Knock It Off: Profiling the Online Storefronts of Counterfeit Merchandise

Matt Der, Lawrence Saul, Stefan Savage, Geoff Voelker, University of California, San Diego

BACKGROUND

- **Counterfeits!**
- **Affiliate programs**
  - Illegitimate business that sells counterfeit goods
  - Manage Web sites that serve as online storefronts
  - Enlist spammers to advertise their storefronts via bulk email
  - Contract out payment & fulfillment services
  - Big business: millions of $ of revenue per month

**Challenges**

1. Acquiring training data
2. Web pages that render very differently are often linked to the same affiliate program

**Data filtering**

- Bottleneck is payment processing
- 95% of spam-advertised pharmaceutical, replica, and software products monetized through only a handful of merchant banks
- If affiliate program can’t process credit cards, business collapses

**Spam value chain** [Levchenko et al, 2011]

- Web pages
- Affiliate program

**Key insight**

100s of thousands of pages

dozens of affiliate programs

**When regular expressions fail**

- Matched
- Not matched

**Previous approach: Regular expressions**

1. \[\{([0-9]{3})\}\] ?650?[- ]?452[- ]6975
2. var\s+SessionType\s*=\s*"SessionName\s*"; var\s+SessionPrefix\s*=\s*"[0-9a-zA-Z\s-P]{32}\s*"
3. var\s+SessionName\s*=\s*"USID";
4. var\s+Session\s*=\s*"\[a-zA-Z0-9\s-\]+";
5. var\s+Session\s*=\s*"[a-zA-Z0-9\s-\]+";
6. /themes/white2/img/logo\.gif
7. You may contact us at \(\{([0-9]{3})\}\] 0-9\, 0-9\, 0-9\(4\), please, keep your order \(1\, 0\, 0\) every time you make a call.
8. \(\{([0-9]{3})\]\) 0-9\, 0-9\, 0-9\, 0-9\(4\)
9. var\s+SessionType\s*=\*\"URL\"; var\s+SessionPrefix\s*=\*\"[0-9a-zA-Z\s-P]{32}\s*"
10. var\s+SessionName\s*=\*\"USID";

EXPERIMENTS

- **Features**
  - Network
    - IP address & ASN of A record
    - IP address & ASN of NS record
  - HTML
    - Bag-of-words

- **EvaPharmacy in 2D**
  - Distances are highly predictive
  - Average 1-nearest neighbor classification accuracy: 99.95%
  - Why so good? Affiliate programs use in-house software to generate storefront templates
  - Discovered & labeled 3,785 new storefronts

- **Nearest neighbor validation tool**

**How much labeling is enough?**

- **Training examples**
  - 33drugs
  - RX-Promo.
  - Avg accuracy (%)

- **Correct classifications**
  - 75 85 93 97 98

- **Balanced accuracy (%)**
  - 1 2 4 16