Explaining Monotonic Ranking Functions

Abraham Gale
Rutgers, the State University of New Jersey
Piscataway, New Jersey
abraham.gale@rutgers.edu

Amélie Marian
Rutgers, the State University of New Jersey
Piscataway, New Jersey
amelie.marian@rutgers.edu

ABSTRACT

Ranking functions are commonly used to assist in decision-making in a wide variety of applications. As the general public realizes the significant societal impacts of the widespread use of algorithms in decision-making, there has been a push towards explainability and transparency in decision processes and results, as well as demands to justify the fairness of the processes. In this paper, we focus on providing metrics towards explainability and transparency of ranking functions, with a focus towards making the ranking process understandable, *a priori*, so that decision-makers can make informed choices when designing their ranking selection process. We propose transparent participation metrics to clarify the ranking process, by assessing the contribution of each parameter used in the ranking function in the creation of the final ranked outcome, using information about the ranking functions themselves, as well as observations of the underlying distributions of the parameter values involved in the ranking. To evaluate the outcome of the ranking process, we propose diversity and disparity metrics to measure how similar the selected objects are to each other, and to the underlying data distribution. We evaluate the behavior of our metrics on synthetic data, as well as on data and ranking functions on two real-world scenarios: high school admissions and decathlon scoring.

PVLDB Reference Format:
Abraham Gale and Amélie Marian. Explaining Monotonic Ranking Functions. PVLDB, 14(4): 640-652, 2021.
doi:10.14778/3436905.3436922

1 INTRODUCTION

Rankings are commonly used to make decisions and allocate resources in a wide variety of applications such as school admissions, job applications, public housing allocations, sport competition judging, and organ donation lists. Decision-making techniques resulting in rankings of objects using multiple criteria have been studied for centuries [18]. However, these techniques were traditionally developed with the decision-maker's interests and constraints in mind, and did not focus on transparency and explainability of the process for the objects/individuals being affected by the outcome of the rankings.

With today's widespread use of algorithms to make decisions in an information-based society, there has been a realization that the outcomes of these algorithms have significant societal impacts, and that the algorithm designers have a responsibility to address the ethical considerations that arise when applying algorithms to individuals, groups, or entities. This has been recognized by several research communities, such as Artificial Intelligence [14, 49], Machine Learning [17], and Data Management [46]. Without transparent and explainable processes, it is not possible to verify whether the outcomes satisfy ethical and fair constraints.

Traditionally, work on ranking algorithms and techniques has either assumed that the ranking function was given and satisfied some required properties, such as monotonicity [16], and considered the ranking function as an oracle, or has focused on designing complex query functions applicable to specific domains [18, 32]. Little attention has been given to making the ranking function itself transparent. In fact, many techniques preprocess the underlying data being ranked, typically via normalization, so that it has desirable properties for the ranking algorithm. The resulting transformation often murks the data and contributes to making the process opaque.

This paper focuses on providing metrics to enable the analysis of ranking functions and the relative impact of individual ranking metrics on the overall ranked results in order to understand the impact of the ranking process *a priori*, based on the ranking functions and data distribution. Our goal is to help decision-makers understand the behavior of their ranking functions, and to provide entities being ranked with some transparent and understandable explanation of the ranking process.

The paper makes the following contributions:

- The design of transparent and understandable participation metrics to clarify the ranking process, by assessing the contribution of each parameter used in the ranking function in the creation of the final ranked outcome, using information about the ranking functions themselves, as well as observations of the underlying distributions of the parameter values involved in the ranking. (Section 4)
- The design of metrics to measure the similarity, or diversity, within the objects selected as part of the ranking process, as well as the disparity of the selected objects with respect to the underlying data distribution. Our diversity and disparity metrics can be used both on parameters involved in the selection (ranking) and on other parameters of the data. We also discuss how this work can be included in and support wider efforts on fairness and diversity of rankings. (Section 5)
- An experimental evaluation of our metrics. We first illustrate the behavior of the metrics using a variety of synthetic data scenarios. Then, we show how our metrics can be used on two real data sets: (a) NYC student data and school admission functions, and (b) decathlon scoring in international competitions to evaluate both the behavior of the ranking

Proceedings of the VLDB Endowment, Vol. 14, No. 4 ISSN 2150-8097.
doi:10.14778/3436905.3436922

functions used for admission and assess whether they reflect the intention of the decision-makers, and to analyse the outcome of the ranking process in terms of diversity and disparity of the selected objects. (Section 6)

In the next section, we present motivations for our work and present a real-life application scenario involving public schools using ranking functions for admissions in NYC. We discuss related work in Section 3 and conclude in Section 7.

2 MOTIVATIONS

Ranking functions are used to make decisions in a wide variety of application domains: public school systems, college rankings [50], affordable housing [3], as well as in complex ranking processes generated from Machine Learning (e.g., search engine results [32], LambdaMart [8]). The prevalence of automated decisions systems has raised numerous questions from the public, and increasingly lawmakers are requiring public decisions systems to be transparent [27]

Consider the example of the NYC school admissions process. NYC middle- and high-school admissions use a matching algorithm [1] similar to the stable marriage algorithm designed by Gale-Shapley [21], and used by medical schools. A typical school matching process consists of three parts: the schools rank students based on some desired criteria (academic, geographic, demographic), the students list schools in order of preference, and the matching algorithm, handled by a clearinghouse and designed to optimize some notion of utility, produces the rankings. One common approach is to focus on strategy-proof techniques that optimize students' choices while guaranteeing stable matches. Several school-choice matching algorithms have been proposed [2]. Such matching algorithms have been well studied, but the literature assumes their input (students' and schools' ranked lists) as given. Some attention has been given to improve the quality of matches by providing better information to students so that their choice lists better reflect their preferences [4, 12], but to our knowledge, there has not been much focus on the way schools rank their students.

Transparency in such a scenario is critical. The NYC public school system has 1.1 million students, and every year, 160,000 rising middle- and high-schoolers are sorted into schools. Without a transparent and explainable process, families are losing trust in the system.

Even when made public, some fully detailed, published school admission rubrics (ranking functions) raise more questions than they answer. Table 1 shows the high school admission function of a NYC public school (School A).

In addition to the transparency requirement, ranking functions used for public policy must satisfy additional criteria to ensure equity, fairness, and other legal considerations. For instance, NY State Law requires that state scores may not be the "sole, primary or major factor" in admissions [45], which has been interpreted by schools and districts as meaning that at most half of the points (50%) in the school rubrics can be derived from state scores. Of course, a ranking function may abide by the 50% limit on the weights of state scores, while still using these scores as the major factor possibly unknowingly, because of the underlying data distribution. In addition, these rubrics are typically designed by non-experts, who do not realize the impact of the underlying data distribution. In this particular school example, the school is located in a NYC district [40] where 45% of the students have a test scores above 4, and 70% have test scores above 3. The coarseness of the point allocation to the state score category does not allow to differentiate among students, in effect using test scores as a filter. The weight given to a small number of absences and latenesses is also disproportionate to the likely intent of the school administrators. Furthermore, the school has around 100 seats, for 1,800 applicants; because this is a school district with many high-performing students, the school ended up admitting only students who scored 100 on the rubric, having to resort to a lottery among these top-ranked students.

CATEGORY		POINTS
Course grades		**35**
ELA, Math, Science, Social Studies (each)	A+,A (95-100)	8.75
	A- (90-94)	7.75
	B+,B (85-89)	6.75
	B- (80-84)	5
	Pass	3.5
	Fail	0
State Scores		**35**
ELA, Math (each)	4.01−4.5	17.5
	3.5−4.00	16
	3−3.49	14.5
	2.5−2.99	12
	2.0−2.49	8
	under 2.0	0
Attendance and Punctuality		**30**
Absences Latenesses (each)	0−2 days	15
	3−5 days	12
	6−8 days	9
	9−10 days	6
	11−15 days	2
	16+ days	0

Table 1: *Example School Admission Rubric (School A)*

Given these observations, it seems critical to design human-understandable ranking functions that take into account real-life constraints (e.g., fairness requirements, bounds on the use of some parameters) and that can shared with non-technical audiences so they know what to expect in the ranking process. We will investigate how to simplify complex ranking processes and analyze the behavior of these processes to create *explainable* ranking functions that address the needs of the decision makers.

Transparency and accountability should be required of all public decision systems. Without these, there cannot be fairness and equity. We propose metrics to allow for accountability of the ranking systems by making transparency an integrated part of the process.

3 RELATED WORK

Ranking functions have been widely studied in the literature. The Multi Criteria Decision Analysis community focuses on making decisions from a set of possibly conflicting criteria [18, 52]; common types of functions include weighted-sum and products [48]. These techniques are typically aimed at experts, and provide complex computation, often hidden in black-box algorithms, with little possibility of explanation. Often, the ranking functions are applied on normalized parameters, which allows the decision making system to better control for variations in the underlying parameters, but results in opaque decisions for the candidates. Normalization functions used in decision-making are based on statistical analysis of the parameters; the choice of the normalization function is domain-specific. Common such functions include z-score , vector, or logarithmic normalization [25, 48]. Most of these normalization functions were designed for and by expert users and assume some understanding of statistics and math. While this is a realistic assumption in some ranking scenarios –we can expect medical professionals who design and use organ wait lists to be fluent in statistics, it is not reasonable to expect in every case. Individuals who are concerned by public policy decisions that are based on rankings cannot be expected to know introductory-level statistics to understand, and therefore trust, the mechanisms that will assign them to schools, public housing, or will decide the amount of public support they are eligible for.

Ranking functions are widely used in Information Retrieval [32, 42]. More recently, the Information Retrieval community has focused on learning-to-rank approaches [29, 31]. However such techniques produce complex ranking functions, that are impossible to explain to a non-expert; for example, LambdaMART [8], a state-of-the-art learning-to-rank algorithm based on gradient boosted decision trees.

In the Data Management community, there has been a significant focus on the optimization of ranking (top-k) queries [7, 33], based on the seminal work by Fagin et al. [16]. A survey can be found in [23]. This work typically focuses on the efficiency of the ranking process, and assumes that the ranking function is already known and that it satisfies some monotonicity properties. Some of this work has looked at the impact of changes in the data distribution [10] or uncertainty in the ranking function [43]; however, these authors did not focus on the impact of ranking parameters on the ranking outcome.

Several measures have been proposed to compare the outcomes of ranking processes. The Spearman ρ [44], and Kendall τ [26] are the most commonly used metrics to compute rank correlation. More recently, distance measures have been proposed by the Database community [15]. These focus on comparing the outputs of ranking processes. In contrast, we focus on the behavior of the ranking functions before the ranking takes place, by analysing the impact of different data distributions on the ranking functions.

Recently, there has been a lot of discussion in the research community and in the media on the impact of algorithms in societal issues and on the inherent bias in many algorithms, including ranking algorithms. Recent work have looked at how to include fairness and diversity into ranking techniques [9, 41, 55] or in Machine Learning settings [35]. Our work is complementary to these approaches: by providing tools to explain ranking processes, we can design more ethical ranking functions.

Explainability and transparency have been at the forefront of many works in Artificial Intelligence (e.g., [13]) and Machine Learning (e.g., [53]). This has been driven in part by political regulations that call for "right to explanation" [22]. Work that aim to explain rankings have mostly focused on *a posteriori* explanations of the results. Most of these work focus on feature selection to explain the contribution of each individual features to the final ranking outcome, in a process similar to sensitivity analysis [11, 47]. In contrast, we focus on making the process and parameter importance transparent so that the information is shared *a priori*. The meaning of explainability and how it is understood by system designers, users, and stakeholders is still the subject of current interdisciplinary research, such as work in interpretability of AI [30]. We focus on capturing both an understanding of what the users want, and explaining the resulting functions to stakeholders in a trustworthy and informative way [34].

4 MEASURING THE CONTRIBUTION OF RANKING PARAMETERS

Many real-world ranking decisions are made using aggregation functions on multiple parameters. Decision-makers often assume that the weight of a parameter (or the number of points associated with the parameter) in the function are adequate proxies of the importance of the parameter in the final decision. We aim at defining a set of metrics that more accurately capture the true impact, or participation, of each parameter in the ranking decision. This impact depends on the ranking function design itself, but also, importantly, on the distribution of the underlying data, and the correlation between parameters, information that is often overlooked by decision-makers.

In [20], we introduced some preliminary metrics, disqualifying power and importance. In this paper, we revisit these metrics, discuss their limitations, and introduce new *participation* metrics that better measure the contributions of each parameter towards the top-k answer.

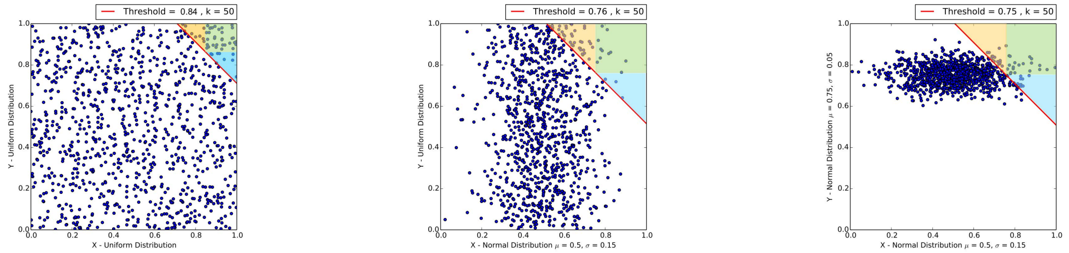
Any ranking decision is dependent on k, the number of selected objects. The relative contributions of each parameter is therefore dependent on this value, and our importance and participation functions are defined w.r.t. k. Note that in some applications k is fixed and known in advance (e.g., the top-10 participants to a local sport competition will qualify for the regional competition), in some others k is not known to the decision-makers in advance (e.g., in our school example, because the students are assigned to schools through a matching algorithm that includes their choices, schools do not know how far down their lists they will admit students). In our experimental evaluation (Section 6.2), we will investigate how parameter participation evolves as k varies.

Formally, we define a ranking function f over a set of P ranking parameters $p_1, ..., p_P$, over an object o as $f(o) = f(p_1, ..., p_P)$. Typically, a ranking process will select the k best objects, or the k objects with the highest $f(o)$ values its answer.

In this section, we focus on *monotonic* ranking functions; a monotonic ranking function f over a set of n P ranking parameters as any function $f(P)$ such that if $p_a \geq p_b$ then $f(p_a, p_2, ...p_n) \geq f(p_b, p_2...p_n)$. Monotonicity is a reasonable property of ranking functions [16], and is widely assumed to hold in the ranking literature, as it ensures that objects with lower scores in a parameter p_i cannot "leap-frog" an object with a higher p_i score, everything else being equal.

To illustrate our metrics, we consider as an example the following weighted-sum ranking function f over a set of P ranking parameters $p_1, ..., p_P$, with weights $W_1, ..., W_P$ such that $\sum_{i=1}^{P} W_i = 1$, over an object o as $f(o) = \sum_{i=1}^{P} W_i * p_i(o)$, where $p_i(o)$ is the value of parameter p_i for object o.

Figure 1 shows the behavior of a simple weighted-sum ranking function over two parameters values $X(o)$ and $Y(o)$ (denoted X and Y for simplicity), $f(o) = 0.5X + 0.5Y$, used to identify the top-50 objects out of 1,000 objects, depending on the underlying distributions of X and Y. We observe that the score of the top 50^{th}



(a) X **and** Y **both follow uniform distributions in** $[0, 1]$**.**

(b) X **is normally distributed** ($\mu = 0.5$,$\sigma = 0.15$)**,** Y **follows a uniform distribution in** $[0, 1]$**.**

(c) X **and** Y **are both normally distributed** ($\mu_X = 0.5$, $\sigma_X = 0.15$, $\mu_Y = 0.75$, $\sigma_Y = 0.05$)**.**

Figure 1: ***Top-k thresholds*** **(*score*** $= 0.5X + 0.5Y$**)** ***based on the underlying distribution of values (N=1000, k=50),*** X ***and*** Y ***are independent variables.***

object (defined as the threshold at 50, red line in Figure 1) varies depending on the underlying distributions of X and Y. This in turn has an impact on the minimum score required in for each dimension (parameter) for an object to qualify as being in the top-50, which we define as the floor value. This threshold will become the basis most of our proposed metrics, we define it as follows:

Definition 4.1. Threshold at k Given a ranking function f, over a set of P ranking parameters $p_1, ..., p_P$, applied to a set of objects O, we compute the threshold value T_k as the k^{th} highest $f(o)$ value for all objects $o \in O$.

This can also be understood as the lowest ranking score ($f(o)$) that an object o can have and still qualify for the top-k. For instance, if 1,000 objects are distributed uniformly in both X and Y as shown in Figure 1(a), the sum $0.5X + 0.5Y$ would follow a triangular distribution:

$$f_{(0.5X+0.5Y)}(x) = \begin{cases} x & 0 \leq x \leq 0.5 \\ 1 - x & 0.5 \leq x \leq 1 \end{cases}$$

From which we can trigonometrically estimate the value of the threshold at $k = 50$ (95^{th}*percentile*), T_{50}, as 0.84.

To account for cases where the output of the ranking function is not over the same domain as the ranking parameters $p_1, ..., p_P$, we define the parameter-wise threshold as:

Definition 4.2. Parameter-wise Threshold at k Given a ranking function f, over a set of P similarly scaled ranking parameters $p_1, ..., p_P$, applied to a set of objects O, We compute the parameter-wise threshold value T_{kp} as the maximum real value of each P_n of a hypothetical object O_h where each of the parameters of O_h have equal value and $f(O_h) = T_k$.

This definition is useful in many real-life scenarios, ranking from weighted-sum ranking functions where the sum of the weights is greater than 1; to point-based ranking functions, to more complex ranking functions.

For example consider a multiplicative function, if $x, y \in [0, 5]$ and $f(x, y) = x^2 \times y^3$, then $f(x, y) \in [0, 3125]$. Consider a scenario where and $T_k = 32$ then we compute T_{kp} as $32 = T_{kp}^2 \times T_{kp}^3$, $32 = T_{kp}^5$, $T_{kp} = 2$. Similarly, if $x, y \in [0, 1]$ and $f(x, y) = x + y^3$, $f(x, y) \in [0, 2]$. In a scenario where $T_k = 0.208$ then we compute T_{kp} as $0.208 = T_{kp}^3 + T_{kp}$, $T_{kp} = 0.2$ since 0.2 is the largest real solution to that equation.

As we describe in Section 6.1.3, the output of the scoring function for decathlon competitions is the sum of 10 inputs, so $T_{kp} = \frac{1}{10}T_k$.

Definition 4.3. Parameter Floor at k Given a threshold T_k a parameter p and a ranking function f, the floor at k of p, noted $floor_k(p)$, is the lowest value an object o' can have in p that would still allow for o' to qualify in the top-k assuming all the other values are maximized, that is for $f(o') \geq T_k$.

For instance, the floor at 50 for X if the objects are distributed uniformly in both X and Y as shown in Figure 1(a), would be:

$$floor_{50}(X) = \frac{T_{50} - W_Y}{W_X} = \frac{0.84 - 0.5}{0.5} = 0.68$$

which geometrically corresponds to the intersection between $f(o) = 0.5X + 0.5Y = T_{50} = 0.84$, and $Y = 1$.

Figures 1(a-c) show the threshold values for various underlying distributions of X and Y. The computed $floor_{50}(X)$ for the distributions of Figures 1(a-c) are 0.72, 0.52, and 0.5, respectively. For the examples of Figure 1, the values for $floor_{50}(Y)$ are the same as $W_X = W_Y$.

We can use the floor value to define the *disqualifying power* of each parameter of the scoring function.

Definition 4.4. Disqualifying power of a Parameter at k Given a parameter floor $floor_k(p)$ for parameter p, the disqualifying power of p at k, $DQ_k(p)$, represents the percentage of objects $o \in O$ for which the value of o for p, $p(o)$ is lower than $floor_k(p)$. Intuitively, $DQ_k(p)$ is the percentile rank of $floor_k(p)$ in p's distribution.

The disqualifying power can be computed from the data, if available *a priori*, or estimated from knowledge of the underlying distribution. For instance, in Figure 1(b), Y is uniformly distributed in $[0, 1]$ and $floor_{50}(Y) = 0.52$, the disqualifying factor of Y at 50, $DQ_{50}(Y)$, is then estimated to be $DQ_{50}(Y) = 0.52$. Similarly, in the same Figure 1(b), X follows a normal distribution ($\mu = 0.5$, $\sigma = 0.15$), from which we can estimate $DQ_{50}(X) = 0.5517$ (*z-value* $= (0.52 - \mu)/\sigma = (0.52 - 0.5)/0.15 = 0.13$). Figure 1(c) exhibits distributions that are not centered on the same values, which result in more variations in disqualifying power between X

and Y: $DQ_{50}(X) = 0.5517$ as X follows the same distribution as in Figure 1(b), but $DQ_{50}(Y) \approx 0$. As disqualifying power value of 0 means that the parameter is not enough, by itself to disqualify the object from the top-k. In the example of Figure 1(c), all of the values of $Y(o)$ are large enough for the objects to be part of the top-k. In that particular scenario, it is the value $X(o)$ that accounts for most of the top-k decision for object o, as a high X value compensates for even the lowest Y value in the data set.

The disqualifying power directly indicates the significance of each parameter discussed above by taking into account the distribution from which it is drawn. Intuitively, a parameter is more significant if it has more power to disqualify objects. However, as we decrease k/N or increase the number of parameters in the ranking function, we are less likely to observe positive disqualifying power values, limiting the usefulness of the metric. In addition, decision-makers need to be able to compare the usefulness, or impact, of each parameter *relative to each other*, as well as *relative to their importance in the final decision.*

To address these concerns, we define the participation of each parameter in the final top-k ranking. In [20], we presented a preliminary metric, called importance, to assess the contribution of the parameter to the final ranking.

Definition 4.5. Importance of a Parameter at k Given a ranking function f, over a set of P ranking parameters $p_1, ..., p_P$, applied to a set of objects $o \in O$, and a threshold value T_k, we compute $I_k(p)$, the Importance of a parameter p at k, as the percentage of objects in the top-k answers (i.e, with $f(o) \geq T_k$) such that the value $p(o) \geq T_k$. If we only have distributions and not values this can be expressed by the conditional probability $\Pr(p(o) \geq T_k \mid f(o) \geq T_k)$

Importance of a parameter p expresses the percentage of objects that dominate an idealized object o' that would be exactly on the threshold, with all parameter values equal to the threshold, for p. If p's value falls behind this object for many other objects in the top-k answer, it follows that objects are being selected as part of the top-k despite their low values for p. On the other hand, if values of p almost always exceed the value of p for o', we see that p is contributing to these objects' selections, making p an important parameter in the ranking.

For example, in Figure 1(b), only 11 of the top-50 objects have values higher than the threshold $T_{50} = 0.76$ for *both* X and Y (green region). The rest of the 39 objects in the top-50 are qualified because one of their values (X or Y) compensates for a lower value in the other parameter. For these distributions, most of the remaining objects qualify thanks to a high value of Y (36 objects, orange region), whereas only 3 objects qualify thanks to a high value of X (blue region). For these particular distributions of X and Y, we see that for $k = 50$, Y dominates the ranking, despite the underlying scoring function $f = 0.5X + 0.5Y$ giving the same importance to both X and Y. We can compute the importance of X and Y in Figure 1(b) for f, as $I_k(X) = (11+3)/50 = 0.28$, and $I_k(Y) = (11+36)/50 = 0.94$. In Figure 1(c), we can see that the relative importance of X and Y is more balanced, with 20 objects in the green region, 19 in the orange region, and 11 in the blue region, resulting in importance values: $I_{50}(X) = (20 + 11)/50 = 0.62$, and $I_{50}(Y) = (20 + 19)/50 = 0.78$. The independent uniform distributions of Figure 1(a) result in equal importance for X and Y $I_{50}(X) = I_{50}(Y) = (8 + 34)/50 = 0.84$ with 34 objects in the common green region and 8 objects each in the orange and blue regions.

The importance metric has some limitations. First, the sum of all parameter importance is often greater than one, making it more difficult for decision-makers to compare the contributions of each parameter. Furthermore, because an object can contribute to the importance of several parameters, if it has a high value for each of them, some top-k objects end up weighting more in the overall importance of all parameters than others. To address this last issue we define a new metric, participation, which divided the contribution of each object o in the top-k among the parameters for which its value $p(o)$ is greater than the threshold T_k:

Definition 4.6. Participation of a Parameter at k Given a ranking function f, over a set of P ranking parameters $p_1, ..., p_P$, applied to a set of objects $o \in O$, and a parameter-wise threshold value T_{kp}, let $S_k(o)$ be the set of parameters such that the value $p_i(o) \geq T^{kp}$ the participation of p at k is then defined as $A_k(p)$ as:

$$A_k(p_i) = \frac{1}{|O|} \sum_{o \in O} \begin{cases} \frac{1}{|S_k(o)|} & if p_i \in S_k(o) \\ 0 & otherwise \end{cases}$$

For example, in in Figure 1, objects in the green area have values higher then T_k for both X and Y. Their contribution to the participation of X and Y is then split among the two parameters ($S_k = 2$). We can compute the participation of X and Y in Figure 1(b) for f, as $A_k(X) = ((11/2) + 3)/50 = 0.17$, and $A_k(Y) = ((11/2) + 36)/50 = 0.83$. Participations for X and Y in Figures 1(a) and (c), can be similarly computed. For Figure 1(c), $A_k(X) = ((20/2) + 11)/50 = 0.42$, and $A_k(Y) = ((20/2) + 19)/50 = 0.58$. For Figure 1(a), $A_k(X) = ((34/2) + 8)/50 = 0.5$, and $A_k(Y) = ((34/2) + 8)/50 = 0.5$; note that the participation metric correctly identifies that both X and Y contribute equally to the top-k answer in Figure 1(a), but that Y contributes more, despite the weights of X and Y being the same, in Figures 1(b-c).

As we increase the weight of a parameter its participation increases as the threshold value and its slope changes. The participation captures the fact that different underlying distributions' contributions respond differently to weight increases. For example, the contribution of normal distributions respond much slower than that uniform ones. This allows participation to accurately represent the degree to which each parameter participates in the final decision.

However, one issue with the participation as defined above is that it assigns some contribution to parameters as long as that parameter value for a selected point is higher than the threshold. In some cases, this can happen by chance, or because of some correlation, for a parameter that is not or very trivially, involved in the ranking. Consider again the example of Figure 1(c). Imaging that there is a third parameter Z, distributed uniformly, involved in a very marginal way in the ranking, so that now our function $f = 0.49X + 0.49Y + 0.02Z$. Assume that this is not enough to impact the final result and that the top-50 selected objects, and T_k are the same as in Figure 1(c). Because Z is distributed uniformly, 25% of all objects have values above the threshold for Z. The participation of X and Y have then to be recomputed[1] as:$A_k(X) = (((11 * 0.25)/3) +$

[1] This is an approximated computation assuming 25% of the points in each region have values for Z above T_k, for illustrative purposes

$((11*0.75)/2)+((3*0.25)/2)+(3*0.75))/50 = 0.15$ to account for all sets of {X,Y,Z}, and $A_k(Y) = (((11*0.25)/3) + ((11*0.75)/2) + ((36*0.25)/2) + (36*0.75))/50 = 0.73$. The participation of Z is $A_k(Z) = (((11*0.25)/3)+((36*0.25)/2)+((3*0.25)/2))/50 = 0.12$ (no points qualifies thanks to Z alone). The participation of Z, which does not contribute directly to the selection ends up being almost as high as that of X because selected objects tend to also have a high value for Z.

For this reason, for ranking functions that include parameter weights, we introduce a weighted version of the participation metric: $WA_k(p)$ that accounts for the impact of the parameter p in the selection of the top-$kobjects$

Definition 4.7. Weighted Ranking Function A weighted ranking function $f(P : W)$ where $P : W$ is a list of ranking parameter, weight pairs $p_1 : W_1, ..., p_P : W_P$, has a well defined value for any given set $P : W$. Where P is a set of parameters and W is a corresponding set of weights.

Definition 4.8. Weighted Participation of a Parameter at k Given a weighted ranking function f, over a set of $P : W$ ranking parameter weight pairs applied to a set of objects $o \in O$, and a parameter-wise threshold value T_{kp}, let $S(o)$ be the set of parameters such that the value $p_i(o) \geq T_{kp}$, let $S(o) : W$ be the set of ranking parameter weight pairs such that the parameter is in $S(o)$ and has the maximum possible (or if that is not defined than dataset maximum) value and W is that parameter normal weight. Weighted Participation of a Parameter at k is then defined as:

$$A_k^W(p_i) = \frac{1}{|O|} \sum_{o \in O} \begin{cases} \frac{f(P_i, W_i)}{f(S(o):W)} & if p_i \in S_k(o) \\ 0 & otherwise \end{cases}$$

where $S_k^W(o) = \sum_{p_j \in S_k(o)} W_j$.

In our example above, the weighted participation of X, Y and Z are $WA_k(X) = (((11*0.25*0.49)/1) + ((11*0.75*0.49)/0.98) + (3*0.25*0.49/0.51) + (3*0.75))/50 = 0.17$, $WA_k(Y) = (((11*0.25*0.49)/1)+((11*0.75*0.49)/0.98)+((36*0.25*0.49)/0.51)+(36*0.75))/50 = 0.82$, $WA_k(Z) = (((11*0.25*0.02)/1) + ((36*0.25*0.02)/0.51) + ((3*0.25*0.02)/0.51))/50 = 0.01$. These value are very similar to that of the original two parameters scenario of Figure 1(c), and accurately identify that Z only plays a marginal role in the selection while Y has a much more important contribution despite having the same weight as X in the ranking function.

The weighted participation accurately measures the fact that parameters that are unweighted do not participate in the decision, while a parameter with a weight of 1 will get full participation of 1.

Note that the popular points functions, and other similar functions, can easily be converted to weighted-sum functions by considering the maximum points for each parameter as the weight of that parameter. For example, School A allocates 35 points for grades, 35 points for scores, and 30 points for attendance, for a total of 100. The corresponding weighted-sum function $0.35*grades+0.35*scores+0.3*attendance$. If a student receives a 3.0 on both their Math and ELA state tests, their points for that rubric would be 29 out of a maximum 35. If the same student has perfect grades and attendance, their score would be: $0.35*\frac{35}{35}+0.35*\frac{29}{35}+0.3*\frac{30}{30} = 0.35+0.29+0.30 = 0.96$ (96 points).

The participation metric can be used to modify the weights of the ranking function to match with the intended behavior of the decision maker. We presented heuristics to adjust the ranking functions weights in [20].

5 MEASURING THE SIMILARITY OF SELECTED OBJECTS

We now turn our focus on the analysis of the type of results that are produced by the ranking functions. In the previous section, we measured the contribution of each parameter towards the answer selection. We now discuss how to measure the similarity, or diversity, within the selected objects as well as the disparity of the selected answers with respect to the underlying object distribution.

The metrics presented in this section apply to the output of the ranking process. As such they do not depend on the type of ranking function used, and can be computed on the output of any ranking function (monotonic not non-monotonic) or selection function.

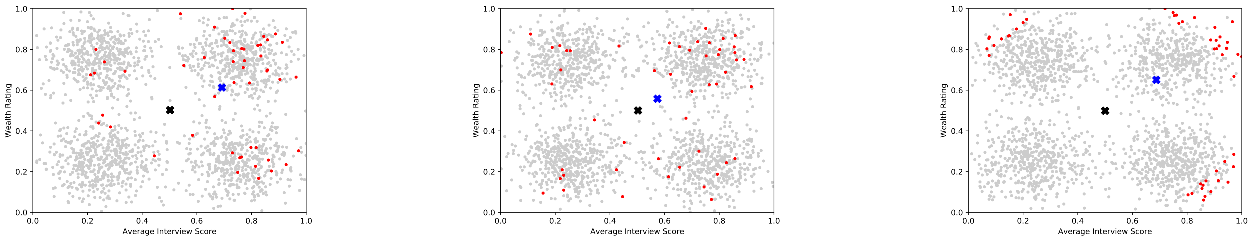
An interesting aspect of our metrics is that they can be used to measure the diversity and disparity of selected objects over parameters that were not involved in the ranking function f. So if a school selects students based on their Math test scores and grades, we can measure how the decision choices impacts the distribution of Math test scores (used in the ranking), the distribution of ELA test scores (not used in the ranking), or the ratio of boys to girls (not used in the ranking) in the set of selected students. Our metrics could then help decision-makers to assess the fairness and diversity of their ranking functions.

5.1 Diversity

Measuring the diversity among a group of selected objects has been extensively studied on recommendation systems and search results. In both domains, users are offered a list of objects/documents that best match their needs. A critical aspect is to provide enough variety so that objects/documents are not all similar. A typical example is that of a web search query for "jaguar," which should ideally return a variety of web sources on the animal, the car, the operating system, the sports team, and not have all results on one of these domains only. Maximizing the utility, or precision, of the results is therefore not always the best approach, as it tends to lead to homogeneous results. Instead, recommendation systems and search engines also focus on maximizing diversity within the set of recommendations [6]. Many work uses various pairwise metrics to measure diversity [28, 36], including euclidean or cosine distance [51, 58].

In a decision-making scenario, similar goals are often desirable. For example, a company using a ranking function for hiring decisions would be better served by selecting a mix of employees with diverse backgrounds rather than employees who all share the same expertise and knowledge.

In addition, the pairwise approaches to measure similarity that are discussed above work well in settings where the number of selected objects k is low; a recommendation system rarely recommends more than 10-15 objects [37]. However, it does not scale well to larger selections as the computation becomes inefficient, taking $O(k^2)$ time.



(a) Disparity vector = (0.19, 0.11)

(b) Disparity vector = (0.07, 0.06)

(c) Disparity vector = (0.19, 0.15)

Figure 2: ***Three different top-k selection scenarios. The red points are selected and the gray points are not selected. The black cross is the centroid of the entire distribution, the blue cross is the centroid of the selected set (red points).***

To alleviate this problem, we propose a diversity metric that is inspired from the popular k-means clustering algorithm. The goal of k-means is to minimise the square euclidian distance from the centroids in each cluster.

We use the ratio of the distance between the selected k objects and their centroid with that of the whole distribution distance to its centroid as a measure of the diversity of the selection. As we will show experimentally in Section 6, this distance approximates well how similar the top-k selected objects are to each other, across the selected dimensions. To be able to compare diversity of different selections, we use a normalization factor of $\frac{|O|}{\sum_{o \in O} ||Q_O - o||}$, where Q_O is the centroid of the whole distribution.

We then define diversity as the Normalized Centroid Distance, as follows:

Definition 5.1. Normalized Centroid Distance Given a set of objects O and a selection K of k objects in O, let $\vec{o_P}$ be the vector representing the objects values in a set of parameters P. Let $\vec{Q_S^P}$ be the centroid of a set of points S over a set of parameters P, $\vec{Q_S^P} = \frac{1}{|S|} \sum_{o \in S} \vec{o_P}$ we then define the Normalized Centroid Distance as:

$$C_P(O, K) = \frac{|O|}{|K|} \frac{\sum_{o \in K} ||\vec{Q_K^P} - \vec{o_P}||}{\sum_{o \in O} ||\vec{Q_O^P} - \vec{o_P}||}$$

Figure 2 shows selected candidates (in red) from a distribution of all candidates (in gray) plotted along two dimensions: wealth, representing the income of the candidates, and average interview score, representing the scores the candidates received in an admission interview. The black cross represents the centroid of the whole set of (gray and red) candidates, and the blue cross represents the centroid of the set of selected (red) candidates. Figure 2 shows three different selection scenarios.

The overall centroid distance in all three figures is (the average distance of all the grey and red points from the black cross) 0.37. Figure 2(a) shows a relatively typical case for weighted sum functions; the decision is biased toward higher values in both wealth and interview score. This leads the points to be concentrated in the top right quadrant and leading to a selected centroid distance (the average distance of the red points from the blue cross) of only 0.30 and a Normalized Centroid Distance $C_P(O, K)$ of 0.83. Figure 2 (b) shows selected points scattered throughout the distribution, resulting in a selected centroid distance of 0.36 and a Normalized Centroid Distance $C_P(O, K)$ of 0.99. Figure2 (c) shows an extreme case where selected points have high wealth, or interview scores, or both. This results in a larger selected centroid distance of 0.45 and a Normalized Centroid Distance $C_P(O, K)$ of 1.24.

A low Normalized Centroid Distance means that the selected objects are more similar to each other, over the set of parameters P than the objects in the overall distribution are. In Figure2 (a) we can see the the selected (red) objects are less spread out than the gray objects. When $C_P(O, K)$ is close to 1, the selected objects are as diverse, with respect the the parameters in P as the underlying distribution (Figure2 (b)). A value of $C_P(O, K)$ greater than 1, as in Figure2 (c), shows selection that have larger spreads than the set of candidates. A extreme case of a value of $C_P(O, K)$ equal to 0 can happen if all the selected objects are identical along the compared dimensions. For example, if a school only admits students with perfect scores.

5.2 Disparity

In addition to the diversity of results *within* the selected objects, we are interested in identifying how similar to the underlying distributions the selected objects are. This idea is related to "group fairness" in rankings, which has been explored in recent work [5, 9, 57]. These work focus on designing algorithm using a fairness criteria, maximizing some notion of utility while satisfying a minimum fairness requirement. Our approach differs from theirs in that we are interested in defining a metric that allows decision-makers, who are often not data or statistics experts, to quickly evaluate whether their ranking selection function results in some disparate impacts on some subsets of the parameters dimensions. Our disparity metric can they be used to correct, or improve, the ranking function, or to assess the impact of some modification (e.g., removing the SAT admission requirement) on the disparity of the selected objects.

We define disparity as the following vector:

Definition 5.2. Disparity Given a set of O objects and a selection K of k objects in o, Let $\vec{Q_O^P}$ be the centroid of a O over a set of parameters P as defined above, and let $\vec{Q_K^P}$ be the centroid of the K over the same set of parameters. We define the disparity $\vec{Q_D^P}$ as the $|p|$ dimensional disparity vector where $\vec{Q_D^P} \equiv \vec{Q_K^P} - \vec{Q_O^P}$.

The disparity represents the vector from the black cross to the blue cross in Figure 2. The norm of this vector yields a scalar measure of the size of this difference and has values of 0.22, 0.09, and 0.24 respectively for Figures 2(a),(b), and (c), showing that the selection of Figure 2(b) is closer to the underlying distribution than the other two selection. Each dimension of the vector measures the size and direction of the difference, or bias in the selection, for that dimension. The vector values for Figures 2(a),(b), and (c) are (0.19,0.11), (0.07,0.06), and (0.19,0.11) respectively, which means that all three selections select candidates with higher interview scores and wealth than would be expected from a random selection. The selection in Figure 2(c) has the highest wealth disparity.

Our disparity metric (as well as our other metrics) can be used in conjunction with recent work on fairness. For instance, Asudeh et. al. [5] design algorithms for fair rankings using a *fairness oracle*. A possibility is to use our disparity vectors, using thresholds for each dimensions, or for the norm, as the binary *fair/unfair* assessment. A strength of our disparity metric is that it can separate the disparity in each dimensions, including those not used in the actual ranking selection. Our metrics can also be included in ranking analysis systems such as [56].

6 EXPERIMENTAL EVALUATION

6.1 Experimental Settings

All experiments were performed using Python 3.6.

6.1.1 Synthetic Data Sets. We generated synthetic uncorrelated data distributions using Numpy. Unless otherwise noted, we performed our experiments on data drawn from the four following distributions: (1) p_1 follows a uniform distribution in $[0, 1]$; (2) p_2 follows a Normal distribution ($\mu = 0.5, \sigma = 0.15$); (3) p_3 follows a Normal distribution ($\mu = 0.5, \sigma = 0.05$); (4) p_4 follows a Normal distribution ($\mu = 0.75, \sigma = 0.05$).

We explore various weighted functions and selection sizes. Our default parameters are: *Size of the distribution N:* 10,000; *Number of selected objects k:* 500; *Number of selection parameters p:* 4, with underlying distribution as detailed above; *Ranking function f:* We use a weighted-sum scoring function with 4 equi-weighted parameters.

6.1.2 NYC High School Data. We evaluate our metrics using real student data from NYC high schools, which we received through a NYC Data Request [38], and for which we have secured IRB approval.

The data used in this paper consists of the grades, test scores, absences, and demographics of around 80,000 7th graders each for the 2016-2017 and 2017-2018 academic years. NYC high schools use the admission matching system described in Section 2 when students are in the 8th grade; the various parameters used for ranking students therefore are from their 7th grade report cards. We do not have information about latenesses, when necessary we used the number of absences as a proxy for lateness.

We consider the admission selection process that two real NYC high schools used for admission in the years 2017 and 2018 (on students from the data sets described above). School A uses the point-based ranking system, described earlier in this paper in Table 1. School B, also a real NYC school, uses a weighted-sum function $f = 0.55 * GPA + 0.45 * TestScores$, where GPA is the normalized average of the students' math, ELA, science, and social studies grades, and $TestScores$ is the normalized average of the math and ELA state test scores.

Both schools are located in a NYC district, which includes around 2,500 students, where students come from higher income families than the overall NYC student population and have typically higher grades and scores. Because students tend to apply to schools close to their homes, and because some NYC schools give some geographical priority to students (by considering them in different admission priority groups) we also report on our metrics at the district level, in addition to the city level. We apply our metrics on the whole set of students, as we do not have specific information as to who applied to the schools, how they ranked the school in the matching process, and which offers the students received. For conciseness, we only report on the 2016-2017 dataset when results for both datasets are similar.

6.1.3 Decathlon Data. We also evaluate our metrics on another real-data set over a different domain (sport). We retrieved data from several Decathlon sporting competitions [54]: the Olympics from 2000-2016, the Decastar from 2006-2019, the IAAF World Athletics Championships from 1991-2019, the Hypo-Meeting from 2006-2019, and the Multistars from 2010-2019. The dataset contains 1537 records, each containing the points that an athlete earned for each of the ten events in one competition, as well as their overall score and biographical information. For each athlete we were able to extract their birthdate, performance and country of origin.

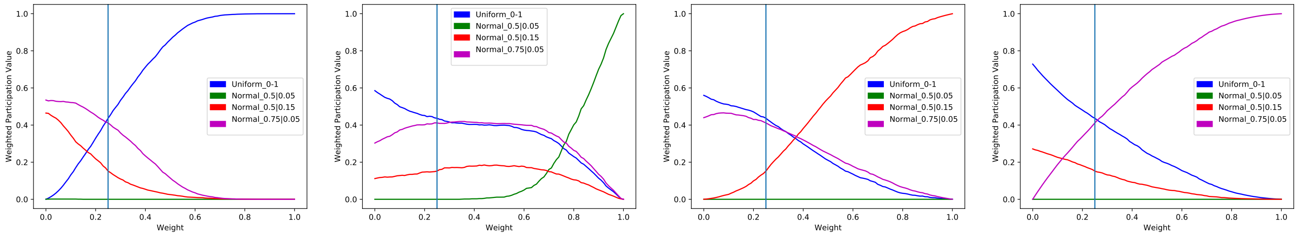
Decathlon is scored using a point ranking function where each of the ten events earns the athlete a number of points, with a calculation specific to each event [24]. The points from all ten events are then summed up to produce the final score of the athlete for that competition.

6.2 Experimental Results

6.2.1 Synthetic Data Experiments.

Weighted Participation. Figure 3 shows how our proposed weighted participation changes as we increase the weight of each parameter in the ranking function. In Figure 3(b) and (d) we see that low-means distributions do not participate in the selection unless they have a high weight. This is particularly true for p_2 which has a low mean and a small standard deviation. In contrast, both p_1 and p_4 quickly account for a high weighted participation as they have a relatively large number of high-scored values. When all parameters are weighted equally (vertical line, same values for all four subfigures), there is a significant difference in the participation of each parameter to the selection, based on its underlying distribution and the probability it produced values above the threshold T_k.

Figure 5 shows how the weighted participation metrics vary, for an equi-weighted ranking function over four parameters, each with a different underlying distribution. First, despite the fact that the parameters are weighted equally in the ranking function, their participation to the selection is very different. The parameters whose distributions follow a uniform distribution p_1, or a normal distribution with a high mean p_4, contribute more to the selection, as they produce more points with high values. In contrast, the parameters that follow a normal distribution with a lower mean, p_2 and p_3 contribute much less than their weight in f, especially when



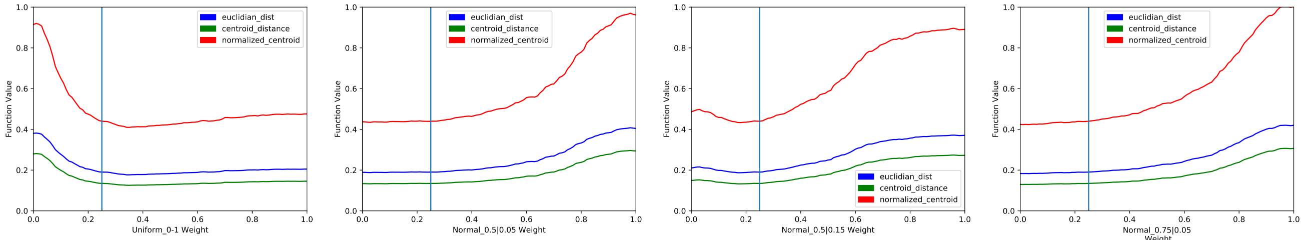
(a) Varying the weight of p_1 (uniform distribution).

(b) Varying the weight o f p_2 (Normal distribution, $\mu = 0.5, \sigma = 0.05$)

(c) Varying the weight of p_3 (Normal distribution, $\mu = 0.5, \sigma = 0.15$)

(d) Varying the weight of p_4 (Normal distribution, $\mu = 0.75, \sigma = 0.05$)

Figure 3: Weighted Participation of each parameter p as a function of the weight of each parameter P. For each plot, we vary the weight W_i of one parameter p_i, the other parameters weights are then computed as $\frac{1-W_i}{3}$.



(a) Varying the weight of p_1 (uniform distribution).

(b) Varying the weight of p_2 (Normal distribution, $\mu = 0.5, \sigma = 0.05$)

(c) Varying the weight of p_3 (Normal distribution, $\mu = 0.5, \sigma = 0.15$)

(d) Varying the weight of p_4 (Normal distribution, $\mu = 0.75, \sigma = 0.05$)

Figure 4: Distance Metrics: Euclidean Distance, Centroid Distance, and Normalized Centroid Distance (Diversity) as a function of the weight of each parameter P. For each plot, we vary the weight W_i of one parameter p_i, the other parameters weights are then computed as $\frac{1-W_i}{3}$.

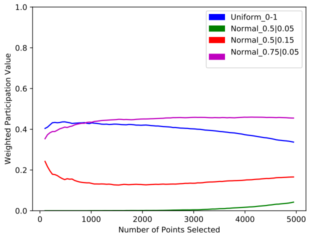


Figure 5: Weighted Participation as a function of k

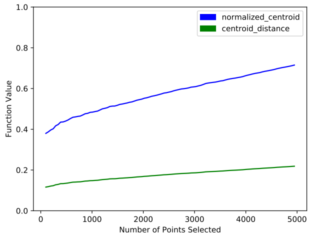
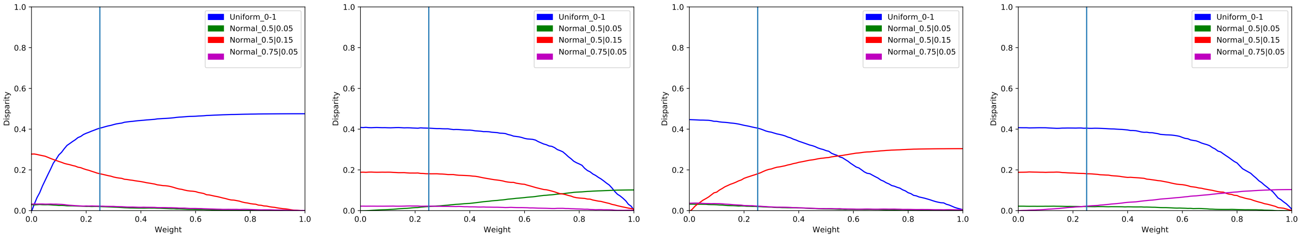


Figure 6: Centroid and Normalized Centroid Distance (Diversity) as a function of k

their standard deviation is low, as all the values for the parameter are similar, and relatively low. As we increase k the behavior vary slightly, depending on which distribution has more values above T_k, and would eventually converge to $A_N^W(p_i) = 0.25$ for all p_i when $k = N$.

Diversity. Figure 4 shows how our proposed Diversity metrics (Normalized Centroid Distance and Centroid Distance) vary as we change the weights of the the parameters compared to the existing Euclidean distance. We vary each parameter weight separately to show how each individually account for the diversity of the selection. Much of the diversity in the selected objects comes from the uniform distribution since it is more spread than other distributions. This explains why, across all the plots, diversity is highest when the weight of the uniform distribution approaches 0. Since all the parameters contribute to the diversity, the point indicated by the light blue line where they are all equal is close to a local minimum as the resulting selected objects tend to have high values in all parameters. We see that the Centroid distance closely approximate the pairwise Euclidean distance, justifying our choice to select it, rather than the $O(k^2)$ Euclidean distance. The normalization factor of our Normalized Centroid Distance scales it to be a more useful an expressive metric of diversity.

Figure 6 shows how the Normalized Centroid distance and Centroid distance vary as we raise k from 1% to 50% of N. By definition, when k is 100%, the normalized centroid distance is 1 since the two distributions become identical. This makes the normalized version of the metric more insightful as it expresses some comparison between the selected k objects diversity and the underlying distribution. It also shows nicely how selecting more objects increases diversity; an interesting result in applications where k is not bounded.



(a) Varying the weight of p_1 (Uniform distribution)

(b) Varying the weight of p_2 (Normal distribution, $\mu = 0.5, \sigma = 0.05$)

(c) Varying the weight of p_3 (Normal distribution, $\mu = 0.5, \sigma = 0.15$)

(d) Varying the weight of p_4 (Normal distribution, $\mu = 0.75, \sigma = 0.05$)

Figure 7: Disparity as a function of the weight of each parameter P. For each plot, we vary the weight W_i of one parameter p_i, the other parameters weights are then computed as $\frac{1-W_i}{3}$.

School A (Points)		Math Score	ELA Score	Grades	Absences
City N=60,865	Floor	17.5 points	17.5 points	35.0 points	30.0 points
k = 304 (0.5%)	Weighted Participation	0.175	0.175	0.0875 0.0875 0.0875 0.0875	0.300
City N=60,865	Floor	14.5 points	14.5 points	32.0 points	27.0 points
k = 3043 (5%)	Weighted Participation	0.181	0.157	0.076 0.081 0.082 0.082	0.340
District N=2,376	Floor	15.5 points	15.5 points	33.0 points	28.0 points
k = 238 (10%)	Weighted Participation	0.180	0.165	0.083 0.083 0.084 0.086	0.320

School B (Weighted Sum)		Math Score	ELA Score	Grades	T_k
City N=61,127	Floor	3.95	3.95	94.98	97.2375
k = 306 (0.5%)	Weighted Participation	0.190	0.096	0.714	
City N=61,127	Floor	3.28	3.28	88.89	93.8875
k = 3,056 (5%)	Weighted Participation	0.150	0.106	0.745	
District N=2,378	Floor	3.47	3.47	90.61	94.8375
k = 238 (10%)	Weighted Participation	0.204	0.090	0.706	

Table 2: ***Floor and Weighted Participation for the NYC high schools data***

Disparity. Figure 7 shows disparity on the parameters used for ranking on our synthetic dataset. Since all the parameters are positively weighted, the disparity never gets far bellow 0. In addition, in Figure 7(a) we see that as we increase the weight of the uniform distribution, we rapidly approach 0.5 disparity. This can easily be explained by the fact that the best objects in for p_1 (uniform distribution) are likely to be close to 1, and the mean of the entire distribution is 0.5. Figure 7 (b-d) are especially interesting, where we see the impact of the standard deviation on the disparity. When p_2 (resp. p_4) dominates the selection (has a high weight W_2), objects are selected in the top-5% of p_2, and tend to be very close to the mean of the distribution itself, leading to small disparity scores.

6.2.2 Real-data experiments: NYC School Admissions. We now use our metrics to analyse the behavior of school admission ranking functions in the NYC data set. The floor and weighted participation of the parameters used in the ranking functions of the two Schools A and B are given in Table 2 for three different scenarios per school: selecting the top 0.5% and top 5% of students citywide, and selecting the top 10% of students district-wide. Because the data set has a lot of missing information (e.g., some students leave the school system, some opt out of tests) we only report on the ranking of N students who have values for all the ranking function parameters. This number is slightly different for School A and School B as their ranking functions use a different set of parameters.

Table 2 shows that the ranking function of school B offers opportunities for students to compensate a (relatively) lower test score or grade with a higher grade or test score in another subject and still qualify to be selected, whereas the coarse point ranking of School A, and the fact that it penalizes for every less-than-excellent grade or score but does not reward extremely high grades and scores, does not allow for much variation. In fact, in the first scenario, only students with perfect scores on the point scale are selected.

While the participation of each parameters for School A is generally close to that of their relative points in the ranking function, there are interesting variations for School B, whose weighted participation of grades (above 70%) exceeds significantly the weight the decision-maker had assigned to grades (55%). For both schools, the Math scores account for more than the ELA scores, despite their weights (or points) being equal in the school rubrics.

The disparity vector for the ranking functions used by Schools A and B is given in Table 3 for the three scenarios highlighted above. Note that while the schools have around 100 available seats each, because the school admission is handled by a matching algorithm (Section 2), schools typically go deeper in their ranked list.

School A (Points)	GPA	Math Score	ELA Score	Low-Income	Sex	ELL	Special Ed	norm
City N=60,865; k = 304 (0.5%)	0.167	0.348	0.293	-0.328	0.145	-0.107	-0.195	0.643
City N=60,865; k = 3043 (5%)	0.154	0.332	0.274	-0.243	0.121	-0.106	-0.196	0.576
District N=2,376; k = 238 (10%)	0.084	0.177	0.165	-0.167	0.124	-0.036	-0.173	0.375
School B (Weighted Sum)	GPA	Math Score	ELA Score	Low-Income	Sex	ELL	Special Ed	norm
City N=61,127; k = 306 (0.5%)	0.177	0.377	0.329	-0.354	0.191	-0.107	-0.204	0.705
City N=61,127; k = 3,056(5%)	0.160	0.346	0.294	-0.313	0.124	-0.107	-0.196	0.629
District N=2,378; k = 238 (10%)	0.088	0.189	0.181	-0.222	0.179	-0.036	-0.174	0.435

Table 3: *Disparity vectors for the NYC high schools data*

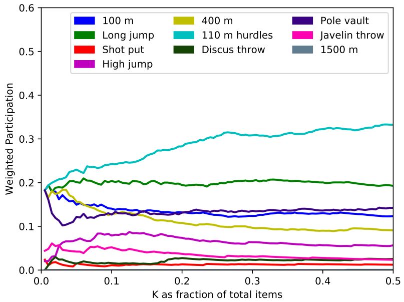


Figure 8: Participation of the ten Decathlon events on the score as a function of k.

A first observation is that both schools select students with higher GPA and scores than the student population as a whole (disparity score between 0.084 and 0.377). This is not surprising as both schools use grades and scores in their rankings. The disparity is more important when selecting from the citywide population, than when only students from the district are selected as that particular district has a higher ratio of high-performing students than the city.

When looking at demographics dimensions, which are not involved in the rankings, we can see some interesting disparities: both schools select fewer low-income, English-language learners, and special education students than are present in the candidate populations (school or district). The disparity is especially marked for low-income. This is in part because a large percentage of NYC public school students qualify as low income (around 70%). In contrast, while the schools select little to no ELL students in all scenarios, these represent approximately 11% of the city population, and 4% of the district population, limiting the maximum potential disparity for that dimension to these values. Interestingly, both schools select more girls than boys (in our settings, a positive disparity value means more girls are selected compared to their proportion in the underlying data).

Finally, the *norm* column shows the norm of the vector, a measure of the overall disparity over all the columns of Table 3. Both schools selections are more disparate citywide than district-wide, which can be explained by the fact that the district has more students with high test scores and grades than the city, the selection of students from the district is then more similar to the underlying set of candidates.

6.2.3 Real-data experiments: Decathlon Competitions. In Figure 8, we look at how each of the ten decathlon event participates in the final scores of the top-performers. We see that the 110m hurdles event has the largest impact, regardless of k, as it is the event with the widest range of performances. At the very top-level, the fast running events (110m hurdles, 400m, 100m and long jump) dominate the scoring. In contrast, shot put and 1500m have a very low participation, as athletes tend to have similar scores in these. Our results are consistent with recent analysis of decathlon scoring in the news, which showed that running events have a larger impact on the outcome of the competition [19].

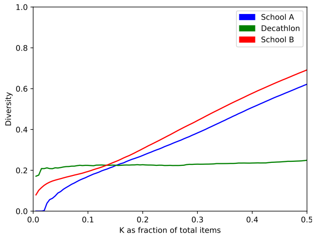


Figure 9: Diversity for NYC Schools and Decathlon datasets as a function of k.

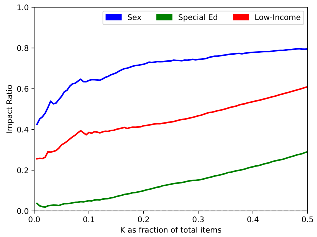
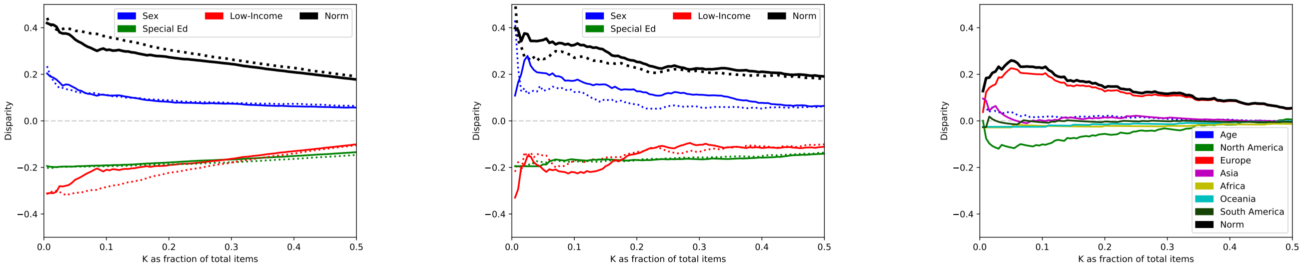


Figure 10: Impact ratio as a function of k (School A,city-wide)

Figures 9 and 11 show the diversity and disparity measures as a function of k for both the NYC Schools and Decathlon datasets. We report on the metrics behavior from selecting the best object, according to the ranking function, to selecting half of the dataset.

In Figure 9, we see how the diversity of the selected items evolves as the number of selected items increased. For this experiment, diversity is computed with respect to the parameters involved in the ranking function, so we are measuring how diverse selected students are in terms of grades and scores, and athletes are in terms of their individual events performances. By definition, our diversity metric converges to 1 as more of the data set is selected. Comparing the behavior of diversity for the two scenario is interesting: the diversity of athletes stays pretty low, even as we select half the data set, suggesting that most strong athletes have similar profiles (and that a few low-performing athletes have very different results). In contrast, while the diversity of selected students is low for low values of k (the best students will excel in every dimension), as more students are selected, we see a greater diversity of profiles.

Figure 11 shows the disparity of the selection, based on attributes not involved in the ranking, for both real-world scenarios. In Figure 11(a), we compare the disparity of School A and B admission functions on the sex, low-income status (poverty), and disability status of the selected students (citywide data). We see that both schools have similar disparities, which become less marked as more



(a) Disparity, measured on sex, students with disabilities, and poverty parameters on School A (solid line) and School B (dotted line) (city-wide data).

(b) Disparity, measured on sex, students with disabilities, and poverty parameters on School A in 2017-2018 (solid line) and 2016-2017 (dotted line) (district-wide data).

(c) Disparity, measured on age and geographic provenance on the Decathlon dataset

Figure 11: *Disparity (based as non-ranking parameters) for NYC Schools and Decathlon datasets as a function of k.*

students are selected. In Figure 11(b), we compare the disparity of School A between the two school years. As we noted before, School A tends to mostly select students from its district, so we only consider district students in this plot. We see that the disparity is similar to that of the citywide dataset, and while there are small year-to-year variations, the patterns of disparities stay the same. Finally. Figure 11(c) shows the disparity on the Decathlon data, where we looked at the geographical provenance and age of the participants. The data shows that high-performers are disproportionally European, with an age slightly older than average. North American athletes are under-represented. For all three plots of Figure 11, we report on the overall disparity (Norm).

6.2.4 Comparison with Existing Disparity Metrics. We compared our disparity metric with two other metrics used to identify disproportionality in the outcome of decision-making processes. The first metric is the popular impact ratio [59] shown in Figure 10 for the School A function on citywide data, as a function of k. Impact ratio is a simple ratio of the probability of an object of the *protected class* being selected to the probability of an object not in the *protected class* being selected. High values are desirable; an impact ratio of 1 means there is no disparity. In practice, and in legal applications, a threshold is set (e.g., 0.8) under which the decision is said to be biased. An issue with impact ratio, is that it treats all parameters equally, regardless of their incidence in the data set. Therefore, it tends to overstate the bias towards *protected classes* which have low numbers of objects, such as students with disabilities in our dataset. In contrast, our measure takes into account how many students are impacted by the disparity.

A more recent metric used in ML systems is the Normalized discounted KL-divergence. This metric uses a logarithmic discount measure to express the likelihood that a complete ranking is fair [55] KL-divergence is a very non-linear metric, something that is slightly more fair can have orders of magnitude less KL-divergence. While this is a very desirable trait for the domains for which the metric was designed, learning fair rankings, it makes it less human-readable, a number that diverges quickly will not allow them to accurately gauge how different two unfairness values are. For example, on our School A dataset for the entire city, the boy-girl disproportionality is about half that of the low-income disproportionality. This is reflected by a Disparity value which is around twice as large in absolute value (0.12 vs. -0.33 when k=304 or 0.12 vs -0.24 when k=3042) (Table 3). The KL-divergence for the same setting, however, is ten times as small for sex as it is for low-income (0.03 vs 0.23), which does not accurately identify the disparity when read by decision-makers and stakeholders.

6.2.5 Discussions with real users. We have been collaborating with one NYC School District as part of their NY State Integration Plan [39]. After numerous discussions with school administrators on their design process for their school ranking functions, a few things became clear: (1) administrators often have no idea of the distribution of students' grades and scores and make ad hoc decisions that seem reasonable to them ("I just give the same weight to Math and ELA"), (2) they would welcome a system that would show them the impact of their choices; we presented them with a protoype interface and the response was enthusiastic ("This is exactly what we have been asking for for years!"), (3) they are very aware of the disproportionate impacts of their decisions, but are at a loss as to how to address them. We plan to develop a full interface, using the metrics proposed in this paper to provide explanations to school administrators.

7 CONCLUSIONS

We proposed a set of metrics to explain the expected behaviors of ranking processes. Our goal is to make ranking decision-processes more transparent and explainable, both for decision-makers and for the entities being ranked, especially in the context of public policy decision systems. Our metrics provide information to decision-makers so that they can understand the impact of their ranking choices depending on the underlying distribution of data.

We performed experiments on synthetic data sets to study the behavior of our metrics, and analysed real ranking decision processes. We showed that the contribution of each parameters in the ranking selection does not always match the weight, or the number of points, the decision-maker had allocated to it. In addition, our proposed disparity metric can be used to measure bias of decisions on multiple data dimensions.

REFERENCES

[1] A. Abdulkadiroğlu, P. A. Pathak, and A. E. Roth. The new york city high school match. *American Economic Review*, 95(2):364–367, 2005.

[2] A. Abdulkadiroğlu and T. Sönmez. School choice: A mechanism design approach. *American economic review*, 93(3):729–747, 2003.

[3] Affordable housing online. https://affordablehousingonline.com/housing-help/What-Does-It-Mean-Preferences.

[4] K. F. Ajayi, W. H. Friedman, and A. M. Lucas. The importance of information targeting for school choice. *American Economic Review*, 107(5):638–43, 2017.

[5] A. Asudeh, H. Jagadish, J. Stoyanovich, and G. Das. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1259–1276, 2019.

[6] K. Bradley and B. Smyth. Improving recommendation diversity. In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*, pages 85–94. Citeseer, 2001.

[7] N. Bruno, L. Gravano, and A. Marian. Evaluating top-k queries over web-accessible databases. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 369–380. IEEE, 2002.

[8] C. J. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical report, June 2010.

[9] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[10] S. Chaudhuri, L. Gravano, and A. Marian. Optimizing top-k selection queries over multimedia repositories. *IEEE Transactions on Knowledge and Data Engineering*, 16(8):992–1009, 2004.

[11] S. B. Cohen, E. Ruppin, and G. Dror. Feature selection based on the shapley value. In *IJCAI*, volume 5, pages 665–670, 2005.

[12] S. R. Cohodes, S. Corcoran, J. Jennings, and C. Sattin-Bajaj. NYC High School Admissions Study, 2017. http://www.nychighschooladmissionstudy.com.

[13] M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg. Building explainable artificial intelligence systems. In *AAAI*, pages 1766–1773, 2006.

[14] V. Dignum. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, 20(1):1–3, Mar 2018.

[15] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on discrete mathematics*, 17(1):134–160, 2003.

[16] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *Journal of computer and system sciences*, 66(4):614–656, 2003.

[17] ACM Conference on Fairness, Accountability, and Transparency. https://facctconference.org/.

[18] J. Figueira, S. Greco, and M. Ehrgott. *Multiple Criteria Decision Analysis: State of the Art Surveys*. International Series in Operations Research & Management Science. Springer, 2005.

[19] FiveThirtyEight. The scoring for the decathlon and heptathlon favors running over throwing. https://fivethirtyeight.com/features/the-scoring-for-the-decathlon-and-heptathlon-favors-running-over-throwing/.

[20] A. Gale and A. Marian. Metrics for explainable ranking functions. In *Proceedings of the 2nd International Workshop on ExplainAble Recommendation and Search (EARS 2019)*, 2019.

[21] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[22] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[23] I. F. Ilyas, G. Beskales, and M. A. Soliman. A survey of top-k query processing techniques in relational database systems. *ACM Computing Surveys (CSUR)*, 40(4):11, 2008.

[24] J. Jablonsky. Multicriteria analysis of classification in athletic decathlon. *Multiple Criteria Decision Making*, 7:112–120, 2012.

[25] A. Jahan and K. L. Edwards. A state-of-the-art survey on the influence of normalization techniques in ranking: Improving the materials selection process in engineering design. *Materials & Design (1980-2015)*, 65:335–342, 2015.

[26] M. G. Kendall. Rank correlation methods. 1948.

[27] J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable algorithms. *U. Pa. L. Rev.*, 165:633, 2016.

[28] M. Kunaver and T. Požrl. Diversity in recommender systems–a survey. *Knowledge-Based Systems*, 123:154–162, 2017.

[29] H. Li. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10):1854–1862, 2011.

[30] Z. C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

[31] T.-Y. Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[32] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[33] A. Marian, N. Bruno, and L. Gravano. Evaluating top-k queries over web-accessible databases. *ACM Transactions on Database Systems (TODS)*, 29(2):319–362, 2004.

[34] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[35] M. Mitchell, D. Baker, N. Moorosi, E. Denton, B. Hutchinson, A. Hanna, T. Gebru, and J. Morgenstern. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 117–123, 2020.

[36] J. Möller, D. Trilling, N. Helberger, and B. van Es. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977, 2018.

[37] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686, 2014.

[38] NYC DOE. Doing research in or about new york city public schools. https://infohub.nyced.org/reports-and-policies/research/doing-research-in-new-york-city-public-schools.

[39] NY State Integration Plan NYCD2. https://www.district2nyc.org/nysip.html.

[40] N. OpenData. Nyc opendata - education. https://data.cityofnewyork.us/browse?category=Education.

[41] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791. ACM, 2008.

[42] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[43] M. A. Soliman, I. F. Ilyas, D. Martinenghi, and M. Tagliasacchi. Ranking with uncertain scoring functions: semantics and sensitivity measures. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 805–816. ACM, 2011.

[44] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[45] N. State Senate. Assembly bill a8556d. https://www.nysenate.gov//legislation/bills/2013/A8556.

[46] J. Stoyanovich, S. Abiteboul, and G. Miklau. Data, responsibly: Fairness, neutrality and transparency in data analysis. In *International Conference on Extending Database Technology*, 2016.

[47] M. ter Hoeve, A. Schuth, D. Odijk, and M. de Rijke. Faithfully explaining rankings in a news recommender system. *arXiv preprint arXiv:1805.05447*, 2018.

[48] C. Tofallis. Add or multiply? a tutorial on ranking and choosing with multiple criteria. *INFORMS Transactions on Education*, 14(3):109–119, 2014.

[49] J. Torresen. A review of future and ethical perspectives of robotics and ai. *Frontiers in Robotics and AI*, 4:75, 2018.

[50] US News College Rankings. https://www.usnews.com/best-colleges.

[51] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116, 2011.

[52] M. Velasquez and P. T. Hester. An analysis of multi-criteria decision making methods. *International Journal of Operations Research*, 10(2):56–66, 2013.

[53] L. S. Whitmore, A. George, and C. M. Hudson. Explicating feature contribution using random forest proximity distances. *arXiv preprint arXiv:1807.06572*, 2018.

[54] World athletics. https://www.worldathletics.org/.

[55] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, page 22. ACM, 2017.

[56] K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, H. Jagadish, and G. Miklau. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1773–1776. ACM, 2018.

[57] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017.

[58] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, and Z. Li. Drn: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, pages 167–176, 2018.

[59] I. Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, 2017.