Challenges for ACL

ACL Presidential Address 2017

Joakim Nivre
(A)CL is booming!
(A)CL is booming!
(A)CL is booming!
The future is bright!
Challenges for ACL

Equity and Diversity

Publishing and Reviewing

Good Science
Can you spot the mistake?

ACL Presidents 2009–2017
Can you spot the mistake?

ACL Presidents 2009–2017

Only four of them are wearing funny hats?
Can you spot the mistake?

Has Jack Lemmon really been ACL president?

ACL Presidents 2009–2017
Can you spot the mistake?

ACL Presidents 2009–2017
Can you spot the mistake?

I didn’t spot it three years ago!

ACL Presidents 2009–2017
A biased world?
A biased world?

The Matthew/Matilda effect in science

Margaret W. Rossiter, *Social Studies of Science*, 1993
A biased world?

The Matthew/Matilda effect in science

Margaret W. Rossiter, Social Studies of Science, 1993

Nepotism and sexism in peer-review

Christine Wennerås and Agnes Wold, Nature, 1997
A biased world?

The Matthew/Matilda effect in science

Margaret W. Rossiter, *Social Studies of Science*, 1993

Nepotism and sexism in peer-review

Christine Wennerås and Agnes Wold, *Nature*, 1997

Awards and prizes in the US, 1990s and 2000s

Anne E. Lincoln, Stephanie Pincus, Janet Bandows Koster and Phoebe S. Leboy, *Social Studies of Science*, 2012
A biased world?

The Matthew/Matilda effect in science

Margaret W. Rossiter, *Social Studies of Science*, 1993

Nepotism and sexism in peer-review

Christine Wennerås and Agnes Wold, *Nature*, 1997

Awards and prizes in the US, 1990s and 2000s

Anne E. Lincoln, Stephanie Pincus, Janet Bandows Koster and Phoebe S. Leboy, *Social Studies of Science*, 2012

Nationality, ethnicity, religion, gender, sexual orientation, …
Why does it matter?
Why does it matter?

Equity

“The acceptance or rejection of claims entering the lists of science is not to depend on the personal or social attributes of their protagonist; [...] race, nationality, religion, class, and personal qualities are as such irrelevant.”

Why does it matter?

Equity

“The acceptance or rejection of claims entering the lists of science is not to depend on the personal or social attributes of their protagonist; [...] race, nationality, religion, class, and personal qualities are as such irrelevant.”


Diversity

“Decades of research by organizational scientists, psychologists, sociologists, economists and demographers show that socially diverse groups (that is, those with a diversity of race, ethnicity, gender and sexual orientation) are more innovative than homogeneous groups.”

What is ACL doing?
What is ACL doing?

A new nominating committee
What is ACL doing?

A new nominating committee

Revised selection criteria for ACL fellows
What is ACL doing?

A new nominating committee

Revised selection criteria for ACL fellows

Promoting a large and diverse pool of nominations
What is ACL doing?

A new nominating committee

Revised selection criteria for ACL fellows

Promoting a large and diverse pool of nominations

Resources for promoting diversity and preventing bias
LAST CALL FOR AREA CHAIRS – A CALL FOR DIVERSITY!

[kmnyn  Uncategorised]

Dear readers:

[TL;DR: We need a diverse set of candidates! Nominate someone (yourself) for area chair here!]
Hello Association for Computational Linguistics Parents!

Thank you very much for your interest in the Association for Computational Linguistics children’s program. Our goal is to provide your children with a program they want to attend, while providing you with that critical “peace of mind” feeling so you can attend your event activities.

KiddieCorp is pleased to provide a children’s program during the July 2017 Annual Meeting. KiddieCorp is in its thirty-first year of providing high-quality children’s programs and youth services to conventions, trade shows and special events. We take caring for your children very seriously. KiddieCorp has enjoyed a long-time partnership with the American Academy of Pediatrics, which has helped to establish KiddieCorp as a premier provider of event children’s program services.

ACTIVITIES

Activities include exciting themes, arts & crafts, group games, music & movement, board games, story time, dramatic play, etc. We provide activities appropriate for each age group, using safe and sturdy equipment. Children can make their own choices within KiddieCorp’s program.
Women and Underrepresented Minorities in Natural Language Processing

Stronger together

WiNLP workshop

Mission
We’re here to help promote and support ideas and voices of women and other underrepresented groups in Natural Language Processing (NLP).

Contact
To contact us: winip-chairs@googlegroups.com. Follow us on Twitter and Facebook!

Read more »
Talk about it!
The ACL Publishing Model
The ACL Publishing Model

- We like conferences
The ACL Publishing Model

- We like conferences
- We use anonymous peer review (almost everywhere)
The ACL Publishing Model

• We like conferences
• We use anonymous peer review (almost everywhere)
• And then there is this thing called arXiv …
Peer review

• First used by *Philosophical Transactions of the Royal Society*, London, 1665
• Standard in scientific journals from the mid-20th century
• Meant to guarantee scientific quality – “organized scepticism”
Peer review
• First used by *Philosophical Transactions of the Royal Society*, London, 1665
• Standard in scientific journals from the mid-20th century
• Meant to guarantee scientific quality – “organized scepticism”

Double-blind peer review
• Popularized by sociology journals in the 1950s
• Less common in natural sciences and engineering
• Meant to reduce author bias – the Matthew/Matilda effect
Single versus Double Blind Reviewing at WSDM 2017

Andrew Tomkins  
Google  
Mountain View, CA, USA  
atomkins@gmail.com

Min Zhang  
Tsinghua University  
Beijing, China  
z-m@tsinghua.edu.cn

William D. Heavlin  
Google  
Mountain View, CA, USA  
bheavlin@google.com

ABSTRACT

In this paper we study the implications for conference program committees of using single-blind reviewing, in which committee members are aware of the names and affiliations of paper authors, versus double-blind reviewing, in which this information is not visible to committee members. WSDM 2017, the 10th ACM International Conference on Web Search and Data Mining, performed a controlled experiment in which each paper was reviewed by four committee members. Two of these four reviewers were chosen from a pool of committee members who had access to author information; the other two were chosen from a disjoint pool who did not have access to this information. This information asymmetry persisted through the process of bidding for papers, reviewing papers, and entering scores. Reviewers in the single-blind condition typically bid for 22% more papers, and preferentially bid for papers from top institutions. Once papers were allocated to reviewers, single-blind reviewers were significantly more likely than their double-blind counterparts to recommend for acceptance papers from famous authors and top institutions. The estimated odds multipliers are 1.66 for famous authors and 1.61 and 2.10 for top universities and companies respectively, so the result is tangible. For female authors, the associated odds multiplier of 0.79 is not statistically significant in our study. However, a meta-analysis places this value in line with that of other experiments, and in the context of this larger aggregate the gender effect is also statistically significant.

1 INTRODUCTION

The scientific peer-review process dates back to the 1600’s, and is generally regarded as a cornerstone of the scientific method. The details of its implementation have been scrutinized and explored across many academic disciplines.

Peer review has many dimensions. At the present time, there is a growing body of research that addresses whether it is effective, whether it is fair, and whether it is susceptible to manipulation. A key aspect of the research is the role of the reviewer.

Each side of the question. Terminology is not completely uniform across the sciences, but following common usage in computer science, we refer to single-blind reviewing as the practice of making reviewers aware of author identity but not the other way around. In double-blind reviewing, neither party is aware of the identity of the other.

Numerous anecdotal studies argue for one form or the other of peer review, often based on observations of findings before and after switching models. A much smaller number of researchers have performed controlled studies of the effects of the two models. Notable among these is the work of Rebecca Blank from 1991 [1], who performed a beautiful controlled study in reviewing papers submitted to the American Economic Review over a two-year period from 1987 to 1989. We discuss this and other related work in some detail below.

The current work came about when two of the authors of this paper were asked to co-chair the program of WSDM 2017, the 10th International ACM Conference on Web Search and Data Mining. WSDM has for its entire history employed single-blind reviewing. We were asked to consider switching to double-blind this year. Upon a review of the literature, we discovered that earlier controlled experiments in the journal setting missed many key aspects of the standard WSDM reviewing process, while many discussions of conferences switching between reviewing methods were uncontrolled experiments in the sense that the switch took place from one year to the next, introducing an analytically intractable set of possible confounding factors. Hence, we decided to perform an experiment in order to make an informed recommendation to the chairs of WSDM 2018, and to offer our findings to the rest of the community.

We now summarize some differences between conference and journal reviewing processes. As a backdrop, we observe that the accelerated pace of computer science in recent decades has led to the ascendance of academic conferences as a primary means for dissemination of new results. The level of formal methodology and scientific rigor presented in conferences, however, is lower than that of a scientific journal. As a result, the impact of the outcome of a conference is lower than that of the publication of an academic journal.

The process of reviewing has a number of dimensions that can be exploited for research. The dominant model is the individual paper review, in which a single reviewer is assigned to each paper and provides a rating or a recommendation. It is possible to design a more centralized, less individualized model, in which a panel is assigned to each paper and provides a rating or a recommendation. This model is analogous to the way in which a referee is assigned to a conference paper and provides a rating or a recommendation. It is possible to design a more centralized, less individualized model, in which a panel is assigned to each paper and provides a rating or a recommendation. This model is analogous to the way in which a referee is assigned to a conference paper and provides a rating or a recommendation.
Odds multipliers from double-blind to single-blind review:

- Famous authors: 1.66
- Top universities: 1.61
- Top companies: 2.10
701 reviewers participated in a survey in June 2017
30% claimed they could identify the authors of a paper
17% contributed 196 unique guesses
Are we overusing peer review?

- Increased volumes lead to reviewer fatigue and lower quality
- Conferences lose their role as a discussion forum for new ideas
- Scientific gatekeeping should be left to journals?

![ACL submissions graph](image)
Are we overusing peer review?

- Increased volumes lead to reviewer fatigue and lower quality
- Conferences lose their role as a discussion forum for new ideas
- Scientific gatekeeping should be left to journals?

Ideas to consider

- Prescreening to reduce reviewer load (and increase quality)
- Journal-style reviewing with rolling deadline for long papers
- Abstract submission for short papers/posters
Are we overusing peer review?

• Increased volumes lead to reviewer fatigue and lower quality
• Conferences lose their role as a discussion forum for new ideas
• Scientific gatekeeping should be left to journals?

Ideas to consider

• Prescreening to reduce reviewer load (and increase quality)
• Journal-style reviewing with rolling deadline for long papers
• Abstract submission for short papers/posters

How to combine our model with preprints on arXiv?
Dear readers!

As comments published in response to our last blog illustrate, the issue of arXiv and blind reviewing is controversial. The only part on which everybody seems to agree is the inadequacy of the current policy. We are rejecting papers for non-anonymised submissions where the authors genuinely forgot to remove their names, but let in papers previously submitted to arXiv, if they are declared during submission time. While this declaration supposedly gives reviewer an option to avoid seeing author names, in practice it is almost impossible. As a result, a large fraction of submitted papers are reviewed under different conditions than the rest. If we continue with the current policy, arXiv will be a death warrant for double-blind ACL reviewing.
FAST

Speeds up scientific advances?

Threatens scholarly thoroughness?
Anyone can publish, anyone can read and discuss

Papers can be revised with version control
PEER REVIEW

Openness undermines double-blind review

Should preprints be cited in peer reviewed work?
ACL Survey on Preprint Publishing and Reviewing

• Run during three weeks in June 2017
• 623 complete responses
• Full report available at: https://www.aclweb.org/portal/
ACL Survey on Preprint Publishing and Reviewing

22% upload to preprint servers always or often
- More likely user: graduate student, male
- Less likely user: academic researcher, female

27% cite preprints often or very often
- Frequent users are more likely to cite often
88% consider double-blind reviewing important

- 65% consider it more important than preprint publishing
- 9% consider preprint publishing more important
ACL Survey on Preprint Publishing and Reviewing

87% would submit to ACL if preprints were banned
  • 5% would probably stop submitting to ACL

Would you still submit?

- Definitely yes: 70%
- Probably yes: 17%
- Maybe: 8%
- Probably not: 4%
- Definitely not: 0%
ACL Survey on Preprint Publishing and Reviewing

How would you like to see ACL’s reviewing model working in the future?

- No reviewing
- Open reviewing
- Single-blind reviewing
- Status quo
- Discourage preprints
- Ban preprints
The following are some actions suggested by members of the community. Please indicate which, if any, you would like to see implemented.

- Lobby preprint servers to allow papers to be anonymously uploaded
- Make available author guidelines for citing preprint papers
- Have a separate track at ACL for preprint papers
- Journal-style reviewing with a rolling deadline (with CL and/or TACL)
Conclusions – Survey
Conclusions – Survey

• Strong support for double-blind reviewing in community
Conclusions – Survey

• Strong support for double-blind reviewing in community
• Weak support for completely banning preprints
Conclusions – Survey

- Strong support for double-blind reviewing in community
- Weak support for completely banning preprints
- Temporarily anonymous preprints is worth looking into
Conclusions – Survey

- Strong support for double-blind reviewing in community
- Weak support for completely banning preprints
- Temporarily anonymous preprints is worth looking into
- Guidelines for citing preprints are needed
Conclusions – Survey

• Strong support for double-blind reviewing in community
• Weak support for completely banning preprints
• Temporarily anonymous preprints is worth looking into
• Guidelines for citing preprints are needed
• Many are concerned about reviewing quality
Conclusions – Survey

• Strong support for double-blind reviewing in community
• Weak support for completely banning preprints
• Temporarily anonymous preprints is worth looking into
• Guidelines for citing preprints are needed
• Many are concerned about reviewing quality
• We have to work both long-term and short-term
ACL Business Meeting

Wednesday, August 2, 13:00-14:30

- More detailed analysis of the survey
- Invited quick-fire position statements
- Open discussion about publishing and reviewing
Good Science
Good Science

“Measurement as a virtue in itself”
Good Science

“Measurement as a virtue in itself”

“Lots of numbers with very small differences”
“Good Science”

“Measurement as a virtue in itself”

“Lots of numbers with very small differences”

“What are the research questions?”
Experimental Science
Experimental Science

- Experiments are run to test hypotheses
Experimental Science

• Experiments are run to test hypotheses
• Hypotheses are tentative theoretical explanations

  morphological segmentation facilitates syntactic parsing

  system A outperforms system B on data set C
Experimental Science

- Experiments are run to test hypotheses
- Hypotheses are tentative theoretical explanations
  - morphological segmentation facilitates syntactic parsing
  - system A outperforms system B on data set C
- Validating hypotheses requires repeated testing
Reproducible Science
Reproducible Science

Replicability

• Repeating the same experiment with the same result
• Necessary to establish the reliability of measurements
• Enables benchmarking and “fair” comparisons – fast science
Reproducible Science

Replicability

• Repeating the same experiment with the same result
• Necessary to establish the reliability of measurements
• Enables benchmarking and “fair” comparisons – fast science

Reproducibility

• Repeating a similar experiment with a similar result
• Necessary to establish the validity of hypotheses
• Requires repeated testing and comparative analysis – slow science
Reproducible Science

Replicability

• Repeating the same experiment with the same result
• Necessary to establish the reliability of measurements
• Enables benchmarking and “fair” comparisons – fast science

Reproducibility

• Repeating a similar experiment with a similar result
• Necessary to establish the validity of hypotheses
• Requires repeated testing and comparative analysis – slow science

We need both!
Reproducible Science

Replicability
• Repeating the same experiment with the same result
• Necessary to establish the reliability of measurements
• Enables benchmarking and "fair" comparisons – fast science

We need both!

Reproducibility
• Repeating a similar experiment with a similar result
• Necessary to establish the validity of hypotheses
• Requires repeated testing and comparative analysis – slow science

We need diversity!
Everything is connected …
Everything is connected ...

Publishing and reviewing

- Our traditional publishing model has a bias towards fast science
- Reinforced by increasing reviewer loads
- Accelerated by preprint publishing
Everything is connected …

Publishing and reviewing

• Our traditional publishing model has a bias towards fast science
• Reinforced by increasing reviewer loads
• Accelerated by preprint publishing

Equity and diversity

• We need diversity and innovation in research and publishing
• A more inclusive and diverse community is more likely to give us that
Keep up the good work!