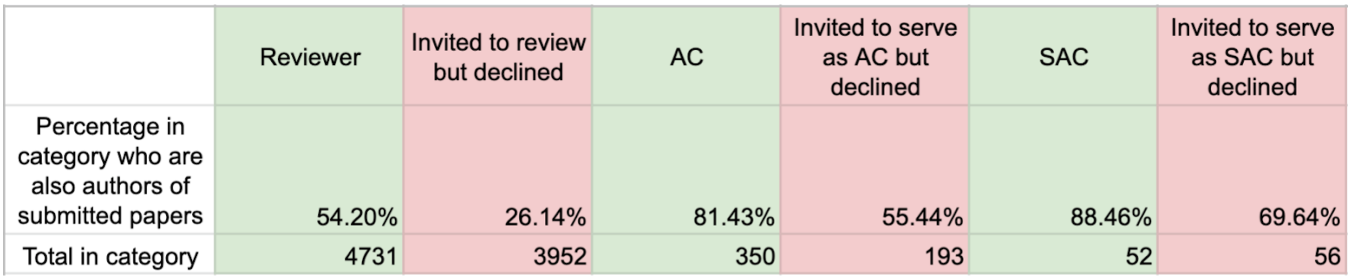
What we learned from NeurIPS 2019 data

Neural Information Processing Systems Conference
Dec 10 · 9 min read

NeurIPS has quadrupled in the last five years. This year, we had 6,743 submissions after filtering (down to 6,614 at notification time), amounting to more than 20,000 reviews written by more than 4,500 reviewers. The overall acceptance rate this year was 21.6%, corresponding to 1428 accepted papers. Given this overwhelming growth, we decided to take a look at the data and see what we can learn from it. This post is not a proposal for a new system. It's about facilitating and informing a discussion about one.

Part I: Dissecting the NeurIPS community

There were 15,920 authors of submitted papers. (We made a careful attempt to merge profiles to keep a single entity per person in this analysis.) Three quarters were not on the invitation list for the program committee in any capacity — as reviewers, area chairs (ACs), or senior area chairs (SACs). Slightly over 70% of the remaining quarter accepted our invitation to serve on the program committee, which is a good sign. Moreover, the majority of those reviewing also submitted papers — also a good sign.

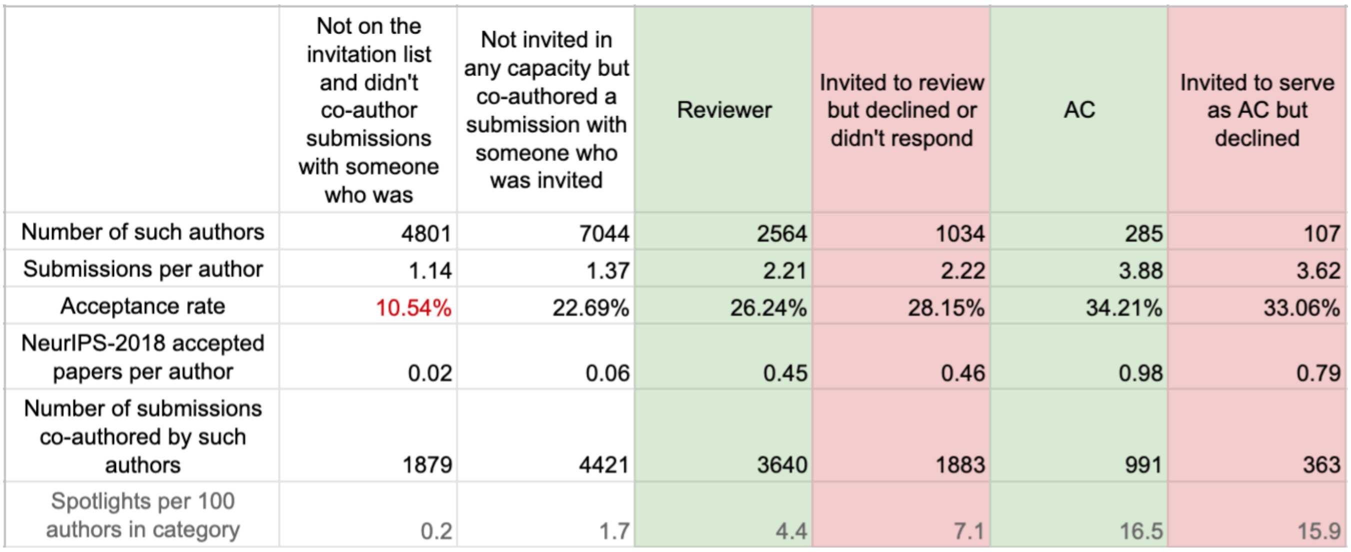


	Reviewer	Invited to review but declined	AC	Invited to serve as AC but declined	SAC	Invited to serve as SAC but declined
Percentage in category who are also authors of submitted papers	54.20%	26.14%	81.43%	55.44%	88.46%	69.64%
Total in category	4731	3952	350	193	52	56

So, does NeurIPS have a free-rider problem? Not a critical one. As the table shows, only about a quarter of those who didn't accept our invitation to review submitted papers to NeurIPS 2019. About 10% of these authors, however, submitted five or more papers. Altogether, there were only 769 papers (out of 6,743) with at least one author invited to serve on the committee but none of the invited authors contributing to the reviewing process.

Let's now take a closer look at the three quarters who were not on the invitation list for the program committee. About 40% of these authors also didn't co-author a submission with anyone who was on the invitation list. The acceptance rate in this category (responsible for close to 30% of all submissions) was the lowest — only about one in ten. Interestingly, the initial interest in these papers during the bidding phase was just as strong as in papers from any other category. We will come back to this point later in this post.

As expected, those invited to serve in a senior role both submitted more papers and had a higher acceptance rate on average. The table below gives statistics for area chairs. The numbers were somewhat higher for senior area chairs — 5.24 submissions per SAC with acceptance rate of 34.78%.

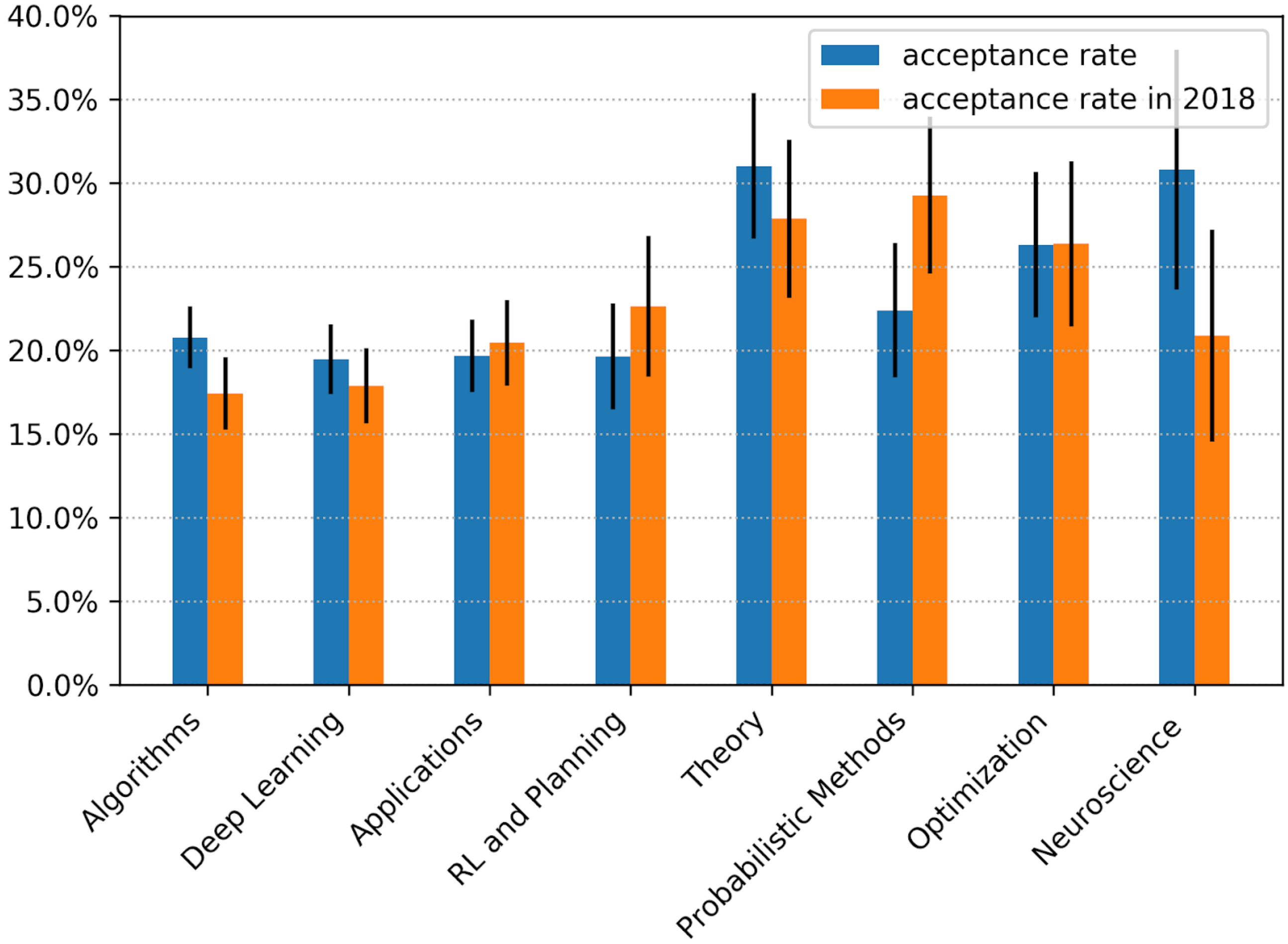


	Not on the invitation list and didn't co-author submissions with someone who was	Not invited in any capacity but co-authored a submission with someone who was invited	Reviewer	Invited to review but declined or didn't respond	AC	Invited to serve as AC but declined
Number of such authors	4801	7044	2564	1034	285	107
Submissions per author	1.14	1.37	2.21	2.22	3.88	3.62
Acceptance rate	10.54%	22.69%	26.24%	28.15%	34.21%	33.06%
NeurIPS-2018 accepted papers per author	0.02	0.06	0.45	0.46	0.98	0.79
Number of submissions co-authored by such authors	1879	4421	3640	1883	991	363
Spotlights per 100 authors in category	0.2	1.7	4.4	7.1	16.5	15.9

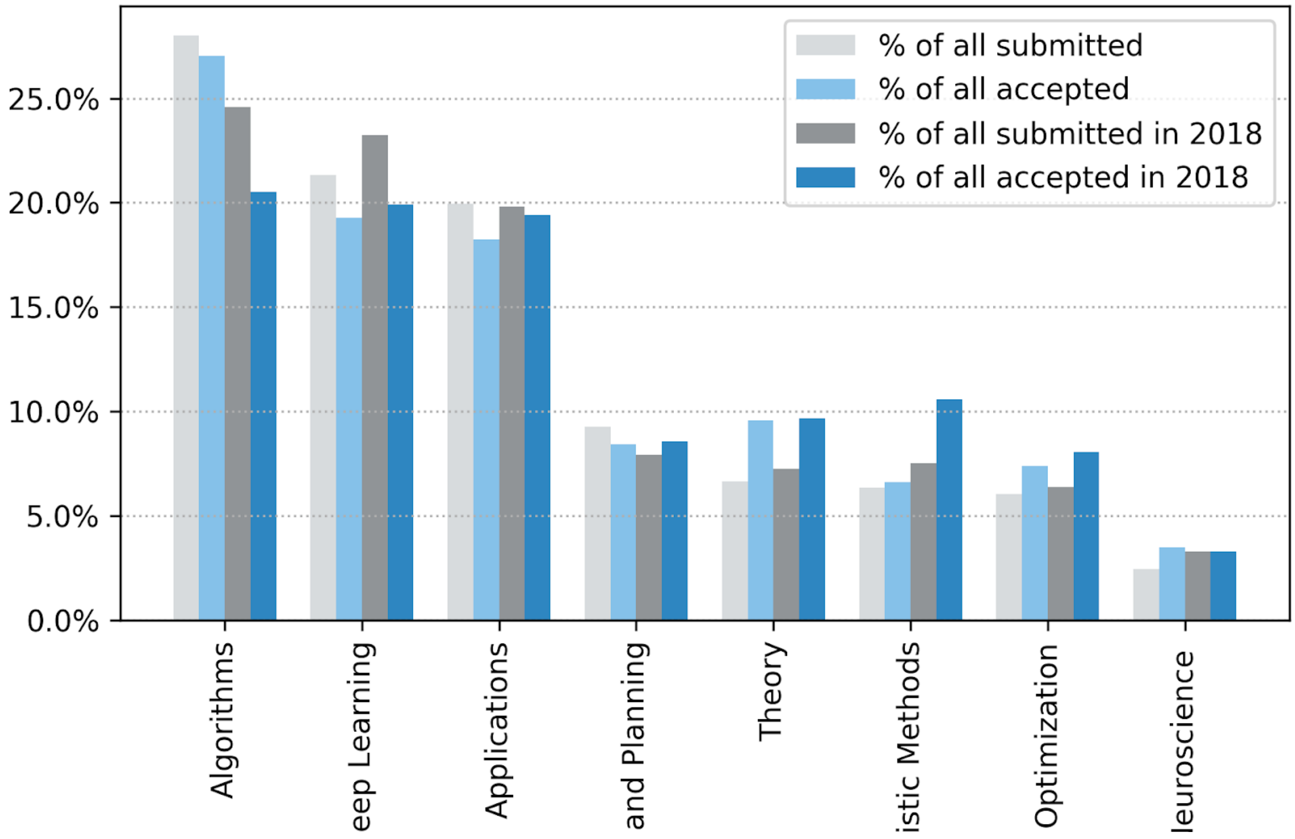
The fraction of reviewers/ACs from academia was around 70%.

There were 85 authors with at least 10 submissions. Only six of them are women (7%). Our provisional estimate of the overall fraction of women authors of submitted papers is 13%, almost twice the fraction among prolific submitters. The average acceptance rate for prolific submitters was 24.7%, slightly higher than the 21.6% base rate across all submissions.

Finally, here is a breakdown of acceptance rates by primary subject area, in comparison with 2018. This plot is ordered by the number of submissions in each area (see below in this post for a plot on submissions by area). As one explanation for the (statistically significant) differences in acceptance rates we see between the first and last four subject areas, it is not too surprising that the subject areas with the most submissions will have a larger fraction of lower-quality submissions as well.



Here is also a composition of submitted and accepted papers by subject area, in comparison with 2018.



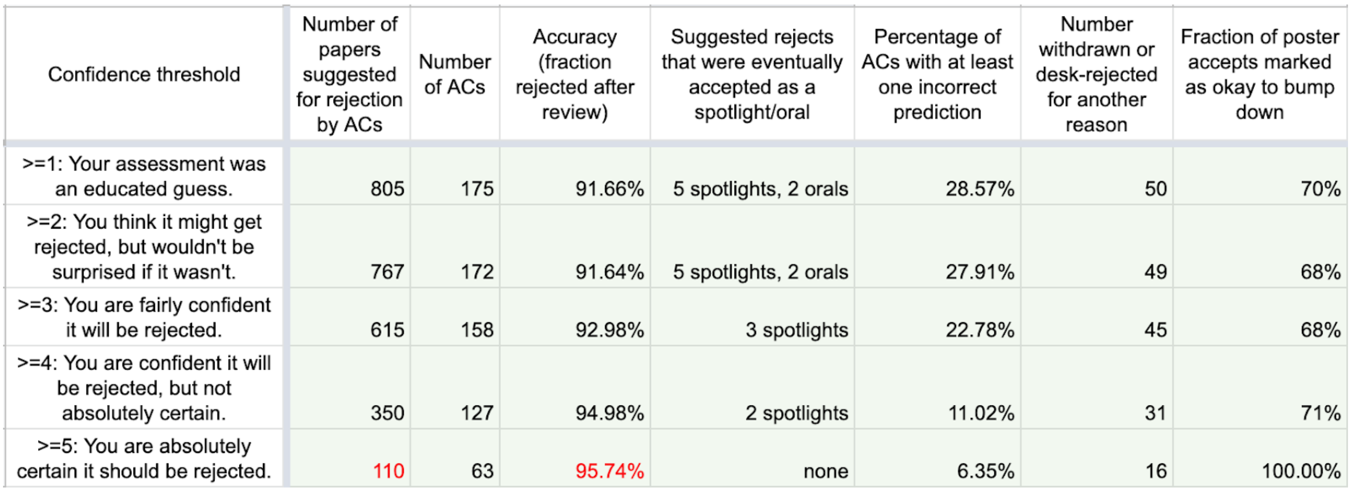
Part II: Speculative experiments on reducing or limiting the number of submissions

There have been many discussions about changing the NeurIPS reviewing model to better handle the growing number of submissions. Let's have a little fun and use the NeurIPS 2019 data to estimate the consequences of some of the proposals we've heard.

Editorial screening

As an experiment, we wanted to measure the ability of ACs to predict, before seeing reviews, which of their assigned submissions were going to be rejected (for example, due to their insufficient novelty or poor scholarship). The question is whether NeurIPS should consider allowing ACs to reject submissions without review in order to reduce the reviewing load — such editorial screening is common practice at top journals.

We asked each AC to provide what they believed to be the bottom 25% of their assigned submissions (5 submissions for most ACs), together with their confidence in each assessment. We received 808 reject recommendations from 50% of our ACs. The table below shows the number of papers recommended for rejection at different confidence levels, as well as the corresponding accuracy — percentage actually rejected after review. Thresholding at confidence level 5 (last row) appears safe but doesn't yield any appreciable reduction in the number of submissions. Even if one accounts for the fact that only 50% of our ACs participated in this experiment, thresholding at confidence level 5 would eliminate the need to review only <4% of all submissions.

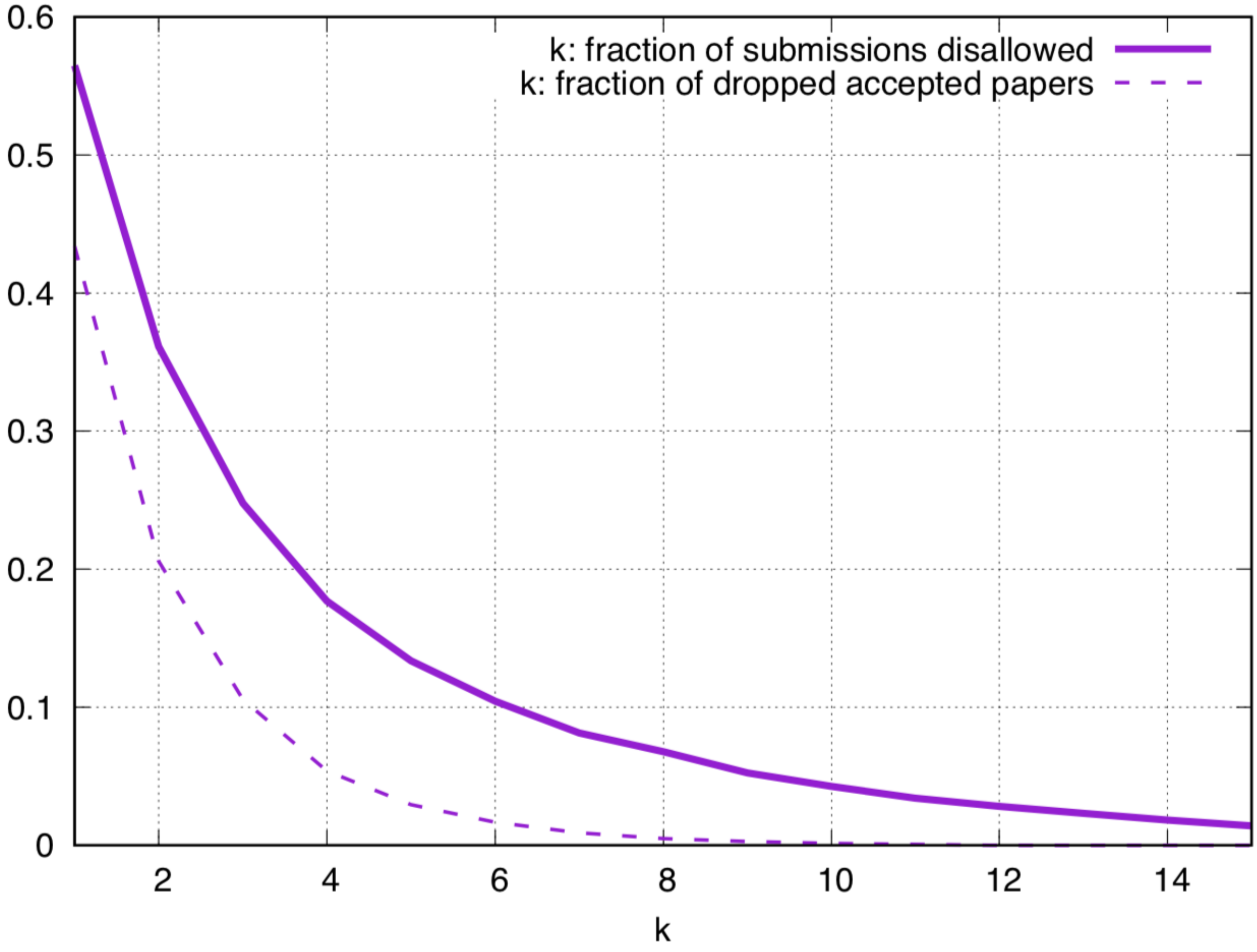


Confidence threshold	Number of papers suggested for rejection by ACs	Number of ACs	Accuracy (fraction rejected after review)	Suggested rejects that were eventually accepted as a spotlight/oral	Percentage of ACs with at least one incorrect prediction	Number withdrawn or desk-rejected for another reason	Fraction of poster accepts marked as okay to bump down
>=1: Your assessment was an educated guess.	805	175	91.66%	5 spotlights, 2 orals	28.57%	50	70%
>=2: You think it might get rejected, but wouldn't be surprised if it wasn't.	767	172	91.64%	5 spotlights, 2 orals	27.91%	49	68%
>=3: You are fairly confident it will be rejected.	615	158	92.98%	3 spotlights	22.78%	45	68%
>=4: You are confident it will be rejected, but not absolutely certain.	350	127	94.98%	2 spotlights	11.02%	31	71%
>=5: You are absolutely certain it should be rejected.	110	63	95.74%	none	6.35%	16	100.00%

Capping the number of submissions

Another often mentioned proposal is to cap the number of papers that anyone can submit. The AAAI conference even introduced a cap of 15 submissions per author for 2020 (see their Call for Papers).

The plot below shows how allowing everyone to co-author only *k* submissions (X axis) would have affected the total submission count at NeurIPS 2019. The Y axis plots the resulting reduction in submission volume. For the purpose of this thought experiment — as we don't know which submissions each author would have chosen to keep given this policy — we gave each author the hindsight of keeping their accepted submissions, up to the allowed *k* at random. If the author had any remaining submission budget, it was filled with their randomly selected rejected submissions.



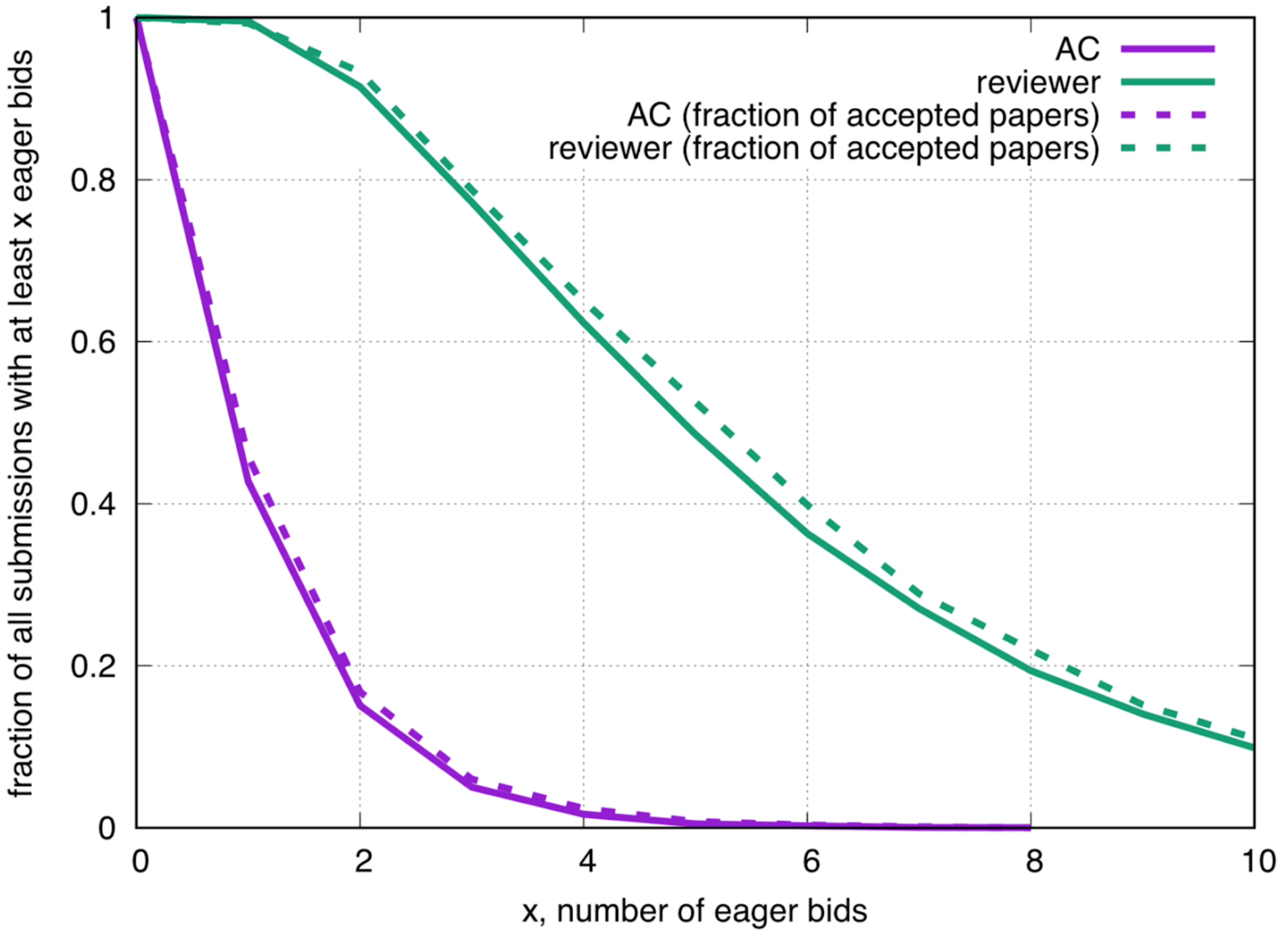
Capping submissions at 15 (as adopted by AAAI-2020) would have eliminated only <100 submissions, 1.5% of the total. Capping submissions at 10 would have removed 4.3% of the total (saving ~850 reviews) with no significant impact on the resulting program.

In summary, perhaps some combination of editorial screening and capping submissions can give a sufficient reduction to make a difference, but more thought about methods of doing this is needed before putting a practice in place.

Supply-and-Demand Reviewing

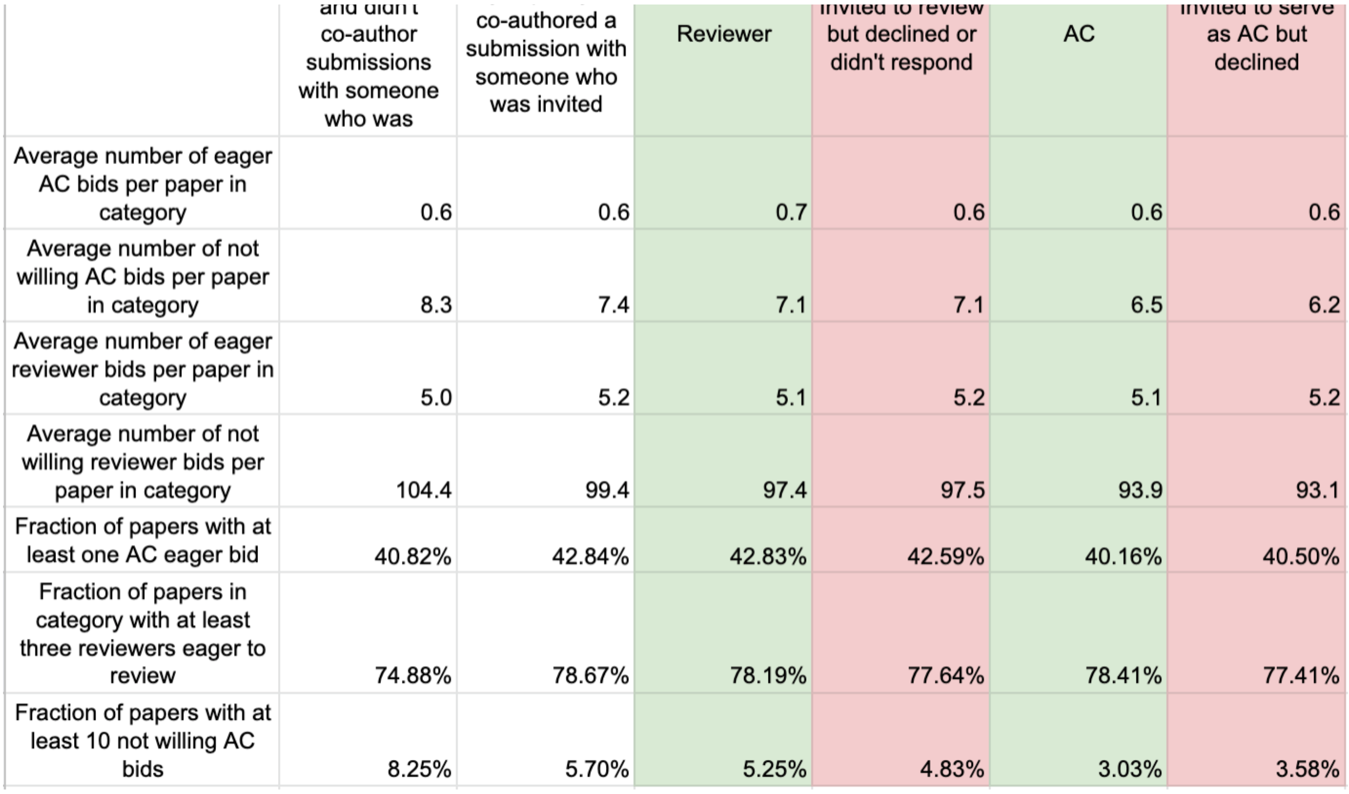
Another proposal (here and here) is to use a market system to control reviewing. Only submissions that gather sufficient interest from reviewers are reviewed.

The analysis below shows that bids — the way they are implemented currently — are a poor predictor of acceptance. Accepted papers had 5.4 eager bids from reviewers on average (0.72 eager bids from ACs), compared to 5.1 for a rejected paper (0.64 from ACs). Thus a naive policy where only submissions with at least three eager bids are reviewed would have eliminated about a quarter of all submissions but it would have also eliminated a quarter of all accepted papers.



The table below breaks this down by author categories, showing that eager bids are remarkably flat across categories.

	Not on the invitation list and didn't	Not invited in any capacity but		Invited to review		Invited to serve



	and didn't co-author submissions with someone who was	co-authored a submission with someone who was invited	Reviewer	invited to review but declined or didn't respond	AC	invited to serve as AC but declined
Average number of eager AC bids per paper in category	0.6	0.6	0.7	0.6	0.6	0.6
Average number of not willing AC bids per paper in category	8.3	7.4	7.1	7.1	6.5	6.2
Average number of eager reviewer bids per paper in category	5.0	5.2	5.1	5.2	5.1	5.2
Average number of not willing reviewer bids per paper in category	104.4	99.4	97.4	97.5	93.9	93.1
Fraction of papers with at least one AC eager bid	40.82%	42.84%	42.83%	42.59%	40.16%	40.50%
Fraction of papers in category with at least three reviewers eager to review	74.88%	78.67%	78.19%	77.64%	78.41%	77.41%
Fraction of papers with at least 10 not willing AC bids	8.25%	5.70%	5.25%	4.83%	3.03%	3.58%

On the flip side, this suggests that all good papers have a good chance of being discovered (consistent with observations in Yann LeCun's proposal here).

Open Dissemination of Submissions under Review

A majority (54%) of all submissions were posted on arXiv; 21% of these submissions were seen by at least one reviewer. The acceptance rate in this latter category was 34%, significantly higher than the base rate of 21.6%. For comparison, the acceptance rate for submissions that were not posted was 17%.

Unfortunately it's hard to disentangle cause and effect. One obvious possibility is that papers released on arXiv are of higher quality, since the authors judged them ready to be shared publicly. Another is that this reflects a bias of single blind review, where well known authors are both more likely to have their arXiv paper read and to induce a bias towards a positive evaluation of their work.

Part III: Review quality

Reviewer Assignment

What's a good proxy for review quality that we can objectively measure? One proxy suggested to us was whether an assigned reviewer is cited in the paper. So what is the fraction of NeurIPS 2019 submissions with at least one cited reviewer?

We extracted reference pages from submission files to find out. As it turns out, less than one third of all submissions was reviewed by someone cited in the paper. As expected, being cited in the submission does correlate with confidence. The average confidence of a NeurIPS review was 3.75, with about half of all reviews rated as 4 (confident in the assessment but not absolutely certain). The average confidence of a cited reviewer was slightly above 4, with close to 30% rated 5 (absolutely certain about the assessment, very familiar with the related work) — almost twice the rate in the general review pool.

While we had certainly hoped to see higher numbers, 40.6% of all submissions had at least one review with confidence rating of 5, and 94.7% had a review with confidence rating of at least 4.

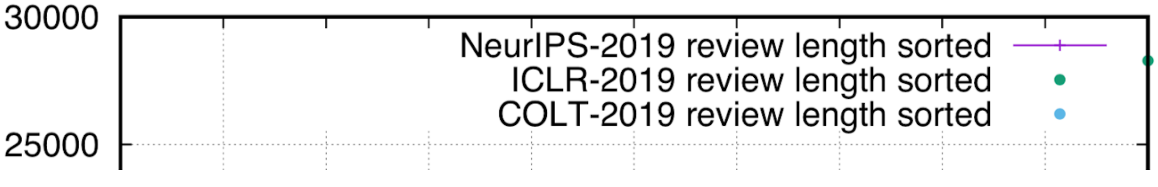
In an effort to improve the assignment, we worked with CMT to allow ACs to recruit external reviewers for specific papers they were handling. If an AC couldn't find a good match in the general pool, they could send a paper-specific invitation to an external reviewer. More than 40% of ACs used this feature, sending close to 400 invitations (almost 80% of which were accepted). ACs could also manually adjust automatically generated assignments for papers in their stack, hand-picking from the non-conflicted general pool. While most ACs largely kept the assignment they received, 10% of our ACs reassigned at least a third of their assignment — averaging at least one reviewer per paper they were handling.

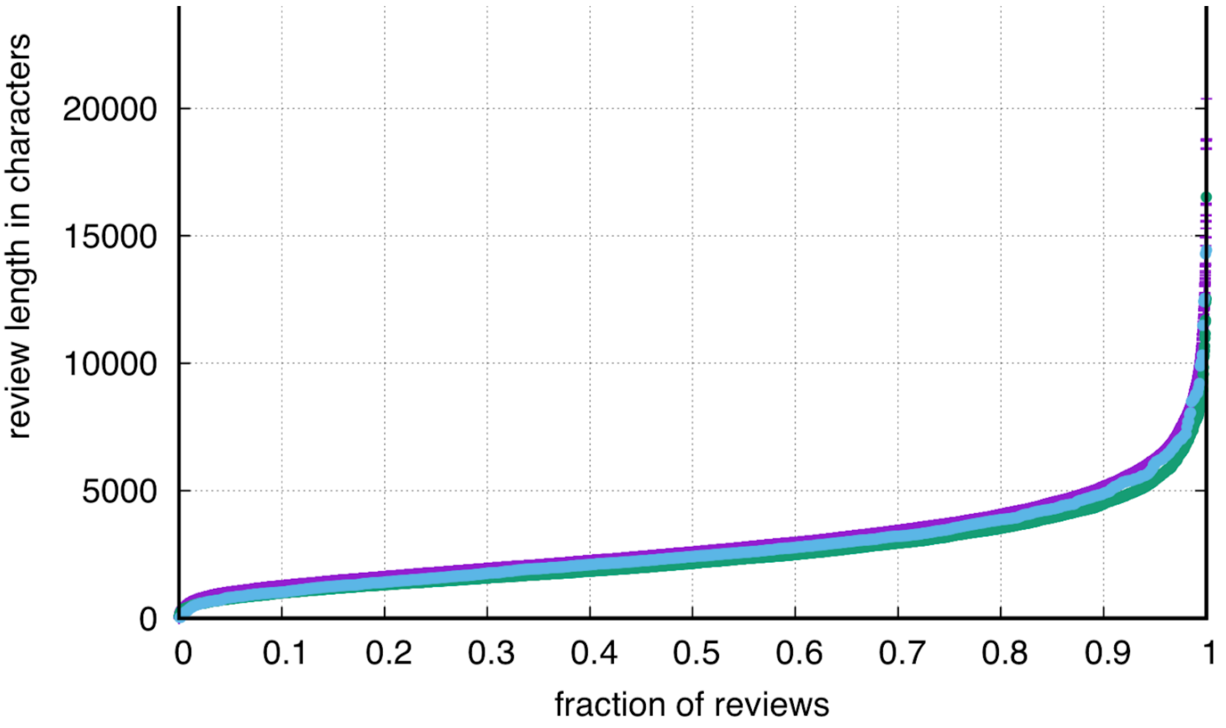
Were ACs more satisfied with reviews if they hand-picked the reviewers? The answer is yes, even though most of these external reviewers were junior. The fraction of reviews rated as "exceeding expectations" grew by a third and the fraction rated as "failed expectations" more than halved in the hand-picked pool.

Distribution of Review Length

Given frequent complaints about short reviews at NeurIPS, we looked at the distribution of review lengths for NeurIPS 2019, ICLR 2019, and COLT 2019.

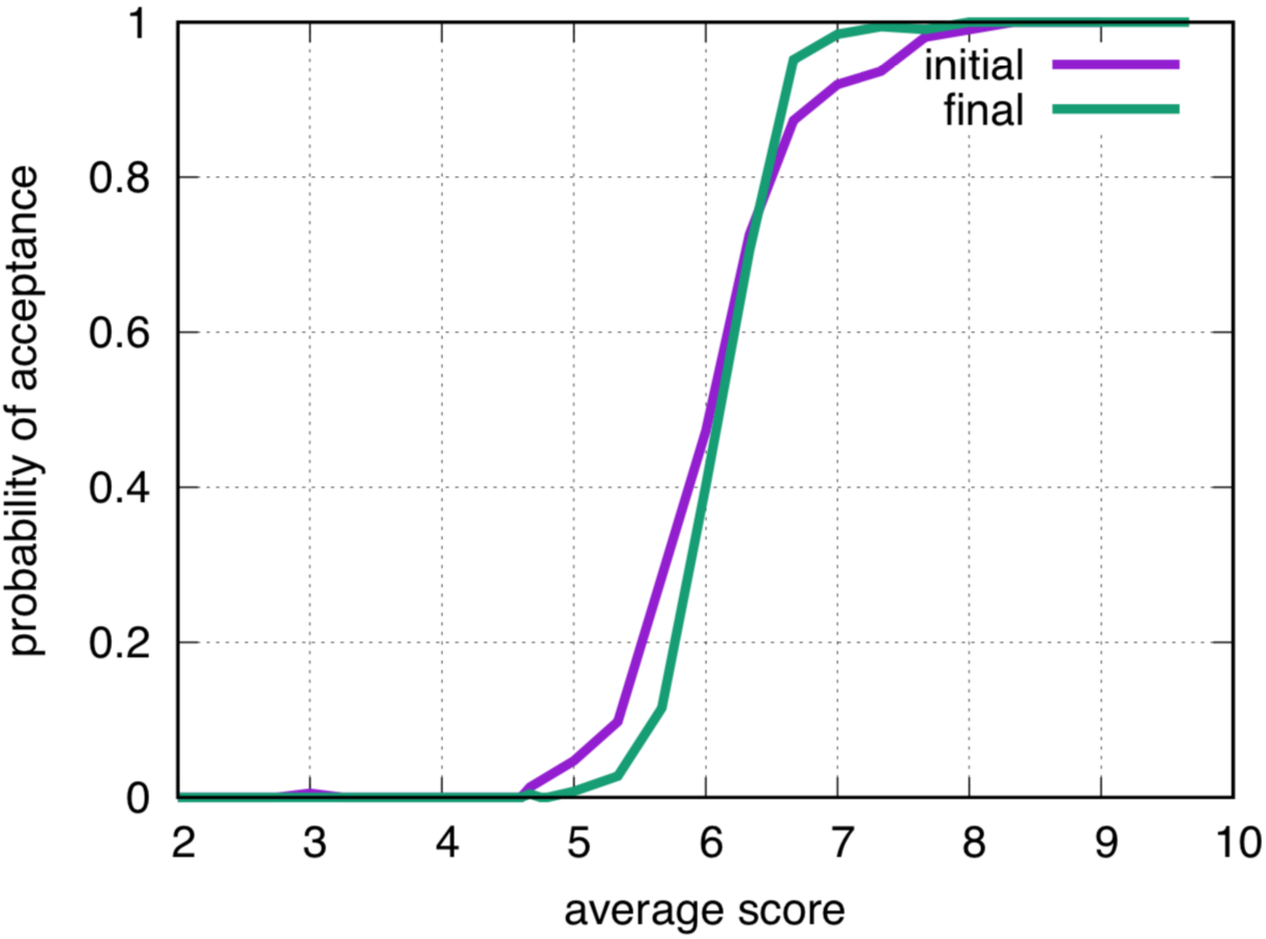
A lognormal distribution of human-generated text length is to be expected but it's interesting to see that the parameters match so closely across very different conferences and different review forms.





Rebuttals, Discussions, and Acceptance Statistics

As an author composing your rebuttal, you probably want to know the probability of your paper getting accepted given its initial scores. What are the odds of your rebuttal shifting the outcome?



About 20% of initial scores changed during the discussion phase, translating into at least one score changing for about 50% of all submissions. As decisions were being reached, the average variance went down from 1.27 (pre-rebuttal) to 0.89 (notification time).

We also compared the following engagement metrics between 2018 and 2019: The average number of comments per paper during discussion period, the average number of people participating, the average number of characters in the discussion posts. All numbers went up, indicating better overall engagement in this part of the review process. Most significantly, the average length of the discussion thread per paper increased by 10%.

Summary

Though the data still leaves a lot of questions unanswered, we personally notice the following takeaways:

1. **No free-loader problem:** Relatively few papers are submitted where none of the authors invited to participate in the review process accepted the invitation
2. **Unclear how to rapidly filter papers prior to full review:** Allowing for early desk rejects by ACs is unlikely to have a significant impact on reviewer load without producing inappropriate decisions. Likewise, the eagerness of reviewers to review a particular paper is not a strong signal, either.
3. **No clear evidence that review quality as measured by length is lower for NeurIPS:** NeurIPS is surprisingly not much different from other conferences of smaller sizes when it comes to review length.
4. **Impact of engagement in rebuttal/discussion period:** Overall engagement seemed to be higher than in 2018.

Although I’m sure this will not end our neverending enthusiasm in debating new reviewing models, hopefully this post can help further focus our future discussions on the topic.

Alina Beygelzimer, Emily Fox, Florence d’Alché-Buc, Hugo Larochelle
NeurIPS 2019 Program Chairs