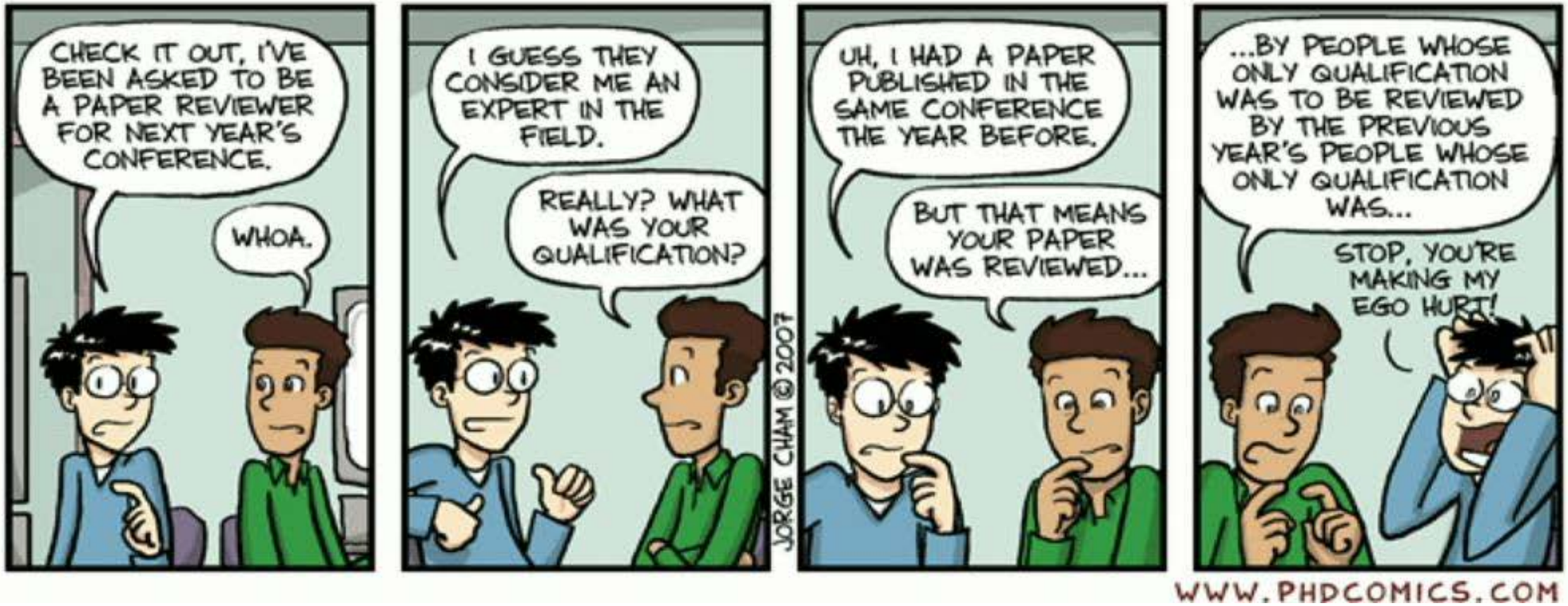
Battling Demons in Peer Review

Nihar B. Shah

Machine Learning Department and Computer Science Department

Carnegie Mellon University



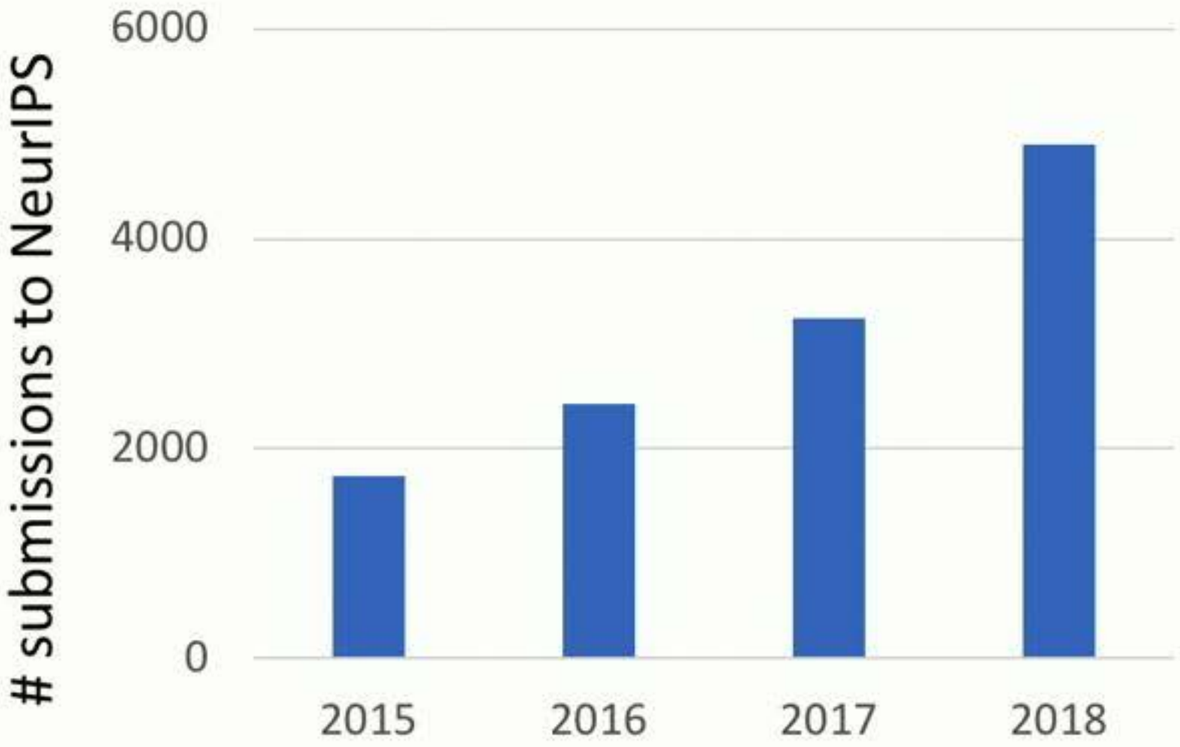
Challenge across many research fields

- **Drummond Rennie (Nature, 2016):**

  "Peer review ... is a human system. Everybody involved brings ***prejudices****,* ***misunderstandings*** *and gaps in knowledge, so no one should be surprised that peer review is often* ***biased*** *and* ***inefficient****. It is occasionally* ***corrupt****, sometimes a charade, an open temptation to plagiarists. Even with the best of intentions, how and whether peer review identifies high-quality science is unknown. It is, in short,* ***unscientific****."*

- **Overwhelming desire for improvement**
  [surveys by Smith 2006, Ware 2008, Mulligan et al. 2013]

Tremendous growth



Several thousands of submissions, 40% increase per year

Tackle systematic problems in peer review

using principled and practical approaches

Subjectivity

Miscalibration

Biases

Strategic behavior

Noise

Subjectivity

Miscalibration

Biases

Strategic behavior

Noise

UNFAIR

Subjectivity

Miscalibration

Biases

Strategic behavior

Noise

Subjectivity

Miscalibration

Biases

Strategic behavior

Noise

Detail

Some detail

Brief overview

Many other applications

Hiring

Admissions

A/B testing

yelp
tripadvisor

Online ratings

Crowdsourcing

Peer grading

...

Subjectivity

with

Ritesh Noothigattu

Ariel Procaccia

Differing opinions about relative importance of criteria

[Kerr et al. 1977, Mahoney 1977, Bakanic et al. 1987, Hojat et al. 2003, Church 2005, Lamont 2009]

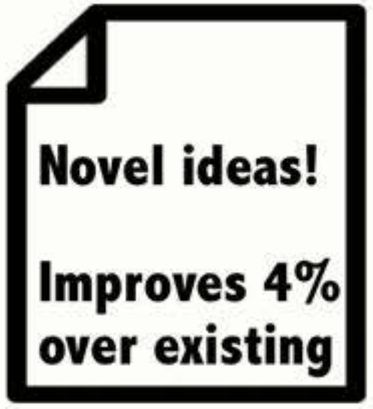
Novelty is not useful unless improvement by at least 10%

Novelty is extremely important

Differing opinions about relative importance of criteria

[Kerr et al. 1977, Mahoney 1977, Bakanic et al. 1987, Hojat et al. 2003, Church 2005, Lamont 2009]

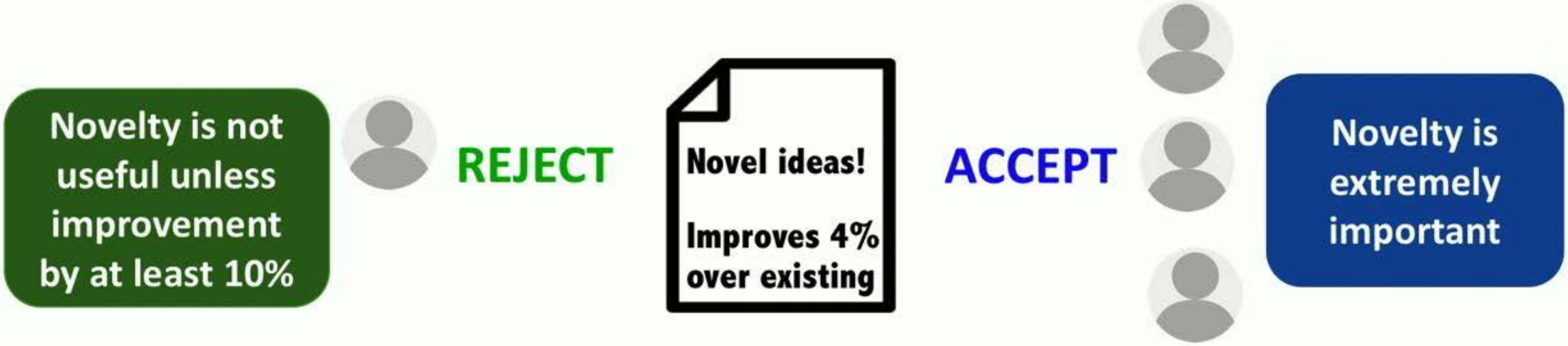
Novelty is not useful unless improvement by at least 10%



Novelty is extremely important

Differing opinions about relative importance of criteria

[Kerr et al. 1977, Mahoney 1977, Bakanic et al. 1987, Hojat et al. 2003, Church 2005, Lamont 2009]



Differing opinions about relative importance of criteria

[Kerr et al. 1977, Mahoney 1977, Bakanic et al. 1987, Hojat et al. 2003, Church 2005, Lamont 2009]

Novelty is not useful unless improvement by at least 10%

REJECT

Novel ideas!

Improves 4% over existing

ACCEPT

Novelty is extremely important

How to ensure that every paper is judged by the same yardstick?

Problem setting

- Reviewers asked to judge papers on **k criteria**
  - E.g. (IJCAI 17): Originality, Relevance, Significance, Writing, Technical
- And an **overall score**
- Reviewer i gives to paper j:
  - Criteria scores $x_{ij} \in [0,1]^k$
  - Overall score $y_{ij} \in [0,1]$

Problem setting

- Reviewers asked to judge papers on **k criteria**
  - E.g. (IJCAI 17): Originality, Relevance, Significance, Writing, Technical
- And an **overall score**
- Reviewer i gives to paper j:
  - Criteria scores $x_{ij} \in [0,1]^k$
  - Overall score $y_{ij} \in [0,1]$
- Each reviewer has a monotonic **(subjective) mapping** from criteria scores in $[0,1]^k$ to overall score in $[0,1]$

Problem setting

- Reviewers asked to judge papers on **k criteria**
  - E.g. (IJCAI 17): Originality, Relevance, Significance, Writing, Technical
- And an **overall score**
- Reviewer i gives to paper j:
  - Criteria scores $x_{ij} \in [0,1]^k$
  - Overall score $y_{ij} \in [0,1]$
- Each reviewer has a monotonic **(subjective) mapping** from criteria scores in $[0,1]^k$ to overall score in $[0,1]$

Need a common mapping for all papers

Data-driven approach: Learn a mapping

Data-driven approach: Learn a mapping

- Obtain $\left(x_{ij}, y_{ij}\right) \in \ [0,1]^k \times [0,1]$ for every review (i, j)

Data-driven approach: Learn a mapping

- Obtain $\left(x_{ij}, y_{ij}\right) \in \ [0,1]^k \times [0,1]$ for every review (i, j)
- Learn a mapping $\hat{f}: [0,1]^k \rightarrow [0,1]$ from this data

Data-driven approach: Learn a mapping

- Obtain $(x_{ij}, y_{ij}) \in [0,1]^k \times [0,1]$ for every review (i,j)
- Learn a mapping $\hat{f}: [0,1]^k \rightarrow [0,1]$ from this data
- For every (i,j), replace overall score y_{ij} with $\hat{f}(x_{ij})$

Data-driven approach: Learn a mapping

- Obtain $(x_{ij}, y_{ij}) \in [0,1]^k \times [0,1]$ for every review (i, j)
- Learn a mapping $\hat{f}: [0,1]^k \rightarrow [0,1]$ from this data
- For every (i, j), replace overall score y_{ij} with $\hat{f}(x_{ij})$

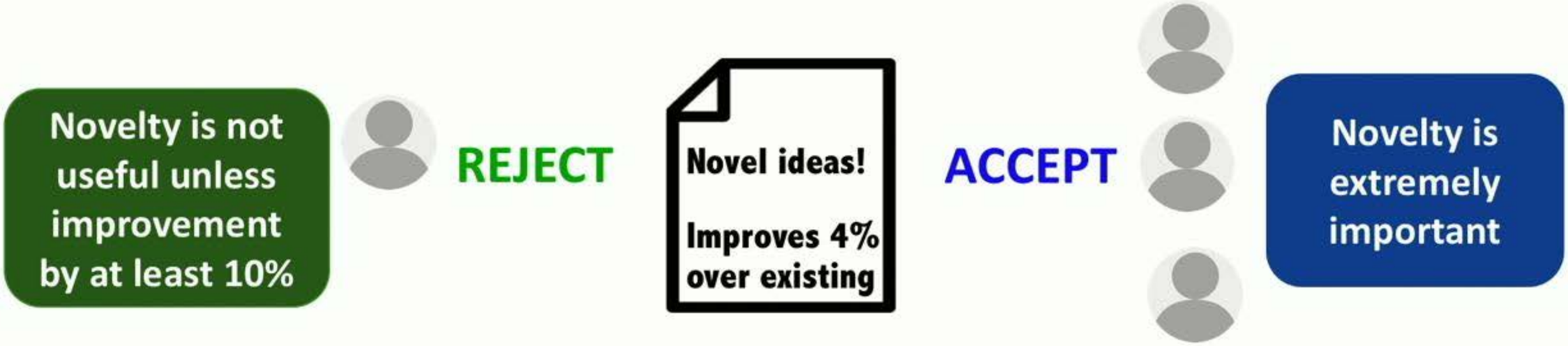
Problem setting

- Reviewers asked to judge papers on **k criteria**
  - E.g. (IJCAI 17): Originality, Relevance, Significance, Writing, Technical
- And an **overall score**
- Reviewer i gives to paper j:
  - Criteria scores $x_{ij} \in [0,1]^k$
  - Overall score $y_{ij} \in [0,1]$
- Each reviewer has a monotonic **(subjective) mapping** from criteria scores in $[0,1]^k$ to overall score in $[0,1]$

Need a common mapping for all papers

Differing opinions about relative importance of criteria

[Kerr et al. 1977, Mahoney 1977, Bakanic et al. 1987, Hojat et al. 2003, Church 2005, Lamont 2009]



Framework

Framework

For this talk: Suppose all papers reviewed by all reviewers

Framework

For this talk: Suppose all papers reviewed by all reviewers

Papers

Reviewers

Criteria scores

Papers

Reviewers

Overall scores

Framework

For this talk: Suppose all papers reviewed by all reviewers

Papers

Reviewers

$$\begin{pmatrix} [.8\ .9\ .9] & [.2\ .3\ .1] & [.8\ .6\ .1] \\ [.9\ .1\ .4] & [.2\ .7\ .4] & [.9\ .6\ .1] \\ [.4\ .1\ .4] & [.1\ .3\ .2] & [.8\ .9\ .2] \\ [.9\ .5\ .4] & [.2\ .3\ .1] & [.7\ .8\ .2] \\ [.3\ .2\ .1] & [.4\ .7\ .9] & [.8\ .9\ .3] \end{pmatrix}$$

Criteria scores

Papers

Reviewers

$$\begin{pmatrix} \quad & \quad & \quad \end{pmatrix}$$

Overall scores

Framework

For this talk: Suppose all papers reviewed by all reviewers

Papers

Reviewers
$$\begin{pmatrix} [.8\ .9\ .9] & [.2\ .3\ .1] & [.8\ .6\ .1] \\ [.9\ .1\ .4] & [.2\ .7\ .4] & [.9\ .6\ .1] \\ [.4\ .1\ .4] & [.1\ .3\ .2] & [.8\ .9\ .2] \\ [.9\ .5\ .4] & [.2\ .3\ .1] & [.7\ .8\ .2] \\ [.3\ .2\ .1] & [.4\ .7\ .9] & [.8\ .9\ .3] \end{pmatrix}$$

Criteria scores

Papers

Reviewers
$$\begin{pmatrix} .9 & .4 & .6 \\ .2 & .4 & .7 \\ .4 & .6 & .6 \\ .4 & .6 & .9 \\ .2 & .3 & .8 \end{pmatrix}$$

Overall scores

Framework

For this talk: Suppose all papers reviewed by all reviewers

$$\begin{pmatrix} f([.8\ .9\ .9]) & f([.2\ .3\ .1]) & f([.8\ .6\ .1]) \\ f([.9\ .1\ .4]) & f([.2\ .7\ .4]) & f([.9\ .6\ .1]) \\ f([.4\ .1\ .4]) & f([.1\ .3\ .2]) & f([.8\ .9\ .2]) \\ f([.9\ .5\ .4]) & f([.2\ .3\ .1]) & f([.7\ .8\ .2]) \\ f([.3\ .2\ .1]) & f([.4\ .7\ .9]) & f([.8\ .9\ .3]) \end{pmatrix} = \begin{pmatrix} .9 & .4 & .6 \\ .2 & .4 & .7 \\ .4 & .6 & .6 \\ .4 & .6 & .9 \\ .2 & .3 & .8 \end{pmatrix}$$

Framework

For this talk: Suppose all papers reviewed by all reviewers

$$\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \underbrace{\left\| \begin{pmatrix} f([.8\ .9\ .9]) & f([.2\ .3\ .1]) & f([.8\ .6\ .1]) \\ f([.9\ .1\ .4]) & f([.2\ .7\ .4]) & f([.9\ .6\ .1]) \\ f([.4\ .1\ .4]) & f([.1\ .3\ .2]) & f([.8\ .9\ .2]) \\ f([.9\ .5\ .4]) & f([.2\ .3\ .1]) & f([.7\ .8\ .2]) \\ f([.3\ .2\ .1]) & f([.4\ .7\ .9]) & f([.8\ .9\ .3]) \end{pmatrix} - \begin{pmatrix} .9 & .4 & .6 \\ .2 & .4 & .7 \\ .4 & .6 & .6 \\ .4 & .6 & .9 \\ .2 & .3 & .8 \end{pmatrix} \right\|_{p,q}}_{\text{L(p,q) loss}}$$

Framework

For this talk: Suppose all papers reviewed by all reviewers

$$\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \underbrace{\left\| \begin{pmatrix} f([.8\ .9\ .9]) & f([.2\ .3\ .1]) & f([.8\ .6\ .1]) \\ f([.9\ .1\ .4]) & f([.2\ .7\ .4]) & f([.9\ .6\ .1]) \\ f([.4\ .1\ .4]) & f([.1\ .3\ .2]) & f([.8\ .9\ .2]) \\ f([.9\ .5\ .4]) & f([.2\ .3\ .1]) & f([.7\ .8\ .2]) \\ f([.3\ .2\ .1]) & f([.4\ .7\ .9]) & f([.8\ .9\ .3]) \end{pmatrix} - \begin{pmatrix} .9 & .4 & .6 \\ .2 & .4 & .7 \\ .4 & .6 & .6 \\ .4 & .6 & .9 \\ .2 & .3 & .8 \end{pmatrix} \right\|_{p,q}}_{\text{L(p,q) loss}}$$

$\mathcal{F} = set\ of\ all\ monotonic\ functions$

Choice of loss function

- $p \in [1, \infty], \quad q \in [1, \infty]$

Choice of loss function

- $p \in [1, \infty], \quad q \in [1, \infty]$

$$\begin{pmatrix} f([.8\ .9\ .9]) & f([.2\ .3\ .1]) & f([.8\ .6\ .1]) \\ f([.9\ .1\ .4]) & f([.2\ .7\ .4]) & f([.9\ .6\ .1]) \\ f([.4\ .1\ .4]) & f([.1\ .3\ .2]) & f([.8\ .9\ .2]) \\ f([.9\ .5\ .4]) & f([.2\ .3\ .1]) & f([.7\ .8\ .2]) \\ f([.3\ .2\ .1]) & f([.4\ .7\ .9]) & f([.8\ .9\ .3]) \end{pmatrix} - \begin{pmatrix} .9 & .4 & .6 \\ .2 & .4 & .7 \\ .4 & .6 & .6 \\ .4 & .6 & .9 \\ .2 & .3 & .8 \end{pmatrix}$$

Choice of loss function

- $p \in [1, \infty], \quad q \in [1, \infty]$

$$\begin{pmatrix} f([.8\ .9\ .9]) & f([.2\ .3\ .1]) & f([.8\ .6\ .1]) \\ f([.9\ .1\ .4]) & f([.2\ .7\ .4]) & f([.9\ .6\ .1]) \\ f([.4\ .1\ .4]) & f([.1\ .3\ .2]) & f([.8\ .9\ .2]) \\ f([.9\ .5\ .4]) & f([.2\ .3\ .1]) & f([.7\ .8\ .2]) \\ f([.3\ .2\ .1]) & f([.4\ .7\ .9]) & f([.8\ .9\ .3]) \end{pmatrix} - \begin{pmatrix} .9 & .4 & .6 \\ .2 & .4 & .7 \\ .4 & .6 & .6 \\ .4 & .6 & .9 \\ .2 & .3 & .8 \end{pmatrix}$$

L_p norm *

Choice of loss function

- $p \in [1, \infty], \quad q \in [1, \infty]$

$$
\begin{pmatrix}
f([.8\ .9\ .9]) & f([.2\ .3\ .1]) & f([.8\ .6\ .1]) \\
f([.9\ .1\ .4]) & f([.2\ .7\ .4]) & f([.9\ .6\ .1]) \\
f([.4\ .1\ .4]) & f([.1\ .3\ .2]) & f([.8\ .9\ .2]) \\
f([.9\ .5\ .4]) & f([.2\ .3\ .1]) & f([.7\ .8\ .2]) \\
f([.3\ .2\ .1]) & f([.4\ .7\ .9]) & f([.8\ .9\ .3])
\end{pmatrix}
-
\begin{pmatrix}
.9 & .4 & .6 \\
.2 & .4 & .7 \\
.4 & .6 & .6 \\
.4 & .6 & .9 \\
.2 & .3 & .8
\end{pmatrix}
$$

L_p norm *

L_p norm *

Choice of loss function

- $p \in [1, \infty], \quad q \in [1, \infty]$

$$
\begin{pmatrix}
f([.8\ .9\ .9]) & f([.2\ .3\ .1]) & f([.8\ .6\ .1]) \\
f([.9\ .1\ .4]) & f([.2\ .7\ .4]) & f([.9\ .6\ .1]) \\
f([.4\ .1\ .4]) & f([.1\ .3\ .2]) & f([.8\ .9\ .2]) \\
f([.9\ .5\ .4]) & f([.2\ .3\ .1]) & f([.7\ .8\ .2]) \\
f([.3\ .2\ .1]) & f([.4\ .7\ .9]) & f([.8\ .9\ .3])
\end{pmatrix}
-
\begin{pmatrix}
.9 & .4 & .6 \\
.2 & .4 & .7 \\
.4 & .6 & .6 \\
.4 & .6 & .9 \\
.2 & .3 & .8
\end{pmatrix}
\begin{matrix}
L_p \text{ norm} \quad * \\
L_p \text{ norm} \quad * \\
L_p \text{ norm} \quad * \\
L_p \text{ norm} \quad * \\
L_p \text{ norm} \quad *
\end{matrix}
$$

Choice of loss function

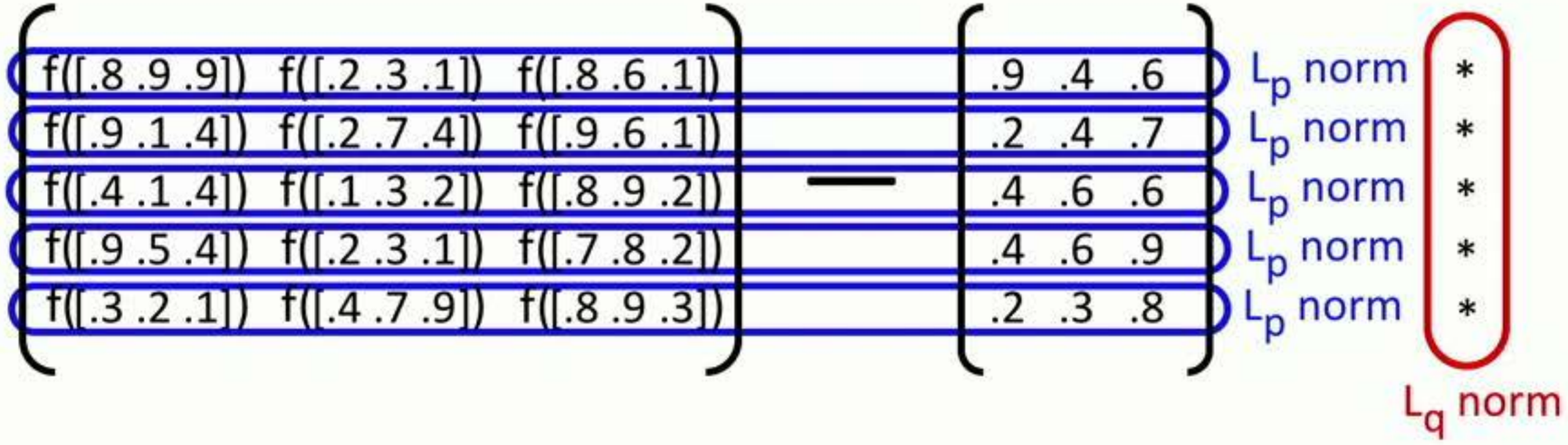
- $p \in [1, \infty], \quad q \in [1, \infty]$

$$
\begin{pmatrix}
f([.8\ .9\ .9]) & f([.2\ .3\ .1]) & f([.8\ .6\ .1]) \\
f([.9\ .1\ .4]) & f([.2\ .7\ .4]) & f([.9\ .6\ .1]) \\
f([.4\ .1\ .4]) & f([.1\ .3\ .2]) & f([.8\ .9\ .2]) \\
f([.9\ .5\ .4]) & f([.2\ .3\ .1]) & f([.7\ .8\ .2]) \\
f([.3\ .2\ .1]) & f([.4\ .7\ .9]) & f([.8\ .9\ .3])
\end{pmatrix}
-
\begin{pmatrix}
.9 & .4 & .6 \\
.2 & .4 & .7 \\
.4 & .6 & .6 \\
.4 & .6 & .9 \\
.2 & .3 & .8
\end{pmatrix}
\begin{matrix}
L_p \text{ norm} \\
L_p \text{ norm} \\
L_p \text{ norm} \\
L_p \text{ norm} \\
L_p \text{ norm}
\end{matrix}
\begin{pmatrix}
* \\ * \\ * \\ * \\ *
\end{pmatrix}
$$

L_q norm

Choice of loss function

- $p \in [1, \infty], \quad q \in [1, \infty]$



- Used in many applications

  [e.g., Ding et al. 2006, He and Cichocki 2008, Nie et al. 2010, Kong et al. 2011, Rahimpour et al. 2017...]

- Different L(p,q) losses used in different applications

Choice of loss function

- $p \in [1, \infty], \quad q \in [1, \infty]$

$$
\begin{pmatrix}
f([.8\ .9\ .9]) & f([.2\ .3\ .1]) & f([.8\ .6\ .1]) \\
f([.9\ .1\ .4]) & f([.2\ .7\ .4]) & f([.9\ .6\ .1]) \\
f([.4\ .1\ .4]) & f([.1\ .3\ .2]) & f([.8\ .9\ .2]) \\
f([.9\ .5\ .4]) & f([.2\ .3\ .1]) & f([.7\ .8\ .2]) \\
f([.3\ .2\ .1]) & f([.4\ .7\ .9]) & f([.8\ .9\ .3])
\end{pmatrix}
-
\begin{pmatrix}
.9 & .4 & .6 \\
.2 & .4 & .7 \\
.4 & .6 & .6 \\
.4 & .6 & .9 \\
.2 & .3 & .8
\end{pmatrix}
\begin{matrix}
L_p \text{ norm} \\
L_p \text{ norm} \\
L_p \text{ norm} \\
L_p \text{ norm} \\
L_p \text{ norm}
\end{matrix}
\begin{matrix}
* \\ * \\ * \\ * \\ *
\end{matrix}
$$

L_q norm

- Used in many applications

  [e.g., Ding et al. 2006, He and Cichocki 2008, Nie et al. 2010, Kong et al. 2011, Rahimpour et al. 2017...]

- Different L(p,q) losses used in different applications

Which L(p,q) loss function to use?

Axiomatic approach

- Approach is popular in economics and social choice theory
- Identify scenarios that is easy to reason about
- Establish necessary conditions (or “axioms”)

Axiomatic approach

- Approach is popular in economics and social choice theory
- Identify scenarios that is easy to reason about
- Establish necessary conditions (or "axioms")

Any paper gets the same criteria scores from all reviewers.

Paper 1: $x_{11} = x_{21} = x_{31} = \cdots := x_1$
Paper 2: $x_{12} = x_{22} = x_{32} = \cdots := x_2$
$\vdots$

Axiomatic approach

- Approach is popular in economics and social choice theory
- Identify scenarios that is easy to reason about
- Establish necessary conditions (or "axioms")

Any paper gets the same criteria scores from all reviewers.

Paper 1: $x_{11} = x_{21} = x_{31} = \cdots := x_1$
Paper 2: $x_{12} = x_{22} = x_{32} = \cdots := x_2$
$\vdots$

Three natural axioms

Axiom 1: Consensus

For some $x \in [0,1]^k$ and $y \in [0,1]$, if all reviewers map x to y then $\hat{f}(x) = y$.

Axiom 1: Consensus

For some $x \in [0,1]^k$ and $y \in [0,1]$, if all reviewers map x to y then $\hat{f}(x) = y$.

Axiom 2: Dominance

For any papers a and b, if the vector of overall scores received by paper a in sorted order is pointwise $\geq$ the corresponding vector for paper b, then $\hat{f}(x_a) \geq \hat{f}(x_b)$.

Axiom 1: Consensus

For some $x \in [0,1]^k$ and $y \in [0,1]$, if all reviewers map x to y then $\hat{f}(x) = y$.

Axiom 2: Dominance

For any papers a and b, if the vector of overall scores received by paper a in sorted order is pointwise $\geq$ the corresponding vector for paper b, then $\hat{f}(x_a) \geq \hat{f}(x_b)$.

Axiom 3: Strategyproofness

No reviewer can bring the learnt overall scores closer to her/his own opinion by strategic manipulation. For any reviewer i, let $(y_{i1}, \dots, y_{im})$ be overall scores she/he gives if honest. Let $\hat{f}$ denote learnt mapping in that case. Let $(y'_1, \dots, y'_m)$ be any other overall scores and $\hat{g}$ be the associated learnt mapping. Then we need:

$$\left\| \left(\hat{f}(x_1), \dots, \hat{f}(x_m) \right) - (y_{i1}, \dots, y_{im}) \right\| \leq \left\| \left(\hat{g}(x_1), \dots, \hat{g}(x_m) \right) - (y_{i1}, \dots, y_{im}) \right\|$$

Theorem

L(1,1) is the only L(p,q) loss that satisfies the three axioms.

Theorem

L(1,1) is the only L(p,q) loss that satisfies the three axioms.

- Strategyproofness violated when $q \in (1, \infty]$
- Consensus violated when $p = \infty$ and $q = 1$
- Dominance violated when $p \in (1, \infty)$ and $q = 1$

Theorem

L(1,1) is the only L(p,q) loss that satisfies the three axioms.

- Strategyproofness violated when $q \in (1, \infty]$
- Consensus violated when $p = \infty$ and $q = 1$
- Dominance violated when $p \in (1, \infty)$ and $q = 1$

Paradoxical!

Dominance violated under L(2,1) loss

Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

- Overall scores:

	Paper 1	Paper 2
Rev. 1	0	0
Rev. 2	1	0
Rev. 3	0	z < 1

Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

- Overall scores:

	Paper 1	Paper 2
Rev. 1	0	0
Rev. 2	1	0
Rev. 3	0	z < 1

Paper 1 dominates paper 2

Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

- Overall scores:

	Paper 1	Paper 2
Rev. 1	0	0
Rev. 2	1	0
Rev. 3	0	z < 1

Paper 1 dominates paper 2

$$\boxed{\text{want } \hat{f}(x_1) \geq \hat{f}(x_2)}$$

Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

- Overall scores:

	Paper 1	Paper 2
Rev. 1	0	0
Rev. 2	1	0
Rev. 3	0	z < 1

Paper 1 dominates paper 2

$$\boxed{\text{want } \hat{f}(x_1) \geq \hat{f}(x_2)}$$

Fermat point of a triangle: Point with smallest total Euclidean distance from the 3 vertices

Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

- Overall scores:

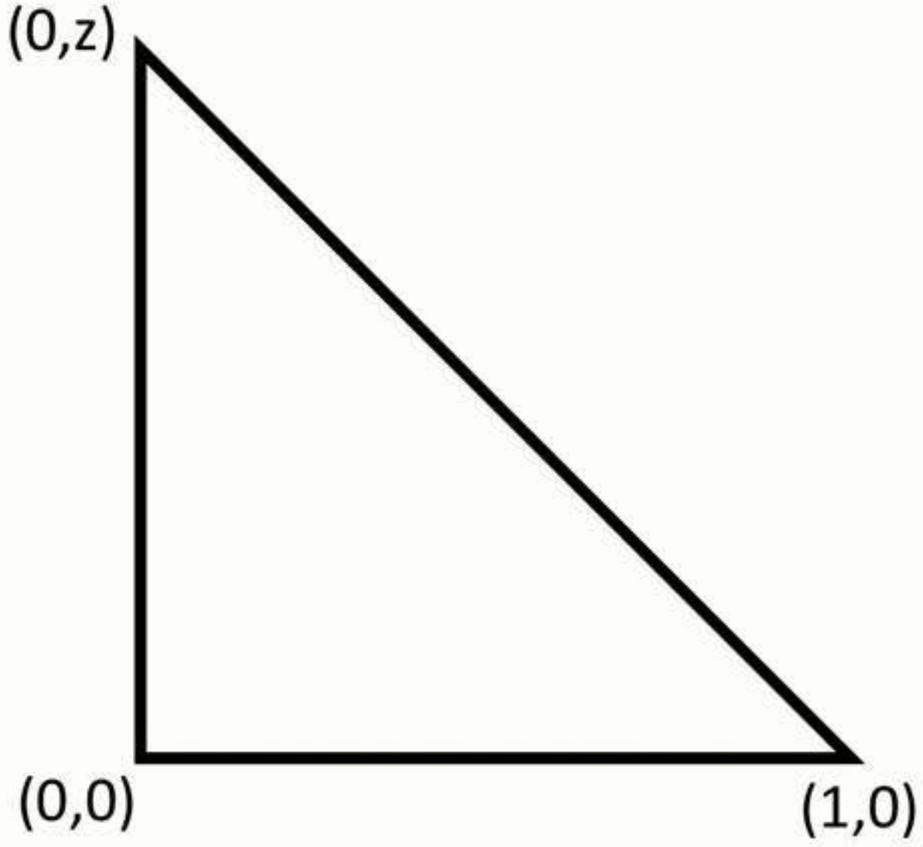
	Paper 1	Paper 2
Rev. 1	0	0
Rev. 2	1	0
Rev. 3	0	z < 1

Paper 1 dominates paper 2

$$\boxed{\text{want } \hat{f}(x_1) \geq \hat{f}(x_2)}$$

Fermat point of a triangle: Point with smallest total Euclidean distance from the 3 vertices

$\left(\hat{f}(x_1), \hat{f}(x_2)\right)$ is exactly the Fermat point of:



Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

- Overall scores:

	Paper 1	Paper 2
Rev. 1	0	0
Rev. 2	1	0
Rev. 3	0	z < 1

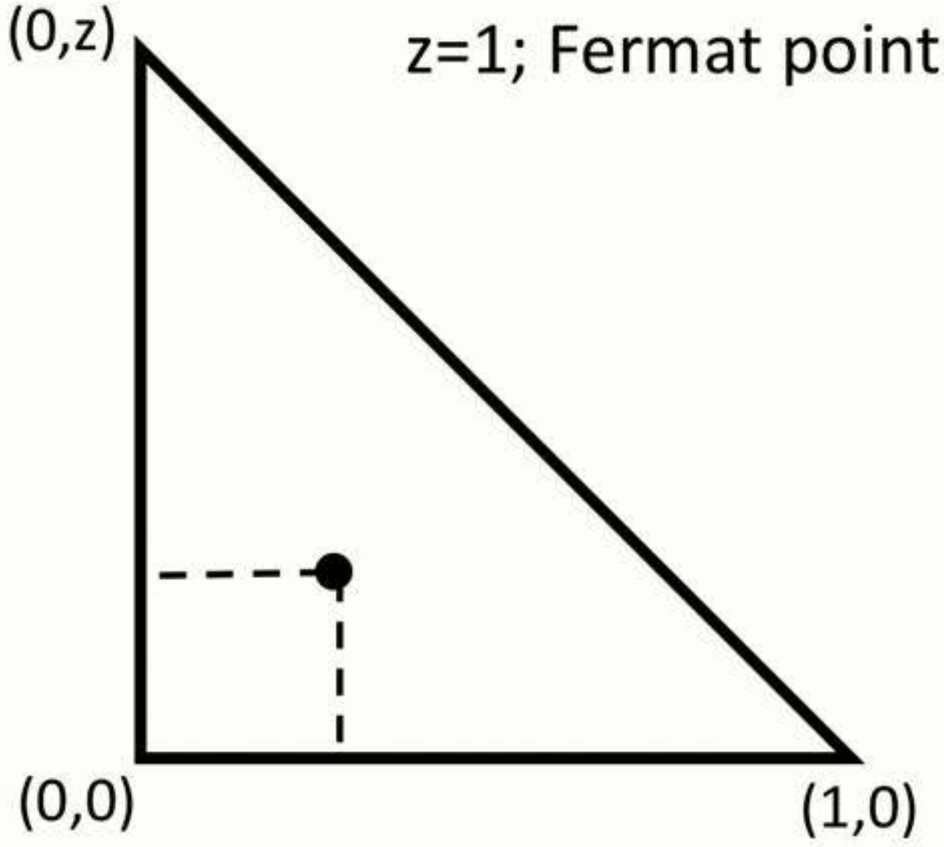
Paper 1 dominates paper 2

$$\boxed{\text{want } \hat{f}(x_1) \geq \hat{f}(x_2)}$$

Fermat point of a triangle: Point with smallest total Euclidean distance from the 3 vertices

$\left(\hat{f}(x_1), \hat{f}(x_2)\right)$ is exactly the Fermat point of:

z=1; Fermat point: (.20,.20)



Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

- Overall scores:

	Paper 1	Paper 2
Rev. 1	0	0
Rev. 2	1	0
Rev. 3	0	z < 1

Paper 1 dominates paper 2

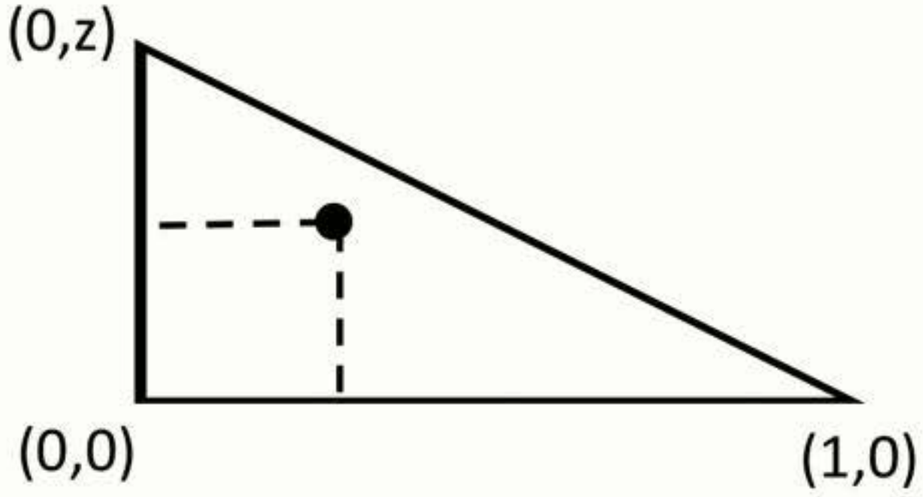
$$\boxed{\text{want } \hat{f}(x_1) \geq \hat{f}(x_2)}$$

Fermat point of a triangle: Point with smallest total Euclidean distance from the 3 vertices

$\left(\hat{f}(x_1), \hat{f}(x_2)\right)$ is exactly the Fermat point of:

z=1; Fermat point: (.20,.20)

z=½



Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

- Overall scores:

	Paper 1	Paper 2
Rev. 1	0	0
Rev. 2	1	0
Rev. 3	0	z < 1

Paper 1 dominates paper 2

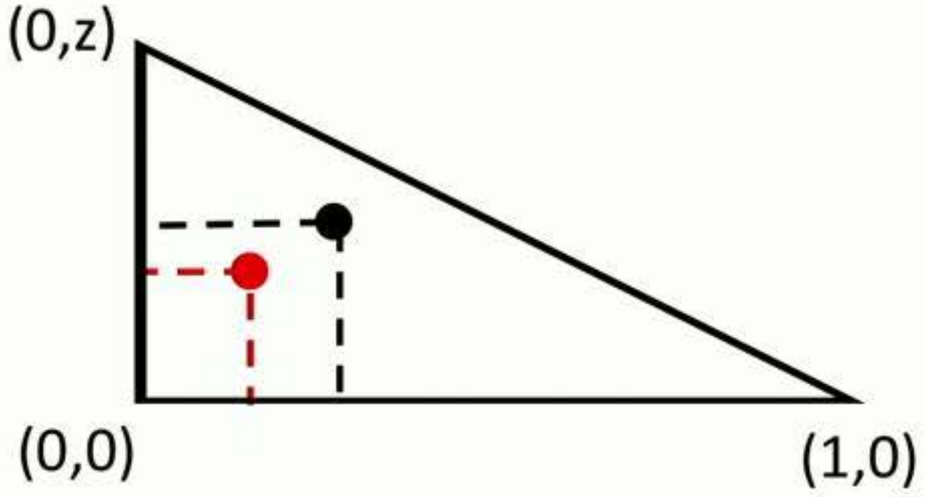
$$\boxed{\text{want } \hat{f}(x_1) \geq \hat{f}(x_2)}$$

Fermat point of a triangle: Point with smallest total Euclidean distance from the 3 vertices

$\left(\hat{f}(x_1), \hat{f}(x_2)\right)$ is exactly the Fermat point of:

z=1; Fermat point: (.20,.20)

z=½; Fermat point: (.12,.15)



Dominance violated under L(2,1) loss

- 2 papers, 3 reviewers, k=2 criteria
- Criteria scores x_1 = [¼, ¾], x_2 = [¾, ¼]

- Overall scores:

	Paper 1	Paper 2
Rev. 1	0	0
Rev. 2	1	0
Rev. 3	0	z < 1

Paper 1 dominates paper 2

want $\hat{f}(x_1) \geq \hat{f}(x_2)$

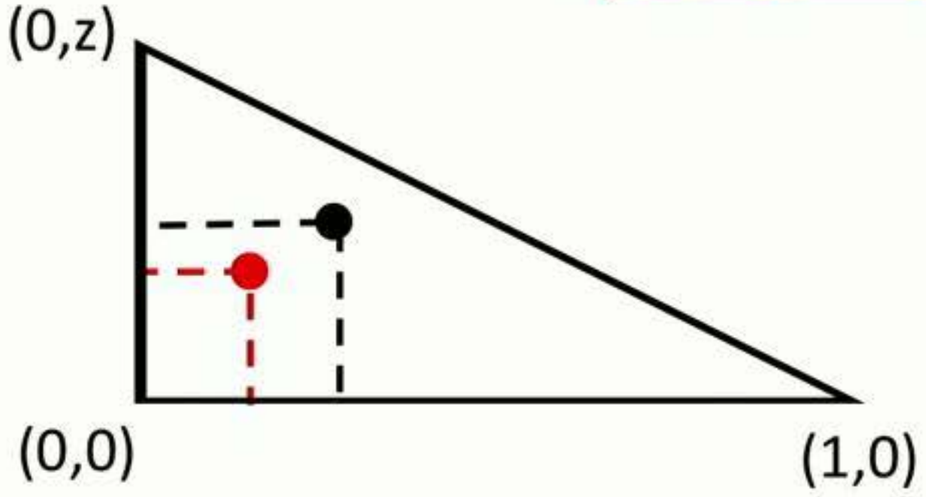
Fermat point of a triangle: Point with smallest total Euclidean distance from the 3 vertices

$\left(\hat{f}(x_1), \hat{f}(x_2)\right)$ is exactly the Fermat point of:

z=1; Fermat point: (.20,.20)

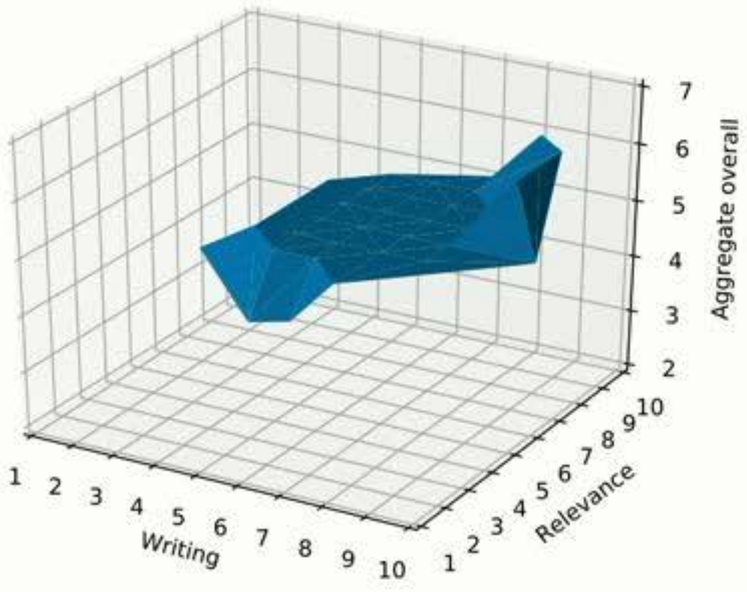
z=½; Fermat point: (.12,.15)

$\hat{f}(x_1) < \hat{f}(x_2)$



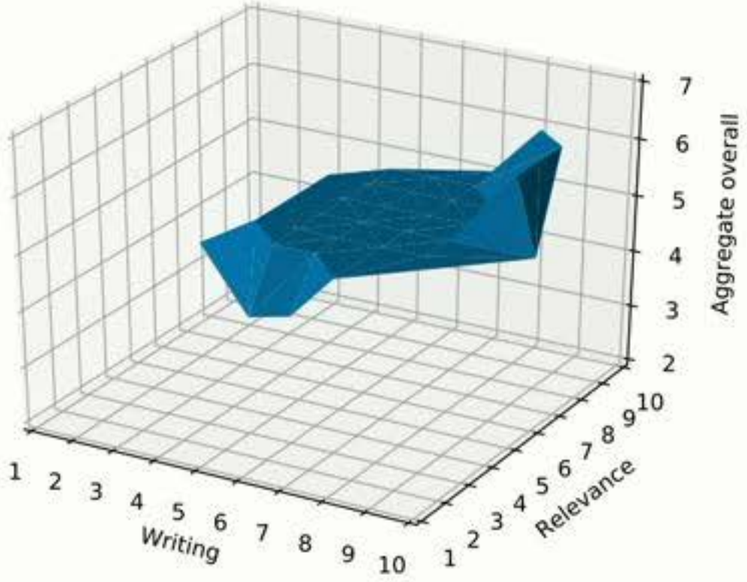
IJCAI 2017

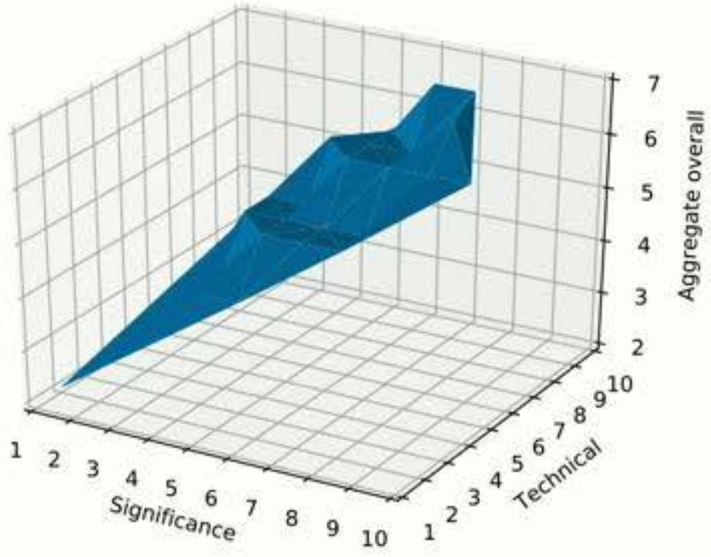
IJCAI 2017



- Writing and Relevance: Really bad - significant downside, really good - appreciated, in between - irrelevant.

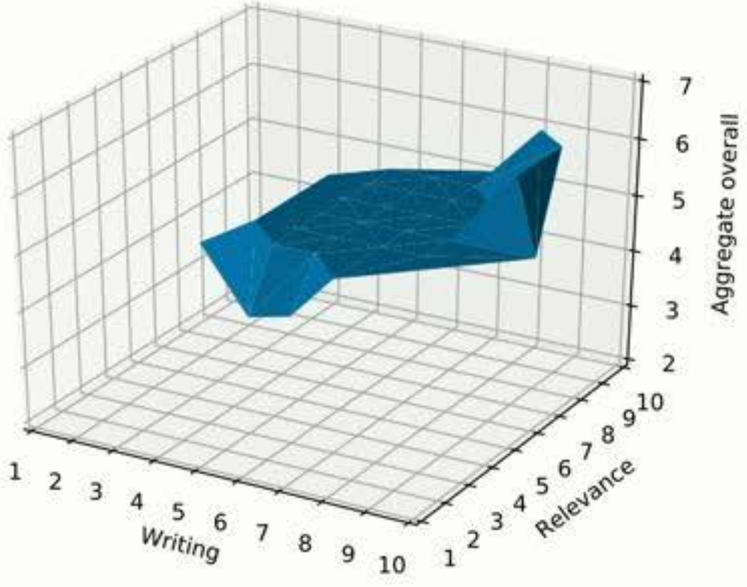
IJCAI 2017

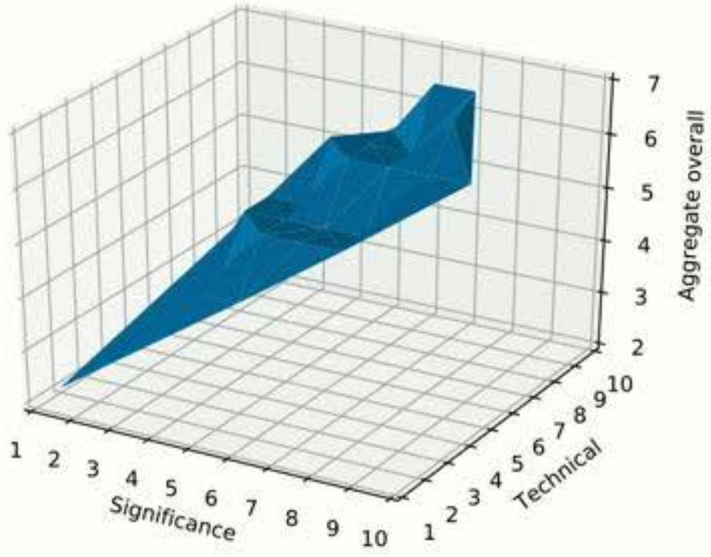


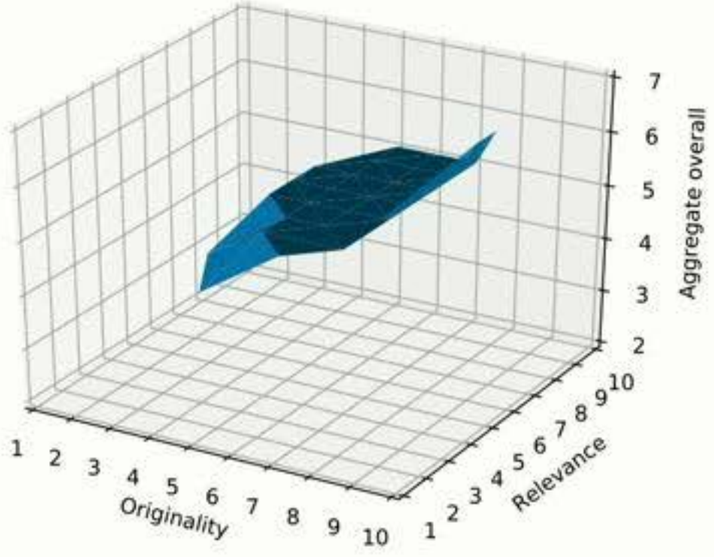


- Writing and Relevance: Really bad - significant downside, really good - appreciated, in between - irrelevant.
- Technical quality and Significance: high influence; the influence is approximately linear.

IJCAI 2017







- Writing and Relevance: Really bad - significant downside, really good - appreciated, in between - irrelevant.
- Technical quality and Significance: high influence; the influence is approximately linear.
- Originality: moderate influence.

Miscalibration

with

Jingyan Wang

Best Student Paper Award at AAMAS 2019
Best Paper Nominee

Miscalibration in ratings

Miscalibration in ratings

Mitliagkas et al. 2011

"A raw rating of 7 out of 10 in the absence of any other information is potentially useless."

Miscalibration in ratings

Mitliagkas et al. 2011
"A raw rating of 7 out of 10 in the absence of any other information is potentially useless."

Ammar et al. 2012
"The rating scale as well as the individual ratings are often arbitrary and may not be consistent from one user to another."

Miscalibration in ratings

Mitliagkas et al. 2011
"A raw rating of 7 out of 10 in the absence of any other information is potentially useless."

Ammar et al. 2012
"The rating scale as well as the individual ratings are often arbitrary and may not be consistent from one user to another."

Freund et al. 2003
"[Using rankings instead of ratings] becomes very important when we combine the rankings of many viewers who often use completely different ranges of scores to express identical preferences."

Two approaches in the literature

1 **Assume simplified models for calibration**

[Paul 1981, Flach et al. 2010, Roos et al. 2011, Baba and Kashima 2013, Ge et al. 2013, Mackay et al. 2017]

- Did not work well – NIPS 2016 program chairs.
- Langford (ICML 2012 program co-chair): *"We experimented with reviewer normalization and generally found it significantly harmful."*

Two approaches in the literature

1 **Assume simplified models for calibration**

[Paul 1981, Flach et al. 2010, Roos et al. 2011, Baba and Kashima 2013, Ge et al. 2013, Mackay et al. 2017]

- Did not work well – NIPS 2016 program chairs.
- Langford (ICML 2012 program co-chair): *"We experimented with reviewer normalization and generally found it significantly harmful."*

2 **Use rankings**

[Rokeach 1968, Freund et al. 2003, Harzing et al. 2009, Mitliagkas et al. 2011, Ammar et al. 2012, Negahban et al. 2012]

- Use rankings induced by ratings or directly collect rankings
- Commonly believed to be the best option if no assumptions on calibration

Two approaches in the literature

1 **Assume simplified models for calibration**

[Paul 1981, Flach et al. 2010, Roos et al. 2011, Baba and Kashima 2013, Ge et al. 2013, Mackay et al. 2017]

- Did not work well – NIPS 2016 program chairs.
- Langford (ICML 2012 program co-chair): *"We experimented with reviewer normalization and generally found it significantly harmful."*

2 **Use rankings**

[Rokeach 1968, Freund et al. 2003, Harzing et al. 2009, Mitliagkas et al. 2011, Ammar et al. 2012, Negahban et al. 2012]

- Use rankings induced by ratings or directly collect rankings
- Commonly believed to be the best option if no assumptions on calibration

? **Is it possible to do better than rankings with essentially no assumptions on the calibration?**

Canonical 2x2 setting

Canonical 2x2 setting

$z_A^* \neq z_B^* \in [0,1]$

Canonical 2x2 setting

$z_A^* \neq z_B^* \in [0,1]$

Calibration function: $f_1 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_1(z_i^*)$

Canonical 2x2 setting

$z_A^* \neq z_B^* \in [0,1]$

Calibration function: $f_1 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_1(z_i^*)$

Calibration function $f_2 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_2(z_i^*)$

Canonical 2x2 setting

$z_A^* \neq z_B^* \in [0,1]$

Calibration function: $f_1 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_1(z_i^*)$

Calibration function $f_2 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_2(z_i^*)$

- Adversary chooses z_A^*, z_B^* and strictly monotonic f_1, f_2

Canonical 2x2 setting

$z_A^* \neq z_B^* \in [0,1]$

Calibration function: $f_1 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_1(z_i^*)$

Calibration function $f_2 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_2(z_i^*)$

- Adversary chooses z_A^*, z_B^* and strictly monotonic f_1, f_2
- One paper assigned to each reviewer at random

Canonical 2x2 setting

$z_A^* \neq z_B^* \in [0,1]$

Calibration function: $f_1 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_1(z_i^*)$

Calibration function $f_2 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_2(z_i^*)$

- Adversary chooses z_A^*, z_B^* and strictly monotonic f_1, f_2
- One paper assigned to each reviewer at random
- Let $\boldsymbol{y_j}$ denote score given by reviewer $j \in \{1, 2\}$

Canonical 2x2 setting

$z_A^* \neq z_B^* \in [0,1]$

Calibration function: $f_1 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_1(z_i^*)$

Calibration function $f_2 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_2(z_i^*)$

- Adversary chooses z_A^*, z_B^* and strictly monotonic f_1, f_2
- One paper assigned to each reviewer at random
- Let $\boldsymbol{y_j}$ denote score given by reviewer $j \in \{1, 2\}$
- **Goal: Given (assignment, $\boldsymbol{y_1}$, $\boldsymbol{y_2}$) estimate if $\boldsymbol{z_A^* > z_B^*}$ or $\boldsymbol{z_B^* > z_A^*}$**

Canonical 2x2 setting

$z_A^* \neq z_B^* \in [0,1]$

Calibration function: $f_1 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_1(z_i^*)$

Calibration function $f_2 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_2(z_i^*)$

- Adversary chooses z_A^*, z_B^* and strictly monotonic f_1, f_2
- One paper assigned to each reviewer at random
- Let $\boldsymbol{y_j}$ denote score given by reviewer $j \in \{1, 2\}$
- **Goal: Given (assignment, $\boldsymbol{y_1}$, $\boldsymbol{y_2}$) estimate if $\boldsymbol{z_A^* > z_B^*}$ or $\boldsymbol{z_B^* > z_A^*}$**
  - Eliciting rankings is vacuous

Impossibility

Theorem

No deterministic estimator has a success probability better than random guessing.

Impossibility

Theorem

No deterministic estimator has a success probability better than random guessing.

Is it possible to do better than random guessing?

Inspirations and connections

- Stein's phenomenon
- Empirical Bayes
- Cover's envelope problem

Estimator

With probability $\frac{1 + |y_1 - y_2|}{2}$ pick paper which received higher score

Theorem

The estimator strictly outperforms random guessing.

Estimator

With probability $\frac{1 + |y_1 - y_2|}{2}$ pick paper which received higher score

Theorem

The estimator strictly outperforms random guessing.

- **Ratings > rankings even if calibration is arbitrary/adversarial**
- **Building block for more general applications**

Impossibility

Theorem

No deterministic estimator has a success probability better than random guessing.

Canonical 2x2 setting

$z_A^* \neq z_B^* \in [0,1]$

Calibration function: $f_1 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_1(z_i^*)$

Calibration function $f_2 : [0,1] \rightarrow [0,1]$

Given paper $i \in \{A, B\}$, outputs $f_2(z_i^*)$

- Adversary chooses z_A^*, z_B^* and strictly monotonic f_1, f_2
- One paper assigned to each reviewer at random
- Let $\boldsymbol{y_j}$ denote score given by reviewer $j \in \{1, 2\}$
- **Goal: Given (assignment, $\boldsymbol{y_1}$, $\boldsymbol{y_2}$) estimate if $\boldsymbol{z_A^* > z_B^*}$ or $\boldsymbol{z_B^* > z_A^*}$**
  - Eliciting rankings is vacuous

Impossibility

Theorem

No deterministic estimator has a success probability better than random guessing.

Is it possible to do better than random guessing?

Biases

with

Ivan Stelmakh

Aarti Singh

Single blind versus double blind

- Gender/race/fame/… biases? Lot of debate!
- “Where is the evidence (of bias in my research community)?”

Single blind versus double blind

- Gender/race/fame/… biases? Lot of debate!
- "Where is the evidence (of bias in my research community)?"

How to rigorously test for biases in peer review (while ensuring "good" review process)?

A remarkable experiment!

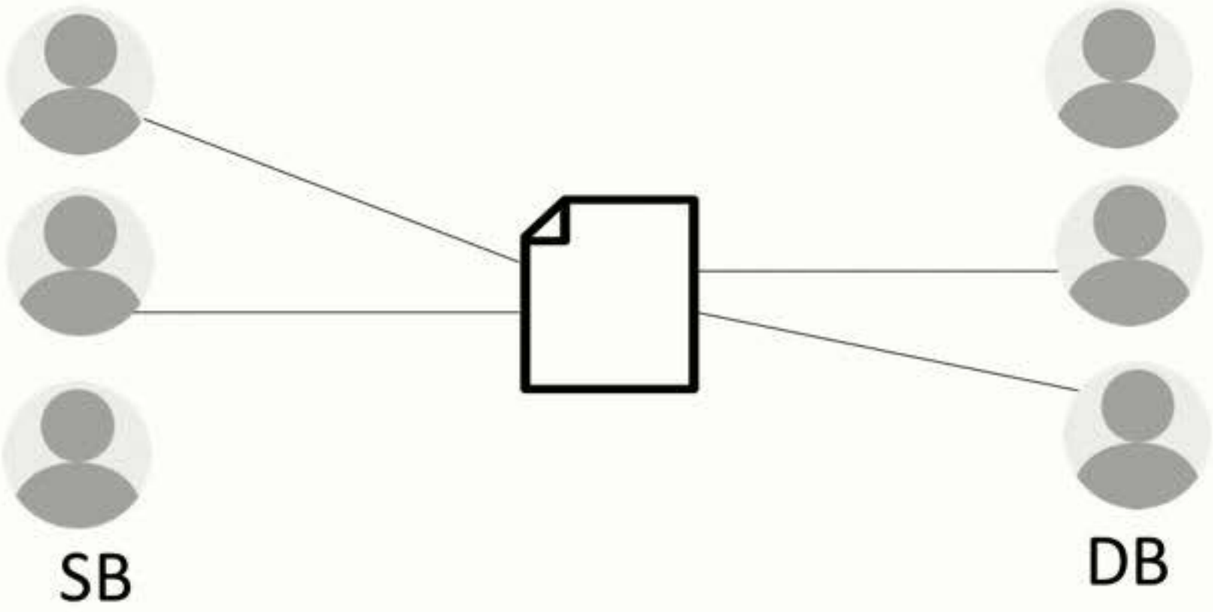
WSDM 2017 (Tomkins, Zhang, Heavlin)

A remarkable experiment!

WSDM 2017 (Tomkins, Zhang, Heavlin)

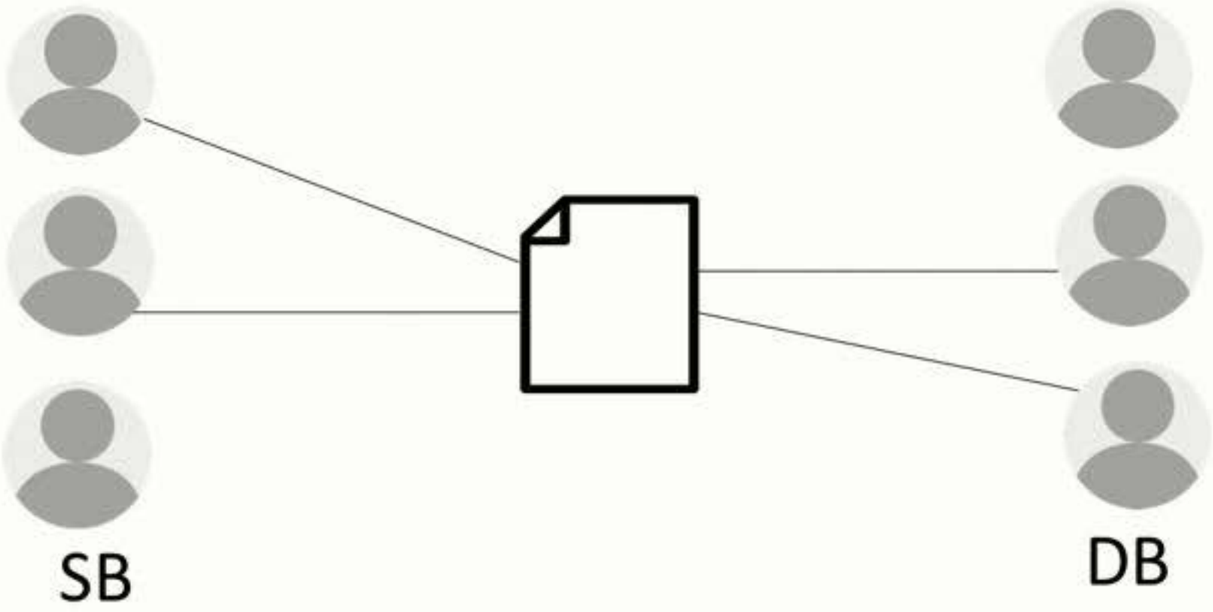
A remarkable experiment!

WSDM 2017 (Tomkins, Zhang, Heavlin)



A remarkable experiment!

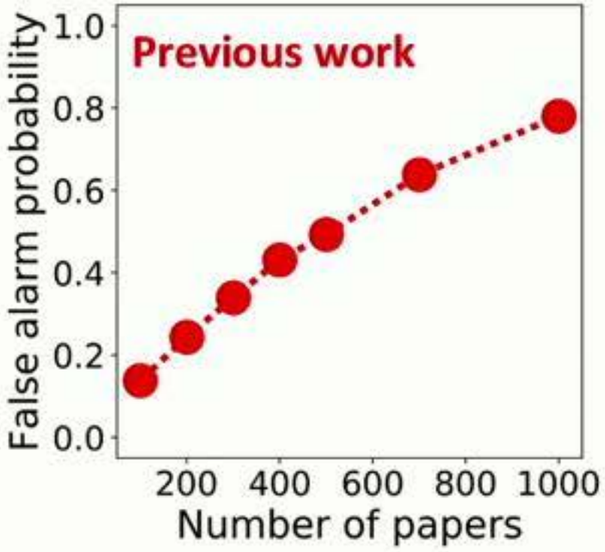
WSDM 2017 (Tomkins, Zhang, Heavlin)



Our negative results: We identify a number of issues in the experimental setup and testing procedure which can lead to spurious (false) positives

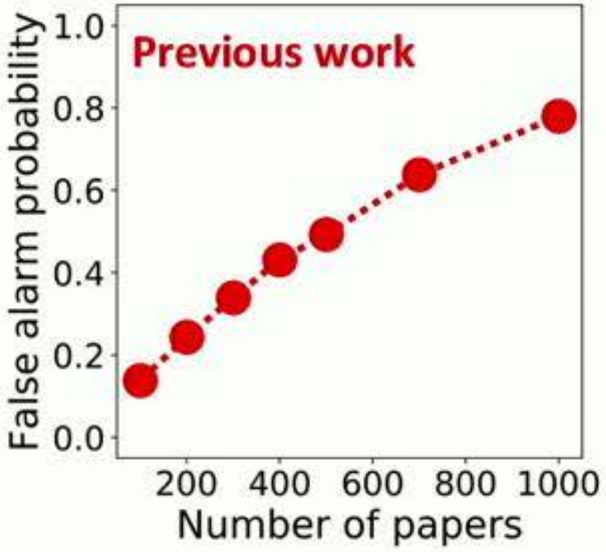
Testing for biases

False alarm probability specified to be ≤ 0.05



Testing for biases

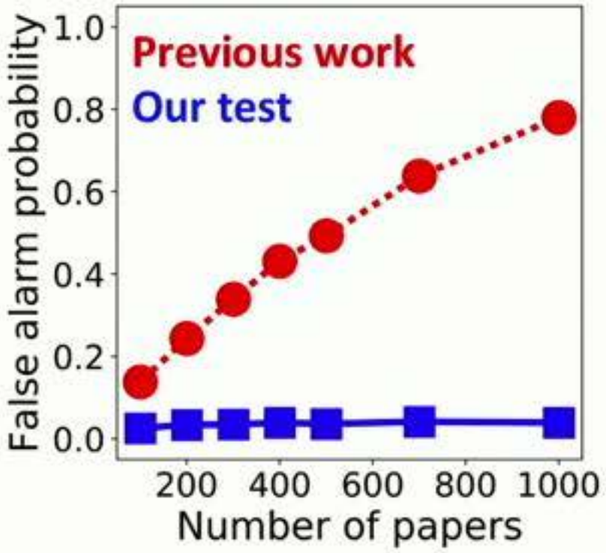
False alarm probability specified to be ≤ 0.05



- **Positive results:**
  - A testing setup (with minimal changes to peer review processes)
  - Statistical tests
  - Strong, rigorous guarantees

Testing for biases

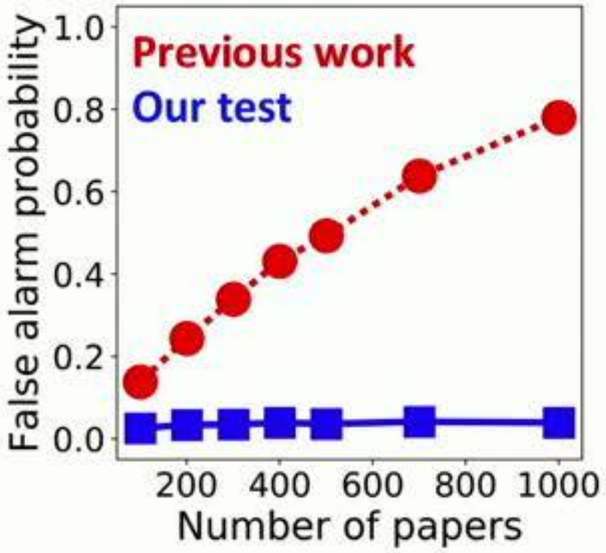
False alarm probability specified to be ≤ 0.05



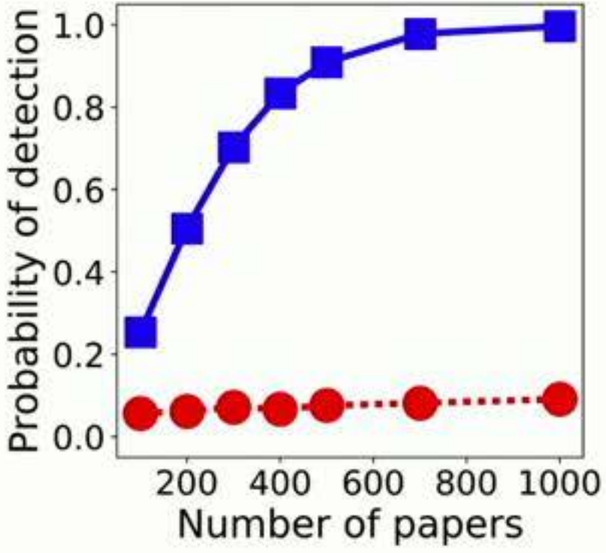
- **Positive results:**
  - A testing setup (with minimal changes to peer review processes)
  - Statistical tests
  - Strong, rigorous guarantees

Testing for biases

False alarm probability specified to be ≤ 0.05

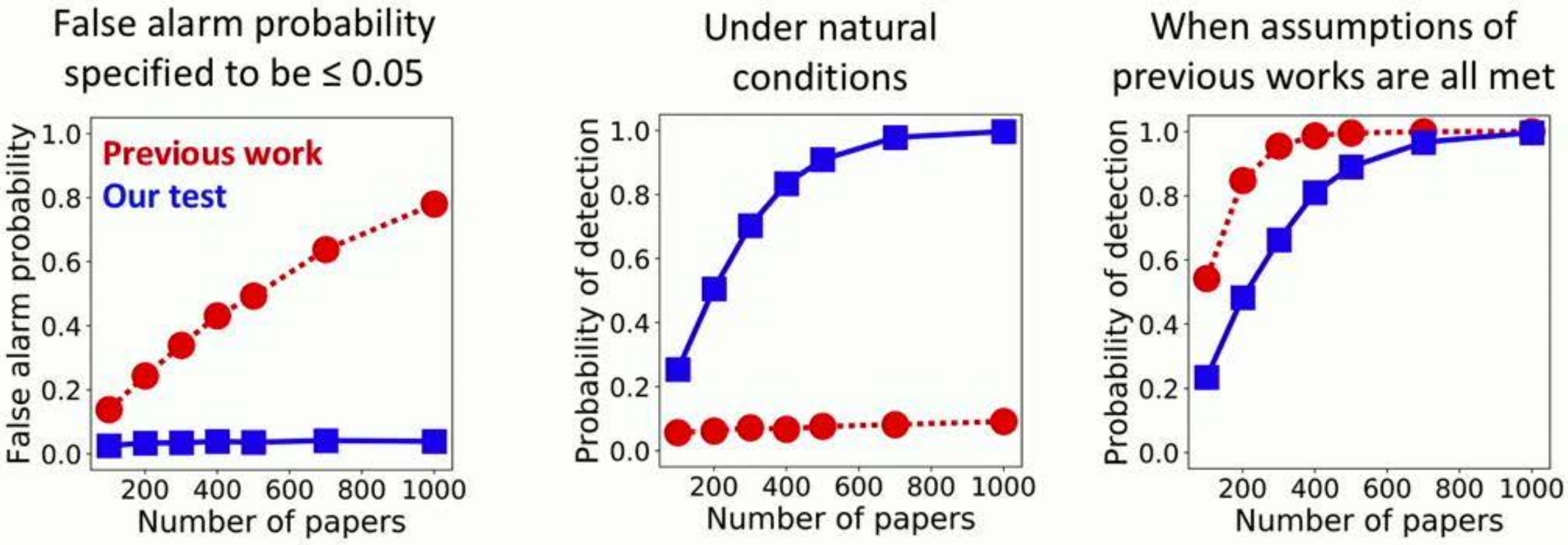


Under natural conditions



- **Positive results:**
  - A testing setup (with minimal changes to peer review processes)
  - Statistical tests
  - Strong, rigorous guarantees

Testing for biases



- **Positive results:**
  - A testing setup (with minimal changes to peer review processes)
  - Statistical tests
  - Strong, rigorous guarantees

Strategic behavior

with

Han Zhao

Yichong Xu

Xiaofei Shi

Motivation

Giving lower scores to other papers will improve my own relative score! Ha ha ha ha!

Motivation

Giving lower scores to other papers will improve my own relative score! Ha ha ha ha!

Balietti et al. (PNAS, 2016):

"competition incentivizes reviewers to behave strategically, which reduces the fairness of evaluations and the consensus among referees"

Also [Anderson et al. 2007, Langdford 2008 (blog), Akst 2010, Thurner and Hanel 2011...]

Motivation

Giving lower scores to other papers will improve my own relative score! Ha ha ha ha!

Balietti et al. (PNAS, 2016):

"competition incentivizes reviewers to behave strategically, which reduces the fairness of evaluations and the consensus among referees"

Also [Anderson et al. 2007, Langdford 2008 (blog), Akst 2010, Thurner and Hanel 2011...]

How to make peer review strategyproof?

Strategyproofness

- ✓ Framework for strategyproof peer review
  - For any reviewer, the decisions on papers conflicted with her/him are provably independent of the reviews given by her/him

Strategyproofness

✓ Framework for strategyproof peer review

- For any reviewer, the decisions on papers conflicted with her/him are provably independent of the reviews given by her/him

✗ Negative results of impossibility

Strategyproofness

✓ Framework for strategyproof peer review

- For any reviewer, the decisions on papers conflicted with her/him are provably independent of the reviews given by her/him

✗ Negative results of impossibility

ICLR 2016 empirical evaluation

✓ **Conditions for strategyproofness are indeed satisfied!**

Noise

with

Ivan Stelmakh

Aarti Singh

Noise

- Poor reviews due to inappropriate choice of reviewers

Noise

- Poor reviews due to inappropriate choice of reviewers
- Automated assignment: Toronto paper matching system (TPMS) and others

Noise

- Poor reviews due to inappropriate choice of reviewers

- Automated assignment: Toronto paper matching system (TPMS) and others
  - Unfair, especially to interdisciplinary or niche papers
  - Assign all very relevant reviewers to one paper and all irrelevant reviewers to another
  - No guarantees on overall process – how well does it help to identify the good papers?

Noise

- Poor reviews due to inappropriate choice of reviewers

- Automated assignment: Toronto paper matching system (TPMS) and others
  - Unfair, especially to interdisciplinary or niche papers
  - Assign all very relevant reviewers to one paper and all irrelevant reviewers to another
  - No guarantees on overall process – how well does it help to identify the good papers?

How to assign reviewers to papers ensuring fairness and accuracy?

Reviewer assignment

PeerReview4All assignment algorithm

Reviewer assignment

PeerReview4All assignment algorithm

Fairness: No paper is disadvantaged in favor of a paper that already has more advantage.

Reviewer assignment

PeerReview4All assignment algorithm

Fairness: No paper is disadvantaged in favor of a paper that already has more advantage.

Accuracy: Optimal recovery of good papers (under standard statistical models for noise in peer review)

Reviewer assignment

PeerReview4All assignment algorithm

Fairness: No paper is disadvantaged in favor of a paper that already has more advantage.

Accuracy: Optimal recovery of good papers (under standard statistical models for noise in peer review)

ICLR 2016

Fairness (assignment quality for worst-off paper)	↑ 25%
Average quality (TPMS objective quality)	↓ 2%

Conclusions

- **Urgent need to revamp and automate peer review**

  Observations & open problems: ***"Design and Analysis of the NIPS 2016 Review Process,"*** Shah, Tabibian, Muandet, Guyon, von Luxborg

- **Principled and practical approaches**
  - Impact!

- **Papers available on arXiv and my website**

  Short survey: **tinyurl.com/PeerReviewCMU**

tinyurl.com/PeerReviewCMU

Thank you!