



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

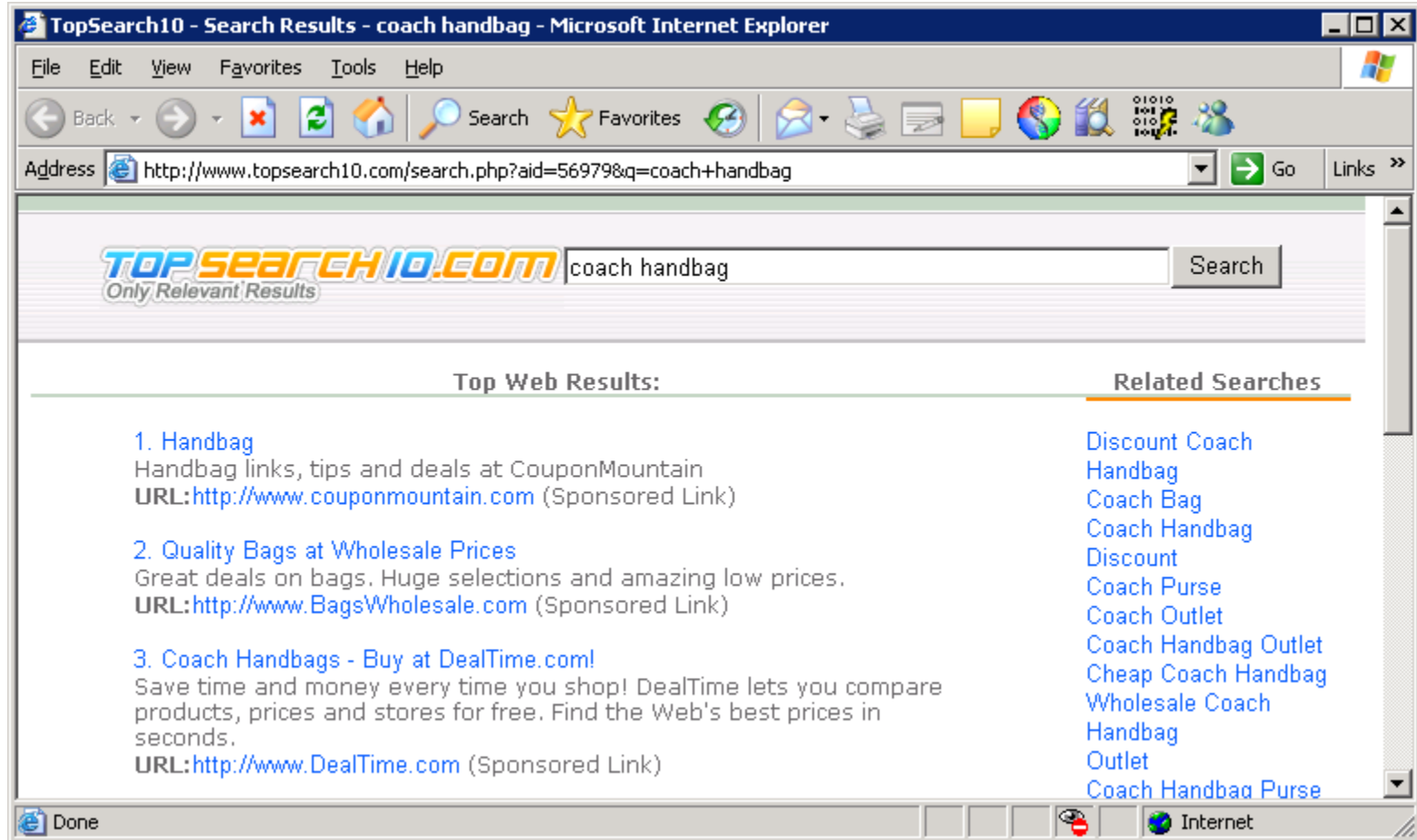
Stateful Static Crawling

- 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

<http://coach-handbag-top.blogspot.com/> script execution led to redirection to topsearch10.com





Link spam from a spammed forum


DOG LEAF BBS - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address <http://cc.msnsocache.com/cache.aspx?q=4693596052619&lang=en-US&mkt=en-US&FORM=CVRE4> Go Links >>

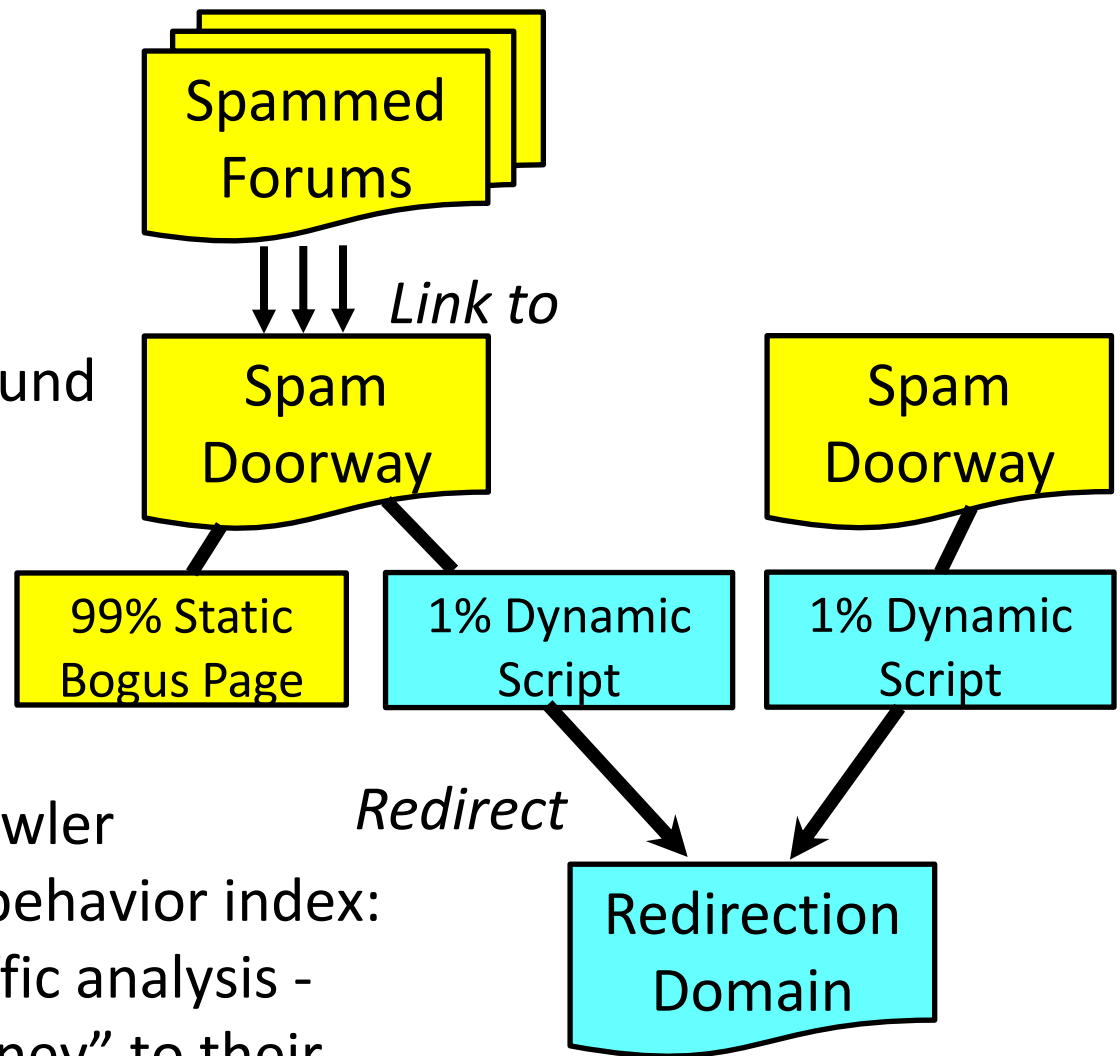
無題 投稿者: [payday loan](#) 投稿日: 2006/10/23(Mon) 02:06 No.867  

 <http://faxless-payday-loan-new.blogspot.com> faxless payday loan http://quick-payday-loan-new.blogspot.com quick payday loan http://no-faxing-payday-loan-new.blogspot.com no faxing payday loan http://payday-advance-loan-new.blogspot.com payday advance loan http://bad-credit-payday-loan-new.blogspot.com bad credit payday loan http://instant-payday-loan-new.blogspot.com instant payday loan http://money-tree-payday-loan-new.blogspot.com money tree payday loan http://no-teletrack-payday-loan-new.blogspot.com no teletrack payday loan http://payday-loan-uk-new.blogspot.com payday loan uk http://david-yurman-jewelry-new.blogspot.com david yurman jewelry http://jewelry-appraiser-new2.blogspot.com jewelry appraiser http://jewish-jewelry-top.blogspot.com jewish jewelry http://amber-silver-jewelry-best.blogspot.com amber silver jewelry http://baltic-amber-silver-jewelry2.blogspot.com baltic amber silver jewelry http://jewelry-boxes2.blogspot.com jewelry boxes http://zales-jewelry-top.blogspot.com zales jewelry http://tiffanys-jewelry-new2.blogspot.com tiffanys jewelry http://pandora-jewelry-new2.blogspot.com pandora jewelry http://lia-sophia-jewelry-new2.blogspot.com lia sophia jewelry http://kays-jewelry-new2.blogspot.com kays jewelry http://jareds-jewelry-new2.blogspot.com jareds jewelry http://jared-galleria-new2.blogspot.com jared galleria http://gordons-jewelry-new2.blogspot.com gordons jewelry http://friedmans-jewelry-new2.blogspot.com friedmans jewelry http://brighton-jewelry-new2.blogspot.com brighton jewelry http://coach-handbag-top.blogspot.com coach handbag http://gucci-handbag-new.blogspot.com gucci handbag http://replica-handbag-new.blogspot.com replica handbag http://prada-handbag-new2.blogspot.com prada handbag http://coach-handbag-top.blogspot.com

Internet

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/>

```
<Script>  
eval('windo'+w.loc+'ati'+on.r+'eplac'+e('ht'+tp:+'/'+ww'+w.on+'lineg'+p.com+'/'));  
</Script>
```

🚩 TopMobile10.com: %74%6F%70%6D%6F%62%69%6C%65%31%30%2E%63%6F%6D = topmobile10.com

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

```
<script>eval(unescape('%74%6F%70%6D%6F%62%69%6C%65%31%30%2E%63%6F%6D%2F%73%65%61%72%63%68%2E%70%68%70%3F%61%69%64%3D%33%30%36%32%35%26%73%61%69%64%3D%76%33%26%71%3D%72%69%6E%67%74%6F%6E%65%22%3B'))</script>
```

Advanced Obfuscation

ABC Searcher.com:

<http://georgia-homeowners-insurance-9.blogspot.com/>

```
<script>
function ebbcd(a,r)
{
if(r){var ei2=[],s="ab";s+="c";s+="d";s+="ef";s+="gh";s+="ijk";s+="lmn";
s+="op";s+="q";s+="rs";s+="t";s+="u";s+="vw";s+="xy";s+="z";g=t="/";k="\\"
for(i=0;i<s.length;i++)ei2[s.charAt(i)]=s.charAt((i+13)%26);
for(i=0;i<s.length;i++)ei2[s.charAt(i).toUpperCase()]=s.charAt((i+13)%26).toUpperCase();
s="";for(i=0;i<a.length;i++){b=a.charAt(i);s+=(b>='A'&&b<='Z'||b>='a'&&b<='z'?ei2[b]:b);}return s;}
eval(ebbcd('qbpzhzrag',1)'+.+ebbcd('ybpngvba=',1)+k+ebbcd('uggc:'+g+t+a+'',1))
}
```

<http://hilaryduff-z.blogspot.com/>

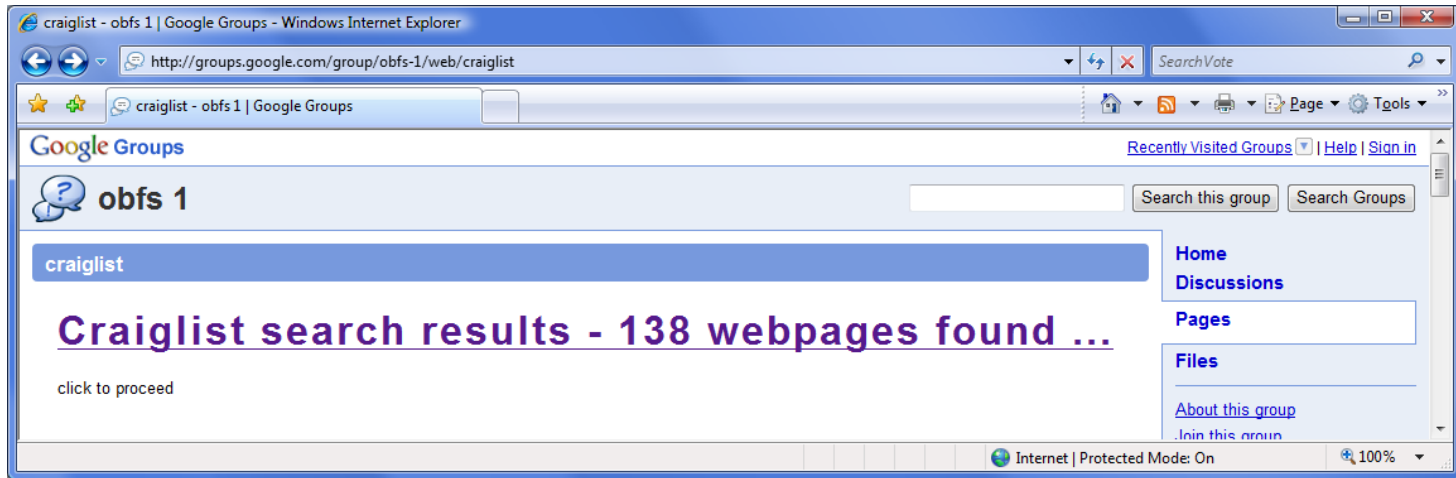
```
<SCRIPT LANGUAGE="JavaScript">
<!--
function Decode()var temp="",i,c=0,out="";var str="60!115!99!114!105!112!116!62!13!10!32!102!117!110!99!116!105!111!110!32!82!40!4!1!23!13!10!32!118!97!114!32!82!10!102!6!1!00!111!109!10!1110!116!46!114!10!1102!10!1114!114!10!1114!59!13!10!32!32!13!10!32!105!102!32!40!82!10!1102!46!105!110!100!10!1120!79!102!40!39!46!103!111!111!103!108!10!1146!39!4!133!6!1145!49!32!124!124!32!82!10!1102!46!105!110!100!10!1120!79!102!40!39!46!109!115!110!46!39!4!133!6!1145!49!32!124!124!32!82!10!102!46!105!110!100!10!1120!79!102!40!39!46!12!1197!104!111!111!1146!39!4!133!6!1145!49!32!124!124!32!82!10!1102!46!105!110!100!10!1120!79!102!40!39!46!97!111!108!46!39!4!133!6!1145!49!32!124!124!32!82!10!1102!46!105!110!100!10!1120!79!102!40!39!46!97!15!107!46!39!4!133!6!1145!49!32!124!124!32!82!10!1102!46!105!110!100!10!1120!79!102!40!39!46!97!108!116!97!118!105!115!116!97!46!39!4!133!6!1145!49!4!113!10!32!32!123!32!100!111!99!117!109!10!110!110!116!46!119!114!105!116!10!140!39!60!115!99!114!105!112!116!32!108!97!110!103!117!97!103!10!116!134!106!97!118!97!115!99!114!105!112!116!34!62!119!105!110!100!111!39!43!39!119!46!108!111!99!97!116!105!111!110!16!1134!104!116!112!58!47!47!19!119!119!46!116!111!12!97!100!117!108!116!49!48!46!99!111!109!47!115!10!197!114!99!104!46!112!104!112!63!97!105!100!6!153!49!5!56!55!38!113!6!1104!105!108!97!114!12!1143!100!117!102!102!34!60!47!115!39!43!39!99!14!105!112!116!62!39!4!1125!13!10!13!10!10!108!115!10!132!123!13!10!100!111!99!117!109!10!1110!116!46!119!114!105!116!10!140!39!60!115!99!114!105!112!116!34!62!119!105!110!100!111!39!43!39!119!46!108!111!99!97!116!105!111!110!16!1134!104!116!116!12!58!47!47!119!119!119!46!116!111!112!97!100!117!108!116!49!48!46!99!111!109!47!115!10!197!114!99!104!46!112!104!112!63!97!105!100!6!153!53!52!56!49!38!113!6!1104!105!108!97!114!12!143!100!117!102!102!34!60!47!115!39!43!39!99!114!105!112!116!62!39!4!113!10!125!13!10!32!125!13!10!32!32!13!10!32!82!40!4!159!13!10!32!32!13!10!60!47!83!99!114!105!112!116!62!";j=str.length;while(c<=str.length-1){while(str.charAt(c)!='')temp=temp+str.charAt(c++);c++;out=out+String.fromCharCode(temp);temp=""document.write(out);}
//-->
</SCRIPT><SCRIPT LANGUAGE="JavaScript">
<!--
Decode();
//-->
```

Challenges in Scalable Dynamic Crawling

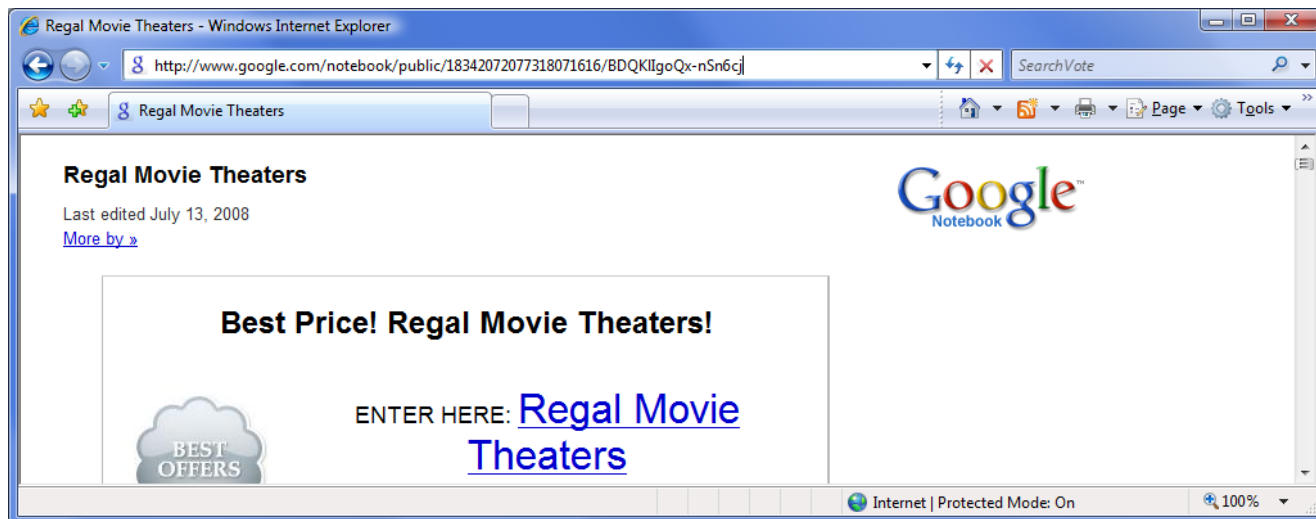
- Asynchronous, stateful crawling
- Caching of shared secondary-URL pages
- UI/screen display issues
- Enhanced index containing visual information
- Re-trained ranking components

Non-redirection Spam Examples

<http://groups.google.com/group/obfs-1/web/craigslist>



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



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

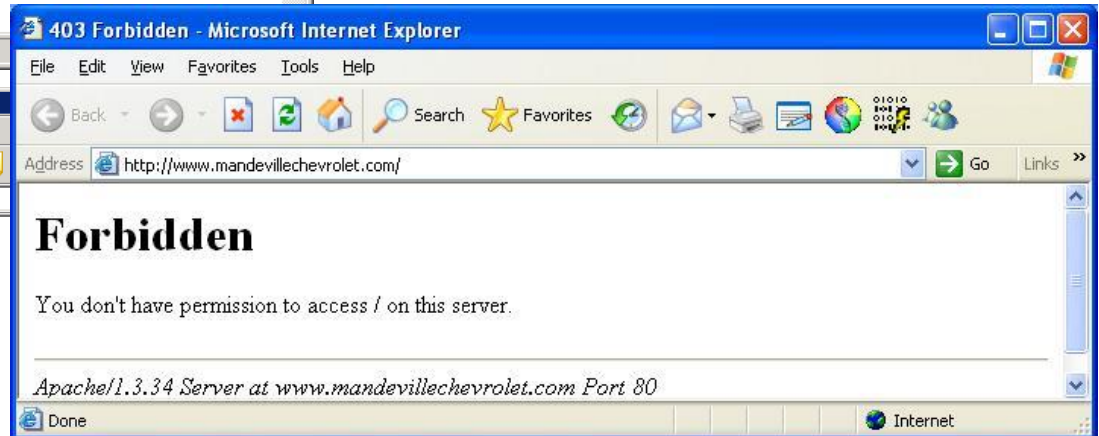
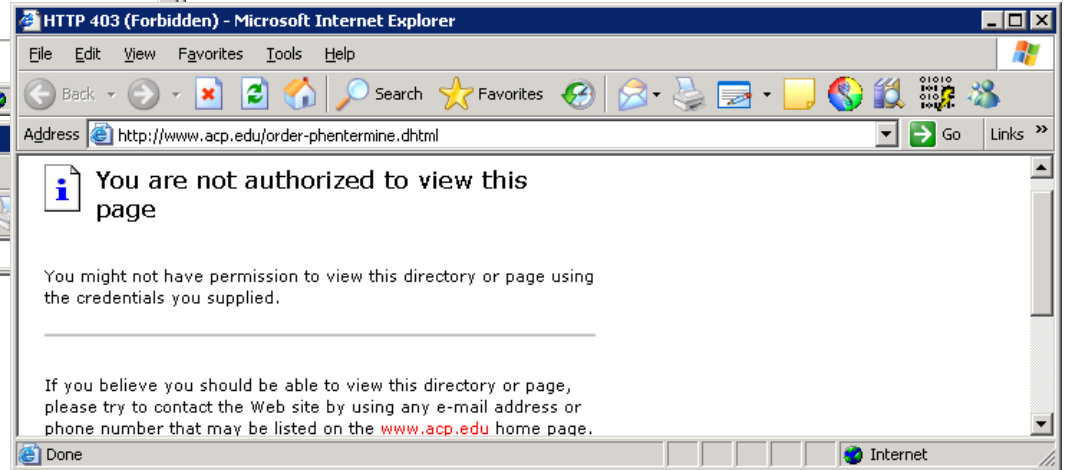
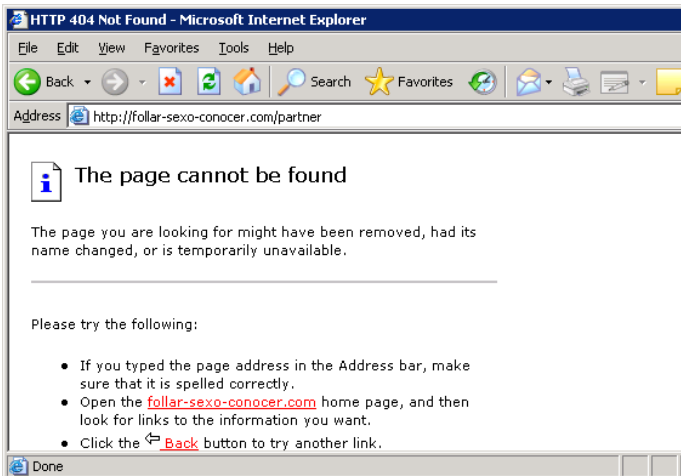
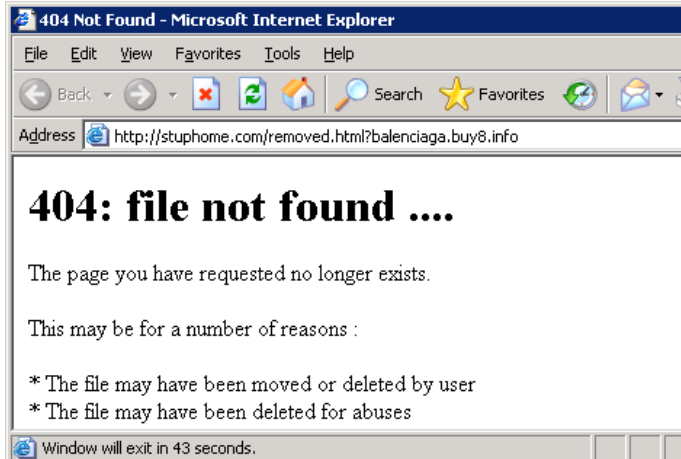
- Client-side script checks local browser's **document.referrer** variable

```
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('accoona')===-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 }
```

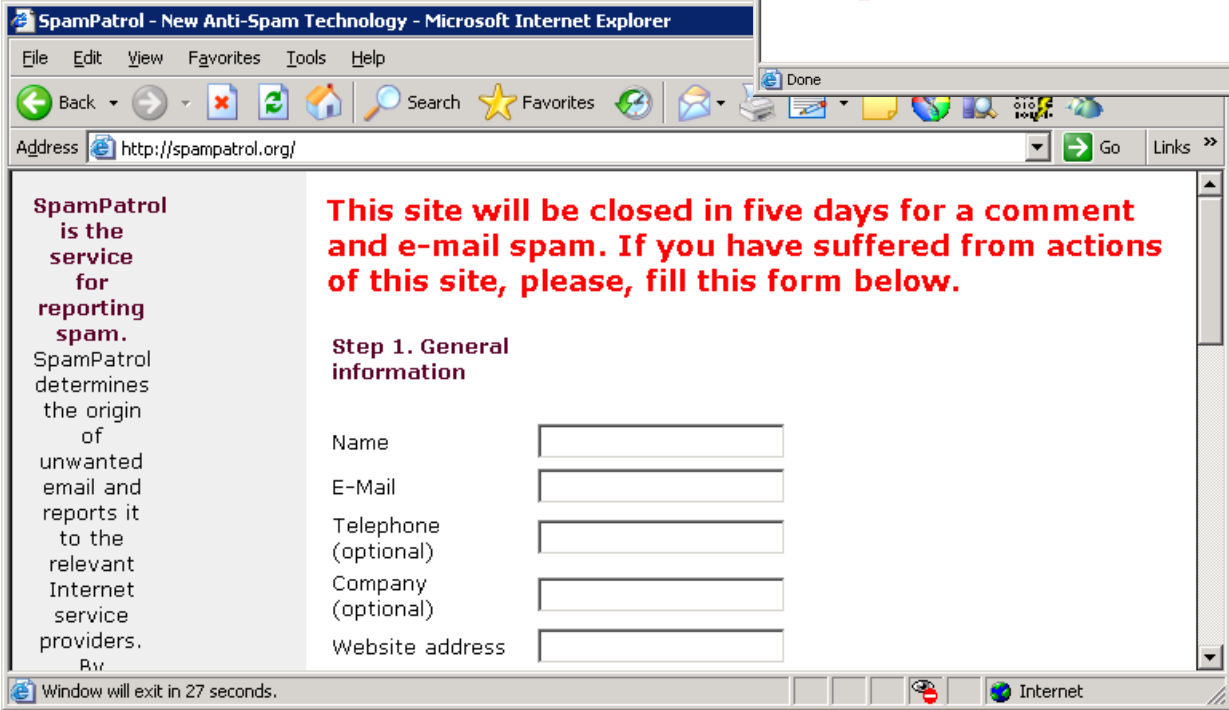
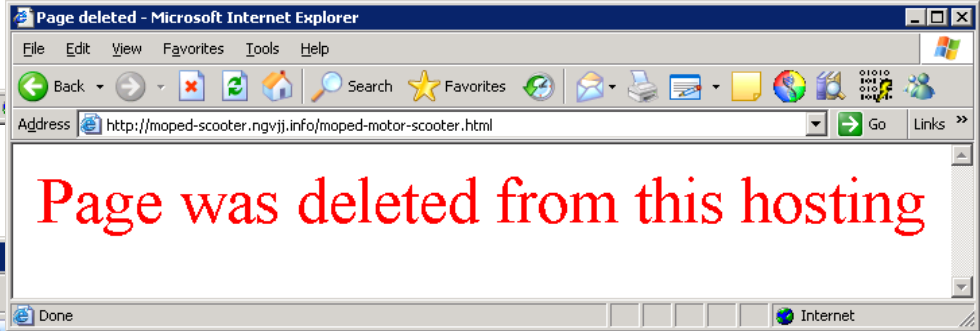
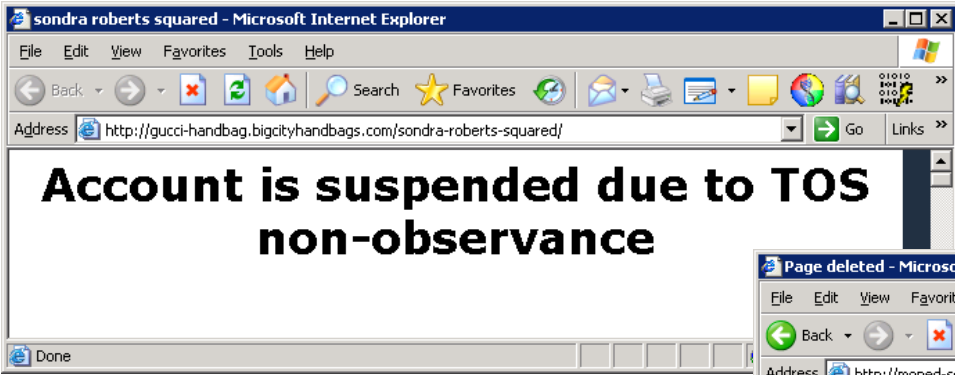
- Advanced cloaking

```
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



Cloaking Examples



Combo cloaking with obfuscation

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

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

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

```
function kqqw(s) {
```

```
    var Tqqe=String("qwertyuioplkjhgfdsazxcvbnmQWERTYU  
    IOPLKJHGFDSA ZXCVBNM_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 Spyware tracks
- 95 Adult-oriented websites tracks
- 3 Programs with probable keylogging activity

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

for as low as \$59.95 demo direct download

for as low as \$59.95 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.

Your computer is still vulnerable to new attacks !!!



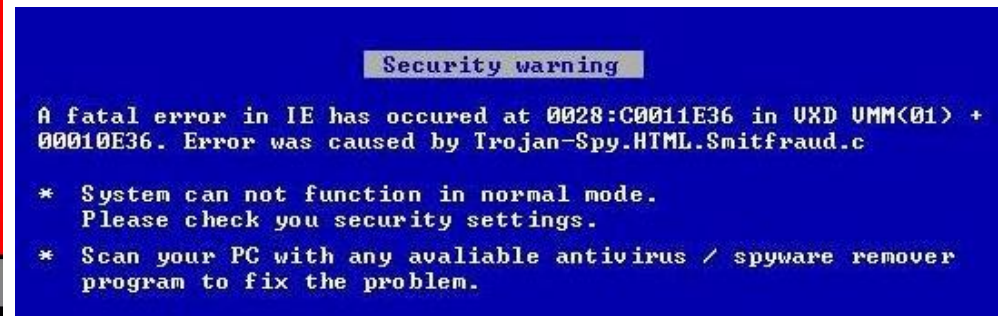
VIRUS ALERT!
YOUR PC IS INFECTED!

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 be like: loosing all the data, data might be stolen, your secrets might be exposed.

PROTECT YOUR PC!
REMOVE ALL VIRUSES NOW!

[Removal instructions](#)



Security warning

A fatal error in IE has occurred at 0028:C0011E36 in UXD UMM<01> + 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 to fix the problem.

Strider HoneyMonkey

HoneyMonkey - Wikipedia, the free encyclopedia - Windows Internet Explorer

http://en.wikipedia.org/wiki/Honeymonkey

File Edit View Favorites Tools Help

HoneyMonkey - Wikipedia, the free encyclopedia

You can [support Wikipedia](#) by making a tax-deductible donation. [Log in / create account](#)

[article](#) [discussion](#) [edit this page](#) [history](#)

HoneyMonkey

From Wikipedia, the free encyclopedia
(Redirected from [Honeymonkey](#))

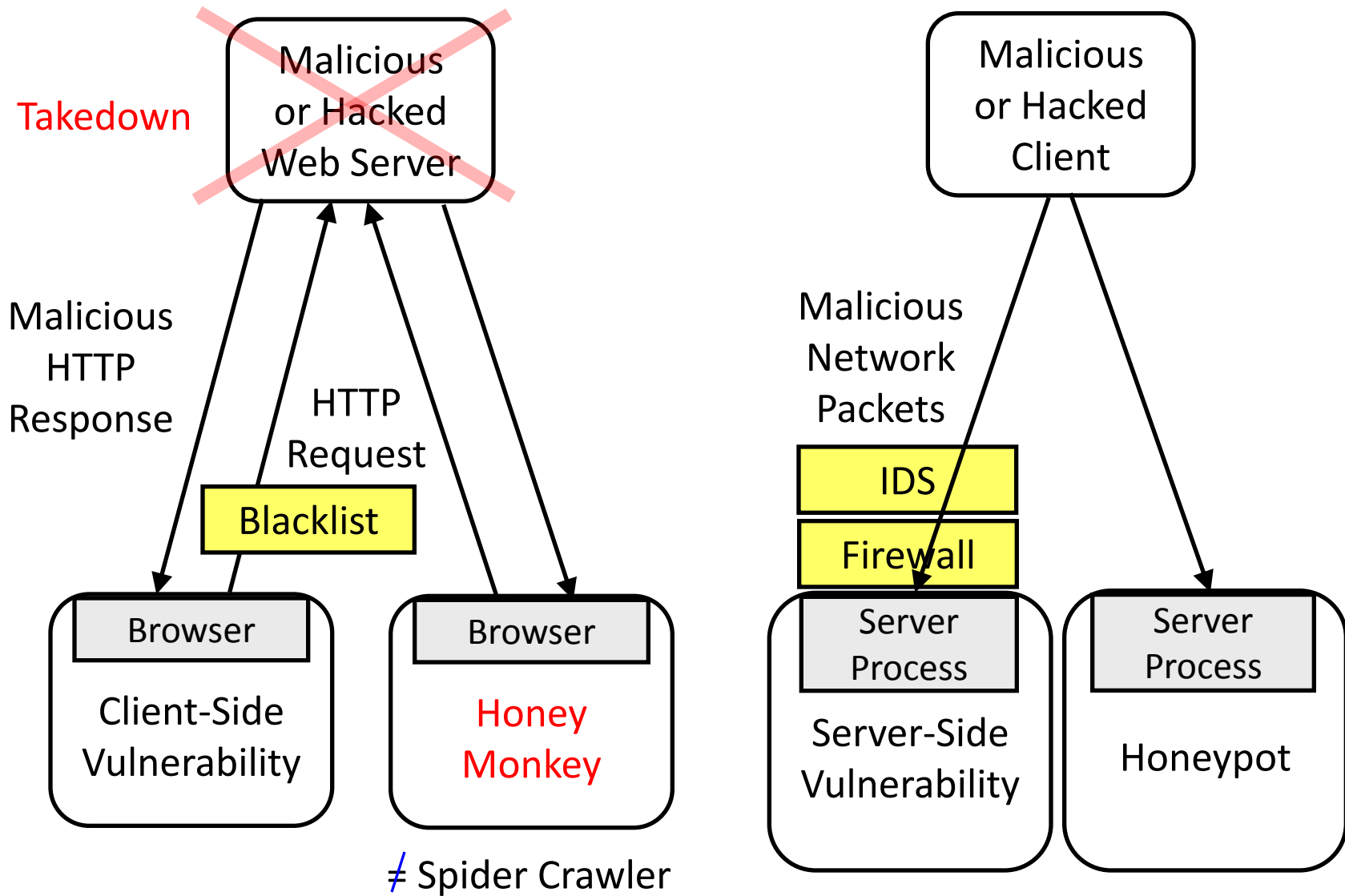
HoneyMonkey, short for **Strider HoneyMonkey Exploit Detection System**, is a [Microsoft Research honeypot](#). The implementation uses a network of computers to [crawl](#) the [World Wide Web](#) searching for [websites](#) that use [browser exploits](#) to install [malware](#) on the HoneyMonkey computer. 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 honeymonkeys it is possible to find open [security holes](#) that aren't yet publicly known but are exploited by attackers.

Done Internet | Protected Mode: On 100%

HoneyMonkey versus HoneyPot

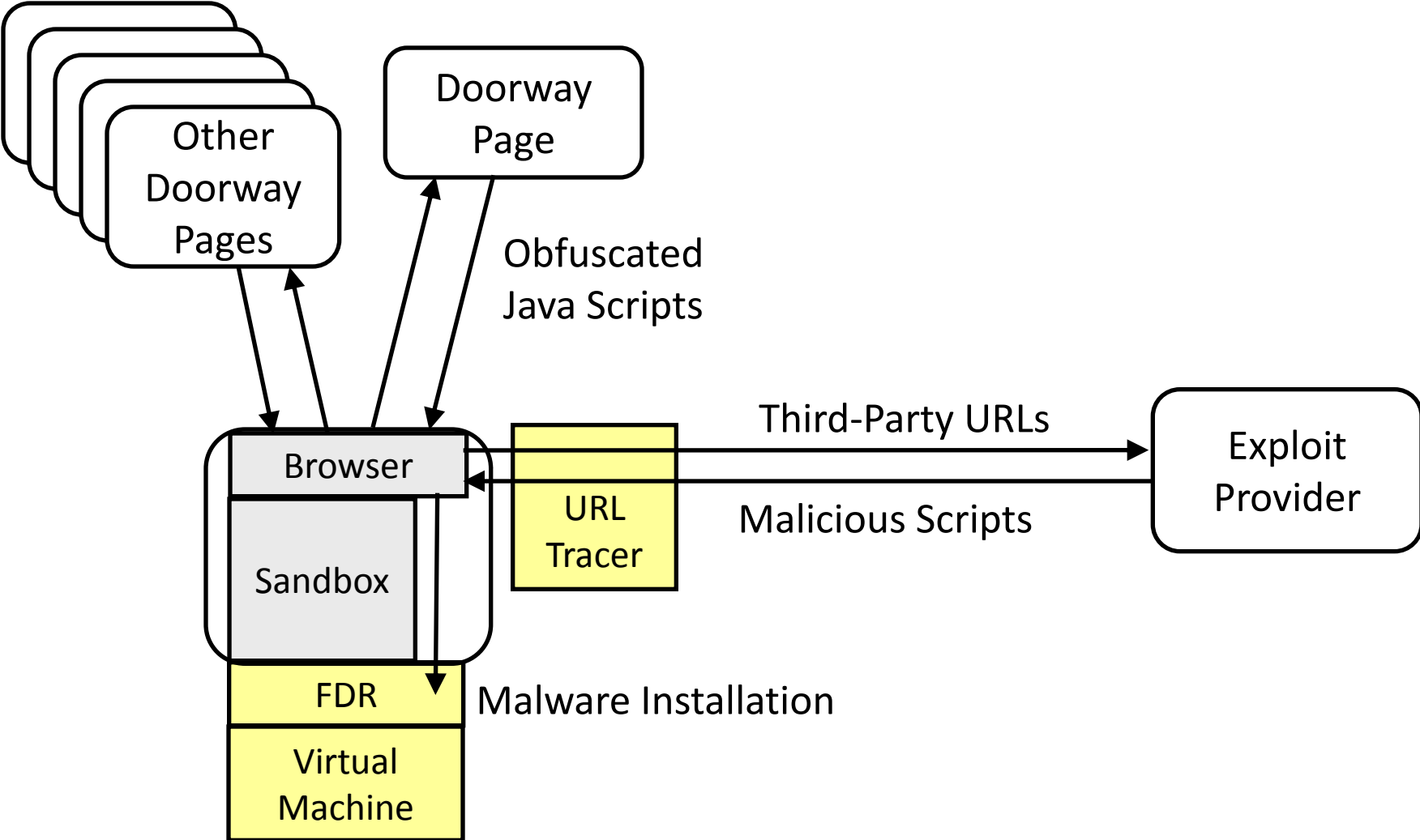
- Traditional honeypots
 - Passive, server-side honeypots that mimic vulnerable network services
 - Wait for attacks from malicious client machines
- Strider HoneyMonkeys
 - Active, client-side honeypots that mimic humans using vulnerable client software (e.g., browsers)
 - Seek attacks from malicious 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

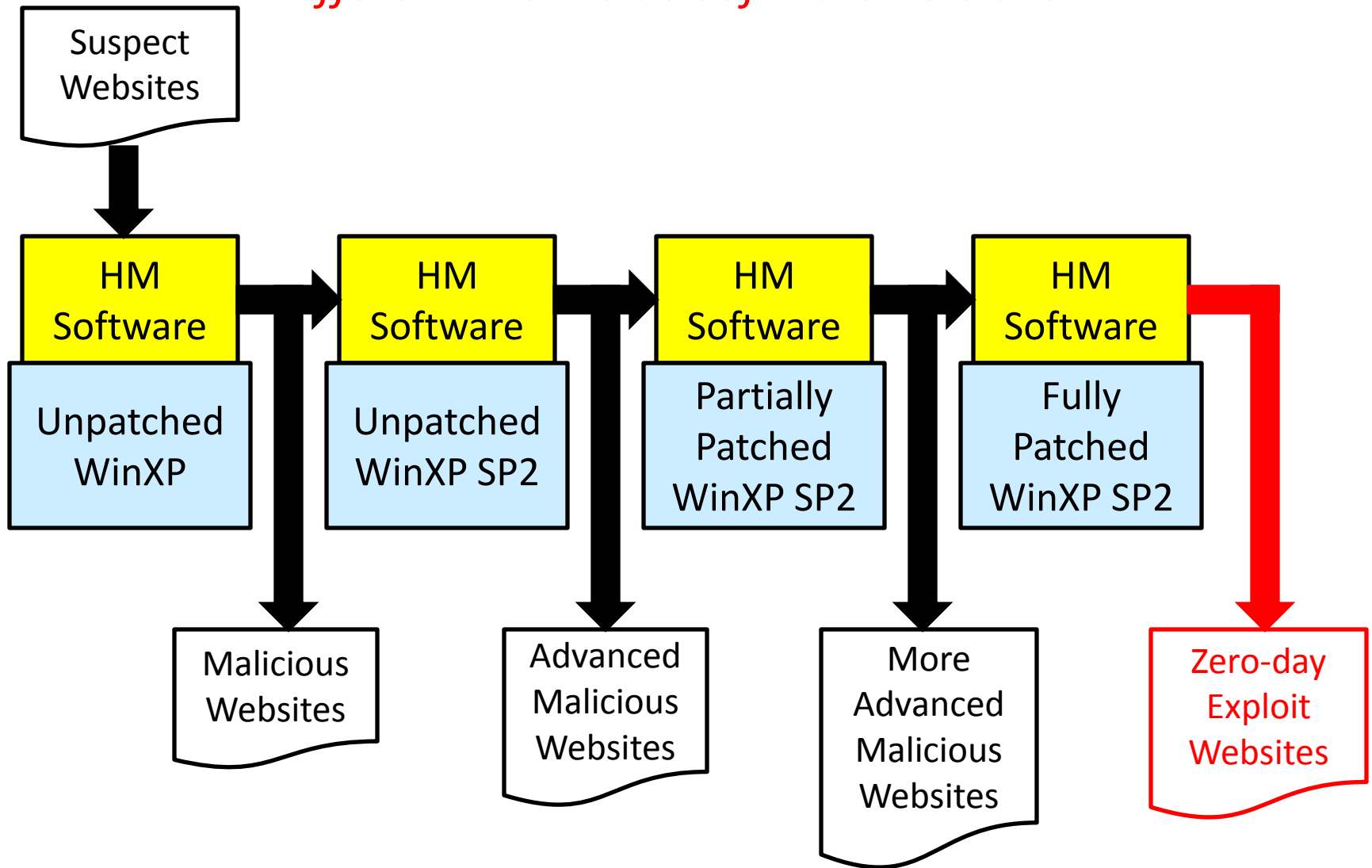
- Behavior-based detection: *detect software installation following a successful vulnerability exploit*
 - Wait a couple of minutes and check for files created outside IE sandbox
 - Strider systems management approach to cybersecurity
- Can capture both known-vulnerability and zero-day exploits
- Use 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



The HoneyMonkey Pipeline

Different Client-side Software Versions



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/>
 - “Detecting Stealth Web Pages That Use Click-Through Cloaking,” MSR-TR-2006-178, December 2006
 - <http://research.microsoft.com/research/pubs/view.aspx?type=Technical%20Report&id=1224>
 - “Spam Double-Funnel: Connecting Web Spammers with Advertisers,” in *Proc. WWW*, May 2007
 - <http://research.microsoft.com/research/pubs/view.aspx?type=Technical%20Report&id=1269>
 - “A Quantitative Study of Forum Spamming Using Context-based Analysis,” in *Proc. NDSS*, February 2007
 - <http://research.microsoft.com/research/pubs/view.aspx?type=Technical%20Report&id=1219>
 - “Strider Search Ranger: Towards an Autonomic Anti-Spam Search Engine,” in *Proc. ICAC*, June 2007
 - http://research.microsoft.com/searchranger/ICAC2007_Autonomic_Anti-Spam_Search_Engines_camera-ready.doc
- **Strider HoneyMonkey** <http://research.microsoft.com/honeymonkey/>
 - “Automated Web Patrol with Strider HoneyMonkeys: Finding Web Sites That Exploit Browser Vulnerabilities,” in *Proc. NDSS*, 2006
 - http://research.microsoft.com/honeymonkey/NDSS_2006_HoneyMonkey_Wang_Y_camera-ready.pdf
- **Strider Typo-Patrol** <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>

BACKUP

Spam Double-Funnel

