FINS Auditing Framework: Group Fairness for Subset Selections

Kathleen Cachel
Worcester Polytechnic Institute
Worcester, MA, USA
kcachel@wpi.edu

Elke Rundensteiner
Worcester Polytechnic Institute
Worcester, MA, USA
rundenst@wpi.edu

ABSTRACT

Subset selection is an integral component of AI systems that is increasingly affecting people's livelihoods in applications ranging from hiring, healthcare, education, to financial decisions. Subset selections powered by AI-based methods include top-k analytics, data summarization, clustering, and multi-winner voting. While group fairness auditing tools have been proposed for classification systems, these state-of-the-art tools are not directly applicable to measuring and conceptualizing fairness in selected subsets. In this work, we introduce the *first comprehensive auditing framework, FINS,* to support stakeholders in interpretably quantifying group fairness across a diverse range of subset-specific fairness concerns. FINS offers a family of novel measures that provide a flexible means to audit group fairness for fairness goals ranging from item-based, score-based, and a combination thereof. FINS provides one unified easy-to-understand interpretation across these different fairness problems. Further, we develop guidelines through the FINS Fair Subset Chart, that supports auditors in determining which measures are relevant to their problem context and fairness objectives. We provide a comprehensive mapping between each fairness measure and the belief system (i.e., worldview) that is encoded within its measurement of fairness. Lastly, we demonstrate the interpretability and efficacy of FINS in supporting the identification of real bias with case studies using AirBnB listings and voter records.

CCS CONCEPTS

• **Information systems → Data analytics**.

KEYWORDS

subset selection, algorithmic fairness, machine learning fairness

ACM Reference Format:
Kathleen Cachel and Elke Rundensteiner. 2022. FINS Auditing Framework: Group Fairness for Subset Selections. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3514094.3534160

1 INTRODUCTION

The task of selecting a subset of items (i.e., individuals or objects) is integral to AI-enabled decision-making systems including problem classes such as shortlisting items, returning results for a top-k query, data summarization, clustering, multi-winner voting and more. Subsets are selected to determine who is interviewed for the job [21], demarcate congressional districts [52], determine who's products are featured in online marketplaces [58], and who is admitted to a clinical trial [60]. Thus it is imperative we develop solutions to audit bias that may be created or perpetuated in the selection of a specified subset. *Fairness algorithmic auditing* refers to stakeholders investigating closed-box AI systems for fairness issues. A critical open problem for algorithmic auditing is the design of measures that interpretably quantify fairness for a chosen subset.

Motivation: Auditing for Fairness. Subset selection picks or demarcates a set of items as part of a decision-making process. This could be a hiring committee creating a shortlist of candidates to interview or a screening portal displaying the best k applicants based on their calculated company fit score [20]. Yet, recent reporting [22, 49] has revealed that increasingly popular AI-powered recruiting tools often do not select women candidates for recruiter shortlists due to gaps in employment; a thorny issue that has only grown more devastating in light of the COVID-19 pandemic [45]. Likewise, similar instances of algorithmic hiring discrimination have occurred with veterans [22, 28] and formerly-incarcerated individuals [22, 28]. Additionally, subset selection tasks such as returning image results (data summarization), placement of community-based COVID-19 testing locations (centroid-clustering), and electing government representatives (multi-winner voting) have been shown to exhibit unfairness towards gender and racial groups [34, 53, 54]. While the development of AI-powered applications for consequential domains is increasing, *no methodology for evaluating the outcomes of subset selections with respect to fairness exists to date.*

State-of-the-Art. The algorithmic bias and discrimination literature has two general conceptualizations of fairness: (1.) *individual fairness* defined as treating similar individuals similarly [23], and (2.) *group fairness* stating that groups be treated similarly [50]. The primary focus of group fairness auditing methodologies has been on classification systems [5, 7, 18, 24, 31, 31, 36, 41, 51, 56, 61]. Likewise, guidance on choosing fairness notions from a problem setting and values-based perspective has solely focused on fairness in classification [32, 43, 47].

For subset selection problems, no auditing framework for different fairness issues currently exists. Further, we lack any subset-specific guidance rationalizing the applicability of different fairness objectives. Recent research aiding decision-makers and auditors is limited to studies analyzing the effect of a bias intervention known as the Rooney Rule [13, 37] and metrics for the quantification of the social notions of diversity and inclusion [46]. The Rooney Rule is an intervention which enforces that the subset has at least one member of an under-privileged group. Thus, from the auditing perspective, no tools exist to support the detection of fairness issues

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534160

in subset selection. *This means that to date there are no avenues for holding unfair subset selectors accountable.*

Challenges. The aforementioned auditing toolkits and measures are not applicable to subset selection. Fairness defined for classification assumes access to ground-truth information (i.e., the class assigned to an object) [5, 7, 18, 24, 31, 31, 36, 41, 51, 56, 61]. Subset selection does not have an equivalent notion of ground-truth. Conceptually, fair top-k ranking measures [29, 63] are closer to the problem of subset selection. However, they focus on auditing the *ordering* of items, whereas in subsets the *selection* is evaluated. Further, set-based fairness contains additional alternate objectives such as selecting a subset with a fair scoring of items. Corresponding auditing methodologies must encompass a wide range of fairness notions including those that consider fairness based on score.

A second challenge in the design of a subset-specific fair auditing methodology is to create one unified conceptual framework that can quantify the existing fairness objectives in subset selection (namely, equal and proportional presence of items) along with many additional novel fairness notions. The literature on fairness has prompted a flurry of research designing fair algorithms for subset selection tasks such as top-k queries [44, 48, 59, 62], clustering [1, 3, 4, 6, 8, 15, 16, 16, 30, 39, 42, 57], data summarization [12, 33, 38], and multi-winner voting [10, 14, 40]. However, these works solve one critical problem, namely, the creation of methods to achieve exact equal or proportional representation of items or their scores. The few subset-scoring strategies for equal or proportional group-representation employ disparate approaches making them challenging to interpret across fairness problems: risk difference [44], KL-divergence [12], and the Gini index [14]. Without careful design, numeric measures for diverse fairness concerns may have too distinct ranges or interpretations; thus inhibiting meaningful comparisons of different fairness problems.

Proposed Approach. In this work, we introduce a framework for auditing subset selections for group fairness, called FINS (short for Fairness IN Subset selection). FINS is the first auditing tool designed to quantify critical variants of group fairness in subset selection tasks spanning problems from top-k queries, multi-winner voting, data summarization, to clustering. FINS is agnostic both to the number of groups defined by the protected attribute and to the choice and cardinality of protected attributes; working on single and intersectional attributes [19] alike. Our FINS methodology is strategically designed to achieve two critical properties. (1.) To quantify fairness within a compact range [0,1] whereby a value of 0 represents maximum unfairness and 1 indicates the fairness notion is perfectly satisfied. Further, the value itself represents the positive outcome received by the least favored group as a proportion of the most favored group's positive outcome. (2.) To apply this easy-to-understand interpretation and metric formulation to a diverse set of fairness notions for a broad class of subset selection problems. In this way, we empower auditors to quickly grasp and easily compare different fairness problems in the audited subset.

FINS designs eleven group-based fairness measures, collectively representing three very different conceptual group fairness goals in subsets. We categorize the different fairness goals as *score-based*, *item-based*, and *item-based while dependent on item score*. We refer to these three goals as flavors of fairness. Further, we are the first to propose and quantify several fair subset notions; including the class of fairness notions that are item-based and dependent on item score in the context of subset selection. These metrics introduce fairness conceptualizations to the subset selection task aligned with Calibration [51] and Equality of Opportunity [31] from seminal work in fair classification.

Effectively utilizing FINS in practice depends on correctly linking the auditing goal with a relevant FINS metric(s). To tackle this, we provide guidance via the design of the Fair Subset Chart (FSC for short); a decision diagram, which guides auditors in selecting a metric(s) based on how the auditors wish to conceptualize fairness. The FSC is intentionally sensitive to the auditor's assumptions about the data and the task they audit. Utilizing the two worldviews (i.e., WAE and WYSIWYG; axiomatic belief systems) proposed in the seminal work by Friedler et al. [26], we provide the first discussion of moral belief systems in fair subset selection along with our proposed mapping of FINS fairness notions to worldviews. Further, we capture this information in the FSC. We then introduce the FSC as a tool for determining the worldview of a preconceived fairness audit. This allows for the determination of worldview in settings where auditing practitioners may not be familiar with the conceptual worldview framework created by Friedler et al. [26].

Finally, we demonstrate the broad applicability and the ease-of-use of our FINS framework along with its ability to diagnose bias with real-world case studies on AirBnB property listings and North Carolina voting records. Our contributions include:

(1.) We define a comprehensive set of fairness measures and conceptual notions based on items, scores, and a combined approach applicable to a diverse range of subset problems.
(2.) We formulate these diverse FINS measures via a unified approach, offering auditors one unified easy-to-understand interpretation.
(3.) We present the FINS Fair Subset Chart as a concise reference providing guidance to auditors for selecting a fairness measure(s) appropriate for their task. The chart presents worldviews and values associated with each measurement of fairness.
(4.) We package our FINS framework into an open-source package of diagnostic tools to further research as well as provide a valuable resource to practitioners.

FINS scope. FINS is designed to evaluate the fair selection of a subset of items. For fairness measures based on the presence of different groups in the subset see Section 4. For fairness considering groups along with a score (utility judgement) associated with items please see Section 5. For fairness solely based on the relative scoring of groups in the subset see Section 6. FINS assumes that once a selection of items is made, all the selected items are un-ordered in the result set. That is, FINS audits the (binary) choice of items (and not their rankings). If the subset needs to be ordered, fair ranking metrics or algorithms could be added as a post-processing step.

2 BACKGROUND AND NOTATION

Subset selection is a ubiquitous task that forms the basis of many AI systems and human-AI decision-making processes. We focus our work on auditing the broad problem of selecting a subset of items (i.e, people, objects or entities) from a larger pool of items. In

Symbol	Representation
X	Pool of items: $x_i ....x_{\|X\|}$
S	Audited subset of items
$G_{p:i}$	Group (i.e, set) of items with value i in protected attribute p
$X_{a:v}$	Set of items with value v in attribute a
s_{x_i}	Score of item x_i
$X_{l<s_{x_i}<u}$	Set of items whose score s_{x_i} is > l and < u

Table 1: Key Notation

particular, we formalize how a subset can be measured for various group fairness issues.

To conduct fairness analysis of a selected subset $S = x_1, x_2, ...x_{|S|}$ against the larger pool of items X, each item x_i is associated with a set of attributes $\mathcal{A}$. One or more of those attributes are categorical **protected attributes** $p^i \in \mathcal{P}$ such as gender, race, nationality, or a combination thereof. We refer to p as the protected attribute chosen by the application, regardless of if it is a single attribute (`race`) or combination (`race` and `gender`). For each possible value j of the protected attribute p, there is a **group** $G_{p:j}$ composed of items in X that have the same value j for the protected attribute p. For instance, $G_{gen:wom.}$ is the group of all people items that have the value `woman` for the `gender` protected attribute. If the application has multiple protected attributes, we recommend and have designed FINS to support, auditing each attribute independently as well as their combinations for intersectional fairness [19].

For fairness audits that compare the choice of subset S to the larger pool of items, pool X (with $|X| >> |S|$) is assumed to be given. Likewise, for audits that want to consider a score associated with each item, we denote this score of x_i as s_{x_i}. The key notation used throughout this work is summarized in Table 1.

3 FINS DESIGN CHOICES

Group fairness is concerned with the fair treatment of groups, yet "fair" can be intuited in many ways. We propose a conceptual framework that classifies group fairness measures based on the fairness goal of the audited task. This allows us to design FINS as a generally applicable subset auditing framework that unifies fundamentally different group fairness objectives within one overarching approach. We categorize these different fairness objectives as representing different *flavors of fairness* for subsets. Our framework designs families of measures for the following flavors of fairness:

(1.) *Item-based - i.e, selecting a subset with a fair presence (number of items) per group.* The positive outcome is "presence" (or inclusion) in the subset. This encompasses audited problems where items may not have scores associated with them (e.g., some instances of multi-winner voting or clustering) and problems in which fairness is desired to be score blind. In short, they answer the question "Are groups fairly present in the subset?".

(2.) *Item-based dependent on score - i.e., selecting a subset with a fair presence (number of items) per group while considering the scores associated with items.* Here the positive outcome is also presence in the subset. This encompasses audited problems where items have an associated score typically conceptualized as "relevance" or "utility" (e.g., a top-k query problem or most instances of multi-winner voting). These measures help auditors answer questions such as "Are groups fairly present in the subset when considering an extra constraint on item scores?"

(3.) *Score-based - i.e., selecting a subset with a fair scoring of items per group.* Here the positive outcome is derived from the score each item in the subset has or receives. This includes audited subset problems where the score is something tangible such as a resource or distance (e.g., some instances of clustering or subsetting geographic regions). These measures answer "Are groups fairly scored in the subset?"

With the fairness flavors in place, we make three strategic design choices in the FINS framework:

- We propose two auditing entities for every fairness notion. A *group-based metric* at the granularity of a group that is a building block for the corresponding *fairness measure* which is quantified across all groups.
- Every fairness measure is formulated as a ratio between the smallest and largest corresponding group-based metrics. The value itself represents what *proportion of the most favored group's positive outcome is received by the least favored group.*
- We assure all measures are *easy to explain and human-readable.* Each fairness measure range from 0 to 1 - with the worst value being 0 and 1 the best value indicating the fairness notion is perfectly satisfied.

FINS creates *one unified interpretation that enables auditors to quickly grasp and trade-off different fairness concerns* in a subset and use the same toolkit to audit different subset problem types.

4 ITEM-BASED FAIRNESS MEASURES

We now introduce metrics that capture the fairness flavor of item-based fairness objectives. These notions are only based on the presence of each group in the subset. Each fairness notion proposed contains two corresponding metrics at the granularity of the group and at the granularity of the protected attribute.

4.1 Proposed Statistical Parity and Balance Measures

Our first two measures, *S:Parity* and *S:Balance* quantify how well a specified subset achieves proportional presence and equal presence of groups respectively. These are pre-existing fairness concepts in subset-selection, which we formulate via the FINS design. Neither metric considers scores of items, making them generally applicable to all subset selection tasks.

***S:Parity* Measure**. Achieving a proportional presence or representation of groups in the specified protected attribute p is akin to satisfying statistical parity [50]. A requirement stipulating items receive a proportional share of the positive outcome regardless of their group membership in a protected attribute. *S:Parity* quantifies statistical parity for the attribute - as opposed to having a user compare every group's selection rate directly.

More precisely, we quantify *statistical parity* by comparing the highest group selection rate with the lowest group selection rate. The group-based metric for *S:Parity* is the **selection rate** *SelectRt* of a group $G_{p:j}$ defined as:

$$SelectRt(G_{p:j}, S) = |S \cap G_{p:j}|/|G_{p:j}| \quad (1)$$

Using $SelectRt$, we design *S:Parity* as:

$$S\text{:}Parity(S,p) = \frac{\min SelectRt(G_{p:j},S)}{\max SelectRt(G_{p:i},S)}, \forall G_{p:j}, G_{p:i} \quad (2)$$

When *S:Parity* equals 1, then all groups have the same selection rate. When *S:Parity* < 1, then the measure quantifies the selection rate of the least favored group as a proportion of the selection rate of the most favored group.

***S:Parity* to Audit for Disparate Impact.** If a subset has a low *S:Parity* value, it is critical to consider if disparate impact has occurred. Disparate impact is legal theory used in the United States to determine if *unintended discrimination* has occurred [2]. Disparate impact arises when a process has resulted in drastically different outcomes for different groups, even if no information about the protected attributes of individuals is known [2]. The measurement of disparate impact is typically operationalized through the "four-fifths" or "80%" rule championed by US Equal Employment Opportunity Commission (EEOC). The "80%" rule states that an unprivileged group must receive a proportion of the positive outcome that is at least 80% of the proportion received by the most privileged group [17]. Thus, if *S:Parity* < 0.80, then disparate impact is present.

***S:Balance* Measure**. The next proposed measure is designed to quantify how well a subset achieves *equal presence of all groups*. For example, having a hiring committee comprised of an equal number of members from all roles. In contrast to statistical parity, balance is unaffected by the total number of items per group in the larger pool of items X. We define *S:Balance* based on the group treatment measure **group proportion** $PropOfS$ of the subset S, which is:

$$PropOfS(G_{p:j},S) = |S \cap G_{p:j}|/|S|. \quad (3)$$

Utilizing $PropOfS$, we define the *S:Balance* measure as:

$$S\text{:}Balance(S,p) = \frac{\min\{PropOfS(G_{p:j},S)\}}{\max\{PropOfS(G_{p:i},S)\}}, \forall G_{p:j}, G_{p:i} \quad (4)$$

When *S:Balance* equals 1, then all groups have an equal presence in the subset. However, when *S:Balance* < 1, then the measure quantifies the ratio between $PropOfS$ of the least favored group and $PropOfS$ of the most favored group. This offers the interpretation of what percentage of subset spots held by the most favored group are held by the least favored group.

4.2 Conditioning *S:Parity* and *S:Balance*

S:Parity and *S:Balance* quantify a fair presence of groups in the subset. Inspired by conditional statistical parity in fair classification which allows for including "legitimate factors" in the prediction [18]; we design simple to understand a variants of *S:Parity* and *S:Balance* called *S:Conditioned Parity* and *S:Conditioned Balance* respectively. These metrics provide auditors with the ability to condition fairness on on an additional characteristic (i.e., a value in an attribute that is not the specified protected attribute).

Consider the task of selecting a gender-balanced administrative committee at a university. *S:Balance* measures if the committee achieves gender balance. However, the people on the committee comprise both students and faculty. *S:Conditioned Balance* (and its proportional sibling *S:Conditioned Parity*) is our proposed innovation to facilitate auditing gender balance for students and faculty while not decreasing (making worse) the fairness measure due to there being more faculty on the committee than students.

S:Conditioned Parity. *S:Conditioned Parity* measures statistical parity conditional on group members sharing an attribute value $a : v$ in addition to their group defined by protected attribute p. The group based metric for *S:Conditioned Parity* is the **conditioned selection rate** $CSelectRt$ of a group $G_{p:j}$ as:

$$CSelectRt(G_{p:j},S,X,a,v) = \frac{|S \cap G_{p:j} \cap X_{a:v}|}{|G_{p:j} \cap X_{a:v}|} \quad (5)$$

Then using $CSelectRt$, we design *S:Conditioned Balance* as:

$$S\text{:}Conditioned\ Parity(S,X,p,a,v) = \frac{\min CSelectRt(G_{p:j},S,X,a,v)}{\max CSelectRt(G_{p:j},S,X,a,v)}, \forall G_{p:j}, G_{p:i} \quad (6)$$

The value iteself is the conditioned selection rate of the least favored group as a proportion of the most favored group's conditioned selection rate. If the conditioned attribute value is `employee`: `previous_inter` then *S:Conditioned Parity* audits if all groups are proportionally present in the subsets when constrained to interns.

S:Conditioned Balance. *S:Conditioned Balance* measures the fairness notion that the subset should contain an equal number of items from each group that share a value v in an additional attribute a. The group-based metrics for *S:Conditioned Balance* is the **conditioned group proportion** $CPropOfS$ of the subset:

$$CPropOfS(G_{p:j},S,p,a,v) = |S \cap G_{p:j} \cap X_{a:v}|/|S| \quad (7)$$

Utilizing $CPropOfS$, we define *S:Conditioned Balance* as:

$$S\text{:}Conditioned\ Balance(p,a=v) = \frac{\min CPropOfS(G_{p:j},S,a,v)}{\max CPropOfS(G_{p:i},S,a,v))}, \forall G_{p:j}, G_{p:i} \in X \quad (8)$$

S:Conditioned Balance quantifies, when conditioning on attribute value pair $a : v$, what percentage of spots held by the most favored group is held by the least favored group.

5 ITEM-BASED AND SCORE DEPENDENT FAIRNESS MEASURES

The next measures consider an item's score into the assessment of fairness. While we continue to cast the positive outcome as presence in the subset, these measures allow auditors to control for various score-based factors. Continuing with our methodology, each proposed fairness notion contains one metric at the granularity of the group and another at the granularity of the protected attribute.

5.1 Proposed Qualified Parity and Balance Measures

The new measures *S:Qualified Parity* and *S:Qualified Balance* respectively quantify the proportional and equal presence of groups for *group members that are deemed qualified*. Conceptually, the fairness measures align with the fair classification requirement of Equality of Opportunity [31], in that fairness is dependent on qualification (in classification this would be a true positive label). To customize the motivation of Equality of Opportunity for subset auditing, we operationalize qualification via an auditor-specified

minimum score value q. In conjunction with qualification, we consider not only proportional presence but also equal presence. For instance, q might represent a minimum test score that a university requires, but simply having a score of q does not guarantee admission. q might also represent the minimum score for being in the top $x\%$ of the pool, thereby facilitating auditing fair group treatment scoped to the top $x\%$ of items.

S:Qualified Parity. *S:Qualified Parity* measures the fairness concept that all groups should be *proportionally present* in the subset when considering items that are qualified. The group based metric for *S:Qualified Parity* is the **qualified selection rate** $QSelectRt$ of a group $G_{p:j}$ defined as:

$$QSelectRt(G_{p:j}, S, X, q) = \frac{|S \cap G_{p:j} \cap X_{s_{x_i} \geq q}|}{|G_{p:j} \cap X_{s_{x_i} \geq q}|}. \tag{9}$$

Then using our ratio-based design, we define *S:Qualified Parity* as:

$$\textit{S:Qualified Parity}(S, X, p, q) = \frac{\min QSelectRt(G_{p:j}, S, X, q)}{\max QSelectRt(G_{p:j}, S, X, q)}, \quad \forall G_{p:j}, G_{p:i} \tag{10}$$

The *S:Qualified Parity* value is interpreted to represent the least favored group's $QSelectRt$ as a proportion of the $QSelectRt$ of the most favored group.

S:Qualified Balance. *S:Qualified Balance* measures the fairness concept that all groups should be *equally present* in the subset when considering items that are qualified. The group based metric for *S:Qualified Balance* is the **qualified proportion** of the subset $QPropOfS$ of a group $G_{p:j}$ as:

$$QPropOfS(G_{p:j}, S, q) = |S \cap G_{p:j} \cap X_{s_{x_i} \geq q}|/|S| \tag{11}$$

This then allows us to define *S:Qualified Balance* as:

$$\textit{S:Qualified Balance} = \frac{\min CPropOfS(G_{p:j}, S, a, v)}{\max CPropOfS(G_{p:i}, S, a, v))}, \quad \forall G_{p:j}, G_{p:i} \in X \tag{12}$$

The *S:Qualified Balance* value captures, when considering qualified group members, what percentage of spots held by the most favored group is held by the least favored group.

5.2 Calibrated Parity and Balance Measures

Building on the methodology of the previous qualified fair measures, the proposed *S:Calibrated Parity* and *S:Calibrated Balance* metrics respectively audit a subset for proportional or equal presence of groups from score bins distributed between the minimum and maximum score values. Conceptually these measures are aligned with calibration, a fairness criteria introduced in probabilistic classification [51], which requires a classifier produce outcomes that are independent of protected attributes after controlling for the estimated likelihood of the classification outcome. There is no estimated likelihood of being selected in subset selection, but our *S:Calibrated Parity* design captures the sentiment that items with similar scores should be treated similarly across groups. *S:Calibrated Balance* introduces an equal presence version to capture additional subset-specific fairness objectives. Calibrated fairness is ideal for subsets with score-based diversity (i.e., a spectrum of scores as opposed to "the best" scores), and for analyzing if items with similar scores are treated similarly regardless of group membership.

S:Calibrated Parity. *S:Calibrated Parity* measures the fairness concept that across the whole distribution of scores, all auditor-specified score bins should have similar selection rates across groups. Thus, the group-based metric for *S:Calibrated Parity* is the **bin selection rate** $BinSelectRt$ of a group $G_{p:j}$

$$BinSelectRt(G_{p:j}, S, X, l, u) = \frac{|S \cap G_{p:j} \cap X_{l < s_{x_i} < u}|}{|G_{p:j} \cap X_{a:v}|} \tag{13}$$

Then using $BinSelectRt$, we define *S:Calibrated Parity* as:

$$\textit{S:Calibrated Parity}(S, X, p, l_k, u_k) = \min\left(\frac{\min BinSelectRt(G_{p:j}, S, X, l_k, u_k)}{\max BinSelectRt(G_{p:i}, S, X, l_k, u_k)}\right), \quad \forall G_{p:j}, G_{p:i} \forall k \text{ bins represented by } l_k, u_k \tag{14}$$

The inner expression calculates the ratio between the highest and the lowest $BinSelectRt$ from the bin. This is performed over all bins. The smallest such ratio is called the *S:Calibrated Parity* value. This value captures, when considering the bin where the difference in group selection rates is largest, what percentage of spots held by the most favored group is held by the least favored group.

S:Calibrated Balance. *S:Calibrated Balance* measures the fairness concept that across the whole distribution of scores, for any auditor-specified score bin the subset should have the same number of items from each group. Thus, the group-based metric for *S:Calibrated Balance* is the **bin proportion** of the subset $BinPropOfS$ for a group $G_{p:j}$:

$$BinPropOfS(G_{p:j}, S, l, u) = |S \cap G_{p:j} \cap X_{l < s_{x_i} < u}|/|S| \tag{15}$$

Then using the FINS ratio-based design, we define *S:Calibrated Balance* as:

$$\textit{S:Calibrated Balance}(S, X, p, l_k, u_k) = \min\left(\frac{\min BinPropOfS(G_{p:j}, S, l, u)}{\max BinPropOfS(G_{p:i}, S, l, u))}\right), \quad \forall G_{p:j}, G_{p:i} \in X \forall k \text{ bins represented by } l_k, u_k \tag{16}$$

S:Calibrated Balance captures what percentage of spots held by the most favored group is held by the least favored group when considering the score bin where the difference in the number of items in the subset per group is largest.

5.3 Proposed Group Fair Relevance Measure

Unlike prior measures which haven both equal or proportional variations, we design *S:Relevance Parity* to audit a subset for group presence that is strictly proportional to the average score of the group. This represents the fairness concept that groups should be represented proportional to their average score (i.e., score-based relevance). While it does not capture individual fairness [23], it does have similar sentiment-based underpinnings, in that we design *S:Relevance Parity* to audit if *similar groups are treated similarly.*

The group-based metric for *S:Relevance Parity* is the **relevance rate** $RelRt$ for a group $G_{p:j}$:

$$RelRt(G_{p:j}, S) = \frac{|S \cap G_{p:j}|}{(\sum_{\forall x_i \in G_{p:j}} s_{x_i}/|G_{p:j}|)} \quad (17)$$

Employing our ratio-based design, we define *S:Relevance Parity* as:

$$S\text{:Relevance Parity}(S, p) = \frac{\min RelRt(G_{p:j}, S, X)}{\max RelRt(G_{p:j}, S, X)}, \quad \forall G_{p:j}, G_{p:i} \quad (18)$$

The *S:Relevance Parity* value is interpreted to represent the least favored group's relevance rate as the proportion of the relevance rate of the most favored group.

6 SCORE-BASED FAIRNESS METRICS

The next measures, unlike prior measures, quantify a subset's fair scoring of items. The conceptualization and thus measurement of fairness is not based on the presence of each group in the subset, but rather the scores each group has in the subset. In this case, the score is less likely to be a value of relevance or utility and more likely to be a distance or resource quantity. For instance, when subsetting a geographic region into communities for community-based COVID-19 testing, the scores might be the distance of each household to the testing center.

This notion is conceptually aligned with a fair centroid-based clustering objective, called socially fair k-means [30] or equitable group representation [1], where the algorithmic goal is to create clusters that have a similar average objective function value for every group. Our work is complimentary to these works, in that our *S:Score Parity* can be used to audit this objective. We also propose a equal-score version named *S:Score Balance*.

6.1 Proposed Score Parity and Balance Measures

The proposed *S:Score Parity* and *S:Score Balance*, respectively, audit a subset for proportional or equal group scoring.

S:Score Parity. *S:Score Parity* measures the concept that the group-total score in the subset should be proportional to the number of items per group in the subset. The group-based metric for *S:Score Parity* is the **average subset score** *AvgScoreInS* of a group $G_{p:j}$:

$$AvgScoreInS(G_{p:j}, S) = \frac{\sum_{\forall x_i \in S \cap G_{p:j}} s_{x_i}}{|G_{p:j}|} \quad (19)$$

This allows us to define *S:Score Parity* as:

$$S\text{:Score Parity}(S, p) = \frac{\min AvgScoreInS(G_{p:j}, S)}{\max AvgScoreInS(G_{p:j}, S)}, \forall G_{p:j}, G_{p:i} \quad (20)$$

The *S:Score Parity* value captures the least favored group's average subset score as a proportion of the average subset score of the most favored group.

S:Score Balance *S:Score Balance* measures the fairness concept that the group-total scores in the subset should be equal. The group-based metric for *S:Score Balance* is the **proportion of the subset total score** *PropScoreInS* of a group $G_{p:j}$:

$$PropScoreInS(G_{p:j}, S) = \frac{\sum_{\forall x_i \in S \cap G_{p:j}} s_{x_i}}{\sum_{\forall x_k \in S} s_{x_k}} \quad (21)$$

Thus, using our ratio-based design, we define *S:Score Balance* as:

$$S\text{:Score Balance}(S, p) = \frac{\min PropScoreInS(G_{p:j}, S)}{\max PropScoreInS(G_{p:j}, S)}, \forall G_{p:j}, G_{p:i} \quad (22)$$

Figure 1: Dataset from [34] of 12 images for the query "childcare worker" and three, to be audited, summaries of 4 images. Labels are from [34], W denotes gender = woman, M denotes gender = man, and * denotes multiple_ppl = yes.

	(a.)	(b.)	(c.)
S:Balance (gender)	1	.33	1
S:Conditioned Bal. (gender, multiple_ppl= yes)	0	1	1
S:Balance (gender X multiple_ppl)	0	0	1

Table 2: Audit results of summaries *a*, *b*, and *c* from Figure 1.

The *S:Score Balance* value represents the total score of the least favored group as a percentage of the most favored group's total score.

7 AUDITOR GUIDANCE FOR UTILIZING FINS

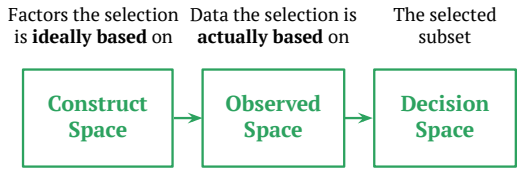
When utilizing our FINS framework auditors have two critical decisions to make. First, how to define a group in their audit. Second, what fairness measure(s) to utilize. Here we provide guidance on how to effectively make these decisions. Further, we present the first discussion of the values and assumptions beneath fair subset selection metrics. Our metric choice guidance, expressed as a decision diagram, facilitates understanding the conceptualization of fairness and its corresponding assumptions in a fairness audit.

7.1 Defining Groups

In many cases how to partition items into groups is clear to auditors based on the application. For instance, settings with one protected attribute and no other attributes yield groups that partition the protected attribute p. However, cases in which there are multiple protected and/or other attributes may require more consideration.

Example. We now illustrate via an example how the fairness audit performed using a given measure is incumbent on the choice of group definition. Figure 1 illustrates a data summarization task with 12 images for the search "childcare worker". The dataset is from Kay et al. [34] which studied Google Image search results for gender representation proportional to real-world representation of various occupational queries.

Every image is labeled by human participants [34] for the protected attribute `gender` = {`woman`,`man`} and an additional attribute `multiple_people` = {`yes`, `no`}. We audit each data summary of 4 items for `gender` balance (whereby we quantify the objective of having an equal presence of both genders in the subset). Table 2



WYSIWYG: Observed space can be a good decision-making proxy for the construct space.

WAE: The observed space is a distorted proxy of the construct space, in which all groups are assumed similar.

Figure 2: Worldviews proposed by Friedler et al. [26], which we describe in the context of subset selection.

illustrates how the definition of groups affects the fairness notion. We consider three different audits applied to the three potential subsets (a.), (b.) and (c.) in Figure 1.

The first audit defines groups using the `gender` protected attribute, we see that both subsets (a.) and (c.) exhibit `gender` balance. The second audit defines groups using the `gender` protected attribute while conditioning on `multiple_people` = `yes` (conceptually this seeks gender balance when multiple people are in the image). In Table 2, we see that both subsets (b.) and (c.) exhibit conditioned gender balance. The third audit redefines the protected attribute to be p = `gender` X `multiple_people`. In Table 2 we see that only subset (c.) is perfectly fair in this regard.

This illustrates that three different approaches yield three different evaluations. Critically, the example shows employing conditioned balance where groups are defined by the protected attribute p and by sharing a value v in attribute a is not the same as redefining groups from a combination of the protected attribute p and attribute a (i.e., p X a). This leads to two takeaways. First, in the case of multiple protected attributes, we recommend auditing for both *all protected attributes independently* and *their combined intersection.* When groups account for an additional attribute value through conditioned fairness this *choice should answer the question – which entities must be treated fairly*?

7.2 Choosing a Fairness Measure via the Guidance of the FINS Fair Subset Chart

The FINS framework encompasses numerous strategies of auditing subsets for fairness. To aid auditors in selecting the most appropriate fairness measure with the problem they are auditing, we design the FINS Fair Subset Chart (FSC for short). The FSC in Figure 3 provides a few targeted questions about the audit setting and fairness goal to derive the recommendation of a specific metric. This allows auditors with potentially multiple fairness goals to navigate their many choices and quickly discover an apppropriate fairness measure(s).

The FSC explicitly avoids guidance in the form of pairing different subset selection tasks (e.g., multi-winner voting or top-k selection) to specific measures. Instead, we design the chart to (1.) determine the fairness flavor applicable for the audited problem and (2.) recommend a metric(s) based on how the auditors are conceptualizing fairness. This allows for a values-based audit as opposed to an overly prescriptive and potentially misaligned rigid framework. Implicit in the choice of how to measure fairness is a moral perspective on what fairness even means in a data-driven context. Seminal work in fair classification by Friedler et al. [26] proposed that fairness (and thus judgments on fairness) in data-driven decisions encodes values and assumptions about the world the data models. They present two axiomatic belief systems termed worldviews [26]: "What You See is What You Get" (WYSIWYG) and "We're All Equal" (WAE). We describe WYSIWYG and WAE in Figure 2. Few works discuss and link classification specific fairness constraints to worldviews [26, 32, 47]. In fact, we are the first to pursue this critical discussion in the context of subset selection. We do so by utilizing the FSC to explicitly highlight the worldviews corresponding to FINS fair auditing measures. Thus, in addition to serving as an auditor's guidance in metric choice, the FSC also captures our proposed mapping of FINS fairness measures to Friedler et al.'s worldviews [26]. It is thus a *simple yet powerful means to discern the worldview of a fairness audit,* without any assumed or necessitated familiarity with Friedler et al.'s [26] conceptual framework.

In Figure 3, the metrics in our FINS framework are marked by our proposed variations of equal or proportional fairness goals (blue and red regions), are binned by the fairness flavor as define in Section 3 (yellow row) and by Friedler et al.'s worldviews (grey row). Below, we discuss the mapping between metric-choice guidance provided in the FSC to fairness flavors and worldviews.

Score-based and WYSIWYG. The FSC suggests these measures when items have an associated score and auditors seek to measure if a subset has a group-fair scoring. This is particularly applicable to problems where the score of an item represents a resource or a distance. As fairness is conceptualized through scores, these metrics embody the "WYSIWYG" worldview proposed by Friedler et al. [26].

Item-based, score dependent and WYSIWYG.The FSC recommends these measures when items have an associated score and auditors seek to measure fairness that accounts for the item associated score. To the best of our knowledge, our framework is the first to propose fairness measures conceptually aligned with Equality of Opportunity [31], Calibration [51], and Conditional Statistical Parity [18] (all previously formulated for classification) that have equal and proportional presence formats. These measures are applicable to settings in which the auditor seeks to consider fairness while also considering utility or relevance of items. The choice to factor item score into the measurement of fairness is indicative of "WYSIWYG" [26] as this is an explicit use (and thus conviction) in the construct space. As there are many ways to account for item score, the FSC asks auditors to choose how to incorporate the score and whether "fair" embodies proportional or equal group presence.

Item-based and WAE. The FSC points to these measures when the items do not have an associated score and when auditors thus chose not to factor item scores into the measurement of fairness. These measures are applicable to almost all settings as they do not require scores to be available or be associated apriori with the items. This lack of scores and/or explicit score-blindness is indicative of the "WAE" worldview, [26] as the measurement of fairness does not utilize the construct space. Here the FSC asks auditors if they would like to condition on an additional attribute value and to choose between proportional and equal group presence.

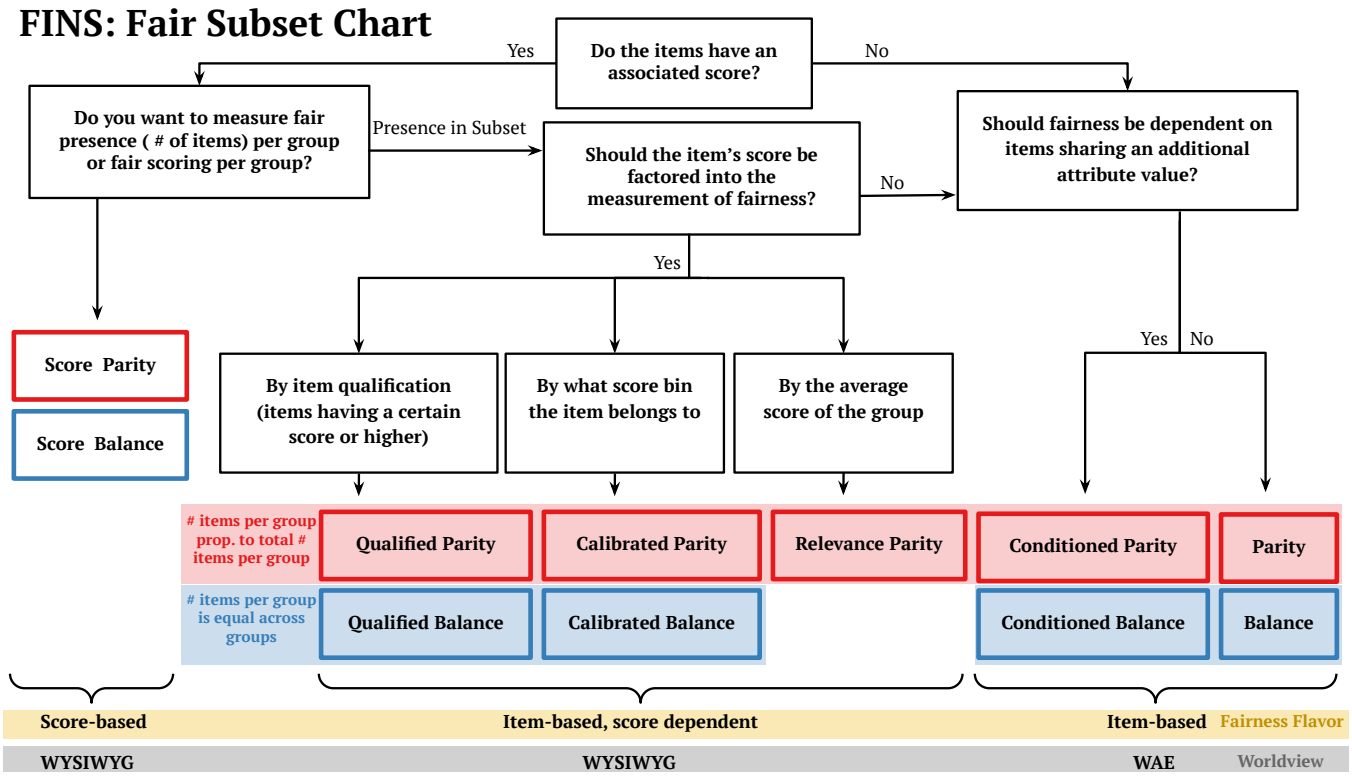


Figure 3: Fair Subset Chart (FSC): guidance for auditors to choose measure(s) applicable to their problem setting and audit goals. FSC highlights the corresponding fairness flavor and worldview [26] of each notion.

Broad Applicability. We note that the same subset can be audited with multiple metrics whereby each metric assesses for a different fairness issue. In fact, our framework is designed to allow for consistent interpretation across different measures. Thus, we recommend that the FSC be used not as a choice of one fairness concept but rather as a set of applicable possibilities.

8 EXPERIMENTAL EVALUATION VIA CASE STUDIES

We illustrate the use of the FSC and our proposed metrics through case studies with AirBnB listings and North Carolina voter registration records; both examining whether our measures indicate any bias and demonstrating their interpretability.

8.1 AirBnB Case Study

As AirBnB hosts can have multiple listings (sometimes hundreds in the same locality), there is increasing concern among "mom and pop" hosts that the platform is advantaging hosts that are clearly running large businesses, because commercial hosts generate more revenue for the company [11]. In this vein, we aim to understand the following **fairness questions** in AirBnB data with regards to the selection of the top 50 listings in each locality:

q1: Does the top-50 subset advantage professional hosts over other hosts relative to how popular their listings are?

q2: Does the top-50 subset advantage professional host over other hosts for listings that are comparably popular?

To conduct the above audit, we used AirBnB datasets from three diverse regions for the localities of Bangkok, Berlin, and New Zealand[1]. We utilize the "reviews per month" variable as each listing's associated popularity score. We only consider listings with > 0 reviews per month. The sizes of our datasets are Bangkok $|X| = 10,418$, Berlin $|X| = 14,716$, and New Zealand $|X| = 34,042$ listings. Based on prior multi-listing AirBnB analysis [35], we create three host categories: `single` (host has one property), `small` (two or three properties), and `professional` (four or more properties).

Guided by the FSC, we see that both questions are concerned with fairness that considers scores (i.e., popularity). To address question q1, we employ *S:Relevance Parity* to quantify if hosts are selected relative to how popular they are. To address q2, we employ qualified fairness in both its proportional and equal presence versions to quantify if qualified (i.e., highly popular) listings are treated fairly across host types. To measure "highly popular", we set q to be the average score for the locality. Figure 4 presents *S:Relevance Parity*, *S:Qualified Parity*, and *S:Qualified Balance*, and their corresponding group-based metrics for the three localities.

Studying question q1, we see that the *S:Relevance Parity* measures of all localities are $<< 1$. This indicates that the selection of listings is not proportional to the average popularity of the host

[1]Downloaded from http://insideairbnb.com/get-the-data.html.

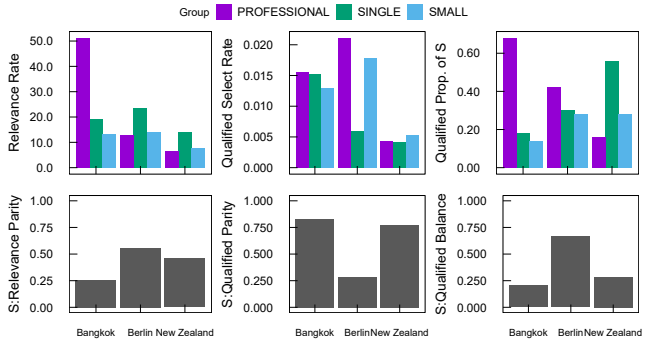


Figure 4: AirBnB locality subsets evaluated for relevance parity, qualified parity, and qualified balance along with their group-based measures for host groups.

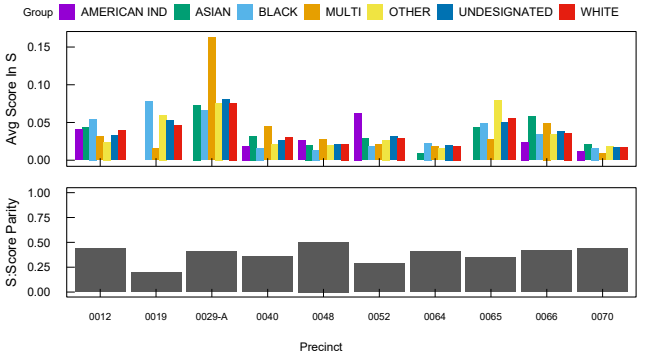


Figure 5: The 10 precincts in Burke County with the lowest *S:Score Parity* values, and the *AvgScoreinS* for racial groups.

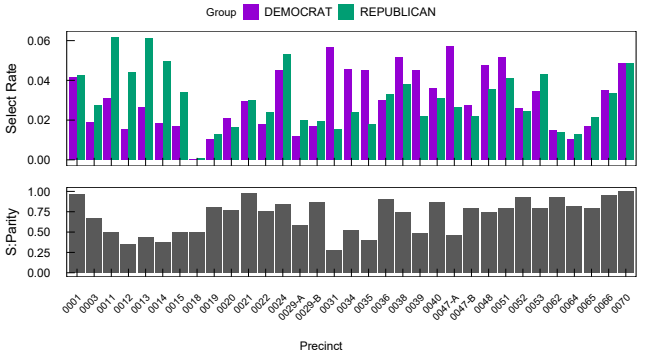


Figure 6: The *S:Parity* per precinct in Burke County and the *SelectRt* for political party group.

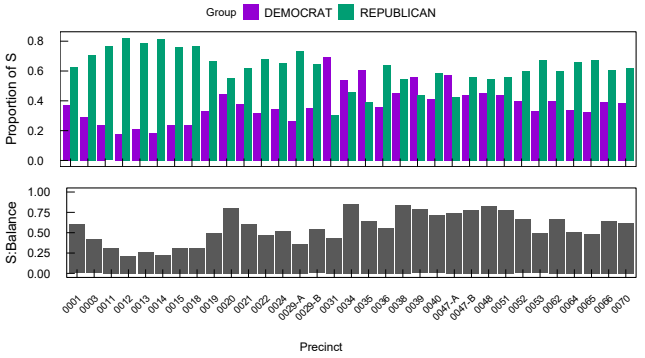


Figure 7: The *S:Balance* per precinct in Burke County and the *PropofS* for political party groups.

type. Ideally, the relevance rates would be similar across host types indicating host-group listings are selected proportional to their average popularity. Instead, in Bangkok, `professional` hosts are over-selected compared to their popularity. The *S:Relevance Parity* value illustrates that `small` hosts have a relevance rate roughly 25% of `professional` hosts. In Berlin and New Zealand, `single` hosts are over-selected compared to their average popularity. Yet the relevance rates are relatively closer, yielding higher *S:Relevance Parity* measures than Bangkok. We observe, that on the whole, listings are not selected relative to their group's popularity.

Studying question q2, we examine both *S:Qualified Parity* and *S:Qualified Balance*. Examining *S:Qualified Parity*, Bangkok and New Zealand have higher values indicating that across popular listings different hosts have a more comparable chance of being selected. This is further illustrated by the relative similarity of the qualified select rates of each host group. However, in Berlin, the *S:Qualified Parity* value is around .28. This indicates the selection rate of the least favored group (`small` hosts in this case) is only 28% of the qualified selection rate of `professional` hosts, the group with the largest qualified selection rate. Thus, if proportional presence of popular listings is the fairness objective, this is closer to being achieved in Bangkok and New Zealand than in Berlin.

Finally, we examine *S:Qualified Balance* to answer if popular listings are equally present in the subset across host groups. Largely, this objective is not achieved. In Bangkok and Berlin, popular `professional` hosts make up a larger share of the subset compared to popular `small` hosts. The *S:Qualified Balance* values indicate that this unfairness is greatest in Bangkok where popular professional hosts have only 20% of the spots in the campaign held by popular `small` hosts. *S:Qualified Balance* is a higher .66 in Berlin. New Zealand resembles Bangkok in terms of unfairness but the qualified proportions of the subset S indicates that in New Zealand `single` hosts are over-represented compared to `professional` hosts. Thus, equal presence of popular listings is not achieved.

8.2 NC Voter Records Case Study

Utilizing voter registration records from the state of North Carolina, we study the following **fairness questions:**

q1: Do certain race groups have further to travel to reach their precinct's polling place?

q2: Do subsetted precinct regions exhibit indication of gerrymandering (i.e., bias towards a particular political party)?

To conduct the above audit, we used voter records from Burke County NC [2], which contain `race` and `political party` attributes

[2]Dataset from: https://www.ncsbe.gov/results-data/voter-registration-data

for every voter. To measure how far an individual must travel to reach their designated polling place, we used the US Census GeoCoding API [9, 55] to geocode to longitude and latitude both each individual's address and the address of each precinct's polling location. Then for every individual, we calculated a distance score as the euclidean norm between their address and corresponding polling place. After removing all addresses that could not be geocoded and individuals registered as deceased, our Burke County dataset includes 42, 938 voter records for 32 precincts.

Guided by the FSC, we see q1 is concerned with fairness measures not based on fair scoring (i.e, distance to polls) per group. We thus employ *S:Score Parity* since we are interested in *average* distances between comparable. For q2, since the auditing problem does not associate scores with voters (distance is no longer relevant here) and gerrymandering bias is entirely dependent on the presence of political party voters in each precinct,the FSC recommends *S:Parity* and *S:Balance*. We note that *S:Parity* and *S:Balance* are related to two different types of gerrymandering - namely *cracking* and *packing*.

Cracking entails selecting district subsets such that a specific political party is a minority in each subset [27]. Interestingly, this is akin to achieving *S:Parity* close to 1 for most districts, since this indicates statistical parity or proportional representation is achieved. Thus, if all political parties are represented proportionally (i.e, *S:Parity* is perfectly achieved), then the minority party in the state is a minority in all districts. It is thus unlikely to win and that district becomes less competitive for the majority. Packing on the other hand is to select district subsets so that the one party has a severe advantage in one or a handful of districts, and a severe disadvantage in all the other districts [27]. The party is "packed" into one or two districts. This can be detected utilizing *S:Balance* . If all districts are unbalanced (*S:Balance* close to 1) and one party has an advantage in a few districts, but a disadvantage in the rest then packing could be at play. On the whole, gerrymandering must be a trend across district subsets, simply analyzing one district does not yield any significant evaluation. For our audit, we treat precincts as selected subsets and examine all precincts in Burke County.

Studying question q1, Figure 5 displays the *S:Score Parity* of the 10 precincts with the lowest *S:Score Parity* values [3]. For this set of precincts, we observe that *S:Score Parity* $< .5$ indicating a substantial difference in distance to polling places across racial groups. The value of *S:Score Parity* indicates the average distance of the group with the smallest travel as a percentage of the average distance of the group with the most travel. For instance, for precinct 0019, which has *S:Score Balance* = .25, this indicates that `multi-racial` voters have 25% of the average distance that `black` voters have to reach their polling place. This is followed by `white` voters with the second lowest average score (travel distance). Across the precincts, *S:Score Parity* highlights significant differences in travel distance to polling centers across racial groups.

Studying question q2, we scope our dataset to `political party` = {`democrat`, `republican` }, Figure 6 shows *S:Parity* across precincts. When auditing for cracking, we look for consistently high *S:Parity* values across precincts. On the whole, we see a large number of precincts with these values $> .75$. However, we also see some precincts where *S:Parity* is much lower – indicting select rates that are vastly different between `democrats` and `republicans`. We observe that there is not an overarching trend indicating cracking is at play. This study demonstrates the use of *S:Parity* to audit for this type of bias (or lack thereof).

Figure 7 shows *S:Balance* across precincts. In auditing for packing we would expect to see *S:Balance* values to be consistently very low. On the whole, we see multiple but not a majority of precincts with low *S:Balance* values. Thus cracking does not immediately appear to be an issue. However, if we examine the group-based measure, proportion of S, for the low *S:Balance* precincts ($< .5$) we see that almost all have significantly more `republicans` than `democrats` except for precinct 0031, where there are more `democrats`. This corresponds to the second component of packing - i.e., one party having a majority in all but one district. With our observations we can conclude that there is not a strong enough trend for packing. However, if `republicans` had a stronger majority of voters in all districts, our proposed measures would have detected this bias.

9 RELATED WORK

We approach subset selection from the perspective of algorithmic auditing (i.e., interpretable diagnosis of different fairness issues), while prior work designs algorithms to achieve fair outcomes. The first line of related research is algorithms that place items into the subset such that the subset has a fair presence of groups. In this instance, being in the set is a positive outcome.

These include fair top-k queries [44, 48, 59, 62], multi-winner voting [10, 14, 40], clustering [3, 3, 4, 6, 15, 16, 16, 39, 57], and data summarization [12, 33, 38]. We can classify these works into three categories, methods that achieve a exact form of (1.) proportional (to the dataset) group presence in the subset [3, 4, 8, 15, 39, 40], (2.) equal group presence [16, 33, 38] or (3) employ lower or upper bound quotas on the number of items per group [6, 10, 12, 14, 16, 25, 44, 48, 57, 59, 62]. The later can potentially achieve either (1.), (2.), or something else application-desired. To the best of our knowledge these algorithms do not consider the subset-based fairness conceptualizations we propose of calibrated, qualified, conditioned, or relevance fairness which we model with proportional and equal presence.

The second line of related work is clustering algorithms that feature fairness characteristics related to our proposed *S:Score Parity* measure [1, 30, 42]. These algorithms are designed to select multiple subsets (i.e, a clustering) with the objective that groups have similar clustering costs. Our work differs in that we define a fairness concept akin to this objective that is broadly applicable (i.e., interpretable and usable beyond integration into the objective function of centroid-clustering). Also, we provide a complementary novel fairness notion for equal group-based scores in a subset.

10 CONCLUSION

In this work we present the first comprehensive auditing framework, FINS, for evaluating subset selections for differing notions of group fairness. Our FINS framework yields a powerful open-source diagnostic toolkit generally applicable to any subset selection problem in which an understanding of bias is needed. FINS is available as a python library at https://github.com/KCachel/fins or via PyPi at https://pypi.org/project/finsfairauditing/.

[3] See https://github.com/KCachel/FINS-Experiments for all precinct results.

ACKNOWLEDGMENTS

This work is supported by NSF grant IIS-2007932.

REFERENCES

[1] Mohsen Abbasi, Aditya Bhaskara, and Suresh Venkatasubramanian. 2021. Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 504–514.
[2] Civil Rights Act. 1964. Civil Rights Act of 1964. (1964).
[3] Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. 2019. Clustering without over-representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 267–275.
[4] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. 2019. Scalable fair clustering. In *International Conference on Machine Learning*. PMLR, 405–413.
[5] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.
[6] Suman K Bera, Deeparnab Chakrabarty, Nicolas J Flores, and Maryam Negahbani. 2019. Fair algorithms for clustering. *arXiv preprint arXiv:1901.02393* (2019).
[7] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).
[8] Matteo Böhm, Adriano Fazzone, Stefano Leonardi, and Chris Schwiegelshohn. 2020. Fair clustering with multiple colors. *arXiv preprint arXiv:2002.07892* (2020).
[9] Branson Fox and Christopher Prener. 2021. slu-openGIS/censusxy: censuxy v1.0.1. https://doi.org/10.5281/ZENODO.4549749
[10] Robert Bredereck, Piotr Faliszewski, Ayumi Igarashi, Martin Lackner, and Piotr Skowron. 2018. Multiwinner elections with diversity constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[11] Olivia Carville. 2020. Airbnb's 3.1 Billion IPO Hinges on Hosts Who Make Rentals Feel Like Home. https://www.bloomberg.com/news/articles/2020-12-09/airbnb-s-3-1-billion-ipo-hinges-on-hosts-who-make-rentals-feel-like-home
[12] Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. 2018. Fair and diverse DPP-based data summarization. In *International Conference on Machine Learning*. PMLR, 716–725.
[13] L Elisa Celis, Chris Hays, Anay Mehrotra, and Nisheeth K Vishnoi. 2021. The Effect of the Rooney Rule on Implicit Bias in the Long Term. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 678–689.
[14] L Elisa Celis, Lingxiao Huang, and Nisheeth K Vishnoi. 2017. Multiwinner voting with fairness constraints. *arXiv preprint arXiv:1710.10057* (2017).
[15] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2018. Fair clustering through fairlets. *arXiv preprint arXiv:1802.05733* (2018).
[16] Ashish Chiplunkar, Sagar Kale, and Sivaramakrishnan Natarajan Ramamoorthy. 2020. How to solve fair k-center in massive data models. In *International Conference on Machine Learning*. PMLR, 1877–1886.
[17] US Equal Employment Opportunity Commission et al. 1979. Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *US Equal Employment Opportunity Commission: Washington, DC, USA* (1979).
[18] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.
[19] Kimberle Crenshaw. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.* 43 (1990), 1241.
[20] B Dattner. 2016. A scorecard for making better hiring decisions. *Harvard Business Review* (2016).
[21] Rebecca Deczynski. 2021. Hiring Is a Pain-These 5 Companies Say A.I. Can Make It Better. *Inc.* (Oct 2021). https://www.inc.com/rebecca-deczynski/artificial-intelligence-hiring-companies-labor-shortage-great-resignation.html
[22] Kathryn Dill. 2021. Companies Need More Workers. Why Do They Reject Millions of Résumés? https://www.wsj.com/articles/companies-need-more-workers-why-do-they-reject-millions-of-resumes-11630728008
[23] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[24] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.
[25] Till Fluschnik, Piotr Skowron, Mervin Triphaus, and Kai Wilker. 2019. Fair knapsack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1941–1948.
[26] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).
[27] John N Friedman and Richard T Holden. 2008. Optimal gerrymandering: sometimes pack, but never crack. *American Economic Review* 98, 1 (2008), 113–44.
[28] Joseph B Fuller, Manjari Raman, Eva Sage-Gavin, and Kristen Hines. 2021. Hidden Workers: Untapped Talent. *Harvard Business School, September* (2021).
[29] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2221–2231.
[30] Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. 2021. Socially fair k-means clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 438–448.
[31] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.
[32] Corinna Hertweck, Christoph Heitz, and Michele Loi. 2021. On the moral justification of statistical parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 747–757.
[33] Matthew Jones, Huy Nguyen, and Thy Nguyen. 2020. Fair k-centers via maximum matching. In *International Conference on Machine Learning*. PMLR, 4940–4949.
[34] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3819–3828.
[35] Qing Ke. 2017. Sharing means renting? An entire-marketplace analysis of Airbnb. In *Proceedings of the 2017 ACM on web science conference*. 131–139.
[36] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[37] Jon Kleinberg and Manish Raghavan. 2018. Selection problems in the presence of implicit bias. *arXiv preprint arXiv:1801.03533* (2018).
[38] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. 2019. Fair k-center clustering for data summarization. In *International Conference on Machine Learning*. PMLR, 3448–3457.
[39] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. 2019. Guarantees for spectral clustering with fairness constraints. In *International Conference on Machine Learning*. PMLR, 3458–3467.
[40] Jérôme Lang and Piotr Skowron. 2018. Multi-attribute proportional representation. *Artificial Intelligence* 263 (2018), 74–106.
[41] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[42] Yury Makarychev and Ali Vakilian. 2021. Approximation Algorithms for Socially Fair Clustering. *arXiv preprint arXiv:2103.02512* (2021).
[43] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. On the applicability of ML fairness notions. *arXiv preprint arXiv:2006.16745* (2020).
[44] Anay Mehrotra and L Elisa Celis. 2021. Mitigating Bias in Set Selection with Noisy Protected Attributes. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 237–248.
[45] Clair Cain Miller. 2021. The Pandemic Created a Child-Care Crisis. Mothers Bore the Burden. https://www.nytimes.com/interactive/2021/05/17/upshot/women-workforce-employment-covid.html
[46] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 117–123.
[47] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867* (2018).
[48] Zafeiria Moumoulidou, Andrew McGregor, and Alexandra Meliou. 2020. Diverse Data Selection under Fairness Constraints. *arXiv preprint arXiv:2010.09141* (2020).
[49] Caitlin Mullen. 2021. *AI use in hiring means women with employment gaps get overlooked* (Sep 2021). https://www.bizjournals.com/bizwomen/news/latest-news/2021/09/ai-hiring-women-employment-gaps.html?
[50] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568.
[51] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).
[52] Richard J Powell, Jesse T Clark, and Matthew P Dube. 2020. Partisan gerrymandering, clustering, or both? A new approach to a persistent question. *Election Law Journal: Rules, Politics, and Policy* 19, 1 (2020), 79–100.
[53] Benjamin Rader, Christina M Astley, Karla Therese L Sy, Kara Sewalk, Yulin Hswen, John S Brownstein, and Moritz UG Kraemer. 2020. Geographic access to United States SARS-CoV-2 testing sites highlights healthcare disparities and may bias transmission estimates. *Journal of travel medicine* (2020).

[54] Representation2020.com. 2016. Women Winning. https://www.representwomen.org/women_winning#voting_systems
[55] RPVote. 2020. eiCompare. https://github.com/RPVote/eiCompare
[56] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).
[57] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. 2019. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*. Springer, 232–251.
[58] Yash Raj Shrestha and Yongjie Yang. 2019. Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms* 12, 9 (2019), 199.
[59] Julia Stoyanovich, Ke Yang, and HV Jagadish. 2018. Online set selection with fairness and diversity constraints. In *Proceedings of the EDBT Conference*.
[60] Karen Taylor, Francesca Properzi, and Maria Joao Cruz. 2020. Intelligent clinical trials: transforming through AI-enabled engagement. (2020).
[61] Sriram Vasudevan and Krishnaram Kenthapadi. 2020. LiFT: A Scalable Framework for Measuring Fairness in ML Applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2773–2780.
[62] Yanhao Wang, Francesco Fabbri, and Michael Mathioudakis. 2021. Fair and Representative Subset Selection from Data Streams. In *Proceedings of the Web Conference 2021*. 1340–1350.
[63] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*. 1–6.