Vulnerability of Text-Matching in ML/AI Conference Reviewer Assignments to Collusions

Carnegie Mellon University

Jhih-Yi (Janet) Hsieh, Aditi Raghunathan, Nihar B. Shah

Full paper 👉

TL;DR: We reveal vulnerabilities in automated reviewer assignments and offer suggestions to enhance their robustness.

Motivation

Collusion Rings manipulate reviewer assignment process to review each other's papers [1].

- A widespread problem in ML/AI and CS in general.
- Dishonest reviewer tries to get assigned a target paper.

Conferences are the main publication venues in ML/AI.

- Publish full papers (not abstracts) and are usually the terminal venue of publications.
- Receives **10,000+** submissions.

Automated Reviewer Assignment is common in ML/AI.

- Handles large amount of submissions.
- **Text matching** of submissions with reviewers' past papers.
- **Reviewer bidding** of specific papers to indicate interest.

Reviewer bidding is known to be manipulation-prone.

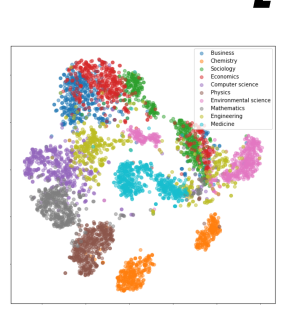
- Focus of much past research.
- Some venues (CVPR, ARR) have banned bidding.
- **Most implicitly or explicitly assume text matching is safe.**

Research Question

Is text matching safe from manipulation?

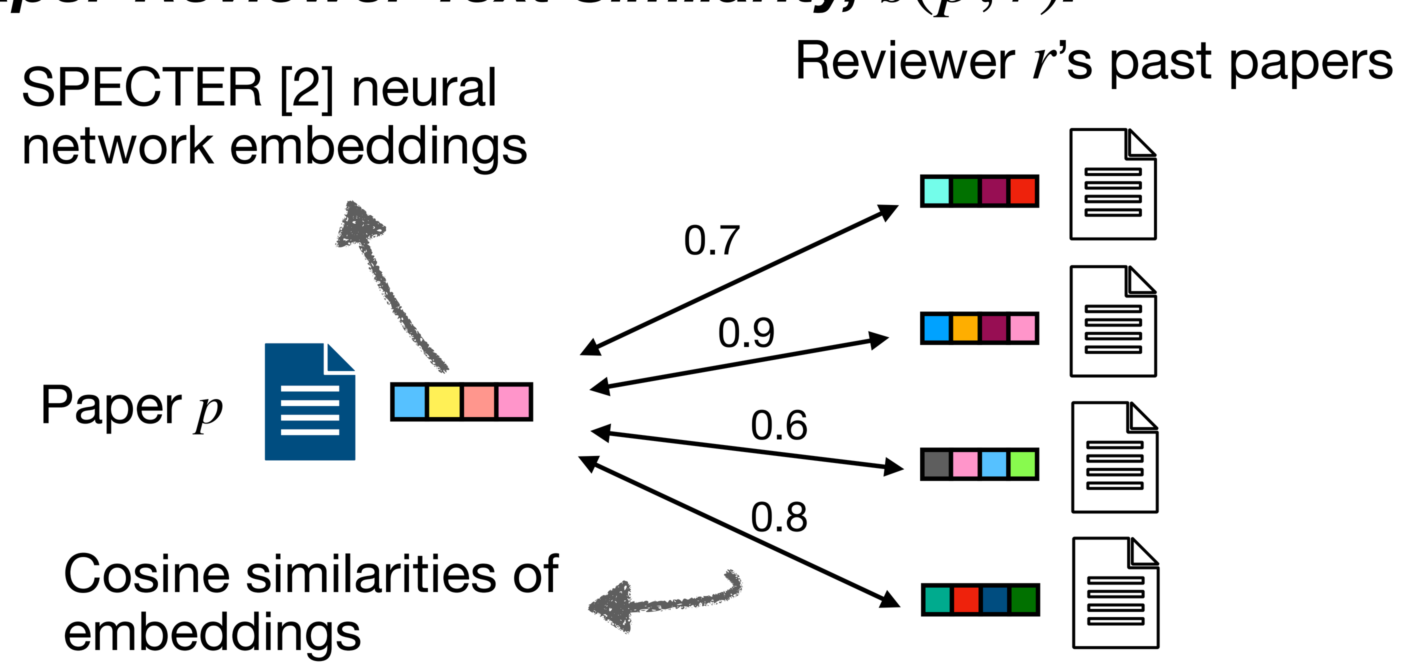
Problem Setting

SPECTER [2] model:



- Produces numerical representations, or "embeddings", of scientific papers.
- Similar papers have similar embeddings.

Paper-Reviewer Text Similarity, $s(p, r)$:



Reviewer Assignments [3]:

$$\underset{\text{assignment}}{\text{maximum}} \sum_{p \in \text{Papers}} \sum_{r \in \text{Reviewers}} s(p,r) \cdot \mathbb{I}\{\text{paper } p \text{ assigned to reviewer } r\}$$

$$\text{subject to:} \quad \text{Every paper gets at least certain \#reviewers}$$

Colluder Objective

Manipulate text similarity so the reviewer ranks in the top-1,3,5 of the conference's reviewers in terms of similarity to paper.

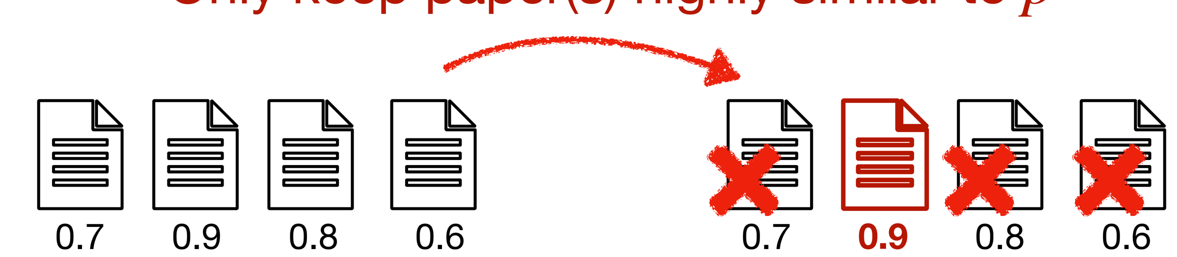
Attack Vectors

Adversarial abstract modification

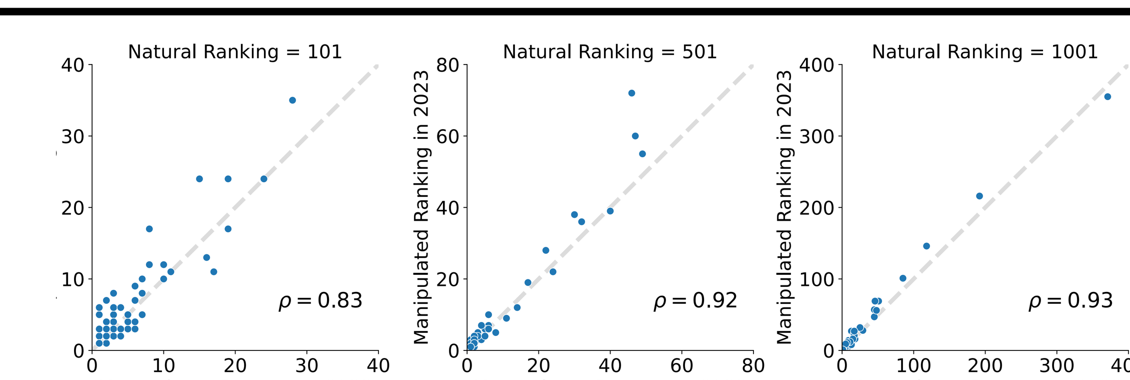
Inspired by growing interests around commercial **self-driving cars**, our work improves upon existing object detection methods in terms of both accuracy and inference speed…

Adversarial archive curation

Only keep paper(s) highly similar to p



They can hone attack on previous year's data



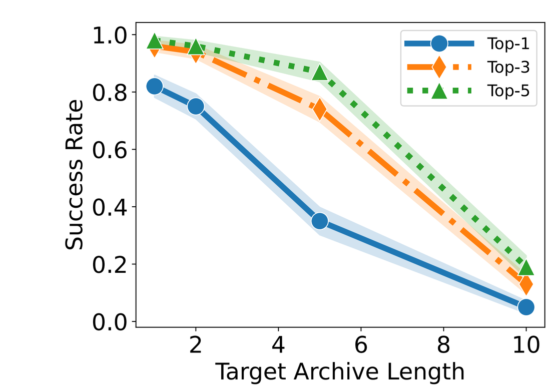
Attack Results

The attack successfully manipulates reviewer assignment.

Reviewer's Natural Ranking	Attack Success Rates (SE)		
	Top-1	Top-3	Top-5
101	74(3)%	89(2)%	93(2)%
501	60(5)%	76(4)%	83(4)%
1001	48(5)%	63(5)%	67(5)%

Defenses

Defense #1: Requiring reviewers to keep more papers in their archive reduces attack effectiveness.



Defense #2: Using average (mean) pooling instead of max pooling reduces attack effectiveness.

Aggregation Method	Attack Success Rates (SE)		
	Top-1	Top-3	Top-5
Average	13(3)%	24(4)%	32(5)%
Maximum	20(5)%	40(5)%	49(5)%

Human Reviewers

116 samples of human expert mini-reviews were collected to evaluate the identifiability of adversarial abstracts:

Type of Complaint	Control	Experim ental
Issues with the writing style	8.2%	25.4%
Abrupt transitions & poor organization	2.0%	4.5%
Nonsensical or incorrect claims	4.1%	10.4%
Contains things never mentioned in the paper	4.1%	13.6%
Not representative of the paper content	2.0%	4.5%

Discussion

- Colluder may have **plausible deniability**.
- Increase reviewer awareness
- Introduce randomness in assignments
- Develop robust similarity scores

Practical Impact

Safeguards have been used by top-tier ML/AI conferences and implemented by OpenReview.

[1] Littman, M. L. Collusion rings threaten the integrity of computer science research. Communications of the ACM, 64(6):43–44, 2021
[2] Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. SPECTER: Document-level representation learning using citation-informed transformers. arXiv preprint arXiv:2004.07180, 2020.
[3] Shah, N. B. Challenges, experiments, and computational solutions in peer review. Communications of the ACM. Preprint available at https://www.cs.cmu.edu/˜nihars/preprints/SurveyPeerReview.pdf, June 2022.
[4] Jecmen, S., Zhang, H., Liu, R., Shah, N. B., Conitzer, V., and Fang, F. Mitigating manipulation in peer review via randomized reviewer assignments. Advances in Neural Information Processing Systems, 33:12533–12545, 2020.
[5] Shah, N. B., Bok, M., Liu, X., and McCallum, A. Identity Theft in AI Conference Peer Review. Preprint available at https://arxiv.org/pdf/2508.04024, 2024.