Extracting & Exploring Threat Intel on Open Sourced Documents using Natural Language Processing

Mayo YAMASAKI
NTT-CERT, NTT Secure Platform Labs
Overview

Developing a Threat Knowledge Extraction System by Using NLP.

Set of Unstructured Documents

RIG Exploit Kit targets Adobe Flash Player exploit (CVE-2015-8651).

Knowledge Graph

Malware
RIG Exploit Kit

Product
Adobe Flash Player

targets

CVE
CVE-2015-8651

attributed to
About NTT-CERT

- The CSIRT of NTT group
- Department of NTT Secure Platform Labs
- Activities
  - Incident Response
  - Product Evaluation
  - Forensics & Malware Analysis
  - Vulnerability Reporting
  - Training & Education
  - OSINT
  - Etc.

www.ntt-cert.org
What is OSINT?

“Open-source intelligence (OSINT) is intelligence that is produced from publicly available information and is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement”

- United States Department of Defense
Our OSINT Activities

Passive Info Source
- News/Release
- Vendor Report
- Vulnerability & Attack Alert

Active Info Source
- BBC / SNS
- Personal Blog
- Other Info

Quality Requirement Group Review
- Collector
- Collector
- Collector

Output Requirement Group Review
- Reporter
- Reporter
- Reporter

Service Requirement
- PoC
- WEB Portal

Constituency
Problem

- Collecting & Storing **Unstructured Documents**
  
  - Hard to Search and Understand Threat Intel
  - Dependency on Knowledge of Team Members
Solution

Automatic Threat Knowledge Extraction System

1. Crawler
   Collect pages & extract text

2. Knowledge Extractor
   Extract structure using NLP

3. Error Checker
   Check extraction error by human

4. Knowledge Database
   Store knowledge in graph DB

5. Knowledge Viewer
   Search & show knowledge
Knowledge Extraction in NLP

Set of Unstructured Documents

RIG Exploit Kit targets Adobe Flash Player exploit (CVE-2015-8651).

Knowledge Graph

```
Malware
  RIG Exploit Kit

Product
  Adobe Flash Player

CVE
  CVE-2015-8651

attributed to

targets
```
Related Work
### Knowledge Extraction Tasks

#### Tasks of Non Security Domain

<table>
<thead>
<tr>
<th>Task</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC-4 '92</td>
<td>Attack, Kidnapping, Bombing, Arson, etc.</td>
</tr>
<tr>
<td>CoNLL '03</td>
<td>Person, Organization, Location, MISC.</td>
</tr>
<tr>
<td>BioNLP '11</td>
<td>Gene Expression, Protein catabolism, etc.</td>
</tr>
<tr>
<td>ScienceIE '17</td>
<td>Process, Task, Material.</td>
</tr>
</tbody>
</table>

#### Tasks of Security Domain

<table>
<thead>
<tr>
<th>Task</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joshi+ '13</td>
<td>Software, Hardware, Attack mean, etc.</td>
</tr>
<tr>
<td>Jones+ '15</td>
<td>Software, Vender, CVE, Version, etc.</td>
</tr>
<tr>
<td>Ramnani+ '17</td>
<td>Vulnerability, Thret Actor, IoC, TTP, etc.</td>
</tr>
</tbody>
</table>
Our Approach

Previous Threat Knowledge Extraction

- Low accuracy
- Hard to evaluate accuracy

Our Approach by using Supervised Learning

- High accuracy
- Easy to evaluate accuracy
Our Approach
Our Task Overview

- **Input**
  - Set of Sentences

- **Output**
  - Knowledge Graph

- **Three Sub Tasks**
  - Entity Extraction
  - Relation Extraction
  - Combining Graphs

\[ S_i \text{ is a sentence} \]

**Task Overview**

\[
\begin{align*}
S_1 & \rightarrow \text{Entity Extractor} \\
S_2 & \rightarrow \text{Relation Extractor} \\
S_3 & \rightarrow \text{Combiner} \\
\end{align*}
\]

Knowledge Graph
How to Define Threat Structure

- Using STIX 2.0 for Knowledge Extraction
  - Adaptation for Ambiguous Structure in Text
    - **Missing Data** (e.g. Unknown Identity)
    - **Required Binary Relation** (e.g. Unstructured Property)

Input Sentence:

\[ X \text{ campaign targets a governmental organization} \]

Desirable Output:

![Diagram showing the relationship between Campaign, targets, and Industry]
Entity Extraction

- Extracting Subsequences of Words as Entities

<table>
<thead>
<tr>
<th>RIG</th>
<th>EK</th>
<th>targets</th>
<th>Adobe</th>
<th>Flash</th>
<th>Player</th>
<th>exploit</th>
<th>(CVE-2015-8651)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malware</td>
<td>Malware</td>
<td>O</td>
<td>Product</td>
<td>Product</td>
<td>Product</td>
<td>O</td>
<td>O Cve O O</td>
</tr>
</tbody>
</table>

- Multiclass Classification
- Entity Classes:
  { AttackPattern, Campaign, Cve, Domain, Hash, Identity, Industry, Ip, Malware, Product, Region, Role, ThreatActor, Time, Version, O }
## Relation Extraction

- Extracting Relation between Entities

<table>
<thead>
<tr>
<th>entity1</th>
<th>entity2</th>
<th>relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rig Exploit Kit</td>
<td>Adobe Flash Player</td>
<td>targets</td>
</tr>
<tr>
<td>Rig Exploit Kit</td>
<td>CVE-2015-8651</td>
<td>targets</td>
</tr>
<tr>
<td>Adobe Flash Player</td>
<td>Rig Exploit Kit</td>
<td>0</td>
</tr>
<tr>
<td>Adobe Flash Player</td>
<td>CVE-2015-8651</td>
<td>0</td>
</tr>
<tr>
<td>CVE-2015-8651</td>
<td>Rig Exploit Kit</td>
<td>0</td>
</tr>
<tr>
<td>CVE-2015-8651</td>
<td>Adobe Flash Player</td>
<td>attributed-to</td>
</tr>
</tbody>
</table>

- Multiclass Classification
- Relation Classes:
  - {attributed-to, aliases, indicate, observed-in, uses, targets}

![Task Overview Diagram]

---

Copyright©2017 NTT corp. All Rights Reserved.
Combining Graphs

Combining Graphs Extracted from each Sentences

<table>
<thead>
<tr>
<th>ID</th>
<th>relation(arg1, arg2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>targets(Rig Exploit Kit, Adobe Flash Player)</td>
</tr>
<tr>
<td>2</td>
<td>targets(Rig Exploit Kit, CVE-2015-8651)</td>
</tr>
<tr>
<td>3</td>
<td>attributed-to(CVE-2015-8651, Adobe Flash Player)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>arg1</th>
<th>arg2</th>
<th>Is combining?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID1</td>
<td>ID2</td>
<td>YES</td>
</tr>
<tr>
<td>ID1</td>
<td>ID2</td>
<td>NO</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Binary Classification
- Classes: Same Entity or Not

Task Overview

- Entity Extractor
- Relation Extractor
- Combiner
- Knowledge Graph
Developing Labeled Data

- Labeling 200 WEB documents (about 10,000 sentences)
- Labeling documents by 5 peoples using a tool
Policy for Labeling Documents

- Creating a Guideline Document with Case Studies
  - E.g.1 Masked domains are labeled as domain.
    
    /reallstatistics[..]info/Domain
  
  - E.g.2 Malware types aren't attack pattern.
    
    /Key logging/AttackPattern
    /Keylogger/O

- Force Restriction
  - E.g. Relations are defined only between specific entities by "brat" annotation tool.

- Checking All Labeled Data by Supervisor

- Hiring Cyber Security Domain Experts
Stats of Labeled Data

Labeled Data for Entity Extraction

- Word Count
- Entity Count
- Entity Type Count

Labeled Data for Relation Extraction

- Relation Count
- Relation Type

Copyright©2017 NTT corp. All Rights Reserved.
Experiments & Results
# Experiment of Entity Extraction

## Extraction by CRF (Conditional Random Field)

### Input

<table>
<thead>
<tr>
<th>RIG</th>
<th>Exploit</th>
<th>Kit</th>
<th>targets</th>
<th>Adobe</th>
<th>Flash</th>
<th>Player</th>
<th>exploit (CVE-2015-8651)</th>
</tr>
</thead>
</table>

### Feature

- $X_0$
- $X_{t-1}$
- $X_t$
- $X_{t+1}$
- $X_T$

### Predication

<table>
<thead>
<tr>
<th>Malware</th>
<th>Malware</th>
<th>Malware</th>
<th>O</th>
<th>Product</th>
<th>?</th>
</tr>
</thead>
</table>

### Output

- Hyper Parameters: Decision by Random Search
Result of Entity Extraction

- Average of 3 F Scores of Predicting Labels for each Words
- Training and Developing Model by 80% of Dataset
- Testing Model by Rest 20% of Dataset
Experiments of Relation Extraction

- Extraction by Linear SVM (Support Vector Machine)

**Sentence**

<table>
<thead>
<tr>
<th>RIG</th>
<th>Exploit</th>
<th>Kit</th>
<th>targets</th>
<th>Adobe</th>
<th>FLash</th>
<th>Player</th>
<th>exploit (CVE-2015-8651)</th>
</tr>
</thead>
</table>

**Input Pair**
(Rig Exploit Kit, Adobe Flash Player)

**Feature**

\[ X_0, X_1, X_2, \ldots, X_N \]

**Predication**

\[ y \]

**Output**

**targets**

- Features (Rink + 2010): Our Entity Labels, Form, POS, Entity Labels for News, Hypernym on WordNet, Distance, Dependency Tree.
- Hyper Parameters: Decision by Grid Search.
Results of Relation Extraction

- Average of 3 F Scores of Predicting Relation
  - Training and Developing Model by 80% of Dataset
  - Testing Model by Rest 20% of Dataset
Experimental Result of Combining

- Average of 3 F Scores of Extracting Entity & Relation
  - Combining Results with Naive Rule
    - Rule: If words and labels of two entities are same, we define these entities are same.
  - Training and Developing Model by 80% of Dataset
  - Testing Model by Rest 20% of Dataset

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity Extraction</td>
<td>0.85</td>
<td>0.74</td>
<td>0.79</td>
</tr>
<tr>
<td>Relation Extraction</td>
<td>0.66</td>
<td>0.76</td>
<td>0.71</td>
</tr>
</tbody>
</table>
In **February 2016**, exploits for **Silverlight** based on **CVE-2016-0034** found their way into **Angler EK** a little more than a month after Microsoft issued a patch for the vulnerability.
Similar gate on 185.118.164.42 leads to more Angler EK traffic on 2016-04-25.
Exploring Threat Intel on the WEB

- Collect About 25,000 WEB Documents by Crawling
- Extract and Convert STIX Data (995 SDOs & 684 SROs)
Conclusion
Conclusion

■ Summary
- Developing Threat Knowledge Extraction System using Supervised Learning and Labeled Dataset
- Developing Entity Extractor of about 80% F Score
- Developing Relation Extractor of about 70 % F Score

■ Future Work
- Examination of Baseline Score for Production
- Analysis of Massive Knowledge Graph