arXiv:1901.09735v3 [cs.CY] 17 May 2019

"And We Will Fight For Our Race!" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan

Alexandros Mittos[⋆], Savvas Zannettou[†], Jeremy Blackburn[‡], Emiliano De Cristofaro[⋆]

[⋆]University College London, [‡]Cyprus University of Technology, [†]University of Alabama at Birmingham

a.mittos@ucl.ac.uk, sa.zannettou@edu.cut.ac.cy, blackburn@uab.edu, e.decristofaro@ucl.ac.uk

Abstract

Recent progress in genomics has enabled an emerging market for "direct-to-consumer" genetic testing. Nowadays, companies like 23andMe and AncestryDNA provide affordable health, genealogy, and ancestry reports, and have already tested tens of millions of customers. At the same time, far-right groups have also reportedly taken an interest in genetic testing, using them to attack minorities and prove their genetic "purity." In this paper, we present a quantitative measurement study shedding light on how genetic testing is being discussed on Web communities in Reddit and 4chan. We collect 1.3M comments from both platforms, posted over 27 months, using a set of 280 keywords related to genetic testing. We then use NLP and computer vision tools to identify trends, themes, and topics of discussion. Our analysis shows that genetic testing attracts a lot of attention on Reddit and 4chan, with discussions often including highly toxic language expressed through hateful, racist, and misogynistic comments. In particular, on 4chan's politically incorrect board (/pol/), content from genetic testing conversations involves several alt-right personalities and openly antisemitic rhetoric, often conveyed through memes. Finally, we find that discussions build around user groups, from technology enthusiasts to communities promoting fringe political views.

1 Introduction

Over the past decade, researchers have made tremendous progress toward understanding the human genome, i.e., the complete set of an individual's DNA, which encodes all of the information needed to build and maintain that organism. With increasingly low costs, millions of people can afford to learn about their genetic make-up, not only in diagnostic settings, but also to satisfy their curiosity about traits, wellness, or discover their ancestry and genealogy.

A number of companies have successfully marketed *direct-to-consumer (DTC)* genetic tests: individuals purchase a kit (typically around $100), mail it back with a saliva sample, and receive online reports after a few days. DTC companies offer a wide range of services, from romantic match-making [28] or identification of athletic skills [70] to reports of health risks (e.g, likelihood of developing Parkinson's), wellness (e.g., lactose intolerance), carrier status (e.g., hereditary hearing loss), traits (e.g., eyes color).

Popular products also include genetic *ancestry* tests, which promise a way to discover one's ancestral roots, building on patterns of genetic variations common in people from similar backgrounds [53]. However, these are subject to limitations, e.g., results differ from provider to provider due to different control groups [67]. AncestryDNA alone has tested more than 10M customers as of January 2019 [2].

Alas, increased interest in self-administered genetic tests, and in particular ancestry, has also been accompanied by media reports of far-right groups using it to attack minorities or prove their genetic "purity" [11, 61]. This prompts concerns of a new wave of scientific racism [63]. For example, white nationalists chug milk at gatherings to demonstrate the ability of people of white color to digest lactose better than people of other ethnicities [40]. Also, statements from Donald Trump led Senator Warren to publicly confirm her Native American ancestry via genetic testing [44].

All of this comes at a time when racism, hate, and antisemitism on platforms like 4chan, Gab, and certain communities on Reddit is on the rise [30, 41]. Thus, these trends are particularly worrying, also considering how technology has been disrupting society in previously unconsidered ways [37]; the fact that racist, misogynistic, and dangerous behavior festers and spreads on the Web at an unprecedented scale, eventually making its way into the real world, prompts the need for a thorough understanding of how these genetic testing tools are being used and misusedused in online discussions. As genetics-based arguments for discrimination [4], and even genocide (e.g., the Holocaust), have been made in the past, *without* hard, statistical data to back them up, the potential abuse of genetic testing results this should not be overlooked.

While other aspects of genetic testing have been studied (e.g., how they affect one's perception of racial identity [57, 66]), we are interested in the relation between genetic testing and online hate. This is a topic that has not been thoroughly studied by the scientific community, despite, as discussed earlier, increasingly worrisome indications of far-right groups exploiting genetic testing for racist rhetoric.

With this motivation in mind, we identify and address the following research questions: (1) What is the overall prevalence of genetic testing discourse on social networks like Reddit and 4chan? (2) In what context do users discuss genetic testing? (3) Is genetic testing associated with far-right views, racist ideologies, hate speech, and white supremacy? (4) If yes, in what context? Can we identify specific themes?

We compile and use a set of 280 keywords related to genetic testing to extract all available posts and comments from Reddit and 4chan. We collect 7K threads from the politically incorrect (/pol/) board of 4chan (consisting of 1.3M posts) from June 30, 2016 to March 13, 2018, and 77K comments from Reddit related to genetic testing from January 1, 2016 to March 31, 2018, and analyze them along several axes to understand how genetic testing is being discussed online. Then, we rely on natural language processing and machine learning tools, including (i) Latent Dirichlet Allocation (LDA) [9] to identify topics of discussion, (ii) word embeddings [46] to uncover words used in a similar context across datasets, (iii) Google's Perspective API [58] to measure toxicity in texts, and (iv) Perceptual Hashing to assess the imagery and memes shared in posts.

Overall, the *main* findings of our study include:

1. Genetic testing is often discussed on /pol/ and on subreddits associated with hateful, racist, and sexist content. These communities discuss genetic testing in a highly toxic manner, often suggesting its use to marginalize or even *eliminate* minorities.
2. The analysis of images posted on /pol/ highlights the recurrent presence of alt-right personalities and antisemitic memes along with genetic testing discussions.
3. Word embeddings analysis reveals that certain subreddits use ethnic terms in conjunction with genetic testing keywords in the same way as /pol/, which may be an indicator of 4chan's fringe ideologies spilling out on more mainstream Web communities.
4. Genetic testing on Reddit is being discussed in a variety of contexts, e.g., g breeds, debating the validity of evidence in real crimes crime evidence, and issues related to children (e.g., adoption, pregnancy), among others. This indicates how mainstream genetic testing has recently become.
5. Reddit users are not uniformly interested in all aspects of genetic testing, rather, they form groups ranging from enthusiasts (i.e., people who are interested in or have undergone genetic testing), to people who use genetic keywords exclusively in subreddits that discuss fringe political views.

2 Related Work

Genetic Testing & Society. Panofsky and Donovan [57] analyze 70 discussion threads on the far-right website Stormfront.org, where at least one user posted ancestry test results. They group posters based on whether they consider their results good and bad, and study how other Stormfront users react: if the posters receive "bad news," they tend to question the validity of genetic genealogy science, trying to reinterpret their results to fit their views on races. Mittos et al. [48] conduct an exploratory study of the Twitter discourse on genetic testing, examining 300K tweets related to genetic testing, and find that those who are interested in genetic testing appear to be tech-savvy and interested in digital health in general. They also find sporadic instances of users using genetic testing in a racist context, and others who express privacy concerns. There is also a non-negligible amount of users who share screenshots of their ancestry test results without considering the possible privacy implications.

Chow-White et al. [18] examine 2K tweets containing the keyword '23andMe' spanning one week. They calculate their sentiment and find out that the positive tweets outnumber the negative, while users appear overall enthusiastic about the company's services. Roth and Ivemark [66] study how ancestry testing affects ethnic and racial identities by conducting 100 interviews with people who have white, black, Hispanic/Latino, Native American, and Asian ancestry. They find instances of consumers not accepting test results, and instead focus on estimates based on social appraisals and aspirations. Overall, they suggest genetic ancestry testing may reinforce race privilege. Clayton et al. [21] conduct a meta-analysis of 53 studies involving 47K people around perceptions of genetic privacy, highlighting how survey questions are often phrased poorly, thus leading to possible misinterpretations of the results. They also show that not enough attention was paid to influential factors, e.g., participants' sociocultural backgrounds. Finally, Couldry and Yu [22] discuss how DTC genetic companies, such as 23andMe, influence the public toward sharing their genetic data by claiming that the abundance of data will improve people's lives in the long term, despite a body of work showing that genetic data cannot be securely anonymized.

Overall, research in this area mostly relies on qualitative studies examining the societal effects of genetic testing [14, 19, 23, 39, 52] and lacks quantitative large-scale measurements. To the best of our knowledge ours is the first large-scale, quantitative measurement study, using Reddit and 4chan. We examine trends, themes, and topics of discussion around genetic testing, and explore how communities related to the alt-right exploit genetic testing for sinister purposes.

Online Hate. Researchers have also studied hate speech on mainstream social networks like Twitter [24, 55, 64, 68], Reddit [16, 55], Facebook [6, 26], YouTube [56], and Instagram [42]. Closer to our work is research on fringe communities in 4chan and Reddit [7, 30, 41, 77, 78]. Specifically, Bernstein et al. [7] study 5M posts on the random (/b/) board to examine how anonymity and ephemerality work in 4chan. They find that most threads expire in less than 5 minutes, while over 90% of the posts are anonymous. Then, Hine et al. [41] focus on /pol/, by studying 8M posts collected over two and a half months. Their content analysis reveals that, while most URLs point to YouTube, a non-negligible amount link to right-wing websites. They also find evidence of organized "raids" against YouTube users, where users collectively post hateful comments on videos they disapprove. Zannettou et al. [78] explore how mainstream and fringe online communities on Twitter, Reddit, and 4chan influence each other with respect to disinformation and hateful propaganda. Among other things, they find that racist memes are very common in /pol/ and Gab, and that /pol/ and the /r/The_Donald subreddit are the most influential Web communities with respect to the dissemination of memes. Finkelstein et al. [30] study antisemitism on /pol/ and Gab, revealing that antisemitic content increases in those networks after major political events, such as the "Unite the Right" rally or the 2016 US elections. Also, they leverage word embeddings to identify terminology associated with antisemitic content. Overall, accurately identifying hateful content remains an open problem due to being largely context-dependent; Chandrasekharan et al. [15] study how Reddit's decision to ban several subreddits that violated anti-harassment policy affected hate speech on the platform. They examine 100M posts and comments from two banned subreddits, namely r/fatpeoplehate and r/CoonTown, and measure the generated hate speech by its users before and after the ban. They find that the ban had a positive effect on the platform as the users who continued posting drastically reduced their hate speech usage. while prior work identifies and/or measures hate on fringe platforms, we examine whether genetic testing, a seemingly harmless topic, is being discussed in a toxic manner.

Exploratory Studies on Reddit. Another line of work has, similar to ours, performed quantitative studies using Reddit data. De Choudhury and De [25] look Reddit conversations about mental health, aiming to understand language attributes of online self-disclosure and factors driving support in online posts. They show that Reddit fills a gap between online health forums and social networks like Facebook and Twitter, as users explicitly share personal information on their mental health, and use Reddit for self-expression, even for seeking diagnosis or treatment information. Other studies analyze how users behave in specific subreddits. Kasunic and Kaufman [43] focus on a specific subreddit called /r/RoastMe, where users post photos of themselves and invite others to ridicule them. They find that the RoastMe community relies on a specific set of norms, such as highly valuing caustic comments but also being concerned about the potential psychological harm of the participants. Nobles et al. [54] study /r/STD to understand how users seek health information on sensitive and stigmatized topics, using 1.8K posts from 1.5K users. They find that most posts crowd-source information about non-reportable STDs, focusing on treatment, symptoms, as well as aspects of social and emotional impact.

Another line of work has studied the /r/The_Donald subreddit. Flores-Saviaga et al. [31] analyze 16M comments spanning two years to examine the characteristics of political troll communities. They find that /r/The_Donald subscribers spend energy educating their community on certain events and that they use various socio-technical tools to mobilize other subscribers. Finally, Mills [47] compares /r/The_Donald to /r/SandersForPresident, a subreddit broadly supporting the 2016 presidential candidate Bernie Sanders, exploring whether rapidly formed subreddits exhibit collective intelligence. Mills finds that these communities are very effective on pursuing their agendas and that Trump supporters more often tend to clash with other communities and Reddit administrators.

3 Datasets

In this section, we present the methodology used to obtain the datasets used in our study.

Genetic Testing Keywords. To extract relevant comments and posts we compile a list of 280 keywords related to genetic testing. First, we use the list of 268 DTC companies offering DNA tests over the Internet between 2011 and 2018 (e.g., 23andme, AncenstryDNA, Orig3n) obtained from [59]. We then add 12 more keywords: ancestry testing/test, genetic testing/test, genomic testing/test, genomics, genealogy testing/test, dna testing/test, and GEDMatch (an open data personal genomics database and genealogy website [36]).

Reddit	**Genetic Testing**	**Random**	***4chan***	**Genetic Testing**	**Random**
Comments	77,184	204,713	Threads	6,986	19,530
Subreddits	3,734	12,616	Posts	1,306,671	760,691
Users	48,096	165,127	Posts/T (Mean)	186.5	37.9
			Posts/T (Median)	183	5
			Images	338,540	206,830

Table 1: Overview of the Reddit and 4chan datasets.

Reddit Dataset. Reddit is a social news aggregation and discussion website, where users post content (e.g., images, text, links) which gets voted up or down by other users. Users can also add comments to the posts, and comments can also be voted up or down and receive replies. Top submissions appear on the front page, and top comments at the top of the post. Content on Reddit is organized in communities created by users, *"subreddits,"* which are usually associated with areas of interest (e.g., movies, sports, politics). As of January 2019, Reddit has more than 330M monthly active users and 14B visits, which makes it the fifth most visited site in the US [60].

We gather all Reddit comments from January 1, 2016 to March 31, 2018 (2B comments in 473K subreddits) via the monthly releases from pushshift.io.[1] We then use the 280 genetic testing keywords as search terms to extract all comments possibly related to genetic testing. This results in a dataset of 77K comments posted in 4.6K subreddits, as summarized in Table 1. For comparison, we also obtain a set of 204K random comments unrelated to genetic testing.

4chan. 4chan is an imageboard website with virtually no moderation. An "Original Poster" (OP) creates a thread by posting an image and a message. Content is organized in subcommunities, called boards (as of January 2019, there are 70 of them), with various topics of interest (e.g., anime, sports, adult, politics, etc.). Others can post in the OP's thread, with a message or an image. On 4chan, users do not need a registered account to post content. We focus on a the politically incorrect board (/pol/), which has been shown to include a high volume of racist, xenophobic, and hateful content [41]. We choose /pol/ as we study how genetic testing is being discussed in communities that have been associated with alt-right ideologies.

We collect 1.9M threads posted on /pol/ from June 30, 2016 to March 13, 2018. Once again, we use the 280 keywords as search terms on each thread: if we find a keyword anywhere in it, we get the *whole thread*. This is slightly different from what we do for Reddit. On 4chan, each discussion is structured as a single-threaded entity where the OP submits an image on which other users respond. There is no official method of responding to a certain comment other than the original one, whereas, on Reddit a user may reply to a specific comment creating a

[1] https://files.pushshift.io/reddit/

new branch of answers. Also, 4chan threads do not contain titles, thus, it is difficult to understand the context of a discussion without reading the whole thread. In the end, we extract 6.9K threads containing 1.3M posts. For comparison, we also get a random sample of 19K threads, with 760K posts. The 4chan dataset is summarized in Table 1, where we report the mean and median number of posts per thread, and the total number of images. Note that, while the threads with genetic testing keywords have 338,540 images, later on we study only images shared in the *posts* containing those keywords (6,375).

Remarks. We look at Reddit and 4chan's politically incorrect board (/pol/) as opposed to mainstream platforms (i.e., Facebook, Twitter) for several reasons. First, Facebook's structure and API do not permit the uniform collection of all available comments containing specific keywords, thus we would have to cherry-pick specific groups possibly leading to a biased dataset, while Twitter's API only allows the collection of 1% of all tweets which would greatly limit the dataset's amplitude. Meanwhile, Reddit's self-organizing content (i.e., subreddits) permits us to easily extract the context in which genetic testing is being discussed, while *all* comments ever posted in the platform are available online (see Section 3). Moreover, we are interested in the hateful and racist connotations of genetic testing discourse.

Ethics. Our study was approved by the ethics committee at UCL. Also, as the content posted in 4chan is anonymous, we make no attempt to de-anonymize users. Overall, we follow standard ethical guidelines [65].

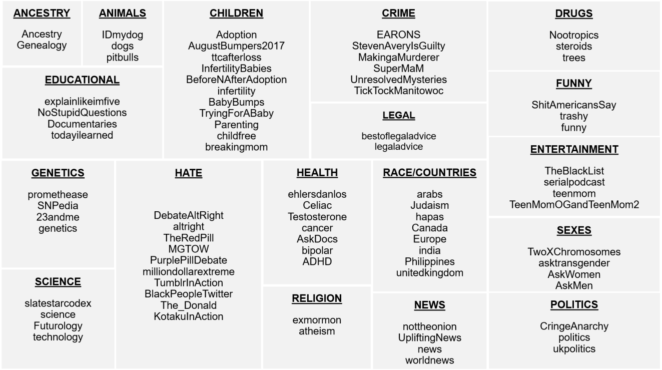
4 Genetic Testing Discussions on Reddit

In this section, we study the prevalence of genetic testing comments on Reddit. We start by identifying the subreddits with the highest number of comments related to genetic testing and thematically grouping them. Then, we use Google's Perspective API [58], a publicly available tool geared to identify toxic comments, to measure the toxicity of each group. We also use Latent Dirichlet Allocation (LDA) for basic topic modeling, aiming to extract the most prominent topics of discussion for each group. Finally, we examine comments in which users express privacy concerns.

4.1 Methodology

Subreddits selection & grouping. We extract all the subreddits where genetic testing comments have been posted to, but discard subreddits if they either have less than 1,000 comments overall or less than 100 comments with one of the keywords. This yields a list of 114 subreddits; see Table 2, which reports the normalized number of genetic testing related comments.

We group the subreddits into categories to study them based on (broad) discussion topics. We first turn to redditlist.com, a website reporting various subreddits metrics (e.g., number of subscribers, growth, etc.) and thematic tags, however, tags are available only for very popular subreddits, and most of the subreddits in our list do not have them. Thus, we have two annotators browse the subreddits and assign up to five tags based on their thematic content. We then create a dictionary based on all the tags, and pick one tag which represents each subreddit best according to the annotators' judgment (the tag is reported in Table 2). Finally, we group them based on this tag, which leads to 18 categories plus a generic one, labeled as "other" (which includes 25 subreddits). We report the subreddits in each category, except "other," in Figure 1.



ANCESTRY
Ancestry
Genealogy

ANIMALS
IDmydog
dogs
pitbulls

CHILDREN
Adoption
AugustBumpers2017
ttcafterloss
InfertilityBabies
BeforeNAfterAdoption
infertility
BabyBumps
TryingForABaby
Parenting
childfree
breakingmom

CRIME
EARONS
StevenAveryIsGuilty
MakingaMurderer
SuperMaM
UnresolvedMysteries
TickTockManitowoc

DRUGS
Nootropics
steroids
trees

EDUCATIONAL
explainlikeimfive
NoStupidQuestions
Documentaries
todayilearned

FUNNY
ShitAmericansSay
trashy
funny

LEGAL
bestoflegaladvice
legaladvice

ENTERTAINMENT
TheBlackList
serialpodcast
teenmom
TeenMomOGandTeenMom2

GENETICS
promethease
SNPedia
23andme
genetics

HATE
DebateAltRight
altright
TheRedPill
MGTOW
PurplePillDebate
milliondollarextreme
TumblrInAction
BlackPeopleTwitter
The_Donald
KotakuInAction

HEALTH
ehlersdanlos
Celiac
Testosterone
cancer
AskDocs
bipolar
ADHD

RACE/COUNTRIES
arabs
Judaism
hapas
Canada
Europe
india
Philippines
unitedkingdom

SEXES
TwoXChromosomes
asktransgender
AskWomen
AskMen

SCIENCE
slatestarcodex
science
Futurology
technology

RELIGION
exmormon
atheism

NEWS
nottheonion
UpliftingNews
news
worldnews

POLITICS
CringeAnarchy
politics
ukpolitics

Figure 1: Subreddits with genetic testing related comments, listed in Table 2grouped into categories based on their thematic topics(excluding a generic 'other' category).

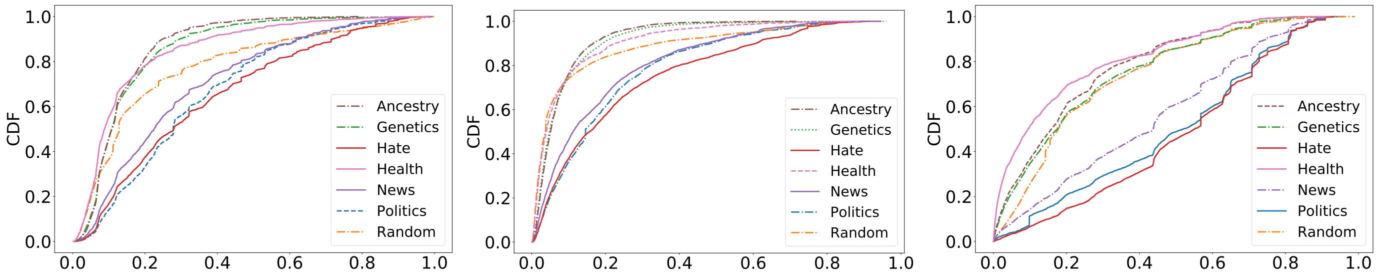
Note that, while the content of most subreddits can be intuitively guessed from the name (e.g., /r/23andMe is about the company 23andMe), that is not always the case. For instance, /r/AdviceAnimals is not about advice on animals, but on humans, and /r/trees is a subreddit about marijuana. Also, we opt to assign a separate 'Ancestry' category rather than 'Genetics', since the former includes subreddits that do not necessarily deal with genetic testing.

Prevalence of genetic testing comments. Unsurprisingly, the top five subreddits with most genetic testing comments are directly related to genetic testing/ancestry. Subreddits like /r/SNPedia or /r/Ancestry have a high fraction of comments with at least one genetic testing keyword; respectively, 10% and 7%. We also find genetic testing to be relatively popular in subreddits about dog breed identification (/r/IDmydog, 1%), children (/r/Adoption, 1%), entertainment (/r/TheBlackList, 0.6%), health (/r/ehlersdanlos, 0.7%), and crime (e.g., /r/EARONS, 0.3%). Note that, by contrast, in the random dataset, only 6 out of 204K comments (0.003%) include a genetic testing keyword. Naturally, these percentages depict conservative lower bounds as: 1) comments can be replied to by other comments, thus creating different branches of discussion, and 2) one can comment on a topic about genetic testing without using a keyword. However, our approach provides ample data points for our analysis.

Topics and toxicity. In the rest of this section, we analyze the 19 categories of subreddits in terms of the topics being discussed as well as the toxicity of the comments therein, using, respectively, LDA and Google's Perspective API [58]. The API returns three values between 0 and 1, pertaining to: 1) Toxicity, i.e., how rude, disrespectful, or unreasonable a comment is likely to be; 2) Severe Toxicity, which is similar to toxicity but only focuses on the "most toxic" comments; and 3)

	Subreddit	Gen Test	Total	Percent.	Tag
		Comms	Comms		
1	r/promethease	347	2,580	13%	Genetics
2	r/SNPedia	184	1,774	10%	Genetics
3	r/23andme	4,150	44,225	9%	Genetics
4	r/Ancestry	190	2,793	6%	Ancestry
5	r/Genealogy	3569	95,205	3%	Ancestry
6	r/genetics	347	11,741	2%	Genetics
7	r/Adoption	610	40,667	1%	Children
8	r/IDmydog	175	14,429	1%	Animals
9	r/ehlersdanlos	340	47,303	0.7%	Health
10	r/TheBlackList	288	43,127	0.6%	Entertainment
11	r/Celiac	171	41,444	0.4%	Health
12	r/Testosterone	306	83,997	0.3%	Health
13	r/serialpodcast	745	213,958	0.3%	Entertainment
14	r/EARONS	155	48,613	0.3%	Crime
15	r/StevenAveryIsGuilty	357	126,689	0.2%	Crime
16	r/cancer	172	68,037	0.2%	Health
17	r/dogs	1,627	803,094	0.2%	Animals
18	r/MakingaMurderer	1,198	624,641	0.1%	Crime
19	r/SuperMaM	139	73,997	0.1%	Crime
20	r/Nootropics	613	331,434	0.1%	Drugs
21	r/DebateAltRight	298	169,354	0.1%	Hate
22	r/AugustBumpers2017	120	71,825	0.1%	Children
23	r/ttcafterloss	223	141,992	0.1%	Children
24	r/UnresolvedMysteries	966	667,940	0.1%	Crime
25	r/InfertilityBabies	156	111,862	0.1%	Children
26	r/BeforeNAfterAdoption	107	81,078	0.1%	Animals
27	r/TickTockManitowoc	443	364,725	0.1%	Crime
28	r/pitbulls	108	103,844	0.1%	Animals
29	r/infertility	427	423,863	0.1%	Children
30	r/arabs	128	157,054	0.08%	Race/Countries
31	r/BabyBumps	973	130,1608	0.07%	Children
32	r/altright	108	166,436	0.06%	Hate
33	r/Judaism	178	299,667	0.06%	Race/Countries
34	r/AskDocs	193	385,831	0.05%	Health
35	r/TryingForABaby	192	411,263	0.04%	Children
36	r/slatestarcodex	123	273,357	0.04%	Science
37	r/bipolar	164	396,899	0.04%	Health
38	r/MensRights	399	993,039	0.04%	Sexes
39	r/bestoflegaladvice	144	362,868	0.03%	Legal
40	r/steroids	320	825,647	0.03%	Drugs
41	r/legaladvice	1,081	2,851,210	0.03%	Legal
42	r/hapas	128	368,467	0.03%	Race/Countries
43	r/science	782	2,666,213	0.03%	Science
44	r/ADHD	168	576,203	0.03%	Health
45	r/changemyview	538	1,908,120	0.02%	Other
46	r/TheRedPill	270	1,044,079	0.02%	Hate
47	r/confession	182	710,132	0.02%	Other
48	r/teenmom	203	824,312	0.02%	Entertainment
49	r/TeenMomOGandTeenMom2	133	565,612	0.02%	Entertainment
50	r/Parenting	194	829,177	0.02%	Children
51	r/childfree	350	1,531,152	0.02%	Children
52	r/MGTOW	365	1,625,881	0.02%	Hate
53	r/relationship_advice	309	1,383,111	0.02%	Other
54	r/ShitAmericansSay	122	547,506	0.02%	Comedy
55	r/relationships	1,853	8,538,031	0.02%	Other
56	r/JUSTNOMIL	359	1,790,725	0.02%	Other
57	r/TrueReddit	102	557,598	0.01%	Other
58	r/tifu	390	2,191,142	0.01%	Other
59	r/TwoXChromosomes	488	2,753,369	0.01%	Sexes
60	r/breakingmom	101	609,366	0.01%	Children
61	r/Advice	157	1,021,798	0.01%	Other
62	r/PurplePillDebate	210	1,421,805	0.01%	Hate
63	r/aww	799	5,671,423	0.01%	Other
64	r/history	142	1,054,177	0.01%	Other
65	r/raisedbynarcissists	163	1,214,553	0.01%	Other
66	r/milliondollarextreme	109	895,032	0.01%	Hate
67	r/asktransgender	158	1,307,753	0.01%	Sexes
68	r/exmormon	288	2,444,535	0.01%	Religion
69	r/nottheonion	313	2,898,542	0.01%	News
70	r/MapPorn	114	1,063,518	0.01%	Other
71	r/explainlikeimfive	388	3,741,174	0.01%	Educational
72	r/Futurology	278	2,689,784	0.01%	Science
73	r/NoStupidQuestions	198	1,943,855	0.01%	Educational
74	r/AskWomen	324	3,328,046	<0.01%	Sexes
75	r/UpliftingNews	114	1,214,761	<0.01%	News
76	r/Documentaries	130	1,386,157	<0.01%	Educational
77	r/todayilearned	1,185	13,088,194	<0.01%	Educational
78	r/conspiracy	469	5,281,831	<0.01%	Other
79	r/news	1,717	19,386,087	<0.01%	News
80	r/ireland	138	1,615,105	<0.01%	Race/Countries
81	r/TumblrInAction	216	2,563,058	<0.01%	Hate
82	r/depression	103	1,277,435	<0.01%	Health
83	r/askscience	101	1,289,247	<0.01%	Science
84	r/fatlogic	120	1,543,070	<0.01%	Hate
85	r/IAmA	242	3,521,706	<0.01%	Other
86	r/technology	268	4,072,195	<0.01%	Technology
87	r/AdviceAnimals	372	5,906,232	<0.01%	Other
88	r/Showerthoughts	477	8,034,239	<0.01%	Other
89	r/trashy	110	1,897,268	<0.01%	Funny
90	r/BlackPeopleTwitter	209	3,762,278	<0.01%	Hate
91	r/OldSchoolCool	142	2,593,419	<0.01%	Other
92	r/canada	231	4,341,997	<0.01%	Race/Countries
93	r/CringeAnarchy	217	4,101,269	<0.01%	Politics
94	r/AskMen	195	3,805,036	<0.01%	Sexes
95	r/The_Donald	1,251	28,360,073	<0.01%	Hate
96	r/worldnews	845	20,224,373	<0.01%	News
97	r/europe	219	5,275,810	<0.01%	Race/Countries
98	r/atheism	108	2,626,435	<0.01%	Religion
99	r/AskReddit	5,421	132,899,306	<0.01%	Other
100	r/india	127	3,141,858	<0.01%	Race/Countries
101	r/KotakuInAction	109	2,811,180	<0.01%	Hate
102	r/pics	543	15,528,294	<0.01%	Other
103	r/politics	1,517	46,270,193	<0.01%	Politics
104	r/personalfinance	150	4,671,327	<0.01%	Other
105	r/Philippines	102	3,245,641	<0.01%	Race/Countries
106	r/unitedkingdom	105	3,595,982	<0.01%	Race/Countries
107	r/ukpolitics	125	4,348,955	<0.01%	Race/Countries
108	r/trees	113	4,009,217	<0.01%	Drugs
109	r/WTF	131	5,609,346	<0.01%	Other
110	r/videos	319	13,934,560	<0.01%	Other
111	r/funny	321	15,792,122	<0.01%	Funny
112	r/gifs	111	9,032,723	<0.01%	Other
113	r/movies	111	11,810,334	<0.01%	Other
114	r/nba	183	23,109,676	<0.01%	Other

Table 2: List of subreddits sorted by normalized number of genetic testing comments.



(a) toxicity (b) severe toxicity (c) inflammatory

Figure 2: CDFs of Google's Perspective API toxicity metrics on the genetic testing comments for the three most and the three least toxic subreddit categories.

Inflammatory, which focuses on texts intending to provoke or inflame. In Figure 2, we plot the CDFs of the toxicity of the comments for the three most and the three least toxic subreddits (we also compare to the random dataset as a baseline). We run two-sample Kolmogorov-Smirnov (KS) tests between the distribution of each category and the random dataset: in all cases, we reject the null hypothesis that they come from a common parent distribution ($p < 0.01$). We note that the two-sample KS test is non-parametric and thus robust in terms of different sample sizes. While we acknowledge this might not be a perfect sampling, it is unlikely that any sampling method would result in perfectly balanced datasets. Also, recall that we are primarily interested in the overall comparison of content related (and unrelated) to genetic testing, thus this is appropriate for our purposes. Overall, the comments originating from subreddits related to genetics, ancestry, and health are less toxic than a random baseline, while comments in news, politics, and "hateful" subreddits are remarkably more toxic.

Remarks. We choose to use Google's Perspective to identify hateful content as other methods, e.g., hate speech detection libraries [24], are primarily trained on short texts with a limited number of training samples. Whereas, our datasets contain numerous lengthy comments which may span several thousand characters; thus, the Perspective API should perform better. In the rest of the section, we report a few representative comments for each category based on our topic analysis – i.e., interesting examples including words extracted as a topic.

4.2 Racism

Remarkably, 10/114 subreddits in our sample are categorized as hateful as they are broadly associated with hateful content. Some are clearly associated with the alt-right [73] (e.g., /r/altright, /r/DebateAltRight, and /r/The_Donald), sexism, or racism. For instance, /r/TheRedPill includes misogyny and toxic behavior towards women [45], while /r/MGTOW, Men Going Their Own Way, is a forum for men who reject romantic relationships with women, and was identified as a supremacist group by the Southern Poverty Law Center [71]. Other subreddits in this group include /r/milliondollarextreme, an American sketch satire show associated with alt-right and anti-semitism [69] which was banned in September 2018, as well as /r/KotakuInAction, which is associated with GamerGate-related toxicity [17]. Also, /r/BlackPeopleTwitter makes fun of tweets purporting to originate from African Americans.

With this in mind, we set to study the relation between genetic testing and racism on Reddit. Our Perspective API analysis (see Figure 2) shows that the category related to hate is the most toxic, and some of the subreddits (e.g., /r/DebateAltRight, /r/altright) have among the highest number of comments including genetic testing keywords in this category of subreddits. In this context, the LDA modeling gives us insight on how these fringe communities discuss genetic testing; see Table 3. Users often discuss their desire to get tested (e.g., dna, test, would, like, know), while others argue on issues related to paternity (e.g., paternity, father, support). Although we find similar topics in genetics/ancestry and parenting subreddits, here they are being expressed in a much more toxic/inflammatory manner; as evidenced by Figure 2. For example, a user writes in /r/TheRedPill: "Would get a DNA test on those kids ASAP. I don't know why all men don't do them secretly as soon as the kids are born."

Other topics are related to ancestry results (e.g., african, jewish, american, european) as well as race in general (e.g., white, black, race), which are not as widely discussed in genetics/ancestry subreddits (see Tables 4, 5). Again, the conversations exhibit clear racist connotations; for example, a user writes in /r/DebateAltRight: "The Jews know who Jews are [...] It doesn't require genetic testing [...] We whites know who whites are. Non-whites know who whites are. Anyone with eyes knows who whites are. And we will fight for our race!" Finally, we find topics related to sexual crimes (e.g., 'rape', 'women', 'evidence', 'sex'), homosexuality and gender (e.g., 'gay', 'gender'), and insurance (e.g., 'insurance', 'company', 'health').

Overall, genetic testing is a relatively popular topic of discussion in subreddits associated with fringe political views. When looking at the comments with the highest toxicity, we find some disturbing content, including instances of xenophobia (e.g., "Can you be Alt-Right and have non-white friends?", receiving the reply "No, as a member of the Alt-Right you have to DNA test all of your friends and if they're not 100% White then you report them to your local Atomwaffen," referring to a neo-nazi terrorist organization [72]). Some users explicitly advocate using genetic testing to eliminate groups of non-white ancestry (e.g., "You know with pre-implantation genetic testing we can breed out non-white ancestry fairly easily [...]").

Topic	Category: Hate
1	dna (0.069), test (0.055), get (0.017), would (0.016), like (0.014), testing (0.013), know (0.012), one (0.011), think (0.009), take (0.008)
2	child (0.037), men (0.023), women (0.022), father (0.019), woman (0.015), support (0.014), man (0.014), paternity (0.014), birth (0.011), get (0.008)
3	white (0.034), people (0.021), african (0.016), black (0.015), european (0.013), race (0.013), ancestry (0.011), like (0.008), american (0.007), genetic (0.006)
4	jewish (0.028), native (0.017), american (0.015), israel (0.015), trump (0.013), clinton (0.010), jews (0.009), cherokee (0.007), citizenship (0.007), indian (0.007)
5	rep (0.027), dem (0.027), act (0.012), gay (0.007), body (0.007), gender (0.006), use (0.004), vote (0.004), proper (0.003), russia (0.003)
6	testing (0.023), genetic (0.022), data (0.008), insurance (0.008), company (0.007), health (0.007), consent (0.007), paternity (0.006), companies (0.005), google (0.005)
7	rape (0.021), women (0.012), lie (0.010), man (0.010), police (0.008), case (0.007), false (0.007), evidence (0.007), sex (0.006), point (0.005)
8	genetic (0.016), human (0.006), even (0.006), testing (0.006), would (0.006), race (0.006), medical (0.006), differences (0.005), social (0.005), could (0.004)
9	youtube (0.010), talk (0.008), islamic (0.007), gedmatch (0.005), watch (0.005), working (0.005), video (0.005), dude (0.004), coast (0.004), saliva (0.004)
10	people (0.009), would (0.008), women (0.008), genetic (0.006), like (0.006), men (0.006), good (0.006), think (0.006), one (0.006), want (0.006)

Table 3: LDA analysis of the Hate subreddits.

4.3 Category Analysis

Next, we select a few categories of subreddits and analyze them via topic modeling and the toxicity metrics, aiming to better understand how users perceive genetic testing in each context. To ease presentation, we only do so on interesting or unexpected categories.

Genetics & Ancestry. As mentioned, the subreddits with the highest ratio of genetic testing keywords (see top five subreddits in Table 2) are directly related to genetic testing and ancestry. This is confirmed by LDA (see Tables 4 and 5). In fact, even in the genetics category, the discussion is dominated by ancestry (e.g., european, ashkenazi, african) and family (e.g., family, father, mother). We also observe that the open personal genomics database and genealogy website, GEDmatch [36], is one of the topics with the greatest weights (0.054); see Table 4. GEDmatch allows users to upload their genetic data obtained from DTC genetic testing companies to identify potential relatives who have also uploaded their data. Interestingly, in December 2018, US police forces declared that GEDmatch helped them find suspects in 28 cold murder and rape cases [38]. Overall, as shown in Figure 2, the subreddits about genetics and ancestry attract far less toxic comments than the random Reddit sample, and are the least toxic categories among the rest in our dataset. In particular, we observe extremely low levels of inflammatory content.

Crime Investigations. As mentioned, genetic testing appears to be discussed frequently in subreddits falling in the crime category, e.g., /r/EARONS, the East Area Rapist/Original Night Stalker, a.k.a. the Golden State Killer [49]. We also find subreddits covering (often controversial) discussions about Steven Avery, who was wrongly convicted of sexual assault and attempted murder (which inspired Netflix's documentary Making a Murderer); e.g., /r/StevenAveryIsGuilty seems to firmly believe Avery was justly convicted, while /r/TickTockManitowoc does not. The LDA analysis confirms how discussion in this category revolves around investigation, police, and evidence (e.g., blood, sample, vial, evidence); see Table 6. A user writes: "Similarly, why didn't we get more impact out of the DNA test on the key? Specifically, one non-courtroom interviewee makes the point that TH DNA should have been all over the key, because she had owned it for many years. The fact that only SAs DNA was found seems to be evidence that it was in fact wiped/disinfected. Why wasn't this a bombshell to be used in court?", while another says: "I am not convinced the DNA matched Teresa. I think they were probably random bones from a cadaver. Read about the DNA testing. It only matches in 7 of 15 locations." The toxicity and inflammatory levels of the content of this category are similar to the random dataset, which, combined with the LDA results, suggest that genetic testing here is discussed for informational reasons.

Parenting. Users also discuss genetic testing in the context of children, pregnancy, and parenting; e.g., in /r/Parenting, /r/Adoption, /r/TryingForABaby, /r/infertility. From the LDA analysis (see Table 7), we find that users often discuss topics related to the identity of the father or child support (e.g., father, support, lawyer), but also health and the characteristics of their child (e.g., ultrasound, gender, embryos). For example, a user is trying to support a woman admitting to having difficulties conceiving a child by saying: "Get a 2nd opinion. And just remember the DNA test was normal. Be calm. My uterus is sending yours good vibes." Once again, the subreddits in this category contain low levels of toxicity.

Animals. Reddit users also use genetic testing keywords in subreddits related to animals, and more specifically those related to dogs. For instance, /r/IDmydog, which is the 8th ranked subreddit in terms of genetic testing comments, focuses on identifying dog breeds from pictures. For example, a user writes: "Turns out, Cherry's DNA test came out as half purebred Miniature Schnauzer, one-eighth Chihuahua, and the rest possibly unknown terrier mix." /r/dogs and /r/pitbulls focus on discussion about dogs and pitbulls respectively. This is also confirmed by LDA (e.g., breeds, terrier, mixed); see see Table 8. An interesting topic of discussion is related to dog breeds banned in certain countries [3], and how one can be identified through DNA testing. For example, a user writes: "Why don't you get a DNA test done and see what he really is? If just by some chance he's not a banned breed, you can show that to your vet and get them to change the breed listed on record and then you can use that to show potential landlords if they say he looks like something that he's not." Once again, this category has similar levels of toxicity and inflammatory content to the random dataset.

Other categories. Genetic testing is also discussed in educational contexts (e.g., /r/explainlikeimfive, /r/ NoStupidQuestions), to learn about science (e.g., /r/science, /r/futurology), discuss their health (e.g., /r/celiac, /r/cancer), or in the context of drugs (/r/Nootropics, /r/steroids). User also use words related to genetic testing in a legal context (/r/legaladvice), to discuss subjects related to their cultural background (e.g., /r/arabs, /r/judaism), and to discuss religion (e.g., /r/exmormon). Finally,

Topic	Category: Genetics
1	dna (0.021), family (0.015), know (0.013), would (0.013), test (0.013), father (0.012), one (0.011), great (0.011), dad (0.009), mother (0.009)
2	european (0.023), ancestry (0.023), dna (0.017), african (0.015), results (0.014), people (0.014), native (0.013), american (0.012), eastern (0.011), german (0.009)
3	chromosome (0.031), haplogroup (0.031), ashkenazi (0.021), jewish (0.019), confidence (0.015), maternal (0.012), paternal (0.011), chromosomes (0.011), also (0.011), line (0.010)
4	genetic (0.021), testing (0.014), test (0.011), would (0.011), information (0.007), like (0.007), people (0.007), results (0.007), get (0.006), know (0.006)
5	data (0.028), snps (0.020), one (0.013), snp (0.013), snpedia (0.011), gene (0.011), genome (0.010), raw (0.009), promethease (0.008), variant (0.008)
6	blood (0.035), hair (0.023), eyes (0.018), type (0.017), cells (0.015), skin (0.015), blue (0.012), dark (0.011), brown (0.010), saliva (0.009)
7	asian (0.055), chinese (0.039), wegene (0.032), south (0.025), results (0.020), east (0.016), korean (0.014), japanese (0.014), southeast (0.013), customers (0.012)
8	sample (0.031), results (0.018), weeks (0.017), received (0.014), time (0.013), kit (0.013), samples (0.012), extraction (0.011), process (0.011), people (0.011)
9	gedmatch (0.054), dna (0.044), data (0.033), ancestry (0.026), results (0.023), raw (0.020), upload (0.016), use (0.015), get (0.013), also (0.012)
10	ancestry (0.025), promethease (0.023), health (0.022), data (0.019), get (0.017), reports (0.017), report (0.014), new (0.011), results (0.011), ancestrydna (0.010)

Table 4: LDA analysis of the Genetics subreddits.

Topic	Category: Ancestry
1	match (0.029), dna (0.026), matches (0.025), one (0.016), cousins (0.014), shared (0.013), share (0.011), cousin (0.011), related (0.011), gedmatch (0.010)
2	dna (0.020), family (0.019), test (0.018), great (0.012), father (0.012), know (0.011), mom (0.011), would (0.011), mother (0.010), side (0.010)
3	native (0.085), american (0.076), cherokee (0.018), ancestry (0.014), indian (0.011), nbsp (0.009), family (0.009), tribe (0.009), claim (0.008)
4	dna (0.026), ancestry (0.018), results (0.011), irish (0.009), people (0.009), european (0.008), like (0.008), african (0.008), ethnicity (0.008), british (0.008)
5	william (0.019), youtube (0.016), watch (0.016), african (0.014), norwegian (0.013), sub (0.011), saharan (0.011), middle (0.009), census (0.008)
6	dna (0.062), test (0.049), testing (0.020), father (0.020), would (0.019), autosomal (0.013), family (0.012), get (0.012), line (0.011), haplogroup (0.010)
7	ancestry (0.049), gedmatch (0.045), ftdna (0.028), dna (0.026), upload (0.024), results (0.024), test (0.023), matches (0.022), get (0.018), data (0.017)
8	jewish (0.031), european (0.023), asian (0.020), europe (0.018), east (0.017), eastern (0.017), italian (0.015), results (0.015), ancestry (0.014), ashkenazi (0.013)
9	dna (0.037), ancestry (0.018), ancestrydna (0.016), test (0.015), testing (0.013), data (0.010), tests (0.009), results (0.008), tree (0.008), information (0.007)
10	tree (0.029), find (0.018), family (0.017), people (0.016), trees (0.013), ancestry (0.012), see (0.012), records (0.012), matches (0.010), search (0.009)

Table 5: LDA analysis of the Ancestry subreddits.

we find genetic testing words in subreddits related to entertainment programs (e.g., /r/TheBlackList), comedy (e.g., /r/funny), and issues related to gender (e.g., /r/AskMen, /r/AskWomen).

4.4 Privacy Concerns

We also examine comments where users discuss privacy concerns in the context of genetic testing. We extract comments which include both a genetic testing keyword and the word 'privacy' from our Reddit dataset, getting 560 comments (0.7% of all comments). Obviously, this set is a conservative sample, as it is possible for a user to discuss issues related to privacy without specifically mentioning the word 'privacy'. Then, we use LDA to identify the context in which users discuss issues related to privacy; see Table 9.

The most common subreddits in which privacy issues are being discussed are /r/genealogy, /r/news, and /r/23andMe. We find that Reddit users express privacy concerns on the use of genetic testing (e.g., dna, data, information, privacy, gender). Specifically, a topic of discussion is the potential misuse of genetic information by employers, while another topic focuses on paternity tests and whether children have the right to know their biological father. Finally, several users discuss the privacy issues stemming from a bill passed by the Republican Party on March 8, 2017 which allows companies to ask for their employees' genetic test results [5].

4.5 User Analysis

We also examine the overlap in users discussing genetic testing among all 114 subreddits in our sample. We do so to examine whether subreddits that have common interests have also similar user base. For instance, we want to assess if users that post on /r/23andMe, also post on /r/ancestry. To do so, we extract the set of users that posted in each subreddit and calculate the pairwise Jaccard Index scores between the set of users in each subreddit. Next, we create a complete graph where nodes are the subreddits and edges are weighted by the Jaccard Index. We then run the community detection algorithm in [10], which provides a set of communities based on the graph's structure.

Figure 3 shows the resulting graph: nodes that have the same color are part of the same community. The main observations are the following: 1) there are high Jaccard Index scores between the nodes in the same community, i.e., there is a substantial overlap of users that posted in all subreddits within the community. 2) Genetic testing subreddits (e.g., /r/genetics, /r/promethease, /r/ancestry, /r/23andMe) are part of the same community (pink nodes) as scientific and education ones (e.g., /r/askscience, /r/science, /r/futurology), highlighting that "enthusiasts" are also active on scientific subreddits. 3) Subreddits associated with sexist content essentially share the same users (e.g., /r/MGTOW, /r/TheRedPill, /r/PurplePillDebate, lower left in olive green); also, users who discuss genetic testing in /r/The_Donald are also active in other alt-right subreddits like /r/AltRight, /r/DebateAltRight (mint green nodes).

Additionally, we find communities with subreddits focused on the geopolitical aspects of genetic testing (see light blue nodes on the top left) like /r/europe, /r/canada, /unitedkingdom, and /r/ukpolitcs, as well as subreddits about personal advice (light blue nodes on the bottom right) like /r/advice, /r/parenting, /r/legaladvice, /r/bestoflegaladvice. Other communities are centered around conceiving children (e.g., /r/infertility, /r/tryingforababy, /r/babybumps, orange nodes on the bottom right side), crime investigation (e.g., /r/MakingaMurderer, /r/StevenAveryIsGuilty, orange nodes on the top left side), and animals (e.g., /r/dogs, /r/IDmydog, /r/pitbulls, pink nodes on top right side).

Overall, Reddit users are not uniformly interested in every aspect of genetic testing, but rather specific communities focus on specific aspects thereof. For example, we find groups ranging from genetic testing enthusiasts, i.e., those who are inter-

Topic	Category: Crime
1	dna (0.041), would (0.020), testing (0.019), think (0.016), people (0.012), like (0.011), test (0.011), know (0.010), could (0.009), get (0.009)
2	blood (0.060), dna (0.043), testing (0.023), test (0.019), sample (0.013), vial (0.012), samples (0.012), tested (0.010), lab (0.009), tests (0.009)
3	found (0.016), murder (0.014), police (0.013), case (0.010), years (0.009), later (0.009), dna (0.008), man (0.007), went (0.007), convicted (0.006)
4	dna (0.054), test (0.020), evidence (0.019), testing (0.011), would (0.011), bullet (0.010), could (0.009), one (0.008), case (0.007), found (0.007)
5	one (0.011), would (0.007), control (0.007), lab (0.006), test (0.006), like (0.006), case (0.006), evidence (0.005), science (0.005), say (0.005)
6	evidence (0.023), avery (0.020), testing (0.016), dna (0.014), case (0.013), court (0.009), allen (0.008), trial (0.008), would (0.008), state (0.007)
7	father (0.031), family (0.023), mother (0.012), son (0.012), dad (0.011), related (0.011), adam (0.011), cousin (0.010), cousins (0.009), different (0.008)
8	said (0.019), fire (0.016), family (0.012), hobbs (0.008), brendan (0.007), barb (0.006), sketch (0.005), monday (0.005), richard (0.005), death (0.004)
9	avery (0.029), blood (0.017), would (0.017), evidence (0.017), found (0.015), key (0.011), garage (0.010), car (0.008), trailer (0.007), police (0.007)
10	bones (0.049), bone (0.035), remains (0.029), found (0.023), human (0.019), fragments (0.017), burn (0.016), pit (0.015), body (0.014), teresa (0.013)

Table 6: LDA analysis of the Crime subreddits.

Topic	Category: Children
1	testing (0.023), genetic (0.020), weeks (0.016), back (0.014), pregnancy (0.012), first (0.012), loss (0.010), results (0.010), pregnant (0.009), get (0.009)
2	genetic (0.035), testing (0.017), child (0.017), children (0.014), people (0.012), health (0.012), would (0.011), kids (0.009), medical (0.008), life (0.008)
3	know (0.019), like (0.019), want (0.014), would (0.014), get (0.013), really (0.012), feel (0.011), time (0.010), think (0.009), even (0.009)
4	child (0.050), dna (0.030), test (0.028), father (0.023), support (0.020), kid (0.016), dad (0.011), lawyer (0.011), paternity (0.010), get (0.009)
5	insurance (0.025), testing (0.013), get (0.013), genetic (0.012), doctor (0.012), labcorp (0.011), blood (0.009), pay (0.009), test (0.009), covered (0.008)
6	dna (0.030), test (0.020), family (0.019), parents (0.013), ancestry (0.012), also (0.011), birth (0.011), adoption (0.011), find (0.011), get (0.010)
7	weeks (0.034), genetic (0.029), scan (0.024), girl (0.022), testing (0.022), ultrasound (0.021), boy (0.016), baby (0.014), week (0.013), gender (0.012)
8	test (0.022), dna (0.016), name (0.012), back (0.012), got (0.008), came (0.008), said (0.008), little (0.008), son (0.007), chow (0.006)
9	ivf (0.017), embryos (0.014), testing (0.011), one (0.010), genetic (0.010), pgs (0.010), dog (0.010), sperm (0.010), embryo (0.009), transfer (0.009)
10	genetic (0.032), testing (0.030), test (0.024), would (0.015), risk (0.012), baby (0.011), done (0.011), also (0.009), results (0.008), back (0.008)

Table 7: LDA analysis of the Children subreddits.

Topic	Category: Animals
1	pit, breeds, breed, bull, amp, bulls, terrier, dogs, mixed, jpg
2	dna, test, mix, like, dog, get, know, really, could, would
3	dogs, genetic, still, testing, breed, thing, even, dog, issues, pedigree
4	health, genetic, breed, dogs, breeder, testing, breeding, breeders, dog, puppy
5	dog, breed, dna, banned, type, test, prove, one, court, pit

Table 8: LDA analysis of the Animals subreddits.

Topic	Category: Privacy
1	dna, privacy, information, data, genetic, testing, ancestry, would, like, people
2	would, child, test, people, one, privacy, father, think, know, right
3	dna, genetic, health, information, employers, bill, wellness, sequencing, would, testing
4	dna, act, table, formatted, view, article, privacy, genetic, gender, testing
5	cancer, breast, congress, house, trump, bill, genetic, republicans, act, laws

Table 9: LDA analysis of the comments including 'privacy'.

ested in or have undergone genetic testing, to people who discuss genetic testing exclusively in subreddits with educational and scientific content, to those who use genetic testing terminology exclusively when discussing fringe political views.

4.6 Take-Aways

Our Reddit analysis shows that genetic testing is discussed in a variety of contexts which is an indicator of how mainstream it has recently become. For instance, they discuss it in the context of issues related to their children, pets, or health, or to debate on their cultural heritage. More interestingly, Reddit users are not uniformly interested in every aspect of genetic testing, rather, they form groups ranging from genetic testing enthusiasts to individuals with fringe political views. Thus, we observe a dichotomy in the type of users interested in genetic testing: some users are interested in typical uses of genetic testing, others discuss their use in worrying ways. Specifically, we find evidence of toxic language displaying clear racist connotations, and of groups of users using genetic testing to push racist agendas, e.g., to eliminate or marginalize minorities. This is also particularly worrying since Reddit is a mainstream platform (5th most visited site in the US [60]).

5 Genetic Testing Discussions on /pol/

In this section, we study the prevalence of genetic testing comments on 4chan's politically incorrect board (/pol/). We first conduct a general characterization of the threads containing genetic testing keywords where we, similarly to the previous section, use Google's Perspective API to measure the toxicity of the contents, and LDA modeling to extract the most prominent topics of discussion. Then, we use Perceptual Hashing [50] and DBSCAN clustering [29] to study the imagery and memes of the dataset.

5.1 General Characterization

Thread Activity. We begin by measuring the number of posts in threads where genetic testing keywords appear, aiming to examine whether these threads attract more or less activity than "usual." On /pol/, there is a limit on how many threads can simultaneously be active: whenever a new one is created, that with the oldest last post is purged. There is also a "bump" limit that prevents a thread from never being purged. As per [41], the majority of threads attract only a few posts before being archived, while some, often covering controversial or popular topics, get many posts and possibly hit the bump limit.

In Figure 4, we plot the CDF of the number of posts per thread, for both the genetic testing threads and our random sample. The former have an order of magnitude more posts than the latter (the median is 183 and 5 posts, respectively), which is an indicator that genetic testing is often discussed in long-lasting/interesting threads and may attract more attention

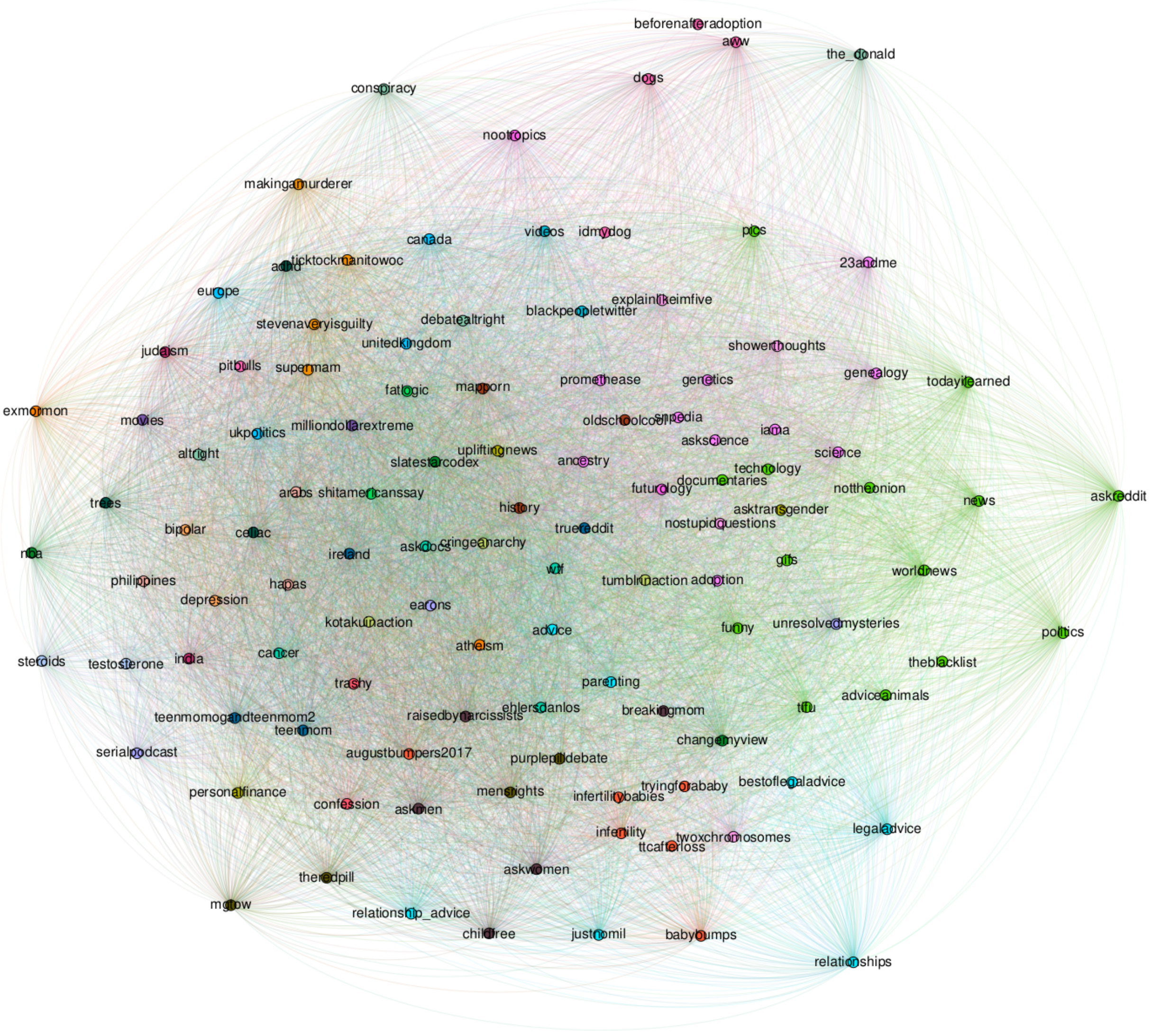


Figure 3: Graph depicting the Jaccard Index of the users whose comments include genetic testing keywords for each subreddit.

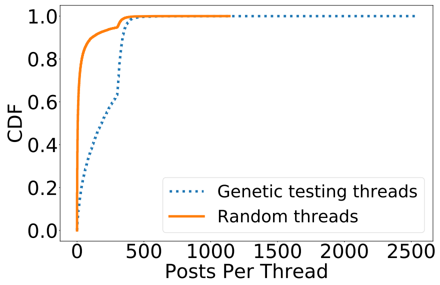


Figure 4: CDF comparing 4chan threads with genetic testing keywords and random threads in terms of number of posts.

by users. We also run a two-sample Kolmogorov-Smirnov test on the distributions and we reject the null hypothesis that they come from a common parent distribution ($p < 0.01$).

Toxicity & Hate. We also measure hate and toxicity in /pol/ threads by computing: 1) the percentage of hate words, and 2) the toxicity/inflammatory levels. For the former, we use a dictionary of hate words compiled by and available from hatebase.org, as used in [41]; for the latter, we again rely on the Perspective API. However, we find no major differences between the genetic testing threads and the random sample—which is not surprising as /pol/ is known for their high level of hate speech [41]—thus, we omit related plots to ease presentation.

Topic Modeling. We then use LDA modeling to identify the most prominent topics of discussion; see Table 10. Similar to Reddit users, 4chan users use keywords suggesting their intention to get tested (e.g., would, get, dna, test). Several topics are related to ancestry, which is also among the words with the highest weights (0.048); for instance, users often discuss the ancestral background of the American population (e.g., american, african, european, white), others debate the cultural connection of modern humans to ancient civilizations (e.g., egyptians, greeks, roman), and the facial traits of modern europeans (e.g., german, irish, eyes, hair). Interestingly, another prominent topic of discussion is related to Lauren Southern, an Inter-

Topic	4chan
1	ancestry (0.048), african (0.046), european (0.023), white (0.015), american (0.012), north (0.011), americans (0.010), population (0.008), south (0.008), europeans (0.008)
2	youtube (0.030), watch (0.028), jewish (0.020), king (0.013), company (0.010), lauren (0.010), tut (0.009), monkey (0.008), igenea (0.007), haplogroup (0.006)
3	ancient (0.023), modern (0.020), egyptians (0.015), egypt (0.012), years (0.009), national (0.008), egyptian (0.008), greeks (0.008), roman (0.007), saharan (0.007)
4	women (0.015), children (0.015), woman (0.011), men (0.010), man (0.009), genes (0.009), kids (0.009), child (0.008), two (0.008), birth (0.008)
5	genetic (0.030), data (0.022), ancestrydna (0.014), information (0.014), health (0.013), company (0.012), testing (0.011), research (0.011), use (0.008), send (0.007)
6	back (0.022), got (0.021), european (0.020), family (0.020), german (0.013), took (0.012), irish (0.011), hair (0.011), came (0.011), eyes (0.010)
7	dna (0.063), test (0.042), white (0.024), like (0.017), people (0.015), would (0.012), genetic (0.012), one (0.011), get (0.011), even (0.010)
8	gedmatch (0.024), raw (0.014), creation (0.008), human (0.007), far (0.007), data (0.007), got (0.007), son (0.006), run (0.006), forum (0.006)
9	screw (0.016), tweet (0.010), bill (0.010), tea (0.010), news (0.010), reddit (0.009), look (0.007), fda (0.005), search (0.005), guy (0.005)
10	companies (0.018), pay (0.016), child (0.015), order (0.015), racists (0.014), support (0.012), testing (0.011), adding (0.011), admit (0.011), law (0.011)

Table 10: LDA analysis of /pol/.

Entity	Clusters (%)	Entity	Clusters(%)
/pol/	15 (6.9%)	Video	3 (1.4%)
Lauren Southern	15 (6.9%)	Jewish people	3 (1.4%)
23andMe	13 (6.0%)	Logo	3 (1.4%)
Pepe the Frog	9 (4.1%)	White	3 (1.4%)
United States of America	8 (3.7%)	Shaun King	2 (0.9%)
Richard Spencer	5 (2.3%)	Screenshot	2 (0.9%)
Genetic	4 (1.8%)	4chan	2 (0.9%)
Meme	4 (1.8%)	The Holocaust	2 (0.9%)
Europe	3 (1.4%)	Race	2 (0.9%)
Greece	3 (1.4%)	Adolf Hilter	2 (0.9%)

Table 11: Top 20 entities with the most clusters.

net personality associated with the alt-right, whose popularity rose after being detained in Italy for trying to block a ship rescuing refugees [20] (e.g., lauren, jewish, youtube). Other conversations likely relate to how genetic testing companies use their data (e.g., genetic, data, use, research), as well as legal issues related to child support (e.g., child, birth, support, law).

5.2 Image Analysis

Next, we look at the images and memes that are shared in /pol/ posts including genetic testing keywords. We use the open source image analysis pipeline introduced by [77]. The pipeline uses Perceptual Hashing [50] and DBSCAN [29] to group together images that are visually similar. We run the pipeline on the 6,375 images included in *posts* where at least one genetic testing keyword appears; as discussed earlier, this is in contrast to the textual analysis where we look at whole threads. (Recall from Table 1 that the total number of images in threads containing genetic testing keywords is 338,540.) We obtain 215 clusters including 543 total images; the other 5,832 images are labeled as noise by the clustering algorithm and thus we discard them. This high noise ratio mirrors findings in [77] and is likely due to 4chan users creating a lot of original content [41]. Also, our dataset only includes a few thousand images, thus not a lot of images are visually similar.

We annotate each cluster using Google's Cloud Vision API[2], specifically, we calculate the medoid of each cluster (i.e., its "representative" image) following the methodology by [77], and use that image to query the API. This returns a set of meaningful entities, which are obtained by searching labeled images across the Web, along with their confidence scores. The exact methodology for extracting the entities is not known, however, upon manual examination, we can confirm that the API is indeed able to extract fine-grained entities. For instance, given an image with Donald Trump, the API returns an entity called "Donald Trump" and not generic labels like "man" or "politician."

For each cluster, we extract the entity with the highest confidence score and analyze the top 20 entities, as reported in Table 11. The most popular entries are /pol/ itself and Lauren Southern with 6.9% of all clusters. The latter is particularly interesting as it adds to the evidence that discussions about genetic testing frequently involve alt-right celebrities. In fact, pictures of American white-supremacist Richard Spencer [75] (6th most popular with 2.3% of all clusters), and Carl Benjamin, a YouTuber known for his misogynistic involvement in the GamerGate controversy [8], are also popular.

We also find several clusters related to: 1) 23andMe (6.0%), e.g., screenshots of genetic testing results from 23andMe or images with the 23andMe logo, 2) memes including Pepe the Frog (4.1%), a 4chan-popularized hate symbol [1], and 3) geographic images related to, e.g., the US (3.7%), Europe (1.4%), or Greece (1.4%). The latter is likely mirroring discussions about the connection of modern humans to ancient civilizations; see topic 6 in Table 10. We also find imagery related to the Jewish community (1.4%), as well as the Holocaust (0.9%) and Hitler (0.9%), suggesting that, on 4chan, genetic testing terms and Nazi-related imagery are used together for the dissemination of hateful and antisemitic content.

We also examine the entities in Table 11 more closely to shed light on the context in which images are being discussed. Specifically, we extract text from the posts appearing alongside the images and use LDA modeling on the posts of each entity separately. We set LDA to produce only three topics per entity given the limited number of posts per entity. Among other things, we find that posts containing images related to 23andMe (see Table 12) actually include discussions with racial connotations; for instance, whether test results show signs of African ancestry (e.g., ancestry, percent, african), or whether people with Jewish heritage are behind the company (e.g., jewish, company, results). For example, a user writes: "Can a genetics company founded by a Jew be trusted?" Similarly, posts with images annotated as United States of America (see Table 13) reveal discussions on the ancestral background of the American population (e.g., americans, ancestry, african, whites). A user writes: "Less than 5% of White Americans have even negligible amounts of African DNA".

Cluster visualization. Finally, we provide a visualization of the clusters in Figure 6. Nodes in the graph represent clusters, while edges represent the Jaccard Index between clusters (as

[2] https://cloud.google.com/vision/

Topic	Entity: 23andMe
1	dna (0.050), ancestry (0.035), tests (0.024), results (0.018), one (0.018), percent (0.016), african (0.016), got (0.014), would (0.014), could (0.014)
2	could (0.030), jewish (0.030), even (0.023), pol (0.023), people (0.023), also (0.023), company (0.016), test (0.016), results (0.016), markers (0.016)
3	white (0.039), genetic (0.034), test (0.034), heritage (0.022), european (0.022), dna (0.018), jew (0.018), like (0.018), nigger (0.014), still (0.014)

Table 12: LDA analysis of the texts in the /pol/ posts with imagery annotated as '23andMe'.

Topic	Entity: United Stated of America
1	white (0.044), ancestry (0.038), americans (0.031), self (0.028), african (0.021), european (0.018), even (0.018), whites (0.018), race (0.018), american (0.018)
2	white (0.039), roman (0.024), people (0.021), whites (0.018), full (0.018), empire (0.016), citizenship (0.016), held (0.016), admixture (0.016), like (0.016)
3	sargon (0.042), get (0.037), spencer (0.032), enoch (0.032), like (0.027), anyone (0.027), think (0.022), say (0.017), would (0.017), even (0.017)

Table 13: LDA analysis of the texts in the /pol/ posts with imagery annotated as 'United Stated of America'.

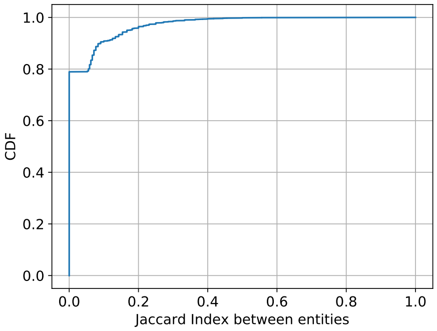


Figure 5: CDF of the clusters' Jaccard Index scores using the set of entities returned by the Cloud Vision API.

per the entities returned by the Cloud Vision API). To ease presentation, we only report edges where the Jaccard Index is greater than a threshold. To select this threshold, we plot the CDF of all the Jaccard Index scores in Figure 5, which shows that 80% of clusters are completely disjoint (Jaccard Index equal to 0). As a result, we decide to select a 0.2 threshold. This corresponds to selecting 4.1% of the edges with the highest Jaccard Index, allowing us to understand the *main* connections between clusters. Then, we perform community detection, using the approach presented in [10]. This considers the structure of the graph and decomposes it into a set of communities, where each community includes a set of highly inter-connected nodes. The resulting graph is presented in Figure 6, with each color representing a different community. For each community, we have manually inspected the images in the clusters and added a high-level description as well as a representative image.

The figure highlights the presence of two tightly-knit communities (bottom right): the green community includes images with logos of genetic testing companies, while the light red community covers images with screenshots of genetic testing results. We also find communities with images related to Haplogroups and Genealogy Trees, as well as others related to the alt-right (top of the graph). In fact, a few communities exhibit clear racial connotations (pink), e.g., a cluster including an image from National Geographic predicting how the average American woman will look like in 2050 [32], which, unsurprisingly, attracted numerous posts on 4chan. Finally, a few communities are related to hateful memes like Pepe the Frog and the Happy Merchant, a caricature of a manipulative Jew used on 4chan in racist contexts [30].

5.3 Takeaways

Overall, we find that genetic testing is a rather popular topic of discussion in 4chan's /pol/, appearing in long/active threads. Also, genetic testing topics are often accompanied by images and memes with clear racial or hateful connotations. While the presence of highly toxic content in /pol/ is unsurprising, the specific content which accompanies threads related to genetic testing is worrying. We find imagery with prominent figures of the alt-right movement (e.g., Lauren Southern, Richard Spencer), antisemitic memes (e.g., Pepe the Frog, Happy Merchant), and topics of discussion using words with racial/hateful meaning (e.g., jewish, monkey, nigger), which may be an indicator that groups adjacent to the alt-right are using genetic testing to bolster their ideology.

6 Language Analysis

Although they both provide discussion platforms, Reddit and 4chan operate in different ways: e.g., the former requires registration, while the predominant mode of operation on the latter is via anonymous and ephemeral posting. Also, Reddit supports an infinite number of user-driven sub-communities, while on 4chan conversations are organized around a few dozens boards, with images playing a key role. Naturally, they also attract different sets of users and content, e.g., 4chan is typically identified as a fringe community, while, Reddit, though also hosting fringe communities, is overall a mainstream site (5th most visited in the US).

Our analysis of genetic testing on the two platforms thus far has highlighted that genetic testing is a subject which is discussed frequently; on Reddit in subreddits ranging many aspects of the every day life of the users, on 4chan, in threads that attract an order of magnitude more posts. At the same time, on both platforms, fringe political groups express their wish to marginalize minorities using genetic testing. Next, we provide a comparison of the *language* used in the context of conversations that are likely to include genetic testing. To do so, we turn to word embeddings, specifically, word2vec [46]. Word2vec models are trained on large corpora of text, and generate a high-dimensional vector for each word that appears in the corpus; words that are used in similar context also have a closer mapping to the high-dimensional vector space. This al-

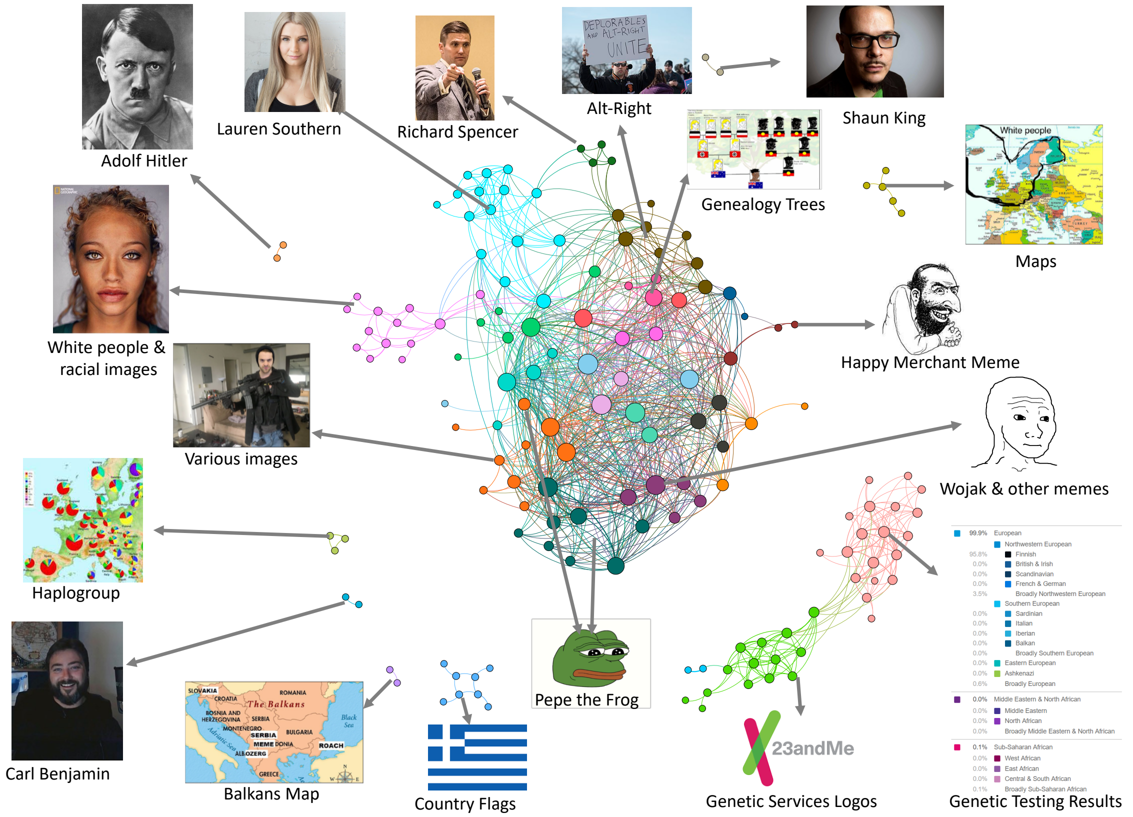


Figure 6: Visualization of the image clusters with manual annotation.

lows us to study which words are used in similar contexts.

Methodology. We train a separate word2vec model, as per the implementation provided by [62], for each of the 19 groups of subreddits (see Figure 1) and 4chan's /pol/, using all of the posts made between January 1, 2016 and March 31, 2018, and June 30, 2016 and March 13, 2018, respectively. We preprocess each corpus as follows: 1) we remove special symbols, punctuation, URLs, and numbers; 2) we tokenize each word that appears on each post; and 3) we perform stemming on the words using the Porter algorithm. Next, we train word2vec models for each community on all the pre-processed posts and all words that appear at least 100 times in each corpus. We use a *context window* equal to 7, i.e., the model considers a context of up to 7 words ahead and behind the current word.

Vocabulary. Table 14 reports the number of words that are considered in each word2vec model. Vocabulary sizes vary greatly, e.g., from 122 in the Ancestry subreddits to 46K in Race/Culture subreddits. This is due to the fact that we only consider words that appear at least 100 times.

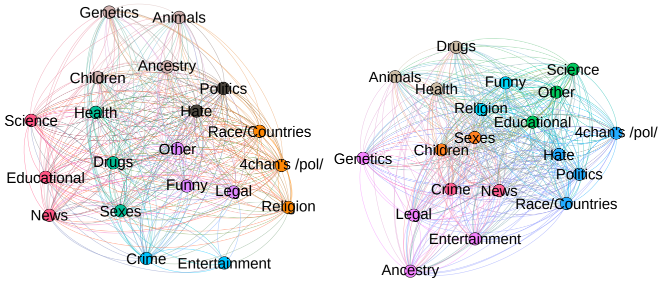
Training. To assess how each community discusses topics related to ethnicity and genetic testing words, for each word2vec model, we get the 10 most similar words for two groups of seed words: 1) 91 genetic testing keywords obtained from the list of 280 keywords (the other 189 including multiple words so we do not consider them) 2) a hand-picked set of words, namely, “white,” “black,” “jew,” “kike,” “ancestry,” “dna,” and “test.” The latter are added aiming to assess whether ethnic terms (e.g., “white”) and genetic testing keywords (e.g., “dna”) are used in different contexts than the set of genetic keywords (e.g., “23andMe”).

Group	# of Words in Vocabulary	Group	# of Words in Vocabulary
4chan's /pol/	31,337	Hate	40,223
Ancestry	122	Health	11,101
Animals	8,065	Legal	4,655
Children	15,858	News	32,097
Crime	11,649	Politics	41,057
Drugs	7,858	Race/Countries	46,978
Educational	23,151	Religion	12,431
Entertainment	7,743	Science	18,341
Funny	5,641	Sexes	20,743
Genetics	1,178	Other	24,767

Table 14: Words that are in the vocabulary of the word2vec models trained for each group of subreddits and /pol/.

Visualization. We calculate the *similarity* of all the possible combinations of word2vec models using the Jaccard Index scores of all the similar words for all the seed words. Then, we create two complete graphs (see Figure 7), one for each set of seed keywords, where nodes are the trained word2vec models and edges are weighted by the Jaccard Index score between the similar words for all the seed words. Once again, we use the community detection algorithm by [10].

By looking at the communities, we gather some interesting insights about how communities compare to each other. When using the genetic testing keywords as seeds (Figure 7(a)), we find that communities about genetics, ancestry, animals, and



(a) Genetic Testing Keywords

(b) Selected Keywords

Figure 7: Graph representation of the word2vec models, using as seeds: (a) all the genetic testing keywords, (b) the terms "white," "black," "jew," "kike," "ancestry," "dna," and "test."

children discuss genetic testing in very similar contexts (light brown nodes). Similarly, we find a cluster with subreddits with scientific, educational, and news content (red nodes on the left), and another related to health, drugs, and sexes (green nodes). Interestingly, the subreddits in the hate category discuss genetic testing in a similar manner as the political ones (brown nodes); this is not entirely surprising also considering that these categories have the two highest toxicity levels (cf. Figure 2). Also, /pol/ users seem to discuss genetic testing in a context similar to subreddits related to race/countries and religion (orange nodes). This may be because /pol/ frequently discusses Judaism (with references to Israel and the Jewish community), as well as other religions [30].

When using the set of hand-picked seed words (Figure 7(b)), /pol/ is similar to the hateful subreddits, as well as the subreddits about politics and race/countries (blue nodes). In other words, Hate, Politics, Race/Countries subreddits, and /pol/, use ethnic terms in conjunction with genetic testing keywords in similar contexts. Overall, the fact that that certain subreddits share language characteristics with /pol/ is particularly worrying as it may be an indicator of 4chan's fringe ideologies propagating into more mainstream media.

7 Discussion & Conclusion

Direct-to-consumer (DTC) genetic testing is one of the first revolutionary technologies with the potential to transform society by improving people's lives. Nowadays, citizens of most developed countries have easy and affordable access to a wealth of informative reports, which allow them to better understand themselves, learn about their health and their cultural heritage, and find lost relatives [12]. However, this new technology also harbors societal dangers as it is used by fringe groups as "evidence" on which to build discrimination and prejudice, and potentially increase ethnic sectarianism. Considering how information has become increasingly misused on the Web, the potential abuse of genetic testing on online platforms is not be underestimated.

Nevertheless, prior work on this topic has mostly been limited to relatively small (qualitative) studies [57, 66], which discuss how DTC genetic testing may have a negative societal impact due to their results possibly reinforcing the concept of racial privilege. In that respect, our analysis furthers this line of research by taking a large-scale, data-driven approach, which provides new insight into both the breadth and depth of the issue (of which hate speech is an important aspect). We believe that our findings broaden the discussion around DTC genetic testing and its potential misuse in furthering hateful rhetoric and ideology as we provide quantitative evidence for the prior qualitative work.

More specifically, we shed light on online discussions about genetic testing on two social networking sites, Reddit and 4chan's politically incorrect board (/pol/), which are known to provide a platform to fringe and alt-right communities. We analyzed 1.3M comments spanning 27 months using a set of 280 keywords related to genetic testing as search terms, relying on a mix of tools including Latent Dirichlet Allocation, Google's Perspective API, Perceptual Hashing, and word embeddings to identify trends, themes, and topics of discussion. Our analysis showed that genetic testing is frequently discussed on both platforms. For instance, on /pol/, we find an order of magnitude increase in activity on threads related to genetic testing when compared to a random sample. Interestingly, images appearing along genetic testing conversations often include alt-right personalities and antisemitic memes. On Reddit, genetic testing is discussed in a wider variety of contexts, however, while there are communities building around the more positive aspects (e.g., health, cultural heritage, etc.), we also found others where conversations include racist, hateful, and misogynistic content.

Overall, we uncovered evidence of genetic testing indeed being misused in online discussions, further ingraining and empowering genetics-based prejudice, discrimination, and even calls for genocide. For instance, comments on both /pol/ and a set of "hateful" subreddits often contain highly toxic language, with users even suggesting leveraging genetic testing tools to further marginalize or even eliminate minorities. In fact, word embeddings showed that /pol/ and certain subreddits share worrying language characteristics, which may be an indicator of 4chan's fringe ideologies spilling out to more mainstream platforms.

Our findings are particularly timely as recent events indicate that those interested in societal disruption have successfully seized upon technological innovations and used them in ways that were not intended by their creators. More specifically, information has been increasingly weaponized, including by state actors, to sew racial discontent [74] and even instigate public health crises [13]. In this context, recent efforts have been made by law enforcement to understand and address such campaigns [51]. Thus, we ought to reflect on the practical implications of our findings and how they affect future work in this area. Considering that previous qualitative studies [57, 66] demonstrate how the commercialization of genetic testing may have a negative societal impact, and since our study provides quantitative data on the matter, the next natural step is to examine whether genetic ancestry testing has an (indirect) effect on the levels of racism and discrimination online. Naturally, such a correlation is not easy to identify and it may require a mixed-methods methodological approach (e.g., interviews with people adjacent to the far-right), but our work arguably provides a stepping stone toward this.

Finally, we note that platforms like Facebook and Twitter have begun to be held accountable when their services enable harmful behavior [76]; if there are strong indications that DTC genetic ancestry testing exacerbates online discrimination, we believe that the DTC industry should also consider the potential abuse of their services and attempt to find ways of minimizing this behavior. In future work, we plan to build tools that automatically distinguish healthy from toxic comments about genetic testing. Currently, a number of techniques (e.g., [24, 26, 27, 33–35]) are available that can be used to identify hateful/toxic comments, using machine learning models trained on annotated datasets. We plan to use similar methods on the dataset built in this work to train models that identify toxic comments specifically in the context of genetic testing, confident that this will yield better accuracy than generic ones.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie "Privacy&Us" and "ENCASE" projects (Grant Agreement No. 675730 and 691025), as well as a Google Faculty Award. We also gratefully acknowledge the support of the NVIDIA Corporation, for the donation of the two Titan Xp GPUs used for our experiments.

References

[1] ADL. Pepe the Frog. https://www.adl.org/education/references/hate-symbols/pepe-the-frog, 2019.

[2] AncestryDNA. Ancestry Company Facts. https://www.ancestry.com/corporate/about-ancestry/company-facts, 2019.

[3] Anna Jones. Brief Summary of Breed Specific Legislation. https://www.animallaw.info/intro/breed-specific-legislation-bsl, 2017.

[4] BBC. James Watson: Scientist loses titles after claims over race. https://www.bbc.co.uk/news/world-us-canada-46856779, 2019.

[5] S. Begley. House republicans would let employers demand workers' genetic test results. https://www.statnews.com/2017/03/10/workplace-wellness-genetic-testing/, 2017.

[6] A. Ben-David and A. Matamoros-Fernandez. Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain. *IJOC*, 10:1167–1193, 2016.

[7] M. Bernstein, A. Monroy-Hernández, D. Harry, P. André, K. Panovich, and G. Vargas. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. *ICWSM*, 2011.

[8] J. Bish. Vice News. Examining the Right Wing British Blowhards Using YouTube to Prove Everybody Wrong. https://bit.ly/2qN4SMG, 2016.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[10] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast Unfolding of Communities in Large Networks. *JSTAT*, 2008(10):P10008, 2008.

[11] E. Boodman. White Nationalists Are Flocking To Genetic Ancestry Tests – But Many Don't Like Their Results. https://read.bi/2DEaQYY, 2016.

[12] K. S. Borrelli. PressConnects. DNA Tales: These People Found Long-Lost or Never-Known Relatives. https://bit.ly/2FxDye2, 2018.

[13] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10), 2018.

[14] T. Caulfield and A. L. McGuire. Direct-To-Consumer Genetic Testing: Perceptions, Problems, and Policy Responses. *Annual Review of Medicine*, 63:23–33, 2012.

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

[16] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert. The Bag of Communities. In *CHI*, pages 3175–3187, 2017.

[17] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying. In *WWW 2017*, 2017.

[18] P. Chow-White, S. Struve, A. Lusoli, F. Lesage, N. Saraf, and A. Oldring. 'warren buffet is my cousin': Shaping public understanding of big data biotechnology, direct-to-consumer genomics, and 23andme on twitter. *Information, Communication & Society*, 21(3):448–464, 2018.

[19] E. Christofides and K. O'Doherty. Company Disclosure and Consumer Perceptions of the Privacy Implications of Direct-To-Consumer Genetic Testing. *New Genetics and Society*, 35(2):101–123, 2016.

[20] M. Claxton. Abbotsford News. Former Langley Libertarian candidate detained in Italy. https://bit.ly/2PUIQWC, 2017.

[21] E. Clayton, C. Halverson, N. Sathe, and B. Malin. A Systematic Literature Review of Individuals' Perspectives on Privacy and Genetic Information in the United States. *PLoS ONE*, 13(10), 2018.

[22] N. Couldry and J. Yu. Deconstructing Datafication's Brave New World. *New Media & Society*, 20(12):4473–4491, 2018.

[23] B. Darst, L. Madlensky, N. Schork, E. Topol, and C. S. Bloss. Perceptions of Genetic Counseling Services in Direct-To-Consumer Personal Genomic Testing. *Clinical genetics*, 84(4):335–339, 2013.

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

[25] M. De Choudhury and S. De. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In *ICWSM*, 2014.

[26] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *CEUR Workshop*, pages 86–95, 2017.

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

[28] DNARomance. Online Dating Based On Science. https://www.dnaromance.com/, 2018.

[29] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, 1996.

[30] J. Finkelstein, S. Zannettou, B. Bradlyn, and J. Blackburn. A Quantitative Approach to Understanding Online Antisemitism. *CoRR*, abs/1809.01644, 2018.

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

[32] A. Froelich. True Activist. This is What Americans Will Look like by 2050. https://bit.ly/2vpAIEH, 2014.

[33] B. Gambäck and U. K. Sikdar. Using Convolutional Neural Networks to Classify Hate-Speech. In *Workshop on Abusive Language Online*, 2017.

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

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

[36] GEDmatch. https://en.wikipedia.org/wiki/GEDmatch, 2019.

[37] Y. Gorodnichenko et al. Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection. National Bureau of Economic Research, 2018.

[38] E. M. Greytak, C. Moore, and S. L. Armentrout. Genetic Genealogy for Cold Case and Active Investigations. *Forensic Science International*, 2019.

[39] K. E. Hann, M. Freeman, L. Fraser, J. Waller, et al. Awareness, Knowledge, Perceptions, and Attitudes Towards Genetic Testing for Cancer Risk Among Ethnic Minority Groups: A Systematic Review. *BMC public health*, 17(1):503, 2017.

[40] A. Harmon. New York Times. Why White Supremacists Are Chugging Milk (and Why Geneticists Are Alarmed). https://nyti.ms/2Afg4Ho, 2018.

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

[42] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *SocInfo*, 2015.

[43] A. Kasunic and G. Kaufman. ”At Least the Pizzas You Make Are Hot”: Norms, Values, and Abrasive Humor on the Subreddit r/RoastMe. In *ICWSM*, 2018.

[44] A. Linskey. The Boston Globe. Warren Releases Results of DNA Test. https://bit.ly/2Chey99, 2018.

[45] S. Marche. The Guardian. Swallowing the Red Pill: A Journey to the Heart of Modern Misogyny. https://bit.ly/2Chey99, 2016.

[46] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*, 2013.

[47] R. A. Mills. Pop-up Political Advocacy Communities on Reddit.com: SandersForPresident and The Donald. *AI and Society*, 33(1):39–54, 2018.

[48] A. Mittos, J. Blackburn, and E. De Cristofaro. "23andMe Confirms: I'm Super White" Analyzing Twitter Discourse On Genetic Testing. *arXiv:1801.09946*, 2018.

[49] M. Molteni. Wired. The Creepy Genetics Behind the Golden State Killer Case. https://bit.ly/2HYECJE, 2018.

[50] V. Monga and B. L. Evans. Perceptual Image Hashing Via Feature Points: Performance Evaluation and Tradeoffs. *IEEE Transactions on Image Processing*, 2006.

[51] R. S. Mueller. Report On The Investigation Into Russian Interference In The 2016 Presidential Election. US Department of Justice, 2019.

[52] D. E. Nielsen, S. Shih, and A. El-Sohemy. Perceptions of Genetic Testing for Personalized Nutrition: A Randomized Trial of DNA-based Dietary Advice. *Lifestyle Genomics*, 7(2):94–104, 2014.

[53] NIH. What Is Genetic Ancestry Testing? https://ghr.nlm.nih.gov/primer/dtcgenetictesting/ancestrytesting, 2019.

[54] A. L. Nobles, C. N. Dreisbach, J. Keim-malpass, and L. E. Barnes. "Is This an STD? Please Help!" Online Information Seeking for Sexually Transmitted Diseases on Reddit. In *ICWSM*, 2018.

[55] A. Olteanu, C. Castillo, J. Boy, and K. R. Varshney. The Effect of Extremist Violence on Hateful Speech Online. In *ICWSM*, 2018.

[56] R. Ottoni, E. Cunha, G. Magno, P. Bernadina, W. Meira, and V. Almeida. Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination. In *WebSci*, 2018.

[57] A. Panofsky and J. Donovan. When Genetics Challenges a Racist's Identity: Genetic Ancestry Testing among White Nationalists. https://osf.io/preprints/socarxiv/7f9bc/, 2017.

[58] Perspective. https://www.perspectiveapi.com/, 2019.

[59] A. M. Phillips. Data on Direct-to-Consumer Genetic Testing and DNA Testing Companies. 10.5281/zenodo.1175800, 2018.

[60] Reddit. https://www.redditinc.com/press, 2019.

[61] E. Reeve. Vice News. White Nonsense. https://bit.ly/2DhP90h, 2016.

[62] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *NLPFrameworks*, 2010.

[63] D. Reich. New York Times. How Genetics Is Changing Our Understanding of 'Race'. https://nyti.ms/2pUxFOw, 2018.

[64] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. Meira. Characterizing and Detecting Hateful Users on Twitter. In *ICWSM*, 2018.

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

[66] W. D. Roth and B. Ivemark. Genetic Options : The Impact of Genetic Ancestry Testing on Consumers' Racial. *American Journal of Sociology*, 124(1):150–184, 2018.

[67] T. H. Saey. What I Actually Learned About My Family After Trying 5 DNA Ancestry Tests. https://bit.ly/2zaUIKy, 2018.

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

[69] D. Sims. The Battle Over Adult Swim's Alt-Right TV Show. https://bit.ly/2g06PPK, 2016.

[70] SoccerGenomics. Unlock The Player Within You. https://www.soccergenomics.com/, 2018.

[71] SPLC. Male Supremacy. https://www.splcenter.org/fighting-hate/extremist-files/ideology/male-supremacy, 2017.

[72] SPLC. Atomwaffen Division. https://www.splcenter.org/fighting-hate/extremist-files/group/atomwaffen-division, 2019.

[73] L. Stack. New York Times. Alt-Right, Alt-Left, Antifa: A Glossary of Extremist Language. https://nyti.ms/2uGOTV5, 2017.

[74] L. G. Stewart, A. Arif, and K. Starbird. Examining trolls and polarization with a retweet network. In *WSDM*, 2018.

[75] C. Welch and S. Ganim. CNN. White Supremacist Richard Spencer: 'We reached tens of millions of people' with video. https://cnn.it/2T7z5D8, 2016.

[76] Q. Wong. Facebook's Privacy Mishaps: Zuckerberg Could Be Held Accountable, Report Says. https://cnet.co/2VDJUlu, 2019.

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

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