2023 CNS Lab – Research Seminar

Acing the IOC Game

: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence

Presenter : **Seungho Song**

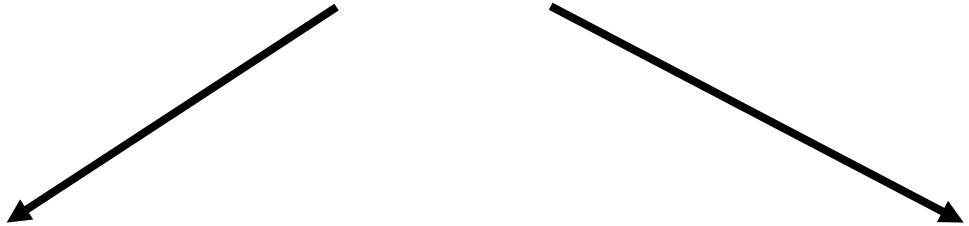
CNS Lab

Introduction

CTI (Cyber Threat Intelligence)

: Evidence-based knowledge, including context, mechanisms, indicators, implications and actionable advice, about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject's response to that menace or hazard

-> Cyber Threat임을 판단함에 있어 사용될 수 있는 증거 기반 지식



Visibility into fast-evolving threat landscape

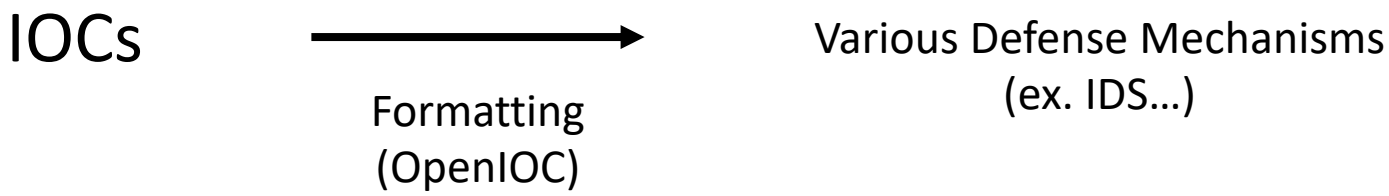
Identify early signs of attacks

Introduction

IOC (Indicators Of Compromise)

: Forensic artifacts of an intrusion
such as virus signatures, Ips/domains of botnets, MD5 hashes of attack files, etc

-> 여러 침입사고의 흔적들을 일정한 포맷으로 정리 해 놓은 문서 or 파일



CNS Lab

Introduction

OpenIOC

: extensible XML schema created by Mandiant to record
technical characteristics that identify a known threat, an attacker's methodology, or other evidence of a compromise.

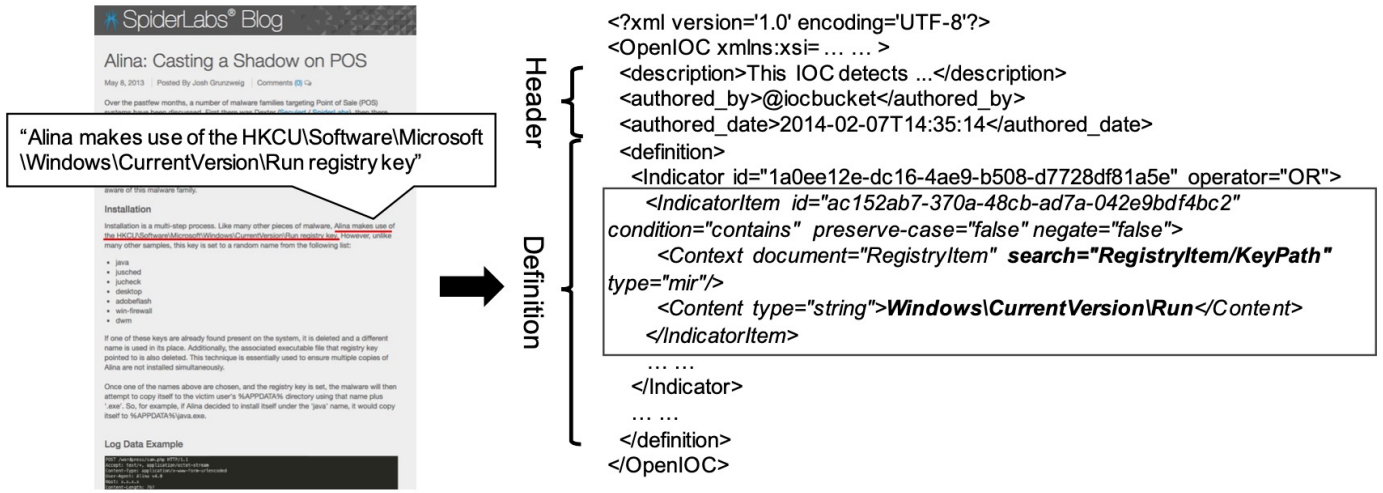
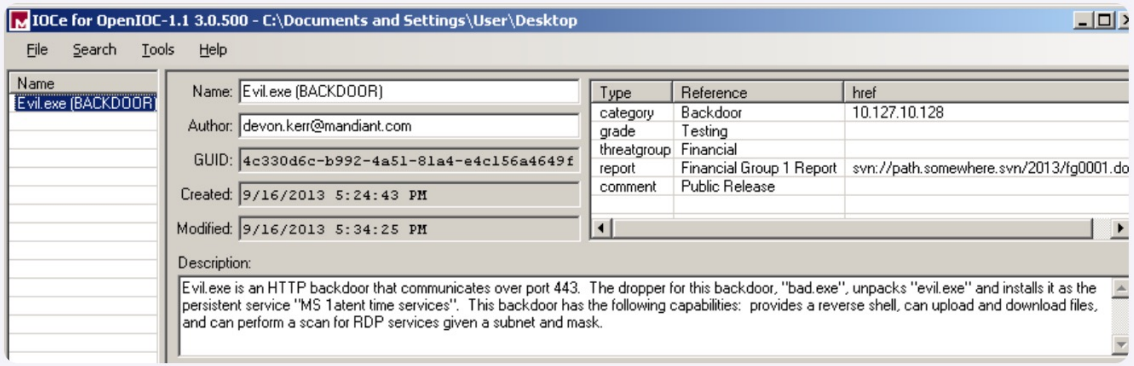


Figure 1: Example of OpenIOC schema.

Introduction

OpenIOC



References: Within the IOC, references can find information like the name of an investigation or case number, comments and information on the maturity of the IOC such as Alpha, Beta, Release, etc. This data can help you to understand where the IOC fits in your library of threat intelligence. One common use for references is to associate an IOC with a particular threat group. It is not uncommon for certain references to be removed from IOCs when sharing IOCs with third parties.

Type	Reference	href
category	Backdoor	10.127.10.128
grade	Testing	
threatgroup	Financial	
report	Financial Group 1 Report	svn://path.somewhere.svn/2013/fg0001.docx
comment	Public Release	

Source : https://www.mandiant.com/resources/blog/openioc-basics

Introduction

Problems of traditional IOC finding

traditional : extracted from blacklist sites

: Problem comes from the effective gathering of such information,
which entails significant burdens for timely analyzing a large amount of data.

- Ineffective
- Covering with small number of IOC classes (just URL, domain, IP and MD5)
- Relation between IOCs is not revealed
- No context information is provided (ex. The criminal group behind malfeasance)

CNS Lab

Introduction

Solution : collect IOC from open-source articles

: IOCs from articles in technical blogs and posts in forums are more favorable
to security practitioners and extensively harvested, since comprehensive descriptions of the attack are often found there.

Features

- Informal
- Natural languages
- Need to be analyzed semantically

Analyzing manually by
security analysts

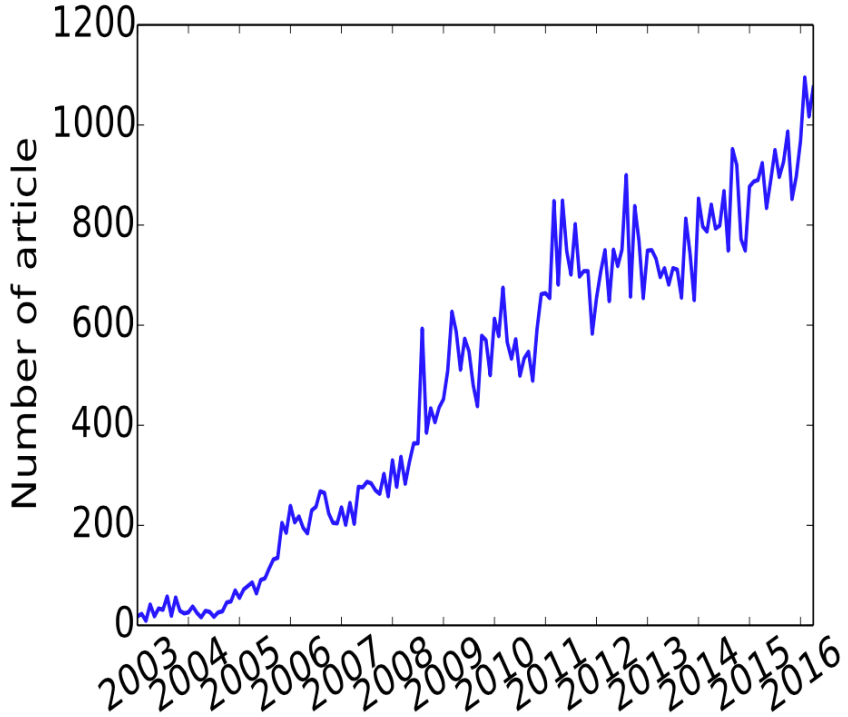
Standard IOC format
(ex. OpenIOC)

CNS Lab

Introduction

Problems of manually analyzing IOC from open-source articles

: unaffordable amount



The number of the articles per month

-> Need to AUTOMATE

the identification and extraction of valuable CTI involved

CNS Lab

Introduction

Challenges to automate IOCs from natural-language texts

1. High **false positives**
   - Mistaking non-IOCs for IOC

2. Even for a confirmed IOC, we **need to know its context**

CNS Lab

Introduction

Challenges to automate IOCs from natural-language texts

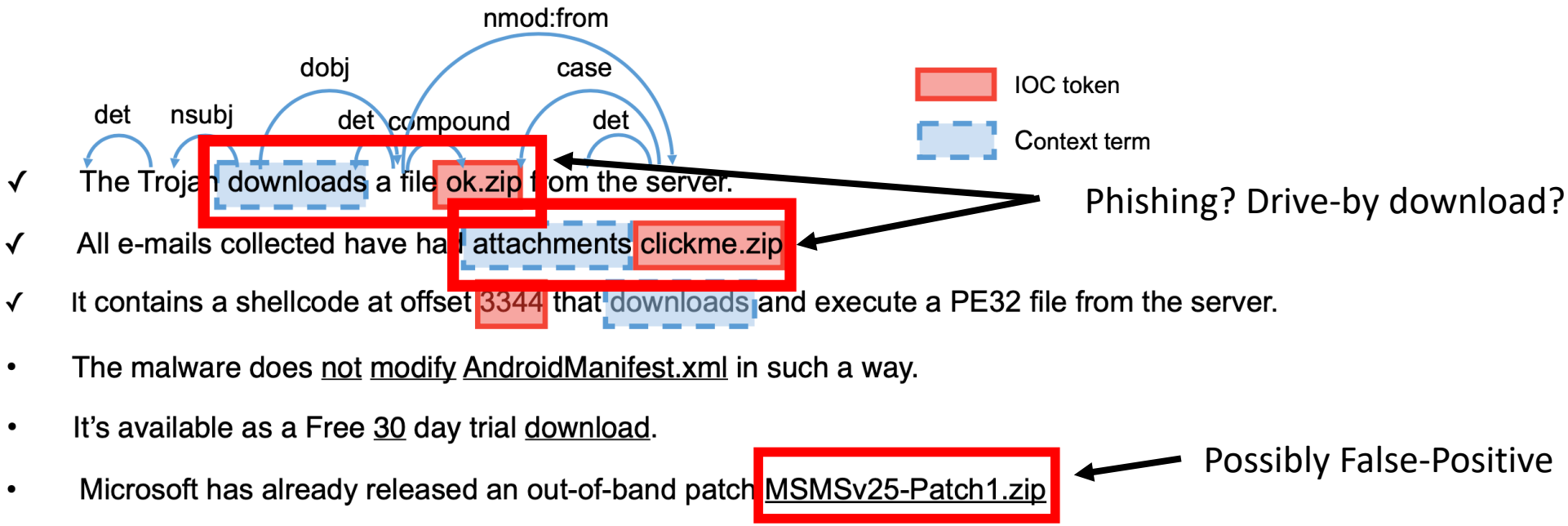


Figure 2: Examples of sentences with/without IOC.

CNS Lab

Introduction

Limits of existing NLP technique

- Highly Domain-sensitive
  - Those designed for one domain hardly work well on the other domain
- Unsatisfying performance
  - Typically low accuracy and coverage

Table 4: Accuracy and coverage comparison of iACE, AlienVault OTX and self-trained Stanford NER.

Tool	Precision	Recall
iACE	98%	93%
AlienVault OTX	72%	56%
Stanford NER	71%	47%

CNS Lab

iACE (IOC AutomatiC Extractor)

- An innovation solution for fully automated IOC extraction
- Providing a better understanding of the relationship, impact, evolution and quality of IOCs from technical blogs

iACE (IOC AutomatiC Extractor)

- IOCs are documented with context term (ex, attachment, malware, registry key, etc..)
- IOCs are connected to the context term through stable grammatical relation

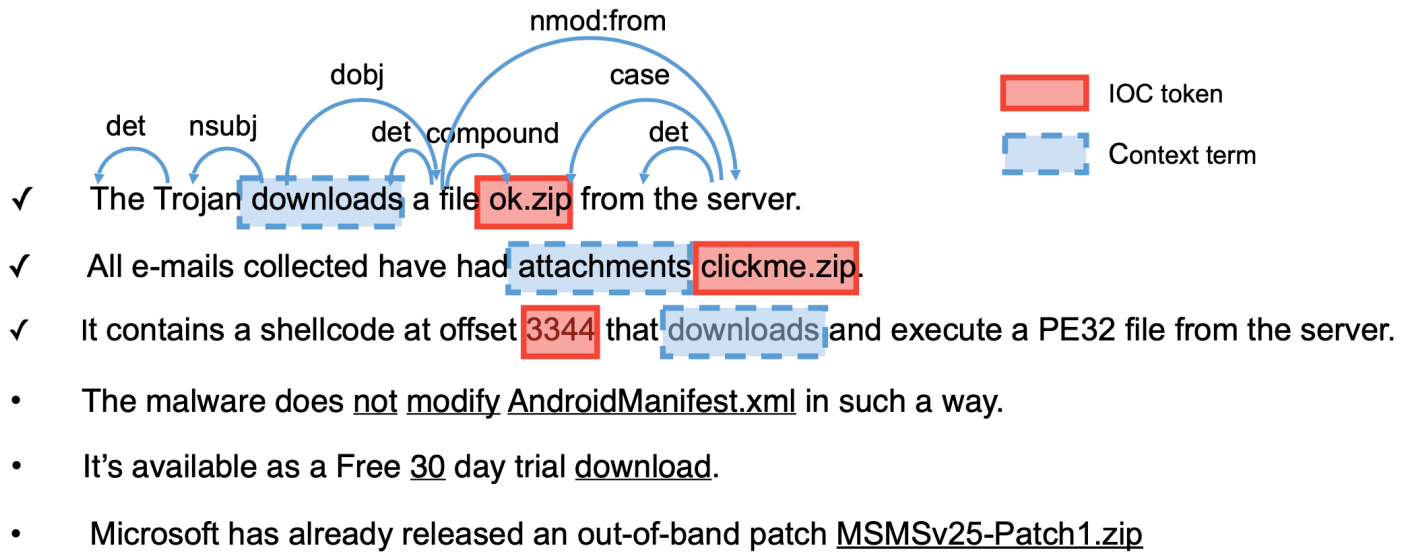


Figure 2: Examples of sentences with/without IOC.

CNS Lab

iACE (IOC AutomatiC Extractor)

Acing the IOC Game

- Search the known “anchors” (context terms and IOC-like strings) to locate IOC candidates
- Check the similarity of their relations to see if they are to the IOC-relations we have learned.

iACE (IOC AutomatiC Extractor)

Architecture

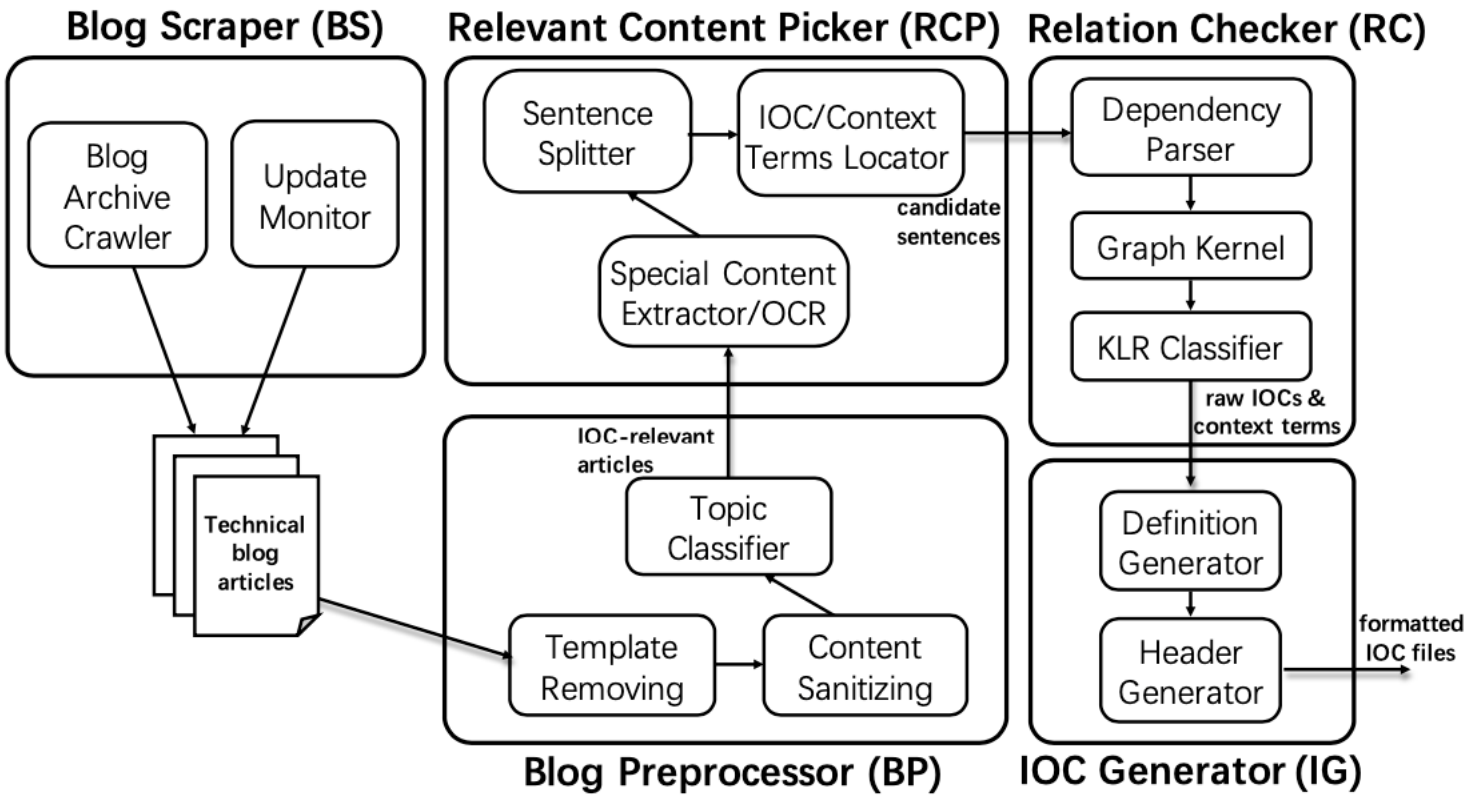


Figure 3: The architecture of iACE.

iACE (IOC AutomatiC Extractor)

1. Blog Scraper (BS)

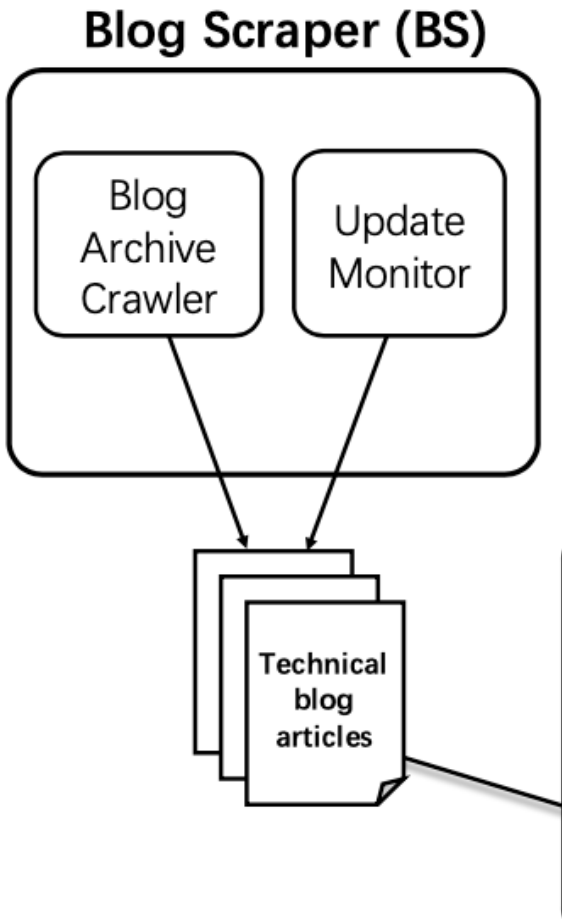
: Automatically collects technical articles from different technical blogs, removing irrelevant information from individual blog pages

(Ex. Advertisement)

Features

- Breadth-first crawling (use Beautifulsoup, python crawling library)
  - Starting from its homepage to explore each link it discovers, until no new link can be found

```
HTML  복사  캡션
<body>
  <div>
    <p>
      <a>...SECTOR 1...</a>
    </p>
  </div>
  <h1>
    <a>...SECTOR 2...</a>
  </h1>
  <div>
    ...SECTOR 3...
  </div>
</body>
```



iACE (IOC AutomatiC Extractor)

1. Blog Scraper (BS)

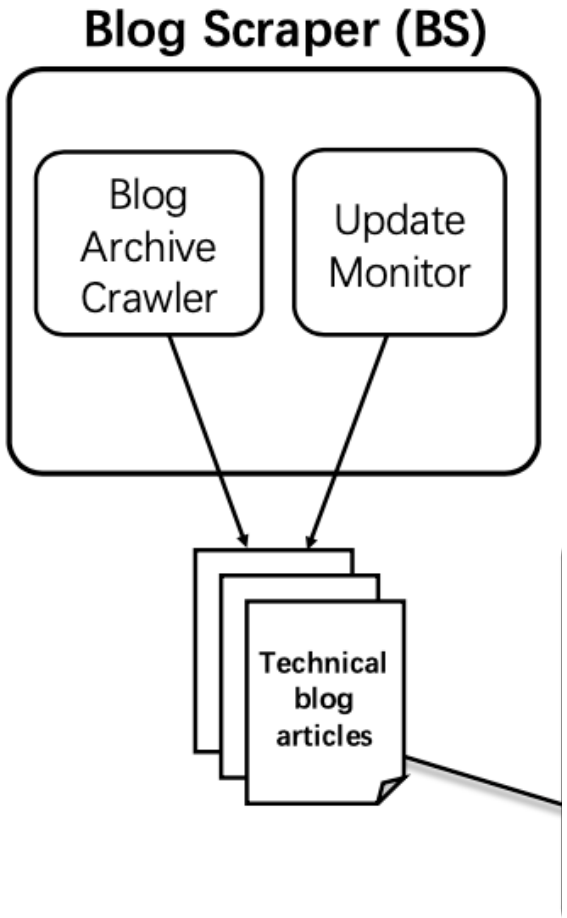
: Automatically collects technical articles from different technical blogs, removing irrelevant information from individual blog pages

(Ex. Advertisement)

Features

- Article template matching
  - Usually, the articles posted on the blog are all framed within the same HTML template, which is very different from other pages such as login, contact, …
  - Also, on any blog site, more article pages are hosted than other types of pages.

Blog Scrapper -> compares each page's DOM tree
to find out unnecessary page and dropped from dataset



iACE (IOC AutomatiC Extractor)

2. Blog Preprocessor (BP)

: Inspect articles using NLP techniques to filter out those unlikely to contain IOCs.

Features

- Template removal
  - Composite importance (SIGKI
    : compares all pages' DOM tre
    to find out the nodes with th
    characterized by their largely

Unnecessary component == large

AVA Discovery View: Surfacing Authentic Moments

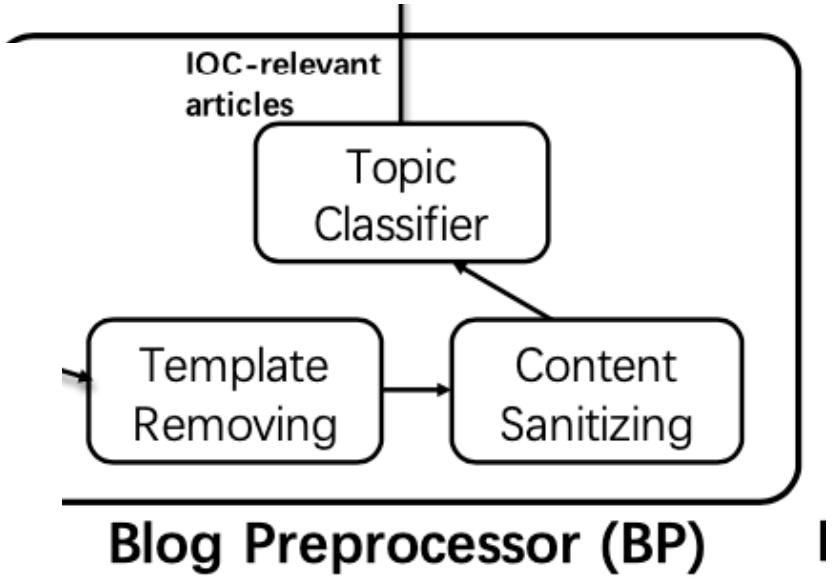
Netflix Technology Blog · Follow
Published in Netflix TechBlog · 10 min read · Aug 18

127 1

By: Hamid Shahid, Laura Johnson, Tiffany Low

Synopsis

At Netflix, we have created millions of artwork to represent our titles. Each artwork tells a story about the title it represents. From our testing on promotional assets, we know which of these assets have performed well and which ones haven't. Through this, our teams have developed an intuition of what visual and thematic artwork characteristics work well for what genres of titles. A piece of promotional artwork may resonate more in certain regions, for certain genres, or for fans of particular talent. The complexity of these factors makes it difficult to determine the best creative strategy for upcoming titles.



iACE (IOC AutomatiC Extractor)

2. Blog Preprocessor (BP)

: Inspect articles using NLP techniques to filter out those unlikely to contain IOCs.

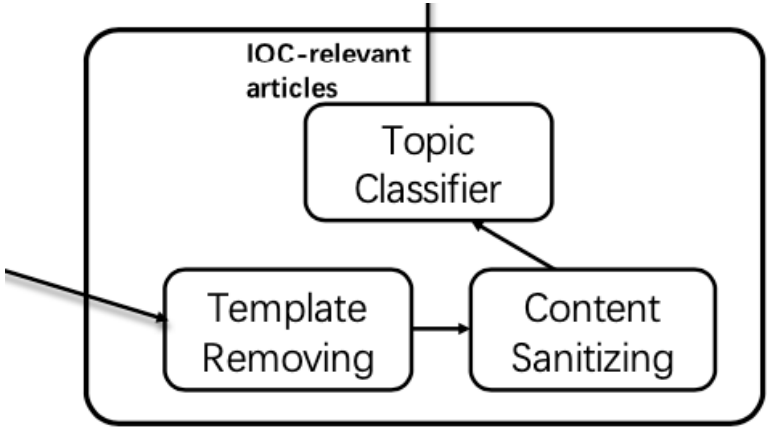
Features

- Text Converting
  - There's still some content cannot be directly analyzed by text-based approach, including images and embedded PDF files.
    - Tesseract : img to text
    - Ghostscript : pdf to text
- Topic classifier

  : Start looking into article's content, first removing those not including any IOCs.
  - Topic words : TOP 20 most likely to appear in IOC articles.
  - Article length : IOC articles length tends to be longer than non-IOC articles.
  - Dictionary-word density

    : IOC articles tend to have a lower dictionary-word density because most IOCs are non-dictionary words

Ex. Articles about Product Promotion, SW Update



Blog Preprocessor (BP)

iACE (IOC AutomatiC Extractor)

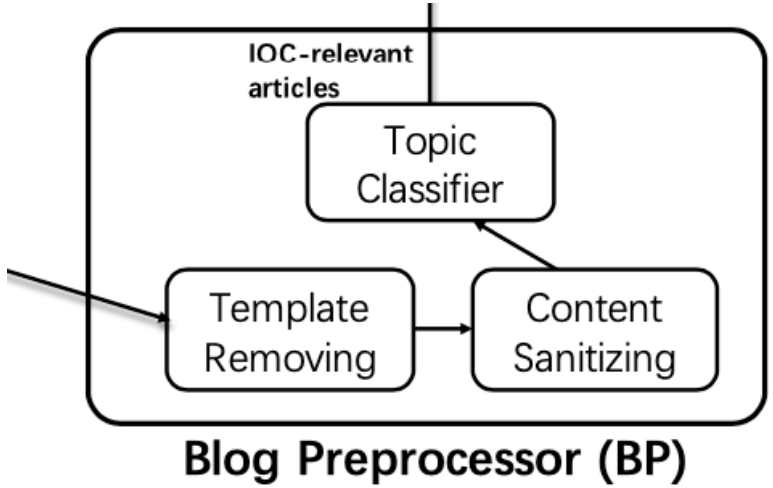
2. Blog Preprocessor (BP)

: Inspect articles using NLP techniques to filter out those unlikely to contain IOCs.

Model Performance

	Precision	Recall
DS-Labeled (Train sets)	98%	100%
	Accuracy	Coverage
DS-Unknown (Test sets)	96%	99%

DS-Labeled : including 150 IOC + 300 non-IOC articles
DS-Unknown : 500 randomly selected articles



iACE (IOC AutomatiC Extractor)

3. Relevant Content Picker (RCP)

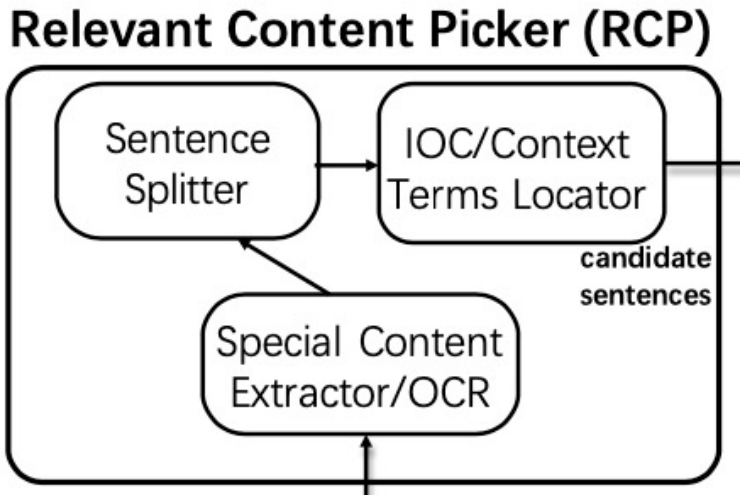
: RCP parses the HTML content of each article,
breaking the text into sentences and detecting tables and lists

Features

- Regular exp match
  - 19 different classes, including IP / hast / int / float ..
- Sentence splitter
  - NLTK
- Context term match
  - 600 OpenIOC terms
  - Semanticlink
  - 5,283 context terms

Table 2: Examples of regexes.

General type	Regex
IPv4	(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.)\{3\}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)(/([0-2][0-9]|3[0-2]|[0-9]))?
hash	\b[a-fA-F\d]{32}\b|\b[a-fA-F\d]{40}\b|\b[a-fA-F\d]{64}\b
int number	((?<=[\s([\s\(\[\{:\+\-=\\])|^)(\+|-)?\d+(?!([a-zA-Z0-9]|\.\S))
float number	((?<=[\s\(\[\{: \+\-=\\])|^)(\+|-)?(\d+)?\.(\d+)(?!([a-zA-Z0-9]|\.\S))

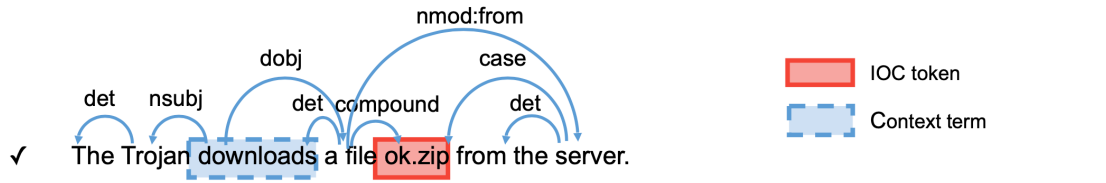


iACE (IOC AutomatiC Extractor)

4. Relation Checker (RC)

: Analyzes its grammatical structure connecting the context anchors (=context terms) and the IOC anchor (matched by regex), using a learned model to determine whether **the latter is indeed an IOC**

Coincidence of relevant tokens does not necessarily indicate the presence of an IOC-related description.



✓ The Trojan downloads a file ok.zip from the server.

✓ All e-mails collected have had attachments clickme.zip.

✓ It contains a shellcode at offset 3344 that downloads and execute a PE32 file from the server.

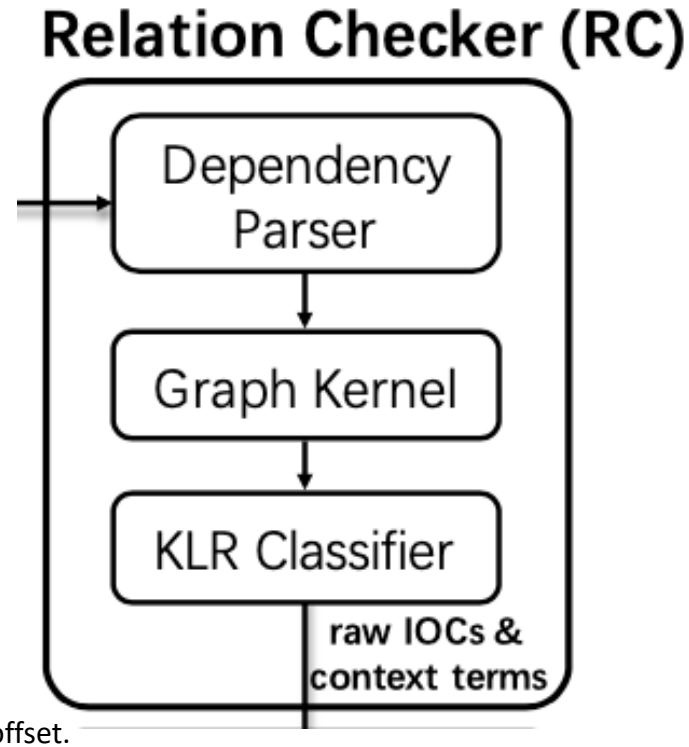
• The malware does not modify AndroidManifest.xml in such a way.

• It's available as a Free 30 day trial download.

• Microsoft has already released an out-of-band patch MSMSv25-Patch1.zip

Figure 2: Examples of sentences with/without IOC.

Actually, It's not offset.
It's PE32 file.



iACE (IOC AutomatiC Extractor)

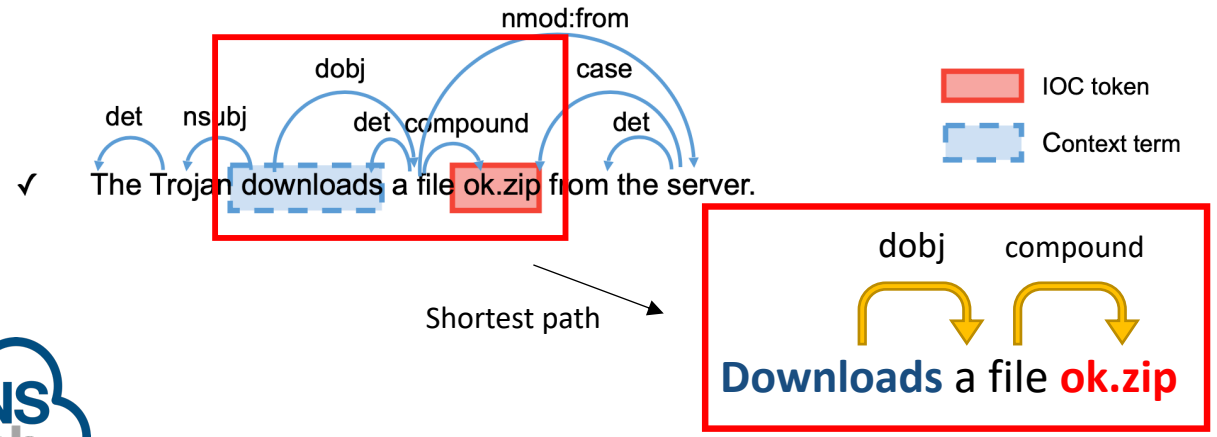
4. Relation Checker (RC)

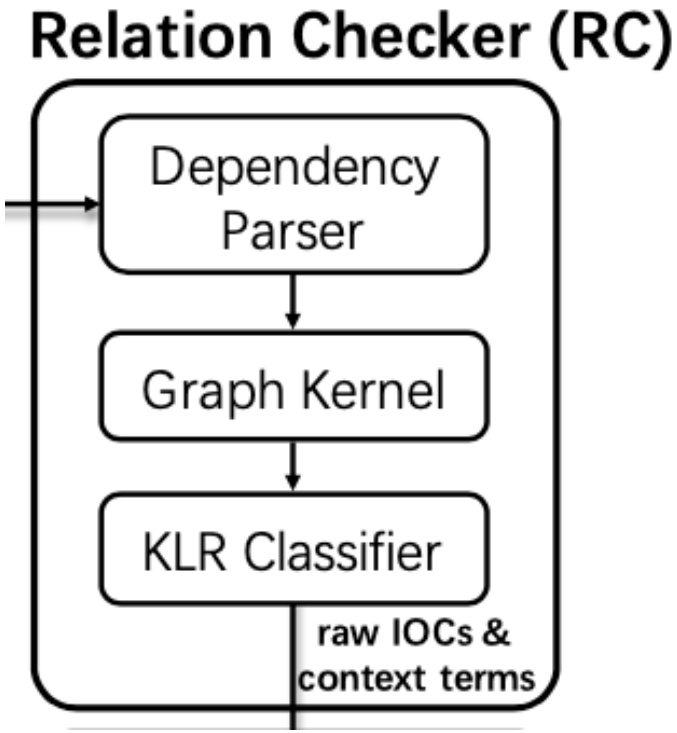
- Dependency Parser

  : transform a sentence into a dependency graph (DG)

- Using dependency to represent the grammatical relation between two tokens
- Extracting the shortest path between the “anchors”

  -> more relevant to understanding of the relations between the anchors





iACE (IOC AutomatiC Extractor)

4. Relation Checker (RC)

- Graph Mining and KLR Classifier

- A similarity measure of directed weighted graph

Direct Product Kernel

: function that measures the similarity of two graphs

by counting the number of all possible pairs of arbitrarily long random walks with identical label sequences

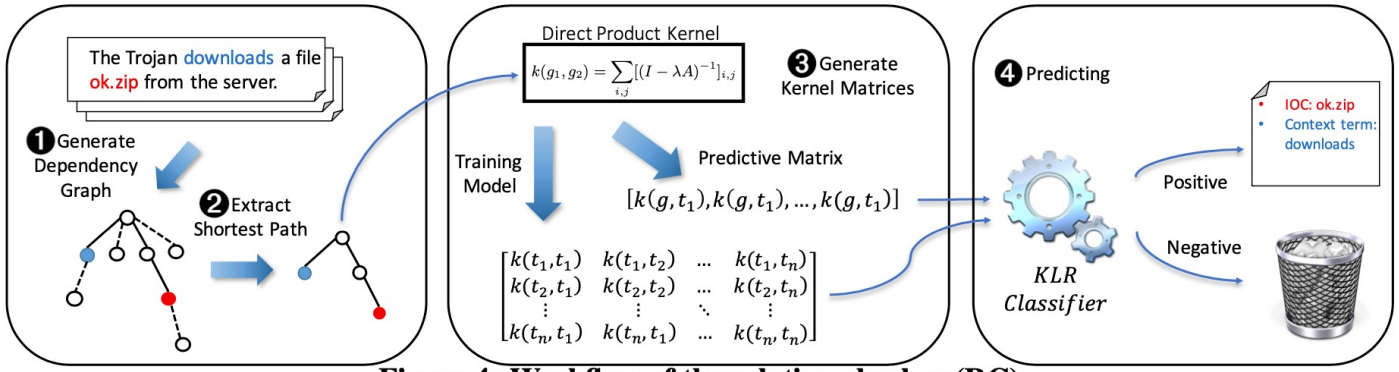
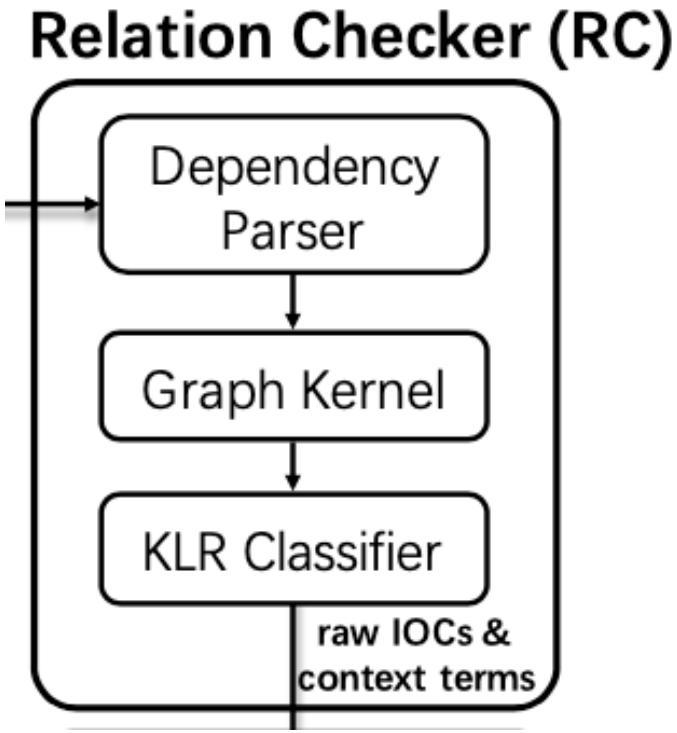


Figure 4: Workflow of the relation checker (RC).



iACE (IOC AutomatiC Extractor)

4. Relation Checker (RC)

- Graph Mining and KLR Classifier

KLR Classifier
- DS-Labeled : 1500 positive and 3000 negative instances
- Model is learned to work on a kernel vector.

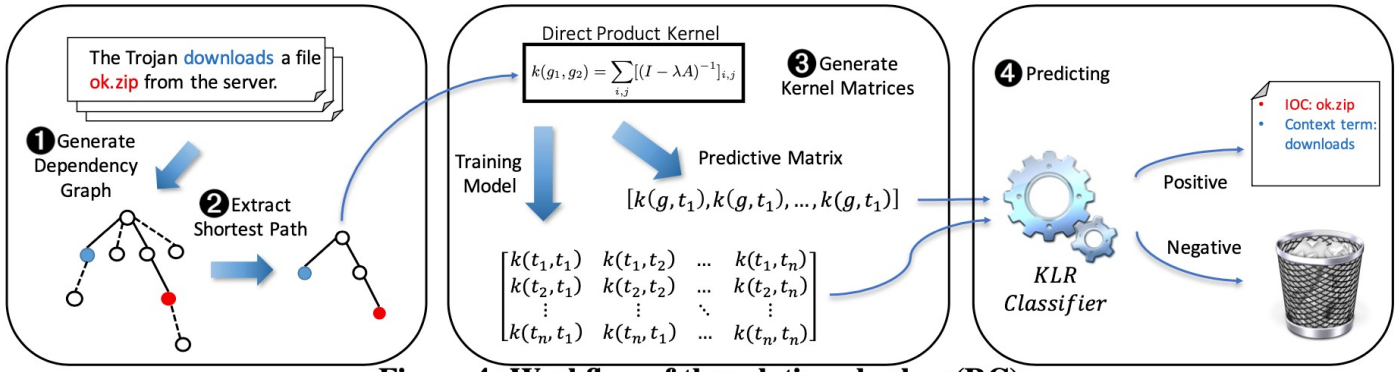
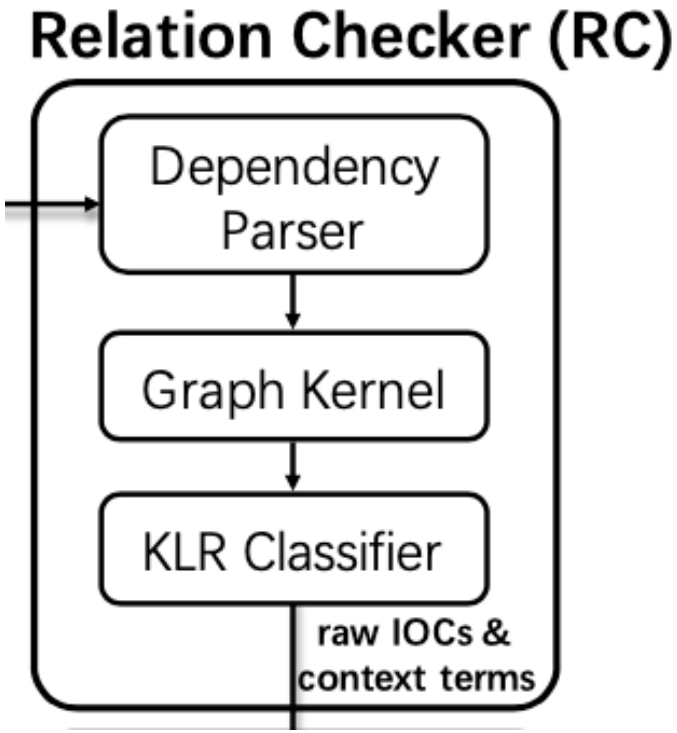


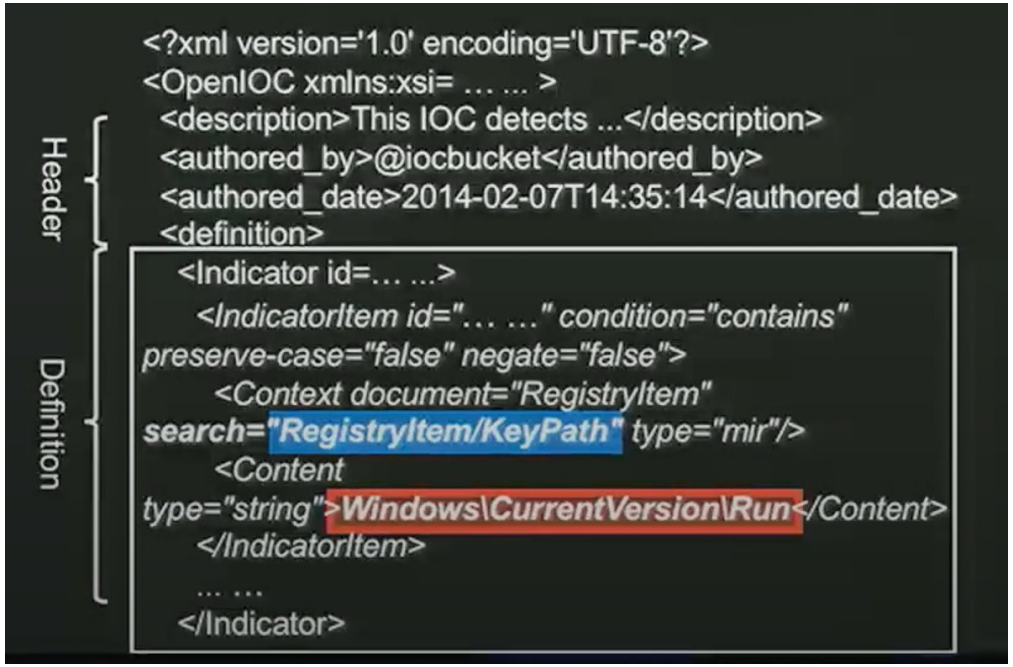
Figure 4: Workflow of the relation checker (RC).

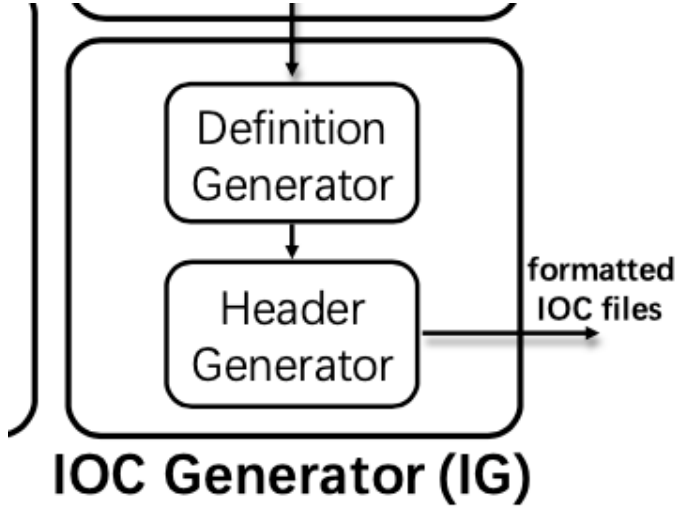


iACE (IOC AutomatiC Extractor)

5. IOC Generator (IG)

: automatically convert the CTI content of a technical blog to the OpenIOC record.





Source : https://www.youtube.com/watch?v=FS46VLGepAk&t=612s

Evaluation

Settings

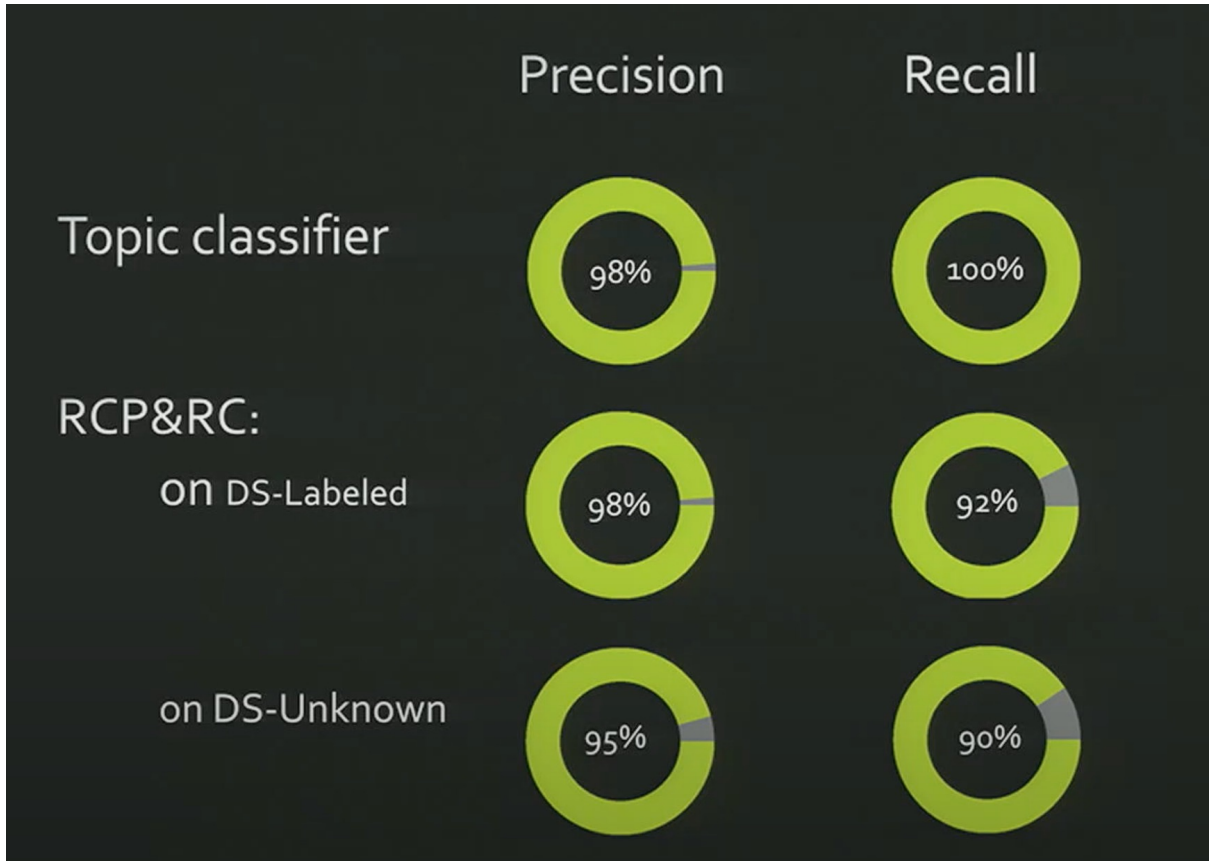
- DS-Labeled:
  - 450 articles from Iocbucket, Openiocdb, and manual gathering
  - 1,500 true IOC sentences, 3000 false IOC sentences
- DS-Unknown:
  - 45 security-related technical blogs
  - 71K articles
  - 2003.04 ~ 2016.05

Source : https://www.youtube.com/watch?v=FS46VLGepAk&t=612s

Evaluation

Results

- Model Performance



Source : https://www.youtube.com/watch?v=FS46VLGepAk&t=612s

Evaluation

Results

- Compare with state-of-the-art alternatives (in 2016)

Table 4: Accuracy and coverage comparison of iACE, AlienVault OTX and self-trained Stanford NER.

Tool	Precision	Recall
iACE	98%	93%
AlienVault OTX	72%	56%
Stanford NER	71%	47%

CNS Lab

Evaluation

Results

- Running time

Table 5: Running time at different stages.

Stage	average time (ms/article)	Std Deviation (ms)
BP	5	-
RCP	20	2.5
RC	252.5	37.5
IG	2.5	-
total	278.6	-

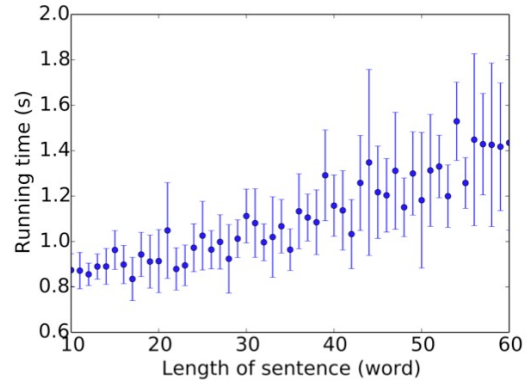


Figure 6: Average running time on the sentences in different lengths.

Analysis

1. Landscape

- These open-source technical blogs are indeed a gold mine for threat intelligence gathering
  - 45 blogs -> 71,000 articles -> 900K IOCs and their context

Table 6: Top-10 iocterms with the largest number of IOCs.

iocterm	# of IOCs	general type
PortItem/remoteIP	18,243	IP
RegistryItem/ValueName	12,572	string
UrlHistoryItem/HostName	12,020	URL
RegistryItem/Path	11, 777	string
FileDownloadHistoryItem/SourceURL	10,324	URL
ServiceItem/serviceDLLmd5sum	9,908	hash
PortItem/remotePort	9,831	int
FileDownloadHistoryItem/FileName	8,720	string
Email/ReceivedFromIP	8,225	IP
FileItem/FileExtension	8,100	string

found by iACE

CNS Lab

Analysis

2. Understanding Threats

- Clustering all the articles into 527 clusters
  - Two arti
- Believe that
  whose scale

before, which was confirmed by our additional search for related literature across the Internet. Table 7 provides the information about those clusters. Most interestingly, we found that for the IOC "132.248.49.112", a shared IP reported by 19 blogs, turned out to point to a C&C campaign's name server. We believe that all the independently documented attacks actually belonged to a massive campaign whose scale was not known before. Note that this corre-

3	30	215	960	0
4	28	897	1,302	0
5	25	897	1,222	0

Discussion

Limits

1. Error / missing analysis

- Limitations of underlying tools affects accuracy.
  - Tesseract : performance issue
  - State-of-the-art dependency parser still can't maintain its accuracy when the sentence is too long.
- Shortest-path : still there are sentences too long for the parser to understand the dependencies between words correctly.
- Author's mistakenly / deliberately misspell URL
- Or, active attacker can compromise the blog websites and inject fake IOCs.

2. Other intelligence sources and standards

- What is less clear is the technique's effectiveness on less formal sources, like technical forums.

CNS Lab

Conclusion

iACE (Ioc AutomatiC Extractor)

- An innovation solution for fully automated IOC extraction
- Providing a better understanding of the relationship, impact, evolution and quality of IOCs from technical blogs

CNS Lab