No more Reviewer #2: Subverting Automatic Paper-Reviewer Assignment using Adversarial Learning

Thorsten Eisenhofer, Erwin Quiring, Jonas Möller, Doreen Riepel, Thorsten Holz, Konrad Rieck

RUHR UNIVERSITÄT BOCHUM

RUB

ICSI INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Technische Universität Berlin

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Papers and Reviews

Peer Review

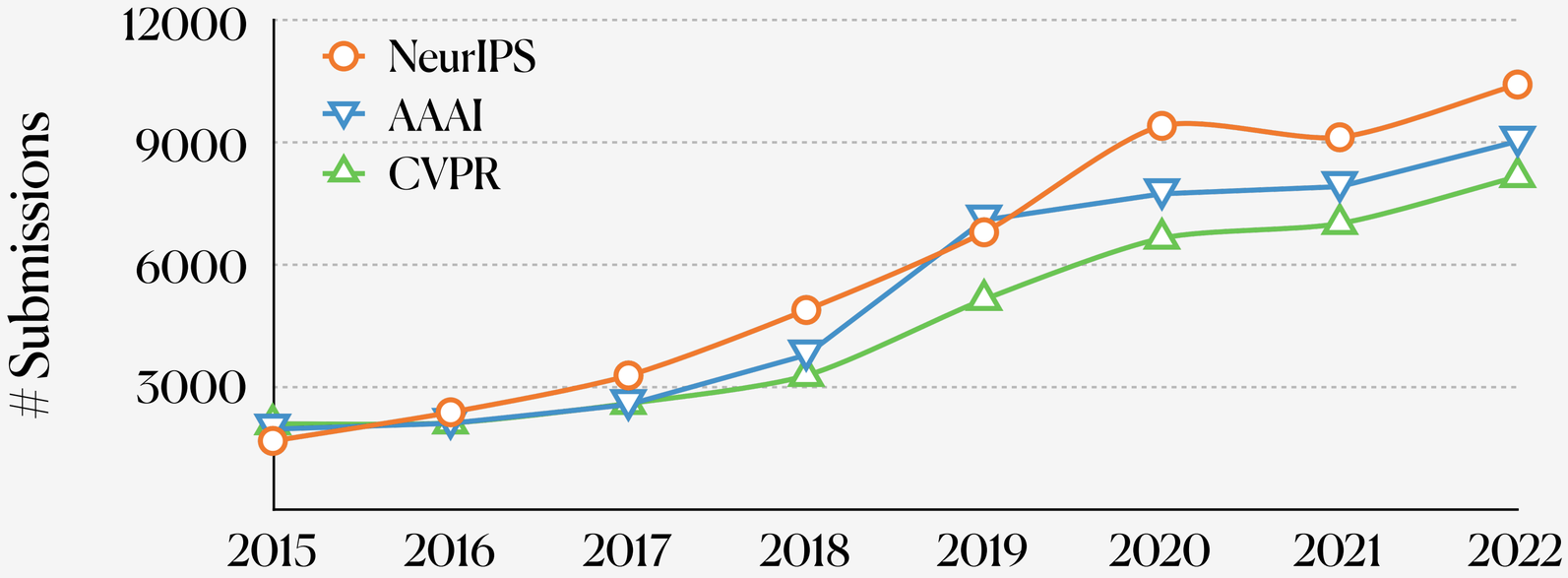
- Independent evaluation of scientific papers
- Main instrument for quality control

Initial Step: Paper-Reviewer Assignment

- Assignment of qualified reviewers to each paper
- Good match of topic (paper) and expertise (reviewer)

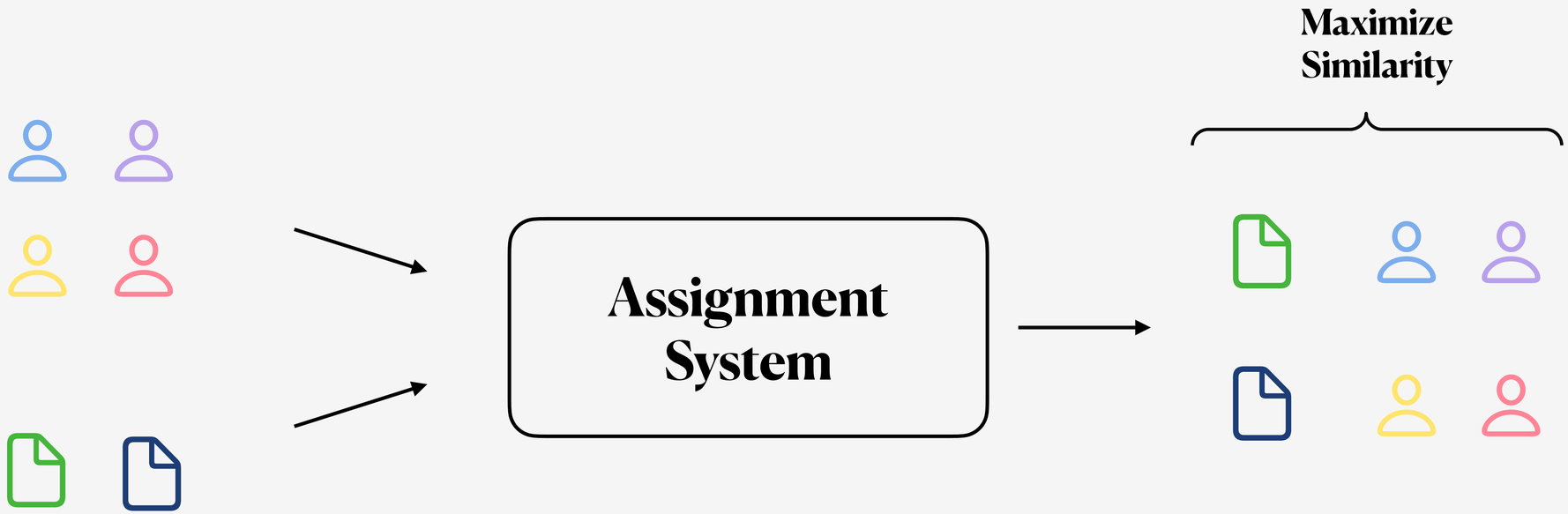
Assignment Process

10.000 submissions.
Reading each paper's
title (~3s) takes 8 hours!



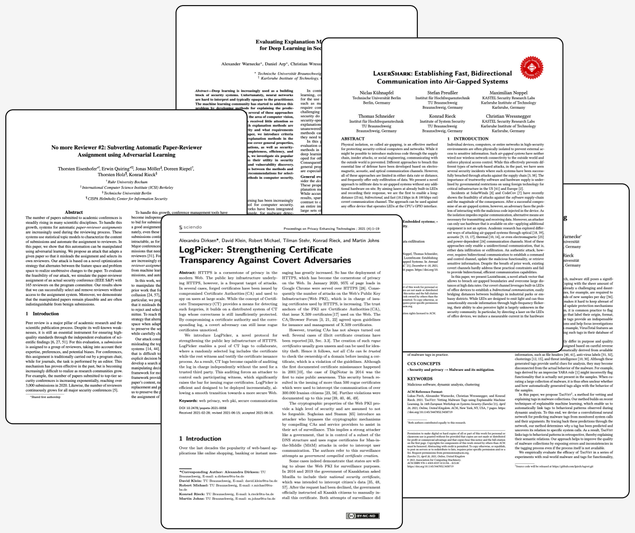
Manual bidding increasingly impossible

Automatic Assignment Systems



Use ML to distill submissions and reviewer expertise

Topic Modeling



$\mathbf{Corpus\ D} = \{$ $,$ $, ..., $ $\}$

Pre-processing

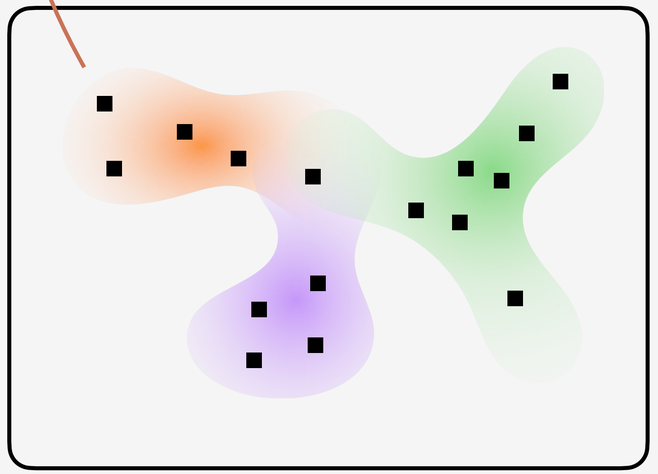
$\Phi(\rho(\square))$

Feature Extractor

61× “attack” → $x =$ [... 61 ... 76 ... 0 ...]

Vocabulary

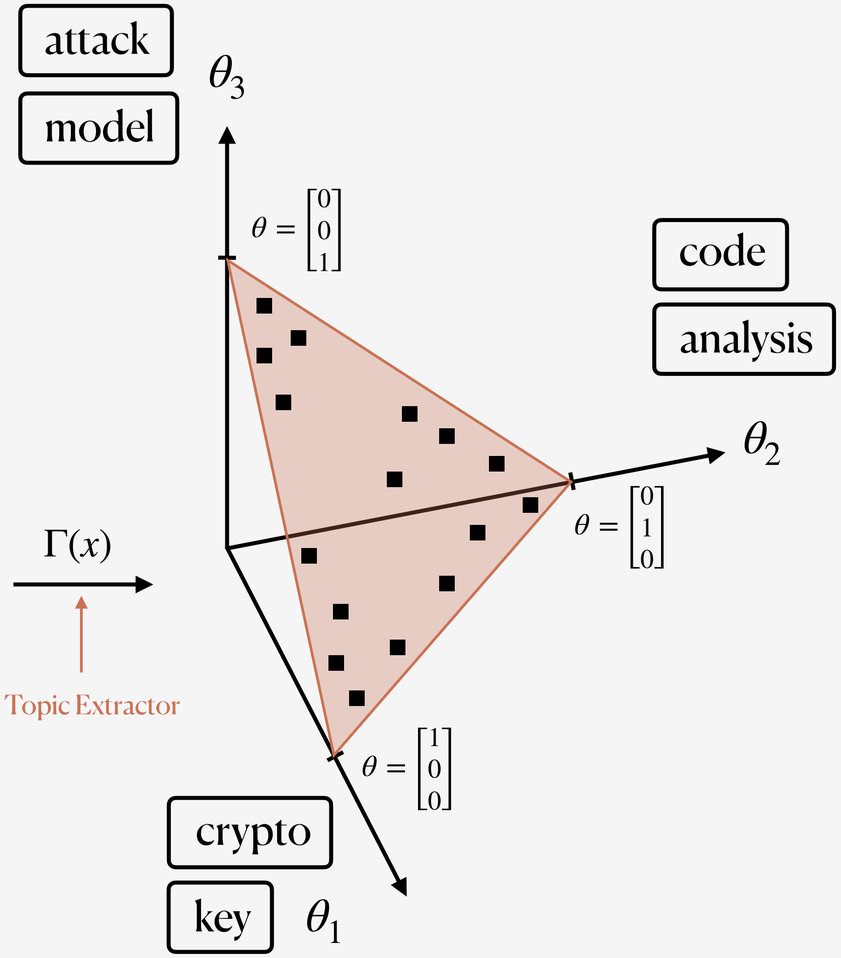
$\in \mathbb{N}_0^{|V|}$



Feature Space

$\Gamma(x)$

Topic Extractor



Topic Space

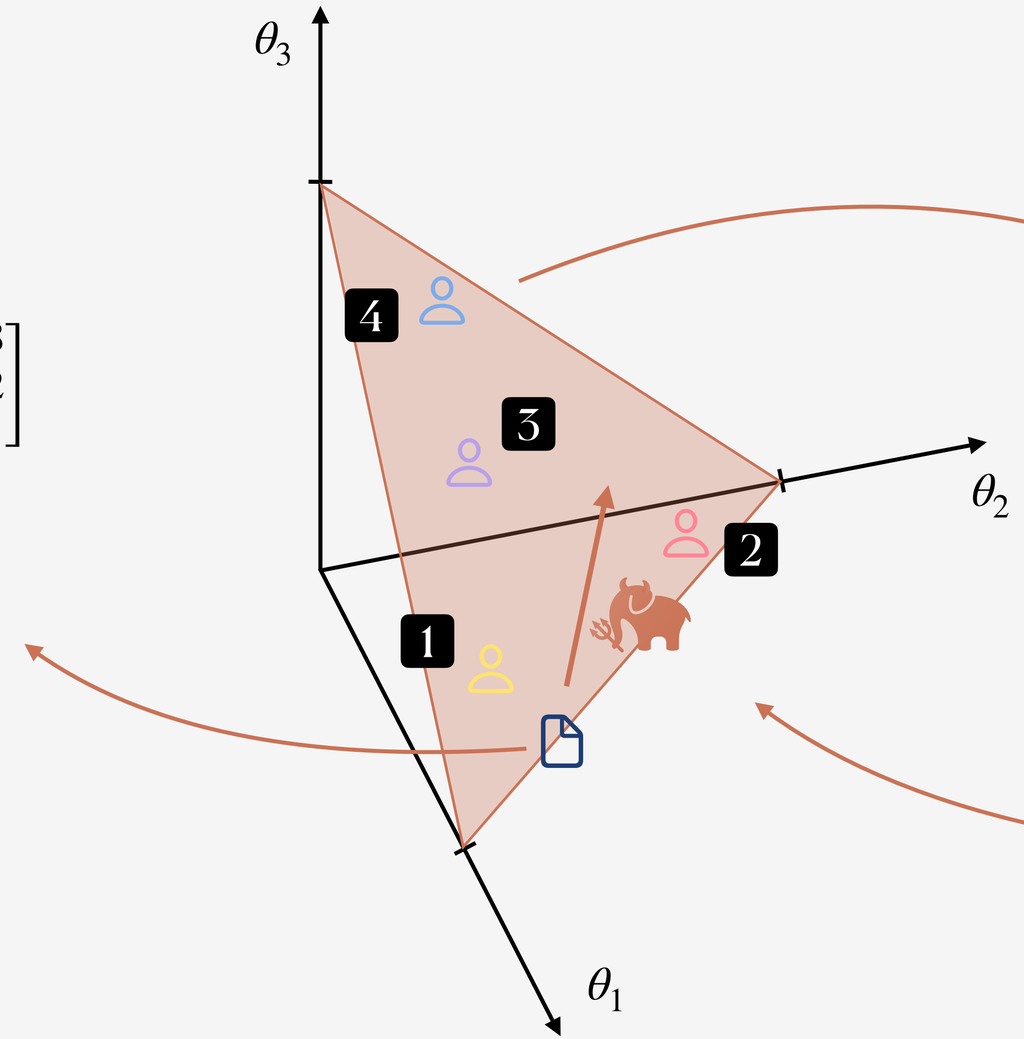
Topic Modeling

Submission Topics $\theta_{\text{doc}} = \begin{bmatrix} 0.8 \\ 0.2 \\ 0 \end{bmatrix}$

80% key crypto

20% code analysis

0% attack model



Reviewer Topics $\theta_{\text{rev}} = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix}$

10% key crypto

20% code analysis

70% attack model

Need to project changes back into the problem space!

Goal: Manipulate submission to pick our own reviewers

Problem-space

Problem-space transformations to add/remove words from input file

Format-/ and encoding-level

Hidden Box

Homoglyphs

u+0061 u+0430

a ≠ a

Text-level

Reference addition

Language models

Synonyms

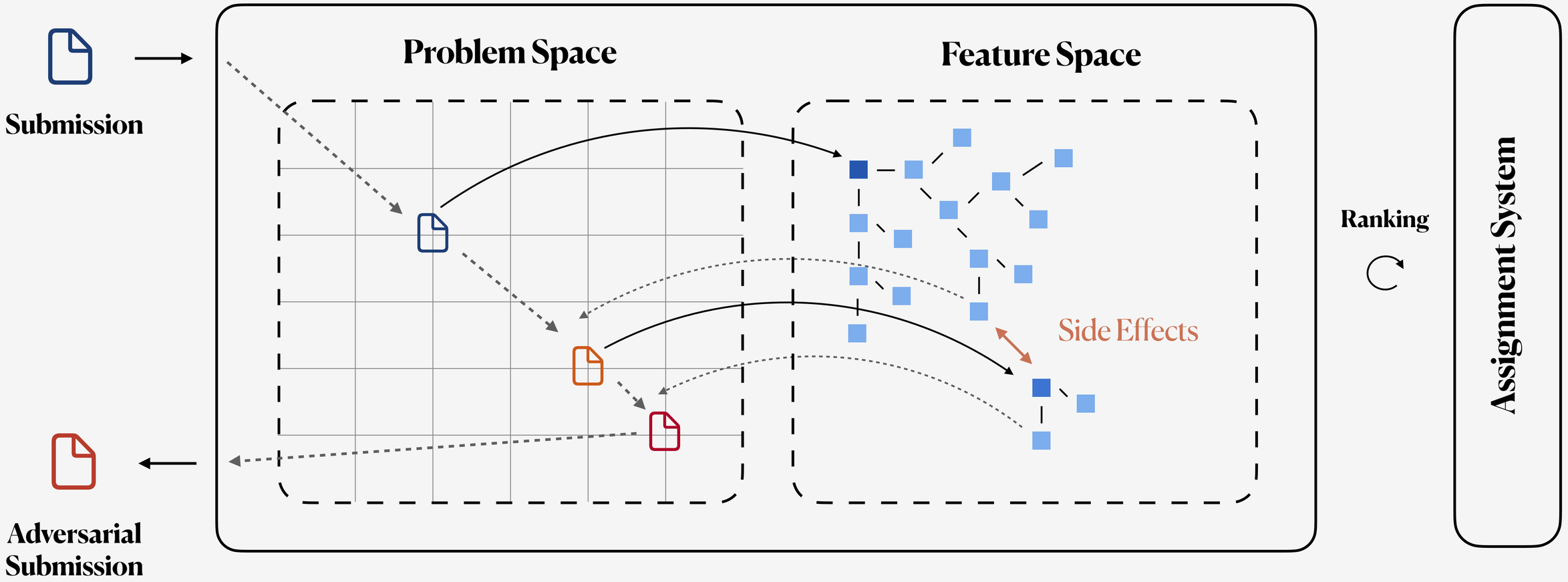
Spelling mistakes

Chain several transformations

Constraints

is plausible and semantic correct

Hybrid Search Strategy



Evaluation

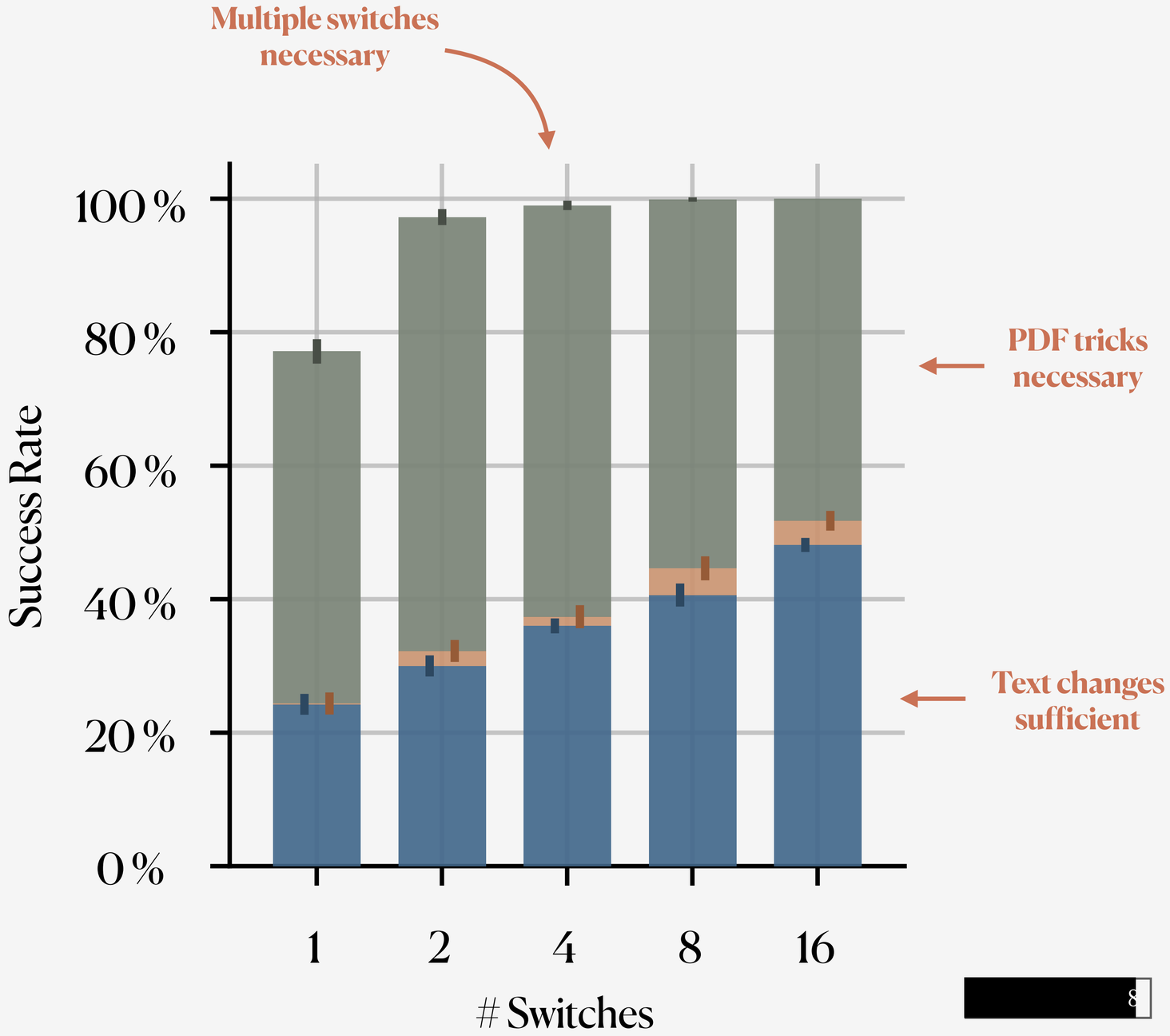
Simulation of IEEE S&P' 20

Format-level

Text-level

Encoding-level

More results in the paper!



Take Aways

New attack against automatic reviewer-paper assignment

- Hybrid attack strategy in feature space and problem space
- Minimal and unobtrusive transformations of papers

Broader perspective

- Decisions based on learning models inherently insecure
- More to explore off the beaten path of adversarial learning

More at github.com/rub-syssec/adversarial-papers

Thank you!

Fin