Quantifying Maliciousness in Alexa Top-Ranked Domains

Paul Royal

Barracuda Labs

BARRACUDA LABS

Agenda

- Drive-by Downloads (DDLs)
  - Definition, distribution
- Quantifying Maliciousness
  - Motivations, design approach
- Experimentation
  - System specification, operation
  - Estimating impact
- Analysis
  - Case studies, screenshots
- Conclusion

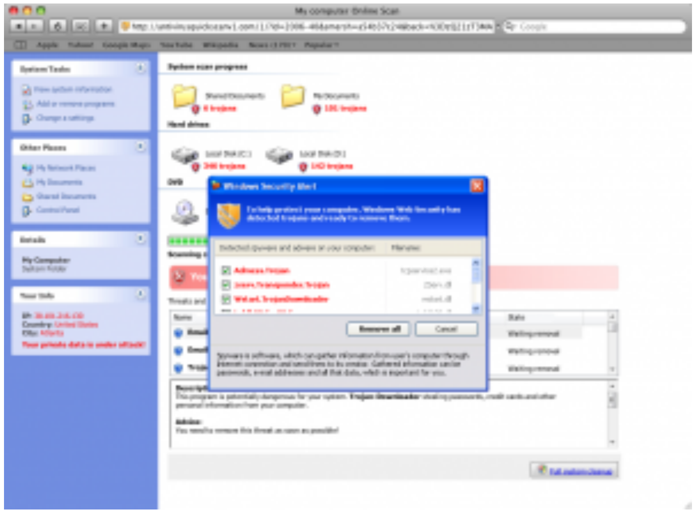
Drive-by Downloads (DDLs)

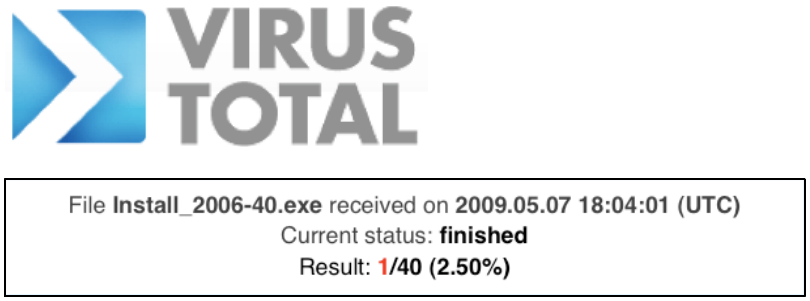
Drive-by Download Definition

- An attack wherein malicious content is served to the web browser or its plugins
  - Intended to occur without user’s knowledge
  - If successful, results in arbitrary code execution
    - Executed code retrieves payload (e.g., malware binary)
- Facilitating a drive-by download
  - Email (e.g., links in fake airline ticket messages)
  - Search Engine Optimization (malicious websites in search results)
  - Compromising a popular, legitimate website

Website Compromise Examples

- USAToday.com ad network compromised in May 2009
- Ad for Roxio Creator 2009 bundled with malicious javascript
  - Code activated without hovering over or clicking on ad
  - Redirected users to Rogue AV website




VIRUS TOTAL

File **Install_2006-40.exe** received on **2009.05.07 18:04:01 (UTC)**
Current status: **finished**
Result: **1/40 (2.50%)**

Examples Cont’d

- PBS.org compromised in September 2009
  - Curious George section served visitors malicious javascript
  - Javascript iframed into exploit site
    - Exploit site targeted browser plugins (e.g., Acrobat Reader via CVE-2008-2992, CVE-2009-0927, and CVE-2007-5659, Apple QuickTime via CVE-2007-0015)
  - Compromised systems were used to build a botnet that was subsequently rented out by cyber criminals
    - “Send a message to ICQ #559156803; stats available under ststst02.“

Examples Cont'd

- Amnesty International UK website compromised in December 2011
- Malicious javascript inserted into front page
  - Iframed into exploit site that targets Java web plugin (CVE-2011-3544)
  - Payload contained properties of targeted malware
    - Campaign likely created by nation-state to spy on human rights activists

Quantifying Maliciousness

Motivations

- Drive-by downloads are one of the most popular ways to get malware onto systems
- Need a way to begin systematically quantifying the prevalence of the problem
  - Identification of maliciousness should be as generic as possible
- Measurement methodologies should be transparent and reproducible

Sourcing Websites

- Given their reach, we decided to collect daily lists of top-ranked sites
- For our initial broad study, used a source that generalizes popularity to the greatest extent possible
  - Some bias (e.g., popularity according to a given country) still inevitable

Detecting Maliciousness

- Given the breadth of coverage offered, we decided to employ a blackbox approach for identifying maliciousness
  - With a blackbox approach, knowledge of an event's occurrence is prioritized
    - Removes dependence on prior knowledge of specific vulnerabilities and exploits
- Blackbox measurement can be coupled with post-experimentation whitebox analysis of results to achieve depth of knowledge

Detecting Maliciousness Cont’d

- Our blackbox experimentation approach leveraged heavyweight virtualization
  - Created a virtual machine (VM) with ubiquitously targeted software components
  - Constructed automated system that executed many VMs simultaneously
    - Browser within each VM forced to visit a website
    - Network-level behaviors of the VM recorded
    - Drive-by downloads heuristically identified
  - Manual, post-experimentation whitebox analysis used to confirm maliciousness/remove false positives

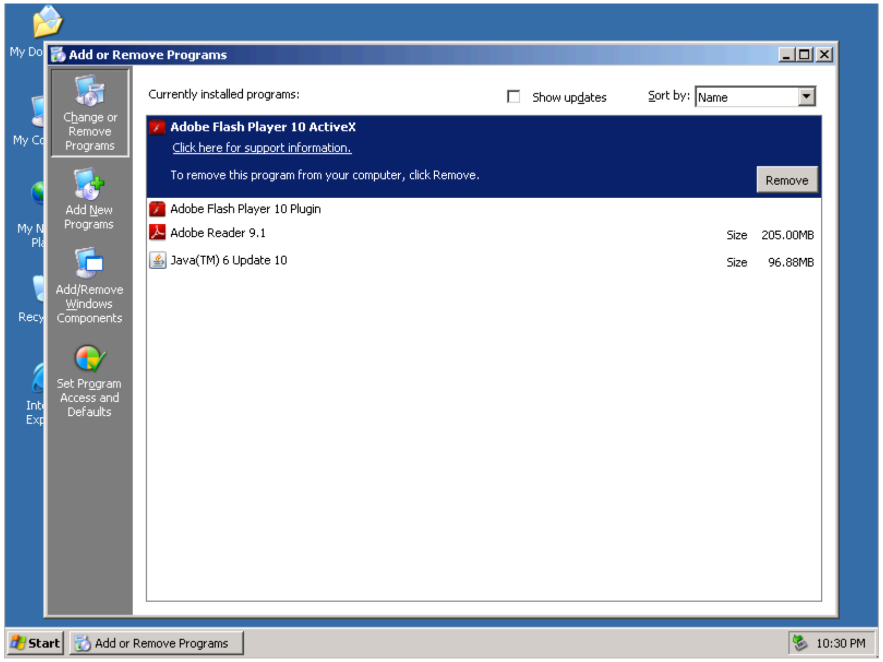
Experimentation

System Specification

- Input Source
  - Daily list of Alexa top 25,000 websites
    - Domains only (no path elements)
- URL Processing Node (1U)
  - Server that will process URLs by executing many virtual machines simultaneously
  - SuperMicro system with 24 cores and 32GB memory
    - Debian Linux and KVM virtualization container
- Database Node (2U)
  - Runs database software and houses session artifacts (e.g., DDL session packet capture files)
  - SuperMicro system with 8 cores, 8GB of memory and six disks
    - Debian Linux and PostgreSQL

Virtual Machine Configuration

- Windows XP SP2
  - No additional patches
- Internet Explorer 6
  - Acrobat Reader 9.1
  - Flash Player 10.0
  - Java 1.6 web plugin



Add or Remove Programs
Change or Remove Programs
Add New Programs
Add/Remove Windows Components
Set Program Access and Defaults
Remove
Start
Add or Remove Programs
10:30 PM

Currently installed programs: □ Show updates Sort by: Name

Adobe Flash Player 10 ActiveX Click here for support information. To remove this program from your computer, click Remove.		
Adobe Flash Player 10 Plugin		
Adobe Reader 9.1	Size	205.00MB
Java(TM) 6 Update 10	Size	96.88MB

System Operation

- On the processing node, a process is instantiated that spawns a series of threads
- Each thread continuously does the following
  - Queries the database for an unprocessed URL
    - Row-level locking used to manage concurrency
  - Starts a sterile, isolated VM that is used to process the URL
    - Begins recording VM network traffic just before VM invocation
    - A bootstrap script inside the VM accesses the URL and forces a browser to visit it
  - Allows the VM to execute for a short period of time
    - Enough time for the browser to visit the URL and potentially get compromised
  - Terminates the VM, then examines network traffic to heuristically determine whether a drive-by download occurred

Heuristic DDL Identification

- Looked for the following attributes in a single ethernet frame
  - MZ header, PE header, and one or more string attributes (e.g., “This program”, “DOS”)
- Would normally result in lots of false positives
  - However, given the input source (domains without path), very effective
  - February 2012
    - Two false positives
      - Both of these served malware, but via social vectors
  - May 2012
    - No false positives

Estimating Impact

- For each DDL site, we needed to conservatively estimate affected users
- Alexa publishes the popularity of a site as a percent of all visits
  - To derive the hard number, we leveraged a popular website’s visitor statistics
    - For example, in February 2012, Wikipedia recorded 15.756 billion views, which comprised 0.5416% of total Alexa views
    - Working backward, Alexa estimates (15,756 * 1,000,000)/(29 * (0.5416/100)) = ~100.31 billion views each day
- Use Alexa-estimated views per user to determine affected users

Estimating Impact Cont'd

- For a set of affected users, we needed to conservatively estimate the subset that were successfully compromised
  - Used visitor statistics to exclude incompatible or exploit-resistant platforms (e.g., those using Chrome or Mac OS X)
    - Narrows prospective candidates to 50.81% of total
- Then, we leveraged Java's status as the most popular mechanism of exploitation
  - 73% of users have the Java web plugin installed (Adobe)
  - 42% of those use a version vulnerable to exploitation (Qualys)
- Thus, as an initial conservative estimate, only 42% of 73% of 50.81%, or 15.57% of users served malicious content are likely to be successfully compromised

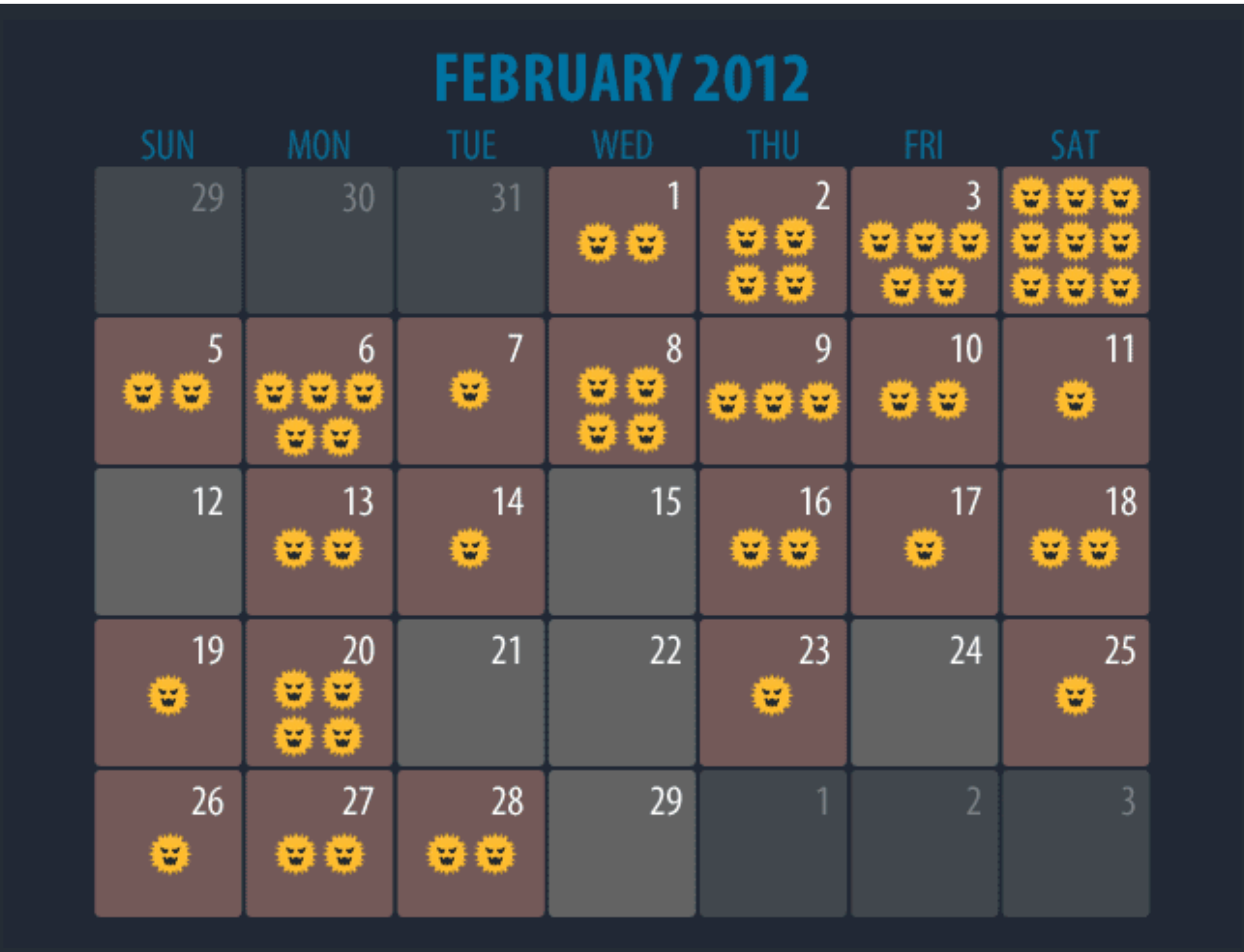
Analysis

BARRACUDA LABS

Case Study: February 2012

- Alexa top 25,000 domains were collected and analyzed each day
- When visited, 58 of these sites resulted in a drive-by download
  - Malicious content served by at least one top-ranked site 73% of the days in February
- Employing previously-described estimations
  - 10.541 million users served malicious content
  - 1.642 million users likely successfully compromised

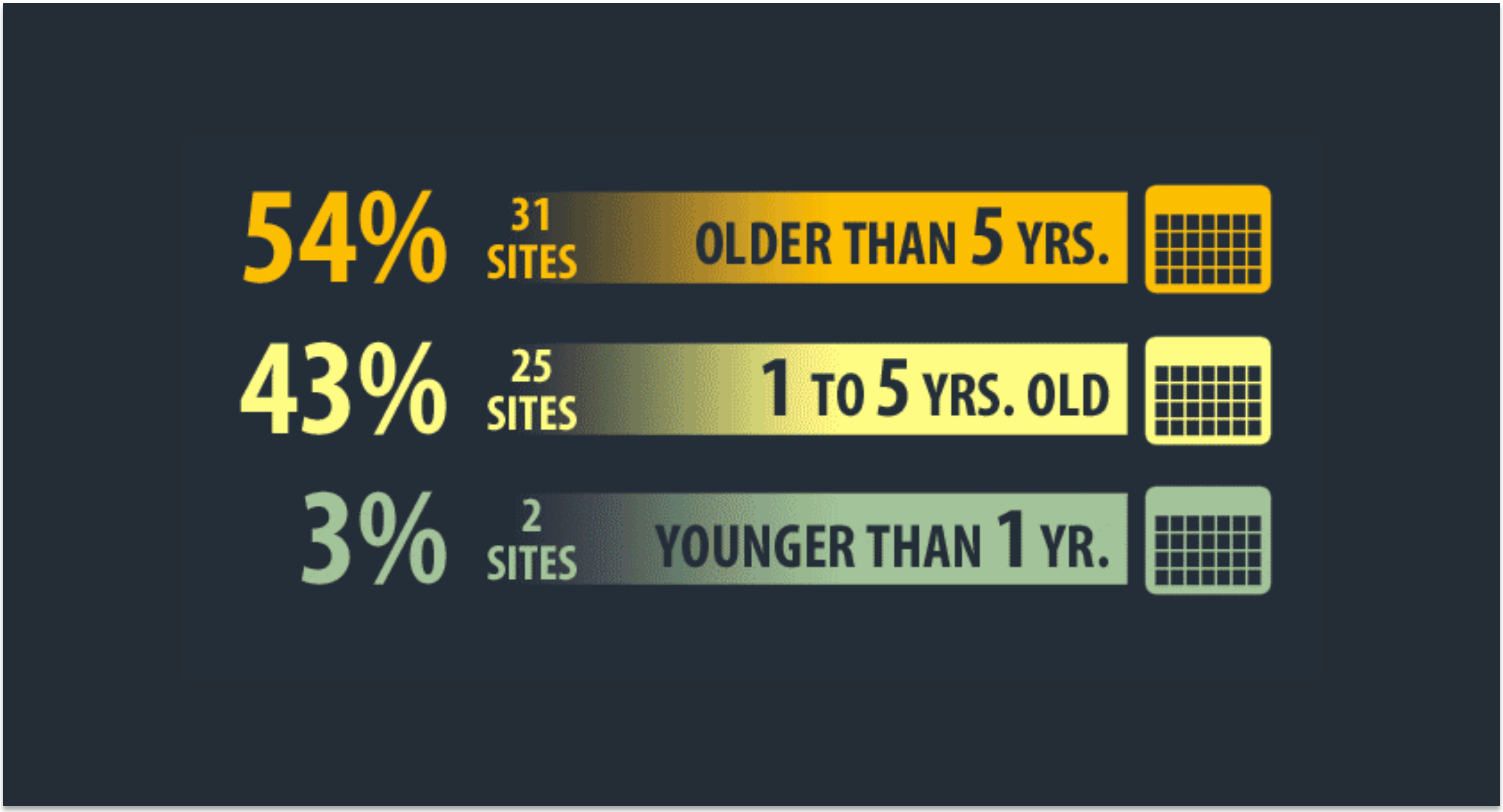
Top-Ranked Site DDL Calendar



FEBRUARY 2012

SUN	MON	TUE	WED	THU	FRI	SAT
29	30	31	1 ●●	2 ●●●●	3 ●●●●●	●●●●●●●●●
5 ●●	6 ●●●●●	7 ●	8 ●●●●	9 ●●●	10 ●●	11 ●
12	13 ●●	14 ●	15	16 ●●	17 ●	18 ●●
19 ●	20 ●●●●	21	22	23 ●	24	25 ●
26 ●	27 ●●	28 ●●	29	1	2	3

Top-Ranked DDL Site Age



54% 31 SITES OLDER THAN 5 YRS.

43% 25 SITES 1 TO 5 YRS. OLD

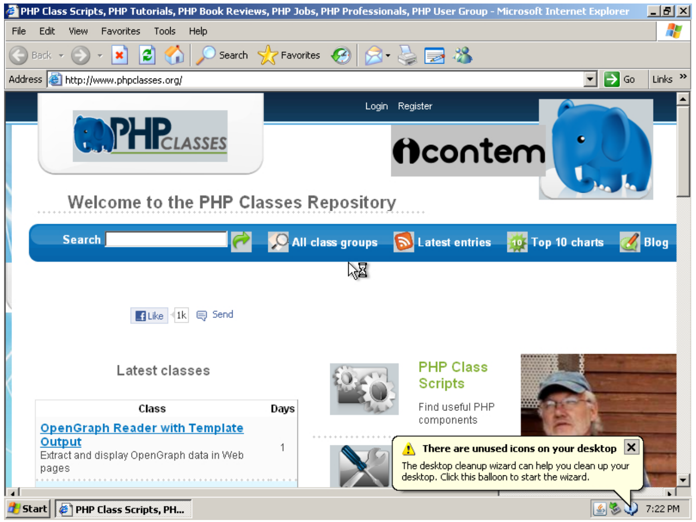
3% 2 SITES YOUNGER THAN 1 YR.

Top-Ranked DDL Sites in February 2012

Domain	Alexa Rank	DDL Served	Affected Views	Affected Users	Likely Compromised
free-tv-video-online[.]me	1,293	2/13/2012	5,366,895	745,402	116,121
bigresource[.]com	2,023	2/6/2012	1,243,916	894,903	139,411
myplaycity[.]com	2,823	2/1/2012	2,126,695	553,827	86,277
gaytube[.]com	3,190	2/3/2012	2,537,990	362,570	56,482
filmaffinity[.]com	3,228	2/1/2012	2,477,800	334,838	52,162
webconfs[.]com	3,684	2/6/2012	802,526	480,555	74,862
liilas[.]com	3,782	2/8/2012	2,437,674	243,767	37,975
peb[.]pl	3,832	2/25/2012	1,274,011	326,669	50,890
java2s[.]com	4,405	2/2/2012	842,653	374,512	58,343
gtbank[.]com	4,716	2/13/2012	1,916,032	319,339	49,748
pornrabbit[.]com	5,373	2/28/2012	772,432	292,588	45,580
fourhourworkweek[.]com	5,575	2/4/2012	642,021	298,614	46,519
feedage[.]com	6,374	2/2/2012	912,874	190,182	29,627
phpclasses[.]org	6,523	2/8/2012	892,811	212,574	33,116
abidjan[.]net	6,871	2/6/2012	782,463	217,351	33,860
hindilinks4u[.]net	7,946	2/19/2012	601,895	171,970	26,790
seeklogo[.]com	8,283	2/4/2012	782,463	170,101	26,499
studenti[.]it	10,213	2/6/2012	581,832	153,114	23,853
statshow[.]com	10,233	2/4/2012	541,705	193,466	30,139
seoforums[.]org	10,314	2/3/2012	581,832	149,188	23,241
wpbag[.]com	10,929	2/5/2012	732,305	107,692	16,777
quotationspage[.]com	10,964	2/9/2012	331,042	170,640	26,583
arabianbusiness[.]com	11,005	2/11/2012	591,863	128,666	20,044
mediafiremoviez[.]com	11,628	2/27/2012	601,895	139,976	21,806
…	…	…	…	…	…
			Totals	10,541,378	1,642,173

Screenshots for February 2012

- phpclasses[.]org
  - PHP developer help site
  - Alexa Rank 6,523
  - Served DDL February 8, 2012


PHP Class Scripts, PHP Tutorials, PHP Book Reviews, PHP Jobs, PHP Professionals, PHP User Group - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.phpclasses.org/
Go Links
Login Register
PHP CLASSES
contem
Search
All class groups
Latest entries
Top 10 charts
Blog
Like 1k Send
Start
PHP Class Scripts, PH...
7:22 PM

Welcome to the PHP Classes Repository

Latest classes

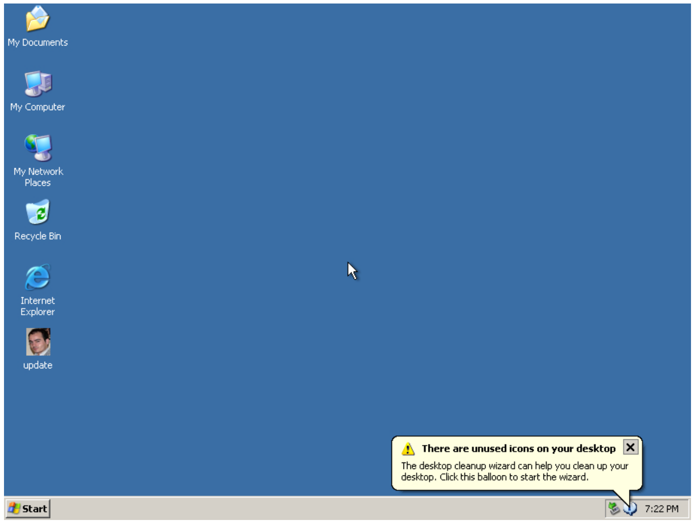
Class	Days
OpenGraph Reader with Template Output Extract and display OpenGraph data in Web pages	1

PHP Class Scripts

Find useful PHP components

There are unused icons on your desktop

The desktop cleanup wizard can help you clean up your desktop. Click this balloon to start the wizard.



There are unused icons on your desktop

The desktop cleanup wizard can help you clean up your desktop. Click this balloon to start the wizard.

Case Study: May 2012

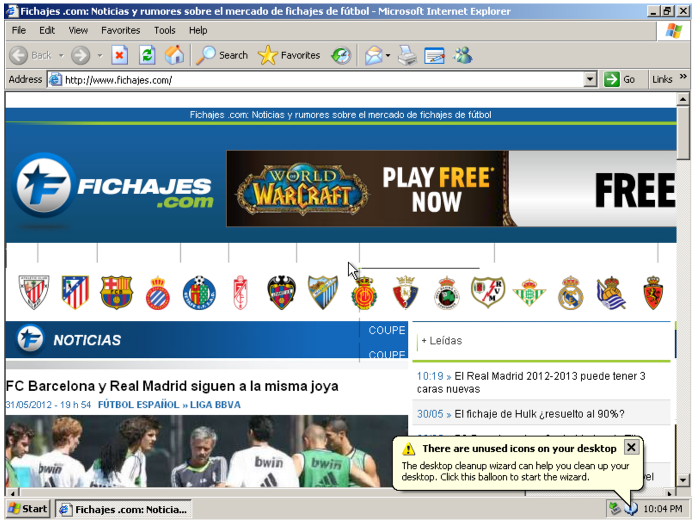
- When visited, 39 of the Alexa top 25,000 resulted in a drive-by download
  - Malicious content served by at least one site 84% of the days in May
  - 7.881 million users served malicious content
  - 1.228 million users likely successfully compromised
- For the May 2012 study, functionality was added to the system that examines recurring maliciousness
  - Most sites (72%) compromised for a single day, others for a week or more
  - Average period of compromise just over 36 hours

Top-Ranked DDL Sites in May 2012

Domain	Alexa Rank	First Served	Days Served	Affected Views	Affected Users	Likely Compromised
dealextreme.com	1,191	05/28/12	1	8,175,737	704,804	109,796
rlslog.net	1,703	05/08/12	4	5,774,427	1,178,455	183,584
funpatogh.com	3,313	05/20/12	1	1,895,968	390,921	60,899
iconarchive.com	3,370	05/24/12	1	2,467,768	304,662	47,461
heraldm.com	4,442	05/09/12	8	1,259,041	740,612	115,374
tehparadox.com	5,733	05/13/12	1	1,274,010	215,933	33,638
incgamers.com	6,033	05/18/12	1	591,863	197,287	30,734
pornrabbit.com	6,203	05/19/12	5	1,107,863	479,594	74,712
nulledscripts.it	7,414	05/31/12	1	112,353	92,854	14,465
larepublica.pe	7,874	05/19/12	1	431,357	196,071	30,544
goldesel.to	9,006	05/05/12	1	953,000	132,361	20,619
caclubindia.com	9,243	05/06/12	2	722,273	240,758	37,506
gabfirethemes.com	9,371	05/29/12	1	702,210	130,038	20,257
thedirty.com	10,503	05/30/12	1	423,332	132,291	20,608
aqori.com	10,749	05/27/12	1	480,512	57,893	9,018
bustnow.com	10,787	05/01/12	3	649,544	282,410	43,994
cssglobe.com	11,511	05/06/12	2	466,467	212,031	33,030
oneclickmoviez.com	12,510	05/13/12	1	491,547	104,584	16,292
iransalamat.com	14,532	05/18/12	3	431,858	226,104	35,223
mondespersistants.com	15,828	05/18/12	2	1,218,836	100,730	15,692
fotoflexer.com	16,051	05/26/12	1	238,751	119,375	18,596
xxvideo.us	16,859	05/27/12	1	213,672	101,748	15,850
goodinfohome.com	16,890	05/18/12	1	315,994	75,236	11,720
di.com.pl	17,576	05/14/12	1	236,745	91,055	14,184
…	…	…	…	…	…	…
				Totals	7,881,423	1,227,774

Screenshots for May 2012

- fichajes[.]com
  - Soccer news website
  - Alexa Rank 17,845
  - Served DDL May 31, 2012



Fichajes .com: Noticias y rumores sobre el mercado de fichajes de fútbol - Microsoft Internet Explorer

Address http://www.fichajes.com/

Fichajes .com: Noticias y rumores sobre el mercado de fichajes de fútbol

FC Barcelona y Real Madrid siguen a la misma joya

31/05/2012 - 19 h 54 FÚTBOL ESPAÑOL » LIGA BBVA

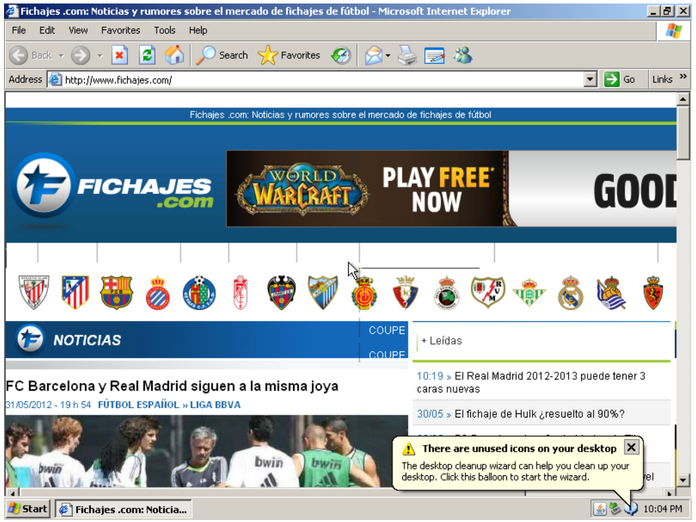
10:19 » El Real Madrid 2012-2013 puede tener 3 caras nuevas

30/05 » El fichaje de Hulk ¿resuelto al 90%?

There are unused icons on your desktop

The desktop cleanup wizard can help you clean up your desktop. Click this balloon to start the wizard.

File Edit View Favorites Tools Help
Back Search Favorites
Go Links
FICHAJES .com
WORLD WARCRAFT PLAY FREE NOW
FREE
NOTICIAS
COUPE
+ Leídas
bwin
Start
Fichajes .com: Noticia...
10:04 PM



Fichajes .com: Noticias y rumores sobre el mercado de fichajes de fútbol - Microsoft Internet Explorer

Address http://www.fichajes.com/

Fichajes .com: Noticias y rumores sobre el mercado de fichajes de fútbol

FC Barcelona y Real Madrid siguen a la misma joya

31/05/2012 - 19 h 54 FÚTBOL ESPAÑOL » LIGA BBVA

10:19 » El Real Madrid 2012-2013 puede tener 3 caras nuevas

30/05 » El fichaje de Hulk ¿resuelto al 90%?

There are unused icons on your desktop

The desktop cleanup wizard can help you clean up your desktop. Click this balloon to start the wizard.

File Edit View Favorites Tools Help
Back Search Favorites
Go Links
FICHAJES .com
WORLD WARCRAFT PLAY FREE NOW
GOOD
NOTICIAS
COUPE
+ Leídas
bwin
Start
Fichajes .com: Noticia...
10:04 PM

May 2012 DDL Properties

- Performed extensive whitebox analysis to measure additional attributes
  - Hypothesized that most DDLs for top-ranked sites would come from ad networks
    - Per analysis, only 46.1% of DDLs arrived via ad networks
      - More than half of were the result of direct website compromise
  - Use of Java in DDLs matched expectation
    - 87.1% of DDLs included one or more exploits for Java
      - Java in the browser should be disabled and only enabled when needed

Conclusion

- Most people assume that it is safe to visit popular, long-lived websites
- Multiple, month-long studies were conducted to systematically evaluate this intuition
- Results indicate that even the mainstream, popular web is not a safe place

Please fill out your feedback forms.

Questions?

DDL Site Details, Data
bit.ly/bhad12bn