Fairness in Social Media Platforms: Modeling Behavior and Designing Interventions

Salima Jaoua
University of Zürich
Zürich, Switzerland
jaoua@ifi.uzh.ch

Machine learning is increasingly used for decision making that determines who is recommended, hired, or funded. These algorithms, trained on historical data, often amplify social biases and have a meaningful impact on the user's lives. In particular, social media platforms provide millions of professional content creators with sustainable incomes. Their income is largely influenced by their number of views and followers, which depends on the platform's recommender system (RS). So, as with regular jobs, it is important to ensure that RSs distribute revenue fairly. These recommender systems can perpetuate biases related to gender, race, and more generally to underrepresented groups (Li et al. 2023). They also often lead to convergence in users' tastes and favor popular items, a phenomenon referred to as the Matthew effect ("rich-get-richer"). My research addresses these challenges by focusing on theoretical design interventions for recommender systems that can enhance fairness while maintaining user satisfaction. By modeling and understanding user behavior, I develop mechanisms that can more accurately mitigate bias amplification. An important component of my approach involves the use of simulations to validate the theoretical models (Gilbert and Troitzsch 2005). Indeed, real-world datasets are limited to the outcomes of the specific recommendations users were shown, making it difficult to infer how users might behave with different content. Simulations are very useful not only for predicting user behavior, but also for studying emergent social phenomena and revealing important research questions. The combination of theoretical modeling and agent-based modeling allows me to explore the long-term impacts of recommendation strategies on the multiple stakeholders (Abdollahpouri et al. 2020).

Research Project

In my current research, I develop a pairwise comparison approach to recommender systems to improve the fairness towards content creators. An important issue for RSs, widely tackled by researchers, is the fairness of recommendations which can impact multiple stakeholders. In the context of social media, both viewers and creators represent the stakeholders in this multi-sided platform, where the popularity bias is problematic for both parties. The notion of individual fairness (IF) in RSs extends beyond social media applications to diverse domains including but not limited to e-commerce, streaming services and job matching platforms (Wang and Wang 2022). IF ensures that creators with similar content quality receive similar audience engagement (Dwork et al. 2012). Doing so helps remunerate creators according to the quality of their work, encourages them to continue posting good content, and reinforces their trust in the system.

However, both empirical (Salganik, Dodds, and Watts 2006) and theoretical (Ionescu, Hannák, and Pagan 2023) work show that individual fairness is difficult to achieve because early random differences in popularity increase over time through social influence and recommendation mechanisms. This effect makes popular creators account for most of the platform's engagement, thus reducing the overall diversity of content. This remains true even when increasing exploration in the RS or even when considering the simple case of one community of users which follow only by maximizing the quality of content they consume, and thus agree on their evaluation of creators. This decision-making process is known in Psychology as the behavior of maximizers (Schwartz et al. 2002). Maximizers aim to make the optimal choice and will search through many alternatives until they find what they believe is the best option. To address the issue of popularity bias, we keep this simplistic scenario in which previous work shows that the rich-get-richer effect occurs (Ionescu, Hannák, and Pagan 2023).

We started by showing that IF is achievable for all content creators; however, it becomes impractical in real-world scenarios. Building on this solution, we developed a feasible intervention: ordered pairwise comparison. More precisely, we consider all the possible ordered pairs of content creators: given n content creators, there are $n(n-1)$ ordered pairs. We divide the users into $n(n-1)$ groups of equal sizes, and recommend each user the corresponding pair of creators within the first two round. Through comparisons, we efficiently approximate the quality of content creator. Hence our approach strives to provide a more direct solution by assessing a ranking of creators by quality only requiring two exploration steps, to gain more information from users' feedback. We show theoretically that the system reaches a fair state after only two time steps. Moreover, there

is a high change that fairness is maintained in the next state. We also included simulations to evaluate the effectiveness of our intervention in the context of a new community cold start problem. Our RS design leads more often to fair outcomes for all content creators (i.e., in over 80% of the simulation runs). To make sure our intervention does not degrade the overall user experience we also consider user satisfaction, which shows that our approach does not harm user satisfaction (unlike exploration-enhancing interventions, such as transiting to random recommendations).

As with any modeling approach, results are currently limited to the model we study. Most importantly, future work should empirically check the extent to which users' decision-making depends on the order of recommendations. Further efforts are necessary to adapt this intervention to more complex datasets; this will require conducting experiments to check cross-session differences in viewer decision-making styles, evaluating this intervention on real-world datasets, and adapting it for other recommendation algorithms. Using an agent-based model that combines learnings from Psychology, Complex Networks, and Recommender Systems allowed us to understand the causal link between cross-session behavior and resulting unfairness, ultimately leading to the proposed intervention.

Next Steps

Following up on this work, there are multiple directions which I plan to explore for my thesis work.

First, one of the main challenges I encountered during this project was understanding user behavior, which lies at the core of addressing fairness in recommender systems. User interactions are deeply entangled with algorithmic outcomes, and amplified trough time because of feedback loops. This makes it difficult to distinguish whether certain patterns arise from the system itself or from the behavior of users interacting with it. This highlighted the importance of designing experiments that isolate and clarify these mechanisms. By doing so, we can better understand how user behavior contributes to unfairness and develop more effective interventions to mitigate it.

Another important aspect is understanding how content creators engage with platforms and perceive fairness. Prior work on music streaming services has shown how artists interpret fairness in algorithmic recommendations, emphasizing the need for more transparent and equitable recommender systems that reduce popularity bias (Dinnissen and Bauer 2023). Similar studies have examined drivers strategies in ride-hailing and other domains. We aim to conduct a similar study understand how content creators interact with the platform and adapt their strategies in response to algorithmic recommendations. By identifying and modeling the theoretical effects of these strategies on the platform and the user behavior, we would gain many information that can inform the design of fairer and more transparent recommender systems.

Finally, we aim to study the network evolution of Bluesky, as it is decentralized social media platform that allows users to choose or even customize their own recommender systems. This unique feature allows multiple RSs to coexist and interact, and thus enabling the formation of communities with shared values and the creation of safer communities. It is also very useful for researcher that can study how different recommendation algorithms influence network structure, content visibility, and user experience over time. By understanding these dynamics, we also hope to build and test our designs on the platform.

Acknowledgments

The author would like to thank Dr. Stefania Ionescu and Professor Anikó Hannák for their support and valuable discussions during this work.

References

Abdollahpouri, H.; Adomavicius, G.; Burke, R.; Guy, I.; Jannach, D.; Kamishima, T.; Krasnodebski, J.; and Pizzato, L. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30(1): 127–158.

Dinnissen, K.; and Bauer, C. 2023. Amplifying Artists' Voices: Item Provider Perspectives on Influence and Fairness of Music Streaming Platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, 238–249.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. 214–226.

Gilbert, N.; and Troitzsch, K. 2005. *Simulation for the Social Scientist*. McGraw-Hill Education (UK).

Ionescu, S.; Hannák, A.; and Pagan, N. 2023. The role of luck in the success of social media influencers. *Applied Network Science*, 8(1): 46.

Li, Y.; Chen, H.; Xu, S.; Ge, Y.; Tan, J.; Liu, S.; and Zhang, Y. 2023. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5): 1–48.

Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *science*, 311(5762): 854–856.

Schwartz, B.; Ward, A.; Monterosso, J.; Lyubomirsky, S.; White, K.; and Lehman, D. R. 2002. Maximizing versus satisficing: Happiness is a matter of choice. *Journal of personality and social psychology*, 83(5): 1178.

Wang, X.; and Wang, W. H. 2022. Providing Item-side Individual Fairness for Deep Recommender Systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 117–127.