# Phishing URLs Analysis Report

# Prepared by: Wateen Al Mubarak

# Computer Science – Cybersecurity Track

# King Khalid University

# 2025

# Phishing URLs Analysis Report

## 1. Introduction

Phishing is one of the most widespread and impactful cyber threats, targeting online users by creating fake websites that mimic legitimate services. As defined by Whittaker et al. [[1]](#_6._References) , a phishing web page is "any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewers would only trust a true agent of the third party."

## This report aims to analyze the URL data used in phishing attacks by identifying the most targeted domains and the most used top-level domains (TLDs) in these attacks.

## 2. Methodology

**2.1 Data Source**

A dataset from Mendeley Data , a platform owned by Elsevier that provides open datasets for the academic and research community, was used in this analysis. The dataset includes URLs collected from reliable sources in the field of cybersecurity and is accurately classified as phishing sites , making it a trusted foundation for analyzing the most targeted phishing domains and TLDs[[2]](#_6._References).

**2.2 Tools Used**

## The analysis was conducted using Python and data analysis libraries, including:

## - Pandas: For loading and analyzing the data.

## - Matplotlib & Seaborn: For data visualization.

## - urllib.parse: For parsing URLs and extracting domains and TLDs.

**2.3 Analytical Steps**

- Extracting domain names (**Domains**) from phishing URLs and calculating the frequency of each domain.

- Extracting top-level domains (**TLDs**) from URLs and analyzing the most targeted TLDs.

- Creating **Graph** to Display the most targeted and TLDs used in phishing attacks.

- Validating the data to ensure accuracy.

## 3. Results & Analysis

**Results:**

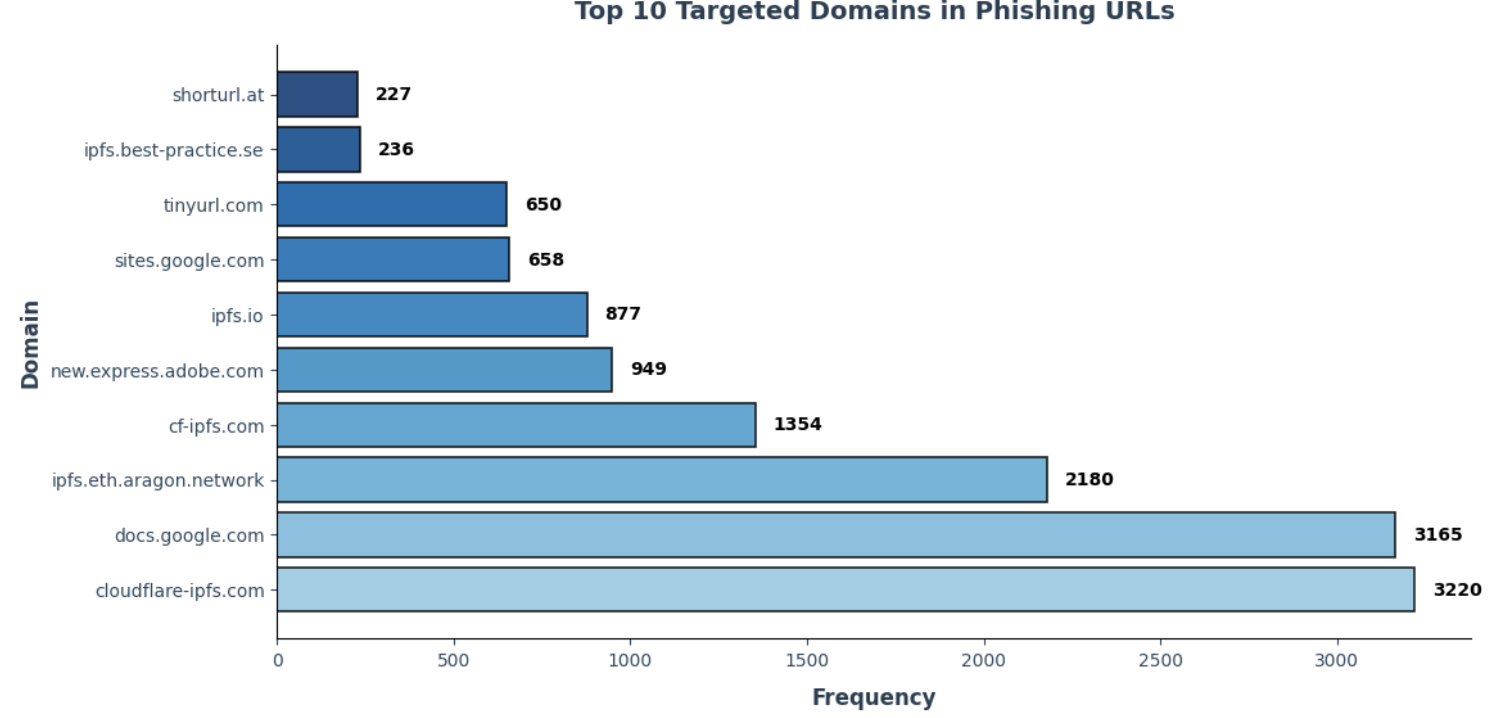
### 3.1 Most Targeted Domains in phishing

After analyzing phishing data through URLs, it was found that some domains are repeatedly targeted in these attacks. The top 10 most targeted domains are as follows:

A screenshot of a computer

AI-generated content may be incorrect.

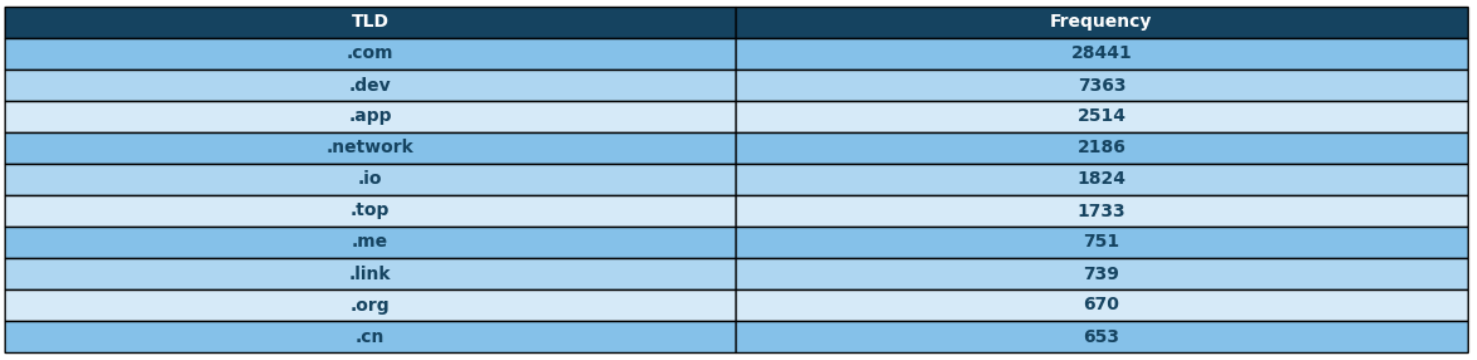
*[Table: Most Targeted Domains]*



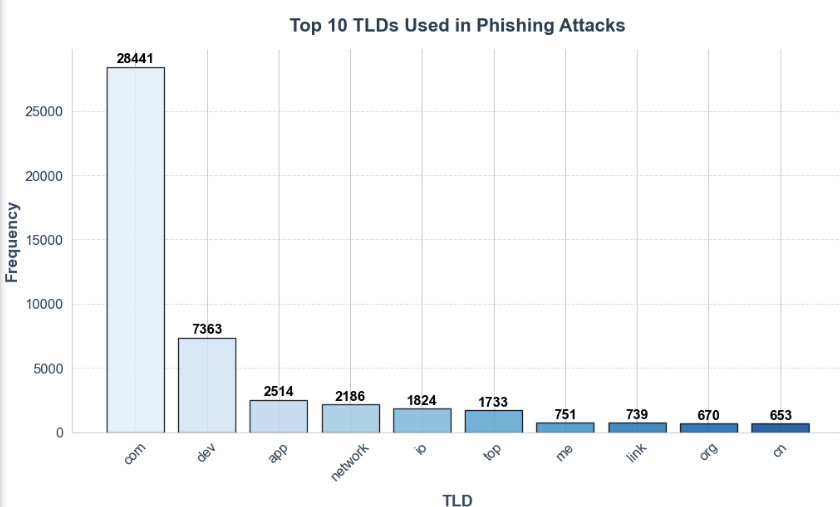
*[Bar Chart: Most Targeted Domains]*

### 3.2 Most Used TLDs in Phishing Attacks

After analyzing the phishing URLs, it was found that certain TLDs are more frequently used in phishing attacks. The 10 most commonly used TLDs are:



*[Table: Most Used TLDs]*



*[Bar Chart: Most Used TLDs]*

**Analysis:**

Phishing has become more sophisticated due to the use of artificial intelligence to create more realistic fake websites, a topic discussed in the DeepFish [[3]](#_6._References) model. This model is designed to generate phishing URLs using artificial intelligence. The model takes human-designed URLs as input and generates new phishing URLs as output. These AI-generated URLs are capable of easily bypassing mainstream phishing detection systems. Thus, we can predict that AI will be used in highly sophisticated and malicious attacks called **Offensive AI. “**Offensive AI refers to the malicious use of AI technologies to carry out cyber attacks[[4]](#_6._References)**”**.

## 4. Data Validation

## The extracted data was verified by comparing it with the initial values in the dataset. The extracted values match the frequency of domains and TLDs, which confirms the validity of the data.

## 5. Conclusions & Recommendations

**Conclusions:**

The analysis results indicated that **legitimate domains** such as **docs.google.com** and **cloudflare-ipfs.com** are highly targeted in phishing attacks.

The **".com"** **TLD** is the most commonly used in phishing URLs, followed by **".dev"** and **".app"**.

**Recommendations:**

Enhancing Security Auditing:

ISPs and web hosting companies should implement (URL scanning) mechanisms such as:

1. **Blacklist**: Comparing URLs against databases containing known malicious links .

2. **Sandboxing**: Isolating links in a controlled environment to analyze their behavior before granting access .

3. **Real-time URL Inspection**: Analyze links as users attempt to access them and block them if they are deemed suspicious. For example, PhishHaven [[3]](#_6._References) is a real-time AI-generated phishing URL detection system. It is designed to detect phishing URLs generated by a DeepPhish [[3]](#_6._References) model.

## 6. References

[1]. R. B. Basnet, A. H. Sung, and Q. Liu, "Learning to Detect Phishing URLs," *IJRET: International Journal of Research in Engineering and Technology*, vol. 3, no. 9, pp. 11-18, 2014. Available: [Learning to Detect Phishing URLs](https://www.researchgate.net/publication/273302231_LEARNING_TO_DETECT_PHISHING_URLS).

[2]. Mendeley Data. "Phishing Dataset." [Online]. Available: [Phishing URL dataset - Mendeley Data](https://data.mendeley.com/datasets/vfszbj9b36/1)

[3]. M. Sameen, K. Han, and S. O. Hwang, ["PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System,"](https://ieeexplore.ieee.org/document/9082616) *IEEE Access*, vol. 8, pp. 2991–3103, Apr. 2020, doi: 10.1109/ACCESS.2020.2991403.

[4] J. Hill, ["Understanding Offensive AI vs. Defensive AI in Cybersecurity,"](https://abnormalsecurity.com/blog/offensive-ai-defensive-ai) *Abnormal Security Blog*, Dec. 16, 2024. [Online]. Available: https://abnormalsecurity.com/blog/offensive-ai-defensive-ai. [Accessed: Mar. 21, 2025].