Adversarial Web Crawling with Strider Monkeys

Yi-Min Wang

Director, Cyber-Intelligence Lab
Internet Services Research Center (ISRC)
Microsoft Research
Search Engine Basics

• Crawler
  – Crawling policy
• Page classification & indexing
• Static ranking
• Query processing
• Document-query matching & dynamic ranking
  – Diversity
• **Goals of web crawling**
  – Retrieve web page content seen by browser users
  – Classify and index the content for search ranking

• **What is a monkey?**
  – Automation program that mimics human user behavior
Stateless Static Crawling

• Assumptions
  – Input to the web server: the URL
    • Stateless client
  – Output from the web server: page content in HTML
    • Static crawler ignores scripts
Stateful Static Crawling

• We all know that Cookies affect web server response

• HTTP User-Agent field affects response too
  – Some servers may refuse low-value crawlers
  – Some spammers use crawler-browser cloaking
    • Give crawlers a page that maximizes ranking (=traffic)
    • Give users a page that maximizes profit
Dynamic Crawling

• Simple crawler-browser cloaking can be achieved by returning HTML with scripts
  – Crawlers only parse static HTML text that maximizes ranking/traffic
  – Users’ browsers additionally execute the dynamic scripts that maximize profit
    • Usually redirect to a third-party domain to server ads

• Need browser-based dynamic crawlers to index the true content
Search Spam Example: Google search “coach handbag”
Spam Doorway URL = http://coach-handbag-top.blogspot.com

Coach Handbag - Handbags & Wallets - Compare Prices, Reviews and ...
Coach Handbag - 28 results like the COACH gray signature bag, American West Painted Pony small coach bag, COACH Signature Patchwork Shoulder Tote (Replica ... www.nextag.com/coach-handbag/search-html - 90k - Cached - Similar pages

coach handbag
Purses e. g. As if coach handbag she had received some other flexible material. And our coach handbag bank tt in the issue of paper cloth, thin plastic or ...
coach-handbag-top.blogspot.com/ - 22k - Cached - Similar pages

eBay - coach handbag, Women's Accessories, Handbags, Fragrances ...
Buy coach handbag, Women's Accessories, Handbags items on eBay. Find a huge selection of Fragrances, Vintage items and get what you want now!
search.ebay.com/coach-handbag - 114k - Cached - Similar pages

Coach Handbags Outlet | BonafideLuxury
Authentic Coach handbags outlet. Discount Coach bags, purses, wallets, belts and woman accessories.
www.bonafideluxury.com/ - 6k - Cached - Similar pages

Result Page: 1 2 3 4 5 6 7 8 9 10 Next
http://coach-handbag-top.blogspot.com/ script execution led to redirection to topsearch10.com

1. Handbag
   Handbag links, tips and deals at CouponMountain
   URL: http://www.couponmountain.com (Sponsored Link)

2. Quality Bags at Wholesale Prices
   Great deals on bags. Huge selections and amazing low prices.
   URL: http://www.BagsWholesale.com (Sponsored Link)

3. Coach Handbags - Buy at DealTime.com!
   Save time and money every time you shop! DealTime lets you compare products, prices and stores for free. Find the Web’s best prices in seconds.
   URL: http://www.DealTime.com (Sponsored Link)
Italy further discount 1 minimum 5 pieces 2 discount coach handbag of paper, bags that they had physically swung her style coach handbag 3 dimensions 300 x 60mm 4 different material a lock coach handbag Sachets are also be bluntly called a money bag is a bag coach handbag may also made of paper, bag. Factories. Many styles and coach handbag cards. A suitcase. Cheap disposable paper bags to the coach handbag sale, terms 1 minimum 5 pieces 2 payment at all. The coach handbag Italian retail trade as if she had received some other coach handbag european countries and colors of nice models, ask for coach handbag use in pockets of his payment, within 12 24 hours. Fromports coach handbag meaning that is similar to the italian retail trade as coach handbag metaphorically bruised as a clutch. A murse or simply coach handbag by a lock. Sachets are england, france, italy, greece, coach handbag germany, the packing list. All the buyer's purchase amount coach handbag sale from small objects in pockets of paper, as an alternative coach handbag to putting small objects in the issue of sheets of clothing coach handbag to buyers' requirement a clutch a purse and pakistan coach handbag. We are also be bluntly called a paper bags that is a coach handbag good reputation among bags has a famous handbag shirt coach handbag shorts, pants, jeans, sunglasses, dresses, shoe, watches, coach handbag sports shoes. Collection summer 2006. Available lots coach handbag of nice models, ask for shoppers, and australia. Specifications coach handbag 1 from the british so a murse or none at our products coach handbag have got a purse all a pocketbook is a purse is often coach handbag called a bag sometimes a paper as 'handbagging' her that coach handbag they had physically swung her being described in the coach handbag buyer's payment bank it in pockets of debate and other coach handbag european countries such a clutch a smaller alternative coach handbag to 30 off on the order, note after the order to confirm the coach handbag the retail trade as if she had received some other flexible coach handbag material. And colors and our products have got a small coach handbag ones, for free or none at our products have got a money coach handbag and our website, our bank receipt of a bag is a convenience coach handbag for shoppers, and discussion led to large ones for the coach handbag british so a murse or for a.
Link spam from a spammed forum
Changing the Battleground

- Traditional static crawler
- Static page-content index: spammers’ preferred battleground

New Battleground

- Strider dynamic crawler
- Dynamic program-behavior index:
  - Redirection traffic analysis - “Follow the Money” to their headquarters - our preferred battleground
Simple Obfuscation

**OnlineGp.com:**

http://blackjack-score.blogspot.com/

```html
<script>
    eval('window.wloc=window.onr+"http://www.on+linegp+p.com/+"');
</script>
```

**TopMobile10.com:**

http://d-6010-nokia-ringtone.blogspot.com/

```html
<script>eval(unescape('%74%6F%70%6D%6F%62%69%6C%65%31%30%2E%63%6F%6D = topmobile10.com')
```

```html
</script>
```
Advanced Obfuscation
Challenges in Scalable Dynamic Crawling

• Asynchronous, stateful crawling
• Caching of shared secondary-URL pages
• UI/screen display issues
• Enhanced index containing visual information
• Re-trained ranking components
Non-redirection Spam Examples

http://www.google.com/notebook/public/18342072077318071616/BDQKIIgoQx-nSn6cj

Regal Movie Theaters
Last edited July 13, 2006
More by »

Best Price! Regal Movie Theaters!

ENTER HERE: Regal Movie Theaters
Interactive Dynamic Crawling

• **Click-Through Cloaking (CTC)**
  • Search click-through visitors and direct visitors see different pages

• **Server-side cloaking**
  • Web server checks **Referer** field of HTTP header
  • Advanced cloaking: distinguish spam investigation-style queries from regular search queries
    – Ignore click-throughs from "url:" (or "info:"), "link:”, "linkdomain:”, “site:” queries
  • But client can fake HTTP **Referer** field
Client-side cloaking

Client-side script checks local browser’s `document.referrer` variable

```javascript
var url = document.location + ""; exit=true;
ref=escape(document.referrer);
if (((ref.indexOf('search')===-1) && (ref.indexOf('google')===-1) && (ref.indexOf('find')===-1) && (ref.indexOf('yahoo')===-1) && (ref.indexOf('aol')===-1) && (ref.indexOf('msn')===-1) && (ref.indexOf('altavista')===-1) && (ref.indexOf('ask')===-1) && (ref.indexOf('alltheweb')===-1) && (ref.indexOf('dogpile')===-1) && (ref.indexOf('excite')===-1) && (ref.indexOf('netscape')===-1) && (ref.indexOf('fast')===-1) && (ref.indexOf('seek')===-1) && (ref.indexOf('find')===-1) && (ref.indexOf('searchfeed')===-1) && (ref.indexOf('about.com')===-1) && (ref.indexOf('dmoz')===-1) && (ref.indexOf('accoloana')===-1) && (ref.indexOf('crawler')===-1)) { exit=false; } if (exit) { p=location; r=escape(document.referrer);
location='http://ppcan.info/mp3re.php?niche=Evans, Sara&ref='+r }
```
• Advanced cloaking

```javascript
Function is_se_traffic() {
    if ( document.referrer ) {
        if ( document.referrer.indexOf("google")>0
            || document.referrer.indexOf("yahoo")>0
            || document.referrer.indexOf("msn")>0
            || document.referrer.indexOf("live")>0
            || document.referrer.indexOf("search.blogger.com")>0
            || document.referrer.indexOf("www.ask.com")>0 )
        {
            if ( document.referrer.indexOf( document.domain )<0
                && document.referrer.indexOf( "link%3A" )<0
                && document.referrer.indexOf( "linkdomain%3A" )<0
                && document.referrer.indexOf( "site%3A" )<0 )
            {
                return true;
            }
        }
    }
    return false;
}
```
Cloaking Examples

404 Not Found

The page you have requested no longer exists.

This may be for a number of reasons:

* The file may have been moved or deleted by user
* The file may have been deleted for abuses

HTTP 403 (Forbidden) - Microsoft Internet Explorer

You are not authorized to view this page

You might not have permission to view this directory or page using the credentials you supplied.

If you believe you should be able to view this directory or page, please try to contact the web site by using any e-mail address or phone number that may be listed on the www.acp.edu home page.

403 Forbidden - Microsoft Internet Explorer

Forbidden

You don't have permission to access / on this server.

Apache/1.3.34 Server at www.mandevillechevrolet.com Port 80
Cloaking Examples

Account is suspended due to TOS non-observance

Page was deleted from this hosting

This site will be closed in five days for a comment and e-mail spam. If you have suffered from actions of this site, please, fill this form below.

Step 1. General information

Name
E-Mail
Telephone (optional)
Company (optional)
Website address
Combo cloaking with obfuscation

mywebpage.netscape.com/superphrm2/order-tramadol.htm

• Extract **document.referrer** and report it to spam server (which hides cloaking logic)

```html
<script> var params="f=pharmacy&cat=tramadol";

function kqqw(s) {
    var Tqqe=String("qwertyuioplkjhgfdsazxcvbnmQWERTYUIOPLKJHGFDSAZXCVBNM_1234567890");
    var tqqr=String(s); var Bqqt=String('"');
    var lqqy,pqqu,Yqqi=tqqr.length;
    for ( lqqy=0; lqqy<Yqqi; lqqy+=2) {
        pqqu=Tqqe.indexOf(tqqr.charAt(lqqy))*63;
        pqqu+=Tqqe.indexOf(tqqr.charAt(lqqy+1));
        Bqqt=Bqqt+String.fromCharCode(pqqu);
    }
    return(Bqqt);
}

eval(kqqw('wKwVwLw2wXwJwCw1qXw4wMwDw1wJqGqHq8qHqSqHw_ ... Bw1qHqSqHq0qHqFq7')); </script>
```
Strider SearchMonkey

• Find an appropriate query string
  – Anchor text, URL ‘subcomponent’, etc.
• Submit the query to a search engine
• Insert the to-be-crawled URL, if necessary
• Click through the URL
Comprehensive Dynamic Crawling

• **Drive-by downloads**
  – When a vulnerable browser visits a malicious Web page
    • Code Obfuscation
    • Third-Party Redirection
      – Content Provider → Exploit Provider
    • Vulnerability Exploits
    • Malware Installation
  – All without any user interaction
Drive-by Download Examples

DANGER: SPYWARE

Full system scan results:
- 3 Spyware infections
- 27 Software tracks
- 95 Adult-oriented websites tracks
- 3 Programs with probable keystlogging activity

Windows recommends you the following software products to keep your PC safe:

for as low as $99.95
demo direct download

Warning!

Your computer might be infected with spyware or adware !!!

Strange homepage, popups, loss of important data and unstable functioning are the sure signs that you are infected.

Click here to get the latest spyware removal software.

VIRUS ALERT!
YOUR PC IS INFECTED!

IT HAS BEEN DETECTED THAT YOUR PC HAS AT LEAST 3 DANGEROUS VIRUSES!
TO KNOW FOR SURE YOU URGENTLY NEED TO RUN AN ANTIVIRUS TEST ON YOUR PC!

The consequences of spyware and virus presence on your pc might belike, losing all the data, data might be stolen, your secrets might be exposed.

PROTECT YOUR PC!
REMOVE ALL VIRUSES NOW!

Security warning

A fatal error in IE has occured at 0028:C00011E36 in VXD UNM(O1) + 00010E36. Error was caused by Trojan-Spy.HTML.Smitfraud.c

* System can not function in normal mode.
  Please check you security settings.
* Scan your PC with any available antivirus / spyware remover program tofix the problem.
HoneyMonkey, short for Strider HoneyMonkey Exploit Detection System, is a Microsoft Research honeypot. The implementation uses a network of computers to crawl the World Wide Web searching for websites that use browser exploits to install malware on the HoneyMonkey computer. A snapshot of the memory, executables and registry of the honeypot computer is recorded before crawling a site. After visiting the site, the state of memory, executables, and registry is compared to the previous snapshot. The changes are analyzed to determine whether the visited site installed malware onto the honeypot computer.

HoneyMonkey is based on the honeypot concept, with the difference that it actively seeks websites that try to exploit it. The term was coined by Microsoft Research in 2005. With honeymoons it is possible to find open security holes that aren’t yet publicly known but are exploited by attackers.
HoneyMonkey versus Honeypot

• **Traditional honeypots**
  – Passive, server-side honeypots that mimic vulnerable network services
    • Wait for attacks from malicious client machines

• **Strider HoneyMonkeys**
  – Active, client-side honeypots that mimic humans using vulnerable client software (e.g., browsers)
    • Seek attacks from malicious server machines
  – Need a list of URLs to visit
  – Precise attribution for detected exploits
    • Exploit URLs can be passed on to others for repro
Key Idea: Black-box Exploit Detection

- Behavior-based detection: detect software installation following a successful vulnerability exploit
  - Wait a couple of minutes and check for files created outside IE sandbox
  - Strider systems management approach to cybersecurity
- Can capture both known-vulnerability and zero-day exploits
- Use a pipeline of Virtual Machines (VMs)
  - Each with a different patch level to determine the strength of an exploit
  - Ease clean-up & automation
HoneyMonkey Black-box Exploit Detection

- Other Doorway Pages
- Doorway Page
- Obfuscated Java Scripts
- Third-Party URLs
- Malicious Scripts
- FDR
- Virtual Machine
- URL Tracer
- Malware Installation
- Exploit Provider
The HoneyMonkey Pipeline

Different Client-side Software Versions

Suspect Websites

HM Software
HM Software
HM Software
HM Software

Unpatched WinXP
Unpatched WinXP SP2
Partially Patched WinXP SP2
Fully Patched WinXP SP2

Malicious Websites
Advanced Malicious Websites
More Advanced Malicious Websites
Zero-day Exploit Websites
SUMMARY

• Adversarial web crawling
  – “Web program”, not just “web page”
  – “Program behavior”, not just “page content”
  – Dynamic crawling – *scripts & redirections*
  – Interactive crawling – *mimic targeted audience*
  – Stateful crawling – *client machine state as input*
  – Comprehensive web program behavior analysis
Publications

• **Strider Search Ranger** [http://research.microsoft.com/searchranger/](http://research.microsoft.com/searchranger/)

• **Strider HoneyMonkey** [http://research.microsoft.com/honeymonkey/](http://research.microsoft.com/honeymonkey/)

• **Strider Typo-Patrol** [http://research.microsoft.com/URLTracer](http://research.microsoft.com/URLTracer)
  - “Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting,” in *Proc. SRUTI*, July 2006
    - [http://www.usenix.org/events/sruti06/tech/wang.html](http://www.usenix.org/events/sruti06/tech/wang.html)
Spam Double-Funnel

- O(10^7) spam pages
- Layer #1
- O(10^4) redirection domains
- Layer #2
- O(10^0) IP-address block
- Layer #3
- O(10^0) syndicators
- Layer #4
- O(10^5) advertisers
- Layer #5

Diagram:
- Doorway
- Aggregators
- Redirection Domain
- Layer #1
- Layer #2
- Layer #3
- Layer #4
- Layer #5