The Effect of Population and “Structural” Biases on Social Media-based Algorithms – A Case Study in Geolocation Inference Across the Urban-Rural Spectrum

Isaac Johnson*, Connor McMahon[†], Johannes Schöning[‡], Brent Hecht*
*Northwestern University, Evanston, USA
[†]GroupLens Research, University of Minnesota, Minneapolis, USA
[‡]University of Bremen, Bremen, Germany
isaacj@u.northwestern.edu, mcmahon250@umn.edu,
schoening@uni-bremen.de, bhecht@northwestern.edu

ABSTRACT

Much research has shown that social media platforms have substantial population biases. However, very little is known about how these population biases affect the many algorithms that rely on social media data. Focusing on the case study of geolocation inference algorithms and their performance across the urban-rural spectrum, we establish that these algorithms exhibit significantly worse performance for underrepresented populations (i.e. rural users). We further establish that this finding is robust across both text- and network-based algorithm designs. However, we also show that some of this bias can be attributed to the design of algorithms themselves rather than population biases in the underlying data sources. For instance, in some cases, algorithms perform badly for rural users even when we substantially overcorrect for population biases by training exclusively on rural data. We discuss the implications of our findings for the design and study of social media-based algorithms.

Author Keywords

Algorithmic accountability; geolocation inference; population bias; social media.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

INTRODUCTION

As social media adoption has increased dramatically, research that takes advantage of this publicly-available, real-time stream of information has also become quite prevalent. In addition to facilitating new discoveries about human behavior in the social sciences and related fields (e.g., [1,12,13]), social media has also been a major catalyst in the development of new intelligent algorithms. For instance, researchers have used social media to develop new recommender systems (e.g., [38,59]), summarize the character of cities (e.g., [9,66]), and infer the location of Internet users (e.g., [34,41,45]), among other applications (e.g., [4,30,49]).

Recently, concerns about population bias in social media have been the subject of much discussion (most notably in *Science* [55]). Social media population bias, or the notion that different demographic groups may participate in social media platforms at different rates, has been found to be endemic to most social media datasets. Much work has gone into quantifying and understanding these biases, with significant biases being found along gender (e.g., [51]), age (e.g., [39]), race (e.g., [46]), income (e.g., [44]), education (e.g., [39]), and urban/rural lines (e.g., [28]).

Many researchers who use social media to understand human behavior have recognized the importance of correcting for social media population biases in their studies (e.g., [14,55]). They have also begun the critical work of developing best practices for doing so (e.g., [55]). However, the same is not true in literature on social media-based algorithms: even though it has been hypothesized for several years that population bias would affect social media-based algorithms (e.g., [28,46]), little research has been done to investigate this hypothesis (let alone identify remediating measures).

The goal of this paper is to help address this gap in the literature. Since the space of social media-based algorithms and social media population biases is extensive, we initiated our exploration with a focused but important case study: examining the performance of Twitter geolocation inference algorithms across the U.S. urban-rural spectrum. Due to the rising import of geographic information, geolocation inference algorithms – usually focusing on Twitter data – have attracted widespread interest across computer science and related fields (e.g., [10,15,23,25,33,45,52]). The aim of these algorithms is to predict the location of a Twitter user or her tweets using implicit signals. This is typically done by examining the content of the user’s tweets and/or the geographic configuration of her explicitly encoded social ties.

CHI 2017, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4655-9/17/05.
http://dx.doi.org/10.1145/3025453.3026015

Similarly, U.S. rural-urban biases in social media have also attracted a growing amount of attention in the literature (e.g., [22,28,31,67]). This is likely in part due to the relatively large effect sizes involved: people who live in rural areas often participate in social media at a fraction of the rate of their urban counterparts and contribute orders of magnitude less content. Twitter is no exception to this trend (e.g., [31,61,67]). The urban-rural population-bias effect sizes are of particular interest given the significant size of the rural population: over 46 million people live in rural counties in the United States (close to double the population of 18-24 year-olds) [36].

Rather than focusing on a single geolocation algorithm to assess the effect of urban/rural population bias on algorithm effectiveness, we study two well-known geolocation inference algorithms: that of Priedhorsky et al. [52] and that of Jurgens [33]. We chose these two algorithms because they span the range geolocation inference algorithm design, allowing us to more robustly evaluate the effect of population bias. Priedhorsky et al.'s and Jurgens' algorithms fundamentally differ in two key ways: methodological paradigm and problem definition. With respect to paradigm, Priedhorsky et al.'s algorithm is representative of text-based approaches to geolocation inference (e.g., [6,27,37,43,52,53]) while Jurgens' is representative of network-based approaches (e.g., [3,10,33,45,54]). Similarly, Priedhorsky et al.'s algorithm seeks to solve the "geolocate a tweet" version of the problem, while Jurgens' algorithm addresses the "geolocate a user" version of the problem.

As we will describe below, we find that regardless of methodological paradigm or problem definition, geolocation algorithms underperform for rural users. In some cases, the effect sizes are dramatic. For instance, the text-based algorithm is able to geolocate urban tweets within 100km of their true location at a rate 2.3x greater than is the case for rural tweets.

However, a major result in this paper is that population bias is not the only driver of population-variable accuracy in social media-based algorithms. Rather, we find evidence that design choices in social media-based algorithms can also have a powerful biasing effect. That is, our results suggest that algorithmic bias is a function to a large degree of both population bias in the underlying social media dataset and *structural bias* that arises from algorithmic design choices. We also observe early evidence that text-based algorithms may be more liable to structural bias than network-based algorithms. In particular, we found that when population bias was removed through balanced resampling and oversampling, our network-based algorithm showed much improved performance on rural users, but the same methods were less effective in our text-based algorithm.

This paper has several implications for those who develop and study social media-based algorithms. For instance, our research provides additional weight to calls (e.g., [57]) to consider the design of algorithms rather than just the output of these algorithms when evaluating intelligent technologies for bias. In other words, our results directly motivate further research on algorithmic design decisions that avoid structural bias. Moreover, as we will discuss, our results point to specific solutions to structural bias in the geographic algorithm domain. Similarly, our results also highlight the dangers of global evaluation metrics for social media-based algorithms, providing a data point that shows that these metrics can hide poor performance for certain populations and establish perverse incentives to reduce performance for these populations even further when they are minorities.

In addition to showing that both population and structural bias can result in uneven performance by social media algorithms, this research also makes more specific contributions to the large body of research on geolocation inference algorithms. Namely, by demonstrating that these algorithms' performance is geographically variable in a systemic fashion, we show that it is likely that these algorithms have introduced additional bias into the studies and systems that have used their output. This raises the stakes for quickly identifying solutions and establishing best practices. It also suggests that researchers who use these algorithms should be careful to audit their output as we have done here prior to incorporating them into their larger systems.

Finally, it is important to note before continuing that while we find important biases that are robust across two separate algorithms, this is a case study on Twitter geolocation inference rather than an exhaustive survey of population bias's effect on all social media-based algorithms. Given their import, identifying potential biases in social media-based algorithms is a serious matter, as is developing means of addressing them. However, because our findings our limited to a single algorithm family (albeit an important one), our findings should be interpreted as an important preliminary step in these directions rather than the definitive answer to the associated questions.

RELATED WORK

This work builds on research in three main areas: 1) characterization of population bias in social media, 2) social media-based geolocation algorithms, and 3) the literature on algorithmic accountability. We discuss these areas and their relationship to our work below.

Population Bias in Social Media

Population bias is a well-studied issue in social media. Since 2005, the Pew Research Center has conducted annual surveys of social media usage in the United States. These surveys [51] show that social media participation rates vary extensively across gender, race, education, socioeconomic status, and urban-rural lines.

Augmenting the Pew findings, many researchers have investigated this problem by analyzing the geographic distribution of posts across social media sites and making demographic inferences from census data or from users' self-reported information. The demographic dimension of analysis we choose for our study – the urban-rural spectrum – has been the focus of some of this work. For instance, Hecht and Stephens [28] studied Foursquare, Flickr, and Twitter and found a consistent pro-urban bias (e.g., there are 3.5 times more Twitter users per capita in urban areas than rural areas). Similarly, Malik et al. [44] demonstrated that there are higher densities of tweets in urban, young, and rich areas. Gilbert, Karahalios, and Sandvig [21] found substantial differences in behavior between urban and rural users in Myspace. Many other papers have studied population bias issues across other geographies, platforms, and demographics, and, to our best knowledge, some form of population bias has consistently appeared in all of this work.

A few researchers have attempted to explicitly account for population bias in their own studies. Culotta [14] demonstrated that tweet-based models of public health indicators saw improved performance when their Twitter dataset was balanced for race and gender by county. Landeiro and Culotta [37] examined how to control for shifts in the magnitude of population bias within input data to classification algorithms. Finally, Pavalanathan and Eisenstein [50] have explored how people of different ages and genders tweet differently and how this relates to performance in geolinguistic algorithms. They balanced their tweet samples based on the total population by county but did not find that this significantly impacted their algorithm.

We are motivated by the above work and the questions that it raises. While Culotta saw improved precision when controlling for specific population biases, Pavalanathan and Eisenstein did not when controlling for more general population density, but still found vast disparities in performance of the algorithm across age and gender. These divergent results raise the question of how algorithmic bias arises and is manifest in social media-based algorithms as well as the relationship between algorithmic bias and population bias in social media. Our results explain the conflicting findings in this motivating work by delineating the concept of structural bias and showing how it can counteract or exacerbate population bias. Moreover, by also examining network-based algorithms in our research, we are able to speak more broadly to social media-based algorithms in general and not just a single class of algorithms.

Geolocation Inference Algorithms

A large portion of the research and applications associated with Twitter data has a geographic component. However, this research is limited by the fact that only 1-2% of tweets are geolocated [10]. As a result, geolocation inference algorithms for Twitter, which attempt to uncover tweet and user locations that have not been explicitly disclosed, have become a very common direction of study [18]. Similar algorithms have been developed for other social media platforms as well (e.g., Flickr [19] and Tumblr [11])

There are two main classes of Twitter geolocation inference algorithms [34]: text-based, which predict the location of a tweet based on its content, and network-based, which predict a home location for a user based on their connections to other users.

Text-based geolocation algorithms rely on the tendency for language usage to vary as a function of geography (e.g., [65]). By extracting lexical features and concepts local to an area from the text during a training phase (e.g., sports teams, regional vernacular, a town name), these algorithms can build models that predict the geographic location of a new tweet based on its content. These algorithms generally either attempt to model text features as an explicitly spatial process (e.g., [52,53,64]) or treat the geolocation problem as a classification problem among administrative units (e.g., cities) (e.g., [8,25,43]).

Network-based geolocation algorithms rely on the social network in which social media posts are typically embedded. In these algorithms, explicitly encoded network ties or user interactions are used to build an egocentric social network for the user whose location is desired. Any known locations of the user's neighbors are combined to predict the location of the user [34]. This approach leverages a fundamental principle in human geography that, in general, interaction decreases with distance (e.g., [24]), meaning that connected users are likely close geographically [3].

Though our goal is to probe algorithmic bias within social media-based algorithms in general, selecting geolocation algorithms as our case study has two additional benefits: both text-based (e.g., [6,27,37,43,52,53]) and network-based (e.g., [3,10,33,45,54]) geolocation inference have been a major area of interest in the past few years, and our work can help lead to more equal and effective approaches in the future. Second, because of this robust literature on both text- and network-based approaches, we are able to explore bias in social-media based algorithms that draw on two of the most prominent methodological paradigms, improving generalizability and affording cross-paradigmatic comparisons.

Algorithmic Accountability

Our research builds on the growing literature on algorithmic accountability (e.g., [2,56,57]), in which algorithms are probed for discrimination or other societally undesirable outcomes. Our work extends the accountability literature to include well-known geolocation inference algorithms (and the rural-urban divide). Additionally, much of the algorithmic accountability literature has focused on detecting algorithmic bias when faced with a black-box system (e.g. [7,35,60,63]). Our research focuses on algorithms with a published description, open-source code, and accessible data, enabling us to investigate biases at a more detailed level, determine the potential causes of these biases with more certainty, and begin to learn how to mitigate these biases. In the discussion section, we highlight how our findings related to structural bias emphasize the importance of lower-level analyses (as per Sandvig et al. [57]) and open-source implementations in understanding and reducing algorithmic bias.

METHODS AND DATA

In this section, we describe our two focus geolocation inference algorithms and the datasets they use in more detail. In general, our approach to working with these algorithms was to replicate the choices and approach taken in the corresponding papers. In the few cases when this was not possible, we deferred to other best practices in the geolocation literature as is explained below.

Text-based Algorithm

We selected the text-based algorithm from Priedhorsky et al. [52] for our analysis. We chose this algorithm because it is representative of many text-based algorithms in its general approach (e.g., it calculates a geographic layer for each term and utilizes a standard set of Twitter text and metadata fields), has made an impact since it was published (e.g., [25]), and its source code has been made available by its authors[1], which minimizes the risk of implementation error and provides us with full control over the algorithm and its inputs.

The algorithm is trained on a set number of tweets with known locations and builds Gaussian mixture models (GMMs) for tokens in the text of the tweet, its user's time-zone, self-reported location field, and specified language. The GMMs capture the probability that a given token originated from an area based on the training data. A prediction for a given tweet is made by tokenizing it, weighting and combining the individual GMMs for each token in the tweet, and making a prediction based on the highest probability area in the resulting GMM.

Network-based Algorithm

We selected the algorithm by Jurgens [33] for our network-based algorithm. We chose this algorithm because its performance is in line with other state-of-the-art approaches [34]. Additionally, like Priedhorsky et al.'s algorithm, code for Jurgens' algorithm has been provided by the author[2], minimizing implementation risk and allowing for direct manipulation.

Jurgens' algorithm builds a bi-directional mention network by generating an edge between two users only when both users have mentioned each other in a tweet. A mention occurs when a user includes another user's username in a tweet using the "@" notation (e.g., "President @barackobama will be speaking tonight"). Starting with a training set of users with known locations, Jurgens' algorithm iteratively propagates the location of the known users to any of their neighbors who have not been successfully located, inferring the location of a newly located user in each iteration as the median of their previously located neighbors. Jurgens repeats this process for five iterations and we do the same in our analyses.

Social Media Datasets

We built our tweet dataset for our text-based algorithm following standard practices in the text-based geolocation inference community (e.g. [6,16,37]). More specifically, we utilized the Twitter Streaming API with a bounding box configured to the contiguous United States, which, like many geographic studies of geotagged content in the U.S. (e.g. [39,47])[3], was the geographic extent of our analyses. In line with prior work (e.g. [6,16,37]), we left open our tweet collector for one month. Only tweets with coordinates in the contiguous United States were retained, resulting in a dataset of 51.2 million tweets from 1.6 million unique users from October 2014. All of these tweets are explicitly geotagged with the latitude-longitude coordinates from where the tweet originated, which is necessary to provide ground truth of the algorithm. One important exception to our approach relative to that of Priedhorksy et al. was that we used geographically bounded version of the Streaming API rather than the random Streaming API. We made this exception for a simple reason: we required a much larger dataset in our study region in order to have sufficient data in rural areas for our experiments.

Network-based approaches utilize different techniques to collect data than text models and, as such, it was important to collect a separate dataset to be in accordance with standard practices in the network-based domain. To gather data for our network model, we adopted the methodology used by Jurgens et al. [34], which involved building a mention network from a dataset of randomly collected tweets (our dataset started with 99 million tweets from 26 million users from August and September 2015). We further restricted this dataset to only consider tweets from users we could geolocate to the United States, which narrowed our dataset down to 3.2 million tweets from 1.2 million users as described below, of which 113K comprised the ground truth of our final mention network (see below for more information about ground truth development). Though this is smaller than that used by Jurgens et al., it is in line with other network-based algorithm studies [34].

Ground Truth Data

As noted above, we selected the urban-rural divide as our focus demographic dimension because it corresponds to a well-studied population bias that exists in most forms of social media [28]. To categorize locations along the urban/rural spectrum for our second demographic variable, we follow standard practice in the literature that looks at urban/rural issues in online communities [28,32].

[1] https://github.com/reidpr/QUAC

[2] https://github.com/networkdynamics/geoinference

[3] Geographic methods often assume a contiguous region, and this is the case for the methods we employ here. Moreover, given their populations, it is highly unlikely that the inclusion of Alaska and Hawaii would have significant altered our results.

Specifically, we utilize the U.S. National Center for Health Statistics' Urban-Rural Classification Scheme for Counties [29], which assigns each county in the United States an ordinal code from 1 (most urban) to 6 (most rural).

Using these codes, it is straightforward to obtain urban/rural data for our text-based algorithm's ground-truth dataset. Namely, since this dataset consists entirely of geotagged tweets in the contiguous United States, we simply perform a reverse geocoding operation that labels each tweet with the county in which it is located. We then assign each county's urban/rural code to the tweet.

While text-based algorithms predict the location of each tweet individually, as noted above, network-based approaches typically seek to locate a given user (and assign all of that user's tweets to that "home" location). There are two methods used in the literature for determining a user's home location for use as ground truth in these algorithms (see Johnson et al. [31] for a complete overview of ground truth identification techniques). Jurgens [33] and McGee, Caverlee, and Cheng [45] respectively define a user's home location as the geometric median of their tweets given a minimum of five (three) geotagged tweets within 30 kilometers (50 miles) of each other. The other method takes advantage of the user's self-reported location field. This method has the drawback of being quite noisy [27] and though Jurgens et al. [34] found the location field to lead to lower overall precision, it is a method that has been used routinely in the literature [40,54]

We first evaluated the geometric median method for our dataset but found the resultant dataset to be prohibitively small, with only 20,000 users and a miniscule number of rural users. As such, we turned instead to geocoding the location field through an approach that has been shown to significant reduce the noise in this data source (albeit with reduced recall) [26]: using a Wikipedia-based geocoder. We leveraged the geocoder in WikiBrain [58], which resulted in a much larger dataset of 1.2 million users from which we built our mention network.

In line with previous results on population bias, we found that, relative to our census data, the most urban users (NCHS class 1) were highly overrepresented in our datasets (130% and 210% relative to their proportion in the overall population for the text-based and network-based datasets respectively). The most rural users (NCHS class 6) were accordingly underrepresented (45% and 24% relative proportion for the text-based and network-based datasets respectively). Lastly, it is important to note that rural/urban labels were used for evaluation purposes only and are not provided to the algorithm with training or testing data.

Evaluation Framework

When evaluating the two algorithms, we followed the procedure outlined by the two corresponding papers as closely as possible to measure the bias present in typical performance. We use five-fold validation for each network-based model. For the text model, we build and test five models for each condition. We constrain the test tweets in each text-based model to those from the days following the dates of the tweets that comprise the training data, ensure no overlap in users between training and testing phases, and use training set sizes (30,000) equal to those used by Priedhorsky et al. However, because we were limited to the Streaming API rather than the higher-volume "gardenhose" API for collecting our Twitter data, we limited the size of our network model training datasets to 24,000 users (selected anew for each fold out of the 113,000 who comprised our ground truth data), which is smaller than that used by Jurgens. We examined training datasets up to 40,000 for our baseline model though and found similar trends.

In all cases, we define algorithm precision as the percentage of predictions within 100km of the actual location. We tested different values for the distance (20km, 50km, 200km, 500km) as well as defining a true positive as a prediction lying within the same county as the ground truth (much research aggregates tweets to counties in order to leverage census data). All of these definitions led to similar patterns of results. We also tested the text-based algorithm with a similar dataset of tweets from June 2015 and saw similar trends. We only report recall (the percentage of users or tweets for which an algorithm could provide a location) for the network-based models because the recall was consistently around 100% for the text-based models.

In line with the NCHS classifications [29], we combine data from counties with NCHS codes of 1 and 2 ("large metropolitan counties") into one "urban" class and counties from NCHS codes of 5 and 6 ("nonmetropolitan counties") into one "rural" class. We use these definitions of urban and rural for the rest of the paper. We restrict our reporting and discussion to just these urban and rural classes, though the results for NCHS categories 3 and 4 generally fell between those for the urban and rural classes.

RESULTS

We structured our exploration of the effect of population bias on social media-based algorithms by asking three cascading research questions. Our first question was whether the two geolocation algorithms exhibit biased performance in favor of urban areas (**RQ1: Is there an algorithmic bias in the direction of the population bias?).** Our next step was to inquire whether any identified bias was due to population bias (**RQ2: Is any bias due to population bias?**). We did this by examining the change in algorithmic bias when we adjusted the training dataset so that it contained a representative urban/rural sample of the general population.

Finally, given a positive answer to RQ2, we planned a third research question whose objective was to investigate whether any remaining bias could be eliminated by training solely on the algorithmically disadvantaged population (**RQ3: Can any remaining underperformance for a specific population be fixed by training solely on data from that population?**). This is equivalent to building a

	% of Training Data		Precision (Recall)		
Text-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	63.8%	9.0%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.3%	15.0%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	**Urban**	**Rural**	**Urban**	**Rural**	**Overall**
Typical (Baseline)	75.0%	5.2%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.3%	15.0%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

Table 1. Urban-rural results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses, except the text-based models which had recall always around 100%.
***The network-based rural-boosted model was trained on 6000 tweets due to lack of data.**

separate algorithm customized (trained) specifically for rural users and tweets and another for urban users and tweets. The goal of RQ3 was to identify whether any bias that remained after adjusting for any population imbalances was inherent to the algorithm.

Below, we use our three research questions to structure our discussion of our results. To determine the significance of changes in precision of the algorithm, we adopt the methods used by Compton et al. [10], which set confidence intervals as the average precision ±1.5 times the interquartile range for the folds. Comparisons are only made when confidence intervals do not overlap unless otherwise noted.

RQ1: Is there algorithmic bias?

Examining the precision of each algorithm in urban and rural contexts (Table 1, rows labeled "Typical (Baseline)"), a clear pattern of bias emerges. Both algorithms perform worse for rural users than for urban users, with the magnitude of the bias being greatest in the text-based algorithm: this algorithm is able to accurately locate urban users within 100km at a rate approximately 2.3 times greater than that for rural users. The equivalent precision number for the network model is 1.3x, and recall is 1.6x better for urban users than rural users in the network model as well.

Figure 1 on the following page shows the precision of the text-based algorithm by county in the contiguous United States. It demonstrates the depth of this bias – almost all high precision clusters center in urban areas around cities.

In addition to motivating further inquiry as to whether this algorithmic bias arises from population bias in the underlying dataset or other factors (i.e. our RQ2 and RQ3), this result has important implications in and of itself. Namely, geolocation inference algorithms have served as inputs to systems and studies and our results establish for the first time that the use of these well-known geolocation inference algorithms from the literature will inject further population bias into any geographically referenced dataset of Twitter users. For instance, in the case of the text-based algorithm, 2.3x more urban users than rural users will be "put on the map" correctly. We return to this point in the discussion section.

RQ2: Is the observed bias due to population bias?

The cells in the "Typical (Baseline)" rows and "% of Training Data" column in Table 1 indicate that, as noted above and in line with a number of previous studies on Twitter, our unadjusted training data has significant underlying population biases. For instance, we can see that 63.8% of our text-based urban/rural ground-truth dataset can be classified as urban according to our definition, but only 9.0% of our dataset can be classified as rural. The actual census proportions would be 55.3% and 15.0% respectively, indicating a strong urban bias in our dataset.

Our goal with this research question was to determine whether correcting for this population bias in the training datasets is the primary cause of the biases we observed in our RQ1 analyses. As pointed out by Pavalanathan and Eisenstein [50] and Burger et al. [5], more data about a group in prediction tasks generally improves accuracy of the algorithm. Therefore, we expect that by adjusting for population bias, we can induce greater and perhaps equal performance for underrepresented populations (i.e. rural populations). Though rural areas still would have a much lower number of tweets even within a population-balanced dataset, maintaining a consistent number of tweets per person in an area as training data should capture an equal percentage of an area's location-indicative words (to use Han, Cook, and Baldwin's [25] vocabulary).

To answer our second research question, we performed a simple modification of the datasets on which we trained each algorithm: instead of resampling the datasets at random for each fold of training the algorithm as we did in RQ1, we resampled them such that they were representative of the demographics of the general population with respect to the rural/urban divide. In other words, we generated training

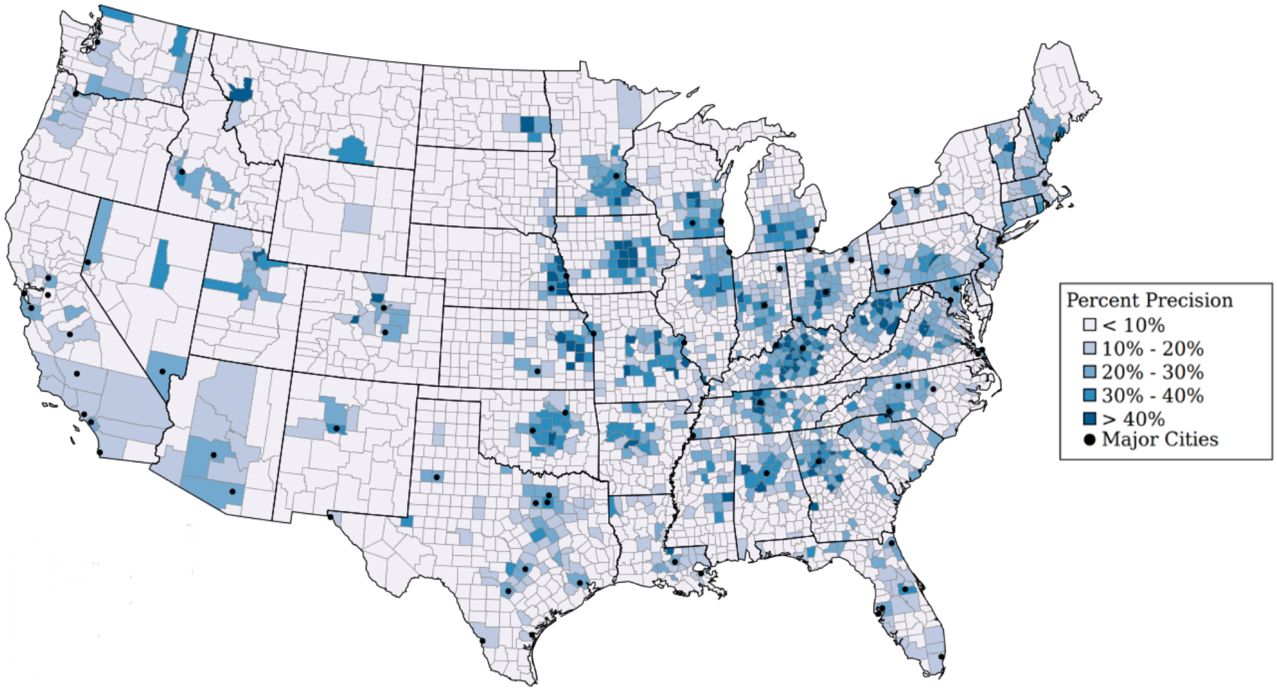

Figure 1. Text-based geolocation baseline precision by county (percentage of tweets originating from each county correctly geolocated to within 100km of each tweet's true location).

datasets without population bias. We did not adjust how we sampled data for testing each fold (i.e. it remained random).

The results of the evaluation of each algorithm under this population-balanced condition can be found in Table 1 in the rows labeled "Population Bias Balanced". These rows tell a relatively straightforward story: while using a representative sample of the general population significantly changed the results for the network-based model with respect to the urban/rural divide, balancing the population had only a negligible effect on the text-based model. In other words, the text-based model remained significantly urban-biased, even when using a training set that removes population bias in the underlying Twitter dataset. On the other hand, the network-based model was significantly less biased when trained on a population-balanced dataset.

Like was the case for RQ1, this result both motivates the investigation of our subsequent research question (RQ3) and has implications on its own. Whereas there is evidence that addressing the effects of demographic bias in social media-based social science research can be done by simply resampling the underlying dataset [14], our findings suggest the same cannot be said for social media-based algorithms, at least in the case of our text-based model. This is another subject to which we attend below.

RQ3: Can we fix biases through oversampling?

Even when the representation of rural Twitter users has been boosted to match that of the general population, rural users still make up a much smaller relative proportion of the training set and therefore have a much smaller corresponding absolute number of training samples. In this experiment, we sought to see if this imbalance in absolute number of tweets can explain the algorithmic bias we observed above.

We did this by training and testing on each demographic group separately (i.e. separate models for rural and for urban). If separate models perform equally well (e.g., rural precision is as high as urban precision), then we would know that the algorithmic bias in our algorithm arises solely from population characteristics of the input dataset. If bias still exists even in separate models, then we would know that there are structural biases within the algorithms that prevent equal performance for these demographics, no matter their representation in the training dataset.

The results of this experiment can be seen in the rows labeled "Urban Boosted" and "Rural Boosted" in Tables 1. Of particular note is the performance of the text-based algorithm. Even when training and testing solely on tweets from rural users, the text-based model cannot geolocate these tweets as well as it can for tweets from urban users when using a simple random sample or population-adjusted training set (let alone using an only-urban model). While we do see a large and significant improvement in the rural case, performance still falls short of that of urban tweets for all of our models that contained any urban training data at all. We also note that we tried boosting the absolute number of training samples by up to two times the number used in Priedhorsky et al., and the results were consistent: the rural-only model still had a precision less than the random and population-adjusted urban precision. In other words, no matter the training data, there appears to be something in the design of the text-based algorithm that prevents it from performing well for rural users relative to urban users.

The story for the network-based model is different, with it appearing to also suffer from structural biases, but not to the same degree as the text-based model. With respect to precision, the model trained only on rural users actually outperforms the model trained only on urban users. However, the recall remains lower than any of the urban models.

These results have important implications for the design and application of social media-based algorithms. We begin our Discussion section below by highlighting these implications.

DISCUSSION

Algorithmic Bias = Population Bias + Structural Bias + Ɛ

The results to our second and third research questions suggest a nuanced understanding of the mechanisms behind algorithmic bias. Namely, while we saw that some algorithmic bias could be explained by population bias in the underlying training sets (e.g., the gap in performance between urban and rural users in the network-based algorithm), not all of it could. In fact, in the case of the text-based algorithm, even dramatically overcorrecting for population bias by training solely on rural users did not make the algorithm perform as well as it typically does for urban users.

These findings support the notion that algorithmic bias must be understood as a function of both population bias and structural bias inherent to the algorithm's design (as well as other factors that have yet to be discovered). In other words, there is something about the nature of some algorithms that inherently biases them towards lower performance for certain populations.

A Closer Look at Structural Bias

Examining the structure of the two algorithm families under consideration, a number of hypotheses emerge for their differing amounts of structural bias along urban/rural lines. Network-based algorithms build an egocentric network, so a prediction for a given user is affected directly by her/his social network neighbors and potentially indirectly by other nearby users (e.g. neighbors of neighbors) through multiple stages of inference. This means that addition or subtraction of users in one part of the network largely does not impact a given user elsewhere in the network even though it can boost precision and recall amongst their more immediate social network neighbors. Indeed, within our mention network, we find very high homophily: 90% of edges between users with known ground-truth locations are of the same type (i.e. for a user known to be in an rural class 6 county, there is a 90% chance that any of their social network neighbors whose location is known a priori are also in rural class 6 counties) Boosting data for rural users therefore greatly increases the likelihood that a rural user in the testing dataset will have neighbors that are located without affecting most urban users because they would not be closely linked through the network to the rural users.

Text-based algorithms, on the other hand, see much greater dependencies between users. Toponyms and language that have broad usage across the country will be skewed towards being located in urban areas with their higher density of users. Furthermore, when examining the average number of words per tweet that can be identified as geographic Wikipedia concepts (i.e., words tied directly to place) through wikification algorithms implemented by Sen et al. [58], we see that the most urban tweets (NCHS code = "1") have 25% more "wikifiable" words per tweet than the most rural tweets (NCHS code = "6"). Since location-specific words such as these are key to how the text-based algorithm operates, some of the urban advantage may come from differing language patterns and topics of conversation in tweets across urban-rural lines [16].

Another common design decision in text-based algorithms that could be a cause of structural urban/rural bias relates to the fixed distance parameters central to many of these algorithms' low-level functionality. These distance parameters, which manifest as grid-cell sizes or geographic probability distribution ranges, fail to take into account the varying distances at which the use of a term is predictive in urban versus rural areas. For instance, when someone tweets a name of a high school mascot in an urban area, that name is predictive for a smaller area than the same situation in a rural area (i.e. rural areas have significantly larger school districts by area). This problem can be understood in geostatistics terms: the use of a fixed-distance parameter assumes a fixed range of spatial autocorrelation for tweet usage, which likely is not true across urban/rural lines. In more general terms, this means that employing fixed distance parameters will fail to capture the full predictive power of each rural tweet in text-based algorithms, and rural areas already have fewer tweets per capita to begin with.

Achieving Parity

Though lower overall recall and precision is a barrier to implementation of network-based algorithms, our work indicates that it may be easier to achieve parity within network-based algorithms by boosting data collection efforts around underrepresented populations. Although more work should be done to confirm this effect in other types of social media algorithms, researchers and designers for whom equity is a top priority may want to consider utilizing network-based methods when doing Twitter-based geolocation.

It is also important to note that the contribution of structural bias to algorithmic bias we have identified here adds weight to the argument of Sandvig et al. [57] and others (e.g., [62]) that algorithmic accountability work needs to consider algorithms at levels deeper than simply inputs and outputs and that algorithmic accountability research teams have people with "algorithmic skills that allow a facility with the relevant [algorithmic] ideas" [57]. These skills will be necessary to identify and address bias at the structural level.

Similarly, the notion of structural bias highlights the importance of open-source implementations of algorithms. Open-source affords the ability to understand and design fixes for these algorithmic biases because we are able to examine and manipulate how these algorithms function. Our results show that we cannot rely solely on adjusting the data going into the algorithm to achieve parity.

A Tradeoff Between Equity and Effectiveness

There is one column in Table 1 that we have yet to discuss in detail: the "Overall" column. This column indicates the performance on the entire randomized population (i.e. data in an unmodified proportion of users and tweets). In other words, the "Overall" column reports the precision (and recall) a researcher or developer would expect to achieve if she were to apply the model listed in each row to all of Twitter.

There is a clear trend in the "Overall" column: for models in which rural performance is better, the performance on overall population gets worse. From the perspective of a researcher or developer, this means that in order to improve rural accuracy, one has to reduce overall accuracy. Furthermore, for the text-based algorithm, we also tested a wider variety of urban-rural training data proportions to understand the responsiveness of the algorithm to smaller shifts in data. In doing so, we found that peak performance (21.0% overall precision at 100 km) came in a model that removed a quarter of the rural training data and boosted urban accordingly. Rural users performed very poorly in this model (6.8% precision), but the slight increase in precision for urban users (23.5% precision) along with their inherently higher proportions in the testing data was sufficient to boost the overall precision.

Our results demonstrate that, at least in the case of Twitter geolocation, that there is a clear trade-off between equity and effectiveness, a result for which there is evidence in other algorithmic accountability contexts as well (e.g., [35,48]). An important corollary to this tradeoff is that using single measures of precision and recall for an algorithm can gloss over very real, non-random variation in algorithmic performance for different groups of people. Until better algorithms can be developed that do not compromise equity (or equality) for effectiveness (e.g., by addressing structural bias), algorithm designers should conduct and report a more thorough examination of performance across different populations as has been advocated by many in the algorithmic accountability community, especially when past work has suggested that there may be population bias or structural bias.

Privacy

We have conducted this research with the assumption that higher precision and recall is a desirable outcome for a given population. While this is a valid assumption in an application of geolocation like public health tracking, this is not always the case for social media-based algorithms and for geolocation specifically. It is important to note that in this light, there may be benefits to being "disadvantaged" by geolocation algorithms: our results suggest that rural users are harder to "find" in an automated fashion, preserving privacy. Geolocation inference is already employed by at least some social surveillance firms, locating tweets based on the language and metadata [17]. An interesting direction of research (e.g. [25]) is to invert the goals of this paper and attempt to find ways reduce the "geolocatability" of a person or a population (i.e. defend against "inference attacks" [42]).

LIMITATIONS AND FUTURE WORK

In this paper, we necessarily binned users into categories along the urban and rural spectrum. There is without a doubt a tremendous amount of diversity in the people and behaviors present within each of these categories, and further research may want to address this. Categorizing users based on behavior and content as opposed to demographic labels could provide additional insight into who is likely to be affected by these algorithmic biases.

Following best practices in the Twitter geolocation literature (e.g. [27,33,34,52]), for our ground truth data, we depended on explicitly geotagged tweets for our text-based algorithm (i.e. tweet location) and a very conservative (i.e. precision-focused rather than recall-focused) geocoding of the location field for our network-based algorithm (i.e. user home location). While doing so was critical to our goal of evaluating bias in the algorithms as they were published, it is possible that these mechanisms may disproportionally remove people of one demographic relative to another demographic. Although developing a ground truth through other means (e.g. a survey) would be a major research project in its own right, examining Twitter geolocation algorithms through this lens would be a useful addition to the literature.

Another area of future work would be expanding the focus of this study (the contiguous United States) to other cultures and geographic contexts. It is known that different cultures use social media differently (and have their own population biases) so it is not clear how our results would extend to these areas. Furthermore, building on our understanding of how different populations use social media (e.g., variation of the use of mentions across cultures [20]) will enable better prediction of where algorithmic biases might arise. Along the same lines, we sought to choose representative algorithms, but different algorithms may perform differently and introduce their own structural biases.

Finally, now that this paper has established the role of structural bias, a very important direction of future work is finding ways to reduce or eliminate it in important algorithms. We expect that doing so could lead to an interesting and fruitful line of work.

CONCLUSION

This research improves our understanding of algorithmic biases in social media-based algorithms. We demonstrated the degree to which these algorithmic biases arise from both population biases in the training data and structural biases inherent to the algorithms themselves. Through the implementation of both a text-based and a network-based algorithm for geolocation inference, we found that network-based approaches may be less susceptible to structural biases. We also discussed the implications of our findings for designers and users of social-media based algorithms. These implications include (1) the need for more work developing algorithms that avoid the structural biases we observed here and (2) that global evaluation metrics can mask significant underperformance for certain populations in these algorithms.

ACKNOWLEDGEMENTS

We would like to thank our colleagues at GroupLens at the University of Minnesota for their support and brainstorming time. This project was funded in part by the U.S. National Science Foundation (CAREER IIS-1552955 and IIS-1526988), the Volkswagen Foundation, and a University of Minnesota College of Science and Engineering Graduate Fellowship.

REFERENCES

1. Saeed Abdullah, Elizabeth L. Murnane, Jean M.R. Costa, and Tanzeem Choudhury. 2015. Collective Smile: Measuring Societal Happiness from Geolocated Images. In *CSCW*. https://doi.org/10.1145/2675133.2675186
2. Mike Ananny, Karrie Karahalios, Christian Sandvig, and Christo Wilson. 2015. Auditing Algorithms from the Outside: Methods and Implications. In *ICWSM*.
3. Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*.
4. Saeideh Bakhshi, David A. Shamma, and Eric Gilbert. 2014. Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram. In Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14), 965–974. https://doi.org/10.1145/2556288.2557403
5. John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *EMNLP*.
6. Miriam Cha, Youngjune Gwon, and H. T. Kung. 2015. Twitter Geolocation and Regional Classification via Sparse Coding. In *ICWSM*.
7. Le Chen, Alan Mislove, and Christo Wilson. 2015. Peeking Beneath the Hood of Uber. 495–508. https://doi.org/10.1145/2815675.2815681
8. Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You Are Where You Tweet : A Content-Based Approach to Geo-locating Twitter Users. *CIKM*. https://doi.org/10.1145/1871437.1871535
9. Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. 2011. Exploring Millions of Footprints in Location Sharing Services. *ICWSM* 2011.
10. Ryan Compton, David Jurgens, and David Allen. 2014. Geotagging one hundred million twitter accounts with total variation minimization. In *IEEE BigData*.
11. Ryan Compton, Craig Lee, Jiejun Xu, Luis Artieda-moncada, Tsai-ching Lu, Lalindra De Silva, and Michael Macy. 2013. Using publicly visible social media to build detailed forecasts of civil unrest. 1–11.
12. Justin Cranshaw, Jason I Hong, and Norman Sadeh. 2012. The Livehoods Project : Utilizing Social Media to Understand the Dynamics of a City. *ICWSM*: 58–65.
13. Aron Culotta. 2014. Estimating county health statistics with twitter. In *JSM Proceedings*, 1335–1344. https://doi.org/10.1145/2556288.2557139
14. Aron Culotta. 2014. Reducing Sampling Bias in Social Media Data for County Health Inference. *JSM Proceedings*.
15. Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop: HIAI*.
16. Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. *EMNLP*. https://doi.org/10.1038/nrm2900
17. Benjamin Elgin and Peter Robison. 2016. How Despots Use Twitter to Hunt Dissidents. *Bloomberg Technology*. Retrieved from https://www.bloomberg.com/news/articles/2016-10-27/twitter-s-firehose-of-tweets-is-incredibly-valuable-and-just-as-dangerous
18. David Flatow, Mor Naaman, Ke Eddie Xie, Yana Volkovich, and Yaron Kanza. 2015. On the Accuracy of Hyper-local Geotagging of Social Media Content. In *WSDM*. https://doi.org/10.1145/2684822.2685296
19. Andrew Gallagher, Devashree Joshi, Jie Yu, and Jiebo Luo. 2009. Geo-location inference from image content and user tags. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, 55–62.
20. Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. 2013. Cultural dimensions in twitter: Time, individualism and power. *ICWSM* 13.
21. Eric Gilbert, Karrie Karahalios, and Christian Sandvig. 2008. The Network in the Garden : An Empirical Analysis of Social Media in Rural Life. *CHI*: 1603–1612.
22. Eric Gilbert, Karrie Karahalios, and Christian Sandvig. 2010. The Network in the Garden: Designing Social Media for Rural Life. *American Behavioral Scientist* 53, 9: 1367–1388. https://doi.org/10.1177/0002764210361690
23. Mark Graham, Scott A. Hale, and Devin Gaffney. 2014. Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional*

Geographer 0, 0: 1–11. https://doi.org/10.1080/00330124.2014.907699

24. T. Hagerstrand. 1968. Innovation diffusion as a spatial process. 334 pp.
25. Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*: 451–500.
26. Brent Hecht and Darren Gergle. 2010. On the "localness" of user-generated content. *CSCW*: 229. https://doi.org/10.1145/1718918.1718962
27. Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *CHI*.
29. Brent Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. In Eighth International AAAI Conference on Weblogs and Social Media.
29. DD Ingram and SJ Franco. 2014. 2013 NCHS urban-rural classification scheme for counties. *Vital Health Statistics* 2, 166.
30. Yushi Jing, David Liu, Dmitry Kislyuk, Andrew Zhai, Jiajing Xu, Jeff Donahue, and Sarah Tavel. 2015. Visual Search at Pinterest. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15), 1889–1898. https://doi.org/10.1145/2783258.2788621
31. Isaac L. Johnson, Subhasree Sengupta, Johannes Schöning, and Brent Hecht. 2016. The Geography and Importance of Localness in Geotagged Social Media. In *2016 CHI Conference on Human Factors in Computing Systems*, 515–526. https://doi.org/10.1145/2858036.2858122
32. Isaac Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at Home on the Range: Peer Production and the Urban/Rural Divide. *CHI*.
33. David Jurgens. 2013. That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. *ICWSM* 13: 273–282.
34. David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *ICWSM*.
35. Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 3819–3828. https://doi.org/10.1145/2702123.2702520
36. Lorin D. Kusmin. 2016. *Rural America At A Glance: 2015 Edition*. United States Department of Agriculture. Retrieved from http://www.ers.usda.gov/media/1952235/eib145.pdf
37. Virgile Landeiro and Aron Culotta. 2016. Robust text classification in the presence of confounding bias. In *Thirtieth AAAI Conference on Artificial Intelligence*. Retrieved May 17, 2016 from http://www.aaai.org/Conferences/AAAI/2016/Papers/02Landeiro12445.pdf
38. Géraud Le Falher, Aristides Gionis, and Michael Mathioudakis. 2015. Where Is the Soho of Rome? Measures and Algorithms for Finding Similar Neighborhoods in Cities. In *ICWSM*.
39. Linna Li, Michael F. Goodchild, and Bo Xu. 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science* 40, 2: 61–77. https://doi.org/10.1080/15230406.2013.777139
40. Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *SIGKDD*.
41. Xutao Li, Tuan-Anh Nguyen Pham, Gao Cong, Quan Yuan, Xiao-Li Li, and Shonali Krishnaswamy. 2015. Where You Instagram?: Associating Your Instagram Photos with Points of Interest. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (CIKM '15), 1231–1240. https://doi.org/10.1145/2806416.2806463
42. J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. 2009. Inferring Private Information Using Social Network Data. In *WWW '09: 2009 International World Wide Web Conference*.
43. Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2014. Home Location Identification of Twitter Users. *ACM TIST* 5, 3: 1–21. https://doi.org/10.1145/2528548
44. Momin M. Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population Bias in Geotagged Tweets. In *ICWSM*.
45. Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. 2013. Location prediction in social media based on tie strength. In *CIKM*, 459–468. https://doi.org/10.1145/2505515.2505544
46. Alan Mislove, Sune Lehmann, Yong-yeol Ahn, Jukka-pekka Onnela, and J Niels Rosenquist. Understanding the Demographics of Twitter Users. *ICWSM*: 554–557.
47. Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. 2013. The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one* 8, 5: e64417. https://doi.org/10.1371/journal.pone.0064417
48. Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, New York.
49. Aditya Pal, Amac Herdagdelen, Sourav Chatterji, Sumit Taank, and Deepayan Chakrabarti. 2016. Discovery of Topical Authorities in Instagram. In *WWW*.
50. Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and Consequences in Geotagged Twitter Data. *EMNLP*.

51. Andrew Perrin. 2015. Social Media Usage: 2005-2015. *Pew Research Center*.
52. Reid Priedhorsky, Aron Culotta, and Sara Y. Del Valle. 2014. Inferring the Origin Locations of Tweets with Quantitative Confidence. *CSCW* 29: 997–1003.
53. Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *EMNLP-CoNLL*.
54. Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2013. Where's@ wally?: a classification approach to geolocating users based on their social ties. In *Hypertext*, 11–20.
55. Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213: 1063–1064. https://doi.org/10.1126/science.346.6213.1063
56. Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination*.
57. Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2015. Can an Algorithm be Unethical? In *65th Annual Meeting of the International Communication Association*.
58. Shilad Sen, Toby Jia-Jun Li, WikiBrain Team, and Brent Hecht. 2014. WikiBrain: Democratizing Computation on Wikipedia. In *OpenSym* (OpenSym '14), 27:1–27:10. https://doi.org/10.1145/2641580.2641615
59. Börkur Sigurbjörnsson and Roelof Van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *WWW*.
60. Gary Soeller, Karrie Karahalios, Christian Sandvig, and Christo Wilson. 2016. MapWatch: Detecting and Monitoring International Border Personalization on Online Maps. In *Proceedings of the 25th International Conference on World Wide Web* (WWW '16), 867–878. https://doi.org/10.1145/2872427.2883016
61. Monica Stephens. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal* 78, 6: 981–996. https://doi.org/10.1007/s10708-013-9492-z
62. Suresh Venkatasubramanian. 2016. Algorithmic Fairness: From social good to a mathematical framework. Retrieved September 17, 2016 from https://algorithmicfairness.wordpress.com/2016/04/15/keynote-at-icwsm/
63. Jacob Thebault-Spieker, Loren G. Terveen, and Brent Hecht. 2015. Avoiding the South Side and the Suburbs: The Geography of Mobile Crowdsourcing Markets. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 265–275. https://doi.org/10.1145/2675133.2675278
64. Benjamin P. Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *ACL*.
65. Wilbur Zelinsky. 1980. North America's Vernacular Regions. *Annals of the Association of American Geographers* 70, 1: 1–16. https://doi.org/10.1111/j.1467-8306.1980.tb01293.x
66. Danning Zheng, Tianran Hu, Quanzeng You, Henry Kautz, and Jiebo Luo. 2015. Towards Lifestyle Understanding: Predicting Home and Vacation Locations from User's Online Photo Collections. In *ICWSM*.
67. Kathryn Zickuhr and Aaron Smith. 2011. 28% of American Adults Use Mobile and Social Location-Based Services. *Pew Internet and American Life Project*.