A Quantitative Study of Twitter Discourse On Genetic Testing

Alexandros Mittos[†], Jeremy Blackburn[‡], Emiliano De Cristofaro[†]
[†]University College London [‡] University of Alabama at Birmingham
alexandros.mittos.16@ucl.ac.uk, blackburn@uab.edu, e.decristofaro@ucl.ac.uk

ABSTRACT

The availability of cheap sequencing technologies and the rise of direct-to-consumer testing are bringing genetic testing to the masses. As a result, news, experiences, and views on genetic testing are increasingly more often shared and discussed online and on social networks such as Twitter. However, to the best of our knowledge, there is no understanding of how customers and users discuss them on social media. In this paper, we start addressing this gap by presenting the first large-scale analysis of Twitter discourse on genetic testing. We search Twitter for thirteen carefully selected keywords related to genetic testing companies and genomics initiatives and collect 300K tweets from more than 100K users. We then analyze these tweets along several axes, with the goal of understanding who tweets about genetic testing, what they talk about, and how they use Twitter for that. Our results confirm certain trends, e.g., that people who have shown interest in genetic testing appear to be overall interested in digital health and technology, but also point to a number of surprising aspects. For instance, we find that marketing efforts as well as announcements, such as the FDA's suspension of 23andMe's health reports, influence the type and the nature of users' engagement. Finally, we discuss ethical and ideological questions emerging from our study, as we find evidence of groups utilizing genomic testing to push racist agendas and of users expressing privacy concerns.

1 INTRODUCTION

In 1990, the Human Genome Project was kicked off with the goal of producing the first complete sequence of a human genome; at a cost of almost $3 billion, it was completed 13 years later [22]. Since then, costs have dropped at a staggering rate: by 2006, high-quality sequencing of a human genome cost $14 million, and, by 2016, private individuals could have their genomes sequenced for about $1,500 [26]. Such a rapid progress prompts hopes for a new era of "personalized medicine," where diagnosis and treatment can be tailored to patients' genetic makeup, thus becoming more preventive and effective [4]. This has also encouraged initiatives to sequence large numbers of individuals for research purposes. In 2015, the US announced the Precision Medicine Initiative, aiming to collect genetic and health data from 1M citizens. In the UK, the Genomics England project is sequencing 100K patients for research on cancer and rare diseases.

Furthermore, a number of companies have emerged that offer *direct-to-consumer* (DTC) genetic testing. Rather than visiting a lab or a clinic, customers purchase a collection kit for a few hundred dollars (or less), deposit a saliva sample, and mail it back. After a few days, without interacting with doctors or genetics experts, they receive a report with information about genetic health risks (e.g., susceptibility to Alzheimer's, breast/ovarian cancer, etc.), wellness information (e.g., lactose intolerance), and/or ancestry and genealogy information. Today, there are possibly hundreds of DTC companies; naturally, some more reputable than others [32]. Very successful companies include 23andMe and AncestryDNA: the former provides reports on carrier status, health, and ancestry, while the latter focuses on genealogy and ancestry. As of November 2017, 23andMe has 3M and AncestryDNA 6M customers.[1]

Traditionally, health-related issues were communicated to patients primarily by doctors and clinicians—the advent of direct-to-consumer genetic testing changes this substantially. Individuals can now learn potentially life-changing results with a few clicks of the mouse, without contacting a medical professional. Also, as results are delivered electronically, they are more easily shared with others. Overall, the rise of participatory sequencing initiatives as well as affordable DTC services means that genetic testing increasingly involves and is available to the general population. As with other aspects of the digital health revolution, this results in discussion, sharing of experiences, and molding of perceptions around genetic testing to be increasingly online and social. However, while the research community has studied the interlinked relationship between health and social networks in great detail (see Section 2), to the best of our knowledge, there is no understanding of how customers and users in general discuss their views on and their experiences with genetic testing on social media.

In this paper, we aim to address this gap by presenting the first large-scale analysis of Twitter discourse on genetic testing. Starting from 13 keywords related to DTC genetics companies and genomics initiatives, we search and crawl all available tweets containing these keywords that were posted between January 1, 2015 and July 31, 2017. We collect 302K tweets from 113K users, and analyze them along several axes, aiming to understand *who* tweets about genetic testing, *what* they talk about, and *how* they use Twitter for that.

We present a general characterization of our datasets (Section 3), then, we analyze the tweets content-wise, studying the most common hashtags/URLs and measuring sentiment (Section 4). Next, we perform a user-based analysis, looking at their profile and location, and assessing whether they are likely to be social bots [39] (Section 5). We also select a random sample of 15K users and analyze their latest 1K tweets to study their interests.

Main findings. Overall, our analysis shows that:

(1) Users interested in genetic testing are generally interested in digital health and technology, but the discourse around genetic testing may be dominated by users that seemingly have a vested interest (e.g., specialist journalists, medical professionals, and entrepreneurs) in its success.
(2) There is a clear distinction in the marketing efforts undertaken by different companies, which naturally influence the type and the nature of users' engagement on Twitter.

[1] See https://mediacenter.23andme.com/company/about-us/, http://ancstry.me/2iD4ITy

(3) Although the majority of the users in our datasets appear not to be bots, each keyword attracts different percentages of tweets originating from automated content publishers.
(4) There are also ethical and ideological questions at play; in particular, we find evidence of groups utilizing genomic testing to push racist agendas and of users expressing privacy concerns.
(5) Two DTC companies, 23andMe and AncestryDNA, are talked about the most. Even though 23andMe has less than half as many customers as AncestryDNA, it has over 4 times as many tweets, with high volumes around dates related to its failure to get FDA approval [38]; interestingly, the attention also makes unrelated privacy concerns resurface.

2 RELATED WORK

User perspectives on genetic testing. There are several *qualitative* studies in literature analyzing users' perspectives around genetic testing. Goldsmith et al. [16] conduct a systematic review of 17 studies conducted in 6 different countries. They find that, although participants appear to be interested in the health-related aspects of testing, they also express concerns about privacy and reliability. Covolo et al. [12] review 118 articles, aiming to systematize perceptions of users on DTC genetic testing. They find that users are mainly drawn to genetic testing by the potential to monitor and improving their health, especially if they work in the biotechnology industry or are at a risk of cancer. Caulfield et al. [8] analyze the controversy around Myriad Genetics and their attempt to patent the BRCA gene which is associated with predisposition to breast cancer. They study related newspapers references, finding that the majority of them demonstrate negative sentiment.

Closer to our work would be *quantitative* studies using social media, however, to the best of our knowledge, the only relevant work is by Chow-White et al. [10]. They look at one week's worth of tweets containing the word '23andMe' and perform a simple sentiment analysis, finding that positive tweets outnumber negative ones, and that people tend to be enthusiastic about it. Compared to ours, their analysis, besides relying on a much smaller dataset (2K vs 324K tweets, collected over 1 week vs 2.5 years), only studies one company and only sentiment, whereas we study 10 companies and 3 initiatives, conducting both a content and a user-based analysis.

Health in social networks. Social networks like Twitter have been used extensively to study health and health-related issues, e.g., to measure and predict depression. De Choudhury et al. [13] identify 476 users self-reporting depression, collect their tweets, and study them in terms of 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. [11] study tweets related to various mental disorders, comparing those from users diagnosed with mental illnesses to a control group, and finding various differences in their language as well as evidence of information relevant to mental health disorders in social media. Paul et al. [31] gather public health information from Twitter, collecting 1.63M tweets and discovering statistically significant correlations between Twitter and official health statistics. Abbar et al. [3] analyze the nutritional behavior of US citizens: they get 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. [33] 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.

Analyzing discourse on social media. Finally, another line of work studies discourse and sentiment in Twitter. Cavazos-Rehg et al. [9] 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. Lerman et al. [23] conduct an emotion analysis on tweets from Los Angeles: using public demographic data, they show 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. Burnap et al. [7] study Twitter response 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.

3 KEYWORD DATASET

In this section, we introduce the methodology used to gather tweets related to genomic and genetic testing, and present a general characterization of the resulting datasets.

3.1 Data Collection

We collect tweets containing keywords related to (1) direct-to-consumer (DTC) genetic testing companies, and (2) public genome sequencing initiatives, using these keywords as search queries and crawling all results from January 1, 2015 to July 31, 2017.

DTC genetic testing companies. We start from the list of 36 (as of October 2017) DTC genetic testing companies maintained by the International Society of Genetic Genealogy (ISOGG) [20]. Although non-exhaustive, this provides a representative sample of the DTC ecosystem. 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 does not provide exact results with quotes.

Genomics initiatives. Besides tweets related to for-profit companies, we also want to measure discourse around public sequencing initiatives and related concepts. Thus, we select three more keywords: PrecisionMedicine, PersonalizedMedicine, and GenomicsEngland. Personalized Medicine is a concept related to advances in genomics which hope 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 to sequence the genome of 1M individuals announced by President Obama in 2015 to understand how a person's genetics,

	Tweets	Users	RTs	Likes	Official	Photos	Quotes	Hashtags	URLs	Top 1M
23andMe	132,597	64,014	72,848	149,897	1.31%	6.14%	3.49%	27.23%	68.68%	75.40%
AncestryDNA	29,071	16,905	16,266	47,249	7.08%	8.79%	2.69%	54.29%	75.50%	49.68%
Counsyl	3,862	1,834	2,716	4,255	3.49%	6.98%	4.64%	44.01%	83.94%	74.97%
DNAFit	2,118	844	1,336	2,508	15.34%	18.74%	5.37%	57.22%	78.94%	79.18%
FamilyTreeDNA	2,794	1,205	1,196	3,111	4.36%	19.97%	6.62%	34.18%	36.47%	69.21%
FitnessGenes	2,142	773	908	2,809	16.29%	18.47%	9.40%	44.53%	56.76%	71.28%
MapMyGenome	1,568	704	4,488	3,726	15.30%	13.13%	4.99%	53.63%	80.35%	64.30%
PathwayGenomics	1,544	579	1,968	2,521	2.13%	18.51%	6.11%	61.01%	76.55%	68.12%
Ubiome	14,420	6,762	9,223	13,991	2.71%	4.37%	2.85%	27.85%	73.28%	64.19%
VeritasGenetics	1,292	497	1,443	2,526	6.65%	17.07%	17.07%	46.13%	58.28%	71.95%
Genomics England	7,009	1,863	19,772	18,756	19.68%	17.80%	11.58%	61.19%	69.18%	48.82%
Personalized Medicine	20,302	4,631	19,085	15,514	–	6.93%	7.55%	99.93%	87.42%	71.98%
Precision Medicine	83,329	13,012	118,043	128,303	–	8.56%	10.41%	99.88%	83.39%	77.16%
Total	302,048	113,624	269,292	395,166	2.26%	7.75%	5.92%	56.54%	74.77%	71.80%
Baseline	163,260	131,712	282,063,006	486,960,753	–	41.20%	12.07%	23.48%	45.49%	89.57%

Table 1: Our keyword dataset, with all tweets from Jan 1, 2015 to Jul 31, 2017 containing keywords related to genetic testing.

environment, and lifestyle can help determine the best approach to prevent or treat disease [27]. Genomics England is a similar UK initiative to sequence 100K genomes primarily for 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 search engine's incorrectness.

Crawl. We use a custom Python script to collect all tweets from January 1, 2015 to July 31, 2017 returned as search results using the 10 DTC keywords as well as the 3 keywords related to genomics initiatives. The crawler, run with self-imposed throttling to avoid issues for the site operators over August–September 2017, collects, for each 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 110K from 19K users for the 3 initiatives.

Note that the keyword search also returns user accounts that match that keyword: e.g., tweets containing 23andMe, #23andMe, or @23andMe, but also those made by the @23andMe account. For consistency, we discard the latter, only keeping tweets that *include* the keyword. When relevant, we analyze these tweets separately.

Baseline. We also crawl a set of 163,260 random English-language tweets, from the same January 2015 to July 2017 period (approximately 170 per day), which serves as a baseline set for comparisons.

3.2 General Characterization

Our "keyword dataset" is summarized in Table 1. From left to right, the table reports the total number of tweets, unique users, retweets, and likes for each of the 13 keywords (as well as the baseline). We also quantify the percentage of tweets made by the official accounts of each company or initiative (where applicable), as well as the percentage of tweets including photos, quoted tweets, hashtags, and URLs, and how many URLs in the Alexa Top 1M.

DTC vs Initiatives. Overall, we 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. 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.

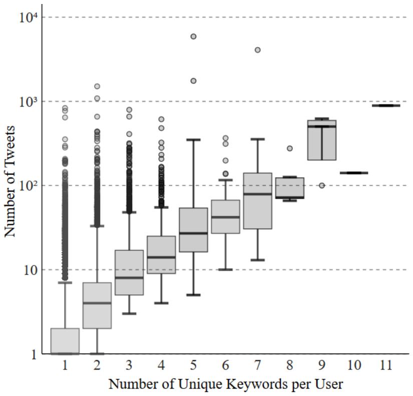


Figure 1: Number of tweets per user as a function of the number of unique keywords they tweeted about.

Only around 50% of URLs linked from tweets related to Genomics England or AncenstryDNA 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 to a very large number of marketing/spam URLs tweeted along with the keyword; we discuss this further in Section 4.

Number of tweets. When looking at the sheer number of tweets per keyword, we find that 23andMe is by far the most popular topic, with one order of magnitude more tweets than any other company (130K in total, around 140/day, from 64K users); AncestryDNA is a distant second (30K tweets from 16.9K users). Given their large customer bases, it is not surprising these two topics generate the most tweets. What *is* surprising, however, is that 23andMe has 4.6 times as many tweets as AncenstryDNA even though AncestryDNA has twice the customers as 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 although we do not report a plot due to space limitations. 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 researchers and companies actively promoting the initiatives as hashtags. The presence of these heavily "invested" users becomes more apparent when we plot the number of tweets as a function of the number of unique keywords a user posts about (Figure 1): 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.

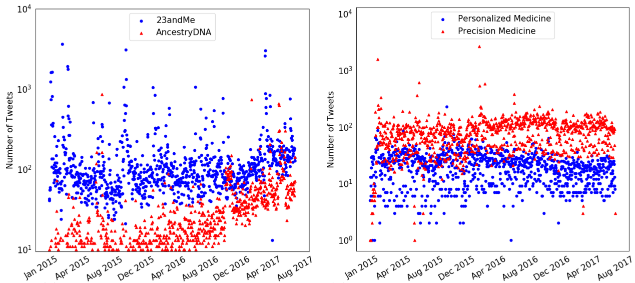
Retweets and Likes. The total number of retweets and likes per tweet in the baseline is substantially higher than for tweets related to genetic testing due to outliers, i.e., viral tweets or tweets posted by famous accounts (e.g., a tweet by @POTUS44 on January 11, 2017 has 875,844 retweets and 1,862,249 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. We collect tweets posted up to July 2017 starting in late-August 2017, allowing ample time for retweets and likes to occur, considering that previous work [21] indicates that 75% of retweets happen within 24 hours and 85% happen within a month.

Official accounts tweets. We also focus on tweets with 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 [28] and has a Twitter account (created in February 2017) that has posted only a few tweets (75 as of October 22, 2017), so we do not consider it.

Tweets made by the official accounts of some companies account for very low percentages (e.g., 1% for 23andMe) but higher for others (e.g., DNAfit, Fitnessgenes, and MapMyGenome for 15%). This is due to some DTC companies using their name in their tweets more than others, possibly in a hashtag (e.g., #AncestryDNA), as we discuss in Section 4.2. (Recall we only collect tweets from official accounts if they also contain the corresponding keywords).

Hashtags, photos, and quotes. Table 1 shows that around a quarter of 23andMe's and Ubiome's tweets have hashtags (slightly more than 23% for the baseline); for most other keywords, it is above 40%. For Personalized and Precision Medicine, we find that almost all tweets have the keyword in the form of hashtag (99%). For Genomics England, this only happens 61% of the time, since a lot of tweets include @GenomicsEngland. We perform a more detailed hashtag analysis in Section 4.2.

We then find that the percentage of tweets with photos vary from 4% in Ubiome to almost 20% in FamilyTreeDNA. Anecdotally, we notice that photos often contain text, i.e., are used to overcome the 140-character limit and comment on issues related to the company (e.g., [30]). We also look at "quotes", i.e., tweets including the URL of another tweet: for most keywords, percentages are lower than the baseline, except for VeritasGenetics (mostly due to the official account), although less so for the initiatives. Possibly, users tweeting about genomics initiatives tend to be discuss more with each other, by commenting on relevant tweets.



(a) 23andMe and AncestryDNA (b) Personalized and Precision Medicine

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

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 (Jan 1, 2015–July 31, 2017) for the two most popular companies and the two most popular genomics initiatives. On average, there are 145 and 30 tweets per day for 23andMe and AncestryDNA keywords, respectively. While the former stays relatively flat over time, the latter increases steadily in 2017 (Figure 2(a)). This may be the result of AncestryDNA's aggressive promotion strategies (see Section 4). We also find a number of outliers for 23andMe, mostly around February 20 and October 19, 2015, and April 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 [38]. In fact, out of the 132K 23andMe tweets in our dataset, 20K are posted around those dates. As for Personalized and Precision Medicine (Figure 2(b)), the volume of tweets stays relatively flat (21 and 83 tweets/day on average, respectively). 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 [18].

4 CONTENT ANALYSIS

In this section, we present a content analysis of the tweets related to genetic testing. We perform sentiment analysis, then, we study hashtags and URLs included in the tweets to extract topics of interest. We also conduct Latent Dirichlet Allocation (LDA) [6] and Term Frequency-Inverse Document Frequency (TF-IDF) [34] analysis, but the results are somewhat inconclusive, so we do not include them due to space limitation.

4.1 Sentiment Analysis

We perform sentiment analysis using the SentiStrength tool [36], which is designed to work on short texts. The tool outputs two scores, one positive, in [1, 5], and one negative, in [−1, −5] range. We calculate the sum value of the positive+negative scores for every

		Without Official Accounts		Only Official Accounts	
Keyword	WH	Top 3 Hashtags	KH	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 2: 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.

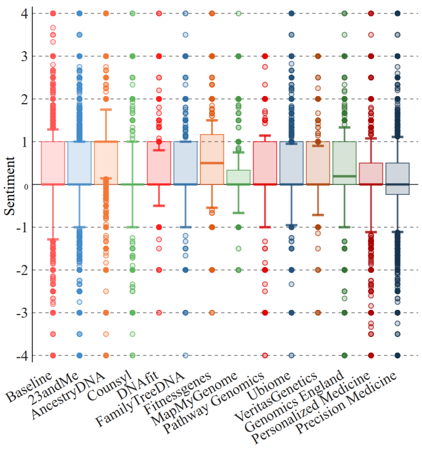


Figure 3: Sentiment scores of the keyword dataset.

tweet, then, we collect *all* tweets with that keyword from the *same* user, and output the mean sentiment score of the user.

In Figure 3, 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 ($p = 0.34$), indicating no significant difference in the sentiment expressed.

In general, we notice that 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 towards the initiative. Tweets about Counsyl are very neutral, while Ubiome tweets seem to be the most positive. Note that we further explore tweets with particularly positive or negative sentiment in Section 6.

4.2 Hashtag Analysis

In Table 2, we report the top three hashtags for every keyword while differentiating between tweets made by regular users and those by official accounts. We also account for the percentage of tweets with at least one hashtag (WH) and that of tweets including the keyword as a hashtag (e.g., #23andMe), abbreviated as KH.

We observe a few unexpected hashtags in 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," between May and June 2017. We find 3.5K tweets, from distinct users, posting the very same tweet (most likely through a "share" button): "I believe I've discovered my @ancestry ! Discover yours for the chance to win an AncestryDNA Kit. #sweepstakes journeythroughhistorysweeps.com". 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 much closer to their main expertise/business. Similarly, those for genomics initiatives are almost exclusively related to genetic testing (this is consistent outside the top 3: the top 10 hashtags include #digitalhealth, #genetics, and #lifestylemedicine).

Finally, the percentage of keyword hashtags (KH), not counting official accounts, range from 12% for 23andMe to 25% for AncestryDNA and 22% for DNAfit. This might be related to promotion from the companies themselves: for the official accounts, we find that AncestryDNA and DNAfit heavily promote their brands using hashtags (46% and 40%, respectively).

4.3 URL Analysis

Next, we 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

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

Table 3. There are quite a few URL shorteners in our dataset (e.g., bit.ly appears in 7.8% of the 23andMe tweets), so we first extract the top 10 domains for each keyword and identify those that *only* provide URL shortening services, then, we "unshorten" the URLs and use them in our analysis instead.

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 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 in Section 4.2, 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 users interested in these initiatives appear to be extensively scheduling posts, thus potentially being more tech-savvy. We also find myriad.com, the domain of Myriad Genetics, which discovered the BRCA1 gene and tried to patent it [8].

5 USER ANALYSIS

In this section, we shed light on the users who have shown interest in genetic testing and perform a user-level analysis on their profiles and on whether they are social bots. We also select a random sample of the users tweeting about the two most popular DTC companies, and analyze their latest tweets to study what their interests are.

5.1 User Profiles

We start by analyzing the characteristics of the users posting tweets about genetic testing, i.e., those in Table 1. In Figure 4, we plot the distribution of the number of followers, following, likes, and tweets.

Followers. Accounts tweeting about genomics initiatives have a median number of followers similar to the 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, we observe a relatively low number of unique users (cf. Table 1), we believe accounts tweeting about them are fewer but more "popular." Overall, there are much fewer outliers than the baseline, which is not surprising since we do not expect many mainstream accounts to tweet about genetic testing. Some big outliers appear for 23andMe and AncestryDNA, which, upon manual examination, turn out to be Twitter accounts of newspapers or known technology websites/companies, reflecting how the two most popular companies also get significant more press coverage.

Following. Conversely, the median number of following in our dataset is usually higher than the baseline (Figure 4(b)). This is particularly evident for users tweeting about PathwayGenomics. Whereas, for the genomics initiatives, we observe a behavior closer to the baseline. On the other hand, the average number of following is higher for the baseline, due to a great deal of outliers. Overall, this suggests that the users who are interested in DTC genetic testing might be interested in getting more information off Twitter by following 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 confirmed by Botometer [39]. Also, FamilyTreeDNA appears to have users liking more tweets than others. However, these accounts appear not to be bots, as we discuss below.

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 (cf. Section 3.2) but follow more accounts, we believe that they are somewhat more *passive* than the average Twitter user, possibly using Twitter to get information but actively engaging less than others.

Geographic Distribution. Finally, we estimate the geographic distribution of the users via the location field in their profile (when

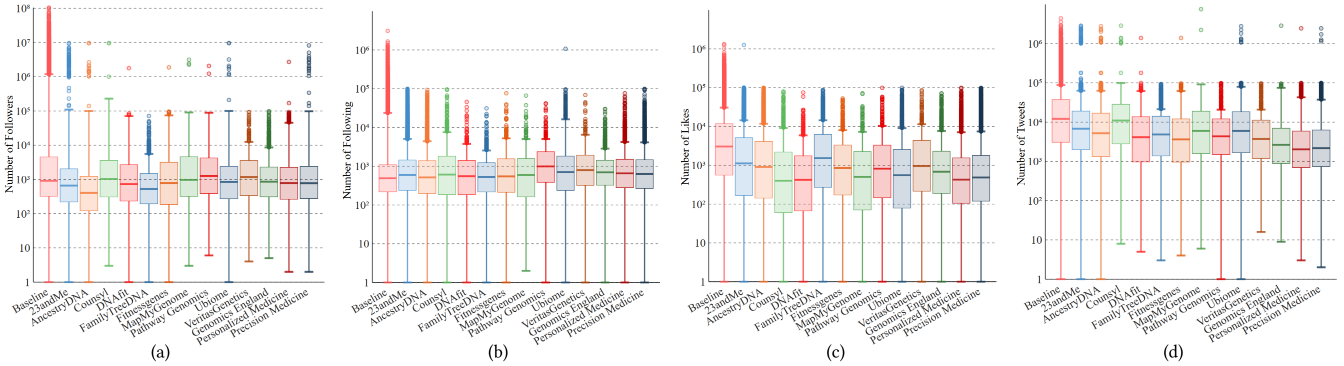


Figure 4: Boxplots with number of (a) followers, (b) following, (c) likes, and (d) tweets, per user profile (y-axis is in log-scale).

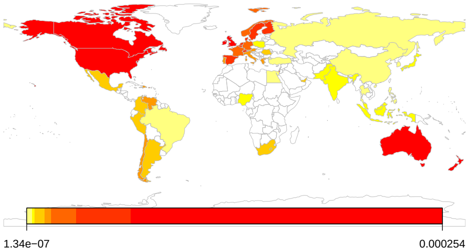


Figure 5: Geolocation of Twitter profiles, normalized by Internet using population per country.

available). This 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 (this happens for 7.5% of the profiles), ambiguous (e.g., Paris, France vs Paris, Texas), or fictitious (e.g., "Hell"). Nevertheless, as done in previous work [24], we use this field to estimate where most of the tweets are coming from. To this end, we use the Google Maps Geolocation API [17], which allows to derive the country from a text containing a location. 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 [35], 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.

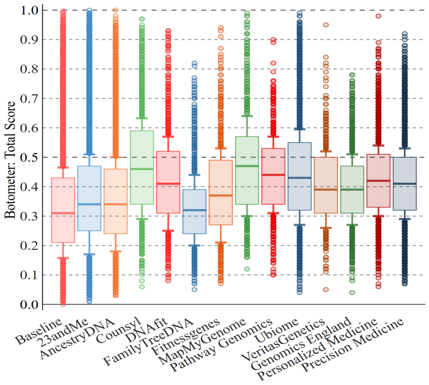


Figure 6: Botometer scores for the keyword dataset.

5.2 Social Bot Analysis

Next, we shed light on the presence of social bots in our datasets. We rely on Botometer [39], a tool developed by the Observatory on Social Media, which, given a Twitter handle, returns the probability of it being a social bot. Varol et al. [39] define social bots as accounts controlled by software, algorithmically generating content and establishing interactions, highlighting how they often perform useful functions (e.g., dissemination of news), but might also emulate human behavior for nefarious or unethical activities.

In Figure 6, we plot the distribution of Botometer scores for all keywords. First, 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 ($p = 0.10$). Next, we see 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 genomic 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. For FamilyTreeDNA this might be an artifact of the relatively low number of tweets (2K users), the scores suggest there might be

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

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%). That said, 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.

5.3 Analyzing a Sample of Users' Last 1K Tweets

Next, we 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 people who have shown 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.

Crawling the samples. We select a random 20% sample of the users that have posted at least one tweet with keywords 23andMe or AncestryDNA (resp., 12.2K/64K and 3.3K/16.9K users) and crawl their latest 1K tweets if their account is still active.[2] This yields a dataset of 12M tweets, outlined in Table 4. 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 4 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. We also observe relatively high percentages of tweets with hashtags (63%) and URLs (around 80%). Finally, note that 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 in Section 5.1. 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.

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 5: 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 6: The top 10 domains of the users' tweets dataset.

Hashtag analysis. We then conduct a hashtag analysis on tweets in Table 4. In Table 5, 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 also 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 (see Section 4) that AncestryDNA extensively uses advertising and marketing campaigns on Twitter.

URL analysis. We also perform a URL analysis, as in Section 4.3, reporting, in Table 6, 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 contests and giveaways, as well as giveaway.amazon.com, an Amazon site to organize promotional sweepstakes. Given the presence of these sites, we postulate this might be due to a large presence of bots, however, Botometer [39] indicates these accounts are not. Therefore, this behavior might in fact be related to the fact that AncestryDNA, through their marketing campaigns, attract Twitter users who are generally active in looking for deals and sweeps.

6 CASE STUDIES

In this section, we take a closer look at "negative" tweets, following the sentiment analysis presented in Section 4.1. More specifically,

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

we select, from the keyword dataset (Table 1), all tweets from users who yield a total sentiment score ≤ -3, getting 3,605 tweets from 3,209 unique users. We then proceed to manually examine those containing keywords 23andMe or AncestryDNA (1,725 and 167 tweets respectively), and discover that several of them contain themes related to racism, hate, and privacy fears.

Racism. Considering the "ethnic" breakdown provided by ancestry reports [2], it is not totally surprising to repeatedly find tweets associated with racism and users disapproving of multi-cultural and multi-ethnic values. For instance, we find the tweet "@23andMe Get this race mixing shit off my time line!!" (Mar 23, 2017) in response to a video posted by 23andMe about ancestry, by a user with more than 3K followers self-describing as a "Yuge fan for Donald Trump." Another user tweets, "@*** 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" (Jan 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' posts: "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).

Although we leave it to future work to perform an in-depth analysis of racism in genetic testing related tweets, we opt to assess whether racism may be systematic, e.g., appearing also in tweets not scored as negative. To this end, we search for the presence of hateful words in our datasets, relying on the Hatebase dictionary, a crowdsourced list of around 1K terms that indicate hate when referring to a third person [1]. Like previous work [19], we remove words that are ambiguous or context-sensitive. Naturally, this is far from perfect 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 [37].

Nonetheless, we do find instances of hate speech along, e.g., with anti-semitic tweets including: "as long as there are khazar milkers to cause people to demand my 23andme results, i will always be here to shitpost" (Nov 19, 2016), or "@*** i would be pleased if you posted your 23andme so i can confirm your khazar milkers are indeed genuine" (Dec 23, 2016). Note that "Khazar milkers" refers to an anti-semitic theory on the origin of Jewish people from the 1900s [15]. 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 origins of Jewish people has been of interest in the genetics community for years [29], and the Khazar theory has been refuted as recently as 2013 [5]. Although it is extremely unlikely 23andMe intentionally spread this theory, the alt-right has seized upon this "scientific" confirmation of their anti-semitic beliefs, incorporating it into their collection of misleading/factually incorrect talking points. In particular, "khazar milkers" was allegedly coined by the "@***" user mentioned above, and is used to imply a sort of succubus quality of Jewish women.

Privacy. We also identify, among the most negative tweets, themes related to fears of privacy violation and data misuse. Examples include "Is it me? Does the idea of #23andMe seem a bit sinister? Do they keep the results? Who owns the results? Who owns 23andMe?" (Jan 1, 2016), "same thing with 23andMe and similar companies. Indefinitely stored data with possible sinister future uses? #blackmirror" (Nov 13, 2016), and "Why does this scare the hell out of me? How can our privacy ever be assured?" (Feb 27, 2016)

We follow up by searching for 'privacy' and 'private' in our keyword dataset. This returns 1,991 tweets, mostly from 23andMe and Precision Medicine (1.1K and 625, resp.), which we proceed to examine both manually and from a temporal point of view (i.e., measuring daily volumes). Overall, we find that privacy in the context of genetic testing appears to be a theme discussed recurrently on social media and a concern far from being addressed. This is not entirely unexpected, considering that both the DTC market and the genomics landscape are evolving relatively fast, with regulation and understanding of data protection as well as informed consent often lagging behind, as also highlighted in prior work [14, 25, 32].

One interesting finding is that one of the peaks in tweets related to 23andMe and privacy occurs on October 19, 2015 (with 152 tweets including 23andMe and privacy/private). As discussed in Section 3.2, this a relevant date w.r.t. the FDA revoking their approval for 23andMe's health reports, which yields a peak in 23andMe tweets overall. However, the FDA ruling had nothing to do with privacy, yet, it put 23andMe in the spotlight, possibly causing privacy concerns to resurface. In fact, privacy and 23andMe discussions periodically appear in our dataset, even beyond tweets with negative sentiment, e.g., "I want to do #23andme but don't want a private company owning my genetic data. Anyone heard of any hacks to do it anonymously?" (Jul 13, 2017), "@23andMe ur privacy policy describes how there is no privacy. How about u not share any data at all. I pay u and u send the results. Period" (Dec 8, 2015), "Should we be concerned about data collection and privacy with direct to consumer DNA testing companies like 23andme?" (Apr 19, 2017).

7 CONCLUSION

This paper presented the first large-scale analysis of Twitter discourse on genetic testing. We examined more than 300K tweets related to genetic testing and 12M tweets posted by Twitter accounts that have shown interest in genetic testing. We found that users tweeting about genetic testing are generally interested in digital health and technology, but the overall discourse around genetic testing seems to be dominated by users that might have a vested interest in its success. Two DTC companies, 23andMe and AncestryDNA, are talked about the most; even though the former has less than half as many customers as the latter, it has over 4 times as many tweets, with high volumes around dates related to its failure to get FDA approval [38]. Moreover, we noticed a clear distinction in the marketing efforts undertaken by different companies, which naturally influence users' engagement on Twitter. We also discussed ethical and ideological issues, as we found evidence of groups utilizing genomic testing to push racist agendas as well as users expressing privacy concerns.

As part of future work, we plan to extend our analysis to other social platforms and health forums/websites, as well as to perform in-depth qualitative studies of racism and privacy issues.

REFERENCES

[1] 2017. Hatebase database. https://www.hatebase.org/. (2017).

[2] 23andMe. 2017. Ancestry Composition. https://permalinks.23andme.com/pdf/samplereport_ancestrycomp.pdf. (2017).

[3] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You Tweet What You Eat: Studying Food Consumption Through Twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3197–3206.

[4] Euan A Ashley. 2016. Towards Precision Medicine. *Nature Reviews Genetics* 17, 9 (2016).

[5] 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), 859–900.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* (2003).

[7] 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), 206.

[8] Timothy Caulfield, Tania Bubela, and CJ Murdoch. 2007. Myriad and the mass media: the covering of a gene patent controversy. *Genetics in Medicine* 9, 12 (2007).

[9] 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. *Journal of Studies On Alcohol And Drugs* 76, 4 (2015), 635–643.

[10] Peter Chow-White, Stephan Struve, Alberto Lusoli, Frederik Lesage, Nilesh Saraf, and Amanda Oldring. 2017. 'Warren Buffet Is My Cousin': Shaping Public Understanding of Big Data Biotechnology, Direct-To-Consumer Genomics, and 23andMe on Twitter. *Information, Communication & Society* (2017), 1–17.

[11] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals In Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 51–60.

[12] Loredana Covolo, Sara Rubinelli, Elisabetta Ceretti, and Umberto Gelatti. 2015. Internet-Based Direct-to-Consumer Genetic Testing: A Systematic Review. *Journal of medical Internet research* 17, 12 (2015).

[13] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. *ICWSM* 13 (2013), 1–10.

[14] Emiliano De Cristofaro. 2014. Genomic privacy and the rise of a new research community. *IEEE Security & Privacy* 12, 2 (2014), 80–83.

[15] 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/. (August 2017).

[16] Lesley Goldsmith, Leigh Jackson, Anita O'connor, and Heather Skirton. 2012. Direct-to-Consumer Genomic Testing: Systematic Review of the Literature on User Perspectives. *European Journal of Human Genetics* 20, 8 (2012), 811.

[17] Google. 2017. The Google Maps Geolocation API. https://developers.google.com/maps/documentation/geolocation. (2017).

[18] Harley, Liz. 2016. White House hosts Precision Medicine Initiative Summit. http://www.frontlinegenomics.com/white-house-hosts-precision-medicine-initiative-summit/. (2016).

[19] 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*. 92–101.

[20] International Society of Genetic Genealogy. 2017. List of DNA testing companies. https://isogg.org/wiki/List_of_DNA_testing_companies. (2017).

[21] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What Is Twitter, A Social Network Or A News Media?. In *Proceedings of the 19th International Conference on World Wide Web*.

[22] 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).

[23] Kristina Lerman, Megha Arora, Luciano Gallegos, Ponnurangam Kumaraguru, and David Garcia. 2016. Emotions, Demographics and Sociability in Twitter Interactions. In *Tenth International AAAI Conference on Web and Social Media*.

[24] Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. 2011. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *CHI*.

[25] Deborah Mascalzoni, Andrew Hicks, Peter Pramstaller, and Matthias Wjst. 2008. Informed consent in the genomics era. *PLoS Medicine* 5, 9 (2008).

[26] National Human Genome Research Institute. 2017. The Cost of Sequencing a Human Genome. https://www.genome.gov/sequencingcosts/. (2017).

[27] National Intstitute of Health. 2016. What is the Precision Medicine Initiative? https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative. (24 October 2016).

[28] National Intstitute of Health. 2017. All of Us. https://allofus.nih.gov/. (2017).

[29] Harry Ostrer. 2017. How 23andMe Fell For Anti-Semitic 'Khazar' Canard. http://forward.com/opinion/382244/how-23andme-fell-for-anti-semitic-khazar-canard/. (September 2017).

[30] Brian Pardy. 2017. Tweet: FamilyTreeDNA Privacy. http://archive.is/AUj6L. (2017).

[31] Michael J Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. *ICWSM* 20 (2011), 265–272.

[32] Andelka M Phillips. 2016. Only a Click Away? DTC Genetics for Ancestry, Health, Love? and More: A View of the Business and Regulatory Landscape. *Applied & translational genomics* 8 (2016), 16–22.

[33] 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 *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*. 55–63.

[34] Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval. (1986).

[35] Internet Live Stats. 2017. Internet Users by Country (2016). http://www.internetlivestats.com/internet-users-by-country/. (2017).

[36] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment Strength Detection In Short Informal Text. *Journal Of The American Society for Information Science and Technology* 61, 12 (2010), 2544–2558.

[37] Twitter. 2017. Hateful conduct policy. https://support.twitter.com/articles/20175050. (2017).

[38] US Food & Drug Administration. 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. (6 April 2017).

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