Reducing Population-level Inequality Can Improve Demographic Group Fairness: a Twitter Case Study

Avijit Ghosh, Tomo Lazovich, Kristian Lum, Christo Wilson

Northeastern University

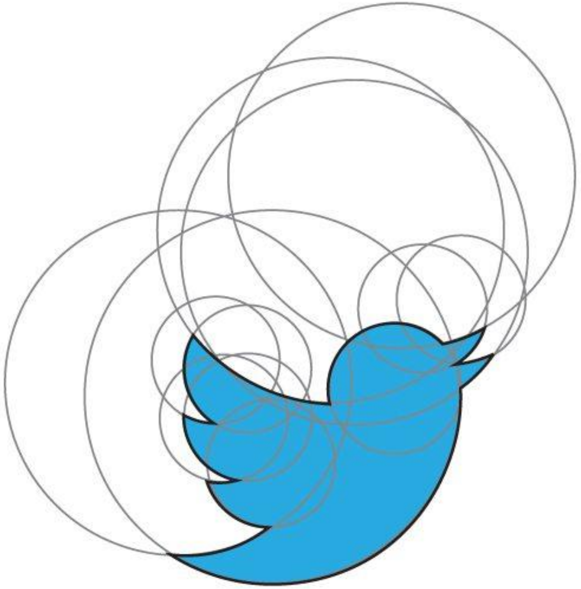
FAccTRec 2024

Overview

A Twitter Case Study

- We explore the relationship between demographic-free inequality metrics and standard demographic bias metrics in the context of engagement inequality on Twitter.

- Findings suggest that inequality metrics can serve as useful proxies for average group-wise disparities in content recommendation scenarios.



Fairness Metrics at Scale

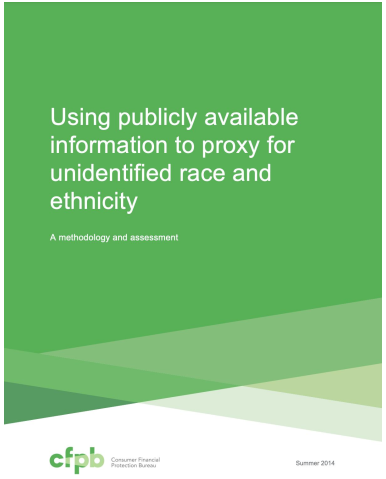
Most proposed fairness metrics have a **caveat**: they require the **knowledge of protected group membership**.

With respect to demographic groups, this has hurdles:

- Difficult for large datasets
- Might be outright illegal based on context
- Privacy concerns

Demographic Classification

Unfortunately, a common workaround is to use **demographic classifiers** that infer the race/gender or other sensitive attribute from people's name, image, zip code, or other information.

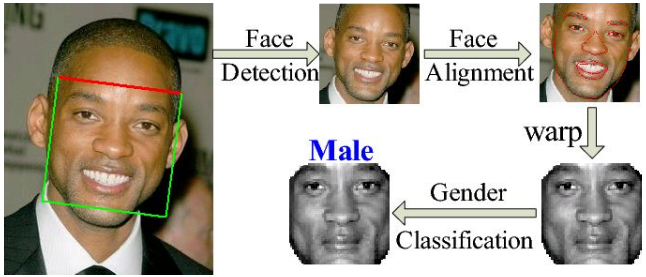


Using publicly available information to proxy for unidentified race and ethnicity

A methodology and assessment

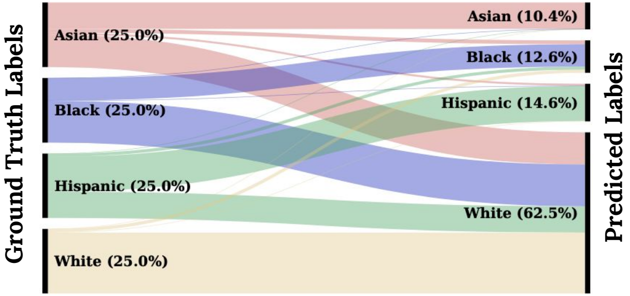
cfpb Consumer Financial Protection Bureau

Summer 2014

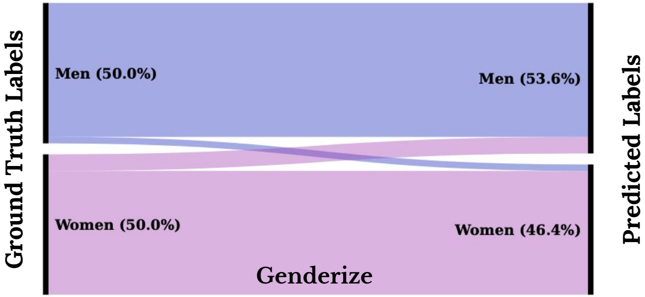


Which is often incorrect!

Name based

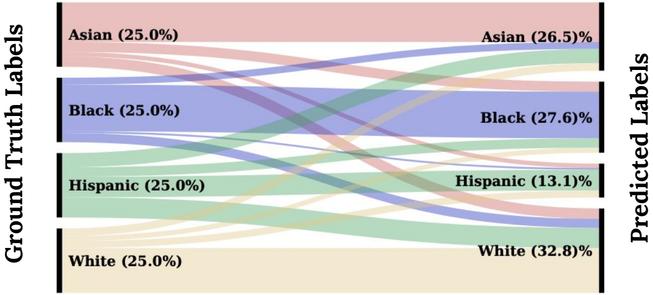


EthCNN

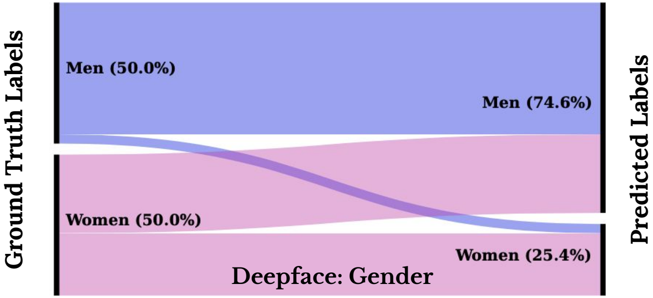


Genderize

Face based

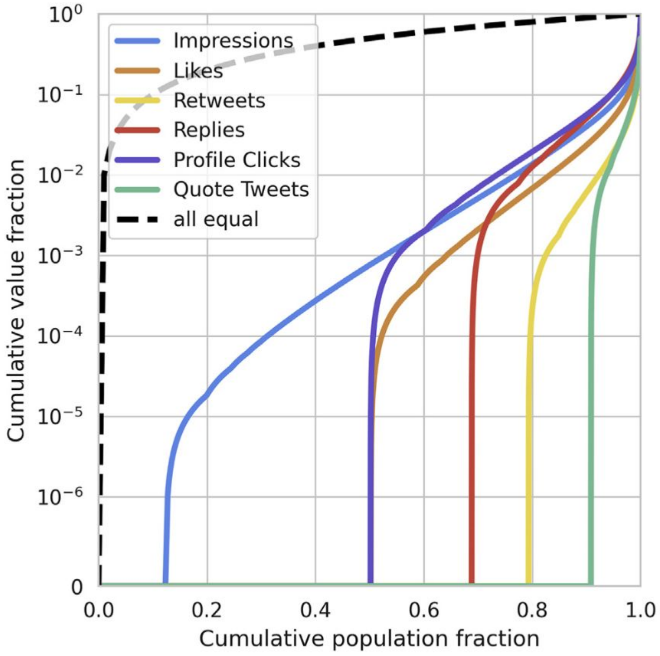


Deepface: Ethnicity



Deepface: Gender

Demographic-Free Inequality Metrics



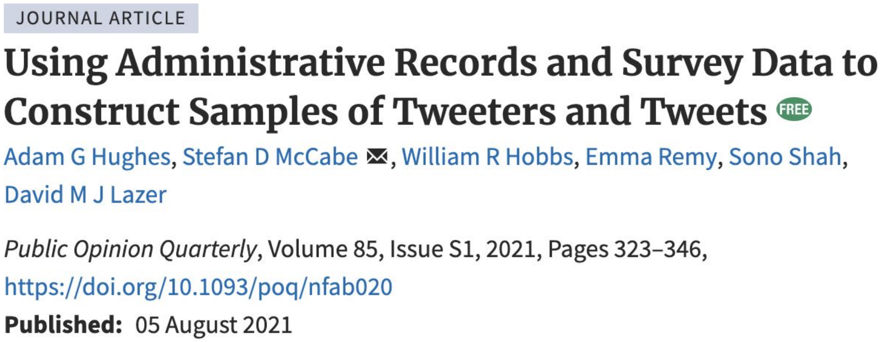
Top 1% of authors receive 80% of all views of Tweets

- **Economic Inequality Metrics: Wealth Inequality**
- **First proposed to measure recommendation bias in Lazovich et. al (2022)**
- **Advantages:** No need for demographic data, measure overall system fairness
- **Unknown:** May not directly translate to demographic fairness
- Conceptually appealing for measuring **system-wide inequality**
- Potential for use in experimentation and **bias mitigation** strategies

Methodology

Data Collection and Analysis Approach

- **269 million tweets** from **174,600 unique authors** who authored at least one tweet in 2021.
- Accounts matched to real users from public voter records provided by data vendor TargetSmart.
- Subset of data collected by Northeastern through Twitter api before the block.



JOURNAL ARTICLE

Using Administrative Records and Survey Data to Construct Samples of Tweeters and Tweets FREE

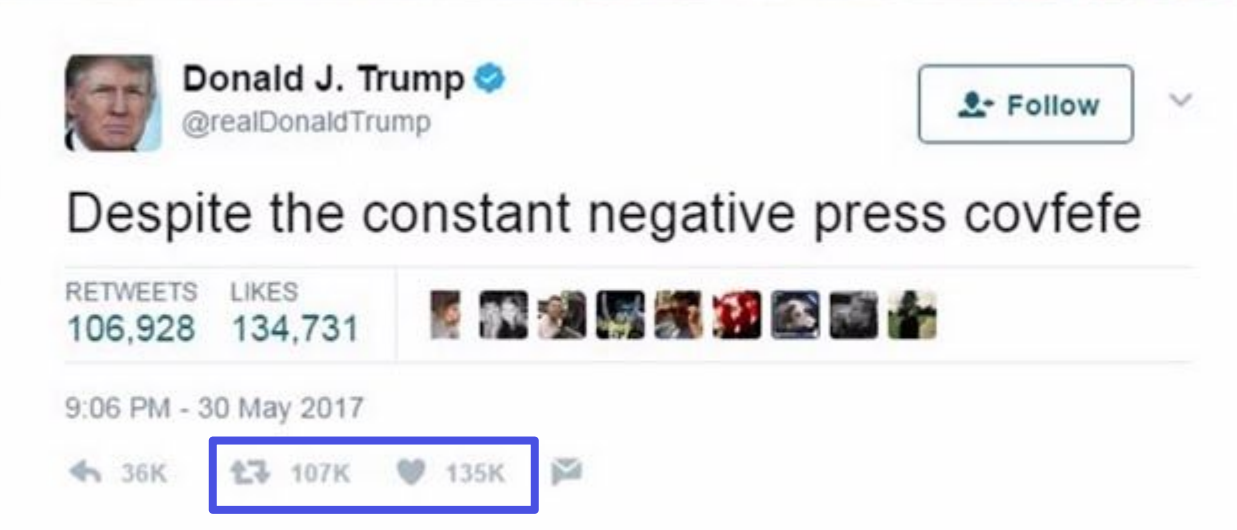
Adam G Hughes, Stefan D McCabe ✉, William R Hobbs, Emma Remy, Sono Shah, David M J Lazer

Public Opinion Quarterly, Volume 85, Issue S1, 2021, Pages 323–346, https://doi.org/10.1093/poq/nfab020

Published: 05 August 2021

Methodology

Data Collection and Analysis Approach



Donald J. Trump
@realDonaldTrump

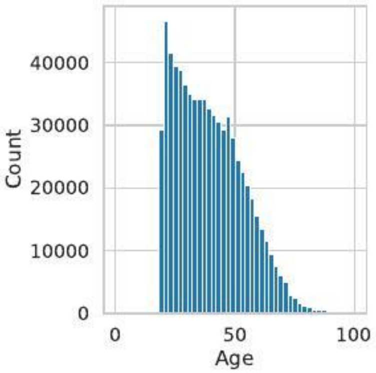
Despite the constant negative press covfefe

RETWEETS 106,928 LIKES 134,731

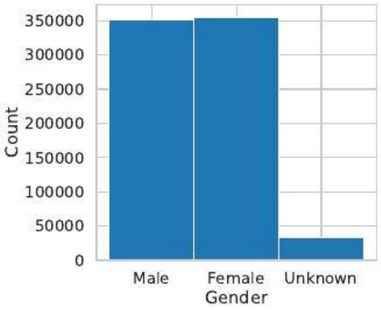
9:06 PM - 30 May 2017

Likes + Retweets = Engagements

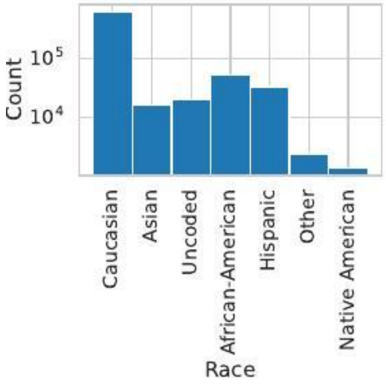
Dataset Demographics



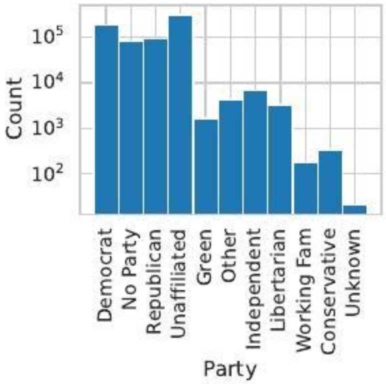
Age Distribution



Gender Distribution



Race Distribution



Political Affiliation

Inequality Metrics Chosen

Gini Coefficient:

- **Purpose:** Measures inequality within a population. Compares the average absolute difference between individuals' engagement to the population mean.
- **Interpretation:** Values range from 0 (perfect equality) to 1 (maximum inequality).

Top 1% Share (T1PS):

- **Purpose:** Measures how much of the total engagement is held by the top 1% of individuals.
- **Interpretation:** A higher value indicates more concentration in the top 1%.

$$\text{Gini} = \frac{\sum_{p=1}^{N} \sum_{q=1}^{N} |E_p - E_q|}{2N \sum_{j=1}^{N} E_j}$$

$$\text{T1PS} = \frac{\sum_{j \in T99} E_j}{\sum_{p=1}^{N} E_p}$$

Demographic Disparity Metrics

Two General Categories of Disparity Metrics:

- **Average Differences:** Metrics that focus on the overall differences between groups, such as **Statistical Parity Difference** and **Equal Opportunity Difference**.
- **Extremes of Disparity:** Metrics that focus on the worst-case group disparities, such as **Disparate Impact**.

Demographic Disparity Metrics

- **Mean Absolute Deviation (MAD):**
  - **Purpose:** Measures the average disparity in engagements received by different demographic groups by comparing each group's average engagement to the overall population's mean engagement.
  - **Interpretation:** A value of 0 means all groups receive the same engagements, while higher values indicate larger demographic disparities.

- **Inverse Min/Max (IMM):**
  - **Purpose:** Measures the worst-case disparity between the group with the highest and lowest average engagements.
  - **Interpretation:** A value of 0 means equal engagements for the most and least engaged groups (maximum fairness), while higher values show more disparity.

$$\mathrm{MAD} = \frac{\sum_{k\in G} |\overline{E_{G_k}} - \overline{E}|}{|G|}$$

$$\mathrm{IMM} = 1 - \frac{\min_{k\in G} \overline{E_{G_k}}}{\max_{k\in G} \overline{E_{G_k}}}$$

Demographic Disparity Metrics

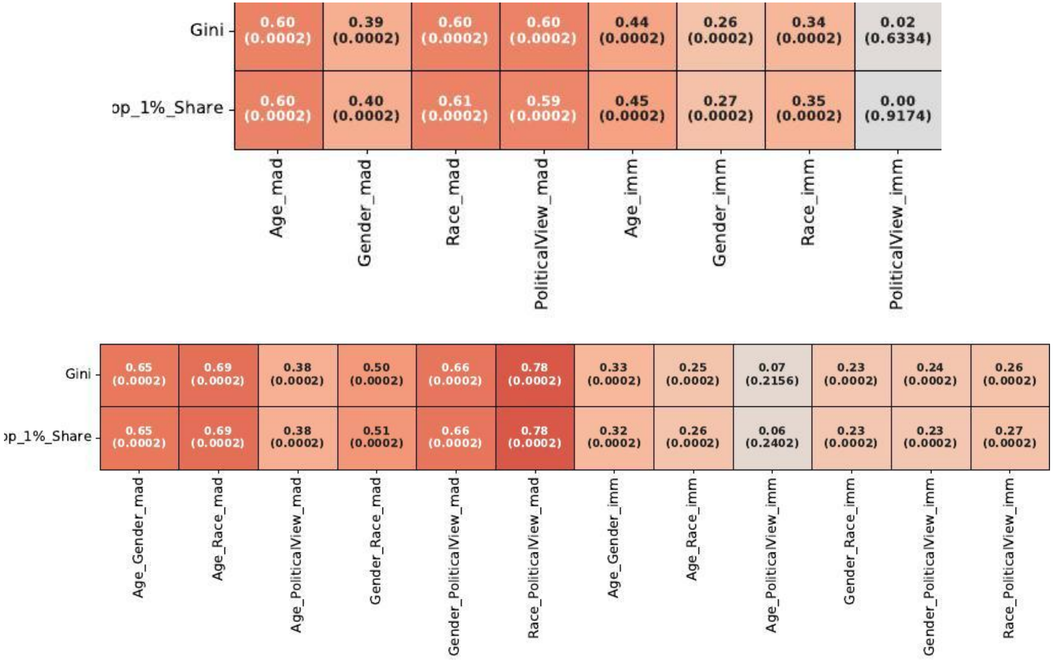
- **Mean Absolute Deviation (MAD):** *Statistical Parity Difference, Equal Opportunity Difference*
  - **Purpose:** Measures the average disparity in engagements received by different demographic groups by comparing each group's average engagement to the overall population's mean engagement.
  - **Interpretation:** A value of 0 means all groups receive the same engagements, while higher values indicate larger demographic disparities.

- **Inverse Min/Max (IMM):** *Disparate Impact*
  - **Purpose:** Measures the worst-case disparity between the group with the highest and lowest average engagements.
  - **Interpretation:** A value of 0 means equal engagements for the most and least engaged groups (maximum fairness), while higher values show more disparity.

$$\mathrm{MAD} = \frac{\sum_{k \in G} \left|\overline{E_{G_k}} - \overline{E}\right|}{|G|}$$

$$\mathrm{IMM} = 1 - \frac{\min_{k \in G} \overline{E_{G_k}}}{\max_{k \in G} \overline{E_{G_k}}}$$

Results: Correlation Analysis

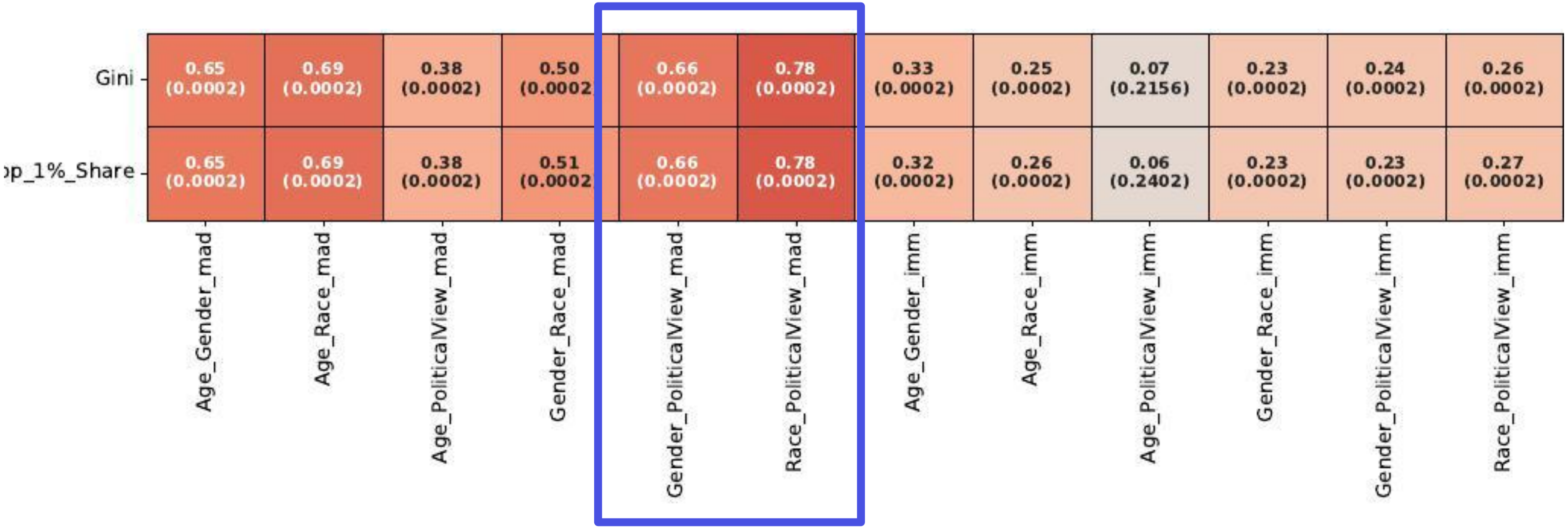


	Age_mad	Gender_mad	Race_mad	PoliticalView_mad	Age_imm	Gender_imm	Race_imm	PoliticalView_imm
Gini	0.60 (0.0002)	0.39 (0.0002)	0.60 (0.0002)	0.60 (0.0002)	0.44 (0.0002)	0.26 (0.0002)	0.34 (0.0002)	0.02 (0.6334)
op_1%_Share	0.60 (0.0002)	0.40 (0.0002)	0.61 (0.0002)	0.59 (0.0002)	0.45 (0.0002)	0.27 (0.0002)	0.35 (0.0002)	0.00 (0.9174)

	Age_Gender_mad	Age_Race_mad	Age_PoliticalView_mad	Gender_Race_mad	Gender_PoliticalView_mad	Race_PoliticalView_mad	Age_Gender_imm	Age_Race_imm	Age_PoliticalView_imm	Gender_Race_imm	Gender_PoliticalView_imm	Race_PoliticalView_imm
Gini	0.65 (0.0002)	0.69 (0.0002)	0.38 (0.0002)	0.50 (0.0002)	0.66 (0.0002)	0.78 (0.0002)	0.33 (0.0002)	0.25 (0.0002)	0.07 (0.2156)	0.23 (0.0002)	0.24 (0.0002)	0.26 (0.0002)
op_1%_Share	0.65 (0.0002)	0.69 (0.0002)	0.38 (0.0002)	0.51 (0.0002)	0.66 (0.0002)	0.78 (0.0002)	0.32 (0.0002)	0.26 (0.0002)	0.06 (0.2402)	0.23 (0.0002)	0.23 (0.0002)	0.27 (0.0002)

Correlation Matrix

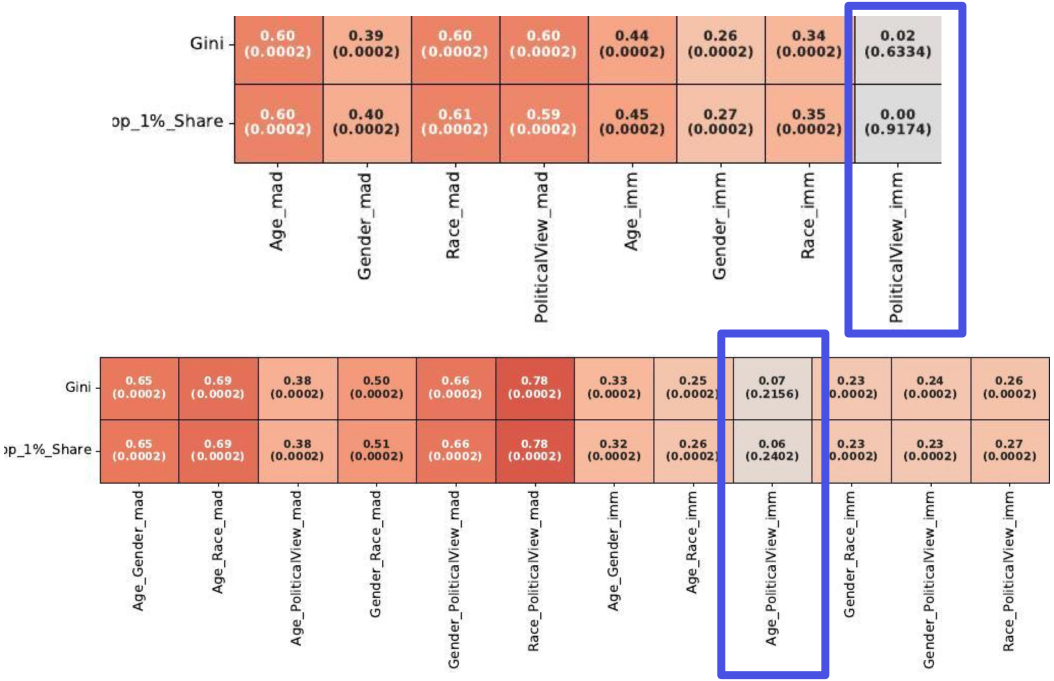
Results: Correlation Analysis



	Age_Gender_mad	Age_Race_mad	Age_PoliticalView_mad	Gender_Race_mad	Gender_PoliticalView_mad	Race_PoliticalView_mad	Age_Gender_imm	Age_Race_imm	Age_PoliticalView_imm	Gender_Race_imm	Gender_PoliticalView_imm	Race_PoliticalView_imm
Gini	0.65 (0.0002)	0.69 (0.0002)	0.38 (0.0002)	0.50 (0.0002	0.66 (0.0002)	0.78 (0.0002)	0.33 (0.0002)	0.25 (0.0002)	0.07 (0.2156)	0.23 (0.0002)	0.24 (0.0002)	0.26 (0.0002)
op_1%_Share	0.65 (0.0002)	0.69 (0.0002)	0.38 (0.0002)	0.51 (0.0002	0.66 (0.0002)	0.78 (0.0002)	0.32 (0.0002)	0.26 (0.0002)	0.06 (0.2402)	0.23 (0.0002)	0.23 (0.0002)	0.27 (0.0002)

Intersectional Metrics exhibit higher correlation (high of 0.78 vs high of 0.6 marginal)

Results: Correlation Analysis



	Age_mad	Gender_mad	Race_mad	PoliticalView_mad	Age_imm	Gender_imm	Race_imm	PoliticalView_imm
Gini	0.60 (0.0002)	0.39 (0.0002)	0.60 (0.0002)	0.60 (0.0002)	0.44 (0.0002)	0.26 (0.0002)	0.34 (0.0002)	0.02 (0.6334)
op_1%_Share	0.60 (0.0002)	0.40 (0.0002)	0.61 (0.0002)	0.59 (0.0002)	0.45 (0.0002)	0.27 (0.0002)	0.35 (0.0002)	0.00 (0.9174)

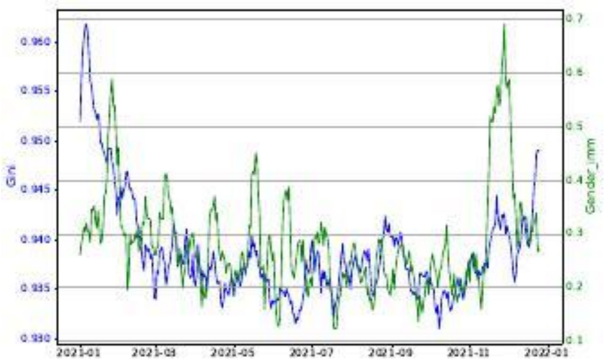
	Age_Gender_mad	Age_Race_mad	Age_PoliticalView_mad	Gender_Race_mad	Gender_PoliticalView_mad	Race_PoliticalView_mad	Age_Gender_imm	Age_Race_imm	Age_PoliticalView_imm	Gender_Race_imm	Gender_PoliticalView_imm	Race_PoliticalView_imm
Gini	0.65 (0.0002)	0.69 (0.0002)	0.38 (0.0002)	0.50 (0.0002)	0.66 (0.0002)	0.78 (0.0002)	0.33 (0.0002)	0.25 (0.0002)	0.07 (0.2156)	0.23 (0.0002)	0.24 (0.0002)	0.26 (0.0002)
op_1%_Share	0.65 (0.0002)	0.69 (0.0002)	0.38 (0.0002)	0.51 (0.0002)	0.66 (0.0002)	0.78 (0.0002)	0.32 (0.0002)	0.26 (0.0002)	0.06 (0.2402)	0.23 (0.0002)	0.23 (0.0002)	0.27 (0.0002)

Political View shows low correlation

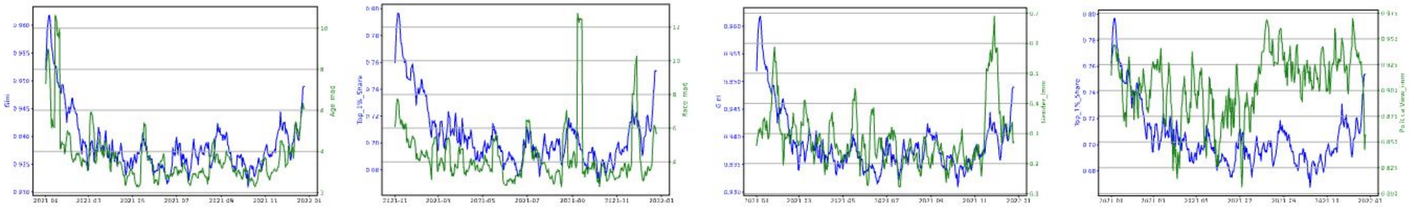
Time Series Analysis

- Daily tracking of inequality metrics and demographic bias metrics throughout 2021.

- Visualized correlations over time.



Time Series Analysis



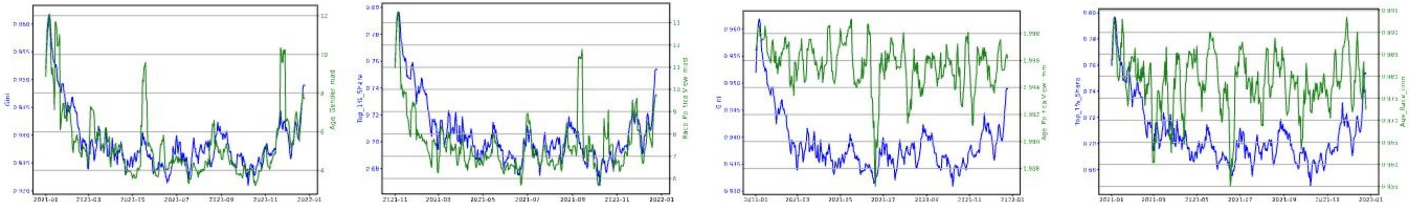
(a) GINI vs. Age (MAD)

(b) T1PS vs. Race (MAD)

(c) GINI vs. Gender (IMM)

(d) T1PS vs. Political View (IMM)

Figure 4: Daily tracking of Inequality Metrics (blue) and Marginal Bias Metrics (green) over 2021.



(a) GINI vs. Age & Gender (MAD)

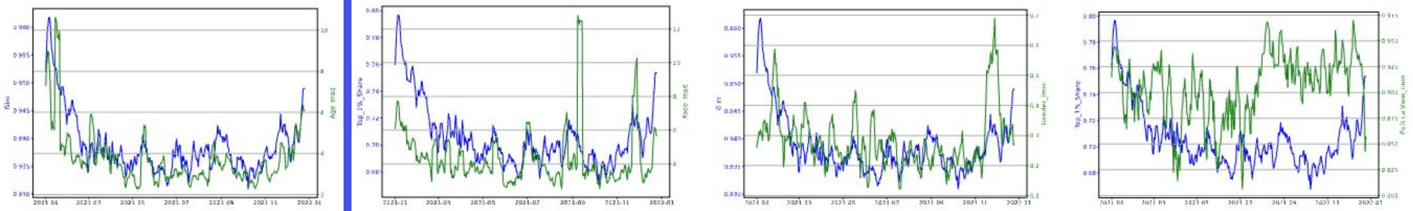
(b) T1PS vs. Race & Political View (MAD)

(c) GINI vs. Age & Political View (IMM)

(d) T1PS vs. Age & Race (IMM)

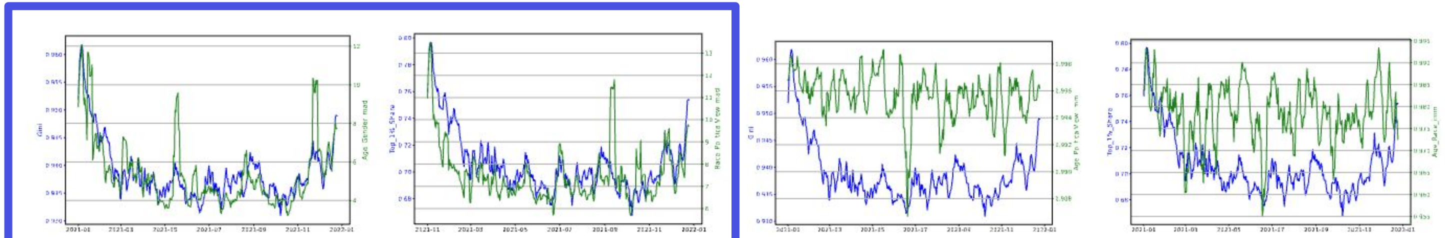
Figure 5: Daily tracking of inequality metrics (blue) and intersectional bias metrics (green) over 2021.

Time Series Analysis



(a) GINI vs. Age (MAD) (b) T1PS vs. Race (MAD) (c) GINI vs. Gender (IMM) (d) T1PS vs. Political View (IMM)

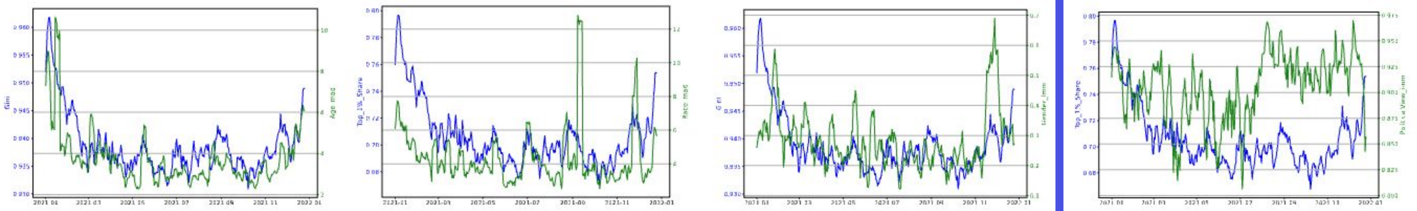
Figure 4: Daily tracking of Inequality Metrics (blue) and Marginal Bias Metrics (green) over 2021.



(a) GINI vs. Age & Gender (MAD) (b) T1PS vs. Race & Political View (MAD) (c) GINI vs. Age & Political View (IMM) (d) T1PS vs. Age & Race (IMM)

Metric pairs with **higher Spearman's correlations** exhibit **tighter correspondence** in time series plots.

Time Series Analysis



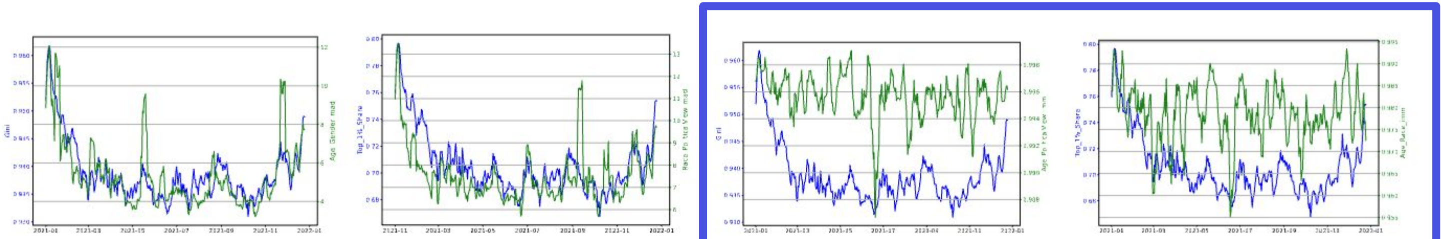
(a) GINI vs. Age (MAD)

(b) T1PS vs. Race (MAD)

(c) GINI vs. Gender (IMM)

(d) T1PS vs. Political View (IMM)

Figure 4: Daily tracking of Inequality Metrics (blue) and Marginal Bias Metrics (green) over 2021.



(a) GINI vs. Age & Gender (MAD)

(b) T1PS vs. Race & Political View (MAD)

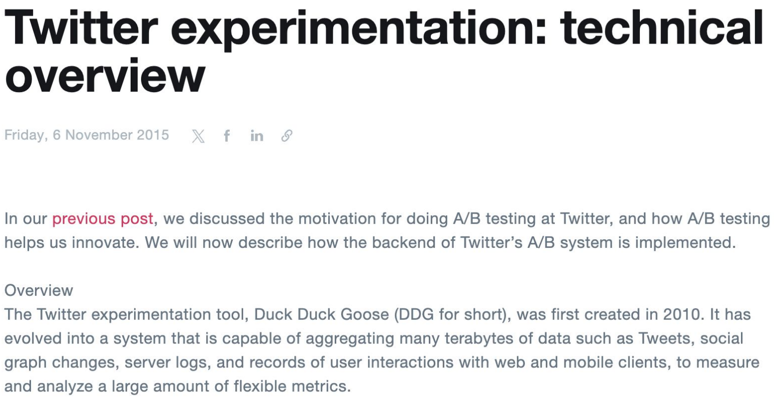
(c) GINI vs. Age & Political View (IMM)

(d) T1PS vs. Age & Race (IMM)

Metric pairs with **lower Spearman's correlations** exhibit **little/no correspondence** in time series plots.

Limitations and Future Work

- **Engagement vs. impression inequality:** Analyze impression data for direct platform insights
- **Dataset coverage:** Extend analysis to global Twitter user base and other platforms
- **Influencer effects:** Separate natural popularity differences from demographic disparities
- **Causal experiments:** Perform A/B tests to determine impact on demographic disparities



Twitter experimentation: technical overview

Friday, 6 November 2015

In our previous post, we discussed the motivation for doing A/B testing at Twitter, and how A/B testing helps us innovate. We will now describe how the backend of Twitter's A/B system is implemented.

Overview

The Twitter experimentation tool, Duck Duck Goose (DDG for short), was first created in 2010. It has evolved into a system that is capable of aggregating many terabytes of data such as Tweets, social graph changes, server logs, and records of user interactions with web and mobile clients, to measure and analyze a large amount of flexible metrics.

https://blog.x.com/engineering/en_us/a/2015/twitter-experimentation-technical-overview

Thank you! Questions?

Link to paper