D1.1 Survey on Bias Metrics

Christos Karanikolopoulos[1], **Panagiotis Papadakos**[1, 2], **Glykeria Toulina**[1], **Spyridon Tzimas**[1], **Panayiotis Tsaparas**[1]

[1] Department of Computer Science and Engineering, University of Ioannina, Greece
[2]Institute of Computer Science (ICS) - Foundation for Research and Technology - Hellas (FORTH), Greece

1. Introduction

The AI revolution of the past decade has led to a world where many decisions that affect human lives are assisted by or deferred entirely to algorithmic systems trained on massive amounts of data. These decisions may be at an individual level, ranging from simple ones, like where to dine, which movie to watch, who to follow, what article to read, or what information to consume, to more important ones, such as what school to apply to, what career to follow, or what treatment to receive. Algorithms also affect decisions at organizational, institutional or societal level that have to do with the operation of financial institutions (determining who should get a loan), judiciary system (affecting sentencing decisions), academics (influencing admissions), or law enforcement (e.g., face recognition systems for suspect identification).

Given the critical role that algorithms play in our lives, there is increased concern as to whether the decisions of these algorithms are ethical and just. These concerns are not unfounded. There is a stream of empirical evidence that suggests that algorithms may exhibit *biases* in their decisions. For example, the COMPASS system, which determines the risk of recidivism, was shown to be biased towards African-American inmates, while Google Ads was shown to be more likely to show ads for low-paid jobs to women than men. There are several such examples, where automated systems are shown to exhibit bias against specific groups of individuals in very diverse settings.

The term bias is a loaded term. For our purposes, following the definition in [43] we will define bias as "the inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair". On the flip side, we can also consider the fairness of an algorithmic system, which can be generally defined as the lack of bias. These two concepts are tightly coupled, as one excludes the other. Formally defining them for algorithmic systems has proved to be a challenging task, that depends on the specific task and the goal of the definition. In the following, we will survey some of the different metrics of algorithmic bias and fairness for specific tasks of interest to the project, aligning with other relevant works [50], [29].

The report is structured as follows. In Section 2, we will present some high-level types of bias and fairness, while in Section 3 we provide bias and fairness definitions for classification tasks, which can be applied and extended to many different settings. In Section 4 we will review bias and fairness metrics for Large Language Models. In Section 5, we will review bias and fairness metrics for clustering algorithms. Finally, in Section 6 we will review bias and fairness metrics for network analysis algorithms.

2. Overview

We now give a general overview of the different approaches in measuring bias and fairness. Depending on the exact class of problems we consider (e.g., classification, clustering, etc), these approaches produce different types of metrics. For the following, we consider both bias and fairness as one, since we view one to be the flip side of the other. Fairness implies lack of bias, while bias implies lack of fairness.

We identify the following three broad approaches in defining fairness [43, 27, 25]:

- **Group Fairness:** In group fairness, we assume that our data can be partitioned into two or more groups. Usually, these groups are defined based on a sensitive (protected) attribute, such as gender, race, religion, etc. Group fairness requires that the different groups are treated equally, or proportionally to their representation in the data. In some cases, we can also define group fairness by focusing on some protected or minority group and requiring an appropriate treatment of that group.
- **Individual fairness:** The principle behind individual fairness is that similar individuals should receive similar treatment. This approach assumes that we can define a distance or similarity in the input space, and a corresponding distance or similarity in the output space, and it requires that these two are related. Individual fairness focuses on individual instances rather than groups.
- **Causal fairness:** Similar to group fairness, causal fairness assumes the existence of one or more sensitive attributes, and defines fairness using causal models [44]. A commonly used definition is *Counterfactual fairness* which requires that algorithmic results for an instance remain the same when we consider a *counterfactual* of the instance, where the values of the sensitive attributes have been flipped to different values. Counterfactual fairness combines aspects from both group fairness since it assumes the existence of groups, and from individual fairness, since it defines fairness with respect to specific instances.

In the following we will provide definitions of specific metrics that measure fairness (or bias) for specific problems. We will consider group fairness definitions, that is, we assume that instances are partitioned into groups based on sensitive attributes. These are the kind of definitions we are interested in our project.

3. Classification Bias

The origins of algorithmic and machine learning fairness can be traced to the problem of classification. Classification, in its simplest form binary classification, is a fundamental decision-making tool, that is applied widely in several automated (or semi-automated) decision systems. For example, classification algorithms are used for determining if someone should get a loan, if they are eligible for parole, if they should be shown an ad, if they will receive a specific recommendation, if they will be admitted to a specific school, etc. All these decisions affect people's lives to a greater or lesser extent, and therefore it is important that they are fair and unbiased.

Classification was thus one of the first tasks for which fairness was defined and quantified. The metrics defined for classification fairness influence to a large extend the fairness metrics for other tasks, so we begin our exposition with classification. There is a wide variety of metrics, in this survey, we present the most commonly used ones. Most other metrics are variations of those.

For the following definitions we adopt some of the notation used in [8] and [55]. We are given a dataset, where for each data instance x we have a set of features (attributes) which are used (some of them or all of them) for classification. The attributes include a sensitive attribute G which partitions the instances into groups. For simplicity, we will assume two groups. That is, the attribute G takes two values $\{g, \bar{g}\}$. We will assume that the value g corresponds to the *protected* group, that is, the group we want the classification algorithm to treat fairly. We will refer to the group $\bar{g}$ as the *complement* group.

The data instances also have an additional attribute Y which is the class label that we want to predict. Without significant loss of generality, we assume a binary classification task, that is, our class label takes values $\{0, 1\}$. We assume that 1 correspond to a positive outcome (for example, a job offer, or a loan approval) while 0 to a negative outcome. We will use $\hat{Y}$ to denote the decision of our classification model, which is again a binary value $\{0, 1\}$.

3.1. Output-based definition

The first definition we present considers only the output of the classifier.

Statistical Parity/Group fairness: A classifier satisfies statistical parity fairness if the probability that an instance receives a positive outcome is the same for the two groups. That is,

$$P[\hat{Y} = 1|G = g] = P[\hat{Y} = 1|G = \bar{g}]$$

Statistical parity implies that overall, the fraction of instances with positive outcome from each group will mirror the fraction in the input data. That is, if the protected group is 20% of the overall population, then 20% of the instances that are labeled positive by the classifier will be from the protected group g.

Statistical parity fairness is a natural notion of fairness, and the underlying intuition can be extended to tasks beyond classification, such as ranking, recommendation, or selection. In this case we require *representation fairness*, that is, that the different groups are proportionally represented into the output of the algorithm. We will see such definitions when discussing clustering fairness, and Pagerank fairness.

3.2. Error-based definitions

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	TP	FP
$\hat{Y} = 0$	FN	TN

Figure 1. Confusion matrix

The next definitions of fairness we consider take into account the *errors* of the classifier. Given the true class label Y and the predicted class label $\hat{Y}$ we can create the *confusion matrix*, as shown in the figure below. Given the confusion matrix, we can estimate different metrics for the performance of the classifier. For example, the *accuracy* of the classifier is estimated as $acc = \frac{TP+TN}{N}$ where N is the total number of instances.

Another quantity of interest is the *True Positive Ratio (TPR)* defined as

$$TPR = \frac{TP}{TP + FN} = P[\hat{Y} = 1|Y = 1]$$

This is the fraction of truly positive instances ($Y = 1$) that are classified as positive by the model. We can also think of this as the probability that the classifier will correctly classify a truly positive instance. Closely related is the *False Negative Ratio (FNR)*, which is defined as

$$FNR = \frac{FN}{TP + FN} = P[\hat{Y} = 0|Y = 1] = 1 - TPR$$

Similarly, we can define the *False Positive Ratio (FPR)* as

$$FPR = \frac{FP}{FP + TN} = P[\hat{Y} = 1|Y = 0]$$

This is the fraction of truly negative instances ($Y = 0$) that are mistakenly classified as positive. We can also think of this as the probability that the classifier will incorrectly classify as positive a truly negative instance. Closely related is the *True Negative Ratio (TNR)*, which is defined as

$$TNR = \frac{TN}{FP + TN} = P[\hat{Y} = 0|Y = 0] = 1 - FPR$$

Given the confusion matrix and the different types of error or success rates, we can define different notions of fairness that aim to achieve equality for these error or success rates when conditioning on the group membership.

Equal Opportunity/TPR Balance: A classifier satisfies Equal Opportunity fairness if the True Positive Ratio (TPR) conditioned on the group membership is the same for the two groups. That is,

$$P[\hat{Y} = 1|Y = 1, G = g] = P[\hat{Y} = 1|Y = 1, G = \bar{g}]$$

Note that equality of TPR also implies equality of False Negative Ratio, that is

$$P[\hat{Y} = 0|Y = 1, G = g] = P[\hat{Y} = 0|Y = 1, G = \bar{g}]$$

The idea of Equal Opportunity fairness is to extend the Statistical Parity fairness to instances that are supposed to get the positive outcome. That is, among the instances with positive true label, the probability they get a predicted positive outcome should not depend on the group membership. For example, in a job recruiting algorithm, the probability that a deserving male candidate gets an offer should be the same as the probability that a deserving female candidate (e.g., with similar qualifications) gets an offer.

Predictive Equality/FPR Balance: A classifier satisfies Predictive Equality fairness if the False Positive Ratio (FPR) conditioned on the group membership is the same for the two groups. That is,

$$P[\hat{Y} = 1|Y = 0, G = g] = P[\hat{Y} = 1|Y = 0, G = \bar{g}]$$

Note that equality of FPR also implies equality of True Negative Ratio, that is

$$P[\hat{Y} = 0|Y = 0, G = g] = P[\hat{Y} = 0|Y = 0, G = \bar{g}]$$

Predictive Equality fairness looks at the flip side of Equal Opportunity, and it asks that the errors that favor instances that do not deserve a positive outcome should be equal among the two groups. For example, in a job recruiting algorithm, the probability that an undeserving male candidate gets an offer should be the same as the probability that an undeserving female candidate gets an offer.

Equalized Odds/Disparate Mistreatment: This definition combines the two definitions above. A classifier satisfies Equalized Odds fairness if both the True Positive Ratio and the False Positive Ratio (FPR) conditioned on the group membership are the same for the two groups. That is,

$$P[\hat{Y} = 1|Y = i, G = g] = P[\hat{Y} = 1|Y = i, G = \bar{g}], \text{ for } i = 0, 1$$

This definition requires that when making a positive (negative) decision, the fraction of deserving instances for the two groups, and the fraction of undeserving instances for the two groups that get this outcome are both the same for the two groups. For the recruiting algorithm, this would imply that the algorithm has equal probability of making a correct offer to a qualified male or female candidate, or an mistaken offer to non-qualified male or female candidate.

Predictive Parity: This definition looks at the errors that the classifier makes when predicting a positive outcome, and requires that they are the same for the two groups. That is,

$$P[Y = 1|\hat{Y} = 1, G = g] = P[Y = 1|\hat{Y} = 1, G = \bar{g}]$$

Note that this implies that:

$$P[Y = 0|\hat{Y} = 1, G = g] = P[Y = 0|\hat{Y} = 1, G = \bar{g}]$$

In our example, this would mean that the probability of the classifier making an error when offering a job should be the same for both men and women.

3.3. Score-based definitions

We will also consider classifiers that instead of outputting a class label, they output a *probability score* for each instance, which corresponds to the probability that the instance should get the positive label. These scores can be used to make a binary decision (by thresholding this score), or as the end product of the classifier.

Let $S(x)$ denote the score of the classifier for instance x, and S the random variable with the score. A classifier is *well calibrated* if the probability that an instance gets a positive outcome, when given score s is s, for any value of s. That is,

$$P[Y = 1|S = s] = s \,\text{forall}, s \in [0, 1]$$

We can now define calibration fairness where we require that the classifier has similar probabilities for each group.

Calibration Fairness: A classifier satisfies Predictive Parity fairness if

$$P[Y = 1|S = s, G = g] = P[Y = 1|S = s, G = \bar{g}], \text{forall}\, s \in [0, 1]$$

We can make this definition more strict by requiring that the classifier is also *well-calibrated* for the two groups. That is,

$$P[Y = 1|S = s, G = g] = P[Y = 1|S = s, G = \bar{g}] = s, \text{forall}\, s \in [0, 1]$$

4. LLMs

In this section, we focus on bias measuring metrics that are commonly used in LLMs [29]. Specifically, we describe embedding-based, probability-based, and metrics over the generated text.

4.1. Embedding-based Metrics

These metrics measure bias by computing the distances of the embeddings of neutral words to words that are associated with protected attributes. For example, the distance of the embedding of the neutral word 'doctor' is computed over the embeddings of the gender-associated words 'man' and 'woman'. Despite the fact that most LLMs do not provide access to embeddings, the authors in [57] propose efficient and effective methods to create them by using as training data synthetic data generated by the LLMs.

Word Embedding Metrics.

The Word Embedding Association Test (WEAT) metric [15], measures the associations between two sets of target words representing protected attributes of social groups like gender (e.g., male and female names) with two sets of words that are considered neutral attributes (e.g., love, hate, beautiful, ugly). The hypothesis is that there is no difference between the two sets of target words in terms of their similarity to the two sets of neutral attributes. WEAT is the normalized measure that denotes how separated the two distributions are. Formally, given the two sets of protected attribute words (P_1, P_2) of equal size and the two sets of neutral attribute words (N_1, N_2), the test statistic is :

$$s(P_1, P_2, N1, N_2) = \sum_{w_p \in P_1} s(w_p, N_1, N_2) - \sum_{w_p \in P_2} s(w_p, N_1, N_2) \tag{1}$$

, where

$$s(w_p, N1, N_2) = mean_{w_n \in N_1} cos(\vec{w_p}, \vec{w_n}) - mean_{w_n \in N_2} cos(\vec{w_p}, \vec{w_n}) \tag{2}$$

, where $cos(\vec{w_p}, \vec{w_n})$ denotes the cosine similarity of the embedding vectors of the protected word w_p and the neutral word w_n. Finally, the WEAT metric measures the effect size

$$WEAT(P_1, P_2, N_1, N_2) = \frac{mean_{w_p \in P_1} s(w_p, N_1, N_2) - mean_{w_p \in P_2} s(w_p, N_1, N_2)}{std_{w_p \in P_1 \cup P_2} s(w_p, N_1, N_2)} \quad (3)$$

Sentence Embedding Metrics. Most of these metrics are adaptations of the WEAT metric for sentences, where the embeddings are not static but learned in the context of sentences. For example, the Sentence Encoder Association Test (SEAT) [39] uses the same equation as in Eq. 3 over template sentences, where the empty slots in the templates are replaced with protected and neutral attribute words. SP_1, SP_2, SN_1, SN_2 denote the two pairs of sets of protected and neutral sentences. loThe embeddings are computed using the classification [CLS] token.

$$SEAT(SP_1, SP_2, SN_1, SN_2) = \frac{mean_{s_p \in SP_1} s(s_p, SN_1, SN_2) - mean_{s_p \in SP_2} s(s_p, SN_1, SN_2)}{std_{s_p \in SP_1 \cup SP_2} s(s_p, SN_1, SN_2)} \quad (4)$$

The Contextualized Embedding Association Test (CEAT) [32] proposes an approach that considers all the different contexts in which words can appear to compute their effect size, by generating sentences with combinations of the protected and neutral attribute words. However, since the number of combinations can be rather large and depends on the size of the word sets, the model proposes to randomly sample subsets of the embeddings and calculate a distribution of effect sizes. Formally, the combined effect size is the weighted mean of the distribution of random effects, as shown in the following formulae:

$$CEAT(P_1, P_2, N_1, N_2) = \frac{\sum_{i=1}^{N} u_i * WEAT(P_{1i}, P_{2i}, N_{1i}, N_{2i})}{\sum_{i=1}^{N} u_i} \quad (5)$$

, where u_i is derived from the variance of the random-effects model.

The Sentence Bias Score [24] was used for male and female gender bias, but can be expanded to any other (even non-binary) protected attributes. It uses the cosine similarity between neutral word embeddings and the protected attribute direction vector (in this case gender) to estimate word-level bias. Then it sums up the bias of all the words in the sentence, normalizing it with respect to the length of the sentence and to the contextualized semantic importance of each word. The bias score keeps the estimations of gender bias towards the male and female directions separated. Formally,

$$BiasScore(s) = \sum_{w \in s, w \notin P} cos(\vec{w}, \vec{p}) * I_w \quad (6)$$

, where $\vec{p}$ is the protected attributes direction, previously identified in the vector space from multiple words per attribute (e.g., words relating to male or words relating to female gender), P is a list of protected attribute words in the same language as the encoder, I_w is the semantic importance of each word in the sentence according to the encoder, and $\vec{w}$ is the embedding of a word of the input sentence.

Discussion. Regarding the embedding-based metrics, there are various references in the bibliography criticizing their effectiveness and consistency, pointing that it is preferable to measure bias on the corresponding tasks [21]. For example, [14] mentions that there might be inconsistencies between bias in representations and bias in the task at hand. Additionally, the effectiveness of the metrics depends on various choices like which are the templates, the seed words capturing the protected attributes, and the type of used embeddings (static or contextualized) [21].

4.2. Probability-based Metrics

These metrics are based on the predicted probabilities that are associated by the LLMs for the protected attribute words or neutral attribute words, based on specific prompts and templates given as input.

Masked Token Methods

These metrics are used in masked language models (MLMs) like BERT [22] by masking a word in a sentence. The MLM then predicts the missing word. For example, the Discovery of Correlations (DisCo) [58] approach uses templates (e.g., `"[P] is a [MASK]"`), where the `[P]` slot is filled with a protected attribute word and `[MASK]` is predicted by the model. By taking the top-3 predicted words, the metric computes the differences in the predicted words for the different social groups represented by the protected attribute words. A predicted word is supplied preferentially for one gender over another when the χ^2 metric rejects a null hypothesis of equal prediction rate.

Log-Probability Bias Score (LPBS) [38] tries to normalize the predicted probability of a protected attribute word using the prior bias of the model towards predicting this specific protected attribute word. Again, the metric uses templates for MLMs of the form `"[MASK] is a nurse"`, where `nurse` is a neutral attribute word and `[MASK]` is used for computing the probability p_{w_p} of a sentence where the `[MASK]` is replaced with a protected attribute word like `he`. The prior probability p_{prior} is computed by using the template `"[MASK] is a [MASK]"` that removes the neutral attribute and by predicting the probability of the sentence `"he is a [MASK]"`. Formally, for binary social groups protected attributes of a sentence S:

$$LPBS(S) = \log \frac{p_{w_{p1}}}{p_{prior1}} - \log \frac{p_{w_{p2}}}{p_{prior2}} \tag{7}$$

A variation of the LBPS for non-binary protected attributes is the Categorical Bias Score (CBS), which is described in [2].

Pseudo-Log-Likelihood Metrics

These metrics are based on the Pseudo-Log Likelihood (PLL) [48, 56] metric used in MLMs, that measures the probability of generating a token given the other words in a sentence. Similarly with previous approaches, a word is masked out, replaced by the `[MASK]` special token. Formally, if Θ denotes the model's parameters, and S a sentence

$$PLL(S) = \sum_{w \in S} \log P(w|S_{\backslash w}; \Theta) \tag{8}$$

The Context Association Test (CAT) Score proposed in [40] for the StereoSet dataset. Each sentence is paired with a stereotype, antistereotype and meaningless sentences. The metric computes the probability of protected attribute related tokens conditioned on neutral tokens, by masking and predicting the protected attribute tokens. Formally, if N the set of neutral words in a sentence S and P the protected attribute words,

$$CAT(S) = \sum_{w_p \in P} \log P(w_p|P_{\backslash w_p}, N; \Theta) \tag{9}$$

A variation of the CPS is the CrowS-Pairs Score proposed in [41] for the CrowS-Pairs datasets. Given a pair of sentences, one stereotyping and one that is not, the metric computes the probability of neutral tokens conditioned on protected attribute related tokens, by masking and predicting the neutral tokens. This is the opposite of CAT addressing that there might be an imbalance in the training data for the protected attribute tokens. The metric proposes to control this frequency imbalance to condition on the protected tokens when estimating the likelihoods of the neutral tokens. Formally, if N the set of neutral words in a sentence S and P the pair of protected attribute words,

$$CPS(S) = \sum_{w_n \in N} \log P(w_n|N_{\backslash w_n}, P; \Theta) \tag{10}$$

The All Unmasked Likelihood (AUL) [35] metric uses an unmasked sentence and the model predicts all the tokens in the sentence. Masking tokens can change the context of the input.

Additionally, removing one token at a time does not guarantee that the rest of the words are not biased. In AUL, the model has all the information to make the prediction of each token, improving the accuracy of the bias evaluation. Formally,

$$AUL(S) = \frac{1}{|S|} \sum_{w \in S} \log P(w|S;\Theta) \tag{11}$$

A variation is the AUL with Attention Weights (AULA) [35], that augments the AUL metric with the importance of each token based on its attention weight. Formally, considering that α_i denotes the attention weight of the token w_i the metric is computed by

$$AULA(S) = \frac{1}{|S|} \sum_{w_i \in S} \alpha_i log P(w_i|S;\Theta) \tag{12}$$

Given any of the above score functions denoted as f, and N pairs of stereotypical S_s and antistereotypical S_a sentences, the bias score of an LLM model M is given by the following function:

$$bias(M) = \frac{\mathbb{I}(f(S_s) > f(S_a))}{N} \tag{13}$$

, where $\mathbb{I}$ is the indicator function, which returns 1 if its argument is True and 0 otherwise. An ideal model should achieve a score of 0.5 when considering all sentences.

Finally, the Language Model Bias (LMB) [7] was proposed for the RedditBias dataset. The metric measures how much likelier is for the LLM to generate a stereotypically biased phrase compared to a corresponding inversely biased phrase, where terms of a protected group are replaced by terms of the inverse group over all the corresponding combinations. The metric is based on the mean perplexity differences between biased expressions and their counterparts. Outlier pairs with very high perplexity are removed to reduce noise, depending on the mean perplexity of the sample and the standard deviation. A two tailed Student t-test indicates the presence of bias.

Discussion. The effectiveness of probability-based metrics have been questioned on the downstream tasks by [21]. The fact that most of these metrics depend on masked templates with low diversity and limited target words can hinder their generalization and reliability. Finally, all of these metrics assume binary social groups which is not always the case.

4.3. Generated text-based Metrics

When the interaction with an LLM is limited to just the generated text and there is no access to the probabilities and embeddings, the only way to evaluate a model for bias is by evaluating its generated text. Usually, the models are given prompts that can contain biases and can lead to generated text that also contains biases. Such metrics include the comparison of the distributions of bias-associated tokens, by using auxiliary classification models that classify the generated text to the bias classes of interest, or by using lexicons that contain a set of biased words, potentially associated with a bias score, and compute the bias score of the generated text.

Distribution Metrics

These metrics compare the distributions of tokens in the LLM generated text for the various social groups. In the following, we discuss the most prominent ones.

Social Group Substitutions (SGS) [46] is based on the assumption that the distributions of the various tokens should be identical for various groups. The metric uses an invariance metric ψ like exact match [46], that is 1 when all characters of two texts match and 0 in any other case. Formally,

$$SGS(\hat{Y}) = \psi(\hat{Y}_i, \hat{Y}_j) \tag{14}$$

, where $\hat{Y}_i$ and $\hat{Y}_j$ are the two LLM generated texts based on input that depends on different values of a protected attribute (e.g., gender).

Co-Occurence Bias Score [11] restricts the focus only to words that co-occur with a set of words related to specific values of a protected attribute. Specifically, given a token w and two set of protected attribute words P_1 and P_2 the bias score for each word in a corpus of generated texts is computed as:

$$Co - occurenceBiasScore(w) = \log \frac{P(w|P_1)}{P(w|P_2)} \tag{15}$$

The score is zero for all words w that co-occur equally for each set of protected attribute words.

Demographic Representation [10] counts how many times a token w associated with a specific group appears in a generated text Y. Formally, for each protected group G_i associated with a set of protected attribute words P_i, and a set of generated texts $\mathbb{Y}$ the count is:

$$DR(G_i) = \sum_{w_p \in P_i} \sum_{\hat{Y} \in \hat{\mathbb{Y}}} C(w_p, \hat{Y}) \tag{16}$$

, where $C(w, \hat{Y})$ is the count of word w in the generated text Y. The vector of counts of all groups, normalized to a probability distribution, can then be compared to a reference distribution probability like the uniform distribution, using metrics like KL divergence, and Wasserstein distance, etc.

Stereotypical Associations [10] is similar to Demographic Representation but measures bias associated with a specific term w. Specifically,

$$ST(w) = \sum_{w_p \in P_i} \sum_{\hat{Y} \in \hat{\mathbb{Y}}} C(w_p, \hat{Y}) \mathbb{I}(C(w, \hat{Y}) > 0) \tag{17}$$

$\mathbb{I}$, the indicator function returns 1 when its arguments are true. The vectors of stereotypical associations of groups are normalized and compared to a reference distribution probability.

Classifier Metrics

The classifier-based metrics use an external classifier to identify any kind of bias in the generated output of LLMs for prompts that are similar but associated with different social groups. Most of the reported metrics have been used in the bibliography for toxicity, but they can be generalized for other bias classification tasks. The Expected Maximum Toxicity (EMT) [30] report the worst-case generations over the generated texts, the Toxicity Probability (TP) [30] which reports the probability of generating at least one toxic text with a toxicity score larger than a threshold (e.g. 0.5), and the Toxic Fraction (TF) [10] which is the fraction of generated texts that are toxic. The above metrics can be adapted to any kind of bias classifier. Formally, considering the classifier function $c : \bar{Y} \to [0, 1]$:

$$EMT(\hat{\mathbb{Y}}) = max_{\bar{Y} \in \hat{\mathbb{Y}}} c(\bar{Y}) \tag{18}$$

$$TP(\hat{\mathbb{Y}}) = P(\sum_{\bar{Y} \in \hat{\mathbb{Y}}} \mathbb{I}(c(\hat{Y}) \geq 0.5) \geq 1) \tag{19}$$

$$TF(\hat{\mathbb{Y}}) = \mathbb{E}_{\bar{Y} \in \hat{\mathbb{Y}}}[\mathbb{I}(c(\hat{Y}) \geq 0.5)] \tag{20}$$

Score Parity [49] measures how consistently a classifier for a specific protected attribute classifies the LLM generated language. Specifically, given the scoring function $c : \hat{Y} x P \to [0, 1]$, where P a protected attribute,

$$ScoreParity(\hat{\mathbb{Y}}) = |\mathbb{E}_{\mathbb{Y} \in \hat{\mathbb{Y}}}[c(\mathbb{Y}_i, i)|P = i] - \mathbb{E}_{\mathbb{Y} \in \hat{\mathbb{Y}}}[c(\mathbb{Y}_j, j)|P = j]| \tag{21}$$

Wasserstein-1 distance between the classified distributions has also been used as in the Counterfactual Sentiment Bias [33].

Lexicon Metrics

The lexicon-base metrics use a precompiled set of biased words or assign a bias score to them. These words are then taken into consideration in a word-level analysis of the generated output. For example, HONEST [42] measures how many top-k completions in templates prompts contain biased words that are in the precompiled lexicon Lex. Formally,

$$HONEST(\hat{\mathbb{Y}}) = \frac{\sum_{\hat{Y}\in\hat{\mathbb{Y}}}\sum_{y_k\in\hat{Y}_k}\mathbb{I}_{Lex}(\hat{y})}{\hat{\mathbb{Y}} * k} \tag{22}$$

The Psycholinguistic Norms [23] and the Gender Polarity [23] use a lexicon with words associated with numeric ratings that measure their affective meaning (e.g. fear) or bias score correspondingly. The metrics compute the weighted average of the psycholinguistic and bias scores of all words in the texts. Formally, for gender bias,

$$GenderPolarity(\hat{\mathbb{Y}}) = \frac{\sum_{\hat{Y}\in\hat{\mathbb{Y}}}\sum_{y_k\in\hat{Y}_k} sign(bias((\hat{y})) * bias(\hat{y})^2}{\sum_{\hat{Y}\in\hat{\mathbb{Y}}}\sum_{y_k\in\hat{Y}_k}|bias(\hat{y})|} \tag{23}$$

Discussion.

An important issue with metrics that measure the generated text as discussed in [3] is the various model parameters such as the decoding parameters or the size of the generated text. So it is important to report the results along with the various parameters. Word associations with protected attributes might not be a reliable proxy for the task at hand and lexicon-based metrics might not be able to capture biases that emerge in sentences and phrases [14]. Finally, classifiers impose their own biases, due to the datasets on which they have been trained and might not be able to capture the dynamicity of the evolution of bias [45].

5. Clustering

Clustering algorithms are unsupervised methods, which aim to separate the data instances into clusters, such that objects in the same cluster are more similar, or close, to each other, dissimilar, or far, from the rest of the objects and their clusters. Clustering is an important data mining procedure, and it is often used as a data processing step for summarization, dimensionality reduction, data analysis, etc. It is thus important to ensure a fair, balanced, and transparent environment in clustering techniques and clustering outputs. Similarly to the classification case, clustering fairness definitions can be divided into group fairness and individual fairness. We now consider the different metrics for these two approaches. Our exposition follows the survey in [18] that organize and categorize the field.

For the following, we assume that we have as input a set of points $X = x_1, x_2, .., x_n$. The output of the clustering is a partition of the points $C = C_1, C_2, ..C_k$, $C_i \subseteq X$, $C_i \cap C_j = \emptyset$ into k clusters. The value of k may be an input to the clustering algorithm, or it may be a value decided by the algorithm.

5.1. Clustering Group Fairness

Group fairness assumes that the input points $X = x_1, x_2, .., x_n$ are partitioned into m groups (colors) $G = g_1, g_2, ..g_m$, $g_i \subseteq X$, $g_i \cap g_j = \emptyset$. These groups are defined by protected attributes associated with the input points. For example, if the points represent individuals these may be attributes like gender, race, religion, etc. For simplicity, for the following we will assume that we have two groups (colors) g_1 and g_2. For a subset of points $Y \subseteq X$, we use Y_{g_1}, Y_{g_2} to denote the set of points in Y that are colored g_1 or g_2. Group fairness requires that all groups should be treated equally in the output clustering. We now present some clustering group fairness metrics.

Balance: The notion of balance, first defined in [20] requires that the clustering produces clusters where the groups are equally or proportionally represented. For a non-empty subset of points

$\emptyset \neq Y \subseteq X$, we define the balance of Y as:

$$\text{balance}(Y) = \min \left\{ \frac{|Y_{g_1}|}{|Y_{g_2}|}, \frac{|Y_{g_2}|}{|Y_{g_1}|} \right\} \in [0, 1] \tag{24}$$

A perfectly balanced subset would have an equal number of points from the groups g_1 and g_2, resulting in a balance value of 1.

Given the definition of a balance of a set of points, we define the **balance of a clustering** $C = \{C_1, \ldots, C_k\}$ as the minimum balance value over all clusters of C. That is,

$$\text{balance}(C) = \min_{C_i \in C} \text{balance}(C_i)$$

The definition of balanced was extended for the case of more than two groups in [47].

Bounded Representation: The Bounded Representation fairness [9] considers the representation ratio of each color (group) in the clusters, and bounds it by the values $a(g_i), b(g_i)$, which are parameters to the definition, for each color g_i.

Formally, for a cluster $C_i \in C$ and the ratio of appearances of the color $g_i \in g$ being $freq(C_i, g_i)$, we define:

$$a(g_i) \leq freq(C_i, g_i) \leq b(g_i) \tag{25}$$

This equation requires that every cluster has at least $a(g_i)$ fraction of nodes with color g_i, and at most $b(g_i)$ fraction of nodes of the same color g_i. There are also fairness definitions that consider an upper bound only [1].

Fair Representation of centers: Another definition of fairness considers specifically the k-Centers algorithm and defines the fairness with respect to the center choice [37]. The intuition of this definition is to have proportional representation of different groups (colors) in the center selection. For every color in g, there must be at least k_i centers of this color, thus, avoiding over-representation or under-representation of any group in the set of selected centers.

Proportionality: Proportionality fairness [17] is defined for center-based clustering. The definition of proportionality states that a clustering $C = \{C_1, \ldots, C_k\}$ with k clusters and X the set of all points, is fair, if for every subset $\frac{|X|}{k}$ of points, there does not exists a center that is closer to all its members than their center.

Fairness with outliers: There are also fairness definitions for clustering with outliers [6]. The output of the algorithm in this case is set of outlier points, and a clustering of the remaining points, after outliers have been removed. Fairness is studied for the k-center problem. The main idea is that the outliers should not belong to a single group, thus depleting the points of the group that are being clustered. Otherwise, the algorithm is considered unfair.

To define fairness, for each group g_i of the g groups, we define a parameter m_i. Let A denote the set of points to be clustered, after the removal of the outliers. We require that $|A \cap g_i| \geq m_i$ for all groups g_i. Different m_i values can capture a plethora of fairness scenarios.

Social Fairness Cost: Social Fairness Cost [31] defines a measure for fairness for the popular k-means algorithm. Recall that the k-means algorithm, with input X produces a clustering $C = \{C_1, ..., C_k\}$, with centers $\{c_1, ..., c_k\}$ that minimizes the cost

$$O(C, X) = \sum_{x \in X} \min_{c_i \in C} ||x - c_i||^2$$

The fair version of k-means objective looks at the cost for the different groups in the data. That is,

$$\Phi(C, X) = \max_{g_i \in g} \frac{O(C, X_{g_i})}{|X_{g_i}|}$$

Since social fairness refers to a cost, it needs to be minimized (unlike balance, which might be maximized). Therefore, lower social fairness cost indicates fairer clustering.

Maximum Fairness Cost: Another notion of fairness is the Maximum Fairness Cost(MFC) [19]. It assumes a parameter m_i for each group $g_i \in G$, which is the ideal fraction of points from group i in each cluster. Let P_{C_j,g_i} be the actual fraction of points from group g_i in cluster $C_j \in C$. Then the Maximum Fairness Cost MFC is defined as

$$MFC = \max_{C_j \in C} \sum_{g_i \in [G]} |P_{C_j,g_i} - m_i| \tag{26}$$

This metric has also been applied to Hierarchical clustering, where we evaluate the metric at each level of the hierarchy.

5.2. Clustering Individual Fairness

Individual fairness is based on the principle "treat similar individuals similarly" [26], or that each individual should receive fair treatment without disrespecting the needs of others. For the following we assume a distance metric $\psi_{x_i,x_j} \in [0,1]$ between the points in the input data X that captures their similarity. This value can be different from the distance metric used for clustering the data.

Probabilistic Pairwise Fairness: In [13] they assume a randomized strategy for assigning points to clusters. Let $\varphi(x_i) : X \to C$ denote the assignment of points to clusters, sampled from a distribution $\mathcal{D}$ over all possible assignments. Their notion of fairness requires that the assignment separates the two points x_i, x_j with probability at most ψ_{x_i,x_j}. Specifically,

$$\Pr_{\varphi \sim \mathcal{D}}[\varphi(x_i) \neq \varphi(x_j)] \leq \psi_{x_i,x_j} \tag{27}$$

Distributional Individual Fairness: The definition in [5] assumes again a probabilistic assignment over the clusters in C. For a point x_i let ϕ_{x_i} denote the probability distribution of the assignment of x_i over the clusters in C. Given the metric D that measures the statistical proximity of two distributions, fairness requires that for each $x_i \in X$:

$$D(\phi_{x_i}, \phi_{x_j}) \leq \psi_{x_i,x_j} \tag{28}$$

α -Equitable k-Center: This fairness definition [16] is defined for the k-center objective but can also be applied to any center-based clustering. For every point $x_i \in X$ we assume a set of other points $S_{x_i} \subseteq X$, which are close (similar) to x_i. Abusing the notation, let $C\{c_1, c_2, \ldots, c_k\}$ denote the k centers produced by the clustering algorithm, and let $\varphi : X \to C$ denote the assignment of points to cluster centers. Given a value $\alpha \geq 1$, and a distance function d the fairness definition requires that $\forall x_i \in X, \forall x_j \in S_i$

$$d(x_i, \varphi(x_i)) \leq \alpha * d(x_j, \varphi(x_j)) \tag{29}$$

The smaller the α is, the more fair the achieved separation.

A center in my neighborhood: This definition [34] deviates from the principle "treat similar individuals similarly". The definition assumes cluster centers. Similar to the previous definition, let $C\{c_1, c_2, \ldots, c_k\}$ denote the k centers produced by the clustering algorithm, and let $\varphi : X \to C$ denote the assignment of points to cluster centers. For each point x_i we assume a parameter r_{x_i}, which is the acceptable radius for point x_i. The fairness definition requires that for each point x_i:

$$d(x_i, \varphi(x_i)) \leq r_{x_i} \tag{30}$$

Kleindessner et al individual fairness: This definition [36] requires that the distance of a point from the cluster is smaller or equal than the distance of a point from another cluster. Specifically: $\forall C_i \in C, \forall x \in C_i$:

$$\frac{1}{|C_i|-1}\sum_{x'\in C_i} d(x,x') \leq \frac{1}{\sum_{j\neq i}|C_j|}\sum_{x'\in C'_j} d(x,x') \tag{31}$$

6. Network Analysis

In our modern society, networks are all around us and influence our daily lives: social networks, transportation networks, supply chain networks, power grid networks, etc. In the era of Big Data, these networks produce massive amounts of data which we desire to analyse in a regular basis in order to make informed decisions regarding their evolution. Given the increasing interest in network analysis expressed by the numerous diverse application domains, a number of algorithms have been proposed over the years to perform the various network analysis tasks. However, most of these algorithms do not consider fairness. As the end-user applications may be life-changing, such as loan approval, or mission-critical, such as disaster response, ensuring that the algorithms used in these applications treat all groups and individuals in a fair manner is paramount.

Ensuring fairness in the context of network data is challenging, because network data is not independent and identically distributed (i.i.d.), so fairness notions established for i.i.d. data do not directly apply to network data. Moreover, each network analysis task introduces its own unique sources of bias, driving researchers in the field to propose specialized metrics of said bias. In this report, we focus on the fairness of two specific network tasks: Diffusion maximization and PageRank. For further reading, we point to the survey by Dong et al. [25].

In what follows, we are given a network in the form of a (directed) graph $G = (V, E)$. We assume that the nodes are associated with a sensitive attribute, e.g. gender or race for the case of a social network, that has P distinct values which we identify with the integers of $[P]$. Let $\{V_i\}_{i\in[P]}$ be the partition of V into P groups with respect to the sensitive attribute where V_i denotes the group of nodes for which the value of this attribute is i. For most of the discussion below, we assume a binary sensitive attribute.

6.1. Diffusion Maximization Fairness

Here, we consider a discrete-time random process for diffusing information in the network with the following properties: A subset S of V, which is called the *seed set*, is being selected to initially possess the information. A node that possesses the information is said to be *active*, otherwise it is said to be *inactive*. Inactive nodes can be activated, but active nodes cannot be deactivated. Nodes can only be activated by their neighbors as a result of passing information through the edges of the network. For example, in the *independent cascade* model, in every round of the random process, every active node activates each one of its inactive neighbors with independent probability p. This implies that if an inactive node has d active neighbors, then it is activated with probability $1-(1-p)^d$. The random process terminates after a round in which no new nodes were activated. In the *diffusion maximization* problem, we are given the graph G and a budget K and we are asked to find a seed set S such that $|S| \leq K$ and the expected number of active nodes of V after the random process termination[1], which is called the *spread*, is maximized.

For subsets S, T of V and a subgraph H of G, we let $U_{H,T}(S)$ denote the spread among the nodes of T through the edges of H with seed set S and we set $N_{H,T}(S) = U_{H,T}(S)/|T| \in [0,1]$ which is the ratio of the nodes of T that are contained in the expected set of active nodes for the same instance. We also set $S_i = S \cap V_i$ for every $i \in [P]$.

6.1.1. Group Fairness

Maximin Fairness Tsang et al. [52] propose two notions of group fairness in the composition of the expected set of active nodes. Their first notion, *maximin fairness*, is based on the Rawlsian

[1] As all expectancies in this section are considered after the random process termination, we will hereafter omit this clarification for brevity.

Theory of Justice principle that inequalities are to be arranged to the greatest benefit of the least advantaged. Maximin fairness requires that

$$N^{\min}_{G,V}(S) = \max_{S' \subseteq V : |S'| \leq K} N^{\min}_{G,V}(S'), \tag{32}$$

where $N^{\min}_{G,V}(S) = \min_{i \in [P]} N_{G,V_i}(S)$ is the *Maximin Fairness* metric.

Group Rationality The second notion of Tsang et al. [52], *group rationality*, is based on the authors' view that no group should prefer to be given its fair set of seeds and diffuse information only in its own subnetwork instead of contributing in the diffusion of information in the whole network. For every $i \in [P]$, we let K_i denote the share of the budget that would be fairly allocated to V_i according to demographic parity, that is, we set $K_i = K|V_i|/|V|$. Group rationality requires that the following constraints are satisfied.

$$\forall i \in [P] : U_{G,V_i}(S) \geq \max_{S'_i \subseteq V_i : |S'_i| \leq K_i} U_{G[V_i],V_i}(S'_i) \tag{33}$$

For a group for which the respective constraint is satisfied, internal diffusion of information is not preferable under any choice of seed set where the group is being allocated its fair share of the budget. The *Diversity Constraints* metric is the ratio of the P constraints of Eq. (33) that is violated.

Demographic Parity Stoica et al. [51] consider group fairness in the composition of the seed set as well as in the composition of the expected set of active nodes. For the seed set, their definition of group fairness requires that the following equalities hold.

$$\forall i, j \in [P] : \frac{|S_i|}{|V_i|} = \frac{|S_j|}{|V_j|}$$

It is not difficult to see that this is equivalent to requiring that the composition of the seed set is according to demographic parity, that is:

$$\forall i \in [P] : \frac{|S_i|}{|S|} = \frac{|V_i|}{|V|}$$

To measure bias in the composition of the seed set, the authors implicitly consider the metric

$$\max_{i \in [P]} \left| \frac{|S_i|}{|S|} - \frac{|V_i|}{|V|} \right|,$$

which we call the *Maximum Demographic Disparity in Seeding* metric. For the expected set of active nodes, their definition of group fairness requires that the following equalities hold.

$$\forall i, j \in [P] : N_{G,V_i}(S) = N_{G,V_j}(S)$$

As before, it is not difficult to see that this is equivalent to requiring that the composition of the expected set of active nodes is according to demographic parity, that is:

$$\forall i \in [P] : \frac{U_{G,V_i}(S)}{U_{G,V}(S)} = \frac{|V_i|}{|V|}$$

To measure bias in the composition of the expected set of active nodes, the authors implicitly consider the metric

$$\max_{i \in [P]} \left| \frac{U_{G,V_i}(S)}{U_{G,V}(S)} - \frac{|V_i|}{|V|} \right|,$$

which we call the *Maximum Demographic Disparity in Diffusion* metric.

Ali et al. [4] introduce another metric for measuring bias in the composition of the expected set of active nodes. This metric is proposed based on the legal concept of *disparate impact* which refers to policies that result in a substantially different treatment of members of a group despite appearing neutral at face value. The *Maximum Disparity in Normalized Utilities* metric is $\max_{i,j \in [P]} |N_{G,V_i}(S) - N_{G,V_j}(S)|$.

6.1.2. Individual Fairness

Maximin Fairness Fish et al. [28] propose *maximin fairness* as a notion of individual fairness in the composition of the expected set of active nodes. Maximin fairness requires that Eq. (32) is satisfied in this context as well, but the *Maximin Fairness* metric is $N^{\min}_{G,V}(S) = \min_{u \in V} N_{G,\{u\}}(S)$.

6.2. PageRank Fairness

A important network analysis task is to rank the nodes of the network with respect to their importance. *Google*'s celebrated *PagePank* algorithm [12] accomplishes this task by performing a special kind of random process on the network graph, which is called a random walk with restarts. A *(1^{st}-order) random walk* on the graph G is a *(1^{st}-order) Markov chain* on V with transition probability matrix $\mathbf{P}$ such that $\mathbf{P}_{uv} \neq 0$ if and only if $uv \in E$. As $\mathbf{P}$ is not row-stochastic if a node of G has no outgoing edges, we treat such nodes differently: for every node $u \in V$ with no outgoing edges, we define $\mathbf{P}_{uv} = 1/|V|$ for all $v \in V$, that is, we define the corresponding row to be the transposed uniform probability vector. A *random walk with restarts* on G is a generalization of the previous notion where we are also allowed to restart the random walk with probability γ from a node chosen according to a probability vector $\mathbf{v}$ — this action is also referred to as a *jump*. An algorithm that takes a graph as input and performs a random walk with restarts on the graph is called a *PageRank* algorithm. *Google*'s original PagePank algorithm performs the random walk with restarts with

- $\mathbf{P}$ being the row-normalized adjacency matrix of the input graph rectified as discussed above,
- $\gamma = 0.15$ and
- $\mathbf{v}$ being the uniform probability vector.

The stationary probability vector $\mathbf{p}$ of a random walk with restarts on G with transition probability matrix $\mathbf{P}$, restart probability γ and jump probability vector $\mathbf{v}$ satisfies the equation

$$\mathbf{p}^T = (1-\gamma)\mathbf{p}^T\mathbf{P} + \gamma\mathbf{v}^T \tag{34}$$

and thus $\mathbf{p}$ can be computed by solving Eq. (34) analytically. It is also known that the probability vector sequence $(\mathbf{p}_t)_{t \in \mathbb{N}}$ where $\mathbf{p}^T_{t+1} = (1-\gamma)\mathbf{p}^T_t\mathbf{P} + \gamma\mathbf{v}^T$ for all $t \in \mathbb{N}$ converges to the vector $\mathbf{p}$ independently of the choice of initial probability vector $\mathbf{p}_0$.

6.2.1. Group Fairness

Tsioutsiouliklis et al. [53] initiate the study of bias in PageRank by considering the case of a binary sensitive attribute of which one value identifies a protected group. The authors propose the following two definitions of PageRank fairness, where $\phi \in (0, 1)$ is a parameter.

ϕ-fairness A PageRank algorithm is ϕ-fair if $\mathbf{p}$ allocates probability mass ϕ to the nodes of the protected group.

Local ϕ-fairness A PageRank algorithm is locally ϕ-fair if every transposed row of $\mathbf{P}$ allocates probability mass ϕ to the nodes of the protected group. This means that every node of G distributes its probability mass in a ϕ-fair manner in every iteration.

The authors also prove that if a RageRank algorithm is locally ϕ-fair, then it is also ϕ-fair. Tsioutsiouliklis et al. [54] use the probability mass that $\mathbf{p}$ allocates to the nodes of the protected group to measure bias in $\mathbf{p}$ and they refer to it as the *PageRank Ratio*.

7. Conclusion

In this report we presented a comprehensive survey of the different metrics for algorithmic bias and fairness. We presented the general approaches for defining of fairness, namely, Individual Fairness, Group Fairness, and Casual Fairness. Then, we presented metrics for specific Machine Learning problems, specifically, Classification, Large Language Models, Clustering, and Network Analysis.

References

[1] Sara Ahmadian et al. "Fair Hierarchical Clustering". In: *CoRR* abs/2006.10221 (2020). arXiv: 2006.10221. URL: https://arxiv.org/abs/2006.10221.

[2] Jaimeen Ahn and Alice Oh. "Mitigating language-dependent ethnic bias in BERT". In: *arXiv preprint arXiv:2109.05704* (2021).

[3] Afra Feyza Akyürek et al. "Challenges in measuring bias via open-ended language generation". In: *arXiv preprint arXiv:2205.11601* (2022).

[4] Junaid Ali et al. "On the Fairness of Time-Critical Influence Maximization in Social Networks". In: *IEEE Transactions on Knowledge and Data Engineering* 35.3 (2023), pp. 2875–2886. DOI: 10.1109/TKDE.2021.3120561.

[5] Nihesh Anderson et al. *Distributional Individual Fairness in Clustering*. 2020. arXiv: 2006.12589 [cs.LG]. URL: https://arxiv.org/abs/2006.12589.

[6] Sayan Bandyapadhyay et al. *A Constant Approximation for Colorful k-Center*. 2019. arXiv: 1907.08906 [cs.DS]. URL: https://arxiv.org/abs/1907.08906.

[7] Soumya Barikeri et al. "RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models". In: *arXiv preprint arXiv:2106.03521* (2021).

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[9] Suman Bera et al. "Fair Algorithms for Clustering". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/fc192b0c0d270dbf41870a63a8c76c2f-Paper.pdf.

[10] Rishi Bommasani, Percy Liang, and Tony Lee. "Holistic evaluation of language models". In: *Annals of the New York Academy of Sciences* 1525.1 (2023), pp. 140–146.

[11] Shikha Bordia and Samuel R Bowman. "Identifying and reducing gender bias in word-level language models". In: *arXiv preprint arXiv:1904.03035* (2019).

[12] Sergey Brin and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine". In: *Computer Networks and ISDN Systems* 30.1 (1998). Proceedings of the Seventh International World Wide Web Conference, pp. 107–117. ISSN: 0169-7552. DOI: https://doi.org/10.1016/S0169-7552(98)00110-X. URL: https://www.sciencedirect.com/science/article/pii/S016975529800110X.

[13] Brian Brubach et al. "A pairwise fair and community-preserving approach to k-center clustering". In: *International conference on machine learning*. PMLR. 2020, pp. 1178–1189.

[14] Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. "On the independence of association bias and empirical fairness in language models". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 370–378.

[15] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[16] Darshan Chakrabarti et al. *A New Notion of Individually Fair Clustering: α-Equitable k-Center*. 2022. arXiv: 2106.05423 [cs.LG]. URL: https://arxiv.org/abs/2106.05423.

[17] Xingyu Chen et al. “Proportionally Fair Clustering”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 1032–1041. URL: https://proceedings.mlr.press/v97/chen19d.html.

[18] Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. “An overview of fairness in clustering”. In: *IEEE Access* 9 (2021), pp. 130698–130720.

[19] Anshuman Chhabra and Prasant Mohapatra. *Fair Algorithms for Hierarchical Agglomerative Clustering*. 2023. arXiv: 2005.03197 [cs.LG]. URL: https://arxiv.org/abs/2005.03197.

[20] Flavio Chierichetti et al. “Fair clustering through fairlets”. In: *Advances in neural information processing systems* 30 (2017).

[21] Pieter Delobelle et al. “Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models”. In: *arXiv preprint arXiv:2112.07447* (2021).

[22] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[23] Jwala Dhamala et al. “Bold: Dataset and metrics for measuring biases in open-ended language generation”. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 862–872.

[24] Tommaso Dolci, Fabio Azzalini, and Mara Tanelli. “Improving gender-related fairness in sentence encoders: A semantics-based approach”. In: *Data Science and Engineering* 8.2 (2023), pp. 177–195.

[25] Yushun Dong et al. “Fairness in Graph Mining: A Survey”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.10 (2023), pp. 10583–10602. DOI: 10.1109/TKDE.2023.3265598.

[26] Cynthia Dwork et al. *Fairness Through Awareness*. 2011. arXiv: 1104.3913 [cs.CC]. URL: https://arxiv.org/abs/1104.3913.

[27] Emilio Ferrara. “Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies”. In: *Sci* 6.1 (Dec. 2023), p. 3. ISSN: 2413-4155. DOI: 10.3390/sci6010003. URL: http://dx.doi.org/10.3390/sci6010003.

[28] Benjamin Fish et al. “Gaps in Information Access in Social Networks?” In: *The World Wide Web Conference*. WWW ’19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 480–490. ISBN: 9781450366748. DOI: 10.1145/3308558.3313680. URL: https://doi.org/10.1145/3308558.3313680.

[29] Isabel O Gallegos et al. “Bias and fairness in large language models: A survey”. In: *arXiv preprint arXiv:2309.00770* (2023).

[30] Samuel Gehman et al. “Realtoxicityprompts: Evaluating neural toxic degeneration in language models”. In: *arXiv preprint arXiv:2009.11462* (2020).

[31] Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. *Socially Fair k-Means Clustering*. 2020. arXiv: 2006.10085 [cs.LG]. URL: https://arxiv.org/abs/2006.10085.

[32] Wei Guo and Aylin Caliskan. “Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 122–133.

[33] Po-Sen Huang et al. “Reducing sentiment bias in language models via counterfactual evaluation”. In: *arXiv preprint arXiv:1911.03064* (2019).

[34] Christopher Jung, Sampath Kannan, and Neil Lutz. *A Center in Your Neighborhood: Fairness in Facility Location*. 2019. arXiv: 1908.09041 [cs.DS]. URL: https://arxiv.org/abs/1908.09041.

[35] Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. “Debiasing isn’t enough!–On the Effectiveness of Debiasing MLMs and their Social Biases in Downstream Tasks”. In: *arXiv preprint arXiv:2210.02938* (2022).

[36] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. *A Notion of Individual Fairness for Clustering*. 2020. arXiv: 2006.04960 [stat.ML]. URL: https://arxiv.org/abs/2006.04960.

[37] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. "Fair k-Center Clustering for Data Summarization". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 3448–3457. URL: https://proceedings.mlr.press/v97/kleindessner19a.html.

[38] Keita Kurita et al. "Measuring bias in contextualized word representations". In: *arXiv preprint arXiv:1906.07337* (2019).

[39] Chandler May et al. "On measuring social biases in sentence encoders". In: *arXiv preprint arXiv:1903.10561* (2019).

[40] Moin Nadeem, Anna Bethke, and Siva Reddy. "StereoSet: Measuring stereotypical bias in pretrained language models". In: *arXiv preprint arXiv:2004.09456* (2020).

[41] Nikita Nangia et al. "CrowS-pairs: A challenge dataset for measuring social biases in masked language models". In: *arXiv preprint arXiv:2010.00133* (2020).

[42] Debora Nozza, Federico Bianchi, Dirk Hovy, et al. "HONEST: Measuring hurtful sentence completion in language models". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. 2021.

[43] Eirini Ntoutsi et al. "Bias in data-driven artificial intelligence systems—An introductory survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.3 (2020), e1356.

[44] Drago Plecko and Elias Bareinboim. *Causal Fairness Analysis*. 2022. arXiv: 2207.11385 [cs.AI]. URL: https://arxiv.org/abs/2207.11385.

[45] Luiza Pozzobon et al. "On the challenges of using black-box apis for toxicity evaluation in research". In: *arXiv preprint arXiv:2304.12397* (2023).

[46] Pranav Rajpurkar et al. "Squad: 100,000+ questions for machine comprehension of text". In: *arXiv preprint arXiv:1606.05250* (2016).

[47] Clemens Rösner and Melanie Schmidt. *Privacy preserving clustering with constraints*. 2018. arXiv: 1802.02497 [cs.CC]. URL: https://arxiv.org/abs/1802.02497.

[48] Julian Salazar et al. "Masked language model scoring". In: *arXiv preprint arXiv:1910.14659* (2019).

[49] Anthony Sicilia and Malihe Alikhani. "Learning to generate equitable text in dialogue from biased training data". In: *arXiv preprint arXiv:2307.04303* (2023).

[50] Karolina Stanczak and Isabelle Augenstein. "A survey on gender bias in natural language processing". In: *arXiv preprint arXiv:2112.14168* (2021).

[51] Ana-Andreea Stoica and Augustin Chaintreau. "Fairness in Social Influence Maximization". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. San Francisco, USA: Association for Computing Machinery, 2019, pp. 569–574. ISBN: 9781450366755. DOI: 10.1145/3308560.3317588. URL: https://doi.org/10.1145/3308560.3317588.

[52] Alan Tsang et al. "Group-Fairness in Influence Maximization". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 5997–6005. DOI: 10.24963/ijcai.2019/831. URL: https://doi.org/10.24963/ijcai.2019/831.

[53] Sotiris Tsioutsiouliklis et al. "Fairness-Aware PageRank". In: *Proceedings of the Web Conference 2021*. WWW '21. Ljubljana, Slovenia: Association for Computing Machinery, 2021, pp. 3815–3826. ISBN: 9781450383127. DOI: 10.1145/3442381.3450065. URL: https://doi.org/10.1145/3442381.3450065.

[54] Sotiris Tsioutsiouliklis et al. “Link Recommendations for PageRank Fairness”. In: *Proceedings of the ACM Web Conference 2022*. WWW ’22. Virtual Event, Lyon, France: Association for Computing Machinery, 2022, pp. 3541–3551. ISBN: 9781450390965. DOI: 10.1145/3485447.3512249. URL: https://doi.org/10.1145/3485447.3512249.

[55] Sahil Verma and Julia Rubin. “Fairness definitions explained”. In: *Proceedings of the International Workshop on Software Fairness*. FairWare ’18. Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 1–7. ISBN: 9781450357463. DOI: 10.1145/3194770.3194776. URL: https://doi.org/10.1145/3194770.3194776.

[56] Alex Wang and Kyunghyun Cho. “BERT has a mouth, and it must speak: BERT as a Markov random field language model”. In: *arXiv preprint arXiv:1902.04094* (2019).

[57] Liang Wang et al. “Improving text embeddings with large language models”. In: *arXiv preprint arXiv:2401.00368* (2023).

[58] Kellie Webster et al. “Measuring and reducing gendered correlations in pre-trained models”. In: *arXiv preprint arXiv:2010.06032* (2020).