Investigating Large-Scale Internet Abuse Through Web Page Classification

Matt Der
09.14.2015

UCSD CSE
Computer Science and Engineering
Internet abuse

email spam

phishing

click fraud

search poisoning

domaining?

malicious advertising
Internet abuse

email spam

phishing

click fraud

search poisoning

malicious advertising
Scaling attacks + defenses

- Internet scale attacks $\rightarrow$ Internet scale defenses
- Attackers automate + replicate their scams
  - Enables automated defenses
  - Commonalities can link them to same attacker

Tradeoff: stealth - vs - scale
Classifying Web crawls

- Empirical, data-driven security: Web crawls
- Classify large sets of Web pages into categories
  - Abusive? Type of abuse? Attacker behind abuse?
Role for machine learning

- Some domain expertise required, but a fully manual approach **does not scale**
Role for machine learning

- Some domain expertise required, but a fully manual approach **does not scale**
- **Thesis:** Automated machine learning tools to aid security efforts
Three Applications

1. Affiliate Program Identification

2. SEO Abuse on Counterfeit Luxury

3. Uses (and Abuses) of New TLDs
Spam-advertised storefronts
Affiliate programs

- Illegitimate business that sells counterfeit goods: millions of $$$ of revenue per month
- Manage Web sites that serve as online storefronts
- Enlist spammers to advertise their storefronts
- Contract out payment & fulfillment services
Tracking affiliate programs

**DO NOT:** detect + filter spam

**DO:** track + target affiliate programs

Identify affiliate programs by their storefront sites
Classification problem

100s of thousands of storefronts and dozens of affiliate programs.
Let’s try it!

1. EvaPharmacy
2. Pharmacy Express
3. RX-Promotions
4. Online Pharmacy
5. GlavMed
6. World Pharmacy
7. Greenline
8. RX Partners
9. RX Rev Share
10. Canadian Pharmacy
11. 33Drugs
12. ED Express
13. RXCash
14. MediTrust
15. MaxGentleman
16. PH Online
17. Dr. Maxman
18. Club-first
19. Stallion
20. MAXX Extend
21. Viagrow
22. HerbalGrowth
23. US HealthCare
24. ManXtenz
25. VigREX
26. Swiss Apotheke
27. Stud Extreme
28. Ultimate Pharmacy
29. Virility
Let’s try it!

1. EvaPharmacy
2. Pharmacy Express
3. RX-Promotions
4. Online Pharmacy
5. GlavMed
6. World Pharmacy
7. Greenline
8. RX Partners
9. RX Rev Share
10. Canadian Pharmacy
11. 33Drugs
12. ED Express
13. RXCash
14. MediTrust
15. Max Gentleman
16. PH Online
17. Dr. Maxman
18. Club-first
19. Stallion
20. MAXX Extend
21. Viagrow
22. HerbalGrowth
23. US HealthCare
24. ManXtenz
25. VigREX
26. Swiss Apotheke
27. Stud Extreme
28. Ultimate Pharmacy
29. Virility
Let’s try it!

1. EvaPharmacy
2. Pharmacy Express
3. RX-Promotions
4. Online Pharmacy
5. GlavMed
6. World Pharmacy
7. Greenline
8. RX Partners
9. RX Rev Share
10. Canadian Pharmacy
11. 33Drugs
12. ED Express
13. RXCash
14. MediTrust
15. MaxGentleman
16. PH Online
17. Dr. Maxman
18. Club-first
19. Stallion
20. MAXX Extend
21. Viagrow
22. HerbalGrowth
23. US HealthCare
24. ManXtenz
25. VigREX
26. Swiss Apotheke
27. Stud Extreme
28. Ultimate Pharmacy
29. Virility
<html><head>
<title>
Cvs pharmacy - buy generic pills, discount cheap tabs, order on line
</title>
<meta http-equiv="content-type" content="text/html; charset=UTF-8">
<link href="/themes/yellow/css/style_1.css" rel="stylesheet" type="text/css">
<script type="text/javascript" src="/themes/yellow/js/main.js"></script>
<script type="text/javascript" src="/themes/yellow/js/menus.js"></script>
<script type="text/javascript" src="/themes/card.js"></script>
<script>
var SessionType = "URL";
var SessionPrefix = "af6a4620cd8e28f167514a5c02017819";
var SessionName = "USID";
</script>
</head>
HTML would help...

GlavMed

<html><head>
<title>Medspharmacysupport</title>
<meta http-equiv="content-type" content="text/html; charset=UTF-8">
<link href="/themes/medic/css/style_1.css" rel="stylesheet" type="text/css">
<script type="text/javascript" src="/themes/medic/js/main.js"></script>
<script type="text/javascript" src="/themes/medic/js/menus.js"></script>
<script type="text/javascript" src="/themes/card.js"></script>

var SessionType = "URL";
var SessionPrefix = "eaac0da7c9d7a69d8d4fa7c147b94879";
var SessionName = "USID";
</head>
HTML would help...

GlavMed

```html
<html><head>
<title> Drugstores Canada </title>
<meta http-equiv="content-type" content="text/html; charset=UTF-8">
<link href="/themes/white/css/style_1.css" rel="stylesheet" type="text/css">
<!--[if lte IE 7]><link rel="stylesheet" type="text/css" href="/themes/white/css/for_IE.css" /><![endif]-->
<script type="text/javascript" src="/themes/white/js/main.js"></script>
<script type="text/javascript" src="/themes/white/js/menus.js"></script>
<script type="text/javascript" src="/themes/card.js"></script>
<var SessionType = "URL">
<var SessionPrefix = "232859d472f31abba47917ce2b07ed27">
<var SessionName = "USID">
</var>
</head>
```
Previously: heroic labeling effort

• Security experts labeled **178k storefronts**
  – Inspect HTML for signatures, encode with regular expressions
  – Painstaking manual process: ~200 person-hours

• **NOT** once-and-for-all effort
  – Storefronts change over time, regexs go stale
Machine learning approach

• HTML code: good source of features
  – Affiliate programs use in-house software engines to generate storefront templates
  – Storefront Web pages are replicated

• Instead of regex signatures, bag-of-words on HTML
  – Automatic feature extraction!
HTML bag-of-words

- Extract + count “words” from HTML tags + content to capture distinctive artifacts
- Encode HTML elements as words: 
  \((\text{tag}, \text{attribute}, \text{value})\) triplets

**HTML**

```html
<img src="example.jpg" alt="Example pic"
     height="50" width="100">
```

**Words**

```
img:src=example.jpg
img:alt=Examplepic
img:height=50
img:width=100
```
Data set

- 44 affiliate programs
- 178k labeled examples
- 34k features (PCA: 200)

A handful of affiliate programs dominate
Proof-of-concept experiment

- Are these HTML features enough to distinguish affiliate programs:
  - From one another?
  - From noise in the spam Web crawl?
    - 43k unlabeled Web pages → “other” class

- Favorable setting: plenty of labeled examples
Proof-of-concept experiment

45-way nearest neighbor classification

1. EvaPharmacy
2. Pharmacy Express
3. RX-Promotions
4. Online Pharmacy
5. GlavMed
6. World Pharmacy
7. Greenline
8. RX Partners
9. RX Rev Share
10. Canadian Pharmacy
11. 33Drugs
12. ED Express
13. RXCash
14. MediTrust
15. MaxGentleman
16. PH Online
17. Dr. Maxman
18. Club-first
19. Stallion
20. MAXX Extend
21. Viagrow
22. HerbalGrowth
23. US HealthCare
24. ManXtenz
25. VigREX
26. Swiss Apotheke
27. Stud Extreme
28. Ultimate Pharmacy
29. Virility
30. WatchShop
31. Affordable Acc.
32. One Replica
33. Ultimate Repl.
34. Prestige Repl.
35. Diamond Repl.
36. Luxury Repl.
37. Distinction Repl.
38. Exquisite Repl.
39. Swiss Repl.
40. EuroSoft
41. Royal Software
42. Authorized SW
43. Soft Sales
44. OEM Soft Store
45. “Other”
Proof-of-concept experiment

45-way nearest neighbor classification

Avg. accuracy = 99.9%
HTML distances are highly predictive

![Distance to nearest neighbor in Eva](chart)

- **Eva**
- **Other affiliates**
- **Unlabeled**

# of examples

Distance to nearest neighbor in Eva
HTML distances are highly predictive

Eva pts are extremely close to their nearest Eva neighbor ...
HTML distances are highly predictive

Eva pts are extremely close to their nearest Eva neighbor ... ... and non-Eva pts are far away from all of Eva
Mimicking an operational deployment

• Experts must label **some** storefronts, but how many?
Mimicking an operational deployment

- Experts must label *some* storefronts, but how many?
- **Learning from scratch:** only small initial seed of labeled storefronts
Classification in operational setting

![Graph showing balanced accuracy vs. number of training examples per class]

- Avg accuracy: 75%, 85%, 93%, 97%, 98%
- Balanced accuracy: 99.95%

- All (200 hours)
One-shot learning

Singly labeled storefront

Correctly classified storefronts

33drugs

RX-Promotions
Summary

Abuse

Machine Learning

HTML bag-of-words + 45-way nearest neighbor
Three Applications

1. Affiliate Program Identification

2. SEO Abuse on Counterfeit Luxury

3. Uses (and Abuses) of New TLDs
Counterfeit luxury stores

• Similar scams as before to monetize traffic
• Different advertising vector: search
Search engine optimization (SEO)

• E-commerce sites want traffic
• Search traffic dictated by search ranking
• Search Engine Optimization (SEO) are techniques to influence search rank
  – **White Hat SEO** (good practices)
    • sitemaps, friendly URLs, fast load times
  – **Black Hat SEO** (deceive search engines + users)
    • keyword stuffing, link farms, hidden text
• Attackers organize into **black hat SEO campaigns** to manipulate search rank + acquire targeted traffic
• **Poison search results** to lure users to counterfeit storefronts
• Key mechanism: cloaking
Attacker → SEO → Doorway
Attacker

SEO

HTTP GET /doorway

Doorway

Search Engine
Web Crawler
Attacker

SEO

HTTP GET /doorway

Doorway

PageRank

Louis Vuitton bags jewelry shoes

Search Engine Web Crawler
1. Attacker uses SEO techniques to manipulate search engine rankings.

2. HTTP GET request to the doorway page.

3. User searches for "cheap lv" on Google.
1. **SEO**
2. **Doorway**
3. **PageRank**
4. **Counterfeit Store**

**Search Engine Web Crawler**

**User**

**HTTP GET /doorway**

**Louis Vuitton bags jewelry shoes**

**“cheap lv”**
Ecosystem analysis

• Holistic understanding of counterfeit luxury
  – Beyond detecting search poisoning
  – Track + target multiple SEO campaigns (52)
  – Multiple victimized brands (16)
Classify stores by campaign

thousands of storefronts

dozens of campaigns

Classify stores by campaign
Classify stores by campaign

thousands of storefronts

dozens of campaigns

key

jsus

php?p=
HTML signatures

• Storefront Web pages less templated
  – Campaigns target multiple brands
• Also have associated doorway Web pages
• Storefronts + doorways must have *some* clues resulting from auto-generation ...
Bag-of-words + Logistic regression

- Same HTML bag-of-words
  - Many irrelevant features
  - Few relevant features
- Euclidean distance between HTML vectors is less predictive
- L1-regularized logistic regression
  - Learn distinctive features of each campaign automatically!
Logistic regression

**Linear model:** \( z = w \cdot x + b \)

- \( x \) : feature values
- \( w \) : feature weights

- Learn a model \((w, b)\) for each campaign
- L1-regularization: most weights are 0!
  - Non-zero weights are meaningful features
- Highly interpretable models + predictions
Bootstrap classification

0 initial supervision

labeled examples
1 train classifier

labeled examples
Bootstrap classification

1. train classifier

2. make predictions
Bootstrap classification

1. train classifier
2. make predictions
3. validate predictions

1. Issue queries on Google
2. Click on each search result
3. Crawl storefront pages

unlabeled examples

labeled examples

validate predictions

make predictions
Bootstrap classification

1. train classifier
2. make predictions
3. validate predictions

unlabeled examples

labeled examples
Bootstrap classification

1. train classifier
2. make predictions
3. validate predictions

Iterate!
Distinctive features: msvalidate

<table>
<thead>
<tr>
<th>s/d</th>
<th>weight</th>
<th>feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>37.087</td>
<td><code>img:src=includes/templates/uggsootsale/buttons/english/button_buy_now.gif</code></td>
</tr>
<tr>
<td>s</td>
<td>15.443</td>
<td><code>li:class=curselt_li</code></td>
</tr>
<tr>
<td>d</td>
<td>13.878</td>
<td><code>a:target=_parent</code></td>
</tr>
<tr>
<td>s</td>
<td>11.874</td>
<td><code>variable</code></td>
</tr>
<tr>
<td>s</td>
<td>8.498</td>
<td><code>div:class=get_to_cart</code></td>
</tr>
<tr>
<td>d</td>
<td>6.845</td>
<td><code>div:class=listingProductImage</code></td>
</tr>
<tr>
<td>s</td>
<td>6.033</td>
<td><code>img:width=160</code></td>
</tr>
<tr>
<td>d</td>
<td>5.486</td>
<td><code>speedy</code></td>
</tr>
<tr>
<td>s</td>
<td>5.466</td>
<td><code>speedy</code></td>
</tr>
<tr>
<td>s</td>
<td>5.443</td>
<td><code>img:title=BuyNowonsale</code></td>
</tr>
</tbody>
</table>
# Making predictions

<table>
<thead>
<tr>
<th>Domain</th>
<th>Most likely camp.</th>
<th>Next likeliest camp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>louisvuittonicon.com</td>
<td>msvalidate 0.999</td>
<td>biglove 0.004</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>s/d</th>
<th>score</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>6.742</td>
<td>a:target=_parent</td>
</tr>
<tr>
<td>s</td>
<td>3.833</td>
<td>img:src=includes/templates/uggsootsale/buttons/english/button_buy_now.gif</td>
</tr>
<tr>
<td>s</td>
<td>1.873</td>
<td>a:target=_parent</td>
</tr>
<tr>
<td>s</td>
<td>0.878</td>
<td>div:class=get_to_cart</td>
</tr>
<tr>
<td>s</td>
<td>0.878</td>
<td>vuitton</td>
</tr>
<tr>
<td>s</td>
<td>0.624</td>
<td>img:width=160</td>
</tr>
<tr>
<td>s</td>
<td>0.563</td>
<td>img:title=BuyNowonsale</td>
</tr>
<tr>
<td>d</td>
<td>0.405</td>
<td>img:width=160</td>
</tr>
<tr>
<td>s</td>
<td>0.303</td>
<td>img:alt=BuyNowonsale</td>
</tr>
<tr>
<td>d</td>
<td>0.185</td>
<td>vuitton</td>
</tr>
</tbody>
</table>
### Making predictions

<table>
<thead>
<tr>
<th>Domain</th>
<th>Most likely camp.</th>
<th>Next likeliest camp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>louisvuittonicon.com</td>
<td>msvalidate 0.999</td>
<td>biglove 0.004</td>
</tr>
</tbody>
</table>

```html
1. `<li class="alt"><a target="_parent" href="http://www.louisvuittonicon.com/artsy-mm-p-8.html"><img src="bzm_cache/d/d2306bd4d2739495a2c343df70438eac.image.160x160.jpg" alt="Artsy MM sale" title="Artsy MM on sale" width="160" /></a></li>
2. `<li class="alt"><a target="_parent" href="http://www.louisvuittonicon.com/capucines-mm-p-1727.html"><img src="bzm_cache/7/7798657cd4a73eb8c620a7439c2c94d7.image.160x160.jpg" alt="Capucines MM sale" title="Capucines MM on sale" width="160" /></a></li>
3. `<li class="alt"><a target="_parent" href="http://www.louisvuittonicon.com/chaine-wallet-mm-p-2164.html"><img src="bzm_cache/e/e8866690f68225a1fe7ff6d79a65ecbc.image.160x160.jpg" alt="Chaine Wallet sale" title="Chaine Wallet on sale" width="160" /></a></li>
4. `<li class="alt"><a target="_parent" href="http://www.louisvuittonicon.com/delightful-monogram-gm-p-29.html"><img src="bzm_cache/5/5a1e4c7f91208262195ea75083dcd2.image.160x160.jpg" alt="Delightful Monogram GM sale" title="Delightful Monogram GM on sale" width="160" /></a></li>
```
Making predictions

<table>
<thead>
<tr>
<th>Domain</th>
<th>Most likely camp</th>
<th>Next likeliest camp</th>
</tr>
</thead>
<tbody>
<tr>
<td>cheapestlouisvuittonoutlets.com</td>
<td>msvalidate 0.999</td>
<td>biglove 0.005</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>s/d</th>
<th>score</th>
<th>feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>7.644</td>
<td>a:target=_parent</td>
</tr>
<tr>
<td>s</td>
<td>2.239</td>
<td>img:src=includes/templates/uggsootsale/buttons/english/button_buy_now.gif</td>
</tr>
<tr>
<td>s</td>
<td>2.065</td>
<td>a:target=_parent</td>
</tr>
<tr>
<td>s</td>
<td>1.042</td>
<td>vuitton</td>
</tr>
<tr>
<td>s</td>
<td>0.513</td>
<td>div:class=get_to_cart</td>
</tr>
<tr>
<td>s</td>
<td>0.364</td>
<td>img:width=160</td>
</tr>
<tr>
<td>s</td>
<td>0.329</td>
<td>img:title=BuyNowonsale</td>
</tr>
<tr>
<td>d</td>
<td>0.237</td>
<td>img:width=160</td>
</tr>
<tr>
<td>d</td>
<td>0.219</td>
<td>vuitton</td>
</tr>
<tr>
<td>s</td>
<td>0.177</td>
<td>img:alt=BuyNowonsale</td>
</tr>
</tbody>
</table>
Making predictions

<table>
<thead>
<tr>
<th>Domain</th>
<th>Most likely camp</th>
<th>Next likeliest camp</th>
</tr>
</thead>
<tbody>
<tr>
<td>cheapestlouisvuittonoutlets.com</td>
<td>msvalidate 0.999</td>
<td>biglove 0.005</td>
</tr>
</tbody>
</table>

```html
<div class="get_to_cart">
  <a target="_parent" href="http://www.cheapestlouisvuittonoutlets.com/?products_id=38396&amp;action=buy_now">
    <img src="includes/templates/uggsootsale/buttons/english/button_buy_now.gif" alt="Buy Now sale" title="Buy Now on sale" width="60" height="15"/>
  </a>
</div>
```

s 0.329  img:alt=BuyNowonsale
d 0.237  img:width=160
d 0.219  vuitton
s 0.177  img:alt=BuyNowonsale
## Rounds of classification

<table>
<thead>
<tr>
<th>Round</th>
<th># Stores</th>
<th># Labeled</th>
<th># Campaigns</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4,432</td>
<td>491</td>
<td>28</td>
</tr>
<tr>
<td>1</td>
<td>4,432</td>
<td>497</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>4,432</td>
<td>557</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>5,690</td>
<td>570</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>7,484</td>
<td>828</td>
<td>52</td>
</tr>
</tbody>
</table>

(86.2% CV accuracy)

> have i told you recently... how much i love the classifier? it’s been working quite well!

---

**Classified:**
- 1.6 M poisoned search results (58%)
- 11 K doorway domains (42%)
- 828 storefront sites (11%)
4.4 User Traffic

For a small number of stores, we were also able to collect user traffic data that directly measures an SEO campaign's effectiveness at attracting traffic via PSRs.

4.3.2 Transactions

As surveyed in Section 2 and described in depth in our previous publication, we discovered the supplier site from the packing slip of two of our purchases. Upon visiting the site, we noticed it contains a scrolling list of fulfilled orders and a mechanism to lookup shipment data from a supplier partnering with one of the largest SEO campaigns peddling counterfeit Louis Vuitton.

5. RESULTS

Using this mechanism, we collected over 279K shipping records in a single ping records for valid order numbers in bulk (20 orders at a time). Each record contains a timestamp and information regarding current status of the order. We filter out orders that are returned by the customer. From country data listed in the records, we received 12 knock offs of low to medium quality, all shipped from China. From the bank identification numbers (BINs) in our transactions, we found that our purchases were processed through Amazon Payments Inc.

We further use our order data to study the relationship among storefronts and suppliers. We collected longitudinal data from a supplier partnering with one of the largest SEO campaigns peddling counterfeit Louis Vuitton. As a result, we are able to fetch visitor data for each store by visiting the publicly accessible default AWStats URL (e.g., http://<site>/awstats/).

The remainder of the areas represents active PSRs, where the filled areas are attributed to specific SEO campaigns and the unfilled area is the remainder unclassified.

In this section we use our crawler data to characterize the account for over 81% of orders.

If we combine these with the countries from Western Europe (41K), these regions and Australia, with 90k, 57K, and 39K orders, respectively. If we combine the orders placed through the supplier between July 02, 2013 and March 28, 2014. In summary, 256K orders successfully converted user traffic into transactions, we found that our purchases were processed through Amazon Payments Inc.

We further use our order data to study the relationship among storefronts and suppliers. We collected longitudinal data from a supplier partnering with one of the largest SEO campaigns peddling counterfeit Louis Vuitton. As a result, we are able to fetch visitor data for each store by visiting the publicly accessible default AWStats URL (e.g., http://<site>/awstats/).

The remainder of the areas represents active PSRs, where the filled areas are attributed to specific SEO campaigns and the unfilled area is the remainder unclassified.

Classification over time

---

Classification over time

---

Classification over time

---

Classification over time
Contributions

Overview

Search + Seizure:

• Data Collection

Brand Holders have the most incentive and are in the most natural position. underground markets (monetize using deceptive means e.g., sales of illicit goods).

• Method to acquire

Interventions intended to disrupt SEO Campaigns exist. In particular, Search Engines and Search Engine Optimization are designed to disrupt SEO and highlight online counterfeit luxury goods market through e-commerce sites posing as authentic storefronts.

• Quantitative and qualitative measurements of interventions at the granularity of campaigns

Attempt to understand and measure the effectiveness of the current state of defenses and the market as opposed to individual search results and storefronts.

• Passively measure effects of interventions.

Examine supply side using real purchases.

Approximate order volume using purchase pairs.

ML clustering { doorways, stores }.

Crawl Google Search Results for 16 verticals { LV, Moncler, Five Finger, Clarisonic, Abercrombie, Beats By Dre, Beats By Dre, Beats Solo, Solo HD, Studio, Pro At Best Price, Fr Beats By Dre Pro: Counterfeits are too dumb to purchase.}

Hot Beats By Dr Dre Cheap Solo HD Sr" www.falkpr.com/ "Recently our store can offer you Beats By Dre Cheap, l Dre Studio, Beats By Dre Pro prices in the 2013 years.

Summary

Machine Learning

HTML bag-of-words + Logistic regression

img:src=includes/templates/uggsootsale/...
Three Applications

1. Affiliate Program Identification

2. SEO Abuse on Counterfeit Luxury

3. Uses (and Abuses) of New TLDs
Domain names

- Example: insurance.com
  - Top-level domain (TLD): com
  - Second-level domain: insurance
- Domain Name System (DNS) maps human-readable strings to machine addresses
  - insurance.com → 70.42.23.110
- Domain names are unique & exclusive resource; first-come, first-served
- Desirable ones are scarce!
New gTLD Program

- total # of TLDs
- 1046
- 10/1/13
- 8/20/15
- singles
- camera
- clothing
- voyage
- party
- science
- mormon
- money
- lol
- cars
- soccer
- coupons
- vet
- dentist
- lawyer
- attorney
100 largest new TLDs

As of 02-03-15; excluding 2 largest
TLD debate

Stated goals: “enhancing competition and consumer choice, and enabling the benefits of innovation”

+ **PROS**
  - democratize Internet
  - promote innovation
  - serve communities

- **CONS**
  - defensive & speculative registrations
  - value to Internet at large?
  - potential for abuse
Domain name abuse

Cybersquatting
Buying a domain to profit from someone else’s trademark

Typosquatting
Gain unintentional direct traffic from typos
(e.g., gooogle.com, googel.com)

Homograph attack
Spoofing an existing domain name w/ an indistinguishable
ASCII representation (e.g., PayPal.com)

Domain parking
Monetizing an undeveloped domain by serving automatically spun ads
and/or reselling the domain at a profit
The Domain Name you’ve entered is not available. It has been taken down as a result of dispute resolution proceedings pursuant to the Uniform Rapid Suspension System (URS) or .us Rapid Suspension System (usRS) Procedure and Rules.

For more information relating to the URS, please visit: http://newgtlds.icann.org/en/applicants/urs

For more information relating to the usURS, please visit: http://www.neustar.us/policies

lockheedmartin.global
holidayinn.vegas
netflix.social
...
Google on domain parking

"Parked domains are placeholder sites with little unique content, so Google doesn’t typically include them in search results."

$497M / yr from typosquatters on the top 100,000 Web sites [Moore + Edelman, 2010]
Usage patterns in new TLDs

- Is usage consistent with claimed goals of TLD expansion?
- Crawl 4.1M domains in 480 TLDs
  - Expect lots of replication

PARKED

UNUSED

SUSPEND
How to classify 4.1M domains??
1) k-means cluster a subset
2) Label homogeneous clusters
3) Add labeled data to training set
4) Nearest neighbor classification
5) Validate + expand labeled set
5) Validate + expand labeled set

Iterate!
Cluster visualization tool

Label visually homogenous clusters \textit{in bulk}

<table>
<thead>
<tr>
<th>Cluster 36 [19059 pages (11.40%), 19059 unique (100.00%)] [tightness = 2.608525e-07] [\textit{top} [club (7.30%), xyz (7.27%), guru (6.47%)]]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top</strong></td>
</tr>
<tr>
<td><img src="image1" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image2" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image3" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image4" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image5" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image6" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image7" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image8" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image9" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image10" alt="Cluster 36" /></td>
</tr>
<tr>
<td><strong>Bottom</strong></td>
</tr>
<tr>
<td><img src="image11" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image12" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image13" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image14" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image15" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image16" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image17" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image18" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image19" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image20" alt="Cluster 36" /></td>
</tr>
<tr>
<td><strong>Random</strong></td>
</tr>
<tr>
<td><img src="image21" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image22" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image23" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image24" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image25" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image26" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image27" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image28" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image29" alt="Cluster 36" /></td>
</tr>
<tr>
<td><img src="image30" alt="Cluster 36" /></td>
</tr>
</tbody>
</table>
Classification results

<table>
<thead>
<tr>
<th>Class</th>
<th>Domains</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parked</td>
<td>1,140,193</td>
<td>27.3%</td>
</tr>
<tr>
<td>Unused</td>
<td>868,086</td>
<td>21.7%</td>
</tr>
<tr>
<td>Error</td>
<td>912,096</td>
<td>21.9%</td>
</tr>
<tr>
<td>Junk</td>
<td>470,664</td>
<td>11.3%</td>
</tr>
<tr>
<td>Suspended</td>
<td>11,007</td>
<td>0.3%</td>
</tr>
<tr>
<td>Content</td>
<td>687,505</td>
<td>16.5%</td>
</tr>
</tbody>
</table>

Registration intent

- **Primary**: unique Web presence  **14.6%**
- **Defensive**: protect existing presence  **39.7%**
- **Speculative**: profit (domain parking)  **45.6%**
TLD’s intended purpose

• Among content-rich Web pages: what *kind* of content?
• Are registrants embracing spirit of new TLDs? Are TLDs serving target communities?

**Document relevance:**

is `foo.bike` actually about bikes?
Learning about bikes

Does foo.bike contain words related to bikes?

1) Harness Google Search + PageRank to collect Web pages (definitely) about bikes

2) Learn a bike topic: distribution of words
Search + crawl Web pages

Search queries:

- bike
- site:wikipedia.org+bike
- site:about.com+bike
- site:ask.com+bike
- site:answers.com+bike
- related:http://en.wikipedia.org/wiki/bike
Wheels: The wheels on a hybrid bike are a true combination of what you find on road and mountain bikes. Wider, like a mountain bike for greater stability and durability, but then with a higher recommended air pressure that puts them in the same level as a road bike when it comes to inflation level. The higher air pressure allows them to go faster by reducing rolling resistance. Think about how a properly inflated basketball bounces compared to one that is even slightly flat. Same concept. The rims and spokes on hybrids are lighter too like a road bike, since the assumption is that you won't be doing the rougher off-road riding that mountain biking entails. Frame: Most hybrid bike frames are made of lightweight aluminum or steel (also called "cromoly"), due to the strength and durability the materials offers and their (relatively) low price.
Collocations

• Count **collocations** (co-occurrences) to estimate topic distribution

\[
P( w \mid \text{bike} ) = \frac{\text{count}(w, \text{bike})}{N}
\]

• Measures how closely associated word w is to “bike”
<table>
<thead>
<tr>
<th>audio</th>
<th>bike</th>
<th>clothing</th>
<th>christmas</th>
<th>email</th>
<th>photography</th>
<th>realtor</th>
</tr>
</thead>
<tbody>
<tr>
<td>audio</td>
<td>bike</td>
<td>clothing</td>
<td>christmas</td>
<td>email</td>
<td>photography</td>
<td>realtor</td>
</tr>
<tr>
<td>digital</td>
<td>mountain</td>
<td>womens</td>
<td>tree</td>
<td>address</td>
<td>art</td>
<td>estate</td>
</tr>
<tr>
<td>sound</td>
<td>bicycle</td>
<td>fashion</td>
<td>day</td>
<td>free</td>
<td>digital</td>
<td>realtor</td>
</tr>
<tr>
<td>format</td>
<td>bikes</td>
<td>wear</td>
<td>holiday</td>
<td>send</td>
<td>camera</td>
<td>home</td>
</tr>
<tr>
<td>video</td>
<td>size</td>
<td>mens</td>
<td>december</td>
<td>account</td>
<td>wedding</td>
<td>find</td>
</tr>
<tr>
<td>recording</td>
<td>riding</td>
<td>women</td>
<td>eve</td>
<td>mail</td>
<td>photographers</td>
<td>realtors</td>
</tr>
<tr>
<td>books</td>
<td>right</td>
<td>clothes</td>
<td>season</td>
<td>service</td>
<td>photographers</td>
<td>agent</td>
</tr>
<tr>
<td>music</td>
<td>road</td>
<td>traditional</td>
<td>trees</td>
<td>message</td>
<td>history</td>
<td>become</td>
</tr>
<tr>
<td>file</td>
<td>ride</td>
<td>dress</td>
<td>traditions</td>
<td>addresses</td>
<td>film</td>
<td>ask</td>
</tr>
<tr>
<td>files</td>
<td>shop</td>
<td>accessories</td>
<td>lights</td>
<td>accounts</td>
<td>technology</td>
<td>association</td>
</tr>
</tbody>
</table>
Relevance scoring

Document relevance $= P(w \mid d) \cdot P(w \mid TLD)$
Relevant or not

- Not relevant
- Relevant

audio

- clothing

threshold
For the fraction of Web pages containing content, intended purposes of TLDs are largely being embraced.
Summary

“Abuse”

PARKED

SUSPEND

UNUSED

Machine Learning

k-means clustering + Nearest neighbor classification

P( w | bike ): bike mountain bicycle :
Conclusions

• Symbiosis of **security + machine learning**
  – Modern Internet abuse demands scalable defenses
  – Web crawls produce rich, big data for ML to model
Automation + replication

- A lot of Web content is automatically generated
- Attackers must scale their scams
HTML signatures

**Automated** HTML bag-of-words feature extraction; No manual hunting + tuning!

- Highly templated: HTML distance is predictive

- Handful of characteristic features: L1-regularized logistic regression

```
img:src=includes/templates/uggsootsale/…
```
Bootstrapped classification

Overview

Search + Seizure: David Y. Wang

Quantitative and qualitative measurements of interventions at the granularity of campaigns

Attempt to understand and measure the effectiveness of the current state of defenses

We deconstruct the ecosystem rife with abuse.

Partnerships exist between SEO campaigns (acquire traffic surreptitiously) and ML clustering { doorways, stores

Nov 2013 – Now.

Brand Holders have the most incentive and are in the most natural position.

Third party companies (e.g., Google) label suspicious sites as penalized.

Current interventions can potentially disrupt SEO campaigns, but as implemented they do not have any tangible, lasting effects.

Ecosystem rife with abuse.

Insufficient coverage of campaigns.

Search Engines (e.g., Google) label suspicious sites as penalized.

Slow seizure rates leave months to monetize.

MarkMonitor

Third party companies (e.g., Google) label suspicious sites as penalized.

Five Finger

MarkMonitor

Third party companies (e.g., Google) label suspicious sites as penalized.

Keywords

Sept 01

Feb 01

Mar 01

Apr 01

Nov 01

Dec 01

Jan 02

Feb 02

Mar 02

Apr 02

May 02

Jun 02

Jul 02

Aug 02

Sept 02

Oct 02

Nov 02

Dec 02

Jan 03

Feb 03

Mar 03

Apr 03

May 03

Jun 03

Jul 03

Aug 03

Sept 03

Oct 03

Nov 03

Dec 03

Jan 04

Feb 04

Mar 04

Apr 04

May 04

Jun 04

Jul 04

Aug 04

Sept 04

Oct 04

Nov 04

Dec 04

Jan 05

Feb 05

Mar 05

Apr 05

May 05

Jun 05

Jul 05

Aug 05

Sept 05

Oct 05

Nov 05

Dec 05

Jan 06

Feb 06

Mar 06

Apr 06

May 06

Jun 06

Jul 06

Aug 06

Sept 06

Oct 06

Nov 06

Dec 06

Jan 07

Feb 07

Mar 07

Apr 07

May 07

Jun 07

Jul 07

Aug 07

Sept 07

Oct 07

Nov 07

Dec 07

Jan 08

Feb 08

Mar 08

Apr 08

May 08

Jun 08

Jul 08

Aug 08

Sept 08

Oct 08

Nov 08

Dec 08

Jan 09

Feb 09

Mar 09

Apr 09

May 09

Jun 09

Jul 09

Aug 09

Sept 09

Oct 09

Nov 09

Dec 09

Jan 10

Feb 10

Mar 10

Apr 10

May 10

Jun 10

Jul 10

Aug 10

Sept 10

Oct 10

Nov 10

Dec 10

Jan 11

Feb 11

Mar 11

Apr 11

May 11

Jun 11

Jul 11

Aug 11

Sept 11

Oct 11

Nov 11

Dec 11

Jan 12

Feb 12

Mar 12

Apr 12

May 12

Jun 12

Jul 12

Aug 12

Sept 12

Oct 12

Nov 12

Dec 12

Jan 13

Feb 13

Mar 13

Apr 13

May 13

Jun 13

Jul 13

Aug 13

Sept 13

Oct 13

Nov 13

Dec 13

Jan 14

Feb 14

Mar 14

Apr 14

May 14

Jun 14

Jul 14

Aug 14

Sept 14

Oct 14

Nov 14

Dec 14

Jan 15

Feb 15

Mar 15

Apr 15

May 15

Jun 15

Jul 15

Aug 15

Sept 15

Oct 15

Nov 15

Dec 15

Jan 16

Feb 16

Mar 16

Apr 16

May 16

Jun 16

Jul 16

Aug 16

Sept 16

Oct 16

Nov 16

Dec 16

Jan 17

Feb 17

Mar 17

Apr 17

May 17

Jun 17

Jul 17

Aug 17

Sept 17

Oct 17

Nov 17

Dec 17

Jan 18

Feb 18

Mar 18

Apr 18

May 18

Jun 18

Jul 18

Aug 18

Sept 18

Oct 18

Nov 18

Dec 18

Jan 19

Feb 19

Mar 19

Apr 19

May 19

Jun 19

Jul 19

Aug 19

Sept 19

Oct 19

Nov 19

Dec 19

Jan 20

Feb 20

Mar 20
Impact of classification

- Large-scale measurement
  - classified
  - 4.1M domains in 480 TLDs

- Inform defensive interventions
  - key
  - jsus
  - php?p=
Thank you!

- Co-chairs
- Committee members
- Co-authors
Thank you, everyone!

Questions?