SEVEN MONTHS' WORTH OF MISTAKES
A Longitudinal Study of Typosquatting Abuse

NDSS, February 9th 2015

Pieter Agten  Wouter Joosen  Frank Piessens  Nick Nikiforakis*

iMinds-DistriNet, KU Leuven, Belgium
* Stony Brook University, New York, USA
BACKGROUND
``Typosquatting is the act of purposefully registering a domain name that is a mistype of another domain name."
``Typosquatting is the act of purposefully registering a domain name that is a mistype of another domain name."

For example youuve.com instead of youtube.com
WHAT IS TYPOSQUATTING?

``Typosquatting is the act of purposefully registering a domain name that is a mistype of another domain name."

For example youtuve.com instead of youtube.com

Why?

- Anticipate typos in browser's address bar
- In order to steal hits from legitimate websites
Typosquatting has been known and studied for 15+ years
Typosquatting has been known and studied for 15+ years, but no *content-based* longitudinal studies have been published.
Typosquatting has been known and studied for 15+ years, but no content-based longitudinal studies have been published.

We studied the typosquatting domains of the 500 most popular websites over a period of 7 months.
Typosquatting has been known and studied for 15+ years, but no *content-based* longitudinal studies have been published.

We studied the typosquatting domains of the 500 most popular websites *over a period of 7 months*.

**Longitudinal** in addition to static observations:

- Are typosquatting domains changing hands?
- Does their content change over time?
OUR STUDY: APPROACH

1. Take Alexa top 500
1. Take Alexa top 500

2. Generate all potential typosquatting domains
1. Take Alexa top 500

2. Generate all potential typosquatting domains
   Based on models of [Wang-2006]
   - Missing dot www.example.com
   - Character-omission www.exmple.com
   - Character-permutation www.examlpe.com
   - Character-substitution www.ezample.com
   - Character-duplication www.exaample.com
1. Take Alexa top 500

2. Generate all potential typosquatting domains
   Based on models of [Wang-2006]
   - Missing dot: www.example.com
   - Character-omission: www.exmple.com
   - Character-permutation: www.examlpe.com
   - Character-substitution: www.ezample.com
   - Character-duplication: www.exaample.com

3. From April 1st, 2013 until October 31st, 2013:
   - 3.4 million visits
   - 900GB
   - Visits each domain daily using PhantomJS
   - Download WHOIS information weekly

4. Cluster and categorize collected pages
   - 8,102 clusters
   - Initial automatic clustering based on visual similarity
   - Second stage manual validation and content-based categorization

5. Analyze categorized clusters
1. Take Alexa top 500

2. Generate all potential typosquatting domains
   - 28,179 potential
   - 17,172 active

3. From April 1st, 2013 until October 31st, 2013:
   - Visit each domain daily using PhantomJS
   - Download WHOIS information weekly
1. Take Alexa top 500

2. Generate all potential typosquatting domains
   - 28,179 potential
   - 17,172 active

3. From April 1st, 2013 until October 31st, 2013:
   - Visit each domain daily using PhantomJS
   - Download WHOIS information weekly
   - 3.4 million visits
   - 900 GB
OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains
   - 28,179 potential
   - 17,172 active

3. From April 1st, 2013 until October 31st, 2013:
   - 3.4 million visits
   - 900 GB
     - Visit each domain daily using PhantomJS
     - Download WHOIS information weekly

4. Cluster and categorize collected pages
   - Initial automatic clustering based on visual similarity
   - Second-stage manual validation and content-based categorization
OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains
   - 28,179 potential
   - 17,172 active

3. From April 1st, 2013 until October 31st, 2013:
   - 3.4 million visits
   - 900 GB
     - Visit each domain daily using PhantomJS
     - Download WHOIS information weekly

4. Cluster and categorize collected pages
   - Initial automatic clustering based on visual similarity
   - Second-stage manual validation and content-based categorization
   - 8,102 clusters
OUR STUDY: APPROACH

1. Take Alexa top 500

2. Generate all potential typosquatting domains
   - 28,179 potential
   - 17,172 active

3. From April 1st, 2013 until October 31st, 2013:
   - 3.4 million visits
   - 900 GB
   - Visit each domain daily using PhantomJS
   - Download WHOIS information weekly

4. Cluster and categorize collected pages
   - 8,102 clusters
   - Initial automatic clustering based on visual similarity
   - Second-stage manual validation and content-based categorization

5. Analyze categorized clusters
RESULTS
95% of the Alexa Top 500 have at least 1 malicious typo domain

Most malicious registrations: adultfriendfinder.com (132)
95% of the Alexa Top 500 have at least 1 malicious typo domain
Most malicious registrations: adultfriendfinder.com (132)

31% of the Alexa Top 500 have at least 1 defensive registration
Most defensive registrations: huffingtonpost.com (57)
CATEGORICAL DISTRIBUTION OVER TIME

- Number of domains

- Ad parking (51%)
- No content (12%)
- Aff. abuse (9%)
- Hit stealing (5%)
- Error (4%)
- Other (18%)
CATEGORY DISTRIBUTION OVER TIME

Make Money Advertising Amazon Products
Earn up to 15% in referrals by advertising Amazon products.

Advertise products on your web page
People follow the links to Amazon
Earn up to 15% when they buy

Ad parking (51%)  Aff. abuse (9%)  Error (4%)
No content (12%)  Hit stealing (5%)  Other (18%)
Category distribution over time

- Ad parking (51%)
- No content (12%)
- Aff. abuse (9%)
- Hit stealing (5%)
- Error (4%)
- Other (18%)
NUMBER OF TYPOSQUATTING DOMAINS OVER TIME

Total known malicious typo domains
Active malicious typo domains

Number of domains

Apr  May  Jun  Jul  Aug  Sep  Oct  Nov

9000 10000 11000 12000 13000 14000
NUMBER OF CATEGORY TRANSITIONS OVER TIME

Week of 2013

Number of transitions
### LARGEST TYPOSQUATTING PAGE HOSTERS

Table: Number of malicious domains per autonomous system

<table>
<thead>
<tr>
<th>Network</th>
<th>Registered owner</th>
<th>Malicious domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>208.73.210.0 /23</td>
<td>Oversee.net</td>
<td>2,405 (18%)</td>
</tr>
<tr>
<td>199.59.243.96 /28</td>
<td>Bodis</td>
<td>1,741 (13%)</td>
</tr>
<tr>
<td>82.98.86.160 /27</td>
<td>Sedo</td>
<td>1,388 (10%)</td>
</tr>
<tr>
<td>69.43.161.128 /25</td>
<td>Castle Access</td>
<td>1,216 (9%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>6,750 (50%)</strong></td>
</tr>
</tbody>
</table>
CONCLUSION
CONCLUSION

We performed the first content-based longitudinal typosquatting study, which shows that:

- Typosquatting is still actively practiced today
- Many malicious, but only few defensive registrations
- Typo domains are changing hands and changing content
- Only 4 hosters account for 50% of the typosquatting domains

Many other results can be found in the paper

Our entire dataset is available at
https://distrinet.cs.kuleuven.be/software/typos15/