CERIAS

The Center for Education and Research in Information Assurance and Security

PURDUE UNIVERSITY

Schema and Script based analysis to identify spam email attacks

REU Student: Shauna Lynn Soto (ssoto2@emich.edu)
Professors: Dr. Victor Raskin & Dr. Julia Taylor

Problem Statement

As both users and researchers have seen, determining spam email and preventing the attack of spam email before reaching the users is a difficult task. Many techniques in research and within business have been developed to try to address the issue. The goal of this project is to develop schema- and script- based analysis to help understand the ways spam email is developed and find common correlations so machines can better identify spam email and possibly enhance current techniques.

Common Ground	***Authentication of Digital Communication***
Problem Space	***Acknowledging Spam email based on characteristics***
Specific Problem	***Analysis of text body in emails to determine if it Spam based or Authentic***

Exploring The World Of Spam

Spam has been an ongoing task for industry to resolve. Many techniques have been developed to resolve them. These areas included:

- **Email Header Analysis**
- **Link Analysis**
- **Keyword Analysis**
- **Link + Keyword Analysis**

Even with ongoing development in spam research, and legislation like the CAN-SPAM Act of 2003 we continue to see spam make it into user inboxes.

The goal of this research so far has been a two step development. Step 1) is to create a schema from spam email to find possible common links found in spam email. Step 2) deep review of the body of a spam email to try to develop a script to help machines better detection of spam email.

Method

Stage 1: Spam Review and Schema Creation
The objective was to create a schema based on information found with a set spam emails within a spam email corpus. The goal was to find common issues within spam email, and immediately it was easy to see that most spam email when viewed in straight HTML shows the laziness, and/or knowledge base of the spammer. The spammer fails in meeting best practices that are commonly known in industry

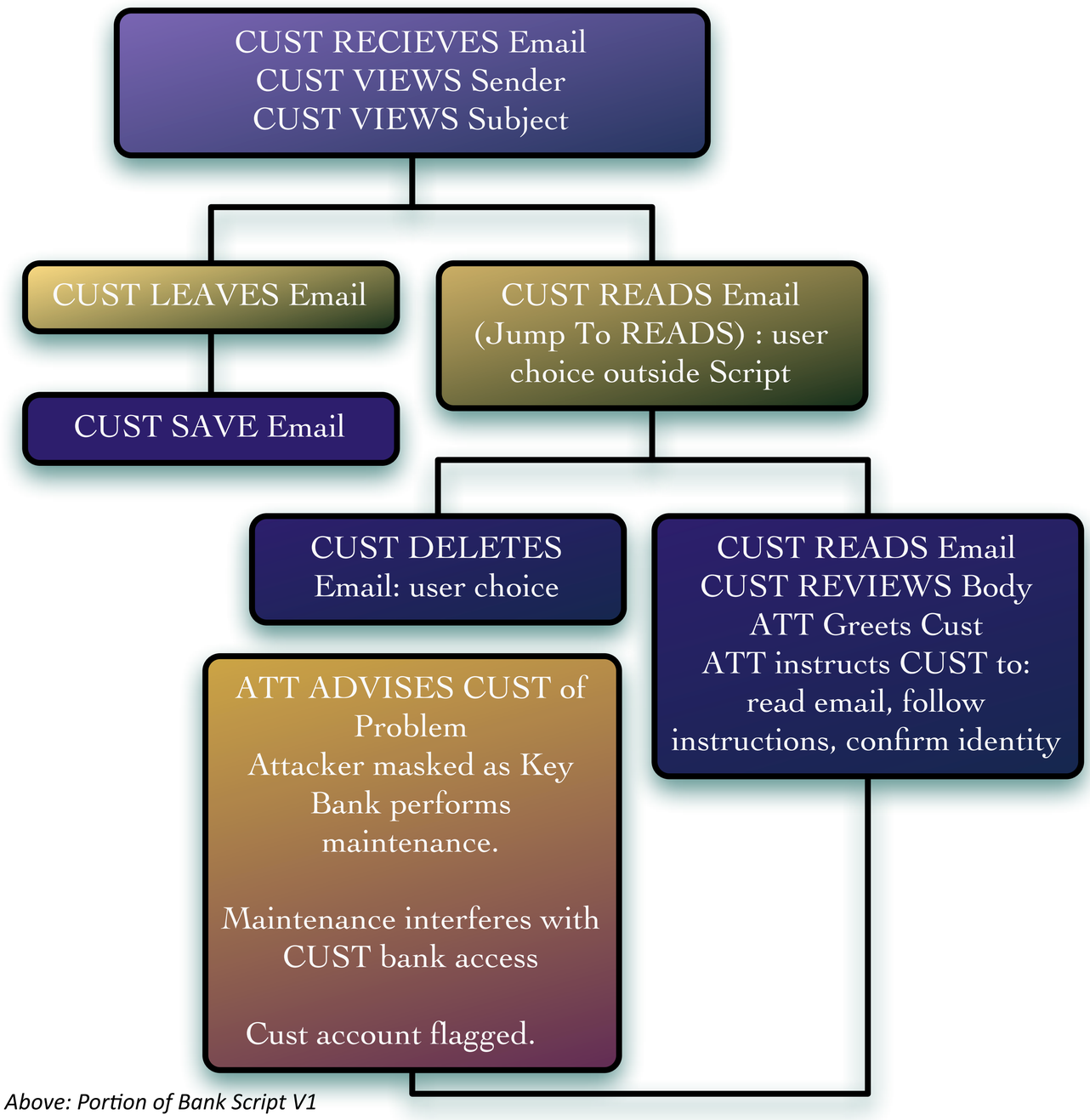
HTML
- Title
  - Default : Default, New1, ect
  - Empty : lacks title tag
  - Missing

Clip Art
- Image
  - Included : Art included in the email
  - Server Based : Art for email is being pulled from webserver

Image Included
- Filename
  - Related Name : ex) chaselogo.gif for Chase bank
  - Unrelated Name : ex) brad.gif for a bank

<Strong>
- Bolded Text
  - Salutation : Emphases on Salutation
  - InBody : Word, Sentence, or Phrase in Emphases

Stage 2: Developing Script
Re-Review over set of emails first reviewed and start to establish a situational based script for how the spammer creates email. The goal is to establish a bottom up approach when doing analysis and script creation. That way we will start out at a very specific level and seek to establish overtime and general script for how spammer manipulates within a spam email.



Above: Portion of Bank Script V1

Current Results And Conclusions

I am breaking down related scripts about Banks, Ebay, etc. Once that has been established it, will need to be compared. If this does resolve spam then I would suspect spammer would seek to employ a level of sincerity by doing large scale spear-phishing attacks on groups within social networks. By attack with this method the spammer can be more establish in the content and better tailor emails.

This material is based upon work supported by the National Science Foundation under grant #1062970 Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

CERIAS

Discovery Park e-Enterprise Center