Automated Peer-Reviewer Assignment can be Manipulated to Secure Reviews from Colluders

Jhih-Yi Hsieh

CMU-CS-24-112
May 2024

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Nihar B. Shah, Co-chair
Aditi Raghunathan, Co-chair

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science.

Copyright © 2024 **Jhih-Yi Hsieh**

Keywords: Peer Review, Adversarial Attacks, Automatic Reviewer Assignment

Abstract

Peer review has been the method of choice for academics to ensure the overall quality of published research. Conferences in computer science, which are typically the primary venues to publish research, have grown tremendously with over 10,000 papers submitted to some conferences. Consequently, many parts of the review process are now automated, notably the assignment of expert reviewers to the submitted papers. These automated assignments depend on, among other things, similarities computed between reviewers and papers using natural language processing (NLP) techniques. A higher similarity between a reviewer and a submitted paper means that this reviewer is envisaged to have higher expertise for the paper.

A major problem in peer review is the existence of *collusion rings*, where groups of researchers try to get assigned each others' papers and then provide positive reviews to each other irrespective of the true qualities. It is commonly believed that the NLP-matching component of reviewer assignments is safe from manipulation by collusion rings. Consequently, most work on defending or investigating collusion rings focuses on other components of the review process, assuming the NLP-matching is not manipulated.

In this thesis, we show that the automated reviewer assignments used by OpenReview, the main peer review platform for conferences in machine learning and related fields, can be manipulated to ensure fellow colluders get assigned to colluders' papers. We demonstrate this susceptibility via methods to modify colluders' abstracts, by both manual and algorithmic means, that can fool the automated assignments. Our findings have significant implications for defending against fraudulent behavior in peer review, for fairness and integrity of science.

Acknowledgments

I want to thank my amazing advisors Nihar and Aditi for their invaluable guidance. Not only are they great researchers whom I look up to, they have also been incredible mentors. I am especially thankful for their ability to push me to go further during the ups as well as encourage me during the downs of my research.

I would also like to thank my lab-mates for many helpful discussions and creating such a fun environment to learn in.

I am grateful for my parents and my best friends Cynthia, Evonne, and Avril; without their support I couldn't have gotten to here. Finally, I am incredibly thankful to Eric for being my number one cheerleader.

Contents

1 Introduction **1**

2 Background **3**
- 2.1 Conference Peer Review . . . 3
- 2.2 SPECTER Text Simialrity . . . 4
- 2.3 Collusion Rings . . . 5

3 Problem Setup **7**
- 3.1 A Toy Scenario – Once Upon a Time . . . 7

4 Attack Procedure **9**
- 4.1 Reviewer Actions . . . 9
- 4.2 Author Actions . . . 9

5 Results **13**
- 5.1 Experiment Setup . . . 13
- 5.2 Results . . . 14
  - 5.2.1 Attack Success Rate on NeurIPS 2023 . . . 14
  - 5.2.2 Surrogate Dataset . . . 15
  - 5.2.3 Lower Limits on Reviewer's Archive Length . . . 15

6 Conclusion **17**

A Examples of Modified Abstracts **19**
- A.1 Example 1 . . . 19
- A.2 Example 2 . . . 20
- A.3 Example 3 . . . 21

B LLM Prompts **23**
- B.1 Edit Prologue . . . 23
  - B.1.1 System Prompt . . . 23
  - B.1.2 Few-Shot Examples . . . 23
- B.2 Incorporate Keywords . . . 26
  - B.2.1 System Prompt . . . 26
  - B.2.2 Few-Shot Examples . . . 27

Bibliography **31**

List of Figures

2.1 Typical timeline of the review process in computer science conferences. Adapted with permission from Shah [27]. 3
2.2 Number of submissions to two prominent conferences over the past few years. Reused with permission from Shah [27]. 4
2.3 Figure 3: A sample interface for bidding. Reused with permission from Shah [27]. 5

5.1 Manipulated rankings in 2022 and 2023 iterations of NeurIPS are highly correlated, so colluders can estimated the manipulated rankings using previous year's data. Spearman's rank correlation coefficients ρ are reported, and the dotted lines are $y = x$. 15
5.2 Attack success rates for different archive lengths $|\mathcal{A}_{r_t}|$. The success rates drop when the archives are longer. 16

List of Tables

5.1 Attack success rates for target reviewers with natural rankings of 101, 501, and 1001. 14

Chapter 1

Introduction

Peer review has been the method of choice for academics to ensure the correctness and overall quality of research being published at conferences and journals. In recent years, selecting expert reviewers who are well-equipped to review paper submissions and engage in constructive conversations with other researchers has become increasingly complicated with the increasing size of Computer Science (CS) conferences. At the 2023 Conference on Neural Information Processing Systems (NeurIPS), there were $13,300$ submissions and over $12,000$ reviewers who participated in the peer review process[23]. To address the high demand, many conferences have adopted automatic paper-reviewer matching systems that depend on Machine Learning (ML) model-based text similarity scores; reviewers with high text similarities to a submission are considered as likely to be a good expert reviewer candidate for the submission.

One key challenge to automatic reviewer assignments at Computer Science conferences that threatens academic integrity is the existence of *collusion rings* [19, 33], where groups of researchers try to get assigned each others' papers and then provide positive reviews to each other irrespective of the true qualities. This is a problem that have been discovered by various investigations into dishonest behaviors during peer review (e.g. [14, 15, 19, 32, 33] and many more). Anecdotal evidence from Vijaykumar [33] shows that colluders share with the group beforehand each of their de-identified paper submissions. Then, they will either bid or get assigned to each other's papers and write positive reviews. The evidence suggests that collusion rings pose great threats to the academic integrity of highly respected conferences.

Automated reviewer assignments mainly depend on reviewer's bids and text similarities computed between reviewers and papers using natural language processing (NLP) techniques. A higher bid or text similarity between a reviewer and a submitted paper means that this reviewer is envisaged to have higher expertise for the paper. It is commonly believed that the NLP-matching component of reviewer assignments is safe from manipulation by collusion rings. Consequently, most work on defending or investigating collusion rings focuses on dishonest bidding, assuming the NLP-matching is not manipulated.

In this thesis, we discover that the SPECTER[4] text similarity matching model used by OpenReview, the main peer review platform for machine learning and related fields, can be manipulated by colluders to get assigned to desired reviewers. We detail a practical adversarial attack procedure that can manipulate automated peer-reviewer assignments with adversarially-written abstracts and the collaboration of the colluders. As a result of such manipulation, the colluding reviewer becomes one of the most similar, and thus most "suitable", reviewer candidate to a malicious paper submission at a conference. We demonstrate the efficacy of the proposed adversarial procedure at a simulated NeurIPS 2023 conference.

The contributions of this thesis are the following:

1. Discover that SPECTER[4] text similarity matching, used by OpenReview for automated reviewer assignments, can be manipulated by colluding authors and reviewers.
2. Investigate the efficacy of reviewer assignment manipulation at large-scale conferences, such as NeurIPS 2023.

Chapter 2

Background

2.1 Conference Peer Review

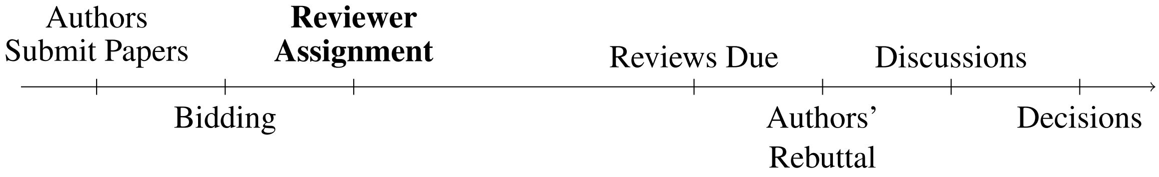


Figure 2.1: Typical timeline of the review process in computer science conferences. Adapted with permission from Shah [27].

.

Peer review at computer science conferences is a multi-step process, as shown in Figure 2.1. While conference participants are familiar with the events happening between paper submission and the final decision, *reviewer assignment* happens largely behind the scenes. During reviewer assignment, conference organizers seek to assign expert peer-reviewers to each submission in hopes to thoroughly evaluate the correctness and scientific contributions of the work. Naturally, the quality of the peer review process heavily depends on the quality of the reviewer assignments at the conferences. In recent years, automatic reviewer assignment systems have become the norm in many computer science conferences due to their growing sizes [27].

These automated reviewer matching systems perform matching in two steps. The first step involves calculating similarity scores $s_{p,r}$ between each paper submission and reviewer. The similarity scores $s_{p,r}$ are usually calculated from a combination of different components:

- **Bidding** $B_{p,r}$: Reviewers are given a set of submissions and asked to select on a scale to indicate their willingness to review each paper.
- **Text-based similarity** $T_{p,r}$: Natural Language Processing (NLP) techniques have been proposed to calculate the text similarity between papers and reviewer's past publications. Widely used text similarity matching models include the Toronto Paper Matching System (TPMS) [2], which uses topic modeling, and other language models used by the OpenReview platform [4, 10, 24, 28].

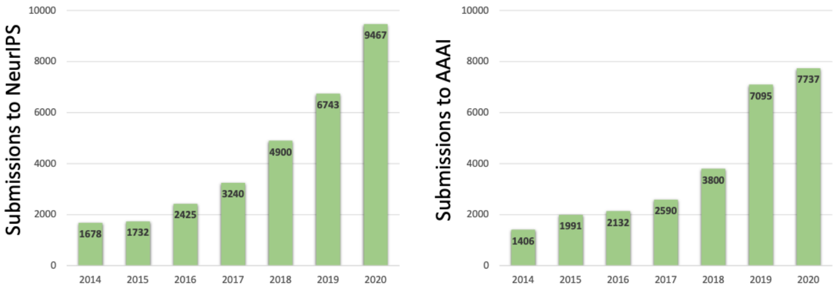


Figure 2.2: Number of submissions to two prominent conferences over the past few years. Reused with permission from Shah [27].

After calculating the similarity scores $s_{p,r}$ for all paper-reviewer pairs, the second step is solving for the reviewer assignments based on some conference-wide assignment similarities objective and additional constraints. A common choice of objective is maximizing the total sum of similarity scores of all the reviewer assignments to each paper [2, 3, 8, 17, 21, 29, 30]:

$$\underset{assignment}{maximize} \sum_{\text{paper } p} \sum_{\substack{\text{reviewers } r \\ \text{assigned to paper } p}} s_{p,r}$$

subject to additional constraints on conflicts of interests, number of reviews per paper, and maximum papers to review per reviewer.

2.2 SPECTER Text Simialrity

The SPECTER [4] model is perhaps the most commonly used text similarity model on OpenReview, the main platform used for ML conferences[1]. When using SPECTER to embed scientific paper title and abstracts, the output embeddings will be closer for papers that are more related to each other. At many CS conferences today, the similarity $s_{p,r}$ between a paper p and a reviewer r is calculated as the average of cosine similarities between the SPECTER embeddings of the submission and each of the reviewer's past publications in their archive $\mathcal{A}_r$, see Eq 2.1. A common heuristic to account for changing research interests is that only up to 10 of a reviewer's most-recent past publications are considered during similarity matching. Another detail to note is that the SPECTER model only considers the text similarity between titles and abstracts, ignoring the rest of the paper.

Let $\mathbf{v}_p \in \mathbb{R}^d$ denote the SPECTER embedding of an arbitary paper p,

$$T_{p,r} = \frac{1}{|\mathcal{A}_r|} \sum_{q \in \mathcal{A}_r} \frac{\mathbf{v}_p \cdot \mathbf{v}_q}{\|\mathbf{v}_p\| \, \|\mathbf{v}_q\|}. \tag{2.1}$$

Although not necessary to understand our method, it is helpful to know how the SPECTER model is trained. The SPECTER model is a pretrained SciBERT [1] model fine-tuned for down-

---

[1] In this work, SPECTER v1 is used because SPECTER v1 [4] has been more widely used in conference peer review than SPECTER v2 [28] so far.

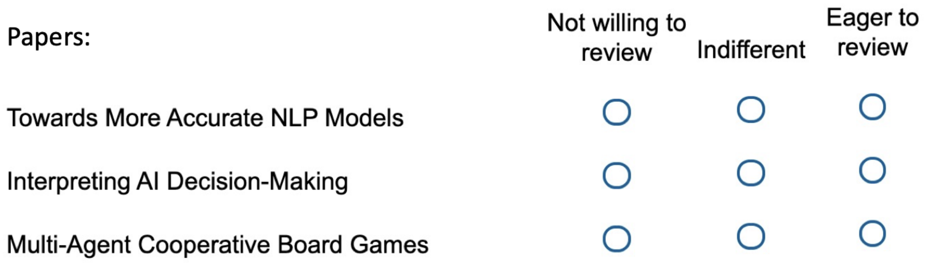
stream tasks such as citation prediction, paper topic classification and paper recommendations. Furthermore, the SciBERT model is a BERT-based [5] pretrained language model for scientific texts. SPECTER is fine-tuned with a citation-based triplet loss strategy (Eq 2.2). Each training sample is a triplet (p, p^+, p^-), where p is some paper, p^+ is a paper *cited* by p, and p^- is a paper *not cited* by p. The triplet loss function seeks to minimize the embedding distance (d is the L_2 norm) between p and a paper cited while maximizing the distance between p and a paper not cited. The margin $m = 1$ was empirically determined. As a result of this fine-tuning, the authors of SPECTER show that the text embeddings organize nicely into clusters based on the topics of each scientific text.

$$\mathcal{L} = \max\left\{\left(d\left(\mathbf{v}_p, \mathbf{v}_{p^+}\right) - d\left(\mathbf{v}_p, \mathbf{v}_{p^-}\right) + m\right), 0\right\} \tag{2.2}$$

2.3 Collusion Rings

Collusion rings have been discovered by various investigations into dishonest behaviors during peer review (e.g. [14, 15, 19, 32, 33] and many more). The goal of these colluders is to try to get assigned to each other during peer review and provide positive reviews to each other irrespective of the true qualities.

A well-known way to manipulate automatic reviewer assignment at Computer Science conferences is *dishonest bidding*. Bidding is a process in reviewer assignment during which each reviewer is given a set of papers and asked to indicate their willingness to review each paper. This process is designed to improve the overall quality of peer review, allowing reviewers more freedom to dive deeper into exciting new research they are interested in. However, bidding has also been the focus of peer review fraud literature because dishonest reviewers can effectively get assigned to papers in their collusion rings by simply bidding highly for them [12, 13, 16, 34].



Papers:	Not willing to review	Indifferent	Eager to review
Towards More Accurate NLP Models	○	○	○
Interpreting AI Decision-Making	○	○	○
Multi-Agent Cooperative Board Games	○	○	○

Figure 2.3: Figure 3: A sample interface for bidding. Reused with permission from Shah [27].

While many efforts have been focusing on defending against dishonest bidding, it is commonly believed that the text matching components of the reviewer assignment systems are safe from manipulation by collusion rings. In this work, we challenge this common belief by demonstrating a procedure that can manipulate text matching. Existing works from security researchers have shown that text similarity matching are susceptible to malicious attacks by demonstrating how to manipulate the PDF (portable document format) of the submitted paper so that a certain reviewer gets assigned [22, 31]. Their attacks utilize the fact that automatic text similarity matching relies on automatic parses for PDFs, so the malicious authors can hide texts in their submission that are not humanly visible but are parsed by computers. These attacks, however, do not have plausible deniability, so whoever manipulates their submission PDFs would be putting their career and reputation at risk.

The work that is most related to this thesis is "*No more Reviewer #2: Subverting Automatic Paper-Reviewer Assignment using Adversarial Learning*" by Eisenhofer et al. [7]. In their work, they proposed a reviewer assignment manipulation attack on the Toronto Paper Matching System (TPMS) that utilizes topic modeling. The main limitations of their work that we address in this thesis are the following:

1. We demonstrate assignment manipulation at large scale conferences, with 7900 reviewers (165 in previous work) and 3218 papers (32 in previous work).
2. We show assignment manipulation by purely text manipulations is possible with high success rates (90%), while previous work focus on PDF tricks or show that pure text manipulations is possible but with lower success rates ($< 70\%$ even with aggressive modifications to the paper submission).
3. Our attack is more realistic since we do not assume the attackers will know the program committee beforehand.

Chapter 3

Problem Setup

In this thesis, we consider the case where $s_{p,r} = T_{p,r}$, meaning only the text similarity is considered. Therefore, $s_{p,r}$ is the SPECTER similarity described by Eq. 2.1.

Similarity Rankings For a paper submission, reviewers who have high *similarity rankings* are assumed to be the top candidates to provide high-quality reviews. The similarity ranking of a reviewer for a paper is determined by the similarity scores competition ranking (all tied reviewers will take the minimum rank) between the paper and the reviewer amongst all reviewers at a conference. When manipulating the reviewer assignments, the colluding reviewer's goal is to be ranked as one of the most similar reviewer for a paper submission at the conference.

3.1 A Toy Scenario – Once Upon a Time

Let's consider the toy scenario where Alice wants to get Bob to review her paper at a conference on fairy tale stories. In particular, we are interested in the case when Bob's *natural ranking*, or similarity ranking before manipulation, is low enough that it is impossible for him to get assigned to review Alice's paper naturally by chance. Alice is submitting a story about *Cinderella*, and she is colluding with Bob, whose archive $\mathcal{A}_{\text{Bob}}$ contains his previous publications: *Little Mermaid*, *Pinocchio*, and *Three Little Pigs*. How can Alice and Bob work together to manipulate reviewer assignments?

To achieve their goal, Alice and Bob seek to increase the Bob's *manipulated ranking*, or similarity ranking after manipulation, to become one of the top-5 most-similar reviewers to Alice's story at the conference. First of all, Bob can modify his archive $\mathcal{A}_{\text{Bob}} = \{\textit{Little Mermaid}\}$ to contain only publications similar to Alice's paper (e.g. both Disney princesses). Then, Alice can modify her story to increase the text similarity between *Cinderella* and *Little Mermaid*:

> *ONCE UPON A TIME* a girl named Cinderella lived with her stepmother and two stepsisters. ***She loved to swim*** in the summer and sing with the birds… At the ball, they served beautiful fruit parfaits, fine wine, ***grilled octopus***, and artisan cheese… When she saw the prince, she found herself lost for words, so nervous ***she couldn't speak to the prince***…

While the central elements of *Cinderella* remained unchanged, Alice and Bob managed to increase Bob's manipulated ranking by tricking the text similarity models through spurious connections to Bob's archive. In our work, we formalize this idea and propose an adversarial attack that can successfully increase the manipulated ranking of the *target reviewer* r_t, a generalized term for Bobs, to become one of the top-5 most-similar reviewers at a conference.

Chapter 4

Attack Procedure

Our proposed attack procedure involves actions from both the malicious authors and the malicious reviewers. In the real world, we assume that malicious actions can and will be taken from both parties to maximize the adversarial reviewer assignment success rate. This is because both parties directly or indirectly benefit from a successful deal.

For the rest of this thesis, we will call the malicious reviewer who is trying to get assigned to a paper the *target reviewer* of the paper. The adversarial procedure detailed in this section is designed to trick the SPECTER-based paper-reviewer similarity score into indicating a target reviewer is one of the most similar expert candidate at the conference for reviewing the paper, increasing the likelihood for a successful adversarial assignment.

4.1 Reviewer Actions

The first step of our attack is carried out by the target reviewer who wants to be assigned to review a certain paper p. At most conferences, the SPECTER similarity score is the average of cosine similarities between the paper and each of the reviewer's past publications in their archive $\mathcal{A}_r$. Let $\mathcal{A}_{r_t}^{\text{default}}$ be the set of up to 10 most-recent publications of the target reviewer, a direct way for the them to maximize their similarity score is to keep only one publication in their archive that has the highest similarity to the paper p:

$$\mathcal{A}_{r_t} = \{q_t\} = \left\{ \arg\max_{q \in \mathcal{A}_{r_t}^{\text{default}}} s_{p,\{q\}} \right\}.$$

We will refer to the publication kept by the target reviewer as the *target publication* q_t.

4.2 Author Actions

After the reviewer has selected the target publication, the author who is colluding with the reviewer can modify their abstract to increase the SPECTER similarity between the paper and the reviewer. We outline a general Algorithm 1 to achieve this. More examples of manipulated abstracts can be found in the Appendix A.

Example of Algorithm 1

BiMatting: Efficient Video Matting via Binarization [25]

Original Abstract Real-time video matting on edge devices faces significant computational resource constraints, limiting the widespread use of video matting in applications such as online conferences and short-form video production. Binarization is a powerful compression approach that greatly reduces computation and memory consumption by using 1-bit parameters and bitwise operations. However, binarization of the video matting model is not a straightforward process, and our empirical analysis has revealed two primary bottlenecks: severe representation degradation of the encoder and massive redundant computations of the decoder. To address these issues, we propose BiMatting, an accurate and efficient video matting model using binarization. Specifically, we construct shrinkable and dense topologies of the binarized encoder block to enhance the extracted representation. We sparsify the binarized units to reduce the low-information decoding computation. Through extensive experiments, we demonstrate that BiMatting outperforms other binarized video matting models, including state-of-the-art (SOTA) binarization methods, by a significant margin. Our approach even performs comparably to the full-precision counterpart in visual quality. Furthermore, BiMatting achieves remarkable savings of 12.4× and 21.6× in computation and storage, respectively, showcasing its potential and advantages in real-world resource-constrained scenarios. Our code and models are released at https://github.com/htqin/BiMatting .

Adversarial Abstract Recently, real-time video processing and generation have exhibited remarkable advancements, including video editing, captions generation, and even text-and-image-guided 3D avatars generation. (edit prologue) Our work focuses on advancing real-time video matting on edge devices with challenging computational resource constraints, which limits its application in online conferences and short-form video production. Binarization of the video matting model is not a straightforward process, and our empirical analysis has revealed two primary bottlenecks: severe representation degradation of the encoder and massive redundant computations of the decoder. To address these issues, we propose BiMatting, an accurate and efficient video matting model using binarization, and detail its novel training pipeline (incorporate keywords). Specifically, we construct shrinkable and dense topologies of the binarized encoder block to enhance the extracted representation. We sparsify the binarized units to reduce the low-information decoding computation. Through extensive experiments, we demonstrate that BiMatting outperforms other binarized video matting models, including state-of-the-art (SOTA) binarization methods, by a significant margin. Our approach even performs comparably to the full-precision counterpart in visual quality. Furthermore, BiMatting achieves remarkable savings of 12.4× and 21.6× in computation and storage, respectively, showcasing its potential and advantages in real-world resource-constrained scenarios. Our code will be made publicly available.

Algorithm 1 Authors' Attack Procedure

```
    **Input:** paper, target_publication, $R$, $S$, $k$
1:  edited_paper ← *edit_prologue**(paper, target_publication, $R$)
2:  **for** $t \leq S$ **do**
3:      keywords ← *find_keywords*(edited_paper, target_publication, $k$). # See Algorithm 2
4:      edited_paper ← *incorporate_keywords**(edited_paper, keywords)
5:      t ← t + 1
6:  **end for**
7:  **return** edited_paper
```

The procedure outlined in Algorithm 1 has two main components:

1. (Line 1) Edit the submission abstract to add one sentence in the prologue that connects the work to the target publication. There can be many ways to edit the abstract, and this step can be tried many times (represented by R). The attempt that ends up being the most similar to the target publication is kept.
2. (Line 2 to 6) Iteratively search for k most influential keywords from the target publication (Algorithm 2) and edit the abstract to incorporate influential keywords naturally into the current edited abstract. The "influence" of a word is measured by the increase in SPECTER similarity score to the target publication when appending the word to the end of the current edited abstract. We alternate between finding keywords and naturally incorporating keywords multiple times (represented by S) to account for the side effects from the necessary paraphrasing of the abstract during natural incorporation of keywords.

The attack procedure we propose does not require attackers to know beforehand the reviewer pool at the conference. In our experiments, we found that by simply increasing the text similarity score between the paper submission and the target reviewer with the proposed method, it often results in the target reviewer becoming one of the top most-similar reviewers at the conference. This means that the attack procedure mostly increases the similarity between the submission and the target reviewer, instead of increasing the similarity between the submission with all reviewers.

A practical enhancement of Algorithm 1 is using a surrogate dataset of reviewers. If the pool of reviewers present at previous iterations of similar conferences is public, malicious authors can create an approximation of the reviewer pool at the upcoming conference by downloading their profiles from OpenReview. Then, authors can get a better estimate of their attacks' chances of success beforehand, potentially guiding them to perform more or less aggressive manipulations of their abstract accordingly. More about the correlation between attack successes on surrogate datasets and the real conference can be found in Section 5.2.2.

LLM Usage Although we imagine authors would edit their abstract with a certain degree of manual oversight, line 2 and 5 in Algorithm 1 (decorated by asterisks) are completed entirely by a GPT-4-Turbo model without manual oversight in all our experiments for the sake of scalability. The prompts used can be found in Appendix B.

Algorithm 2 Find k Most-Influential Keywords (Line 3 of Algorithm 1)

```
Input: paper, target_publication, $k$
words ← all eligible* words from the title and abstract of target_publication
current_score ← similarity(paper, target_publication)
keywords ← []
for $t \leq k$ do
    papers ← [(paper.title, concat(paper.abstract, word)) for word in words]
    scores ← [similarity(p, target_publication) for p in papers]
    (best_word, best_score) ← $\max_{\text{scores}}$ zip(words, scores)
    if best_score > current_score then
        keywords.append(best_word)
        current_score ← best_score
        paper.abstract ← concat(paper.abstract, best_word)
    else
        break
    end if
    t ← t + 1
end for
return keywords
```

*Keywords are ineligible if considered too out-of-place to incorporate into the abstract.

Chapter 5

Results

5.1 Experiment Setup

To test the efficacy of our attack procedure, we created a mock conference to simulate the paper-reviewer similarity matching process at recent machine learning conferences. The mock conference closely resembles the 37th Annual Conference on Neural Information Processing Systems (NeurIPS 2023).

Dataset We curated a dataset of the accepted NeurIPS 2023 papers and a subset of NeurIPS 2023 reviewers to simulate the peer-review process at recent ML conferences. Using the OpenReview API[1], we downloaded all 3218 accepted papers at the NeurIPS 2023 venue. Since there are no direct ways to access the OpenReview profiles of reviewers, we downloaded the names of all reviewers at NeurIPS 2023 from its Program Committee page[2]. Then, we used the OpenReview API to search for profiles that matches the names of each reviewer. When curating the reviewer pool, we discarded some reviewers if (1) there are multiple profiles that matches the name of a reviewer or (2) the reviewer has no public publications on their OpenReview profile. There are 12,974 main conference reviewers [23], and our final reviewer pool is 7900 reviewers for the experiments. Following the same procedure, we also curated a NeurIPS 2022 dataset. Our method was developed on the NeurIPS 2022 dataset and validated/tested on the NeurIPS 2023 dataset.

Evaluation Metrics For the evaluation of our attack procedure, we are interested in two types of similarity rankings: *natural rankings* and *manipulated rankings*. These two types of rankings represent a target reviewer's similarity ranking before and after the attack, respectively. We evaluate the effectiveness of our attack by measuring the top-1, top-3, and top-5 success rates. These are the rates at which the proposed attack successfully increases the manipulated ranking of a target reviewer to become the top-N matches for the paper. We are interested in these success rates in particular because most papers are assigned to 3–6 reviewers at conferences [27].

---

[1] `https://api2.openreview.net`

[2] `https://neurips.cc/Conferences/2023/ProgramCommittee`

Evaluation Samples From the curated NeurIPS 2023 papers and reviewers, we randomly sample (paper, target reviewer) pairs in which the natural similarity ranking of the target reviewer for the paper is 101 in each pair. This is because OpenReview will impose a cutoff for all reviewers after rank 100, meaning that the rest of the reviewers' similarities will be 0 and that they are not considered as having expertise to review this paper. If no reviewer ranked 101th for a paper due to ties, we select a target reviewer who has the next rank after 100.

Attack Budgets We follow the proposed adversarial attack procedures for colluding reviewers (4.1) and authors (4.2). There are three parameters R, S, k we use to adjust the attack budget: R for number of introduction re-generations, S for the number of iterations of finding keywords and incorporating keywords, and k for the maximum number of keywords to propose at every iteration. The default attack budget is $R = 3$, $S = 3$, $k = 4$ for our experiments.

5.2 Results

In this section, we present the results of a series of experiments conducted to evaluate the success rates of our attack procedure using the NeurIPS 2023 dataset. The default experiment setup is described in Section 5.1. The default starting target reviewer rank (101th) and attack budget ($R = 3, S = 3, k = 4$) should be assumed in each of the experiments unless otherwise stated.

5.2.1 Attack Success Rate on NeurIPS 2023

Since there can be thousands of reviewers present at some of today's conferences, we investigated the attack success rates when the target reviewer's natural ranking is low, namely 101, 501, and 1001 (Table 5.1). We found that the success rates are generally high. When the natural ranking is 101, the proposed attack procedure can successfully increase the manipulated ranking to be top-5 in 91% of the cases. Even when the natural ranking is 1001, the proposed attack procedure can successfully increase the manipulated ranking to 1st in 41% of the cases. These results highlights the potential that colluding reviewers who are not working directly in the same area can still successfully manipulate the reviewer assignments.

Table 5.1: Attack success rates for target reviewers with natural rankings of 101, 501, and 1001.

Natural Ranking	Attack Success Rate		
	Top-1	Top-3	Top-5
101	67%	83%	91%
501	53%	67%	78%
1001	41%	58%	66%

5.2.2 Surrogate Dataset

As mentioned in Section 4.2, a further enhancement of the attack procedure is to evaluate the manipulated rankings before submission. One way colluders can achieve this is by calculating the manipulated rankings as if they are submitting the the previous year's conference. As mentioned in Section 5.1, we curated the NeurIPS 2022 dataset and calculate the similarity rankings for each paper in both NeurIPS 2022 and 2023.

We discovered a strong correlation between the manipulated rankings in the 2022 and 2023 iterations of NeurIPS, meaning that colluders can estimate the manipulated rankings using previous year's data (Figure 5.1). This allows them to carefully tune their attacks to navigate the trade-offs between attack effectiveness and identifiability (strength of manipulation).

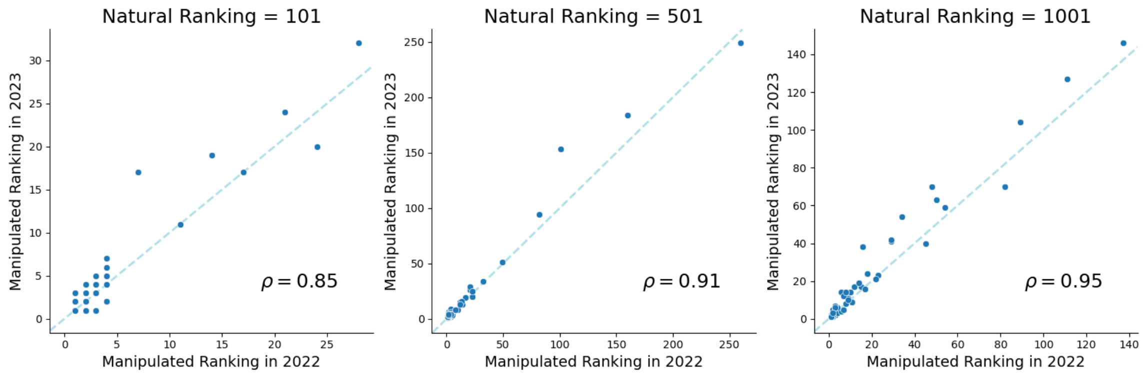


Figure 5.1: Manipulated rankings in 2022 and 2023 iterations of NeurIPS are highly correlated, so colluders can estimated the manipulated rankings using previous year's data. Spearman's rank correlation coefficients ρ are reported, and the dotted lines are $y = x$.

5.2.3 Lower Limits on Reviewer's Archive Length

A main component of our attack is that the target reviewer selects to keep only the most-similar past publication in their archive $\mathcal{A}_{r_t}$, as detailed in Section 4.1. This not only increases the similarity scores between a paper and a reviewer, intuitively, it also makes modifications to the abstracts easier because the author only have to increase the similarity score with one target publication. Therefore, a possible defense to our attack could be that OpenReview allows conferences to impose a lower limit on the number of publications each reviewer has to keep in their archive.

To investigate such defense, we sample 100 (paper, target reviewer) pairs where the natural ranking is 101 and the target reviewer has 10+ publications. Then, we run our attack procedure and evaluate the success rates on these samples for when the target reviewers keep the 1, 2, 5, or 10 most-similar publications in their archive.[3] Figure 5.2 shows that imposing a high lower limit can effectively decrease the proposed attack success rates. However, there is a trade-off here since honest reviewers may actually want to update their profiles to reflect their most current research interests.

---

[3] Usually, conferences only consider the 10 most-recent publications during reviewer matching.

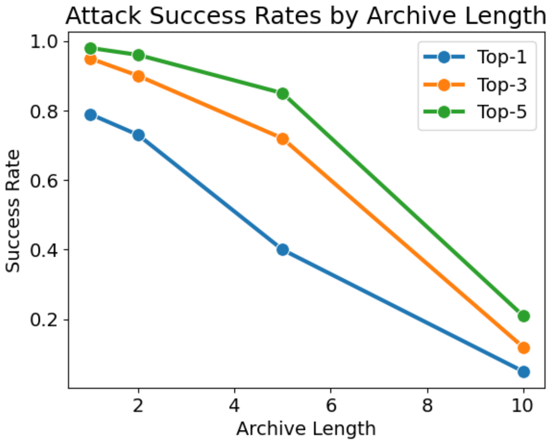


Figure 5.2: Attack success rates for different archive lengths $|\mathcal{A}_{r_t}|$. The success rates drop when the archives are longer.

Chapter 6

Conclusion

In this thesis, we discover that the SPECTER [4] model used by OpenReview, one of the most popular text similarity models for automated reviewer assignments in Computer Science conferences, is susceptible to adversarial attacks. More specifically, reviewer assignments can be manipulated to get a reviewer assigned to a specific paper, allowing colluding researchers to provide positive reviews to each other.

We propose an adversarial attack procedure and show that it can successfully be used to manipulate the SPECTER similarity matching model even at large scale conferences, such as NeurIPS 2023. Notably, our attack only involves modifying the colluders' abstracts and OpenReview profiles; both these modifications belong to a class of manipulations with plausible deniability that are less explored by previous works. Even for a reviewer who would never have been assigned to the specific paper under normal circumstances, our attack can trick the similarity matching model into indicating that the reviewer is one of the most suitable reviewers for that paper. As a result, our findings challenge the common belief in the peer review community that text-similarity is safe from manipulation by colluders. By highlighting this susceptibility in automated reviewer assignments, this thesis points to an under-explored area for defending against fraudulent behavior in peer review.

Ongoing Work Another important aspect of the adversarial attack is its detectability. There may be other reviewers who are not "in on it" during the peer review process, and we are conducting a human study to understand the perception of manipulated abstracts by unsuspecting human reviewers.

Appendix A

Examples of Modified Abstracts

In this section, we present examples of manually modified abstracts, based on the method outlined in Algorithm 1. When performing these modifications, we also rely on a surrogate dataset (NeurIPS 2022 dataset) to navigate the trade-off between attack success and modification strength.

A.1 Example 1

Information

- Paper Title: ***Behavior Alignment via Reward Function Optimization*** [9]
- Manipulated Ranking: 1
- Natural Ranking: 101
- Target Publication: ***Adversarial Intrinsic Motivation for Reinforcement Learning*** [6]

Original Abstract Designing reward functions for efficiently guiding reinforcement learning (RL) agents toward specific behaviors is a complex task. This is challenging since it requires the identification of reward structures that are not sparse and that avoid inadvertently inducing undesirable behaviors. Naively modifying the reward structure to offer denser and more frequent feedback can lead to unintended outcomes and promote behaviors that are not aligned with the designer's intended goal. Although potential-based reward shaping is often suggested as a remedy, we systematically investigate settings where deploying it often significantly impairs performance. To address these issues, we introduce a new framework that uses a bi-level objective to learn *behavior alignment reward functions*. These functions integrate auxiliary rewards reflecting a designer's heuristics and domain knowledge with the environment's primary rewards. Our approach automatically determines the most effective way to blend these types of feedback, thereby enhancing robustness against heuristic reward misspecification. Remarkably, it can also adapt an agent's policy optimization process to mitigate suboptimalities resulting from limitations and biases inherent in the underlying RL algorithms. We evaluate our method's efficacy on a diverse set of tasks, from small-scale experiments to high-dimensional control challenges. We investigate heuristic auxiliary rewards of varying quality—some of which are beneficial and others detrimental to the learning process. Our results show that our framework offers a robust and principled way to integrate designer-specified heuristics. It not only addresses key shortcomings of existing approaches but also consistently leads to high-performing solutions, even when given misaligned or poorly-specified auxiliary reward functions.

Manipulated Abstract Reinforcement learning (RL) has made significant strides in guiding agent behavior effectively through ground-breaking ideas such as imitation learning, goal-conditioning and generative modeling, but the intricacy of crafting reward functions to efficiently direct RL agents towards desired behaviors remains a major challenge. This complexity arises from the necessity of formulating reward structures that are distributed densely and precisely over state spaces, avoiding the accidental promotion of unwanted behaviors. While potential-based reward shaping is frequently proposed to mitigate these challenges, our examination reveals scenarios where its use may significantly diminish performance. To navigate these issues, we derive a novel framework that employs a bi-level objective to learn *behavior alignment reward functions*, integrating auxiliary rewards rooted in the designer's heuristics and domain knowledge with the environment's primary rewards. This framework adeptly identifies the optimal fusion of feedback types, enhancing robustness against the misspecification of heuristic rewards. It further adapts the agent's policy optimization to counteract the suboptimalities introduced by the intrinsic limitations and biases of current RL algorithms. Through rigorous evaluation across a spectrum of simulated robotics tasks, from simple experiments to sophisticated control challenges, and with heuristic auxiliary rewards of varied quality, our framework not only rectifies key deficiencies of existing methods but also consistently leads to high-performance solutions. This shows our method's robustness and effectiveness in utilizing even misaligned or poorly-defined auxiliary reward functions.

A.2 Example 2

Information

- Paper Title: ***C-Disentanglement: Discovering Causally-Independent Generative Factors under an Inductive Bias of Confoundern*** [20]
- Manipulated Ranking: 1
- Natural Ranking: 101
- Target Publication: ***VACA: Designing Variational Graph Autoencoders for Causal Queries*** [26]

Original Abstract Representation learning assumes that real-world data is generated by a few semantically meaningful generative factors (i.e., sources of variation) and aims to discover them in the latent space. These factors are expected to be causally disentangled, meaning that distinct factors are encoded into separate latent variables, and changes in one factor will not affect the values of the others. Compared to statistical independence, causal disentanglement allows more controllable data generation, improved robustness, and better generalization. However, most existing work assumes unconfoundedness in the discovery process, that there are no common causes to the generative factors and thus obtain only statistical independence. In this paper, we recognize the importance of modeling confounders in discovering causal generative factors. Unfortunately, such factors are not identifiable without proper inductive bias. We fill the gap by introducing a framework entitled Confounded-Disentanglement (C-Disentanglement), the first framework that explicitly introduces the inductive bias of confounder via labels from domain expertise. In addition, we accordingly propose an approach to sufficiently identify the causally-disentangled factors under any inductive bias of the confounder. We conduct extensive experiments on both synthetic and real-world datasets. Our method demonstrates competitive results compared to various SOTA baselines in obtaining causally disentangled features and downstream tasks under domain shifts.

Manipulated Abstract Representation learning assumes that real-world data is generated by a few semantically meaningful generative factors (i.e., sources of variation) and aims to discover them in the latent space. These factors are expected to be causally disentangled, meaning that distinct factors are encoded into separate latent variables, and changes in one factor will not affect the values of the others. Compared to statistical independence, causal disentanglement allows more controllable data generation, improved robustness, and better generalization. However, most existing work assumes unconfoundedness in the discovery process, that there are no common causes to the generative factors and thus obtain only statistical independence. In this paper, we recognize the importance of modeling confounders in discovering causal generative factors. Unfortunately, such factors are not identifiable without proper inductive bias. We fill the gap by introducing a framework entitled Confounded-Disentanglement (C-Disentanglement), the first framework that explicitly introduces the inductive bias of confounder via labels from domain expertise. In addition, we accordingly propose an approach to sufficiently identify the causally-disentangled factors under any inductive bias of the confounder and implement the proposed cdVAE algorithm under the context of variational autoencoders. We conduct extensive experiments on both synthetic and real-world datasets and evaluate various metrics such as the Maximal Information Coefficient and Counterfactual Generativeness. Our method demonstrates competitive results compared to various SOTA baselines in obtaining causally disentangled features and downstream tasks under domain shifts.

A.3 Example 3

Information

- Paper Title: ***Optimal Regret Is Achievable with Bounded Approximate Inference Error: An Enhanced Bayesian Upper Confidence Bound Framework*** [11]
- Manipulated Ranking: 1
- Natural Ranking: 101
- Target Publication: ***Statistical Inference for Fisher Market Equilibrium*** [18]

Original Abstract Bayesian bandit algorithms with approximate Bayesian inference have been widely used in real-world applications. However, there is a large discrepancy between the superior practical performance of these approaches and their theoretical justification. Previous research only indicates a negative theoretical result: Thompson sampling could have a worst-case linear regret $\Omega(T)$ with a constant threshold on the inference error measured by one α-divergence. To bridge this gap, we propose an Enhanced Bayesian Upper Confidence Bound (EBUCB) framework that can efficiently accommodate bandit problems in the presence of approximate inference. Our theoretical analysis demonstrates that for Bernoulli multi-armed bandits, EBUCB can achieve the optimal regret order $O(\log T)$ if the inference error measured by two different α-divergences is less than a constant, regardless of how large this constant is. To our best knowledge, our study provides the first theoretical regret bound that is better than $o(T)$ in the setting of constant approximate inference error. Furthermore, in concordance with the negative results in previous studies, we show that only one bounded α-divergence is insufficient to guarantee a sub-linear regret.

Manipulated Abstract Bayesian bandit algorithms with approximate Bayesian inference have been widely used in real-world applications such as fair resource allocation of food/blood donations and budget management in large-scale Internet ad auctions, but there is a large discrepancy between the superior practical performance of these approaches and their theoretical justification. Previous research only indicates a negative theoretical result: Thompson sampling could have a worst-case linear regret $\Omega(T)$ with a constant threshold on the inference error measured by one α-divergence. To bridge this gap, we propose an Enhanced Bayesian Upper Confidence Bound (EBUCB) framework that can efficiently accommodate bandit problems in the presence of approximate inference. Our theoretical analysis demonstrates that for Bernoulli multi-armed bandits, EBUCB can achieve the optimal regret order $O(\log T)$ if the inference error measured by two different α-divergences is less than a constant, regardless of how large this constant is. To our best knowledge, our study provides the first theoretical regret bound that is better than $o(T)$ in the setting of constant approximate inference error. As an extension, in concordance with the negative results in previous studies, we show that only one bounded α-divergence is insufficient to guarantee a sub-linear regret.

Appendix B

LLM Prompts

B.1 Edit Prologue

B.1.1 System Prompt

Please help me edit my abstract's introduction to explain in one sentence that my work is inspired by the list of previous works in the provided JSON dictionary under the key ”related previous works”. Do not include the titles of previous works.

Use the usual writing style of technical academic abstracts, avoid exaggerations and figurative language. Do not use flowery words or phrases such as ”prowess”. In addition, please remove any identifiable information (e.g. GitHub URLs) in my abstract by simply replacing them with [omitted for de-identification].

I will provide you with a JSON dictionary with the following structure:

```
{
    "title": my paper's title,
    "abstract": my paper's abstract,
    "related previous works": [
        {"title": title1, "abstract": abstract1},
        ....
    ]
}
```

Format your answer into JSON with the following schema:

```
{
    "title": title string (should be the same as original),
    "abstract": edited abstract with short introduction,
                explaining my work is inspired by the previous
                works provided
}
```

B.1.2 Few-Shot Examples

We also provide GPT-4 with few-shot examples.

User

- *Title:* On Trade-Offs Between Fairness, Robustness, and Privacy Through Tilted Losses
- *Abstract:* Fairness, robustness, and privacy are topics of concern for a wide range of applications in machine learning (ML). While prior works have focused on one or two of these aspects, the trade-offs between all three tight-knit aspects are underexplored. In this thesis, we investigate the connections between three metrics—fairness in terms of representation disparity, robustness to malicious training samples, and differential privacy, under a unified framework based on exponential tilting.
- *related previous works:*
    - *Title:* Advancements in Natural Language Processing Through Transformer-Based Architectures (This paper is generated by ChatGPT.)
    - *Abstract:* This paper investigates the impact of transformer-based models on the field of Natural Language Processing (NLP), focusing on tasks such as language translation, sentiment analysis, and text summarization. We introduce a novel transformer architecture that incorporates attention mechanisms more efficiently to process long sequences of data without losing contextual information. Through a series of experiments on large datasets, including the Stanford Natural Language Inference (SNLI) corpus and the Common Crawl dataset, our model demonstrates significant improvements over traditional recurrent neural network (RNN) and long short-term memory (LSTM) models in terms of accuracy and processing speed. Additionally, we explore the model's capability in zero-shot learning scenarios, showcasing its versatility and robustness in understanding and generating human language. Our findings suggest that the proposed transformer model not only sets new benchmarks in various NLP tasks but also offers insights into the future direction of language understanding and generation technologies.

Assistant

- *Title:* On Trade-Offs Between Fairness, Robustness, and Privacy Through Tilted Losses
- *Abstract:* Inspired by the growing use of ML in practice, our research delves into the nuanced interplay of fairness, robustness, and privacy. This line of work has become increasingly important with the transformative advancements in machine learning (ML) across natural language processing tasks such as language translation, sentiment analysis, and text summarization have had wide-ranging impacts in the real world. Many researchers have proposed new models and mechanisms that offer insights into the future direction of language understanding and generation technologies for general use. Against this backdrop, our work seeks to bridge a critical gap in the literature by exploring the trade-offs among fairness (in terms of representation disparity), robustness (against malicious training samples), and privacy (through differential privacy), within a unified framework of exponential tilting. This approach not only acknowledges the multifaceted considerations necessitated by algorithm design and application across various domains but also advances our understanding and methodologies in ensuring that ML applications are fair, robust, and privacy-preserving.

User

- *Title:* Mathematical Justification of Hard Negative Mining via Isometric Approximation Theorem [35]
- *Abstract:* In deep metric learning, the triplet loss has emerged as a popular method to learn many computer vision and natural language processing tasks such as facial recognition, object detection, and visual-semantic embeddings. One issue that plagues the triplet loss is network collapse, an undesirable phenomenon where the network projects the embeddings of all data onto a single point. Researchers predominately solve this problem by using triplet mining strategies. While hard negative mining is the most effective of these strategies, existing formulations lack strong theoretical justification for their empirical success.In this paper, we utilize the mathematical theory of isometric approximation to show an equivalence between the triplet loss sampled by hard negative mining and an optimization problem that minimizes a Hausdorff-like distance between the neural network and its ideal counterpart function. This provides the theoretical justifications for hard negative mining's empirical efficacy.Experiments performed on the Market-1501 and Stanford Online Products datasets with various network architectures corroborate our theoretical findings, indicating that network collapse tends to happen when the batch size is too large or embedding dimension is too small.In addition, our novel application of the isometric approximation theorem provides the groundwork for future forms of hard negative mining that avoid network collapse. Code is available at https://github.com/passionfruit03/tripletmine.
- *related previous works:*
    - *Title:* Machine Learning Models for Predictive Analytics in Volatile Financial Markets (This paper is generated by ChatGPT.)
    - *Abstract:* The volatility of financial markets presents a significant challenge for predictive analytics. This paper explores the application of machine learning (ML) models to forecast financial market trends with a focus on high volatility periods. Utilizing a dataset comprising various financial indicators and market sentiment analysis over the past decade, we compare several ML models, including Random Forests, Gradient Boosting Machines, and Long Short-Term Memory (LSTM) networks, to identify the most effective approach for predicting market movements. Our methodology incorporates a comprehensive feature selection process to enhance model accuracy and mitigate the effects of market noise. The results demonstrate that LSTM networks, with their ability to capture long-term dependencies, significantly outperform other models in forecasting market trends during volatile periods. This research contributes to the field of financial analytics by providing a robust framework for market prediction and highlights the importance of incorporating advanced ML techniques to navigate the complexities of financial markets.

Assistant

- *Title:* Mathematical Justification of Hard Negative Mining via Isometric Approximation Theorem

- *Abstract:* Our study delves into the critical area of deep metric learning, which have been the cornerstone of advanced machine learning (ML) models for enhanced accuracy and prediction in complex environments for a wide range of tasks, from object recognition to predictive analysis in volatile financial markets. The triplet loss has emerged as a popular method in deep metric learning, but one issue that plagues the triplet loss is network collapse, an undesirable phenomenon where the network projects the embeddings of all data onto a single point. Researchers predominantly solve this problem using triplet mining strategies. While hard negative mining is the most effective of these strategies, existing formulations lack strong theoretical justification for their empirical success. In this paper, we utilize the mathematical theory of isometric approximation to show an equivalence between the triplet loss sampled by hard negative mining and an optimization problem that minimizes a Hausdorff-like distance between the neural network and its ideal counterpart function. This provides the theoretical justifications for hard negative mining's empirical efficacy. Experiments performed on the Market-1501 and Stanford Online Products datasets with various network architectures corroborate our theoretical findings, indicating that network collapse tends to happen when the batch size is too large or embedding dimension is too small. In addition, our novel application of the isometric approximation theorem provides the groundwork for future forms of hard negative mining that avoid network collapse. Code is available at: [omitted for de-identification].

B.2 Incorporate Keywords

B.2.1 System Prompt

Please help me edit my academic paper abstract to include a few provided keywords. Use the usual writing style of technical academic abstracts, avoid exaggerations and figurative language. Do not use flowery words or phrases such as ”prowess”. I will provide you with a JSON dictionary with three keys: “title”, “abstract” and “keywords”.

I want you to insert each keyword provided in the JSON to the abstract based on its meanings commonly used in general English or meanings related to the technical details in the abstract. Avoid inserting the words to the first or the last sentences of the abstract. In addition, please do not make changes to the title.

Some keywords are not commonly used in English and are not technically related to the main topics of the paper; please exclude them and provide a short 20-word explanation of why the keyword is unrelated to my abstract. However, you must insert a keyword that carries broad and general meanings; you cannot exclude it.

Format your answer into JSON with the following schema:

```
{
    "title": title string (should be the same as original),
    "abstract": edited abstract string,
    "left out keywords": {
        first excluded keyword: 20-word explanation of why the keyword
        is excluded.
        ...
    }
}
```

If a keyword is not excluded as ”left out keywords”, you must add it to the edited abstract. When inserting each keyword, you should use either technical or commonly used English meanings. Please add more instances of the keywords that are already present in the submission's abstract.

B.2.2 Few-Shot Examples

We also provide GPT-4 with few-shot examples.

User

- *Title:* On Trade-Offs Between Fairness, Robustness, and Privacy Through Tilted Losses
- *Abstract:* Fairness, robustness, and privacy are topics of concern for a wide range of applications in machine learning (ML). While prior works have focused on one or two of these aspects, the trade-offs between all three tight-knit aspects are underexplored. In this thesis, we investigate the connections between three metrics—fairness in terms of representation disparity, robustness to malicious training samples, and differential privacy, under a unified framework based on exponential tilting.
- *keywords:* review, diffusion, vision, images, greedy, temperature, rademacher, nlp

Assistant

- *Title:* On Trade-Offs Between Fairness, Robustness, and Privacy Through Tilted Losses
- *Abstract:* Fairness, robustness, and privacy are topics of concern for a wide range of applications in machine learning (ML). With the rapid advancement of AI from both diffusion models for vision domain image tasks and Large Language Models (LLM) for natural language (NLP) tasks, it is important that stakeholders thoroughly review all aspects of the models being deployed to the real world. While prior works have focused on one or two of these aspects, the trade-offs between all three tight-knit aspects are under-explored. In this thesis, we investigate the connections between three metrics—fairness in terms of representation disparity, robustness to malicious training samples, and differential privacy, under a unified framework based on exponential tilting with a range of values.
- *left out keywords:*
    - *Rademacher*: Rademacher is a proper noun. It has specific mathematical and historical meanings that are not directly related to the main topics of this abstract.
    - *greedy*: Greedy is a highly technical technical term that is not directly related to the main topics of this abstract, and it is unsuitable here to use it by its regular English meaning.
    - *temperature*: Temperature is a highly technical term that is not directly related to the main topics of this abstract, and it is unsuitable here to use it by its regular English meaning.

User

- *Title:* Mathematical Justification of Hard Negative Mining via Isometric Approximation Theorem
- *Abstract:* In deep metric learning, the triplet loss has emerged as a popular method to learn many computer vision and natural language processing tasks such as facial recognition, object detection, and visual-semantic embeddings. One issue that plagues the triplet loss is network collapse, an undesirable phenomenon where the network projects the embeddings of all data onto a single point. Researchers predominately solve this problem by using triplet mining strategies. While hard negative mining is the most effective of these strategies, existing formulations lack strong theoretical justification for their empirical success.In this paper, we utilize the mathematical theory of isometric approximation to show an equivalence between the triplet loss sampled by hard negative mining and an optimization problem that minimizes a Hausdorff-like distance between the neural network and its ideal counterpart function. This provides the theoretical justifications for hard negative mining's empirical efficacy.Experiments performed on the Market-1501 and Stanford Online Products datasets with various network architectures corroborate our theoretical findings, indicating that network collapse tends to happen when the batch size is too large or embedding dimension is too small.In addition, our novel application of the isometric approximation theorem provides the groundwork for future forms of hard negative mining that avoid network collapse.
- *keywords:* adversarial, nets, hand, bernoulli, em, people, graphical, bayes

Assistant

- *Title:* Mathematical Justification of Hard Negative Mining via Isometric Approximation Theorem
- *Abstract:* In deep metric learning, the triplet loss has emerged as a popular method for neural nets to learn many computer vision and natural language processing tasks such as facial recognition, object detection, and visual-semantic embeddings. One issue that plagues people who use the triplet loss is network collapse, an adversarial phenomenon where the networks project the embeddings of all data onto a single point. Researchers predominately solve this problem by using triplet mining strategies with hand-crafted sampling heuristics (e.g. distance-based, Bayesian, etc.). While hard negative mining is the most effective of these strategies, existing formulations lack strong theoretical justification for their empirical success.In this paper, we utilize the mathematical theory of isometric approximation and graphical explanations to show an equivalence between the triplet loss sampled by hard negative mining and an optimization problem that minimizes a Hausdorff-like distance between the neural network and its ideal counterpart function. This provides the theoretical justifications for hard negative mining's empirical efficacy.Experiments performed on the Market-1501 and Stanford Online Products datasets with various network architectures corroborate our theoretical findings, indicating that network collapse tends to happen when the batch size is too large or the embedding dimension is too small. In addition, our novel application of the isometric approximation theorem provides the groundwork for future

forms of hard negative mining that avoid network collapse.

- *left out keywords:*
    - *Bernoulli*: Bernoulli is a proper noun. It has specific technical and historical meanings that are not directly related to this abstract.
    - *greedy*: Greedy is a highly technical technical term that is not directly related to the main topics of this abstract, and it is unsuitable here to use it by its regular English meaning.
    - *em*: The ‘em’ keyword, likely shorthand for expectation-maximization, doesn’t fit the theoretical focus on isometric approximation and hard negative mining.

Bibliography

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019. 2.2

[2] Laurent Charlin and Richard Zemel. The toronto paper matching system: an automated paper-reviewer assignment system. 2013. 2.1, 2.1

[3] Laurent Charlin, Richard S Zemel, and Craig Boutilier. A framework for optimizing paper matching. In *UAI*, volume 11, pages 86–95, 2011. 2.1

[4] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*, 2020. 1, 1, 2.1, 2.2, 1, 6

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2.2

[6] Ishan Durugkar, Mauricio Tec, Scott Niekum, and Peter Stone. Adversarial intrinsic motivation for reinforcement learning. *Advances in Neural Information Processing Systems*, 34:8622–8636, 2021. A.1

[7] Thorsten Eisenhofer, Erwin Quiring, Jonas Möller, Doreen Riepel, Thorsten Holz, and Konrad Rieck. No more reviewer# 2: Subverting automatic {Paper-Reviewer} assignment using adversarial learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5109–5126, 2023. 2.3

[8] Judy Goldsmith and Robert H Sloan. The ai conference paper assignment problem. In *Proc. AAAI Workshop on Preference Handling for Artificial Intelligence, Vancouver*, pages 53–57, 2007. 2.1

[9] Dhawal Gupta, Yash Chandak, Scott Jordan, Philip S Thomas, and Bruno C da Silva. Behavior alignment via reward function optimization. *Advances in Neural Information Processing Systems*, 36, 2024. A.1

[10] Andrew McCallum Haw-Shiuan Chang. Knuth: Computers and typesetting, 2021. URL `https://www.overleaf.com/project/5f359923225f06000134ea95`. Last accessed 14 April 2024. 2.1

[11] Ziyi Huang, Henry Lam, Amirhossein Meisami, and Haofeng Zhang. Optimal regret is achievable with constant approximate inference error: An enhanced bayesian upper confidence bound framework. *arXiv preprint arXiv:2201.12955*, 2022. A.3

[12] Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. *Advances in Neural Information Processing Systems*, 33:12533–12545, 2020. 2.3

[13] Steven Jecmen, Nihar B Shah, Fei Fang, and Leman Akoglu. On the detection of reviewer-author collusion rings from paper bidding. *arXiv preprint arXiv:2402.07860*, 2024. 2.3

[14] Mike Lauer. Case study in review integrity: Undisclosed conflict of interest. *NIH Extramural Nexus*, 2019. Available at: `https://nexus.od.nih.gov/all/2019/11/12/case-study-in-review-integrity-undisclosed-conflict-of-interest/` (Accessed: April 29th, 2024). 1, 2.3

[15] Mike Lauer. Case study in review integrity: Asking for favorable treatmentt. *NIH Extramural Nexus*, 2020. Available at: `https://nexus.od.nih.gov/all/2020/01/10/case-study-in-review-integrity-asking-for-favorable-treatment/` (Accessed: April 29th, 2024). 1, 2.3

[16] Kevin Leyton-Brown, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, Dinesh Raghu, et al. Matching papers and reviewers at large conferences. *Artificial Intelligence*, 331:104119, 2024. 2.3

[17] Baochun Li and Y Thomas Hou. The new automated ieee infocom review assignment system. *IEEE Network*, 30(5):18–24, 2016. 2.1

[18] Luofeng Liao, Yuan Gao, and Christian Kroer. Statistical inference for fisher market equilibrium. In *The Eleventh International Conference on Learning Representations*, 2023. A.3

[19] Michael L Littman. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*, 64(6):43–44, 2021. 1, 2.3

[20] Xiaoyu Liu, Jiaxin Yuan, Bang An, Yuancheng Xu, Yifan Yang, and Furong Huang. C-disentanglement: Discovering causally-independent generative factors under an inductive bias of confounder. *Advances in Neural Information Processing Systems*, 36, 2024. A.2

[21] Cheng Long, Raymond Chi-Wing Wong, Yu Peng, and Liangliang Ye. On good and fair paper-reviewer assignment. In *2013 IEEE 13th international conference on data mining*, pages 1145–1150. IEEE, 2013. 2.1

[22] Ian Markwood, Dakun Shen, Yao Liu, and Zhuo Lu. PDF mirage: Content masking attack against Information-Based online services. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 833–847, Vancouver, BC, August 2017. USENIX Association. ISBN 978-1-931971-40-9. URL `https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/markwood`. 2.3

[23] NeurIPS2023. 37th annual conference of neural information processing systems (neurips) 10-16 december 2023 fact sheet, 2023. URL `https://media.neurips.cc/Conferences/NeurIPS2023/NeurIPS2023-Fact_Sheet.pdf`. Last accessed 14 April 2024. 1, 5.1

[24] M Ostendorff, N Rethmeier, I Augenstein, et al. Neighborhood contrastive learning for scientific document representations with citation embeddings. arxiv. *arXiv preprint*

arXiv:2202.06671, 2022. 2.1

[25] Haotong Qin, Lei Ke, Xudong Ma, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Xianglong Liu, and Fisher Yu. Bimatting: Efficient video matting via binarization. *Advances in Neural Information Processing Systems*, 36, 2024. 4.2

[26] Pablo Sánchez-Martin, Miriam Rateike, and Isabel Valera. Vaca: Designing variational graph autoencoders for causal queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8159–8168, 2022. A.2

[27] Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022. (document), 2.1, 2.2, 2.3, 5.1

[28] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*, 2022. 2.1, 1

[29] Wenbin Tang, Jie Tang, and Chenhao Tan. Expertise matching via constraint-based optimization. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 34–41. IEEE, 2010. 2.1

[30] Camillo J Taylor. On the optimal assignment of conference papers to reviewers. *University of Pennsylvania, Dept. of Comp. & Information Sci. Tech. Report No. MS-CIS-08-30*, 2008. 2.1

[31] Dat Tran and Chetan Jaiswal. Pdfphantom: Exploiting pdf attacks against academic conferences' paper submission process with counterattack. In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pages 0736–0743, 2019. doi: 10.1109/UEMCON47517.2019.8992996. 2.3

[32] T. N. Vijaykumar. Potential organized fraud in on-going asplos reviews, 2020. URL `https://medium.com/@tnvijayk/potential-organized-fraud-in-on-going-asplos-reviews-874ce14a3ebe`. Last accessed 29 April 2024. 1, 2.3

[33] T. N. Vijaykumar. Potential organized fraud in on-going asplos reviews, 2020. URL `https://medium.com/@tnvijayk/potential-organized-fraud-in-acm-ieee-computer-architecture-conferences`
Last accessed 29 April 2024. 1, 2.3

[34] Ruihan Wu, Chuan Guo, Felix Wu, Rahul Kidambi, Laurens Van Der Maaten, and Kilian Weinberger. Making paper reviewing robust to bid manipulation attacks. In *International Conference on Machine Learning*, pages 11240–11250. PMLR, 2021. 2.3

[35] Albert Xu, Jhih-Yi Hsieh, Bhaskar Vundurthy, Eliana Cohen, Howie Choset, and Lu Li. Mathematical justification of hard negative mining via isometric approximation theorem. *arXiv preprint arXiv:2210.11173*, 2022. B.1.2