Dissecting Non-malicious Artifacts: One IP At A Time

Ido Naor (@idonaor1) & Dani Goland (@danigoland)
Disclaimer

A few people from companies X, Y & Z gave us a “friendly” call to not expose information linking back to them. Hence, we had to redact the juicy content of our presentation.
Employees use online services to educate themselves. Security products are not supposed to use online services to enhance detection. Vendors use data of their clients and later have zero responsibility for it. Hackers steal data and later publish it to online platforms.
Research Scope

Garbage

Gold
Research Goal

Goal
Prove that data is being unwillingly exfiltrated from organizations, and that with simple tools – it can be exposed.

Starting points
- Paid/non-paid services
- Repositories
- Other

Building Blocks
- Malware Research
- Yara
- Data Science
Proof of Concept

• One simple search: “message”
• Mail containing non-malicious artifacts
  • Options:
    • Mail uploaded by employee
    • Mail suspected by security product
    • Mail uploaded by 3rd party
• Example of company intellectual property being leaked
Research Subjects

- Emails
- Code
- Dumps
- Archives
- Documents
- Certificates
- Keys
- Secrets
- Graphs
Step 1: Possible Feeds

- Code repositories / Open source
- Script/Paste repositories
- Malware repositories
- Multi-scanners
- Online sandboxes
- Forums / Social platforms
Step 2: Yara rules

```
rule filter_interesting_eml {
  strings:
    $eml_01 = "From:"
    $eml_02 = "To:"
    $eml_03 = "Subject:"
    $attachment_id = "X-Attachment-Id"
    $mime_type = "Content-Type: multipart/mixed"
  condition:
    all of ($eml_*) and $attachment_id and $mime_type and filesize > 300KB
}
```

```
rule filter_interesting_outlook {
  strings:
    $a1 = {D0 CF 11 E0 A1 B1 1A E1}
    $a2 = "FROM:" nocase
    $a3 = "TO:" nocase
  condition:
    all of them and filesize > 300KB
}
```
Step 3: Collection

- Searched feeds for undetected samples
- List hashes from results
- Downloaded undetected samples
- Filtered emails using Yara
- Sent matches to pipeline

```javascript
var https = require('https');
var AWS = require('aws-sdk');
AWS.config.region = 'us-west-2';
var s3 = new AWS.S3();
let url = "https://<datasource>/...";
exports.handler = function (event, context, callback) {
    https.get(url, res => {
        res.setEncoding("utf8");
        let body = "";
        res.on("data", data => {
            body += data;
        });
        let ids = []
        res.on("end", () => {
            body = JSON.parse(body);
            let notifs = body['notifs'];
            let filtered = {alerts:[]};
            for(let i = 0; i<notifs.length;i++)
            {
                ...
            }
        })
    });
```
General Architecture

- RSS feeds
- Pull Hashes
- Send SNS
- AWS Lambda
- SNS
- Push
- Queue
- Get Tasks
- ec2 instances
- Headers + Parsed Body + Metadata
- Samples & Attach.
- Download Samples
- Samples & Attach.
- Bucket (S3)
- Postgres DB
- EMR
- Elastic Search & Kibana
Analysis Pipeline

Email Queue
- Spam Or Ham
  - TF-IDF → Naïve Bayes
  - Open Source Dataset

Attachment Queue
- Custom Parser
  - Subject & Body
  - Language Detection
  - Domain Names
  - Geo lookup on domains
  - Extract Attachments

Content & Metadata Extraction

NLP Pipeline

Preprocessing
- Stemming
- Stop word Removal
- Tokenization
- POS Tagging

Named Entity Recognition
- Names Of Persons
- Organizations
- Roles (CXO, VP)
- Locations
- Monetary Values

Topic Modeling
- 10 Topics
- Top 10 words/topic

Kibana

NMF = Non-Negative Matrix Factorization
LDA = Latent Dirichlet Allocation
Graph-based
Global Problem
DEMO TIME
QUESTIONS?

(1) Please 🙋

(2) Grab the 🎤

(3) Stand up & speak English

VirusBay Invite Code: SECTSWEDEN2018

https://beta.virusbay.io/
THANK YOU!

VirusBay

VirusBay Invite Code: SECTSWEDEN2018

https://beta.virusbay.io/