Scalable Detection of Promotional Website Defacements in Black Hat SEO Campaigns

Ronghai Yang*, Xianbo Wang*, Cheng Chi, Dawei Wang, Jiawei He, Shiming Pang, and Wing Cheong Lau

*First Co-Authors
Promotional Website Defacement

Goal: advertise illicit goods/services by cheating
Classic Black Hat SEO Tricks

- Keyword stuffing
- Link farming
- Massive modification
- Easily detectable

```html
<html>
  <title> News Article </title>
  <body>
    <p> content of the article </p>
    <a href="https://a.marksix.u">play mark six</a>
    <a href="https://b.marksix.u">gambling game</a>
    <a href="https://c.marksix.u">earn money fast</a>
    <a href="https://d.marksix.u">online mark six</a>
    ... 
  </body>
</html>
```
Existing Solutions

• Profile-based anomaly detection (content inconsistent with profile)
  • Davanzo et al. (2011), Liao et al. (2016)

• Change monitoring (focus on updated segments)
  • Delta (2013)

• Image-based detection
  • Meerkat (2015)
Modern Steathy Defacement
Conspiring Acts of Defacers and Illicit Content Seekers

Defacer attacks and controls a benign web server (benign.com)

Path A: Hint Consumer
Path B: Redirect

Defaced webpage in search results

Consumer tries to search for illicit content
Key Challenges and Solutions

- **Keyword Obfuscation**
- Tiny amount of insertions
DMoS
Our Cloud-Based Defacement Monitoring System

Service Subscription Page

Search Engine Dork Query (Section 3.1.1)

Website Crawlers (Section 3.1.2)

Web Page Acquisition

Web Page Pre-processing

Jargon Normalization

HTML Parser

M4RK SIX (六合彩) => MARK SIX (六合彩)

Tag-aware Machine-learning Models

Defacement Detection

https://yoursite.com
Submit for Monitoring
Jargon: Homophonic/Homomorphc

Similar pronunciation

Similar character shape

六合彩
MARK SIX

六合财
MARC SIX

六台彩
M4RK SIX

六台财
M4RC SIX

* English cases here are for illustration/translation of corresponding Chinese jargons only.
**JNA**

**Jargon Normalization Algorithm**

- **Language Model**
  - (title, “News Article”)
  - (img.alt, “play M4RC SIX …”)
  - (a, s_2')
  - (p, s_M')

- **Homophonic/Homomorphic Jargon Recovery**
  - “play M4RC SIX at …”

- **seed jargons**
  - jargon candidates
  - “M4RK”
  - “MARC”
  - “MARK”

- **recovered sentences**

- **non-fluent sentences**

- **tag-sentence pairs**
Web Page to Tag-Sentence Pairs

```html
<html>
  <title>News Article</title>
  <body>
    <p>
      content of the article
      <img alt="play M4RK SIX at m6m6.com" href="null">
      content of the article
    </p>
  </body>
</html>
```

Convert to tag-sentence pairs:

- `(title, "News Article")`
- `(p, "content of the article ...")`
- `(img.alt, “play M4RK SIX at ...”)`
- `(img.href, “null”)`
THAN: Tag-aware Hierarchical Attention Network

tag-sentence pair

Low-level Attention

High-level Attention

webpage embedding
Attention Weights Visualized
The Whole Pipeline Visualized

- Identify and normalize jargon on raw web page
  
- Assign weights and convert to embedding based on normalized page
Quantifying the Effectiveness of THAN & JNA

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAF</td>
<td>93.32%</td>
<td>9.76%</td>
</tr>
<tr>
<td>Saxe et al. [39]</td>
<td>82.53%</td>
<td>85.62%</td>
</tr>
<tr>
<td>BoW</td>
<td>86.56%</td>
<td>91.87%</td>
</tr>
<tr>
<td>HAN [46]</td>
<td>86.95%</td>
<td>93.86%</td>
</tr>
<tr>
<td>FastText [31]</td>
<td>81.47%</td>
<td>88.21%</td>
</tr>
<tr>
<td>BERT [24]</td>
<td>93.41%</td>
<td>97.64%</td>
</tr>
<tr>
<td>THAN</td>
<td>91.29%</td>
<td>96.89%</td>
</tr>
<tr>
<td>T-BERT</td>
<td>95.66%</td>
<td>98.76%</td>
</tr>
<tr>
<td>DMoS (JNA + THAN)</td>
<td>94.89%</td>
<td>97.40%</td>
</tr>
<tr>
<td>DMoS_v2 (JNA + T-BERT)</td>
<td>97.53%</td>
<td>99.37%</td>
</tr>
</tbody>
</table>

Tag-awareness + JNA help DMoS achieve production-ready performance

* THAN can be replaced with tag-aware BERT for a marginal improvement at a cost of 30x processing time
Five-Month Online Experiment

- Deploy DMoS to real customers and monitored over 38 million web pages on 7000+ websites
- Carefully designed manual verification for ground truth estimation
- 89% precision, 99% recall

Defacements can be found in prestigious websites
Compare Page Dimensions: 960.0x540.0

## Comparison with Commercial URL Checkers

<table>
<thead>
<tr>
<th>Schemes</th>
<th>No. of Legitimate Pages</th>
<th>No. of Defaced Pages</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tencent</td>
<td>235</td>
<td>190</td>
<td>50.00%</td>
<td>6.73%</td>
<td>11.86%</td>
</tr>
<tr>
<td>VirusTotal</td>
<td>235</td>
<td>190</td>
<td>57.14%</td>
<td>15.38%</td>
<td>24.24%</td>
</tr>
<tr>
<td>Baidu</td>
<td>235</td>
<td>190</td>
<td>63.41%</td>
<td>50.49%</td>
<td>56.22%</td>
</tr>
<tr>
<td>DMoS</td>
<td>235</td>
<td>190</td>
<td>98.37%</td>
<td>97.32%</td>
<td>97.85%</td>
</tr>
</tbody>
</table>

DMoS outperforms popular commercial products significantly.
Top-8 Jargons Transformed from “MARK SIX”

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
<th>Equivalence in English</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>六合彩</td>
<td>MARK SIX</td>
<td>620356</td>
</tr>
<tr>
<td>2</td>
<td>六和彩</td>
<td>MARC SIX</td>
<td>145759</td>
</tr>
<tr>
<td>3</td>
<td>6和彩</td>
<td>MARK 6</td>
<td>48991</td>
</tr>
<tr>
<td>4</td>
<td>六合采</td>
<td>MARK SYX</td>
<td>26509</td>
</tr>
<tr>
<td>5</td>
<td>六合财</td>
<td>MARK SIKS</td>
<td>4852</td>
</tr>
<tr>
<td>6</td>
<td>六台彩</td>
<td>M4RK SIX</td>
<td>3378</td>
</tr>
<tr>
<td>7</td>
<td>liuhecai</td>
<td>N.A.</td>
<td>3357</td>
</tr>
<tr>
<td>8</td>
<td>六盒彩</td>
<td>MARCK SIX</td>
<td>3181</td>
</tr>
</tbody>
</table>

Defacers are creative in terms of creating jargons
Frequency of Infected HTML Tags

- Search-ranking influential HTML tags (e.g., a, title, meta) are preferred.
- New tags for evasion, old tricks (e.g., font) are phasing out.
  * `<marquee>` is a **non-standard** HTML tag which cause text to **scroll** automatically.
Other Interesting Findings

• **Stealth**: over 60% defaced pages contain < 10 insertions.

• **Trend in mobile search**: Mobile version of search engines start to attract defacers’ attention.

• **Domain hijack in gov sites**: Many hijacking of expired Chinese government domains in 2019 due to regulation change (all government websites need to use .gov.cn)

• **Shallow depth**: Most of defacements were found in webpages with path depth 1 to 2. Fewer are in home page (depth 0) or deep paths
• We have deployed DMoS to test over 7000 real customers websites (38 million unique web pages) and found 11% contain defacements.

• Although the design of DMoS (mainly JNA) focuses on Chinese promotional defacements, we have demonstrated its applicability for English defacements with proof-of-concept experiments.

• We release a sample English defacement dataset at http://mobitec.ie.cuhk.edu.hk/DMoS/

• DMoS is robust with incomplete initial keyword list: it has generalization power thanks to word embedding and JNA.