Interdisciplinarity, Gender Diversity, and Network Structure Predict the Centrality of AI Organizations

MADALINA VLASCEANU*, New York University, USA
MIROSLAV DUDÍK, Microsoft Research, USA
IDA MOMENNEJAD, Microsoft Research, USA

Artificial intelligence (AI) research plays an increasingly important role in society, impacting key aspects of human life. From face recognition algorithms aiding national security in airports, to software that advises judges in criminal cases, and medical staff in healthcare, AI research is shaping critical facets of our experience in the world. But who are the people and institutional bodies behind this influential research? What are the predictors of influence of AI researchers and research organizations? We study this question using social network analysis, in an exploration of the structural characteristics, i.e., network topology, of research organizations that shape modern AI. In a sample of 149 organizations with 9,987 affiliated authors of published papers in a major AI conference (NeurIPS) and two major conferences that specifically focus on societal impacts of AI (FAccT and AIES), we find that both industry and academic research organizations with influential authors are more interdisciplinary, have a greater fraction of women, are more hierarchical, and less clustered, even when controlling for the size of the organizations. The influence is operationalized as betweenness centrality in co-authorship networks, i.e., how often an author is on the shortest path connecting any pair of authors, acting as a bridge connecting otherwise distant (or even disconneted) members of the network, such as their own co-authors who are not each other's co-author themselves. Using this operationalization, we also find that women have less influence in the AI community, determined as lower betweenness centrality in co-authorship networks. These results suggest that while diverse AI institutions are more influential, the individuals contributing to the increased diversity are marginalized in the AI field. We discuss these results in the context of current events with important societal implications.

Additional Key Words and Phrases: organizational structure, artificial intelligence, gender diversity, interdisciplinarity

ACM Reference Format:
Madalina Vlasceanu, Miroslav Dudík, and Ida Momennejad. 2022. Interdisciplinarity, Gender Diversity, and Network Structure Predict the Centrality of AI Organizations. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3531146.3533069

1 INTRODUCTION

Artificial intelligence (AI) research plays an increasingly important role in society, impacting key aspects of human life. From face recognition algorithms aiding national security in airports [15], to risk assessment systems that advise judges in criminal cases [46] and medical staff in healthcare [59], research in artificial intelligence is shaping critical facets of our experience in the world. At first, this widespread deployment of AI was positively embraced, given classical depictions of artificial intelligence systems as autonomous, rational, and intelligent problem solvers [64]. Recently, however, there has been a surge of interest in questioning the assumptions, decisions, performance, and motivations underlying such algorithms that have rapidly infiltrated our everyday lives [10, 16, 24, 25, 58, 60].

This growing interest was partly fueled by the numerous cases of algorithmic oppression brought to light in recent years [58]. Notable examples are Facebook's job advertisement algorithm targeting users based on their gender, race, and religion, that was found to disproportionately suggest stereotypically feminine jobs (e.g., nurse, secretary) to women and stereotypically masculine jobs (e.g., janitor, taxi driver) to men and members of racial minorities, thus further propagating sexist and racist tendencies in the labor market [1]. Similarly, Amazon's recruitment algorithm

*Research performed as an intern at Microsoft Research.

programmed to automate preexisting hiring practices was found to discriminate against women applicants [27]. In the criminal justice system, algorithms used to predict criminal reoffense were found to disproportionately attribute higher risks to Black defendants and lower risks to white defendants [46]. In the healthcare industry, algorithms used in US hospitals to decide healthcare allocation were found to systematically assign a lower standard of care to Black patients [59]. Consequently, it has been argued that instead of autonomous and rational, artificial intelligence algorithms are actually "a registry of power" designed to serve existing dominant interests [25]. Indeed, previous research has shown how technology can encode political values [78], facilitating biases in the automated machinery affecting us all. Such instances of societal bias making their way into AI systems have turned scholars' attention to the broader socio-technical context in which AI systems are designed and deployed, including the humans responsible for designing, training, and monitoring these AI systems. In line with the feminist standpoint theory [4, 29, 35, 36], recently published books such as Noble's *Algorithms of Oppression* [58] and Crawford's *The Atlas of AI* [25] urge the questioning of widely deployed algorithms, to understand the implications of "*what* is being optimized, and for *whom*, and *who* gets to decide", as well as for whom are algorithms such as Google Search offering "the best information". Similarly, it has been suggested that characteristics of Silicon Valley's programmers (i.e., "even younger, more masculine and more fully committed to working all hours") are mirrored in the algorithms they produce, with profound implications for many people [73].

Building on this body of work, here we examine the question of which individuals and what organizations are influential in AI community through the lens of social network analysis. Specifically, we assess the role of network topology, gender diversity, and interdisciplinarity in AI co-authorship networks. To this end, we are employing social network analysis, a well established tool for mapping the structural composition (e.g., network topology) of a community. Understanding the organizational structure of a social network has been shown to be instrumental in identifying the relative influence of each individual within the community [41, 74] and predicting the spread of memories [23, 54], practices [22], and behaviors [19] through the community. As any other complex system, social networks are usually analyzed in the framework of graph theory, as graphs are used to represent the connections between members of a social network [67]. In this context, each member is represented by a node (or a vertex), and each connection between two nodes is represented by an edge. The main properties of social networks have been identified as the degree of *centrality* [13, 33], *hierarchy* [12], *clustering* [47], as well as the *diversity* of its members [82]. We next review these properties, beginning with centrality.

Centrality is a property of a node in a network, originally introduced to quantify efficiency of problem solving and perception of leadership in human networks [6, 48]. One of the most widely used centrality metrics is *eigenvector centrality* [13], a measure of how many "influential" nodes one is connected to, computed as the weighted sum of both direct and indirect connections of every length, thus taking into account the entire pattern in the network. Another important centrality measure is *betweenness centrality* [33], which indicates how often a person is positioned on the shortest path, or a "geodesic", between other pairs of people in the network. Intuitively, betweenness centrality captures how often an author acts as a bridge connecting otherwise distant (or even disconnected) members of the network, such as their own co-authors who are not each other's co-author themselves. Because of this bridging property, a person with high betweenness is the most able to control the flow of information through the network, which is why betweenness has been suggested to signify power [17, 49]. Moreover, betweenness centrality (but not eigenvector centrality, see [61]) has been shown to significantly predict the publication rates of computer science faculty members at a major US university [38]. Motivated by these observations, we will use betweenness centrality as a proxy for influence in the AI community.

Another network property capturing fundamental features of its structure and dynamics is the degree of *hierarchy*. Most commercial companies and government agencies have a hierarchical topology. Hierarchical structures are known

for their slow information flow but strong authority connections [52]. In organizations, hierarchical structures have been found to create power imbalances, decrease team members' identification with the team [76], and impact collective performance [12, 34]. Here, we are interested in whether the hierarchical patterns of communication in AI collaborations predict influence (i.e., betweenness centrality) in the AI community.

Another property of network structures that predicts their spread of information and influence throughout the community is *network clustering*. Network clustering refers to the degree to which the network is arranged into distinctive, densely connected "cliques" or "clumps" [47]. Network clustering can be quantified with the network clustering coefficient, which is the probability that any two nodes connected to the same node are themselves linked [75]. This property of a network has significant implications for the speed and degree of spread of social influence through the network, such that influence spreads faster and farther through more clustered communities compared to less clustered communities [19]. Here, we are therefore interested whether the degree of clustering of AI organizations also predicts their influence in the AI community.

The final network structure property of interest is the *diversity* of the members of the network. Diversity refers to differences between individuals on any attributes (e.g., demographic characteristics) that may lead to the perception that another person is different from the self [69, 71]. The lack of gender and race diversity in the AI sector has been identified as a factor contributing to the rise of AI systems that replicate societal inequities and that offer poor service to marginalized populations [32, 40, 77]. Diversity also plays an important role in organizational network structures, as increased diversity in organizations has been found to improve decision making through increased creativity and innovation [3, 28, 53]. A recent study assessing the role of diversity in the context of higher academia showed the value of diverse editors in curating law reviews. The authors examined the main law reviews of the 20 most prestigious law schools aggregating data from 60 years, and found that issues published by diverse editors were 23% more cited than their counterparts [21]. Similarly, another recent study found that increased gender diversity on banks' boards reduces the frequency of misconduct incidents [2]. Conversely, other studies have found that diversity of both race and gender negatively impacts team performance [8]. However, a meta-analysis including results from 140 studies concluded that female board representation was positively associated with firms' financial performance [62]. Therefore, we are interested in uncovering both the role of diversity in AI organizations, and the relative influence of researchers contributing to diversity. As a crude first step in our inquiry into the role of diversity in AI, here, we focus our analysis on gender diversity, and in particular on the representation of women (or, more precisely, the researchers with feminine first names). We acknowledge that this approach has substantial limitations in that it omits other forms of gender identity and other axes of human experience that intersect with gender like race, ethnicity, or educational background. In the discussion section, we highlight how the incorporation of these additional dimensions would advance this body of research. We do not wish to reify binary gender, nor perpetuate the narrow conception of diversity as the issue of ''women in tech", which is likely to privilege white women [77].

Lastly, *interdisciplinarity* has always been a critical dimension in AI research since its early days [63], *and* remains crucial today given its potential to make AI more accessible, inclusive, and trustworthy [44]. More broadly, interdisciplinarity has been shown to meaningfully contribute to one's influence in the scientific community. For instance, interdisciplinary publications have been found to achieve more citations [20]. Similarly, an analysis of 17.9 million research articles spanning all scientific fields found that papers using unusual combinations of knowledge are twice as likely to be highly cited [70]. Moreover, prior work has found a strong association between interdisciplinarity in scientific journals and betweenness centrality [50]. Therefore, here, we are also interested in exploring whether more interdisciplinary AI research organizations are more influential in the AI community.

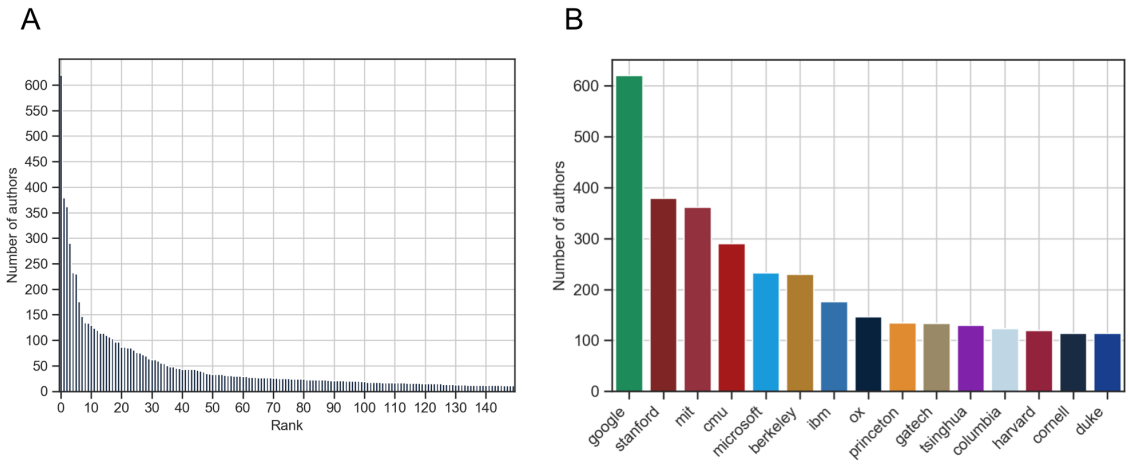


Fig. 1. Panel A: Histogram of the number of authors of the organizations in the dataset (organizations with published papers in NeurIPS 2015–2019, FAccT 2019–2020, and AIES 2018–2020). Panel B: Top 15 most represented organizations in the dataset.

Hypotheses. We first hypothesize that an author's influence in the AI community, quantified as their *betweenness centrality* in co-authorship networks (i.e., how often an author is on the shortest path connecting any pair of authors), is predicted by properties of the structure of the author's organization, namely network topology (e.g., degree of hierarchy, clustering), gender diversity, and interdisciplinarity. Our second hypothesis is that women would have less influence (i.e., lower betweenness centrality) in the AI community.

2 METHODS

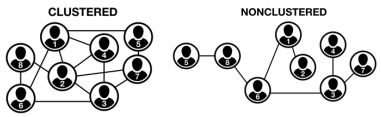
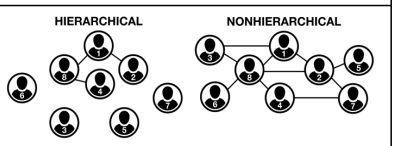
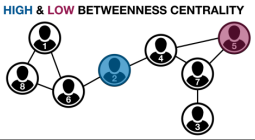
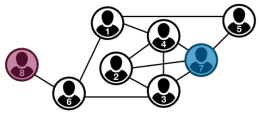
All data and analyses can be found at https://github.com/mvlasceanu/AI_coauthorship_network.

For our analysis, we constructed a dataset from publication metadata at a major AI conference (NeurIPS) and two major conferences that specifically focus on societal impacts of AI (FAccT and AIES). Our choices were driven by the following criteria: (1) include venues that publish highly cited papers in AI, as a proxy for prestige (NeurIPS is among top three AI conferences along a number of citation metrics); (2) include venues with an interdisciplinary focus, to improve the representativeness of our data in terms of interdisciplinarity (FAccT and AIES explicitly encourage interdisciplinary participation and have interdisciplinary organization committees); (3) include venues with available metadata, as a necessary constraint for our analysis.

We scraped 5 years (2015–2019) of NeurIPS proceedings metadata using custom Python scripts accessing the conference website. NeurIPS stopped publishing proceedings metadata in 2020, which is why our data stops at 2019. We scraped 2 years (2019, 2020) of FAccT and 3 years (2018–2020) of AIES proceedings metadata using CrossREF's open source API [45]. These were the only years for which the conferences published their metadata online. In total, we collected information comprising 14,849 author–publication pairs.

Organizations. We used authors' emails to infer their organizational affiliations. In doing so, we mapped different departments within an organization (e.g., @cs.berkeley.edu and @ee.berkeley.edu) to the same organization (e.g., berkeley). We manually cross-referenced this procedure to ensure high accuracy. Using this procedure, we identified the organizational affiliation of 91.4% of the authors in the dataset. Moreover, we used authors' email domains to

Table 1. Illustration of graph measures. The clustering coefficient measures the degree to which nodes in a graph tend to cluster together. The hierarchical coefficient captures the degree of hierarchy in a complex network. Betweenness centrality indicates the extent to which a node is positioned on the shortest path ("geodesic") between other pairs of nodes in the network. Eigenvector centrality is the weighted sum of direct and indirect connections of every length, thus taking into account the entire pattern in the network.

MEASURE	EXAMPLE
CLUSTERING COEFFICIENT	
HIERARCHICAL COEFFICIENT	
BETWEENNESS CENTRALITY	
EIGENVECTOR CENTRALITY	

classify whether they are part of an academic versus an industry organization. To differentiate *academic* from *industry* organizations we used the domains in researchers' emails, such that ".edu" or ".ac" domains were labeled as academic, and ".ai" or ".com" were labeled as industry. We identified unique authors by concatenating first names, last names, and organizations and assigned each such unique person an individualized id. Using this procedure, we identified 9,987 unique authors in 149 organizations (see Figure 1).

Organizational structure. Organizational structure is operationalized as the network connectivity among the members of an organization (examples of organizations include Google, Stanford University, or MIT). A pair of authors (nodes) is connected by an edge if they have co-authored any publication in the conference proceedings that we scraped. Using the connectivity network, we computed the clustering and hierarchical coefficients, which are graph measures frequently used in network science to describe the properties of a group of interconnected nodes, in our case researchers. The clustering coefficient $\bar{C}$ of a network G quantifies the degree to which nodes in G tend to cluster together [75]:

$$\bar{C} = \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \frac{\#\text{ of edges in } G(\mathcal{N}_i)}{\binom{|\mathcal{N}_i|}{2}}, \tag{1}$$

where $\mathcal{N}$ is the set of nodes of G, $\mathcal{N}_i$ is the set of neighbors of node i, and $G(\mathcal{N}_i)$ is the subnetwork of G induced by $\mathcal{N}_i$. This graph measure, alongside others, is illustrated in Table 1.

The hierarchical coefficient captures the degree of hierarchy in a complex network [55]. It is defined as

$$\text{HierCoeff} = \frac{1}{\binom{|\mathcal{N}|}{2}} \sum_{\substack{i,j\in\mathcal{N}:\\ 0<d(i,j)<\infty}} \frac{1}{d(i,j)}, \tag{2}$$

where $\mathcal{N}$ is the set of nodes of G as before, and $d(i,j)$ is the length of the shortest path from i to j. In studying this graph property, we conceptualize the analyzed network as only a sample of a larger underlying network. Accordingly, it is natural to expect that a sample of a hierarchical organization will be characterized by many disconnected components (because hierarchical networks are characterized by few lateral connections). This intuition is reflected in our illustration of a large hierarchical coefficient in Table 1.

Centrality metrics. Centrality metrics are operationalized using the co-authorship graph inferred from the scraped publications (Figure 2). Using the adjacency matrix of the co-authorship graph, we calculated the betweenness centrality and the eigenvector centrality of each author. Betweenness centrality indicates the extent to which a node (in our case a researcher) is positioned on the shortest path ("geodesic") between other pairs of nodes in the network [33]:

$$\text{BetCentrality}(i) = \sum_{j,k\in\mathcal{N}\setminus\{i\}} \frac{\sigma_{jk}(i)}{\sigma_{jk}}, \tag{3}$$

where σ_{jk} is the total number of shortest paths from node j to node k and $\sigma_{jk}(i)$ is the number of those paths that pass through i. (See Table 1 for an illustration of nodes with high and low betweenness centrality.)

Eigenvector centrality assesses whether a node is connected to "influential" nodes in the network, and this in turn quantifies how "influential" the node itself is [13]. The relative influence scores x_i, which are referred to as the eigenvector centrality scores, are assigned to each node such that, for a suitable constant $\lambda > 0$,

$$\text{EigCentrality}(i) = x_i = \frac{1}{\lambda} \sum_{j\in\mathcal{N}_i} x_j = \frac{1}{\lambda} \sum_{j\in\mathcal{N}} a_{ij} x_j, \tag{4}$$

where a_{ij} are entries of the adjacency matrix, equal to 1 if the node i is connected with node j and zero otherwise. The values x_i, can be seen to form an eigenvector of the adjacency matrix. Among all the eigenvectors, eigenvector centrality uses the unique eigenvector with non-negative values (which is the eigenvector associated with the top eigenvalue).

Centrality measures of an organization are obtained by taking an average centrality of all its nodes (authors).

Gender diversity. Gender diversity of an organization is operationalized as the fraction of the number of researchers with traditionally feminine first names out of the researchers with either traditionally feminine or traditionally masculine first names. Thus, an organization where all members have traditionally masculine first names (or the names that are neither traditionally masculine or feminine) has the smallest score 0, whereas the organization where all members have traditionally feminine first names (or the names that are neither traditionally masculine or feminine) has the largest score 1. Note that in the latter case, the organization itself might not be gender diverse, but, since the AI field skews heavily towards men, we expect that such organizations would contribute to an increase of gender diversity of the field. Therefore, we use this one-sided operationalization (this only affects 4 out of 149 organizations in our data, whose gender diversity values are greater than 0.5).

In order to label author names, we used the publicly available dataset provided as part of the Gender Guesser Python package[1]. The dataset comprises over 45,000 names with labels "neutral", "male", "female", "mostly male", or "mostly

[1] https://raw.githubusercontent.com/lead-ratings/gender-guesser/master/gender_guesser/data/nam_dict.txt

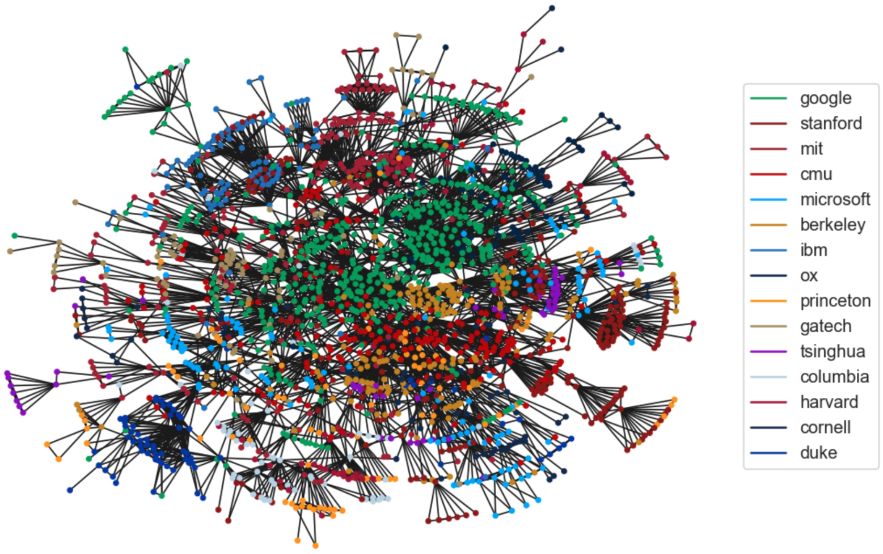


Fig. 2. Co-authorship connectivity graph. Only showing the connectivity graph of the top 15 organizations according to the number of authors with published papers in NeurIPS 2015–2019, FAccT 2019–2020, and AIES 2018–2020, omitting connected components with fewer than 20 nodes.

female" provided, and manually checked, by native speakers from various countries [65]. According to this collection, 47.28% of author names were labeled as "male", 2.84% as "mostly male", 7.12% as "female", 0.88% "mostly female", 8.23% "neutral", and 33.65% as "unknown" (not listed in the dataset). In our analyses, gender diversity is calculated as the ratio of the authors whose first names were labeled as "female" to the authors whose first names were labeled as either "male" or "female", so first names without a binarized gendered perception are treated as "unknown".

In the discussion section, we discuss limitations of this method, such as focus on just one type of minorized gender identity, uneven coverage across author nationalities and ethnicities, and reliance on an automated tool rather than self-identification.

Interdisciplinarity. To infer organizational interdisciplinarity, we clustered the paper titles using the clustering approach of Yin and Wang [80], developed for clustering of short texts (like titles). The approach is based on the probabilistic mixture model of Nigam et al. [57], which posits that each title belongs to one of the clusters, referred to as topics, described by a separate distribution over the words in vocabulary. After assigning each paper to one of the topics, we calculated the interdisciplinarity of each author as the entropy of the author's topic distribution. The interdisciplinarity of an organization is then calculated as the average interdisciplinarity of its authors. Thus, an organization with authors publishing in multiple topics has a higher interdisciplinarity than an organization where most authors publish on a single topic.

Organizational size. The organizational size is operationally defined as the number of authors from that organization in our dataset.

Table 2. Hypothesis 1—Results. Coefficients of linear regression models fitted on all organizations, as well as on subsets of large and small organizations. Betweenness centrality is predicted as a function of interdisciplinarity, gender diversity, degree of hierarchy, and degree of clustering, while controlling for the organization size.

	Model coefficients (β)		
Explanatory variable	All organizations	Large organizations	Small organizations
Interdisciplinarity	0.54^{**} ($p < 0.001$)	0.50^{**} ($p < 0.001$)	0.55^{**} ($p < 0.001$)
Gender diversity	0.16^{*} ($p = 0.007$)	0.32^{**} ($p = 0.001$)	0.08 ($p = 0.337$)
Hierarchy	0.18^{*} ($p = 0.017$)	0.30^{*} ($p = 0.007$)	0.12 ($p = 0.261$)
Clustering	-0.15^{*} ($p = 0.034$)	-0.30^{*} ($p = 0.018$)	-0.09 ($p = 0.339$)
Log of org. size	0.15 ($p = 0.066$)	0.32^{*} ($p = 0.003$)	0.27 ($p = 0.431$)

3 RESULTS

3.1 Hypothesis 1: Predictors of AI Organization's Influence

Our first hypothesis was that influence of an organization in the AI community (i.e., betweenness centrality in co-authorship networks) is predicted by organizational structure, measured as the organization's network topology (e.g., degree of hierarchy, clustering), gender diversity, and interdisciplinarity.

We fitted a linear model with the average betweenness centrality of organizations as the response variable, and gender diversity, interdisciplinarity, hierarchical coefficient, and clustering coefficient as explanatory variables, while also controlling for organizational size by including the log transformation of the organizational size as an additional regressor. We used the log transformation, because it showed a higher correlation with response than raw organization size ($\rho = 0.43$ versus $\rho = 0.38$) and was therefore deemed more suitable for linear modeling. All the regressors as well as the response variable were standardized to allow relative comparisons of regression coefficients.

The resulting regression coefficients are shown in Table 2 (in the "All organizations" column). The model has $R^2 = 0.50$, so it explains half of the observed variance in the response variable, and, as the table shows, all of the explanatory variables of interest are significant: interdisciplinarity with the coefficient $\beta = 0.54$ ($SE = 0.03$, $p < 0.001$), gender diversity with $\beta = 0.16$ ($SE = 0.06$, $p = 0.007$), hierarchy with $\beta = 0.18$ ($SE = 0.08$, $p = 0.017$), clustering with $\beta = -0.15$ ($SE = 0.07$, $p = 0.066$). Thus, we found that organizations with central authors are more interdisciplinary, more gender diverse, more hierarchical, and less clustered.

For comparison, we also fitted a linear model with the same regressors, but this time predicting the average eigenvector centrality of organizations. The resulting model fitted the data very poorly ($R^2 = 0.03$) and none of the explanatory variables significantly predicted the response.

Returning back to the linear model of betweenness centrality, we diagnosed the assumptions underlying linear modeling by constructing CCPR (component and component-plus-residual) plots. These plots show, for each explanatory variable, how the response (with the predictions of all other variables subtracted) varies with that explanatory variable (see Figure 3, top row). Evenly distributed errors and linear relationship would suggests that the linear modeling approach is appropriate. In our plots, the linear relationship seems approximately justified, but the distribution of errors is somewhat uneven. This means that the magnitudes of coefficients β should be interpreted with caution, perhaps with the exception of the coefficient for interdisciplinarity, which appears most robust.

The CCPR plots in particular show that prediction errors are larger for smaller organizations. This is perhaps not surprising, since our explanatory variables are obtained by averaging over the authors in the organization, and so their

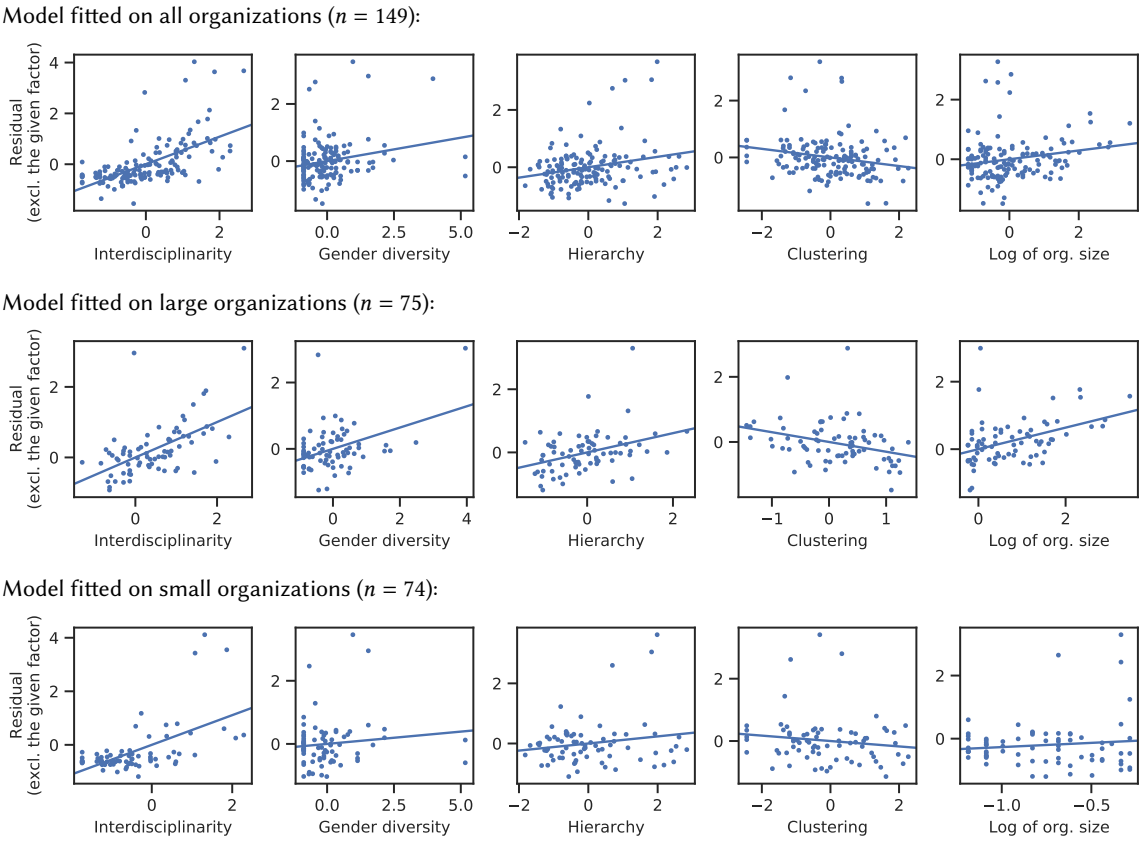


Fig. 3. CCPR plots for the linear regression fit of the betweenness centrality (response variable) as a function of interdisciplinarity, gender diversity, hierarchical coefficient, clustering coefficient, and log organization size (explanatory variables). The y-axis shows the residual of the linear regression model without the coefficient for the variable that varies on the x-axis. The explanatory variables as well as the response are standardized.

values have higher variance for smaller organizations. To compensate for the different levels of errors between large and small organizations, and to simultaneously check the robustness of our findings, we therefore refitted our model separately on *large organizations* (those with size equal or greater than the median size, which is 25 in our case) and *small organizations* (those with the size smaller than median).

The resulting regression coefficients are shown in Table 2. As we see, the model fitted on large organizations has similar properties as the overall model, but the overall fit is better: the model has $R^2 = 0.59$, so it explains 59% of the variance in the response variable. As before, all of the explanatory variables of interest are significant, and their effects are generally more pronounced: interdisciplinarity has the coefficient $\beta = 0.50$ ($SE = 0.11$, $p < 0.001$), gender diversity $\beta = 0.32$ ($SE = 0.09$, $p = 0.001$), hierarchy $\beta = 0.30$ ($SE = 0.11$, $p = 0.007$), and clustering $\beta = -0.30$ ($SE = 0.12$, $p = 0.018$). Thus, we found that large organizations with central authors are more interdisciplinariy, more hierarchical, more gender diverse, and less clustered. Also looking at the CCPR plots (Figure 3, middle row), we see that some of the issues with the uneven error distribution are alleviated.

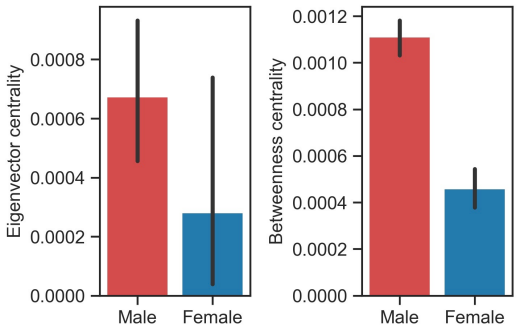


Fig. 4. Hypothesis 2—Results. Centrality metrics (eigenvector centrality in the left panel, betweenness centrality in the right panel) of authors with traditionally masculine names ("male"; red) and traditionally feminine names ("female"; blue).

In contrast, when running the same model for small organizations, we found that only *interdisciplinarity* ($\beta = 0.55$, $SE = 0.12$, $p < 0.001$) is a significant predictor of betweenness centrality (Table 2), and the fit is generally worse ($R^2 = 0.42$). This is also illustrated in the CCPR plots (Figure 3, bottom row) that show only weak dependence between the other explanatory variables and the response.

3.2 Hypothesis 2: Women Yield Less Influence in AI Community

Our second hypothesis was that women have lower betweenness centrality in AI co-authorship networks. We confirm this hypothesis and observe that authors with traditionally masculine first names have higher betweenness centrality than authors with traditionally feminine first names, with a highly significant difference between average betweenness of the two cohorts: difference = 0.086, $SE = 0.014$, $p < 0.001$ (using t-test; Figure 4). Note that this difference is not significant for eigenvector centrality ($p = 0.238$; see Figure 4).

4 DISCUSSION

Summary. In our investigation of the structure of AI researcher organizations using co-authorship networks, we hypothesized that influence in the AI community is predicted by the network topology, gender diversity, and interdisciplinarity of AI organizations. We found that influential AI organizations (i.e., organizations with high betweenness centrality authors) are more hierarchical, more gender diverse, and less clustered, when controlling for the size of the organizations. Moreover, we found that authors with traditionally feminine names remain less central in AI co-authorship networks, even though having more such authors is correlated with higher centrality of a given AI organization. These results suggest that while diverse AI institutions are more influential, the individuals contributing to the increased diversity are marginalized in the AI field.

Implications. The finding that gender diversity is associated with higher centrality is consistent with previous research suggesting that workforce diversity is essential for sustained competitive advantage as it increases creativity and innovation [5]. This result is also consistent with previous work identifying the positive role of diversity in academia and higher education [18, 21] and in the financial sector [2, 62]. Such evidence has profound implications across various domains, a noteworthy one being the higher education landscape in the United States. In this context such findings

provide empirical evidence relevant, for instance, in the current Supreme Court case deliberating the benefits of diversity in higher academia [51].

Also aligning with prior work, the result that diverse individuals are marginalized in artificial intelligence collaborations is consistent with research reporting structural and social inequities in scientific publications. For instance, men are found to be preferentially cited in academic journals, more than would be expected if gender were not a factor in citation choices [30], and in AI field, women authors are found to receive lower scores, have lower acceptance rates, and gather fewer citations than men [68]. The fact that women are consistently marginalized broadly in the scientific sphere but also specifically in the AI field is highly concerning, given overwhelming amounts of empirical evidence showing that gender diversity in teams decreases misconduct [2] and increases collective performance [7, 79], especially in scientific research [56]. Beyond optimizing team performance, increasing gender diversity, or, more broadly, participation of marginalized groups in AI field is a crucial step towards AI systems that do not perpetuate existing patterns of societal injustice [11, 29, 32, 40, 42, 77].

Also in line with our hypotheses, we find that organizations with high betweenness centrality authors in AI co-authorship networks tend to be more interdisciplinary. This result aligns with prior work showing that betweenness centrality is an indicator of interdisciplinarity in scientific journals [50], and that interdisciplinary publications achieve more citations [20]. Our finding also provides additional support for the notion that interdisciplinarity constitutes a critical dimension in AI research. Other researchers have noted that interdisciplinary is critical in AI, especially when achieved by integrating ethics principles in machine learning [44]. Indeed, a growing number of scholars are urging researchers, especially those in the AI field, to incorporate such principles into their work and to assume responsibility for the outcomes of their research, rather than writing them off as beyond the scope of their work [9, 37]

The result that organizations with high betweenness centrality authors tend to have more hierarchical patterns of collaboration is not surprising since most commercial companies and government agencies favor a hierarchical topology. Hierarchical structures are known for their strong authority connections [52], which, in the case of scientific collaborations in AI, translates into an aggregation of power in the form of centrality in the field. However, contrary to prior research showing that social influence spreads faster and farther through more clustered communities [19], here, we found that organizations with high betweenness centrality authors tend to be *less clustered*. Future research further exploring the causal connection between network clustering and centrality metrics could provide additional insight into these apparently contradictory results. For instance, future experimental work could programmatically manipulate the degree of clustering of a community [23, 72] to observe the causal impact clustering might have on the collective influence of a group. Such an investigation could shed light on organizational structures needed to achieve high centrality in one's community.

Limitations and future directions. First, our research relies entirely on correlational analyses, which, while informative in exploring associations between variables such as organizational topologies, diversity, and interdisciplinarity, cannot speak to the causal direction of these uncovered relationships. Future experimental work is needed to understand the causal mechanisms behind these connections.

Moreover, the organizational network structure mapped here using machine learning conference proceedings makes the assumption that co-authorship links capture collaborations. However, this method does not capture the directionality of these connections (e.g., the power relationships between authors), because the undirected graph formalism assumes symmetrical connections. Future work could increase the complexity of our model by using richer author attributes, such

as author order, which author is corresponding, organizational reporting relationships (e.g., manager, full professor), or citation counts.

Furthermore, here, we use betweenness centrality in a co-authorship network of AI publications as a proxy for influence in the AI community. While our operationalization of influence is not the most frequently used measure of determining impact in the research community [14], we chose it over citation-based operationalizations such as the h-index [39], the g-index [31] or the hm-index [66], which have been criticized for being associated with researcher attributes such as gender [26, 81], an imbalance that has been increasing over time [30]. Nevertheless, even though less popular than citation-based metrics, betweenness centrality has been used in prior work to capture influence or power in a community [17, 38, 49, 61],

Also, in this study, we explore the correlates of binarized gender diversity in AI organizations based on an automated tool (namely, Gender Guesser). We acknowledge shortcomings of this approach, such as relying on outside labels rather than self-identification, uneven coverage of different nationalities and ethnicities, and reducing the wide range of gender expression and gender identity to only one type of minorized gender identity. All of these have been condemned for silencing the voices of the marginalized people with rich intersectional identities and in particular those who fall outside the gender binary [43]. We fully agree with this sentiment, and our approach was merely a first crude step towards assessing the role of gender diversity in influential AI organizations, which we are interested in incrementally improving with the use of metrics that are based on self-identification, and by adding additional intersectional dimensions contributing to diversity, such as ethnicity, race, and educational backgrounds. Adding these dimensions would greatly benefit this field of study, and would add to previous research showing for instance that more innovative teams are composed of members with diverse educational and occupational backgrounds [8].

Finally, here, we restricted our analyses to only a few AI conference proceedings (NeurIPS, FAccT, AIES), given practical limitations such as available metadata. This investigation could however be extended to include additional AI conferences such as ICLR (International Conference on Learning Representations) or ICML (International Conference on Machine Learning), for a more complete and representative co-authorship network of AI researchers.

While acknowledging these limitations, we believe our work meaningfully contributes to ongoing debates regarding the role of diversity in organizations [51], and to emerging efforts to encourage AI organizations to incorporate ethics principles into their agendas [25]. We have shown that influential AI organizations are more interdisciplinary, have a greater fraction of women, and are also more hierarchical and less clustered, but that despite the significant value diversity adds to AI organizations, the individuals contributing to the increased diversity are marginalized in the field. These findings may, for instance, provide an additional incentive to individuals in positions of power to take steps towards reducing organizational gender imbalances, and encouraging the incorporation of ethics in AI research.

REFERENCES

[1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[2] Francesca Arnaboldi, Barbara Casu, Angela Gallo, Elena Kalotychou, and Anna Sarkisyan. 2021. Gender diversity and bank misconduct. *Journal of Corporate Finance* (2021), 101834.

[3] Karen A Bantel and Susan E Jackson. 1989. Top management and innovations in banking: Does the composition of the top team make a difference? *Strategic management journal* 10, S1 (1989), 107–124.

[4] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1301–1310.

[5] Nigel Bassett-Jones. 2005. The paradox of diversity management, creativity and innovation. *Creativity and innovation management* 14, 2 (2005), 169–175.

[6] Alex Bavelas. 1948. A mathematical model for group structures. *Human organization* 7, 3 (1948), 16–30.

[7] Julia B Bear and Anita Williams Woolley. 2011. The role of gender in team collaboration and performance. *Interdisciplinary science reviews* 36, 2 (2011), 146–153.

[8] Suzanne T Bell, Anton J Villado, Marc A Lukasik, Larisa Belau, and Andrea L Briggs. 2011. Getting specific about demographic diversity variable and team performance relationships: A meta-analysis. *Journal of management* 37, 3 (2011), 709–743.

[9] Samy Bengio, Alina Beygelzimer, Kate Crawford, Jeanne Fromer, Iason Gabriel, Amanda Levendowski, Deborah Raji, and Marc'Aurelio Ranzato. 2021. NeurIPS 2021 Ethics Guidelines. *NeurIPS Blog* (August 23, 2021). https://blog.neurips.cc/2021/08/23/neurips-2021-ethics-guidelines/

[10] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity.

[11] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021).

[12] Peter M Blau. 1972. Interdependence and hierarchy in organizations. *Social science research* 1, 1 (1972), 1–24.

[13] Phillip Bonacich. 1972. Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology* 2, 1 (1972), 113–120.

[14] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. 2008. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and technology* 59, 5 (2008), 830–837.

[15] Kevin W Bowyer. 2004. Face recognition technology: security versus privacy. *IEEE Technology and society magazine* 23, 1 (2004), 9–19.

[16] Meredith Broussard. 2018. *Artificial unintelligence: How computers misunderstand the world.* MIT Press.

[17] Ronald S Burt. 1992. *Structural holes.* Harvard university press.

[18] Lesley G Campbell, Siya Mehtani, Mary E Dozier, and Janice Rinehart. 2013. Gender-heterogeneous working groups produce higher quality science. *PloS one* 8, 10 (2013), e79147.

[19] Damon Centola. 2010. The spread of behavior in an online social network experiment. *science* 329, 5996 (2010), 1194–1197.

[20] Shiji Chen, Clément Arsenault, and Vincent Larivière. 2015. Are top-cited papers more interdisciplinary? *Journal of Informetrics* 9, 4 (2015), 1034–1046.

[21] Adam Chilton, Justin Driver, Jonathan S Masur, and Kyle Rozema. 2022. Assessing Affirmative Action's Diversity Rationale. *Columbia Law Review* 122, 2 (2022).

[22] Nicholas A Christakis and James H Fowler. 2007. The spread of obesity in a large social network over 32 years. *New England journal of medicine* 357, 4 (2007), 370–379.

[23] Alin Coman, Ida Momennejad, Rae D Drach, and Andra Geana. 2016. Mnemonic convergence in social networks: The emergent properties of cognition at a collective level. *Proceedings of the National Academy of Sciences* 113, 29 (2016), 8171–8176.

[24] Kate Crawford. 2013. The hidden biases in big data. *Harvard business review* 1, 4 (2013).

[25] Kate Crawford. 2021. *The Atlas of AI.* Yale University Press.

[26] Blaise Cronin and Kara Overfelt. 1995. Brief Communication E-journals and Tenure. *Journal of the American Society for Information Science (1986-1998)* 46, 9 (1995), 700.

[27] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October* 9 (2018), 2018.

[28] Carsten KW De Dreu and Michael A West. 2001. Minority dissent and team innovation: The importance of participation in decision making. *Journal of applied Psychology* 86, 6 (2001), 1191.

[29] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data feminism.* The MIT Press.

[30] Jordan D Dworkin, Kristin A Linn, Erin G Teich, Perry Zurn, Russell T Shinohara, and Danielle S Bassett. 2020. The extent and drivers of gender imbalance in neuroscience reference lists. *Nature neuroscience* 23, 8 (2020), 918–926.

[31] Leo Egghe. 2006. Theory and practise of the g-index. *Scientometrics* 69, 1 (2006), 131–152.

[32] Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. *Big Data & Society* 9, 1 (2022), 20539517221082027.

[33] Linton C Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry* (1977), 35–41.

[34] Lindred L Greer, Bart A de Jong, Maartje E Schouten, and Jennifer E Dannals. 2018. Why and when hierarchy impacts team effectiveness: A meta-analytic integration. *Journal of Applied Psychology* 103, 6 (2018), 591.

[35] Sandra G Harding. 2004. *The feminist standpoint theory reader: Intellectual and political controversies.* Psychology Press.

[36] Nancy Hartsock. 1983. The feminist standpoint: Developing the ground for a specifically feminist historical materialism. In *Discovering reality.* Springer, 283–310.

[37] Brent Hecht, Lauren Wilcox, Jeffrey P. Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, Danish Contractor, and Cathy Wu. 2018. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. First published on *ACM Future of Computing Academy Blog* (March 29, 2018). arXiv:2112.09544

[38] Victoria A Hill. 2008. Collaboration in an academic setting: Does the network structure matter. *Center for the Computational Analysis of Social and Organizational Systems* (2008).

[39] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* 102, 46 (2005), 16569–16572.

[40] Ayanna Howard and Charles Isbell. 2020. Diversity in AI: The Invisible Men and Women. *MIT Sloan Management Review* 62, 2 (2020).

[41] Matthew O Jackson. 2014. Networks in the understanding of economic behaviors. *Journal of Economic Perspectives* 28, 4 (2014), 3–22.

[42] Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583, 7815 (2020), 169.

[43] Os Keyes, Chandler May, and Annabelle Carrell. 2021. You Keep Using That Word: Ways of Thinking about Gender in Computing Research. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[44] Remy Kusters, Dusan Misevic, Hugues Berry, Antoine Cully, Yann Le Cunff, Loic Dandoy, Natalia Díaz-Rodríguez, Marion Ficher, Jonathan Grizou, Alice Othmani, et al. 2020. Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities. *Frontiers in Big Data* 3 (2020), 45.

[45] Rachael Lammey. 2016. Using the Crossref Metadata API to explore publisher content. *Sci Ed* 3, 3 (2016), 109–11.

[46] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9, 1 (2016).

[47] Bibb Latané and Martin J Bourgeois. 1996. Experimental evidence for dynamic social impact: The emergence of subcultures in electronic groups. *Journal of Communication* (1996).

[48] Harold J Leavitt. 1949. *Some effects of certain communication patterns upon group performance.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[49] Jongshin Lee, Yongsun Lee, Soo Min Oh, and B Kahng. 2021. Betweenness centrality of teams in social networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 31, 6 (2021), 061108.

[50] Loet Leydesdorff. 2007. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology* 58, 9 (2007), 1303–1319.

[51] Adam Liptak. 2021. As Harvard Case Looms at Supreme Court, Study Tests Value of Diversity. *New York Times* (2021).

[52] Luis López, Jose FF Mendes, and Miguel AF Sanjuán. 2002. Hierarchical social networks and information flow. *Physica A: Statistical Mechanics and its Applications* 316, 1-4 (2002), 695–708.

[53] Poppy Lauretta McLeod, Sharon Alisa Lobel, and Taylor H Cox Jr. 1996. Ethnic diversity and creativity in small groups. *Small group research* 27, 2 (1996), 248–264.

[54] Ida Momennejad, Ajua Duker, and Alin Coman. 2019. Bridge ties bind collective memories. *Nature communications* 10, 1 (2019), 1–8.

[55] Enys Mones, Lilla Vicsek, and Tamás Vicsek. 2012. Hierarchy measure for complex networks. *PloS one* 7, 3 (2012), e33799.

[56] Mathias Wullum Nielsen, Sharla Alegria, Love Börjeson, Henry Etzkowitz, Holly J Falk-Krzesinski, Aparna Joshi, Erin Leahey, Laurel Smith-Doerr, Anita Williams Woolley, and Londa Schiebinger. 2017. Opinion: Gender diversity leads to better science. *Proceedings of the National Academy of Sciences* 114, 8 (2017), 1740–1742.

[57] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine learning* 39, 2 (2000), 103–134.

[58] SU Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.

[59] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[60] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown.

[61] John F Padgett and Christopher K Ansell. 1993. Robust Action and the Rise of the Medici, 1400-1434. *American journal of sociology* 98, 6 (1993), 1259–1319.

[62] Corinne Post and Kris Byron. 2015. Women on boards and firm financial performance: A meta-analysis. *Academy of Management Journal* 58, 5 (2015), 1546–1571.

[63] F Rosenblat. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review* 65, 6 (1958), 386–408.

[64] Stuart Russell and Peter Norvig. 2002. Artificial intelligence: a modern approach. (2002).

[65] Lucía Santamaría and Helena Mihaljević. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* 4 (2018), e156.

[66] Michael Schreiber. 2008. A modification of the h-index: The hm-index accounts for multi-authored manuscripts. *Journal of Informetrics* 2, 3 (2008), 211–216.

[67] John Scott. 1988. Social network analysis. *Sociology* 22, 1 (1988), 109–127.

[68] David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. 2020. An Open Review of OpenReview: A Critical Analysis of the Machine Learning Conference Review Process. *arXiv preprint arXiv:2010.05137* (2020).

[69] Harry C Triandis, Lois L Kurowski, and Michele J Gelfand. 1994. Workplace diversity. (1994).

[70] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. *Science* 342, 6157 (2013), 468–472.

[71] Daan Van Knippenberg, Carsten KW De Dreu, and Astrid C Homan. 2004. Work group diversity and group performance: an integrative model and research agenda. *Journal of applied psychology* 89, 6 (2004), 1008.

[72] Madalina Vlasceanu, Michael J Morais, Ajua Duker, and Alin Coman. 2020. The synchronization of collective beliefs: From dyadic interactions to network convergence. *Journal of Experimental Psychology: Applied* 26, 3 (2020), 453.

[73] Judy Wajcman. 2019. How silicon valley sets time. *New Media & Society* 21, 6 (2019), 1272–1289.

[74] Stanley Wasserman, Katherine Faust, et al. 1994. Social network analysis: Methods and applications. (1994).

[75] Duncan J WATTS. 2003. Networks, dynamics and the small world phenomenon. *Amer. J. Sociology* 105, 2 (2003), 50–59.

[76] Ned Wellman, JM Applegate, John Harlow, and Erik W Johnston. 2020. Beyond the pyramid: alternative formal hierarchical structures and team performance. *Academy of Management Journal* 63, 4 (2020), 997–1027.

[77] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems. *AI Now* (2019).

[78] Langdon Winner. 1980. Do artifacts have politics? *Daedalus* 109, 1 (1980), 121–136.

[79] Anita Williams Woolley, Ishani Aggarwal, and Thomas W Malone. 2015. Collective intelligence and group performance. *Current Directions in Psychological Science* 24, 6 (2015), 420–424.

[80] Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 233–242.

[81] Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 408–427.

[82] Kevin JS Zollman. 2010. The epistemic benefit of transient diversity. *Erkenntnis* 72, 1 (2010), 17.