Detecting Lookalike Domains

MSc Thesis
Introduction

11th August 2020

Patrick Frischknecht
Lookalike Domains

Domain names that look, read, or sound similar to other legitimate domains

facebook.com => faceboook.com
The Threats
Your money works better.

Sign Up for Free

Own a business? Open a business account
...lookalike domains are not just about phishing!

- Trademark abuse
- Social engineering
- Benign purpose: additional reach
Flightradar24 is a Swedish Internet-based service that shows real-time aircraft information on a map. It includes flight tracks, origins, and destinations, flight numbers, aircraft types, positions, altitudes, headings, and speeds. It can also show time-lapse replays of previous tracks and historical flight data by airline, aircraft, aircraft type, area, or airport. It aggregates data from multiple sources but, outside of the...
Detection Methods
Website features

URL-related

- Domain name [our focus]
- Domain history
- WHOIS information

Content-related

- HTML code
- Text
- Images [our focus: logos]
URL-related: Domain name squatting

- Attacker tries to impersonate another domain using a similar “lookalike” domain

  facebook.com => faceb00k.com

- Lookalikes reduce the chance a user recognizes the wrong page

Three squatting types are commonly used: homographs, typosquatting, and combosquatting
Homographs

Using homographs, characters look like others:

facebook.com => faceb00k.com
apple.com => äpple.com

Detection:

Generative: generate all homograph domains and check
Threshold on string edit distance
**Typosquatting**

Assuming typos like missing or mistyped characters on the keyboard:

- facebook.com => facebool.com
- apple.com => aple.com

Detection:

- Generative: generate all homograph domains and check
- Threshold on “fat finger distance”
Combosquatting

Assuming word recombinations:

facebook.com => facebook-site.com

apple.com => applestore.com

Detection:

Partial matching (?)
Difficulty: Partial matching

attack.com => contains “att”, really a lookalike?

bikeandride.com => contains “ikea”
Difficulty: Ambiguity

How to handle false positives and ambiguity?

Check the website content!
Logo Detection

Is template present in image?
The Idea
Project overview

- User provides at least:
  logo image and domain name

- Our tool does:
  1) generate a list of lookalike domains
  2) validate the list and capture a screenshot of each website
  3) search for similarities in the screenshot
  4) report the results for manual review
The Current State + Difficulties
Project overview

Our tool currently does:

1) **generate** a list of **lookalike domains**

2) **validate** the list and **capture a screenshot** of each website

3) **search for similarities** in the screenshot

4) **report** the results for manual review
Difficulty: Logo detection

Experienced poor matching performance on websites

=> looking for better algorithms
SIFT with sliding window

- detects and matches feature vectors in two images

(used in other papers)
The Next Steps
Recap

...lookalike domains are not just about phishing!
- Trademark abuse
- Social engineering
- Benign purpose: additional reach

Project overview
- User provides at least:
  - logo image and domain name
- Our tool does:
  1) generate a list of lookalike domains
  2) validate the list and capture a screenshot of each website
  3) search for similarities in the screenshot
  4) report the results for manual review

Website features
URL-related
- Domain name [our focus]
- Domain history
- WHOIS information

Content-related
- HTML code
- Text
- Images [our focus: logos]