Spam Double-Funnel: Connecting Web Spammers with Advertisers

Yi-Min Wang, Ming Na - Microsoft Research
Yuan Niu, Hao Chen - University of California

WWW, 2007

HY 558
Mar. 2010
S. Loutou

Introduction

- Spammers use questionable Search Engine Optimization (SEO) techniques to promote their spam links into top research results
- Focusing on redirection path
- Propose:
  - A 5 layer, double-funnel model for describing end-to-end redirection spam
  - Methodology for analyzing the layers
  - Identify prominent domains on each layer using 2 sets of commercial keywords (targeting spammers ans advertisers)

Techniques used by spammers

- ✓ Search spammers use questionable SEO to promote their links
- Stuffing keywords
- Link farms
- Comment spamming
- Click-through cloaking
- ✓ Redirection spam: Web pages that redirect browsers to visit known spammer-controlled third party domains
- ✓ Syndication: used by redirection spam pages where they participate in pay-per-click programs

Strider Search Ranger System 1/4: Web Patrol with Search Monkeys

- Spammers use crawler-browser cloaking techniques
- To defend, Search Monkeys visit each web page with a full-fledged popular browser which executes all client-side scripts
- Newer cloaking techniques: Serve spam content only to users who click through search results
- Monkeys mimic the click-through

Strider Search Ranger System 2/4: Follow the money through Redirection Tracking

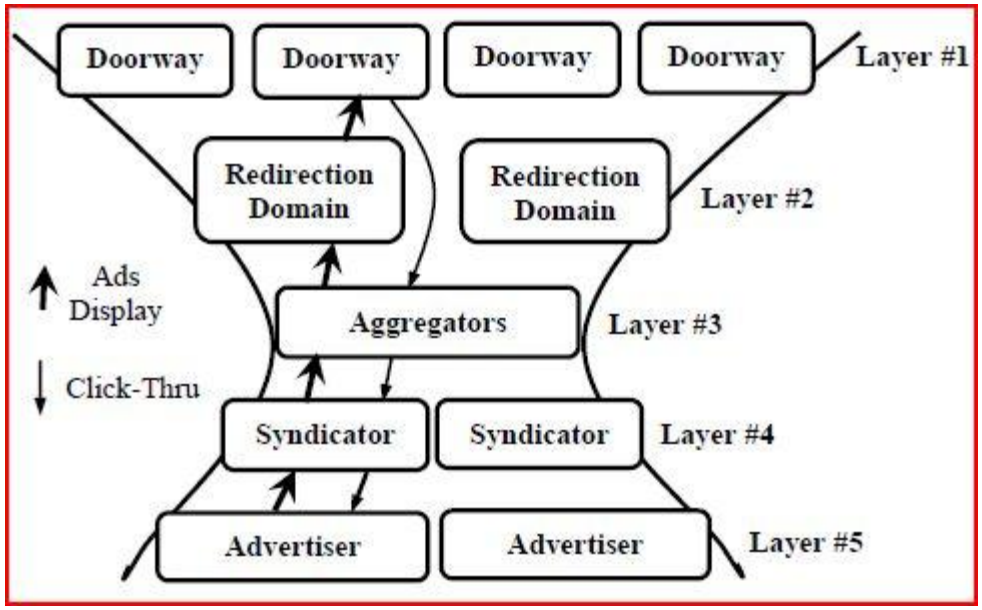
- Who is behind spam activities?
- Use the Strider URL Tracer to intercept browser redirection traffic at the network layer to record all redirection URLs
- Track both ads-fetching traffic and ads click-through traffic

Strider Search Ranger System 3/4: Similarity-based Grouping for Identifying Large-scale Spam

- Focus on monitoring search results of popular queries targeted by spammers to obtain a list of URLs with the high spam densities
- Analyze the similarity between the redirections from these pages to identify related pages
- “Backward propagation of distrust”

Spam Double-Funnel

- Advertising Syndication Business:
  Publishers – Syndicators – Advertisers
- To survive spam detection and blacklisting spammers:
  - Doorways pages
  - Redirection Domains



Spammer-targeted Keywords

- Redirection Spammers often use their targeted keywords as the anchor text of their spam links at public forums

➢ Collect keywords by extracting all the anchor text from a large number of spammed forums and ranking their frequencies. (June to August 2006)

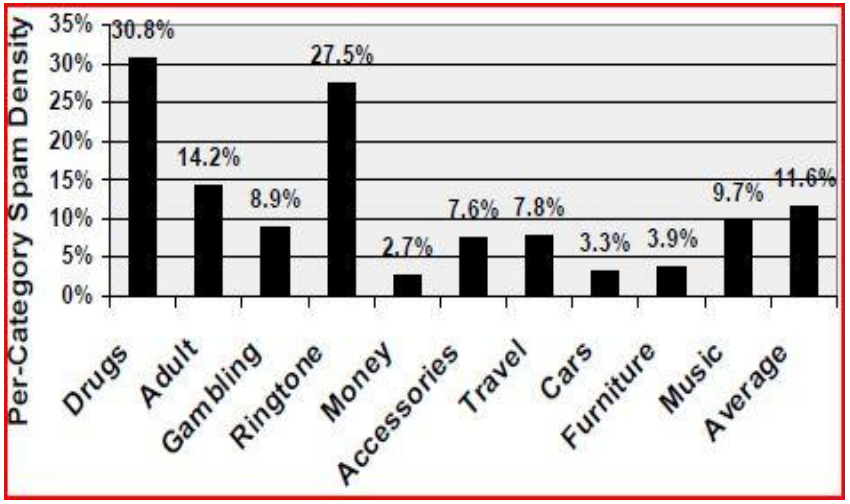
➢ Top-5: *phentermine, viagra, cialis, tramadol, xanax*

➢ 10 most prominent categories:
Drugs, Adult, Gambling, Ringtone, Money, Accessories, Travel, Cars, Music, Furniture

Redirection Spam Analysis : Spam Density Analysis

Spammer-targeted

- Submit 1000 keywords to the Search Ranger System -> Retrieve the top-50 results from the 3 major search engines



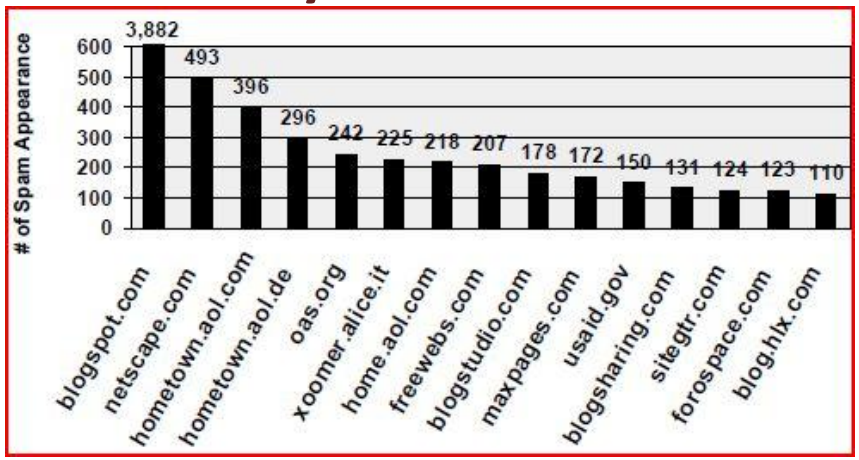
Advertiser-targeted

- Scan 95.753 unique URLs and identified 6.153 of them as spam
- This accounts for 5.8% of all top-50 appearances.
- Lower than 11.6 for the other benchmark. Why?

Redirection Spam Analysis : Double-Funnel Analysis: **Layer #1** 1/3

Spammer-targeted

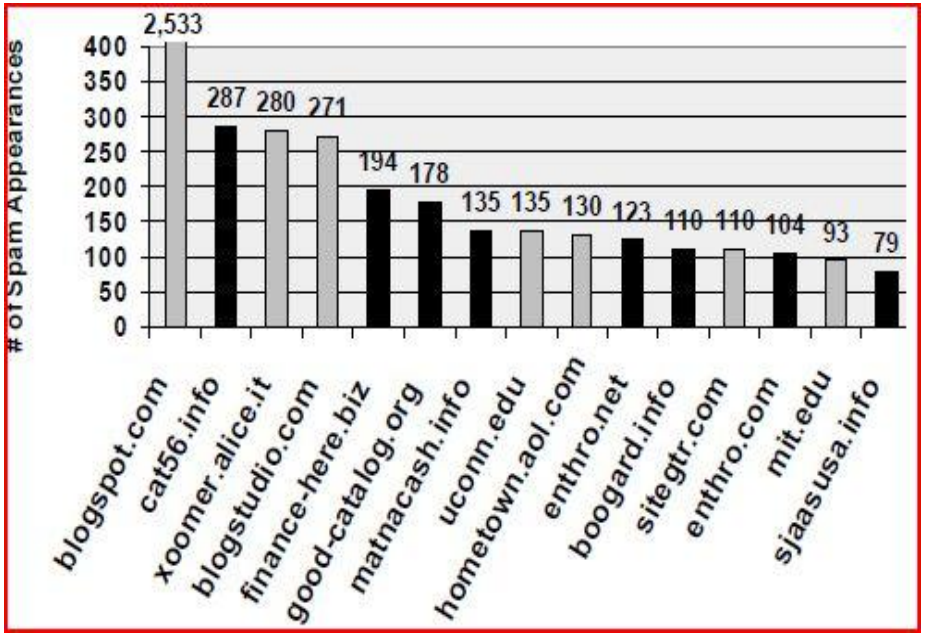
- **Doorway Domains**



- None of them only spam, hence cannot be blacklisted

Advertiser-targeted

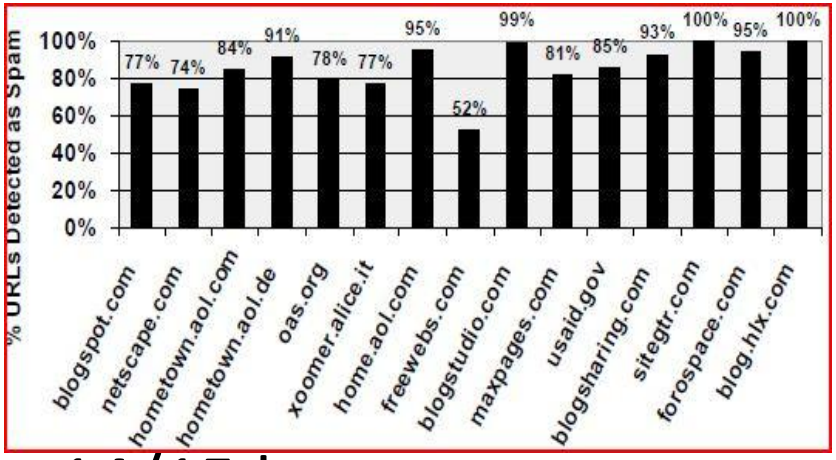
- **Doorway Domains**



- 4 .info domains

Redirection Spam Analysis : Double-Funnel Analysis: **Layer #1** 2/3

- Spam percentages of top doorway domains



- 14/15 have a percentage >74%

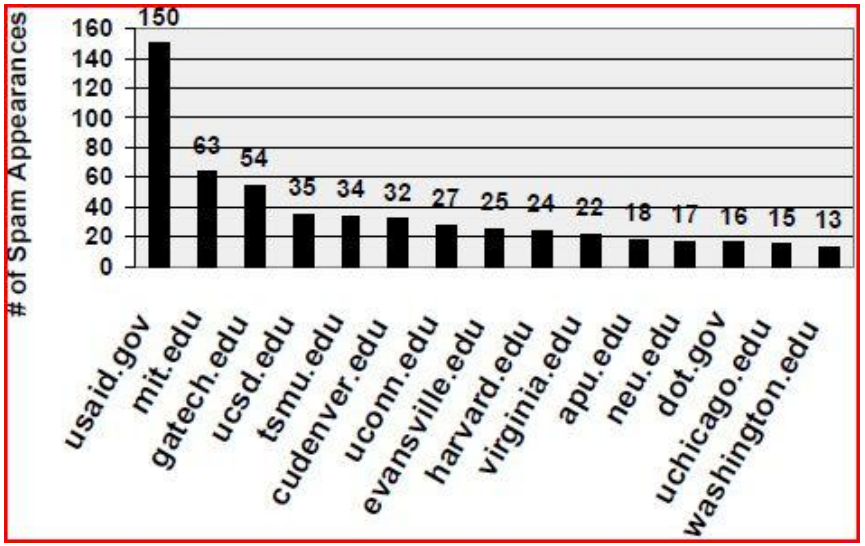
- Spam percentages for Top-Level Domains

- 68% spam percentage for .info (63% for the other benchmark) -> An order of magnitude higher than .com

TLD	.com	.org	.net	.biz	.info
Spam %	4.1%	11%	12%	53%	68%

Redirection Spam Analysis : Double-Funnel Analysis: **Layer #1** 3/3

- **(spammers-targeted)**
- **Spam in .gov and .edu**

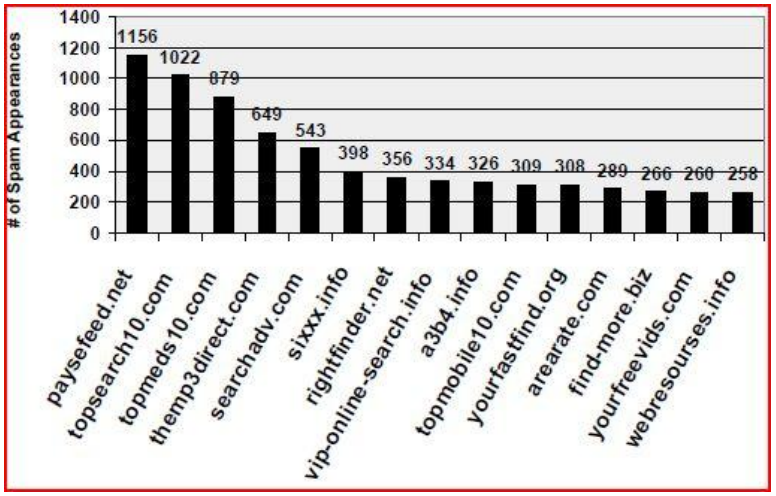


- URLs divided:
  - Universal redirectors
  - Unprotected upload areas
  - Home page-like directories

Redirection Spam Analysis : Double-Funnel Analysis: **Layer #2**

Spammers-targeted

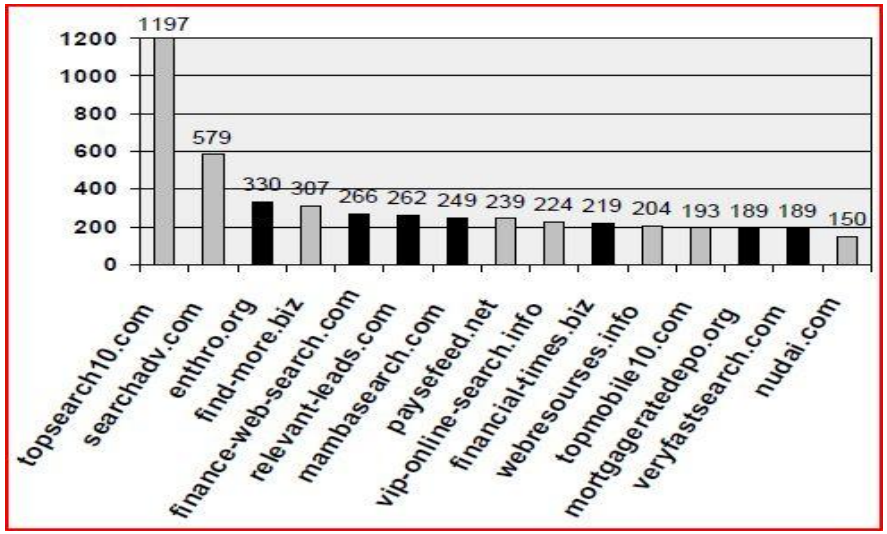
- **Redirection Domains**
- Top-15 redirection domains ranked by the number of spam doorway URLs that redirected to them



- 12 syndication-based, 2 pornographic ads, 1 commerce
- #1,#2,#3,#5,#10 resided on the same IP block
- 2 shared the same proxy registrant
- Major spammer groups own multiple top redirection domains

Advertisers-targeted

- Top-15 redirection domains: All syndication-based



- Drugs and adult spammers are replaced by money spammers

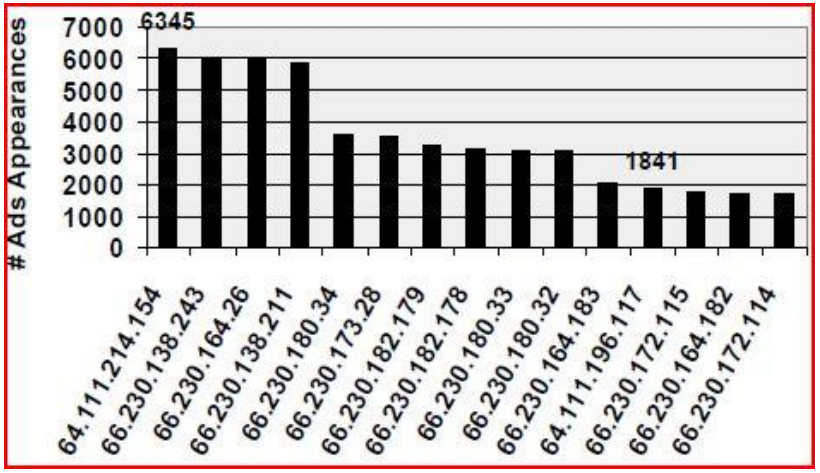
Redirection Spam Analysis : Double-Funnel Analysis: The bottom 3 layers

- Focus on redirection spam pages
- 2 types of analysis:
  - For layers 3 & 5, **page analysis**
    Extracting target advertiser URLs as well as their associated click-through URLs from ads-portal pages without visiting the ads
  - For layer 4, **click-through analysis**
    randomly selecting and visiting one ad from each portal page and recording all resulting redirection traffic (Necessary because the domain names of intermediate syndicators did not appear in the content of ads-portal pages)

Redirection Spam Analysis :
Double-Funnel Analysis: **Layer #3 (Aggregators)**

Spammers-targeted

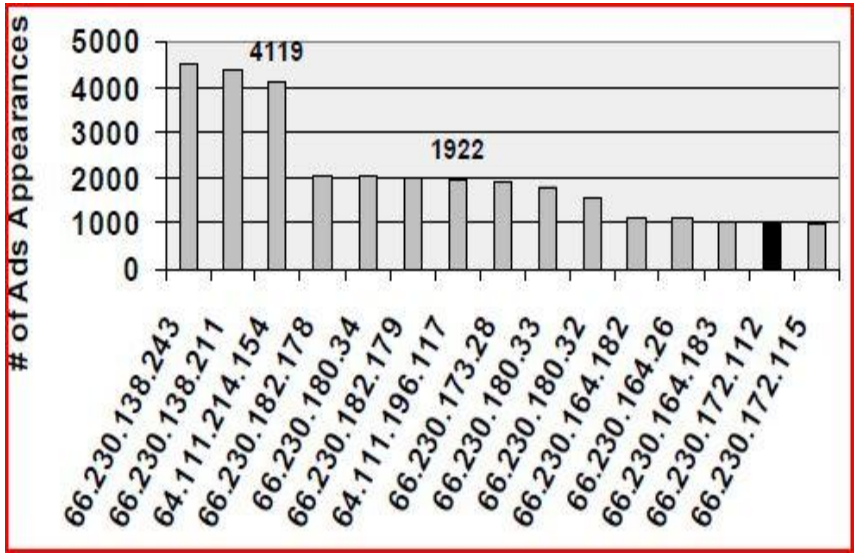
- Top-15 click through traffic receiver domains



- All of them are in 2 groups:
  66.230.128.0-*66.230*.191.255
  64.111.192.0-*64.111*.223.255
- The 2 IP blocks share the same *Whois* record

Advertisers-targeted

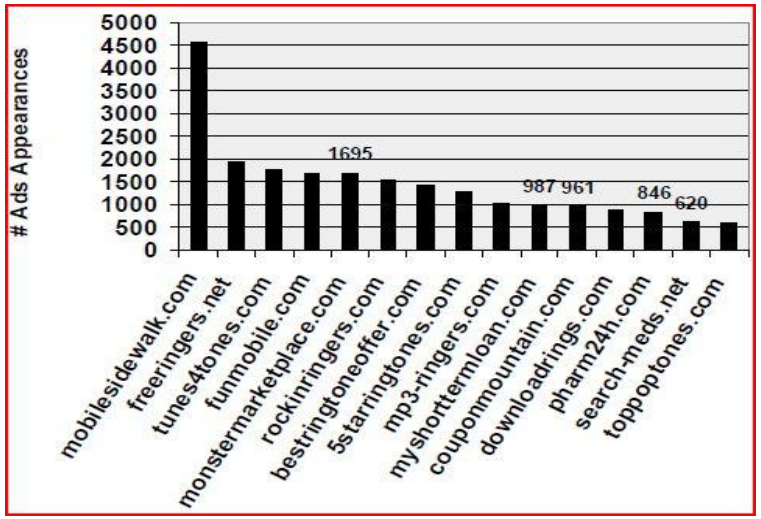
- 66.230 and 64.111 IP blocks contain again dominating receiver domains for spam-ads click-through traffic



Redirection Spam Analysis :
Double-Funnel Analysis: **Layer #5 (Advertisers)**

Spammers-targeted

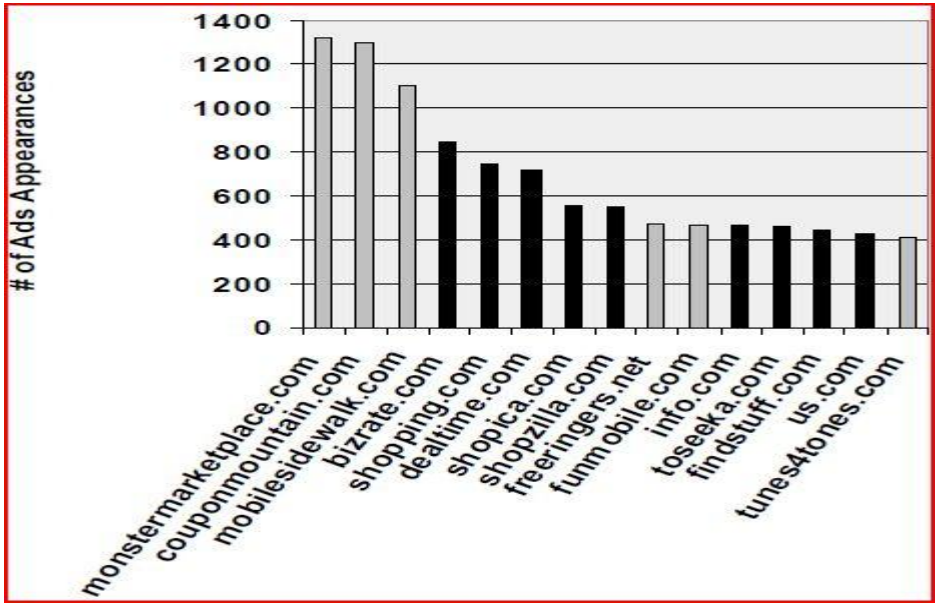
- The advertisers' domain names were often displayed either as anchor text or in the status bar upon mouse-over



- Top-15 advertisers by the number of ads appearances on spam pages (ringtones!!)

Advertisers-targeted

- Top-15 advertisers which are significantly different



- This benchmark's keywords better match these shopping websites

Redirection Spam Analysis : Double-Funnel Analysis: **Layer #4 (Syndicators)**

Spammers-targeted

- A handful of syndicator domains had significant presence in the redirection chains
  -> The major middlemen between spam-traffic aggregators and the advertisers

Advertisers-targeted

- The two benchmarks share the same top-3 syndicators, despite the fact they had only 15% overlap in the list of keywords and very different top-advertisers

Other Common Spam

- Many syndication-based spammers who do not use client-side browser redirections to fetch ads, share the same bottom half of the double-funnel with redirection spammers
- Although they fetch ads on the server side, they also funnel the click-through traffic from their pages into the same IP blocks

·Blog farms ·Parasite Ads-Portal Farms

Conclusion

Presented a 5-layer double-funnel model in which

- Ads from merchant advertisers are funneled through syndicators, aggregators & redirection domains to get displayed on doorway pages
- Click-through traffic from these spam ads is funneled in the reverse direction to reach advertisers
- Middle layers provide the infrastructure for converting spam traffic to money