

GOOSE: GET OUT OF SITE

CHUNGNAM UNIVERSITY 장성근



CONTENTS

- INTRO
- BACKGROUND
- METHOD
- CONCLUSION



INTRO

가짜 웹사이트 공격, '네이버' 사칭 가장 많아..."한 달 간 695건 적발"

| 삼성·카카오·쿠팡 등 국민이 많이 사용하는 웹사이트 악용 공격 횡행

컴퓨팅 | 입력 :2024/03/26 09:41



이한열 기자 | □ □ 기자 페이지 구독 □ 기자의 다른기사 보기



[웨비나] 시각화된 OKR로 빠르게 전략을 실현하는 스마트 워크플로우를 구상해보세요!

가짜 웹사이트 공격이 최근 두드러지는 가운데 해커들이 네이버를 가장 많이 사칭해 개인정보를 탈취하고 있는 것으로 드러났다.

26일 로그프레스 발간한 3월 CTI(Cyber Threat Intelligence) 월간 리포트에 따르면

국민 대부분이 이용하는 서비스와 주요 기업 웹사이트를 이용한 해킹 공격이 최근 증가한 것으로 나타났다.



Home / Blog / *Wait, You're Not Google: The Threat of Typosquatting and What to Do About It*

Wait, You're Not Google: The Threat of Typosquatting and What to Do About It

Posted on November 23, 2016



INTRO

POWERMARC

[Platform >](#)

[Tools >](#)

[Pricing](#)

[MSP Program](#)

[Services >](#)

[Resources >](#)

[Company >](#)

[Sign In](#)

[Sign Up](#)

[Contact U](#)

[Home](#) > [Blog](#) > [What is Typosquatting in Cybersecurity](#)

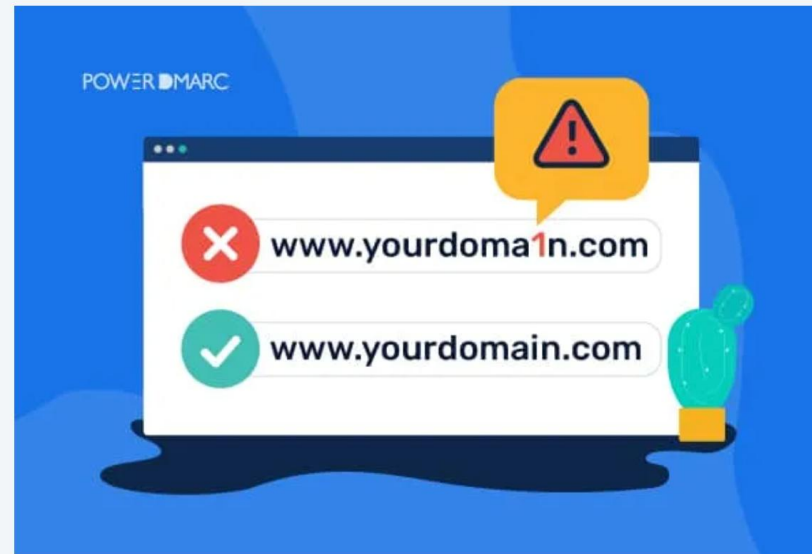
What is Typosquatting in Cybersecurity

Blog

Typosquatting is the use of misspelled domain names to deceive users into thinking that the site they're trying to access is legitimate. The result? Users are directed to sites with malware or phishing attempts, which can lead to identity theft and other serious problems.




by Ahona Rudra | July 26, 2022 | Reading Time: 6 min



GOAL

to establish a system that analyzes similar domains and detects phishing sites.

◆ 팀 구성

역할	주요 작업	담당자 수	담당자
Domain discovery and data collection	WHOIS retrieve, DNS <u>alayze</u>	3	성근, 이주, 가은
OSINT-based site investigation	<u>phising</u> site distinction	3	효주, 수인, 경빈
Automation and Visualization	dash-board <u>develope</u>	2	정은, 채연, 소진
Report writing and evaluation	Organizing results, writing reports, preparing presentations	1	성근 

BACKGROUND

similarweb

Products Customers Our Data Pricing Resources Contact sales Get started Login

December 2024 All traffic

Ranking Trending Last updated: January 1, 2025 Export Excel

Rank	Website	Category	Rank Change	Avg. Visit Duration	Pages / Visit	Bounce Rate
1	google.com	Computers Electronics and Technology > Search Engines	=	00:10:21	7.96	29.22%
2	youtube.com	Arts & Entertainment > Streaming & Online TV	=	00:20:31	11.03	23.45%
3	facebook.com	Computers Electronics and Technology > Social Media Networks	=	00:10:56	11.93	31.43%
4	instagram.com	Computers Electronics and Technology > Social Media Networks	=	00:08:40	11.66	35.62%
5	x.com	Computers Electronics and Technology > Social Media Networks	▲ 1	00:12:17	12.35	33.28%
6	whatsapp.com	Computers Electronics and Technology > Social Media Networks	▼ 1	00:14:13	7.07	54.19%
7	wikipedia.org	Reference Materials > Dictionaries and Encyclopedias	=	00:03:19	3.17	53.7%
8	reddit.com	Computers Electronics and Technology > Social	▲ 1	00:06:00	4.56	43.33%

SEMURSH

Features Pricing Resources Company App Center Enterprise

EN Log In Sign Up

Open .Trends

Explore the world's most visited websites
Open .Trends unwraps top websites across the web— just select an industry and location.

All Industries South Korea Find top websites

Top websites in South Korea (All Industries)

December 2024

	Domain	Visits ↑	Desktop share	Mobile share	MoM	YoY	Main Traffic Source		
🔗	🇻🇹 youtube.com	9.17B	7.97%	730.83M	92.03%	8.44B	↑104.14%	↑21.7%	Direct
🔗	🇸🇬 google.com	3.43B	20.34%	697.37M	79.66%	2.73B	↑63.19%	↑44.34%	Direct
🔗	🇳🇷 naver.com	2.11B	28.21%	594.02M	71.79%	1.51B	↑50.77%	↑44.53%	Direct
🔗	🇳🇵 newtoki466.com	540.96M	1.36%	7.37M	98.64%	533.59M	↑172.58%	—	Direct



BACKGROUND

DNS 분석

1) naver.com

- A 레코드 조회 (IP 주소 확인)

223.130.192.247

223.130.200.236

223.130.192.248

223.130.200.219

- MX 레코드 조회 (메일 서버 정보)

20 mx6.mail.naver.com.

20 mx5.mail.naver.com.

20 mx4.mail.naver.com.

메인 도메인

1) 포털 및 검색 엔진

- naver.com → 네이버 (한국 포털) n3ver.
- duckduckgo.com → 덕덕고 (프라이버시 보호 검색 엔진)
- msn.com → 마이크로소프트 뉴스 및 포털
- nate.com → 네이트 (한국 포털)

2) IT 및 클라우드 서비스

- live.com → 마이크로소프트 이메일 및 클라우드
- microsoftonline.com → 마이크로소프트 온라인 서비스
- microsoft.com → 마이크로소프트 공식 사이트
- office.com → 마이크로소프트 오피스 관련 서비스
- sharepoint.com → 마이크로소프트 협업 플랫폼
- canva.com → 디자인 및 그래픽 툴

3) 소셜 미디어 및 커뮤니티

- linkedin.com → 비즈니스 및 인맥 네트워킹
- pinterest.com → 이미지 및 아이디어 공유
- fandom.com → 팬 커뮤니티 및 위키
- t.me → 텔레그램 링크
- quora.com → 질문/답변 플랫폼
- facebook.com → 페이스북
- instagram.com → 인스타그램

METHOD: OSINT-based

1. First verification using google safe browsing API
2. By using only DNS lookup, if a phishing site actually exists, it can be classified as a safe site, and is determined by combining it with a typosquatting check.



Google safe browsing API

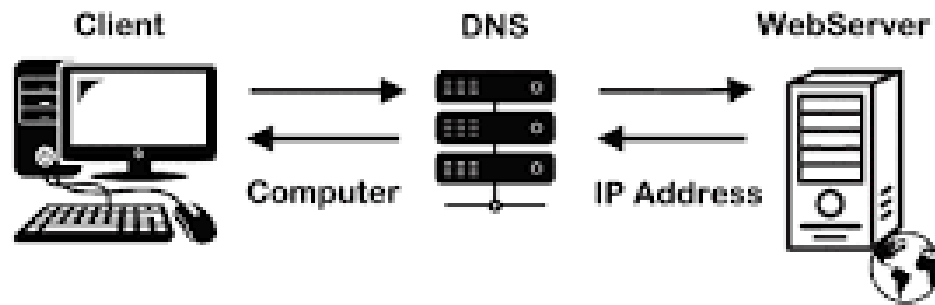


- When a user searches a specific URL, it is notified whether the URL is malicious.
- If the viewed URL is in the list of sites blocked by Google, it is determined to be a malicious site and a response is returned.
- If it is not on the blacklist, it responds as safe, so there are limits.



DNS search

- System for converting domain names to IP addresses
- Check whether the entered domain actually exists



Standards for verifying harmful sites

- Typosquatting detection criteria
- Levenshtein distance (analyzes similarity to official domains)
- TLD modulation
- Required libraries: `pip install python-Levenshtein tldextract`



Standards for verifying harmful sites

4. Risk Score Calculation (0 to 1)

- A risk score between 0 and 1 is assigned (higher scores indicate a higher likelihood of phishing or malicious domains).

- **Factors affecting the score:**

1. **Levenshtein Distance-based Similarity Score (50% weight)**

- Measures how similar the input domain is to an official domain.
- Formula:

$$\text{similarity score} = 1 - \left(\frac{\text{levenshtein distance}}{\text{max length}} \right)$$

- The smaller the Levenshtein distance, the lower the score.

2. **TLD Change Score (30% weight)**

- If the TLD is changed, **+0.3 points**, otherwise **0 points**.

3. **Altered Character Score (10% weight)**

- **If numbers are included → +0.1 points**
- **If hyphens (-) are included → +0.1 points**

4. **Homoglyph Attack Detection (20% weight)**

- Detects character-to-number substitution patterns:

- 0 ↔ 0

- 1 ↔ 1, l ↔ 1

- 5 ↔ S

- 8 ↔ B

- 9 ↔ g

- If detected, **+0.2 points**.

Final Score Calculation:

final score = (1 − similarity score) × 0.5 + TLD change score + altered chars score + homoglyph penalty

5. Site Classification Criteria

1. **Primary Verification:** Uses Google Safe Browsing API for an initial check.

- If the domain is in Google's blacklist, it is classified as malicious.
- If not, it is considered safe (even if it may still be a phishing site).

2. **Typosquatting Detection:**

- A DNS query alone may falsely classify a phishing site as safe if it exists.
- Combining DNS lookup with typosquatting detection improves accuracy.

DEMO

