Analyzing Genetic Testing Discourse on the Web Through the Lens of Twitter, Reddit, and 4chan

ALEXANDROS MITTOS, University College London
SAVVAS ZANNETTOU, Max Planck Institute for Informatics
JEREMY BLACKBURN, Binghamton University
EMILIANO DE CRISTOFARO, University College London

Recent progress in genomics has enabled the emergence of a flourishing market for direct-to-consumer (DTC) genetic testing. Companies like 23andMe and AncestryDNA provide affordable health, genealogy, and ancestry reports, and have already tested tens of millions of customers. Consequently, news, experiences, and views on genetic testing are increasingly shared and discussed on social media. At the same time, 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 set to study the genetic testing discourse on a number of mainstream and fringe Web communities. We do so in two steps. First, we conduct an exploratory, large-scale analysis of the genetic testing discourse on a mainstream social network such as Twitter. We find that the genetic testing discourse is fueled by accounts that appear to be interested in digital health and technology. However, we also identify tweets with highly racist connotations. This motivates us to explore the connection between genetic testing and racism on platforms with a reputation for toxicity, namely, Reddit and 4chan, where we find that discussions around genetic testing often include highly toxic language expressed through hateful and racist comments. In particular, on 4chan's politically incorrect board (/pol/), content from genetic testing conversations involves several alt-right personalities and openly anti-semitic rhetoric, often conveyed through memes.

1 INTRODUCTION

The sequencing of the first human genome was completed in 2003 after 13 years and an investment of approximately $3 billion [54]. Today, due to breakthroughs in biology and bioinformatics, we can *fully* sequence genomes for less than $1,500 [63] and genotype it (i.e., look for specific markers) for even less [48]. This rapid progress in genomics is paving the way to *personalized medicine*, a concept advocating for diagnosis and treatment to be tailored to patients' genetic features, aiming to make healthcare more preventive and effective. It also enables *public initiatives* to sequence large numbers of genomes and build large bio-repositories for research purposes; for instance, the All Of Us research program in the US [65] and the Genomics England project [43] in the UK are sequencing the genomes of, respectively, 1M and 100K volunteers.

Moreover, a number of companies have successfully marketed *direct-to-consumer (DTC)* genetic testing. Using this service, 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 [33] or identification of athletic skills [91] to reports of health risks (e.g, likelihood of developing Parkinson's), wellness (e.g., lactose intolerance), carrier status (e.g., hereditary hearing loss), and traits (e.g., eyes color). Another popular service includes 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 [66]. As of July 2019 AncestryDNA alone has tested more than 15M customers [4]. However, these tests are reportedly subject to limitations, e.g., results differ from provider to provider due to different control groups [87].

Affordable DTC products and participatory sequencing initiatives make genetic testing increasingly more accessible and available to the general population. Like with other aspects of digital

Authors' addresses: Alexandros Mittos, a.mittos@ucl.ac.uk, University College London; Savvas Zannettou, szannett@mpi-inf.mpg.de, Max Planck Institute for Informatics; Jeremy Blackburn, jblackbu@binghamton.edu, Binghamton University; Emiliano De Cristofaro, e.decristofaro@ucl.ac.uk, University College London.

health, this leads to social media attracting discussions, sharing of experiences, and molding of perceptions around genetic testing, thus, becoming a key platform for related news and marketing efforts. However, while the research community has analyzed in great detail the interlinked relationship between health and social networks such as Twitter and Reddit, to the best of our knowledge, genetic testing discourse on social media has not been adequately studied.

Furthermore, 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" [12, 80] mirroring concerns of a new wave of scientific racism [82], e.g., white nationalists were recently taped chugging milk at gatherings to demonstrate the ability of white people to better digest lactose [47].

Research Questions. With this motivation, we set to answer the following research questions:

RQ1) Progress in genomics is bringing a revolutionary technology like genetic testing to the masses, along with its potential of improving lives [5]. However, is this potential impact being reflected on *mainstream* social networks, and if so, how? Specifically, what is the genetic testing discourse about? Can we identify certain keywords, themes, or companies dominating the discussion? What can we discern about the online narrative on genetic testing? What is the overall sentiment on genetic testing? Do users express positive or negative emotions when they discuss genetic testing?

RQ2) Who posts about genetic testing? Which users are particularly active and what do they talk about? Are there particular groups of users with an interest in genetic testing's success?

RQ3) Genetic testing is often a controversial topic, frequently associated with racism [58] and ethical concerns [77]. Are these controversies echoed on *fringe* social networks? Are there instances of users discussing genetic testing in the context of racist or hateful ideologies? If so, can we identify certain communities where this behavior is systematic?

Technical Roadmap. First, we conduct an exploratory, large-scale analysis of the genetic testing discourse on a mainstream social network such as Twitter. Using 10 keywords related to DTC genetics companies and 3 to genomics initiatives, we search and crawl all available tweets containing these keywords posted between January 1, 2015 and July 31, 2017. We collect 302K tweets from 113K users and analyze them content-wise, studying the most common hashtags/URLs and measuring sentiment. We also examine the most negative tweets in our dataset, finding a number of tweets related to racism and hate speech.

This leads us to the second part of our work, where we explore the connection between genetic testing and racism. To do so, considering that Twitter employs a strict hateful conduct policy that results in the deletion of tweets related to racism [97], we move our attention to two platforms that are only loosely moderated and are often associated with high degrees of toxicity: Reddit and 4chan. Furthermore, we expand our set of keywords by an order of magnitude, aiming to capture as many hateful instances as possible. More specifically, we collect 77K comments from Reddit related to genetic testing from January 1, 2016 to March 31, 2018, and 7K threads from the politically incorrect (/pol/) board of 4chan (consisting of 1.3M posts) from June 30, 2016 to March 13, 2018. Our analysis focuses on instances of hateful and toxic speech, using natural language processing and machine learning tools, including (i) Latent Dirichlet Allocation (LDA) [10] to identify topics of discussion, (ii) word embeddings [59] to uncover words used in a similar context across datasets, (iii) Google's Perspective API [74] to measure toxicity in texts, and (iv) Perceptual Hashing to assess the imagery and memes shared in posts.

Main Findings. Overall, we present a comprehensive overview of the DTC genetic testing discourse online, starting from an exploratory analysis of discussions and then moving onto racism and toxic conversations. Our main findings include:

(1) ***Topics.*** Genetic testing on Reddit is being discussed in a wide variety of contexts, while, on /pol/, the analysis of images highlights the recurrent presence of popular alt-right personalities and "popular" anti-semitic memes. By contrast, on Twitter, users tweeting about genetic testing seem overall interested in digital health and technology, with coverage boosted by mainstream news (e.g., 23andMe's issues with the FDA [35]) and announcements (e.g. President Obama announcing the All of Us program [42]). Overall, the topic of genetic testing appears to be popular on all three platforms.
(2) ***Users.*** Reddit users are not uniformly interested in all aspects of genetic testing, rather, they form groups ranging from enthusiasts (e.g., those who are interested in or have undergone genetic testing), to people who use genetic keywords exclusively in subreddits that discuss fringe political views. Whereas, on Twitter, the conversation is often dominated by those with a vested interest in its success, e.g., journalists, medical professionals, entrepreneurs, etc. On the latter, we also find evidence of large marketing efforts undertaken by different companies, which naturally influence the type and the nature of users' engagement; e.g., we find promotional hashtags (e.g., #sweepstakes) in 1 out 8 tweets containing the keyword AncestryDNA.
(3) ***Racism.*** On /pol/ and Reddit, genetic testing is often associated with hateful, racist, and sexist content. Discussion is highly toxic, with users suggesting that genetic testing can be used to marginalize or even *eliminate* minorities. Word embeddings analysis also reveals that certain subreddits use ethnic terms in conjunction with genetic testing keywords in the same way as /pol/, which may suggest that fringe ideologies are spilling out to more mainstream communities. In fact, even though sentiment on Twitter is mostly positive around public initiatives and neutral for DTC companies, we do find evidence of groups using genetic testing in connection to racist and anti-semitic agendas on Twitter as well.

Paper Organization. The rest of the paper is organized as follows. In Section 2, we review prior work, then, in Section 3, we describe the methodology used to obtain our datasets. In Section 4, we present the results of an exploratory analysis around the genetic testing discourse on Twitter, while in Section 5, we study the connection between genetic testing and racism focusing on Reddit and 4chan's /pol/. Finally, Section 6 concludes the paper with a broader discussion of our findings.

2 RELATED WORK

We now describe relevant prior work, reporting on the following areas: 1) societal effects of genetic testing; 2) health on social networks; 3) hate speech on social media; and 4) exploratory studies on Twitter, Reddit, and 4chan.

2.1 Genetic Testing & Society

Roth and Ivemark [86] interview users to 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 that genetic ancestry testing may reinforce race privilege. Clayton et al. [25] 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. Also, Couldry and Yu [27] 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 [44, 88].

Panofsky and Donovan [70] 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 or 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 [71], 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. Finally, Chow-White et al. [22] 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.

Overall, most of the research in this area relies on qualitative studies examining the societal effects of genetic testing [16, 23, 28, 45, 64], and somewhat lacks quantitative large-scale analysis. To the best of our knowledge ours is the largest, quantitative measurement study spanning three social networks, namely, Twitter, 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.

2.2 Health in social networks

Twitter has been extensively used to study health and health-related issues, e.g., to measure and predict depression. De Choudhury et al. [31] identify 476 users self-reporting depression, collect their tweets, and study their engagement, emotion, and use of depressive language. By comparing to a control group, they extract significant differences, and build a classifier to predict the likelihood of an individual's depression. Coppersmith et al. [26] study tweets related to various mental disorders, while Paul et al. [73] gather public health information from Twitter, discovering statistically significant correlations between Twitter and official health statistics.

Abbar et al. [2] analyze the nutritional behavior of US citizens: they collect 892K tweets by 400K US users using food-related keywords and find that foods match obesity and diabetes statistics, and that Twitter friends tend to share the same preferences in food consumption. Prasetyo et al. [76] study how social media can effect awareness in health campaigns. Focusing on the Movember charity campaign, they collect more than 1M tweets, using the keyword 'Movember', and uncover correlations between the visitors of the Movember website and popular Twitter users, but none between tweets and donations. Finally, Cavazos-Rehg et al. [17] study drinking behaviors on Twitter: using keywords related to drinking (e.g., drunk, alcohol, wasted), they collect 10M tweets and identify the most common themes related to pro-drinking and anti-drinking behavior.

2.3 Quantitative Studies of Twitter, Reddit, and 4chan

Researchers have extensively studied social networks like Twitter, Reddit, and 4chan, producing a number of exploratory studies of discourse and overall user behavior.

Twitter. Lerman et al. [55] conduct an emotion analysis on tweets from Los Angeles: using public demographic data, they find that users with lower income and education levels, and who engage with less diverse social contacts, express more negative emotions, while people with higher income and education levels post more positive messages. Chatzakou et al. [20] study the GamerGate controversy[1] on Twitter, collecting a dataset of tweets containing keywords indicating abusive

[1] https://en.wikipedia.org/wiki/Gamergate_controversy

behavior. They compare the characteristics of the related Twitter profiles to a baseline, finding that users tweeting about GamerGate are more technologically savvy and active, and that their tweets are more negative. Burnap et al. [15] study Twitter responses to a terrorist attack occurred in Woolwich in 2013. Using 'Woolwich' as a keyword search, they collect 427K tweets, finding that opinions and emotional factors are predictive of size and survival of information flows.

Reddit. De Choudhury and De [30] look at Reddit conversations about mental health, aiming to understand language attributes of online self-disclosure and factors driving support in online posts. They show that 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 et al. [52] 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. [67] study /r/STD to understand how users seek health information on sensitive and stigmatized topics. 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.

Flores-Saviaga et al. [38] 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 [60] 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.

4chan. Bernstein et al. [8] study 5M posts on 4chan's random board (/b/) to examine anonymity and ephemerality on 4chan. They find that most threads expire in less than 5 minutes, while over 90% of the posts are anonymous. Hine et al. [49] study 8M posts from /pol/ 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 of.

Overall, Twitter, Reddit, and 4chan have been analyzed along several axes focusing on a multitude of topics, however, to the best of our knowledge, our work is the first large-scale, multi-platform quantitative study on genetic testing discourse.

2.4 Online Hate

Researchers have also studied hate speech on mainstream social networks like Twitter [29, 68, 83, 89], Reddit [19, 68], Facebook [7, 32], YouTube [69, 84], and Instagram [50]. Chatzakou et al. [21] explore cyberbullying on Twitter. They rely on a dataset of 1.6M tweets to study the properties that characterize bullying, and build a machine learning classifier which identifies users exhibiting aggressive behavior. Founta et al. [39] focus on abuse detection. They leverage a deep learning architecture which takes into consideration various features (e.g., metadata of posts, prior posts, account settings, social network, popularity) and propose a tool that is able to capture several facets of abusive behavior, i.e., cyberbullying, hateful and offensive content, and sarcasm.

Zannettou et al. [102] explore how mainstream and fringe online communities on Twitter, Reddit, and 4chan influence each other with respect to disinformation and hateful propaganda. Then, Zannettou et al. [101] detect and study racist and hateful memes, and their propagation, on 4chan, Gab, Reddit, and Twitter. Among other things, they find that racist memes are very common on /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. Then, in [37], Zannettou et al. study anti-semitism on /pol/ and Gab—a Twitter clone known for attracting users who are banned from Twitter—revealing that anti-semitic 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 anti-semitic content.

Chandrasekharan et al. [18] study how Reddit's decision to ban 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, accurately identifying hateful content remains an open problem due to being largely context-dependent.

3 DATASETS

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

3.1 Twitter Dataset

Twitter is a social networking and microblogging service where users can post short messages known as "tweets." These tweets can, in return, be retweeted and liked by other users. As of February 2020, Twitter has 330 million active users [103].

Genetic Testing Keywords. We start from a list of 36 DTC genetic testing companies compiled by the International Society of Genetic Genealogy[2]. We use each company's name as a search keyword; if the search returns less than 1,000 tweets, we discard it. In the end, we collect tweets for 10 companies: 23andMe, AncestryDNA, Counsyl, DNAFit, FamilyTreeDNA, FitnessGenes, MapMyGenome, PathwayGenomics, Ubiome, and VeritasGenetics. We opt for keywords not separated by spaces (e.g., VeritasGenetics) rather than quoted search (e.g., "Veritas Genetics") since we notice that companies are primarily discussed via hashtags or mentions, and because Twitter's search engine, at the time of the collection, did not provide exact results with quotes.[3]

Besides tweets related to for-profit companies, we also want to study discourse related to public sequencing initiatives and related concepts. Thus, we select three more keywords: PrecisionMedicine, PersonalizedMedicine, and GenomicsEngland. Personalized Medicine aims to make diagnosis, treatment, and care of patients tailored and optimized to their specific genetic makeup. Precision Medicine conveys a similar concept, but also refers to the initiative sequencing the genome of 1M individuals announced by President Obama in 2015 to understand how a person's genetics, environment, and lifestyle can help determine the best approach to prevent or treat disease [42]. Genomics England is a similar UK initiative with 100K volunteers, primarily focusing on cancer and rare disease research. Once again, we search for keywords not separated by spaces (e.g., PrecisionMedicine) since these concepts are mostly discussed via hashtags and because of the incorrectness of the search engine.

Dataset. We use a Python library called Tweepy that is a wrapper of Twitter's official API to collect all posted tweets from January 1, 2015 to July 31, 2017 returned as search results using the 10 DTC keywords and the 3 keywords related to genomics initiatives. Our script collects, for each

[2]https://isogg.org/wiki/List_of_DNA_testing_companies

[3]For instance, when searching for tweets using the quoted search "Family Tree DNA," we expect to only get the tweets that include these exact words, in that order, including the spaces. However, we notice that the following tweet appears: "Celebrate Valentine's Day & give the gift of Family Finder for only $59. Sale ends February 14th! http://familytreedna.com"

	Tweets	Users	RTs	Likes	Official	URLs	Top 1M
23andMe	132,597	64,014	72,848	149,897	1.31%	68.68%	75.40%
AncestryDNA	29,071	16,905	16,266	47,249	7.08%	75.50%	49.68%
Counsyl	3,862	1,834	2,716	4,255	3.49%	83.94%	74.97%
DNAFit	2,118	844	1,336	2,508	15.34%	78.94%	79.18%
FamilyTreeDNA	2,794	1,205	1,196	3,111	4.36%	36.47%	69.21%
FitnessGenes	2,142	773	908	2,809	16.29%	56.76%	71.28%
MapMyGenome	1,568	704	4,488	3,726	15.30%	80.35%	64.30%
PathwayGenomics	1,544	579	1,968	2,521	2.13%	76.55%	68.12%
Ubiome	14,420	6,762	9,223	13,991	2.71%	73.28%	64.19%
VeritasGenetics	1,292	497	1,443	2,526	6.65%	58.28%	71.95%
Genomics England	7,009	1,863	19,772	18,756	19.68%	69.18%	48.82%
Personalized Medicine	20,302	4,631	19,085	15,514	–	87.42%	71.98%
Precision Medicine	83,329	13,012	118,043	128,303	–	83.39%	77.16%
Total	302,048	113,624	269,292	395,166	2.26%	74.77%	71.80%
Baseline	163,260	131,712	282,063,006	486,960,753	–	45.49%	89.57%

Table 1. Overview of the Twitter dataset.

tweet, its content, the username, date and time, the number of retweets and likes, as well as the URL of the tweet. It also visits the profile of the users posting each tweet, collecting their location (if any), the number of followers, following, tweets, and likes. Overall, we collect a total of 191K tweets from 94K users for the 10 DTC companies and 111K from 19K users for the 3 initiatives, as summarized in Table 1. Note that to download this information one needs the prior consent of Twitter. When applying for an API key, Twitter asks the applicant to explain their reasons for doing so. We describe in our form our intended methodology and our application is accepted. We also collect a set of 163,260 random English tweets, from the same January 2015 to July 2017 period (approx. 170 per day), which serves as a baseline set for comparisons. This set originates from a pre-existing dataset of tweets collected using Twitter's "Sample Stream" API which returns a random sample of all public tweets.[4]

We remark that the keyword search returns accounts that match that keyword, e.g., tweets including 23andMe, #23andMe, or @23andMe, but also those posted by the @23andMe account. For consistency, we discard the latter, analyzing them separately when relevant. Note that our dataset includes tweets from users who discuss their opinions on genetic testing, but also blog posts, ads, and news articles. As our goal is to discover how genetic testing is reflected through the lens of Twitter, we choose not to discard any of the above subsets in an attempt to "clean" the dataset, or to focus only on certain kinds of profiles. We also make our Twitter dataset available to the public.[5]

3.2 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, called "subreddits," which are usually associated with areas of interest (e.g., movies, sports, politics). As of February 2020, Reddit has more than 430M monthly active users, receives

[4]See https://developer.twitter.com/en/docs/tweets/sample-realtime/
[5]https://github.com/amittos/genetic-testing-twitter-dataset

Reddit	Genetic Testing	Random
Comments	77,184	204,713
Subreddits	3,734	12,616
Users	48,096	165,127

Table 2. Overview of the Reddit dataset.

4chan	Genetic Testing	Random
Threads	6,986	19,530
Posts	1,306,671	760,691
Posts/Thread (Mean)	186.5	37.9
Posts/Thread (Median)	183	5
Images	338,540	206,830

Table 3. Overview of the 4chan dataset.

21B monthly visits, and consists of more than 1.8M subreddits [1]. This makes it the fifth most visited site in the US [78].

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 [75]. 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 [41]).

Dataset. We gather all Reddit comments from January 1, 2016 to March 31, 2018 (2B comments in 473K subreddits) via the publicly available monthly releases of pushshift.io.[6] 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 2. For comparison, we obtain a 0.0001% subset of all Reddit comments posted between January 1, 2016 to March 31, 2018. To do so, we use a simple Python script that randomly selects 0.0001% of all comments for each month of the same time period (recall that all Reddit comments are publicly available by pushshift.io). This results in a set of 204K random comments unrelated to genetic testing.

3.3 4chan Dataset

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 July 2019, there are 70 of them), with various topics of interest (e.g., 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 the politically incorrect board (/pol/), which has been shown to include a high volume of racist, xenophobic, and hateful content [49]. We choose /pol/ to examine how genetic testing is being discussed in communities that have been associated with alt-right ideologies.

Dataset. To create our genetic testing dataset we turn to the publicly available 4chan dataset by Papasavva et al. [72]. We collect 1.9M threads posted on /pol/ from June 30, 2016 to March 13, 2018. We employ the list of genetic testing keywords we used for the Reddit dataset 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 flat, single-threaded entity where the OP submits an image (perhaps with text) which other users then respond to. 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. Also, titles are optional for 4chan threads and not regularly used, thus, it is difficult to understand the context

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

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 3, 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).

3.4 Important Remarks

We note that the methodologies invoked for collecting and studying the three datasets differ significantly. This is unavoidable as the three platforms possess unique properties (e.g., length of text, use of images, language tropes, anonymity of users) that drastically change the complexion of their data. While we do compare the parts of the datasets where appropriate, we note that our aim is *not* a direct comparison of online platforms on the topic genetic testing. Rather, the datasets are complimentary to one another and are used to provide a comprehensive overview of different aspects of genetic testing discourse from a quantitative point of view.

Ethics. Our study was approved by the ethics committee at University College London. Furthermore, note that the content posted on 4chan is anonymous and we make no attempt to de-anonymize users. Overall, we follow standard ethical guidelines [85].

4 ANALYZING THE GENETIC TESTING DISCOURSE: A TWITTER STUDY

In this section, we present the results of an exploratory, large-scale analysis of the genetic testing discourse online using Twitter. To do so, we study the dataset presented in Section 3.1 on multiple axes, such as the most used hashtags and URLs, the sentiment, and the nature of the users who tweet about genetic testing.

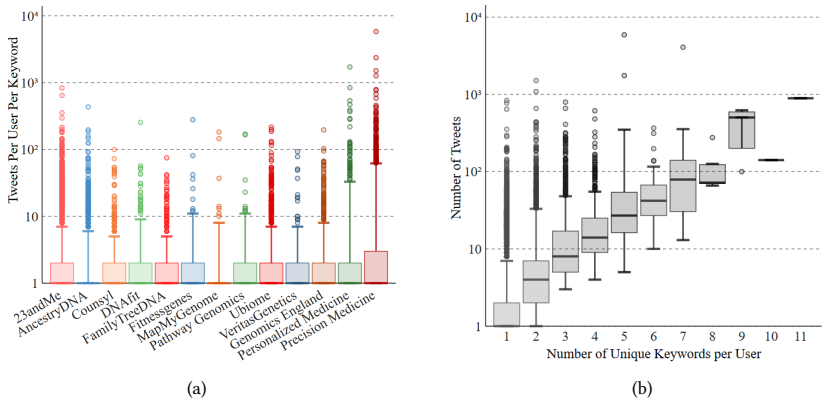
4.1 General Characterization

We start by presenting a general characterization of the tweets in our dataset. Simple statistics of our keyword-based dataset are reported in Table 1. From left to right, the table lists the total number of tweets, unique users, retweets, and likes for each of the 13 keywords and the random baseline. We also quantify the percentage of tweets made by the official accounts of each company or initiative, as well as the percentage of tweets including media (images and videos), quoted tweets, hashtags, and URLs, and how many of them are in the Alexa Top 1M[7].

Number of tweets. 23andMe is by far the most popular keyword, with one order of magnitude more tweets than any other company (130K in total, around 140/day, from 64K distinct users); AncestryDNA is a distant second (30K tweets from 16.9K users). Given their large customer bases, this should not come as a surprise. However, it *is* surprising that 23andMe has 4.6 times as many tweets as AncenstryDNA even though AncestryDNA, at the time of measurement, had over twice the number of customers of 23andMe. The least popular companies are MapMyGenome, PathwayGenomics, and VeritasGenetics, with less than 2K tweets each over our 2.5 year collection period. Among the initiatives, Precision Medicine generates a relative high number of tweets (83K from 13K users), much more so than Personalized Medicine (20K tweets).

Tweets per user. For each keyword, we also measure the number of tweets per user (see Figure 1(a)). We find that the median for every keyword is 1; i.e., 50% of users tweet about a given DTC company or initiative only once. However, we do find differences in the outliers for different keywords. For instance, there are several highly engaged users tweeting about Personalized Medicine and Precision Medicine. Manual examination of these users indicates that most of them are medical

[7]https://www.alexa.com/topsites



(a) (b)

Fig. 1. Number of tweets (a) per keyword, and (b) per user as a function of the number of unique keywords they tweeted about. Note the log scale in y-axis.

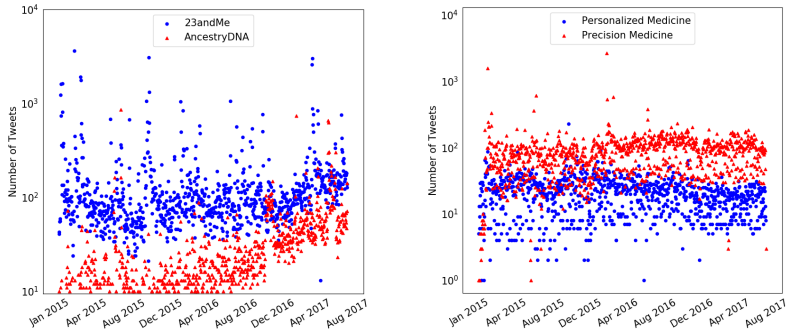
researchers and companies actively promoting the initiatives as hashtags. The presence of these heavily "invested" users becomes more apparent when we look at the number of tweets as a function of the number of unique keywords a user posts about, as plotted in Figure 1(b): 95% of them post about only one keyword, and those that post in more than one tend to post *substantially* more tweets about genetic testing in general; in some cases, orders of magnitude more tweets.

We also find differences between tweets about DTC genetic testing companies and those about genomics initiatives. The majority of the latter come from a smaller set of users compared to the former, i.e., a few very dedicated users drive the discussion about genomics initiatives. This is clear from Figure 1(a), which plots the number of tweets per user for each keyword: Personalized/Precision Medicine have more outliers than most of the DTC genetic companies (although the median for all keywords is 1). We also find these tweets are more likely to contain URLs (87% and 83% of tweets, respectively) than most companies, and even more so when compared to the baseline (45%).

This suggests that tweets about these topics often include links to news and/or other external resources. Only around 50% of URLs linked from tweets related to Genomics England or AncestryDNA are in the Alexa top 1M, compared to 60–75% for other keywords. For Genomics England, this is due to many URLs pointing to genomicsengland.co.uk itself. For AncenstryDNA, whose official site at ancestry.com *is* in the top 1M, it appears to be due a large number of marketing URLs tweeted along with the keyword; which we discuss later on.

Retweets and Likes. The total number of retweets and likes per tweet in the baseline is substantially higher than for tweets related to genetic testing due to outliers, i.e., viral tweets or tweets posted by famous accounts (e.g., a tweet by @POTUS44 on January 11, 2017 has 875,844 retweets and 1,862,249 likes). However, the median for retweets and likes in the baseline dataset mirrors that of tweets in our keywords dataset, with values between 0 and 1. Note that, although the number of retweets and likes per tweet could be influenced by how old the tweets are, this is not really the case in our dataset. Starting in late-August 2017, we collect tweets posted up to July 2017. This allows ample time to capture likes and retweets since preivous work [53] indicates that 75% of retweets happen within 24 hours and 85% happen within a month.

Official accounts tweets. We also look at the tweets including a given keyword (e.g., Ubiome) made by the corresponding official account (e.g., @Ubiome). There are no official accounts for Personalized and Precision Medicine, however, the Precision Medicine initiative is now called All Of Us and has a Twitter account (created in February 2017) that has posted only a few tweets (224



(a) 23andMe and AncestryDNA

(b) Personalized/Precision Medicine

Fig. 2. Number of tweets per day. Note the log scale in y-axis.

as of April 4, 2018), so we do not consider it. The percentage of tweets made by the official accounts of most companies including the name of the company as keyword is unsurprisingly very low (e.g., 1% for 23andMe). However, it is higher for others (e.g., 15% for DNAfit, Fitnessgenes, and MapMyGenome), due to the fact that these companies actually add their names in their tweets as a hashtag (e.g., #AncestryDNA). In fact, we find that hashtags are used quite predominantly for several DTC keywords, in some cases 40% of tweets have hashtags vs 23% for baseline tweets.

Temporal analysis. Finally, we analyze how the volume of tweets changes over time. In Figure 2, we plot the number of tweets per day in our dataset (between Jan 1, 2015–July 31, 2017) for the two most popular companies (23andMe/AncestryDNA) and the two most popular genomics initiatives (Personalized/Precision Medicine). On average, there are 145 and 30 tweets per day for 23andMe and AncestryDNA keywords, respectively. While the former is relatively constant, the latter increases steadily in 2017 (Figure 2(a)). This may be the result of AncestryDNA's aggressive promotion strategies (see Section 4.2). We also find a number of outliers for 23andMe, mostly around Feb 20 and Oct 19, 2015, and Apr 6, 2017, which are key dates related to 23andMe's failure to get FDA approval for their health reports in 2015, then obtained in 2017 [35]. In fact, 20K/132K 23andMe tweets are posted around those dates. As for Personalized and Precision Medicine (Figure 2(b)), the volume of tweets stays relatively flat. There are outliers for Precision Medicine too, e.g., 2,628 tweets on February 25, 2016, when the White House hosted the Precision Medicine Initiative summit [46].

4.2 What Are The Tweets About?

Next, we analyze the content of the tweets related to genetic testing, studying hashtags and URLs included in them and performing a simple sentiment analysis.

Hashtag Analysis. In Table 4, we report the top three hashtags for every keyword, while differentiating between tweets made by regular users and those by official accounts. We also quantify the percentage of tweets with at least one hashtag (WH) and that of tweets including the keyword as a hashtag (KH), e.g., #23andMe.

We find a few unexpected hashtags among the DTC tweets, e.g., #sweepstakes (AncestryDNA), #startup (Fitnessgenes), #vote (Ubiome), #shechat, and #appguesswho (MapMyGenome). AncestryDNA's top hashtag, #sweepstakes (12%), is related to a marketing campaign promoting a TV series, "America: Promised Land." There are 3.5K tweets, from distinct users, with the very same content (most likely due to a "share" button): "I believe I've discovered my @ancestry! Discover yours for the chance to win an AncestryDNA Kit. #sweepstakes journeythroughhistorysweeps.com."

Keyword	– Without Official Accounts – WH	Top 3 Hashtags	KH	– Only Official Accounts – Top 3 Hashtags	KH
23andMe	27.09%	dna (3.58%), genetics (2.07%), tech (1.96%)	12.46%	23andMestory (6.67%), genetics (6.35%), video (5.19%)	9.74%
AncestryDNA	75.48%	sweepstakes (12.38%), dna (4.90%), genealogy (4.86%)	25.94%	dna (11.74%), ancestry (5.92%), familyhistory (5.07%)	46.88%
Counsyl	45.24%	getaheadofcancer (2.64%), cap (1.93%), medical (1.94%)	3.08%	acog17 (6.18%), womenshealthweek (5.15%), teamcounsyl (5.15%)	0%
DNAFit	55.30%	diet (4.19%), fitness (3.72%), crossfit (3.54%)	22.91%	dna (5.33%), fitness (3.71%), generictogenetic (3.48%)	40.37%
FamilyTreeDNA	29.31%	dna (14.24%), genealogy (13.42%), ancestryhour (3.18%)	10.86%	geneticgenealogy (5.55%), ftdnasuccess (4.44%), ftdna (3.33%)	56.66%
FitnessGenes	72.19%	startup (5.93%), london (5.73%), job (5.59%)	18.22%	fitness (5.85%), dna (4.32%), gtsfit (2.79%)	45.29%
MapMyGenome	54.98%	shechat (7.94%), appguesswho (5.32%), genomepatri (4.22%)	15.80%	genomepatri (7.28%), knowyourself (4.04%), genetics (2.02%)	0%
PathwayGenomics	55.85%	coloncancer (6.91%), genetictesting (3.29%), cancer (2.85%)	3.34%	dnaday16 (9.67%), ashg15 (9.67%), health (3.22%)	19.35%
Ubiome	28.57%	microbiome (13.23%), tech (2.14%), vote (2.07%)	6.61%	microbiome (24.48%), bacteria (4.76%), meowcrobiome (2.72%)	6.12%
VeritasGenetics	57.16%	brca (3.92%), genome (3.62%), genomics (3.32%)	4.22%	brca (11.82%), liveintheknow (11.82%), wholegenome (10.75%)	0%
Genomics England	62.05%	genomes100k (14.84%), genomics (7.72%), raredisease (5.24%)	1.77%	genomes100k (32.45%), raredisease (19.49%), genomics (18.71%)	0%
Personalized Medicine	–	precisionmedicine (22.74%), genomics (9.77%), pmcon (8.37%)	–	–	–
Precision Medicine	–	genomics (6.70%), personalizedmedicine (5.49%), cancer (4.89%)	–	–	–

Table 4. Top 3 hashtags for each keyword, along with the percentage of tweets with at least a hashtag (WH) as well as that of of "keyword hashtags" (KH), e.g., #23andMe.

Keyword	Without Official Accounts	Only Official Accounts
23andMe	23andMe.com (7.33%), techcrunch.com (3.09%), fb.me (2.48%)	23me.co (50.88%), 23andMe.com (21.13%), instagram.com (5.40%)
AncestryDNA	journeythroughhistorysweeps.com (15.18%), ancestry.com (13.94%), ancstry.me (6.67%)	ancstry.me (74.11%), youtube.com (3.27%), ancestry.com.au (2.88%)
Counsyl	techcrunch.com (8.42%), businesswire.com (5.30%), bioportfolio.com (4.46%)	businesswire.com (14.78%), counsyl.com (13.91%), medium.com (5.21%)
DNAFit	fb.me (15.81%), instagram.com (14.65%), dnafit.com (2.99%)	fb.me (11.74%), dnafit.com (10.52%), dnafit.gr (2.83%)
FamilyTreeDNA	familytreedna.com (11.31%), myfamilydnatest.com (4.28%), fb.me (4.17%)	familytreedna.com (76.56%), abcn.ws (3.12%), instagram.com (1.56%)
FitnessGenes	instagram.com (14.77%), fitnessgenes.com (8.48%), workinstartups.com (6.29%)	fitnessgenes.com (31.11%), instagram.com (4.44%), pinterest.com (4.44%)
MapMyGenome	yourstory.com (11.84%), owler.us (11.44%), mapmygenome.in (9.18%)	mapmygenome.in (42.12%), youtu.be (14.35%), indiatimes.com (3.70%)
PathwayGenomics	paper.li (11.96%), atjo.es (10.82%), pathway.com (3.31%)	pathway.com (23.07%), nxtbook.com (3.84%), drhoffman.com (3.84%)
Ubiome	techcrunch.com (9.30%), bioportfolio.com (4.83%), ubiomeblog.com (4.21%)	ubiomeblog.com (34.32%), igg.me (26.07%), ubiome.com (6.60%)
VeritasGenetics	veritasgenetics.com (10.97%), technologyreview.com (5.01%), buff.ly (2.30%)	veritasgenetics.com (75.67%), biospace.com (1.35%), statnews.com (1.35%)
Genomics England	genomicsengland.co.uk (33.85%), youtube.com (1.98%), buff.ly (1.64%)	genomicsengland.co.uk (98.03%), peoplehr.net (0.58%), campaign-archive1.com (0.21%)
Personalized Medicine	instagram.com (8.78%), myriad.com (2.54%), buff.ly (2.32%)	–
Precision Medicine	buff.ly (2.92%), instagram.com (2.27%), nih.gov (1.87%)	–
Baseline	instagram.com (4.18%), fb.me (3.44%), youtu.be (2.72%)	–

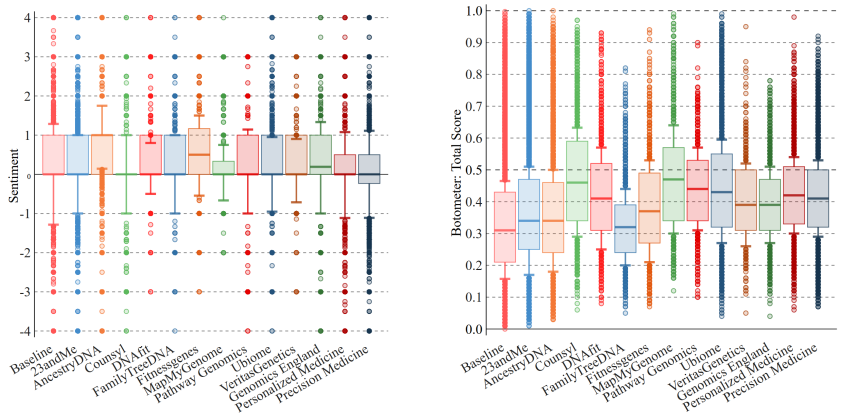
Table 5. The top 3 domains per keyword, without official accounts and only considering the official accounts.

We also find hashtags like #feistyfrugal and #holidaygiftguide in the AncestryDNA top 10 hashtags, which confirms how AncestryDNA uses Twitter for relatively aggressive marketing campaigns. Moreover, in the Fitnessgenes tweets, we find hashtags like #startup, #london, and #job due to a number of tweets advertising jobs for Fitnessgenes, while #shechat appears in tweets linking to an article related to women in business about MapMyGenome's founder.

By contrast, top hashtags for official accounts' tweets are closer to their main expertise/business. Similarly, those for genomics initiatives are pretty much always related to genetic testing, and this is actually consistent besides top 3. (The top 10 hashtags include, e.g., #digitalhealth, #genetics, and #lifestylemedicine). Finally, the percentage of tweets with the keyword appearing as a hashtag (KH), range from 12% for 23andMe to 25% for AncestryDNA even when excluding official accounts, which might be the by-product of promotion campaigns. When looking at tweets by official accounts KH values go up for some companies, e.g., AncestryDNA heavily promotes their brand using hashtags (46% KH).

URL Analysis. We also analyze the URLs contained in the tweets of our dataset. Recall that the ratio of tweets containing URLs, as well as the percentage of those in the Alexa top 1M domains, are reported in Table 1. Once again, we distinguish between tweets from the official accounts and report the top 3 (top-level) domains per keyword in Table 5. If we discover URL shortener services in our dataset (e.g., bit.ly, goo.gl, TinyURL, ow.ly) we "unshorten" the URLs and use those in our analysis instead. We also note that all reported cases of fb.me lead to Facebook posts.

Among the top URLs shared by the official accounts, we find, unsurprisingly, their websites, as well as others leading to other domains owned by them, e.g., 23me.co, ancestry.com.au, and ancstry.me. A few companies also promote news articles about them or related topics, e.g., top



(a) Sentiment Analysis (b) Botometer

Fig. 3. Sentiment and Botometer scores of the keyword dataset.

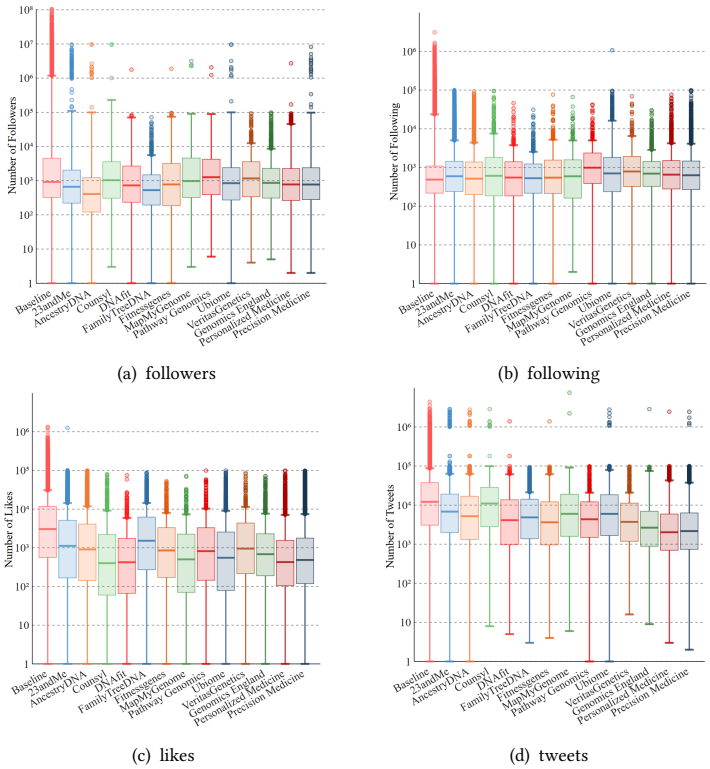
domains for Counsyl and MapMyGenome include businesswire.com and indiatimes.com, while DNAfit seems more focused on social media with its top domain being Facebook. As discussed previously, the domain journeythroughhistorysweeps.com appears frequently in AncestryDNA tweets. Then, note that techcrunch.com, a blog about technology, appears several times, as it often covers news and stories about genetic testing. We also highlight the presence of owler.us, an analytics/marketing provider sometimes labeled as potentially harmful by Twitter, as one of the top domains for MapMyGenome. Finally, for genomics initiatives, we notice buff.ly, a social media manager, suggesting that interested users appear to be extensively scheduling posts, thus potentially being more tech-savvy.

Sentiment Analysis We perform sentiment analysis using SentiStrength [96], which is designed to work on short texts. The tool outputs two scores, one positive, in $[1, 5]$, and one negative, in $[-1, -5]$. We calculate the sum value of the positive+negative scores for every tweet, then, collect *all* tweets with that keyword from the *same* user, and output the mean sentiment score.

In Figure 3(a), we report the distribution of sentiment across the different keywords. The vast majority of tweets have neutral sentiment, ranging from 0 to 1 scores. We run pair-wise two-sample Kolmogorov-Smirnov tests on the distributions, and in most cases reject the null hypothesis that they come from a common distribution at $\alpha = 0.05$. However, we are *unable* to reject the null hypothesis when comparing the baseline dataset to the PathwayGenomics dataset ($p = 0.77$) and when comparing DNAfit to Ubiome ($p = 0.34$). In general, the genomics initiatives, and in particular Personalized Medicine and Precision Medicine, have many outliers compared to most DTC genetic companies, suggesting more users who reveal strong feelings for or against these concepts. Genomics England, however, has a median above zero, indicating generally positive sentiment. Tweets about Counsyl are very neutral, while Ubiome tweets seem to be the most positive.

4.3 Who Tweets About Genetic Testing?

In this subsection, we shed light on the accounts tweeting about genetic testing. After a general characterization of the profiles, we look for the presence of social bots [98]. Then, we select a random sample of users tweeting about the two most popular DTC companies and analyze their latest 1,000 tweets to understand their interests. We start by analyzing the profiles tweeting about



(a) followers (b) following

(c) likes (d) tweets

Fig. 4. Boxplots with statistics per user profile (note the log-scale in y-axis).

genetic testing: in Figure 4, we plot the distribution of the number of their followers, following, likes, and tweets.

Followers. Accounts tweeting about genomics initiatives have a median number of followers similar to baseline, while for the DTC companies the median is always lower, except for Counsyl, MapMyGenome, PathwayGenomics, and VeritasGenetics (see Figure 4(a)). Also considering that, for these four companies, there is a relatively low number of unique users (see Table 1), we believe accounts tweeting about them are fewer but more "popular." There are fewer outliers than the baseline, which is not surprising since we do not expect many mainstream accounts to tweet about genetic testing. Some outliers appear for 23andMe and AncestryDNA, which, upon manual examination, turn out to be Twitter accounts of newspapers or known technology websites, reflecting how the two most popular companies also get more press coverage.

Following. The median number of 'following' (i.e., the accounts followed by the users in our dataset) is usually higher than baseline for DTC companies but similar for genomics initiatives (Figure 4(b)). This suggests that users interested in DTC genetic testing might want to get more information off Twitter and/or from more accounts.

Likes. We then measure the number of tweets each profile has liked (Figure 4(c)). This measure, along with the number of tweets, depicts, to a certain extent degree, a level of engagement. We find that, for all keywords, profiles like fewer tweets than baseline users. There is one interesting outlier for 23andMe (@littlebytesnews), who liked more than 1M tweets; this is likely to be a bot, as also

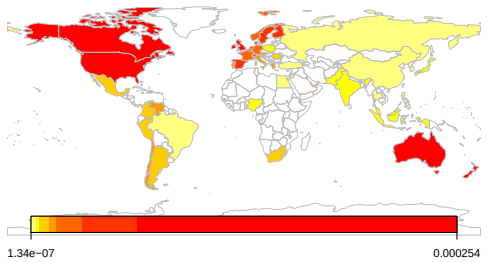


Fig. 5. Geolocation of Twitter profiles, normalized by Internet using population per country.

confirmed by Botometer [98]. Also, FamilyTreeDNA appears to have users liking more tweets than others. However, these accounts appear not to be bots, as we discuss later.

Tweets. We also quantify the number of tweets each account posts (Figure 4(d)). As with the number of likes, users in our datasets are less "active" than baseline users. There are interesting outliers above 1M tweets, which are due to social bots. We also find more tweets from Counsyl's users, seemingly mostly due to a large number of profiles describing themselves as "promoters" of science/digital life, technology enthusiasts, and/or influencers. Finally, users tweeting about genomics initiatives appear to be even less active, with a lower median value of tweets than the rest. Also considering that these users tweet more about the same keyword (as discussed in Sec. 4.1) but follow more accounts, we believe that they are more *passive* than the average Twitter user, using Twitter to get information but actively engaging less than others.

Geographic Distribution. We then estimate the geographic distribution of the users via the location field in their profile. Note that i) the corpus is biased as the used keywords are in English (e.g. Precision/Personalized Medicine), and ii) the location is self-reported, and users use it in different ways, adding their city (e.g., Miami), state (e.g., Florida), and/or country (e.g., USA). In some cases, entries might be empty (7.5% of the tweets in our dataset), ambiguous (e.g., Paris, France vs Paris, Texas), or fictitious (e.g., "Hell"). Nevertheless, as done in previous work [57], we use this field to estimate where most of the tweets are coming from. We use the Google Maps Geolocation API, which allows us to derive the country from a text containing a location.[8] The API returns an error for 6.6% of the profiles, mostly due to fictitious locations.

We find that the top 5 countries in our dataset are mostly English-speaking ones: 69.1% of all profiles with a valid location are from the US, followed by the UK (8.6%), Canada (4.5%), India (2.1%), and Australia (1.4%). We then *normalize* using Internet-using population estimates [51], and plot the resulting heatmap, with the top 50 countries, in Figure 5. The maximum value is obtained by the US (i.e., 0.000254 users per Internet user), with 72.8K unique users, out of an estimated Internet population of 286M, posting tweets in our dataset. This suggest that US users dominate the conversation on genetic testing on Twitter.

We also perform a geolocation analysis broken down to specific keywords. Unsurprisingly, the top country of origin for Genomics England is the UK, as it is for DNAfit, which is based in London. Similarly, the top country for India-based company MapMyGenome tweets is India. Overall, we find that tweet numbers are in line with the countries where the DTC companies are based or operate – e.g., 23andMe health reports are available in US, Canada, and UK, while AncestryDNA also operates in Australia – as well as where the genomics initiatives are taking place.

[8]https://developers.google.com/maps/documentation/geolocation

Social Bot Analysis Next, we investigate the presence of social bots in our datasets, using the Botometer (botometer.iuni.iu.edu), a tool that, given a Twitter handle, returns the probability of it being a "social bot," i.e., an account controlled by software, algorithmically generating content and establishing interactions [98].

In Figure 3(b), we plot the distribution of Botometer scores for all keywords. We compare the distributions using pairwise 2 sample KS tests, and reject the null hypothesis at $\alpha = 0.05$ for all datasets *except* Counsyl and MapMyGenome ($p = 0.29$), DNAfit and VeritasGenetics ($p = 0.17$) and PrecisionMedicine and VeritasGenetics ($p = 0.10$). We also find that all median scores are higher than the baseline (between 0.35 and 0.5 vs 0.3). This is not entirely surprising since we expect many blogs, magazines, and news services covering genetic testing, and these are likely to get higher scores than individuals since they likely automate their activities. However, about 80% of the accounts in our dataset have scores lower than 0.5 and 90% lower than 0.6 (i.e., it is unlikely they are bots). We also find the two most popular keywords, 23andMe and AncestryDNA, as well as FamilyTreeDNA, somewhat stand out: accounts tweeting about them get the lowest Botometer scores. Although for FamilyTreeDNA this might be an artifact of the relatively low number of tweets (2K users), the scores suggest there might be more interaction/engagement from "real" individuals and/or fewer tweets by automated accounts about 23andMe and AncestryDNA.

We then look at accounts with Botometer scores *above* 0.7, finding that, for most DTC keywords, they account for 3–5% of the users; not too far from the baseline (2%) and the genomics initiatives (1.5–2%). Counsyl and MapMyGenome have more than 10% of users with scores above 0.7. We also quantify *how many* tweets are posted by (likely) social bots: almost 15% of all PathwayGenomics tweets come from users with score 0.7 or above (4.5% of all users), while for all other keywords social bots are not responsible for a substantially high number of tweets in our datasets.

4.4 What Do Users Tweet About Otherwise?

We then focus on the users tweeting about the two most popular companies – i.e., 23andMe and AncestryDNA – and study their last 1K tweets, aiming to understand the characteristics of the accounts who show interest in genetic testing. We only do so for 23andMe and AncestryDNA as these companies have the highest numbers of tweets and users, and thus, are more likely to lead to a representative and interesting sample.

Data Crawl. We select a random 20% sample of the users who have posted at least one tweet with keywords 23andMe/AncestryDNA (resp., 12.2K/64K and 3.3K/16.9K users) and crawl their latest 1K tweets if their account is still active.[9] This yields a dataset of 12M tweets, outlined in Table 6. For comparison, we also get the last 1K tweets of a random sample of 5K users from the keyword dataset's baseline users. Note that statistics in Table 6 refer to the latest 1K tweets of the user sample, while those in Table 1 to tweets with a given keyword.

The numbers of retweets and likes per tweet are, once again, lower than the baseline. However, users tweeting about AncestryDNA receive, for their last 1K tweets, one order of magnitude more likes than those tweeting about 23andMe. Moreover, we observe relatively high percentages of tweets with hashtags (63%) and URLs (around 80%). How far back in time the 1,000th tweet appears varies across users, depending on how often they tweet. We measure the time between the most recent and the 1,000th tweet, and find that baseline users are more "active" than the users who have tweeted about 23andMe and AncestryDNA, in line with what discussed previously. In particular, AncestryDNA users appear to post less: for half of them, it takes at least 359 days to tweet 1K tweets compared to 260 for the baseline and 287 for 23andMe.

[9]We find 575 and 61 inactive accounts, resp., for 23andMe and AncestryDNA.

	Tweets	Users	RTs	Likes	Hashtags	URLs	Top 1M
23andMe	9,534,302	12,227	9,077,066	3,501,053	24.40%	63.62%	81.43%
AncestryDNA	2,466,443	3,320	1,399,804	22,001,065	34.21%	63.64%	78.86%
Total	12,000,745	15,547	10,476,870	25,502,118	26.41%	63.62%	80.89%
Baseline	4,208,967	5,035	139,551,104	342,052,546	17.47%	41.24%	88.41%

Table 6. Summary of the users' tweets dataset, with last 1K tweets of a 20% sample of 23andMe and AncestryDNA users.

23andMe	AncestryDNA	Baseline
tech (1.07%)	giveaway (3.31%)	gameinsight (0.55%)
news (1.06%)	sweepstakes (2.01%)	trecru (0.34%)
health (0.58%)	win (2.01%)	btsbbmas (0.33%)
business (0.48%)	genealogy (1.01%)	nowplaying (0.30%)
healthcare (0.43%)	tech (0.63%)	android (0.28%)
digitalhealth (0.40%)	ad (0.51%)	androidgames (0.27%)
startup (0.39%)	entry (0.51%)	ipad (0.26%)
socialmedia (0.34%)	promotion (0.48%)	trump (0.24%)
viral (0.34%)	perduecrew (0.47%)	music (0.21%)
technology (0.34%)	contest (0.44%)	ipadgames (0.20%)

Table 7. The top 10 hashtags of the users' tweets dataset.

23andMe	AncestryDNA	Baseline
fb.me (4.00%)	instagram.com (6.78%)	fb.me (5.85%)
instagram.com (3.06%)	fb.me (5.48%)	instagram.com (4.42%)
youtu.be (2.18%)	techcrunch.com (4.42%)	youtu.be (2.94%)
buff.ly (2.17%)	youtu.be (4.04%)	twittascope.com (0.58%)
techcrunch.com (1.53%)	wn.nr (1.79%)	tmblr.co (0.56%)
lnkd.in (1.02%)	woobox.com (1.51%)	buff.ly (0.54%)
mashable.com (0.65%)	giveaway.amazon.com (1.17%)	fllwrs.com (0.40%)
entrepreneur.com (0.63%)	buff.ly (1.08%)	gigam.es (0.33%)
nyti.ms (0.62%)	swee.ps (0.80%)	soundcloud.com (0.32%)
reddit.com (0.55%)	twittascope.com (0.41%)	vine.co (0.30%)

Table 8. The top 10 domains of the users' tweets dataset.

Hashtag analysis. We conduct a hashtag analysis on tweets in Table 6. In Table 7, we report the top 10 hashtags of the users' last 1K tweets. For 23andMe, we find several hashtags related to health in the top 10; also considering that the top 30 include #pharma, #cancer, and #biotech, it is likely that users who have shown interest in 23andMe are also very much interested in (digital) health, which is one of the primary aspects of 23andMe's business. This happens to a lesser extent for AncestryDNA results: while top hashtags include #genealogy (4th), they also include #giveaway, #sweepstakes, #win, #ad, #promotion, #perduecrew, and #contest, suggesting that these users are rather interested in promotional products. This is line with our earlier observation that AncestryDNA extensively uses advertising and marketing campaigns on Twitter.

URL analysis. In Table 8, we report the top 5 domains of the three sets. Over the last 1K tweets, users tweeting about 23andMe and AncestryDNA share a substantial number of links to techcrunch.com, a popular technology website; i.e., users who have tweeted at least once about these companies have an interest about subjects related to new technologies. In fact, the top 10 list of 23andMe's set of tweets also include lnkd.in, mashable.com, and entrepreneur.com. For AncestryDNA, we find wn.nr, another website related to contests and sweeps. There are thousands of tweets like "Enter for a chance to win a $500 Gift Card! wn.nr/DRRrZq #MemorialDaySweeps #Entry". We also note the presence of woobox.com, a marketing campaign website, responsible for organizing giveaways, as well as giveaway.amazon.com, an Amazon site organizing promotional sweepstakes. Botometer scores indicate these accounts are not actually bots, hence this might be related to the fact that AncestryDNA, through their marketing campaigns, attract Twitter users who are generally active in looking for deals and sweeps.

4.5 Instances of Racism

Finally, we are interested in tweets with extremely negative emotion. We select all tweets with genetic testing keywords from users who yield a total sentiment score below -3, obtaining 3,605 tweets from 3,209 unique users. We then manually examine those with keywords 23andMe or

AncestryDNA (1,725 and 167, respectively), and find several of them containing themes related to racism and hate.

In particular, the "ethnic" breakdown provided by ancestry reports[10] seems to spur several instances of negative-sentiment tweets associated with racism and disapproval of multi-cultural/multi-ethnic values. For instance, a user with more than 3K followers self-describing as a "Yuge fan for Donald Trump", tweets: "Get this race mixing shit off my time line!!" (March 23, 2017) in response to a 23andMe video about ancestry. Another posts: "I wanna do that 23andMe so bad! I'm kinda scared what my results will be tho lmao I'm prob like half black tbh"(January 13, 2017), and gets a response: "I was too just do it and never tell anyone if you're a halfbreed haha". Also, a user identifying as 'American Fascist' tweets: "I'd like to get the @23andMe kit but, I'm worried about the results. Just my luck, I'd have non-white/kike ancestors. #UltimateBlackpill" (May 30, 2017).

To assess whether genetics-based racism is a systematic theme on Twitter, we search for the presence of hateful words using the hatebase.org dictionary, a crowdsourced list of 1K terms that indicate hate when referring to a third person, removing words that are ambiguous or context-sensitive, as done by previous work [49]. Naturally, this is a best-effort approach since hateful terms might be used in non-hateful contexts (e.g., to refer to oneself), or, conversely, racist behavior can occur without hate words. Also, Twitter might be removing tweets with hate words as claimed in their hateful conduct policy.[11] Nonetheless, we do find instances of hate speech, e.g., anti-semitic tweets such as: "as long as there are khazar milkers to cause people to demand my 23andMe results, i will always be here to shitpost" (November 19, 2016), or "@*** i would be pleased if you posted your 23andMe so i can confirm your khazar milkers are indeed genuine" (December 23, 2016).

Note that "Khazar milkers" refers to an anti-semitic theory on the origin of Jewish people from the 1900s [36]. In a nutshell, it posits that Ashkenazi Jews are not descendant from Israelites, but from a tribe of Turkic origin that converted to Judaism. 23andMe issued ancestry reports that suggested Ashkenazi Jews in a given haplogroup were descendant from a single Khazarian ancestor. Understanding the ancestry of Jewish people has been of interest to the genetics community for years, and the Khazar theory has been refuted repeatedly [6]. Nonetheless, the alt-right has exploited it to corroborate their anti-semitic beliefs [79], and incorporate it into their collection of misleading/factually incorrect talking points.

However, we have already seen reports of genetic ancestry testing being used by white nationalists as a means of genetic discrimination, and the fact that we find such tweets in our dataset serves as our main motivation for studying the discourse on genetic testing on other social networks which allow for longer texts and are less susceptible to moderation, specifically, Reddit and 4chan.

4.6 Take-Aways

Overall, our characterization shows that highly engaged users drive the discussion around public genomics initiatives, which is particularly influenced, at least in terms of volumes, by important announcements such as the one made by President Obama. As for direct-to-consumer (DTC) genetic testing, the conversation is, as expected, dominated by the two most popular companies: 23andMe and AncestryDNA. However, it is interesting that the former generates 4 times more tweets even though the latter had, at the time, more than twice the customers. Some of this "popularity" seems to be due to 23andMe's controversy around FDA approval.

Furthermore, a large part of the genetic testing discourse appears to be generated from news and technology websites, and from tech-savvy users who rely on services to schedule social media posts. Also, sentiment around DTC companies is overall neutral, but positive for the genomics

[10] E.g. https://permalinks.23andMe.com/pdf/samplereport_ancestrycomp.pdf
[11] https://support.twitter.com/articles/20175050

initiatives, however, tweets about DTC companies include a lot of strongly opinionated users (both positive and negative). In general, we find several social media marketing strategies at play, with some companies employing traditional giveaways, others promoting mainly third-party articles about the company, and others focusing their efforts across multiple social media platforms. For instance, AncestryDNA is quite active in this context, with one particular hashtag (#sweepstakes) found in 1 out of 8 AncestryDNA tweets which may have an impact on how "regular" users engage in tweeting about genetic testing.

Then, we find that tweets related to genetic testing mostly come from users in the US and in general in English-speaking countries, but not necessarily by very popular accounts. The majority thes users are not bots, so Twitter conversation is, to some extent, "genuine." However, promotion and marketing campaigns end up attracting different kinds of users, and yield different levels of engagement. Furthermore, users tweeting about genetic testing appear less active than a random baseline, however, they are more likely to be interested in technology and digital health subjects.

Finally, our analysis uncovers evidence of tweets with hateful and racist content despite Twitter enforcing a hateful conduct policy which should lead to account suspension [97]. This finding serves as our motivation to study other social networks with relaxed moderation policies to identify whether genetic testing is being systematically used to promote racist ideologies; see next section.

5 STUDYING THE WEAPONIZATION OF GENETIC TESTING THROUGH REDDIT AND 4CHAN

Motivated by our findings around genetic testing and racism on Twitter as well as media reports of far-right groups using genetic testing to attack minorities [12, 80], we then study the relationship between genetic testing and racist views. In particular, we focus on Reddit and 4chan's politically incorrect board due to their relaxed moderation policies, their reputation for toxicity, and the fact that they allow for longer texts compared to Twitter.

5.1 Reddit

We make use of our Reddit dataset (see Section 3.2) 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 [74], 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 perform a user analysis in terms of overlap across subreddits

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 1,000 comments overall or fewer than 100 comments with one of the keywords. This yields a list of 114 subreddits, which we list in Table 15 (see Appendix), along with 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 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 Figure 6.

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

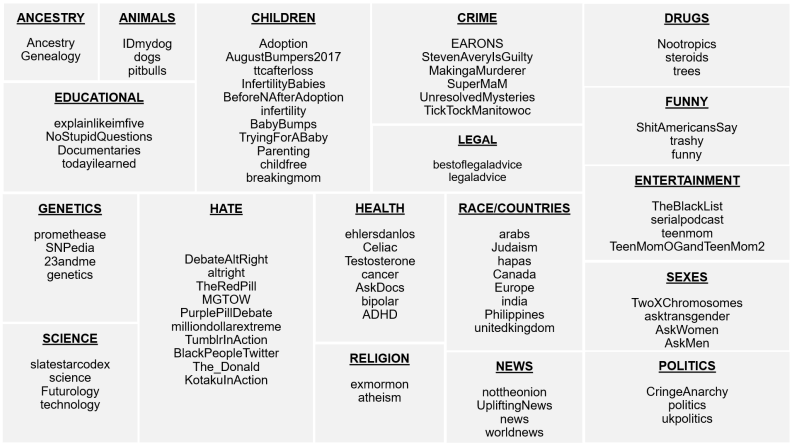
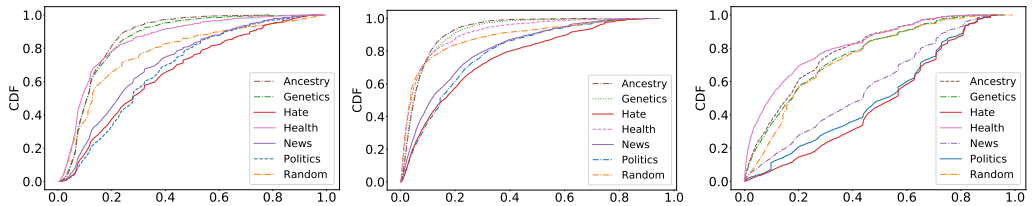


Fig. 6. Subreddits with genetic testing related comments, grouped based on their thematic topics (excluding a generic 'other' category).



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

Fig. 7. CDFs of Google's Perspective API toxicity on the genetic testing comments for the three most/least toxic subreddit categories.

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%). 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. Next, we analyze the category of subreddits denoted as hateful in terms of the topics being discussed using LDA. We also compare the toxicity of the comments therein with the rest of the categories using Google's Perspective API [74].

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 Figure 7, 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

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 9. LDA analysis of the Hate subreddits.

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.

Overall, we choose to use Google's Perspective to identify hateful content as other methods, e.g., hate speech detection libraries [29], 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.

Genetic Testing on Hateful Subreddits. 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 [94] (e.g., /r/altright, /r/DebateAltRight, and /r/The_Donald), sexism, or racism. For instance, /r/TheRedPill includes misogyny and toxic behavior towards women [56], 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 [92]. Other subreddits in this group include /r/milliondollarextreme, an American sketch satire show associated with alt-right and anti-semitism [90] which was banned in September 2018, as well as /r/KotakuInAction, which is associated with GamerGate-related toxicity [20]. Also, /r/BlackPeopleTwitter makes fun of tweets purporting to originate from African Americans. Our Perspective API analysis (see Figure 7) 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 9. 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 shown by Figure 7. 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). 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

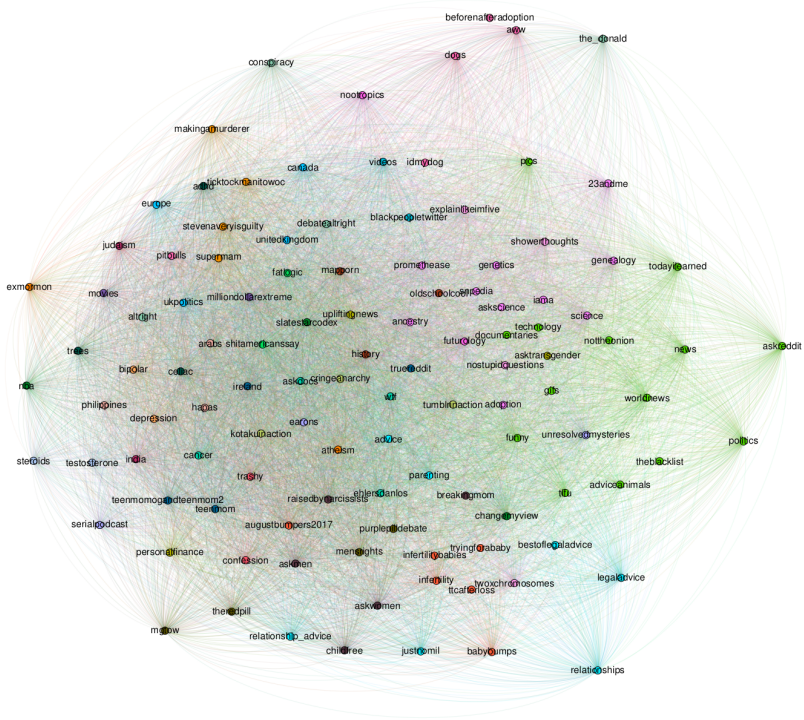


Fig. 8. Graph depicting the Jaccard Index of the users whose comments include genetic testing keywords for each subreddit.

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., "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 [93]). Finally, 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 [...]".

Grouping Users Based on Their Interests. 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. 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 [11], which provides a set of communities based on the graph's structure.

Figure 8 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 interested 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.

5.2 4chan

We now study genetic testing comments on 4chan's politically incorrect board (/pol/) which is shown to host users expressing fringe political views [49]. 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 [61] and DBSCAN clustering [34] to study imagery and memes in the dataset. Finally, we provide a comparison of the language used between Reddit and 4chan on genetic testing using word embeddings, specifically, word2vec [59].

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 [49], 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 9, 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 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 then measure hate and toxicity in /pol/ threads by computing: 1) percentage of hate words, and 2) toxicity/inflammatory levels. For the former, we use a dictionary of hate words compiled by and available from hatebase.org, similar to Section 4.5; for the latter, we again rely on the Perspective API. However, we find no major differences between the genetic testing

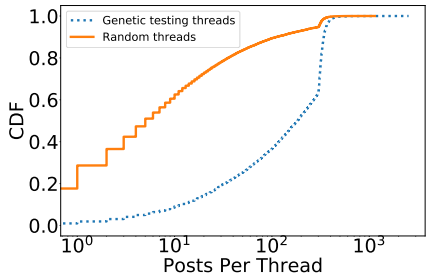


Fig. 9. CDF of number of posts in 4chan threads with genetic testing keywords vs random threads.

	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/.

threads and the random sample—which is not surprising as /pol/ is known for its high level of hate speech [49]—thus, we omit related plots to ease presentation.

Topic Modeling. We also use LDA modeling to identify the most prominent topics of discussion; see Table 10. 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 [24]. 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 open source image analysis pipeline introduced in [101] which uses Perceptual Hashing [61] and DBSCAN [34] 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 are labeled as noise by the clustering algorithm and thus we discard them. This high noise ratio mirrors findings in [101] and is likely due to 4chan users creating a lot of original content [49].

We annotate each cluster using Google's Cloud Vision API.[12] We calculate the medoid of each cluster (i.e., its "representative" image) following the methodology by [101], and use that image to

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

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.

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'.

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 [99] (6th most popular with 2.3% of all clusters), and Carl Benjamin, a YouTuber known for his misogynistic involvement in the GamerGate controversy [9], 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 [3], 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 anti-semitic 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

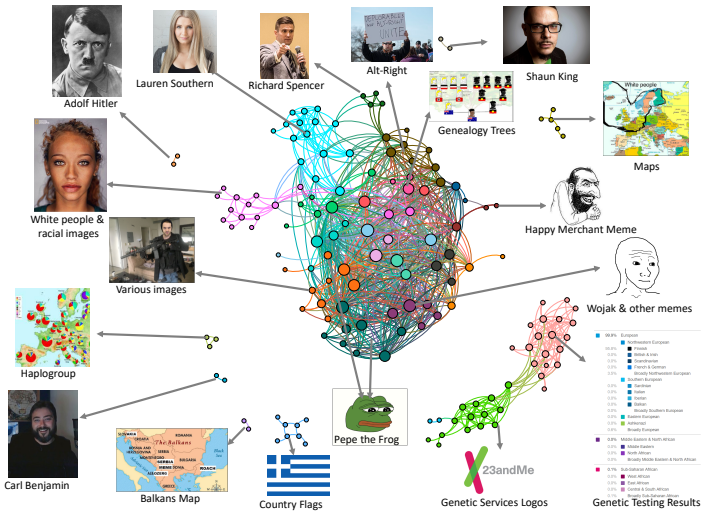


Fig. 10. Visualization of the image clusters with manual annotation

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 10. 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 [11]. 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 10, 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 [40], 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 [37].

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 in the vocabulary of the word2vec models trained for each group of subreddits and /pol/.

5.3 A Language Comparison of Reddit & 4chan

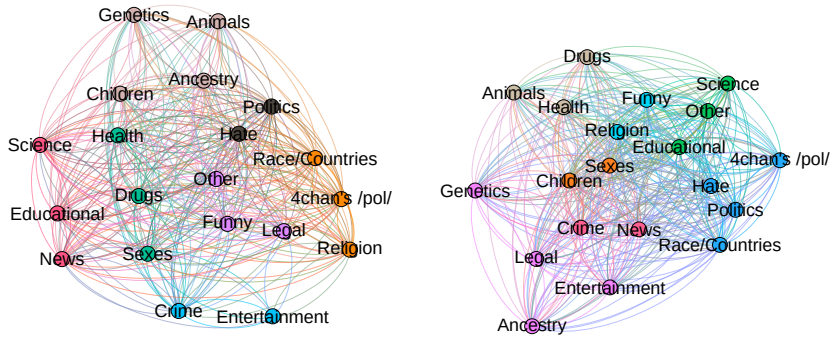
Although they both provide discussion platforms, Reddit and 4chan operate in different ways: e.g., the former requires registration, while the predominant mode of operation on the latter is via anonymous and ephemeral posting. Naturally, they also attract different sets of users and content, e.g., 4chan 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 [59]. 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 [81], for each of the 19 groups of subreddits (see Figure 6) 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 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 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, we use the methodology described above and for each word2vec model, we get the 10 most similar words (based on the cosine similarities obtained from the word2vec model) 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").



(a) Genetic Testing Keywords (b) Selected Keywords

Fig. 11. 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."

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 11), 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 [11].

When using the genetic testing keywords as seeds (Figure 11(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. Figure 7). 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 [37].

When using the set of hand-picked seed words (Figure 11(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.

5.4 Take-Aways

Our Reddit analysis shows that genetic testing is discussed in a variety of contexts (subreddits) which in itself is another indicator of how mainstream it has become. Interestingly, 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 focus in typical uses of genetic testing, others discuss their use in worrying ways, which is in line with our Twitter findings. More importantly, we find that toxic language around genetic testing is more prevalent on Reddit than on Twitter, often including users who use genetic testing to push racist agendas, e.g., to eliminate or marginalize

minorities, which is particularly worrying since Reddit is a mainstream platform (5th most visited site in the US [78]).

On 4chan, we find that genetic testing is a rather popular topic of discussion, often appearing in long/active /pol/ 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 on /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), anti-semitic 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.

Finally, our word embeddings analysis reveals that certain subreddits use ethnic terms in conjunction with genetic testing keywords similarly to /pol/, which somewhat confirms how 4chan's fringe ideologies spill out on more mainstream Web communities like Reddit.

6 DISCUSSION

6.1 Summary

Direct-To-Consumer (DTC) genetic testing is a revolutionary technology 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 [13]. 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.

This paper, to the best of our knowledge, is the first, large scale, multi-platform quantitative study on genetic testing. Our work encompasses two steps. The first consists of an exploratory, large-scale analysis of the genetic testing discourse through the lens of Twitter. The second explores the connection between genetic testing and racism on Reddit and 4chan's /pol/.

Our analysis shows that genetic testing is a popular topic of discussion on all three networks, mainly discussed by tech enthusiasts and people with a vested interested in its success. However, our Twitter analysis also unveils the existence of tweets with racist connotations. These trends are quite pronounced on Reddit and 4chan, where we find that discussions around genetic testing often include highly toxic language expressed through hateful and racist comments. Specifically, on 4chan's politically incorrect board (/pol/), content from genetic testing conversations involves several alt-right personalities and openly anti-semitic rhetoric, often conveyed through memes.

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. Specifically, information has been increasingly weaponized, including by state actors, to sew racial discontent [95] and even instigate public health crises [14]. In this context, recent efforts have been made by law enforcement to understand and address such campaigns [62]. Thus, we ought to reflect on the practical implications of our findings and how they affect future work in this area.

6.2 High-Level Cross-Platform Examination

Next, we attempt to provide some comparison across the three social networks we have analyzed with respect to genetic testing discourse. This is not without challenges: 1) the three platforms differ in several functionality aspects, and 2) because we have looked at different aspects of the discourse and through a slightly different set of keywords (see Section 3). Re. 1), tweets may or may not be responses to other tweets, while on Reddit, a comment follows an original post and each

comment can be replied to. On 4chan, all comments are responses to an original post but there is only one thread of discussion. Also, a tweet contains a maximum of 280 characters, while, on Reddit and 4chan, there is no restriction. Furthermore, Twitter and Reddit require the creation of a profile, while, 4chan comments can be and in most cases are anonymous. Re. 2), we have actually attempted to conduct the same analysis on Twitter, in terms of keywords and dates, as the one done on Reddit/4chan, using a dataset containing 1% of all tweets (using Twitter's streaming API) from 2015, which only yields a set of only 35K total tweets. A preliminary study of this dataset, unfortunately, did not yield conclusive findings due to its relative limited size, and thus we do not include it.

For these reasons, it would arguably be impossible to conduct a one-to-one comparison of our three datasets. In fact, our aim is not to do so, nor to compare how the users of each platform differ with respect to their views on genetic testing, but rather to reason around some of the aspects of DTC genetic testing discourse we extract via our quantitative analysis. As a result, we set to perform a high-level cross-platform examination, focusing on three axes: topics, user engagement, and hateful content.

Topics. On Twitter, our hashtag and URL analyses show that a large part of the genetic testing discourse is generated from news and technology websites. We also find several social media marketing strategies at play, with some DTC companies employing traditional giveaways, and others promoting third-party articles about their brands. On the other hand, we find that the genetic testing discourse on Reddit is centered around 18 "mega-topics," ranging from educational and scientific content to politics, religion, crimes, and a set of subreddits strongly associated with hateful content. Interestingly, our /pol/ analysis shows that genetic testing is a popular topic of discussion, often accompanied by images and memes with clear racial or hateful connotations.

Users. On Twitter, the conversation around genetic testing is often dominated by users with a vested interest in its success, such as journalists, medical professionals, and entrepreneurs. Conversely, on Reddit, users who discuss genetic testing tend to form distinct groups ranging from enthusiasts (e.g., those who are interested in or have undergone genetic testing), to people who use genetic keywords exclusively in subreddits that discuss fringe political views. Due to the fact that most 4chan posts are anonymous, we obviously are unable to go down to the user-level. Overall, we observe a dichotomy in the type of users interested in genetic testing among all three datasets: some focus in typical uses of genetic testing (Twitter and most subreddits), while others discuss their use in worrying ways (subreddits associated with hateful content and 4chan's /pol/).

Hateful Content. Toxic content surrounding genetic testing conversations is rather sparse on Twitter, but not inexistent, despite Twitter's conduct policy which should lead to account suspension [97]. Whereas, on Reddit and /pol/, our analysis shows that genetics-based racism is rather systematic. Specifically, we uncover 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 subreddits associated with hateful content often contain highly toxic language, with users even suggesting leveraging genetic testing tools to further marginalize or even eliminate minorities. We also find that images appearing along genetic testing conversations often include alt-right personalities and anti-semitic memes. Word embeddings reveal 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.

6.3 Limitations

Our study has a few limitations. First, the three platforms differ drastically both qualitatively (type of content) and quantitatively (number of users and posts). Second, our datasets do not span the same time periods. Namely, our Twitter dataset starts on January 1, 2015 and ends on July 31, 2017, our Reddit dataset spans from January 1, 2016 to March 31, 2018, and our 4chan dataset spans from June 30, 2016 to March 13, 2018. Third, we use a different set of keywords for collecting the Twitter and the Reddit/4chan datasets. Specifically, the set of keywords used for the Reddit/4chan datasets is one order magnitude larger than the Twitter dataset.

6.4 Future Work

As part of future work, we plan to leverage the rapid progress of machine learning algorithms to train a classifier that would be able to detect harmful users in the context of genetic testing discourse. Furthermore, considering that previous qualitative studies [70, 71, 86] 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 [100]; 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.

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 donating two Titan Xp GPUs used in our experiments.

REFERENCES

[1] 2020. Reddit Metrics. https://redditmetrics.com/history.
[2] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You Tweet What You Eat: Studying Food Consumption Through Twitter. In *CHI*.
[3] ADL. 2019. Pepe the Frog. https://www.adl.org/education/references/hate-symbols/pepe-the-frog.
[4] AncestryDNA. 2019. Ancestry Company Facts. https://www.ancestry.com/corporate/about-ancestry/company-facts.
[5] Euan A Ashley. 2016. Towards Precision Medicine. *Nature Reviews Genetics* 17, 9 (2016), 507.
[6] Doron M. Behar, Mait Metspalu, Yael Baran, et al. 2013. No Evidence from Genome-Wide Data of a Khazar Origin for the Ashkenazi Jews. *Human Biology* 85, 6 (2013).
[7] Anat Ben-David and Ariadna Matamoros-Fernandez. 2016. Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain. *IJOC* 10 (2016), 1167–1193.
[8] Michael Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 2011. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. *ICWSM* (2011).
[9] Joe Bish. 2016. Vice News. Examining the Right Wing British Blowhards Using YouTube to Prove Everybody Wrong. https://bit.ly/2qN4SMG.
[10] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.
[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast Unfolding of Communities in Large Networks. *JSTAT* 2008, 10 (2008), P10008.
[12] Eric Boodman. 2016. White Nationalists Are Flocking To Genetic Ancestry Tests – But Many Don't Like Their Results. https://read.bi/2DEaQYY.

[13] Katie Sullivan Borrelli. 2018. PressConnects. DNA Tales: These People Found Long-Lost or Never-Known Relatives. https://bit.ly/2FxDye2.
[14] David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. 2018. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American journal of public health* 108, 10 (2018).
[15] Pete Burnap, Matthew L Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack. *Social Network Analysis and Mining* 4, 1 (2014).
[16] Timothy Caulfield and Amy L McGuire. 2012. Direct-To-Consumer Genetic Testing: Perceptions, Problems, and Policy Responses. *Annual Review of Medicine* 63 (2012), 23–33.
[17] Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina J Sowles, and Laura J Bierut. 2015. Hey Everyone, I'm Drunk. An Evaluation Of Drinking-Related Twitter Chatter. *JSAD* 76, 4 (2015).
[18] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. 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 (2017), 31.
[19] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities. In *CHI*. 3175–3187.
[20] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying. In *WWW 2017*.
[21] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In *Proceedings of the 2017 ACM on web science conference*. ACM, 13–22.
[22] Peter Chow-White, Stephan Struve, Alberto Lusoli, Frederik Lesage, Nilesh Saraf, and Amanda Oldring. 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 (2018), 448–464.
[23] Emily Christofides and Kieran O'Doherty. 2016. Company Disclosure and Consumer Perceptions of the Privacy Implications of Direct-To-Consumer Genetic Testing. *New Genetics and Society* 35, 2 (2016), 101–123.
[24] Matthew Claxton. 2017. Abbotsford News. Former Langley Libertarian candidate detained in Italy. https://bit.ly/2PUIQWC.
[25] EW Clayton, CM Halverson, NA Sathe, and BA Malin. 2018. A Systematic Literature Review of Individuals' Perspectives on Privacy and Genetic Information in the United States. *PLoS ONE* 13, 10 (2018).
[26] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals In Twitter. In *CLPsych*.
[27] Nick Couldry and Jun Yu. 2018. Deconstructing Datafication's Brave New World. *New Media & Society* 20, 12 (2018), 4473–4491.
[28] BF Darst, L Madlensky, NJ Schork, EJ Topol, and Cinnamon S Bloss. 2013. Perceptions of Genetic Counseling Services in Direct-To-Consumer Personal Genomic Testing. *Clinical genetics* 84, 4 (2013), 335–339.
[29] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.
[30] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In *ICWSM*.
[31] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *ICWSM*.
[32] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *CEUR Workshop*. 86–95.
[33] DNARomance. 2018. Online Dating Based On Science. https://www.dnaromance.com/.
[34] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*.
[35] FDA. 2017. FDA allows marketing of first direct-to-consumer tests that provide genetic risk information for certain conditions. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm551185.htm.
[36] Ari Feldman. 2017. 23andMe Backpedals On Khazar Theory But The 'Alt-Right' Eats It Up, Anyway. http://forward.com/news/national/381500/23andme-backpedals-on-khazar-theory-but-the-alt-right-eats-it-up-anyway/.
[37] Joel Finkelstein, Savvas Zannettou, Barry Bradlyn, and Jeremy Blackburn. 2018. A Quantitative Approach to Understanding Online Antisemitism. *CoRR* abs/1809.01644 (2018).
[38] Claudia Flores-Saviaga, Brian C. Keegan, and Saiph Savage. 2018. Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community. In *ICWSM*.
[39] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A Unified Deep Learning Architecture for Abuse Detection. In *Proceedings of the 10th ACM Conference on Web*

Science. ACM, 105–114.
[40] Amanda Froelich. 2014. True Activist. This is What Americans Will Look like by 2050. https://bit.ly/2vpAIEH.
[41] GEDmatch. 2019. https://en.wikipedia.org/wiki/GEDmatch.
[42] Genetics Home Reference. 2019. What is the Precision Medicine Initiative? https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative.
[43] Genomics England. 2019. https://www.genomicsengland.co.uk/.
[44] Melissa Gymrek, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. 2013. Identifying Personal Genomes by Surname Inference. *Science* 339, 6117 (2013), 321–324.
[45] Katie EJ Hann, Madeleine Freeman, Lindsay Fraser, Jo Waller, et al. 2017. Awareness, Knowledge, Perceptions, and Attitudes Towards Genetic Testing for Cancer Risk Among Ethnic Minority Groups: A Systematic Review. *BMC public health* 17, 1 (2017), 503.
[46] Liz Harley. 2016. White House hosts Precision Medicine Initiative Summit. http://www.frontlinegenomics.com/white-house-hosts-precision-medicine-initiative-summit/.
[47] Amy Harmon. 2018. New York Times. Why White Supremacists Are Chugging Milk (and Why Geneticists Are Alarmed). https://nyti.ms/2Afg4Ho.
[48] Helix. 2017. DNA Technologies 101 Genotyping vs Sequencing, and What They Mean For You. https://blog.helix.com/dna-technologies-genotyping-vs-sequencing/.
[49] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*.
[50] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *SocInfo*.
[51] Internet Live Stats. 2017. Internet Users by Country (2016). http://www.internetlivestats.com/internet-users-by-country/.
[52] Anna Kasunic and Geoff Kaufman. 2018. "At Least the Pizzas You Make Are Hot": Norms, Values, and Abrasive Humor on the Subreddit r/RoastMe. In *ICWSM*.
[53] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What Is Twitter, A Social Network Or A News Media?. In *WWW*.
[54] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. 2001. Initial Sequencing and Analysis of the Human Genome. *Nature* 409, 6822 (2001), 860–921.
[55] Kristina Lerman, Megha Arora, Luciano Gallegos, Ponnurangam Kumaraguru, and David Garcia. 2016. Emotions, Demographics and Sociability in Twitter Interactions. In *ICWSM*.
[56] Stephen Marche. 2016. The Guardian. Swallowing the Red Pill: A Journey to the Heart of Modern Misogyny. https://bit.ly/2Chey99.
[57] Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, et al. 2011. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *CHI*.
[58] Medical Press. 2018. US Craze for DNA 'heritage' Tests May Bolster Racism, Critics Warn. https://medicalxpress.com/news/2018-10-craze-dna-heritage-bolster-racism.html.
[59] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*.
[60] Richard A. Mills. 2018. Pop-up Political Advocacy Communities on Reddit.com: SandersForPresident and The Donald. *AI and Society* 33, 1 (2018), 39–54.
[61] Vishal Monga and Brian L. Evans. 2006. Perceptual Image Hashing Via Feature Points: Performance Evaluation and Tradeoffs. *IEEE Transactions on Image Processing* (2006).
[62] Robert S. Mueller. 2019. Report On The Investigation Into Russian Interference In The 2016 Presidential Election. US Department of Justice.
[63] NHGRI. 2018. The Cost of Sequencing a Human Genome. https://www.genome.gov/sequencingcosts/.
[64] Daiva E Nielsen, Sarah Shih, and Ahmed El-Sohemy. 2014. Perceptions of Genetic Testing for Personalized Nutrition: A Randomized Trial of DNA-based Dietary Advice. *Lifestyle Genomics* 7, 2 (2014), 94–104.
[65] NIH. 2017. All of Us. https://allofus.nih.gov/.
[66] NIH. 2019. What Is Genetic Ancestry Testing? https://ghr.nlm.nih.gov/primer/dtcgenetictesting/ancestrytesting.
[67] Alicia L Nobles, Caitlin N Dreisbach, Jessica Keim-malpass, and Laura E Barnes. 2018. "Is This an STD? Please Help!" Online Information Seeking for Sexually Transmitted Diseases on Reddit. In *ICWSM*.
[68] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. 2018. The Effect of Extremist Violence on Hateful Speech Online. In *ICWSM*.

[69] Raphael Ottoni, Evandro Cunha, Gabriel Magno, Pedro Bernadina, Wagner Meira, and Virgilio Almeida. 2018. Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination. In *WebSci*.

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

[71] Aaron Panofsky and Joan Donovan. 2019. Genetic ancestry testing among white nationalists: From identity repair to citizen science. *Social studies of science* (2019).

[72] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. *arXiv preprint arXiv:2001.07487* (2020).

[73] Michael J Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In *ICWSM*.

[74] Perspective. 2019. https://www.perspectiveapi.com/.

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

[76] Nugroho Dwi Prasetyo, Claudia Hauff, Dong Nguyen, Tijs van den Broek, and Djoerd Hiemstra. 2015. On the Impact of Twitter-Based Health Campaigns: A Cross-Country Analysis of Movember. In *EMNPL*.

[77] Presidential Commission for the Study of Bioethical Issues. 2012. Privacy and Progress in Whole Genome Sequencing. https://bioethicsarchive.georgetown.edu/pcsbi/node/764.html.

[78] Reddit. 2020. https://www.redditinc.com/press.

[79] Elspeth Reeve. 2016. Vice News – White Nonsense: Alt-right trolls are arguing over genetic tests they think prove their whiteness. http://bit.ly/2DhP90h.

[80] Elspeth Reeve. 2016. Vice News. White Nonsense. https://bit.ly/2DhP90h.

[81] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *NLPFrameworks*.

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

[83] Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira. 2018. Characterizing and Detecting Hateful Users on Twitter. In *ICWSM*.

[84] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira. 2019. Auditing radicalization pathways on youtube. *arXiv preprint arXiv:1908.08313* (2019).

[85] Caitlin M Rivers and Bryan L Lewis. 2014. Ethical research standards in a world of big data. *F1000Research* 3 (2014).

[86] Wendy D Roth and Biorn Ivemark. 2018. Genetic Options : The Impact of Genetic Ancestry Testing on Consumers' Racial. *Amer. J. Sociology* 124, 1 (2018), 150–184.

[87] Tina Hesman Saey. 2018. What I Actually Learned About My Family After Trying 5 DNA Ancestry Tests. https://bit.ly/2zaUIKy.

[88] Suyash S Shringarpure and Carlos D Bustamante. 2015. Privacy Risks from Genomic Data-Sharing Beacons. *The American Journal of Human Genetics* 97, 5 (2015), 631–646.

[89] Leandro Silva, Mainack Mondal, Denzil Correa, Fabricio Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *ICWSM*.

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

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

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

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

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

[95] Leo G Stewart, Ahmer Arif, and Kate Starbird. 2018. Examining trolls and polarization with a retweet network. In *WSDM*.

[96] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment Strength Detection In Short Informal Text. *JASIST* 61, 12 (2010).

[97] Twitter. 2019. Hateful Conduct Policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[98] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *ICWSM*.

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

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

[101] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *IMC*.

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

Table 15. List of subreddits sorted by normalized number of genetic testing comments.

[102] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *IMC*.

[103] Zephoria. 2020. Top 10 Twitter Statistics – Updated February 2020. https://zephoria.com/twitter-statistics-top-ten/.

A LIST OF SUBREDDITS

In Table 15, we report the list of subreddits sorted by normalized number of genetic testing comments.