Auditing Exposure Bias on Social Media for a Healthier Online Discourse

Nathan Bartley,[1] **Keith Burghardt,** [1] **Kristina Lerman**[1]

[1] Information Sciences Institute
University of Southern California
nbartley@isi.edu

Abstract

Just as they are responsible for curating *what* users see in their personalized feeds, social media platforms also curate *who* you see. This is not a common object of study in platform audit studies, as the unit of analysis tends to be found in the content. With approximately 24 sock puppet accounts on X (formerly known as Twitter), we show that biased simple interactions skew a users' perception of their local network when compared to random simple interactions. We further show that reverse chronological and personalized timelines behave differently, with personalized timelines showing significantly underestimated proportion of pro-science content than reverse chronological timelines. This suggests that personalization mixes with end-user activity to expose users to significantly skewed groups of friends than similar users using reverse chronological timelines. This has implications for the analysis of social media with regard to both the spread of harmful narratives and how social media may impact individuals' mental health.

Introduction

The complex recommender systems that construct personalized timelines in online platforms mediate users' access and attention to information. By design these systems rank and reorder content in users' feeds, influencing the type of content users see and how often they see it. How do these algorithms shape individual's perception of what content is popular within their networks? This is an important question as we collectively grapple with the spread of harmful narratives relating to the trust in science (whether it be regarding the safety of vaccines or extent of climate change), prejudices like misogyny and general misinformation.

Social psychological literature as well as social sensing literature describes the utility and accuracy of humans' social judgements, i.e., their perceptions of their immediate social network connections (Galesic et al. 2021), and there is evidence supporting the claim that humans have generally an unbiased social judgment apparatus that is trying to make sense of the statistically biased world around them. Humans can be generally accurate in inferring properties of their immediate network (Galesic et al. 2018), but tend to show biases when making inferences about the broader population (Galesic, Olsson, and Rieskamp 2012). Given the structural statistical phenomena that appear in networks like the friendship paradox (Feld 1991; Lerman, Yan, and Wu 2016) and homophilic social networks (McPherson, Smith-Lovin, and Cook 2001), it is apparent that people would project their perceptions of their local environment onto larger populations (Lee et al. 2019).

This perception can interact in intricate ways with recommender systems, which themselves have become an object of study for potentially creating and/or reinforcing biases in the content they surface for users. While there is increasing evidence (albeit conflicting) for recommender systems in social networks creating more partisanship (Ribeiro et al. 2020; Chen et al. 2021), there is still a clear utility in these systems for sifting through the cognitively overwhelming amount of information being created on the Internet. Regardless of whether or not social media polarization is caused by selective exposure to information or filtering from algorithmic systems, it is clear that the systems are responsive to user behavior and mediate information exposure to some degree. As such, this study is focused on understanding the difference between algorithmic timelines and how they might skew individual users' perceptions of their networks, specifically the prevalence of certain types of information.

We center our study on the perception of science on social media as recent work in studying the uptake of vaccines suggests that passive usage of social media has been negatively associated with vaccine uptake, an essential part of public health (McKinley and Limbu 2023).

To assess the impact of different algorithmic timelines on perception, we conduct a "sock puppet" audit on Twitter from 2021-2022 with relation to the Covid-19 vaccines. We create 24 artificial accounts called "sock puppets" which follow the same set of 100 pro- and anti-science accounts but each controlled account behaves differently with respect to how they use the platform. Each account will log-in, observe either a reverse chronological or personalized timeline, and will "like" tweets in a biased manner in attempts to drive personalization of their timeline. These pro- and anti-science accounts were labeled in a previous study based on polarization scores primarily based on the URLs that they and similar accounts to them share (Rao et al. 2021).

In this work we answer the following research questions:

- **RQ1.** Do biased sock puppet accounts experience significantly different exposure bias than unbiased sock puppet accounts?
- **RQ2.** Do personalized sock puppet accounts experience different levels of exposure bias than reverse chronological sock puppet accounts?
- **RQ2a.** Does exposure bias stabilize over time?

Related Work

Auditing Algorithmic Systems

There has been a tremendous growth in interest in the past few years in auditing algorithmic bias. Fewer works however have focused on bias in online social networks (OSNs), with even fewer works on the personalized news feeds algorithms one can expect to see on major platforms. In line with Sandvig et al., methods that audit these platforms tend to fall in the following categories: 1) creating artificial agents that scrape data from the interface; 2) recruit real users to install a software browser extension to scrape pages users are exposed to; and 3) simulating agents and interacting with platforms via the platform API (Sandvig et al. 2014).

On YouTube Ribeiro et al. used untrained sock puppet accounts to see how following recommendations can lead some users from moderate to more extreme political videos (Ribeiro et al. 2020). Similar results were found by Ledwich and Zaitsev (Ledwich and Zaitsev 2019). While these are relevant studies, the YouTube context is substantially different from platforms like Twitter with regards to how recommendations work (e.g., there is more social network context given on Twitter and many recommendations are given in order at once).

Meta researchers have been interested in measuring algorithmic bias for political content on the platform for nearly a decade. In Bakshy et al. they studied the kind of ideological slant users get their political news from (Bakshy, Messing, and Adamic 2015). Assessing the diversity of URLs clicked, they found partisan users interacting and getting exposed to cross-ideological content (albeit asymmetrically), stating that user choices are in part responsible for the content they see. Guess et al. showed that for a set of users on Instagram and FaceBook the personalized and chronological feeds had significant differences in what content they were presenting to users (e.g., the proportion of users' feeds coming from friends was significantly higher in the personalized versus chronological timeline). However the feeds did not seem to have an impact on political attitudes or behavior (Guess et al. 2023) in the 2020 election. Our work is more focused on assessing the distribution of traits (i.e., the prevalence of being pro-science) across a user's observed friends, impacting the perceived prevalence of such traits.

Another relevant Meta study is Yan et al. which surveyed FaceBook users' satisfactions with their friend counts via a large user sample and attempts to explain that satisfaction with a weighted perception bias (Yan et al. 2022). Their work extended the local perception bias measure introduced by Alipourfard et al. by incorporating heterogeneous weighted user information, including information from a user embedding of an interaction graph (Alipourfard et al. 2020). Our work is focused on explaining how this local perception bias differs between feed algorithms and biased user behavior types.

Twitter (before X) conducted it's own internal analysis and published two relevant studies. In Lazovich et al. the authors focused on describing the utility of distributional inequality metrics and their utility in studying the outcome of content recommendation systems (Lazovich et al. 2022). Huszar et al. studied the amplification of news sources on Twitter in various countries and identify right-leaning sources and politicians as getting more favorable algorithmic amplification (i.e., reaching more users in a personalized timeline than a reverse chronological one) (Huszár et al. 2022). We do not take a global perspective in our work as we are interested in the individual user perception experience over time on the platform.

For sock puppets on Twitter, two studies are the most relevant. Bandy and Diakopoulos used archetypal sock puppet accounts (accounts that are representative of different communities on Twitter at the time of writing). They found that algorithmic timelines reduce exposure to external news sources (via URLs) when compared to chronological timelines, however the composition of sources is rather stable over time. This suggests a minor role of the algorithm in directing news exposure (Bandy and Diakopoulos 2021).

The second study is from Bartley et al. (Bartley et al. 2021). The authors set up eight accounts that followed the same groups of users and compared the popularity bias and exposure bias between personalized and reverse chronological conditions. They did not find significant difference in exposure bias but found personalized timelines serving more tweets that were ultimately more popular. They also set their sock puppets to not interact with tweets.

A recent study by Vombatkere et al. presented a framework for measuring the amount of personalization a user is experiencing through a personalization score computed over a user trace on platforms like TikTok (Vombatkere et al. 2024). Even though this study is not explicitly about perception bias, this framework could be useful for assessing the extent to which personalization is responsible for perceived biases.

While not explicitly an OSN, Google Search has also been subject to audit studies. Robertson et al. recruited participants to complete a survey and install a browser extension that enabled the authors to collect their query result pages (Robertson et al. 2018). The authors found little supportive evidence for ideological filter bubbles in search engines. Other lines of research in search engines focus on directly controlling for bias contained in the data that works as input data for the search engine algorithmic system (Kulshrestha et al. 2017).

To the last strategy of audits, researchers now are working to simulate partisan users on Twitter (now X) and expose biases in timelines via API (**?**). With simple behavior functionality, they constructed bots to interact with tweets, generate tweets, and choose to follow/unfollow people. They find conservative accounts observing more low-credibility

content, and liberal accounts observing more ideologically moderate content.

Social Media Usage & Effects

There is a wealth of research associating the usage of social media with the prevalence of eating disorders, especially among young women across cultures (Dane and Bhatia 2023). As to the possible mechanisms causing this, young women seem to compare themselves with close and distant peers, but even moreso with influencers, leading to potential body dissatisfaction (Pedalino and Camerini 2022). Social media literacy and other internal factors like body appreciation seem to generally be protective against developing disordered eating and body image issues, suggesting that social comparison mechanisms can be mitigated by how users process the information they get from their personalized social media (Dane and Bhatia 2023). This study is focused on assessing the network information that users get exposed to, before they would generate any perceptions of their networks.

Young men are also susceptible to online social media influence, with usage being associated with risk of body dissatisfaction (Stein, Krause, and Ohler 2021) and possibly associated with the risk of adoption of misogynistic beliefs that can lead to enacting violence (Koester and Marcus 2024).

As it pertains to vaccine uptake and hesitancy, studies often present conflicting results depending on the platform being studied and the type of social media usage. Information-seeking active social media use is associated with vaccine intent, whereas passive exposure is more broadly observed to be negatively associated with intent/uptake (Jabbour et al. 2023; McKinley and Limbu 2023; Mascherini and Nivakoski 2022). This study is focused on passive social media exposure, with minor interactions with the recommender system.

Methods

In this section we describe our experimental methodology and how we set up our sock puppets to audit the two different timeline conditions.

We adopt methods described in Bandy and Diakopoulos and Bartley et al., where we implement Selenium automated accounts, or sock puppets, that log-in to Twitter at scheduled times during the day and are presented with a timeline of tweets recommended by Twitter. We deploy 24 accounts using Selenium split across six conditions (Bandy and Diakopoulos 2021; Bartley et al. 2021):

1. **Random personalized**. The automated account will browse normally under a "For You" timeline, and like any tweet uniformly at random.
2. **Random reverse chronological**. The account will browse under a "Following" timeline, and like any tweet with a uniform probability.
3. **Pro-science personalized**. The account will browse under a "For You" timeline, and like pro-science friends' and their endorsed tweets with a higher probability than anti-science.
4. **Pro-science reverse chronological**. The account will browse under a "Following" timeline, and like any pro-science friends' and their endorsed tweets with a higher probability than anti-science.
5. **Anti-science reverse chronological**. The account will browse under a "Following" timeline, and like any anti-science friends' and their endorsed tweets with a higher probability than pro-science.
6. **Anti-science personalized**. The account will browse under a "For You" timeline, and like any anti-science friends' and their endorsed tweets with a higher probability than pro-science.

Each account was set up to follow 100 total users, 50 of whom are listed as pro-science and 50 of whom are labeled anti-science per Rao et al. These labels are assigned according to the polarization score each user has, which is computed primarily based on the type of URLs shared (Rao et al. 2021).

The accounts were active over a period from 04/12/2021 until 02/02/2022. The sequence of actions a sock puppet would take can be described as follows:

1. A sock puppet would run Selenium and connect to Twitter
2. If necessary the sock puppet would log-in and handle any pop-ups that occur in the user interface
3. A sock puppet would scroll down the screen and like a tweet with fixed probabilities according to their condition until at least 30 tweets were observed. If a user is pro-science polarized they will like friends' tweets that are themselves pro-science with probability 0.35, all others 0.03, to get approximately one like per session. All tweets were recorded including promoted tweets.

Data was collected five times daily (at 9am, 12pm, 3pm, 6pm, 9pm), however accounts were sporadically inactive until remedied. Missing data is largely due to three errors: 1) unexpected runtime error; 2) unexpected changes in HTML/server-side code that prevented gathering data; 3) accounts were suspended. When sock puppet accounts were suspended, new accounts were made and set to run-up again as soon as possible. Similarly, when accounts that the sock puppets follow were suspended or otherwise deleted, each sock puppet was updated with a new friend to maintain an equal balance in friends. The sock puppet accounts were run on four machines that would connect to the Twitter website via a proxy server to control for location information each account would give. We also filter all promoted tweets and only analyze tweets that were cleanly parsed and attributable to an initial friend (i.e., a tweet from user Z that is liked or retweeted by friend Y is given the same label as friend Y).

In parallel we collected tweets and retweets for each of the friends the sock puppets were following and as many friends-of-friends as possible from the Twitter API using the Home Timeline endpoint. We regularly scanned the endpoint for each encountered user and gathered any new activity since the last scan. This "actual" activity data serves to act as a baseline that we compare the timelines against.

Measurements

In this section we describe the metrics we use to evaluate exposure bias. We use the Gini coefficient and the measure of local bias from Alipourfard et al. (Alipourfard et al. 2020). This measure is useful as it reflects a difference between the expected fraction of friends that have a specific trait and the global prevalence of that trait in the network. In other words, this would reflect each user's experience of their own local network compared against the whole network.

To compute the Gini coefficient we use the following strategy:

- For each sock puppet, create a matrix X of dimensions n_{users} x n_{days} the number of total users that have been observed over the course of the study for every day of the study.
- Each position x_{ij} contains the number of tweets viewed by the particular sock puppet of user i on day j.
- The value is then calculated according to the following formula:

$$G = \frac{\sum_{i=1}^{n} \sum_{i=1}^{n} |x_i - x_j|}{2n^2 \bar{x}}$$

To compute the local bias measure we use the following strategy:

- For each sock puppet we keep track of the number of tweets observed from each friend every day of the study in the same way we track it for the Gini coefficient.
- We then binarize the matrix to reduce it to a matrix of the edges observed connecting the sock puppet to its network.
- The local bias measure is calculated as:

$$B_{local} = \mathbb{E}[q_f(X)] - \mathbb{E}[f(X)]$$

where

$$\mathbb{E}[q_f(X)] = \bar{d} \cdot \mathbb{E}[f(U)A(V)|(U,V) \sim \text{Uniform}(E)]$$

$$q_f(v) = \frac{\sum_{u \in F(v)} f(u)}{d_i(v)}$$

$$A(v) = \frac{1}{d_i(v)}$$

where in our study $f(u) \in \{0, 1\}$, such that $f(u) = 1$ if the user is pro-science, and $f(u) = 0$ if the user is anti-science. It is assumed that the true prevalence $\mathbb{E}[f(X)] = 0.50$ as we are looking at the same network for each account and are assessing an over or under representation of pro-science users.

We structure our analysis around two parts: bootstrapped sessions aggregated over time to help understand the overall variability (and potential skew due to missing data) and time series data to observe possible trends over time.

Results

The date coverage of the sock puppets are described in Fig. 1. We observe that the coverage is weakest for the anti-science biased reverse chronological condition, followed by the pro-science biased reverse chronological condition. This is likely due to the semi-regular HTML issues that were more pernicious for the reverse chronological condition accounts.

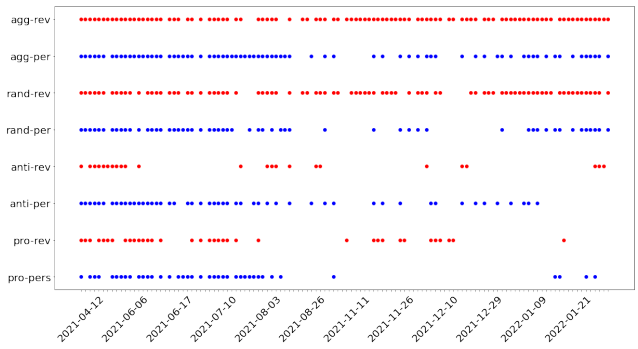

Figure 1: Dates covered by the sock puppet accounts. Agg refers to the aggregated accounts all with the reverse chronological or personalized condition.

The local bias results of the study over this date range are presented in Fig. 2. We observe significant differences between the pairs of conditions (pro and anti, pro and random, anti and random), and all ranges are distinct from the bias computed on the total "actual" activity (i.e., what you would expect if you observed all the friends that tweeted that day).

The Gini coefficient results are presented in Fig. 3. We observe significantly higher personalized Gini in the random personalized condition versus the reverse chronological as well as the Gini coefficient derived from the actual activity of all friends.

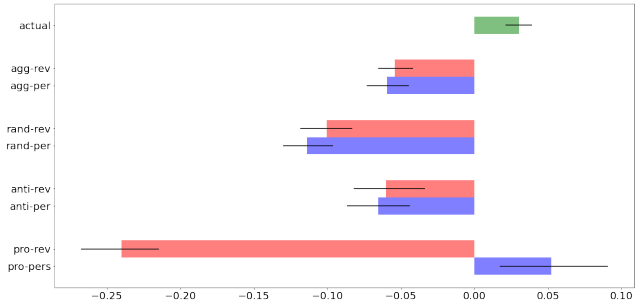

Figure 2: Local bias measurements for audit groups. Bootstrapped confidence intervals computed over 1000 samples of sessions, presented are the mean of the daily means with 95% intervals. Negative implies under-representation of pro-science content, positive implies over-representation. Agg refers to the aggregated accounts under the reverse chronological or personalized conditions. Actual refers to the measure derived from the actual friend activity over the course of the study.

The two measures of exposure bias were also subjected to independent t-tests to assess significance of the differences in the means computed over the course of the study.

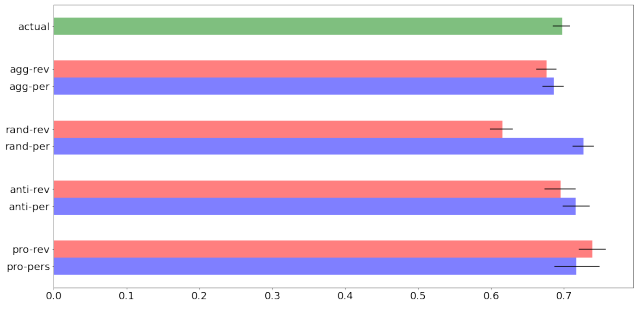

Figure 3: Gini coefficient measurements for audit groups. Bootstrapped confidence intervals computed over 1000 samples of sessions, presented are the mean of the daily means with 95% intervals. Zero implies equality in mean number of tweets observed for each friend, One implies one friend dominates. Agg refers to the aggregated accounts under the reverse chronological or personalized conditions. Actual refers to the measure derived from the actual friend activity over the course of the study.

As all bootstrapped figures use means computed over the course of the study, we also present the relevant time-series values in Fig. 4 and 5. We similarly plot the linear regression line to identify trends in each time-series, and in Fig. 4 we observe a general downward trend for the aggregate personalized timelines, and a general upward trend for the aggregate reverse chronological timeline. Both are significantly different than the trend observed in the measurements on the actual activity. In Fig. 5 we observe a positive trend in the aggregate personalized timeline over time, and a steady Gini coefficient trend for aggregate reverse chronological. As these are computed weekly we observe higher coefficients than in the aggregate plot in Fig. 3.

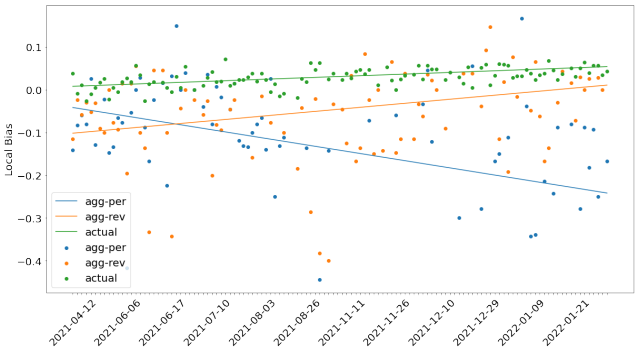

Figure 4: Local bias computed over time. Agg refers to the aggregated accounts under the reverse chronological or personalized conditions. Actual refers to the measure derived from the actual friend activity over the course of the study.

To account for possible correlates in the activity of friends, we independently take the correlation of both the local bias and Gini coefficients versus the daily actual activity of pro-science friends and anti-science friends. We observe in Fig. 6 that the aggregate personalized timeline condition has a significant positive correlation to the anti-science friend activity time-series, similar to the correlation the pro-science friend activity has with the anti-science friend activity timeline. We also observe the actual activity bias measures to be negatively correlated with anti-science friend activity, consistent with anti-science friends having the neutral label 0 and pro-science friends having the positive label 1. Similarly reverse chronological timelines generally have no or weak correlation with either activity time-series.

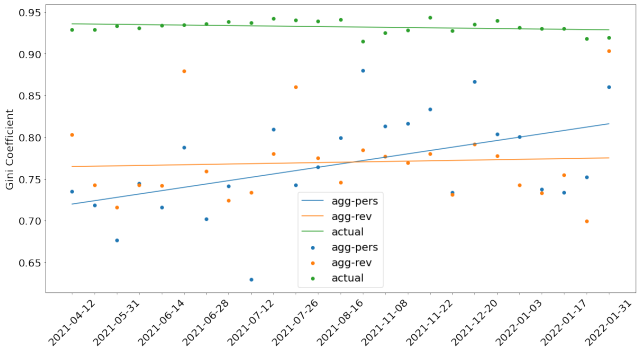

Figure 5: Gini coefficient computed weekly over time. Agg refers to the aggregated accounts under the reverse chronological or personalized conditions. Actual refers to the measure derived from the actual friend activity over the course of the study.

RQ1. To address the first research question of whether the biased sock puppet accounts are significantly different than the random ones, we find that under independent t-tests presented in Table 1 both biased conditions were significantly different from their randomly unbiased counterparts (the only exception being for anti-reverse chronological and random-reverse chronological).

RQ2. To address the second research question of differences between personalized and reverse chronological timelines, under independent t-tests described in Table 1 personalized timelines in all conditions except pro-science biased accounts were significantly different under both metrics to their reverse chronological counterpart. Interestingly, the pro-science reverse chronological timeline feeds were not significantly different than the pro-science personalized timeline in their Gini coefficient per Table 1.

RQ2a. While we observe relatively stable trends for both actual activity based time-series and aggregate reverse chronological time-series in Fig. 4 and 5, we observe nonzero trends for the aggregate personalized timeline suggesting that under personalization no stabilization occurred over the course of the study.

Missing data sensitivity. To analyze the impact of missing data, we use several imputation methods to analyze the difference in the results we observe for RQ2a. We present the results for the median imputed results here in Fig. 7, and the other imputed results in the appendix. We observe nonzero trends in roughly the same direction as we observe

in the non-imputed data in Fig. 4 suggesting a minimal impact of the missing data on the time-varying nature of the results.

Discussion

Correlation in exposure bias measurement and friend activity provides the clearest explanation for the differences between personalized and reverse chronological conditions. We expect the local bias of the actual activity to correlate negatively with anti-science friend activity, and positively with pro-science activity. This is because positive local bias implies more pro-science friends are being observed by sock puppets, and negative implies fewer pro-science friends seen. Similarly, we expect reverse chronological conditions to largely be uncorrelated with the amount of friend activity as it we would expect it to be more correlated with the time each friend is active relative to when the sock puppets log-in and observe them. Interestingly we observe in Fig. 6 a positive correlation with the anti-science friend activity when looking at the aggregate personalized (agg-pers), pro-science biased personalized (pro-pers) and pro-science biased reverse chronological (pro-rev), albeit to different degrees.

Given the positive correlation between pro-science friend activity and anti-science friend activity it may be a spurious correlation. However, as the reverse chronological conditions have either the same or weaker correlation with the friend activity, it seems less likely that the same spurious correlation would affect multiple timeline conditions. A possible explanation is that recorded anti-science activity dropped in the second half of the study period such that any increase in activity would generally lower the fraction of pro-science friends in the inventory of tweets to serve.

We present the activity of the anti-science and pro-science users as recorded by our API sweeps in appendix Fig. 8 and 9. While there is a sharp drop in activity in anti-science activity around 2021-10-24, this can be attributed to both some accounts being temporarily suspended and a brief API rate limit issue we encountered. However, given the time-series measures of local bias and Gini coefficient in Fig. 4 and 5 we do not observe a significant impact on the measures at that time, suggesting an insignificant impact. Therefore we believe the correlations discussed above are not simply spurious.

RQ1. The t-test results show consistent significant differences between biased and random feeds except for the anti-science reverse chronological feeds, and also show that the pro and anti-science feeds experienced significant differences in the direction that we would predict for the local bias measure (i.e., that pro-science personalized would be positive and much higher than random personalized, and anti-science personalized would be negative and more negative than random personalized). Because of this we claim that we can measure the change in users' structured perceptions of who they follow.

It makes sense that biased accounts who interact with similar users would observe more users along that skew, as that seeming measure of utility would be reflected in the personalized algorithm. Similar results have been described in analysis of how user activity and personalization interact in the YouTube video recommender system (Ribeiro, Veselovsky, and West 2023).

RQ2. Given the potential coverage issues across the conditions, we primarily discuss the aggregate personalized and reverse chronological results. We observe significantly lower local bias for the aggregate personalized conditioned compared to the reverse chronological condition which is consistent for all pairs except for the pro-science biased condition (which was flipped) and the anti-science biased condition (which had no significant difference). Similarly we observe a higher Gini coefficient for the aggregate personalized when compared to the aggregate reverse chronological condition. In Fig. 5 we observe an increase in aggregate personalized Gini over time, which correlates inversely with the anti-science friend tweet activity per Fig. 6. Given that both aggregate personalized and reverse chronological both negatively correlate with the anti-science friend activity (and both of these are flipped in sign compared to the actual activity Gini computation), we consider this insignificant.

We claim that given the significant t-test results, as well as the differences in correlation between aggregate personalized and aggregate reverse chronological when compared to each other that there is a measurable difference in exposure bias between the conditions.

RQ2a. We observe relative stability in the actual activity (defined by having little to no trend in their linear regression models) and aggregate reverse chronological condition plotted in Fig. 5. We do not however observe stability for either aggregate timeline over the course of the study according to the local bias measure. This could be the case because we had accounts (both friends and sock puppets) that were suspended and needed to be replaced. Alternatively this could be explained by changes in friend posting behavior or timelines; because users that are out of your network can appear in your timelines, the personalized feeds perhaps saw consistent changes in who they were exposed to. We observe oscillation in the total number of unique friends both conditions saw in appendix Fig. 16, however we observe a drop in the number of new friends seen after 2021-07 in appendix Fig. 17, suggesting it is not new users in the feed.

We observe a possible drop in the total number of usable personalized tweets in appendix Fig. 10 that we do not observe in Fig. 11: we had more difficulties with HTML changes and suspensions in the personalized sock puppets, which could explain the lack of stability over time. Similarly, we observe a drop in the total number of tweets per day for anti-science users around 2021-10-24 in appendix Fig. 8 which might explain the lack of stability for the aggregate personalized condition. We do not report it here, but computing the correlation matrices for each of the measurements until 2021-10-24 (rather than through until 2022-02-02) has some of the correlations going the opposite direction, suggesting the drop in activity impacts the trend overtime, even if it does not seem to affect the trend in the actual activity metrics.

Feed 1	Feed 2	Condition 1	Condition 2	B_{local} t-test Score	p-value	Gini t-test Score	p-value
Personalized	**RevChron**	**Anti**	**Anti**	-309.72	$< 10^{-10}$	97.46	$< 10^{-10}$
Personalized	**RevChron**	**Anti**	**Random**	-369.84	$< 10^{-10}$	136.71	$< 10^{-10}$
Personalized	**RevChron**	**Anti**	**Pro**	-315.89	$< 10^{-10}$	53.15	$< 10^{-10}$
Personalized	**RevChron**	**Anti**	**Aggregate**	-395.39	$< 10^{-10}$	159.60	$< 10^{-10}$
Personalized	**RevChron**	**Pro**	**Anti**	162.14	$< 10^{-10}$	36.36	$< 10^{-10}$
Personalized	**RevChron**	**Pro**	**Random**	180.40	$< 10^{-10}$	55.97	$< 10^{-10}$
Personalized	**RevChron**	**Pro**	**Pro**	175.22	$< 10^{-10}$	1.22	0.22
Personalized	**RevChron**	**Pro**	**Aggregate**	178.11	$< 10^{-10}$	75.45	$< 10^{-10}$
Personalized	**RevChron**	**Random**	**Anti**	-113.35	$< 10^{-10}$	76.47	$< 10^{-10}$
Personalized	**RevChron**	**Random**	**Random**	-152.63	$< 10^{-10}$	118.46	$< 10^{-10}$
Personalized	**RevChron**	**Random**	**Pro**	-110.56	$< 10^{-10}$	26.69	$< 10^{-10}$
Personalized	**RevChron**	**Random**	**Aggregate**	-177.86	$< 10^{-10}$	148.91	$< 10^{-10}$
Personalized	**RevChron**	**Aggregate**	**Aggregate**	-135.51	$< 10^{-10}$	-176.58	$< 10^{-10}$
Actual	**RevChron**	**Actual**	**Anti**	219.85	$< 10^{-10}$	5.17	$< 10^{-10}$
Actual	**RevChron**	**Actual**	**Random**	335.25	$< 10^{-10}$	36.42	$< 10^{-10}$
Actual	**RevChron**	**Actual**	**Pro**	257.08	$< 10^{-10}$	-54.08	$< 10^{-10}$
Actual	**RevChron**	**Actual**	**Aggregate**	352.75	$< 10^{-10}$	68.09	$< 10^{-10}$
Actual	**Personalized**	**Actual**	**Anti**	592.88	$< 10^{-10}$	-116.03	$< 10^{-10}$
Actual	**Personalized**	**Actual**	**Pro**	-37.32	$< 10^{-10}$	-36.92	$< 10^{-10}$
Actual	**Personalized**	**Actual**	**Random**	466.58	$< 10^{-10}$	-94.45	$< 10^{-10}$
Actual	**Personalized**	**Actual**	**Aggregate**	413.57	$< 10^{-10}$	257.12	$< 10^{-10}$
RevChron	**RevChron**	**Pro**	**Anti**	-9.84	$< 10^{-10}$	47.64	$< 10^{-10}$
RevChron	**RevChron**	**Pro**	**Random**	-14.06	$< 10^{-10}$	79.58	$< 10^{-10}$
RevChron	**RevChron**	**Pro**	**Aggregate**	-28.48	$< 10^{-10}$	105.54	$< 10^{-10}$
RevChron	**RevChron**	**Anti**	**Random**	-1.57	0.12	22.82	$< 10^{-10}$
RevChron	**RevChron**	**Anti**	**Aggregate**	-14.20	$< 10^{-10}$	46.45	$< 10^{-10}$
Personalized	**Personalized**	**Pro**	**Anti**	405.17	$< 10^{-10}$	-37.14	$< 10^{-10}$
Personalized	**Personalized**	**Pro**	**Random**	260.23	$< 10^{-10}$	-17.06	$< 10^{-10}$
Personalized	**Personalized**	**Pro**	**Aggregate**	237.83	$< 10^{-10}$	181.45	$< 10^{-10}$
Personalized	**Personalized**	**Anti**	**Random**	-245.66	$< 10^{-10}$	31.62	$< 10^{-10}$
Personalized	**Personalized**	**Anti**	**Aggregate**	-269.78	$< 10^{-10}$	314.28	$< 10^{-10}$

Table 1: **Pairwise Significance Tests**. Data are treated under a independent t-test given the normalcy of the means measured. QQ plots for the aggregate conditions provided in appendix Fig. 18 and 19.

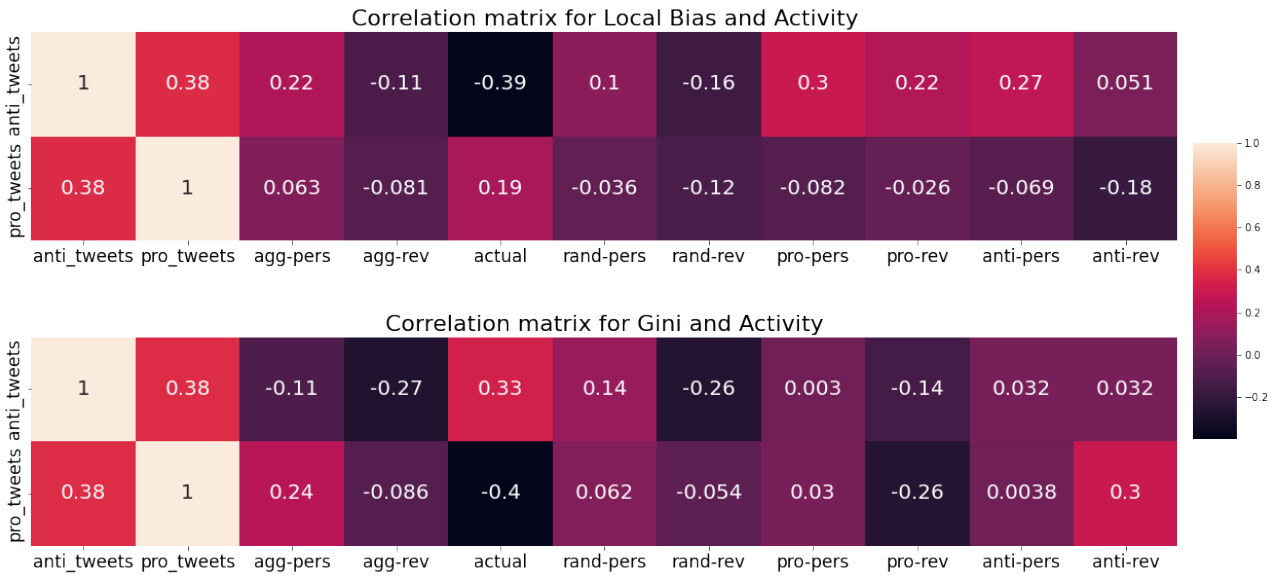

Figure 6: Correlation between time-series matched on days. Correlation is computed between the daily activity of pro- or anti-science friends (pro_tweets, anti_tweets). Agg refers to the aggregated accounts under the reverse chronological or personalized conditions. Actual refers to the measure derived from the total actual friend activity over the course of the study.

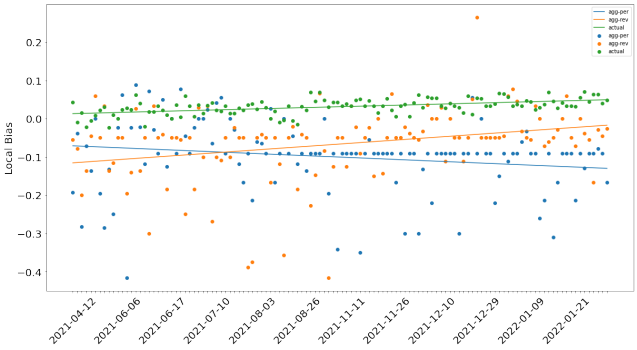

Figure 7: **Median imputed** local bias computed over time. Agg refers to the aggregated accounts under the reverse chronological or personalized conditions. Actual refers to the measure derived from the actual friend activity over the course of the study.

Implications. If we assume significant differences between the personalized and reverse chronological conditions, and more specifically differences between the biased conditions within those conditions, interacting with a fixed-length feed and observing and passively interacting with a biased environment could have downstream consequences on a user's mental and physical health. While differences in algorithms did not seem to result in different behavior in the 2020 election on FaceBook and Instagram (Guess et al. 2023), it remains plausible that they could change users' perceptions of their networks in other areas like gender norms and through other mechanisms like social comparison. Personalized timelines seem to be more sensitive to user behavior (which makes sense by design of the algorithm), but as such may inadvertently facilitate the spread of harmful narratives by making them seem more popular than they may actually be.

Future Work

There is plenty of relevant future work to do, and we envision several extensions:

- Regenerating the analysis with sock puppets following each other would allow for richer control of account activity, as we could have sock puppets generate their own tweets in their own niche network.
- To account for the variability in session lengths that users have on social media platforms, we could run sock puppets for longer sessions. This would ostensibly offer the personalization engine running the algorithmic feeds more signals and opportunities to differentiate itself from the reverse chronological timeline.
- Working with BlueSky and decentralized Mastodon instances would allow us to test different feeds with consenting users, allowing analysis of exposure bias on real user data.
- Data donations from real users of the sessions they experience would be valuable for assessing the variable conditions users experience on the same platform.

Limitations

Labels. The focus of the sock puppets on the activity of friends labeled around their attitudes on science may be limited as their attitudes could have changed over the course of the study. Similarly, attributing the friends' labels to any tweets/users they interact with may be inappropriate as the source tweets/user may be taking a stance opposite to the

friend being followed. The pro/anti label may also be capturing other information, like general levels of activity, network size, and community norms with different kinds of engagement (which would be weighted differently in the ranking algorithm). Having to replace the sock puppets also draws time-varying claims into suspicion as each new account is assumed to start from no personalization.

Missing data. Missing data is a clear limitation for the analysis presented in this study. However, we present a sensitivity analysis which suggests that at least for the trending local bias data that the missing data does not seem to seriously impact the direction of the trends, implying the robustness of our results. Similarly, our use of the bootstrapping approach for both metrics allow us to assess the variability of our measurements due to the missing data.

Limited data. Complementary to the missing data is the study's focus on 100 followed users. While we believe our claims can generalize to the broader Twitter user experience at the time, our study only focuses on English-speaking Twitter users and does not consider other languages or other types of content sent via tweets (e.g., videos, images, external links). Nor do the sock puppets utilize all the features of the platform, e.g. lists, which could complicate the applicability of the conclusions to users who make use of such features.

Relevance to current platforms. Because we base our analysis on data from before Twitter became X, we cannot claim that the conclusions around bias are current, however we can claim that the point-in-time analysis of bias is still accurate. We also can claim the methodology is still relevant for similar platforms (not limited to X) as their underlying infrastructures have not obviously changed significantly as of this writing. Regular analysis of the state of the recommender systems constructing feeds should be performed as both minor and major changes to the systems can impact outcomes.

Conclusion

In this study we show that untrained sock puppet accounts on Twitter (now known as X) over a period of eight months, all following the same accounts, can experience vastly different perceptions of their networks. Accounts that are biased ideologically in interacting with tweets show differences in the feeds that they are presented with. When controlling for algorithmic impacts and as many other factors as possible, the biased accounts observing a reverse chronological feed show less exposure bias than similarly biased accounts observing a personalized feed. We believe that this implies additional scrutiny is needed on the signals that the feed personalization engine uses, as it may be unduly skewing some users' perception more than we would expect considering their activity.

As we study the perception of pro- and anti-science users in one's follower network, we hypothesize that our results would extend to perceptions of other political/ideological and social signals: the use of video, media and other external links that serve political or ideological purposes may drive more engagement for the platform, and as such may be weighted higher in users' feeds than content driven simply by smaller "weight" interactions.

With the open source GitHub of the Twitter/X ML pipeline at https://github.com/twitter/the-algorithm-ml, there is an interesting opportunity to assess how external audits might be assisted by code transparency from the platforms. For the purposes of this study, the pipeline has been helpful for identifying places to investigate (namely, the weights of various interactions) but it is difficult to simulate or gather data for all the components, especially as API access has been paywalled significantly (not to mention the lack of API access to the order in which users observe their personalized timelines). Complex recommender systems mediate our exposure to our online networks, and it remains the case that external research architecture for understanding them is imperative to ensure society has a reasoned and clear knowledge of how these systems may effect us.

References

Alipourfard, N.; Nettasinghe, B.; Abeliuk, A.; Krishnamurthy, V.; and Lerman, K. 2020. Friendship paradox biases perceptions in directed networks. *Nature communications*, 11(1): 707.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130–1132.

Bandy, J.; and Diakopoulos, N. 2021. Curating quality? How Twitter's timeline algorithm treats different types of news. *Social Media+ Society*, 7(3): 20563051211041648.

Bartley, N.; Abeliuk, A.; Ferrara, E.; and Lerman, K. 2021. Auditing algorithmic bias on twitter. In *Proceedings of the 13th ACM Web Science Conference 2021*, 65–73.

Chen, W.; Pacheco, D.; Yang, K.-C.; and Menczer, F. 2021. Neutral bots probe political bias on social media. *Nature communications*, 12(1): 5580.

Dane, A.; and Bhatia, K. 2023. The social media diet: A scoping review to investigate the association between social media, body image and eating disorders amongst young people. *PLOS Global Public Health*, 3(3): e0001091.

Feld, S. L. 1991. Why your friends have more friends than you do. *American journal of sociology*, 96(6): 1464–1477.

Galesic, M.; Bruine de Bruin, W.; Dalege, J.; Feld, S. L.; Kreuter, F.; Olsson, H.; Prelec, D.; Stein, D. L.; and van Der Does, T. 2021. Human social sensing is an untapped resource for computational social science. *Nature*, 595(7866): 214–222.

Galesic, M.; Bruine de Bruin, W.; Dumas, M.; Kapteyn, A.; Darling, J.; and Meijer, E. 2018. Asking about social circles improves election predictions. *Nature Human Behaviour*, 2(3): 187–193.

Galesic, M.; Olsson, H.; and Rieskamp, J. 2012. Social sampling explains apparent biases in judgments of social environments. *Psychological Science*, 23(12): 1515–1523.

Guess, A. M.; Malhotra, N.; Pan, J.; Barberá, P.; Allcott, H.; Brown, T.; Crespo-Tenorio, A.; Dimmery, D.; Freelon, D.; Gentzkow, M.; et al. 2023. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, 381(6656): 404–408.

Huszár, F.; Ktena, S. I.; O'Brien, C.; Belli, L.; Schlaikjer, A.; and Hardt, M. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 119(1): e2025334119.

Jabbour, D.; Masri, J. E.; Nawfal, R.; Malaeb, D.; and Salameh, P. 2023. Social media medical misinformation: impact on mental health and vaccination decision among university students. *Irish Journal of Medical Science (1971-)*, 192(1): 291–301.

Koester, D.; and Marcus, R. 2024. How does social media influence gender norms among adolescent boys?

Kulshrestha, J.; Eslami, M.; Messias, J.; Zafar, M. B.; Ghosh, S.; Gummadi, K. P.; and Karahalios, K. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 417–432.

Lazovich, T.; Belli, L.; Gonzales, A.; Bower, A.; Tantipongpipat, U.; Lum, K.; Huszar, F.; and Chowdhury, R. 2022. Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *Patterns*, 3(8).

Ledwich, M.; and Zaitsev, A. 2019. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.

Lee, E.; Karimi, F.; Wagner, C.; Jo, H.-H.; Strohmaier, M.; and Galesic, M. 2019. Homophily and minority-group size explain perception biases in social networks. *Nature human behaviour*, 3(10): 1078–1087.

Lerman, K.; Yan, X.; and Wu, X.-Z. 2016. The" majority illusion" in social networks. *PloS one*, 11(2): e0147617.

Mascherini, M.; and Nivakoski, S. 2022. Social media use and vaccine hesitancy in the European Union. *Vaccine*, 40(14): 2215–2225.

McKinley, C. J.; and Limbu, Y. 2023. Promoter or barrier? Assessing how social media predicts Covid-19 vaccine acceptance and hesitancy: A systematic review of primary series and booster vaccine investigations. *Social Science & Medicine*, 116378.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1): 415–444.

Pedalino, F.; and Camerini, A.-L. 2022. Instagram use and body dissatisfaction: the mediating role of upward social comparison with peers and influencers among young females. *International journal of environmental research and public health*, 19(3): 1543.

Rao, A.; Morstatter, F.; Hu, M.; Chen, E.; Burghardt, K.; Ferrara, E.; and Lerman, K. 2021. Political partisanship and antiscience attitudes in online discussions about COVID-19: Twitter content analysis. *Journal of medical Internet research*, 23(6): e26692.

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A.; and Meira Jr, W. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 131–141.

Ribeiro, M. H.; Veselovsky, V.; and West, R. 2023. The Amplification Paradox in Recommender Systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1138–1142.

Robertson, R. E.; Jiang, S.; Joseph, K.; Friedland, L.; Lazer, D.; and Wilson, C. 2018. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–22.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014): 4349–4357.

Stein, J.-P.; Krause, E.; and Ohler, P. 2021. Every (Insta) Gram counts? Applying cultivation theory to explore the effects of Instagram on young users' body image. *Psychology of popular media*, 10(1): 87.

Vombatkere, K.; Mousavi, S.; Zannettou, S.; Roesner, F.; and Gummadi, K. P. 2024. TikTok and the Art of Personalization: Investigating Exploration and Exploitation on Social Media Feeds.

Yan, S.; Altenburger, K. M.; Wang, Y.-C.; and Cheng, J. 2022. What does perception bias on social networks tell us about friend count satisfaction? In *Proceedings of the ACM Web Conference 2022*, 2687–2695.

Paper Checklist to be included in your paper

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes we believe it will advance science without violating social contracts, as long as accounts are clearly identified as sock puppets, we follow an IRB, and the assumptions regarding labelling users are clearly and consistently defined.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, we believe we adequately control for as many factors as we can and that the claims adequately reflect the scope and contributions of the paper

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes in the methods section.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we primarily discuss this in the methods and limitations sections.

   (e) Did you describe the limitations of your work? Yes in the limitations and future work section.

(f) Did you discuss any potential negative societal impacts of your work? Yes in both the ethical considerations and limitations sections.

(g) Did you discuss any potential misuse of your work? Yes, in the limitations of the work section

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes we release the scripts for generating results and will make anonymized data available at least upon request.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Yes when we report results

(b) Have you provided justifications for all theoretical results? We provide independent t-tests

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes by considering aggregated data and disaggregated data.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes by addressing missing data

(e) Did you address potential biases or limitations in your theoretical framework? Yes in Limitations

(f) Have you related your theoretical results to the existing literature in social science? Yes both in related work and discussion.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? We have an implications section.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? N/A

(b) Did you include complete proofs of all theoretical results? N/A

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? N/A

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? N/A

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? No

(c) Did you include any new assets in the supplemental material or as a URL? Yes, anonymized data and scripts.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? We discuss this in the ethical statement

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, as we use number IDs and refer to minimial tweet level information.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? We plan to anonymize the data.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? Yes we will create a datasheet and host it at the github

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? N/A

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A

(d) Did you discuss how data is stored, shared, and deidentified? N/A

Ethical considerations & Reproducibility

The study was approved by an IRB. Accounts were made to only interact with tweets via low-frequency likes to minimize potential implications on the final popularity of said tweets. Accounts were explicitly designated in their profile biography as research accounts with an email address pointing to the institution running the study. We release code that was used to generate this analysis and data showing which anonymized users were seen at which time, at https://github.com/bartleyn/curly-octo-fortnight.

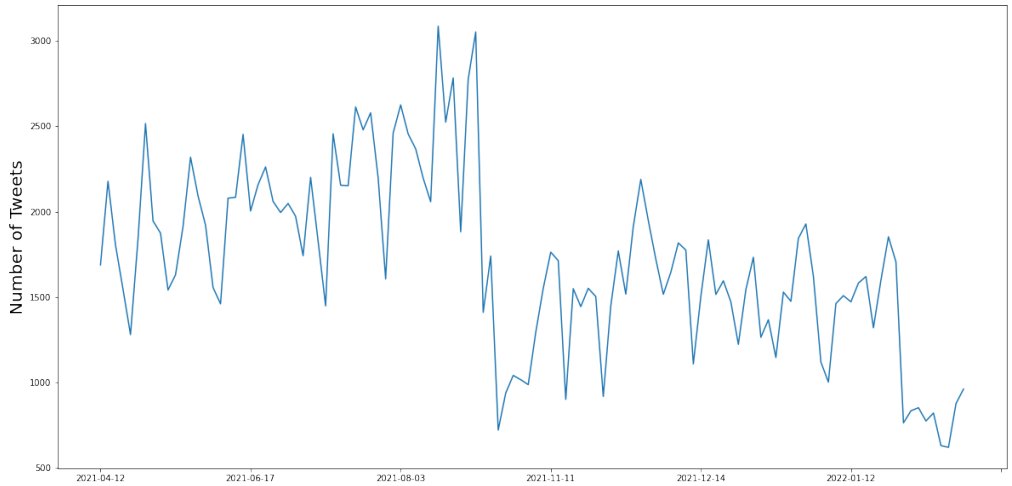

Figure 8: Number of tweets per day of anti-science users. The steep drop of tweets around 2021-10-24 can be explained both by accounts that were temporarily suspended and a brief rate limiting issue with the API.

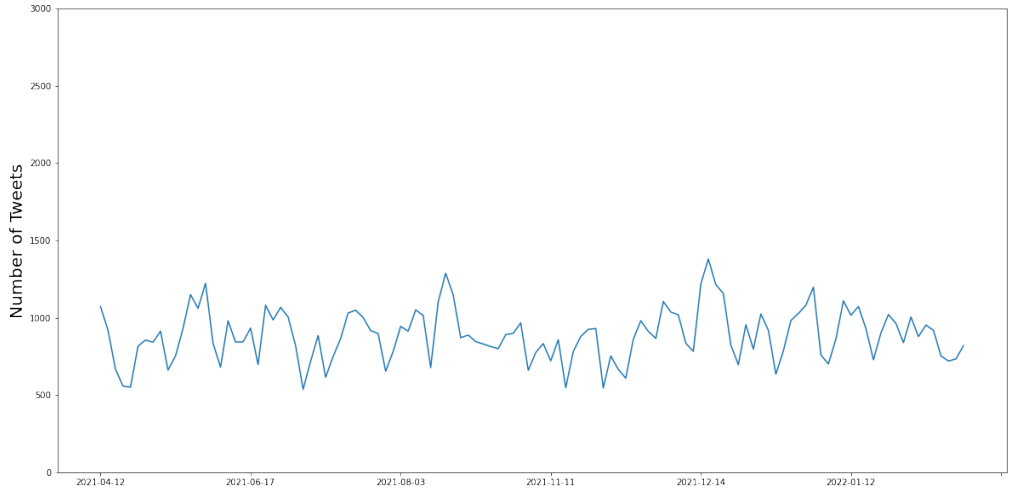

Figure 9: Number of tweets per day of pro-science users.

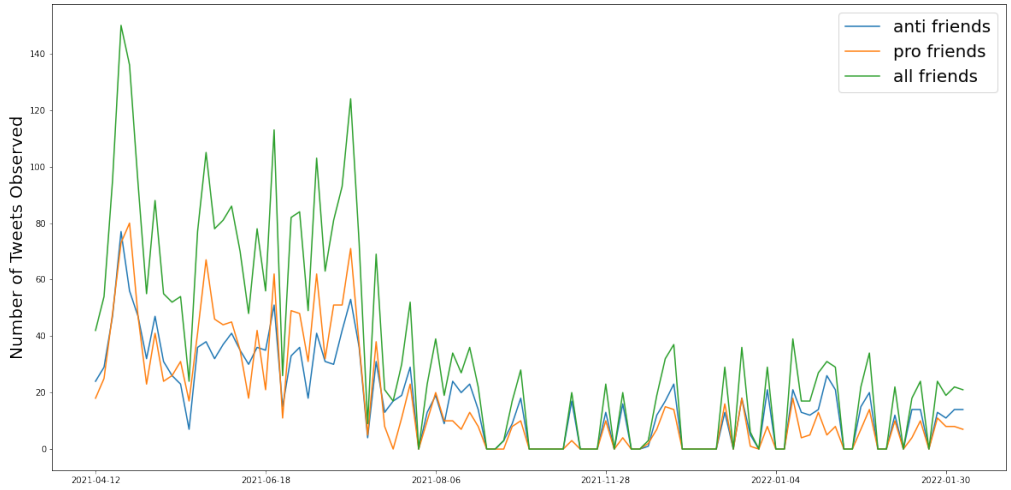

Figure 10: Number of usable personalized tweets seen per day. Other tweets were either promoted or otherwise had a problem in parsing.

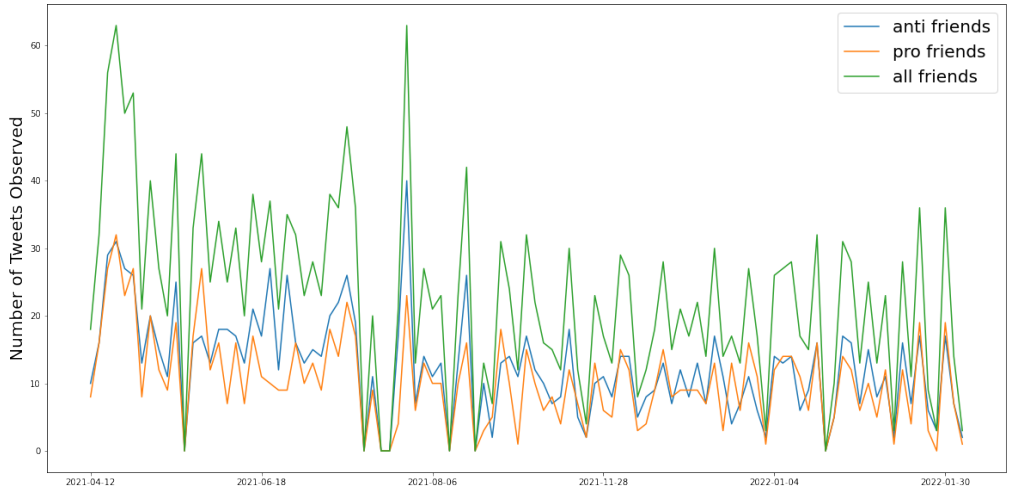

Figure 11: Number of usable reverse chronological tweets seen per day. Other tweets were either promoted or otherwise had a problem in parsing.

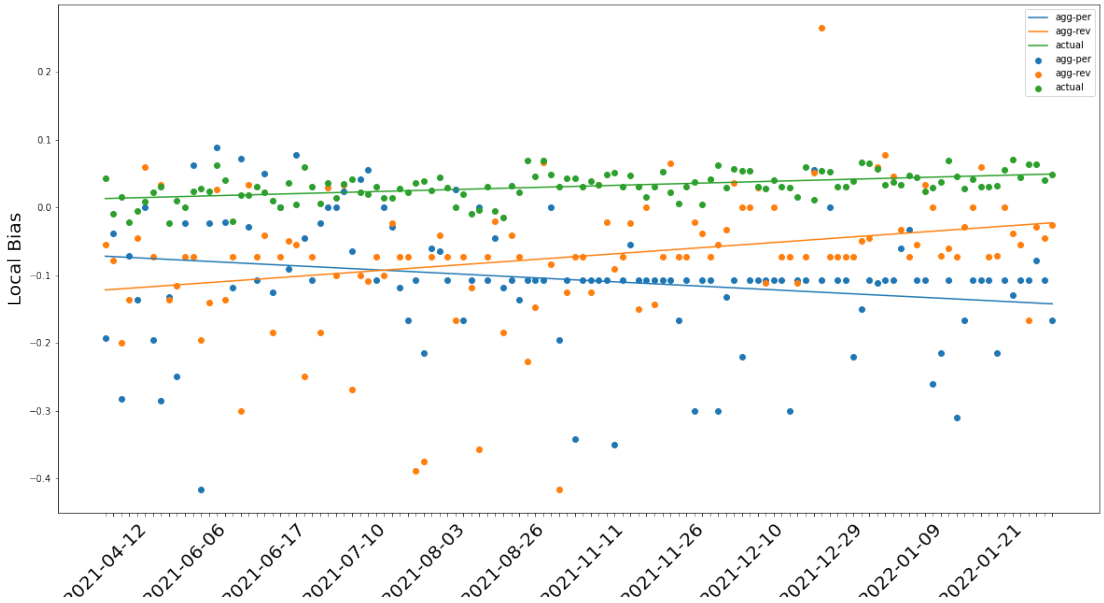

Figure 12: **Mean imputed** local bias computed over time.

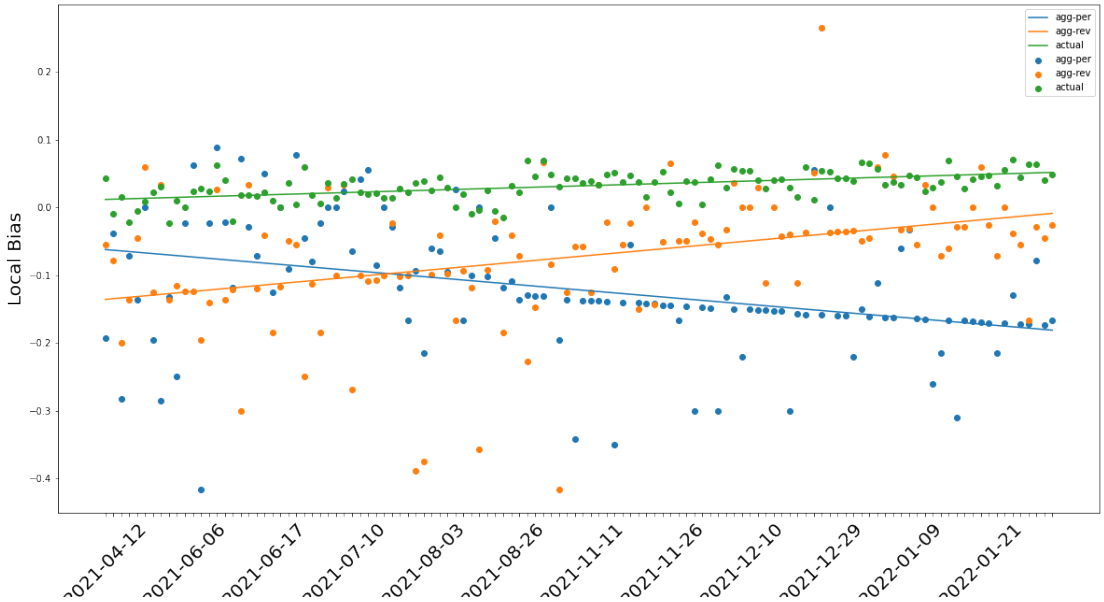

Figure 13: **Most-frequent imputed** local bias computed over time.

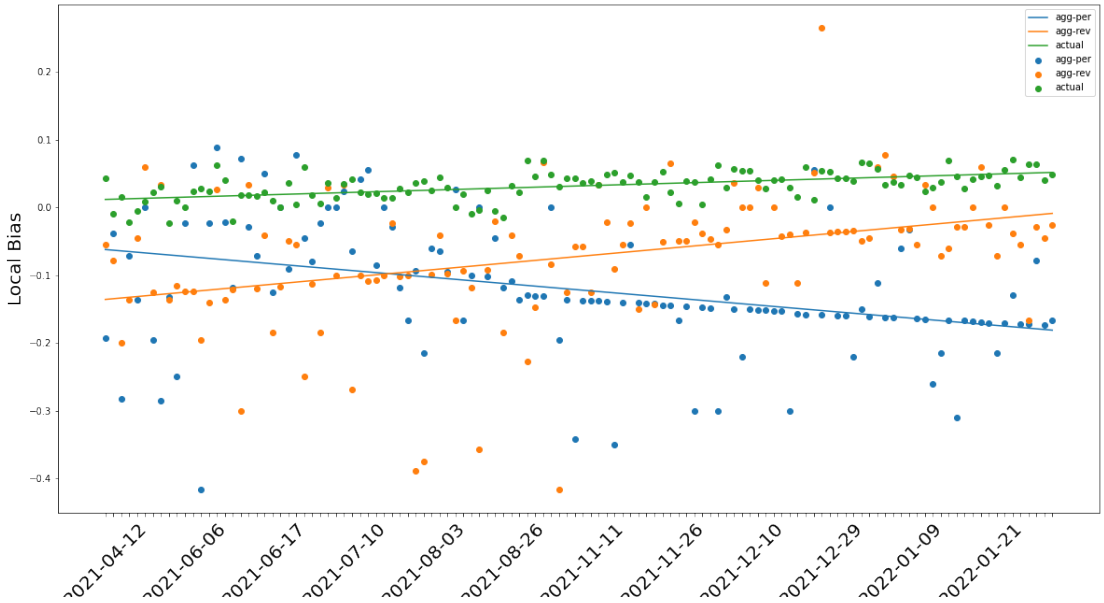

Figure 14: **Iterative imputed** local bias computed over time.

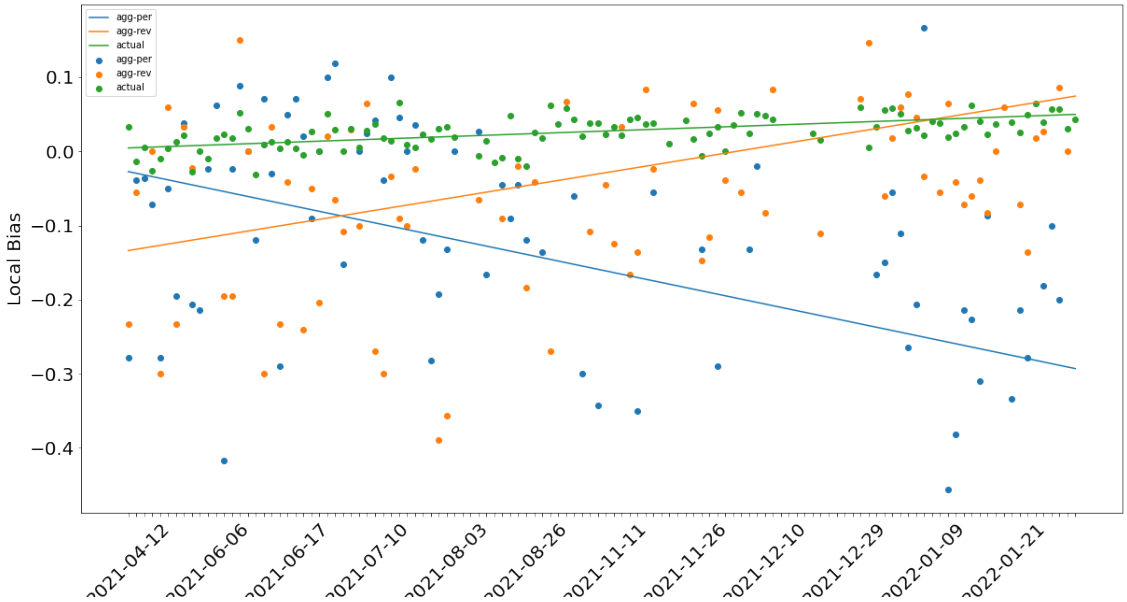

Figure 15: **Gaussian imputed** local bias computed over time.

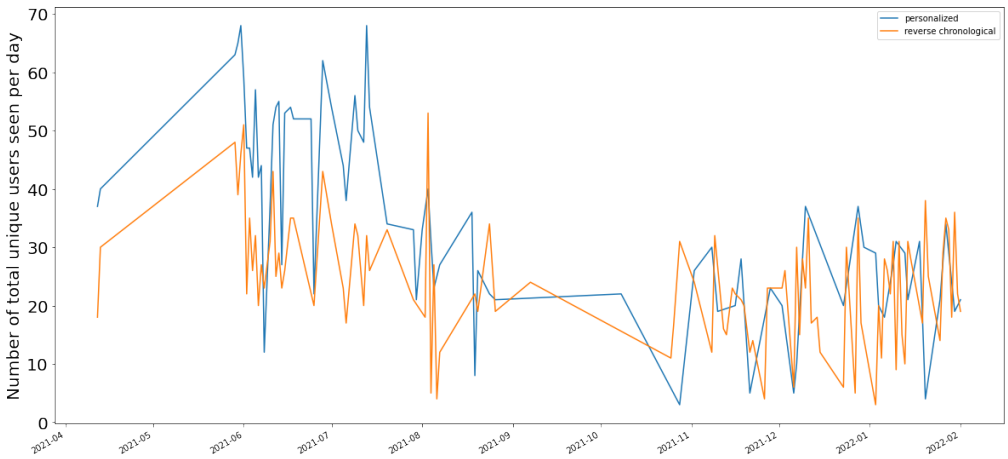

Figure 16: Number of unique friends seen overtime.

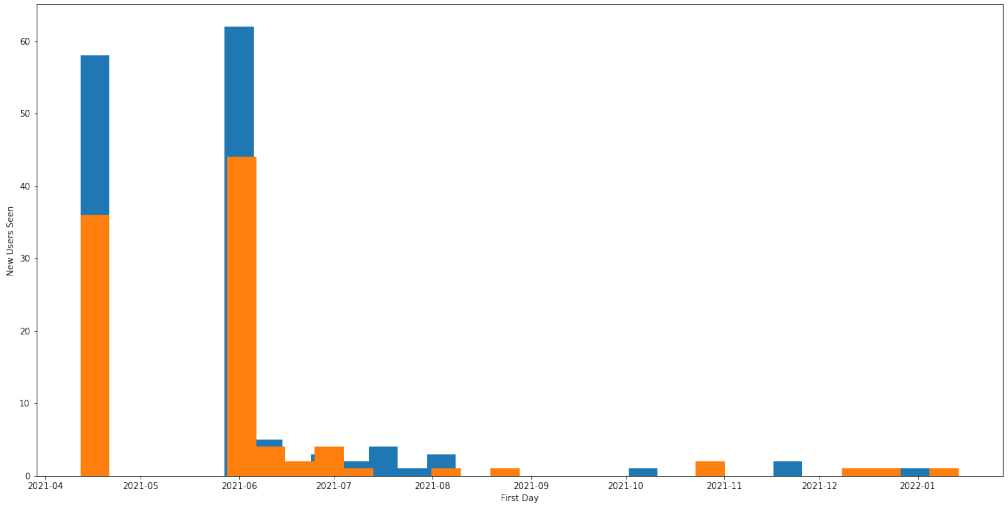

Figure 17: Number of new users exposed to per day.

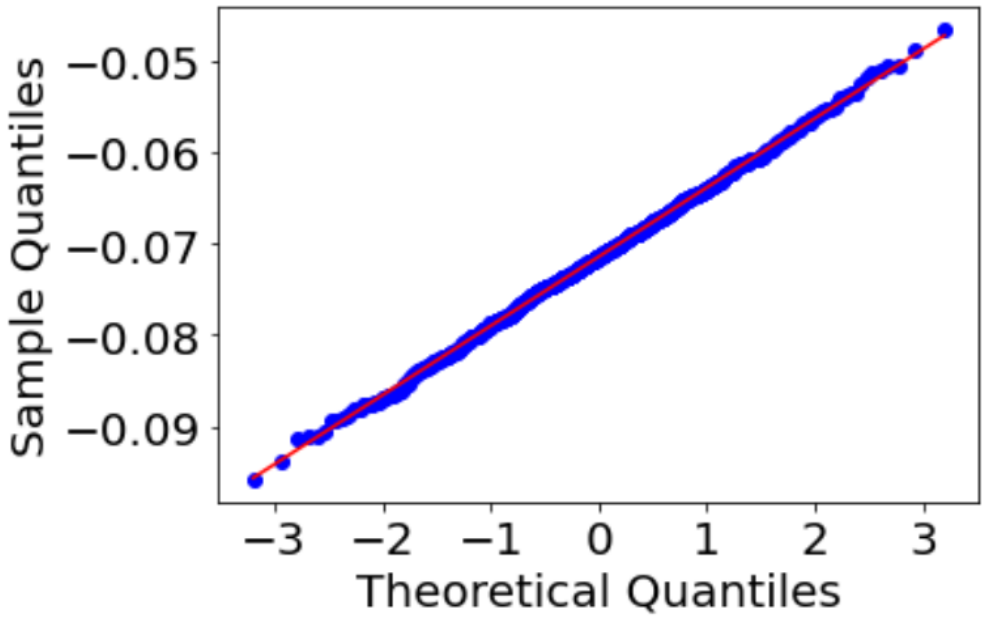

Figure 18: QQ Plot for Aggregate Personalized timeline.

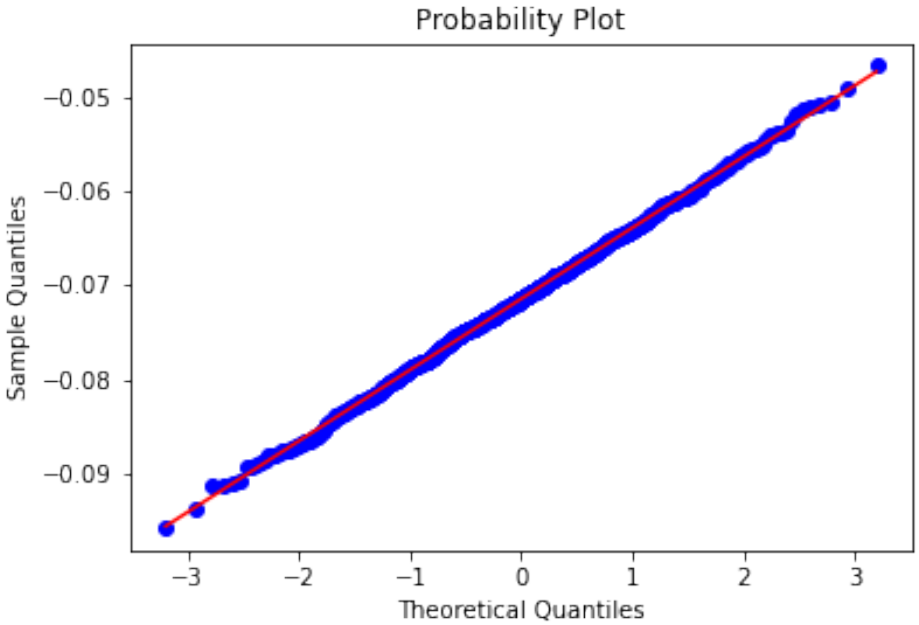

Figure 19: QQ Plot for Aggregate Reverse Chronological timeline.