RTs ≠ Endorsements: Rethinking Exposure Fairness on Social Media Platforms

Nathan Bartley
Kristina Lerman
nbartley@usc.edu
Information Sciences Institute
Marina Del Rey, California, USA

Abstract

Recommender systems underpin many of the personalized services in the online information & social media ecosystem. However, the assumptions in the research on content recommendations in domains like search, video, and music are often applied wholesale to domains that require a better understanding of why and how users interact with the systems. In this position paper we focus on social media and argue that personalized timelines have an added layer of complexity that is derived from the social nature of the platform itself. In particular, definitions of exposure fairness should be expanded to consider the social environment each user is situated in: how often a user is exposed to others is as important as *who* they get exposed to.

ACM Reference Format:
Nathan Bartley and Kristina Lerman. 2024. RTs ≠ Endorsements: Rethinking Exposure Fairness on Social Media Platforms. In *Proceedings of 7th FAccTRec Workshop on Responsible Recommendation (FAccTRec '24).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Solving for fairness in recommender systems is a difficult problem with multiple types of stakeholders, e.g., consumers, producers, and advertisers [10]. As with other algorithmic systems, there has been acknowledgement of harms from recommender systems, both allocative and representative [4]. Many concerns around algorithmic harms stem from the fact that we measure performance of recommender systems by aggregating over the score of each user's individual recommendations rather than a top-down view of the behavior of the system. These measures of bias and harm are useful for informing better recommender system design, but a common theme in these analyses is a tacit abstraction of the domain that the system is situated in [10, 11].

Take for instance popularity bias: the observation that popularity bias in music recommendations can yield material differences in revenue across artists is useful for mitigating those harms in the music recommendation domain and only marginally useful for other domains of recommendations. Popularity bias in social media personalized timelines can also have material consequences for content creators in terms of revenue, however the social context of the recommendations means that a user's engagement with content is also dependent on other factors. These factors include one's own beliefs or ideological alignment to the content; how one perceives the norms about engaging with the content; one's perceived alignment with the user sharing the content; finally, any considerations the individual has for their "imagined audiences" and how they might be perceived for engaging with the content[18]. These factors may be present in other kinds of recommendation domains, however the extent to which they are present depends on the structure and affordances of the platform.

Another aspect of this is how to enforce constraints about allocative harms like exposure fairness in recommender systems. Extensive work has been done to address these concerns, however they tacitly assume the inventory of items is universally accessible to each user, and also do not consider systems that rely on user-generated content (i.e., where consumers are also producers)[1, 8, 9, 22]. Given the network structure underpinning social media platforms, who one user is connected to will have substantial influence on the content they are exposed to. In order to enforce constraints about exposure in social media recommender systems, it is important for such measures of fairness to be designed with users' perceptions, possible cognitive biases and their interactions with other users (not to mention the system itself) in mind.

Given the above caveats, in this position paper we build upon previous arguments that metrics and fairness strategies should be adapted to the domain they are situated in [11]. We discuss the social media personalized timeline context specifically, the social and cognitive factors impacting perception, and what exposure bias entails. We then present a toy example of exposure bias and suggest a means for mitigation.

2 Social sensing & cognitive biases

First we consider the users' perceptions and biases. Human beings are fundamentally social creatures, and it has been shown that we have the capacity to observe our social networks and make inferences about the mental states of others. This capacity for social sensing is susceptible however to cognitive biases: Galesic, Olsson and Rieskamp 2012 describe a model where individuals estimate large scale statistics like average household wealth by using information from their immediate social network [12]. In this they found individuals surveyed tended to be more accurate when making estimates about their immediate network, but were much less accurate when making estimates about broader populations.

We refer to the cognitive sciences to identify relevant socio-cognitive biases. A primary bias is the salience bias: humans tend to pay more attention to unexpected or irregular stimuli [14]. This bias can manifest in identifying minority groups as standing out, often resulting in an overestimation of their size. This is similar to the structural phenomena in networks in which network structure

FAccTRec '24, October 14, 2024, Bari, Italy
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

distorts a user's local perception of the whole network, e.g., the majority illusion and the friendship paradox[16].

The individual's own prior beliefs can influence their social perception as well through the false uniqueness (i.e., underestimating the prevalence of one's views) and false consensus effects (i.e., overestimating the prevalence of one's views). Empirical evidence suggests that Americans with more conservative views on climate change, as well as those with more conservative local norms and exposure to conservative news underestimate support for climate change policies by as much as half [21]. These biases have also been shown to be important to overcoming collective action problems in theoretical games [20].

3 Operationalization of Exposure Bias

As a minority/majority dichotomy in a social network can reflect a wide range of factors from demographic factors like gender, ethnicity and age to differences in opinion on arbitrary matters, we consider *exposure bias* to be a systematic distortion in content visibility and perceived prevalence within a social network. This distortion arises from discrepancies between the "potential network" (active social connections, e.g., who follows who), the "activated network" (the users who engage in content creation or activity and thus can be observed by other users), and the feed-exposed network (the network users actually observe and interact with). The feed-exposed networks, mediated by specific recommender systems, can modulate this bias leading to certain content being either over- or under-represented, which skews its perceived prevalence relative to its actual prevalence. In a sense, minimizing this bias can be seen as enforcing statistical parity in exposure across users in the "global" network[19].

There are multiple ways one may measure exposure bias– previous studies have used the Gini coefficient and local perception bias as introduced in Alipourfard et al., 2020 [3, 5]. These measures can be adjusted to take exposure into consideration, allowing us to compare exposure under different personalized timelines. As described in previous works, it is essential to consider multiple metrics as they complement each other in interpreting their results[17].

An important note in enforcing any measure of exposure fairness is that it may work well in aggregate, however depending on what "global" represents there may be subsets of users with vastly different experiences of their network. To better understand the heterogeneous experiences users may have of their network, we should consider how users interact with each other and with the platform itself.

4 Interpersonal dynamics & utility

Three interpersonal user dynamics have been described in these online "networked publics" that should be considered when building personalized timelines: 1) unless explicitly constrained to friends and followers, users have invisible audiences to their content that may scale in size beyond their control; 2) the different contexts in which content is observed are collapsed (e.g., the designated audience may perceive a joke to be appropriate but others may perceive the same content to be inappropriate); and 3) similar to the first point, content and conversations assuming more private environments may be thrust into more public environments with unintended consequences [6].

These considerations suggest that *who* shows up on your timeline and the social context in which their post is presented can shape users' experiences of the platform. As a practical example, Instagram researchers have suggested that they filter out some rare user IDs from users' feeds as they have found too many rare users can impact users' satisfaction / utility with the system [7]. The measurements of user utility and relevance, tied into an optimization that rewards engagement, may inadvertently optimize a user's feed for niche tastes and specific behaviors. For instance, Jiang et al., 2023 suggests that toxic behavior on platforms like X/Twitter may be reinforced by social approval and the propensity to act in a toxic behavior in their network [13]. If we are not careful about user dynamics when enforcing fairness we may end up optimizing negative engagement between users or facilitating second-order effects like fewer professional connections for marginalized users[2].

5 Toy example & Practical Considerations

Consider the network described in Fig. 1. In this example users are subjected to two different kinds of timelines with limited slots (i.e., each user only observes their timeline for a fixed amount of posts). One can imagine that in aggregate both feeds show each users' posts in a manner that maintains a particular measure of exposure fairness (e.g., the probability of observing a post from a user with the minority trait is proportional to their prevalence in expectation). The problem arises with the network: each user within that network will observe a different prevalence of the trait, and as such may have a different perception of whether or not the system is actually providing fairness in exposure. If the user highlighted in the figure is only exposed to blue users, then this may have consequences on how the user perceives the network as a whole.

To mitigate this exposure concern, practitioners in this may consider the homophilic structure of each user's network when selecting candidates for recommendation[15]: the degree-attribute correlation ρ_{kx} is well-understood to be connected to the strength of these phenomena like the majority illusion[16]. This correlation is defined as follows:

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

where x is the binary attribute, k is the degree of the node (here in-degree), σ_x, σ_k the standard deviations of the binary attribute and in-degree respectively, and $\langle k \rangle$ the average in-degree over all nodes.

Keeping the effective ρ_{kx} close to zero can be a feasible vector for minimizing the distorted perception across users. This may be readily inserted alongside more complicated recommendations based on user history and social dynamics described previously.

6 Conclusion

In this paper we argue that to make adequate and exposure-fair recommendations in social media the platforms need to consider the perceived social environment the system is putting the user into. We suggest an operationalization of exposure bias that considers

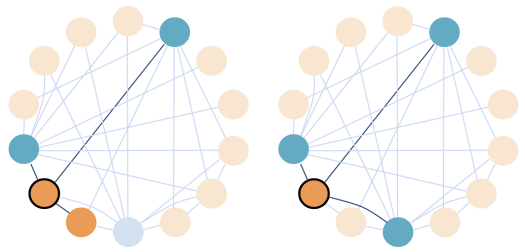

Figure 1: Simple exposure example. On the left the user is exposed to two users with a globally minority trait and one with the majority trait. On the right the same user, under a different feed is exposed to three users with the minority trait.

how a recommender system can modulate a user's perception of their social environment. We also suggest a potential method for minimizing such exposure bias.

Of course, to some degree the user has agency in what and who they choose to interact with, however the platform decides how users can interact with it through design choices and affordances of the platform suggesting they have more liability in how they shape users' experiences. In order to adequately understand exposure fairness on these platforms, we must connect it more to each users' view of their online surroundings.

References

[1] Himan Abdollahpouri and Masoud Mansoury. 2020. Multi-sided exposure bias in recommendation. *arXiv preprint arXiv:2006.15772* (2020).

[2] Nil-Jana Akpinar and Sina Fazelpour. 2024. Authenticity and exclusion: social media recommendation algorithms and the dynamics of belonging in professional networks. *arXiv preprint arXiv:2407.08552* (2024).

[3] Nazanin Alipourfard, Buddhika Nettasinghe, Andrés Abeliuk, Vikram Krishnamurthy, and Kristina Lerman. 2020. Friendship paradox biases perceptions in directed networks. *Nature communications* 11, 1 (2020), 707.

[4] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*. New York, NY, 1.

[5] Nathan Bartley, Keith Burghardt, and Kristina Lerman. 2023. Evaluating Content Exposure Bias in Social Networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. 379–383.

[6] Danah Boyd. 2009. Social media is here to stay... now what. *Microsoft Research Tech Fest* 5 (2009).

[7] Thomas Bredillet. 2023. Large Scale Recommendations at Instagram. VideoRecSys 2023 (2023). https://videorecsys.com/slides/thomas_talk1.pdf

[8] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on fairness, accountability and transparency*. PMLR, 202–214.

[9] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 275–284.

[10] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. 2022. Fairness in information access systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177.

[11] Michael D Ekstrand and Maria Soledad Pera. 2022. Matching Consumer Fairness Objectives & Strategies for RecSys. *arXiv preprint arXiv:2209.02662* (2022).

[12] Mirta Galesic, Henrik Olsson, and Jörg Rieskamp. 2012. Social sampling explains apparent biases in judgments of social environments. *Psychological Science* 23, 12 (2012), 1515–1523.

[13] Julie Jiang, Luca Luceri, Joseph B Walther, and Emilio Ferrara. 2023. Social approval and network homophily as motivators of online toxicity. *arXiv preprint arXiv:2310.07779* (2023).

[14] Rasha Kardosh, Asael Y Sklar, Alon Goldstein, Yoni Pertzov, and Ran R Hassin. 2022. Minority salience and the overestimation of individuals from minority groups in perception and memory. *Proceedings of the National Academy of Sciences* 119, 12 (2022), e2116884119.

[15] Eun Lee, Fariba Karimi, Claudia Wagner, Hang-Hyun Jo, Markus Strohmaier, and Mirta Galesic. 2019. Homophily and minority-group size explain perception biases in social networks. *Nature human behaviour* 3, 10 (2019), 1078–1087.

[16] Kristina Lerman, Xiaoran Yan, and Xin-Zeng Wu. 2016. The" majority illusion" in social networks. *PloS one* 11, 2 (2016), e0147617.

[17] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2021. A graph-based approach for mitigating multi-sided exposure bias in recommender systems. *ACM Transactions on Information Systems (TOIS)* 40, 2 (2021), 1–31.

[18] Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133.

[19] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[20] Fernando P Santos, Simon A Levin, and Vítor V Vasconcelos. 2021. Biased perceptions explain collective action deadlocks and suggest new mechanisms to prompt cooperation. *Iscience* 24, 4 (2021).

[21] Gregg Sparkman, Nathan Geiger, and Elke U Weber. 2022. Americans experience a false social reality by underestimating popular climate policy support by nearly half. *Nature communications* 13, 1 (2022), 4779.

[22] Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. 2022. Joint multisided exposure fairness for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval*. 703–714.