A Quantitative Approach to Understanding Online Antisemitism

Savvas Zannettou,[†] **Joel Finkelstein,**[⋆+] **Barry Bradlyn,**[◇] **Jeremy Blackburn**[‡]
[†]Max-Planck-Institut für Informatik, [⋆]Network Contagion Research Institute, [+]Princeton University
[◇]University of Illinois at Urbana-Champaign, [‡]Binghamton University
szannett@mpi-inf.mpg.de, joel@ncri.io, bbradlyn@illinois.edu, jblackbu@binghamton.edu

Abstract

A new wave of growing antisemitism, driven by fringe Web communities, is an increasingly worrying presence in the socio-political realm. The ubiquitous and global nature of the Web has provided tools used by these groups to spread their ideology to the rest of the Internet. Although the study of antisemitism and hate is not new, the scale and rate of change of online data has impacted the efficacy of traditional approaches to measure and understand these troubling trends.

In this paper, we present a large-scale, quantitative study of online antisemitism. We collect hundreds of million posts and images from alt-right Web communities like 4chan's Politically Incorrect board (/pol/) and Gab. Using scientifically grounded methods, we quantify the escalation and spread of antisemitic memes and rhetoric across the Web. We find the frequency of antisemitic content greatly increases (in some cases more than doubling) after major political events such as the 2016 US Presidential Election and the "Unite the Right" rally in Charlottesville. We extract semantic embeddings from our corpus of posts and demonstrate how automated techniques can discover and categorize the use of antisemitic terminology. We additionally examine the prevalence and spread of the antisemitic "Happy Merchant" meme, and in particular how these fringe communities influence its propagation to more mainstream communities like Twitter and Reddit. Taken together, our results provide a data-driven, quantitative framework for understanding online antisemitism. Our methods serve as a framework to augment current qualitative efforts by anti-hate groups, providing new insights into the growth and spread of hate online.

1 Introduction

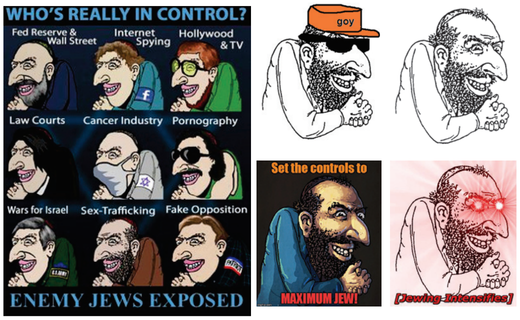
With the ubiquitous adoption of social media, online communities have played an increasingly important role in the real-world. The news media is filled with reports of the sudden rise in nationalistic politics coupled with racist ideology (Sunstein 2018) generally attributed to the loosely defined group known as the alt-right (SPLC 2017a), a movement that can be characterized by the relative youth of its adherents and relatively transparent racist ideology (ADL 2017). The alt-right differs from older groups primarily in its use of online communities to congregate, organize, and disseminate information in weaponized form (Marwick and Lewis 2017), often using humor and taking advantage of the scale and speed of communication the Web makes possible (Flores-Saviaga, Keegan, and Savage 2018; Hine et al. 2017; Zannettou et al. 2017; 2018a; 2018b; Morstatter et al. 2018). Recently, these fringe groups have begun to weaponize digital information on social media (Zannettou et al. 2017), in particular the use of weaponized humor in the form of memes (Zannettou et al. 2018b).

While the online activities of the alt-right are cause for concern, this behavior is not limited to the Web: there has been a recent spike in hate crimes in the United States (CSUSB 2018), a general proliferation of fascist and white power groups (SPLC 2017b), a substantial increase in white nationalist propaganda on college campuses (ADL 2018b). This worrying trend of real-world action mirroring online rhetoric indicates the need for a better understanding of online hate and its relationship to real-world events.

Antisemitism in particular is seemingly a core tenet of alt-right ideology, and has been shown to be strongly related to authoritarian tendencies not just in the US, but in many countries (Dunbar and Simonova 2003; Frindte, Wettig, and Wammetsberger 2005). Historical accounts concur with these findings: antisemitic attitudes tend to be used by authoritarian ideologies in general (Adorno et al. 1950; Arendt 1973). Due to its pervasiveness, historical role in the rise of ethnic and political authoritarianism, and the recent resurgence of hate crimes, understanding online antisemitism is of dire import. Although there are numerous anecdotal accounts, we lack a clear, large-scale, quantitative measurement and understanding of the scope of online semitism, and how it spreads between Web communities.

The study of antisemitism and hate, as well as methods to combat them are not new. Organizations like the Anti-Defamation League (ADL) and the Southern Poverty Law Center (SPLC) have spent decades attempting to address this societal issue. However, these organizations have traditionally taken a qualitative approach, using surveys and a small number of subject matter experts to manually examine content deemed hateful. While these techniques have produced valuable insights, qualitative approaches are extremely limited considering the ubiquity and scale of the Web.

In this paper, we take a different approach. We present an open, scientifically rigorous framework for quantitative analysis of online antisemitism. Our methodology is trans-



(a) (b)

Figure 1: Examples of the antisemitic Happy Merchant Meme.

parent and generalizable, and our data will be made available upon request. Using this approach, we characterize the rise of online antisemitism across several axes. More specifically we answer the following research questions:

1. **RQ1:** Has there been a rise in online antisemitism, and if so, what is the trend?
2. **RQ2:** How is online antisemitism expressed, and how can we automatically discover and categorize newly emerging antisemitic language?
3. **RQ3:** To what degree are fringe communities influencing the rest of the Web in terms of spreading antisemitic propaganda?

We shed light to these questions by analyzing a dataset of over 100M posts from two fringe Web communities: 4chan's Politically Incorrect board (/pol/) and Gab. We use word2vec (Mikolov et al. 2013) to train "continuous bag-of-words models" using the posts on these Web communities, in order to understand and discover new antisemitic terms. Our analysis reveals thematic communities of derogatory slang words, nationalistic slurs, and religious hatred toward Jews. Also, we analyze almost 7M images using an image processing pipeline proposed by (Zannettou et al. 2018b) to quantify the prevalence and diversity of the notoriously antisemitic Happy Merchant meme (Know Your Meme 2018c) (see Fig. 1). We find that the Happy Merchant enjoys substantial popularity in both communities, and its usage overlaps with other general purpose (i.e. not intrinsically antisemitic) memes. Finally, we use Hawkes Processes (Hawkes 1971) to model the relative influence of several fringe and mainstream communities with respect to dissemination of the Happy Merchant meme.

Disclaimer. Note that content posted on both Web communities can be offensive and racist. In the rest of the paper, we present our analysis without censoring any offensive language, hence we inform the reader that the rest of the paper contains language and images that are likely to be upsetting.

2 Related Work

Hate Speech on Web Communities. Several studies focus on understanding the degree of hate speech that exists in various Web communities. (Hine et al. 2017) focus on 4chan's Politically Incorrect board (/pol/); using the Hatebase database they find that 12% of the posts are hateful, hence highlighting /pol/'s high degree of hate speech. (Zannettou et al. 2018a) focus on Gab finding that Gab exhibits half the hate speech of /pol/, whereas when compared to Twitter it has twice the frequency of hateful posts. (Silva et al. 2016) focus on the targets of hate speech by performing a quantitative analysis on Twitter and Whisper, while (Mondal, Silva, and Benevenuto 2017) focus on understanding the prevalence of hate speech, the effects of anonymity, and the forms of hate speech in each community. (ElSherief et al. 2018b) perform a personality analysis on instigators and recipients of hate speech on Twitter, while (ElSherief et al. 2018a) perform a linguistic-driven analysis of hate speech.

Hate Speech Detection. Another line of work is the one that focus on the detection of hate speech on Web communities. A large corpus of previous work aim to detect hate speech using neural networks or traditional machine learning techniques on specific communities (Kwok and Wang 2013; Djuric et al. 2015; Gitari et al. 2015; Vigna et al. 2017; Serra et al. 2017; Founta et al. 2018; Davidson et al. 2017; Gao, Kuppersmith, and Huang 2017; Gao and Huang 2017). (Saleem et al. 2017) approach the problem through the lens of multiple Web communities by proposing a community-driven model for hate speech detection, while (Burnap and Williams 2016) focus on the various forms of hate speech by proposing a set of classification tools that assess hateful content with respect to race, sexuality, and disability. (Magu, Joshi, and Luo 2017) undertake a case study on Operation Google, a movement that aimed to use benign words in hateful contexts to trick Google's automated systems. They build a model that is able to detect posts that use benign words in hateful contexts and analyze the set of Twitter users that were involved. Finally, (Olteanu, Talamadupula, and Varshney 2017) propose the use of user-centered metrics (e.g., users' overall perception of classification quality) for the evaluation of hate speech detection systems.

Antisemitism. (Leets 2002) surveys 120 Jews or homosexual students to assess their perceived consequences of hate speech, to understand the motive behind hate messages, and if the recipients will respond or seek support after the hate attack. (Shainkman, Dencik, and Marosi 2016) use the outcomes of two surveys from the EU and ADL to assess how the level of antisemitism relates to the perception of antisemitism by the Jewish community in eight different EU countries. (Alietti, Padovan, and Lungo 2013) undertake phone surveys of 1.5K Italians on islamophobic and antisemitic attitudes finding that there is an overlap of ideology for both types of hate speech. (Ben-Moshe and Halafoff 2014) use focus groups to explore the impact of antisemitic behavior to Jewish children, concluding that there is a need for more education in matters related to racism, discrimination, and antisemitism. (Bilewicz et al. 2013) make two studies on antisemitism in Poland finding that Jewish conspiracy is the most popular and oldest antisemitic belief.

	/pol/			Gab		
Term	**#posts**	**Rank**	**Ratio**	**#posts**	**Rank**	**Ratio**
jew	1,993,432	13	1.64	763,329	19	16.44
kike	562.983	147	2.67	86,395	628	61.20
white	2,883,882	3	1.25	1,336,756	9	15.92
black	1,320,213	22	0.89	600,000	49	7.20
nigger	1,763,762	16	1.28	133,987	258	36.88
Total	67,416,903	-	0.95	35,528,320	-	8.14

Table 1: Number of posts for the terms "jew," "kike," "white," "black," and "nigger." We also report the rank of each term for each dataset (i.e., popularity in terms of count of appearance) and the ratio of increase between the start and the end of our datasets.

Word	**/pol/**	**Gab**
"jew"	42%	42%
"white"	33%	27%
"black"	43%	28%

Table 2: Percentage of hateful posts from random samples of 100 posts that include the words "jew," "white," and "black."

3 Datasets

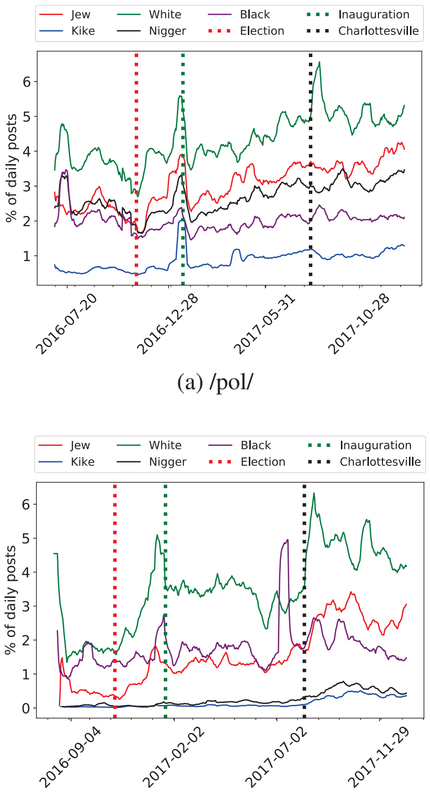
/pol/. 4chan is an anonymous image board that is usually exploited by troll users. A user can create a new thread by creating a post that contains an image. Other users can reply below with or without images and possibly add references to previous posts. The platform is separated to boards with varying topics of interest. In this work, we focus on the Politically Incorrect board (/pol/) as it exhibits a high degree of racism and hate speech (Hine et al. 2017) and it is an influential actor on the Web's information ecosystem (Zannettou et al. 2017). To obtain data from /pol/ posts we use the same crawling infrastructure as discussed in (Hine et al. 2017), while for the images we use the methodology discussed in (Zannettou et al. 2018b). Specifically, we obtain posts and images posted between July 2016 and January 2018, hence acquiring 67,416,903 posts and 5,859,439 images.

Gab. Gab is a newly created social network, founded in August 2016, that explicitly welcomes banned users from other communities (e.g., Twitter). It waves the flag of free speech and it has mild moderation; it allows everything except illegal pornography, posts that promote terrorist acts, and doxing other users. To obtain data from Gab, we use the same methodology as described in (Zannettou et al. 2018a) and (Zannettou et al. 2018b) for posts and images, respectively. Overall, we obtain 35,528,320 posts and 1,125,154 images posted between August 2016 and January 2018.

Ethical Considerations. During this work, we only collect publicly available data posted on /pol/ and Gab. We make no attempt to de-anonymize users and we follow best ethical practices as documented in (Rivers and Lewis 2014).

4 Results

In this section, we present our temporal analysis that shows the use of racial slurs over time on Gab and /pol/, our text-based analysis to understand the use of language with respect to ethnic slurs, and our memetic analysis that focuses on the propagation of the antisemitic Happy Merchant meme. Finally, we present our influence estimation findings that shed light on the influence that Web communities have on each other regarding the spread of antisemitic memes.

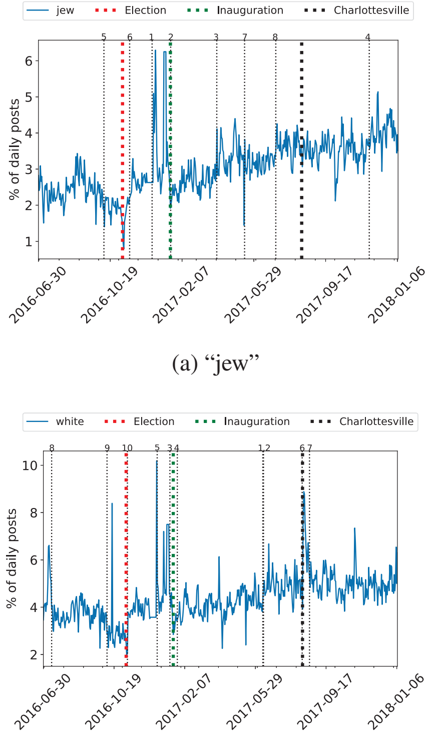


(a) /pol/

(b) Gab

Figure 2: Use of ethnic racial terms and slurs over time on /pol/ and Gab. The vertical lines show three indicative real-world events (not obtained via rigorous time series analysis).

Temporal Analysis. Anecdotal evidence reports escalating racial and ethnic hate propaganda on fringe Web communities (Thompson 2018). To examine this, we study the prevalence of some terms related to ethnic slurs on /pol/ and Gab, and how they evolve over time. We focus on five specific terms: "jew," "kike," "white," "black," and "nigger." We limit our scope to these because while they are notorious for ethnic hate for many groups, these specific words ranked among the the most frequently used ethnic terms on both communities. To extract posts for these terms, we first tokenize all the posts from /pol/ and Gab, and then extract all posts that contain either of these terms. Note that we use the entire dataset without any further filters (e.g., we do not filter posts in other languages). Table 1 reports the overall number of posts that contain these terms in both Web communities, their rank in terms of raw number of appearances in our dataset, as well as the increase in the use of these terms between the beginning and end of our datasets. For the latter,



(a) "jew"

(b) "white"

Figure 3: Percentage of daily posts per day for the terms "jew" and "white" on /pol/. We also report the detected changepoints (see Tables 3 and 4, respectively, for the meaning of each changepoint).

we note that although our computation of this ratio is in principle sensitive to large fluctuations at the ends of the dataset, Fig.2 do not display substantial fluctuations. Other methods, such as rolling averages, give comparable results. We study the effects of fluctuations systematically below. Also, Fig. 2 plots the use of these terms over time, binned by day, and averaged over a rolling window to smooth out small-scale fluctuations. We annotate the figure with three real-world events, which are of great interest and are likely to cause change in activity in these fringe communities (according to our domain expertise). Namely, we annotate the graph with the 2016 US election day, the Presidential Inauguration, and the Charlotesville Rally. We observe that terms like "white" and "jew" are extremely popular in both Web communities; 3rd and 13th respectively in /pol/, while in Gab they rank as the 9th and 19th most popular words, respectively. We see a similar level of popularity for ethnic racial slurs like "nigger" and "kike," especially on /pol/; they are the 16th and 147th most popular words.

We also find an increasing trend in the use of most ethnic terms; the number of posts containing each of the terms except "black" increases, even when normalized for the increasing number of posts on the network overall. Interestingly, among the terms we examine, we observe that the term "kike" shows the greatest increase in use for both /pol/ and Gab, followed by "jew" on /pol/ and "nigger" on Gab. Also, it is worth noting that ethnic terms on Gab have a greater increase in the rate of use when compared to /pol/ (cf. ratio of increase for /pol/ and Gab in Table 1). Furthermore, by looking at Fig. 2 we find that by the end of our datasets, the term "jew" appears in 4.0% of /pol/ daily posts and 3.1% of the Gab posts, while the term "nigger" appears in 3.4% and 0.6% of the daily posts on /pol/ and Gab, respectively. The latter is particularly worrisome for anti-black hate, as by the end of our datasets the term "nigger" on /pol/ overtakes the term "black" (3.4% vs 1.9% of all the daily posts). Taken together, these findings highlight that most of these terms are increasingly popular within these communities, hence emphasizing the need to study the use of ethnic identity terms.

To assess the extent that these terms are used in hateful/racist contexts we perform a small-scale manual annotation. Specifically, we collect 100 random posts from /pol/ and Gab for the words "jew," "white," and "black" and annotate them as hateful/racist or non-hateful/racist. For each of these posts, an author of the paper inspects the post and, according to the tone and terminology used, labels it as being hateful/racist or not. Note that we focus only on these three words, as the two other words (i.e., "kike" and "nigger") are highly offensive racial slurs, and therefore their use make the post immediately hateful/racist. Table 2 report the percentage of hateful/racist posts for the random samples of posts obtained from /pol/ and Gab. We observe that these words are used in a hateful/racist context frequently: in our random sample more than 25% of the posts that include one of the three words is hateful/racist. We also find the least hateful/racist percentage for the term "white" mainly because it is used in several terms like "White House" or "White Helmets", while the same applies for the term "black" (to a lesser extent) and the "Black Lives Matter" movement. Finally, we note a large hateful/racist percentage (42%) for posts containing the term "jew", highlighting once again the emerging problem of antisemitism on both /pol/ and Gab.

We note major fluctuations in the the use of ethnic terms over time, and one reasonable assumption is that these fluctuations happen due to real-world events. To analyze the validity of this assumption, we use changepoint analysis, which provides us with ranked changes in the mean and variance of time series behavior. To perform the changepoint analysis, we use the PELT algorithm as described in (Killick, Fearnhead, and Eckley 2012). We model each timeseries as a set of samples drawn from a normal distribution with mean and variance that are free to change at discrete times. We expect from the central limit theorem that for networks with large numbers of posts and actors, that this is a reasonable model. The algorithm then fits a robust timeseries model to the data by finding the configuration of changepoints which maximize the liklihood of the observed data, subject to a penalty for the proliferation of changepoints. The PELT algorithm thus returns the unique, exact best fit to the observed timeseries data. Subject to the assumptions mentioned above, we are thus confident that the changepoints represent a meaningful aspect of the data. We run the algorithm with a decreasing set of penalty amplitudes. We keep track of the largest penalty amplitude at which each

Rank	Date	Events
1	2016-12-25	2016-12-19: ISIS truck attack in Berlin Germany (Pleitgen et al. 2016).
2	2017-01-17	2017-01-17: Presidential inauguration of Donald Trump (Holland 2017). 2017-01-17: Netanyahu attacks the latest peace-conference by calling it "useless" (Buet, Mclaughlin, and Masters 2017).
3	2017-04-02	2017-04-05: Trump removes Bannon from his position on the National Security Council (Costa and Phillip 2016). 2017-04-06: Trump orders a strike on the Shayrat Air Base in Homs, Syria (Hennigan 2017).
4	2017-11-26	2017-11-29: It is revealed that Jared Kushner has been interviewed by Robert Mueller's team in November (Apuzzo 2017).
5	2016-10-08	2016-10-09: Second presidential debate (Politico 2016). 2016-10-09: A shooting takes place in Jerusalem that kills a police officer and two innocent people (BBC Press 2016).
6	2016-11-20	2016-11-19: Swastikas, Trump Graffiti appear in Beastie Boys' Adam Yauch Memorial Park in Brooklyn (Rielly 2016).
7	2017-05-16	2017-05-16: Donald Trump admits that he shared classified information with Russian envoys (Miller 2017). 2017-05-16: U.S. intelligence warns Israel to withhold information from Trump (Moore 2017).
8	2017-07-02	2017-06-25: The Supreme Court reinstates Trump's travel ban (Wolf and Gomez 2017). 2017-06-29: Trump's partial travel ban comes into effect (BBC 2017).

Table 3: Dates that significant changepoint were detected in posts that contain the term "jew" on /pol/. We sort them according to their "significance" and we report corresponding real-world events that happened one week before/after of the changepoint.

Rank	Date	Events
1 2	2017-06-10 2017-06-11	2017-06-08: Comey testifies about his conversations with Trump about investigations into Flynn (staff 2017). 2017-06-12: A court rejects Trump's appeal to stop the injunction against his travel ban (Levine and Hurley 2017). 2017-06-13: The US Senate interviews Jeff Sessions about Russian interference in the 2016 election (Savage 2017). 2017-06-15: Trump admits he is officially under investigation for obstruction of justice (Shear 2017).
3	2017-01-14	2017-01-17: Presidential inauguration of Donald Trump (Holland 2017).
4	2017-01-24	2017-01-23: Women's March protest (Przybyla and Schouten 2017). 2017-01-25: Trump issues executive order for construction of a wall on the Mexico border (Hirschfeld 2017).
5	2016-12-25	2016-12-19: ISIS truck attack in Berlin Germany (Pleitgen et al. 2016).
6	2017-08-12	2017-08-12: The "Unite the Right" rally takes place in Charlottesville, Virginia (Spencer and Scholberg 2017). 2017-08-13: Trump says there is blame for both sides about the Charlottesville rally (Lemire 2017).
7	2017-08-21	2017-08-17: Steve Bannon resigns as Chief Strategist for the White House (Diamond, Collins, and Landers 2017).
8	2016-07-13	2016-07-08: Fatal shooting of 5 police officers in Dallas by Micha Xavier Johnson (Ellis and Flores 2016). 2016-07-14: Truck attack in Nice, France (BBC 2016). 2016-07-16: The 2016 Republican National Convention (Collinson 2016).
9	2016-10-08	2016-10-09: Second presidential debate (Politico 2016).
10	2016-11-10	2016-11-08: Presidential election of Donald Trump (CNN 2016).

Table 4: Dates that significant changepoint were detected in posts that contain the term "white" on /pol/. We sort them according to their "significance" and we report corresponding real-world events that happened one week before/after of the changepoint.

changepoint first appears. This gives us a ranking of the changepoints in order of their "significance."

To identify real-world events that likely correspond to the detected changepoints, we manually inspect real-world events that are reported via the Wikipedia "Current Events" Portal[1] and happened one week before/after of the changepoint date. The portal provides real-world events that happen across the world for each day. To select the events, we use our domain expertise to identify the real-world events that are likely to be discussed by users on 4chan and Gab, hence they are the most likely events that caused the statistically significant change in the time series.

In /pol/, our analysis reveals several changepoints with temporal proximity to real-world political events for the use of both "jew" (see Fig. 3(a) and Table 3) and "white" (see Fig. 3(b) and Table 4). For usage in the term "jew," major world events in Israel and the Middle East correspond to several changepoints, including the U.S. missile attack against Syrian airbases in 2017, and terror attacks in Jerusalem. Events involving Donald Trump like the resignation of Steve Bannon from the National Security Council, the 2017 "travel ban", and the presidential inauguration occur within proximity to several notable changepoints for usage of "jew" as well. For "white," we find that changepoints correspond closely to events related to Donald Trump like the election, inauguration, and presidential debates. Additionally, several changepoints correspond to major terror attacks by ISIS in Europe, including vehicle attacks in Berlin and Nice, as well as news related to the 2017 "travel ban". In the case of "white," the relationship between online usage and real-world behavior is best illustrated by the Charlottesville rally, which marks the global maximum in our dataset for the use of the term on both /pol/ and Gab (see Fig. 2). For Gab, we find that changepoints in these time series reflect similar kinds of news events to those in /pol/, both for "jew" and "white" (we omit the Figures and Tables due to space constraints). These findings provide evidence that discussion of ethnic identity on fringe communities increases with political events and real-world extremist actions.

Text Analysis. We hypothesize that ethnic terms (e.g., "jew" and "white") are strongly linked to antisemitic and white supremacist sentiments. To test this, we use word2vec, a two-layer neural network that generate word representations as embedded vectors (Mikolov et al. 2013). Specifically, a word2vec model takes as an input a large corpus of text and generates a multi-dimensional vector space where each

[1] https://en.wikipedia.org/wiki/Portal:Current_events

/pol/				Gab			
Word	**Sim.**	**Word**	**Prob.**	**Word**	**Sim.**	**Word**	**Prob.**
(((jew)))	0.802	ashkenazi	0.269	jewish	0.807	jew	0.770
jewish	0.797	jew	0.196	kike	0.777	jewish	0.089
kike	0.776	jewish	0.143	gentil	0.776	gentil	0.044
zionist	0.723	outjew	0.077	goyim	0.756	shabbo	0.014
goyim	0.701	sephard	0.071	zionist	0.735	ashkenazi	0.013
gentil	0.696	gentil	0.026	juden	0.714	goyim	0.005
jewri	0.683	zionist	0.025	(((jew)))	0.695	kike	0.005
zionism	0.681	hasid	0.024	khazar	0.688	zionist	0.005
juden	0.665	talmud	0.010	jewri	0.681	rabbi	0.004
heeb	0.663	mizrahi	0.006	yid	0.679	talmud	0.003

Table 5: Top ten similar words to the term "jew" and their respective cosine similarity. We also report the top ten words generated by providing as a context term the word "jew" and their respective probabilities on /pol/ and Gab.

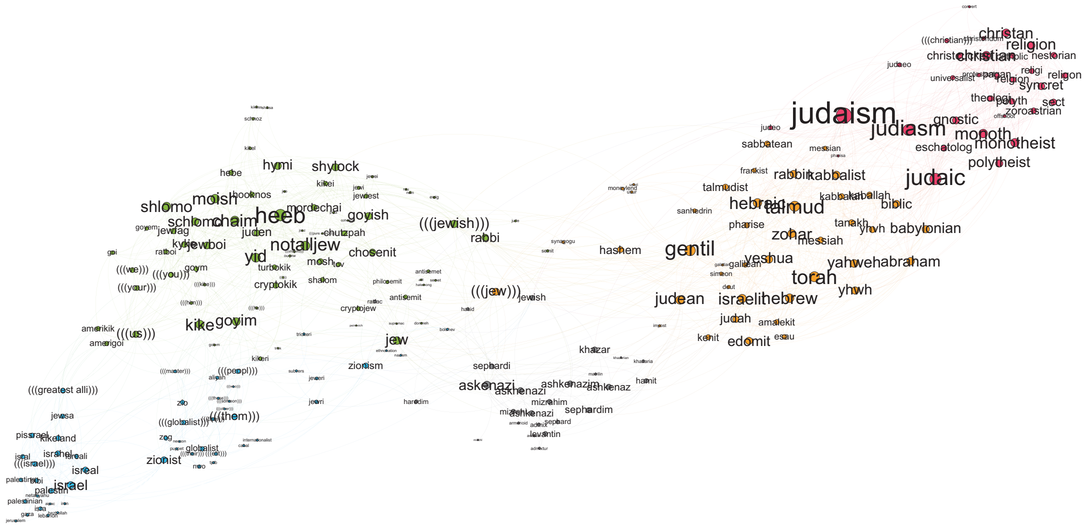


Figure 4: Graph representation of the words associated with "jew" on /pol/. Note that the figure is best viewed in color.

word is mapped to a vector in the space (also called an embedding). The vectors are generated in such way that words that share similar contexts tend to have nearly parallel vectors in the multi-dimensional vector space. Given a context (list of words appearing in a single block of text), a trained word2vec model also gives the probability that each other word will appear in that context. By analyzing both these probabilities and the word vectors themselves, we are able to map the usage of various terms in our corpus.

We train two word2vec models; one for the /pol/ dataset and one for the Gab dataset. First, as a pre-processing step, we remove stop words, punctuation, and we stem every word. Then, using the words of each post we train our word2vec models with a context window equal to 7 (defines the maximum distance between the current and the predicted words during the generation of the word vectors). We elect to slightly increase the context window from the default 5 to 7, since posts on /pol/ tend to be longer when compared to other platforms like Twitter. Also, we consider only words that appear at least 500 times in each corpus, hence creating a vocabulary of 31,337 and 20,115 stemmed words for /pol/ and Gab, respectively. Next, we use the generated word embeddings to gain a deeper understanding of the *context* in which certain terms are used. We measure the "closeness" of two terms (i and j) by generating their vectors from the word2vec models (h_i and h_j) and calculating their cosine similarity ($\cos\theta(h_1, h_2)$). Furthermore, we use the trained models to predict a set of candidate words that are likely to appear in the context of a given term.

We first look at the term "jew." Table 5 reports the top ten most similar words to the term "jew" along with their cosine similarity, as well as the top ten candidate words and their respective probability. By looking to the most similar words, we observe that on /pol/ "(((jew)))" is the most similar term ($\cos\theta = 0.80$), while on Gab is the 7th most similar term ($\cos\theta = 0.69$). The triple parentheses is a widely used, antisemitic symbol that calls attention to supposed secret Jewish involvement and conspiracy (Schama 2018). Slurs like "kike," which is historically associated with general ethnic disgust, rank similarly ($\cos\theta = 0.77$ on both /pol/ and Gab). This suggests that on both Web communities, the term "jew" itself is closely related to classical antisemitic contexts.

When looking at the set of candidate words, given the term "jew," we find the candidate word "ashkenazi" (most likely on /pol/ and 5th most likely on Gab), which refers to a specific subset of the Jewish community. Interestingly, we note that the term “jew” exists in the set of most likely words for both communities, hence indicating that /pol/ and Gab users abuse the term “jew” by posting messages that include the term “jew” multiple times in the same sentence.

To better show the connections between words similar to “jew,” Fig. 4 demonstrates the words associated with “jew” on /pol/ as a graph (we omit the same graph for Gab due to space constraints), where nodes are words obtained from the word2vec model, and the edges are weighted by the cosine similarities between the words (obtained from the trained word2vec models). The graph visualizes the two-hop ego network from the word “jew,” which includes all the nodes that are either directly connected or connected through an intermediate node to the “jew” node. We consider two nodes to be connected if their corresponding word vectors have a cosine similarity that is greater or equal to a pre-defined threshold. To select this threshold, we inspect the CDF of the cosine similarities between all the pair of words that exist in the trained word2vec models (we omit the figure due to space constraints). We elect to set this threshold to 0.6, which corresponds to keeping only 0.2% of all possible connections (cosine similarities). We argue that this threshold is reasonable as all the pairwise pairs of cosine similarities between the words is an extremely large number. To identify the structure and communities in our graph, we run the community detection heuristic presented in (Blondel et al. 2008), and we paint each community with a different color. Finally, the graph is layed out with the ForceAtlas2 algorithm (Jacomy et al. 2014), which takes into account the weight of the edges when laying out the nodes in the space.

This visualization reveals the existence of historically salient antisemitic terms, as well as newly invented slurs, as the most prominent associations to the word “jew.” Keeping in mind that proximity in the visualization implies contextual similarity, we note two close, but distinct communities of words which portray Jews as a morally corrupt ethnicity on the one hand (green nodes), and as powerful geopolitical conspirators on the other (blue). Notably the blue community connects canards of Jewish political power to anti-Israel and anti-Zionist slurs. The three, more distant communities document /pol/’s interest in three topics: The obscure details of ethnic Jewish identity (grey), Kabbalistic and cryptic Jewish lore (orange), and religious, or theological topics (pink).

We next examine the use of the term “white.” We hypothesize that this term is closely tied to ethnic nationalism. To provide insight for how “white” is used on /pol/ and Gab, we use the same analysis as described above for the term “jew.” Table 6 shows the top ten similar words to “white” and the top ten most likely words to appear in the context of “white.” When looking at the most similar terms, we note the existence of “huwhite” ($\cos\theta = 0.78$ on /pol/ and $\cos\theta = 0.70$ on Gab), a pronunciation of “white” popularized by the YouTube videos of white supremacist, Jared Taylor (Urban Dictionary 2017). “Huwhite” is a particularly interesting example of how the alt-right adopts certain language, even language that is seemingly derogatory towards themselves, in an effort to further their ideological goals. We also note the existence of other terms referring to ethnicity, such the terms “black” ($\cos\theta = 0.77$ on /pol/ and $\cos\theta = 0.71$ on Gab), “whiteeuropean” ($\cos\theta = 0.64$ on /pol/), and “caucasian” ($\cos\theta = 0.64$ on Gab). Interestingly, we again note the presence of the triple parenthesis “(((white)))” term on /pol/ ($\cos\theta = 0.75$), which refers to Jews who conspire to disguise themselves as white. When looking at the most likely candidate words, we find that on /pol/ the term “white” is linked with “supremacist,” “supremacy,” and other ethnic nationalism terms. The same applies on Gab with greater intensity as the word “supremacist” has a substantially larger probability when compared to /pol/.

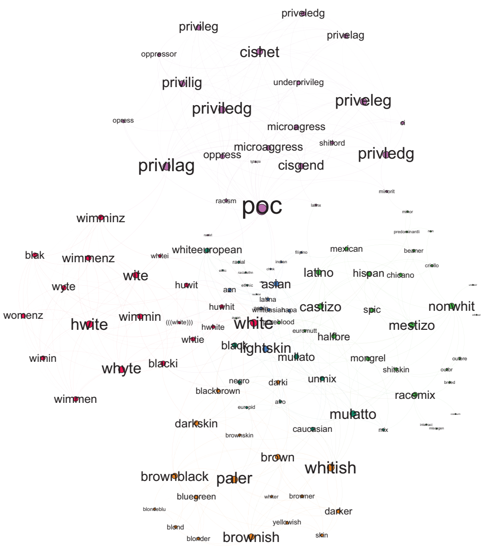


Figure 5: Words associated with “white” on /pol/.

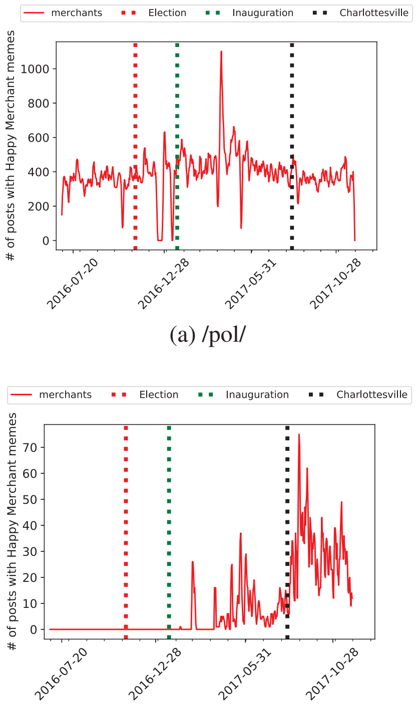
To provide more insight into the contexts and use of “white” on /pol/ we show its most similar terms and their nearest associations in Fig. 5 (using the same approach as for “jew” in Fig. 4, we omit the same graph for Gab due to space constraints). We find six different communities that evidence identity politics alongside themes of racial purity, miscegenation, and political correctness. These communities correspond to distinct ethnic and gender themes, like Hispanics (green), Blacks (orange), Asians (blue), and women (red). The final two communities relate to concerns about race-mixing (teal) and a prominent pink cluster that intriguingly, references terms related to left-wing political correctness, such as microagression and privilege (violet).

Note that we made the same analysis for the rest of the words that we study (i.e., “kike,” “nigger,” and “black”), however, we omit the figures due to space constraints.

Meme Analysis. In addition to hateful terms, memes also play a well documented role in the spread of propaganda and ethnic hate in Web communities (Zannettou et al. 2018b). To detail how memes spread and how different Web communities influence one another with memes, previous work (Zannettou et al. 2018b) established a pipeline that is able to track memes across multiple platforms. In a nutshell, the pipeline uses perceptual hashing (Monga and Evans 2006) and clustering techniques to track and analyze the propagation of memes across multiple Web communities. To achieve

/pol/				Gab			
Word	**Sim.**	**Word**	**Prob.**	**Word**	**Sim.**	**Word**	**Prob.**
huwhit	0.789	supremacist	0.494	black	0.713	supremacist	0.827
black	0.771	supremaci	0.452	huwhit	0.703	supremaci	0.147
(((white)))	0.754	supremist	0.008	nonwhit	0.684	genocid	0.009
nonwhit	0.747	male	0.003	poc	0.669	helmet	0.004
huwit	0.655	race	0.002	caucasian	0.641	nationalist	0.003
hwite	0.655	supremecist	0.002	whitepeopl	0.625	hous	0.003
whiteeuropean	0.644	nationalist	0.002	dispossess	0.624	privileg	< 0.001
hispan	0.631	genocid	0.002	indigen	0.602	male	< 0.001
asian	0.628	non	0.001	negroid	0.599	knight	< 0.001
brownblack	0.627	guilt	0.001	racial	0.595	non	< 0.001

Table 6: Top ten similar words to the term "white" and their respective cosine similarity. We also report the top ten words generated by providing as a context term the word "white" and their respective probabilities on /pol/ and Gab.



(a) /pol/

(b) Gab

Figure 6: Number of posts with images of Happy Merchant meme on /pol/ and Gab. The vertical lines show indicative real-world events (not obtained via changepoint analysis).

this, it relies on images obtained from the Know Your Meme (KYM) site (Know Your Meme 2018d), which is a comprehensive encyclopedia of memes.

In this work, we use this pipeline to study how antisemitic memes spread within and between these Web communities, and examine which communities are the most influential in their spread. To do this, we additionally examine two mainstream Web communities, Twitter and Reddit, and compare their influence with /pol/ and Gab. For Twitter and Reddit, we use the dataset from (Zannettou et al. 2018b), which includes all the posts from Reddit and Twitter, between July 2016 and July 2017, that include an image that is a meme as dictated by the KYM dataset and their processing pipeline. The final dataset consists of 581K tweets and 717K Reddit posts that include a meme. In this work, we focus on the Happy Merchant meme (see Fig. 1) (Know Your Meme 2018c), which is an important hate-meme to study in this regard for several reasons. First, it represents an unambiguous instance of antisemitic hate, and second, it is extremely popular and diverse in /pol/ and Gab (Zannettou et al. 2018b).

We aim to assess the popularity and increase of use over time of the Happy Merchant meme on /pol/ and Gab. Fig. 6 shows the number of posts that contain images with the Happy Merchant meme for every day of our /pol/ and Gab dataset. We further note that the numbers here represent a *lower bound* on the number of Happy Merchant postings: the image processing pipeline is conservative and only labels clusters that are unambiguously the Happy Merchant; variations of other memes that incorporate the Happy Merchant are harder to assess. We observe that /pol/ consistently shares antisemitic memes over time with a peak in activity on April 7, 2017, around the time that the USA launched a missile strike in a Syrian base (Wikipedia 2017). By manually examining a few posts including the Happy Merchant meme on this specific date, we find that 4chan users use this meme to express their belief that the Jews are "behind this attack." On Gab we note a substantial and sudden increase in posts containing Happy Merchant memes immediately after the Charlottesville rally. Our findings on Gab dramatically illustrate the implication that real-world eruptions of antisemitic behavior can catalyze the acceptability and popularity of antisemitic memes on other Web communities.Taken together, these findings highlight that both communities are exploited by users to disseminate racist content that is targeted towards the Jewish community.

Another important step in examining the Happy Merchant meme is to explore how clusters of similar Happy Merchant memes relate to other meme clusters in our dataset. One possibility is that Happy Merchants make-up a unique family of memes, which would suggest that they segregate in form and shape from other memes. Given that many memes evolve from one another, a second possibility is that Happy Merchants "infect" other common memes. This could serve, for

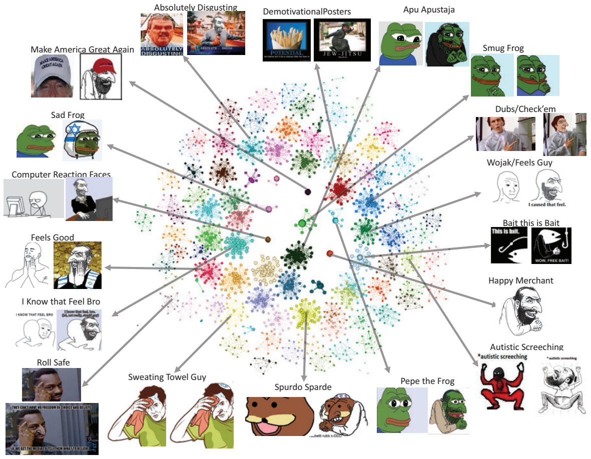
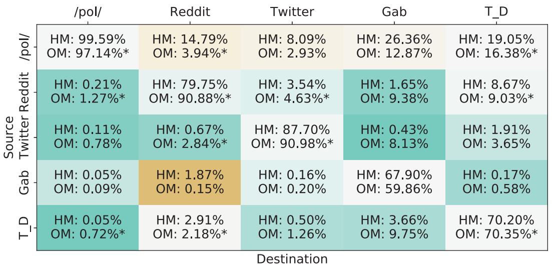
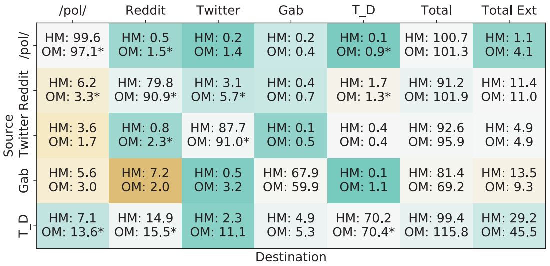


Figure 7: Visualization of a subset of the obtained image clusters with a focus on the penetration of the Happy Merchant meme to other popular memes. The figure is inspired from (Zannettou et al. 2018b).



	/pol/	Reddit	Twitter	Gab	T_D
/pol/	HM: 99.59% OM: 97.14%*	HM: 14.79% OM: 3.94%*	HM: 8.09% OM: 2.93%	HM: 26.36% OM: 12.87%	HM: 19.05% OM: 16.38%*
Reddit	HM: 0.21% OM: 1.27%*	HM: 79.75% OM: 90.88%*	HM: 3.54% OM: 4.63%*	HM: 1.65% OM: 9.38%	HM: 8.67% OM: 9.03%*
Twitter	HM: 0.11% OM: 0.78%	HM: 0.67% OM: 2.84%*	HM: 87.70% OM: 90.98%*	HM: 0.43% OM: 8.13%	HM: 1.91% OM: 3.65%
Gab	HM: 0.05% OM: 0.09%	HM: 1.87% OM: 0.15%	HM: 0.16% OM: 0.20%	HM: 67.90% OM: 59.86%	HM: 0.17% OM: 0.58%
T_D	HM: 0.05% OM: 0.72%*	HM: 2.91% OM: 2.18%*	HM: 0.50% OM: 1.26%	HM: 3.66% OM: 9.75%	HM: 70.20% OM: 70.35%*

Figure 8: Percent of the destination community's Happy Merchant (HM) and non-Happy-Merchant (OM) memes caused by the source community. Colors indicate the percent difference between Happy Merchants and non-Happy-Merchants, while $*$ indicate statistical significance between the distributions with $p < 0.01$.



	/pol/	Reddit	Twitter	Gab	T_D	Total	Total Ext
/pol/	HM: 99.6 OM: 97.1*	HM: 0.5 OM: 1.5*	HM: 0.2 OM: 1.4	HM: 0.2 OM: 0.4	HM: 0.1 OM: 0.9*	HM: 100.7 OM: 101.3	HM: 1.1 OM: 4.1
Reddit	HM: 6.2 OM: 3.3*	HM: 79.8 OM: 90.9*	HM: 3.1 OM: 5.7*	HM: 0.4 OM: 0.7	HM: 1.7 OM: 1.3*	HM: 91.2 OM: 101.9	HM: 11.4 OM: 11.0
Twitter	HM: 3.6 OM: 1.7	HM: 0.8 OM: 2.3*	HM: 87.7 OM: 91.0*	HM: 0.1 OM: 0.5	HM: 0.4 OM: 0.4	HM: 92.6 OM: 95.9	HM: 4.9 OM: 4.9
Gab	HM: 5.6 OM: 3.0	HM: 7.2 OM: 2.0	HM: 0.5 OM: 3.2	HM: 67.9 OM: 59.9	HM: 0.1 OM: 1.1	HM: 81.4 OM: 69.2	HM: 13.5 OM: 9.3
T_D	HM: 7.1 OM: 13.6*	HM: 14.9 OM: 15.5*	HM: 2.3 OM: 11.1	HM: 4.9 OM: 5.3	HM: 70.2 OM: 70.4*	HM: 99.4 OM: 115.8	HM: 29.2 OM: 45.5

Figure 9: Influence from source to destination community of Happy Merchant and non-Happy-Merchant memes, normalized by the number of events in the *source* community, while $*$ indicate statistical significance between the distributions with $p < 0.01$.

instance, to make antisemitism more accessible and common. To this end, we visualize in Fig. 7 a subset of the meme clusters and a Happy Merchant version of each meme. This visualization is inspired from (Zannettou et al. 2018b) and it demonstrates numerous instances of the Happy Merchant infecting well-known and popular memes. Some examples include Pepe the Frog (Know Your Meme 2018e), Roll Safe (Know Your Meme 2018f), Bait this is Bait (Know Your Meme 2018a), and the Feels Good meme (Know Your Meme 2018b). This suggests that users generate antisemitic variants on recognizable and popular memes.

Influence Estimation. While the growth and diversity of the Happy Merchant within fringe Web communities is a cause of significant concern, a critical question remains: How do we chart the influence of Web communities on one another in spreading the Happy Merchant? We have, until this point, examined the expanse of antisemitism on individual, fringe Web communities. Memes however, develop with the purpose to replicate and spread between different Web communities. To examine the influence of meme spread between Web communities, we employ Hawkes processes (Linderman and Adams 2014; 2015), which can be exploited to measure the predicted, reciprocal influence that various Web communities have to each other. Generally, a Hawkes model consists of K processes, where a process is a sequence of events that happen with a particular probability distribution. Colloquially, a process is analogous to a specific Web community where memes (i.e., events) are posted. Each process has a rate of events, which defines expected frequency of events on a specific Web community (for example, five posts with Happy Merchant memes per hour). An event on one process can cause *impulses* on other processes, which increase their rates for a period of time. An impulse is defined by a weight and a probability distribution. The former dictates the intensity of the impulse (i.e., how strong is the increase in the rate of a process), while the latter dictates how the effect of the impulse changes over time. For instance, a weight of 1.5 from process A to B, means that each event on A will cause, on average, an additional 1.5 events on B.

In this work, we use a separate Hawkes model for each cluster of images that we obtained when applying the pipeline reported in (Zannettou et al. 2018b). Each model consists of five processes; one for each of /pol/, The_Donald, the rest of Reddit, Gab, and Twitter. We elected to separate The_Donald from the rest of Reddit, as it is an influential actor with respect to the dissemination of memes (Zannettou et al. 2018b). Next, we fit each model using Gibbs sampling as reported in (Linderman and Adams 2014; 2015). This technique enable us to obtain, at a given time, the weights and probability distributions for each impulse that is active, hence allowing us to be confident that an event is caused because of a previous event on the same or on another process. Table 7 shows the number of events (i.e., appearance of a meme) for each community we study, for both the Happy Merchant meme and all the other memes.

First, we report the percentage of events expected to be attributable from a source community to a destination community in Fig. 8. In other words, this shows the percentage of memes posted on one community which, in the context of our model, are expected to occur in direct response to posts

	/pol/	Reddit	Twitter	Gab	T_D	Total Events	# of clusters
Happy Merchant Meme	43,419	1,443	1,269	376	282	46,789	133
Other Memes	1,530,821	581,244	717,752	44,542	81,665	2,956,024	12,391

Table 7: Events per Web community for the Happy Merchant and all the other memes.

in the source community. We can thus interpret this percentage in terms of the relative influence of meme postings one network on another. We also report influence in terms of efficacy by normalizing the influence that each source community has, relative to the total number of memes they post (Fig. 9). We compare the influence that Web communities exert on one another for the Happy Merchant memes (HM) and all other memes (OM) in the graph. To assess the statistical significance of the results, we perform two-sample Kolmogorov-Smirnov tests that compare the distributions of influence from the Happy Merchant and other memes; an asterisk within a cell denotes that the distributions of influence between the source and destination platform have statistically significant differences ($p < 0.01$).

Our results show that /pol/ is the single most influential community for the spread of memes to all other Web communities. Interestingly, the influence that /pol/ exhibits in the spread of the Happy Merchant surpasses its influence in the spread of other memes. However, although /pol/'s overall influence is higher on these networks, its per-meme efficacy for the spread of antisemitic memes tended to be lower relative to non-antisemitic memes with the intriguing exception of The_Donald. Another interesting feature we observe about this trend is that memes on /pol/ itself show little influence from other Web communities; both in terms of memes generally, and non-antisemitic memes in particular. This suggests a unidirectional meme flow and influence from /pol/ and furthermore, suggest that /pol/ acts as a primary reservoir to incubate and transmit antisemitism to downstream Web communities.

Main Take-Aways.

1. Racial and ethnic slurs are increasing in popularity on fringe Web communities. This trend is particularly notable for antisemitic language.
2. Our word2vec models in conjunction with graph visualization techniques, demonstrate an explosion in diversity of coded language for racial slurs used in /pol/ and Gab. Our methods demonstrate a means to dissect this language and decode racial discourse on fringe communities.
3. The use of ethnic and antisemitic terms on Web communities is substantially influenced by real-world events. For instance, our analysis shows a substantial increase in the use of ethnic slurs including the term "jew" around Donald Trump's Inauguration, while the same applies for the term "white" and the Charlottesville rally.
4. When it comes to the use of antisemitic memes, we find that /pol/ consistently shares the Happy Merchant Meme, while for Gab we observe an increase in the use in 2017, especially after the Charlottesville rally. Finally, our influence estimation analysis reveals that /pol/ is the most influential actor in the overall spread of the Happy Merchant to other communities, possibly due to the large volume of Happy merchant memes that are shared within the platform. The_Donald however, is the most efficient in pushing Happy Merchant memes to other Web communities.

5 Discussion

Antisemitism has been a historical harbinger of ethnic strife (ADL 2018a). While organizations have been tackling antisemitism and its associated societal issues for decades, the rise and ubiquitous nature of the Web has raised new concerns. Antisemitism and hate have grown and proliferated rapidly online, and have done so mostly unchecked. This is due, in large part, to the scale and speed of the online world, and calls for new techniques to better understand and combat this worrying behavior.

In this paper, we take the first step towards establishing a large-scale quantitative understanding of antisemitism online. We analyze over 100M posts from July, 2016 to January, 2018 from two of the largest fringe communities on the Web: 4chan's Politically Incorrect board (/pol/) and Gab. We find evidence of increasing antisemitism and the use of racially charged language, in large part correlating with real-world political events like the 2016 US Presidential Election. We then analyze the *context* this language is used in via word2vec, and discover several distinct facets of antisemitic language, ranging from slurs to conspiracy theories grounded in biblical literature. Finally, we examine the prevalence and propagation of the antisemitic "Happy Merchant" meme, finding that 4chan's /pol/ and Reddit's The_Donald are the most influential and efficient, respectively, in spreading this antisemitic meme across the Web.

Naturally our work has some limitations. First, most of our results should be considered a *lower bound* on the use of antisemitic language and imagery. In particular, we note that our quantification of the use of the "Happy Merchant" meme is extremely conservative. The meme processing pipeline we use is tuned in such a way that many Happy Merchant variants are clustered along with their "parent" meme. Second, our quantification of the growth antisemitic language is focused on two particular keywords, although we also show how new rhetoric is discoverable. Third, we focus primarily on two specific fringe communities. As a new community, Gab in particular is still rapidly evolving, and so treating it as a stable community (e.g., Hawkes processes), may cause us to underestimate its influence.

Regardless, there are several important recommendations we can draw from our results. First, organizations such as the ADL and SPLC should refocus their efforts towards open, data-driven methods. Small-scale, qualitative understanding is still incredibly important, especially with regard to understanding offline behavior. However, resources *must* be devoted to large-scale data analysis. Second, we believe that–regardless of the participation of anti-hate organizations–scientists, and particularly computer scientists, must expend effort at understanding, measuring, and combating online antisemitism and online hate in general. The Web has changed the world in ways that were unimaginable even ten years ago. The world has shrunk, and the Information Age is in full effect. Unfortunately, many of the innovations that make the world what it is today were created with little thought to their negative consequences. For a long time, technology innovators have not considered potential negative impacts of the services they create, in some ways abdicating their responsibility to society. The present work provides solid quantified evidence that the technology that has had incredibly positive results for society is being co-opted by actors that have harnessed it in worrying ways, using the same concepts of scale, speed, and network effects to greatly expand their influence and effects on the rest of the Web and the world at large.

Acknowledgments. SZ received funding from the EU's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie ENCASE project (GA No. 691025). We also acknowledge the support of the NVIDIA Corporation, for the donation of the two Titan Xp GPUs.

References

ADL. 2017. Alt Right: A Primer about the New White Supremacy. https://bit.ly/2WLo9j2.

ADL. 2018a. Anti-Semitism. https://bit.ly/3aKXUfQ.

ADL. 2018b. White Supremacist Propaganda Surges on Campus. https://bit.ly/2VGQaqW.

Adorno, T. W.; Frenkel-Brunswik, E.; Levinson, D. J.; Sanford, R. N.; et al. 1950. The authoritarian personality.

Alietti, A.; Padovan, D.; and Lungo, L. E. 2013. Religious Racism. Islamophobia and Antisemitism in Italian Society.

Apuzzo, M. 2017. Mueller's Prosecutors Are Said to Have Interviewed Jared Kushner on Russia Meeting. https://nyti.ms/2k83owr.

Arendt, H. 1973. *The origins of totalitarianism*, volume 244. Houghton Mifflin Harcourt.

BBC Press. 2016. Jerusalem shooting: Two killed by Palestinian gunman. https://bbc.in/2VHoh20.

BBC. 2016. Nice attack: At least 84 killed by lorry at Bastille Day celebrations. https://bbc.in/2y7yzPN.

BBC. 2017. Trump travel ban comes into effect for six countries. https://bbc.in/3f0VIUC.

Ben-Moshe, D., and Halafoff, A. 2014. Antisemitism and Jewish Children and Youth in Australia's Capital Territory Schools.

Bilewicz, M.; Winiewski, M.; Kofta, M.; and Wójcik, A. 2013. Harmful Ideas, The Structure and Consequences of Anti-S emitic Beliefs in Poland. *Political Psychology* 34(6):821–839.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008(10):P10008.

Buet, P.-E.; Mclaughlin, E.; and Masters, J. 2017. Netanyahu: Paris peace conference is 'useless'. http://cnn.it/2jmLCU3.

Burnap, P., and Williams, M. L. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*.

CNN. 2016. presidential results. https://cnn.it/3cSzF0G.

Collinson, S. 2016. Donald Trump accepts presidential nomination. http://cnn.it/2akveQU.

Costa, R., and Phillip, A. 2016. Stephen Bannon removed from National Security Council. https://wapo.st/2oDddTL.

CSUSB. 2018. Report to the Nation: Hate Crime Rise in U.S. Cities and U.S. Counties in Time of Division and Foreign Interference. https://bit.ly/2xeeF5h.

Davidson, T. J.; Warmsley, D.; Macy, M. W.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.

Diamond, J.; Collins, K.; and Landers, E. 2017. Trump's chief strategist Steve Bannon fired. http://cnn.it/2icjT9v.

Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; and Bhamidipati, N. 2015. Hate Speech Detection with Comment Embeddings. In *WWW*.

Dunbar, E., and Simonova, L. 2003. Individual difference and social status predictors of anti-Semitism and racism US and Czech findings with the prejudice/tolerance and right wing authoritarianism scales. *Int. J. Intercult. Relat.* 27(5):507–523.

Ellis, R., and Flores, R. 2016. Multiple officers killed at Dallas protest over police killings. http://cnn.it/29soAXE.

ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W. Y.; and Belding-Royer, E. M. 2018a. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. In *ICWSM*.

ElSherief, M.; Nilizadeh, S.; Nguyen, D.; Vigna, G.; and Belding-Royer, E. M. 2018b. Peer to Peer Hate: Hate Speech Instigators and Their Targets. In *ICWSM*.

Flores-Saviaga, C.; Keegan, B. C.; and Savage, S. 2018. Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community. In *ICWSM*.

Founta, A.-M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2018. A Unified Deep Learning Architecture for Abuse Detection. *arXiv preprint arXiv:1802.00385*.

Frindte, W.; Wettig, S.; and Wammetsberger, D. 2005. Old and new anti-Semitic attitudes in the context of authoritarianism and social dominance orientation. *Peace and Conflict*.

Gao, L., and Huang, R. 2017. Detecting Online Hate Speech Using Context Aware Models. In *RANLP*.

Gao, L.; Kuppersmith, A.; and Huang, R. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In *IJCNLP*.

Gitari, N. D.; Zuping, Z.; Damien, H.; and Long, J. 2015. A Lexicon-based Approach for Hate Speech Detection. *IJMUE*.

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

Hennigan, W. 2017. Trump Orders Strikes on Syria Over Chemical Weapons. http://time.com/5240164/.

Hine, G. E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*.

Hirschfeld, J. 2017. Trump Orders Mexican Border Wall to Be Built and Plans to Block Syrian Refugees. https://nyti.ms/2ktUvwa.

Holland, S. 2017. Trump, now president, pledges to put 'America First' in nationalist speech. http://reut.rs/2iQMMmK.

Jacomy, M.; Venturini, T.; Heymann, S.; and Bastian, M. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*.

Killick, R.; Fearnhead, P.; and Eckley, I. A. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* 107(500):1590–1598.

Know Your Meme. 2018a. Bait / This is Bait Meme. https://knowyourmeme.com/memes/bait-this-is-bait.

Know Your Meme. 2018b. Feels Good Meme. https://knowyourmeme.com/memes/feels-good.

Know Your Meme. 2018c. Happy Merchant Meme. http://knowyourmeme.com/memes/happy-merchant.

Know Your Meme. 2018d. Know Your Meme Site. http://knowyourmeme.com/.

Know Your Meme. 2018e. Pepe the Frog Meme. http://knowyourmeme.com/memes/pepe-the-frog.

Know Your Meme. 2018f. Roll Safe Meme. http://knowyourmeme.com/memes/roll-safe.

Kwok, I., and Wang, Y. 2013. Locate the Hate: Detecting Tweets against Blacks. In *AAAI*.

Leets, L. 2002. Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *J. of social issues*.

Lemire, J. 2017. Trump blames 'many sides' after violent white supremacist rally in Virginia. https://bit.ly/2wGzG8p.

Levine, D., and Hurley, L. 2017. Another U.S. appeals court refuses to revive Trump travel ban. http://reut.rs/2rSDEFM.

Linderman, S. W., and Adams, R. P. 2014. Discovering Latent Network Structure in Point Process Data. In *ICML*.

Linderman, S. W., and Adams, R. P. 2015. Scalable Bayesian Inference for Excitatory Point Process Networks. *ArXiv 1507.03228*.

Magu, R.; Joshi, K.; and Luo, J. 2017. Detecting the Hate Code on Social Media. In *ICWSM*.

Marwick, A., and Lewis, R. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, Z. 2017. National Security Advisor: Trump's Conversation With Russians Was 'Wholly Appropriate'. http://ti.me/2pR6XVQ.

Mondal, M.; Silva, L. A.; and Benevenuto, F. 2017. A Measurement Study of Hate Speech in Social Media. In *HT*.

Monga, V., and Evans, B. L. 2006. Perceptual image hashing via feature points: performance evaluation and tradeoffs. *IEEE Transactions on Image Processing*.

Moore, J. 2017. US officials warned Israel not to share sensitive material with Trump. https://bit.ly/2wsBtOx.

Morstatter, F.; Shao, Y.; Galstyan, A.; and Karunasekera, S. 2018. From Alt-Right to Alt-Rechts: Twitter Analysis of the 2017 German Federal Election. In *WWW Companion*.

Olteanu, A.; Talamadupula, K.; and Varshney, K. R. 2017. The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection. In *WebSci*.

Pleitgen, F.; Dewan, A.; Griffiths, J.; and Schoichet, C. 2016. Berlin attack: ISIS claims it inspired truck assault at market. http://cnn.it/2gWjnXf.

Politico. 2016. Full transcript: Second 2016 presidential debate. https://politi.co/35gQ7Fl.

Przybyla, H., and Schouten, F. 2017. At 2.6 million strong, Women's Marches crush expectations. http://usat.ly/2jJCzfY.

Rielly, K. 2016. NYC's Adam Yauch Park Vandalized With Swastikas. Until Kids Replaced Them With Hearts. http://ti.me/2fbBHe4.

Rivers, C. M., and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research* 3.

Saleem, H. M.; Dillon, K. P.; Benesch, S.; and Ruths, D. 2017. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *CoRR* abs/1709.10159.

Savage, C. 2017. Highlights from attorney general jeff sessions's senate testimony. https://nyti.ms/2siLKIG.

Schama, S. 2018. (((SEMITISM))) Being Jewish in America in the Age of Trump.

Serra, J.; Leontiadis, I.; Spathis, D.; Stringhini, G.; Blackburn, J.; and Vakali, A. 2017. Class-based Prediction Errors to Detect Hate Speech with Out-of-vocabulary Words.

Shainkman, M.; Dencik, L.; and Marosi, K. 2016. Different Antisemitisms: on Three Distinct Forms of Antisemitism in Contemporary Europe – with a Special Focus on Sweden.

Shear, M. 2017. Trump Attacks Rosenstein in Latest Rebuke of Justice Department. https://nyti.ms/2tuOzUb.

Silva, L. A.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the Targets of Hate in Online Social Media. In *ICWSM*.

Spencer, H., and Scholberg, C. 2017. White Nationalists March on University of Virginia. https://nyti.ms/2vr58UU.

SPLC. 2017a. ALT-RIGHT. https://bit.ly/2KGCgi9.

SPLC. 2017b. The Year in Hate and Extremism. https://bit.ly/2YeLcUa.

staff, P. 2017. James Comey statement to Senate intelligence committee on Trump contact. https://politi.co/2JgCrjt.

Sunstein, C. R. 2018. *Republic: Divided democracy in the age of social media*. Princeton University Press.

Thompson, A. 2018. The Measure of Hate on 4Chan. https://bit.ly/2Uzp1V4.

Urban Dictionary. 2017. Huwhite. https://www.urbandictionary.com/define.php?term=Huwhite.

Vigna, F. D.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *ITASEC*.

Wikipedia. 2017. 2017 shayrat missile strike. https://bit.ly/3bWzxwU.

Wolf, R., and Gomez, A. 2017. Supreme Court reinstates Trump's travel ban, but only for some immigrants. https://usat.ly/2u8w4p5.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *IMC*.

Zannettou, S.; Bradlyn, B.; De Cristofaro, E.; Kwak, H.; Sirivianos, M.; Stringini, G.; and Blackburn, J. 2018a. What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber. In *WWW Companion*.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018b. On the Origins of Memes by Means of Fringe Web Communities. In *IMC*.