Adaptive File Analyzer

NLP combined with Heuristic analysis to detect malicious email attachments

Presented By:

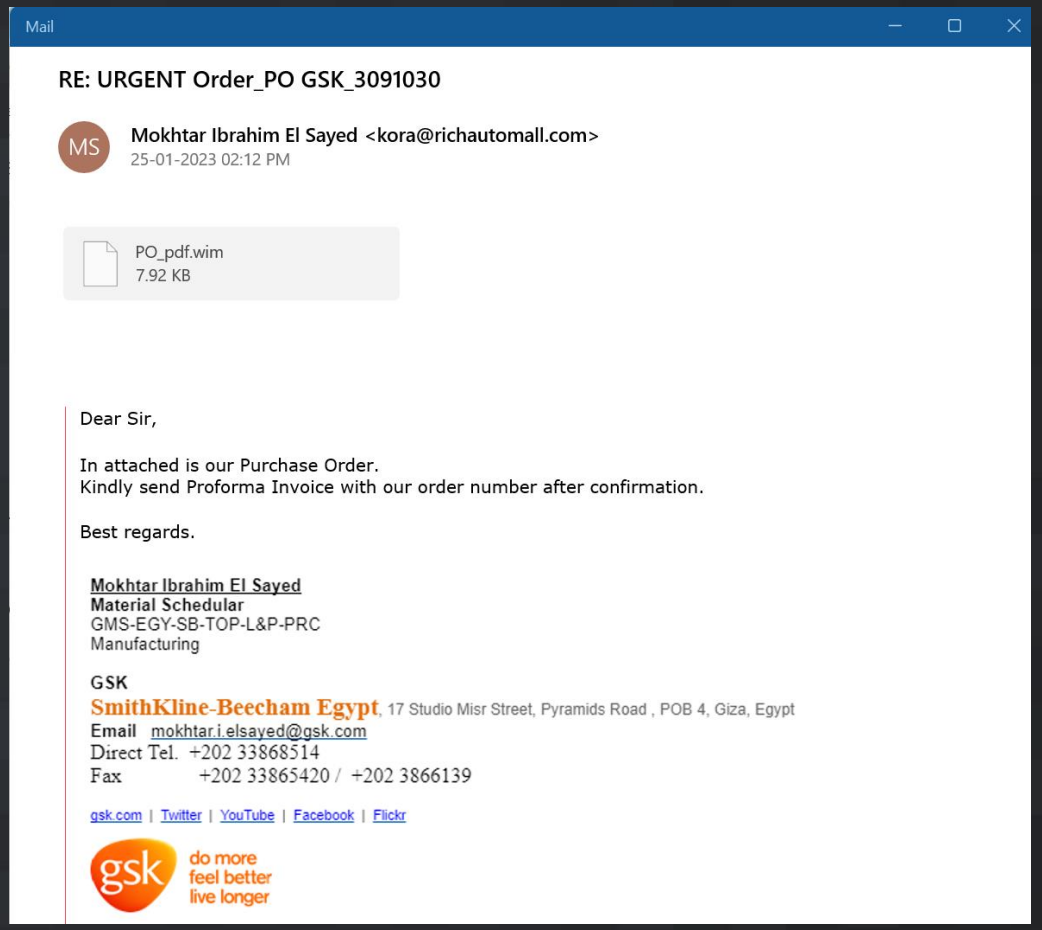
Abhishek Singh & Kalpesh Mantri

cisco | TALOS

Principles

The context of the email is Invoice, and the attachment is in .wim format. Invoices are sent in PDF or Word, not in .wim format so, if we use contextual info, we can label attachments as malicious.

Context of email can be correlated with the deep file parsing results of attachments to determine malicious or benign attachments.



cisco | TALOS

Leveraging Contextual Analysis

More examples:

Embedded Exe in Word with a context of Installer in the body of the email can be a legitimate word attachment

Embedded Exe in Word with a context of Invoice, delivery email, such as DHL, in the body of email is a malicious attachment

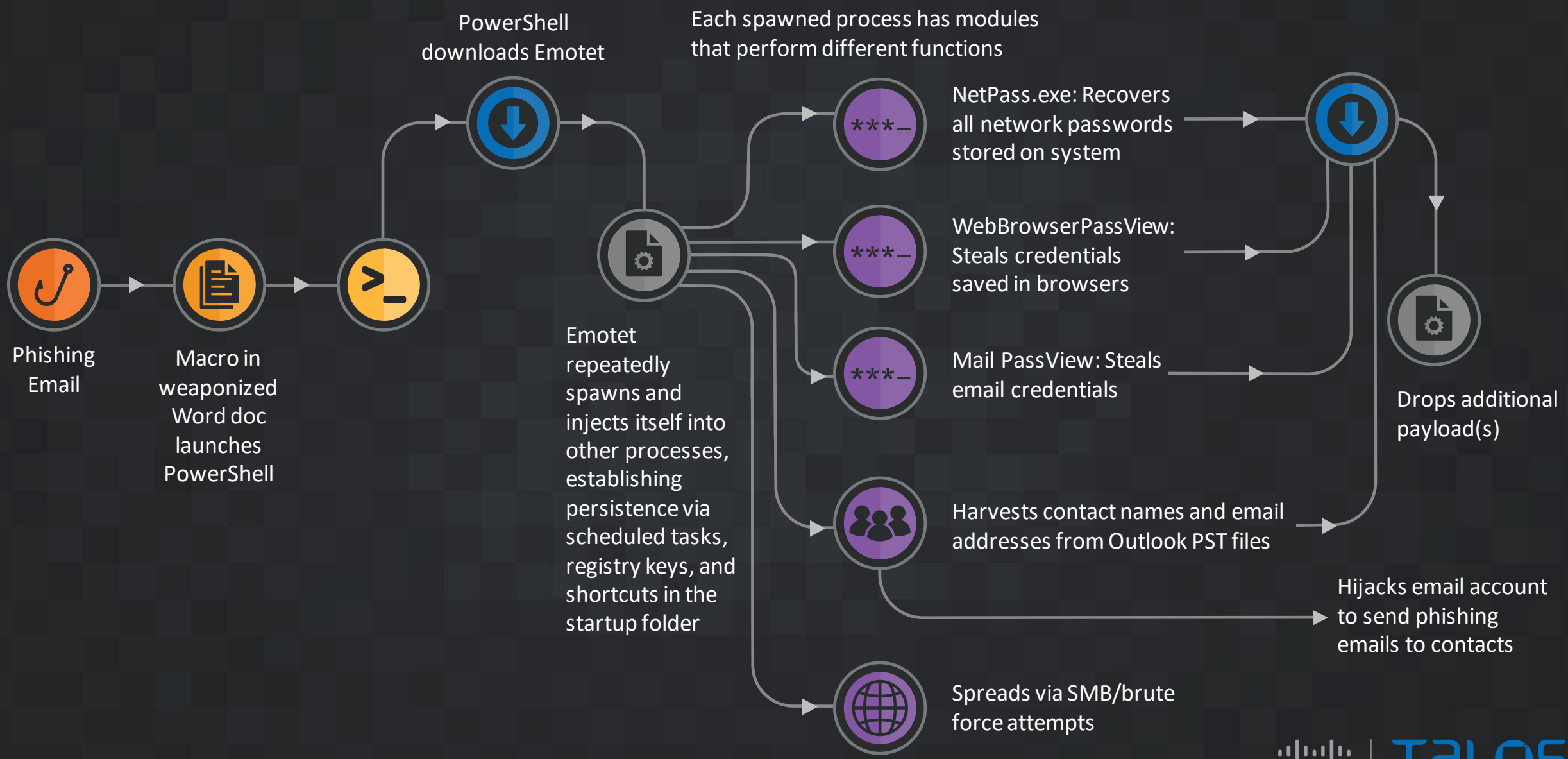
Contextual analysis of email, which can be derived using NLP from the body or subject, can be used as a feature set for correlation with deep file results of the attachments in many cases (**not all**) to classify attachments as malicious or benign.

Problems we are Solving

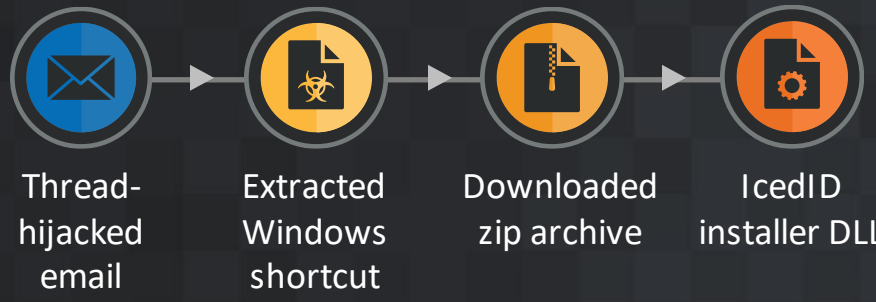
- In-wild malware delivered as email attachments such as Agent Tesla, HTML smuggling, Remcos RAT, IcedID bots etc. is multi-stage malware.
- Analysis of multistage malware means gathering downloaders/droppers, which is not always feasible.
- Malware extensively employs VM evasion techniques such as extended sleep calls, checking for user interaction such as mouse movements, lack of proper environment, etc. https://media.blackhat.com/us-13/US-13-Singh-Hot-Knives-Through-Butter-Evading-File-based-Sandboxes-WP.pdf making it challenging to capture behavior in VM-based environments.
- Besides multi-stage malware, attachments having phishing links to steal passwords are often difficult to detect they employ captcha , redirects making them difficult to detect.

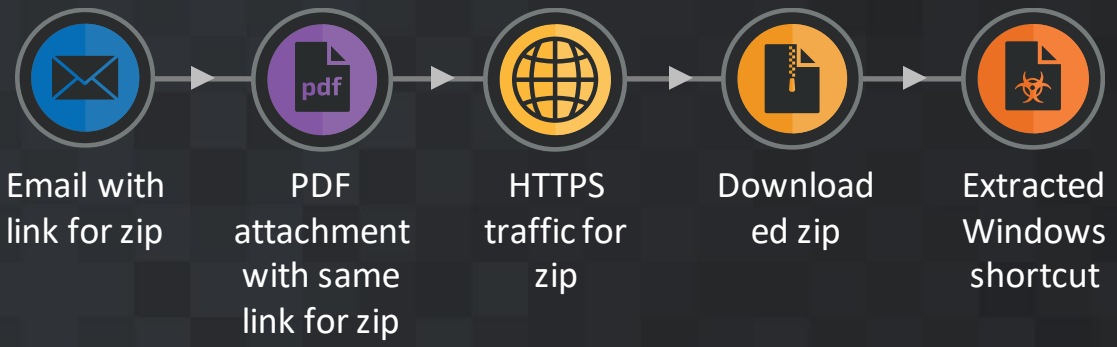
HTML Smuggling

Multi-Stage Malware



Multi-Stage Malware





cisco | Talos

Extracting Context From Email via NLP

- Currently, AFA uses Latent Dirichlet Allocation (LDA) to get the context or topic from the Email.
- It is Fast, so it can scale for the traffic volume and be applied to the body and subject to get context/topic of discussion.

Only these types of emails are considered for correlation since these are extensively employed by malware

Finance Related (Payment / Bank)

Information and communication related (fax, memo, VOIP)

Office Related (Official)

Invoice Related (Bills/Receipts)

Delivery Related (Logistics)

Call to Action Related (CTA)

Designed Principles for AFA to detect Multi-Stage Malware

Extract the context of the email from the body and subject via NLP

Leverage anomaly in the metadata from email headers such as differences in “Mail From: ” and “Reply-To:,” use of free email address, lack of “in-reply to:” header, etc

Perform deep file scanning of the first stage of malware attached to emails

Correlate the above three conditions to determine malicious or benign attachments without relying on malware's second and third stages.

cisco | TALOS

Context from Email and File Parsing Capabilities

OLE
- Has Macro
- Is Encrypted
- Has Shellcodes
- Has Embedded Objects
- Checks for Malware (API Calls, Embedded PE)

HTML
- URLs <a tag>
- URLs <form tag>
- Script Blocks <script tag>
- Has Obfuscation in Script
- Text

PDF
- Count of Structure tags [obj, endobj, stream, endstream, xref, trailer, startxref]
- Count of URL tags [/URI]
- Count of JS tags [/JS, /JavaScript]
- Count of Pages [/Page]
- Encrypted PDF [/Encrypt]
- Count of Launch tags [/AA, /OpenAction, /Launch, /EmbeddedFile]
- Other tags [/ObjStm, /AcroForm, /JBIG2Decode, /RichMedia, /XFA]

Executable
- Directly blocking these extensions

Archive
- Extraction of internal filenames
- Extraction of internal files

One
- Extraction of internal filenames
- Extraction of internal files

- EML:
  - Context Categories Classification Via Document Modeling LDA Algorithm
- [Subject, Body, Name of Files]:
  - Finance (Payment / Bank)
  - Information & Communication (fax, memo, VOIP)
  - Office (Official)
  - Invoice (Bills/Receipts)
  - Delivery (Logistics)
  - Call to Action (CTA)
  - Filename and extensions
  - Encoded headers
  - Anomalies in header

cisco | TALOS

Deep File Parsing: Contextual Tags

HTML

- URLs <a tag>
- URLs <form tag>
- Script Blocks <script tag>
- Has Obfuscation in Script
- Text

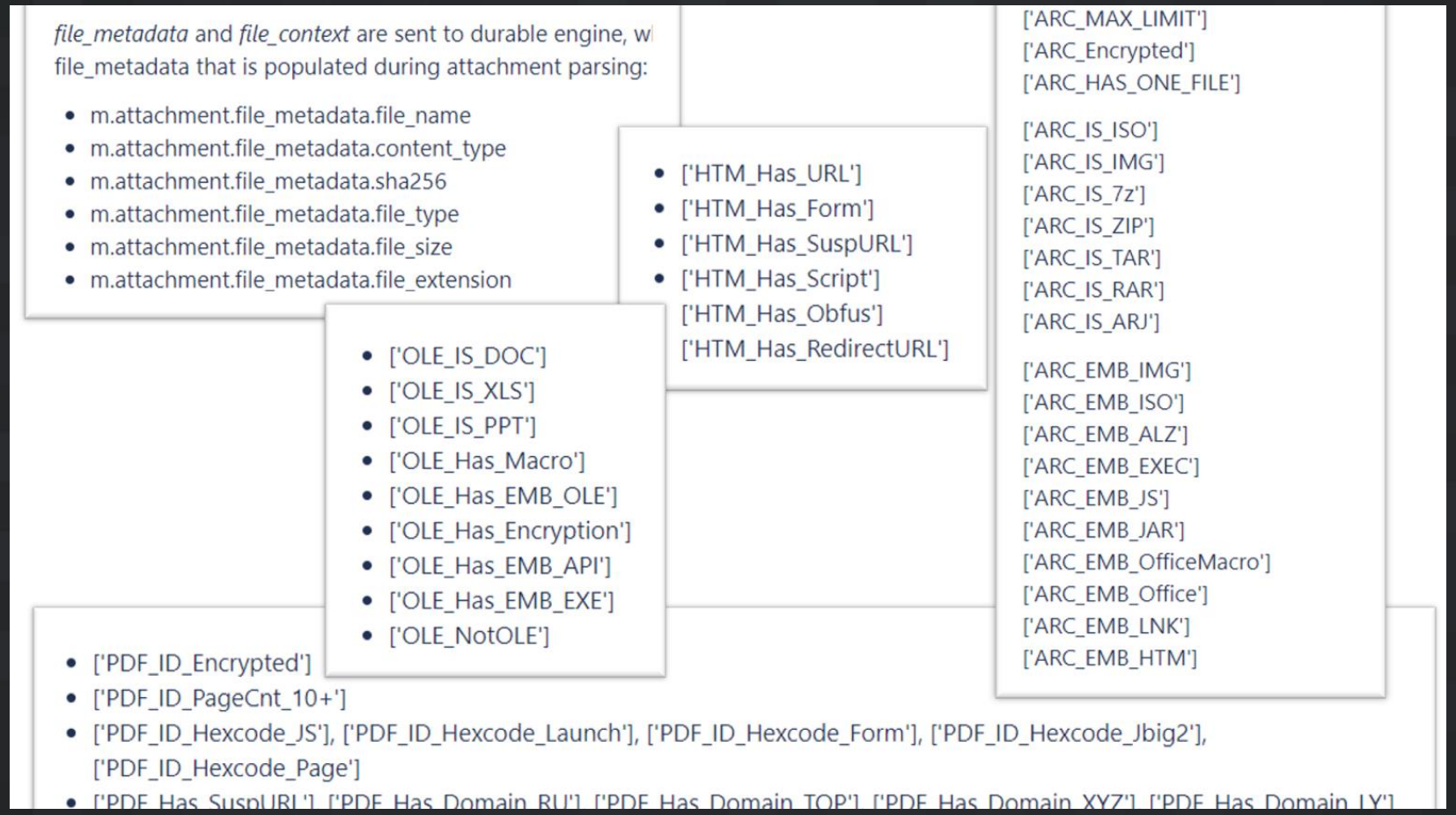
```
def getURLs(self):
    url_href = []
    try:
        a_tag = self.html.find_all('a')
        if len(a_tag):
            for ele in a_tag:
                if hasattr(ele, 'attrs') and ele.attrs.get('href'):
```

```
def getScript(self):
    script_block = ''
    try:
        a_tag = self.html.find_all('script')
        if len(a_tag):
            script_block = ''
            for ele in a_tag:
                if hasattr(ele,
                    script_block
```

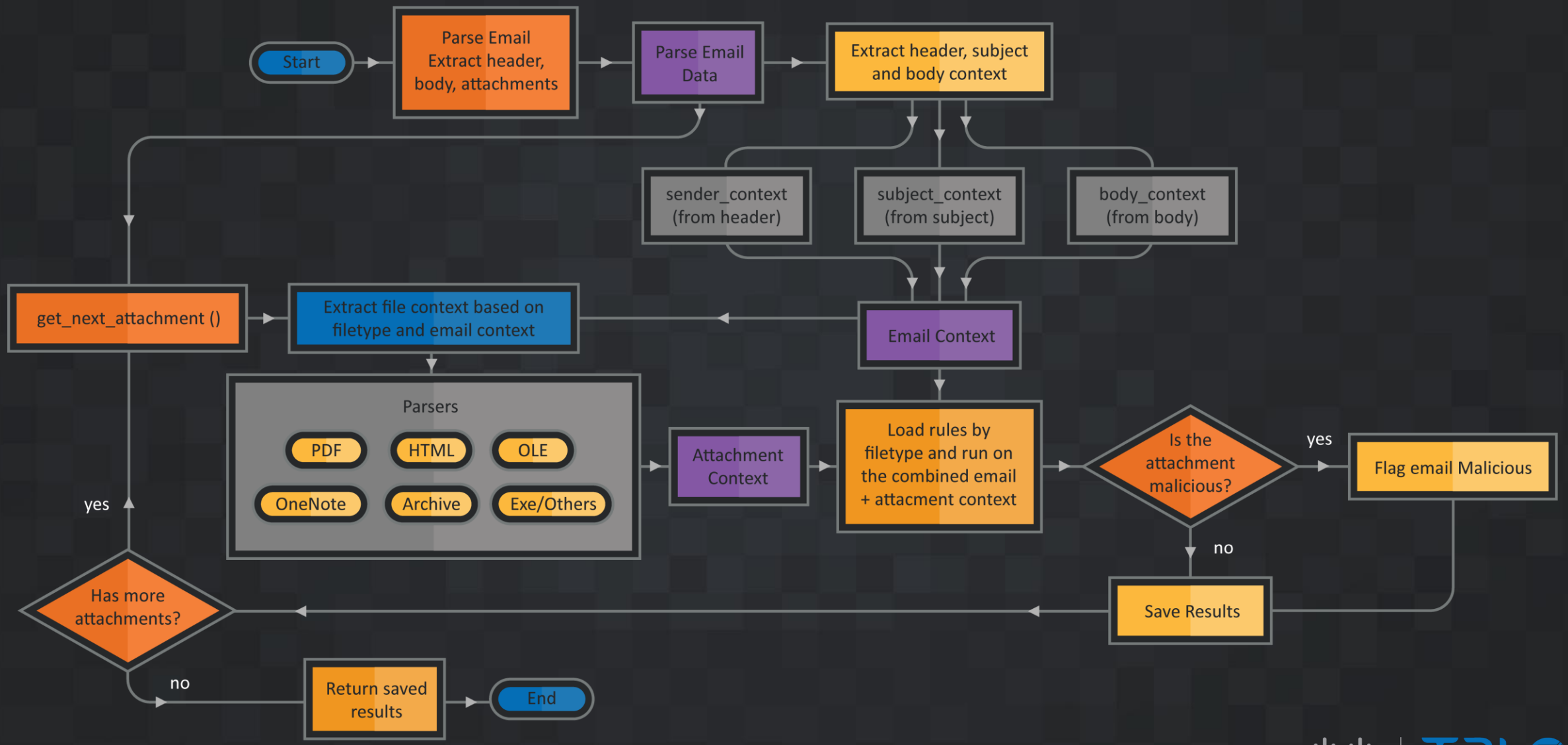
```
ohtm = cParseHtml(data)

checks = ohtm.html.find_all('a')
if checks: html_context += ['HTM_Has_URL']

checks = ohtm.html.find_all('form')
if checks: html_context += ['HTM_Has_Form']

# Extract all URLs from HTML
urls = ohtm.getURLs()
if urls and url_result(urls):
    html_context += ['HTM_Has_SuspURL']

# Extract all Scripts from HTML
sblock = ohtm.getScript()
if sblock:
    html_context += ['HTM_Has_Script']
    hasObfus = hasObfuscation(sblock)
```

cisco | TALOS

Deep File Parsing: Contextual Tags



AFA Design


Start
Parse Email Extract header, body, attachments
Parse Email Data
Extract header, subject and body context
sender_context (from header)
subject_context (from subject)
body_context (from body)
Email Context
get_next_attachment ()
Extract file context based on filetype and email context
Parsers
PDF
HTML
OLE
OneNote
Archive
Exe/Others
Attachment Context
Load rules by filetype and run on the combined email + attacment context
Is the attachment malicious?
yes
Flag email Malicious
no
Save Results
yes
Has more attachments?
no
Return saved results
End

Durable Rules to give Verdict

Contextual Information + File Analysis

```
from durable.lang import *

## Anomaly Rule 11
## Subject Context (INV|FIN) + Body Context (CTA) + Attached (ARC->IMG(1))
@when_all(
    (m.attachment.file_metadata.file_type == 'archive')
    & (m.attachment.file_context.archive_context.anyItem(item == 'ARC_EMB_IMG'))
    & (m.attachment.file_context.archive_context.anyItem(item == 'ARC_HAS_ONE_FILE'))
    & (
        (m.email.subject_context.anyItem(item.matches('INV|FIN')))
        | (m.email.body_context.anyItem(item == 'CTA'))
    )
)
def rule_archive_arc_img_01(c):
    c.s.verdict = 'Malicious: Body Context with ARC->IMG Attachment'
```

cisco | Talos

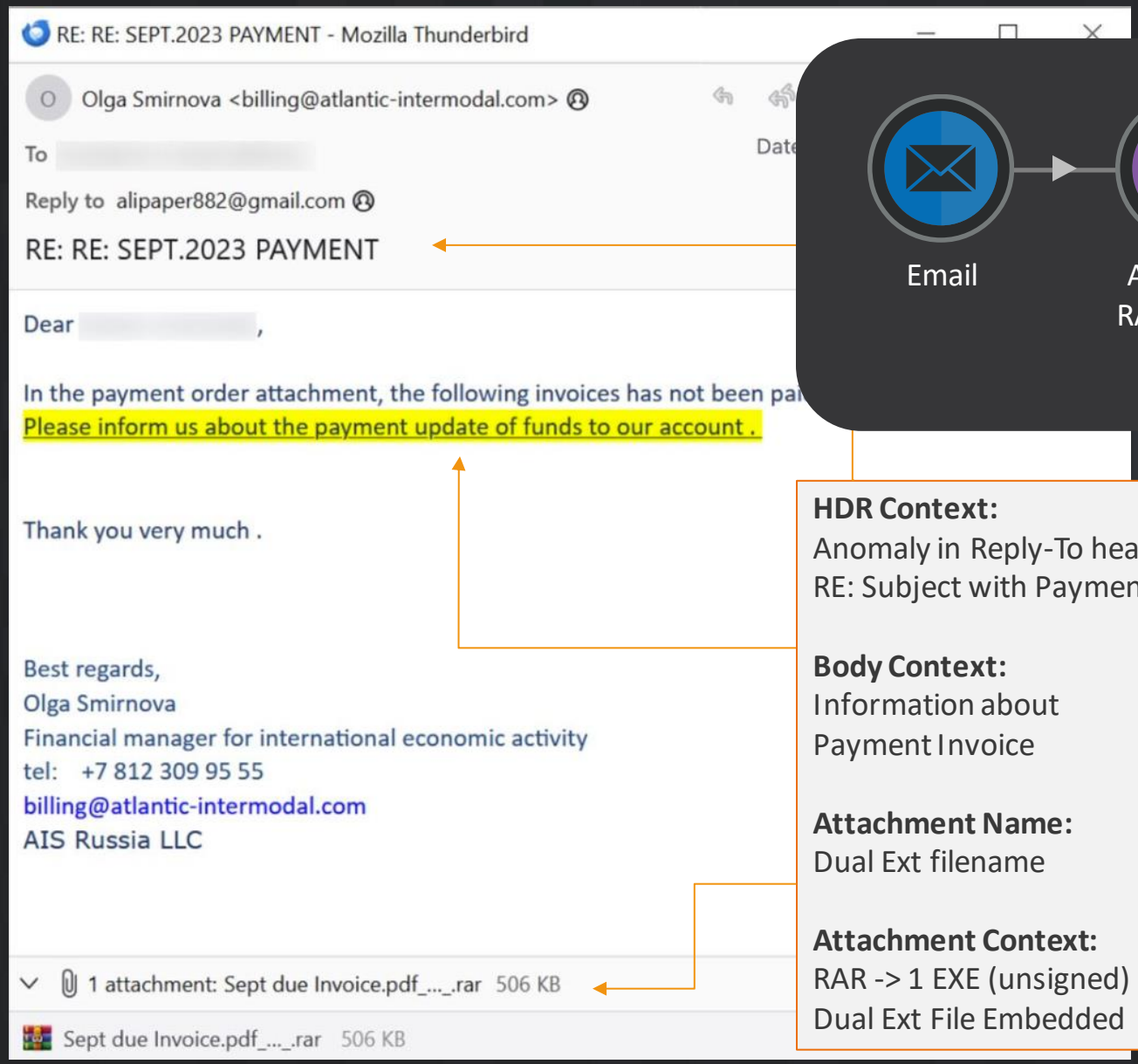
More Examples of Durable Rules

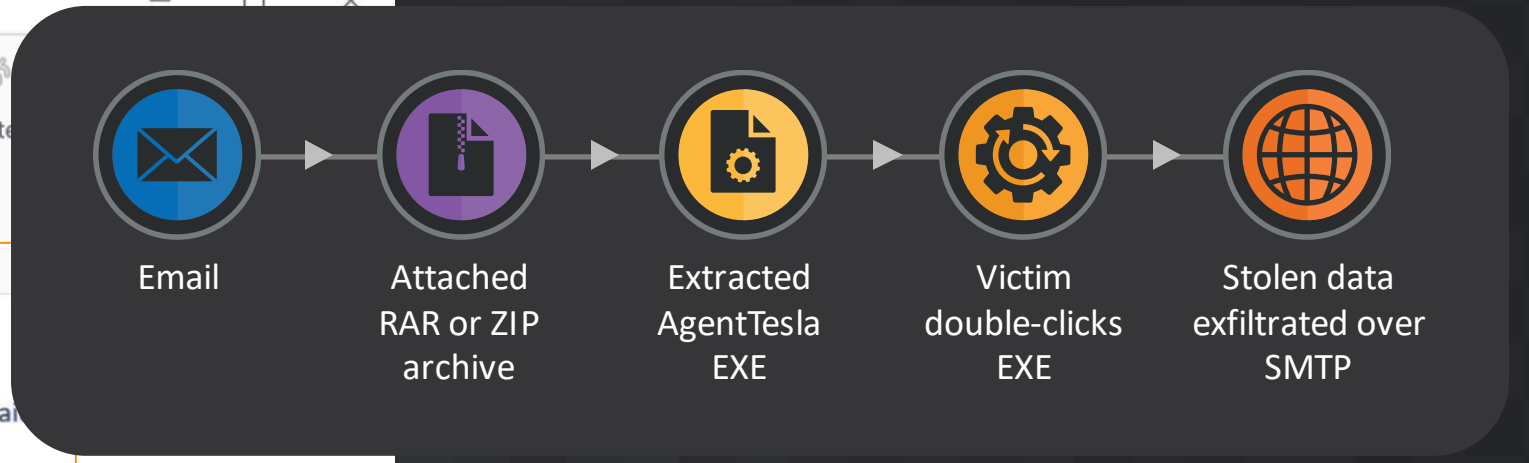
```
##  ( Subject Context (INV|DEL) | Body Context (INV|CTA) )
##  + HDR_Anomaly + 01 Attached (PPT_Macro)
@when_all(
    (m.attachment.file_metadata.file_type == 'ole')
    & (m.email.attachment_count == 1)
    & (m.attachment.file_context.ole_context.anyItem(item == 'OLE_IS_PPT'))
    & (m.attachment.file_context.ole_context.anyItem(item == 'OLE_Has_Macro'))
    & (m.email.sender_context.anyItem(item == 'EML_HDR_Ret_Anomaly'))
    & (
        (m.email.subject_context.anyItem(item.matches('INV|DEL')))
        | (m.email.body_context.anyItem(item.matches('INV|CTA')))
    )
)
def rule_ole_context_with_macro_ppt(c):
    c.s.verdict = 'Malicious: INV Context with Macro PPT'
```

More Examples of Durable Rules

```
##  Subject Context (FIN) + Body Context (FIN) + HTML_with_Obfuscation
@when_all(
    (m.attachment.file_metadata.file_type == 'html')
    & (m.attachment.file_metadata.file_size > 20)
    & (m.attachment.file_context.html_context.anyItem(item == 'HTM_Has_Obfus'))
    & (
        (m.email.body_context.anyItem(item == 'FIN'))
        & (m.email.subject_context.anyItem(item.matches('FIN')))
    )
)
def rule_html_body_obfus_html(c):
    c.s.verdict = 'Malicious: FIN Body Context with HTML Obfuscation'
```

AFA identifying Mal spam pushing AGENT-TESLA



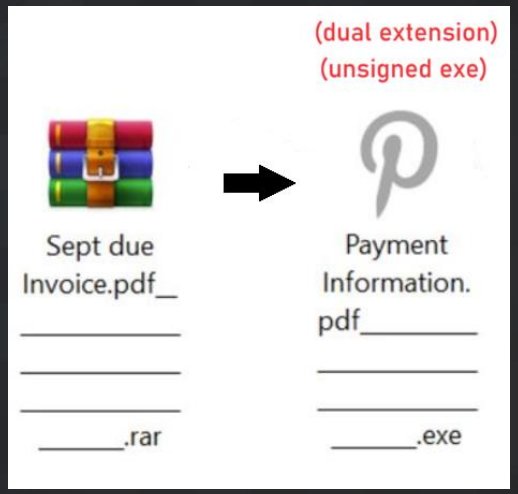


HDR Context:
Anomaly in Reply-To header
RE: Subject with Payment context

Body Context:
Information about Payment Invoice

Attachment Name:
Dual Ext filename

Attachment Context:
RAR -> 1 EXE (unsigned)
Dual Ext File Embedded

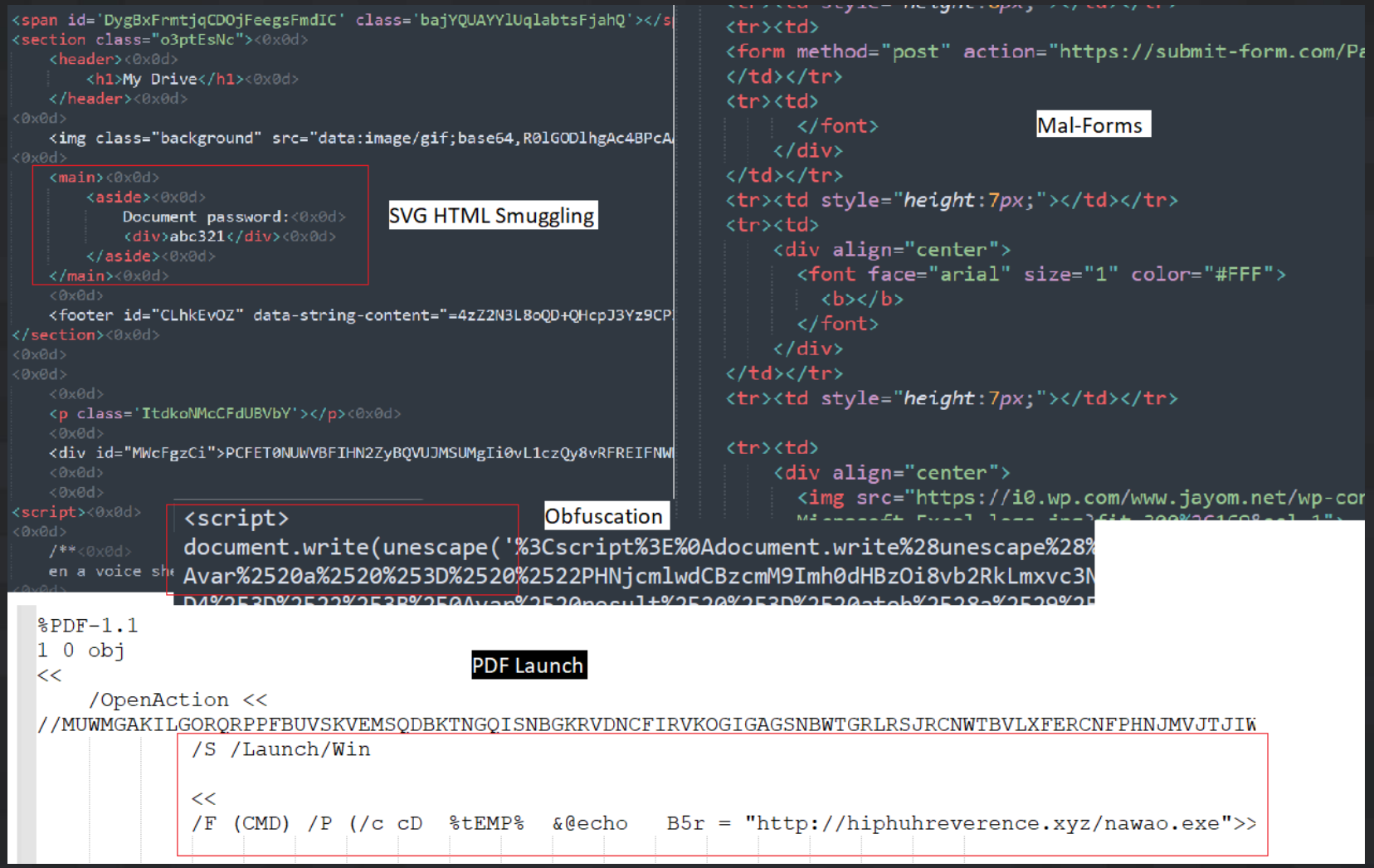


cisco | TALOS

Results of Adaptive File Analyzer

- The engine is rigorously tested with ~800K to 1M samples monthly for True Positives / False Positives.
  - Since Apr to Sep 2023 :
    - 25626 unique SHA256 hits | Were categorized from unseen to malicious on AMP Cloud
    - Out of which 11815 were HTML (Smuggling / Phishing) hashes
- Hits by File Type:
  - HTML, HTM, SHTML - ~46%
  - Archives (ZIP, RAR, ISO, etc.) - ~19%
  - OLE (Macro, Embedded, Malware) - ~10%
  - PDF - ~9%
  - All Other - ~16%
- Type of Malware seen were HTML Smuggling, RATs, Bots, Macro/PDF Downloaders etc.

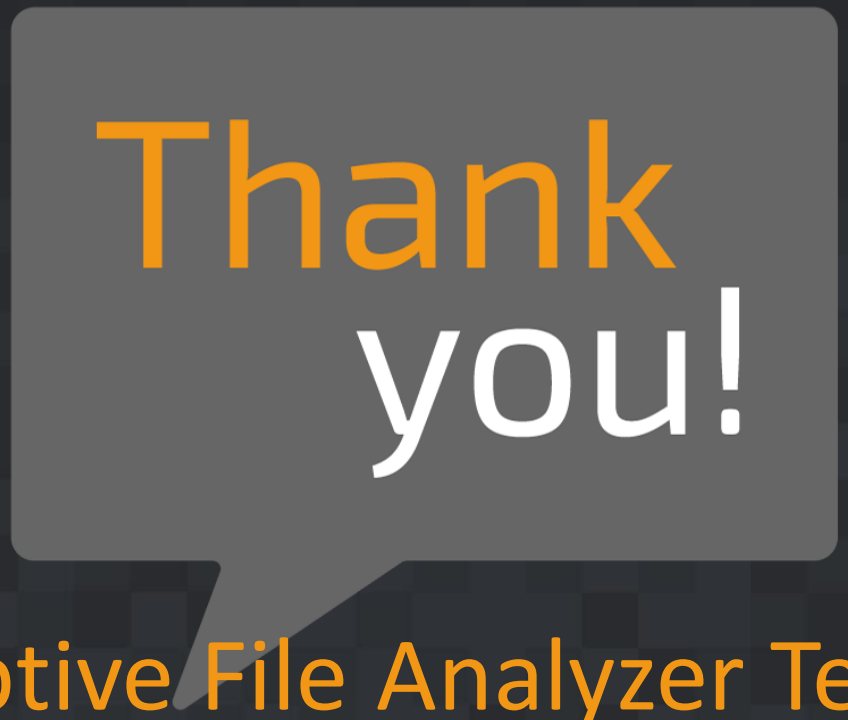
Types of Detections Seen in Live Traffic



Conclusion

1. Results shows that in many cases (not all), Adaptive File Analyzer (AFA) can be used to classify emails as malicious or benign
2. Anomaly pattern rules are the key to detect new campaigns and even unknown malwares
3. Multi-Stage attachments provide a challenge to detect. Capturing every stage of the malware may not be feasible, leading to false negatives by dynamic analysis technology. AFA solves the problem of detecting multi-stage attachments

cisco | TALOS



Adaptive File Analyzer Team

Abhishek Singh & Kalpesh Mantri

TALOSINTELLIGENCE.COM

Acknowledgments:

We want to thank Eric Peterson for his feedback on the design discussion and Shray Kapoor for helping with the implementation.