The glass ceiling in NLP

Natalie Schluter
Department of Computer Science
IT University
Copenhagen, Denmark
natschluter@itu.dk

Abstract

In this paper, we provide empirical evidence based on a rigourously studied mathematical model for bi-populated networks, that a glass ceiling within the field of NLP has developed since the mid 2000s.

1 Introduction

The *glass ceiling* is a powerful metaphor for the unethical, invisible, and yet virtually impenetrable barrier that prevents highly achieving women and minorities from obtaining equal access to senior career opportunities. The existence of a glass ceiling is well documented both in STEM[1] and specifically in Computer Science (Moss-Racusin et al., 2012; Shen, 2013; Larivière et al., 2013; Van der Lee and Ellemers, 2015; Way et al., 2016, for example). To date there has been no published study on this topic for the field of NLP.

In most countries, Computer Science has long been struggling to support female researchers sufficiently: female representation in Computer Science is not only disproportional to the population, but it is lower than the average STEM field. Moreover, as opposed to STEM fields in general, the proportion of women in Computer Science has been on a marked decline for the past two decades (Sax et al., 2017; Williams et al., 2017), placing the entire the tech field in a diversity crisis today.

The discussion of gender representation or even the existence of a glass ceiling is rather more complex for NLP due to its fundamental interdisciplinarity especially across the fields of Linguistics, Computer Science, and Statistics. That is, much mainstream research in NLP follows trends that are heavily situated in one of the main subdisciplines. Can we witness any emergent glass ceiling for female researchers in the wake of an increasing concentration on deep learning engineering techniques applied to NLP problems? What about the preceding Machine Learning wave from the mid 2000s? In this paper we answer this question in the affirmative.

We acquired a gender-annotated co-author dataset covering arguably the most central ACL publication venues for the past 52 years. We carry out basic data analysis over this dataset and the bi-populated (female and male researcher) mentor-mentee network derived from it. We make the following concerning empirical observations:

1. **There is a growing mentor gender gap.** There is a growing disparity between the proportions of female and male NLP researchers who achieve mentor status, with a higher proportion of male researchers becoming mentors, especially since the mid 2000s.
2. **There is a significant time gap to mentor status across genders.** Female NLP researchers must wait a considerable time longer to achieve mentor status than their male colleagues.
3. **In-gender mentorship correlates with future success.** Female NLP researchers who take a male supervisor will have greater difficulty in becoming a mentor than if they take a female supervisor, on average.
4. **Homophily is on the rise.** There is consistently increasing homophily in our field– the preference to establish in-gender mentor-mentee relationships.

Following this analysis, we employ Avin et al. (2015)'s rigorously studied conditions for power inequality and the glass-ceiling effect for complex systems data structured like ours to show that these empirical observations indicate quite precisely **the existence of a glass ceiling effect for the field of NLP**.

[1] Science, technology, engineering and mathematics fields.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2793–2798
Brussels, Belgium, October 31 - November 4, 2018.

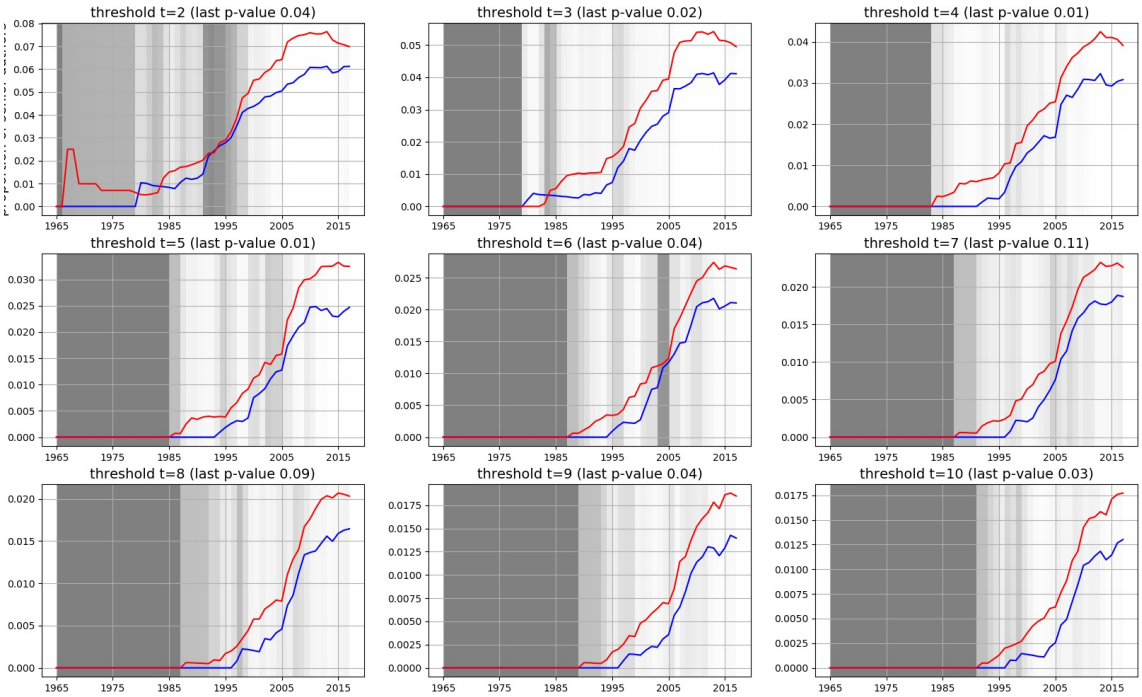

Figure 1: Proportion of male (red line) and female (blue line) mentors from 1966 to 2017, for different thresholds of "mentor seniority". A disparity in these proportions has been increasing since the 1990s, which follows the general field of Computer Science. The whiter the background the more significant the difference in proportions for the corresponding year, with the p-value for 2017 in the title.

2 Acquiring a gender-annotated mentor-mentee network

We scraped all meta-information available from the ACL Anthology[2] for arguably the most central publication venues in NLP. This includes all papers from CoNLL, EACL, TACL, CL, ACL, EMNLP, COLING, ANLP, NAACL, *Sem/SemEval from 1965 to 2017: 19,552 papers in total.

We carried out some normalisation of the author names scraped by lower-casing, normalising for order (first name then last name), removing middle initials and title abbreviations, and removing accents and punctuation, collapsing the extracted list of 18,437 author names to a list of 17,232 author names. Following this, we applied several gendered first-name lists to automatically annotate a large portion of the author names with gender.[3] This resulted in 13,435 automatically annotated author names. Of the remaining 3797 unannotated names, we automatically label as 'unknown' all author names with only an initial standing for the first name, effectively filtering out a further 565 author names. The remaining 3232 author names were annotated by the current authors by manually inspecting the results of Google Image queries for the full name.

The resulting dataset spans 52 years and includes 17,232 authors, of which we labeled 10,382 as male, 5,227 as female and 1,623 whose gender we could not identify. In what remains of our study, we discard these latter authors. This leaves a total of 15,609 researchers.

Power in academia. In our study, we need to account for mentor status–a type of seniority and power. As in many other fields, in NLP it is customary for mentors to take the last-authorship position of papers. Though there can be exceptions to this custom, the assumption of mentor last-authorship is simple, and with this large dataset, we believe it provides a robust approximation of mentorship in the absence of other more precise indications like centralised supervision logs. This method was also adopted by Avin et al. (2015).

We use the assumption of mentor last-authorship to provide an empirical definition of a mentor in our dataset. We say that after t last-authored papers for some threshold $t \in \{1, \ldots, 10\}$, and excluding all sole-author papers,

[2] http://www.aclweb.org/anthology

[3] The lists are discussed in the appendix.

a researcher is considered to hold *mentor standing with seniority threshold* t.

We model the interactions between researchers by creating the bi-populated (for female and male populations) mentor-mentee network. The network's nodes therefore are researchers and there is an edge between two co-authors of a paper in our dataset if and only if one of the co-authors is the last author. This leaves a mentor-mentee network with 14248 nodes, 25211 edges, and average degree 3.539. This network allows us to observe whether the current system of mentor-mentee relationships entails a glass ceiling effect in the modeled community. We now present the results of this analysis.

3 Evidence of a rising gender gap

We provide some basic empirical evidence which could be indicative of the presence of a glass ceiling effect in NLP. In Section 4, we then prove that there is indeed a glass ceiling.

3.1 Growing mentor standing disparity

A researcher who has achieved some seniority is generally eligible to become a mentor and supervise students. As such, the rise to becoming a mentor is a measurable criterion of success for a researcher in academia. Concretely, in some countries, the mentor role is reserved for permanent/tenured faculty (for example, in Denmark). Therefore a barrier to mentor standing for females can lead to an important under-representation of women. This under-representation in turn may perpetuate itself through the lower availability of same-gender advisors for female students, which we show to be of central importance for rising NLP researchers (in Section 3.3).

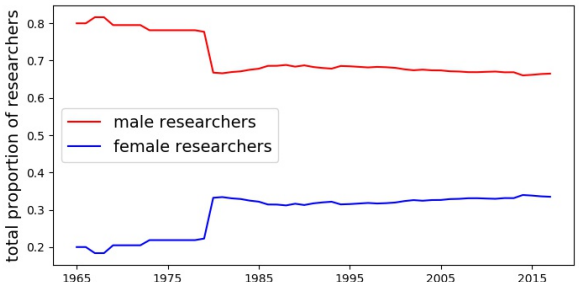

Figure 2: The proportion of female researchers (blue line) in NLP has been gradually increasing since 1965, but seems to be leveling off at around 33.5%, with a very slight decrease since 2014. A corresponding decrease is observed for male NLP researchers (red line).

For thresholds of "mentor seniority" $t \in \{2, \ldots, 10\}$ we examined the proportion of mentors with respect to the pool of researchers of the same gender over time from 1966 to today. Figure 1 shows the resulting time series. Across all thresholds, we observe that the proportion of male supervisors with respect to the total number of male researchers is increasing faster than the proportion of female supervisors within the general pool of female researchers. In fact, the discrepancy between these two proportions seems to slowly close until the early-to-mid 2000s after which it steadily increases again. And in almost all cases this difference in proportions develops into a statistically significant difference (with a 1-sided z-test for proportions, and p-value 0.05). This is despite there being no corresponding development in mentor-mentee proportions as shown in Figure 2.

3.2 Time to seniority gap

We further investigate the subset of female researchers who achieved mentor standing, and compare their difficulty in doing so with that of the respective pool of male researchers. One measurable factor from our dataset is time. Isolating a substantially larger delay to achieving mentor standing for female researchers is one way to use our dataset to measure the difficulty in transitioning female researchers from mentee to mentor standing. We consider the average time it takes to achieve mentor standing between the two populations. For consecutive periods of two years, we compute the average number of years for researchers to achieve mentor standing at threshold $t \in \{2, \ldots, 10\}$. We provide a visualisation of the results in Figure 3. We do a two sample t-test to expose the statistical significance in the non-equality of the respective means. We observe that across all thresholds for mentor standing, female researchers are substantially more delayed than male researchers in becoming mentors. For the most recent numbers, the result is most significant where there is the most data, at seniority $t = 3$, with p-level 0.04. However we note the general whitening of the plots (indicating statistical significance) after the mid-2000s.

3.3 The effects of in-gender supervision

The availability of female mentors been has shown to correlate with mentees' future success–in particular, females in Chemistry who are mentored by

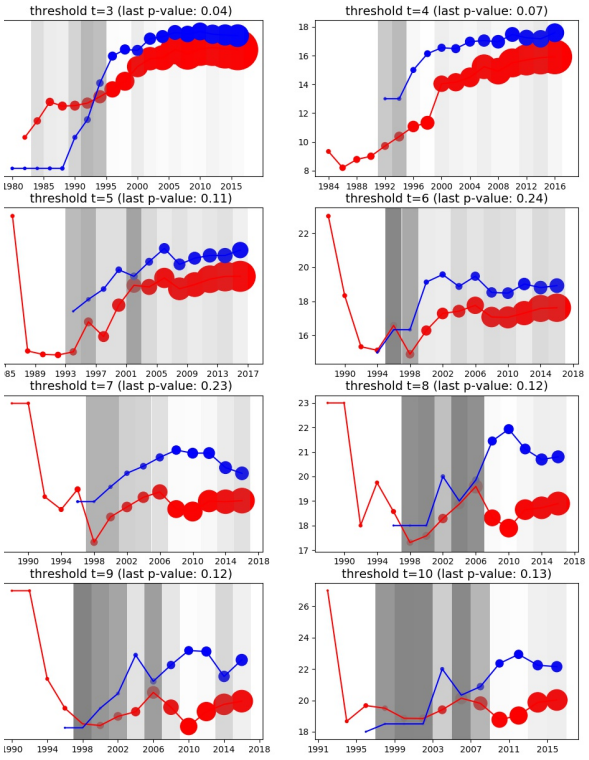

Figure 3: The average time it takes in years, across the past 52 years, for female researchers (blue line) and male researchers (red line) to achieve mentor status in the NLP field. The more significant the difference in means, the whiter the background. The relative number of data points for each year and gender are indicated by the size of the scatter points (blue for females and red for males). The last recorded p-level for each seniority threshold is provided in the corresponding plot's title.

			mentor			
	male	female	male	female	male	female
mentee	$t=2$		$t=3$		$t=4$	
male	12.26	5.69	6.98	4.21	4.95	3.29
female	7.29	10.73	4.96	6.12	3.69	4.11
	$t=5$		$t=6$		$t=7$	
male	3.91	2.72	3.25	2.4	2.64	1.99
female	2.83	3.08	2.31	2.47	2.01	2.1
	$t=8$		$t=9$		$t=10$	
male	2.26	1.74	2.03	1.58	1.85	1.46
female	1.8	1.87	1.61	1.65	1.38	1.4

Table 1: Probabilities (as % here) that a mentee of the row gender, supervised by a mentor of the column gender will achieve mentor standing, for various thresholds t.

female supervisors are considerably more likely to become faculty themselves (Gaule and Piacentini, 2018). In Table 1 we observe a similar trend for in-gender mentorship. In particular, female researchers who have female mentors are much more likely to become mentors themselves. This is a particular problem if, as Sections 3.1 and 3.2 show, the proportion of female NLP mentors is not increasing at the same rate as that of male NLP researchers, possibly due in part to the added delay in achieving mentor status for women. Indeed this delay in access, perpetuated due to the lack of in-gender supervision, can be the result of a glass ceiling in NLP. In the next section we investigate the likelihood of such a glass ceiling.

4 The glass ceiling effect in NLP

In order to understand better how the population of female researchers in NLP can be increasing, but the growth level of seniority/mentor standing still falls significantly below that of the male population and that this gap is widening, we turn to an investigation of power inequality and the glass ceiling effect.

First three key observations can be made of the mentor-mentee network introduced in Section 2 vis-à-vis three well-accepted mechanisms of observed human behavior.

(O1) **Minority-majority partition.** Figure 2 shows the resulting proportion of male and female researchers in NLP through the 52 years. Our network displays a minority-majority partition: the proportion of females has hovered around 33.5% for the past decade now.

(O2) **Homophily** is the is the tendency of individuals to associate with people similar to them.

Easley and Kleinberg (2010) provide the following test for homophily. Given the proportions of male and female ended edges in the network, we should be able to calculate the approximate proportion of mixed edges (the probability that we select a mixed-gendered edge at random). If the true fraction is significantly below the expected amount, the network is exhibiting homophily. Figure 4 shows that homophily is a consistently worsening problem in the NLP community. All numbers are significant with p-value virtually 0 (1-sided z-test for proportions). Note that the plot includes error bars, which are so small they are not visible.

(O3) **The "rich-get-richer" feedback mechanism** describes and explains the process of wealth concentration, by which the future distribution of wealth is predictable from empirical data

based on the current wealth distribution.

In our network, the degree of a node captures its level of social wealth: people may try to connect more often to people who already have many connections, either in order to profit from their social wealth or because they are more visible in the network. In our NLP mentor-mentee network, the average degree for male researcher nodes is 3.356, while for females it is 3.186. Hence our mentor-mentee network exhibits a "rich-get-richer" mechanism in favour of male researchers.

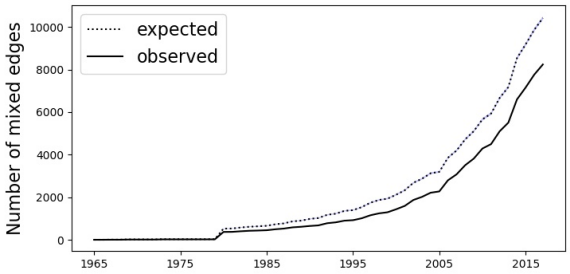

Figure 4: Evidence of ever-increasing homophily. The expected number of mixed-gender interactions is higher than the observed.

The biased preferential attachment model. Avin et al. (2015) extend Barabási and Albert (1999)'s preferential attachment model that was originally based on the "rich-get-richer" feedback mechanism to a *biased preferential attachment model* of mentor-mentee dynamics, $G(n, f, p)$, where there further is (1) a minority-majority partition (the proportion of female nodes is less than half, $f < \frac{1}{2}$) and (2) homophily. The model works as follows, instantiated to our context. Over time, a sequence of bi-populated mentor-mentee networks is constructed, $G_t = (V_t, E_t)$, like the one described in Section 2. $V_t = F_t \cup M_t$ is the set of G_t's nodes, and E_t its edges, where $F_t(M_t)$ is the set of female (male) nodes. G_0 is the empty graph. At each time $t > 0$ a mentee enters the network. The mentee is a female with probability f and a male with probability $m = 1 - f$. Assuming a rich-get-richer mechanism, the mentee chooses a potential mentor according to that mentor's importance in the network: with probability $\frac{\delta_t(u)}{\sum_{v \in V_t} \delta_t(v)}$ where $\delta_t(v)$ is the degree of $v \in V_t$. If this supervision is in-gender, then a relation (edge) is established. However if genders differ, then the relation (edge) is established according to the probability of homophily (p); otherwise (with probability $(1 - p)$) it is rejected and the mentee must restart the process of finding a mentor. Once an advisor for the mentee has been found, t increments to the next time step.

We now introduce definitions and the main theorem established by (Avin et al., 2015) for conditions of the existence of power inequality and a glass ceiling effect in bi-populated networks. Then we empirically check for these conditions in our NLP mentor-mentee network for the main result.

Power inequality definition. The sequence of mentor-mentee networks G_t is said to exhibit a *power inequality effect* for females if the average power of a female node is strictly bounded by the power of a male node: i.e., $\lim_{t\to\infty} \frac{\frac{1}{|F_t|}\sum_{v\in F_t}\delta_t(v)}{\frac{1}{|M_t|}\sum_{v\in M_t}\delta_t(v)} < 1$.

Tail and moment glass ceiling definitions. Let $\text{top}_k(F_t)$ ($\text{top}_k(M_t)$) denote the number of female (male) nodes that have degree of at least k in G_t– this is the group of scholars whose wealth in relations in the network is at level at least k; this wealth of relations is a form of power. The glass ceiling effect for the minority of females describes a process by which the proportion of access to this wealth of relations is limited for females but not for males. Formally, the sequence G_t is said to exhibit a *tail glass ceiling effect* for the female nodes (the minority) if there exists an increasing sequence k_t such that $\lim_{t\to\infty} \text{top}_{k_t}(M_t) = \infty$ and $\lim_{t\to\infty} \frac{\text{top}_{k_t}(F_t)}{\text{top}_{k_t}(M_t)} = 0$. G_t exhibits a *moment glass ceiling* g for the female nodes, if $g = \lim_{t\to\infty} \frac{\frac{1}{|F_t|}\sum_{v\in F_t}\delta_t(v)^2}{\frac{1}{|M_t|}\sum_{v\in M_t}\delta_t(v)^2}$. And if $g = 0$, G_t has a *strong glass ceiling effect*.

The main result: Power inequality and glass ceiling. Avin et al. (2015) proved that if $0 < f < \frac{1}{2}$ and $0 < p < 1$, then for $G(n, f, p)$ produced by the biased preferential attachment model, $G(n, f, p)$ exhibits both power inequality and a tail and strong glass ceiling effects. In observations (O2) and (O3), we identified the conditions $f = 0.335 < 0.5$ and the existence of homophily (i.e., $0 < p < 1$) in our NLP mentor-mentee network. We have therefore shown there to exist power inequality and a glass ceiling in NLP.

5 Concluding remarks

Given our study of the mentee-mentor network for NLP, we have shown that there is a glass ceiling for female researchers in NLP that has taken a hold of the field since the mid-2000s.

References

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, ITCS '15, pages 41–50.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286(5439):509–512.

David Easley and Jon Kleinberg. 2010. *Networks, Crowds and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Patrick Gaule and Mario Piacentini. 2018. An advisor like me? advisor gender and post-graduate careers in science. *Research Policy* (47):805–813.

Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. 2013. Bibliometrics: Global gender disparities in science. *Nature News* 504(7479):211.

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

Linda J. Sax, Kathleen J. Lehman, Jerry A. Jacobs, M. Allison Kanny, Gloria Lim, Laura Monje-Paulson, and Hilary B. Zimmerman. 2017. Anatomy of an enduring gender gap: The evolution of womens participation in computer science. *The Journal of Higher Education* 88(2):258–293. https://doi.org/10.1080/00221546.2016.1257306.

Helen Shen. 2013. Mind the gender gap. *Nature* 495(7439):22.

Romy Van der Lee and Naomi Ellemers. 2015. Gender contributes to personal research funding success in the netherlands. *Proceedings of the National Academy of Sciences* 112(40):12349–12353.

Samuel F Way, Daniel B Larremore, and Aaron Clauset. 2016. Gender, productivity, and prestige in computer science faculty hiring networks. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pages 1169–1179.

Wendy M. Williams, Agrima Mahajan, Felix Thoemmes, Susan M. Barnett, Francoise Vermeylen, Brian M. Cash, and Stephen J. Ceci. 2017. Does gender of administrator matter? national study explores u.s. university administrators' attitudes about retaining women professors in stem. *Frontiers in Psychology* 8:15.