County-level Algorithmic Audit of Racial Bias in Twitter’s Home Timeline

Luca Belli  
Twitter Inc.

Kyra Yee  
Twitter Inc.

Uthaipon Tantipongpipat  
Twitter Inc.

Aaron Gonzales  
Twitter Inc.

Kristian Lum  
Twitter Inc.

Moritz Hardt  
Max Planck Institute for Intelligent Systems, Tübingen

ABSTRACT

We report on the outcome of an audit of Twitter’s Home Timeline ranking system. The goal of the audit was to determine if authors from some racial groups experience systematically higher impression counts for their Tweets than others.

A central obstacle for any such audit is that Twitter does not ordinarily collect or associate racial information with its users, thus prohibiting an analysis at the level of individual authors. Working around this obstacle, we take US counties as our unit of analysis. We associate each user in the United States on the Twitter platform to a county based on available location data. The US Census Bureau provides information about the racial decomposition of the population in each county.

The question we investigate then is if the racial decomposition of a county is associated with the visibility of Tweets originating from within the county. Focusing on two racial groups, the Black or African American population and the White population as defined by the US Census Bureau, we evaluate two statistical measures of bias.

Our investigation represents the first large-scale algorithmic audit into racial bias on the Twitter platform. Additionally, it illustrates the challenges of measuring racial bias in online platforms without having such information on the users.

CCS CONCEPTS

• Human-centered computing → Empirical studies in collaborative and social computing. Social network analysis; Social media; Social networking sites.

KEYWORDS

algorithmic audit, algorithmic bias, recommender system, dataset

1 INTRODUCTION

The content on Twitter’s Home Timeline is by default selected and ordered by personalization algorithms. This raises questions of algorithmic biases, such as: Does algorithmic personalization increase the visibility of some groups relative to others in a manner that aligns with and reinforces persisting inequality? In this paper, we audit Twitter’s Home Timeline ranking algorithm for racial bias. The scope of our study is limited to the United States and focuses on two racial groups: the Black or African American population and the White population, as defined by the US Census Bureau.

Twitter does not generally collect or associate racial information with its users, which poses a methodological challenge to our audit. In fact at no time was any individual user’s actual race or ethnicity known or used as part of the research. Addressing this challenge, we use coarse location information associated with a user that places the primary location of a user within a US county. Counties vary in their racial demographics, raising the possibility that differences in a county’s visibility on Twitter systematically align with its racial demographics. Although there are challenges in using aggregate level data as proxies for individual level disparities [8, 9, 22], county-level data has been used to study racial disparities in a variety of domains, such as law enforcement [19], healthcare [10, 15, 16], and education [18]. There is also growing interest in applying county level data to understanding disparate impact in algorithmic applications [6, 13, 14], where individual level demographic data is often difficult to procure [1].

Choosing counties as our unit of analysis, we study two non-experimental ways of measuring bias. The first considers the percentage of amplified users in each county, i.e., users whose Tweets have higher normalized impression counts than the US median user. We test whether the racial demographics of a county have an effect on the fraction of amplified users in the county. The second analysis splits counties based on whether the county has a greater than median population within a specific racial group. For the two groups of counties, the analysis then compares the distribution of amplified users.

Both analyses suggest that the size of the Black or African American population within a county is not significantly associated with higher or lower normalized impression counts. Both analyses show a small, but statistically significant, negative correlation between the size of the White population within a county and normalized impressions on the Home Timeline.

The contributions of our report are as follows:

• We extend the growing body of work on algorithm audits to include a large-scale study of racial bias on the Twitter platform.
• We show one way to work around general unavailability of user-level racial information in the context of an algorithm audit, while detailing the limitations of this approach.
• Our report provides systematic quantitative information about racial bias on the Twitter platform.
• Along with this report, we release the county-level dataset used for our investigation, hoping to facilitate additional analyses, replication efforts, and discussion. The dataset, for example, allows researchers to extend our analysis to other racial groups, as well as rural versus urban counties.

*Work performed while MH was a paid consultant for Twitter.
2 OBSERVATIONAL DESIGN OF OUR STUDY

At a high level, our goal is to determine whether the racial composition of a US county is associated with higher or lower visibility on Twitter’s Home Timeline. This requires demographic information at the county level as well as a measure of visibility at the Twitter user level. Both are detailed below.

2.1 Definition of county and associated demographics

Our unit of analysis is the county or “county equivalent”, such as Louisiana parishes. For simplicity we are going to use the word county to refer to both. Counties that are part of the study are from the 2018 American Community Survey (ACS) of the US Census Bureau [5]. For the demographic data, we use racial and ethnic data from the same source [5], thus relying on the 2018 ACS categorization of race. Racial and ethnic data in the Census is self-identified, where participants can choose to mark one or more of the following categories: White, Black or African-American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander. There is a separate question for ethnicity, which asks whether or not participants are of Hispanic, Latino, or Spanish origin or not. Hispanic, Latino, or Spanish people may be of any race. For the limitations of this categorization, see Section 2.6.5.

2.2 Time and scope of the data collection

Our study includes all Twitter users who posted at least one Tweet between September 1, 2021 and September 9, 2021, that we were able to associate a location with in the United States. We assign to each user a unique county, which is the one from where they logged in the most frequently between August 10, 2021 and August 24, 2021. We restrict impressions to readers who see Twitter’s Home Timeline, this counts as one impression. We consider unique impressions, meaning that multiple impressions by the same reader count as one. We restrict impressions to readers who see Twitter’s default algorithmically personalized timeline. Twitter users may opt out of personalization and a small fraction of users are randomly excluded from algorithmic personalization, see [12] for more details on location and its limitations. This study includes on the order of tens of millions of unique users.

2.3 Impressions and amplification

Throughout the paper, we use the concept of an impression. When a Tweet is shown to any user in the Home Timeline, this counts as one impression for the author of the Tweet. We consider unique impressions, meaning that multiple impressions by the same reader count as one. We restrict impressions to readers who see Twitter’s default algorithmically personalized timeline. Twitter users may opt out of personalization and a small fraction of users are randomly excluded from algorithmic personalization, see [12] for more details on this aspect.

The number of impressions associated with a user are generally a function of how much content the user authors and how many followers they have. We make the simplifying assumption that the number of impressions scales bilinearly in the number of followers a user has and the number of Tweets the user authors. This assumption motivates the notion of a normalized impression.  

\[
\text{normalized} \text{ impressions} = \frac{\text{time window} \times \text{total impressions}}{\text{total number of unique impressions}}
\]

Here, total impressions is the total number of unique impressions received only via the algorithmically-ranked timeline, total time window is the total number of impressions received in the time frame, and total number of unique impressions is the number of followers at the beginning of the experiment. 

Normalized impressions are what allows us to define what an amplified user is.

\[
\text{amplified users} = \text{normalized} \text{ impressions} \times \text{total number of unique impressions}
\]

Note that by definition of median, there is an equal number of amplified and non-amplified users. As a robustness check, we considered quantiles other than the median in the definition.

2.4 Analysis 1: Effect of racial composition on amplification

For each of the two racial groups we focus on (Black or African American and White), we fit one least squares linear regression model\( Y = \alpha + \beta X \). where \( X \) is the fraction of the county’s population in the given racial group and \( Y \) is the share of amplified users in the county. Both variables range in the interval [0, 1]. The unit of analysis is the county, meaning that each data point corresponds to one county.

We use the regression coefficient \( \beta \) as an observational measure of bias: A positive coefficient indicates that fractional size of the group within a county is associated with a higher share of amplified users. A negative coefficient indicates the opposite. To take into account the fact that different counties have different number of users (e.g. metropolitan areas tend to have more Twitter users), we perform two kinds of regression: an unweighted one (each county has equal weight) and weighted one, where the counties are weighted by the share of Twitter users (not the total population). We report the \( \beta \)-coefficient for each of the four regressions with 95% confidence intervals, as well as the \( R^2 \)-values. In a two variable linear causal model \( X \rightarrow Y \) with no unobserved confounders, the coefficient \( \beta \) can be interpreted as the causal effect of a county’s population in a given racial group on amplification. The county level, a possible confounder could be a property of the county that influences both the county’s racial composition and its amplification on Twitter. A plausible confounder that we did not control for is the county’s urban-rural status, a classification also provided by the US Census Bureau.

We refer to the following features from the Census using the following labels: 
\[ DP05_0037E: \text{white}, \ DP05_0038E: \text{Black}, \ DP05_0077E: \text{White alone, not Hispanic or Latino}, \ DP05_0078E: \text{Black alone, not Hispanic or Latino} \]

See https://api.census.gov/data/2018/acs/acs5/profile/variables.html for complete list of features in ACS 2018. For more information on the conceptualization of race on the Census, see https://www.census.gov/topics/po
2.5 Analysis 2: Distribution of amplified users by county

For our second analysis, we divide the counties into two separate sets, namely counties above and below the median of each racial group we consider. We then consider the histograms of the fraction of amplified users for each of the two cohorts. The difference between the two histograms is another indicator that racial composition of a county is associated with amplification. We report the mean and standard deviation for each distribution, as well as the total variation distance between the two distributions corresponding to the county split.

2.6 Limitations

Our study design has several inherent limitations and some technical limitations specific to the Twitter context.

2.6.1 Location. The county of a user is constructed from available location data. The location of a user has intrinsic variation over time. Some users tag their Tweet with a location, which gives more confidence about their location at the time of writing the Tweet. There is lower precision when trying to infer the location of a user via IP address. As stated above, we are assigning to a user the county they log-in the most from, while the Census data measures the location where people live in. There could be discrepancies between the two: it could be possible that people who live in places with poor internet access will tend to use Twitter more in location with better internet access (e.g. the library, a work location etc), which might be located in a different county than the one where they reside.

A small fraction of users that we aren’t able to locate are dropped from the study.

2.6.2 Data Loss. The 2018 ACS Census data has 3219 counties listed that are part of our study. After assigning users to counties, we notice that 120 counties had no users assigned to them. Note that this need not imply that there are no users in those counties, just that we weren’t able to associate users to them. In some cases this seems expected, as some of those are inhabited by a few hundred people and it’s possible that no one used the platform from any of those. In other cases the size or population alone does not justify the lack of user data, and it’s possible that data loss is happening at a step in the data pipeline.

2.6.3 Level of Granularity. Our analysis focuses on exposure disparity at the county level. From this data, we are not able to verify that our observations also hold at the user level. To illustrate the point, consider a county with a larger than average Black of African American population that has a large fraction of amplified users. Given the lack of demographic data on individual users, we don’t know the race of amplified users within the county. It is possible, in principle, that all amplified users within the county are not Black or African American. Testing this hypothesis would require racial information about individuals.

2.6.4 Disparities between Census and Twitter’s population. When aggregating the users’ population and comparing it with the Census data, we are implicitly assuming that the Census data is representative for Twitter’s users. Research suggests that US-based Twitter’s users are more young and liberal than the general population [24]. However, we do not adjust for how the racial demographics of the general population in a county differ from those who use Twitter within the county.

2.6.5 Definition of race. Following the recommendations in recent work [11, 21], we give a brief description of the sociohistorical context of the racial and ethnic data of the Census to elucidate the limitations of Census data. With strong influence from eugenicists and those looking to quantitatively validate the use chattel slavery, racial categories on the Census were developed in relation to Whiteness [11, 23]. The taxonomy used by the Census is closely tied with the allocation of resources, but has also been used to manage and surveil different populations, and thus the categories have been politically contested throughout its history [11, 23]. For example, in 1930, “Mexican” was added to the Census in part due to increasing anti-Mexican sentiment during the great depression, but the Mexican government lobbied for it to be removed and for Mexicans to be classified as “White” in order to assert Mexicans should not be subject to Jim Crow laws in the US [23]. However, after civil rights, minority groups had more interest in being counted in order to fight discrimination and get access to resources. In 1980, the Census added a specific question that asked first about “Hispanic, Latino, or Spanish origin” followed by a separate race question. Proponents argue two separate questions capture two separate axes of oppression for race and ethnicity and better capture the heterogeneity of the Latino community. However, others argue combining the two questions could lead to a more reflective count of Latinos. More recently, because Arab Americans have been undercounted and disadvantaged in terms of acquiring resources that could help them, some have called for the addition of a MENA category (Middle Eastern and North African). However, given the post 9/11 political climate, others fear it could be used to increase surveillance and policing of this community [23]. These examples illustrate that the Census does not reflect natural or inherent categories, but rather constitutes and reinscribes socially constructed categories [2, 7, 11]. The use of racial/ethnic data here is not meant to reify racial/ethnic categories as natural, but rather study the impact of algorithmic harms on historical marginalized groups [11, 17].

2.6.6 Amplification from other sources. We are only considering Tweets that are scored (i.e. position-manipulated) by the Home Timeline ranking systems. If certain users get more exposure from other sources (for example ads or “Who to Follow” recommendations), such exposure will not be captured in the analysis. Similarly, if an author has a large number of followers who opted into the reverse-chronologically sorted timelines, their normalized impressions may be lower than expected. When computing the ratio, the numerator is not going to include the impressions from those users, while the denominator will stay the same.

---

4For more information about the Home Timeline please refer to https://help.twitter.com/en/using-twitter/twitter-timeline

5More information on how Twitter uses location data can be found on https://twitter.com/en/privacy
3 ANALYSIS AND OBSERVATIONS

In this section, we present the findings of the two biases analyses we conducted. The first is the regression analysis described in Section 2.4. The second is the distribution comparison outlined in Section 2.5.

3.1 Linear relationship between racial composition on amplification

In our first analysis, we show the linear fit between the percentage of the population in a given racial group and the share of amplified users in a county.

Table 4 shows the results of the regression analysis. We run the regression using the percentage of each racial group to predict the share of amplified users. We run both the weighted and unweighted regressions, where the weights are the percentage of Twitter users in that county (as a share of the US total). We do report the regression coefficients, associated 95% confidence interval, and the $R^2$ value (i.e. the share of the variance explained by the linear model).

Figure 1a and Figure 1b visualize the data via scatter plots, where each circle corresponds to county scaled by the size of its population. The y-axis shows the percentage of the population in a racial group, while the x-axis presents the fraction of amplified users.

3.2 Distribution of amplified users by county and racial group

Per each racial group, we split the counties on the US median value, obtaining two groups of almost equal size, see below and Section 2.6.2 for more details. We then measure the distance between the two resulting distributions using the total variation distance. We report the results in Table 2.

There is a caveat to add at this point. While, by definition, there is an equal number of counties in in below and above the median, this does not translate in an equal number of counties when we plot the above histograms (and compute their distances). The reason is that there are 120 counties missing in the final dataset as we could not associate any users to them for reasons described in Section 2.6.2. Out of those, 18 are in the top quantile for the White population and 102 in the bottom one. For the Black population, 90 are in the top quantile and 30 in the bottom one.

3.3 Dataset description

As part of this report, we are releasing the aggregated data used for our analysis. The dataset includes the 3099 counties that we have data for, with one county per row. The columns correspond to features of the county, most provided by the Census data. Out of the 121 columns, only two (twitter_amplified_users_percent and twitter_users_share) are based on Twitter data. We are leaving the Census data in the released dataset for ease of use. The meaning of the two new columns is as follows:

- twitter_amplified_users_percent: share of amplified users in each county. The number is between 0 and 1. For example a value of .65 means that 65% of the users in that county are amplified, i.e. have above median normalized impression.
- twitter_users_share: share of Twitter’s US-based users that are assigned to the county. The number is between 0 and 1. For example a value of 0.014 would mean that 1.4% of all US-based users have been assigned to the county.

4 DISCUSSION

Our analysis is based on counting impressions of Tweets in different counties. As such it gives a coarse representation of visibility on Twitter’s Home Timeline. We see our analysis as providing a piece of quantitative evidence in a growing ecosystem of algorithmic audits of online platforms. Many questions remain.

The analysis is based on roughly a week’s worth of Tweets. It does not begin to answer how possibly small differences in visibility compound over time as expressed in differential growth in the number of followers or diverging usage patterns.

Implicit in our analysis is the assumptions that an unbiased world is one in which each county has the same fraction of amplified users. But deviations from this baseline state of equality may have different causes and interpretations. Counties varies in many ways. Some urban areas might inherently feature more events that people Tweet about. Different groups also use Twitter differently,
with respect to sensitive demographic traits, residential location where people tend to live close to people with similar traits—it might

Timeline.

Tweets from people they do not follow imputed by personalization individual-level racial information. We highlight an unaddressed
suring racial bias on online platforms, especially in the absence of
data may not be as useful as a proxy for bias analysis [20].

history of segregation or where housing is more heterogeneous not generalize well to other countries. In countries without the same
data was to be extended to other countries, analyzing different attributes would likely be more appropriate. Additionally, while
using location based data might make sense for the United States, if this approach of using location and census
race that is unique to the historical and current power structure

as illustrated in a manner relevant to our work by Brock’s account of African American culture on Twitter [3, 4].

Visibility itself is not necessarily a normative goal. Some groups of users may use Twitter for conversations that are not meant to attract a wider audience. Others actively seek a broad audience for their message on Twitter.

When users see Tweets on the Home timeline, they come from a variety of sources: people they directly follow, Topics they follow, Tweets from people they do not follow imputed by personalization algorithms (e.g., based on previous likes), ads, and other kinds of content. It would be interesting to understand what impact each type of suggestion has on the overall amplification. However, it is less clear how to normalize impressions for these different content types. As an example, not all the Tweets liked by someone a user follows will be considered for a suggestion on the user’s Home Timeline.

The methodology we follow does not necessarily generalize to other regions. Because the US census offers a conceptualization of race that is unique to the historical and current power structure of the United States, if this approach of using location and census data was to be extended to other countries, analyzing different attributes would likely be more appropriate. Additionally, while using location based data might make sense for the United States—where people tend to live close to people with similar traits—it might not generalize well to other countries. In countries without the same history of segregation or where housing is more heterogeneous with respect to sensitive demographic traits, residential location data may not be as useful as a proxy for bias analysis [20].

Not least our report gives a glimpse at the challenges of measuring racial bias on online platforms, especially in the absence of individual-level racial information. We highlight an unaddressed
tension between data privacy and algorithmic bias. In other words—the best way to analyze for bias based on a characteristic is to have that information, but the best way to ensure appropriate use is to never collect that data at all.

| Independent variable | Statistic | Coefficient | 95% CI | Weighted | Unweighted |
|----------------------|-----------|-------------|--------|----------|------------|
| Percentage of Black population |          | 0.0524      | [0.0322, 0.0726] | 0.0593    | [0.0358, 0.0827] |
| Percentage of White population |          | -0.1451     | [-0.1609, -0.1294] | -0.0660   | [-0.0862, -0.0458] |

Table 1: Regression coefficients, confidence intervals, and $R^2$ values

| Black | White |
|-------|-------|
| Total variation | 0.172684 | 0.18839 |

Table 2: Measure of the difference between the top and bottom quantile for each race dimension.

| Black | White |
|-------|-------|
| Above Median | Mean | 30.3800 |
| Below Median  | 31.6000 |
| Variance | 3659.1956 | 3028.6000 |
| Std err  | 8.6416 | 7.8618 |

Table 3: Summary statistics of the above histograms.

REFERENCES

[1] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can’t Measure, We Can’t Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 249–260.

[2] Sebastian Bengthall and Bruce D Haynes. 2019. Racial categories in machine learning. In Proceedings of the conference on fairness, accountability, and transparency. 289–298.

[3] André Brock. 2012. From the Blackhand Side: Twitter as a Cultural Conversation. Journal of Broadcasting & Electronic Media 56, 4 (2012), 529–549.

[4] André Brock. 2020. Distributed blackness. In Distributed Blackness. New York University Press.

[5] U.S. Census Bureau. 2018. American Community Survey 5-Year Data. https://www.census.gov/data/developers/data-sets/acs-5year.html

[6] Ho-Chun Herbert Chang, Matt Bui, and Charlton McIlwain. 2021. Targeted Ads and/as Racial Discrimination: Exploring Trends in New York City Ads for College Scholarships. arXiv preprint arXiv:2109.15294 (2021).

[7] Karen E Fields and Barbara J Fields. 2014. Racecraft: The soul of inequality in American life. Verso Books.

[8] David A Freedman. 1999. Ecological inference and the ecological fallacy. International Encyclopedia of the social & Behavioral sciences. 6, 4027-4030 (1999), 1–7.

[9] Arline T Geronimus, John Bound, and Lisa J Neidert. 1996. On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics. J. Amer. Statist. Assoc. 91, 434 (1996), 529–537.

[10] Marian E Gornick, Paul W Eggers, Thomas W Reilly, Renee M Mentnech, Leslye K Fitterman, Lawrence E Kucken, and Bruce C Vladeck. 1996. Effects of race and income on mortality and use of services among Medicare beneficiaries. New England journal of medicine 335, 11 (1996), 791–799.

[11] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In Proceedings of the 2020 conference on fairness, accountability, and transparency. 501–512.

[12] Ferenc Huszár, Sofia Ira Ktena, Conor O’Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic amplification of politics on Twitter. Proceedings of the National Academy of Sciences.
We are presenting the same results as Section 3 but for other racial groups, namely DP05_0078E (Black alone, not Hispanic or Latino) and DP05_0077E (white alone, not Hispanic or Latino). As before we are starting with the scatter plots and the regression coefficients.

\[ \text{RESULTS FOR OTHER RACIAL GROUPS} \]

We are presenting the same results as Section 3 but for other racial groups, namely DP05_0078E (Black alone, not Hispanic or Latino) and DP05_0077E (white alone, not Hispanic or Latino). As before we are starting with the scatter plots and the regression coefficients.

Figure 2: Distribution of share of amplified users between the top and bottom half for each race group.

Figure 3: Distribution of share of amplified users between the top and bottom half for each race group.

Now we present the histograms and their relative statistics.
| Independent variable                                      | Statistic | Weighted  | Unweighted |
|-----------------------------------------------------------|-----------|-----------|------------|
| Percentage of Black population, not Hispanic nor Latino   | Coefficient | 0.0561 | 0.0587     |
|                                                            | 95% CI    | [0.0356, 0.0765] | [0.035, 0.0822] |
|                                                            | $R^2$     | 0.0092 | 0.0076     |
| Percentage of White population, not Hispanic not Latino   | Coefficient | -0.1222 | -0.0830    |
|                                                            | 95% CI    | [-0.135, -0.1102] | [-0.0997, -0.0665] |
|                                                            | $R^2$     | 0.1115 | 0.0300     |

Table 4: Regression coefficients, confidence intervals, and $R^2$ values

|                                | Black alone (no Hispanic or Latino) | White alone (no Hispanic or Latino) |
|--------------------------------|-------------------------------------|-------------------------------------|
| Total Variation                | 0.172727                            | 0.18306                             |
| White alone (no Hispanic or Latino) |                                     |                                     |
| Above Median                   | 31.8000                             | 30.1000                             |
| Below Median                   | 31.6200                             | 30.3600                             |
| Black alone (no Hispanic or Latino) |                                     |                                     |
| Above Median                   | 30.1000                             | 31.6200                             |
| Below Median                   | 2762.7504                           | 3955.7956                           |
| Mean                           | 31.8800                             | 31.6200                             |
| Variance                       | 3445.9856                           | 3955.7956                           |
| Std err                        | 8.3860                              | 8.9850                              |

Table 5: Summary statistics of the above histograms.
