Arthur Liou
CS373

Week 8 Writeup

Prompt: Submitting a write-up of your thoughts, impressions, and any conclusions based on the material from the week. Each week will have its own assignment in the grades page.

For the first part of this week's writeup, I'm reflecting on the topic – Messaging Security. I really loved this week. Not as many slides, but also there was a lot of content and lab work in this week, which I really like for learning because I could play around with the materials and content without having any expectations (group work/hw type) set upon what I'm learning. Some of the terminology and tools I knew and had used before, but there were much more that I didn't know before. It was great learning more about this from a educational cybersecurity standpoint rather than a personal and professional viewpoint. All in all, an excellent week of material!

For what we covered and learned, see my lecture notes below.

Lecture Notes

Lesson 1 – Messaging Security

- Phishing Quiz
- Terminology - Spam/Ham, Strap/Honeypot, Botnet, Snowshoe spam, Phishing vs Spear Phishing, RBL, Heuristics, Bayesian (Statistical), Fingerprinting/Hashing
- History and Evolution of Spam, botnets
- Technology to combat spam: Engines
  - Reputation-driven: IP, message, URL
  - Content-driven: common strings, fixed strings vs variable strings (regular expression), message attributes, combo of strings and attributes
- Tools for Messaging Data – Linux tools, open-source DBs, regex coach, trustedsource.org, spamhaus.org
- Tools for research purposes – Dig (Domain Information Groper, WHOIS
- Demo 1 – Postgres – Exploration.
- Demo 2 – Regex Coach + Group Practice
- Research Techniques for managing the data flood: Samples metadata: Parsing, Grouping, Aggregation, ID of outliers
- Considerations – Human input required, fully automated, combination of auto and human input?, probability scoring vs additive scoring
- Lab 1 – Data Exploration. Total, distinct, average, types of files and their extensions
- Lab 1 – Representative Delegation
- Nominate an individual to present the group's analysis and findings at the end of Lab 2
- Recap; Key Concepts – Data Model – Spam/Ham

Lesson 2 – Lecture Wrap and Classification Lab

- SMTP Conversation – Ham, Spam, Email Header Reading
- Data Model – Spam, Ham
- The Data Scientific Method
- 1. Start with data.
- 2. Develop intuitions about the data and the questions it can answer.
- 3. Formulate your question.
- 4.Leverage your current data to better understand if it is the right
- question to ask. If not, iterate until you have a testable hypothesis.
- 5. Create a framework where you can run tests/experiments.
- 6. Analyze the results to draw insights about the question.
- Classification Lab - The provided message_data table has 100k rows of real-world message meta data. Use the tools and techniques covered to make spam/ham decisions for all records
- SQL Examples. Useful Operators:
- COUNT()
- DISTINCT()
- SPLIT_PART()
- GROUP BY $col
- ORDER BY $col
- Spam is pervasive - Digital & Printed media, Audio/Visual
- Many aspects of Security can be reduced to finding the least common denominator among large data sets
- Automate "Finding the needle"
- Classification accuracy is directly tied to the depth in which we are able to describe samples
- Tools
- Spamhaus RBL
- McAfee RBL
- The Regex Coach
- Trustedsource.org
- Domaintools.net
- Reputationauthority.org
- Yougetsignal.com/tools/web-sites-on-web-server/
- Spamassassin.apache.org
- PostgreSQL