Paper Summary
Detecting Credential Spearphishing Attacks in Enterprise Settings

Sarthak | 2020CS10379

June 2024

Problem Statement

The paper addresses the challenge of detecting credential spearphishing attacks in enterprise settings without relying on email headers, which often result in high false positive rates

Motivation

- Spearphishing requires no technical expertise, doesn't depend on any particular vulnerability, and hence frequently succeeds
- Access to sensitive systems via stolen credentials can lead to substantial breaches, especially with the prevalent use of remote desktops, VPNs, and cloud services
- High false positive rates and insufficient labeled data for training traditional machine learning models due to the rarity of successful attacks render existing methods ineffective

Key Idea

To analyze the fundamental characteristics of spearphishing attacks to derive targeted features and form a feature vector for each email and then apply an unsupervised, non-parametric technique (DAS) for anamoly detection

Framework & Methodology

Feature Extraction for emails

- Lure Features (Attacker impersonates a trusted source) - Depending on the type of lure, features are extracted based on:

  - Detection of new IP address
  - Number of prior logins by the user from the geographic location
  - Number of other employees who have logged in from this location

- Exploit features (User clicks the malicious link) - Features are extracted combining factors like
  - number of prior visits to hostname across all enterprises' network traffics
  - number of days between first employee's visit to the hostname and current email

Leveraging the features for detection - Directed Anomaly Scoring(DAS)

1. Specify B = Alert budget (Number of alerts you are willing to process each day)
2. For each email, assign a suspiciousness score
   - Score(Event X) = number of other events that are as benign (non-suspicious) as X in each dimension
   - Large score = Few other emails are more suspicious than X or X is one of the most suspicious emails
3. Rank events by their suspiciousness score
4. Output the B most suspicious events

Contributions

- Developed a technique to fragment the email into lure and exploitation and extract feature vectors from them to characterize each mail
- Developed an unsupervised, non-parametric, non direction agnostic machine learning model to accurately detect attacks within specific Budget constraint with minimal false positive rates

Strengths

- Rate for True Positives = 89% + detection of previously unnoticed attacks was achieved
- False Positives rate of less than 0.004% was achieved
- Overcomes the requirement of hyperparameter tuning of traditional unsupervised learning

- Immune to direction agnostic and hence generalizable to needle-in-haystack problems with directional features

Weaknesses & Limitations

- HTTPS traffic is not intercepted due to privacy concerns, which means the system will miss spearphishing attacks involving links to HTTPS websites
- Adversaries could boost the reputation of their domains or sender emails by slowly building up a fake email's reputation, thus evading detection
- The detector's effectiveness relies on having a significant amount of historical data (at least 3 months)
- Since the detector has an upper bound on the number of alerts per day, it can give false negatives on days with many attacks