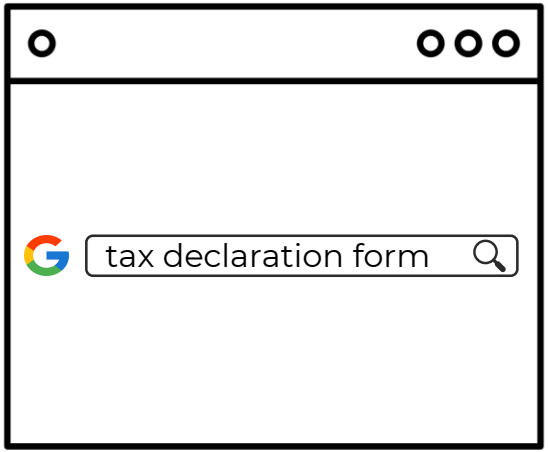
# From Attachments to SEO: Click Here to Learn More about Clickbait PDFs!

*Giada Stivala, Sahar Abdelnabi, Andrea Mengascini, Mariano Graziano, Mario Fritz, Giancarlo Pellegrino*

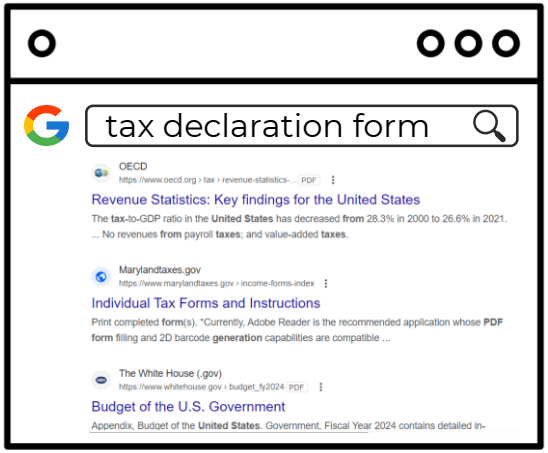
ACSAC 2023

Austin, TX

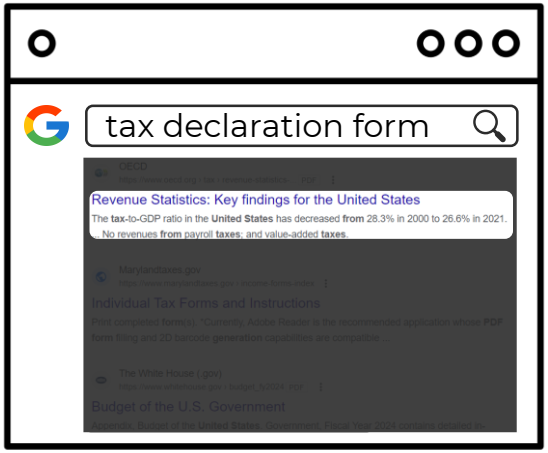
# Execution of a Clickbait PDF Attack



# Execution of a Clickbait PDF Attack

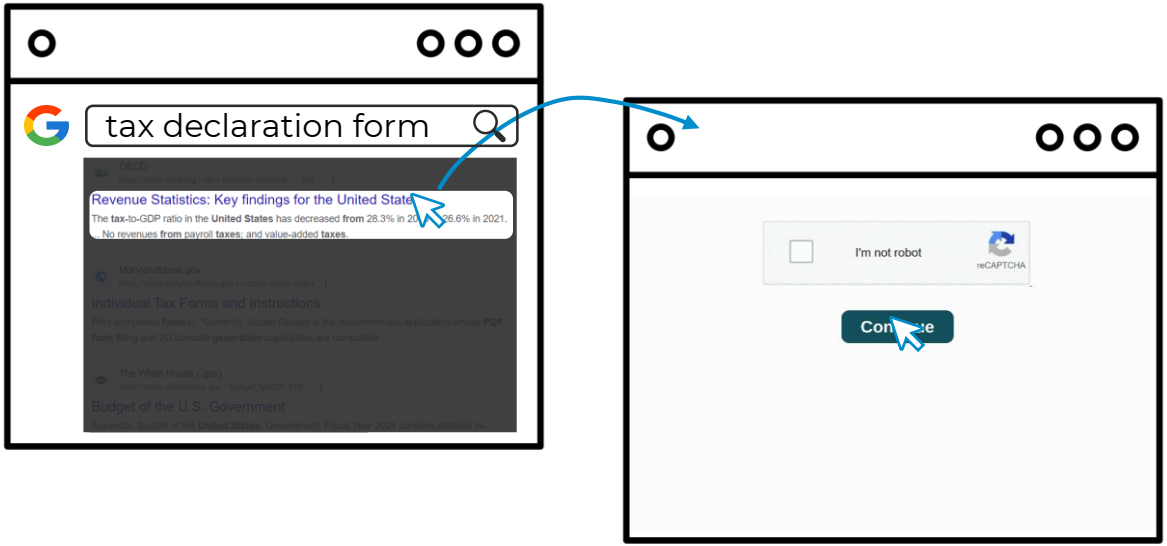


# Execution of a Clickbait PDF Attack

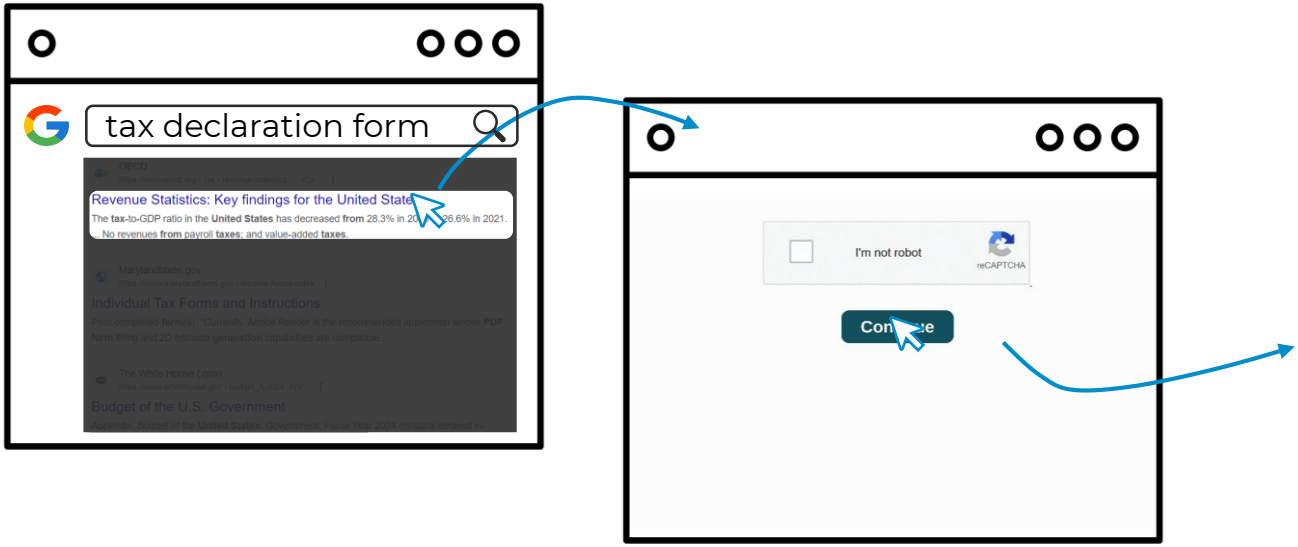


# Execution of a Clickbait PDF Attack

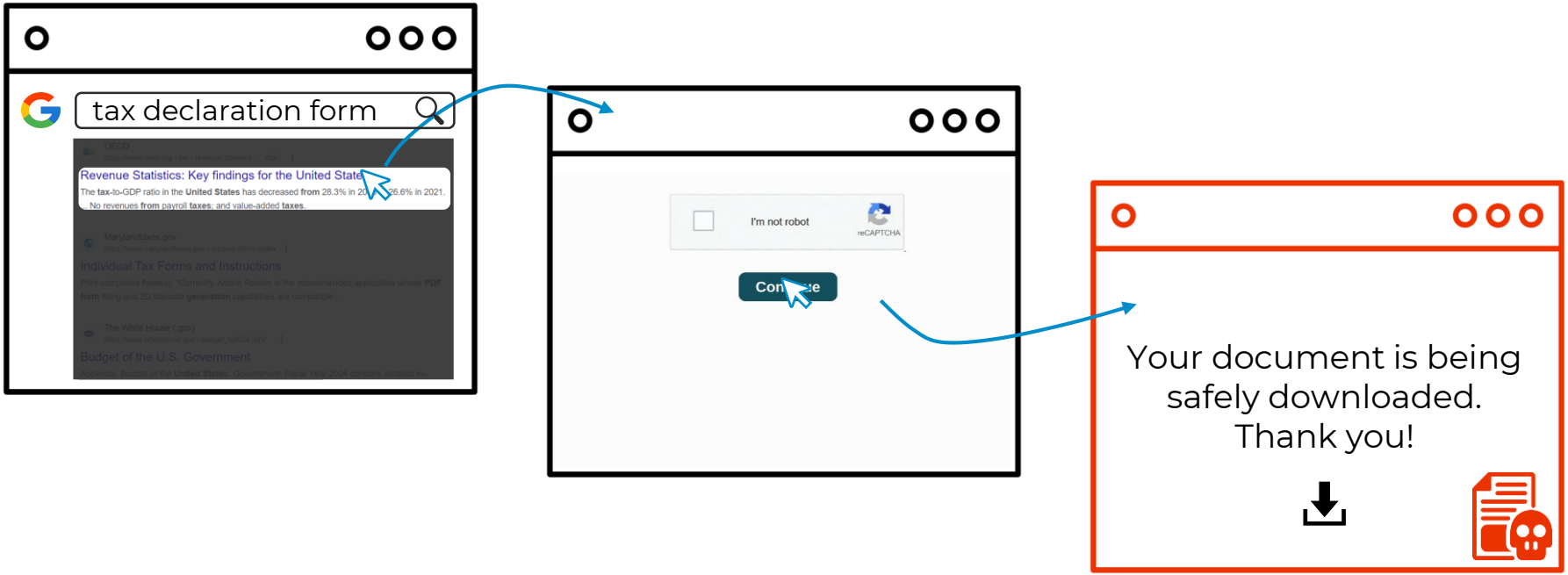
# Execution of a Clickbait PDF Attack



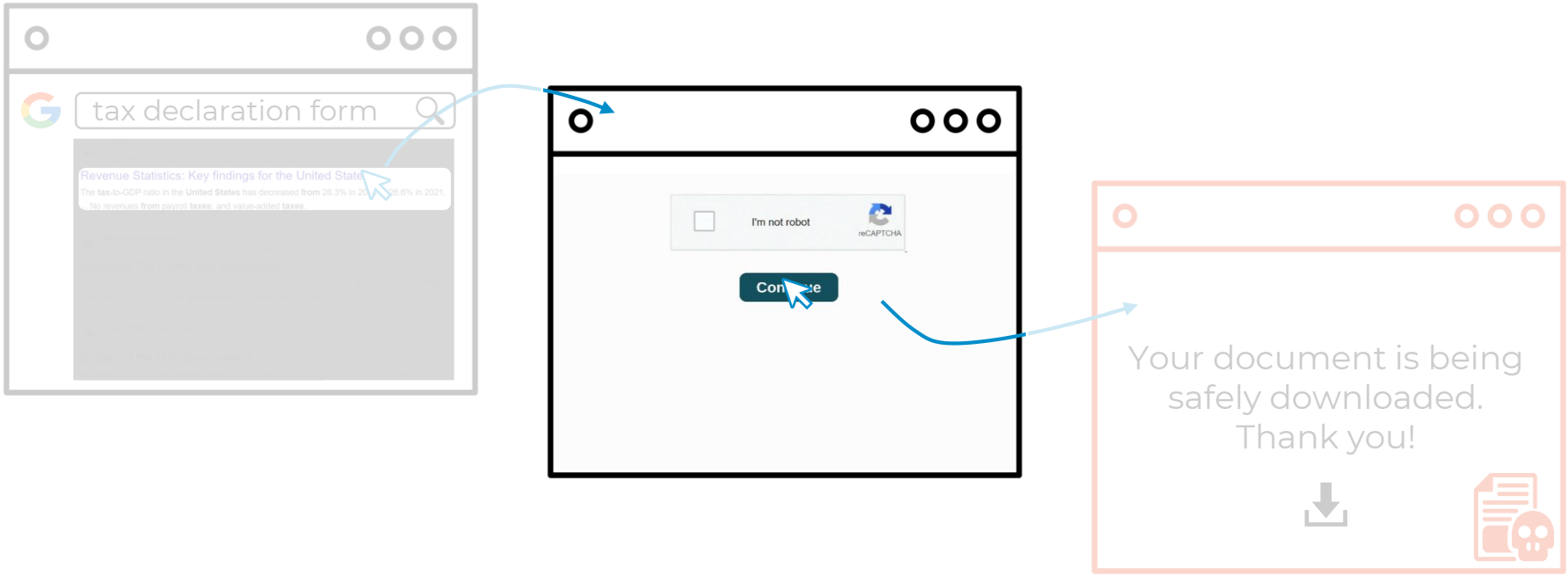
# Execution of a Clickbait PDF Attack



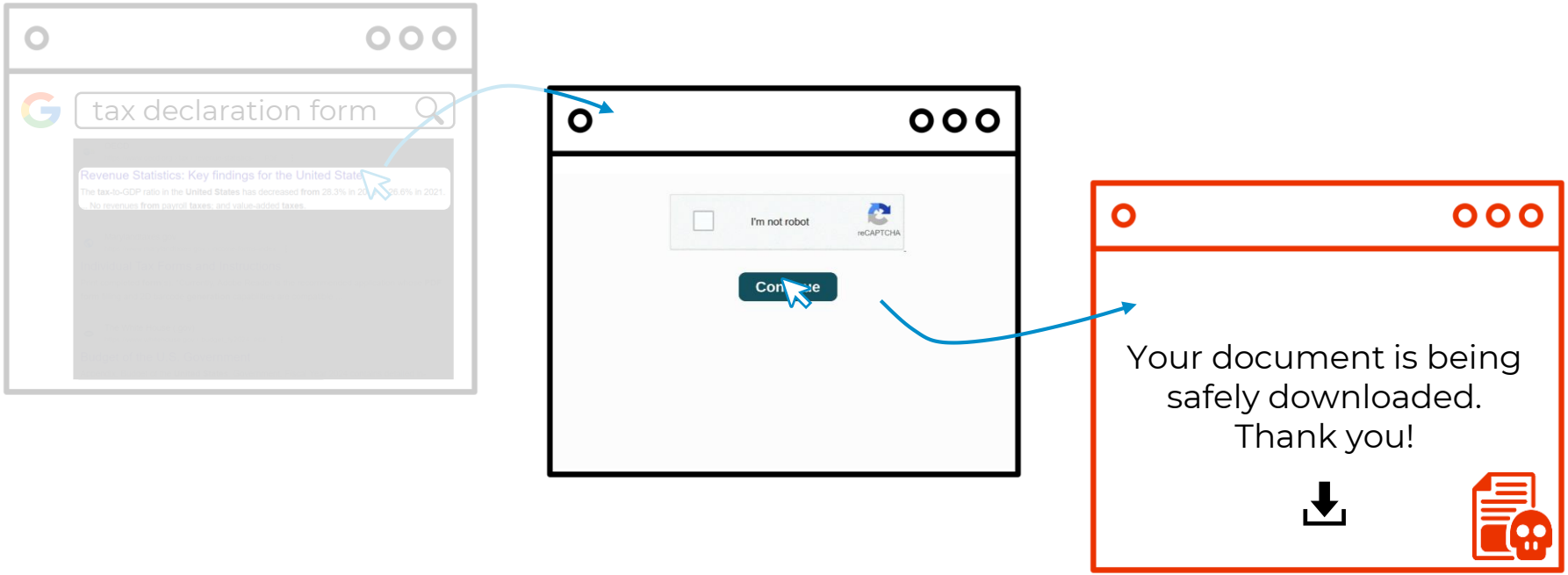
# Execution of a Clickbait PDF Attack



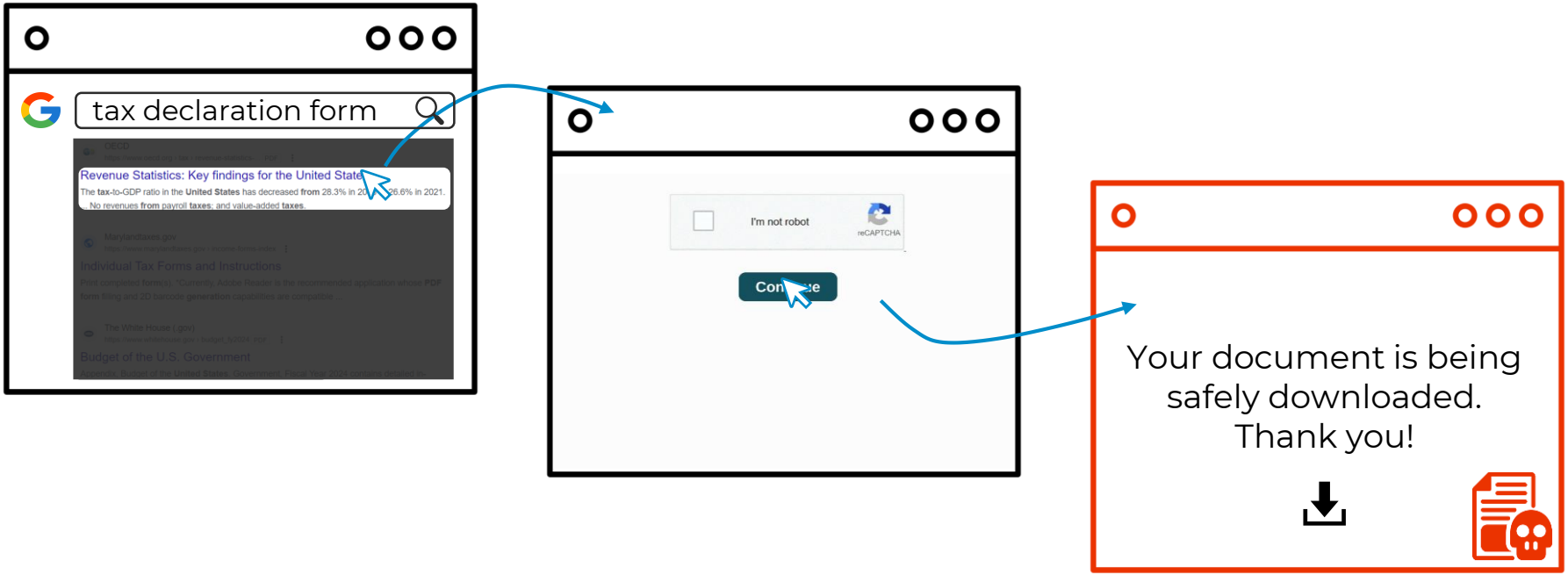
# Execution of a Clickbait PDF Attack



# Execution of a Clickbait PDF Attack

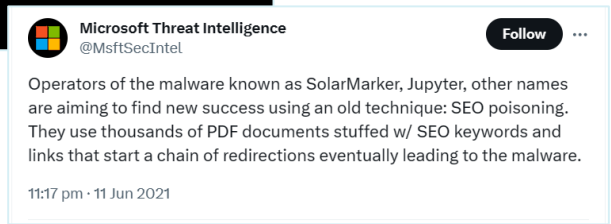


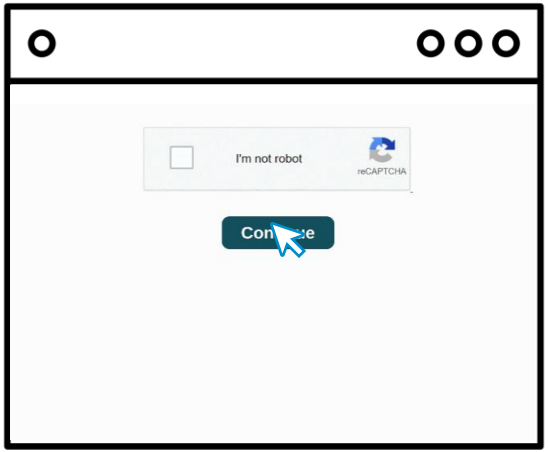
# Execution of a Clickbait PDF Attack



# Scope and Research Questions

Millions of Clickbait PDFs online [1, 2]



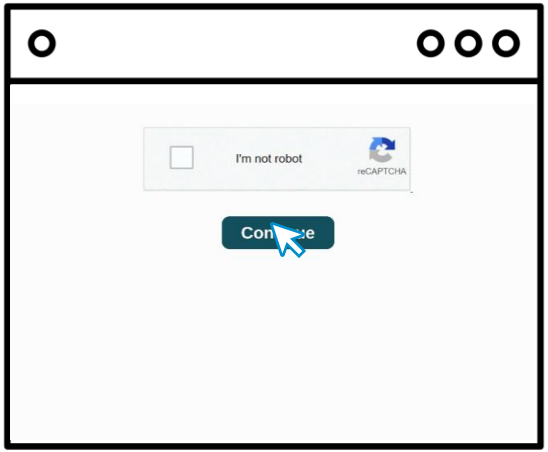

[1] https://unit42.paloaltonetworks.com/phishing-trends-with-pdf-files/
[2] https://twitter.com/MsftSecIntel/status/1403461397283950597

# Scope and Research Questions

Millions of Clickbait PDFs online [1, 2]

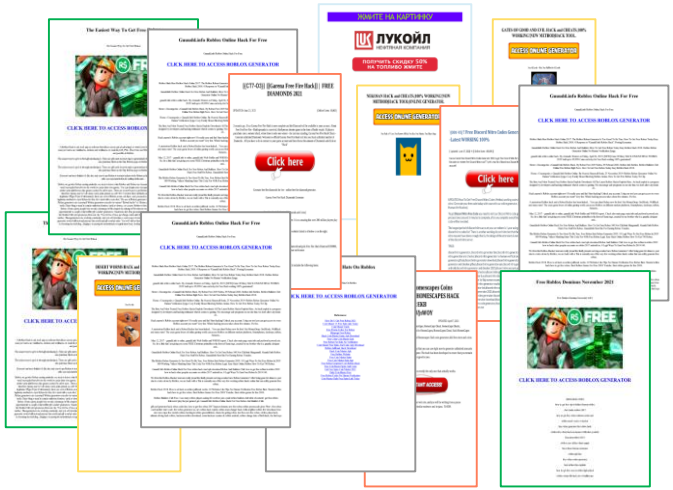
We characterize:

- Volume and duration
- Textual and visual baits
- Web attacks
- Distribution methods




[1] https://unit42.paloaltonetworks.com/phishing-trends-with-pdf-files/
[2] https://twitter.com/MsftSecIntel/status/1403461397283950597

# Data Collection

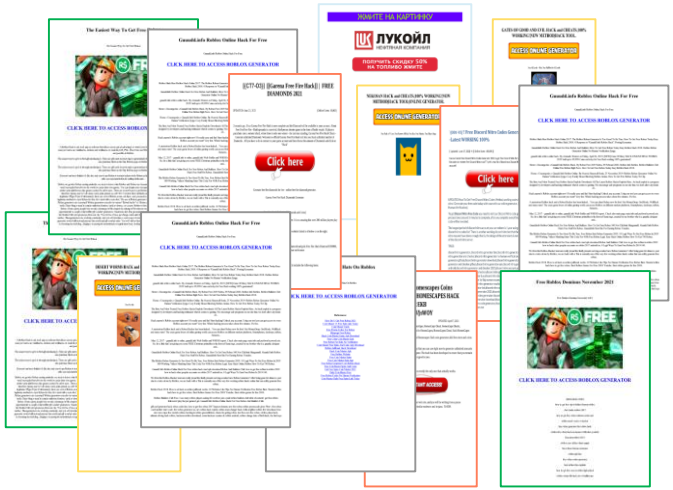



- Collected 176,208 PDFs with unique SHA256, shared by industrial partners

INQUEST LABS

CISCO

# Data Collection

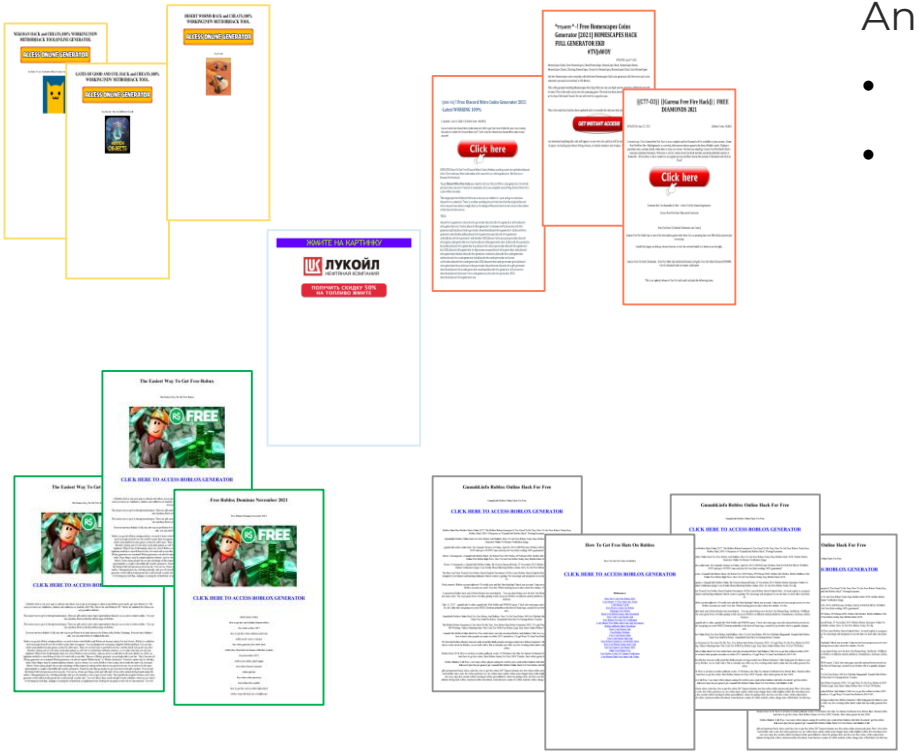



- Collected 176,208 PDFs with unique SHA256, shared by industrial partners
- PDFs do not embed malicious code:
  - Used VirusTotal tags `js-embedded`, `file-embedded`, `exploit`, `cve-xxxx`, and `launch-action`

INQUEST LABS

CISCO

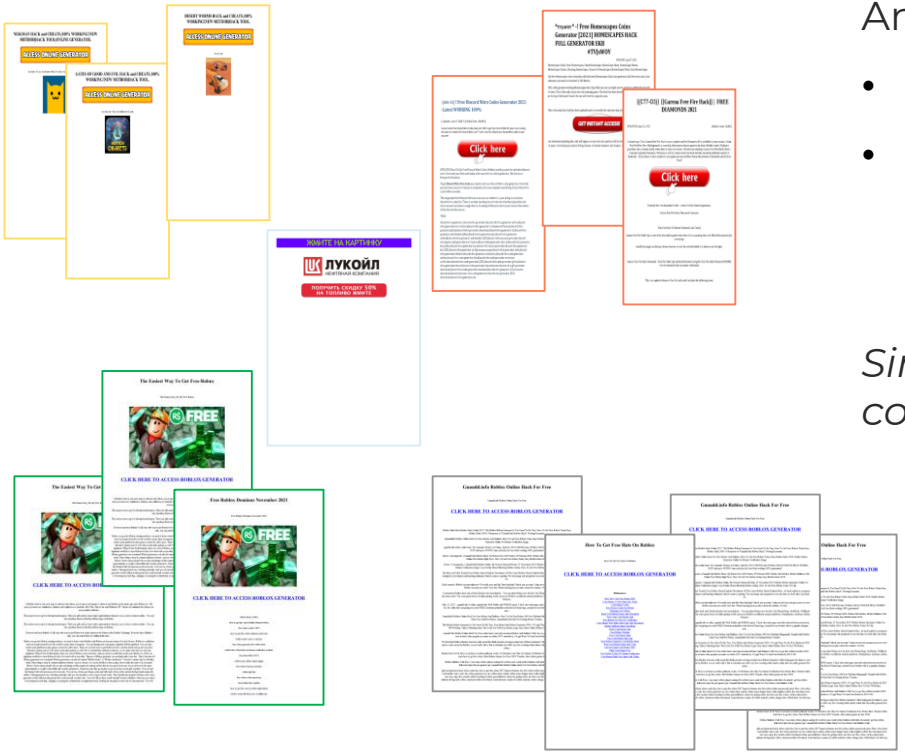
# Characterizing Clickbait PDFs at Scale




Analyze feed of PDFs

- At scale
- Daily

# Characterizing Clickbait PDFs at Scale

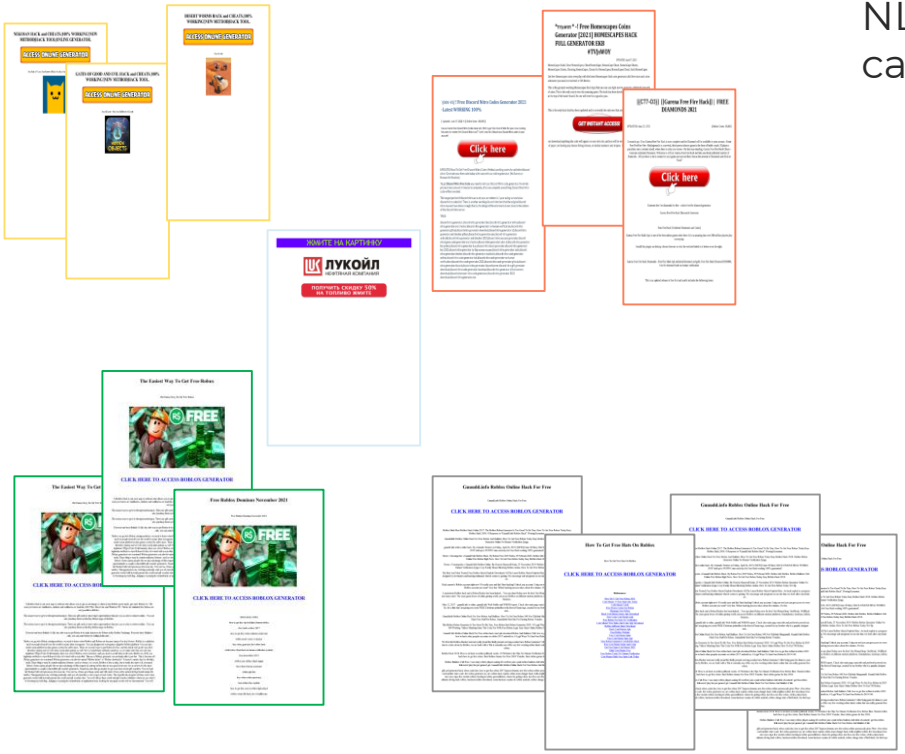



Analyze feed of PDFs

- At scale
- Daily

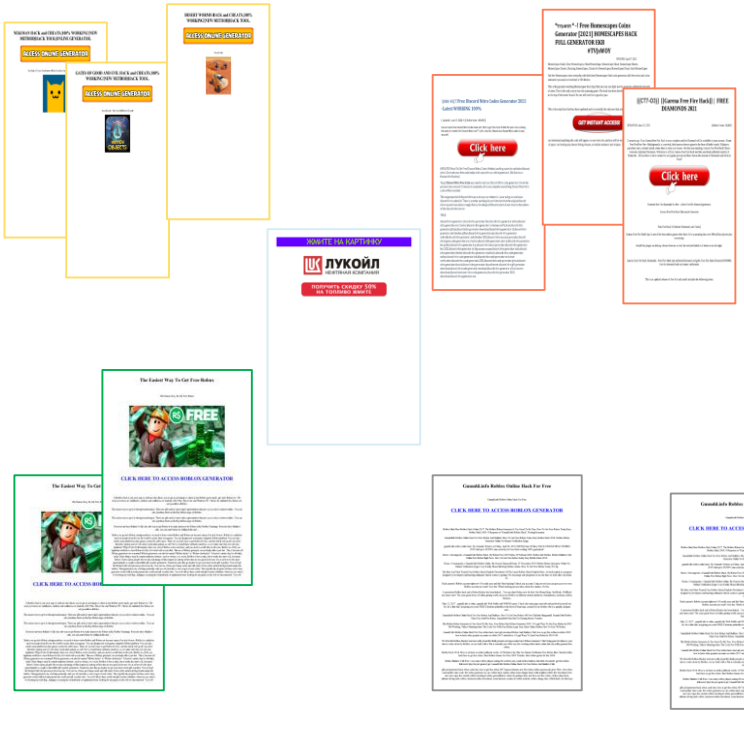
*Similarity metrics to reduce complexity of task*

# Characterizing Clickbait PDFs at Scale



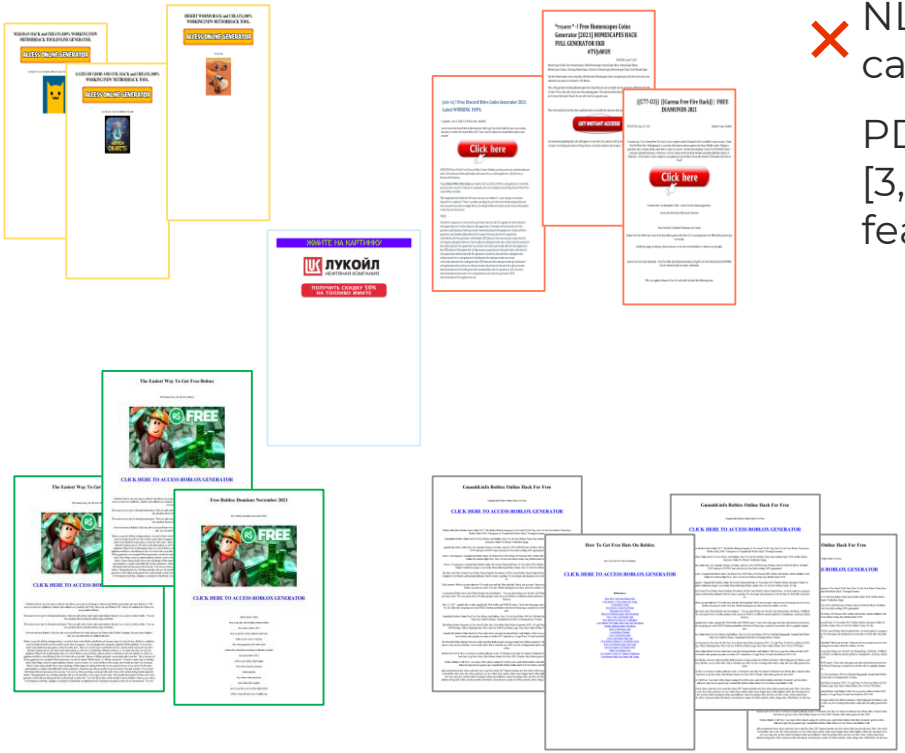

NLP-based clustering does not capture visual characteristics

# Characterizing Clickbait PDFs at Scale




NLP-based clustering does not capture visual characteristics

# Characterizing Clickbait PDFs at Scale

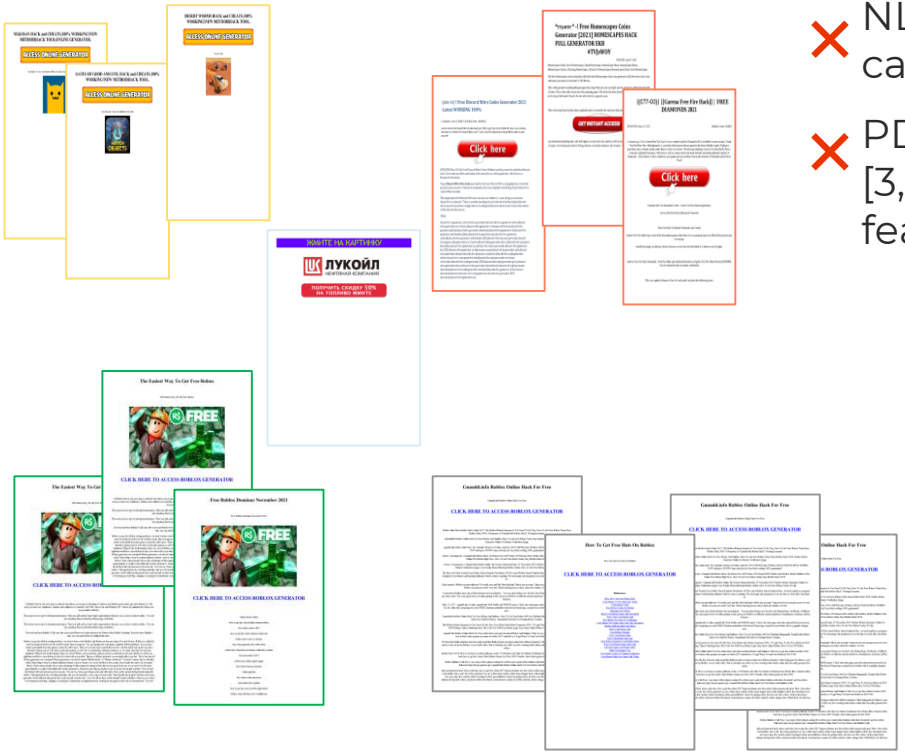



- NLP-based clustering does not capture visual characteristics

- PDF structure-based techniques [3, 4] select unrelated structure features

[3] Šrndic, N. et al. "Detection of Malicious PDF Files Based on Hierarchical Document Structure.“ NDSS 2013
[4] Chen, Yizheng, et al. "On training robust PDF malware classifiers." 29th USENIX 2020

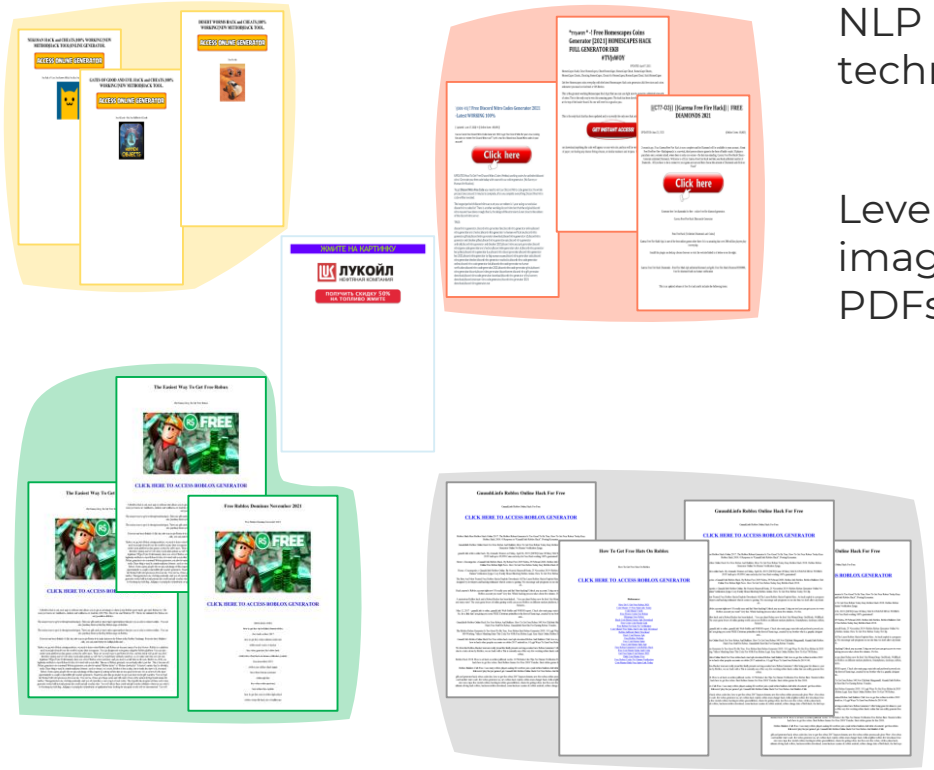
# Characterizing Clickbait PDFs at Scale




- ✗ NLP-based clustering does not capture visual characteristics
- ✗ PDF structure-based techniques [3, 4] select unrelated structure features

[3] Šrndic, N. et al. "Detection of Malicious PDF Files Based on Hierarchical Document Structure.“ NDSS 2013
[4] Chen, Yizheng, et al. "On training robust PDF malware classifiers." 29th USENIX 2020

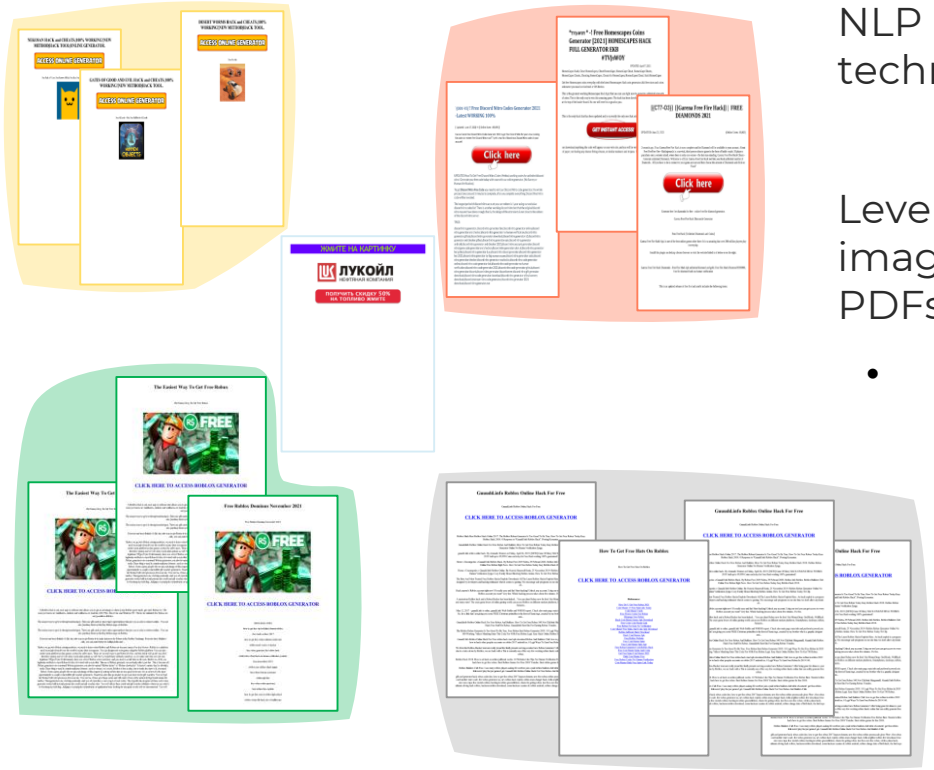
# Characterizing Clickbait PDFs at Scale




NLP and PDF structural similarity techniques ineffective for clustering

Leverage visual similarities in images on the first page of clickbait PDFs

# Characterizing Clickbait PDFs at Scale

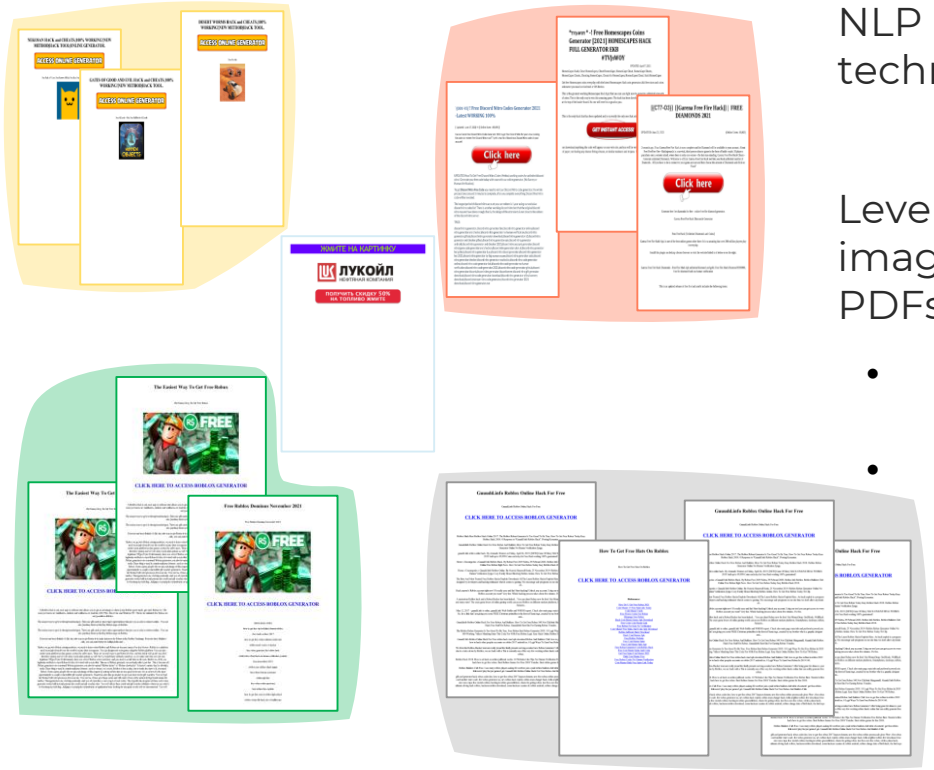



NLP and PDF structural similarity techniques ineffective for clustering

Leverage visual similarities in images on the first page of clickbait PDFs

- Unknown number of unique baits

# Characterizing Clickbait PDFs at Scale

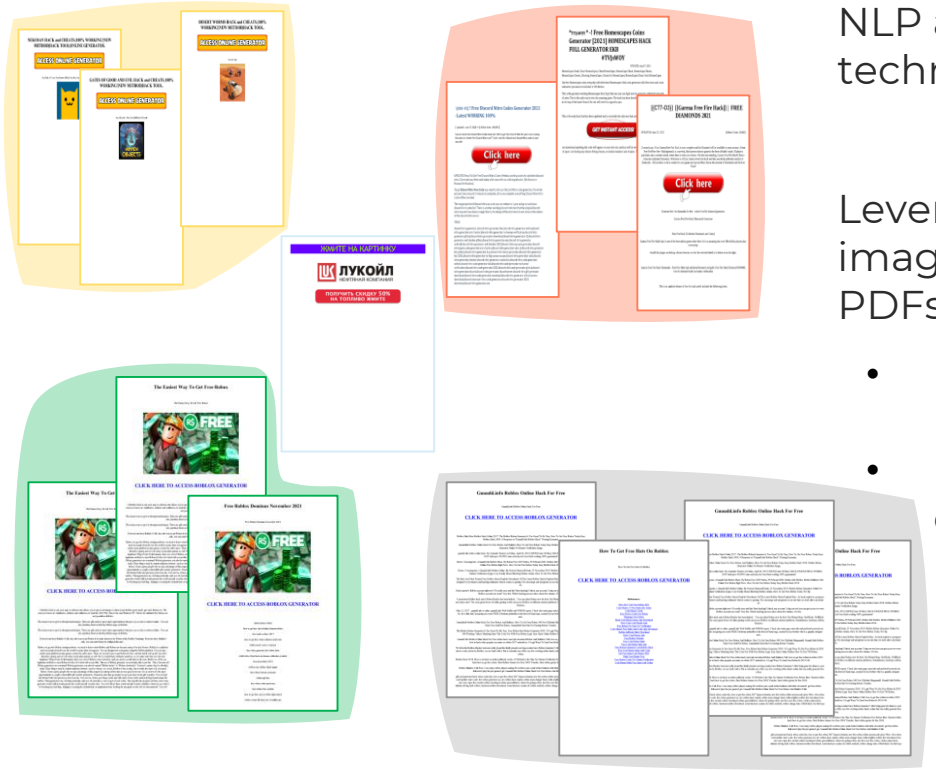



NLP and PDF structural similarity techniques ineffective for clustering

Leverage visual similarities in images on the first page of clickbait PDFs

- Unknown number of unique baits
- Unknown distribution across dataset

# Characterizing Clickbait PDFs at Scale



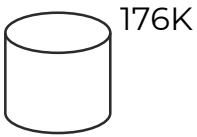

NLP and PDF structural similarity techniques ineffective for clustering

Leverage visual similarities in images on the first page of clickbait PDFs

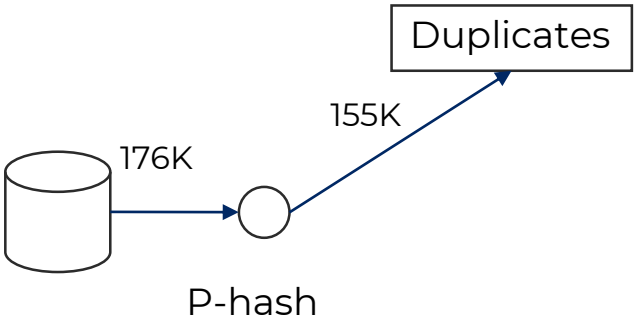
- Unknown number of unique baits
- Unknown distribution across dataset
  - ➢ DeepCluster [5]: CNN with k-means clustering

[5] Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." ECCV 2018

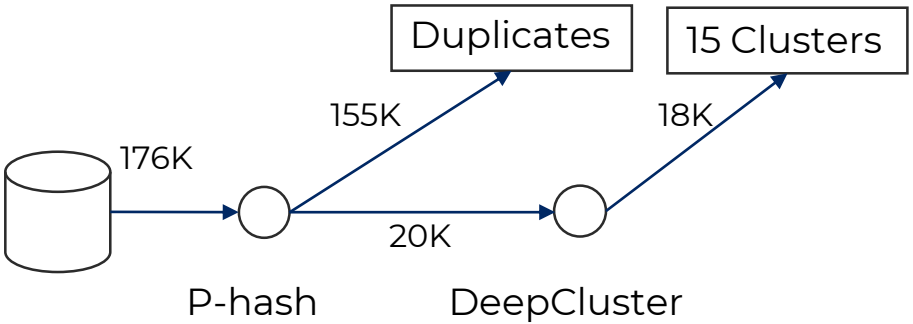
# Manually-Aided Clustering



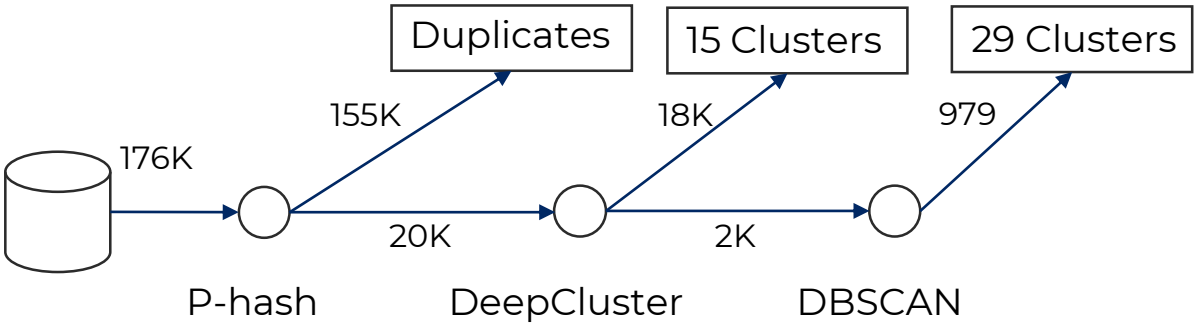
# Manually-Aided Clustering



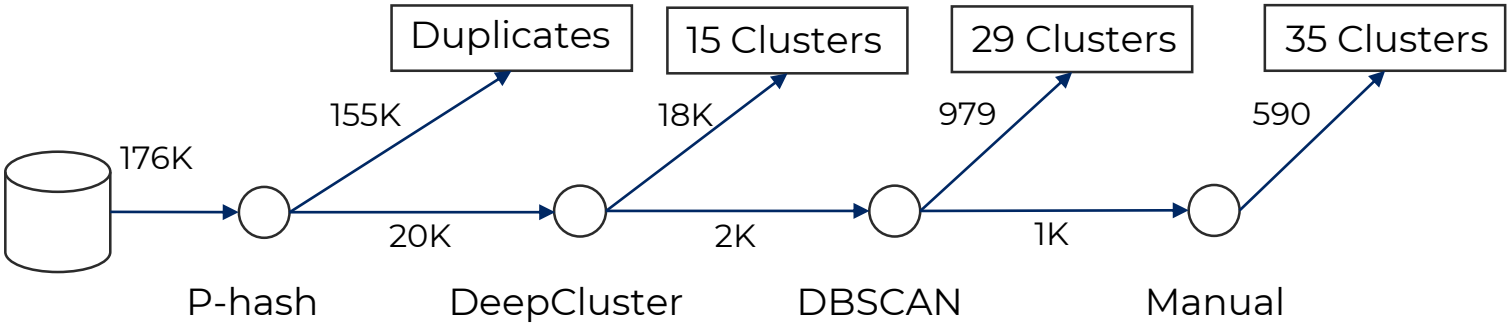
# Manually-Aided Clustering



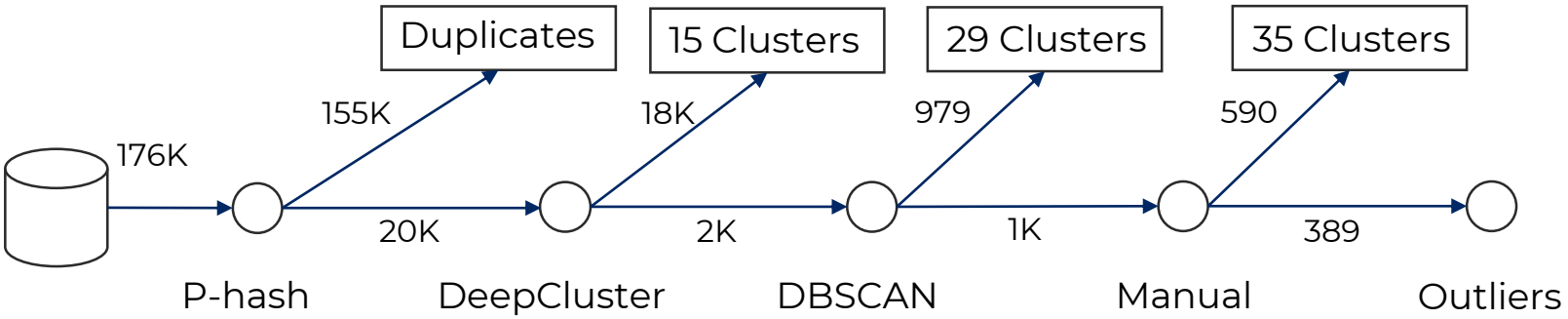
# Manually-Aided Clustering



# Manually-Aided Clustering



# Manually-Aided Clustering



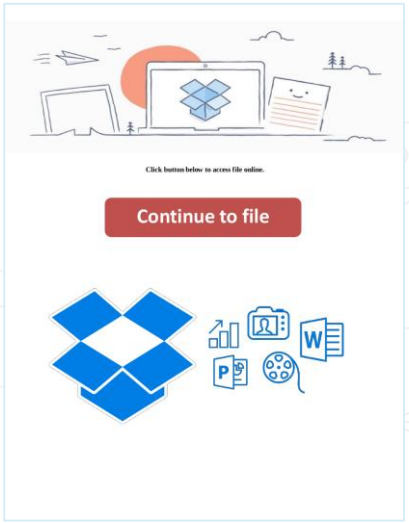

80 clusters, including Outliers

# Manually-Aided Clustering



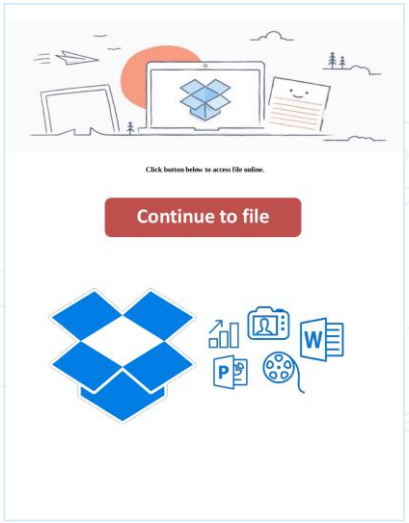

80 clusters, including Outliers

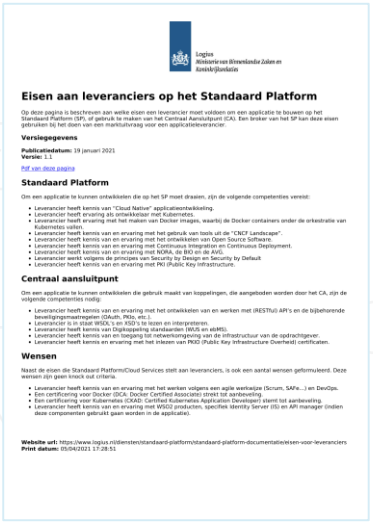
# Manually-Aided Clustering



80 clusters, including Outliers.

# Manually-Aided Clustering

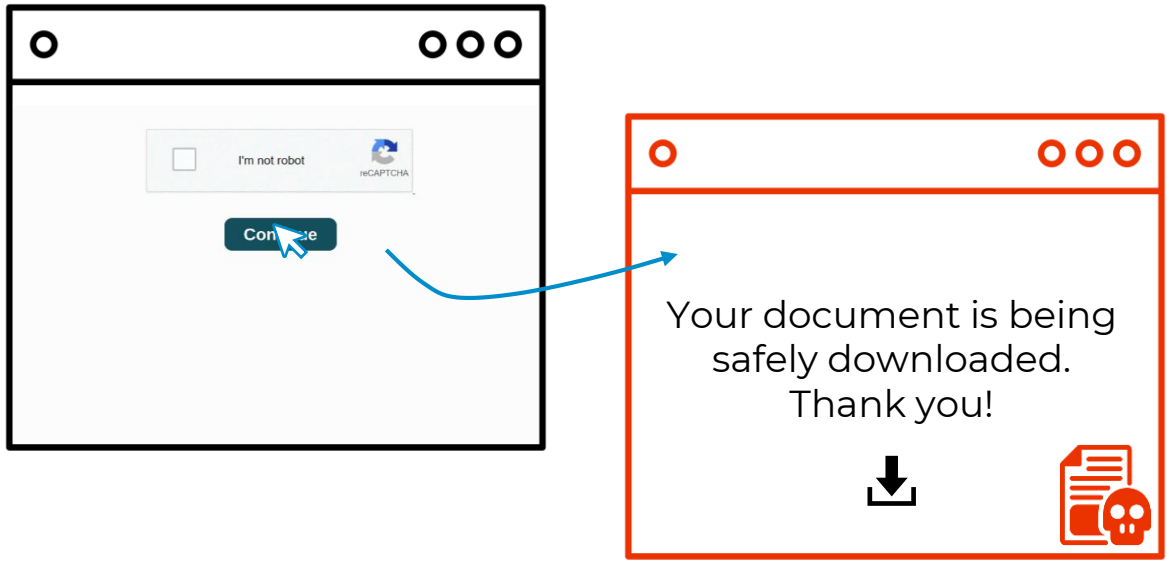





80 clusters, including Outliers

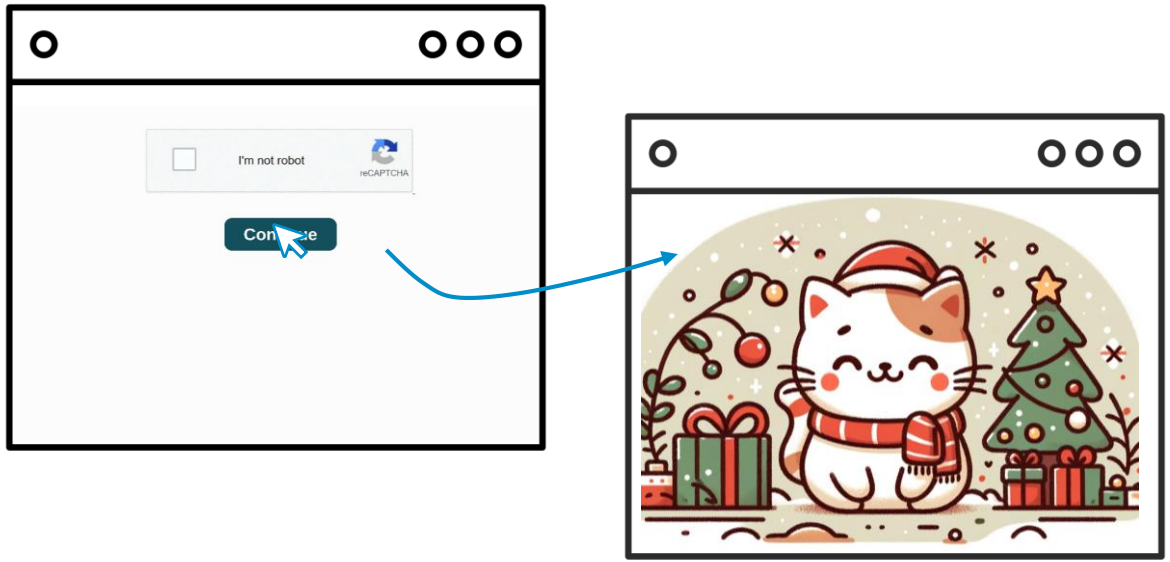
# Determining Maliciousness

- PDF considered as malicious when its first-page URL leads to a Web attack



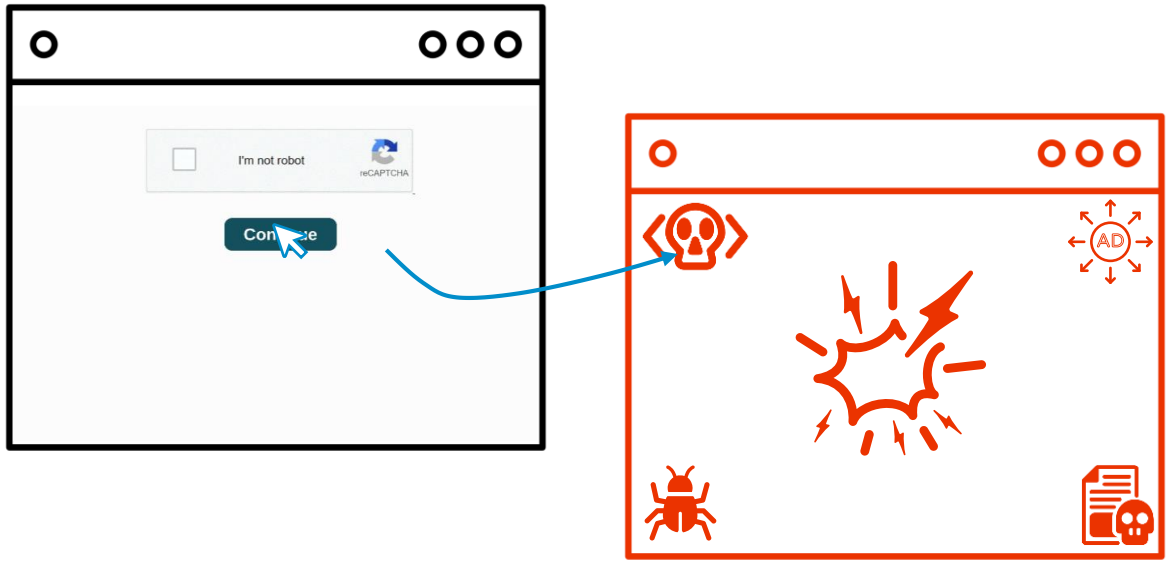
# Determining Maliciousness: Challenges

- PDF considered as malicious when its first-page URL leads to a Web attack
- Challenges: technical (evasion, multiple attacks)



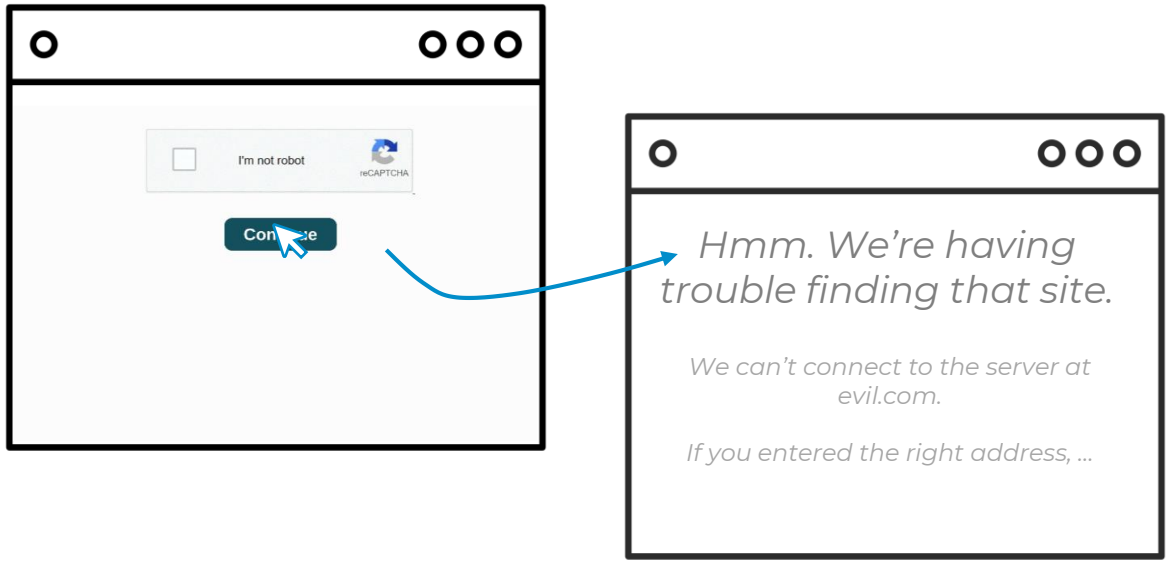
# Determining Maliciousness: Challenges

- PDF considered as malicious when its first-page URL leads to a Web attack
- Challenges: technical (evasion, multiple attacks)

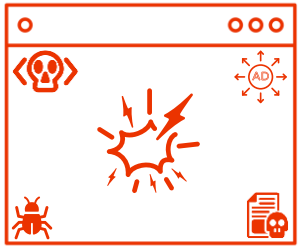


# Determining Maliciousness: Challenges

- PDF considered as malicious when its first-page URL leads to a Web attack
- Challenges: technical (evasion, multiple attacks), temporal (short lifespan)



# Determining Maliciousness: Methodology




Multiple attacks

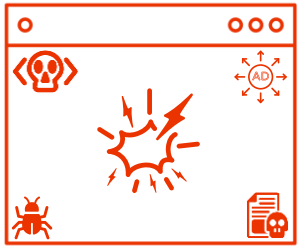
Evasion

Attack page offline

Hybrid approach:

- Third-party URL analysis service w/ historical records
- Manual analysis

# Determining Maliciousness: Methodology



Multiple attacks

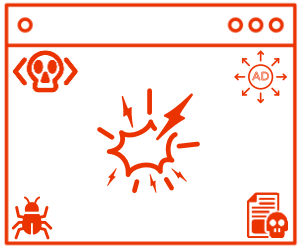
Evasion

Attack page offline

Hybrid approach:

- Third-party URL analysis service w/ historical records
  - VirusTotal Premium: 20K URL analysis reports
- Manual analysis

# Determining Maliciousness: Methodology



Multiple attacks

Evasion

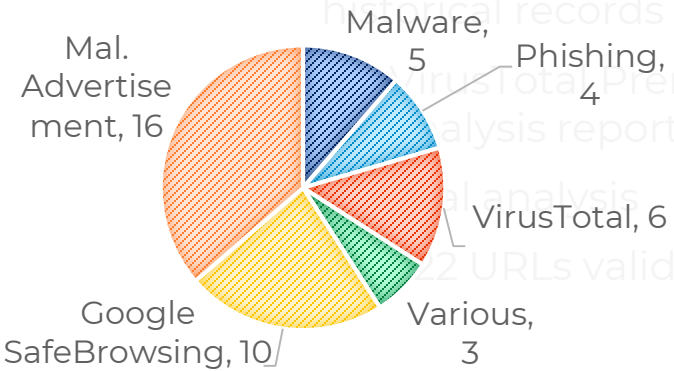
Attack page offline

Hybrid approach:

- Third-party URL analysis service w/ historical records
  - VirusTotal Premium: 20K URL analysis reports
- Manual analysis
  - 722 URLs validated

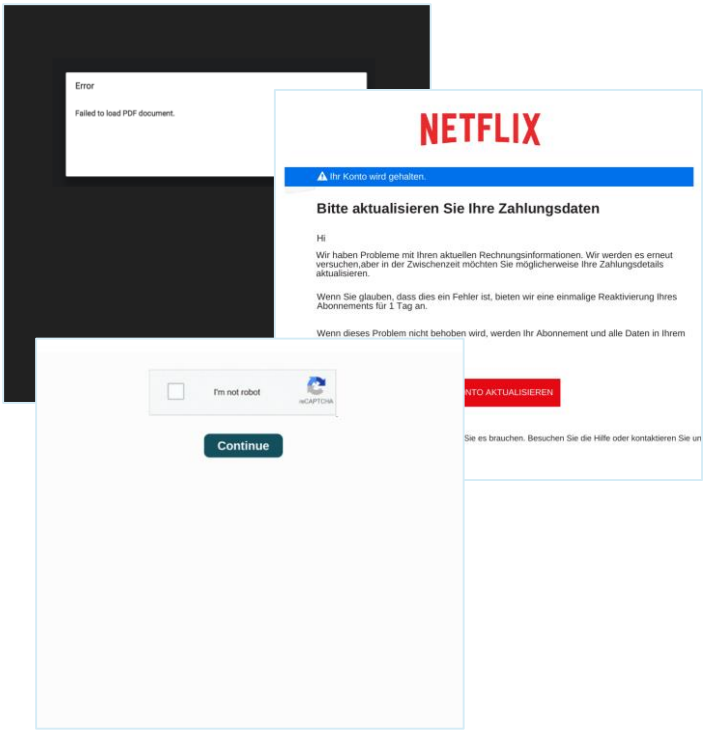
# Determining Maliciousness: Results

44 clusters lead to Web attacks



# Clusters Visual and Volumetric Properties

*Impersonation*




*Promotion*

# Clusters Visual and Volumetric Properties

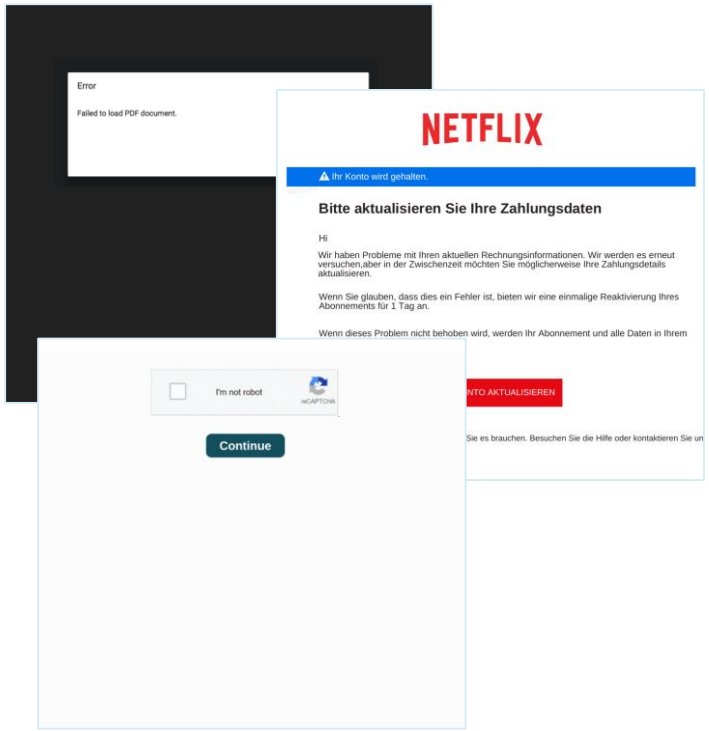
*Impersonation*




*Promotion*

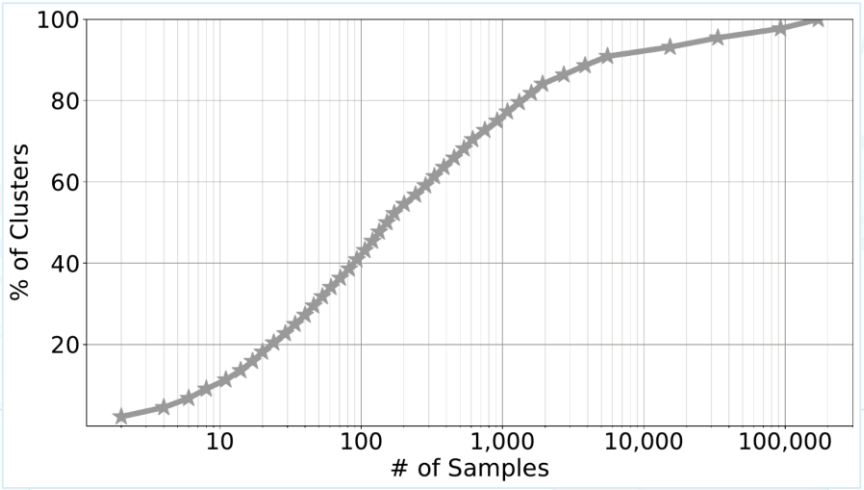
# Clusters Visual and Volumetric Properties

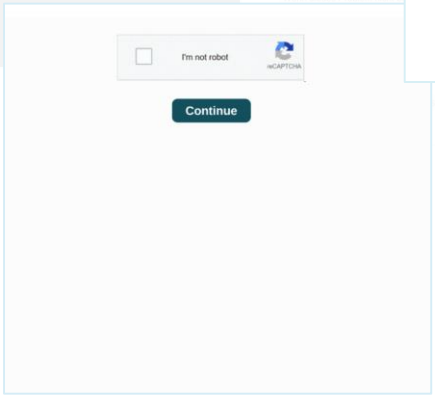
*Impersonation*

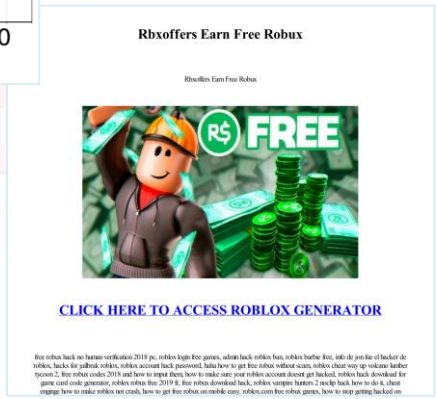



*Promotion*

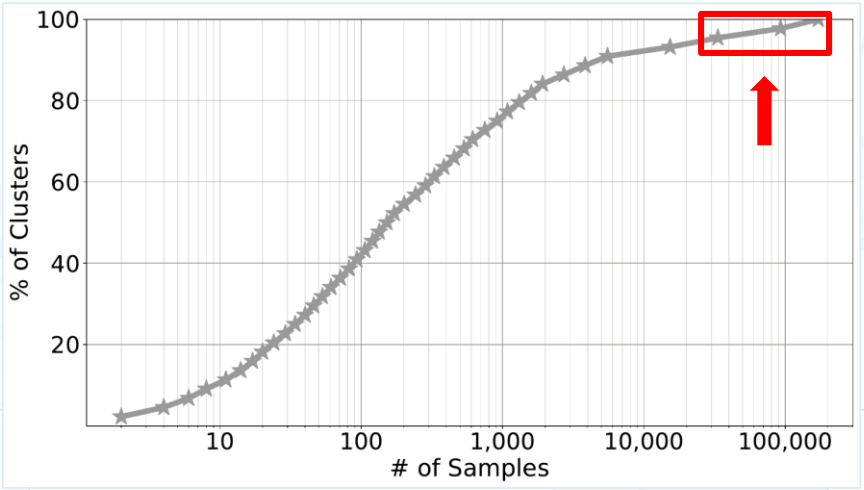
# Clusters Visual and Volumetric Properties

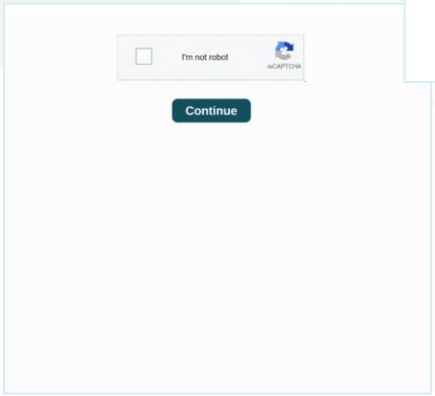


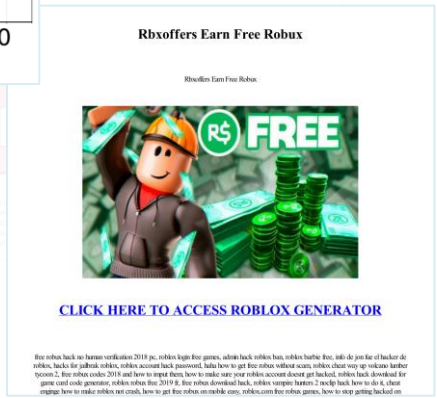





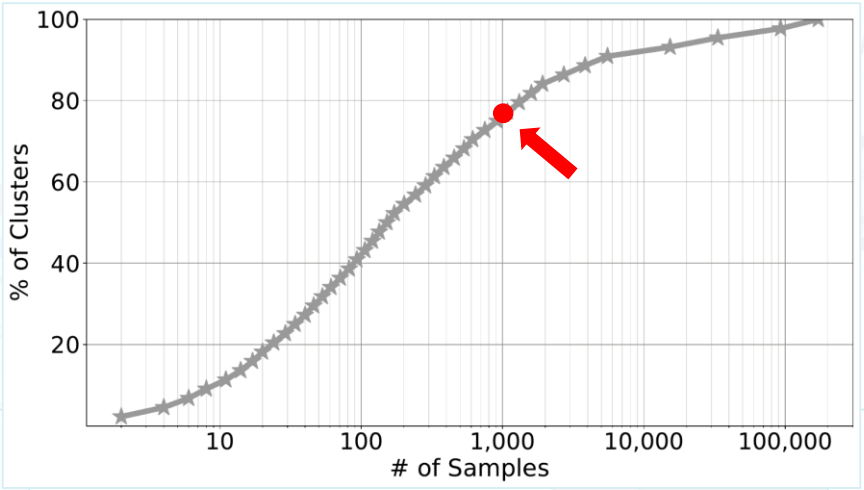
# Clusters Visual and Volumetric Properties

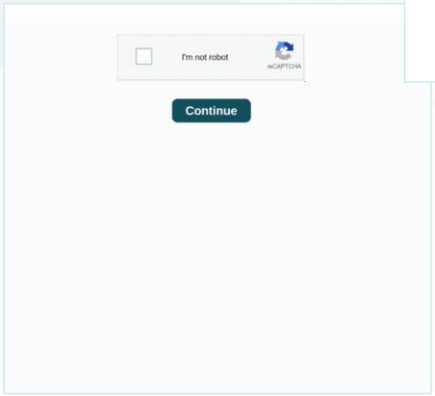


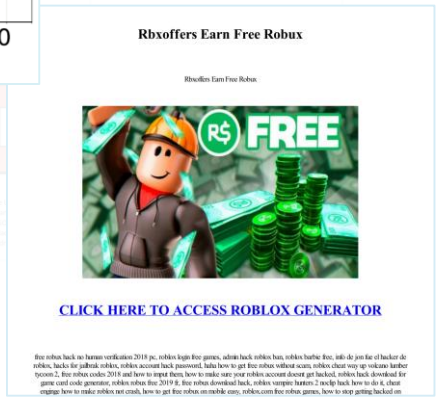





# Clusters Visual and Volumetric Properties





# Distribution Vectors

Indicators of Search Engine Optimization (SEO) attacks:

- Keyword stuffing
- Cross-linking of resources
- Benign websites

First page

Second and last page

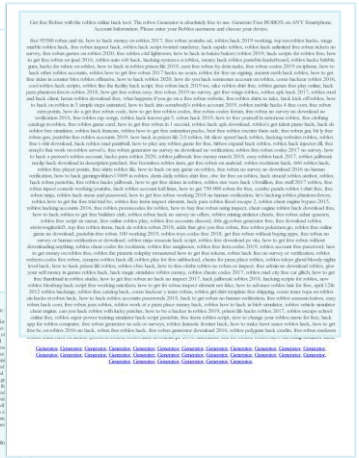
# Distribution Vectors

Indicators of Search Engine Optimization (SEO) attacks:

- Keyword stuffing
- Cross-linking of resources
- Benign websites

First page

Second and last page



# Distribution Vectors

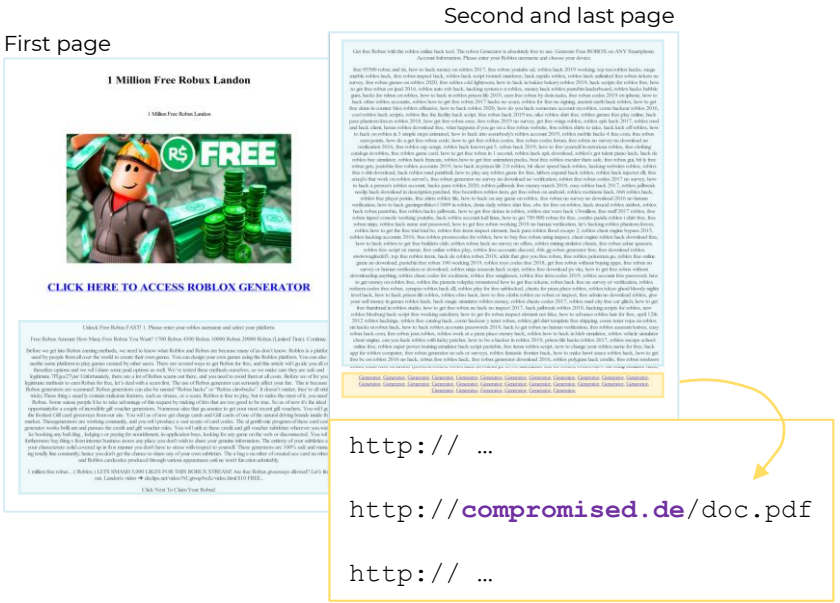
Indicators of Search Engine Optimization (SEO) attacks:

- Keyword stuffing
- Cross-linking of resources
- Benign websites

# Distribution Vectors

Indicators of Search Engine Optimization (SEO) attacks:

- Keyword stuffing
- Cross-linking of resources
- Benign websites

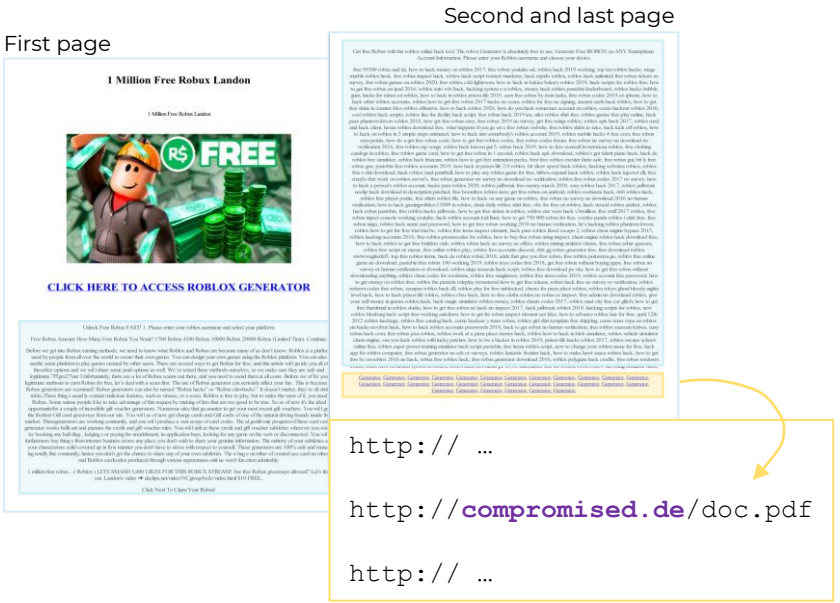


# Distribution Vectors

Indicators of Search Engine Optimization (SEO) attacks:

- Keyword stuffing
- Cross-linking of resources
- Benign websites

*Can clickbait PDFs achieve high search engine rankings?*



# Verifying SEO Distribution: Methodology

Challenges:

- Clickbait PDF titles not popular
- Popular clickbait PDFs taken down

First page

Second and last page

```
http:// …

http://compromised.de/doc.pdf

http:// …
```

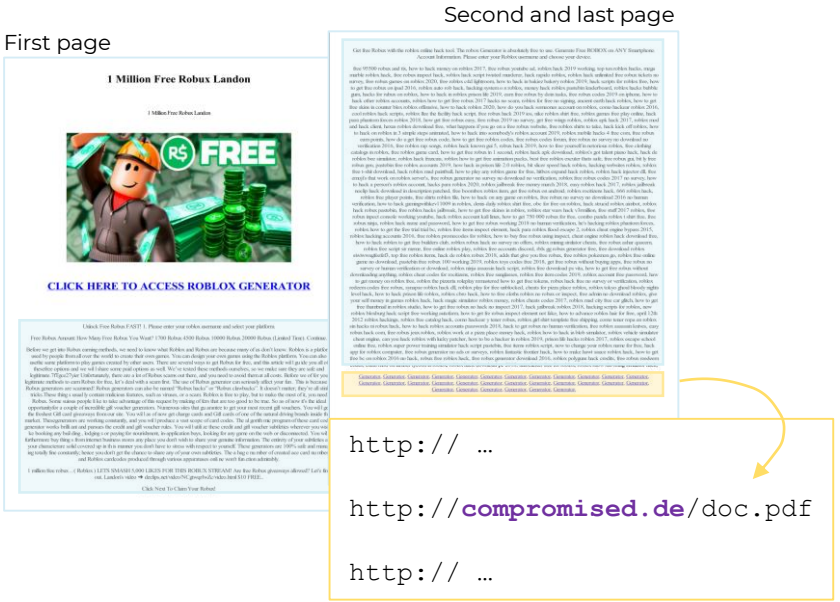
# Verifying SEO Distribution: Methodology

Challenges:

- Clickbait PDF titles not popular
- Popular clickbait PDFs taken down

Solution:

- Find unseen, freshly-uploaded clickbait PDFs



# Verifying SEO Distribution: Methodology

Challenges:

- Clickbait PDF titles not popular
- Popular clickbait PDFs taken down

Solution:

- Find unseen, freshly-uploaded clickbait PDFs

# Verifying SEO Distribution: Methodology

Challenges:

- Clickbait PDF titles not popular
- Popular clickbait PDFs taken down

Solution:

- Find unseen, freshly-uploaded clickbait PDFs
- Use title as search query on popular SEs (Google, Bing)

First page

Second and last page

```
http:// …

http://compromised.de/doc.pdf

http:// …
```

# Verifying SEO Distribution: Methodology

Challenges:

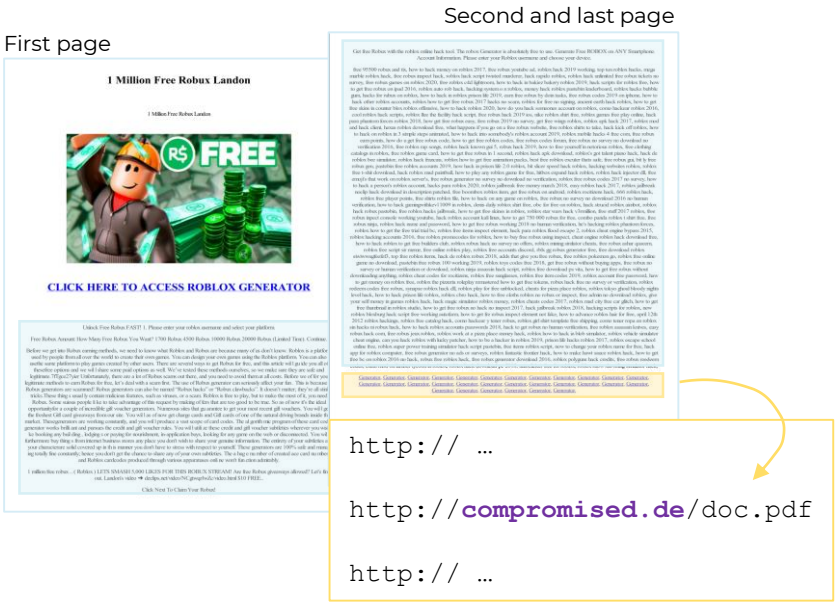
- Clickbait PDF titles not popular
- Popular clickbait PDFs taken down

Solution:

- Find unseen, freshly-uploaded clickbait PDFs
- Use title as search query on popular SEs (Google, Bing)

Success criteria:

- Direct match: URL of originating PDF found
- Cross-link: PDF uses cross-linking



# Verifying SEO Distribution: Results

Challenges:

- Clickbait PDF titles not popular
- Popular clickbait PDFs taken down

Solution:

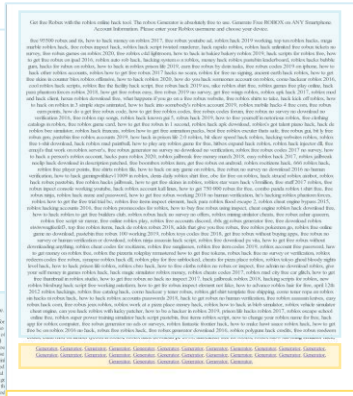
- Find unseen, freshly-uploaded clickbait PDFs
- Use title as search query on popular SEs (Google, Bing)

Success criteria:

- Direct match: URL of originating PDF found
- Cross-link: PDF uses cross-linking

First page

Second and last page



```
http:// …

http://compromised.de/doc.pdf

http:// …
```

- Bing: direct (3K, 60/day) and cross-link match (6K, 104/day)
- Google: cross-link match (925, 16/day)

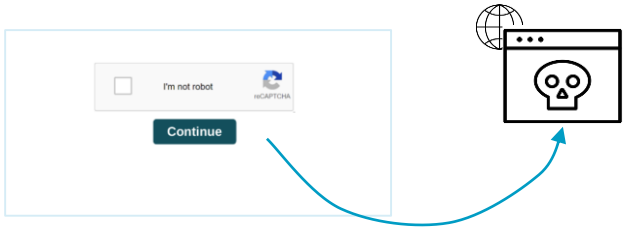
# Takeaways

Many well-defined visual baits

# Takeaways

Many well-defined visual baits

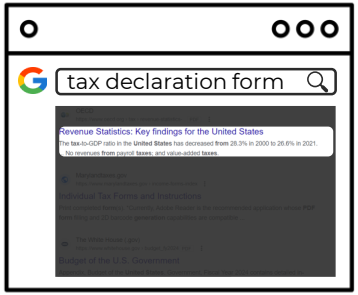
Clickbait PDFs in Web context



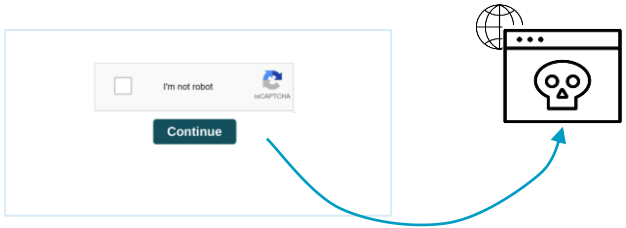
# Takeaways

Many well-defined visual baits

Distributed via SEO attacks



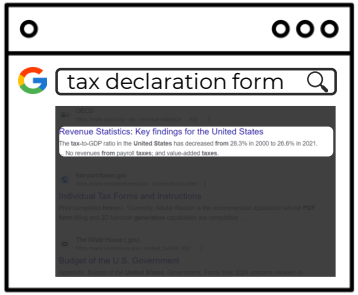

Clickbait PDFs in Web context



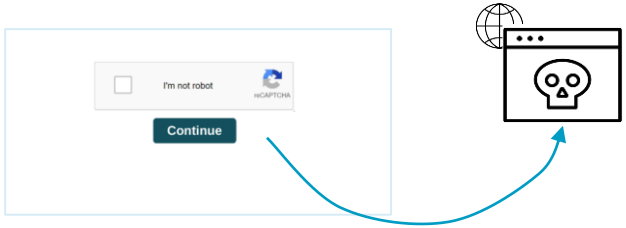
# Takeaways

Many well-defined visual baits

Distributed via SEO attacks




Clickbait PDFs in Web context



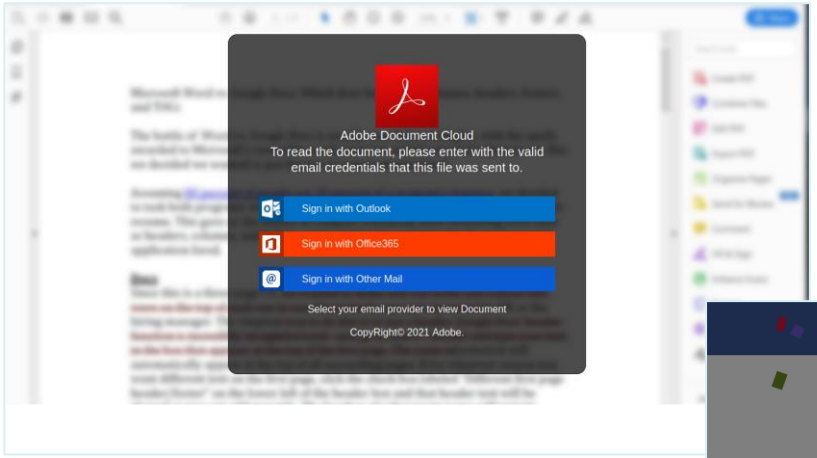

**Thank you! Questions?**

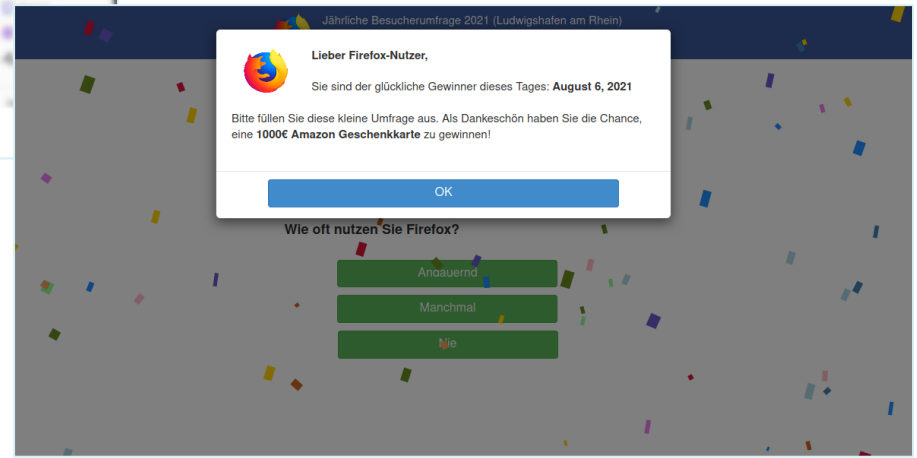
giada.stivala@cispa.de

g_stivala

# Backup Slides

# Example Attacks



# Email as Distribution Vector

Ideally: large dataset of emails available [4, 5]

Practical solution: VirusTotal tags (`email-spam`, `attachment`) and spamtrap

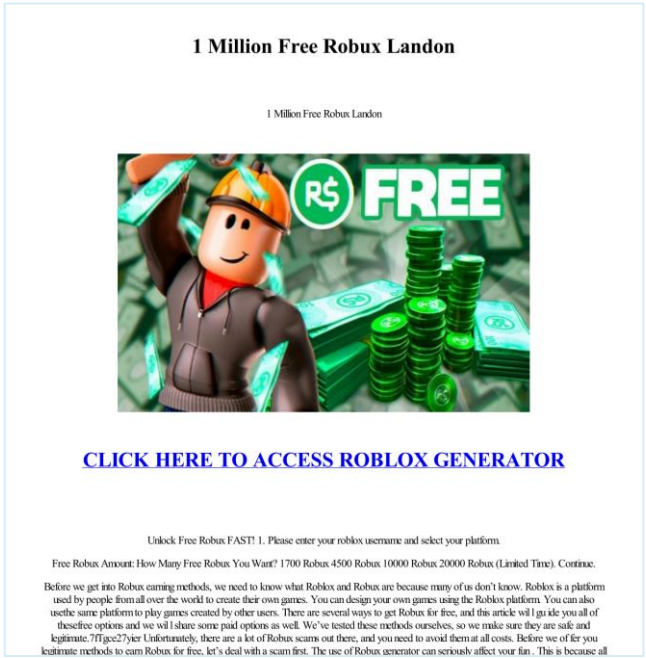
| Spamtrap | VirusTotal tags |
| --- | --- |
| • 68 hits for 19 files<br>• Seven clusters | • 65 files in six clusters<br>• No match for `email-spam` |

➢ Three largest clusters not in matches

➢ Distribution vector, but on a much lower scale

[4] Simoiu, Camelia, et al. "Who is targeted by email-based phishing and malware? measuring factors that differentiate risk." *Proceedings of the ACM Internet Measurement Conference*. 2020.
[5] https://www.mxmaildata.com/

# More Visual Baits



**1 Million Free Robux Landon**

1 Million Free Robux Landon



**CLICK HERE TO ACCESS ROBLOX GENERATOR**

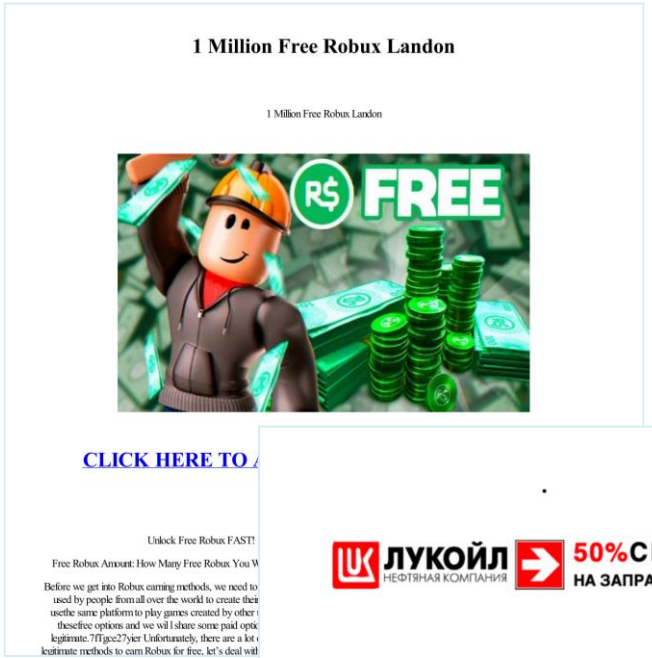
Unlock Free Robux FAST! 1. Please enter your roblox username and select your platform.

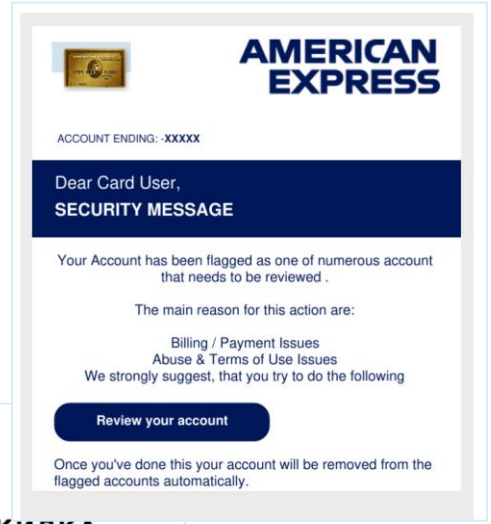
Free Robux Amount: How Many Free Robux You Want? 1700 Robux 4500 Robux 10000 Robux 20000 Robux (Limited Time). Continue.

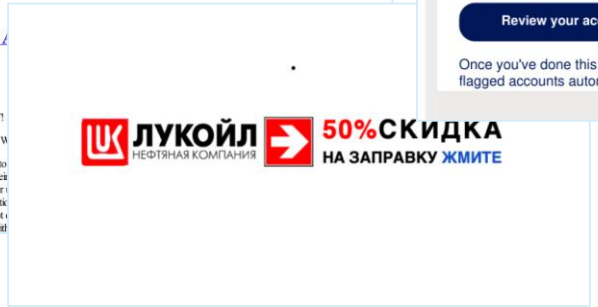
Before we get into Robux earning methods, we need to know what Roblox and Robux are because many of us don't know. Roblox is a platform used by people from all over the world to create their own games. You can design your own games using the Roblox platform. You can also usethe same platform to play games created by other users. There are several ways to get Robux for free, and this article will gu ide you all of thesefree options and we will share some paid options as well. We've tested these methods ourselves, so we make sure they are safe and legitimate.7fTgce27yier Unfortunately, there are a lot of Robux scams out there, and you need to avoid them at all costs. Before we of fer you legitimate methods to earn Robux for free, let's deal with a scam first. The use of Robux generator can seriously affect your fun . This is because all

# More Visual Baits

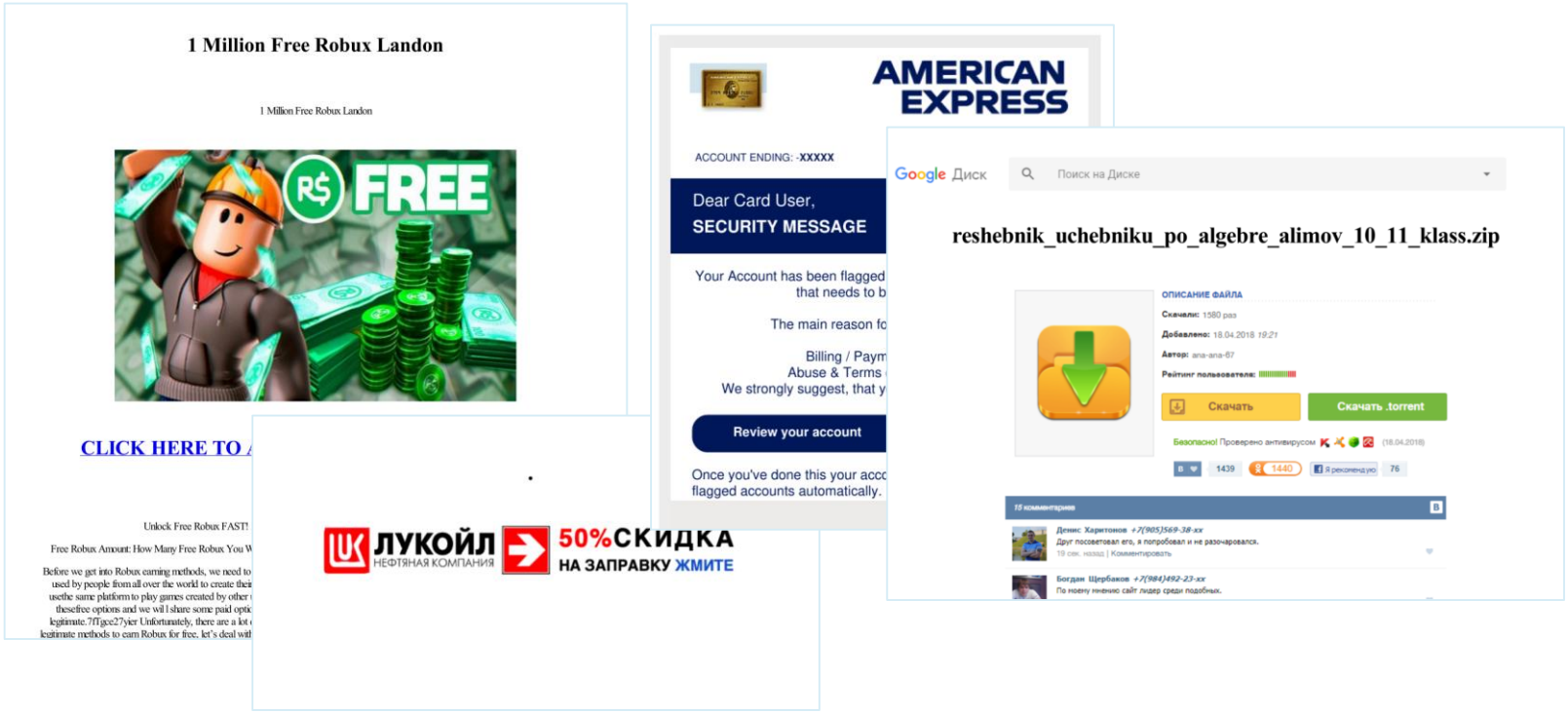
# More Visual Baits



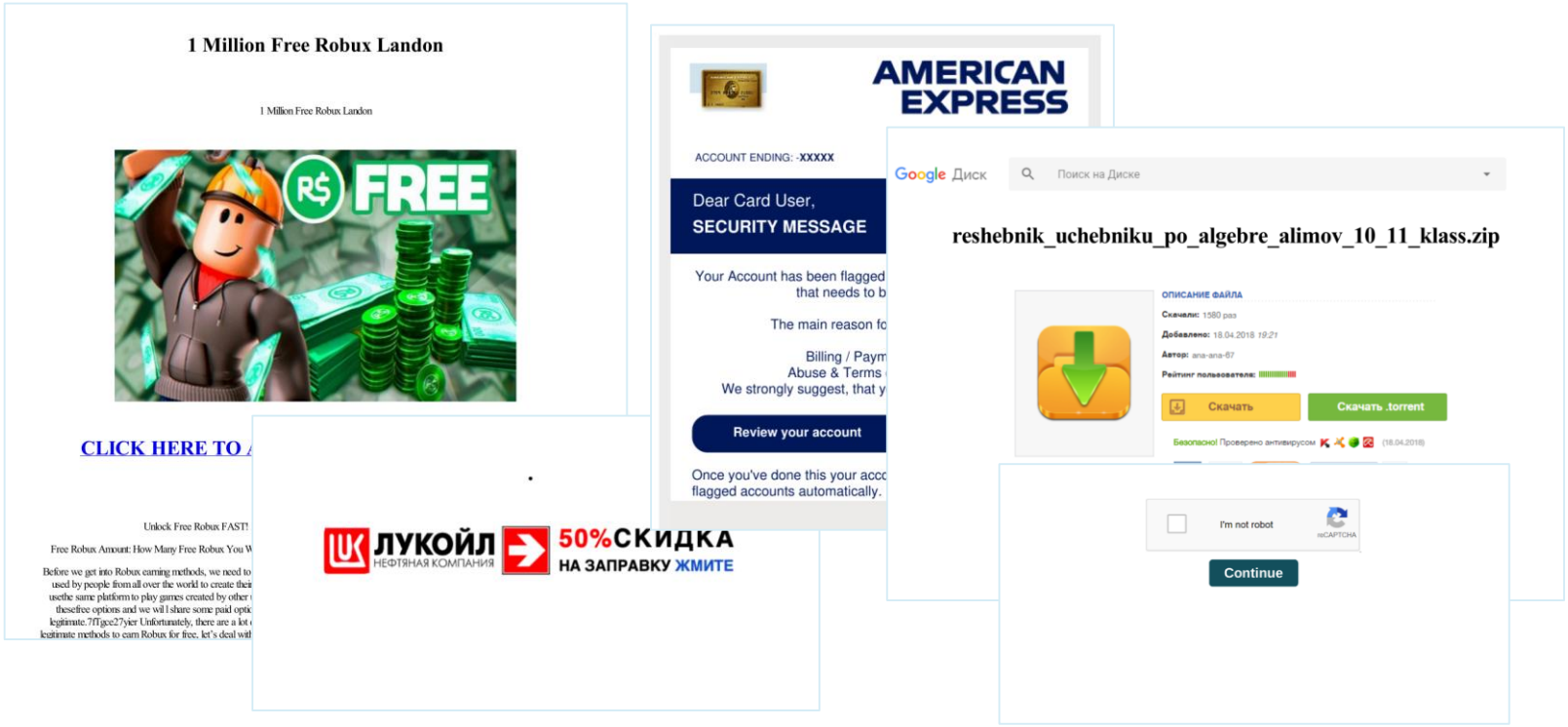



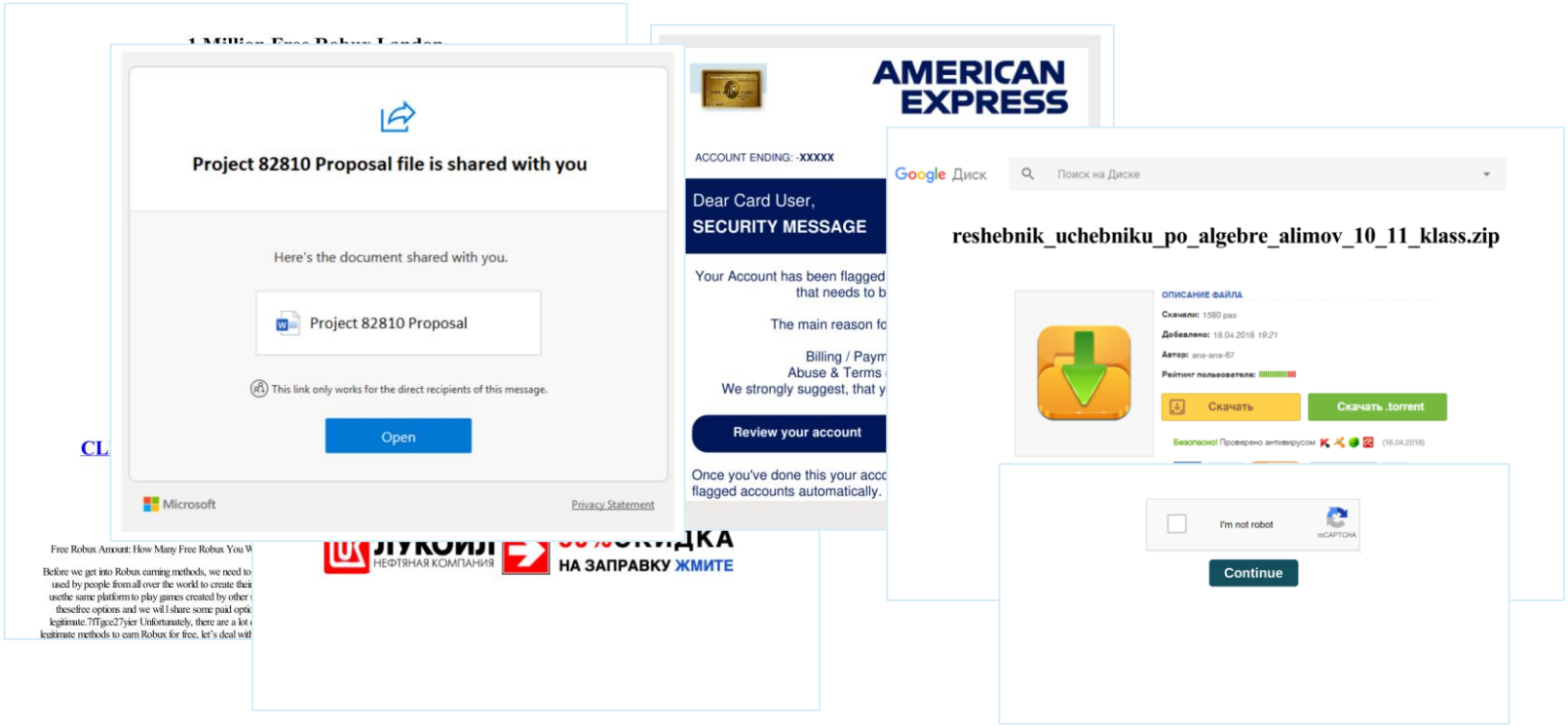
# More Visual Baits



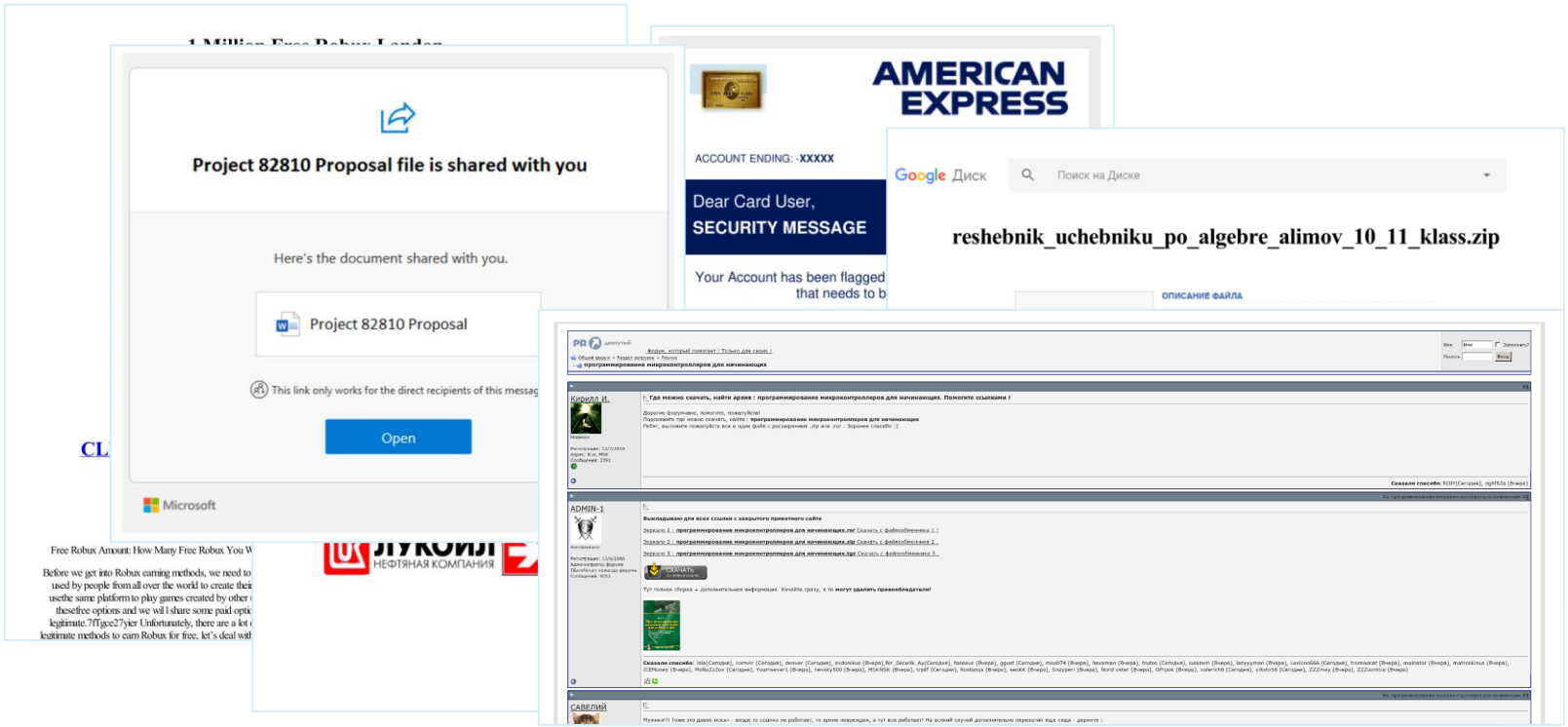
# More Visual Baits



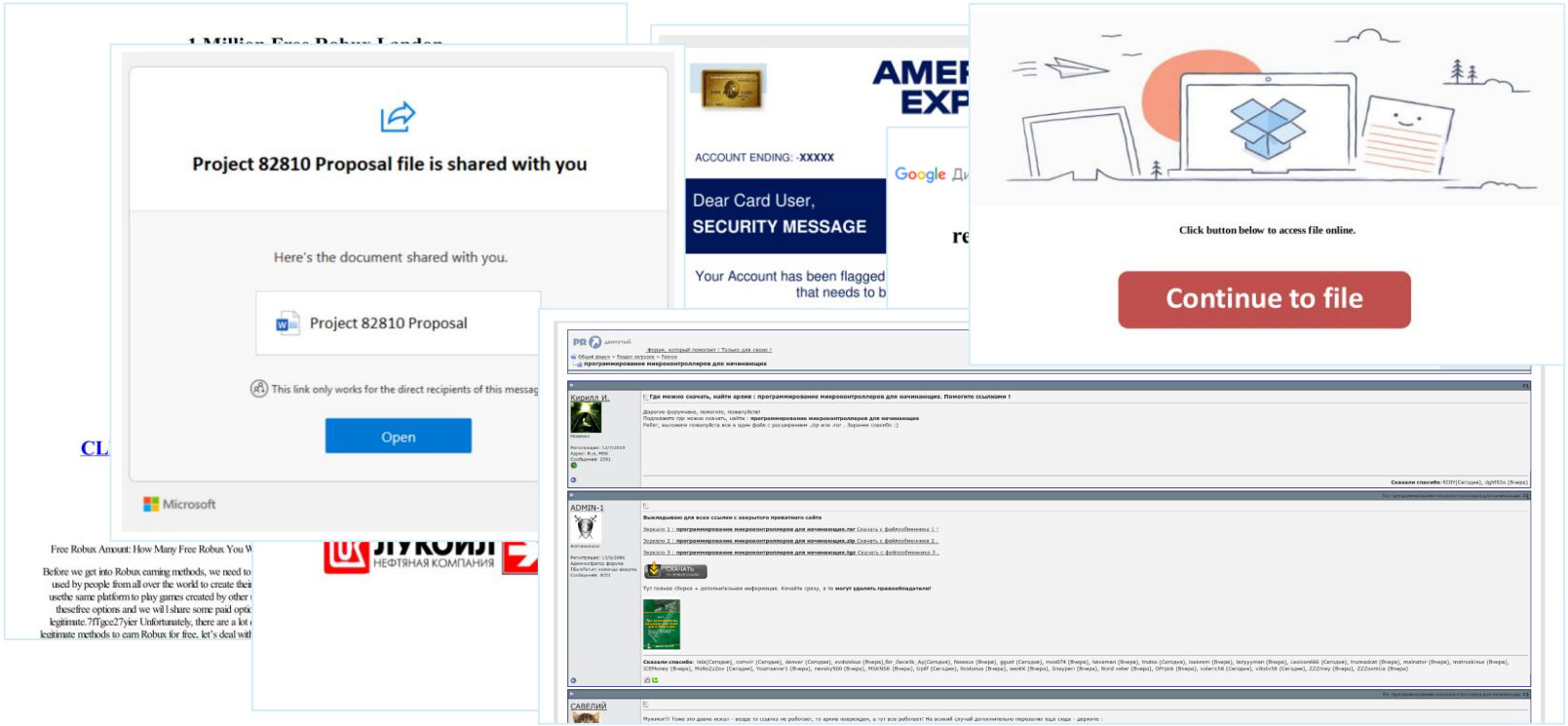
# More Visual Baits



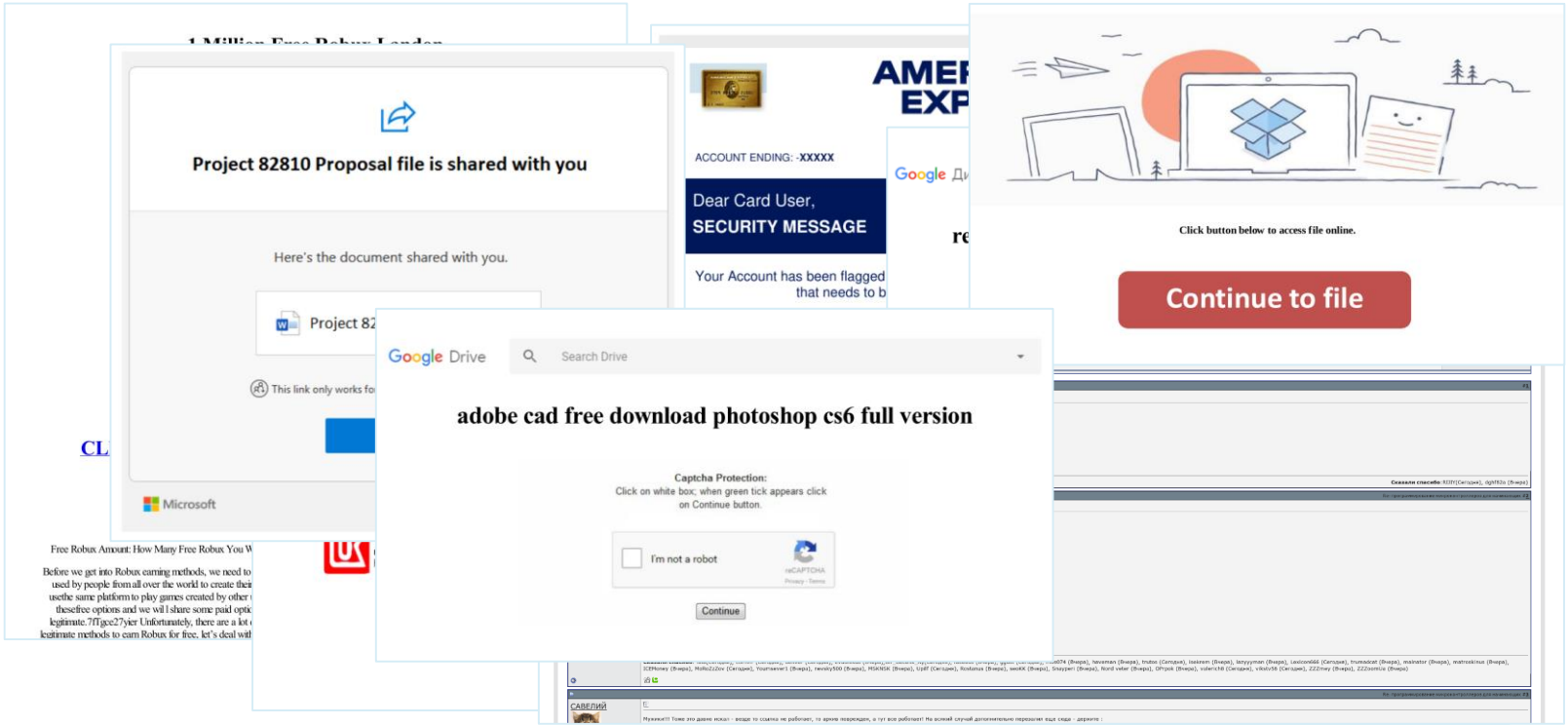
# More Visual Baits



# More Visual Baits

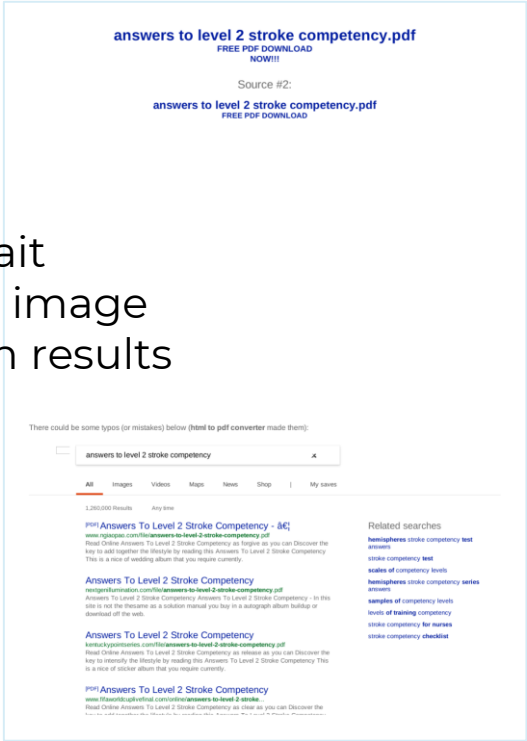


# More Visual Baits

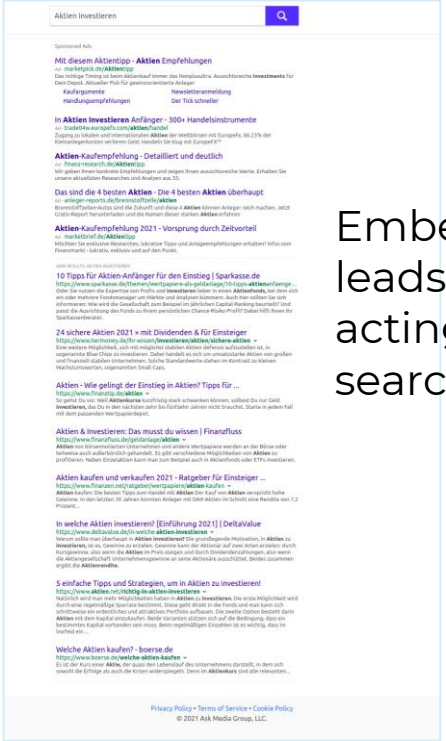


# Fake SE: Visual Bait and Malicious Webpage

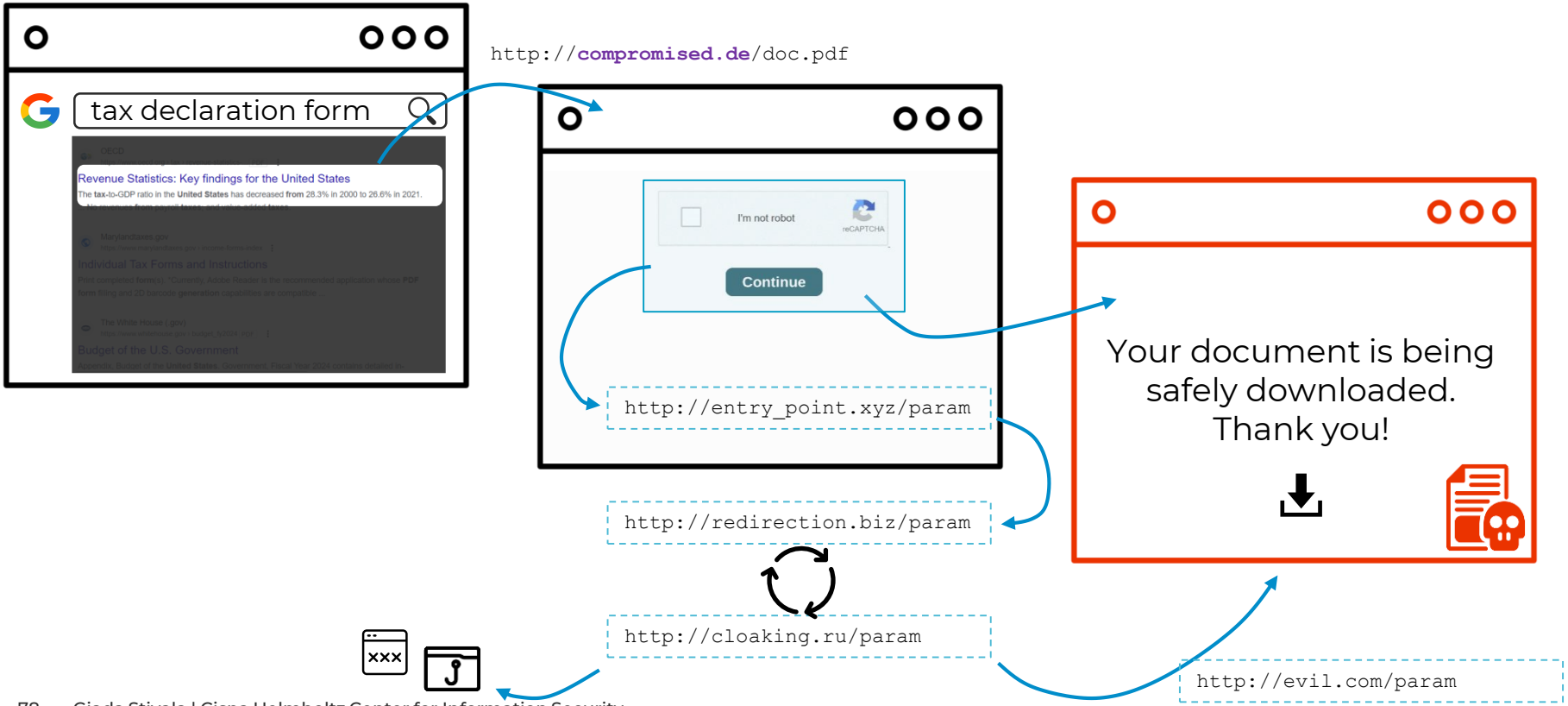
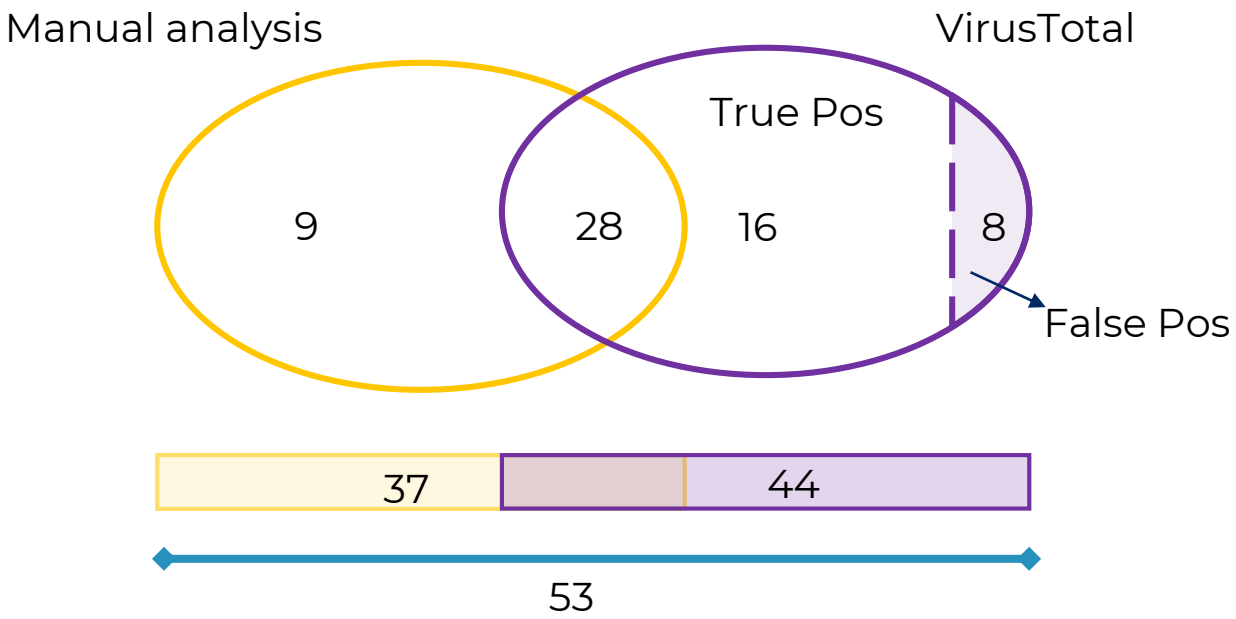
Visual bait embeds image of search results




Embedded link leads to webpage acting as a search engine



# Clickbait PDF Attack Details



# URL Maliciousness: Detailed Results