Using modern NLP algorithms to extract patterns and insights from phishing landing webpages and help develop more efficient identification strategies.

April 2026 | Student: Julien Blanchard | Supervisor: Paul Barry

Introduction

- Current machine learning models offer strong detection capabilities against most phishing attempts, but little is known about the actual content of the phishing landing websites.
- This dissertation explores the textual content of over 10k recent phishing landing websites.
- A combination of traditional and modern Natural Language Processing (NLP) techniques is used.

Literature review

- Most of the research papers on phishing focus on binary classification, or on impact and prevention techniques.
- NLP techniques have evolved fast since 2017. New algorithms have emerged based on the transformers architecture.
- No analysis or dataset of phishing landing pages exists.
- A package has been created to allow for the creation of a phishing landing pages dataset.

Research questions

- What patterns and semantic insights can be drawn from analysing phishing landing pages?
- Can modern algorithms such as BERTopic help us extract meaningful topic clusters from these landing pages?
- Can the insights drawn from traditional and modern approaches help us develop our understanding of the phishing landscape, and improve on the detection of phishing emails?

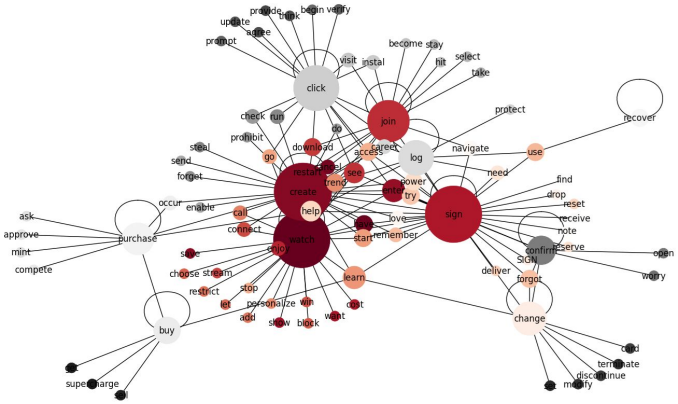
Methodology

- Acquiring a Dataset (Data sources, Data collection).
- Performing Pre-processing (Language filtering, HTML tags, Emojis, Anti-phishing avoidance techniques, tokenisation, stopwords)
- Analysing the dataset (Part-of-speech tags, Bi-grams, Named entities).
- Applying NLP algorithms (Using POS Tags to draw network graphs, BERTopic).

Findings & Results

- Phishing landing webpages mostly revolve around the following topics: Roblox, cryptocurrencies, streaming.
- Just like regular web analysts, threat actors adopt best SEO practices. This could indicate that threat actors expect some level of natural, non-email traffic.
- Threat actors tend to favour free hosting platforms.
- BERTopic allows for the extraction of semantically-coherent topics.

Relationship between verbs within the main dataset



Conclusion

- This first ever analysis of phishing landing webpages reveals patterns that provide a better understanding of current trends.
- This was made possible through a combination of traditional and modern NLP techniques.
- The insights highlighted through this analysis show patterns that could help further improve their classification models.