Fairness Perception from a Network-Centric Perspective

Farzan Masrour, Pang-Ning Tan, Abdol-Hossein Esfahanian
Department of Computer Science and Engineering, Michigan State University
Emails:{masrours, ptan, esfahanian}@msu.edu

***Abstract*—Algorithmic fairness is a major concern in recent years as the influence of machine learning algorithms becomes more widespread. In this paper, we investigate the issue of algorithmic fairness from a network-centric perspective. Specifically, we introduce a novel yet intuitive function known as fairness perception and provide an axiomatic approach to analyze its properties. Using a peer-review network as a case study, we also examine its utility in terms of assessing the perception of fairness in paper acceptance decisions. We show how the function can be extended to a group fairness metric known as fairness visibility and demonstrate its relationship to demographic parity. We also discuss a potential pitfall of the fairness visibility measure that can be exploited to mislead individuals into perceiving that the algorithmic decisions are fair. We demonstrate how the problem can be alleviated by increasing the local neighborhood size of the fairness perception function.**

I. Introduction

The influence of machine learning is pervasive across numerous applications, from healthcare and e-commerce to financial and criminal justice systems. Despite its utility, previous studies have shown that the algorithmic decisions may contain unintended biases that discriminate against certain groups of the population [1], [2], [3]. As a result, the challenge of removing biases from the algorithmic decision-making process has gained significant attention in recent years. In particular, various mathematical formulations of fairness have been proposed. For instance, group fairness metrics such as demographic parity and equalized odds have been developed to assess the degree of prejudice against certain protected groups in the population. While each metric has its own merits, many of them are incompatible with each other [4], [5].

The group fairness metrics are designed to determine the level of equity among different groups of individuals who are harmed by or benefited from the algorithmic-driven decisions. However, the consequences of an unfair decision may extend beyond those individuals who are directly impacted by the decision. In fact, they may elicit negative responses from other individuals who identified themselves to be in the same group as the affected individuals. For instance, hiring discrimination against a qualified member from an underrepresented group not only affects the well being of that individual, but will also have an adverse effect on other members of the underrepresented group who observed such behavior. This example suggests that fairness assessment must take into consideration the perception of other individuals who may not be directly impacted by the algorithmic decisions [6], [7].

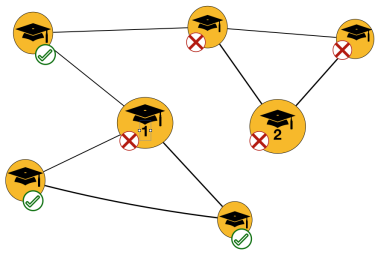


Fig. 1: An example illustration of fairness perception.

Fairness perception is rooted in the social comparison theory. For instance, equity theory [8] argues that "humans do not base their satisfaction on what they receive but rather what they receive in relation to what they think they should receive". The reaction of an individual to the outcome of a decision process is based on the expectation of the individual and this expectation not only depends on one's own outcome but also the outcomes of other individuals they are aligned with, which we refer to as the *reference group*. The choice of the reference group is typically influenced by the similarity measure we use, e.g., we may compare ourselves to our co-workers, friends, and family members. By observing the outcomes of other members in our reference group, this will help shape our expectation about what should be considered a fair outcome.

In this paper, we examine the notion of fairness perception from a network analysis perspective. Networks provide a natural way to represent individuals and their connections to other individuals in the same reference group. For instance, Figure 1 shows a toy example of a network of students applying for college admission to a prestigious university. In this network, two students are linked if they know each other. Suppose the admission committee of the college has decided to accept 3 of the applicants, denoted as nodes with green check marks, and to reject the other 4 applicants. For brevity, we assume all the students have similar qualifications. Consider the two students labeled as 1 and 2, respectively. Although both applicants were rejected, their expectations for admission and perceptions of fairness are very different. Student 1 has a higher expectation of being admitted compared to student 2 since all of his/her friends were accepted. Thus, the perception of fairness for student 1 is different than that for student 2.

This paper introduces the notion of network-centric fairness perception and illustrates its application to peer review process. Peer evaluation of scientific work has a significant

effect on scientific advancement. However, similar to other systems designed by humans, it is potentially biased, favoring certain groups of individuals (e.g., famous researchers from top institutions) [9], [10]. In this study, we show how the proposed function can be used to assess the perception of authors about paper acceptance decisions. An axiomatic approach for analyzing the desirable properties of fairness perception functions is also presented. We then extend our proposed function to a group fairness measure known as fairness visibility and show its relationship to demographic parity under certain mild assumptions. We also describe a potential pitfall of assessing fairness from a local neighborhood perspective. Specifically, it can mislead individuals into thinking that the decision-making process is fair even though the overall decisions are biased toward certain groups of individuals. Finally, we show how to alleviate the problem by expanding the local neighborhood size of the fairness perception function.

II. Quantifying Fairness Perception

Let $G =< V, E, X >$ be an attributed network, where V is set of nodes, $E \subseteq V \times V$ is the set of links (edges), and $X \in \mathbb{R}^{|V| \times d}$ is the feature matrix associated with the nodes. We further assume that $X = (X^{(p)}, X^{(u)})$, where $X^{(p)}$ are the protected attributes and $X^{(u)}$ are the unprotected ones. The set of links can also be represented by an adjacency matrix, A, where $A_{ij} = 1$ if a link exists between nodes i and j. Furthermore, we denote $A^k = \prod_{i=1}^{k} A$, where $A^k_{ij} > 0$ if there exists a path of length k between i and j, and 0 otherwise.

We also assume that each node v is associated with a target outcome, $y_v \in \{0, 1\}$. As an example, in the context of peer review network, each node corresponds to a submitted paper and links between papers are established if the two papers share the same authors or have authors who had previously collaborated with each other. The outcome y_v of a given paper v may indicate whether the paper is acceptable or unacceptable based on the average ratings provided by reviewers.

We assume there exists a decision function $h : V \rightarrow \{0, 1\}$ associated with each node in the network. Let $\mathcal{H}$ be the hypothesis space of all decision functions. Our goal is to learn a decision function $h \in \mathcal{H}$ that is consistent with the set of outcomes Y while satisfying some fairness criterion. From the perspective of peer review network, the decision function h may refer to the final decision whether to accept or reject the paper. The true positive and false positive rates of the binary decision function, h, can be computed as follows:

- **True positive rate**, TPR = $\frac{\sum_v y_v h(v)}{\sum_v y_v}$
- **False positive rate**, FPR = $\frac{\sum_v (1-y_v) h(v)}{\sum_v y_v}$

Our goal is to determine how the final decisions are perceived by the individual nodes in the network. Do they feel that the decisions are biased toward nodes that belong to certain groups? To answer this question, we assume each node v is associated with a fairness perception function, $f(v, h)$, given a decision function h. The function provides a local, albeit myopic, view on individual fairness of the nodes in a network.

A. Axioms for Fairness Perception

We first outline the desirable properties of the fairness perception function, $f(v, h)$, using the following set of axioms. We assume each node $v \in V$ is associated with the following tuple, $(v.X_p, v.X_u, y_v, N(v))$, where $v.X_p$ denotes the value of its protected attribute, $v.X_u$ denotes the value of its other (unprotected) attributes, y_v denotes its target outcome, and $N(v)$ denotes its δ-neighborhood, which is defined as follows:

$$N(v) = \{u \mid \exists k \leq \delta : A^k_{uv} > 0\}. \quad (1)$$

For brevity, we assume $\delta = 1$, unless stated otherwise. Let $G_r = (V_r, E_r, X_r)$ be an ego-network for node r, where $V_r = N(r)$ is the 1-neighborhood of r, $E_r = \{(i, j) \mid i, j \in N(r) \text{ and } (i, j) \in E\}$ and X_r is the feature matrix associated with the attributes of the nodes in V_r. We present a set of axioms on the fairness perception function.

1) **Locality axiom**: If $h(v) = h'(v)$ and $\forall u \in N(v) : h(u) = h'(u)$, where $h, h' \in \mathcal{H}$, then $f(v, h) = f(v, h')$.
2) **Monotonicity axiom**: If $h(v) < h'(v)$ and $\forall u \in N(v) : h(u) = h'(u)$, where $h, h' \in \mathcal{H}$, then $f(v, h) \leq f(v, h')$.
3) **Neighborhood expectation axiom**: If $h(v) = h'(v)$ and $\forall u \in N(v) : h(u) \leq h'(u)$, where $h, h' \in \mathcal{H}$, then $f(v, h) \geq f(v, h')$.
4) **Homogeneity axiom**: Let G_u and G_v be the induced sub-graphs of $V_u = N(u) \cup \{u\}$ and $V_v = N(v) \cup \{v\}$, respectively. If G_u and G_v are isomorphic with respect to the decision function h, then $f(u, h) = f(v, h)$.

For the last axiom, we say that a pair of networks, $G_r = (V_r, E_r, X_r)$ and $G_s = (V_s, E_s, X_s)$, are isomorphic with respect to the decision function h if there exists a bijection function $m : V_r \rightarrow V_s$ such that:

- $\forall u \in V_r : h(u) = h(m(u)), y_u = y_{m(u)}$ and $X_u = X_{m(u)}$.
- $\forall (u_1, u_2) \in E_r : (m(u_1), m(u_2)) \in E_s$

The locality axiom states that the perception of fairness for an individual depends on the decision outcomes for other individuals in its neighborhood. As long as the outcomes for the node and its neighborhood remains unchanged, the fairness perception function should remain the same. The monotonicity axiom suggests that the perception of fairness for an individual never decreases if the decision changes in favor of the individual (assuming the decisions for its neighbors remain unchanged). For example, if a previous decision on the paper was overturned (say from reject to accept), then one should expect the fairness perception to improve (or at least stays the same). In contrast, the neighborhood expectation axiom states that if the number of neighbors with favorable decisions increases, then fairness perception decreases monotonically. This is because, if more individuals in our reference group received favorable decisions, we expect the decision outcome to be favorable for us as well. The increased expectation makes it less likely for us to perceive the decision as fair if our paper is rejected. The fourth axiom ensures consistency of the fairness perception function when applied to different nodes in the network, The axiom states that if two disparate nodes with

similar neighborhoods receive the same decision outcomes, their perception of fairness should be the same.

B. Proposed Network-Centric Fairness Perception

Definition 1 (Network-Centric Fairness Perception): Given a network $G =< V, E, X >$ and a decision function h, the network-centric fairness perception function is defined as:

$$f(v,h) = \begin{cases} 1 & \text{if } \mathbb{E}[h(v)] \leq h(v) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbb{E}[h(v)]$ is the expected value of $h(v)$, which must satisfy the following properties:

1) If $\forall u \in N(v) : h(u) = h'(u)$, then $\mathbb{E}[h(v)] = \mathbb{E}[h'(v)]$.
2) If $\forall u \in N(v) : h(u) \leq h'(u)$, then $\mathbb{E}[h(v)] \leq \mathbb{E}[h'(v)]$.
3) Let G_u and G_v be the the induced sub-graphs based on the node sets $V_u = N(u) \cup \{u\}$ and $V_v = N(v) \cup \{v\}$, respectively. If G_u and G_v are isomorphic with respect to the decision function h, then $\mathbb{E}[h(v)] = \mathbb{E}[h(u)]$.

Our fairness perception function can thus be viewed as a local measure of individual fairness for any given node v in a network. If the decision $h(v)$ is more favorable than expected, then v will perceive the decision as fair. Furthermore, the expected value of the decision outcome, $\mathbb{E}[h(v)]$, depends on the neighborhood of the node v.

Theorem 1: The network-centric fairness perception function given in Eqn. (2) satisfies the locality, monotonicity, neighborhood expectation, and homogeneity axioms.

Proof: The locality and monotonicity properties are proven using the first property. Since $\mathbb{E}[h(v)]$ remains unchanged when $h(u) = h'(u)$ for all the nodes u in the neighborhood $N(v)$, Eqn. (2) suggests that $f(v,h)$ depends only on $h(v)$. If $h(v) = h'(v)$, then $f(v,h) = f'(v,h)$, thereby proving that the locality axiom holds. Similarly, if $h(v) < h'(v)$, then $f(v,h) \leq f(v,h')$, which satisfies the monotonicity axiom. For the neighborhood expectation axiom, the second property states that the expected value monotonically decreases when $h(u) \leq h'(u)$ for all the nodes $u \in N(v)$. Since $\mathbb{E}[h'(v)]$ is larger, then nodes that initially satisfy the inequality $\mathbb{E}[h(v)] \leq h(v)$ may no longer do so since $h'(v) = h(v)$. Thus, $f(v,h) \geq f(v,h')$. Finally, we use the the third property to prove the homogeneity axiom. Let G_u and G_v be the induced sub-graphs based on node sets $V_u = N(u) \cup \{u\}$ and $V_v = N(v) \cup \{v\}$, respectively. Since G_u and G_v are isomorphic with respect to the decision function h and $\mathbb{E}[h(u)] = \mathbb{E}[h(v)]$ holds due to the third property, therefore $f(u,h) = f(v,h)$. ■

We consider the following **neighborhood peer expectation** approach to compute $\mathbb{E}[h(v)]$:

$$\mathbb{E}[h(v)] = \frac{y_v}{k_1}\Big[\sum_{u \in N(v)} y_u h(u)\Big] + \frac{1-y_v}{k_0}\Big[\sum_{u \in N(v)} (1-y_u) h(u)\Big],$$

where $k_0 = \sum_{u \in N(v)}(1 - y_u)$, $k_1 = \sum_{u \in N(v)} y_u$, and $y_u \in \{0, 1\}$. Note that if the target outcome $y_v = 1$, then $\mathbf{E}[h(v)]$ depends only on the first term (i.e., other nodes u in its neighborhood with $y_u = 1$). On the other hand, if $y_v = 0$, then $\mathbf{E}[h(v)]$ depends only on the second term (i.e., other nodes u in its neighborhood with $y_u = 0$).

Intuitively, the neighborhood peer expectation considers the average decision of all its neighbors with the same target outcome. For example, if y_u denotes whether paper u is acceptable (based on its review ratings) and $h(u)$ is its decision for acceptance, then the expected value of $h(u)$ depends on the average decision for other papers in its neighborhood (e.g., papers co-authored by one of the authors or their collaborators) with the same degree of acceptability. Since the expectation is a monotonically increasing function of $h(u)$ for its neighbors, it can be trivially shown that the neighborhood peer expectation satisfies the first two properties given in Definition 1. The third property holds since the bijection function guarantees that the y and h values for the nodes in the neighborhoods of u and v to be the same. Thus, their expected values, $\mathbb{E}[h(u)]$ and $\mathbb{E}[h(v)]$, will also be the same.

III. Fairness Visibility

We now introduce fairness visibility, which extends the fairness perception function to a group fairness measure.

Definition 2 (Fairness Visibility): Let $V_c = \{u \mid u \in V,\ u.X_p = c\}$, i.e., the set of nodes belonging to the protected attribute group c. The fairness visibility of h for group c is defined as follows:

$$FV(V_c) = \frac{\sum_{v \in V_c} f(v,h)}{|V_c|} \tag{3}$$

Note that the fairness visibility for a given group c can be viewed as the average fairness perception of all the nodes that belong to the protected group c. For example, the group c may refer to all the papers written by well-established authors in the peer review network. To determine whether the decision function h is fair, we compare the fairness visibility for different groups of nodes using the definition below.

Definition 3 (Fairness Visibility Parity): The decision function h satisfies fairness visibility parity for V_c and $V_{c'}$ if

$$FV(V_c) = FV(V_{c'}) \tag{4}$$

For example, in a peer review network, we may categorize the papers into two groups, those written by famous authors or those written by less established researchers. If the average fairness perception for both groups of papers are the same, then their fairness visibility parity holds. The larger the disparity, the more biased are the decisions as perceived by the groups.

A standard approach for measuring group fairness is to compute demographic parity, which is defined as follows:

Definition 4 (Demographic Parity): The decision function h satisfies demographic parity for V_c and V_c' if

$$P(h(v) = 1 \mid v \in V_c) = P(h(v) = 1 \mid v \in V_c') \tag{5}$$

Unlike fairness visibility parity, demographic parity is computed for non-relational data since it ignores the neighborhood structure of a node. In the context of peer review network, each probability term in Eqn. (5) corresponds to the acceptance rate of papers that belong to the group c or c'. For brevity, we

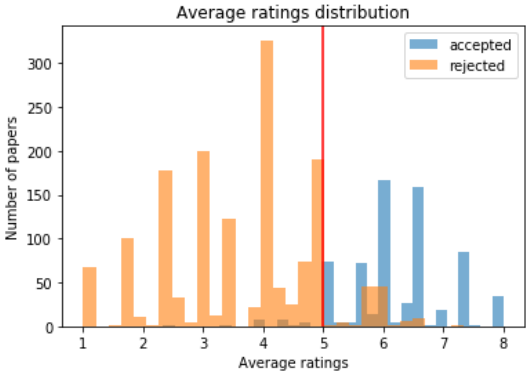


Fig. 2: The average rating distribution of submitted papers. The red line indicates the threshold used for classifying papers as acceptable ($y = 1$) or unacceptable ($y = 0$).

termed $P(h(v) = 1|v \in V_c)$ as the acceptance probability for the group V_c. The theorem below illustrates the relationship between fairness visibility and acceptance probability.

Theorem 2: Assuming the network graph is connected and the decision function h has non-zero true positive and false positive rates, the fairness visibility of group V_c, based on the neighborhood peer expectation, converges to the acceptance probability for V_c as the δ-neighborhood size increases.

Proof: Given a node v, note that $N(v) \to V$ as the δ-neighborhood expands since the network graph is assumed to be connected. Furthermore, if the true positive and false positive rates for h are non-zeros, then eventually $\mathbb{E}[h(v)] > 0, \forall v \in V$ by expansion of δ-neighborhood. It follows that $f(v, h) = 1$ if $h(v) = 1$ and $f(v, h) = 0$ if $h(v) = 0$. Thus $FV(V_c)$ converges to $P(h(v) = 1|v \in V_c)$. ■

Corollary 2.1: Given a connected network G, the decision function h satisfies demographic parity if and only if there exists a positive integer k such that for all $\delta \geq k$, fairness visibility parity holds for h with the given δ-neighborhood.

IV. Application to Peer Review Networks

This section presents a case study on the application of our proposed approach to a peer review network dataset.

A. Data

We constructed a network from the peer review dataset collected for the ICLR 2020 conference from the OpenReview.net website. Specifically, for each submitted paper, we gathered information about its title, abstract, list of authors and their affiliations. In addition, the anonymized reviews and acceptance decision for each reviewed paper are also available. For the ICLR 2020 conference, the number of submitted papers is 2594. However, 382 of the submissions were withdrawn. Our analysis is therefore restricted to only 2212 papers which had been reviewed. We use this information to create a network that contains 2212 nodes, one for each peer-reviewed paper.

The total number of accepted papers, either as oral or poster presentation, is 687 while the number of rejected papers is 1525. Thus, the conference acceptance rate is around 31%.

TABLE I: Summary distribution of acceptable and accepted papers for the ICLR 2020 conference.

		Acceptability	
		$y = 1$	$y = 0$
Acceptance Decision	$h = 1$	589	98
	$h = 0$	117	1408

(a) All papers

		Acceptability	
		$y = 1$	$y = 0$
Acceptance Decision	$h = 1$	94	13
	$h = 0$	12	153

(b) Famous authors

		Acceptability	
		$y = 1$	$y = 0$
$h = 1$		495	85
$h = 0$		105	1255

(c) Non-famous authors

		Acceptability	
		$y = 1$	$y = 0$
Acceptance Decision	$h = 1$	190	34
	$h = 0$	21	328

(d) Top institutions

		Acceptability	
		$y = 1$	$y = 0$
$h = 1$		399	64
$h = 0$		96	1080

(e) Non-top institutions

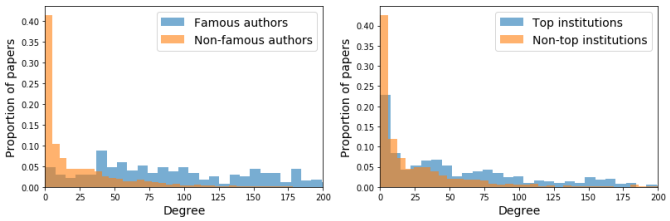
We use the acceptance decision of each paper as the decision function h to evaluate fairness perception. We consider the acceptability of a paper, in terms of its average review ratings, as the target outcome y. Our assumption here is that the reviewers are rational-minded individuals, whose average ratings given to a paper reflect the technical merits and acceptability level of the paper. Figure 2 shows histograms of average review ratings for the accepted and rejected papers. Given that the number of accepted papers is 687, we choose an acceptability threshold of 6 since it gives a number of acceptable papers that has the closest match to the actual number of accepted papers. With this threshold, all papers whose average ratings are larger than 5 are considered acceptable, i.e., $y = 1$. Table I(a) shows a confusion matrix comparing the acceptability of the paper (y) and its acceptance decision (h).

The total number of authors who had submitted papers to the conference was 6953. We were able to extract authorship information for each paper, such as names and email addresses of the co-authors, affiliation, gender, and scholarid by prepossessing the the users profile page on the OpenReview website. Based on this information we classified the submitted papers into groups based on the following "protected" attributes:

- **Famous author papers**: If a paper includes one or more famous authors, its protected attribute value is $X_p = 0$, otherwise $X_p = 1$. We consider the top 500 authors with highest h-index according to Google scholar[1] as famous authors. With this designation, 272 of the submissions were classified as famous author papers.
- **Top institution papers**: If a paper has an author from a top-10 university according to the csrankings.org website[2], then it its protected attribute value is $X_p = 0$, otherwise $X_p = 1$. We found 573 of the submitted papers have at least one author from a top institution.

[1]https://scholar.google.com/citations?view_op=search_authors&hl=en&mauthors=label:machine_learning

[2]http://csrankings.org/#/index?all



(a) Famous author (b) Top institution

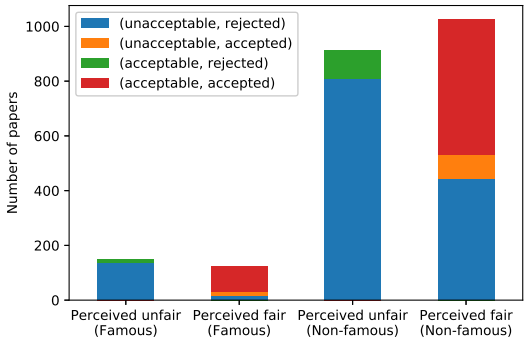
Fig. 3: Degree distribution of nodes in peer-review network.

A breakdown on the number of acceptable and accepted or rejected papers for each group is shown in Tables I(b)-(e). The results given in these tables are consistent with previous research, which had suggested that conference paper acceptance decisions are generally biased in favor of famous authors or papers written by authors from top institutions [9], [10]. In particular, the results suggest that the chance for an acceptable paper to be accepted is significantly higher for papers written by famous authors (88.7%) or authors from a top institution (90.1%) compared to those written by non-famous authors (82.5%) or authors from lower ranked institutions (80.6%). Papers by famous or top institution authors also have a higher chance of getting their unacceptable papers accepted compared to those by non-famous authors or authors from lower ranked institutions, as reflected by their higher false positive rates.

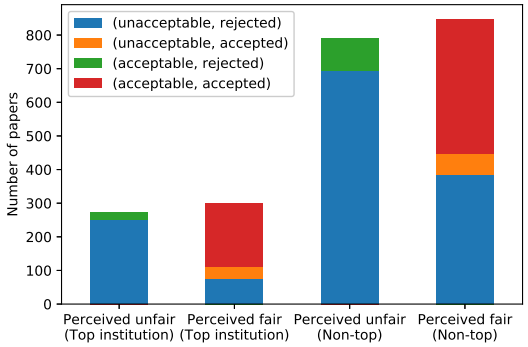
We use the co-authorship information extracted from the authors' profile pages on OpenReview.net to construct the links between the nodes in the network. We consider two papers are linked if they share a common co-author or if the authors have collaborated in the past. Figures 3-(a) and 3-(b) show the degree distribution of the networks based on the famous author and top institution protected attributes. The results suggest that papers by famous authors or authors from top institutions tend to have higher degree (on average) and a heavier tail in their distribution compared to those written by non-famous authors or authors from lower ranked institutions.

B. Fairness Perception

We applied the proposed fairness perception function to the network and evaluated the proportion of papers who perceived the paper acceptance decision to be fair or unfair. The results are shown in Figure 4-(a) for the famous author protected attribute. Despite the fact that papers by famous authors are generally favored (i.e., have higher true positive and false positive rates), the bar chart shown in Figure 4-(a) suggests that the majority of them still perceived the decision to be unfair. According to the fairness perception function, the main source of their discontent is the unacceptable papers that were rejected (i.e., the blue bar), which they believe should have been accepted. For papers by non-famous authors, Figure 4-(a) tells an opposite story as the majority of them perceived the paper acceptance decisions to be fair. Although a significantly large number of them still perceived the decision to reject their unacceptable papers as unfair (see the blue bar for perceived



(a) Famous author



(b) Top institution

Fig. 4: Assessment of network-centric fairness perception

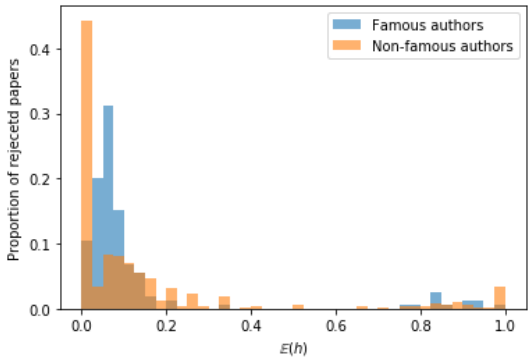
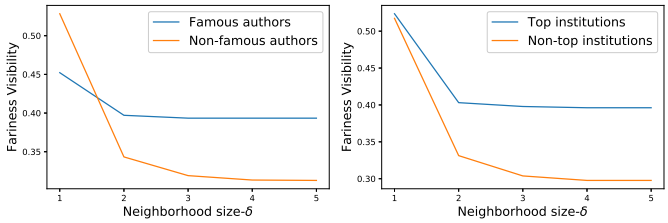


Fig. 5: Comparison of $\mathbb{E}[h(v)]$ for rejected papers by famous and non-famous authors.

unfair), the non-famous author papers are more amenable to accepting the decision to reject their unacceptable papers (see the proportion of blue bar for perceived fair).

The preceding results show a potential pitfall of using the fairness perception function (with neighborhood size $\delta = 1$). Although the analysis of the confusion matrices given in Table I suggests that the decision is biased in favor of papers written by famous authors or top institutions, the non-famous authors or those from non-top-tier institutions still perceived the decisions to be fair! This can be explained as follows. Since our fairness perception function depends on the computation of $\mathbb{E}[h(v)]$, we examine the distribution of $\mathbb{E}[h(v)]$ for rejected papers by famous and non-famous authors. The results are



(a) Famous author (b) Top institution

Fig. 6: Effect of neighborhood size on fairness visibility.

shown in Figure 5. More than 40% of the rejected papers by non-famous authors have an expected value close to 0 compared to around 10% of the rejected papers by famous authors, Based on the definition given in Eqn. (2), the larger the proportion of papers with $\mathbb{E}[h(v)]$ close to zero, the more likely they perceived the decision to be fair. One possible explanation for the famous authors to have fewer proportion of papers with $\mathbb{E}[h(v)]$ close to zero is due to the degree distribution of their nodes (see Figure 3). Since papers by famous authors generally have a higher degree, this increases the number of nodes in their neighborhood, which in turn, results in a higher expected value according to the formula used to compute the neighborhood peer expectation. In contrast, many papers by non-famous authors have low degree nodes, thus producing more nodes with low $\mathbb{E}[h(v)]$.

C. Fairness Visibility

In this section, we will empirically evaluate the theoretical results for fairness visibility, which provides a possible solution to alleviate the potential pitfall of using our fairness perception function. For this experiment, we vary the δ-neighborhood size from 1 to 5 and compute the corresponding fairness visibility measure with respect to the protected attributes. The results are plotted in Figure 6.

For Figure 6-(a), observe that the fairness visibility of papers by famous authors are initially lower than that for papers by non-famous authors when $\delta = 1$. This means that, on average, the papers by famous authors have lower perceived fairness. As δ increases, fairness visibility decreases for both groups of papers. However, the rate of decrease is higher for papers by non-famous authors. According to Theorem 2, under mild assumption, fairness visibility will converge to the acceptance probability of each subgroup of the protected attribute when δ increases. Since the acceptance probability for $X_p = 0$ (famous authors) is higher than that for $X_p = 1$ (non-famous authors), the fairness visibility for famous authors will be higher for larger values of the neighborhood size, δ. This provides a strategy to counter against the potential pitfall of using fairness perception by expanding the neighborhood size δ. Furthermore, it is worth noting that the peer review network is not a connected graph. As a result, the fairness visibility does not converge exactly to the acceptance probability for each group, which is 0.2989 (for non-famous authors) and 0.3933 (for famous authors), when δ is sufficiently large.

A similar observation can be made when analyzing the effect of increasing neighborhood size on fairness visibility using top institution as protected attribute. As shown in Figure 6-(b), increasing δ leads to lower fairness visibility. However, with sufficiently large δ, the fairness visibility for papers by authors from top institutions is higher than that for papers by authors from lower ranked institutions. By setting $\delta = 2$, the fairness visibility provides a good assessment on the true bias of the paper acceptance decisions.

V. Conclusion

This paper presents a novel approach for algorithmic fairness in network data. Motivated by the equity theory in social science, we introduced the concept of fairness perception as a local formulation of fairness and quantified this notion through an axiomatic approach to analyze its properties. We also showed how our proposed network-centric fairness perception function can be extended to a group fairness measure known as fairness visibility. We provided theoretical analysis to demonstrate its relationship to demographic parity. Using a peer-review network as case study, we also examined its utility in terms of assessing the perception of fairness in paper acceptance decisions. We also highlighted a potential pitfall of using fairness visibility measure as it can be exploited to mislead individuals into perceiving that the algorithmic decisions are fair. Finally, we show how to alleviate the problem by increasing the local neighborhood size.

Acknowledgment

This material is based upon work supported by the NSF Program on Fairness in AI in collaboration with Amazon under award #IIS-1939368. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Amazon.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica, May*, vol. 23, 2016.

[2] B. J. Jefferson, "Predictable policing: Predictive crime mapping and geographies of policing and race," *Annals of the American Association of Geographers*, vol. 108, no. 1, pp. 1–16, 2018.

[3] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *SIGKDD*. ACM, 2017, pp. 797–806.

[4] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[5] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im) possibility of fairness," *arXiv preprint arXiv:1609.07236*, 2016.

[6] J. M. Peiró, V. Martínez-Tur, and C. Moliner, *Perceived Fairness*. Dordrecht: Springer Netherlands, 2014, pp. 4693–4696.

[7] L. Shulga and S. Tanford, "Measuring perceptions of fairness of loyalty program members," *Journal of Hospitality Marketing & Management*, vol. 27, no. 3, pp. 346–365, 2018.

[8] J. S. Adams, "Inequity in social exchange," in *Advances in experimental social psychology*. Elsevier, 1965, vol. 2, pp. 267–299.

[9] I. Stelmakh, N. B. Shah, and A. Singh, "Peerreview4all: Fair and accurate reviewer assignment in peer review," *arXiv preprint arXiv:1806.06237*, 2018.

[10] A. Tomkins, M. Zhang, and W. D. Heavlin, "Reviewer bias in single- versus double-blind peer review," *Proceedings of the National Academy of Sciences*, vol. 114, no. 48, pp. 12 708–12 713, 2017.