Algorithmic Fairness for Networked Algorithms

Doctoral Consortium

Ana-Andreea Stoica
Columbia University
astoica@cs.columbia.edu

ABSTRACT

Recent evidence points to the detrimental effects of algorithmic deployment on human datasets, as often times such algorithms mirror and exacerbate existing inequalities in the input data. This work focuses on understanding the disparate effects of algorithms on social inequality and building theory and applications for graph algorithms with ramifications in the way we learn information online and offline. We show that in the case of recommendation algorithms, the most common heuristics that learn connections for providing social recommendations exacerbate disparity between different communities in a bi-populated network by reinforcing certain patterns in the network, such as homophilic behavior. Similar results occur for content recommendation, where we show that minority viewpoints are being further diminished by algorithms that learn relational data and over-recommend a majority viewpoint. On the other hand, algorithms may leverage community affiliation to disperse information in a network in a more effective manner while being more equitable in terms of the demographics reached in certain conditions. For such studies, we find closed-form conditions of the results using graph theoretical models that replicate inequality in social networks and use them to develop a set of algorithms that use network statistics to diffuse information in a feature-aware way, effectively reaching more communities than the status quo heuristics that are blind to sensitive features. Through validation on real-world data, we show that such learning algorithms benefit from being feature-aware in learning relational data in order to mitigate bias.

KEYWORDS

social networks; fairness; inequality; graph theory; influence; clustering; recommendation

ACM Reference Format:
Ana-Andreea Stoica. 2020. Algorithmic Fairness for Networked Algorithms. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION AND RELATED WORK

With the advent of Big Data, automated decision-making becomes ubiquitous in a variety of domains, from the online world to obtaining bail, credit assessment, and access to resources. When social connections become a proxy for income, race, gender, or other sensitive attributes, algorithms that learn biased data features may lead to disparate impact for minority groups regarding fair access to services and opportunities. While relational data becomes powerful in understanding the intricacies of human connection, it unfortunately also brings into the equation historical prejudices that easily get picked up by algorithms learning such data. Even in the cases where sensitive data, such as income or demographics, is obscured, correlations between these variables and others, such as geography, lead to similar unequal outcomes. While the role of technology is under debate in these circumstances, this study aims to unravel the power of computational tools in diagnosing instances of social inequality at large scale, as well as quantifying the effect of certain algorithms that learn from biased data on social inequality.

Recent studies show that such effects can be detrimental in a variety of spaces, from predicting recidivism rates in predictive policing [4], image classification [7], search engines [24], advertising [2, 34] and more recently in public health systems [25]. The main questions raised tackle the fairness and explainability of the implemented methods in order to facilitate compliance with legal obligations [6]. Thus, it is of uttermost importance to understand the disparate effects of algorithms on social inequality and to formulate a fair and explainable framework in designing such tools for prediction.

This work intertwines theoretical underpinnings for explaining such effects and building interventions for algorithmic bias, starting from defining what 'algorithmic fairness' means—what properties of social networks lead to differentiated outcomes and what equitable means in different contexts of a social network—and leading to re-designing learning algorithms to output more equitable results. Throughout this work two components are primary: feature-awareness and behavioral impact of algorithmic design.

Social inequality has been studied in many contexts, including access to opportunities [3], and more recently, in the context of algorithmic output in online settings [12, 14, 23]. Models for explaining the root cause of inequality have been developed to embed both human tendencies for connections [5] as well as human responses to algorithmic output [22]. Different types of solutions have been proposed to mitigate such effects, from addressing individual inequality [11], data representation techniques [13, 36], assessing implicit bias when evaluation different groups [9, 28], to understanding the causal relationships between data features [19, 21]. In many of theses settings, the efficiency of an algorithm seems to come into contrast with the equity of its output, showing a trade-off between fairness and accuracy [8, 13, 16, 20, 35]. We aim to understand the nature (and necessity) of this trade-off in relational data and to show the benefit of structure awareness for diversity-enhancing techniques. Here, we present three contexts for this problem, namely, in recommendation algorithms, influence maximization, and clustering.

This work is supported by the National Science Foundation under Grant No.1514437 and 1761810, as well as by a J.P. Morgan Ph.D. Fellowship.
Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

2 RESULTS

Recommendation systems: The first important concern for algorithmic discrimination is created by algorithms that restrict access to opportunities. The most commonly observed instance of restricted access is the glass ceiling effect, defined by "the unseen, yet unbreachable barrier that keeps minorities and women from rising to the upper rungs of the corporate ladder, regardless of their qualifications or achievements" [10]. First observed in the corporate world, this effect is pervasive in many online social networks whose constituents are subject to systemic inequality. My work proved that the most common algorithms that learn social connections for providing social recommendations exacerbate the glass ceiling effect in a bi-populated network by reinforcing certain patterns in the network, such as homophilic behavior [32]. We formalize such effects into a theoretical model and characterize in closed form through a fixed-point equation the conditions for which this happens. We build on a model of network growth with embedded homophily, different communities, and a preferential attachment dynamics to reproduce the cause of inequality in networks that undergo recommendation and information diffusion [5]. As this model is shown to exhibit inequality between its different groups, we are able to embed recommendations and asymptotically study their effect on group inequality. We validated our results on data collected from a large crawl of Instagram and DBLP, showing the effect of recommendation on people's degrees. We continued this into a study of content recommendation, showing that minority viewpoints are being further diminished by algorithms that learn relational data and over-recommend a majority viewpoint [30]. This work paves the way in the design of fair recommendations that are "aware" of the network structure and can rectify at best or mirror at worst biases that human datasets contain. This is also presented in our work at the Mechanism Design for Social Good (MD4SG) Workshop in 2018.

As a co-organizer of the MD4SG interdisciplinary initiative [1], I lead working groups on different topics related to Machine Learning and inequality, connecting with researchers from AI, Economics, Operations Research, Sociology, Policy, and Law. Through this initiative, I organize monthly colloquia of leaders in these fields, where I had the opportunity to learn about a large array of machine learning methods for prediction and game theoretical models for resource allocation, where algorithmic bias is pervasive through increasing inequality in networks that already exhibit bias in their features [6, 27, 33]. Inspired by this, the following question becomes central to my research: is it possible to design algorithms that actually use such inequality or lack of access as an opportunity for better growth? It is a common assumption that "diverse teams are more efficient", or equivalently "lack of diversity hurts your bottom line" [17, 26], but can an algorithm using social connections detect these diversity gaps and correct for them while being more efficient?

Influence maximization: Our recent work answers this question in the context of information diffusion, which entails the algorithmic selection of nodes chosen due to their advantageous position in a network [18]. These nodes are then used in strategic deployment of an idea, product, or technology, and can collectively trigger efficient diffusion of information through connections in the network, resulting in a massive improvement in awareness and innovation spread. My project focuses on designing fair influence maximization: a set of algorithms that learn the connections between individuals and their position in a social network to optimize the diffusion of a message while avoiding the creation of disparate impact among network participants based on community affiliation [29]. Through a theoretical model of network growth based on the biased preferential attachment model [5] and an influence model based on the independent cascade model, we reproduce inequality in influence maximization. Indeed, we show that classic heuristics that optimize influence based on people's centrality in a network end up reproducing a majority's community advantageous position. This has ramifications in the spread of information among under-represented groups. Our theoretical model allows us to show that such heuristics are actually not Pareto-efficient, and diversity-enhancing policies actually help with the spread of information. Our results find analytical conditions in which algorithms that are aware of the network structure can mitigate inequality within a population by selecting the most promising individuals in a more efficient way. We show that such a condition is common in real networks, and even in opposite cases, the cost of diversity is actually marginal. We develop a set of algorithms that use network statistics to diffuse information in a feature-aware way, effectively reaching more communities than the status quo heuristics that are blind to sensitive features. We show their effectiveness at diffusing a message in the DBLP dataset.

Clustering: Inspired by these, I currently work on algorithms that learn individuals' position and relations in order to cluster them based on their preferences and constraints, with a focus on feature-aware design. In a recent project, we leverage voting mechanisms that embed users' geography, constraints, and preferences, and use them to adapt classical clustering algorithms in order to split people in districts or communities that fairly represent their voices [31]. More specifically, our algorithms improve demographic diversity in segregated clusters by optimizing cluster utility while maintaining competition within clusters, which we show to reduce school segregation in a dataset of public schools from Detroit. Building on this, we work on understanding how people's preferences come into play with traditional clustering algorithms.

Finally, my current interests entail the study of learning algorithms with strategic agents, in cases where a classifier or a ranking algorithm induces a strategic behavior on the population involved. Building on work that shows the applications and social cost of strategic agents in classification on marginalized communities [15, 22], I work on re-designing such algorithms to encapsulate the differentiated cost that different communities bear when reacting to an algorithm that learns their features. This project aims to understand not only the way bias in data creeps into algorithmic design and deployment, but also the reactions that an algorithm elicits that may also contribute to systemic inequality.

These lines of work serve as an inspiration for fair algorithmic design, in which I strive to connect network theory, incentives, and notions of equity, with a focus on explainability of design choices. I hope to continue exploring these both through methodical analysis of current algorithms as well as through leading community efforts for interdisciplinary work.

REFERENCES

[1] Rediet Abebe and Kira Goldner. 2018. Mechanism design for social good. *AI Matters* 4, 3 (2018), 27–34.

[2] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *arXiv preprint arXiv:1904.02095* (2019).

[3] Sigal Alon. 2009. The evolution of class inequality in higher education: Competition, exclusion, and adaptation. *American Sociological Review* 74, 5 (2009), 731–755.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (May 2016).

[5] Chen Avin, Barbara Keller, Zvi Lotker, Claire Mathieu, David Peleg, and Yvonne-Anne Pignolet. 2015. Homophily and the glass ceiling effect in social networks. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 41–50.

[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[8] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).

[9] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[10] David A Cotter, Joan M Hermsen, Seth Ovadia, and Reeve Vanneman. 2001. The glass ceiling effect. *Social forces* 80, 2 (2001), 655–681.

[11] Cynthia Dwork, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness through awareness. *arXiv preprint arXiv:1104.3913* (2011).

[12] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics* 9, 2 (2017), 1–22.

[13] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[14] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1914–1933.

[15] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. ACM, 111–122.

[16] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv preprint arXiv:1610.02413* (2016).

[17] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[18] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 137–146.

[19] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. *arXiv preprint arXiv:1706.02744* (2017).

[20] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[21] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. *arXiv preprint arXiv:1703.06856* (2017).

[22] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 230–239.

[23] Shirin Nilizadeh, Anne Groggel, Peter Lista, Srijita Das, Yong-Yeol Ahn, Apu Kapadia, and Fabio Rojas. 2016. Twitter's Glass Ceiling: The Effect of Perceived Gender on Online Visibility. In *Tenth International AAAI Conference on Web and Social Media*.

[24] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

[25] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[26] Scott E Page. 2008. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies-New Edition*. Princeton University Press.

[27] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[28] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. 2017. The Problem of Infra-marginality in Outcome Tests for Discrimination. *Annals of Applied Statistics* 11 (2017).

[29] Ana-Andreea Stoica and Augustin Chaintreau. 2019. Fairness in Social Influence Maximization. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 569–574.

[30] Ana-Andreea Stoica and Augustin Chaintreau. 2019. Hegemony in Social Media and the effect of recommendations. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 575–580.

[31] Ana-Andreea Stoica, Abhijnan Chakraborty, Palash Dey, and Krishna P Gummadi. 2019. Minimizing Margin of Victory for Fair Political and Educational Districting. *arXiv preprint arXiv:1909.05583* (2019).

[32] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 923–932.

[33] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*. IEEE, 1521–1528.

[34] Sandra Wachter. 2019. Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. *Available at SSRN* (2019).

[35] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1171–1180.

[36] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Sanjoy Dasgupta and David McAllester (Eds.), Vol. 28. PMLR, Atlanta, Georgia, USA, 325–333. http://proceedings.mlr.press/v28/zemel13.html