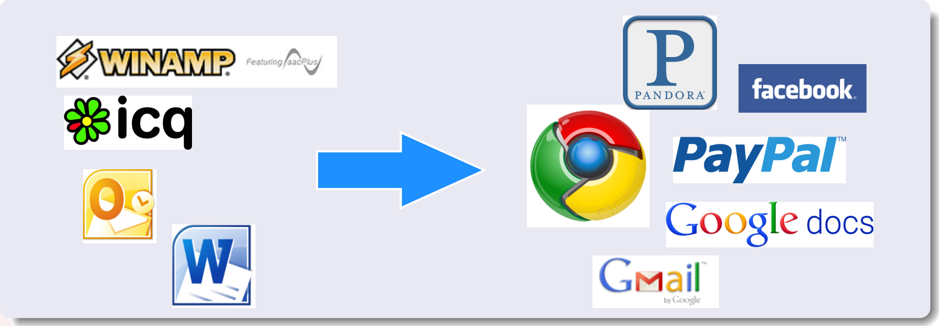
WebPatrol: Automated Collection and Replay of Web-based Malware Scenarios

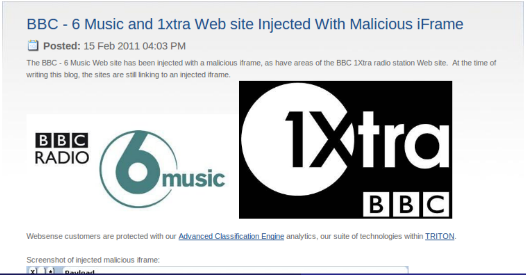
Kevin Zhijie Chen[1,2] (kevinchn@cs.berkeley.edu)
Guofei Gu[3] Jianwei Zhuge[4] Jose Nazario[5] Xinhui Han[1]

[1]Institute of Computer Science and Technology, Peking University
[2]UC Berkeley [3]Texas A&M University
[4]Network Center, Tsinghua University [5]Arbor Networks

March 22

New Era, New Threat





BBC - 6 Music and 1xtra Web site Injected With Malicious iFrame

Posted: 15 Feb 2011 04:03 PM

The BBC - 6 Music Web site has been injected with a malicious iframe, as have areas of the BBC 1Xtra radio station Web site. At the time of writing this blog, the sites are still linking to an injected iframe.

Websense customers are protected with our Advanced Classification Engine analytics, our suite of technologies within TRITON.

Screenshot of injected malicious iframe:

- Web-based services
- Browser-centric
- Web-based malware

What’s web-based malware?

A web-based malware is...

- A program in a browser
- Written in HTML, JavaScript, etc
- An exploit of browser vulnerabilities
- Cause of a drive-by download

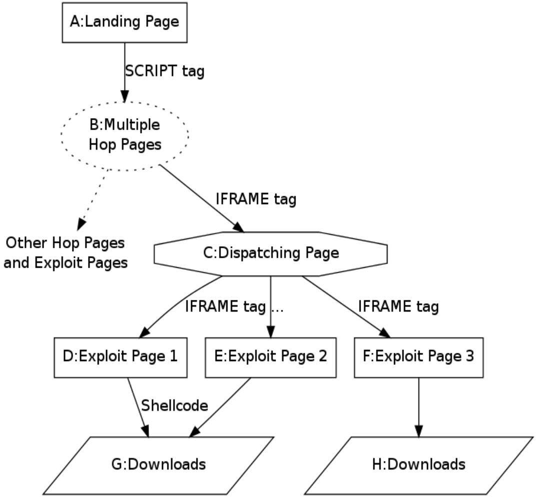
NeoTracePro 3.25 ActiveX Control ”TraceTarget()” b0f [NeoTraceExplorer.dll]

```
<object classid="clsid:3E1DD897-F300-486C-BEAF-711183773554"
 id="NeoTracePro"></object>
<script> ...
 while(Target.length < PwnEIP) Target += "\x0C";
 NeoTracePro.TraceTarget(Target); </script>
```

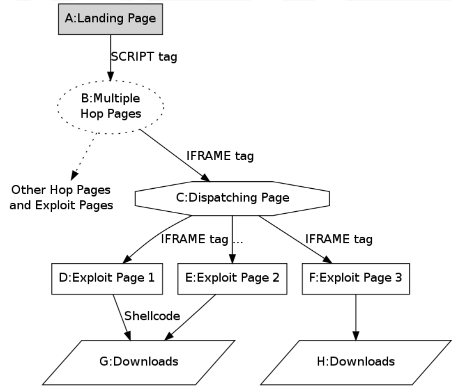
But...

Exploitation is not all

Before the exploitation happens



Before the exploitation happens



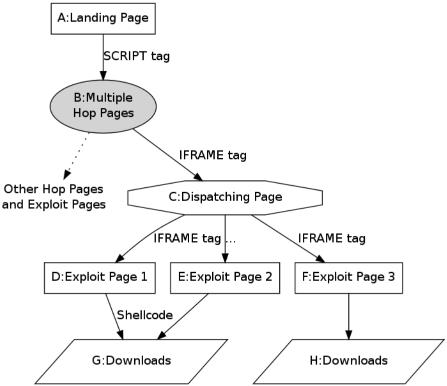
A: Landing page

Websites with large traffic

`http://www.bbc.co.uk/6music/`,

`http://www.pku.edu.cn`

Before the exploitation happens



A:Landing Page

SCRIPT tag

B:Multiple Hop Pages

Other Hop Pages and Exploit Pages

IFRAME tag

C:Dispatching Page

IFRAME tag ...

IFRAME tag

D:Exploit Page 1

E:Exploit Page 2

F:Exploit Page 3

Shellcode

G:Downloads

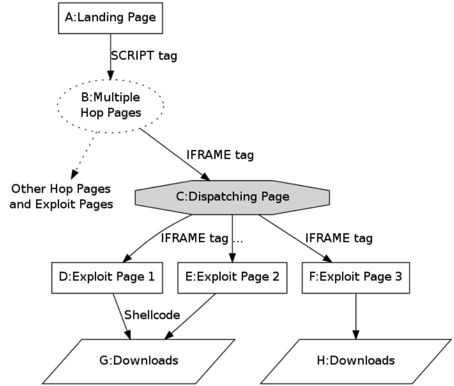
H:Downloads

B: Hop pages

Inline linking through widgets,AD networks, etc

Type	Example
HTML Tags	<iframe src = “foo.html”></iframe>, or script, object, img ...
JS/VB API	clientXmlHttpRequest.open(”GET”, ”test.txt”, true);
Plugins	Com.DloadDS(”http://www.***.com/calc.cab”,”muma.exe”,0);
Shellcodes	URLDownloadToFile(0,”http://foo.com/calc.exe”,”calc.exe”,0,0);

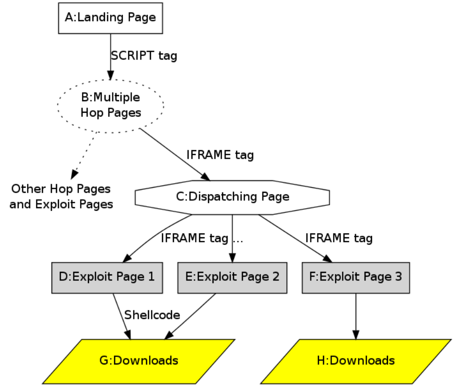
Before the exploitation happens



C: Dispatching pages

Fingerprinting the browser and its plugins, and exposing exploits accordingly.

Before the exploitation happens



D-F: Exploit pages

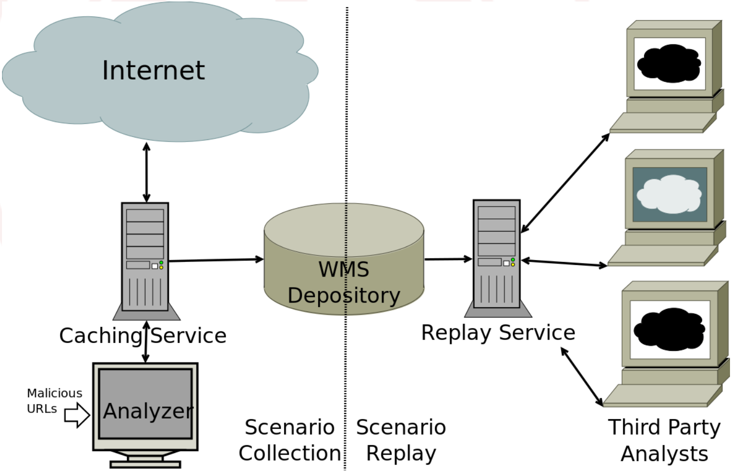
Heap spraying, Buffer overflow, use-after-free, English shellcode, etc.

G,H: Downloads

Download&Exec, joining botnets, stealing information etc.

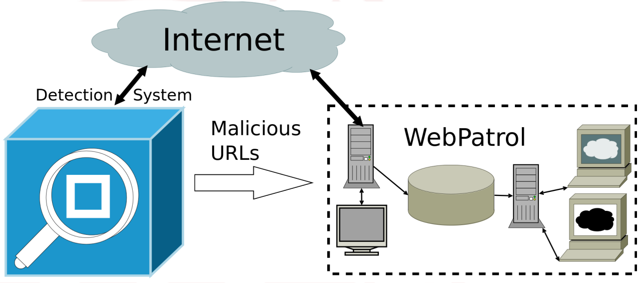
System Architecture

Goal: Collect all malicious webpages inline-linked in the landing page at a certain time (called a *web-based malware scenario, WMS*) and replay them faithfully for different kinds of clients.



Analysis based on replay and System evaluation

Data Collection



- **Time period:** Jan. - May, 2010
- **Data source:**
  - a crawler + *high-interaction* client honeypots
  - ˜ 35,000 websites from CERNET (China Education and Research Network, mostly *.edu.cn)
- **Client environment:** IE(6.0, 7.0), Adobe Reader(8,9), Flash Player, Storm Player etc. on WinXP (SP1, SP2)
- (List of malicious URLs) → WebPatrol

Basic Statistics

26,498 malicious scenarios from 1,248 distinct landing sites.(3.52%)

Google Safe Browsing only labeled 295 landing sites

23.2 days of average lasting time with the longest 132 days

Exploit hosting site changes 4.82 times on average during lifetime

Hot spot for web-based malware but no enough attention received.

Replay Analysis: Vulnerability Trend

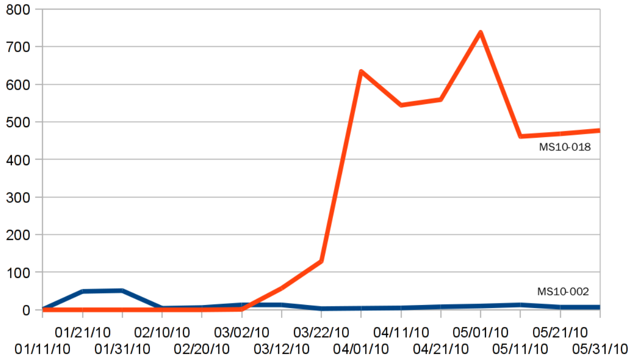
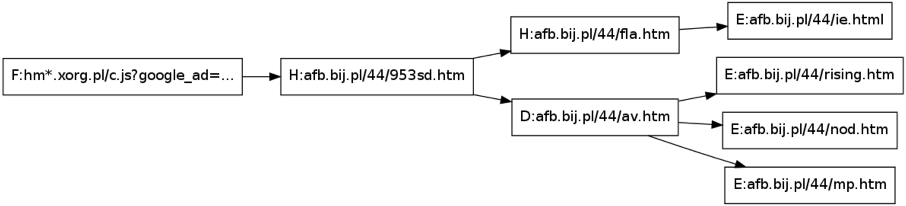


Figure: Number of new WMSs with MS10-002 “Aurora” and MS10-018 Exploits

Also the exploits evolves: 7 variants of MS10-018 exploits found since March 11, 2010 with increasing levels of obfuscation and optimization (for better successful rate)

Replay Analysis: Scenario Evolution

Mar.11:



F:hm*.xorg.pl/c.js?google_ad=...

H:afb.bij.pl/44/953sd.htm

H:afb.bij.pl/44/fla.htm

D:afb.bij.pl/44/av.htm

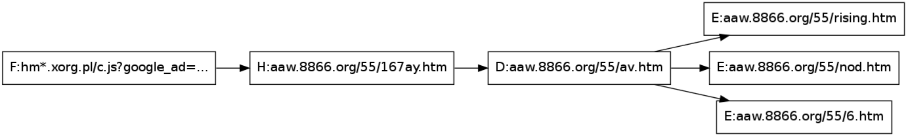
E:afb.bij.pl/44/ie.html

E:afb.bij.pl/44/rising.htm

E:afb.bij.pl/44/nod.htm

E:afb.bij.pl/44/mp.htm

Apr.25:



F:hm*.xorg.pl/c.js?google_ad=...

H:aaw.8866.org/55/167ay.htm

D:aaw.8866.org/55/av.htm

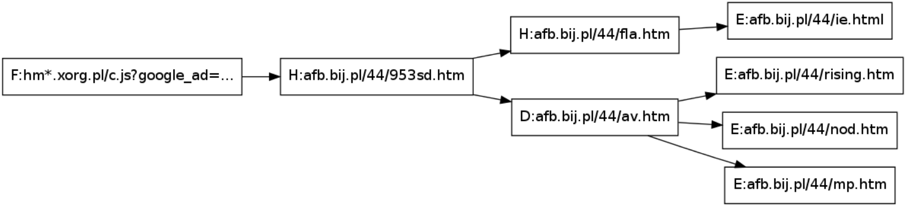
E:aaw.8866.org/55/rising.htm

E:aaw.8866.org/55/nod.htm

E:aaw.8866.org/55/6.htm

Replay Analysis: Scenario Evolution

Mar.11:



F:hm*.xorg.pl/c.js?google_ad=...

H:afb.bij.pl/44/953sd.htm

H:afb.bij.pl/44/fla.htm

D:afb.bij.pl/44/av.htm

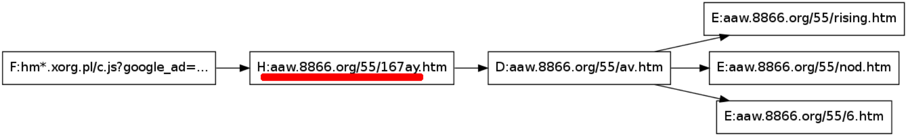
E:afb.bij.pl/44/ie.html

E:afb.bij.pl/44/rising.htm

E:afb.bij.pl/44/nod.htm

E:afb.bij.pl/44/mp.htm

Apr.25:



F:hm*.xorg.pl/c.js?google_ad=...

H:aaw.8866.org/55/167ay.htm

D:aaw.8866.org/55/av.htm

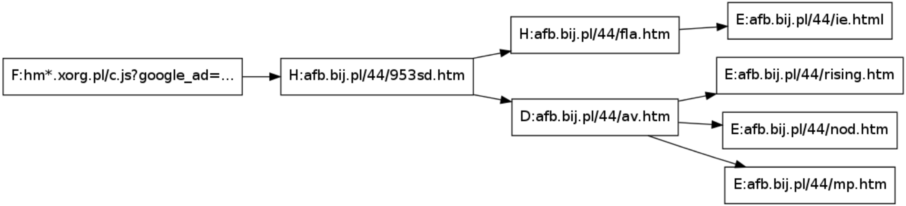
E:aaw.8866.org/55/rising.htm

E:aaw.8866.org/55/nod.htm

E:aaw.8866.org/55/6.htm

Replay Analysis: Scenario Evolution

Mar.11:



F:hm*.xorg.pl/c.js?google_ad=...

H:afb.bij.pl/44/953sd.htm

H:afb.bij.pl/44/fla.htm

D:afb.bij.pl/44/av.htm

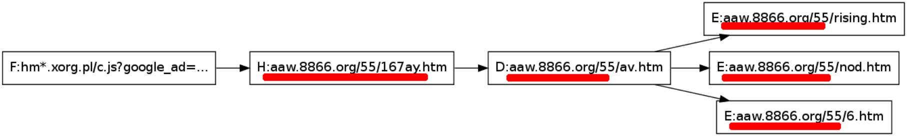
E:afb.bij.pl/44/ie.html

E:afb.bij.pl/44/rising.htm

E:afb.bij.pl/44/nod.htm

E:afb.bij.pl/44/mp.htm

Apr.25:



F:hm*.xorg.pl/c.js?google_ad=...

H:aaw.8866.org/55/167ay.htm

D:aaw.8866.org/55/av.htm

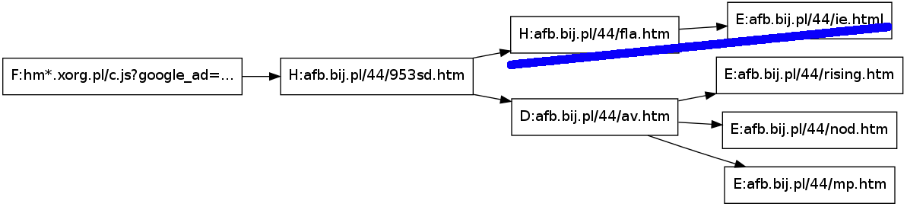
E:aaw.8866.org/55/rising.htm

E:aaw.8866.org/55/nod.htm

E:aaw.8866.org/55/6.htm

Replay Analysis: Scenario Evolution

Mar.11:



F:hm*.xorg.pl/c.js?google_ad=...

H:afb.bij.pl/44/953sd.htm

H:afb.bij.pl/44/fla.htm

D:afb.bij.pl/44/av.htm

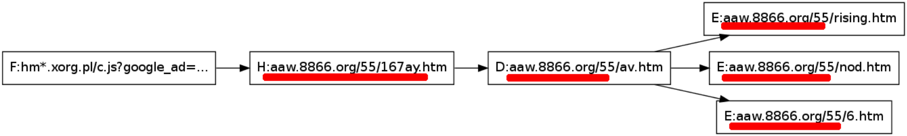
E:afb.bij.pl/44/ie.html

E:afb.bij.pl/44/rising.htm

E:afb.bij.pl/44/nod.htm

E:afb.bij.pl/44/mp.htm

Apr.25:



F:hm*.xorg.pl/c.js?google_ad=...

H:aaw.8866.org/55/167ay.htm

D:aaw.8866.org/55/av.htm

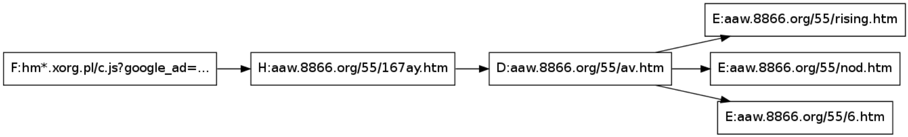
E:aaw.8866.org/55/rising.htm

E:aaw.8866.org/55/nod.htm

E:aaw.8866.org/55/6.htm

Replay Analysis: Scenario Evolution II

Apr.25:



F:hm*.xorg.pl/c.js?google_ad=...

H:aaw.8866.org/55/167ay.htm

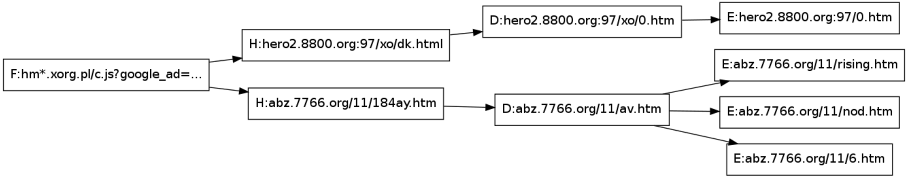
D:aaw.8866.org/55/av.htm

E:aaw.8866.org/55/rising.htm

E:aaw.8866.org/55/nod.htm

E:aaw.8866.org/55/6.htm

May.04:



F:hm*.xorg.pl/c.js?google_ad=...

H:hero2.8800.org:97/xo/dk.html

D:hero2.8800.org:97/xo/0.htm

E:hero2.8800.org:97/0.htm

H:abz.7766.org/11/184ay.htm

D:abz.7766.org/11/av.htm

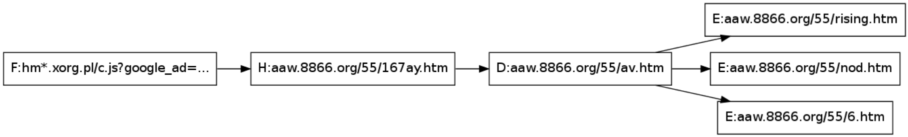
E:abz.7766.org/11/rising.htm

E:abz.7766.org/11/nod.htm

E:abz.7766.org/11/6.htm

Replay Analysis: Scenario Evolution II

Apr.25:



F:hm*.xorg.pl/c.js?google_ad=...

H:aaw.8866.org/55/167ay.htm

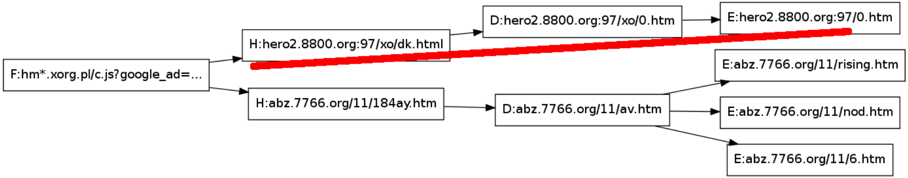
D:aaw.8866.org/55/av.htm

E:aaw.8866.org/55/rising.htm

E:aaw.8866.org/55/nod.htm

E:aaw.8866.org/55/6.htm

May.04:



F:hm*.xorg.pl/c.js?google_ad=...

H:hero2.8800.org:97/xo/dk.html

D:hero2.8800.org:97/xo/0.htm

E:hero2.8800.org:97/0.htm

H:abz.7766.org/11/184ay.htm

D:abz.7766.org/11/av.htm

E:abz.7766.org/11/rising.htm

E:abz.7766.org/11/nod.htm

E:abz.7766.org/11/6.htm

Scenario Collection

The Analyzer: Improved PHoneyC

- *Low-interaction* client honeypot (browser emulation)
- Key modules: HTML Parser, JS engine and plugin emulation
- Our improvement:
    - True DOM support
    - mock ActiveXObject class for universal ActiveX support
    - Download URL extraction and shellcode detection through bytecode instrumentation

The Caching Service: wmPolipo

- Proxy-like caching solution
- Caches **all** traffic (ignores the *private/no-cache/no-store* fields)
- Once collected, never update

Scenario Replay

Replay Service:wmPolipo

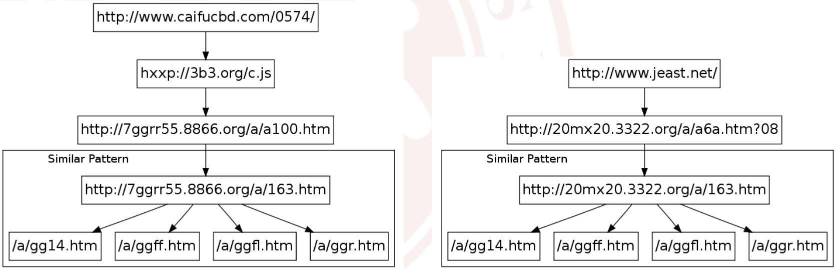
- isolated, works in an offline fashion
- aware of scenarios: URL + scenario ID is the primary key for a resource
- URL-similarity-check for generalizing *randomized URLs*
- Multi-user, multi-cache-directory support

Replay Client:

- All kind of third party tools & platforms
- Only requirement: proxy support

Web-based malware family

Web-based malware exploit kit: Scenarios sharing similar reference sub-graph and directory/file names.



http://www.caifucbd.com/0574/

hxxp://3b3.org/c.js

http://7ggrr55.8866.org/a/a100.htm

Similar Pattern

http://7ggrr55.8866.org/a/163.htm

/a/gg14.htm

/a/ggff.htm

/a/ggfl.htm

/a/ggr.htm

http://www.jeast.net/

http://20mx20.3322.org/a/a6a.htm?08

Similar Pattern

http://20mx20.3322.org/a/163.htm

/a/gg14.htm

/a/ggff.htm

/a/ggfl.htm

/a/ggr.htm

(a) www.caifucbd.com

(b) www.jeast.net

Figure: Infection/Reference Graph of Two Different Scenarios

Pre-Evaluation

Kit ID	Pattern Description	Cnt.	P_i
1	MS10-018_0_htm	713	35.6%
2	MS10-018_xo_dk_html	438	21.9%
3	av_htm_6_7_htm	177	8.8%
4	wm_multiple_pages	124	6.2%
5	wm_IE_html	82	4.1%
6	av_htm_mp_htm	82	4.1%
7	01_x01_htm_jk.htm	75	3.8%
8	GV_hk_series	39	1.9%
9	apt_spa_chinese	36	1.8%
10	index_5_htm	24	1.2%
11	index_nivea_htm	18	0.9%
12	xc15_15index_htm	13	0.7%
13	other	179	9.0%
	total	2000	100%

Table: Percentage of each family in randomly selected 2000 samples

Evaluation: Completeness

Initial Site	KID	All	**WP**	N_{KID}^{WP}	**PHC**	N_{KID}^{PHC}	**HPC**	N_{KID}^{HPC}
dj.csuft.edu.cn	1	5	4	0.80	3	0.60	3	0.60
ecls.ynu.edu.cn	2	5	4	0.80	2	0.40	4	0.80
student.fzu.edu.cn	5	3	2	0.67	2	0.67	3	1.00
ebm.lzu.edu.cn	4	8	8	1.00	3	0.38	4	0.50
cheds.pku.edu.cn	7	7	6	0.86	7	1.00	6	0.86
xsc.ruc.edu.cn	6	14	13	0.93	3	0.21	13	0.93
btzy.nm.edu.cn	8	23	20	0.87	3	0.13	20	0.87
rwxy.zjut.edu.cn	12	3	2	0.67	2	0.67	3	1.00
psy.ntu.edu.cn	3	16	12	0.75	2	0.13	2	0.13
ecls.ynu.edu.cn	9	6	5	0.83	2	0.33	4	0.67
xlzx.sdu.edu.cn	10	21	17	0.81	3	0.14	2	0.10
art.dufe.edu.cn	11	7	6	0.86	3	0.43	5	0.71
ecls.ynu.edu.cn	13	5	4	0.80	2	0.40	2	0.40
abc.hznu.edu.cn	13	6	5	0.83	4	0.67	5	0.83
jwc.sdjzu.edu.cn	13	3	3	1.00	1	0.33	3	1.00
total	-	132	**119**		**42**		**79**	
C	-	100%	**81.9%**		**47.1%**		**65.3%**	

$$C = \sum_{i=1}^{13} (N_i * P_i)$$

Cause of missing nodes

1. Out-going links in different branches

```
if(navigator.userAgent.toLowerCase().indexOf("msie")>0)
{document.write("<EMBED src=iie.swf width=0 height=0>");}
else
{document.write("<EMBED src=fff.swf width=0 height=0>");}
```

2. Limitation of the shellcode detection & emulation module (libemu)
3. Different implementations of the parser and the script engine

Future Work

- Static analysis
- Run the analyzer multiple times with different configurations
- Improvement on libemu (more system API support)

Thank you!

Questions?

kevinchn@cs.berkeley.edu

Attacks on WebPatrol

Sandbox Detection

- All kinds of plugins coexisting
- Incomplete browser emulation
- Different implementations for a specification

but implementing new browser features is fast and easy (adding some Python module).

Other attacks

- DoS attacks
- Vulnerability attacks (e.g. against spidermonkey or PySGML Parser)

Our approach

Automated collection and analysis of web-based malware

	Current practice	Our approach
Collection	1. The downloaded malware binaries, 2. Individual malicious web pages	The complete set of web pages
Analysis	Exploits, downloads, etc.	Complete scenario

Web-base client software exploiting

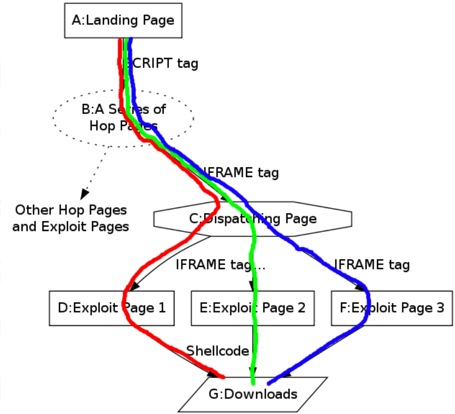
Compared to traditional server-side exploiting malware, web-based malware has the following characteristics.

First, it exploits client-side vulnerabilities, mostly in modern complex browsers and their extensions. Thus, it is more stealthy and evasive because it does not need to send aggressive scanning traffic.

Second, it is pervasive considering the large base of insecure web sites/pages on the Internet.

Finally, it is hard to block because most networks allow web traffic.

Infection Trails and Scenarios



Infection trail

We define a *web infection trail* as a directed path in the graph, starting from μ to some sink node in T.

Web-based malware scenario (WMS)

A directed graph (μ, V, E, T), where

- μ : the initial landing URL, $\mu \in V$
- V : Nodes (pages & other resources)
- E : Edges *(outgoing links)*
- T : Sink nodes ($T \subset V$) *(each indicates a successful web infection/exploitation)*

Malicious pages hosting domains

Domain Name	Registrant	No. of Inject Sites
8800.org	Yaako Ltd.	610
6600.org	Yaako Ltd.	475
3322.org	Yaako Ltd.	255
lookforhosting.com	GoDaddy.com	255
9966.org	Yaako Ltd.	163
caipiaoyuce.info	Yue You	157
chinawordpress.info	Yue You	129
cptiandi.com	Melbourne IT	118
8866.org	Yaako Ltd.	110
2288.org	Yaako Ltd.	54

Table: Top 10 Malicious Hosting Domains Discovered during the Measurement of WMS on CERNET

Replay Analysis: Scenario Evolution II

Date	Hop Pages	Exploit Pages
03.11	*First Hop Pages:* 1.hm*.xorg.pl/c.js?google_ad=... → 2 *Following Hop Pages:* 2.afb.bij.pl/44/953sd.htm → 3,4 3.afb.bij.pl/44/fla.htm → 8 *Dispatching Page:* 4.afb.bij.pl/44/av.htm → 5,6,7	5.afb.bij.pl/44/rising.htm 6.afb.bij.pl/44/nod.htm 7.afb.bij.pl/44/mp.htm ... 8.afb.bij.pl/44/ie.html
04.25	*First Hop Pages:* 1.hm*.xorg.pl/c.js?google_ad=... → 2 *Following Hop Pages:* 2.aaw.8866.org/55/167ay.htm → 4 *Dispatching Page:* 4.aaw.8866.org/55/av.htm → 5,6,7	5.aaw.8866.org/55/rising.htm 6.aaw.8866.org/55/nod.htm 7.aaw.8866.org/55/6.htm ...

Table: Scenario Evolution on cc.njarti.edu.cn/

Replay Analysis: Scenario Evolution III

Date	Hop Pages	Exploit Pages
04.25	*First Hop Pages:* 1.hm*.xorg.pl/c.js?google_ad=... → 2 *Following Hop Pages:* 2.aaw.8866.org/55/167ay.htm → 4 *Dispatching Page:* 4.aaw.8866.org/55/av.htm → 5,6,7	5.aaw.8866.org/55/rising.htm 6.aaw.8866.org/55/nod.htm 7.aaw.8866.org/55/6.htm ...
05.04	*First Hop Pages:* 1.hm*.xorg.pl/c.js?google_ad=... → 2,3 *Following Hop Pages:* 2.abz.7766.org/11/184ay.htm → 4 3.hero2.8800.org:97/xo/dk.html → 8 *Dispatching Page:* 4.abz.7766.org/11/av.htm → 5,6,7 8.hero2.8800.org:97/xo/0.htm → 9	5.abz.7766.org/11/rising.htm 6.abz.7766.org/11/nod.htm 7.aaw.8866.org/55/6.htm 9.hero2.8800.org:97/0.htm ...

Table: Scenario Evolution on cc.njarti.edu.cn

Table of Contents