Long-Term Effects of Fair Ranking

Eric Chen, Ransheng Guan, and Neil Meehan

August 2020

Abstract

The widespread adoption of artificial intelligence systems in decision making today has caused algorithmic fairness to become a growing concern. Due to their omnipresence, ranking systems, used in applications as diverse as advertising, hiring, and dating, have come under scrutiny in recent years. Previous literature has focused on long-term fairness in classification tasks, and we propose a mathematical and computational framework under which long-term fairness dynamics in ranking systems can be modeled. We illustrate the impact of fair ranking by providing simulated results using synthetic data. We also contribute an outline of fair ranking as modeled by Markov chains and examine what fairness may look like in such a setting.

1 Introduction

In 2016, an unusual case appeared before the courts, challenging commonly-held beliefs about the nature of computing. The question itself was simple to pose, albeit much harder to answer: what — if any — was the role of predictive algorithms in criminal justice? A Wisconsin man, Eric L. Loomis, argued that his sentencing was unlawful due to the use of a risk assessment algorithm named COMPAS. At the same time the debate hit the courts, a study by ProPublica found that COMPAS exhibited a startling bias against Black defendants, with Black defendants being incorrectly judged as recidivists, or criminal reoffenders, far more often than white defendants.

With the ever-increasing prevalence of artificial intelligence systems in decision making, their potential impact on society, as well as the ethics of their use, is a growing cause for concern. Given that machine learning algorithms can be difficult to introspect and interpret, they have enormous capacity to perpetuate harm, especially in sensitive, high-stakes situations. Examples closer to everyday life exist as well, fraught with their own pitfalls: ranking candidates for a job interview, deciding which credit applicants to lend to, and many more.

Algorithmic fairness is a field of study concerned with the decisions made by machine learning models and their relationship with what are called protected attributes — things such as sex, race, and age. There is no single definition for what constitutes a "fair" algorithm; many different characterizations and fairness metrics exist. In fact, many impossibility theorems exist that state how competing fairness metrics cannot be mutually satisfied under reasonable conditions [FSV16, KMR16].

Given that most existing research focuses on static fairness in classification problems, we decided to investigate how the concepts of dynamic fairness and group fairness could be applied to ranking systems.

We show this is a valid concern by highlighting a common use case for ranking systems in real life: automated hiring practices. Oftentimes, algorithms will be used to parse resumes and rank candidates by some query-specific metric of fit; we will refer to this as relevance. However, hiring is often subject to the "rich-get-richer" effect, where candidates with previous work experience are preferred over those who do not possess such qualifications. When prior disparities between population subgroups exist, subsequent hiring decisions can compound them and widen inter-group inequities over time. Our research sought to understand the dynamics of such systems and how imposing various fairness constraints can help correct — or exacerbate — such disparities.

To do this, we formulated a mathematical framework for long-term fairness in ranking systems and designed a simulation to model the effects of various fairness-in-ranking constraints on synthetic data. In our experiments, we saw that enforcing these constraints does not necessarily lead to long-term fairness between two groups. These results suggest, similar to analogous findings in classification systems, that we need to better model the long-term effects of ranking systems in order to truly achieve fair outcomes.

2 Literature Review

2.1 An Overview of Fairness

We conducted an extensive literature review to familiarize ourselves with the field and identify areas of interest. Research in algorithmic fairness has been broad and covers many different domains, but most existing literature is concerned with classification in the static setting; many areas outside of classification — such as rankings — have not been looked at extensively. We summarize some common fairness concepts and paradigms below.

- **pre-**, **in-**, and **post-processing**: fairness can be enforced at various points in model development. Pre-processing refers to the practice of debiasing input data so that fairness considerations are model-agnostic. In-processing refers to the enforcing of fairness during model training, such as through the addition of constraints or regularization. Post-processing seeks to avoid bias in already-trained models; in the classification setting, one way to accomplish this is by adjusting the resulting decision thresholds based on group membership.
- **disparate treatment** and **disparate impact**: taken from precedent in U.S. labor law, these refer to the effect of discriminatory practices on disadvantaged groups, with treatment differing from impact by the presence of intent; disparate treatment is explicitly intentional, while disparate impact is not necessarily so [BS16].
- **individual fairness** and **group fairness**: group fairness involves satisfying some fairness criterion with respect to different subgroups (e.g. Black and white defendants, men and women). Independence, separation, and sufficiency are three such criteria that deal with the relationship between the model output, protected characteristic, and true label; these concepts are explained well in [BHN19]. Achieving such constraints is usually equivalent to equalizing some statistical quantity, such as false positive rates, false negative rates, or accuracy, across the subgroups. In contrast, individual fairness is the principle that similar individuals should be treated similarly regardless of their group status; similarity is usually a task-specific metric. Note that individual fairness does not necessarily imply group fairness, and vice versa.
- **static fairness** and **long-term (dynamic) fairness**: the decisions machine learning models make can have the potential to directly affect the population, resulting in a feedback loop that changes not only the populations over time, but also downstream decisions made by the algorithm; this feedback mechanism is what is known as "long-term" fairness. In contrast, static fairness is concerned with the fairness of a model output at a single point in time without focusing on feedback effects.
- **classification** and **ranking**: two of the most common use cases for machine learning algorithms, classification and ranking tasks possess major differences between them that can make it difficult to find direct parallels. While subject selection in classification is usually assumed to be independent, the ranking space is an inherently competitive setting, with higher-ranked subjects necessarily taking precedence over lower-ranked subjects. There is also no directly-equivalent "selection" process in ranking problems; literature on click-through rates seems to be the closest in concept.

The goal of many machine learning models is to maximize utility to the end user by optimizing for some performance metric. However, in algorithmic fairness, this "max-util" strategy must be tempered with fairness considerations; there is usually a tradeoff between utility and fairness.

2.2 Fairness in Classification

Since most fairness research has focused on classification problems, it is important to gain a good understanding of fairness in this setting in order to identify parallels in the ranking space.

Algorithmic fairness has substantial precedent in well-established subjects such as education [HM18], from which it draws many of its group fairness metrics. Most of these are derived from a confusion matrix. The following list is only the tip of the iceberg, and many more are included in [HM18].

- **statistical (demographic) parity**: the selection rate for each group must be equal to its overall population proportion; this is the same as achieving independence.
- **equality of opportunity**: the true positive rates must be equal between groups [HPS16].
- **equalized odds**: the true positive and false positive rates (or equivalently, their ratio) must be equal between groups.

The canonical definition of individual fairness is referred to as **fairness through awareness**: "similar" individuals must be treated "similarly", where similarity is a task-specific metric. Achieving individual fairness is not always as clear-cut as achieving group fairness, as this a similarity metric can be difficult to define.

2.3 Fairness in Ranking

In the ranking space, there is no confusion matrix to derive metrics from. Instead, the conventional wisdom has been that the "best" ranking should satisfy the following *Probability Ranking Principle*: subjects should be ordered by decreasing probability of relevance [SJ18]. Studies such as [AP18] have demonstrated that rankings are subject to a *position bias* in which users pay disproportionately more attention to the top few results. As a consequence, the resulting visibility or *exposure* of each subject decreases at a nonlinear rate. Most utility metrics in ranking are therefore functions of relevance and exposure, with exposure itself being a decreasing, nonlinear function of position; we refer to this function as a discount. However, when exposure does not decrease at the same rate as relevance, such rankings can be deemed unfair. Fairness considerations in ranking are therefore largely concerned with fair exposure, such as the following:

- **top-k constraints**: these ensure that the proportion of protected class members in the top k positions in a ranking is at least p, with p often matching the underlying population proportion of the protected group. We can view this special case as an analogue of demographic parity in classification problems, as both seek to enforce the population proportions in the model outputs. In [ZBC$^+$17], the authors extend this idea by also enforcing this constraint for every $j \leq k$.
- **exposure-merit ratios**: another method of achieving exposure-based fairness is to constrain the exposure-to-merit ratios for each subject or group, where merit is a function of relevance. In [BGW18], the authors enforce equity of attention, where the received attention of each subject or group over time is proportional to its relevance, so that the resulting ratios are equal across groups. Note that the merit function is taken as the identity here. [SJ19] contribute a relaxed version of this principle where the exposure-to-merit ratio for each subject satisfies a one-sided inequality constraint such that the exposure-to-merit ratio for a group with greater merit is less than or equal to that of a group with lesser merit.
- **randomization** and **amortization**: it is often impossible for a single ranking to satisfy desired fairness criteria due to unequal exposure. [SJ18] sought to achieve fairness in expectation, where the objective is to find an optimal distribution over rankings. [BGW18] use an amortization approach to individual and group fairness, where the accumulated exposure over a series of rankings is proportional to each subject's or group's accumulated relevance. This idea of achieving fairness through multiple rankings was something we used heavily in our work.

In addition to hard constraints, there also exist continuous metrics for measuring group fairness in ranking, such as normalized discounted Kullback-Leibler divergence [YS16]. Although interesting and worthy of consideration, we did not explore this concept in our project.

2.4 Long-Term Fairness

Recall the hiring scenario presented in Section 1 in which preexisting disparities between population subgroups get magnified over time. Long-term fairness seeks to move beyond static notions of fairness and instead consider the downstream effects of algorithmic decisions on population groups, as well as the decision algorithm itself.

The seminal paper in this area is [LDR$^+$18], in which the authors develop a one-step feedback model in a hypothetical credit-lending scenario. They suggest three possible outcomes for population groups: improvement, stagnation, and decline. The authors also classify selection strategies, or *policies*, as causing relative harm, relative improvement, or active harm to a specific group by comparing the group outcome to the one resulting from the max-util policy. The authors show that implementing various fairness constraints can lead to all of these outcomes and plot an *outcome curve* showing under what conditions each of these three outcomes can be achieved. Especially notable was the idea that enforcing fairness constraints may even lead to decline.

In [ZL20], a multi-step, online problem is considered in which sequential algorithms are sorted into two classes:

- **P1**: the algorithm seeks to optimize their decision over the entire time horizon. Model decisions are restricted by partial information at every time step, but do not have the capability of directly changing the feature space the population is drawn from. The goal is to understand how fairness constraints impact the decision rule.
- **P2**: the algorithm learns an optimal decision rule for each time step. In this setting, model decisions can transform the underlying feature space. The goal is to understand the downstream effects of constrained and unconstrained model decisions.

Long-term fairness can also be viewed through more of an empirical lens [DSA+20]. In this paper, the authors contribute an open-source software library for running multi-step fairness experiments and highlight some toy examples. Although we developed our own framework for running experiments, the concept of using a simulation was key in our research, and this approach may help lay some groundwork for more rigorous, theoretical claims about long-term fairness.

In general, these findings support the idea that fairness is not necessarily achieved from a one-step process and that larger dynamics are usually at play.

3 Problem Set-Up

In this section, we present a mathematical model to understand the long-term dynamics of ranking systems. At a high level, we consider a scenario where a decision-maker (e.g., a job recruiter) uses rankings to facilitate determining whom to select for a task (e.g., whom to interview). We assume that each individual belongs one of two subgroups: an advantaged group or a disadvantaged group. With the standard learning-to-rank set-up, we model the long-term dynamics that selection or non-selection for the task has on each sub-population.

Formally, let $\mathcal{X}$ be a set of individuals such that each $d \in \mathcal{X}$ belongs to one of two groups: A or B. A query $q \in \mathcal{Q}$ is a set of n individuals: $q = \{d_1, \cdots, d_n : d_i \in \mathcal{X}, i \in [n]\}$. Each individual $d_i \in q$ possesses a query-dependent relevance score $\text{rel}(d_i|q) \in [0,1]$, drawn from a group-specific distribution Σ_j with mean μ_j for $j \in \{A, B\}$. Throughout this paper, we use the convention that A represents the advantaged group and B represents the disadvantaged group, so $\mu_A > \mu_B$. Let $\text{rel}^q = \{\text{rel}(d_i|q) \in \mathbb{R}\}_{i=1}^n$ be the relevance scores for each individual in a query, which is used for ranking the individuals in q. Let p_{d_i} be the probability of success of individual $d_i \in q$ for the task $\mathcal{T}$.

Ranking. Given a query q, define a ranking $r_q : q \to [n]$ such that $r_q(d_i)$ denotes the position of d_i in the ranking and $r_q^{-1}(j)$ denotes the j-th ranked subject, where a smaller number indicates a better position in the ranking. We drop q from r_q when clear from context. Let $\mathcal{R}_n$ be the set of all possible rankings on n items, and let $\Delta(\mathcal{R}_n)$ be the set of all possible distributions over the rankings. Given a query q, let the ranking policy $\tau_R : \mathcal{Q} \to \Delta(\mathcal{R}_n)$ be an algorithm that selects a distribution over rankings given a query q.

Selection. We define a selection policy $\tau_S : \mathcal{R} \to \Delta(P([n]))$ to be an algorithm whose input is a ranking and whose output is a distribution over the power set $\mathcal{P}([n])$.

Feedback Effects. To model feedback effects, we assume that the time horizon consists of K time steps. At each time step $k \in [K]$, we perform the following steps:

1. A query q_k is drawn from a query distribution $Q\big(\Sigma_A(\mu_A^{(k)}), \Sigma_B(\mu_B^{(k)})\big)$, where the relevance scores are sampled according to Σ_A with mean $\mu_A^{(k)}$ and Σ_B with mean $\mu_B^{(k)}$.
2. A ranking r_k is drawn from $\tau_R(q_k)$.
3. A subset s_k of q_k is drawn from $\tau_S(r_k)$.
4. Each selected subject $d_i \in s_k$ succeeds (resp. fails) at the task with probability p_{d_i} (resp. $1 - p_{d_i}$) independently from the other individuals.

Let o_k denote the success and failure outcomes.

5. Each individual $d_i \in q_k$ receives a reward (or penalty) u_i, where

$$u_i(d_i, q_k, r_k, s_k, o_k) = \begin{cases} u_+ & \text{if } d_i \in s_k \text{ and succeeds at } \mathcal{T} \\ u_- & \text{if } d_i \in s_k \text{ and fails at } \mathcal{T} \\ u_0 & \text{if } d_i \notin s_k. \end{cases}$$

Let $G_j^{q_k} \subseteq q_k$ represent the set of individuals in group j in query q_k. Given a query q_k, ranking r_k, selection set s_k, and the success and failure outcomes o_k of the selected individuals, the group change is

$$\Delta\mu_j^{(k)}(q_k, r_k, s_k, o_k) = \frac{1}{|G_j^{q_k}|} \sum_{d_i \in G_j^{q_k}} u_i(d_i, q_k, r_k, s_k, o_k) \tag{1}$$

6. For the next time step $k+1$, we modify the score distributions Σ_j by updating their means so that $\mu_j^{(k+1)} := \mu_j^{(k)} + \Delta\mu_j^{(k)}$.

We propose that long-term fairness in this setting is achieved when the two group distributions Σ_A and Σ_B are equal at the end of the time horizon. In this case, a ranking policy that maximizes utility with no regard to fairness will interleave both groups equally throughout the ranking, giving both groups equal exposure and representation in the ranking.

3.1 Ranking Algorithms τ_R

Given a ranking r of a query q and a set of relevance scores rel_q, let $\Delta(r, \text{rel}_q)$ be a ranking metric (e.g., normalized discounted cumulative gain). We define the utility of a ranking policy $\pi(\cdot|q)$, i.e., distribution over rankings, to be

$$\text{U}(\pi|q) = \mathbb{E}_{r\sim\pi(\cdot|q)}\Delta(r, \text{rel}_q). \tag{2}$$

We now discuss specific ranking policies.

Max-Util. In accordance with the Probability Ranking Principle [SJ18], the max-util ranking should order subjects by decreasing relevance. We can extend this principle to distributions over rankings by maximizing the *expected* utility of a ranking.

In particular, given a query q, our objective is to find an optimal distribution π^* over rankings:

$$\pi^* = \underset{\pi\in\Delta(\mathcal{R}_n)}{\arg\max}\ \text{U}(\pi|q) \tag{3}$$

Note that Equation (3) is maximized by the deterministic ranking policy that always selects the max-util ranking as given by the Probability Ranking Principle.

Top-t. Given a ranking r, a set of integers $\{t_1, t_2, ..., t_n\}$, and a specified proportion p, let p_i be the proportion of disadvantaged (i.e., Group B) subjects in the top t_i positions in r. We say r satisfies the top-t constraint if $p_i \geq p$ for all $i \in [n]$.

Let $\text{supp}(\pi)$ denote the support of a distribution π. Then, the optimal ranking distribution is given by:

$$\begin{aligned} \pi^* = \ & \underset{\pi\in\Delta(\mathcal{R}_n)}{\arg\max} && \text{U}(\pi|q) \\ & \text{subject to} && p_i \geq p \ \ \forall i \in [n]\ \forall r \in \text{supp}(\pi) \end{aligned} \tag{4}$$

Note that this selection policy concentrates all its probability mass on rankings that satisfy the top-t constraint, and maximizing utility should satisfy the *in-group monotonicity* constraint as described in [ZBC+17].

Deserved Exposure. We adapt the formulation from [SJ19]. Intuitively, fairness in ranking is achieved when group exposure is allocated proportionally to group relevance. Let the *deserved exposure* of a group be equal to its exposure-to-relevance ratio.

Let $\mathbf{v}$ be a discount vector encoding position bias, e.g., log discounting, where $\mathbf{v}_i = 1 \ / \ \log_2(i+1)$.

Given a distribution over rankings π, let the exposure of a document $d_i \in q$ be equal to the expected attention that it receives: $\text{exposure}(d_i|\pi) = \mathbb{E}_{r\sim\pi(\cdot|q)}[\mathbf{v}_{r(d_i)}]$. Define the exposure of a group $j \in \{A, B\}$ as the average exposure for all individuals $d_i \in G_j^q$:

$$\text{exposure}(j|\pi) = \frac{1}{|G_j^q|} \sum_{d_i \in G_j^q} \text{exposure}(d_i|\pi)$$

Define the relevance of a group $j \in \{A, B\}$ to be the average relevance of all individuals $d_i \in G_j^q$:

$$\text{rel}(j|q) = \frac{1}{|G_j^q|} \sum_{d_i \in G_j^q} \text{rel}(d_i|q)$$

Define the *disparity*, a measure of how much the deserved exposures differ, as follows:

$$\mathcal{D}(\pi|q) = \max\left(0, \frac{\text{exposure}(A|\pi)}{\text{rel}(A|q)} - \frac{\text{exposure}(B|\pi)}{\text{rel}(A|q)}\right)$$

Given a parameter δ, the optimal distribution under this policy is then given by:

$$\begin{aligned} \pi^* = \ & \underset{\pi \in \Delta(\mathcal{R}_n)}{\arg\max} \quad \text{U}(\pi|q) \\ & \text{subject to} \quad \mathcal{D}(\pi|q) \leq \delta \end{aligned} \tag{5}$$

As noted in [SJ19], the disparity is a one-sided measure since utility maximization already favors the opposite direction, where the deserved exposure of the advantaged group A exceeds the deserved exposure of the disadvantaged group B.

3.2 Selection Algorithms τ_S

In this section, we detail specific selection policies.

Deterministic Top-s. Given a ranking r and a positive integer s, this selection policy simply selects the top s individuals to be in s_k.

Stochastic Top-s. Given a ranking r, a positive integer s, and a discount vector $\mathbf{v}$, we start from the top of the ranking and successively select individuals with probability equal to their received attention. Here, the selection policy mimics a "click model" where higher-ranked individuals are viewed first and selected with greater probability.

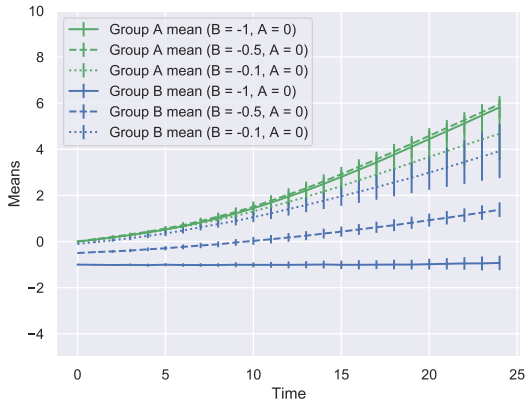
Starting at the top-ranked subject, we accept or reject each individual with probability $\mathbf{v}_i$, where $i \in [n]$ is the rank of the subject. If rejected, we continue on to the next subject. We stop when we have selected s individuals or exhausted all possible subjects.

4 Experiments

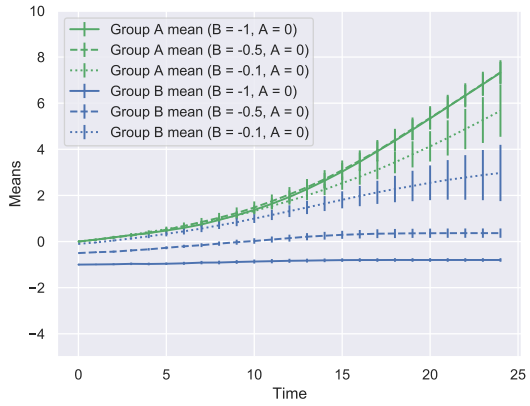
In order to investigate the long-term dynamics of fairness in ranking, we developed a Python framework for running simulations on synthetic data.

4.1 Experimental Set-Up

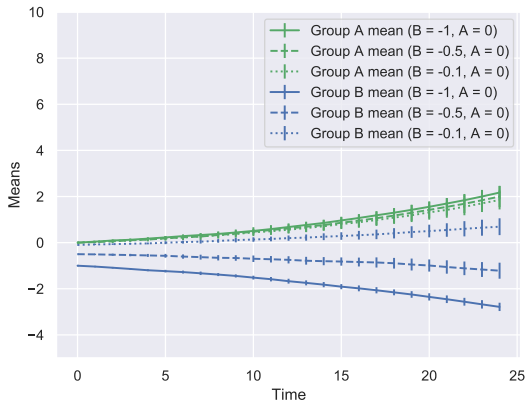
For our experiments, we sampled 10 queries of length $n = 20$ for each of the $K = 25$ time steps. Let $p = .5$ be the proportion of advantaged individuals in each query. To build each query at time step k, we sample $\lfloor np \rfloor$ relevance scores from the advantaged score distribution Σ_A and $\lfloor n(1-p) \rfloor$ scores from the disadvantaged score distribution Σ_B. Both the distributions Σ_A and Σ_B are logit-normals: for each individual in group i, we sample $x_i \sim N(\mu_i^{(k)}, 1)$ and then take $\frac{1}{1+e^{-x_i}}$ to be the relevance score of individual i. However, the framework allows for other distributions with support in $[0, 1]$ (e.g., beta). The probability of success of each individual is equal to their relevance score: $p_{d_i} = \text{rel}(d_i|q)$. The number of individuals (attempted to be) chosen by the selection policy is $s = 10$. We set Δ in Equation (2) to be normalized discounted cumulative gain (NDCG), a metric commonly used in information retrieval papers for measuring ranking quality. Given a ranking r, the discounted cumulative gain of r is given by $\text{DCG}(r) = \sum_{i=1}^{n} \frac{\text{rel}(r^{-1}(i)|q)}{\log_2(i+1)}$. The NDCG of r is then defined as $\frac{\text{DCG}(r)}{\text{DCG}(r^*)}$, where r^* is a ranking that maximizes DCG. The rewards u_i were chosen as follows: $u_+ = 0.5$, $u_- = -0.5$, $u_0 = 0$. In our experiments, we consider both the deterministic and stochastic top-s selection policies; we also compare the max-util ranking policy to the top-t fair ranking policy, which we implemented by adapting the code from [ZBC+17]. We investigate the effect of varying the initial means $\mu_A^{(1)}$ and $\mu_B^{(1)}$ such that $\mu_B^{(1)} < \mu_A^{(1)}$.



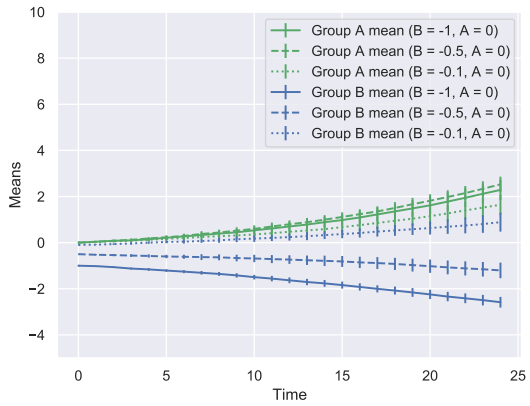
(a) top-t ranking, top-s selection.



(b) max-util ranking, top-s selection.



(c) top-t ranking, top-s stochastic selection.



(d) max-util ranking, top-s stochastic selection.

Figure 1: Experiments on top-t fair ranking policies.

4.2 Simulations

In this section, we describe some of the specific experiments we implemented. We will focus on top-t fairness dynamics, as we did not have sufficient time or computing power to run simulations involving the deserved exposure constraint.

We sought to compare ranking policies and selection policies to understand how they affected the population distributions in the long run. To this end, we tracked how μ_A and μ_B changed over time. For ease of comparison, in the figure, selection policies are constant across rows and ranking policies are constant down columns.

As seen in Figures 1(a) and 1(b), top-s selection tends to enforce a "rich-get-richer" dynamic, with Group A's mean increasing at a faster rate than Group B's mean. Recall that Group A is the advantaged group, while Group B is the disadvantaged group. Max-util ranking seems to exacerbate this effect slightly more than top-t ranking does, which should make intuitive sense. To use the terminology in [LDR$^+$18], note that Group B essentially stagnates if the initial mean μ_B is sufficiently smaller than μ_A. However, in no situation does Group B ever decline.

When we compare the selection policies as seen in Figures 1(a) and 1(c), we see that stochastic selection dampens the rich-get-richer effect somewhat, with Group A's mean increasing at a slower rate than with top-s selection. Also notable is the fact that stochastic selection can actually lead to decline in Group B. We hypothesize that this is due to the probabilistic nature of the stochastic selection policy; while individuals outside of the top-s results will never get selected under deterministic top-s selection, it is entirely possible — even likely — that these individuals will be selected under the stochastic click model. These individuals have lower relevance scores, and are consequently more likely to fail at the task and incur a penalty for the group.

Overall, we see that implementing top-t fairness constraints during ranking do not necessarily offer improvements over a max-util strategy, and may still lead to decline for the disadvantaged group; though we are working under a number of assumptions and with synthetic data, this conclusion could prove to be valuable when evaluating how to approach a problem in the fair ranking space. If we are faced with a situation where the disadvantaged group is significantly worse off from the outset, we may have to consider more drastic fairness constraints or other ways to eradicate bias, since top-t alone will not be enough to make a noticeable impact. We also see that the selection policy tends to have more of an impact on results than the ranking policy itself does, suggesting that we may also need to implement fairness measures in selection algorithms as well as ranking.

4.3 Future Work

We have only conducted preliminary experiments on long-term fairness in ranking. For example, it would be worthwhile to investigate how sensitive outcomes are to parameters such as the population proportions, length of time horizon, and choice of s, the number of people (attempted to be) selected by the selection policy.

In addition, varying rewards and penalties in the feedback step may affect long-term outcomes; in particular, incurring a penalty for individuals who are not selected by the selection policy may greatly exacerbate disparities between the two populations, which we did not explore.

The stochastic ranking policy as described in Section 3.1 is another direction for future research; while the ranking policies in our experiments were mainly deterministic, the additional randomness induced by this ranking policy may introduce dynamics into the system that are worthwhile to model.

Another idea is to include the concept of temporary "interventions"; for instance, is it possible to achieve long-term fairness by starting out with a utility-maximizing strategy, applying a fair ranking policy for some amount of time, and then switching back to max-util? This also ties into the greater question of stability, which seeks to answer if and how fair outcomes can be made unfair, and under what conditions. It may also be valuable to try adapting the simulation framework to work on real datasets instead of synthetic data, if any become available, in order to strengthen any conclusions one may draw from running these experiments.

5 Markov Chain Ranking

In machine learning, Markov chains are often used to model sequential events. In this section, we model long-term ranking using a Markov chain framework, equivalent to viewing the ranking problem from a sequential decision making lens.

5.1 Motivation

We motivate the Markov chain model with an example from sociology, as adapted from [Mon13]. Some sociologists and economists are interested in investigating the intergenerational transitions between social classes (e.g., a parent's social class vs. a child's social class). The relative ease or difficulty of transitioning between social classes as defined by income or occupation is known as social mobility.

Each social class can be considered a state in the Markov model. The transition diagram or matrix gives the probabilities of moving between social classes in consecutive generations. Assuming that social mobility is indeed a Markov chain process, we can also obtain longer relations through matrix algebra.

There are many important properties that Markov chains may satisfy that help us understand fairness in ranking and social mobility:

1. **irreducibility**: there is a non-zero probability of transitioning from any state to any other state.
2. **aperiodicity**: the positivity of the diagonal elements of the powers of the transition matrix has no periodic behavior. This is satisfied, for example, if every state has a self-loop.
3. **double stochasticity**: all rows and columns sum to 1.

We also note that a transition matrix is **diagonally dominant** if, for every row of the matrix, the magnitude of the diagonal entry in a row is larger than or equal to the sum of the magnitudes of all the other (non-diagonal) entries in that row. In the context of ranking, this suggests that the rankings are relatively stable and do not tend to change much between time steps.

5.2 Worked Example

We tested this framework on the U.S. News Rankings dataset using data from 1996 to 2019. U.S. News Rankings is a resource for students to decide where to apply and find their best-fit college. The rankings compare institutions from across the U.S. on 15 diverse measures of academic quality, based on criteria such as faculty resources, expert opinion, financial resources, student excellence, and alumni-giving.

The transition matrix M is defined as

$$M_{ij} = \mathbf{P}(\text{university ranked } i^{th} \text{ moves to } j^{th} \text{ position in next time step}).$$

We can visualize the transition matrix with a heatmap, as seen in Figure 2. We can see that the transition matrix is diagonally dominant with 4 blocks.

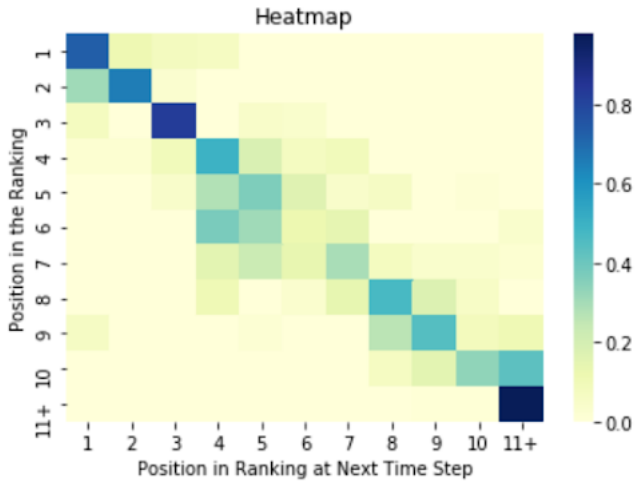


Figure 2: Transition matrix for 1996-2019 U.S. News ranking data.

5.3 Discussion

We now discuss various notions of fairness in this framework. Since it is not fair if one position in the ranking is not reachable from another, we claim that irreducibility should be required for a Markov model

to be fair. Moreover, it is not fair if the ranking can never stay the same from time t to time $t+1$, i.e., if there are no self-loops. This is equivalent to requiring every diagonal entry to be strictly positive; recall that this is a sufficient condition for satisfying aperiodicity. If irreducibility and aperiodicity are both satisfied, then the Markov chain has a unique stationary distribution. If we also ensure that the transition matrix is also doubly stochastic, then the stationary distribution is uniform.

Diagonal dominance is somewhat of a "short-term" unfairness criterion, as it tends to only manifest for a few time steps. Convergence speed, determined by the first eigengap, or difference between the two largest eigenvalues of the transition matrix, measures how fast the transition matrix converges to the stationary distribution.

As seen in Figure 3, the diagonal dominance seen in the transition matrix is much less prominent after 10 time steps and completely absent after 1000 time steps. Markov chains with faster convergence times tend to display diagonal dominance for shorter amounts of time.

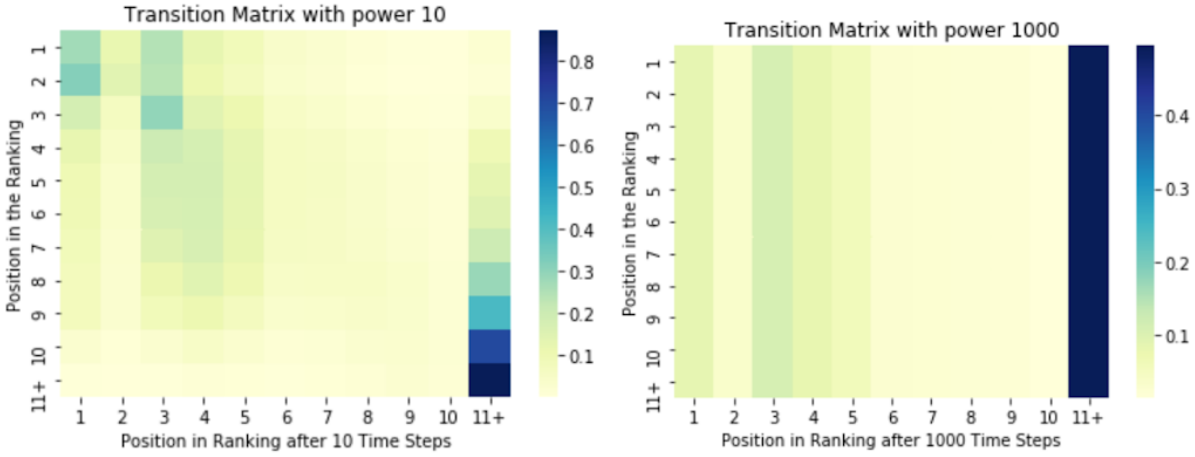


Figure 3: Transition matrix powers for 1996-2019 U.S. News ranking data.

For future work, it may be valuable to explore other definitions of fairness. For instance, we would be interested in seeing how the concept of exposure could be reformulated in this setting. In addition, it may be worthwhile to see how fairness can be achieved through audits or interventions, as mentioned in Section 4.3. Another possibility is to impose constraints on the transition matrix to guarantee some definition of fairness is satisfied. Finally, we could also formulate the problem as a constrained Markov decision process to see if it will give decision-makers a tool to keep rankings fair.

6 Conclusion

In this report, we developed a simulation framework for running fair ranking experiments in Python. We presented a mathematical model of the fair ranking problem as well as the long-term dynamics of repeated ranking and selection. We saw that top-t fair ranking is insufficient for ensuring long-term fairness, and may even exacerbate existing disparities over time. Later on, we digress and explore a Markov chain model for ranking and discuss what fairness would look like in such a setting. It is important to reiterate that our investigation of long-term fairness in ranking has only skimmed the surface; future work will be of the utmost importance if we are to truly understand fairness dynamics in ranking.

References

[AP18] Grigor Aslanyan and Utkarsh Porwal. Direct estimation of position bias for unbiased learning-to-rank without intervention. *CoRR*, abs/1812.09338, 2018.

[BGW18] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. *CoRR*, abs/1805.01788, 2018.

[BHN19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[BS16] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016.

[DSA+20] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[FSV16] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016.

[HM18] Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: Lessons for machine learning. *CoRR*, abs/1811.10104, 2018.

[HPS16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[KMR16] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016.

[LDR+18] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *CoRR*, abs/1803.04383, 2018.

[Mon13] James Montgomery. *Social Mobility*. 2013. `https://www.ssc.wisc.edu/~jmontgom/socialmobility.pdf`.

[SJ18] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. *CoRR*, abs/1802.07281, 2018.

[SJ19] Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. *CoRR*, abs/1902.04056, 2019.

[YS16] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. *CoRR*, abs/1610.08559, 2016.

[ZBC+17] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. FA*IR: A fair top-k ranking algorithm. *CoRR*, abs/1706.06368, 2017.

[ZL20] Xueru Zhang and Mingyan Liu. Fairness in learning-based sequential decision algorithms: A survey, 2020.