Seeding Network Influence in Biased Networks and the Benefits of Diversity

Ana-Andreea Stoica
Columbia University
astoica@cs.columbia.edu

Jessy Xinyi Han
Columbia University
xh2318@columbia.edu

Augustin Chaintreau
Columbia University
augustin@cs.columbia.edu

ABSTRACT

The problem of social influence maximization is widely applicable in designing viral campaigns, news dissemination, or medical aid. State-of-the-art algorithms often select "early adopters" that are most central in a network unfortunately mirroring or exacerbating historical biases and leaving under-represented communities out of the loop. Through a theoretical model of biased networks, we characterize the intricate relationship between diversity and efficiency, which sometimes may be at odds but may also reinforce each other. Most importantly, we find a mathematically proven analytical condition under which more equitable choices of early adopters lead simultaneously to fairer outcomes *and* larger outreach. Analysis of data on the DBLP network confirms that our condition is often met in real networks. We design and test a set of algorithms leveraging the network structure to optimize the diffusion of a message while avoiding to create disparate impact among participants based on their demographics, such as gender or race.

CCS CONCEPTS

• **Theory of computation** → *Graph algorithms analysis*; *Submodular optimization and polymatroids*; *Random network models*; • **Information systems** → *Web crawling*; **Social networks**; • **Human-centered computing** → *Social networks*; *Social network analysis*; • **Computing methodologies** → *Theorem proving algorithms*; *Probabilistic reasoning*.

KEYWORDS

social networks, influence, fairness, graph algorithms

ACM Reference Format:
Ana-Andreea Stoica, Jessy Xinyi Han, and Augustin Chaintreau. 2020. Seeding Network Influence in Biased Networks and the Benefits of Diversity. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3366423.3380275

1 INTRODUCTION

Data-driven algorithms are increasingly affecting high-stakes decisions in our everyday lives, including employment, online visibility, and justice. With their advent, the concern that some automated processes mirror or amplify bias against disadvantaged groups has become commonplace. The first observation regarding such issues is that different groups or communities are unequally represented: typically, minority groups are disproportionately absent from advantageous positions, creating a diversity gap. This immediately leads us to the question: when can an intervention be effective in (partially) restoring that lost diversity? As a social problem, legal interventions and public opinion pressure may be required to alleviate this disparity gap, both raising the stakes in understanding the possible benefits and drawbacks of diversity-enhancing rules.

Quite often, two sides emerge in this dilemma: one that argues, especially in situations with historical prejudice, that awareness of demographics and community-affiliation is essential in restoring some parity as a goal towards eventual fairness, and one that is more cautious, either from being reluctant to identify various groups, or from fear of the ramifications of differential treatment. Some of the arguments against diversity-enhancing interventions may also be more cynical: the interventions might come at a non-negligible cost to efficiency [16]. It is therefore especially critical to identify cases where these views can be reconciled: one in which diversity-enhancing strategies not *only* promote fairness and equal representation but *also* measurably improve the outcome of a process. This is the focus of our paper, where we carefully identify through a theoretical argument backed by empirical validation that diversity-enhancing rules maximize the outreach for the well-known influence maximization problem, showing that being aware of the features of the individuals and the network structure is crucial in obtaining a more equitable and better outcome. Our result is necessarily subtle (it comes from counteracting a form of bias in an emerging graph), and thus the exact mathematical conditions required are non-trivial and depend on the various graph and algorithmic factors. Nevertheless, an evaluation of these techniques on experimental data shows that the main property we prove holds in most practical conditions: we find that improving diversity in seeding has a drastic effect on restoring fairness in the outreach, rarely at a significant cost to efficiency. In this paper, we present the following contributions:

- We provide a model of network growth that embeds social bias through unequal communities and prove the existence of an analytical condition in which diversity acts as a catalyst for efficiency. Our model allows an in-depth analysis of seed selection heuristics with partial network information that choose early adopters based on their network centrality (in this case, their degree), proving that a large enough seed set leads to better diversity and efficiency, while a small seed set may incur a cost of fairness. Through our model and synthetic networks, we show that including sensitive attributes in the input of such algorithms is crucial in the

This material is based upon work supported by the National Science Foundation under Grant No.1514437 and 1761810, as well as by a J.P. Morgan Ph.D. Fellowship.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380275

design of more equitable heuristics and provide an analytical study of the conditions in which disparity in seed selection occurs. (Section 3).
- We analyze diversity-enhancing interventions on seeding using a dataset of Computer Scientists collected from DBLP, confirming that our results qualitatively hold even outside of the strict assumptions embedded in our theoretical model. We find that our theoretical lower bounds on the seed set size are conservative (Section 4).
- Finally, we extend our results to evaluate our findings across multiple centrality measures. For scenarios including a small seed set budget, we find that an alternative diversity-enhancing rule is effective at improving equity in information diffusion at no cost to outreach (Section 5).

Social influence maximization has been a widely-studied problem in online networks, having impactful applications in information diffusion, disease spread, marketing strategies, and many others. A classical method is to pick a set of individuals who will be the "early adopters" (a seed set), who either adopt the desired message or product out of their own will or receive it at no cost. As they share it with their friends, who may or may not adopt it themselves, the process continues into a cascade. In the traditional formulation inspired by marketing, one may aim to maximize the number of people who adopt the product (the outreach). That goal was recently called into doubt in cases where fairness is paramount, e.g., when social influence is leveraged for public health, with a mandate to equally protect members of various groups. It was quickly pointed out that even if algorithms choose seeds based on their network position (e.g., "most central") and seemingly ignore demographics, properties of social networks also encode historical biases and gender artifacts that algorithms can reinforce. The state-of-the-art is to propose parity-restoring clauses inside an optimization problem, which have been reported to navigate a fairness-efficiency trade-off [1, 17, 42].

We aim to understand biases present in the graph through a modeling approach, so that an algorithm may learn how to correct for them. Far from introducing fairness as a top-down objective, we experiment with interventions from the start and examine the conditions under which diversity comes not at a cost, but as a catalyst for maximizing outreach. We focus on situations where information is sparse or the graph must be learned or approximated, as this has been the most active line of work in the recent analysis of influence maximization. In retrospect, all previous results showing that seed diversity comes as a costly intervention with respect to efficiency are built on the specific case of a greedy algorithm with complete information.

2 RELATED WORK

Several online services have recently come under scrutiny for reproducing or amplifying bias against disadvantaged groups in society [15, 18, 34]. Information diffusion is subject to similar concerns, for instance via online advertising [2, 43], search engines [9], or profile recommendations [40]. Leveraging social networks to accelerate information diffusion follows a similar trend [1, 17, 42].

Judiciously choosing a seed set to create a cascading effect and maximize the eventual spread of influence remains a practical challenge. Even when the network topology and the information diffusion properties among nodes are known, most approximate algorithms require costly polynomial steps [22, 23]. This is why seed set selection deployed in practice exploits simpler heuristics based on neighbors or centrality properties [11, 12, 24, 47]. These “rules of thumb” are *a fortiori* indispensable to guide the choice of seeds when the network is only partially known [37] or learned from experiments [6]. Building on a preliminary version of our results [39], no other paper so far has considered fairness in such a context to our knowledge. Recent work [1, 17, 42] starts from the different assumption that a polynomial greedy heuristic is run on a full information network. Like these papers, we find that algorithms that are agnostic to community affiliation or sensitive attributes when maximizing influence may come at a cost to diversity in both the choice of seeds and the population reached. However, unlike this past work, we conclude that diversity-enhancing rules can restore diversity *without decreasing performance*. Indeed, contrary to it, our theoretical analysis and empirical results give conditions under which fairness presents a (modest) gain. We note that our results resonate with a more general line of work concerned with *algorithmic fairness* when processing human data, shedding light onto the tendency of classification algorithms to reproduce or amplify historical prejudices (see [3, 8, 46] and references therein).

Some of the conditions we analyze resemble the these for ensuring equity in output [26, 33], and the diversity-enhancing seeding we define resembles previous techniques for fair personalization [10, 19, 27]. Our results contribute to recent evidence suggesting that sensitive attributes should not be ignored but can be leveraged to simultaneously improve fairness and accuracy [25]. The apparent paradox between fairness and accuracy (or efficiency) is often explained [13, 38] by the presence of an implicit bias when evaluating candidates of certain subgroups [27] according to simple metrics. One may sometimes refine the metric used, leveraging sensitive attributes, ideally so that selection is solely dependent on the true outcome probability.

While diversity has been long studied as having performance-enhancing benefits in many domains [21, 35], showing its strengths beyond moral reasons, theoretical models encompassing diverse agents have shown complicated dynamics of minority suppression due to homophilic behavior and rich-get-richer effects [5, 29]. Our work complements these results as we provide a theoretical foundation for why social network metrics follow spontaneous dynamics that give rise to such effects and when one can provide a strong justification to introduce diversity-enhancing rules.

3 THEORY OF DIVERSITY SEEDS

In understanding the true causes of bias in influence maximization, we introduce a theoretical model that reproduces disparity between different communities in social networks. While emulating the organic growth of communities of individuals with different features or interests, we are able to mathematically analyze the conditions in which bias is firstly created and then captured by algorithms that leverage network centrality, investigating the complex relationship between diversity and optimality.

3.1 Model of biased networks and influence

In studying the effect of information diffusion for different communities, we use an established model of network growth that encodes two unequal populations, namely the **biased preferential attachment model**. Built on the classical Barabasi-Albert model [7] and extended to encompass more than one community [5], the model has been shown to have real applications by encoding patterns of human connections, from community affiliation to homophily and a rich-get-richer effect [4, 5, 28, 40]. In this model, the network grows through the sequential addition of new nodes, satisfying three properties:

- *Community affiliation:* When a node enters the network, it chooses one of two labels: red (R) with probability r, and blue (B) with probability $1 - r$. By setting $r < 0.5$, the red community is the minority. For simplicity, we focus on a network with two communities, but the model immediately generalizes to any finite number of communities.
- *Rich-get-richer:* The new node chooses to connect to some other node according to *preferential attachment* (chooses a node with probability proportional to that node's degree). This step is repeated $d > 1$ times, giving each node an out-degree of d.
- *Homophily:* If the two nodes from the previous step have the same label, an edge is formed; otherwise, the new node accepts the connection with probability ρ, where $0 < \rho < 1$, and the process is repeated until an edge is formed. This models individuals' tendency to be more likely to connect with people belonging to the same group.

Previous studies of this model [5] show that the degree distribution follows a power law with different coefficients: $\beta(R)$ and $\beta(B)$ for the two populations, where $\beta(R) > 3 > \beta(B) > 2$, showing that the minority population encounters a "glass ceiling" effect in acquiring a high number of connections, or equivalently, that minority nodes will be under-represented at the top of the degree hierarchy. With a large body of evidence for its pervasiveness in offline [14] and online [5, 34] networks, the glass ceiling effect captures such inequality due to historical prejudices in top-ranked individuals of a population. This model in particular reproduces the observations [1, 17, 42] that seeds may be chosen disproportionately among members of the majority group (see arguments below), affecting the proportions of minority and majority nodes in outreach from a fairness perspective.

Information diffusion: Assuming the network grows according to the biased preferential attachment model, we use the *independent cascade model* for information diffusion under a seed set budget constraint, where each individual adopts the information transmitted to them from one of their friends with a probability p, independently for every friend, for $0 < p < 1$. We assume an undirected graph in this study. Note that, given the complexity of the graph, any other information diffusion model appears much harder to tackle. As we prove below that even the simplest case of an approximation of the independent cascade is non-trivial, how diversity affects outreach in more complex diffusion processes becomes an interesting avenue for future work.

3.2 A theoretical proof of diversity benefit

We denote by $\phi_V(S)$ the *influence* of a set S over a network V, i.e. the set of nodes activated by the independent cascade model. We can formally define statistical parity in outreach as

Definition 3.1. *For a graph $G = (V, E)$ that adopts information through an independent cascade model from a seed set S, there is statistical parity in the outreach for communities $C_1, \cdots C_k$ if $\forall i \neq j$:*

$$\frac{\mathbb{E}|\{u \in \phi_V(S) | u \in C_i\}|}{|C_i|} = \frac{\mathbb{E}|\{u \in \phi_V(S) | u \in C_j\}|}{|C_j|}. \qquad (1)$$

This notion can ensure that the product or the information being spread reaches the population in a representative way, i.e. reaches the same percentage of each community. Note that, even for a particular fixed seed set S, its outreach (or influence $\phi_V(S)$) may be costly or impossible to compute, for instance when not all pertinent details are known. In such cases, we can require statistical parity in seeding as a proxy towards parity in outreach (i.e., the same condition where $\phi_V(S)$ is replaced by S). We note, however, that parity seeding appears neither necessary nor sufficient.

Similarly, we can define statistical parity for a seed set S, and use this notion as a *tool* to gain more parity in the outreach. Since we do not have the full network information, optimizing directly for parity in the outreach becomes impossible, and thus we use the seed set as a proxy. This is particularly important in news spread, where manipulating distribution of news can lead to misinformation and to amplifying an echo chamber effect, and maybe even more compelling for public health concerns.

Seeding Strategically: While the greedy algorithm provides a good approximation of the best seed set in the independent cascade model [22], it is costly to compute and typically inaccessible in many practical cases, when the network is only partially observed and the nodes are only characterized by their degree or other centrality measures. As mentioned before, a large body of literature focuses on maximizing influence through different simple heuristics [11, 12, 24, 47], obtaining scalable algorithms for seed selection by leveraging network statistics such as degree centrality and distance centrality. For concreteness, we focus in this section on analyzing *degree centrality* before validating our results for diversity in other heuristics in Section 5. Degree centrality is arguably simplistic but already presents an interesting behavior and approximates greedy for a small probability p of diffusion [22]. We hope our study provides a starting point for analyzing other similar heuristics. Note also that the glass ceiling effect proved in [5] implies that seeding based on degree centrality can be arbitrarily far from having statistical parity in the seed set.

Formally, the baseline *agnostic seeding* heuristic based on degree centrality sets a threshold $k(n)$ for degree above which all nodes are chosen as seeds, regardless of their color or community affiliation:

Definition 3.2. *The baseline agnostic seeding defines the seed set of a bi-populated network V of red (R) and blue (B) nodes as $S_{k(n)} = \{v \in V | deg(v) \geq k(n)\}$.*

Figure 1 (left) shows an illustration of that process for two communities of red and blue nodes, where the blue nodes are the majority and have a higher degree. As a given budget for the size of

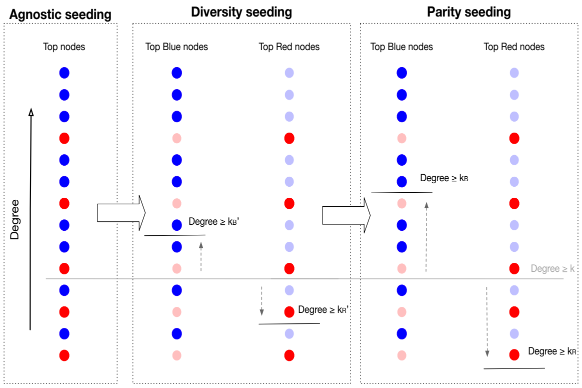

Figure 1: Differentiated thresholds for strategic seeding.

the seed set is equivalent to choosing nodes above a certain degree threshold k, a color-agnostic seeding heuristic would choose nodes from the degree ranking, while a *strategic* one may include the color of the nodes in the choice of seeds.

Through our concept of diversity, we define strategic heuristics based on statistical parity of the seeds, called *parity seeding* (Figure 1 right) by increasing the threshold for the blue nodes and decreasing it for the red nodes in order to achieve the same ratio of red and blue nodes in the seed set as it is in the general population while preserving the seed set budget:

Definition 3.3. *Parity seeding defines the seed set of a bi-populated network V of red (R) and blue (B) nodes based on two differentiated thresholds $k^R(n)$ and $k^B(n)$ as*

$$S^R_{k^R(n)} \cup S^B_{k^B(n)} = \{v \in R | deg(v) \geq k^R(n)\} \cup \{v \in B | deg(v) \geq k^B(n)\}$$

$$\text{such that } |S_{k(n)}| = |S^R_{k^R(n)} \cup S^B_{k^B(n)}| \text{ and } \frac{|S^R_{k^R(n)}|}{|S^B_{k^R(n)} \cup S^B_{k^B(n)}|} = \frac{|R|}{|V|}. \tag{2}$$

We also propose an intermediary variant (Figure 1 center), which we denote *diversity seeding*, by allowing the ratio of the two populations in the seed set to vary between the agnostic one and the parity one, again by setting differentiated thresholds for the blue and red nodes (which need not be the same as in Definition 3.3):

Definition 3.4. *Diversity seeding defines the seed set of a bi-populated network V of red (R) and blue (B) nodes based on two differentiated thresholds $k^R(n)$ and $k^B(n)$ as*

$$S^R_{k^R(n)} \cup S^B_{k^B(n)} = \{v \in R | deg(v) \geq k^R(n)\} \cup \{v \in B | deg(v) \geq k^B(n)\}$$

$$\text{such that } |S_{k(n)}| = |S^R_{k^R(n)} \cup S^B_{k^B(n)}|. \tag{3}$$

Finally, we need to introduce an approximation: since the function $\phi_V(S)$ remains elusive in such complex network, we approximate it in the analysis presented here by assuming influence spread over one hop. This approximation is accurate for small probability of diffusion (e.g. $p = 0.01$), not uncommon in practice. The results for when this assumption does not hold are validated numerically for more general values of p in Sections 4 and 5. Interestingly, it seems that by making this approximation we tend to underestimate the benefits of diversity.

Given these concepts, we find an analytic condition for which strategic seeding improves both efficiency and outreach in the diffusion process:

Theorem 3.1. *In the independent cascade model in a network V that follows the biased preferential attachment model with two communities of red (R) and blue (B) nodes, there exists $k_0(n)$ known in closed form such that the following statements hold:*

(*i*) *When $k(n) \leq k_0(n)$, for a range of differentiated thresholds $k^R(n) < k^B(n)$ chosen to maintain the same expected budget as agnostic seeding, diversity seeding obtains a larger expected outreach while getting closer to outreach parity, i.e.,*

$$\mathbb{E}(|\phi_V(S_{k(n)})|) < \mathbb{E}(|\phi_V(S^R_{k^R(n)} \cup S^B_{k^B(n)})|),$$
$$\textit{with constraint } \mathbb{E}(|S_{k(n)}|) = \mathbb{E}(|S^R_{k^R(n)} \cup S^B_{k^B(n)}|), \textit{ and} \tag{4}$$

$$\frac{\mathbb{E}(|\phi_V(S^R_{k^R(n)} \cup S^B_{k^B(n)}) \cap R|)}{\mathbb{E}(|\phi_V(S^R_{k^R(n)} \cup S^B_{k^B(n)})|)} > \frac{\mathbb{E}(|\phi_V(S_{k(n)}) \cap R|)}{\mathbb{E}(|\phi_V(S_{k(n)})|)}. \tag{5}$$

(*ii*) *When $k(n) \geq k_0(n)$, all diversity seeding heuristics with the same expected budget as agnostic seeding can be closer to outreach parity but always obtain a smaller expected outreach.*

This result illustrates how seeding diversity and seeding efficiency are intricately related: when aiming for selecting too few seeds, diversity hurts overall outreach, intuitively because each chosen seed among the minority group reaches a smaller community. Having less "room to grow" than a typical majority node, minority seeds are far down in the rankings. One should lower the threshold significantly to move towards a fair representation and that comes at a significant cost. However, with more seeds and lower $k(n)$, the situation starts to improve and it becomes less and less costly to bring proportional representation in outreach. As we approach a critical point, which is entirely known although its exact value remains quite complex (see proof below), the cost of equal representation virtually vanishes. Then, remarkably, the opposite trend emerges past this point, where enhancing diversity can increase the outreach both in absolute size and in proportional representation. Intuitively, this is because a seed from a minority group overlaps much less with prior seeds with better rankings, making its individual contribution substantially better than even a slightly higher rank majority node.

Finally, we find that parity seeding is also able to be more equitable and more efficient, given a sufficiently large seed set budget, albeit much larger than for diversity heuristic (we omit the proof here due to lack of space):

Corollary 3.2. *In the independent cascade model in a network V that follows the biased preferential attachment model with two communities of red (R) and blue (B) nodes, there exists $k_1(n)$ known in closed form such that the previous statements hold for the parity seeding heuristic.*

The rest of the section provides a sketch of the proof for Theorem 3.1, limited due to space constraints.

Proof. The intuition behind the proof is two-fold: on one hand, the biased preferential attachment model allows us to compute the probabilities of diffusion for different communities and to note that homophily preserves influence within the group that seeds lie in; on the other hand, we are able to use the diminishing marginal return property of submodular functions in showing the benefits of diversity. In a sense, although the high degree nodes are better connected, they tend to have overlapping influence spheres, leaving minority communities untapped. Note that an upper bound on the degree threshold translates into a lower bound for the seed set size. For the first part of the theorem, we continue with the following lemma, for which we use the biased preferential attachment model dynamics: □

Lemma 3.1. *The expected size of a seed set that includes nodes of degree at least $k(n)$ is:*

$$\begin{aligned}\mathbb{E}(|S_{k(n)}|) &= n \cdot \alpha \cdot d \cdot \frac{\beta(R)-2}{\beta(R)-1} \cdot k(n)^{1-\beta(R)} \\ &+ n \cdot (1-\alpha) \cdot d \cdot \frac{\beta(B)-2}{\beta(B)-1} \cdot k(n)^{1-\beta(B)},\end{aligned} \tag{6}$$

where α is the fraction of edges with one end in R.

Proof. We can write

$$\mathbb{E}(|S_{k(n)}|) = \mathbb{E}(|S_{k(n)} \cap R|) + \mathbb{E}(|S_{k(n)} \cap B|). \tag{7}$$

Knowing the degree distribution of the model, we know that the number of red/blue nodes of degree k is proportional to $k^{-\beta(R/B)}$. Thus, we get

$$\mathbb{E}(|S_{k(n)} \cap R|) = \sum_{k' \geq k(n)} k'^{-\beta(R)} \cdot C_R \tag{8}$$

$$= n \cdot \alpha \cdot d \cdot (\beta(R)-2) \cdot \int_{k(n)}^{\infty} x^{-\beta(R)} dx \tag{9}$$

$$\Rightarrow \mathbb{E}(|S_{k(n)} \cap R|) = n \cdot \alpha \cdot d \cdot \frac{\beta(R)-2}{\beta(R)-1} \cdot k(n)^{1-\beta(R)} \tag{10}$$

$$\mathbb{E}(|S_{k(n)} \cap B|) = \sum_{k' \geq k(n)} k'^{-\beta(B)} \cdot C_B \tag{11}$$

$$= n \cdot (1-\alpha) \cdot d \cdot (\beta(B)-2) \cdot \int_{k(n)}^{\infty} x^{-\beta(B)} dx \tag{12}$$

$$\Rightarrow \mathbb{E}(|S_{k(n)} \cap B|) = n \cdot (1-\alpha) \cdot d \cdot \frac{\beta(B)-2}{\beta(B)-1} \cdot k(n)^{1-\beta(B)} \tag{13}$$

□

Since we want to find $k^R(n)$ and $k^B(n)$ that satisfy our goal in equation 4, we set $k^B(n) = k(n) \cdot x$, for a variable x. In order to preserve the total number of seeds, we may compute the minimum degree $k^R(n)$ in closed form in terms of the network parameters. Writing these as functions of x, we define:

$$F(x) = \mathbb{E}(|\phi_V(S^B_{k^B(x)} \cup S^R_{k^R(x)})|) - \mathbb{E}(|\phi_V(S_{k(n)})|), \tag{14}$$

with the goal of solving for x such that $F(x) > 0$. Knowing that the blue community is majority, we would like to increase diversity and thus to choose $x > 1$ for which $F(x) > 0$. In order to do that, we explicitly write the expected number of influenced nodes as:

$$\begin{aligned}&\mathbb{E}(|\phi_V(S_{k(n)})|) \\ &= n \cdot \mathbb{P}(v \in B) \cdot \mathbb{P}(v \text{ influenced by one of its } d \text{ edges} | v \in B) \\ &+ n \cdot \mathbb{P}(v \in R) \cdot \mathbb{P}(v \text{ influenced by one of its } d \text{ edges} | v \in R).\end{aligned} \tag{15}$$

Then, considering the blue population (and performing the same computations for the red one),

$$\begin{aligned}&\mathbb{P}(v \text{ influenced by one of its } d \text{ edges} | v \in B) \\ &= 1 - (1 - \mathbb{P}(v \text{ influenced by one edge} | v \in B))^d.\end{aligned} \tag{16}$$

We can compute the probability for a node to be influenced by one of its edges as:

$$\begin{aligned}&\mathbb{P}(v \text{ influenced by one edge} | v \in B) = \\ &\mathbb{P}(v \text{ forms edge } B \longrightarrow B | v \in B) \cdot \frac{|\{e=(u,w) | u \in B, w \in S_{k(n)} \cap B\}|}{|\{e=(u,w) | u, w \in B\}|} + \\ &\mathbb{P}(v \text{ forms edge } B \longrightarrow R | v \in B) \cdot \frac{|\{e=(u,w) | u \in B, w \in S_{k(n)} \cap R\}|}{|\{e=(u,w) | u \in B, w \in R\}|}.\end{aligned} \tag{17}$$

Computing and replacing these expressions in our original equation 14, we obtain a closed-form equation which we can solve for k. Using the unequal dynamics provided by the glass ceiling effect, we find that the power law exponent of the degree distribution of the blue community dominates the one for the red community. We obtain a bound k_0 for which: for $k < k_0, \exists x > 1$ s.t. $F(x) > 0$, and for $k > k_0 \forall x > 1, F(x) < 0$, where k_0 is computed as below:

$$k_0 = \left(\frac{2(1-\alpha)}{\left(\frac{1-r}{r}\right)^{\frac{1}{d-1}} - 1} \left(\frac{\frac{1-r}{\alpha\rho+1-\alpha}}{\frac{\rho r}{\alpha+\rho(1-\alpha)} + \frac{1-r}{\alpha\rho+1-\alpha}} \left(\left(\frac{1-r}{r} \right)^{\frac{1}{d-1}} + 1 \right) - 1 \right) \right)^{\frac{1}{\beta_B - 2}} \tag{18}$$

Thus, we conclude with an analytical condition depending on the network parameters for the bound k_0, for which $k < k_0$ allows diversity seeding to achieve a better outreach than agnostic seeding, while for $k > k_0$ the opposite occurs and diversity comes at a cost.

These results show that, given enough seeds, diversity seeding is able to achieve better outreach than being agnostic to community affiliation. Moreover, it is also able to *nudge* the outreach in a more equitable way, promoting diversity in the outreach:

Corollary 3.3. *In the above-mentioned conditions, the minority group R obtains better parity in the outreach:*

$$\frac{\mathbb{E}(\phi_V(S^R_{k^R(n)} \cup \phi_V(S^B_{k^B(n)})) \cap R)}{\mathbb{E}(\phi_V(S^R_{k^R(n)} \cup \phi_V(S^B_{k^B(n)}))} > \frac{\mathbb{E}(\phi_V(S_{k(n)}) \cap R)}{\mathbb{E}(\phi_V(S_{k(n)})}. \tag{19}$$

We only sketch the proof intuition showing that diversity in the outreach also increases: since we are assuming a small probability of diffusion and the two communities are homophilic, the red nodes who are added to the seed set are influencing predominantly other red nodes, while the blue nodes that are being removed from the seed set for achieving better seed parity are losing their influenced blue nodes. Thus, we can show that in expectation, the number

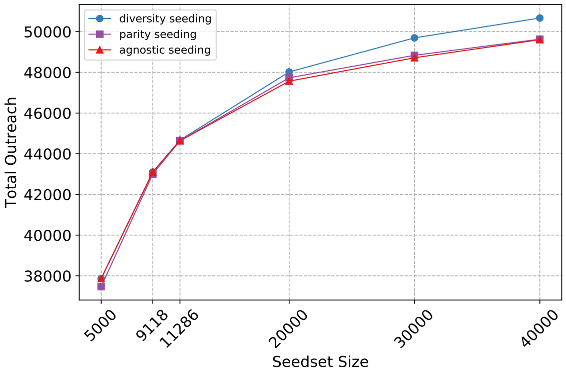

Figure 2: Seed set size versus outreach for $r = 0.186$, $n = 53,307$, $\rho = 0.294$, $\alpha = 0.179$, $d = 3$.

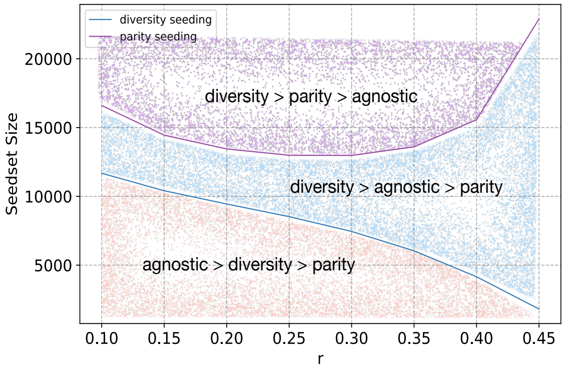

Figure 3: Minimum seed set size budget as r varies between 0 and 0.5, and $n = 53,307$, $\rho = 0.294$, $d = 3$.

of red nodes in the outreach increases, while the number of blue nodes decreases, leading to a more equitable ratio of the red nodes in the outreach.

3.3 How many seeds are enough?

As the theory suggests, given a large enough seed set, diversity can act as a catalyst for increasing outreach. To realistically understand these conditions, we compute the approximate bounds of the size of the seed set emerging from the conditions that let us achieve both fairness and efficiency for a synthetic network.

Using equation 18, we are able to approximate the upper bound of the minimal degree in our seed set. From this, we can approximate the number of nodes with a higher degree than this bound, obtaining the minimum seed set size to achieve both fairness and efficiency. We present this numerical computations for a simulated network that resembles the DBLP dataset we used for empirical evaluation, setting $n = 53,307$ nodes, $r = 0.186$ the ratio of women, $d = 3$ average degree. We infer $\alpha = 0.179$ from a fixed-point equation of the biased preferential attachment model [5], and we compute $\rho = 0.294$ as the homophily parameter associated to DBLP.[1]

Figure 2 illustrates the performance of our three seeding heuristics for this network: while parity seeding needs 11, 286 seeds (or 21% of the population) in order to be better than agnostic seeding, diversity seeding only needs 9, 118 seeds (or 17% of the population). For a smaller seed set, with 5, 000 nodes, both strategic heuristics perform worse than the agnostic one, as theory predicts, yet within a marginal bound. As the seed set size increases, parity seeding leads to a similar outreach as being agnostic, while enforcing diversity seeding leads to a better outreach.

Finally, we vary the minority proportion r to understand the relationship between fairness and network disparity. Figure 3 shows the different phase transitions that occur, for the two necessary seed set budgets for parity seeding and diversity seeding, plotting the necessary seed set size to encompass agnostic seeding as a function of r for diversity seeding (blue line) and parity seeding (purple line). As these budgets show the minimum seed set size for the strategic heuristics to become better than the agnostic one, the red region shows the parameter space for which both strategic heuristics encounter a loss in outreach to agnostic seeding (equivalent to having too few seeds), the blue region shows that for a moderate amount of seeds diversity is able to do better than agnostic but parity is not, and finally the purple region shows that for a large enough seed set, parity seeding is also able to perform better than agnostic seeding (with diversity performing best since it's a relaxed heuristic).

As r increases and the groups become more equal, we need fewer seeds to achieve both better equity and efficiency for diversity seeding, but not necessarily for the parity heuristic. Indeed, as r increases from 0.1 to 0.45, the required budget for the seed set size decreases from 17.8% to 4.3% of the total population for diversity seeding, while parity seeding starts needing more seeds as r approaches 0.5. Intuitively, a more equal population requires fewer seeds that can achieve both fairness and efficiency, but as the population becomes more equal, it becomes harder to achieve strict statistical parity, as we would have more minority nodes in with high degree and thus the minority population would have been already partially influenced. Moving from partially influenced to being influenced in a truly proportional way requires a much higher cost than a relaxed condition on parity as our diversity seeding heuristic ensures, and thus diversity seeding becomes a powerful tool in achieving fairness.

4 DIVERSITY BENEFIT IN PRACTICE?

As the theoretical model provides us with a rigorous analysis of the relationship between diversity and efficiency (when diversity comes at a cost, and when it can improve efficiency), we aim to understand its applicability in real-life scenarios. To such end, we collected a co-authorship dataset of Computer Scientists from DBLP [30], an online database that records most publications in Computer Science. As professional opportunities for Ph.D. students, post-docs, as well as collaborations, are often shared through word of mouth or email

[1] Since rejected edges are lost, computing ρ is not as simple as counting the number of cross-community edges. However, as one can approximate r and α, we can replace them as parameters in the function F from Lemma 4.3 from [5], knowing that F has α as its fixed point. Thus, $F(\alpha) - \alpha$ must be equal to 0 as a function of ρ from the fixed-point equation, so we can obtain an approximate value for ρ.

lists in the Computer Science community, we aim to study the effect of network structure on who receives such opportunities, in a field known to exhibit gender and race imbalance [20, 36, 41, 45].

4.1 Inequality in scientific networks

We collected the first 200, 000 articles from DBLP, recording the listed authors as nodes and creating an undirected edge between any two authors who have co-authored an article. We then extracted the largest connected component of this graph, resulting in 53, 307 nodes and 288, 864 edges.

Perceived gender: similar to previous studies [5, 31], we used the authors' first names in understanding their perceived gender based on the Social Security Data (SSN) in the U.S. between 1940 and 2007. The SSN data provides us with a list of names and genders, from which we kept only those with at least 50 occurrences and which assign one of 'male' or 'female' at least 95% of the time. This created a dictionary of 32, 676 first names-gender with low ambiguity, which allowed us to infer the perceived gender of our graph. Out of the 53, 307 people, about 19% are perceived as female and 81% as male, confirming a large imbalance known in the field.

The imbalance in representation leads to an imbalance in degree, shown in Figure 4, illustrating the percentage of female researchers amongst those with degree at least x, as x varies between 1 and the maximum number of co-authors in the network. The size of the points represents the relative size of the population who has degree higher than the x-axis value. We note that there are always more men than women for every degree in the network. Except for a core of researchers who are very well-connected, the degree distribution resembles a power law, with very researchers of high degree. Although women do reach quite high degrees, the most well-connected researchers are still men and there are very few people with such high degrees. Indeed, choosing a seed set of size 5, 000−10, 000 is equivalent to choosing people of degree at least 8 or 10, for which the fraction of women on the y-axis is 15%. We further notice that the network is quite clustered, the vast majority of edges being intra-community. Thus, while the glass ceiling effect is not extremely pronounced, disparity in influence still exists by choosing only a few top people in the seed set and spreading influence in their community, leading to under-representing women.

Homophily: Beyond the inequality in the degree distribution of the two communities, the data also exhibits a tendency of people to connect more with those of the same perceived gender. Indeed, women have 24.3% of their edges directed at other women, while men connect 83.5% of their edges towards other men. Therefore, although many inter-community edges do exist, both genders exhibit moderate homophily, reproducing the theoretical model from Section 3.

Limitations: this gender inference method is constrained by limitations on gender binarity in the SSN data, as well as the exclusion of people who have more 'ambiguous' names, as defined above. Above all, it is not intended to replace the private identity, but rather to reflect the public perception, which is still a major factor in discrimination [32].

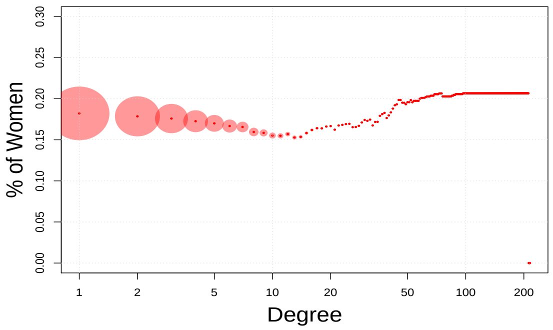

Figure 4: Log-scale plot of degree versus fraction of women among researchers with degree at least x, sized by the relative proportion of population of that degree or larger.

4.2 Practical evaluation of diversity seeding

Algorithms: We implement the three degree heuristics for seed selection on this dataset: *agnostic seeding*, *parity seeding*, and *diversity seeding*. While the agnostic heuristic chooses nodes above a certain degree threshold, the strategic ones essentially choose *differentiated* degree thresholds for each community, just like in the theoretical set-up. To implement diversity seeding, we vary a relaxation parameter ζ that allows us to shift between agnostic and parity seeding and report the best results in terms of outreach as a ζ varies, averaged over 1, 000 iterations.

Experiments: We simulate the cascading process in DBLP for these three seeding algorithms. While theory provides a bound for the seed set size needed to be more fair and more efficient, the model assumptions render it only approximate in real life, where we find that we need even fewer seeds to achieve our goal and that theory only provides a conservative estimate of the seed set size bound. We compute the minimum seed set size budget that theory predicts to achieve both fairness and efficiency in the strategic algorithms to be around 9, 100 nodes (∼ 17% of the population). We experiment with seed sets of 1, 000, 5, 000, and 9, 100 nodes for $p = 0.01$ and $p = 0.1$ (Table 1). We bold the values for the best heuristic for each row and seedset size, and show the increase and decrease of outreach values of the strategic heuristics in green and red arrows, respectively, as compared to the agnostic seeding baseline.

Indeed, choosing a relatively small seed set of 1, 000 nodes, both strategic heuristics perform worse than agnostic seeding, obtaining a lower outreach (Table 1 top) while increasing diversity. However, increasing the seed set size to 5, 000 nodes, we note that the diversity seeding slightly increases the outreach, yet parity seeding one decreases it, while for a seed set size of 9, 100 nodes, both strategic heuristics perform slightly better than the agnostic one. This confirms our theoretical findings that when the seed set size is too low, achieving parity comes at a cost of efficiency, but increasing it diversity comes at no cost to efficiency, but quite the opposite.

Although the outreach increase is marginal, what is impressive is to notice that, for 9, 000 seeds, replacing 100 (for diversity seeding) and 200 (for parity seeding) male seeds with female seeds leads to a

Table 1: Results for the DBLP dataset for $p = 0.01$ and $p = 0.1$, for $1,000$, $5,000$, and $9,100$ seeds.

	1,000 seeds			5,000 seeds			9,100 seeds		
$p = 0.01$	Agnostic seeding	Parity seeding	Diversity seeding	Agnostic seeding	Parity seeding	Diversity seeding	Agnostic seeding	Parity seeding	Diversity seeding
Total outreach	**1,149.15**	↓1,147.874	↓1,149.1	5,410.748	↓5,408.762	↑**5411.191**	9,554.934	↑9,555.559	↑**9,556.349**
F outreach	191.95	↑**210.456**	↑196.6	862.191	↑**1,004.232**	↑892.11	1,581.842	↑**1,776.037**	↑1,679.423
M outreach	**957.2**	↓937.418	↓952.5	**4,548.557**	↓4,404.53	↓4,519.081	**7,973.092**	↓7,779.522	↓7,876.926
F % in outreach	0.167	↑**0.183**	↑0.171	0.15934	↑**0.18567**	↑0.165	0.16555	↑**0.186**	↑0.176
F seeds	165	**185**	170	784	**930**	815	1,490	**1,690**	1,590
$p = 0.1$	Agnostic seeding	Parity seeding	Diversity seeding	Agnostic seeding	Parity seeding	Diversity seeding	Agnostic seeding	Parity seeding	Diversity seeding
Total outreach	3,479.65	↓3,460.37	↑**3,480.2**	9,861	↓9,847.4	↑**9,862.73**	14,343.6	↓14,337.63	↑**14,344.6**
F outreach	612.72	↑**616.9**	↑616.5	1,710	↑**1,810.547**	↑1,745.3	2,498.87	↑**2,642.333**	↑2,534.3
M outreach	**2866.93**	↓2843.47	↓2863.7	**8,151**	↓8,036.861	↓8,117.43	**11,844.725**	↓11,810.3	↓11,695.3
F % in outreach	0.176	↑**0.178**	↑0.177	0.17341	↑**0.1838**	↑0.17695	0.174	↑**0.1843**	↑0.176672
F seeds	165	**185**	175	784	**930**	830	1,490	**1,690**	1,540

decrease in male outreach which is overcome in both heuristics by the increase in female outreach, even if females are less than 20% of the population and have lower degrees.

While the case of $p = 0.01$ can be viewed as an approximation for our one-hop theoretical model (since when the conducting probability is as low as 0.01, the two-hop influence is negligible), as we increase the conducting probability, the one-hop model is no longer an accurate approximation of our simulation.

Table 1 (bottom) illustrates the $p = 0.1$ case, showing a larger outreach achieved by both strategic heuristics than for $p = 0.01$. Indeed, even for $1,000$ seeds, diversity seeding is able to do better than agnostic seeding. Furthermore, there is a smaller loss in the male outreach that was evident for $p = 0.01$, due to more cross-community interaction, since information is able to travel beyond the seeds' immediate neighbors. Although parity seeding achieves a better female ratio in the outreach, diversity seeding gains a better total outreach and a better outreach for males as well, needing far fewer seeds to do so.

5 DIVERSITY SEEDS IN BROADER CONTEXT

While our previous results show a promising direction in choosing seeds fairly without losing efficiency, we present a set of extensions that generalize our claims, including the case of small seed set budget and different centrality measures for seed selection.

Diversity on a small budget: influence problems may often be limited to a small budget for seed set size—perhaps only a small number of free products can be given by a company for promotion, or a small number of vaccines are initially available to the general population. In these cases, even if theory predicts that a sufficiently large seed set can achieve both efficiency and fairness, what to do when that seed set size is simply not available?

To answer this question, we extend our analysis to add parity constraints at a new level: at the neighbors of the seed set. We call this variant of ranking nodes *neighbor seeding*. A node's potential to influence is a function of its neighbors' potential. Thus, more balanced neighborhoods will prevent influence from being restricted to one community and may yield wider diffusion. Here, we leverage the intuition given by networks that exhibit homophily: since nodes with the same label cluster together, influence will also be contained within those clusters. While we may apply this reasoning for ensuring parity at every level of the network, it is not computationally feasible to do so, and thus we compare results for agnostic seeding, diversity seeding, and neighbor seeding.

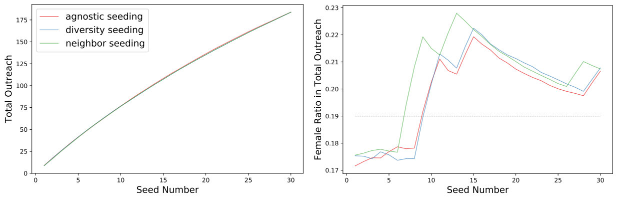

Figure 5: Outreach (left) and female ratio in outreach (right) for the degree heuristics in DBLP, $p = 0.01$.

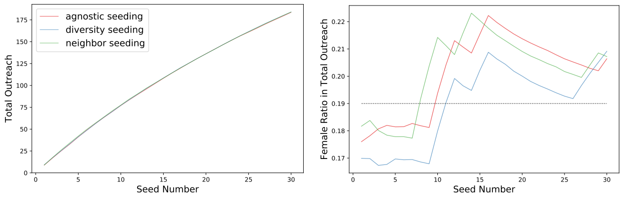

Figure 6: Outreach (left) and female ratio in outreach (right) for the degree discount heuristics in DBLP, $p = 0.01$.

Diversity for different heuristics: We apply these variants to degree and other centrality heuristics, showing that the intuition behind these results is not limited to degree centrality, but it generalizes to other statistical measures of centrality. We thus implement the three seeding heuristics on the DBLP dataset, for a budget of 30 seeds and $p = 0.01$ and $p = 0.1$, comparing their performance in terms of outreach and fairness for different state-of-the-art centrality-based algorithms: degree, degree discount [12], greedy [22], distance centrality [44], and random. The random algorithm, while extremely simplistic, provides a good benchmark for the difference between being completely agnostic to labels and position, and being strategic through our heuristics. We report the results from averaging each cascade $1,000$ times.

In almost all cases, neighbor seeding achieves a better ratio of women in the outreach than diversity seeding (Figures 5–9

Table 2: Female and male outreach for agnostic seeding (AS), diversity seeding (DS), and neighbor seeding (NS) heuristics.

	Degree			Degree discount			Greedy			Distance centrality			Random		
$p = 0.01$	AS	DS	NS	AS	DS	NS	AS	DS	NS	AS	DS	NS	AS	DS	NS
F outreach	37.984	↑**38.174**	↑38.151	37.901	↑**38.484**	↑38.135	42.215	↑42.793	↑**44.251**	6.332	↑**8.528**	↑6.581	8.99	↑18.646	↑**31.433**
M outreach	145.746	↓145.71	↑**145.786**	145.728	↓145.523	↑**145.901**	179.093	↑**180.128**	↓178.67	34.594	↓32.6	↑**35.217**	36.988	↓34.021	↑**122.246**
$p = 0.1$	AS	DS	NS	AS	DS	NS	AS	DS	NS	AS	DS	NS	AS	DS	NS
F outreach	44.494	↓44.481	↑**97.949**	137.856	↑**155.69**	↑148.924	215.441	↑**216.641**	↓207.07	63.498	↑64.967	↑**68.068**	22.414	↑31.503	↑**57.104**
M outreach	170.542	↓170.54	↑**497.613**	694.161	↑**747.224**	↑735.246	997.377	↑1001.032	↑**1006.079**	404.006	↓400.938	↑**417.748**	103.767	↓93.218	↑**229.248**

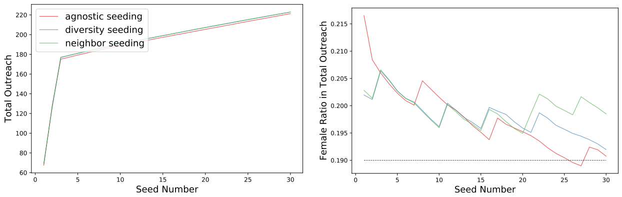

Figure 7: Outreach (left) and female ratio in outreach (right) for the greedy heuristics in DBLP, $p = 0.01$.

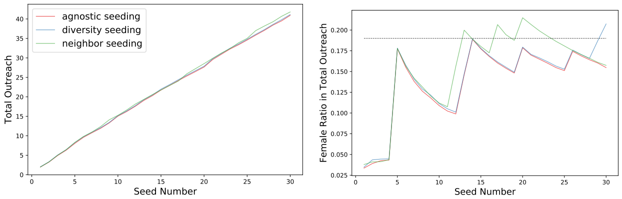

Figure 8: Outreach (left) and female ratio in outreach (right) for the distance centrality heuristics in DBLP, $p = 0.01$.

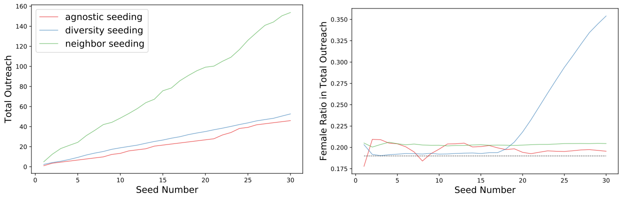

Figure 9: Outreach (left) and female ratio in outreach (right) for the random heuristics in DBLP, $p = 0.01$.

(right)), in some cases considerably surpassing agnostic seeding (Figures 5, 7, 8). The exception occurs for the random algorithm (Figure 9 (left)), for which diversity seeding obtains a better parity ratio in for females in the outreach than neighbor seeding, both of the them surpassing agnostic seeding. However, neighbor seeding is able to obtain a much better influence spread (Figure 9 (right)).

Moreover, neighbor seeding achieves a similar outreach, with a marginal increase, for the other heuristics as well (Figures 5–9 (left)), showing that an increase in outreach diversity does not have to come at a cost of efficiency. Although the outreach increase is again marginal, investigating the male and female outreach gains (or losses) shows the intricacy of our result: as more females are added to the seed set, more are gained in the outreach as well, at a male loss (Table 2 top). What is again impressive is to notice that the female gain in outreach is able to compensate for the male loss, given the large ratio difference between males and females in the network. We obtain similar results for $p = 0.1$, for which we omit the equivalent plots due to space considerations.

Pareto efficiency: Finally, while diversity seeding may reduce the male outreach in some cases, especially for small p (Table 2 top, for $p = 0.01$, diversity seeding slightly decreases the male nodes in the outreach compared to the agnostic heuristic for the degree algorithm, degree discount, distance centrality, and random algorithm), we notice that neighbor seeding increases both the female and male outreach in most cases (Table 2 except for the greedy algorithm). This suggests that the agnostic heuristic is not actually Pareto-efficient for larger values of p (Table 2 bottom). This implies that influencing other nodes is not an all-or-nothing resource and thus exceeds the expectations that female gain in outreach comes at a male loss that the theoretical findings suggest. Intuitively, it may be explained as larger p implies a more than one-hop influence, leading to a better cross-community communication and showing that information diffusion can be done in a win-win equitable way.

6 CONCLUSION

In this paper, we have unraveled the subtle dynamics of network structure in influence maximization, showing that including sensitive features in the input of most natural seed selection algorithms substantially improves diversity but also often leaves efficiency untouched or even provides a small gain. Through a detailed theoretical analysis of biased networks, we show that when the seed set size is sufficiently large, promoting better parity in the seed selection process leads to better parity in the outreach as well since seed set diversity taps into inactivated communities that are hard to reach only from central nodes. On the other hand, when the seed set is too small, fairness comes at a small cost of efficiency. However, as we show again on real-world data, alternative algorithms can extend the benefits of diversity.

Beyond these immediate results, our paper opens future research avenues to analyze more complex algorithms and diffusion processes. Antithetical factors, such as fairness and efficiency, make this a particularly worthwhile area in designing corrections that make the algorithmic output more balanced and justified.

REFERENCES

[1] Junaid Ali, Mahmoudreza Babaei, Abhijnan Chakraborty, Baharan Mirzasoleiman, Krishna P Gummadi, and Adish Singla. 2019. On the Fairness of Time-Critical Influence Maximization in Social Networks. *arXiv preprint arXiv:1905.06618* (2019).

[2] 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 skewed outcomes. *arXiv preprint arXiv:1904.02095* (2019).

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (May 2016).

[4] Chen Avin, Avi Cohen, Pierre Fraigniaud, Zvi Lotker, and David Peleg. 2018. Preferential attachment as a unique equilibrium. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 559–568.

[5] Chen Avin, Barbara Keller, Zvi Lotker, Claire Mathieu, David Peleg, and Yvonne-Anne Pignolet. 2015. Homophily and the glass ceiling effect in social networks. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 41–50.

[6] Eric Balkanski, Nicole Immorlica, and Yaron Singer. 2017. The Importance of Communities for Learning to Influence. In *Advances in Neural Information Processing Systems*. 5864–5873.

[7] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[8] Solon Barocas and Moritz Hardt. 2017. Fairness in Machine Learning Tutorial. *Neural Information Processing Systems* (2017).

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[10] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).

[11] Wei Chen, Chi Wang, and Yajun Wang. 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1029–1038.

[12] Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 199–208.

[13] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[14] David A Cotter, Joan M Hermsen, Seth Ovadia, and Reeve Vanneman. 2001. The glass ceiling effect. *Social forces* 80, 2 (2001), 655–681.

[15] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics* 9, 2 (2017), 1–22.

[16] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[17] Benjamin Fish, Ashkan Bashardoust, Danah Boyd, Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Gaps in Information Access in Social Networks?. In *The World Wide Web Conference*. ACM, 480–490.

[18] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1914–1933.

[19] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv preprint arXiv:1610.02413* (2016).

[20] Catherine Hill, Christianne Corbett, and Andresse St Rose. 2010. *Why so few? Women in science, technology, engineering, and mathematics*. ERIC.

[21] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[22] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 137–146.

[23] David Kempe, Jon Kleinberg, and Éva Tardos. 2005. Influential nodes in a diffusion model for social networks. In *International Colloquium on Automata, Languages, and Programming*. Springer, 1127–1138.

[24] Masahiro Kimura and Kazumi Saito. 2006. Tractable models for information diffusion in social networks. In *European conference on principles of data mining and knowledge discovery*. Springer, 259–271.

[25] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. 22–27.

[26] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[27] Jon Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias.. In *Proc. 9th Conf. on Innovations in Theoretical Computer Science (ITCS)*.

[28] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D Sivakumar, Andrew Tomkins, and Eli Upfal. 2000. Stochastic models for the web graph. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE, 57–65.

[29] Eun Lee, Fariba Karimi, Claudia Wagner, Hang-Hyun Jo, Markus Strohmaier, and Mirta Galesic. 2019. Homophily and minority-group size explain perception biases in social networks. *Nature human behaviour* 3, 10 (2019), 1078–1087.

[30] Michael Ley. 2009. DBLP: some lessons learned. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1493–1500.

[31] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*.

[32] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* 109, 41 (2012), 16474–16479.

[33] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*.

[34] Shirin Nilizadeh, Anne Groggel, Peter Lista, Srijita Das, Yong-Yeol Ahn, Apu Kapadia, and Fabio Rojas. 2016. Twitter's Glass Ceiling: The Effect of Perceived Gender on Online Visibility. In *Tenth International AAAI Conference on Web and Social Media*.

[35] Scott E Page. 2008. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies-New Edition*. Princeton University Press.

[36] Sharon Sassler, Yael Levitte, Jennifer Glass, and Katherine Michelmore. 2011. The missing women in stem? accounting for gender differences in entrance into stem occupations. In *Annual meeting of the Population Association of America Presentation*.

[37] Lior Seeman and Yaron Singer. 2013. Adaptive seeding in social networks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 459–468.

[38] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. 2017. The Problem of Infra-marginality in Outcome Tests for Discrimination. *Annals of Applied Statistics* 11 (2017).

[39] Ana-Andreea Stoica and Augustin Chaintreau. 2019. Fairness in Social Influence Maximization. In *Companion Proceedings of The 2019 World Wide Web Conference*. 569–574.

[40] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 923–932.

[41] Randall Stross. 2008. What has driven women out of computer science. *New York Times* 15 (2008).

[42] Alan Tsang, Bryan Wilder, Eric Rice, Milind Tambe, and Yair Zick. 2019. Group-fairness in influence maximization. *arXiv preprint arXiv:1903.00967* (2019).

[43] Sandra Wachter. 2019. Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. *Available at SSRN* (2019).

[44] Stanley Wasserman, Katherine Faust, et al. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press.

[45] Michele A Whitecraft and Wendy M Williams. 2010. Why aren't more women in computer science. *Making software: What really works, and why we believe it* (2010), 221–238.

[46] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1171–1180.

[47] Honglei Zhuang, Yihan Sun, Jie Tang, Jialin Zhang, and Xiaoming Sun. 2013. Influence maximization in dynamic social networks. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 1313–1318.