# Impacts of Personalization on Social Network Exposure

1st Nathan Bartley
*Information Sciences Institute*
Marina Del Rey, United States
nbartley@isi.edu

2nd Keith Burghardt
*Information Sciences Institute*
Marina Del Rey, United States

3rd Kristina Lerman
*Information Sciences Institute*
Marina Del Rey, United States

***Abstract*—Algorithms personalize social media feeds by ranking posts from the inventory of a user's network. However, the combination of network structure and user activity can distort the distribution of content before the personalization step. To measure this "exposure bias" and how users might perceive the popularity of topics in their network when subjected to personalization, we conducted an analysis using archival X (formerly Twitter) data with a fixed inventory. We compare different ways recommender systems rank-order feeds: by recency, by popularity, based one the expected probability of engagement, and random sorting. Our results suggest that users who are subject to simpler algorithmic feeds experience significantly higher exposure bias compared to those with chronologically-sorted, popularity-sorted and deep-learning recommender models. Furthermore, we identify two key factors for bias mitigation: the effective degree-attribute correlation and session length. These factors can be adjusted to control the level of exposure bias experienced by users. To conclude we describe how this framework can extend to other platforms. Our findings highlight how the interactions between social networks and algorithmic curation shape—and distort—user's online experience.**

***Index Terms*—Recommender systems, Exposure bias, Social networks**

## I. Introduction

Online social networks (OSNs) serve as a critical conduit for the spread of news and information in everyday life as well as in times of emergency, including natural disasters (1). As OSNs have grown in popularity, the volume of user-generated content has exploded. To deal with information overload, OSNs use recommender systems to create personalized feeds for users; the systems sort content and bring useful content to the top of the user's feed. While personalized feeds have great utility, evidence has emerged in recent years that recommender systems can narrow the types of information users are exposed to as they interact with the system, creating echo chambers (2) and potentially increasing partisanship and affective polarization (3; 4). While there is conflicting evidence as to the extent that these recommender systems cause these problems versus how much user choices are responsible (5; 6), it is nonetheless acknowledged that a problem exists.

The problem of affective polarization, in particular, is theorized to have multiple interacting causes, one of which is the use of other peoples' identity signals by users as a cognitive shortcut to establish the credibility or popularity of the information being presented (7). As algorithmic curation in aggregate may result in an ideological asymmetric environment where algorithmic amplification may be over-representing one ideology over another (8), it is essential to understand how visibility and biased exposure can affect the perception of social reality online (which ultimately may create a self-fulfilling prophecy).

As a simple example, some accounts a user follows may be much more active than others, which can make their opinions over-represented in a user's feed. In addition, users themselves vary in how much time they spend on Twitter: some users may have short sessions where they read just a few of all posts, while others may have longer sessions where they read a much larger fraction of new posts. These factors, combined with how the recommender algorithm orders the posts within the user's feed, will affect the information the user sees.

We use the term *exposure bias* to refer to the phenomenon where the content a user sees on their feed is not an accurate representation of the inventory of all available content. While studies have explored various facets of algorithmic personalization, there remains a gap in how different ranking mechanisms—such as recency, popularity, and deep learning based mechanisms— can affect what users get exposed to, and through this their perception of their network. This paper aims to bridge this gap by introducing a framework for measuring exposure bias, incorporating elements of network structure and user-platform interaction dynamics.

In this work we address the following research questions:

RQ1. How do different feed-construction strategies affect exposure bias?

RQ2. Do different length sessions experience different levels of exposure bias?

We explore these questions on an X follower network that contains all tweets posted by the users in the sample. We emulate different feed ranking algorithms and then proceed to measure what each user is exposed to. Moreover, we use three measures of bias that measure a different component of that user's network exposure. We also vary the number of tweets the user sees, i.e., the length of the feed. We find that users under simpler model-based feeds exhibit more exposure bias than deep-learning based, chronological, and popularity-ranked feeds. We also describe the relationship that measures of exposure bias have to the degree-attribute correlation of a network, i.e., how degree of a node correlates

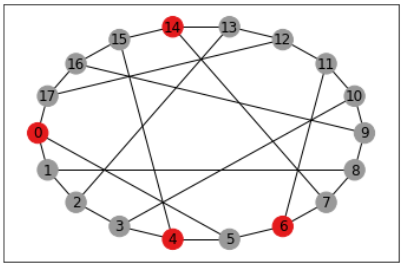



Fig. 1: **Majority Illusion**. Plotted is a graph with 18 nodes and 27 edges, with 4 nodes (0, 4, 14, 6) having an "active" binary trait. Nodes 15, 7, and 5 would experience the majority illusion in that the majority of their connections have the active trait, whereas the minority of the global population has the trait.

with the probability of having an attribute. We finally show that exposure bias can be both affected by the time users spend online as well as the algorithm, which implies the benefits of some algorithms over others at reducing echo chambers should they account for what tweets users observe[1].

## II. Related Work

This study can be situated at the interface between social psychology, network science, and the study/auditing of recommender systems.

### A. Cognitive Biases & Information Diffusion

In the cognitive sciences there have been many studies supporting the idea that humans are prone to multiple cognitive biases. Of particular interest is the salience bias: that humans are prone to paying more attention to stimuli that are remarkable or prominent, i.e., those stimuli that are irregular or unexpected (9). In social psychology this salience bias can manifest in identifying minority groups as standing out, often resulting in an overestimation of their size. This is similar to the *majority illusion*, a statistical bias in which network structure distorts a user's local view of the network (9). For illustrative purposes, we present an example network in Fig. 1 where some nodes experience the illusion due to how the minority trait is distributed across the network.

There is a similar assessment of cognitive biases from the social sensing literature: while people have the neurological capacity to observe and infer properties of their social network and mental states of others, their judgments may be differentially accurate depending on the populations being asked about. Galesic, Olsson and Rieskamp, 2012 detail a social sampling model where individuals sample from their immediate social network to estimate characteristics like average household wealth (10). Individuals surveyed tended to be accurate within their immediate network, but were less accurate for some measurements when judging larger populations. This speaks to the importance of understanding how OSNs mediate our exposure to our networks largely through recommender systems: distorted exposures to one's network can plausibly affect inferences about broader populations.

In the study of social networks these perception and cognitive biases are well understood to be integral to how information diffuses through a network. Kooti, Hodas, and Lerman, 2014, investigated the origins of network paradoxes and found that they have behavioral origins, suggesting that individuals have distorted perceptions of their networks (11). Rodriguez, Gummadi, and Schoelkopf, 2014, studied the effects of cognitive overload on information diffusion and found that exposure to information is less likely to infect a user if they are processing information from an ordered "queue" at higher rates. Our paper focuses on comparing different feed "queue" strategies and is not explicitly concerned with information diffusion/overload (12). Our framework for comparing personalized feeds can be used for assessing impact on information diffusion.

The most relevant work to our study is that of Alipourfard et al., 2020, where they both introduce a primary measure we use in our work and study the local perception bias of various hashtags in Twitter data from 2014 (13). They identify various hashtags that were overrepresented to users relative to the hashtags' global prevalence. A key difference in our work is that we study how constructing feeds in different ways can affect the perception of user traits. We also are more concerned with understanding how sensitive different feeds are to the distribution of the trait itself.

### B. Algorithmic Audits of OSN Recommender Systems

Beginning with Sandvig et al., 2014, there has been a steady and growing interest in auditing algorithms in online systems for discriminatory behavior (14). Algorithmic audits have focused on a wide variety of sectors, ranging from e-commerce platforms (15) to search queries (16; 17) and OSNs (18; 19).

Both Beattie, Taber, and Cramer, 2022 and Ramaciotti Morales and Cointet, 2021 analyzed the friend recommendation engine on Twitter, where the former presents a method for breaking echo chambers via user embeddings (20) and the latter is focused on incorporating ideological positions into similar user embeddings (21). Our work is not concerned with friend recommendations, however they are a very valuable tool for intervention.

From within Twitter, Huszár et al., 2022 looked at the algorithmic amplification of political parties across different countries on Twitter with proprietary data on users (8). They identify right-wing ideological bias under the algorithmic condition, suggesting that users in aggregate may be unduly influenced in how they perceive politics (at least on the platform).

Guess et al., 2023 study Facebook and Instagram feed data to assess the impact of personalization on user attitudes and political behaviors around the 2020 US election. They found no significant impact on behaviors but a significant difference in on-platform exposure to untrustworthy and uncivil content

[1]In an effort for reproducibility we provide a public repository with the simulation and analysis scripts: https://anonymous.4open.science/r/laughing-train-EF82/.

on the platforms (22). Gonzalez-Bailon et al., 2023 study the impact of algorithmic and social amplification in spreading ideological URLs on Facebook, identifying ideological segregation taking place and both the algorithmic/exposure stage as well as the social amplification (23). While these studies do not identify changes in beliefs or behaviors pertaining to the US 2020 presidential election, it is important to note that the effects may not be generalizable across different platforms and are focused on political-related outcomes from 2020 that may not apply in other domains given the relationship to Covid-19.

Donkers and Ziegler, 2021 studied both "epistemic" and "ideological" echo chambers in OSNs and how diversifying recommendations can potentially depolarize discussions (24). They constructed a recommender system based on knowledge graph embeddings, allowed for users with varying propensities for accepting new information. However the utility of their framework, utilizing knowledge graph embeddings, might be limited as they do not consider different methods of ranking the tweets that are proposed to users in their evaluations.

### C. User Perception of Ranked Feeds

Understanding user perceptions of their ranked feeds has been a focus of human-computer interaction work for the past decade. Notably this has been focused primarily on Facebook, where prominent work done by Eslami et al., 2016 identified several "folk theories" for how Facebook's personalized News Feed curated the information users in the study saw (25; 26). FeedVis, the tool they developed, empowers users by presenting which friends they will never see, rarely see, often see alongside other information about their feeds (25). Our study differs from this vein of research as we focus on comparing different recommender systems directly.

## III. Data & Methods

### A. Twitter Data

Starting in March 2014, Alipourfard et al., 2020 (13) queried Twitter to identify accounts followed by each of $5,599$ seed users. These followed accounts are known as *follower graph friends* or *friends* for short. The authors continued to query Twitter for the followed accounts daily through September 2014 to identify any new friends of seed users. This subset of the Twitter follower graph has over 4M users and more than 17M edges.

In addition to follow relations, the authors also collected messages posted by seed users and their friends over this time period, roughly 81.2M tweets. For this study we consider tweets from May 2014 to September 2014. At the time of data collection, Twitter created a feed for each user by aggregating all messages posted by the user's (follower graph) friends and ranking them in reverse chronological order. Given the uniform feed treatment, we are therefore able to reconstruct the feed for each seed user and quantify empirical exposure bias.

To summarize we use the tweets and retweets from $5,599$ seedusers and all of the people they follow at the time of collection in 2014.

### B. Reconstructing Feeds

To address our research questions we make use of the described empirical data. This data is important for this situation because it was collected before Twitter implemented an algorithmic recommender system for constructing feeds in 2016. Given that we know the users all experienced the same chronological feed, we can re-rank and construct new feeds to explore the exposure effects of the different feeds.

With the Twitter data we construct artificial "sessions": for each user $u$ we assume that they browse their feed one time on any calendar day $d$ they have any activity (either tweets or retweets). On each day $d$ we then select all tweets and retweets that friends of $u$ made that day and sort the tweets according to the specific feed construction algorithm. For analysis we only consider sessions with at least ten tweets and at least one unique friend seen.

We construct feeds according to the following strategies:

1) **Popularity.** We rank each of the tweets by their historical total number of retweets.
2) **Reverse Chronological.** We rank each of the tweets by the time they were posted (we assume the user logs in at the end of the day and observes tweets closer to the end of the day first).
3) **Random.** We rank each of the tweets randomly.
4) **Logistic Regression.** Given that the X/Twitter timeline personalization system utilizes a logistic regression model at the candidate-ranking step (27), we utilize a simplified logistic regression model with user-based features to rank tweets by the probability that user will retweet each tweet. For each user, we gather all possible tweets the user could have seen and all their retweets and then we train an individual model on that data. For users without retweets we use a similar user's model for prediction.
5) **Neural Collaborative Filtering (NCF).** We implement a dense neural network meant for personalized recommendations of tweets (28). The model has two parallel embedding layers for user and tweet (item) inputs, which are then concatenated and passed into a dense layer with ReLU activation. The output layer predicts the likelihood of a user retweeting a tweet. The model uses a binary focal cross-entropy loss function.
6) **Wide&Deep.** Similar to the NCF model, we implement the Wide&Deep model as described in Cheng et al., 2016 (29). We chose this model as researchers at Twitter described using a modified Wide&Deep model for ad recommendation in 2020 (30).

### C. Simulating User Attributes

To measure the exposure bias we assign each user in the network a binary random variable $X \in \{0, 1\}$ with a fixed uniform probability $\mathrm{P}(X = 1) = 0.10$. This variable can represent the user's ideology (e.g., liberal vs conservative), status (e.g., verified vs not), or it can represent a one-hot encoding of a belief the user shares. We choose a prevalence of $0.10$ because we want to represent a minority trait of the

| Notation | Description |
|---|---|
| $B_{\text{local}}$ | Local Bias |
| $A(v)$ | Attention function |
| G | Gini coefficient |
| $\rho_{kx}$ | Degree-attribute correlation |
| $\text{M}_i$ | No. of users with majority illusion on day i |

TABLE I: **Notation**.

population and observe how activity and exposure could distort it. We also run the same analyses under $\text{P}(X = 1) = 0.05$ and $\text{P}(X = 1) = 0.50$ to verify that the results are consistent.

After all accounts in the follower graph have been assigned values of the variable, we can then measure its prevalence in each user's feed. This allows the user to estimate the fraction of "activated" or "positive" (i.e., with value $x_i = 1$ ) friends within their network. To assess the relationship between the network structure and this random variable $X$ we follow the attribute-swapping procedure as described in Lerman et al., 2016 (31) to vary the **degree-attribute correlation** $\rho_{kx}$:

$$\rho_{kx} = \frac{P(x=1)}{\sigma_x \sigma_k}[\langle k \rangle_{x=1} - \langle k \rangle]$$

where x is the binary attribute, k is the degree of the node (here in-degree), $\sigma_x, \sigma_k$ the standard deviations of the binary attribute and in-degree respectively, and $\langle k \rangle$ the average in-degree over all nodes. As $\rho_{kx}$ increases we expect the exposure bias to increase.

### D. Measures of Exposure Bias

We use the following metrics to measure exposure biases:

1) Local Perception Bias.

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

2) No. of users experiencing majority illusion per day.

$$\text{M}_i = |\text{fraction of positive friends seen per day} > 0.50|$$

3) Gini Coefficient $G$.

$$G = \frac{\sum_{i=1}^{n}(2i - n - 1)x_i}{n \sum_{i=1}^{n} x_i}$$

For the Gini coefficient $G$ we assume that the tweets are distributed across all of each user's friends, i.e., that if a seed user has 50 friends and only one has tweets that are observed then $G$ is computed over all 50 friends instead of just the one observed friend that day.

We define $B_{\text{local}}$ as the average prevalence of the attribute among a node's friends:

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

$\mathbb{E}[f(X)]$ is the global prevalence of the node attribute f; $\bar{d}$ represents the expected in-degree of a node; $f(U)$ the attribute value $f$ of node $U$. $A(V)$ represents the attention node $V$ pays to any particular friend.

Intuitively, this $B_{\text{local}}$ measure is the difference between the average fraction of activated friends exposed to random users and the true prevalence of the trait (represented by $\frac{|\text{users}_{x=1}|}{|\text{users}|}$). A positive number indicates the average user should expect to see a higher fraction of friends in their feeds with $x = 1$ than the actual global prevalence.

## IV. Results

### A. RQ1 - Difference between Feeds

We report the results of our analysis in Figure 2 and Table II. Of note is the different behavior we observe between the six conditions in Figure 2: $B_{\text{local}}$ for the reverse chronological feed is noticeably higher than the random and popularity feeds, comparable to the Wide&Deep feed, and seems to be lower than the logistic regression feed and NCF feeds. We note that for each of the conditions, when computing $B_{\text{local}}$ we use the same attention function, leaving remaining differences to feed strategy. We report the statistical analysis in Table II where the Popularity feed is lowest according to $B_{\text{local}}$, Random is lowest according to $G$, and NCF is the lowest according to $M_i$. In the logistic regression plot in Fig. 2 (d) we observe that for shorter feeds the bias goes above 1.0, which may be an artifact of how we compute the bias for shorter length sessions.

### B. RQ2 - Difference between Session Lengths

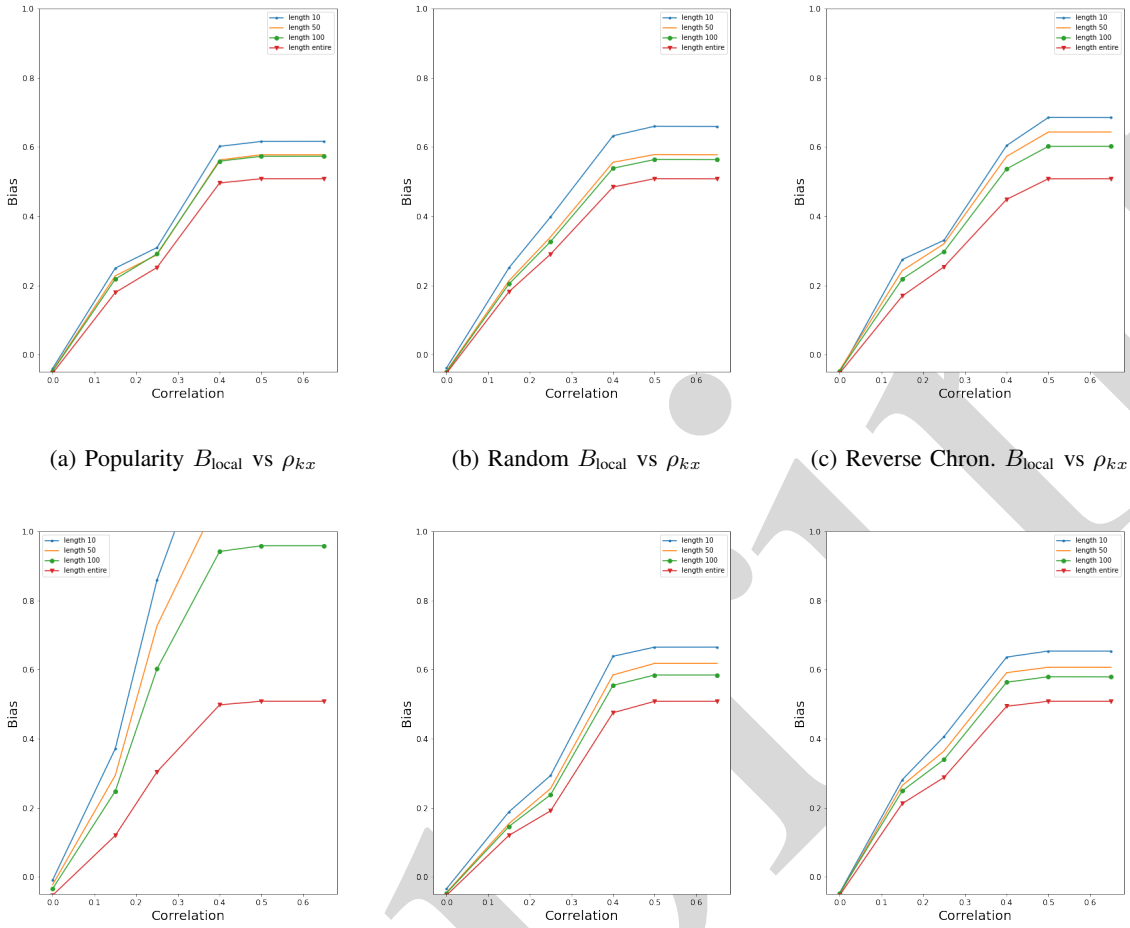
Can users compensate for exposure bias by consuming a larger share of their feeds? We model the attention users pay to their feed through the session length parameter. In the empirical data, we observe that the feed conditions behave similarly in terms of session length (Fig. 2). For $B_{\text{local}}$ the longer sessions show less bias than the shorter sessions (Table III). For $G$, we observe that the session lengths are ordered largely the same, with longer sessions being less biased than shorter ones. In contrast, the shortest feeds had lowest relative $M_i$ compared to the longer feeds.

## V. Discussion

### A. RQ1 - Difference between Feeds

Since the seed user follower graph and inventory is consistent across all conditions, we can treat the results as paired samples and take paired t-tests to establish difference between each pair of feeds. Table II shows that for $B_{\text{local}}$ the popularity feed is less biased than other feeds, but that the order changes with the other measures. We would expect the relationship to $\rho_{kx}$ to disappear in the random feed, making it the least biased across the metrics. This suggests we are not necessarily evaluating an appropriate random baseline. In simulated networks (not reported here), we shuffled the data generated for each seed user in order to eliminate relationship to $\rho_{kx}$. However doing this in the empirical data did not seem to have any significant impact.

Our findings suggest that seed users face greater exposure bias in reverse chronological feeds compared to simple retweet-based feeds, with logistic regression-based feeds showing the highest bias among all. The relatively higher Gini coefficients for chronological and logistic regression feeds suggest lower user diversity, possibly due to a temporal reliance on user activity. Interestingly the two deep learning




(a) Popularity $B_{\text{local}}$ vs $\rho_{kx}$ (b) Random $B_{\text{local}}$ vs $\rho_{kx}$ (c) Reverse Chron. $B_{\text{local}}$ vs $\rho_{kx}$

(d) Logistic Regression $B_{\text{local}}$ vs $\rho_{kx}$ (e) NCF $B_{\text{local}}$ vs $\rho_{kx}$ (f) Wide&Deep $B_{\text{local}}$ vs $\rho_{kx}$

Fig. 2: **Bias $B_{\text{local}}$ versus correlation** $\rho_{kx}$. Mean values are reported across all seed user days; bars are standard error of the mean.

models behave differently, suggesting the NCF is better than others at keeping the fraction of exposed friends that are positive low but higher than the true prevalence than other models.

### *B. RQ2 - Difference between Session Lengths*

When we consider the empirical session lengths, we see a significant difference between each length within each feed condition (Table III). Interestingly $B_{\text{local}}$ reports higher bias for shorter sessions than longer ones; this can be explained by the number of unique friends observed in longer sessions. Longer sessions create a larger universe of edges to compute the expected value over, yielding a value closer to the population estimate. This may also create an artifact in our computation as we observe $B_{\text{local}} > 1.0$ in Fig. 2 (d), which is interesting because we observe the same effect in $P(X = 1) \in \{0.05, 0.50\}$.

Given that both $B_{\text{local}}$ and $G$ agree on the ordering of the feeds, we interpret this as reporting that the same seed users experience significantly more exposure bias in shorter feeds than longer feeds on average. How we compute each metric (especially Gini coefficient) could be resulting in very small standard error when running the t-tests. For $M_i$ we observe a lower relative bias, which follows from shorter feeds being sensitive to the addition and subtraction of individual users.

Our methodology effectively highlights differences between feeds. A common limitation in prior audits is the inconsistent application of treatments. For example, even when some users are subject to chronological feeds, the influence of personalized feeds used by their friends may still impact their experience. Applying uniform treatments in an archival setting, where data is generated under a uniform context, provides confidence in measuring effects between conditions.

### *C. Considerations for Practitioners*

We can modulate the effective correlation $\rho_{kx}$ in a practical manner by choosing network edges to observe to change $\langle k \rangle_{x=1} - \langle k \rangle$. However, this has ethical implications in varying how often different people get observed in expectation. This

| Feed 1 | Feed 2 | $P(X=1)=0.10$ | | |
|---|---|---|---|---|
| | | $B_{\text{local}}$ | G | $M_i$ |
| **Pop.** | **Wide&Deep** | **-137.85**** | 20.99** | -0.81 |
| **Pop.** | **Rand.** | **-41.36**** | 8.17 ** | 1.07 |
| **Pop.** | **RevChron.** | **-162.60**** | -26.44 ** | 5.32 ** |
| **Pop.** | **NCF** | **-44.13**** | 8.21** | 44.47 ** |
| **Pop.** | **LogReg** | **-205.73**** | -44.86 ** | 12.48 ** |
| **Wide&Deep** | **Pop.** | 137.85** | -20.99** | 0.81 |
| **Wide&Deep** | **Rand.** | -7.16** | 7.47** | 1.58 |
| **Wide&Deep** | **RevChron.** | -204.21** | -34.99** | 6.08** |
| **Wide&Deep** | **NCF** | -26.01** | 2.49* | 44.82** |
| **Wide&Deep** | **LogReg** | -213.47** | -46.03** | 12.87** |
| **Rand.** | **Pop.** | 41.36** | **-8.17**** | -1.07 |
| **Rand.** | **Wide&Deep** | 7.16** | **-7.47**** | -1.58 |
| **Rand.** | **REvChron.** | -10.70** | **-8.95**** | 2.61** |
| **Rand.** | **NCF** | -25.40** | **-3.58**** | 43.27** |
| **Rand.** | **LogReg** | -117.02** | **-10.49**** | 10.49** |
| **RevChron.** | **Pop.** | 162.60** | 26.44** | -5.32** |
| **RevChron.** | **Wide&Deep** | 204.21** | 34.99** | -6.08** |
| **RevChron.** | **Rand.** | 10.70** | 8.95** | -2.61** |
| **RevChron.** | **NCF** | -14.37** | 12.66** | 43.70** |
| **RevChron.** | **LogReg** | -213.92** | -39.19** | 8.75** |
| **NCF** | **Pop.** | 44.13** | -8.21** | **-44.47**** |
| **NCF** | **Wide&Deep** | 26.01** | -2.49* | **-44.82**** |
| **NCF** | **Rand.** | 25.40** | 3.58** | **-43.27**** |
| **NCF** | **RevChron.** | 14.37** | -12.66** | **-43.70**** |
| **NCF** | **LogReg** | -104.33** | -20.46** | **-42.86**** |
| **LogReg** | **Pop.** | 205.73** | 44.86** | -12.48 ** |
| **LogReg** | **Wide&Deep** | 213.47** | 46.03 ** | -12.87** |
| **LogReg** | **Rand.** | 117.02** | 10.49** | -10.49** |
| **LogReg** | **RevChron.** | 213.92** | 39.19** | -8.75** |
| **LogReg** | **NCF** | 104.33** | 20.46** | 42.86** |

TABLE II: **Empirical Pairwise Significance Tests**. We treat the seed users in different feed conditions as paired samples and compute a paired t-test over the mean values of the various metrics. Conditions that are bolded are least biased according to the measure. * - p <0.05, ** - p <0.01

| Length 1 | Length 2 | $P(X=1)=0.10$ | | |
|---|---|---|---|---|
| | | $B_{\text{local}}$ | G | $M_i$ |
| 10 | 100 | 245.79** | 76.85** | **-30.25**** |
| 10 | Full length | 222.30 ** | 74.52** | **-28.39**** |
| 100 | 10 | -245.79** | -76.85** | 30.25** |
| 100 | Full length | 195.26** | 43.48** | -1.87 |
| Full length | 10 | **-222.30**** | **-74.52**** | 28.39** |
| Full length | 100 | **-195.26**** | **-43.48**** | 1.87 |

TABLE III: **Empirical Session-Length Pairwise Significance Tests**. We treat the seed users in different feed conditions as paired samples and compute a paired t-test over the mean values of the various metrics computed for each session length. * - p <0.05, ** - p <0.01

should be considered in tandem with measures of individual and group fairness to make feeds robust to biases.

Considering the number of unique friends that a user is exposed to may be a useful measure for platforms to maintain as a measure of overall fitness of the platform; it will have an impact on the measures we present in this study. More specifically, observing more unique friends seen lowers Gini coefficient and minimizes the absolute value of $B_{\text{local}}$.

### *D. Other Platforms*

Since 2014, platforms like X have evolved to show users content from beyond their immediate follower "in-network", introducing a larger "out-of-network" user set for feed algorithms to sample from. For example, a user with friends primarily exhibiting trait $x_i = 0$ could encounter unexpected content from non-followed users with trait $x_i = 1$.

This exposure bias framework is widely applicable to other social network recommender systems, primarily through the idea of a partially observed network. If we consider repeated exposure to specific users and kinds of content as a weak tie in a social network then we can treat exposure to a user's post as someone observing an edge to that user in a partially observed network. That observed network may have a different prevalence of certain traits than what you would expect from either the follower network or total universe of possible users one could observe. This framework can apply to other platforms as follows:

1) **TikTok.** TikTok's For You Page (FYP) explicitly considers user interactions, such as videos shared and accounts followed among other factors in personalizing users' FYP. TikTok describes explicitly diversifying feeds to prevent repetitive exposure to particular users and/or types of content (32). Exposure to different communities on TikTok (so called -Tok communities, e.g., BookTok) and certain users can be analyzed under this exposure bias framework because the exposed prevalence of certain traits within the communities can be compared to larger scale prevalence (e.g., how prevalent the traits are in the users' geographic area).
2) **Facebook.** Facebook's Feed works in a similar way as X: they gather the inventory of posts from friends, pages and groups, then calculate a relevance score and rank order each users' feed with high scoring out-of-network user posts interspersed (33). Given the parallels to how X works, this framework readily applies to Facebook.
3) **Instagram.** Two recommender system components are worth investigating on Instagram: the Explore page and the Instagram feed. The feed works much like Facebook and X as described previously. The Explore page is explicitly designed to show content from accounts that the user does not follow, drawing on information from followed accounts, information from posts that were interacted with and general connections on Instagram. This framework applies more readily to the feed than the Explore page.
4) **YouTube.** The YouTube front page and recommender system has been a central focus of auditing efforts for the kinds of information and news the system recommends for users to watch (3). Signals that are used to drive recommendations include clicks, watchtime, total views, user surveys for "valued watchtime" and other interactions including shares, likes and dislikes (34). If we consider how content creators are related to each other

online (e.g., partnerships, content networks), and how content can be grouped together (e.g., political content being grouped together ideologically), we can construct a user-creator network to assess exposure.

## VI. Limitations & Future work

A primary limitation for this study is the reliance on X data from 2014. The benefit of a vertical slice of data for the subgraph (with uniform feeds) is important, but it is also important to verify that the differences in recommender strategies hold in other networks as well, especially networks that are larger than the ones we analyzed here. We have simulation data of similar scale networks that suggest the results hold, but we do not report the results here.

Another limitation for this study is the assumption of binary user traits. Binary traits can be variable over time and subjective for different users. For example, person A may be perceived as being on the political left by a conservative observer B, however that same person A may be perceived as being conservative by a third liberal observer C. Allowing for dynamic non-binary traits would lead to valuable insights.

Future work could entail extending our feeds to be more complex. For instance, we could study the "out-of-network" feed, where we would prioritize friends and allow for more users to appear in the feed as a supplement. This would get the analysis to be closer to the empirical X self-studies where they describe in-network and out-of-network tweet impressions (8; 35). Similarly, we could attempt to implement all the working components and services of the reported Twitter system, however simulating data with this could end up too contrived to be useful.

In this study, we assume a fixed social graph; future research could explore dynamic social graphs where users change their follows over time. Additionally, applying exposure bias metrics to various network substructures could reveal differences in content consumption between network cliques and highly-connected hubs for instance.

## VII. Conclusion

In this study we have introduced exposure bias, related it to existing cognitive biases, and created metrics to quantify this bias. Our results illustrate the interconnectedness of social network structure, activity, and feed recommendation. We have shown that a mixture of bias metrics can adequately discriminate between different feed conditions. We described these metrics as being able to assess the propensity for individuals in a network to experience a distorted view of their immediate network, where a minority trait may be overrepresented and unduly more salient to the user than would be expected otherwise. Importantly we show that these feeds behave differently as the prevalence of the trait changes.

Huszár et. al., 2022 shows that within an organization it is feasible to observe and record the personalized feeds for large sets of individuals (8). Because of this we believe that both internal and external auditors should be able to use measures of exposure bias. With such measures, audits can be more closely tied to how individual users experience the system on a session-level. It would also allow for interpretable analysis to examine if different communities have disparate experiences with their personalized feeds.

## VIII. Ethical Statement

All data were anonymized prior to analysis. The analysis has minimal risk to user privacy, and analysis is unlikely to involve any ethical risk.

## References


[1] P. Panagiotopoulos, J. Barnett, A. Z. Bigdeli, and S. Sams, "Social media in emergency management: Twitter as a tool for communicating risks to the public," *Technological Forecasting and Social Change*, vol. 111, pp. 86–96, 2016.

[2] D. Nikolov, A. Flammini, and F. Menczer, "Right and left, partisanship predicts (asymmetric) vulnerability to misinformation," *Harvard Kennedy School (HKS) Misinformation Review*, 2021.

[3] M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, and W. Meira Jr, "Auditing radicalization pathways on youtube," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 131–141.

[4] W. Chen, D. Pacheco, K.-C. Yang, and F. Menczer, "Neutral bots reveal political bias on social media," *arXiv preprint arXiv:2005.08141*, 2020.

[5] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[6] M. H. Ribeiro, V. Veselovsky, and R. West, "The amplification paradox in recommender systems," *arXiv preprint arXiv:2302.11225*, 2023.

[7] S. González-Bailón and Y. Lelkes, "Do social media undermine social cohesion? a critical review," *Social Issues and Policy Review*, vol. 17, no. 1, pp. 155–180, 2023.

[8] F. Huszár, S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, and M. Hardt, "Algorithmic amplification of politics on twitter," *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, p. e2025334119, 2022.

[9] R. Kardosh, A. Y. Sklar, A. Goldstein, Y. Pertzov, and R. R. Hassin, "Minority salience and the overestimation of individuals from minority groups in perception and memory," *Proceedings of the National Academy of Sciences*, vol. 119, no. 12, p. e2116884119, 2022.

[10] M. Galesic, H. Olsson, and J. Rieskamp, "Social sampling explains apparent biases in judgments of social environments," *Psychological Science*, vol. 23, no. 12, pp. 1515–1523, 2012.

[11] F. Kooti, N. Hodas, and K. Lerman, "Network weirdness: Exploring the origins of network paradoxes," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2014, pp. 266–274.

[12] M. G. Rodriguez, K. Gummadi, and B. Schoelkopf, "Quantifying information overload in social media and its impact on social contagions," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 170–179.

[13] N. Alipourfard, B. Nettasinghe, A. Abeliuk, V. Krishnamurthy, and K. Lerman, "Friendship paradox biases perceptions in directed networks," *Nature communications*, vol. 11, no. 1, p. 707, 2020.

[14] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, "Auditing algorithms: Research methods for detecting discrimination on internet platforms," *Data and discrimination: converting critical concerns into productive inquiry*, vol. 22, no. 2014, pp. 4349–4357, 2014.

[15] P. Juneja and T. Mitra, "Auditing e-commerce platforms for algorithmically curated vaccine misinformation," in *Proceedings of the 2021 chi conference on human factors in computing systems*, 2021, pp. 1–27.

[16] P. Sapiezynski, W. Zeng, R. E Robertson, A. Mislove, and C. Wilson, "Quantifying the impact of user attentionon fair group representation in ranked lists," in *Companion proceedings of the 2019 world wide web conference*, 2019, pp. 553–562.

[17] M. Tomlein, B. Pecher, J. Simko, I. Srba, R. Moro, E. Stefancova, M. Kompan, A. Hrckova, J. Podrouzek, and M. Bielikova, "An audit of misinformation filter bubbles on youtube: Bubble bursting and recent behavior changes," in *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021, pp. 1–11.

[18] J. Bandy and N. Diakopoulos, "More accounts, fewer links: How algorithmic curation impacts media exposure in twitter timelines," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–28, 2021.

[19] N. Bartley, A. Abeliuk, E. Ferrara, and K. Lerman, "Auditing algorithmic bias on twitter," in *13th ACM Web Science Conference 2021*, 2021, pp. 65–73.

[20] L. Beattie, D. Taber, and H. Cramer, "Challenges in translating research to practice for evaluating fairness and bias in recommendation systems," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 528–530.

[21] P. Ramaciotti Morales and J.-P. Cointet, "Auditing the effect of social network recommendations on polarization in geometrical ideological spaces," in *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021, pp. 627–632.

[22] A. M. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow *et al.*, "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science*, vol. 381, no. 6656, pp. 398–404, 2023.

[23] S. González-Bailón, D. Lazer, P. Barberá, M. Zhang, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Freelon, M. Gentzkow, A. M. Guess *et al.*, "Asymmetric ideological segregation in exposure to political news on facebook," *Science*, vol. 381, no. 6656, pp. 392–398, 2023.

[24] T. Donkers and J. Ziegler, "The dual echo chamber: Modeling social media polarization for interventional recommending," in *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021, pp. 12–22.

[25] M. Eslami, A. Aleyasen, K. Karahalios, K. Hamilton, and C. Sandvig, "Feedvis: A path for exploring news feed curation algorithms," in *Proceedings of the 18th acm conference companion on computer supported cooperative work & social computing*, 2015, pp. 65–68.

[26] M. Eslami, K. Karahalios, C. Sandvig, K. Vaccaro, A. Rickman, K. Hamilton, and A. Kirlik, "First i" like" it, then i hide it: Folk theories of social feeds," in *Proceedings of the 2016 cHI conference on human factors in computing systems*, 2016, pp. 2371–2382.

[27] https://github.com/twitter/the-algorithm/blob/main/src/python/twitter/deepbird/projects/timelines/scripts/models/earlybird/README.md, accessed: 2024-01-01.

[28] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[29] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016, pp. 7–10.

[30] D. Guo, S. I. Ktena, P. K. Myana, F. Huszar, W. Shi, A. Tejani, M. Kneier, and S. Das, "Deep bayesian bandits: Exploring in online personalized recommendations," in *Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 456–461.

[31] K. Lerman, X. Yan, and X.-Z. Wu, "The" majority illusion" in social networks," *PloS one*, vol. 11, no. 2, p. e0147617, 2016.

[32] "How tiktok recomends videos for you," https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you, accessed: 2024-01-01.

[33] "Ranking and content," https://transparency.fb.com/features/ranking-and-content/, accessed: 2024-01-01.

[34] "On youtube's recommendation system," https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/, accessed: 2024-01-01.

[35] T. Lazovich, L. Belli, A. Gonzales, A. Bower, U. Tantipongpipat, K. Lum, F. Huszar, and R. Chowdhury, "Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics," *Patterns*, vol. 3, no. 8, p. 100568, 2022.