"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, [†]Max-Planck-Institut für Informatik, [‡]Binghamton University
{a.mittos, e.decristofaro}@ucl.ac.uk, szannett@mpi-inf.mpg.de, jblackbu@binghamton.edu

Abstract

Progress in genomics has enabled the emergence of a booming 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, alt- and far-right groups have also taken an interest in genetic testing, using them to attack minorities and prove their genetic "purity." In this paper, we present a measurement study shedding light on how genetic testing is being discussed on Web communities in Reddit and 4chan. We collect 1.3M comments posted over 27 months on the two platforms, 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.

Introduction

Over the past decade, researchers have made tremendous progress toward understanding the human genome. 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 to reports of health risks, wellness, hereditary traits, etc. 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 (NIH 2019). AncestryDNA alone has tested more than 16M customers as of Jan 2020 (AncestryDNA 2020).

Alas, increased popularity of self-administered genetic tests has also been accompanied by media reports of far-right groups using it to attack minorities or prove their genetic "purity" (Reeve 2016), mirroring concerns of a new wave of scientific racism (Reich 2018). Also, statements from Donald Trump led Senator Warren to publicly confirm her Native American ancestry via genetic testing (Linskey 2018).

Interest in DTC genetic testing by right-wing communities comes at a time when racism, hate, and antisemitism on platforms like 4chan, Gab, and certain communities on Reddit is on the rise (Hine et al. 2017; Zannettou et al. 2020). Thus, these trends are particularly worrying, also considering how technology has been disrupting society in previously unconsidered ways (Gorodnichenko and others 2018); 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 (mis)used in online discussions. As genetics-based arguments for discrimination (BBC 2019), and even genocide (e.g., the Holocaust), have been made in the past, 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 (Panofsky and Donovan 2017; Roth and Ivemark 2018)), 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/or 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 Jun 30, 2016 to Mar 13, 2018, and 77K comments from Reddit related to genetic testing from Jan 1, 2016 to Mar 31, 2018, and analyze them along several axes to understand how genetic testing is being discussed online. We rely on natural language processing, computer vision, and machine learning tools, including (i) Latent Dirichlet Allocation (LDA) to identify topics of discussion, (ii) word embeddings to uncover words used in a similar context across datasets, (iii) Google's Perspective API (Perspective 2019) 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. Our image analysis on /pol/ shows the recurrent presence of popular alt-right personalities and "popular" 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. Reddit users are not uniformly interested in all aspects of genetic testing, rather, they form groups ranging from enthusiasts to people who use genetic keywords exclusively in subreddits that discuss fringe political views.

Related Work

Genetic Testing & Society. (Panofsky and Donovan 2017) 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. In follow-up work (Panofsky and Donovan 2019), they also look at the relationship between citizen science and white nationalists' use of genetic testing, shedding light on how "repair strategies" combine anti-scientific attacks on the legitimacy of these tests and reinterpretations of them in terms of white nationalist histories. (Mittos, Blackburn, and De Cristofaro 2018) conduct a study of the Twitter discourse on genetic testing, examining 300K tweets, 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.

(Chow-White et al. 2018) 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 2018) interview users to study how ancestry testing affects ethnic and racial identities, also finding instances of consumers not accepting test results and suggesting genetic ancestry testing may reinforce race privilege. (Clayton et al. 2018) 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. Finally, (Couldry and Yu 2018) 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 (Gymrek et al. 2013; Shringarpure and Bustamante 2015).

Overall, most of the research in this area mostly relies on qualitative studies examining the societal effects of genetic testing (Caulfield and McGuire 2012; Darst et al. 2013) and lacks quantitative large-scale measurements.

Online Hate. Researchers have also studied hate speech on mainstream social networks like Twitter (Silva et al. 2016; Mondal, Silva, and Benevenuto 2017; Davidson et al. 2017; Olteanu et al. 2018), Reddit (Olteanu et al. 2018), Facebook (Ben-David and Matamoros-Fernandez 2016), YouTube (Ottoni et al. 2018), and Instagram (Hosseinmardi et al. 2015). Closer to our work is research on fringe communities in 4chan and Reddit. Specifically, (Bernstein et al. 2011) study 5M posts on the random (/b/) board to examine how anonymity and ephemerality work in 4chan, while (Hine et al. 2017) focus on /pol/, 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. Then, (Zannettou et al. 2018) detect and study racist and hateful memes, and their propagation, on 4chan, Gab, Reddit, and Twitter. (Zannettou et al. 2020) 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. (Chandrasekharan et al. 2017) 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. Overall, 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 2014) look Reddit conversations about mental health, while other studies analyze how users behave in specific subreddits. (Nobles et al. 2018) study /r/STD to understand how users seek health information on sensitive and stigmatized topics, using 1.8K posts from 1.5K users. Another line of work has studied the /r/The_Donald subreddit. (Flores-Saviaga, Keegan, and Savage 2018) 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.

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

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

Dataset

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 (Phillips 2018). 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).

Reddit Dataset. Reddit is a social news aggregation and discussion website, where users post content which gets voted up or down by other users. Users can 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 Jan 2020, Reddit has more than 430M monthly active users and 21B visits.

We gather all Reddit comments from Jan 1, 2016 to Mar 31, 2018 (2B comments in 473K subreddits) via the monthly releases from pushshift.io. 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 Jan 2020, there are 80 of them), with various topics of interest (e.g., video games, literature, 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 (Hine et al. 2017). 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 Jun 30, 2016 to Mar 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 new branch of answers. 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.

Remarks. We look at Reddit and 4chan's politically incorrect board (/pol/) as opposed to mainstream platforms (e.g., Facebook or Twitter) as we are interested in the hateful and racist connotations of genetic testing discourse, and these platforms have been previously found to host far-right ideologies (Hine et al. 2017; Olteanu et al. 2018). Finally, note that our study was approved by the ethics committee at UCL.

Genetic Testing Discussions on Reddit

Methodology

Subreddits selection & grouping. We extract all the subreddits where genetic testing comments have been posted to, but discard subreddits if they either have fewer than 1K comments overall or fewer than 100 comments with one of the keywords. This yields a list of 114 subreddits; due to space limitations, we do not report the complete list, however, it is available in the full version of the paper (Mittos et al. 2019), along with the number of comments we extract from them.

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 and thematic tags, however, tags are available only for very popular subreddits. 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. 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 Fig. 1.

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%. (The number of genetic testing comments in each subreddit, as well as the total number of comments, is also reported in the full version of the paper.) 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%). 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, us-

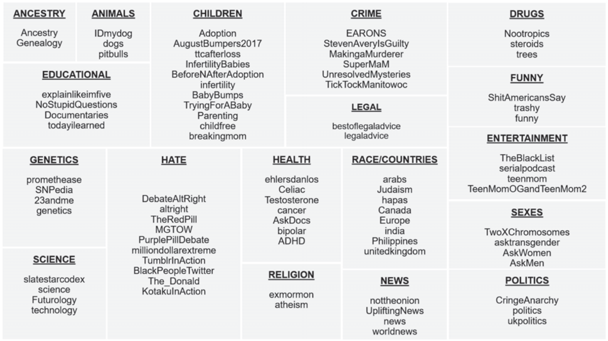
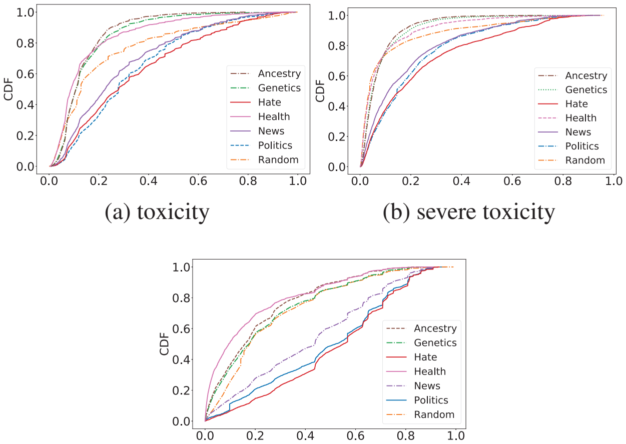


Figure 1: Subreddits with genetic testing related comments, grouped into categories based on their thematic topics.



(a) toxicity

(b) severe toxicity

(c) inflammatory

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

ing LDA and Google's Perspective API (Perspective 2019). 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) Inflammatory, which focuses on texts intending to provoke or inflame. In Fig. 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 (Davidson et al. 2017), are primarily trained on short texts with a limited number of training samples. Whereas, our datasets contain lengthy comments; 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.

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 (Stack 2017) (e.g., /r/altright, /r/DebateAltRight, and /r/The_Donald), sexism, or racism. For instance, /r/TheRedPill includes misogyny and toxic behavior towards women (Marche 2016), while /r/MGTOW, Men Going Their Own Way, is a forum for men who reject romantic relationships with women. 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 Fig. 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 2. 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 other subreddits, here they are being expressed in a much more toxic/inflammatory manner; as evidenced by Fig. 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., 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 Table 3). Again, the conversations exhibit clear racist connotations; e.g., 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!"

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., "[...] 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 (SPLC 2019)). Some users explicitly advocate using genetic testing to elim-

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 2: LDA analysis of the Hate subreddits.

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)
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 3: LDA analysis of the Genetics and Ancestry subreddits.

inate groups of non-white ancestry (e.g., "You know with pre-implantation genetic testing we can breed out non-white ancestry fairly easily [...]").

Category Analysis

Next, we select a few categories of subreddits and analyze them further, aiming to better understand how users perceive genetic testing in each context.

Genetics & Ancestry. As mentioned, the subreddits with the highest ratio of genetic testing keywords are directly related to genetic testing and ancestry. This is confirmed by LDA (see Table 3). 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 (GEDmatch 2019), is one of the topics with the greatest weights (0.054); see Table 3. 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 (Greytak, Moore, and Armentrout 2019). Overall, as shown in Fig. 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. Genetic testing appears to be discussed 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 (Molteni 2018). We also find subreddits covering (often controversial) discussions about Steven Avery, who was wrongly convicted of sexual assault and attempted murder; 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 and evidence (e.g., blood, sample, evidence); see Table 4. 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 4), 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).

Animals. Reddit users also use genetic testing keywords in subreddits related to animals, and more specifically those related to dogs (we omit the results due to space constraints).

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), as well as religion (e.g., /r/exmormon). Finally, 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.,

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)
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 4: LDA analysis of the Crime and Children subreddits.

/r/AskMen, /r/AskWomen).

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 (Blondel et al. 2008), which provides a set of communities based on the graph's structure.

Fig. 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/ancestry, /r/23andMe) are part of the same community (pink nodes) as scientific and education ones (e.g., /r/askscience, /r/science), 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, 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, /r/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).

Take-Aways. Our Reddit analysis shows that genetic testing is discussed in a variety of contexts which in itself is an indicator of how mainstream it has become. For instance, users discuss it in the context of issues related to their children, pets, or health, or to debate on their cultural heritage. More interestingly, they 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 focus 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 worrying since Reddit is a mainstream platform (5th most visited site in the US).

Genetic Testing Discussions on /pol/

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, the one with the oldest last post is purged. There is also a "bump" limit that prevents a thread from never being purged. As per (Hine et al. 2017), 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 Fig. 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 indicates that genetic testing is often discussed in long-lasting/interesting threads and may attract more attention by users. We also run a two-sample KS test on the distributions and we reject the null hypothesis that

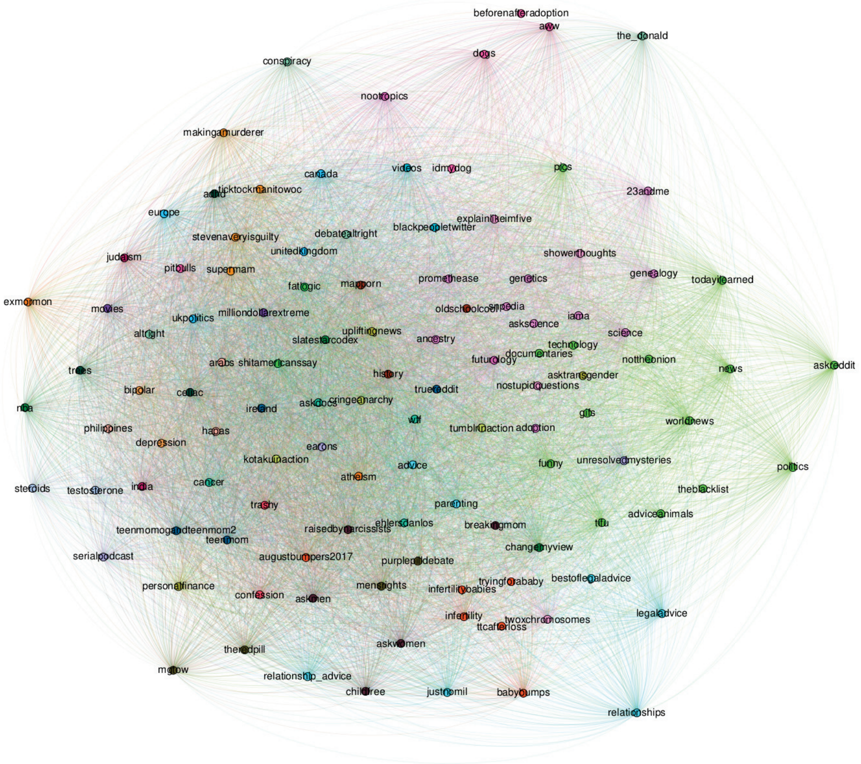


Figure 3: Graph depicting the Jaccard Index of the users whose comments include genetic testing keywords for each subreddit (see (fig 2019a) for an interactive version).

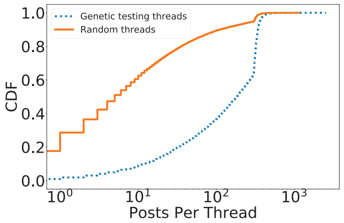


Figure 4: CDF comparing 4chan threads with genetic testing keywords and random threads in terms of number of posts.

they come from a common parent distribution ($p < 0.01$).

Toxicity & Hate. We then 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 (Hine et al. 2017); 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 its high level of hate speech (Hine et al. 2017)—thus, we omit related plots due to space limitations.

Topic Modeling. We also use LDA modeling to identify the most prominent topics of discussion; see Table 5. Similar to Reddit, 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 (e.g., lauren, jewish, youtube), an Internet personality associated with the alt-right, whose popularity rose after being detained in Italy for trying to block a ship rescuing refugees (Claxton 2017). 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).

Image Analysis

Next, we look at the images and memes that are shared in /pol/ posts including genetic testing keywords. We use the image analysis pipeline introduced in (Zannettou et al. 2018) which uses Perceptual Hashing (Monga and Evans 2006) and clustering techniques 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. We obtain 215 clusters including 543 total images; the other 5,832 images

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 5: 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 6: Top 20 entities with the most clusters.

are labeled as noise by the clustering algorithm and thus we discard them. This high noise ratio mirrors findings in (Zannettou et al. 2018) and is likely due to 4chan users creating a lot of original content (Hine et al. 2017).

We annotate each cluster using Google's Cloud Vision API[1], specifically, we calculate the medoid of each cluster (i.e., its "representative" image) following the methodology by (Zannettou et al. 2018), 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 6. The most popular entries are /pol/ itself and Lauren Southern with 6.9% of all clusters. The latter is 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 (Welch and Ganim 2016) (6th most popular with 2.3% of all clusters), and Carl Benjamin, a YouTuber known for his misogynistic involvement in the GamerGate controversy (Bish 2016), are also popular. We also find 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 (ADL 2019), 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 5. 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 6 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 7) 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 7) reveal discussions on the ancestral background of the American population (e.g., americans, ancestry, african, whites).

Cluster visualization. Finally, we provide a visualization of the clusters in Fig. 5. Nodes in the graph represent clusters, while edges represent the Jaccard Index between clusters (as per the entities returned by the Cloud Vision API). To ease presentation, we only consider edges where the Jaccard Index is greater than 0.2, a threshold we select after inspecting the distribution of all the Jaccard Index scores. 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 (Blondel et al. 2008). 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 Fig. 5, 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

[1] 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)
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 7: LDA analysis of the texts in the /pol/ posts with imagery annotated as '23andMe' or 'United Stated of America'.

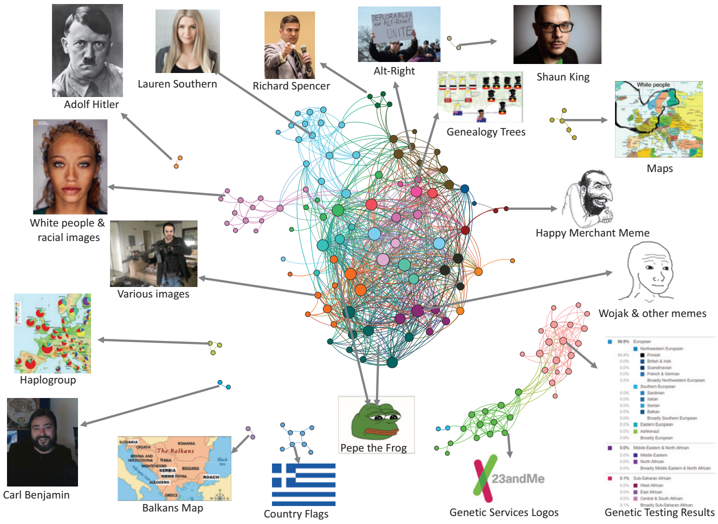


Figure 5: Visualization of the image clusters with manual annotation (see (fig 2019b) for an interactive version).

cluster including an image from National Geographic predicting how the average American woman will look like in 2050 (Froelich 2014), 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 (Zannettou et al. 2020).

Take-aways. Overall, we find that genetic testing is a rather popular topic of discussion in 4chan's /pol/, often 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 very worrying. We find imagery with prominent figures of the alt-right movement (e.g., Lauren Southern, Richard Spencer), antisemitic memes (e.g., Happy Merchant), and topics of discussion using words with racial/hateful meaning (e.g., jewish, nigger), which may be an indicator that groups adjacent to the alt-right are using genetic testing to bolster their ideology.

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. 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 (Mikolov et al. 2013). 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 allows us to study which words are used in similar contexts.

Methodology. We train a separate word2vec model, as per the implementation provided by (Řehůřek and Sojka 2010), for each of the 19 groups of subreddits (see Fig. 1) and 4chan's /pol/, using all of the posts made between Jan 1, 2016 and Mar 31, 2018, and Jun 30, 2016 and Mar 13, 2018, respectively. We pre-process 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

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

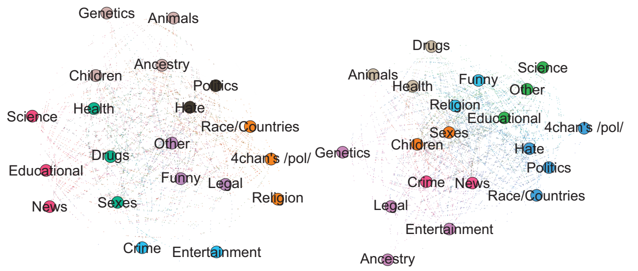
Table 8: Words that are in the vocabulary of the word2vec models trained for each group of subreddits and /pol/.

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 8 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").

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 Fig. 6), 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 (Blondel et al. 2008). When using the genetic testing keywords as seeds (Fig. 6(a)), we find that communities about genetics, ancestry, animals, and 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. Fig. 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 (Zannettou et al. 2020).



(a) Genetic Testing Keywords (b) Selected Keywords

Figure 6: Graph representation of the word2vec models.

When using the set of hand-picked seed words (Fig. 6(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.

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 (Borrelli 2018). 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 (Panofsky and Donovan 2017; Roth and Ivemark 2018), 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 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 (Stewart, Arif, and Starbird 2018) and even instigate public health crises (Broniatowski et al. 2018). In this context, recent efforts have been made by law enforcement to understand and address such campaigns (Mueller 2019). 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 (Panofsky and Donovan 2017; Roth and Ivemark 2018) 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 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 (Wong 2019); 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., (Davidson et al. 2017; Del Vigna et al. 2017; Djuric et al. 2015; Gambäck and Sikdar 2017)) 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 acknowledge the support of the NVIDIA Corporation, for the donation of two Titan Xp GPUs.

References

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

AncestryDNA. 2020. Ancestry Company Facts. https://www.ancestry.com/corporate/about-ancestry/company-facts.

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

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

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

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

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast Unfolding of Communities in Large Networks. *JSTAT*.

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

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

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

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. In *CSCW*.

Chow-White, P.; Struve, S.; Lusoli, A.; Lesage, F.; Saraf, N.; and Oldring, A. 2018. '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.

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

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

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

Darst, B.; Madlensky, L.; Schork, N.; Topol, E.; and Bloss, C. S. 2013. Perceptions of Genetic Counseling Services in Direct-To-Consumer Personal Genomic Testing. *Clinical genetics*.

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

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

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

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

2019a. Jaccard Index Subreddit Visualization. https://emilianodc.com/PAPERS/genetic-racism/index.html#fig3.gexf.

2019b. Visualization of the 4chan Image Clusters. https://emilianodc.com/PAPERS/genetic-racism/index.html#fig6.gexf.

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

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

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

GEDmatch. 2019. https://en.wikipedia.org/wiki/GEDmatch.

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

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

Gymrek, M.; McGuire, A. L.; Golan, D.; Halperin, E.; and Erlich, Y. 2013. Identifying Personal Genomes by Surname Inference. *Science* 339(6117):321–324.

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

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

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

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

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

Mittos, A.; Zannettou, S.; Blackburn, J.; and De Cristofaro, E. 2019. "And We Will Fight For Our Race!'" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan. *arXiv preprint arXiv:1901.09735*.

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

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

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

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

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

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

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

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

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

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

Panofsky, A., and Donovan, J. 2019. Genetic ancestry testing among white nationalists: From identity repair to citizen science. *Social studies of science*.

Perspective. 2019. https://www.perspectiveapi.com/.

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

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

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

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

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

Shringarpure, S. S., and Bustamante, C. D. 2015. Privacy Risks from Genomic Data-Sharing Beacons. *The American Journal of Human Genetics*.

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

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

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

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

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

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

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

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *ICWSM*.