ADVERSARIAL PHISHING ATTACKS & DEFENCES

Prepared by : Archana Sreekumar
Mentored by : Dr Krishnasree Achutan
Collaborator : Gilad Gressel

PROBLEM STATEMENT

What is the problem?
- Adversarial attack on phishing detection models.

Why is it a problem?
- Model fails to protect user from a phishing attack.
- Billions of users are prone to phishing.
- Billions of user accounts compromised.

Technically why does it exist?
- The model can be used as an oracle to launch black box attack.
- Adversary has all the information about the model in-order to launch a white box attack.
- Data injection: Injecting adversarial samples into training data.
- Data Modification: Modifying the data using adversarial examples.
- Logic corruption: Tampering the model algorithm.

Why should you care about the problem?
- Model misclassifies a potential phishing page/URL.
- Can cause legitimate page /URL classified as a phish.
- Stealing of sensitive information.

SOLUTIONS

What are the state-of-art solutions & why are they inadequate?

- Adversarial training
  - The defense is not robust for black-box attacks where an adversary generates malicious examples on a locally trained substitute model.
- Obfuscation of code
  - Deobfuscation techniques to reverse engineer the obfuscated code.
- Feature squeezing
  - Accuracy of model is affected

PROPOSALS

What are our ideas to solve the problem?

- Methods to detect hidden characters.
- Safeguarding the code used.
- Using different forms of keywords(singular/plural).

Thank You