

# Whose Tweets are Surveilled for the Police: An Audit of a Social-Media Monitoring Tool via Log Files

Glencora Borradaile  
School of Electrical Engineering and  
Computer Science  
Oregon State University

Brett Burkhardt  
School of Public Policy  
Oregon State University

Alexandria LeClerc  
School of Electrical Engineering and  
Computer Science  
Oregon State University

## ABSTRACT

Social media monitoring by law enforcement is becoming commonplace, but little is known about what software packages for it do. Through public records requests, we obtained log files from the Corvallis (Oregon) Police Department's use of social media monitoring software called DigitalStakeout. These log files include the results of proprietary searches by DigitalStakeout that were running over a period of 13 months and include 7240 social media posts. In this paper, we focus on the Tweets logged in this data and consider the racial and ethnic identity (through manual coding) of the users that are therein flagged by DigitalStakeout. We observe differences in the demographics of the users whose Tweets are flagged by DigitalStakeout compared to the demographics of the Twitter users in the region, however, our sample size is too small to determine significance. Further, the demographics of the Twitter users in the region do not seem to reflect that of the residents of the region, with an apparent higher representation of Black and Hispanic people. We also reconstruct the keywords related to a Narcotics report set up by DigitalStakeout for the Corvallis Police Department and find that these keywords flag Tweets unrelated to narcotics or flag Tweets related to marijuana, a drug that is legal for recreational use in Oregon. Almost all of the keywords have a common meaning unrelated to narcotics (e.g. broken, snow, hop, high) that call into question the utility that such a keyword based search could have to law enforcement.

As social media monitoring is increasingly used for law enforcement purposes, racial biases in surveillance may contribute to existing racial disparities in law enforcement practices. We are hopeful that log files obtainable through public records request will shed light on the operation of these surveillance tools. There are challenges in auditing these tools: public records requests may go unfulfilled even if the data is available, social media platforms may not provide comparable data for comparison with surveillance data, demographics can be difficult to ascertain from social media and Institutional Review Boards may not understand how to weigh the ethical considerations involved in this type of research. We include in this paper a discussion of our experience in navigating these issues.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

FAT\* '20, January 27–30, 2020, Barcelona, Spain

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6936-7/20/02...\$15.00

<https://doi.org/10.1145/3351095.3372841>

## CCS CONCEPTS

• **Social and professional topics** → **Governmental surveillance; Corporate surveillance; Race and ethnicity**; • **General and reference** → *Empirical studies*;

## KEYWORDS

social media monitoring, surveillance, police, demographics, keywords, audit

### ACM Reference Format:

Glencora Borradaile, Brett Burkhardt, and Alexandria LeClerc. 2020. Whose Tweets are Surveilled for the Police: An Audit of a Social-Media Monitoring Tool via Log Files. In *Conference on Fairness, Accountability, and Transparency (FAT\* '20)*, January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3351095.3372841>

## 1 INTRODUCTION

Law enforcement use of social media monitoring software has been in the news for several years, and usually it is not good news. The ACLU of Northern California reported that MediaSonar, used by the Fresno Police Department, encouraged police to track #BlackLivesMatter and related hashtags to identify “threats to public safety” [34]. After it was revealed that MediaSonar marketed itself as a way for police to “avoid the warrant process,” Twitter cut off the company’s access to their enterprise API [36]. Twitter also cut SnapTrends’ API access after the release of details of law enforcement use of their software; SnapTrends closed shop shortly thereafter [15]. Geofeedia was notably used during the Freddie Gray uprisings to “arrest [protesters] directly from the crowd” aided by social media posts and face recognition technology [23]; shortly after this revelation from the ACLU of Northern California, Facebook, Twitter and Instagram all revoked API access from Geofeedia [35]. Both SnapTrends and Geofeedia are known to have enabled “undercover” accounts that befriend Facebook super-users in order to bypass users’ privacy settings [15]. During a trial period of DigitalStakeout, an agent of the Oregon Department of Justice searched for #BlackLivesMatter, discovered that an Oregon DOJ attorney was tweeting support and wrote a memo describing the posts as “possible threats towards law enforcement” – the agent who wrote the memo was later found to be in violation of state law [14].

The usefulness of social media monitoring has been called into question. Conarck reports that social media monitoring in Jacksonville, FL by Geofeedia “included largely protected free-speech activity and useless miscellanea” [12]. Relevant to the monitoring of social media in Corvallis, OR, in February 2018, an individual was arrested for Tweets threatening a shooting on the Oregon State University’s Corvallis campus. However, the Tweets were not discovered through surveillance of social media but through an

anonymous tip line [43]. Indeed, our work echoes that of Conarck, uncovering that DigitalStakeout uses simple keyword search, at least on the topic of Narcotics, and that almost all the keywords have benign drug meanings that uncover “useless miscellanea.”

Police increasingly utilize social media. A 2015 survey of over 500 US police departments found that 94% of agencies had used social media in some capacity—to notify the public, recruit employees, gather intelligence, manage reputations, or other. The survey found that 89% of agencies had used social media tools to further criminal investigations [24]. Further, a 2016 report by the Brennan Center for Justice identified 151 local and state law enforcement agencies in the United States that have subscribed to social media monitoring services. These jurisdictions partner with a variety of private firms that deliver the monitoring service, including Geofeedia, Media Sonar, SnapTrends, Dataminr, DigitalStakeout, and Babel Street [40]. What is known about social media monitoring technology is mostly gleaned from documents obtained through public records requests but these documents are often limited to marketing and training materials. Meanwhile, the technologies are proprietary, and details of the underlying algorithms are unknown.

In this paper, we seek to understand how social media surveillance software may place certain groups of users under undue scrutiny. Pew Research reports on the racial and ethnic, gender and age biases across the many social media platforms [1]. Sloan and Morgan report further demographic differences (in terms of gender, age, class and language) that exist among Twitter users as to whether they opt to geotag their tweets [45]. We ask: Do these biases combine to unduly focus attention on certain users? Does the software introduce biases that cannot be explained by a disparity in how different groups use social media? We find that the demographics of the users whose Tweets are flagged by DigitalStakeout are representative of the demographics of the Twitter users in the region, but may not reflect that of the residents of the region, with an apparent higher representation of Black and Hispanic people.

To understand law enforcement monitoring of social media, we made public records requests to agencies asking for logs from social media monitoring tools. We show that it is possible to reverse engineer the operation of keyword-based social media monitoring using log files. We also show that we can audit the software [42], using the limited log files, for potential demographic disparities from the use of social media monitoring. Because the data size is small and comes from a single jurisdiction, we are limited in the scope of questions we can answer. However, this study provides a proof of concept and highlights areas for future study.

## 1.1 Overview: From data to defining the research questions

In the summer of 2017, we sent public records requests to 10 agencies listed by the Brennan Center as having (had) access to DigitalStakeout: Allentown Police Department, Alpharetta City Police Department, Corvallis Police Department, Fort Worth Police Department, Georgia Bureau of Investigation, Hillsboro County Sheriff’s Office, Indiana State Police, Oregon State Police, Scottsdale Police Department, and Yakima Police Department. We chose DigitalStakeout as a case study because it is a social media monitoring software package that was not reported to be subject to API restrictions by

social media platforms (as MediaSonar, SnapTrends and Geofeedia were), is still actively used and had the largest number of listed subscribing agencies in the Brennan Center report. Initially these requests were not made with a specific research question in mind, but more generally seeking to understand the use of social media monitoring software. As part of the public records request, we asked for “logs of searches that have been input into DigitalStakeout” and “debug logs produced by DigitalStakeout.”

Several departments have claimed criminal investigatory material exemptions to public records laws (for which we are still seeking research exemptions to that exemption) and at least two agencies did not have records to release: Oregon DOJ did not subscribe to DigitalStakeout after their trial run (and now reports a policy of not subscribing to social media monitoring software) and the Yakima Police Department reports that their officers did not use the software and no longer subscribe. The Corvallis Police Department did furnish logs in the form of .csv files which consist of 7240 links to social media posts, with some additional meta-data. We describe the data in more detail in Section 2.

Upon initial examination of the data, we observed that: more people of color seemed to be represented in the collected social media posts than in Corvallis, and the collected social media posts largely did not seem to be relevant to law enforcement. These observations lead to the research questions:

- (1) Are the demographics of the social media users identified by DigitalStakeout representative of social media users or of the target population? (Section 5)
- (2) How are the social media posts being identified by DigitalStakeout? (Section 6)

At this point, we sought guidance from our IRB on how to responsibly pursue these questions. We describe our procedure for demographic coding in Section 3. An analysis of the racial and ethnic demographics are given in Section 5. A look at the keywords used to flag social media posts is given in Section 6. We describe our navigation of the ethical issues of this work in Section 7.

## 1.2 Related work

While the extent of social media monitoring has been reported in the news, there is little work in the academic literature on the impacts of this. The University of Chicago Crime Lab’s report on using information gathered via social media to identify high school students for social service intervention is an exception, but it is not clear how they are monitoring social media [48]. As far as we know, no work in analyzing the actual tools used for monitoring social media has appeared in the academic literature. The closest related work to ours is that which seeks to understand the algorithms and tools used for predictive policing, recidivism prediction, and face recognition. We discuss work related to the demographic coding of social media users (which we do in this study) in Section 3.

Platforms for predictive policing may incorporate social media (such as Palantir [7]), but the impact of social media in predictive policing decisions has not been studied. A simpler system for predictive policing, PredPol (whose basic algorithm is known and takes as input arrest data and reported incidents) has been the object of academic study. Lum and Isaac demonstrate the existence and describe the potential consequences of feedback loops in PredPol [33]

and Ensign et al. prove why these loops occur and suggest ways to prevent them [18].

More closely related to our work is that of understanding COMPAS, a tool used to predict recidivism and used in parole decisions. COMPAS hit the media after a ProPublica expose argued bias against black defendants [27]. ProPublica published their full data set of defendants, their demographics and abbreviated arrest history, and COMPAS scores, which they obtained through public records requests. This data set allowed several academic teams to follow up with more in-depth statistical analyses and explanations (both supporting and criticizing the original analysis) [10, 16] and the development of theoretically-grounded models to explain and further understand the data [2, 21].

Others have studied face recognition algorithms (which are increasingly being used in policing [11]) and have shown lower accuracies for younger, dark-skinned, or feminine faces [8, 29].

Similar to these studies, we examine how software may introduce bias into law enforcement. However, whereas the previous studies relied on data derived from administrative records and physical appearance, the present study considers how social media users' online actions may expose them to more or less law enforcement attention. In particular, this study relies on public Twitter data. Such data are necessarily consciously curated by individual users, and the resulting public presentations may convey varying degrees of information. Thus, while social media monitoring presents one new method of surveilling citizens (perhaps differentially), it poses new challenges for how to define individual group membership and subsequently measure aggregate levels of surveillance among different sub-populations.

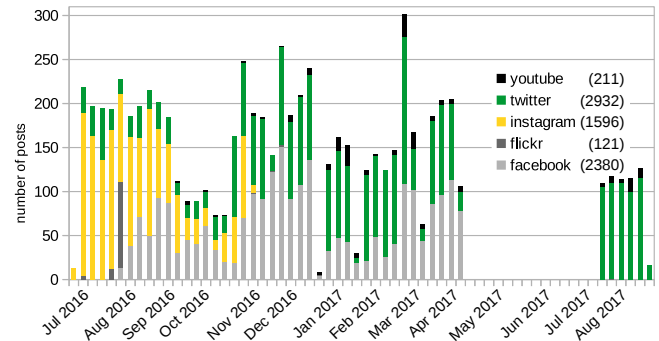
## 2 DESCRIPTION OF THE DATA

The *search logs* used in our audit contain the results of automated searches defined by DigitalStakeout (rather than the police department). They consist of 7240 links to social media posts, with some additional meta-data, over a period of 13 months (with a 3 month gap); see Figure 1. Also furnished by the Corvallis Police Department were additional *use logs* documenting officer-initiated inquiries in DigitalStakeout. Our IRB denied our request to address research questions towards these additional use logs. However, the use logs show that the Corvallis Police Department used DigitalStakeout infrequently and did not access the results of the automated searches defined by DigitalStakeout.

DigitalStakeout did not respond to our request for a demonstration, but the Corvallis Police Department did describe the system to us. DigitalStakeout is provided as a subscription software that the Corvallis Police Department accesses through a web portal. It provides three main ways to navigate social media, all within the predefined geographic region: (1) a map of the region with pins corresponding to recent posts of interest, (2) a search box for searching by name or screen name, and (3) an “intelligence discovery” tool that presents links referencing the geographic region of interest from the last hour. A drop down menu gives access to posts captured by automated searches. The use logs show that the Corvallis Police Department did not access the results of the automated searches.

As we see from Figure 1, DigitalStakeout seems to inconsistently access all social media platforms except for Twitter. From colleagues

FAT\* '20, January 27–30, 2020, Barcelona, Spain



**Figure 1: Number of posts per week in the DigitalStakeout search logs from the Corvallis Police Department categorized by social media platform (total in brackets).**

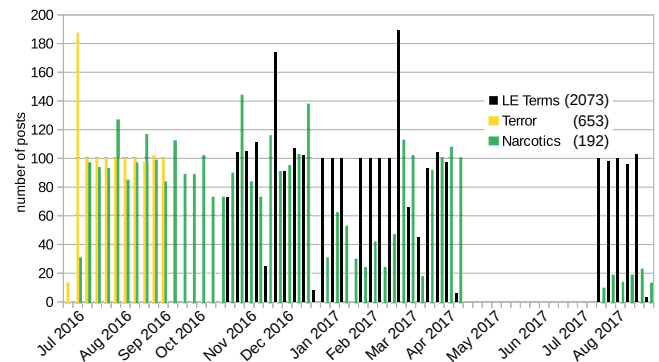
at the Brennan Center, we understand that the access to the Facebook API was pulled for all social-media monitoring platforms in Spring of 2017.

### 2.1 Description of the search logs

The data furnished by the Corvallis Police Department that we study here consists of 83 spreadsheets in comma-separated form, broken into 3 groups corresponding to different sets of search terms: LE (Law Enforcement) Terms, Terror Report, Narcotics. The explanation accompanying the records request was that these are the results of search terms and (according to the Corvallis Police Department) “all the search terms are preset proprietary lists of terms [DigitalStakeout] searches for.” The columns of each spreadsheet include URL and TIME. From July through early September, the Narcotics search logs include keywords for each social media post. (We describe this in more detail in Section 6.)

We note that for *LE Terms* and *Terror Report*, in many weeks exactly 100 search results are listed (Figure 2). Indeed, in each of these weeks, the search logs seem to indicate that the search process collects social media posts until 100 results are returned then deactivates for the remainder of the week.

The vast majority of the social media posts are “unprotected,” arising from public accounts, so we believe DigitalStakeout to only



**Figure 2: Number of Tweets per week in the DigitalStakeout search logs from the Corvallis Police Department categorized by search term set (total in brackets).**

be accessing publicly available data. (We posit that those accounts that are currently not publicly available were made protected in the time since the posts were collected by DigitalStakeout.) Herein, we focus on the subset of 2932 social media posts in the search logs that originate from Twitter, as Twitter’s API allows us to sample comparative data sets (as we describe in Sections 2.3 and 6.1). We only analyze data that is available on the date that we collect comparative samples (Table 1).

	# tweets available		# users available	
	# tweets	07/12/19	# users	05/31/18
Terror	192	101	86	74
Narcotics	653	549	225	195
LE Terms	2073	1442	774	595

**Table 1: Available data for analysis. We only use data that is publicly available on the date that we collect comparative samples.**

## 2.2 Geography of the DigitalStakeout Tweets

In conversation with the Corvallis Police Department, we learned that DigitalStakeout was calibrated to search social media posts originating from Benton County, Oregon (where Corvallis is located), but returned results outside of this county, notably from Benton County, Washington.

We examined the geotags of the Tweets and profile locations of the corresponding users. All the Tweets in the Narcotics and Terror Report sets are geotagged with coordinates which lie within the 5 mile radius of Corvallis. We presume that for these searches, DigitalStakeout is limiting the search of Twitter to Tweets with coordinate geotags in this region. (Note that Tweets can also be tagged with a “place”, which is a more general geographic region.)

On the other hand, the Tweets in the LE Terms data set seem to be collected with Twitter’s “place” search. The argument to a place search is either an ID or a place name. A place ID is a precise identifier for a unique geographic region. Given an ID, Twitter’s place search will return Tweets: with geotags in that region; tagged with a place within that region; or, from a user with a profile location within that region if the Tweet has no place or coordinate geotag. On the other hand, a place name is imprecise and Twitter will match that name to any geographic region that closely matches it. It appears that the LE Terms search was configured with a descriptor relating to Benton County as the LE Terms data set contains Tweets with place and profile locations in Benton City, Washington and Bentonville, Arkansas. Further, the LE Terms data set includes many “retweets”, which in the past had geotags, but no longer do. The current Twitter API will not pull retweets through a place search. The poor configuration of the LE Terms search and change in the Twitter API has prevented us from being able to reproduce DigitalStakeout’s geographical query for the LE Terms search.

## 2.3 A comparative dataset of Corvallis Geotagging Tweeters

To understand the demographics of the users whose Tweets appear in the DigitalStakeout data, we collect a sample of Tweets geotagged within the 5 mile radius of Corvallis, OR, matching the inferred geographical constraint for the Narcotics and Terror Report DigitalStakeout search results. We were restricted to doing so with the public Twitter API, as our interest in studying the demographics of the account users violates Twitter’s Premium API agreements.<sup>1</sup>

Twitter’s public API geotag filter takes as input an input polygon and returns any Tweet whose geotag intersects that polygon. The geotag of a Tweet is either a point (e.g. a GPS coordinate) or a polygon (rectangular, bounding a city, state or country, for example). For our filter, we used the smallest rectangle encompassing Benton Co., OR. We refiltered the collected tweets to those whose geotags were points within Corvallis’s 5 mile radius, and matching the behavior of the Twitter Premium API geotag filter we infer was used to curate the DigitalStakeout data for the Narcotics and Terror Report searches. We collected 1961 Tweets between March 6, 2018 and May 22, 2018. From this, we sampled a set of 949 Twitter accounts to use in our comparative analysis<sup>2</sup>; each account was selected for this final set with probability proportional to their frequency of Tweeting. We call this set of accounts the Corvallis Geotagging Tweeters. Of these 949 Twitter accounts, 102 accounts also appeared in the DigitalStakeout data set.

## 3 DEMOGRAPHIC CODING

There is a body of literature on extracting demographic information from social media users in support of sociological and public health research that would be possible from the wealth of information available on social media platforms [49]. Cesare, Grant and Nsoesie discuss in detail many of the issues involved in inferring the demographic information from social media users, including the issue we dealt with most prominently: “One challenge associated with the prediction of race and ethnicity is the need to create a clear, bounded definition. Racial and ethnic identity is complex and evaluations by others may not match an individuals’ self-identification.” [9]

Cesare et al. also review 60 studies aimed at automatic detection of demographics, either using simple data detection (e.g. from profile descriptions) or matching (e.g. to user names) or machine learning techniques. However, these techniques often limit the metadata they use to just profile photos or names in order predict demographics, and doing so limits the fraction of profiles for which demographics would be determinable. For example, those methods that use profile photos to infer demographics only classify users

<sup>1</sup>Twitter’s Master License Agreement which one needs to sign to gain access to the Twitter Premium API “may not be used [...] to target, segment, or profile individuals based on health (including pregnancy), negative financial status or condition, political affiliation or beliefs [...] racial or ethnic origin”. Of course, Twitter’s current Master License Agreement also seems to preclude social media monitoring itself: “may not be used by [...] any public sector entity (or any entities providing services to such entities) for surveillance purposes, including but not limited to: (a) investigating or tracking Twitter’s users or their Content; and, (b) tracking, alerting, or other monitoring of sensitive events (including but not limited to protests, rallies, or community organizing meetings).”

<sup>2</sup>Initially we sampled 1000 accounts with Tweets within Benton Co.’s bounding box and readjusted the sample upon observing the different configuration for the LE Terms search.

with a profile photo containing a single face that they presume to be the profile owner. In doing so, An and Weber discard 50% of profiles [25] and Messias, Vikatos, and Benevenuto discard 68% of profiles [37]. In our case, since we are dealing with limited data and wish to avoid introducing any biases that may exist in users opting to use a profile photo that is of themselves, these approaches are not appropriate.

The automatic detection methods that rely on machine learning techniques need a training set of data to seed the work. In some cases this is generated from an external source (such as MySpace’s self-reported names and ethnicities [26] or mugshots.com arrest records [44]), or through using a secondary machine learning algorithm as a black box (such as Face++ [37]), or through manual coding much like we describe below. While some groups perform in-house manual demographic coding similar to our own [5, 17], McCormick et al. recommend using Amazon Mechanical Turk for this task (and report on the reliability of doing this) [49]. However, since our IRB determined that our work was human subjects’ research and requested a high level of data security, passing our data to MTurk workers would violate our approved protocols.

In the time since we completed our demographic coding, Preotiuc-Pietro and Unger presented a method which infers race and ethnicity from social media text and only require 100 posts from an account to predict demographics [39]. Their model was robustly trained on a data set the authors built of users who self-report their race/ethnicity through a survey. While the accuracy claims are quite strong, the authors have not responded to a request for access to their method.

### 3.1 Protocol

We coded all the DigitalStakeout accounts and Corvallis Geotagging Tweepers using the following protocol. Before coding, we mixed the two data-sets, removed duplicates, and randomized the order of the accounts. We did this for two reasons. First, this would eliminate any bias that may be introduced from knowing that an account is or is not in the surveilled data. Second, this provides some amount of privacy for the account holder from research scrutiny resulting from having been picked up by a surveillance tool. We discuss this second point further in Section 7.

We used the following publicly-available information to classify the gender and race of users with unprotected Twitter accounts:

- Name on the account (Twitter handle and profile name)
- Profile and banner photo
- Biography section (including links to external pages)
- Recent tweets
- Photos and videos available via the left sidebar

Using this information, coders first indicated whether the account belonged to an individual or an organization (e.g., company, band, school, group, etc.).

For individual accounts, coders classified the users’ gender using the categories {Female, Male, Other (for users who *self-identify* as non-binary, gender fluid, transgender, genderqueer, or third gender), Don’t know (if there is no image or text to indicate gender)}.

Coders then classified the users’ race using the categories: {*White; Hispanic; Latino, or Spanish; Black or African American; Asian; American Indian or Alaska Native; Middle Eastern or North African; Native*

*Hawaiian or Other Pacific Islander; Other (including users who self-identify as multiracial); Don’t know (if there is no image or text to indicate race or ethnicity)}*.

In our protocol, coder’s looked first for positive evidence, such as self-identification, and then relied on photos or language in the absence of self-identification.

As shown below, several demographic categories appeared rarely, if at all, in the Twitter data. For the sake of more robust statistical comparisons, some analyses below collapse these race categories to, for example, {*White; Black; Hispanic; Other; Don’t Know*}.

Gender and race are fluid and socially constructed categories, and there are other possible ways of categorizing the gender and race of users. However, we believe these categories provide a reasonable, though necessarily simplified, reflection of race and gender divisions in the US. Importantly, we determined that different coders following this protocol could reliably classify the race and gender of users. Our protocol is similar to that used in other studies [5, 17, 49].

## 4 INTER-RATER RELIABILITY

We established the reliability of the coding protocol using a sub-sample of 99 accounts from the randomly ordered, mixed data set. Three coders (the authors and an undergraduate research assistant) applied the protocol to this sub-sample.

We measure the reliability of our coding protocol by measuring the inter-rater reliability of our three coders using Krippendorff’s alpha. Krippendorff’s  $\alpha = 1 - D_o/D_e$  where  $D_o$  is the observed disagreement between the coders and  $D_e$  is the disagreement that would be expected by chance. We evaluated the resulting reliability measures using the benchmarking method proposed by Gwet [28] (as implemented in Stata by Klein [13]), indicating which of the Landis and Koch [31] levels of agreement are met with probability at least 95%. As indicated in Table 2, our coders achieved at least a substantial level of agreement on the Landis and Koch scale for all our coding dimensions.

The levels of inter-rater reliability are reported in Table 2. In coding Twitter accounts, the coders used a single label with three choices: {Not accessible (protected, deleted, or suspended), Individual, Organization}. The first of these three choices can be done programmatically, but as we have observed, accounts become inaccessible or accessible over time as accounts are suspended and reinstated or made protected over time. Since coding was not completed simultaneously, we opted to include “Not accessible” in the first feature, with “Individual” and “Organization”. However, the choices for this first feature impact our measure of inter-rater reliability for gender and race/ethnicity. When measuring inter-rater reliability for gender and race/ethnicity, we included all subjects

	Indiv?	Gender	Race & Ethnicity	
# categories	3	4	7	3
# subjects	99	70	70	70
$\alpha$	0.940	0.884	0.703	0.785
reliability	almost perfect		substantial	

Table 2: Inter-rater reliability

that had at least two coders select “Individual” for the first feature. In order to understand the reliability of coding for race and ethnicity, in addition to considering the full set of 7 categories, we also considered a collapsed set of categories consisting of {White, Not White, Don’t Know} where “Not White” consists of all remaining race and ethnicity categories.

With the inter-rater reliability established, one of the three test coders (an undergraduate research assistant), coded the full, randomized, mixed data set. These codes serve as the basis for the analyses below.

## 5 DEMOGRAPHICS: RACE AND ETHNICITY

We report on our coded demographics for race and ethnicity for the Corvallis Geotagging Tweepers and those in the DigitalStakeout data who were coded as “Individuals” in our protocol. We do not include users for whom there were neither images nor text to indicate race or ethnicity in these counts (which were coded “Don’t know” according to our protocol). In Table 3, we reduce the number of categories of race and ethnicity since the number of users coded in several categories were very small; in Table 3, “Other” encompasses several under-represented minorities: {Asian, American Indian or Alaska Native, Middle Eastern or North African, Native Hawaiian or Other Pacific Islander, Other (including users who self-identify as multiracial)}. The full table of demographics is in Appendix A. We summarize gender demographics in Appendix B.

Users in the DigitalStakeout Narcotics and Terror Report data sets are drawn from Twitter in the same way as Corvallis Geotagging Tweepers. We ask, are the users in the Narcotics and Terror Report data sets representative samples of Twitter users who geotag in Corvallis? The p-values reported correspond to a Pearson Chi-squared test between CVI and Narc, Terr, N+T, respectively,

	Corvallis Geotagging	DigitalStakeout		
	Tweepers	Narc.	Terr.	N+T
<i>n</i>	788	148	47	180
White	71.8%	78.4%	83.0%	78.9%
Black	6.5%	7.4%	4.3%	7.2%
Hispanic	11.7%	7.4%	10.6%	7.8%
Other	10.0%	6.8%	2.1%	6.1%
<i>p</i>	-	0.23	0.25	0.13

Note: 15 users appear in both Narc and Terr.

**Table 3: Coded Demographics—Narc. & Terr.**

<i>n</i>	242
White	84.3%
Black	4.1%
Hispanic	5.8%
Other	6.0%

**Table 4: Coded Demographics—LE Terms**

for the race categories given in Table 3. In each comparison, the race distributions differ, with white users appearing at higher rates in the DigitalStakeout sample than in the Corvallis Geotagging Tweepers sample. However, these differences are not statistically significant, a fact due in part to the small number of users in the DigitalStakeout samples. We assume that the demographic distribution of geotagging Twitter users in Corvallis has not changed significantly from when the DigitalStakeout data was collected to when the Corvallis Geotagging Tweepers were sampled.

As described above, the DigitalStakeout LE Terms data set seems to be drawn in a more general way that includes profile locations, and due to presumed poor configuration, includes users that seem to be from outside of Corvallis, OR. Using the Twitter API, we examined the geotags of the Tweets in the LE Terms data set (if available) and user-described profile locations (as recorded on May 31, 2018 through the Twitter API). We classified interpretable account profile locations according to whether they correspond to locations within Corvallis, OR or not. (A profile location is non-interpretable if they did not correspond to mappable locations (such as *the moon* or *bliss*)). The coded demographics of the accounts in the LE Terms data set that have Tweets geotagged or profile locations in Corvallis, OR are given in Table 4 (for the reduced set of categories – full data given in Appendix A). These accounts would represent the same search, but configured for the Corvallis Police Department’s region of interest.

### 5.1 Demographics of the local population

The demographics represented in Table 3 are notably different from that of the Corvallis, OR (pop. 54,462) given in Table 5. The census considers race orthogonal to ethnicity. We give the fraction of the

	only	only or in combo.	Hispanic
White	83.8%	87.5%	3.9%
Black	1.1%	1.8%	0.1%
Native American	0.7%	1.8%	0.2%
Asian	7.3%	9.3%	-
Native HI/Pac. Isl’r	0.3%	0.8%	-
some other race	2.8%	3.2%	2.6%
two or more races	4.0%	N/A	0.6%
Total	100.0%	N/A	7.4%

**Table 5: Corvallis 2010 Census Demographics**

<i>n</i> (arrests)	22,875
White	89.4%
Black	3.5%
American Indian	0.8%
Asian	1.7%
Unknown	4.6%
Total	100.0

**Table 6: Benton County arrests, 2007-2012**

population that identifies as a *single* given race (column “only”), the fraction of the population that identifies as a given race (alone or in combination with any other race), and the fraction of the population that identifies as Hispanic as well as any *single* given race.

Corvallis is also home to Oregon State University (OSU), with Spring 2018 enrollment of 28,568 students, 4,916 of which attended via e-campus alone. OSU reports demographics of their domestic students, but not of the 11.72% of the students who are international [38]. The demographics of the domestic students at OSU is similar to that of Corvallis demographics.

Arrests made by local police represent another relevant point of comparison, as they represent the population of local residents who are formally brought into the criminal justice system. According to data published by Lanfear [32], the demographics of arrests roughly mirror the demographics of the population. In particular, nearly 90% of arrests made by Corvallis Police Department or Benton County Sheriff’s Office from 2007-12 involved a White suspect (see Table 6).<sup>3</sup>

## 5.2 Differences between Twitter users and the broader population

We wish to comment on the apparent difference in demographics between Twitter users and (Table 3 and 4) and Corvallis residents (Table 5). It is impossible to determine the source of these differences, as there are many reasons to expect the differences we see in the demographics of these populations as (i) race is a significant factor for explaining difference in behavior, (ii) externally assigned (coded) demographics are a highly imperfect proxy for self-identified demographics, and (iii) the demographic categories we used for coding Twitter users are not perfectly comparable to Census categories.

To comment further on the first point, we refer to relevant literature and surveys which aim to quantify the demographic factors that play a role in social media use.

There are racial and ethnic differences in what social media platform people use. For example: 28% of Black U.S. adults use LinkedIn versus 13% of Hispanic; 49% of Hispanic U.S. adults use WhatsApp versus 14% of White [1]. While Pew’s report that 24% of White people, 26% of Black people, and 20% of Hispanic people use Twitter does not explain the differences in the demographics of the populations we observe [1], there are two further considerations: First, Pew’s survey is nationwide, and there may be regional differences that compound racial and ethnic differences in social media use. Second, Pew’s survey does not drill down into how people interact with a given social media platform. In particular, there could be racial and ethnic differences in whether people opt to geotag their Tweets. Very few Tweets have geotags (measured at 0.85% in 2013 [46]), and Sloan and Morgan show that prevalence of geotagging varies among users depending on their gender, age, class and language [45].

There are significant differences in Twitter use according to age: 45% of 18-24 year-olds use Twitter compared to only 14% of

those over 50 [1]. In a college town like Corvallis, this issue will be compounded.

Finally, we note that the Twitter users represented in Table 3 are gathered purely based on geotags and that geotags will pick up Tweets from users who are simply visiting the area and not resident in Corvallis.

## 6 KEYWORDS

In order to understand how the social media posts are being identified by DigitalStakeout, we attempt to reverse engineer the search. We do so only for the “Narcotics” search. Of the 101 Tweets that are still available (not deleted) in the “Terror Report” data set, 25 contain videos or images and 57 contain urls – one Tweet contains only an image. Given the limited and type of data and the likelihood that this search is not simply defined by a keyword search, we are not able to explain how the “Terror Report” is generated. For “LE Terms”, as previously noted, we are unable to reproduce the geographic filter.

The Narcotics dataset includes partial meta-data that suggests a simple keyword search is being employed: for the first 2 months of *Narcotics* search results, each social media post is accompanied by a set of keywords that match or closely match a word in the Tweet. This seems to employ Twitter’s keyword search which is more general than exact keyword matching: e.g., searching Twitter for “hop” will return Tweets containing the word “hopped” but not “hope”. We cluster keywords into keyword variant groups if they are variants of each other such as *rock*, *rocked*, *rocking*, *rocks*. We use the simplest version in the group as a “root” representative (although it may not be a formal linguistic root).

There are 39 *known keywords* across 36 variant groups (listed in Table 7) that appear in the meta-data of the Narcotics dataset. The known keywords explain 68% of the available “Narcotics” Tweets. We aim to uncover keywords that explain the remaining Tweets and develop a process that would reliably identify keywords should such meta-data not be available.

### 6.1 A comparative dataset of historical Tweets

To understand how DigitalStakeout identifies Tweets in their Narcotics search, we collected the historical Tweets geotagged within the 5 mile radius of Corvallis over the same time period as the DigitalStakeout data using Twitter’s Premium API. The DigitalStakeout search logs suggest that there are time periods when the searches are not active in addition to the 3 month period for which we have no DigitalStakeout data such as gaps in time between search log files. To most conservatively represent the possible input accessed by DigitalStakeout, we down-sample the set of historical tweets to those with time-stamps between the first and last time stamps in a given search log for the Narcotics search. This is imperfect, as there may be times during the creation of a search log in which the DigitalStakeout software is not active or in which the Twitter API is down.

### 6.2 Reverse engineering keywords

In order to reverse engineer the keywords used by DigitalStakeout, we compare the the (presumed) input to DigitalStakeout to the output from DigitalStakeout:

<sup>3</sup>The data omit Oregon State Police, which has jurisdiction over the Oregon State University campus.



- $T_{in}$  The set of Tweets obtained with the same (presumed) geographical filter over the same period of time as the DigitalStakeout Narcotics search (as described in Section 6.1).
- $T_{out}$  The set of Tweets in the DigitalStakeout Narcotics search log that are still publicly available.

Let  $P$  be the set of words that appear only in DigitalStakeout Tweets; that is, words that appear in a Tweet of  $T_{out}$  but not in a Tweet of  $T_{in} \setminus T_{out}$ .  $P$  is the set of *possible* keywords. If the search is active for all the periods of time that cover  $T_{in}$  and keywords are matched consistently, then a keyword must be in  $P$ . Unfortunately, data is rarely *perfect*. We find that 4 of the 39 known keywords (“yay”, “broken”, “trip”, “tracks”) are not in  $P$ . “Broken” is in 4 Tweets of  $T_{in} \setminus T_{out}$ ; “yay”, “trip”, and “tracks” are each in 1 Tweet of  $T_{in} \setminus T_{out}$ . This could be explained by the Twitter API or DigitalStakeout’s services being down during this time period and not making data available for collection at the time of DigitalStakeout’s collection. That a Tweet with the word “broken” is missed 4 times is not surprising, as “broken” is overall a very high frequency word; indeed, “broken” is contained in 34% of the available Narcotics Tweets.

Each Tweet in the Narcotics set contains at least one word of  $P$ . For a Tweet in the Narcotics set that contains exactly one word  $w$  from  $P$ , we presume that  $w$  is the keyword that returned this Tweet. We call the set of all such words in  $P$  the set of *necessary* keywords and denote it  $N$ . Given perfect data and exact keyword matching, all words in  $N$  must be keywords. For more general keyword matching but otherwise perfect data, all words in  $N$  must be keywords (or variants of keywords).

We call the set of Tweets in the Narcotics set that do not contain a word or variant of a word in  $N$  the set of *unexplained* Tweets. Each Tweet in this set contains at least two words from  $P$  and at least one word from  $P \setminus N$ . Determining which words in  $P \setminus N$  are keywords or derived from keywords is an impossible task. The problem is a hitting set problem: find a subset  $K$  of  $P$  such that every Tweet contains a word of  $K$ . There may be multiple feasible solutions, and no objective to decide between feasible solutions will necessarily correctly reverse engineer the set of keywords (or variants of keywords). However, by examining the frequencies of words across the entire Narcotics dataset, we can determine a set of *likely* keywords: words that appear with higher frequency are more likely to have been keywords for the search.

We aim for an automated and reproducible method for identifying a set  $L$  of likely keyword variant groups as follows. For each unexplained tweet, let  $W \in P \setminus N$  be a subset of words with a common root (a variant group) with highest frequency in the Narcotics dataset. We let  $L$  be the union of such variant groups that explain at least 2 DigitalStakeout Tweets ( $T_{out}$ ) that are not already explained by  $N$ . We use a threshold of 2 to provide some confidence; a higher threshold could be used with a larger data set.

For the Narcotics data,  $|P| = 607$ ,  $|N| = 28$ , and  $|L| = 21$ , where size measures the number of variant groups (e.g. *rock*, *rocked*, *rocking*, and *rocks* count as 1 variant group).  $N$  and  $L$  explain 62% of the Tweets in the Narcotics set. We give the root form of the words in  $N$  and  $L$  in Table 7 along with their frequency: the number of Tweets in  $T_{out}$  that these words (and their variants) appear in. The full list of variants corresponding to these roots are given in the Appendix.

### 6.3 Understanding the keywords

Our method of reverse engineering is relatively robust for the following reasons:

- While  $N$  and  $L$  only explain 62% of the Narcotics Tweets,  $N \cup L \cup \{\text{“broken”}\}$  explains 91% of the Narcotics Tweets. With a larger corpus of data, more noise-resistant methods would be able to capture missed words such as “broken” that were excluded as a possible keyword by relaxing the definition of *necessary*.
- All but 3 of the keywords (keg, malt, melt) in  $N$  and  $L$  are drug terms, according to the DEA Drug Slang list [3] and the Urban Dictionary<sup>4</sup>. Since our methods were oblivious to the meaning of the words, the words we uncover are quite likely to have been keywords.
- $N$  and  $L$  correctly identify 28 of 36 known keyword variant groups. Of the remaining 8, 3 were not discoverable because they do not appear in any available Tweets, 3 have frequency 1 (so a lack of data make them difficult to discover) and the remaining 2 appeared in both  $T_{in}$  and  $T_{out}$  as discussed above.
- Further of the words in  $T_{out}$  that do not appear in  $N \cup L$ , the only obvious drug term is “marijuana” which appeared in only one Tweet as part of a hashtag compounded with other words; this Tweet also contained a necessary keyword.

Note that all the keywords are English-language words or slang. This may bias the search toward English-language users. Among Twitter users geotagging in Corvallis, the effect is not large: 96.0% of the Tweets in  $T_{in}$  and 98.9% of the Tweets in  $T_{out}$  are labeled English-language by Twitter.<sup>5</sup>

Given these keywords, we feel that the Narcotics search results are unlikely to be useful for either risk assessment or sentiment analysis. Of the 56 known, necessary and likely keyword variant groups, 15 are related to marijuana [3] and that recreational marijuana has been legal in Oregon for the entire period of DigitalStakeout data collection for the Corvallis Police Department. Many of the Tweets containing marijuana-related keywords are from one of the many state-regulated marijuana dispensaries in Corvallis. Many of the keywords, although drug-related, are sufficiently general that they pick a lot of Tweets that are unrelated to narcotics. For example, “broken” picks up 178 Tweets from a weather bot that reports “broken clouds” as the forecast (and only 6 other Tweets). Variants of the word “hop” (which pertain to drug use) exclusively pick up Tweets from local breweries (of which Corvallis is home to many). Likewise “bowl” exclusively flags Tweets about the game of bowling or bowls of food. Finally, “party” picks up the variants “#kidsparty”, “#pizzaparty”, and “#birthdayparty” which are unlikely to be related to narcotics.

## 7 RESEARCH ETHICS

When we initially made our public records requests we didn’t know if we would receive any data, never mind what form that data would take, or whether it would include personally identifiable information (PII). Upon receiving the data (which includes PII in the form of links to social media posts that are created by individuals,

<sup>4</sup><https://www.urbandictionary.com/>

<sup>5</sup>98.0% of the Terror Report Tweets and 98.6% of the LE Terms Tweets are labeled English-language by Twitter.



**Table 7: Reverse-engineered and known keywords for the Narcotics search.**

Necessary								Likely						Missed Known	
Root	<i>f</i>	Root	<i>f</i>	Root	<i>f</i>	Root	<i>f</i>	Root	<i>f</i>	Root	<i>f</i>	Root	<i>f</i>	Root	<i>f</i>
snow	54	face	11	trip	4	waste	2	pie	8	indica	4	growweed	2	broken	184
hop	45	cheese	8	burger	4	gang	2	pot	6	mash	4	burn	2	yay	3
high	40	bag	8	cook	3	hustle	2	zone	6	dank	4	keg	2	hookup	1
line	22	jack	7	dope	3	rip	2	bud	6	hip	4	malt	2	stuck	1
party	22	treat	6	blow	3			fade	5	jam	4	melt	2	munchies	1
smoke	14	blast	5	load	3			dabpro	5	angel	3			stash	0
bowl	13	fried	4	wreck	3			bang	5	addict	3			track	0
rock	11	crystal	4	bake	3			deal	5	roll	3			tweed	0

Bold words are roots of known keywords. Frequency *f* is the number of available Narcotics Tweets; *f* = 0 implies the corresponding Tweets are now deleted. Missed Known words are those that were not discovered by reverse engineering.

many of whom associate their account with their real identity) and formulating our research questions, we embarked on gaining IRB approval for our research. We did so prior to pulling comparative data sets through the Twitter API.

Discussions with colleagues regarding this work received mixed opinions as to whether this research is Human Subjects Research at all, with opinions seeming to fall along disciplinary lines. In fact, our IRB took one full-board meeting to decide that question alone. In deciding whether the proposed work falls under Human Research Protection, consider the Office for Human Research Protections' guidelines for deciding "Is an Activity Research Involving Human Subjects?"<sup>6</sup> Although our work does not involve intervention or interaction with individuals, the information does include PII. Deciding whether this research is Human Subjects Research thus comes down to deciding if this information is *private*: "About behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, or provided for specific purposes by an individual and which the individual can reasonably expect will not be made public."

Although by the letter of the law, the data we used (both collected from Twitter and obtained from the Corvallis Police Department) is public (or was at the time of collection) and so may be considered exempt from IRB oversight, one also needs to consider whether someone would expect their data (and the association of their data with a given commercial/state surveillance dataset) to be used for research. While many people are aware of the extent of digital surveillance [41], few would expect their public social media posts to be collected by private company, furnished to a police department and logged, made the subject of a public records request to finally end up in the hands of a researcher. Indeed, Fiesler and Proferes report that "few [Twitter] users were previously aware that their public tweets could be used by researchers, and the majority felt that researchers should not be able to use tweets without consent." [20]

Along these lines, our IRB determined that our research is Human Subjects Research. The permission to pursue our research is under expedited categories 5 & 7 (minimal risk to adults and minimal risk to children – §46.404). We obtained a waiver of informed consent, but only by agreeing to the highest level of data protections.

<sup>6</sup><https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts/index.html#c1>

## 7.1 Waiver of informed consent

We argue that the importance of shedding light on proprietary software being used for policing outweighs the risk to an individual user that may (unknowingly) participate in this research. We also argue that this research would not be possible if informed consent was required. Not only would it likely reduce the available data significantly and introduce more biases, the very act of seeking out consent from people whose Tweets were collected by DigitalStakeout makes a data breach much more likely. We do respect users explicit choices by ignoring accounts marked as protected although these accounts still display name, profile location, profile banner and profile photos.

One cannot forget that Twitter and other social media users are not informed that they are being monitored by DigitalStakeout – it is far from common knowledge that the Corvallis Police Department subscribes to DigitalStakeout. Publication of this and similar research serves as one means of communicating this fact.

## 7.2 Data protections

In order to obtain a waiver of consent, in balancing risk and benefit, our IRB required the highest level of data protections. This involves storing information in a manner that provides access only to authorized individuals. Our data is stored on encrypted drives and shared between the study staff via end-to-end encrypted channels. Prior to demographic coding, DigitalStakeout data was randomly mixed with the Corvallis Geotagging Tweepers data. The demographically coded PII (social media user names) was kept separate from indicators as to their source. All this was to minimize the risk of leaking that a given user was included in a surveillance data set. However, this does preclude sharing the data more broadly for further study: the demographic data with search type (e.g. Narcotics) but without PII (the social media user name) may be enough information to de-anonymize users in the small community of Corvallis, much like the famed ZIP, gender, DOB de-anonymizing observation of Sweeney [47]. Of course, another researcher could request the same information from the Corvallis Police Department as we did. Alternatively, we will work with other researchers in collaboration with our IRB to ensure that further research can be pursued. As part of our currently approved IRB research protocols, we can only share

aggregate data and keywords, but not precise Tweets. Although the keywords could be used to reproduce DigitalStakeout “Narcotics” searches, since the DigitalStakeout searches were not running continuously and we are not publishing precise times during which the searches were active, it is impossible to perfectly reproduce the output of the DigitalStakeout search logs.

## 8 CONCLUSION

This study has shown that we can indeed use log files to audit social media monitoring software and address our research questions. First, we find small but non-significant differences in the race distribution of users flagged by DigitalStakeout and users geotagging their Tweets in Corvallis. Second, we were able to, for the Narcotics search, able to reverse engineer the keywords most likely used and show that this method is robust by comparing to a subset of keywords available through the metadata.

Racial disparities exist throughout the justice system, including in policing. Research has found that non-whites are more likely than whites to be arrested or stopped, net of legally relevant factors like crime type and presence of witnesses [6, 19, 22, 30]. These disparities in police contact contribute to a severe overrepresentation of people of color in US prisons [4, 50]. Given this, we argue that it is important to be able to audit tools used in the justice system (such as has been done for recidivism prediction, face recognition, and predictive policing) for racial disparities. Social media monitoring is simply another avenue for creating disparities, and there are many points at which an inequity could be introduced: including access to social media, adoption of a particular social media platform, interacting with the platform in a way that gives access to monitoring software, and using certain keywords. We have shown that geotagging or setting a profile location are choices that result in access by DigitalStakeout for monitoring. Others have shown that such choices are likely to correlate with demographics [46]. We have also shown that many keywords flag benign Tweets, at least from a risk-assessment perspective; this could draw undue attention from law enforcement.

Whether the purpose of social media monitoring by police is for sentiment analysis or risk assessment, unless the population that is affected mirrors that of the police jurisdiction, the bias will result in a skewed view of the population (if used for sentiment analysis) or undue attention on one subpopulation over another (in the case of risk assessment).

The use of log files is useful in gaining insight into the proprietary tools. We would recommend that log files be required and available for research or other independent evaluation to ensure transparency of the algorithms that are reshaping law enforcement. Policy could help overcome the difficulties of this audit, including lack of data, poorly or incompletely logged data and inaccessibility of data (from all entities involved, including law enforcement agencies, social media monitoring software houses and social media platforms).

*Acknowledgements.* We would like to acknowledge the help of Alexander Guyer and Baigong Zheng with data collection and generation efforts and help from the Civil Liberties Defense Center with pursuing public records requests. We would also like to acknowledge the funding support of the College of Engineering Dean’s Professorship.

## REFERENCES

- [1] Aaron Smith and Monica Anderson. 2018. *Social Media Use in 2018*. Technical Report. Pew Research Center. <http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/>
- [2] Philip Adler, Casey Falk, Sorelle Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54, 1 (Jan. 2018), 95–122. <https://doi.org/10.1007/s10115-017-1116-3>
- [3] Drug Enforcement Administration. 2018. *Slang Terms and Code Words: A Reference for Law Enforcement Personnel*. DEA Intelligence Report DEA PRB 06-13-18-25. Drug Enforcement Administration.
- [4] Michelle Alexander. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New Press, New York.
- [5] Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. 2015. Predicting the Demographics of Twitter Users from Website Traffic Data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, Austin, TX, 72–78.
- [6] Katherine Beckett, Kris Nyrop, and Lori Pfingst. 2006. Race, Drugs, and Policing: Understanding Disparities In Drug Delivery Arrests. *Criminology* 44, 1 (2006), 105–137.
- [7] Sarah Brayne. 2017. Big data surveillance: The case of policing. *American Sociological Review* 82, 5 (2017), 977–1008.
- [8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research (Conference on Fairness, Accountability and Transparency)*, Vol. 81. ACM, New York University, NYC, 1–15. <http://proceedings.mlr.press/v81/buolamwini18a.html>
- [9] Nina Cesare, Christan Grant, and Elaine Nsoesie. 2017. *Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices*. Technical Report 1702.01807. arXiv.
- [10] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (June 2017), 153–163. <https://doi.org/10.1089/big.2016.0047>
- [11] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. 2016. *The Perpetual Line-Up: Unregulated Police Face Recognition in America*. Technical Report. Georgetown Law, Center on Privacy & Technology. <https://www.perpetuallineup.org/>
- [12] Ben Conarck. 2017. Sheriff’s Office’s social media tool regularly yielded false alarms. *The Florida Times-Union* (May 2017). <https://www.jacksonville.com/news/public-safety/metro/2017-05-30/sheriff-s-office-s-social-media-tool-regularly-yielded-false>
- [13] Daniel Klein. 2016. KAPPAETC: Stata module to evaluate interrater agreement. Statistical Software Components S458283. Boston College Department of Economics, revised 01 Feb 2018.
- [14] David Rogers. 2016. *An Internal Report on Oregon’s Illegal Surveillance of Black Lives Matter on Twitter Leaves Us With More Questions Than Answers*. Technical Report. American Civil Liberties Union. <https://www.aclu.org/blog/racial-justice/race-and-criminal-justice/internal-report-oregons-illegal-surveillance-black>
- [15] Dell Cameron. 2016. *Twitter Cuts Ties with SnapTrends, a Social Media Spying Tool for Police*. Blog. The Daily Dot. <https://www.dailydot.com/layer8/twitter-snaprends-geofeedia-social-media-monitoring-facebook/>
- [16] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (Jan. 2018), 1–5. <https://doi.org/10.1126/sciadv.aao5580>
- [17] Ehsan Mohammady and Aron Culotta. 2014. Using County Demographics to Infer Attributes of Twitter Users. In *Proc. of Workshop on Social Dynamics and Personal Attributes in Social Media Proceedings of the Workshop*. Association for Computational Linguistics, Baltimore, Maryland, 7–17.
- [18] Danielle Ensign, Sorelle Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. In *Proc. of Fairness, Accountability, and Transparency in Machine Learning Workshop*, Vol. 81. ACM, New York University, NYC, 1–12.
- [19] Charles R. Epp, Steven Maynard-Moody, and Donald Haider-Markel. 2014. *Pulled over: How Police Stops Define Race and Citizenship*. University of Chicago Press, Chicago.
- [20] Casey Fiesler and Nicholas Proferes. 2018. “Participant” Perceptions of Twitter Research Ethics. *Social Media + Society* 4, 1 (Jan. 2018), 205630511876336. <https://doi.org/10.1177/2056305118763366>
- [21] Sorelle Friedler, Carlos Scheidegger, and Scheidegger Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. (Sept. 2016). <https://arxiv.org/pdf/1609.07236>
- [22] Andrew Gelman, Jeffrey Fagan, and Alex Kiss. 2007. An Analysis of the New York City Police Department’s “Stop-and-Frisk” Policy in the Context of Claims of Racial Bias. *J. Amer. Statist. Assoc.* 102, 479 (Sept. 2007), 813–823. <https://doi.org/10.1198/016214506000001040>
- [23] Geofeedia. 2016. Baltimore County Police Department and Geofeedia Partner to Protect the Public During Freddie Gray Riots. [http://www.aclunc.org/docs/20161011\\_geofeedia\\_baltimore\\_case\\_study.pdf](http://www.aclunc.org/docs/20161011_geofeedia_baltimore_case_study.pdf)

- [24] International Association of Chiefs of Police. 2015. IACP Social Media: Publications. <http://www.iacpsocialmedia.org/resources/publications/>.
- [25] Jisun An and Ingmar Weber. 2016. #greysanatomy vs. #yankees: Demographics and Hashtag Use on Twitter. In *Proc. of the 10th International Conference on Weblogs and Social Media (IWCSM)*. AAAI, Cologne, Germany, 523–526.
- [26] Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. 2010. epluribus: Ethnicity on Social networks. In *Proc. of the Fourth International Conference on Weblogs and Social Media (ICWSM)*. AAAI, Washington, DC, 18–25.
- [27] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks*. Report. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [28] Kilem Gwet. 2014. *Handbook of Inter-Rater Reliability*. Advanced Analytics Press, Maryland, USA.
- [29] Brendan Klare, Mark Burge, Joshua Klontz, Richard Bruegge, and Anil Jain. 2012. Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security* 7, 6 (Dec. 2012), 1789–1801. <https://doi.org/10.1109/TIFS.2012.2214212>
- [30] Tammy Rinehart Kochel, David B. Wilson, and Stephen D. Mastroski. 2011. Effect Of Suspect Race On Officers' Arrest Decisions. *Criminology* 49, 2 (2011), 473–512. <https://doi.org/10.1111/j.1745-9125.2011.00230.x>
- [31] J. Richard Landis and Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (March 1977), 159–174.
- [32] Charles C. Lanfear. 2013. *Exploring a Mental Health Crisis : An Examination of Mental Health Arrests in Benton County, OR*. M.S. Thesis. Oregon State University, Corvallis, OR.
- [33] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (Oct. 2016), 14–19. <https://doi.org/10.1111/j.1740-9713.2016.00960.x>
- [34] Matt Cagle. 2015. *This Surveillance Software is Probably Spying on #BlackLivesMatter*. Report. ACLU of Northern CA. <https://www.aclunc.org/blog/surveillance-software-probably-spying-blacklivesmatter>
- [35] Matt Cagle. 2016. *Facebook, Instagram, and Twitter Provided Data Access for a Surveillance Product Marketed to Target Activists of Color*. Report. ACLU of Northern CA. <https://www.aclunc.org/blog/facebook-instagram-and-twitter-provided-data-access-surveillance-product-marketed-target>
- [36] Michelle McQuigge. 2017. Experts divided on social media surveillance after Twitter pulls plug on Media Sonar. *The Hamilton Spectator* (Jan. 2017). <https://www.thespec.com/news-story/7052284-experts-divided-on-social-media-surveillance-after-twitter-pulls-plug-on-media-sonar/>
- [37] Johnnatan Messias, Pantelis Vikatos, and Fabricio Benevenuto. 2017. White, man, and highly followed: Gender and race inequalities in Twitter. In *Proc. of the International Conference on Web Intelligence (WI'17)*. IEEE/ACM, Leipzig, Germany, 1–9.
- [38] Office of Institutional Research, Oregon State University. 2018. Enrollment Summary - Spring Term. <https://institutionalresearch.oregonstate.edu/enrollment-and-demographic-reports>
- [39] Daniel Preotiuc-Pietro and Lyle Ungar. 2018. User-Level Race and Ethnicity Predictors from Twitter Text. In *Proceedings of the International Conference on Computational Linguistics*. 12.
- [40] Rachel Cohn and Angie Liao. 2016. *Mapping Reveals Rising Use of Social Media Monitoring Tools by Cities Nationwide*. Report. Brennan Center for Justice. <https://www.brennancenter.org/blog/mapping-reveals-rising-use-social-media-monitoring-tools-cities-nationwide>
- [41] Lee Rainie and Mary Madden. 2015. *How People are Changing Their Own Behavior*. Technical Report. Pew Research Center. <http://www.pewinternet.org/2015/03/16/how-people-are-changing-their-own-behavior/>
- [42] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Proc. of 64th Annual Meeting of the International Communication Association - Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. ICA, Seattle, WA, 23.
- [43] Lillian Schrock. 2018. Police arrest person suspected of threatening a shooting at OSU. *Corvallis Gazette Times* (Feb. 2018). [https://www.gazettetimes.com/news/local/police-arrest-person-suspected-of-threatening-a-shooting-at-osu/article\\_91c14e3d-71de-523e-b66e-99fd1966b304.html](https://www.gazettetimes.com/news/local/police-arrest-person-suspected-of-threatening-a-shooting-at-osu/article_91c14e3d-71de-523e-b66e-99fd1966b304.html)
- [44] Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. 2013. Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter. In *Proc. of the 2013 North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*. ACL, Atlanta, GA, 1010–1019.
- [45] Luke Sloan and Jeffrey Morgan. 2015. Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLOS ONE* 10, 11 (Nov. 2015), e0142209. <https://doi.org/10.1371/journal.pone.0142209>
- [46] Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. 2013. Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online* 18, 3 (Aug. 2013), 1–11. <https://doi.org/10.5153/sro.3001>
- [47] Latanya Sweeney. 2000. *Simple Demographics Often Identify People Uniquely*. Working paper 3. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA.
- [48] The University of Chicago Crime Lab. 2019. Connect & Redirect to Respect: Final Report. <https://urbanlabs.uchicago.edu/projects/connect-redirect-to-respect>
- [49] Tyler McCormick, Hedwig Lee, Nina Cesare, Ali Shojajie, and Emma Spiro. 2015. Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods and Research* 46, 3 (2015), 390–421.
- [50] Bruce Western. 2006. *Punishment and Inequality in America*. Russell Sage Foundation, New York.