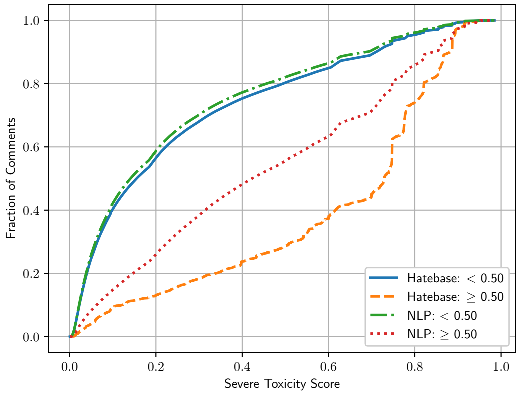


Figure 3: Comparing Dictionary and NLP Classifiers Against Perspective

For our purposes, issues surrounding toxicity classification are somewhat mitigated. First, we are less interested in scoring any *particular* comment, and instead are interested in aggregate trends and the distribution of scores. Second, we compare Dissenter to several baselines which gives us an idea of *relative* differences in scores across users and communities.

Recent efforts by Google have sought to make Perspective's performance more transparent [22], while third-parties have evaluated perspective on canonical datasets and found it to exhibit high precision and recall [30]. As such, Perspective provides a publicly available platform that has been independently validated. Because of the general agreement between classifiers and our use of the classification as a relative metric, we use Perspective for the remainder of our evaluation.

4 RESULTS

This section begins with a high-level characterization of the Dissenter platform, analyzing users and URLs being discussed. We then measure hate speech and toxicity in Dissenter as compared to other platforms, as well as investigate the relationship of toxicity to content and the social network of Dissenter users.

4.1 Dissenter Users

4.1.1 How popular is Dissenter? We first examine the set of Dissenter users we discover using the methodology of §3.1. As noted in §2.2, the *author-id* identifier encodes each account's creation time. Dissenter experienced a steep initial influx of users to the platform, as nearly 79k (77%) joined through the first full month of operation (March, 2019).

Of the more than 101k unique usernames we discover, approximately 47k (47%) commented on at least one URL. Considering only these active users that have made at least one comment, Figure 4 shows that approximately 90% of comments are made by about 14% of active users (7% of total users). The long tail of Figure 4 indicates that many users made a relatively small number of comments. We note that none of the top users fall into the top twenty Gab users by number of followers, score, or PageRank as determined by prior work [42], nor are they prominent in the Dissenter social network as will be shown in §4.4. Finally, we discover approximately 1,300 users who commented on URLs through our Dissenter crawl that

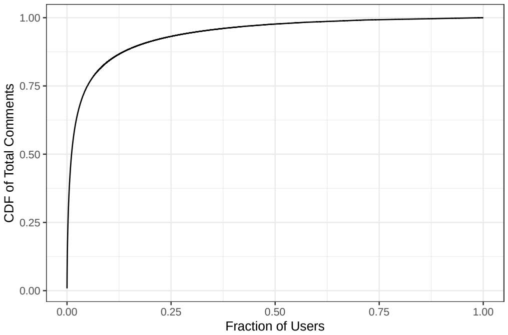


Figure 4: Dissenter Comments and Replies per Active User

did not appear in our enumeration of Gab's users in §3.1. This finding was surprising, as an active Gab account is a prerequisite for creating a Dissenter account. On closer examination, we discovered that these accounts appeared to be deleted by their owners, as their Gab home pages matched the appearance of an account that we test-deleted. Interestingly, these users' Dissenter accounts and comments remain despite the deletion of their Gab account.

Takeaways: Slightly more than half (53%) of Dissenter's users have not commented on a URL or replied to another user's comment. This does not necessarily mean that these users are inactive; users can interact with Dissenter by giving "thumbs up" or "thumbs down" on both the URL and other users' comments, but these actions are transparent to us.

While Dissenter's user base is a strict subset of Gab's, Dissenter is not simply a Gab in miniature. Its core group of users are extremely active on the site, posting thousands of comments in little over a year on web content that they presumably consume beforehand.

We discover more than 1,300 users whose Gab accounts were deleted. The comments left by these users remain on Dissenter, and because their Gab account no longer exists, they are unable to authenticate to delete these posts.

4.1.2 User Characterization. Using the embedded JavaScript data described in §3.2, we are able to more extensively characterize the 47k active users. Two Dissenter users are flagged as "isAdmin": `@a`, Andrew Torba's account, and `@shadowknight412`, which belongs to Rob Colbert, the Gab CTO. Despite the existence of "isModerator", no active accounts we queried had this moderator flag set, although it is possible one or more moderator accounts exist that do not post comments. Eight accounts were banned from the platform; of these, several had no obvious explanation for being banned based on the comments we obtained, while for others the reason is clear. For example, one account is clearly related to a home remodeling company and posted only advertisements for its business, while another posted what appears to be the home address of a federal official and expressed a desire for an "accident" to happen to that individual. Table 1 counts the frequency of all possible account flags across the active user set.

Users can further apply filters to show or hide comments on URLs based on categories of users or comment labels. For instance, "NSFW" is a label that a user may apply when posting a comment, and "pro" is a label that a designates a paid GabPRO account – a status that unlocks additional platform features, such as the removal of ads and ability to upload larger videos. Table 1 summarizes the number of positive responses for each comment filter. Of note, nearly all users choose to see content from "pro", "verified", and "standard" Dissenter accounts; this is unsurprising, as all three of these are applied by default. However, the "NSFW" and "offensive" preferences are by default disabled. We discover 18k comments classified by Dissenter as NSFW (∼10k) or "offensive" (∼8k), though we were unable to determine with certainty what moderation policies or user feedback generates an "offensive" classification. Because only 15% and 7% of active users enable the NSFW and "offensive" filters, respectively, these comments constitute a kind of shadow platform *within* Dissenter.

Takeaways: Artifacts in the Dissenter HTML source show that even in Dissenter, users get banned for speech that is deemed unacceptable. Further, the "view" flags indicate that NSFW and offensive content exists as a shadow overlay on the Dissenter overlay itself, viewable only to a small fraction of Dissenter users.

4.2 Content Analysis

Since Dissenter serves users concerned that their ability to comment directly on the source material might be curtailed, it is only intuitive to examine what content they discuss.

4.2.1 What URLs are being commented on? We discover 588k URLs that have been commented upon according to Dissenter's own unique identifier, the *commenturl-id*. However, this number overcounts unique content in two ways. First, Dissenter differentiates between URLs that differ only in the protocol portion; that is, the HTTP and HTTPS version of URLs will receive different *commenturl-ids*, separate comment pages, and can contain entirely different comment content. We observe 400 distinct URLs that differ only in the protocol part of the URL; another 60 differ only by the presence or absence of a trailing forward-slash character. Second, Dissenter's handling of URLs with HTTP GET query parameters causes unique content over-counting. Many of the URLs we observe contain several GET parameters separated by the "&" character; however, because page content is typically only determined by a single parameter, if at all, it is likely unnecessary to store more than the first key-value pair as part of the URL in the Dissenter system.

A full 97% (571k) of URLs in Dissenter are HTTPS; another 2% (15k) are HTTP, and a small fraction contain browser-specific protocols, such as `chrome://`. Thirteen URLs contain the `file` protocol, indicating that the URL points to a file on the user's file system. While most of these URLs point to Windows "letter drives" like `C:\`, several include file paths that appear to point to legitimate documents on the user's file system.

Of the URLs we discover within Dissenter, the overwhelming majority point to pages under the `.com` TLD (78%); the second-most-frequent TLD is `.uk` (7.5%). While `.be` rounds out the top five TLDs, it most frequently appears a domain hack for YouTube URLs (e.g. `youtu.be/id`) rather than for Belgian content.

In addition to the popular TLDs, Table 2 gives the most popular second-level domains by percentage of URLs. YouTube is by far the most common, comprising about 21% of all URLs we discovered with comments between the `youtube.com` and `youtu.be`

Table 1: User Attribute Flags and Comment View-Filters Enabled for Active Users (*n*=47,165)

User Flags						Comment Filters	
canLogin	47,152 (99.97%)	isBanned	8 (0.02%)	is_investor	137 (0.29%)	pro	47,093 (99.85%)
canPost	47,150 (99.97%)	isAdmin	2 (0.00%)	is_premium	61 (0.13%)	verified	47,103 (99.87%)
canReport	47,158 (99.99%)	isModerator	0 (0.00%)	is_tippable	73 (0.15%)	standard	47,112 (99.89%)
canChat	47,153 (99.97%)	is_pro	1,257 (2.67%)	is_private	1,838 (3.90%)	nsfw	7,094 (15.04%)
canVote	47,152 (99.97%)	is_donor	397 (0.84%)	verified	485 (1.03%)	offensive	3,456 (7.33%)

Table 2: Most Frequently Commented Top-Level Domains (TLDs) and Domains

Top Level Domains				Domain			
.com	455,885 (77.57%)	.au	6,892 (1.17%)	youtube.com	121,928 (20.75%)	foxnews.com	12,196 (2.08%)
.uk	43,808 (7.45%)	.ca	5,490 (0.93%)	twitter.com	40,392 (6.87%)	bitchute.com	12,124 (2.06%)
.org	19,502 (3.32%)	.net	4,787 (0.81%)	breitbart.com	23,705 (4.03%)	zerohedge.com	8,634 (1.47%)
.de	10,257 (1.75%)	.nz	2,979 (0.51%)	bbc.co.uk	16,213 (2.76%)	theguardian.com	8,010 (1.36%)
.be	8,013 (1.36%)	.no	2,928 (0.50%)	dailymail.co.uk	15,752 (2.68%)	youtu.be	7,819 (1.33%)
Other	27,194 (0.05%)	**Total**	587,735 (100%)	Other	320,962 (54.61%)	**Total**	587,735 (100%)

domains. Twitter content is the second most frequent at about 7% of all URLs. With the exception of Bitchute, a video hosting alternative to YouTube oriented towards the same general user base as Gab [37], the remainder of the top URLs are news-oriented websites. In contrast, when we rank domains by median comment volume per URL, YouTube ranks very low, with a median comment count of 1. Domains with the highest comment volumes per URL are typically fringe content with a small number of commented URLs. For instance, `thewatcherfiles.com`, a conspiracy aggregation site, ranks first with 116 comments on one URL about the Jewish Blood Libel; the second highest comment volume domain is `deutschland.de`, with 95 comments on a single URL, most of which express anger about the Muslim diaspora in Europe.

Takeaways: Dissenter comments are typically made on video streaming, social media, or news sites, with YouTube comprising the largest fraction of commented domains. However, domains with the highest comment volume per URL are disjoint from the set of highest commented URL counts and often contain fringe content. Dissenter users can comment on *any* URL, including those that are local to their own file system or are non-existent.

4.2.2 YouTube. YouTube content comprises a large fraction of the URLs commented on in Dissenter; Table 2 shows that ~22% of URLs we obtained with comments are YouTube content. Further, Dissenter comment-pages typically contain little information about the video itself, likely because this information is dynamically generated and thus difficult for Dissenter itself to mine. This creates difficulty in understanding the content at the URL, and is compounded by the fact that YouTube videos also have no inherent bias, making broad generalizations of this content effectively impossible.

Therefore, as described in §3.3, we gather and analyze the content of 128k YouTube URL present in our data. The majority of these YouTube URLs are videos: 125k are labeled video content, along with 2k channels and 1k users. Because YouTube videos may be taken down by the owner or the platform itself for a variety of reasons, we discovered only 109k active video pages, while 16k were unavailable. While the most common reason for video removal was a generic "Video Unavailable" label, 3k videos were listed as private and required permission to view, another 3k were unavailable because the YouTube account that had posted them was terminated, and nearly 400 were removed for violating YouTube's hate speech policy. It is noteworthy that even in the event that YouTube takes action to remove objectionable comment, Dissenter still provides a platform for users to comment on what *was* at that URL, serving as an ersatz digital history for the content that once existed there.

Each active video has a "content-owner", the name of the entity or individual that uploaded the content. Interestingly, Fox News and CNN – generally considered to be on opposite ends of the ideological spectrum – both appear in the top six most commented upon YouTube content producers. 2.4% of all YouTube videos that had a comment were produced by Fox News, as compared with 0.6% for CNN. Normalized by the fraction of all videos produced by each news source, 4.7% of all Fox News videos have at least one Dissenter comment while only 0.5% of CNN videos were commented upon. Slightly more than 10% of the active videos we crawl have their comment functionality disabled on the YouTube platform, reinforcing Dissenter's argument that it provides an outlet for users to express their opinions on content where it would otherwise not be allowed.

Takeaways: YouTube is a sizable fraction of all Dissenter comment URLs; the videos Dissenter users comment on frequently disallow commenting, and are often removed from YouTube altogether for a variety of reasons.

4.2.3 Are there language differences in Dissenter comments? Using the `langid` [10] language identification tool, we classify each of the 1.68M comments and replies in our dataset. Our results indicate that Dissenter comments are overwhelmingly in English (1.57/1.68M or 94%). German is the second most popular language with 31k (2%); this matches our expectations as `.de` is the fourth most-common TLD and first non-English-speaking country's Country-Code TLD (ccTLD) in Table 2. French, Spanish, and Italian complete the top five most-frequent languages, each with less than 0.5% popularity.

Takeaways: The vast majority (94%) of Dissenter comments are in English, with German the only other language achieving >1% representation.

4.3 Dissenter Toxicity

The concept of toxicity online has gained a bit of a spotlight lately. In a nutshell, toxicity is loosely defined as anti-social behavior that causes harm to a community at the social level. Things like harassment, hate speech, personal attacks, and trolling can all be considered toxic, as they reduce the inherent utility of the platform they occur on, as well as harm its underlying community of users. Previous work [28, 42], as well as articles and world events have indicated that Gab is more toxic than the average community.

In fact, at least part of the motivation behind Gab and Dissenter's existence is that its user base was considered too toxic for platforms like Twitter, and that the discussions they have are similarly considered unsuitable by many platforms. This raises several interesting questions that we aim to answer in this section using our content classification techniques in §3.5. Are Dissenter users particularly toxic? What kind of toxicity is exhibited?

4.3.1 NSFW and Offensive Comments. As described in §2.2, a user is presented with the option to label a comment or reply as NSFW when posting. This label prevents the comment from appearing to any user that has not explicitly opted-in to seeing this content. We find ∼10k comments (0.6% of all comments) tagged as NSFW that appear *only* in an authenticated crawl's results with NSFW-viewing enabled compared to an unauthenticated baseline crawl. Although the mechanism for a comment being classified as "offensive" is opaque to us, we similarly discover ∼8k (0.5% of all comments) labeled "offensive" by Dissenter. Posts can be "reported" by users, which informs the reporting individual that the content will be "reviewed as soon as possible." This mechanism is one hypothesis for how "offensive" comments become labeled; alternatively, Dissenter may attempt to automatically label comments based on the presence of certain words in the text.

In order to ensure that we did not obtain false positive classifications due to HTTP timeouts and other errors introduced from the crawling framework, we perform two validation steps. First, we keep track of any URL during our crawl that received a timeout error and mark those URLs for re-crawling at a later time. We repeat this process until all pages have been successfully parsed by our framework. Second, we select a random sample of 100 NSFW and "offensive" comments, and perform a manual validation to ensure that the comment only appears when authenticated and with the proper settings enabled. All 100 comments we manually verified were correctly classified as NSFW or "offensive", although several were posted both as NSFW and also without the label. Because a user cannot see even their own NSFW-labeled comments if they do not have this visibility setting enabled, it is possible that these duplicate posts occurred when a user posted a comment with the NSFW label enabled but did not see it appear, and re-posted without the label believing their first post was unsuccessful.

We find that NSFW content is more toxic than the standard comments and replies posted to Dissenter, and "offensive" comments are much more so. Figure 5 is a CDF of the Perspective scores attributed to NSFW and "offensive" comments vs the entire comment population in the Perspective categories of "OBSCENE", "SEVERE_TOXICITY", and content likely to be rejected by the New York Times' moderation section ("LIKELY_TO_REJECT".) In all three categories, we find the "offensive" content to be significantly more extreme than both the unlabeled comments and replies and the NSFW content. For instance, 80% of the "offensive" comments score > 0.95 in the LIKELY_TO_REJECT category, whereas only 25% of NSFW comments and $< 20\%$ of all comments score this high. The NSFW content is also more extreme than the aggregate Dissenter comments, but to a lesser degree than the "offensive" comments. This indicates that users are correctly using the NSFW label on some of the more extreme content on Dissenter, and that the "offensive" labeling mechanism captures the most radical and toxic content. However, because ∼85% of users do not have the NSFW or "offensive" view-settings enabled, including both Gab's founder and CTO at the time we ingested user settings in §4.1, these comments act as a network within a social network, where the most extreme content on the platform appears.

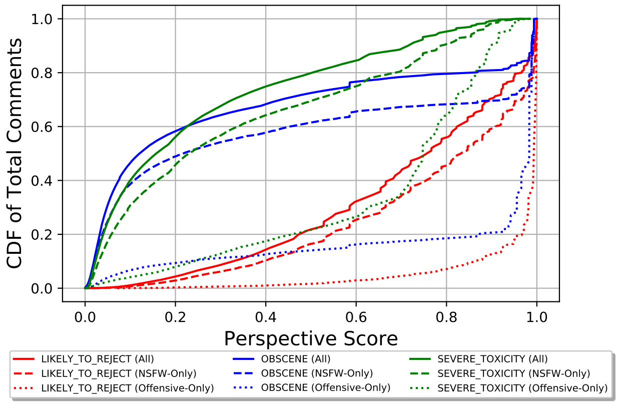


Figure 5: NSFW, Offensive, and Aggregate Comments Comparison

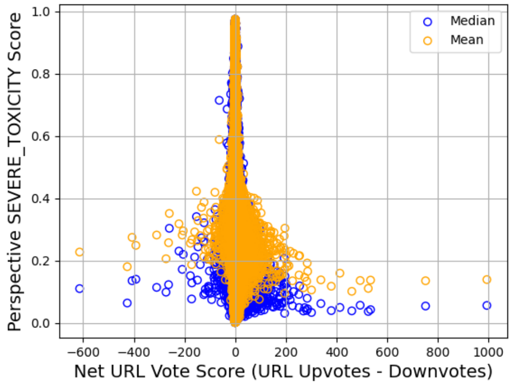


Figure 6: SEVERE_TOXICITY Score Compared to URL Net Dissenter Vote Score

Takeaways: NSFW and "offensive" content comprises a small fraction of the total Dissenter comment corpus; however, their Perspective scores indicate that this content is substantially more extreme than non-tagged content, and is not visible to most Dissenter users with their current settings. The labeling mechanism for "offensive" comments, while unknown, captures the most extreme content on Dissenter, which is substantially more toxic than the total Dissenter comments in aggregate.

Table 3: Overview of baseline toxicity datasets.

Dataset	# comments	# Dissenter users
NY Times	4,995,119	N/A
Daily Mail	14,287,096	N/A
Reddit	13,051,561	35,718

4.3.2 Are URLs with toxic comments up- or down-voted? Dissenter allows authenticated users to cast votes on the URL its members have commented upon by clicking a "thumbs up" or "thumbs down" button. We collect this data for 588k URLs in our crawl, and compare the net vote score (upvotes minus downvotes) to the Perspective SEVERE_TOXICITY scores for comments on these URLs. 104k URLs had a positive net vote score, 64k a negative score, and the majority (420k) had a net score of zero. 581k (99%) have a net vote score $n \in (-10, 10)$. Figure 6 plots the mean and median Perspective SEVERE_TOXICITY score for each URL with its net vote score. 415k URLs have no votes in either direction, contributing to the grouping around the zero net vote score. The zero vote score content exhibits the highest mean and median SEVERE_TOXICITY scores, evidenced by the tall peak of points around $x = 0$. As the net vote score absolute value increases, however, the SEVERE_TOXICITY scores corresponding to those URLs decrease. URLs with negative net vote scores in general having higher SEVERE_TOXICITY scores than their positive counterparts. One possible explanation for this is that Dissenter users down vote a given URL because they disagree with its content. Disagreeing with said content is also likely a trigger for toxic speech.

Takeaways: Significantly up- or down-voted content appears to generate lower comment mean and median toxicities, while comments with net vote scores near zero garner toxicity scores across the spectrum.

4.4 Relative Toxicity

Next, we consider whether Dissenter users are more or less toxic than other users and platforms.

4.4.1 Baseline Datasets. In addition to Dissenter comments, we construct three additional datasets: 1) NY Times, 2) Daily Mail, and 3) Reddit (summarized in Table 3). The NY Times and Daily Mail datasets are comments crawled from their respective sites, acquired from [43]. We chose these two news outlets for a few reasons. First, they are both relatively popular on Dissenter; Daily Mail is the 5th most commented on domain by Dissenter users (see Table 2) and NY Times is the 21st most popular. Next, they are on different sides of the political spectrum. Finally, the Perspective API has a set of models that are trained on NY Times comment moderator decisions, providing additional insight into how the model describes the specific environment in which it was trained.

The Reddit dataset includes comments by accounts on Reddit that we believe are likely to controlled by a corresponding Dissenter user. The construction of the Reddit dataset is a bit more complicated, but allows us to answer some questions related to how Dissenter users' behavior differs from their behavior on other, moderated social media platforms. We construct this dataset by querying Reddit for users matching our known Dissenter usernames. This revealed a large number of Reddit users: more than 56k Dissenter usernames (56%) correspond to a registered Reddit account. Of course, different people might choose the same username on different platforms, so we do not claim that all 56k of these accounts represent the same person on both platforms. While it is a near certainty that there are false positives in this construction, especially for particularly short usernames or usernames based on common words, previous work [29] established a lower bound precision of 0.6 for this type of matching and found it sufficient to describe behavioral trends when studying user migration from Reddit. With these caveats in mind, for each of the 56k identified Reddit accounts, we query Pushshift [16] for all of the comments they made on Reddit.

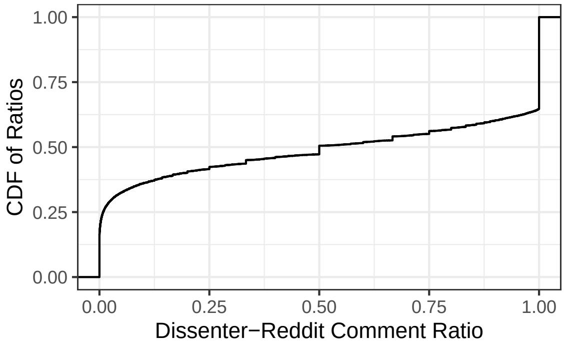
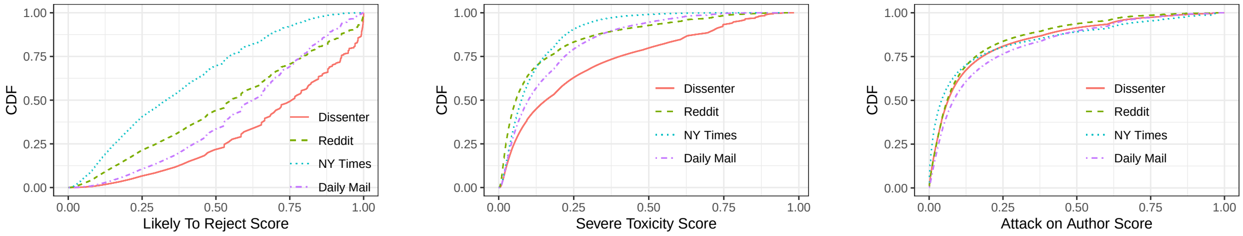


Figure 7: Ratio of Dissenter to Reddit Post Counts

Figure 7 plots the CDF of user "comment ratios," which is defined as $\frac{d}{d+r}$, where d is the number of posts a user has made on Dissenter, and r the analogous count on Reddit. We consider only users that have commented on *at least one* platform so that the ratio is well-defined; this limits the scope to 31k unique usernames. There is a roughly even split between which platform has been used more. The users that have more comments on Dissenter, however, tend to use that service exclusively, with more than a third having commented *only* on Dissenter.

Takeaways: A majority of usernames (~56%) exist on both Dissenter and Reddit. More than a third of users on both platforms post on Dissenter exclusively, as opposed to 20% that post only on Reddit.

4.4.2 Is the fear of censorship warranted? One of the motivating factors behind Dissenter's creation is the belief that moderators are stifling the open discussion of content on their platforms. The Perspective API provides a model that can help us ascertain whether or not this is the case: the LIKELY_TO_REJECT model. This model is trained on decisions from New York Times comment moderators and provides a score indicating whether or not they would reject a given comment on an article from being published. Figure 8a plots the CDF of scores from the LIKELY_TO_REJECT model for comments from Dissenter NY Times, Daily Mail, and Dissenter users' Reddit accounts. From the Figure, we see that over 75% of Dissenter comments receive a LIKELY_TO_REJECT score of 0.50 or more and 50% of comments receive a score above 0.75. While the figure makes the different norms on NY Time and Daily Mail quite obvious, Dissenter comments stand out as being significantly more likely to reject than comments from other platforms. We also note that the Dissenter users' Reddit comments follow a mostly uniform distribution that falls somewhere between the Daily Mail comments and the NY Times comments. Finally, although we do not



(a) LIKELY_TO_REJECT

(b) SEVERE_TOXICITY

(c) ATTACK_ON_AUTHOR

Figure 8: Perspective Model Scores for Dissenter and Related Datasets

show the results for clarity purposes, we found that when looking at just Dissenter comments on NY Times or Daily Mail URLs the LIKELY_TO_REJECT score distributions follow the same shape as the Dissenter curve in Figure 8a.

This result indicates that a substantial chunk of Dissenter comments are indeed considered unsuitable for publishing (at least by NY Times standards) and provides some degree of justification for Dissenter's motivation. It further indicates that this behavior might be associated with Dissenter itself, since Dissenter users' Reddit accounts fall somewhere between comments from NY Times and Daily Mail. Unfortunately, the LIKELY_TO_REJECT model provides no explanation of *why* a comment might have been rejected, but we can get a feeling by looking at the scores from other models.

Takeaways: The Perspective LIKELY_TO_REJECT model indicates that Dissenter comments are significantly more likely to be rejected by comment section moderators, lending support for Dissenter's niche in subverting moderation.

4.4.3 How toxic are Dissenter comments? The SEVERE_TOXICITY model offers another window into the type of comments posted by Dissenter's user. This Perspective model scores content by its ability to cause users to feel like they do not want to participate in further discussion, and is less sensitive to positive uses of profanity (e.g., "Damn, that's cool") than similar toxicity models offered by the Perspective API. A high SEVERE_TOXICITY score for a comment indicates a "very hateful, aggressive, or disrespectful comment" [21].

Like previous work ([20, 25, 42]),we compare the amount of toxic hate speech on Dissenter to similar platforms. Figure 8b plots the CDF of SEVERE_TOXICITY for our datasets. Dissenter comments score the highest in toxicity scores of the four data sources considered; approximately 20% of Dissenter comments have a SEVERE_TOXICITY score ≥ 0.5, about double the fraction of Reddit, the nearest dataset. Dissenter also has the thickest tail of the four datasets. Roughly 10% of Dissenter comments score 0.75 or above, indicating that many comments contain toxic speech.

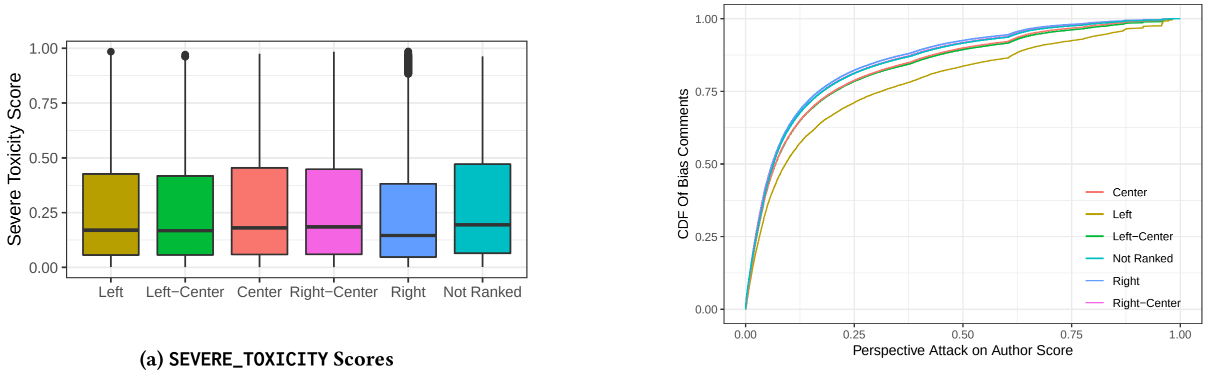
Takeaways: Dissenter comments exhibit substantially higher levels of toxicity than comments on the other platforms we study. SEVERE_TOXICITY measures *very* hateful speech, and is less sensitive to profanity than other toxicity models.

4.4.4 Are Dissenter comments attacking the message or the messenger? There is nothing inherently wrong with dissenting opinions. In fact, dissent is part of healthy debate and discussion. However, ad hominem attacks serve to stifle productive debate, and also have implications when discussing news in particular. Figure 8c plots the ATTACK_ON_AUTHOR Perspective scores for Dissenter users, as well as our baseline datasets. Surprisingly, Dissenter comments do not a display drastically different tendency to contain an attack on the URL's author. However, looking at the full distribution here does not reveal the the full picture.

To gain insight into the type of content that elicits toxic Dissenter user comments, we use the classification of the commented on URLs we discover according to the Allsides media bias rating organization [12]. Allsides uses multiple methodologies for classifying a media outlet's political bias, and categorizes popular news media organizations as "left", "center-left", "center", "center-right", or "right"-leaning. By design, Allsides categorizes the bias of mainstream media organizations and journalists only; therefore, many URLs that Dissenter users comment upon do not have an Allsides bias. For example, 437k unique comments appear on YouTube URLs; YouTube, as a video sharing service, does not have an Allsides bias ranking (intuitively, users can post either left- or right-leaning content on the platform.) Similarly, social media sites do not have an Allsides bias ranking either. Of 1.68M unique comments, approximately 1M fall on URLs with no Allsides ranking. The preponderance of these comments (∼45%) are on video sharing site URLs, primarily YouTube. Another 110k are on social media domains, like Twitter, Facebook, and Gab, and 155k more are on media outlets for which Allsides does not have a bias ranking.

Of the 600k comments on URLs that have an Allsides bias, we find that the underlying media bias has a slight, but significant impact (confirmed via two-sample Kolmogorov-Smirnov; all pairs $p < 0.01$) on the toxicity of the comment on that URL. Using the Perspective SEVERE_TOXICITY scores, we compare the scores according to each Allsides bias category in Figure 9a. From the Figure, we observe that toxicity tends to be higher for more center-leaning URLs; right-leaning URLs exhibit lower SEVERE_TOXICITY than all other bias types. On the other hand, the ATTACK_ON_AUTHOR scores in Figure 9b evince a higher likelihood that left-leaning content will generate comments that are an attack on the author of the article than the other Allsides bias rankings (again confirmed significantly different via two-sample KS test with $p < 0.01$).

Takeaways: Comment Perspective scores exhibit a slight, but statistically significant influence by the underlying media bias. Interestingly, while SEVERE_TOXICITY tends to be higher on more center URLs, ATTACK_ON_AUTHOR is higher on left-leaning URLS and decreases as the media bias moves rightward.



(a) `SEVERE_TOXICITY` Scores

(b) `ATTACK_ON_AUTHOR` Scores

Figure 9: Perspective Scores By Allsides URL Bias Classification

4.5 Social Network Analysis

4.5.1 Are there clusters of hateful users in the Dissenter followers graph? We construct the directed Dissenter social network graph using the data from §3.4, inclusive of 45,524 Dissenter users with at least one comment or reply. Both the in (followers) and out (following) degree distributions fit a power law distribution. The top three users by number of followers have 10,705, 9,588, and 8,183 followers, while the three users following the most other users follow 15,790, 10,646, and 10,625 others. Of note, none of the top ten highest degree users (in or out) are among the most prolific commenters on Dissenter overall. This indicates that Dissenter's user base, while technically a subset of Gab's user base, is not uniformly drawn from Gab's users. In other words, Dissenter seems to appeal to a smaller, more niche groups of users. While it might seem easy to dismiss these users due to their general lack of "popularity" on Gab proper, we believe this is dangerous. Small, extremely niche online communities have repeatedly been shown to harbor hateful and racist activity, have been actively used in disinformation campaigns, and have spawned numerous acts of violence.

Figure 10a shows the relationship between number of users following versus followed. Fully 15,702 of the users have no followers and follow no one. We conjecture that these are Gab users who tried Dissenter, but none of their Gab friends or followers are part of Dissenter. In general, however, the number of Dissenters each user follows is proportional to the number of followers.

Next, we examine the relationship between toxicity and the social network graph. Figure 10b shows the mean and median toxicity among Dissenter users for a given number of followers, while Figure 10c analyzes the relationship by number of followed users. While the toxicity is relatively low for users that are not well-connected in the graph, there are clear outliers with high toxicity and high degree. Of particular note, the mean is larger than the median for small degrees, but is then smaller than the median for higher degrees – indicating that the toxicity is skewed depending on the social network.

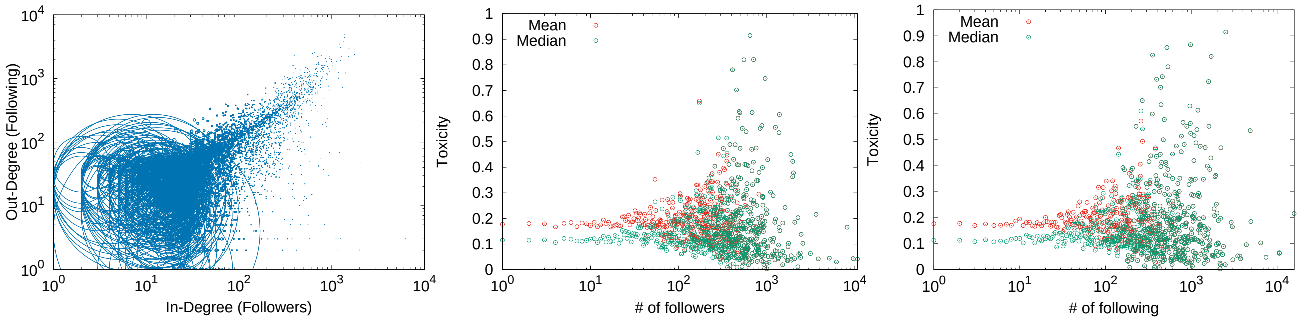
Given the macro-level toxicity properties of the social network, we sought to find the "hateful core," i.e. clusters of users with high toxicity that connect to other users also with high toxicity. To find this core, we induce a subgraph on our social network that includes users a and b iff: i) a and b are mutual followers; ii) a has posted $\geq$ 100 comments or replies; iii) a's median comment toxicity is $\geq$ 0.3. The restriction to at least 100 comments is to ensure that the user is active; many users have surprisingly high degree, but with few (or only one) toxic messages.

The resulting hateful core consists of only 42 users, with six connected components. There is one large connected component, with 32 interconnected users. While 18 of the usernames have an active account on Twitter, we find seven of the 32 whose Twitter accounts were suspended at some point. The users with active Twitter accounts largely exhibited self-professed radical beliefs in their Twitter profiles or posts. Thus, it appears that the most connected, hateful users in Dissenter represent both users than have been banned from Twitter, those who use both platforms, as well as users using Dissenter and not Twitter.

Takeaways: We examine the induced social network of Dissenter users on the Gab social network, and discover a small "hateful core" – a cluster of users that are active on Dissenter and routinely post highly toxic content. We find that a sizable number of these users with Twitter accounts have been suspended.

5 CASE STUDIES

The topics Dissenter users comment about are subject both to the prevailing issues of the day and to the interests and biases of the user base. This results in short-term fluctuations in comment themes intermixed with more stable, long-term topic trends. In this section, we highlight two major events that are reflected in the Dissenter data, one term that has elicited steady interest throughout Dissenter's existence, and a final topic that exhibits both short- and long-term community interest behaviors. In all four instances, we consider the terms' incidence in either a Dissenter comment or reply, or within the URL string itself. We include the URL as part of the analysis in order to capture comments about content that appears in the URL but do not reference the topic in the comment body. Figure 11 displays the incidence of each term as a percentage of the total daily comments from February 2019 to June 2020. This data incorporates an additional two months of Dissenter comments that we crawled in order to gain insight into the final topic, which occurred subsequent to manuscript submission.



(a) Following vs. Followers **(b) Toxicity vs. Followers** **(c) Toxicity vs. Following**

Figure 10: Dissenter Social Network Analysis

The first major news event reflected in Figure 11 is the arrest and subsequent death of Jeffery Epstein, an American financier charged with sex trafficking minor victims on July 8, 2019 [46]. While his arrest results in ~7% of Dissenter comments referencing "Epstein", his death a month later on August 10, 2019 is responsible for a spike of ~22% of all Dissenter comments.

Although the term "impeachment" occurs sporadically throughout Dissenter's first six months, it first crosses the threshold of 5% of daily comments in late September, 2019. This spike coincides with the release of a whistleblower complaint alleging the US President misused his power to solicit interference in the 2020 Presidential Election [15]. Following this initial jump, "impeachment" occurs at an elevated rate throughout the US House's impeachment inquiry, reaching a peak of 26% on December 19, 2019, the day after the House voted to impeach the President [34]. Another period of increased incidence of "impeachment" occurs from mid-January to early February 2020, coinciding with the impeachment trial and acquittal in the US Senate. Subsequent to this event, the term "impeachment" occurs in less than 1% of comments.

In contrast to the terms "Epstein" and "impeachment", which track closely with two major US news events, the term "Jew" appears with regularity throughout Dissenter's existence. Figure 11 shows that "Jew" appears in between 1% and 5% of URLs and comments consistently, breaking above 5% once in the final days of 2019, when the term briefly occurred in more than 10% of the daily totals. This spike followed a mass stabbing event during Hanukkah in which five people were wounded at a rabbi's home in New York [14]. Despite this singular small peak, the term "Jew" represents a long-term community interest, receiving a sustained non-trivial percentage of daily comments even in the absence of relevant current events. Previous work [45] noted that Gab, Dissenter's parent social network, exhibits a high degree of antisemitism.

Finally, Figure 11 shows the aggregate frequency with which the words "Black", "BLM", and a racial slur (not shown) for African-Americans occurs in daily comments. These terms exhibit the same behavior of "Jew" throughout most of Dissenter's timespan, albeit typically several percentage points more frequent, and with more spikes throughout the time period. However, following the death of George Floyd while in police custody in Minneapolis, Minnesota [24], which served as a catalyst for widespread protests throughout the US [18] and the world [17], the percentage of comments referencing these terms drastically increased from approximately 5% of all comments in March, 2020, to nearly 25% of daily comments in mid-June, 2020. This shows that while "Black" and associated terms are of long-term community interest, the frequency of these terms may additionally be propelled by current events, increasing their appearance in Dissenter comments.

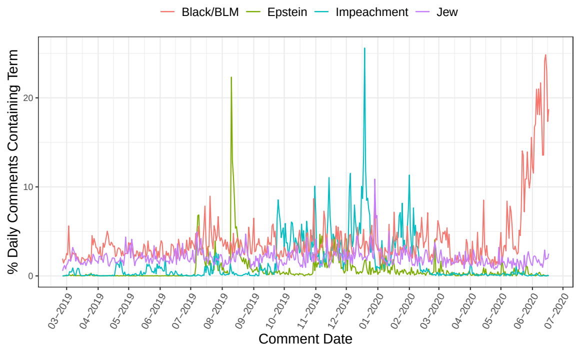


Figure 11: Incidence of Selected Terms in Comments and URLs Over Time

6 RELATED WORK

Hate Speech and Toxicity in User Comments. Zannettou *et al.*'s study "What is Gab" [42] is most closely related to this work, as it examines Dissenter's parent social network. The authors find that Gab attracts users from the "alt-right" and fringe conspiracy communities and that hate speech is prevalent on the site. As [42] occurred more than a year before Dissenter's launch, it does not study the Dissenter comment aggregation system. Lima *et al.* [27] also investigate hate speech on Gab, as well as URLs that are posted on Gab. As with Zannettou's work, [27] predates Dissenter's launch in 2019. Much prior work exists in detecting hate speech occurrences in online social networks. Hine *et al.* study `/pol/`, a community on the discussion-board website 4chan in [25]. Among many characterizations of the discussion board, they attempt to quantify the amount of hate on within the community, and also identify the URLs that users post in `/pol/` comments. However, unlike Dissenter, `/pol/` posts need not include or relate to a URL, and the toxicity of comments vis-à-vis the underlying URL is not studied. Djuric *et al.* study the prevalence of hate speech in Yahoo Finance user comments by using a neural language model trained on low-dimensional text embeddings of the comments [20].

Censorship. Dissenter exists to circumvent user-submitted comment moderation policies, or the inability to comment at all – in other words, *censorship* by Big Tech. Web censorship and censorship detection has a vast body of work; recently, Yadav *et al.* [41] studied web censorship mechanisms employed by Indian ISPs. Proxies are often employed to anonymize web traffic and defeat censorship efforts, and toward understanding the universe of free web proxies, Perino *et al.* develop a distributed active and passive measurement system to measure proxy performance [31], which they characterize in a longitudinal study [32]. In [33], Raman *et al.* discuss Mastodon, a decentralized web microblogging platform that has been forked by Gab in order to avoid being deplatformed by a single service provider. The pressures toward increasing centralization that Raman *et al.* identify could negate the benefit Gab gains from using a decentralized platform, e.g. increased availability, difficulty to censor, and resilience to outages. Because Dissenter users are a subset of Gab's, Dissenter is also impacted by this trend.

7 CONCLUSIONS AND FUTURE WORK

Dissenter is an approach to evading content provider censorship by decoupling the comment system from the underlying content; as opposed to network-level tools (e.g. Tor) that address censorship and privacy at the protocol level, Dissenter is concerned entirely with the application layer. Unlike more general social media (e.g. Twitter, Reddit, and even Gab) user activity on Dissenter is clearly bound to off-site content.

Like its parent social network, Gab, Dissenter claims to support its users' right to free speech; in practice, this manifests itself in toxic content. In fact, our study shows that Dissenter contains more hate-speech than prior work on Gab [42]. However repulsive the content on Dissenter, it clearly fills a need for its user base: Dissenter comments score higher on a machine learning model trained to classify comments as "likely to be rejected by a content moderator" than comments from any other data source we studied.

Several additional interesting security and privacy properties of Dissenter bear discussion. First, a small number of Dissenter comments on non-HTTP(S) scheme URLs are the result of a user using Dissenter to view content on their local file system. These comments leak information about the user's file system and the content they have downloaded. More curious are Dissenter comments on web browser start pages and tabs, e.g. "`chrome://startpage/`". Indeed, any URL is a potential anchor for a Dissenter comment thread, suggesting the possibility for a potential form of covert channel, a hidden conversation within a hidden conversation. The URL need not exist, can use any arbitrary scheme, and could be shared among users wishing to engage in a hidden conversation within the Dissenter platform. As we cannot easily differentiate between web URLs that are no longer responsive versus intentionally fictitious URLs, we leave this investigation for future research.

Second, a concerning aspect of Dissenter is the inability of a content owner to prevent discussion on their content within the Dissenter framework. It is not readily possible to block the Dissenter browser via traditional fingerprinting techniques as it is built on the Brave codebase and does not report a distinct user-agent string. Interestingly, a proactive defense may discourage or even break the current Dissenter model. A content producer could preemptively post comments within Dissenter for the content they own to overwhelm the conversation with positive comments. This has the tangible effect of content publishers being able to potentially affect the way that Dissenter discussions might go. Such proactive approaches are important to investigate further in an environment of active de-platforming.

Finally, the community should be particularly interested in what Dissenter represents moving forward. It is without question that computer scientists not only have a role to play in addressing online safety concerns, but also a *responsibility*. Further, future research in the measurements community is uniquely positioned to have meaningful impact in addressing societal problems of toxicity and hate speech. The work of policy makers and other experts depends on a data-driven understanding of these emergent platforms and networks. As the world faces unprecedented challenges, including increased levels of violence directly tied to the modern information age and fierce debate over censorship and de-platforming, it is incumbent on us to provide methodologies, experimentation, and results to understand what is *really* going on online.

8 ETHICAL CONSIDERATIONS

Considering Dissenter exists as a service designed to circumvent the moderation policies and discontinued comment sections of content providers, a further discussion of the ethical implications of our study is warranted.

By its nature, Dissenter users employ pseudonyms and use the platform to converse anonymously. However, all of the information that we collected in this study is publicly available and is in fact meant for consumption by normal Dissenter users. While we acknowledge that public availability is not a panacea for the potential misuse of data, we believe that the unmitigated risk to society that racist and hateful communities like Dissenter can pose outweigh the potential harm that could come from our study. Specifically, we believe that the type of understanding gained from the present study is not only sufficient but *necessary* to designing solutions that guard against the exploitation of the Internet to harm others.

Some users may purposefully or inadvertently post personally identifiable information (PII). While our analyses do not depend on any PII, we requested a determination from our Institutional Review Board (IRB) to ensure we were acting ethically. Our IRB found that the data we analyze is from publicly available Internet posts where there is no reasonable expectation of privacy, and that our research methods support beneficence and respect for persons. Finally, none of our work violates Dissenter's terms of use policies.

ACKNOWLEDGEMENTS

We thank Ben Zhao for shepherding and the anonymous reviewers for their insightful feedback. This work was partially supported by a Content Policy Research on Social Media Platforms award from Facebook Research. Views and conclusions are those of the authors and should not be interpreted as representing the official policies or position of the U.S. government.

REFERENCES

[1] 2009. Google Sidewiki. https://support.google.com/docs/forum/AAAABuH1jm0mVGy2ICxQUA.

[2] 2019. Comment Page for https://www.washingtontimes.com/news/2019/nov/26/exclusive-rivals-drudge-throne-gain-ground-amid-ag/. https://dissenter.com/discussion/begin?url=https://www.washingtontimes.com/news/2019/nov/26/exclusive-rivals-drudge-throne-gain-ground-amid-ag/.

[3] 2019. Exclusive: As Drudge Report falters amid anti-Trump shift, rivals on right gain ground. https://www.washingtontimes.com/news/2019/nov/26/exclusive-rivals-drudge-throne-gain-ground-amid-ag/.

[4] 2019. Superseding Indictment, United States of America v. Robert Bowers. https://www.justice.gov/usao-wdpa/press-release/file/1125346/download.

[5] 2019. The removal of the Dissenter extension. https://discourse.mozilla.org/t/the-removal-of-the-dissenter-extention/38140.

[6] 2020. Annotate the web, with anyone, anywhere. https://web.hypothes.is/.

[7] 2020. Brave Browser: Secure, Fast & Private Web Browser with Adblocker. https://brave.com.

[8] 2020. Gab Trends. https://trends.gab.com/.

[9] 2020. Github: gab-ai-inc/defiant-browser. https://github.com/gab-ai-inc/defiant-browser.

[10] 2020. Github: saffsd/langid.py. https://github.com/saffsd/langid.py.

[11] 2020. Hatebase. https://www.hatebase.org.

[12] 2020. Media bias ratings. http://www.allsides.com.

[13] 2020. Selenium Browser Automation. https://selenium.dev.

[14] Armstrong, Kevin and Mettler, Katie and Iati, Marisa and Knowles, Hannah. 2019. Knife-wielding man shattered night of celebration, witnesses say, renewing fears of violence against Jews. https://www.washingtonpost.com/nation/2019/12/29/monsey-ny-stabbing-attack/.

[15] Barrett, Devlin and Zapotosky, Matt and Dawsey, Josh and Harris, Shane. 2019. Whistleblower claimed that Trump abused his office and that White House officials tried to cover it up . https://www.washingtonpost.com/national-security/house-intelligence-committee-releases-whistleblowers-complaint-citing-trumps-call-with-ukraines-president/2019/09/26/402052ee-e056-11e9-be96-6adb81821e90_story.html.

[16] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *arXiv preprint arXiv:2001.08435* (2020).

[17] Booth, William and Morris, Loveday. 2020. Protests over death of George Floyd, police killings spread to London, Berlin, Toronto. https://www.washingtonpost.com/world/europe/george-floyd-police-protests-london-berlin-toronto/2020/05/31/cf4485e8-a357-11ea-898e-b21b9a83f792_story.html.

[18] Buchanan, Larry and Bui,Quoctrung and Patel,Jugal K. 2020. Black Lives Matter May Be the Largest Movement in U.S. History. https://www.nytimes.com/interactive/2020/07/03/us/george-floyd-protests-crowd-size.html.

[19] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665

[20] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. 29–30.

[21] Google. 2020. Perspective API. https://perspectiveapi.com.

[22] Google Jigsaw. 2019. Increasing Transparency in Perspective's Machine Learning Models. https://medium.com/the-false-positive/increasing-transparency-in-machine-learning-models-311ee08ca58a.

[23] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 1322–1328.

[24] Hill,Evan and Tiefenthäler, Ainara and Triebert,Christiaan and Jordan,Drew and Willis,Haley and Stein, Robin . 2020. How George Floyd Was Killed in Police Custody. https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html.

[25] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Eleventh International AAAI Conference on Web and Social Media*.

[26] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138* (2017).

[27] Lucas Lima, Julio CS Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 515–522.

[28] L. Lima, J. C. S. Reis, P. Melo, F. Murai, L. Araujo, P. Vikatos, and F. Benevenuto. 2018. Inside the Right-Leaning Echo Chambers: Characterizing Gab, an Unmoderated Social System. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 515–522.

[29] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Proceedigns of the Tenth International AAAI Conference on Web and Social Media (ICWSM '16)*. 10.

[30] John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. 2019. Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 571–576.

[31] Diego Perino, Matteo Varvello, and Claudio Soriente. 2018. ProxyTorrent: Untangling the free HTTP(S) proxy ecosystem. In *Proceedings of the 2018 World Wide Web Conference*. 197–206.

[32] Diego Perino, Matteo Varvello, and Claudio Soriente. 2019. Long-term Measurement and Analysis of the Free Proxy Ecosystem. *ACM Transactions on the Web (TWEB)* 13, 4 (2019), 1–22.

[33] Aravindh Raman, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2019. Challenges in the Decentralised Web: The Mastodon Case. In *Proceedings of the Internet Measurement Conference* (Amsterdam, Netherlands) *(IMC '19)*. Association for Computing Machinery, New York, NY, USA, 217–229. https://doi.org/10.1145/3355369.3355572

[34] Rucker, Phillip and Sonmez, Felicia and Itkowitz, Colby. 2019. Trump is impeached by the House, creating an indelible mark on his presidency. https://www.washingtonpost.com/politics/trump-is-impeached-by-the-house-creating-an-indelible-mark-on-his-presidency/2019/12/18/501bcab2-2105-11ea-a153-dce4b94e4249_story.html.

[35] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[36] Andrew Torba. 2016. Gab: A Social Network that Champions Free Speech, Individual Liberty and the Free Flow of Information. https://gab.com.

[37] Milo Trujillo, Maurício Gruppi, Cody Buntain, and Benjamin D Horne. 2020. What is BitChute? Characterizing the" Free Speech"Alternative to YouTube. *arXiv preprint arXiv:2004.01984* (2020).

[38] Donald Trump. 2020. Executive Order on Preventing Online Censorship. https://www.whitehouse.gov/presidential-actions/executive-order-preventing-online-censorship/.

[39] Twitter. 2020. Hateful conduct policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[40] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[41] Tarun Kumar Yadav, Akshat Sinha, Devashish Gosain, Piyush Kumar Sharma, and Sambuddho Chakravarty. 2018. Where The Light Gets In: Analyzing Web Censorship Mechanisms in India. In *Proceedings of the Internet Measurement Conference 2018* (Boston, MA, USA) *(IMC '18)*. Association for Computing Machinery, New York, NY, USA, 252–264. https://doi.org/10.1145/3278532.3278555

[42] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*. 1007–1014.

[43] Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites. In *WebSci*.

[44] Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM '20)*.

[45] Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 786–797.

[46] Zapotosky, Matt and Merle,Renae and Barrett,Devlin. 2019. Jeffrey Epstein charged with federal sex trafficking crimes involving young girls. https://www.washingtonpost.com/world/national-security/prosecutors-expected-to-unseal-charges-against-jeffrey-epstein/2019/07/08/3dec0fbe-a0db-11e9-b732-41a79c2551bf_story.html.