Security Intelligence
Data Mining

University of Amsterdam
System & Network Engineering (MSc)

Research Project 1

Diana Rusu
Nikolaos Petros Triantafyllidis

Friday, 30/01/2015
Research Question

• How can we effectively use public sources to obtain real time information about security incidents?
Why?

• Critical company infrastructure & confidential data protection

• Recent events…
What?

• A way to use public sources to collect security intelligence

• A system that collects, processes and analyses data to create security incident alerts
System Outline

• Modular and pipelined configurable system

• Data aggregation, preprocessing, analytics, alerting, assessment

• Let’s take a look at each part
Configuration

- Define list of common security threats
- List of clients
- List of Software/Hardware vendors, models, tools
- Attach list of keywords with associated importance (weight) to each threat as well as a list of sources
Example

- DDoS:
  keywords: {[DDoS, 1.5], [Attack, 0.2], [Holland, 0.1], [Strawberries, 0.8]}
- Phishing:
  keywords: {[Phishing, 2], [Site, 0.2], [Nicolas Cage, 0.1], [Jazz, 0.8]}
- Malware
  keywords: {[Virus, 2], [Worm, 1.5], [Trojan, 1.2], [Punk, 0.1]}
- Exploits:
  keywords: {[Exploit, 2], [Bug, 2], [Vulnerability, 1.5], [Nachos, 0.3]}
- Etc.
Example

- DDoS
  sources: {{www.alltheddosdiscussion.nl, targeted], [www.pastebin.com, general]}

- Phishing
  sources: {{www.phishtank.com, targeted], [www.pastebin.com, general]}

- Malware
  sources: {{www.aplacethatviriiilivein.org, targeted], [www.pastebin.com, general]}

- Exploits:
  sources: {{www.allthecves.gov, targeted], [www.reddit.com/r/blackhat, general]}

- Etc.
Example

Vendors: {Oracle, IBM, HP, Microsoft, Apple, Canonical, Ubuntu}

Clients: {Deloitte, ING, AMRO, Rabobank, DUWO}
Aggregation

• Extensible plugin oriented module

• Configurable to include authentication keys, execution intervals, source files, etc.

• Must perform initial cleaning of the data (surrounding HTML, profile pics, etc.)
Demo

- Twitter, Reddit, Pastebin, Phishtank
- Scrapy, Tweepy, Standard Python libs
- RESTful/Streaming APIs, traditional scraping
- 1078827 tweets, 23050 pastes, 6325 reddit posts, 25051 phishing websites
Preprocessing & Warehousing

- Deflate raw data according to the specified keywords
- Score each document according to the occurrence of important words
- Apply certain threshold to store only interesting data to warehouse
Demo

- Input: Bull***t data
- Output:

- DDoS:
  
  {keywords: [[DDoS,4],[Attack,10],[Holland:1]], vendors:[], clients: [[ING,2]], doc_id:1, score: 17}
Mining & Analytics

• Data Mining Tasks
  • Anomaly Detection
  • Association Rule
  • Classification
  • Clustering
  • Regression
  • Summarisation
Association Rule

• Finding relationships between different subsets in the database

• Apriori algorithm
Demo

• Orange Library
Clustering

• Discovering previously unknown structures or groups of subsets that present similarities within the dataset

• k-means algorithm
Demo

• SKlearn Library

Cluster 13: paris charlie hebdo attack http mayor nypd rt victims french honor visited nyc france terror muslim bolsters security jewish cover

Cluster 15: photos leaked upton kate jennifer lawrence nude victoria justice megan fox http seen rt hilarious kardashian kim

Cluster 18: attack http rt titan panic bus hotel tel aviv killed terror amp deadly anxiety people terrorist tripoli video uses
Classification

- Generalising known structures and applying them to new data.
- Requires training set (supervised learning)
Demo

• SKlearn Library

• Support Vector Machine
Anomaly Detection

• Identifying unusual data records that might imply unusual behaviour in the dataset

• Requires training set (supervised learning)

• Example: Unusually high mention of company name in DDoS database
Regression

• Defining a function which models the data with the least error
Summarisation

- Providing a more compact representation of the data set, including visualisation and report generation
- Extracting information and alerting
Example

{alert_id: 00001, subject: "Something is rotten in the state of Denmark", importance: Red, backing_documents:[1,3,4,6]}
Feedback & Reconfiguration

• Assessment provided by Security Analysts to determine the actual threat level

• Calculate the divergence from the estimated threat level

• Propagate to the adjustable parts of the system and reconfigure
Conclusions

• Aggregated Data from 4 sources (~2GB)
• Designed and tested a deflation and scoring model
• Tested some Data Mining tasks and algorithms that produced promising results
• Proposed alerting and feedback extensions
• Each submodule can be integrated in one Software Suite
• We believe that this attempt is feasible
Future Work

• System Implementation
• Integration with CTI-portal
• Testing the algorithms with real world data
• Build training sets
• Implement more Data Mining algorithms
• Explore NLP capabilities
Future Work

• System Extensions:
  • Real Time Alerting
  • Sentiment Analysis
  • Support for non Latin alphabets
  • Automatic Learning & Adjustment
Tools & Sources

- http://www.toonpool.com/user/5624/files/sony_playstation_hack_attack_1262435.jpg
- http://www.scrapy.org
- http://www.tweepy.org
- http://www.python.org
- http://docs.orange.biolab.si/reference/rst/index.html
Thx!

Qs?