Fairness and Explainability in Algorithmic Hiring

Proposers' Names and Contact Information

- (PI) John P. Dickerson, Assistant Professor, Department of Computer Science, University of Maryland, College Park, MD 20742; john@cs.umd.edu. URL: `http://jpdickerson.com`
- (Co-PI) Nicholas Mattei, Assistant Professor, Department of Computer Science, Tulane University, New Orleans, LA 70118; nsmattei@tulane.edu. URL: `http://www.nickmattei.net`

Technical Details

Proposal Abstract. "Hiring is rarely a single decision point, but rather a cumulative series of small decisions." So begins a recent report on *automated hiring processes* released by the non-profit group UpTurn [7], before recommending that digital sourcing firms begin explicitly addressing concerns of fairness and bias. Indeed, at various decision points in the hiring process, algorithms are already used to determine who sees which job advertisements; estimate the expected performance of an applicant; select which applicants to screen more heavily and with whom to match them; and forecast salary and other benefits necessary to ensure a successful offer. Thus, issues of bias or fairness at one stage of this procedure may lead to unexpected or amplified issues later on.

In addition to the difficulty of these decisions on their own, there are a number of regulatory and legal requirements that must be met at each stage of the hiring process. As a recent Facebook settlement[1] showed, the tools, platforms, and techniques developed to streamline hiring can be subtly—or blatantly—illegal. Hence, in order to streamline the hiring process and ensure both fairness and legal compliance, there is a need for novel techniques from data science, artificial intelligence, and machine learning to ensure our algorithms act within the constraints set forth by business process, laws, and norms [18]. In this proposal, we discuss two key factors related to the job hiring pipeline: how to allocate effort (e.g., budget, interview slots) along the hiring pipeline and how to explain the decisions made by our algorithm in a transparent and compliant way. This research directly addresses questions of transparency, constraints, and fairness when working with multi-armed bandit algorithms [3, 4, 14, 19]. Our proposed work will support the following thesis:

Data-driven approaches to measuring and promoting fairness and explainability at a single stage of the hiring process can be extended—in a principled way—to the full, multi-stage hiring process.

Proposed Research. We propose to develop principled, data-driven methods for quantitative talent sourcing that are general-purpose enough to be used in different industries, under different hiring processes and norms, and assuming different definitions of "fairness" and "transparency." To complement that generality, throughout, we will test our approaches on a specific use case: graduate admissions, a form of academic hiring. PI Dickerson already works with real graduate admissions data at Maryland, with the support of his department chair and the Institutional Review Board (IRB); the senior Maryland Ph.D. student who will be funded by this award, Candice Schumann, has already developed the methods to parse application data (including OCR methods for scraping text from scanned recommendation letters and transcripts, feature extraction, and so on). We will validate our algorithmic approaches from this proposal on that real data, which is already collected and stored in an easily-accesible way. (The storage and access patterns are secure, as discussed in the "Data Policy" section below).

First, we will develop a series of algorithms that help hiring committee members gather and aggregate data to make decisions that take fairness into account. Our starting place is prior work by PI Dickerson and colleagues that considers how to **allocate interviewing resources** during graduate admissions [19]. Our key idea is that *structured interviews* can potentially provide more information about applicants and reduce bias [2, 15, 17, 22], but they are too expensive to conduct with every applicant. To address this issue, we developed an algorithm in the multi-armed bandit (MAB) setting in which written application reviews are modeled as *weak pulls* with low cost but low information gain, and interviews are modeled as *strong pulls* with high cost and high information gain. Our algorithm allocates weak and strong pulls to optimize a desired objective. We showed that, in simulation, our algorithm selected a more diverse cohort with similar quality to the actual Maryland graduate admissions process [19].

We propose to extend this idea to a **multi-tiered setting**, e.g., in the graduate admissions case, the process could include an initial screening for minimum qualifications such as GPA, then an application review, followed by a Skype interview, and finally an in-person interview. Figure 1 gives an example tiered hiring process, and shows (in red text) where our proposed interventions fit into the present hiring system. To our knowledge, no multi-armed bandit model exists that captures this multi-tiered setting; we propose to develop a model (extending parts of our prior work [19]), along with theoretical results that can be used in practice to guide various design decisions such as budget, timing, number of simultaneous interviews, and others.

Our presently-developed methods allow for the promotion of diversity in the final cohort of applicants (e.g., graduate students). Dovetailing with this, the **fairness of the review process** is also important. In the MAB setting, we propose to incorporate and expand methods from the fairness in machine learning literature [5, 11] (such as

[1] `https://www.propublica.org/article/facebook-ads-discrimination-settlement-housing-employment-credit`

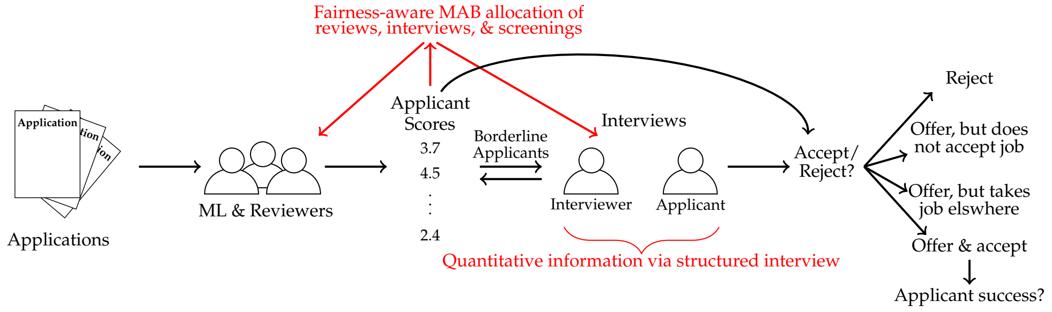

Figure 1: A sample current tiered hiring process (in black) and the interventions of this proposal (in red).

those developed within the silos of fairness of treatment and fairness of outcome) into our single-level and multi-tiered settings, and will explore theoretical metrics such as the impact on overall economic efficiency due to the use of a "fair" objective, and experimental validation on sensitive attributes such as self-reported gender, race, and country of origin that are available in our real data sets.

We note that notions of "fairness, "bias," and "explainability" are (i) definable in many ways [12] and (ii) necessarily different based on application areas, societal norms, and policy-maker preferences. However, in hiring, credit, and housing there are a number of federally protected features that one must not use in the decision making process and also must not use for explanation. Simply removing these features from consideration of our algorithms is not enough and we must actively ensure the fairness criteria is enforced across these features [9]. Thus, we endeavor to remain somewhat definition-agnostic in our modeling work, and then explicitly instantiate a definition when needed (e.g., we plan to use the well-known *equality of opportunity* [13] definition of fairness in our earliest experiments). However, our proposed approaches should generalize to a whole host of fairness or parity measures, so long as the measure of bias/fairness can be written as a linear constraint on conditional moments of predicted distributions over predictions, ground truth, and protected attributes [1].

Approach & Expected Results. We propose a multidimensional approach to tackling issues in the efficient *gathering* and *aggregation* of information by review committee members, which jointly compose part of a decision support system for potential job offer *decisions*. We propose to use the concept of *structured interviews* [8, 21], used widely in industry as well as in some academic programs (e.g., Fisk-Vanderbilt [20]); then, drawing on our prior work (described above), we will cast *tiered* hiring as a combinatorial pure exploration (CPE) problem in the stochastic multi-armed bandit setting [10]. The goal is to select a cohort of applicants after narrowing the pool after successive stages or tiers. Each tier or interviewing stage has an associated strength of arm pull, similar to (indeed, generalizing the concept of) the weak and strong arm pulls discussed earlier, and introduced in our prior work [19]. The strength determines the confidence of the signal generated by the reviewer/interview as well as the cost of performing an arm pull. This allows us to generalize to more than two types (weak and strong) of arm pulls which will make the system more accesible to hiring committees.

After each stage the applicant pool is narrowed. In other words, during each stage K_i arms (i.e., applicants) move on to the next stage (i.e., we remove $K_{i-1} - K_i$ arms), where $n = K_0 > K_1 > \cdots > K_{m-1} > K_m = K$). Therefore, each stage i could be considered a cohort selection problem where K_i applicants need to be selected in order to maximize some objective function. We will first prove results until a linear objective, as is standard in the Top-K MAB literature. Theoretical results may include proving bounds on convergence, or providing principled methods for choosing user-defined parameters (such as the short-list sizes K_i for each stage i). We will empirically validate our models and theoretical results (under both types of objectives) on the Maryland graduate data set that we have on hand, and a second data set that Co-PI Mattei will build at Tulane.

We are proposing a data-driven decision support tool that draws inferences in part based on observed and estimated features of humans—and such tools are increasingly known to result in unexpected or adverse impact on dimensions such as fairness and bias [16]. We acknowledge that both our initial work in this space as well as our proposed extension to the more realistic tiered admissions setting may exacerbate issues of *fairness*. Thus, we also propose to incorporate recent definitions of fairness from the machine learning community into our tiered admissions model, and perform analysis on our real admissions data.

Specifically, we will extend our model to include notions of at least *fair treatment* or *fair outcome* (and possibly others) in the multi-armed bandit setting [14].[2] We overview both notions of fairness now. As in the previous sections, we have a set of arms $a \in A$ where A is partitioned into L groups $A = P_1 \cup P_2 \cup \cdots \cup P_L$, but now corresponding to specific sensitive attribute groups. We will begin by studying self-reported gender, race, and country

[2] We re-emphasize that, throughout, our models will be built to accept a host of fairness and parity measures (subject to the light assumptions discussed earlier.); still, it is important to provide concrete plans for specific definitions of each.

of origin—each of which is present in our real data set.

Roughly, under a fair treatment regimen, an algorithm will not treat individuals (in our case, pulling arms) differently based on sensitive attribute value(s). That is, for arms a and b, ensure $\Pr[\text{pull } a|a \in P_1] = \Pr[\text{pull } b|b \in P_2]$. Conversely, under a fair outcome regimen, an algorithm may enforce so-called equality of opportunity, attempting to equalize false positive rates across groups (i.e., sensitive attributes). We propose to (i) first incorporate concrete instantiations of these two approaches into our tiered model, (ii) measure the "price of fairness" [6]—the relative loss in system efficiency due to using a fair objective—under each of these concrete instantiations, and (iii) validate on our real data sets using three sensitive attributes (self-reported gender, race, country of origin). Our expected outcomes for this project are as follows:

1. A mathematical formulation of *fair treatment* of equals when applied to tiered hiring, formulated as a multi-armed bandit (MAB) problem with varying arm pulls;
2. A working implementation, simulated on real data to which we currently have access and IRB approval, of the proposed algorithm(s) in our public GitHub repository, with documentation; and
3. Public and detailed white paper including data analysis on real-world admission data highlighting the results of our experiments, and a discussion of lessons learned and similarities to traditional talent sourcing.

Budget

The proposed budget includes only direct costs. It covers two students, each for one full semester as well as a full shared summer. Additionally, the budget covers travel for each student to visit the other's university for one week in the initial three months of the proposal, to solidify early-stage alignment of objectives.

Data Policy

Advances made in this project will be published in peer-reviewed venues and/or technical reports in multiple disciplines (e.g., computer science or management science, when appropriate). We intend to release software developed for this project as open source.[3] Application, admissions, and graduate student success data will remain on secure servers at each institution and will not be directly shared with the other institution. No additional admissions access will be granted beyond what would normally be available for admissions work. Only the PIs will have access to student success data, and only for their institution. The PIs will run algorithms on that data on behalf of the graduate students on the project. Our preliminary work in this space has been, and continues to be, supported by the Institutional Review Board (IRB); for any additional work, we will ensure compliance.

Tulane PhD Stipend	$21,881	8 months
Tulane PhD Tuition	–	*Tulane does not charge*
Maryland PhD Stipend	$26,041	8 months
Maryland PhD Tuition	$5,848	7 credits
PhD Travel (2x)	$6,000	2x university visits
Total	**$59,770**	

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, 2018.

[2] Richard D Arvey and James E Campion. The employment interview: A summary and review of recent research. *Personal Psychology*, 35(2):281–322, 1982.

[3] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Using contextual bandits with behavioral constraints for constrained online movie recommendation. In *IJCAI*, 2018.

[4] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Incorporating behavioral constraints in online AI systems. In *AAAI*, 2019.

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018.

[6] Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. The price of fairness. *Operations Research*, 59(1):17–31, 2011.

[7] M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. Technical report, Upturn, 2018.

[8] J. Breaugh and M. Starke. Research on employee recruitment: So many studies, so many remaining questions. *Journal of Management*, 26(3):405–434, 2000.

[9] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *NIPS*, pages 3992–4001, 2017.

[10] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *NIPS*, pages 379–387, 2014.

[11] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

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

[13] Moritz Hardt, Eric Price, , and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[14] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *ICML*, 2017.

[15] Eugene C Mayfield. The selection interview—a re-evaluation of published research. *Personal Psychology*, 17(3):239–260, 1964.

[16] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

[17] Richard A Posthuma, Frederick P Morgeson, and Michael A Campion. Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Personal Psychology*, 55(1):1–81, 2002.

[18] Francesca Rossi and Nicholas Mattei. Building ethically bounded AI. In *AAAI*, 2018.

[19] Candice Schumann, Samsara N. Counts, Jeffrey S. Foster, and John P. Dickerson. The diverse cohort selection problem. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2019. To appear. Preliminary versions appeared at Aligned AI Workshop at NIPS-17 and the Women in Machine Learning (WiML) Workshop at NIPS-17.

[20] Keivan G Stassun, Susan Sturm, Kelly Holley-Bockelmann, Arnold Burger, David J Ernst, and Donna Webb. The Fisk-Vanderbilt Master's-to-PhD Bridge Program: Recognizing, enlisting, and cultivating unrealized or unrecognized potential in underrepresented minority students. *American Journal of Physics*, 79(4):374–379, 2011.

[21] Pelin Vardarlier, Yalcin Vural, and Semra Birgun. Modelling of the strategic recruitment process by axiomatic design principles. In *Procedia–Social and Behavioral Sciences*, pages 374–383, 2014.

[22] Laura Gollub Williamson, James E Campion, Stanley B Malos, Mark V Roehling, and Michael A Campion. Employment interview on trial: Linking interview structure with litigation outcomes. *Journal of Applied Psychology*, 82(6):900, 1997.

[3] Indeed, we have already released the code for our initial paper [19] on GitHub at `https://github.com/principledhiring/SWAP` and will continue pushing updates to that GitHub group and repository as our work progresses.