bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A signal detection theoretic argument against claims of unmeritocratic faculty hiring

Authors: Michael Miuccio[1], Ka-yuet Liu[2,3], Hakwan Lau[1,4], Megan A. K. Peters[1]

Affiliations:
1 Department of Psychology, University of California Los Angeles
Los Angeles, California, 90095-1563
United States

2 Department of Sociology, University of California Los Angeles
Los Angeles, California, 90095
United States

3 California Center of Population Research, University of California Los Angeles
Los Angeles, California, 90095
United States

4 Brain Research Institute, University of California Los Angeles
Los Angeles, California, 90095
United States

Correspondence should be addressed to:
Megan A. K. Peters
Department of Psychology, University of California Los Angeles
1285 Franz Hall, Box 951563
Los Angeles, California 90095-1563

Tel: (323) 596-1093
Email: meganakpeters@ucla.edu

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

To achieve faculty status, graduating doctoral students have to substantially outperform their peers, given the competitive nature of the academic job market. In an ideal, meritocratic world, factors such as prestige of degree-granting university ought not to play a substantial role. However, it has recently been reported that top-ranked universities produced about 2–6 times more faculty than did universities that were ranked lower (Clauset, Arbesman, & Larremore, 2015). It was therefore argued that the academic faculty job market is not purely meritocratic, because it seems unrealistic that students from top-ranked universities could outperform their peers by as much as six times in productivity. Here we dispute the claim that substantially higher rates of faculty production would require substantially (and unrealistically) higher levels of student productivity; a signal detection theoretic argument shows that a high threshold for hiring (due to keen competition) means that a small difference in average student productivity between universities can result in manifold differences in placement rates. Under this framework, the previously reported results are compatible with a purely meritocratic system. As a proof of concept in the field of Psychology, ranks for universities were gathered from the *U.S News and World Report* (2016) and students' productivity was quantified by the impact factors of the journals in which they published. The results are in agreement with our theoretical model. Whereas these results do not necessarily mean that the actual faculty hiring market is purely meritocratic, they highlight the difficulty in empirically demonstrating that it is not so.

Keywords

Signal detection theory, faculty hiring, impact factor, prestige, university ranking, meritocracy

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A signal detection theoretic argument against claims of unmeritocratic faculty hiring

Introduction

Is academia a pure meritocracy? If it is not, what makes it deviate from the ideal? Previous studies have focused on the roles of nepotism, racism, and sexism, and at the institutional level, hiring network structures and prestige of the programs (e.g., Burris, Hudson, & Brien, 2004; Mai, Liu, & González-Bailón, 2015; Merritt & Reskin, 1997; Wennerås & Wold, 1997). A more recent study reexamined the effects of prestige hierarchy on university faculty hiring networks (Clauset et al., 2015). The authors confirmed previous findings that placement rates are highly skewed: the top-10 ranked universities produce up to six times more faculty than universities ranked 21st–30th. Clauset et al. argued that if the system is a pure meritocracy, this imbalance means that students from the top-10 ranked universities should be up to six times more productive than the third 10. Because this is unrealistic, they concluded that the faculty hiring system is, at least partially, a prestige system.

This paper examines the claim that highly imbalanced faculty production rates necessitate substantially different student productivity levels under a meritocracy. Certainly, within any discipline it is difficult to identify a universally accepted measure of productivity. However, we question the notion that a purely meritocratic system implies that substantial differences in placement rates should always be matched with equally substantial increases in student productivity (however it is measured). Curvilinear relationships between productivity and return are not unique to the faculty hiring markets, and have been referred as the “superstar” (Rosen,

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1981) and "winner-take-all" effects (Frank, Cook, & Rosen, 1996). A multitude of mechanisms can generate such non-linearities. Using faculty hiring as an example, our specific contribution to this broader literature is that we show that a simple mathematical model of binary decisions can explain such nonlinearities under a pure meritocracy. A signal detection theory argument can explain the large discrepancies in faculty production while supposing largely similar rates of productivity between top-ranked and lower-ranked universities.

Signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2004) quantifies how binary decisions are made when there is noise in a system. The decision maker uses a criterion to set a threshold, classifying values above that threshold as belonging to one category (*hits* and *false alarms*) and values that fall below it as belonging to the other category (*misses* and *correct rejections*). Setting a high criterion can result in large differences in what passes the threshold in two similar distributions. Figure 1 shows example distributions representing the height of males (red) and females (blue). For the purpose of explanation, let's suppose that 73 inches qualifies you to be considered a tall person. Our criterion is then placed at 73 inches to determine the proportion of males and females who qualify as being tall. With this high criterion set, it becomes clear that the proportion of females who qualify as being tall is many times less than the proportion of males who qualify, despite the fact that the distributions differ only slightly (the means of the distributions only differ by 5 inches).

We applied this argument to faculty production at ranked universities and demonstrated its validity with a single survey study. We illustrated that the productivity of graduate students is very similar between high- and low-tier universities; however, faculty production *appears*

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

disproportionately skewed toward higher-tier schools because of a high set criterion for faculty hiring. We used the meritocratic measurement of Impact Factor Sum (the total number of impact factors from each publication a graduate student has) to objectively quantify productivity, based on findings from van Dijk and colleagues (van Dijk, Manor, & Carey, 2014). After finding a criterion which could represent achieving a faculty position, we found that, while productivity between schools was similar, the probability of being hired as faculty for students who had graduated from a higher-tier university was nearly twice that of students from the lower-tier schools.

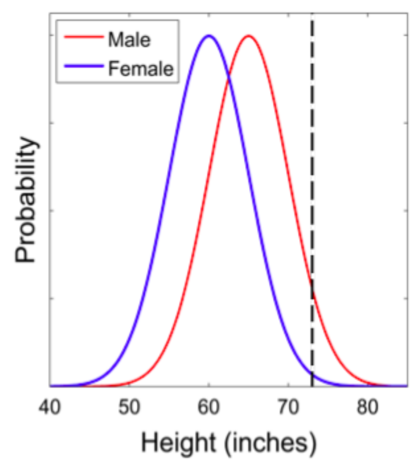

Fig. 1. Illustration of the effect of an extreme criterion on categorization according to Signal Detection Theory. Shown is an arbitrary, imaginary example of height distributions for men and women. The distributions differ in mean by only 5 inches, but if the criterion for labeling an individual as "tall" is set to an extreme value, such as 73 inches (6 ft 1 inch), the proportion of males categorized as "tall" will be *many* times more than the proportion of females categorized as "tall."

Methods

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In the present investigation, we examined the productivity and faculty hiring rates of Psychology students at various universities on the *U.S. News & World Report* (2016) ranking list. It is worth noting that faculty hiring networks created by Clauset et. al (2015) consisted of Business, History, and Computer Science disciplines. These disciplines were chosen because the authors deemed them to have little 'leakage' between the department an individual graduates from and the department where he or she is ultimately hired as faculty. However, there is also little leakage in Psychology, which has the additional benefit of an easily-defined objective meritocratic measure based on the impact factors of peer-reviewed publications for each individual: in contrast to excelling in the fields of Business, History, and Computer Science, which typically relies heavily on publishing books, presenting at conventions, and developing software, excelling in Psychology is more focused on publishing papers in peer-reviewed journals that are assigned these objective impact factor metrics.

For our samples of individuals, we drew from institutions listed on the National Universities Rankings from the *U.S. World and News Report* (2016). From these universities we first collected three samples of individuals in their Psychology departments depending on university rank — rank 1–10, rank 11–20, and rank 21–100 — through a combination of online search of student directories and direct contact with departments. We defined the metric of Impact Factor Sum (IFS) for each individual in this sample as the sum of the impact factors for every publication that individual had authored regardless of authorship order, as indexed in Google Scholar. This method was used as a means to equitably search all students' publications because not all students post their CVs, and data downloaded from large archives (e.g., PubMed) would

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

by definition exclude individuals with no publications and journals not indexed by that engine. We also wanted to quantify productivity for all current students, not graduates. Although other more complete and complex indices are available (e.g., h-index [Hirsch, 2005, 2007], or predictions based on machine learning techniques [Acuna, Allesina, & Kording, 2012]), we elected to use this IFS metric due to its simplicity and close relationship with more complex metrics. Specifically, it has been shown that the perceived quality (i.e., impact factor) of a publication is given more weight in the faculty hiring process than its actual quality (i.e., its citation rate), and that the two most important factors in predicting faculty hiring are impact factor of publications and number of publications (van Dijk et al., 2014); we therefore combined these factors into a single IFS.

We tested for differences among the three samples (rank 1–10, 11–20, and 21–100) with (a) Wilcoxon Rank-Sum (Mann-Whitney U) tests, which are nonparametric tests that do not rely on assumptions of normality (Wilcoxon, 1945), and (b) Kolmogorov-Smirnov nonparametric tests, which test for differences between probability distributions and are sensitive to both mean and distribution shape (Chakravarti, Laha, & Roy, 1967). Because these tests revealed no differences between universities with rank 11–20 and 21–100 in IFS distribution (see Results), we combined data from the lower-tiered universities into a single Lower tier group. We evaluated the similarity between the remaining Higher tier (rank 1–10) and Lower tier (rank below 10) groups using ROC analysis, which plots the *hit rate* versus *false alarm rate* at varying criterion values (Green & Swets, 1966; Macmillan & Creelman, 2004). We used the area under this ROC curve (AUC) as a measure of the similarity between the Higher and Lower tier IFS scores, as it

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

provides a normalized metric of separability of distributions: AUC = 0.5 indicates distributions are identical, and AUC = 1 indicates distributions are completely separable.

Criterion setting

To define the criterion in IFS space for being hired as faculty regardless of graduate university, we estimated the probability of being hired as faculty across a large number of universities. We collected a fourth sample from the Psychology departments across all *U.S. News and World Report* ranks from 1–50, to compare the average number of graduates from each per year to the average number of faculty hires made at those same universities per year. Thus, we can define

$$p(hire) = \frac{\#\ faculty\ hires\ made}{\#\ graduates} \quad (1)$$

To calculate a criterion within the overall IFS distribution, we collected all IFS for all individuals in the Higher and Lower tier groups regardless of university tier. We then used bootstrapping to ensure that our results were not overly sensitive to our particular sample. On each bootstrap loop, a random sample of 1000 IFS data points (with replacement) was drawn from this overall IFS distribution and binned into 30 bins of width 10.52. To this histogram we fitted an exponential function of the form $f(x) = ax^b + cx^d$, which we then normalized so that it would constitute a probability density function over the range of IFS in our sample. We then defined the criterion c, or IFS Cutoff, as the IFS above which the area under the curve (AUC) matched this probability, i.e.

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$p(hire) = H(c) = 1 - F(c, x) = 1 - \int_0^c f(x)dx = \int_c^\infty f(x)dx \quad (2)$$

leading to

$$c = H^{-1}(p(hire)) \quad (3)$$

where the integral initially is taken from 0 to c because an exponential function is undefined at $x < 0$. This process was repeated 1000 times for a total of 10,000,000 samples, leading to 1000 estimates of c.

Probability of being hired conditioned on university tier

To evaluate how the criterion might lead to potentially exaggerated differences in $p(hire)$ if we conditioned on university tier, i.e. $p(hire|tier)$, we again used bootstrapping to ensure that our results were not overly sensitive to our particular sample. As before, on each bootstrap loop, a random sample of 1000 IFS data points (with replacement) was drawn from each of the Higher and Lower tiers, respectively. These were then binned as before (30 bins of width 10.52) and fitted with respective exponential functions of the form $f(x|tier) = ax^b + cx^d$. Following normalization such that the fitted functions constituted probability density functions, we evaluated the AUC above the criterion c on each bootstrap loop for each tier, i.e.

$$p(hire|tier) = H(c|tier) = \int_c^\infty f(x|tier)dx \quad (4)$$

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

This process was repeated 1000 times for a total of 20,000,000 samples (10,000,000 from each of the Higher and Lower tiers), leading to 1000 estimates of $p(hire|Higher\ tier)$ and 1000 estimates of $p(hire|Lower\ tier)$.

Results

We collected Impact Factor Sum (IFS) data on 1871 individuals from 27 institutions. IFS for each individual in our entire sample ranged from 0 (no publications) to 304.637 (many first-, middle-, and last-authorship papers in high-impact journals) ($\mu = 7.619$, $\sigma = 19.480$). These were collected across three groups corresponding to the tier of the university from which an individual had received his or her doctorate: rank 1–10 (n = 607, range: 0–304.637, $\mu = 11.655$, $\sigma = 26.929$), rank 11–20 (n = 532, range: 0-96.17, $\mu = 5.751$, $\sigma = 13.332$), and rank 21-100 (n = 732, range: 0-150.87, $\mu = 5.629$, $\sigma = 14.846$).

Because these samples are not normally distributed, we used nonparametric tests to compare them. Wilcoxon Rank-Sum (Mann-Whitney U) tests revealed significant differences in IFS between universities with rank 1–10 and those with rank 11–20 and 21–100, but no differences between universities with rank 11–20 and 21–100 (Table 1, top three rows). Kolmorogov-Smirnov tests revealed an identical pattern (Table 1, top three rows).

				Wilcoxon Rank Sum / Mann-Whitney U		**Kolmorov-Smirnov**	
	Rank pairing	**n_1**	**n_2**	**U**	**p**	**D**	**p**

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Individual groups	1–10 vs. 11–20	607	532	3.641e5	< .001*	0.134	< .001*
	1–10 vs. 21–100	607	732	4.368e5	< .001*	0.156	< .001*
	11–20 vs. 21–100	532	732	4.575e5	.319	0.054	.318
Merged groups	Higher (1–10) vs. Lower (below 10)	607	1264	6.163e5	< .001*	0.146	< .001*

Table 1. Results of nonparametric comparisons of all university tier groups. All individual groups' pairwise comparisons are significant (denoted with *) except for rank 11–20 vs. 21–100 (top three rows). We therefore collapsed across the two similar groups to create a single pair of groups (bottom row). This pair of groups is used in all further analyses.

We therefore collapsed across the two lower-ranking groups to create a Higher tier sample (n = 607, rank 1–10) and a Lower tier sample (n = 1264, rank below 10). The range of IFS scores for the Higher tier was therefore 0–304.637 ($\mu = 11.655$, $\sigma = 26.929$), and for the Lower tier was 0–150.87 ($\mu = 5.680$, $\sigma = 14.223$). These tiers are significantly different from each other (Table 1, bottom row). However, despite significant differences between the Higher and Lower tiers at the statistical level, the samples appear visually similar (Figure 2a) and ROC analysis revealed AUC = 0.565 (Figure 2b). As completely overlapping distributions would have AUC = 0.5, this result means that the distributions are exceedingly similar to one another despite statistical differences. All IFS data are available in Supplementary Table 1 (available online).

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

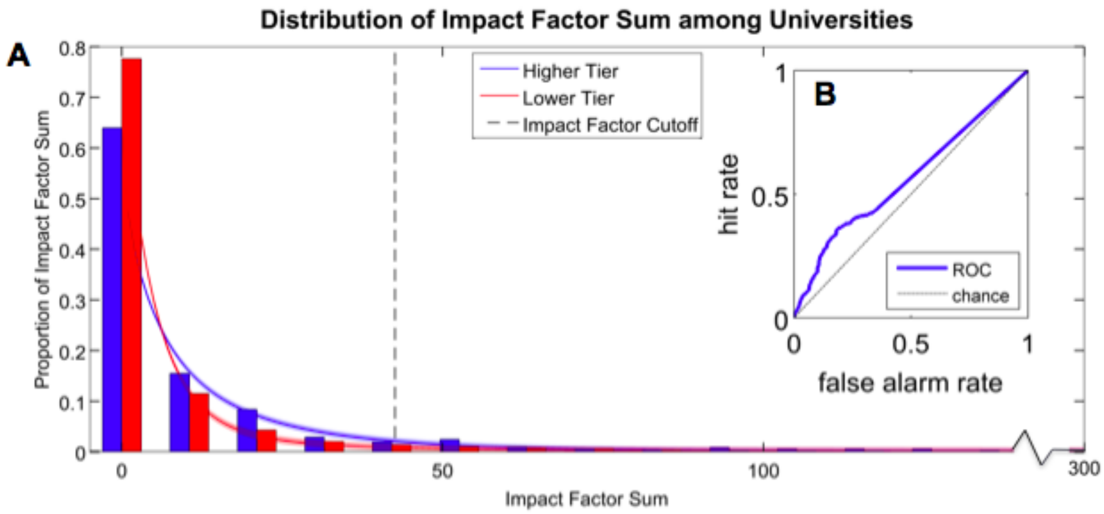

Fig. 2. Distributions of Impact Factor Sum (IFS) across university tier are very similar. Panel (a) shows that productivity (graduate students' IFSs) of the different university tiers (Higher vs. Lower) is very similar while the criterion for getting a faculty position is, as expected, extremely high. The difference between the two curves is minimal, as shown by the ROC curve in (b): the area under the curve (AUC), representing discriminability between Higher and Lower tier universities, is 0.565. This is almost at chance (0.50), showing that the distributions are nearly equivalent. The IFS criterion to be hired as faculty (i.e., IFS Cutoff) was placed at 42.57 in accordance with the reality of faculty production (see Methods for details on criterion setting). We fitted an exponential function to the overall distribution of IFS, which represents productivity of all graduate students regardless of university (see Methods). The percentage of the area under this curve that falls above the criterion, *i.e.,* the probability of a graduate student getting a faculty position after graduating from *any* university, is about 5%.

Criterion setting

In our separate sample of graduates and hires from Psychology departments at 22 institutions, we found that an average of 701.2 students graduated per year, and an average of 35.5 individuals were hired as faculty at those same institutions per year. By Equation 1, this leads to $p(hire)$= 0.0506, meaning that approximately 5% of all graduates with doctorates in Psychology are hired as faculty in any given year. This result is in line with previous reports of faculty hiring rates of

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

about 6.2% (van Dijk et al., 2014). We then used the bootstrapping process described in Methods to calculate 1000 estimates of c, the criterion or IFS Cutoff, corresponding to this $p(hire)$. We found mean c = 42.57 (median = 45.00, σ = 14.012), meaning that any given individual should aim to have a total IFS equal to or exceeding 42.57 if he or she hopes to be hired as a faculty member.

Probability of being hired conditioned on university tier

Despite the similarity between the Lower and Higher tier distributions of IFS, closer inspection of the tail ends of the distributions, above the criterion c, reveals important differences. Figure 3a displays the mean of $f(x|tier)$ for both tiers over all loops of the bootstrapping analysis zoomed in on the region of the IFS Cutoff criterion c, with SEM across bootstrap loops represented by the shaded regions. The Higher tier IFS scores display a strikingly large advantage over the Lower tier IFS scores at the location of the criterion.

This advantage for Higher tier universities is further clarified when we calculate $p(hire|tier)$ for the Lower and Higher tier universities using Equation 4. Figure 3b displays the mean and standard error of $p(hire|tier)$ for the Higher and Lower tier universities, respectively, across all loops of the bootstrapping analysis. From these results we see that graduates of Higher tier universities are significantly more likely to be hired as faculty ($t(999) = 23.06$, $p < .001$), and by a factor of two: simply put, you are twice as likely to be hired as faculty if you receive your doctorate in Psychology from a top-10 university than if you attended any university of lower rank, *based purely on the meritocratic metric of IFS*. This occurs despite the extreme similarity

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in the IFS distributions for Higher and Lower tier universities (AUC = 0.565), as a result of small but significant differences in these distributions.

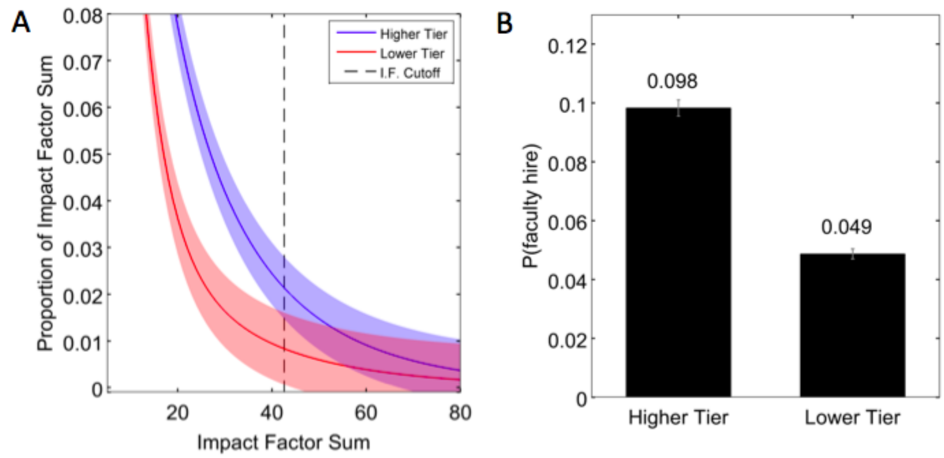

Fig. 3. Zoomed-in view of IFS distributions and probability of faculty hire conditioned on university tier. Panel (a) shows the portion of the graph from Figure 2a nearest the criterion. At this zoom level it is clear that despite their overall similarity, the distributions are quite different at the extreme value of the IFS criterion. Shaded regions indicate the standard error of the mean (SEM) across bootstrapping loops (see Methods) at each IFS value. Panel (b) shows the probability of being hired as faculty conditioned on having graduated from a Lower or Higher tier university, or mean $p(hire|tier)$ across bootstrapped samples (see Methods). The mean probability of being hired after graduating from a Lower tier university is 4.87%, or approximately half the mean probability of being hired as faculty after graduating from a Higher tier university (9.83%). Error bars represent the SEM across all bootstrapping loops.

Asymmetry in hiring rates is a direct consequence of an extreme criterion

We next examined the degree to which the large asymmetry in hiring rates, which has been blamed on a non-meritocratic system, is dependent on the extreme criterion that the current hiring climate has set (regardless of its exact numerical value along the dimension(s) used to quantify productivity). The criterion at IFS ≈ 42 reflects the externally valid hiring rate of ~5–

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

6% (van Dijk et al., 2014), and leads to a hiring asymmetry of a factor of two between Higher and Lower tier universities. However, if the hiring climate were less competitive, with a less extreme criterion along any meritocratic dimension, this asymmetry would dwindle with increasing values for $p(hire)$ and eventually disappear (Figure 4). We demonstrated the decrease in the asymmetry of the $p(hire)$ by using bootstrapping as before, drawing 1000 random samples of 1000 IFS scores from each of the Higher and Lower tier distributions. The criterion was then calculated as above (Equation 3) ranging from 0.05 to 0.85 in intervals of 0.05, and the ratio of $p(hire|Higher\ tier)$ to $p(hire|Lower\ tier)$ was calculated via Equation 4 at each criterion level. The mean of the all $p(hire)$ ratios is shown in Figure 4.

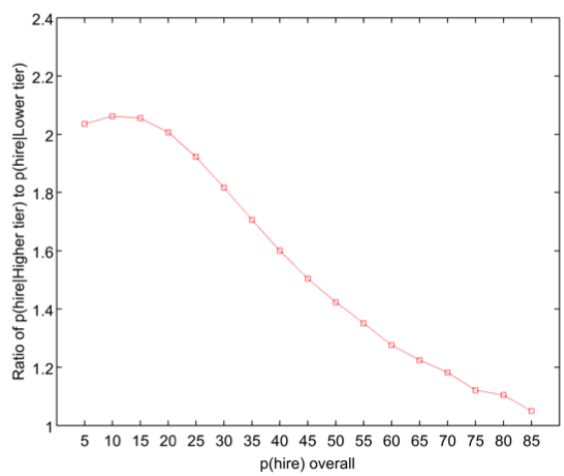

Fig. 4. Less extreme hiring criterion values lead to less pronounced hiring asymmetry between university tiers. By shifting the criterion to more liberal values — from $p(hire)$ = 0.05 to $p(hire)$ = 0.50 — we show that the hiring asymmetry is reduced and ultimately disappears entirely. The appearance of a non-meritocratic system is thus perpetuated by the severe competitiveness of the current hiring climate.

So the appearance of a non-meritocratic system is in fact perpetuated in large part by the extreme difficulty of attaining a faculty position.

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Effect of focusing on more advanced students

We repeated the above bootstrapping analyses excluding individuals with IFS = 0 from both Higher tier and Lower tier samples as before ($n_{Higher} = 335$, $n_{Lower} = 801$), which can be thought of as selecting primarily the more advanced students while removing first- and second-year students who have not published yet. Analyses on this subsample of individuals shows no change in overall findings: the observed similarity between Higher and Lower tier distributions is maintained, albeit a little lower (ROC = 0.6542); a high IFS criterion is set (IFS = 55.16); and resultant asymmetry between Higher and Lower tier faculty hiring rates is similar (mean $p(hire|Higher\ tier)$= 14.61, mean $p(hire|Lower\ tier)$= 2.70). See Supplemental Material for further discussion.

Discussion

Using a relatively objective measurement of productivity (Impact Factor Sum), we have shown that faculty production is highly skewed towards Higher tier universities while productivity between tiers is practically indistinguishable. The high criterion, put in place by an extremely competitive faculty hiring system, is reflective of actual faculty hiring rates (5–6%). While it has been shown that doctoral prestige better predicts faculty placement than productivity (Burris et al., 2004; Miller, Glick, & Cardinal, 2005 but see van Dijk et al., 2014), this high criterion casts doubt on the claim that the differences observed by Clauset et. al (2015) must imply vast (and unrealistic) differences in productivity or a strongly non-meritocratic system. Our model also

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

provides an alternative explanation to why controlling for publication records usually has only modest effects on the prestige-placement relationship (e.g, Headworth & Freese, 2015), without having to resort to unmeasured heterogeneities in the positions or in the candidates.

While the median for the Higher tier and Lower tier universities both equal 0, the Wilcoxon Rank-Sum test between Higher tier and Lower tier revealed that the distributions do differ to a statistically significant degree, ROC analysis showed that the normalized magnitude of this difference was indeed quite slight, at 0.565 — meaning that if a random person is picked from either cohort and you are asked to guess the cohort to which the person belong based on the IFS score, your likelihood of being correct would be just 6.5% above chance (chance = 50%). This similarity between the two cohorts is also reflected by the median score for both distributions: both Higher and Lower tier groups have a median IFS of 0. So the distributions' differences are highlighted primarily in the extreme end of their upper tails.

One possible limitation of our study is the means by which we quantified productivity. We reasoned that since not all students upload their CVs to a university or other website, using CV information would result in a sample biased by those students who add a CV or publication list to their searchable profiles. Instead, we elected to rely on Google Scholar, because (a) individuals do not have control over what appears in the search engine (unlike, for example, the appearance of an individual in NeuroTree) and (b) we wanted a metric by which to quantify all students' productivity as a function of university tier *including those students who had published nothing at all*. By searching for individual students' names we calculated their IFS, or the sums of all impact factor scores for all papers they had published, if any. This metric was favored over

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

h-index (Acuna et al., 2012; Hirsch, 2005, 2007) because of its simplicity and similarity to previously validated approaches. Additionally, h-index is based off of the number of citations an individual's research obtains. The h-index will be much lower for graduate students due to the recency of their publications. Total impact factor score has been shown to be more predictive of fellowship application success than measures based only on first-author publications or number of citations (Wennerås & Wold, 1997). Similarly, it has been demonstrated that perceived impact of publications (impact factor) and number of publications (regardless of authorship order) are most predictive of faculty hiring (van Dijk et al., 2014); these top predictors are combined for our dataset into the single IFS score for each individual.

Yet despite this support for IFS as an objective measure of productivity, it could still be argued that a hiring criterion of IFS ≈ 42 is unrealistic, and reflects a problem with our chosen productivity metric: does a student really need to have some large number of relatively strong publications in order to be hired as faculty? Of course not; there are many other potential measures that can affect hiring decisions (meritocratic or not), and it is the constellation of scores along these dimensions that will ultimately result in the decision of whether to hire a candidate. Because of this, and because the criterion is of course a soft boundary, candidates with IFS < 42 will certainly end up in faculty positions. That is, it is not impossible for a candidate with IFS < 42 to achieve faculty status. That the exact value of the criterion along the dimension we have chosen (IFS) seems unreasonable should not suggest that our approach is flawed, however. It simply suggests that the dimension we have chosen may not be the best or only factor considered in the hiring process. Indeed, it should be possible to evaluate the

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reasonableness of any meritocratic metric (e.g., h-index) by how it ends up adhering to reality: if it is too far off, other metrics — or a combination of metrics — may be more appropriate.

Despite these limitations, it should be recognized that we simplified our analysis by assuming graduates seeking faculty positions who are not hired do not cumulatively add to the faculty applicant pool from year to year. If we had not assumed this, the remaining 95% of graduate students who are not hired would be added to the following years' new applicants, which would result in an even more extreme criterion. It will lead to even more inequality in job placement if the differences in IFS between Higher- and Lower-tier programs remain the same, or becomes larger, a few years after graduation. We expect this cumulative advantage (Bedeian & Hunt, 2010; DiPrete & Eirich, 2006; Headworth & Freese, 2015) will continue to maintain or even exacerbate the differences in IFS between Higher- and Lower-tier programs during the post-doc years. Even if the difference in IFS across programs becomes smaller over time, our results demonstrate that an extremely high criterion will convert any small differences in IFS to large effects on the odds of being hired.

Finally, while the method we used to collect data and the selection of IFS as a measure of productivity may be controversial, the purpose of our study was a proof of concept; the results of the signal detection analysis presented here do not depend on these choices. Our argument holds in any distribution, and so the sample we gathered serves primarily to justify our theoretical argument.

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In sum, we have shown that similarly productive cohorts will produce very different rates of faculty hires simply because of a high hiring threshold. We should note that our results cannot speak to whether or not the current system is a in fact pure meritocracy (see e.g., van Dijk et al., 2014 for discussion of the impact of university rank and gender on faculty hiring). However, our results do make clear that the argument made by Clauset et. al (2015) — that a substantial discrepancy in hiring rates as a function of degree-granting university tier *must* be due to factors beyond the meritocratic — does not hold.

Acknowledgements

This research is partially supported by the National Institutes of Health (NINDS R01 NS088628-01 to HL). We also thank Nikolaus Kriegeskorte for seminal discussion on internet social media which inspired the study.

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature*, *489*(7415), 201–202. http://doi.org/10.1038/489201a

Bedeian, A. G., & Hunt, J. G. (2010). Doctoral Degree Prestige and the Academic Marketplace : A Study of Career Mobility Within the Management Discipline. *Academy of Management & Learning Education*, *9*(1), 11–25.

Burris, V., Hudson, K., & Brien, R. O. (2004). The academic caste system: Prestige hierachies in Phd exchange networks. *American Sociological Review*, *69.2*, 239–264.

Chakravarti, I. M., Laha, R. G., & Roy, J. (1967). *Handbook of methods of applied statistics, volume 2*. John Wiley & Sons.

Clauset, A., Arbesman, S., & Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, *1*(1), e1400005–e1400005. http://doi.org/10.1126/sciadv.1400005

DiPrete, T. a., & Eirich, G. M. (2006). Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments. *Annual Review of Sociology*, *32*(1), 271–297. http://doi.org/10.1146/annurev.soc.32.061604.123127

Frank, R. H., Cook, P. J., & Rosen, S. (1996). The winner-take-all society. *Journal of Economic Literature*, *34*(1), 133–134.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. *Society* (Vol. 1). http://doi.org/10.1901/jeab.1969.12-475

Headworth, S., & Freese, J. (2015). Credential Privilege or Cumulative Advantage? Presitige, Productivity, and Placement in the Academic Sociology Job Market. *Social Forces*.

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

http://doi.org/10.1093/sf/sov102

Hirsch, J. E. (2005). An index to quantify an individual's scientifc research output, *102*(46), 16569–16572.

Hirsch, J. E. (2007). Does the H index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America*, *104*(49), 19193–8. http://doi.org/10.1073/pnas.0707962104

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Mai, B., Liu, J., & González-Bailón, S. (2015). Network Effects in the Academic Market: Mechanisms for Hiring and Placing PhDs in Communication (2007-2014). *Journal of Communication*, *65*(3), 558–583. http://doi.org/10.1111/jcom.12158

Merritt, D. J., & Reskin, B. F. (1997). Sex, race, and credentials: The truth about affirmative action in law faculty hiring. *Columbia Law Review*, *97*(2), 199–311.

Miller, C. C., Glick, W. H., & Cardinal, L. B. (2005). The Allocation of Prestigious Positions in Organizational Science : Accumulative Advantage, Sponsored Mobility, and Contest Mobility. *Journal of Organizational Behavior*, *26*(5), 489–516. http://doi.org/10.1002/job.325

Rosen, S. (1981). The economics of superstars. *The American Economic Review*, *71*(5), 845–858.

van Dijk, D., Manor, O., & Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology*, *24*(11), R516–R517. http://doi.org/10.1016/j.cub.2014.04.039

Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature*, *389*(6649), 326. http://doi.org/10.1038/387341a0

bioRxiv preprint doi: https://doi.org/10.1101/061200; this version posted June 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83.