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

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

Manual bidding increasingly impossible

10,000 submissions. Reading each paper’s title (~3s) takes 8 hours!

# Submissions

- NeurIPS
- AAAI
- CVPR
Automatic Assignment Systems

Use ML to distill submissions and reviewer expertise
**Topic Modeling**

Corpus $D = \{, \, \ldots, \}$

Feature Space

- Pre-processing $\Phi(\rho(D))$
- Feature Extractor

Vocabulary $x = \begin{bmatrix} 61 \\ \vdots \\ 76 \\ \vdots \\ 0 \\ \vdots \end{bmatrix} \in \mathbb{N}^{|V|}_0$

Feature Space

- Topic Extractor $\Gamma(x)$

Topic Space

- $\theta_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$
- $\theta_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$
- $\theta_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

- attack
- model
- code
- analysis
- crypto
- key
Goal: Manipulate submission to pick our own reviewers.

**Submission Topics** \( \theta_b = \begin{bmatrix} 0.8 \\ 0.2 \\ 0 \end{bmatrix} 

80% key crypto  
20% code analysis  
0% attack model

**Reviewer Topics** \( \theta_e = \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!
Problem-space

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

Format-/ and encoding-level

<table>
<thead>
<tr>
<th>Hidden Box</th>
<th>u+0061</th>
<th>u+0430</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homoglyphs</td>
<td>a ≠ a</td>
<td></td>
</tr>
</tbody>
</table>

Text-level

<table>
<thead>
<tr>
<th>Reference addition</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language models</td>
<td>Spelling mistakes</td>
</tr>
</tbody>
</table>

Chain several transformations

Constraints

is plausible and semantic correct
Hybrid Search Strategy

Submission

Adversarial Submission

Problem Space

Feature Space

Side Effects

Assignment System

Ranking
Evaluation

Simulation of IEEE S&P’ 20

- Format-level
- Text-level
- Encoding-level

More results in the paper!

Success Rate

# Switches

Multiple switches necessary

PDF tricks necessary

Text changes sufficient

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!