Quantifying Maliciousness in Alexa Top-Ranked Domains

Paul Royal
Barracuda Labs

BARRACUDA LABS

Agenda

- Background
  - Drive-by downloads
- Quantifying Maliciousness
  - Motivation
- Experimentation Setup
- Analysis
  - Case study highlights
- Conclusion

Background

BARRACUDA LABS

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 campaign
  - Email (e.g., links referencing an online purchase)
  - Search Engine Optimization (malicious content linked from search results)
  - Compromising a popular, legitimate website

Website Compromise Examples

- RollingStone.com served visitors drive-by downloads in June 2013
- Redirections to malicious content via ad network (DoubleClick)
  - Browser, plugins served exploits from site backed by Sweet Orange exploit kit
  - ZeroAccess installed on successful compromise

Examples Cont'd

- PHP.net served drive-by downloads in October 2013
  - Redirections to malicious content the result of direct website compromise
  - Exploits served by Magnitude exploit kit
- Labs shared DDL PCAP to help community, site maintainers confirm details

Quantifying Maliciousness

Motivation

- Drive-by downloads are a popular way to propagate malicious software
- Want to better understand the extent of the problem
  - Measurement requires detection, which should be generic as possible
- Measurement approach should be transparent and reproducible

Scoping Measurement

- Scale of the problem space makes comprehensive measurement difficult
- Our experiments focused on maliciousness in top-ranked sites
  - Represents a subspace of tractable size and significant impact
- Elected to use an openly available list of popular websites
  - Wanted to go beyond country-centric vendor visibility

Detecting Maliciousness

- Sought to identify drive-by downloads in a vulnerability and exploit-independent manner
- Settled on a blackbox approach for identifying maliciousness
  - With a blackbox approach, knowledge of an event’s occurrence is prioritized
- Blackbox identification significantly reduced dependence on prior knowledge of specific vulnerabilities and exploits
- Post-experimentation whitebox analysis can be used to enhance granularity of knowledge

Detecting Maliciousness Cont’d

- Implementation of blackbox approach leveraged heavyweight virtualization
- Created a virtual machine (VM) with ubiquitously targeted software components and established identification process
  - Browser within the VM forced to visit a website
  - Network traffic of the visit is recorded
  - Drive-by downloads heuristically identified from traffic
- Engineered automation harness that operates many such VMs simultaneously
- Manual whitebox analysis used to confirm maliciousness/remove false positives

Experimentation Setup

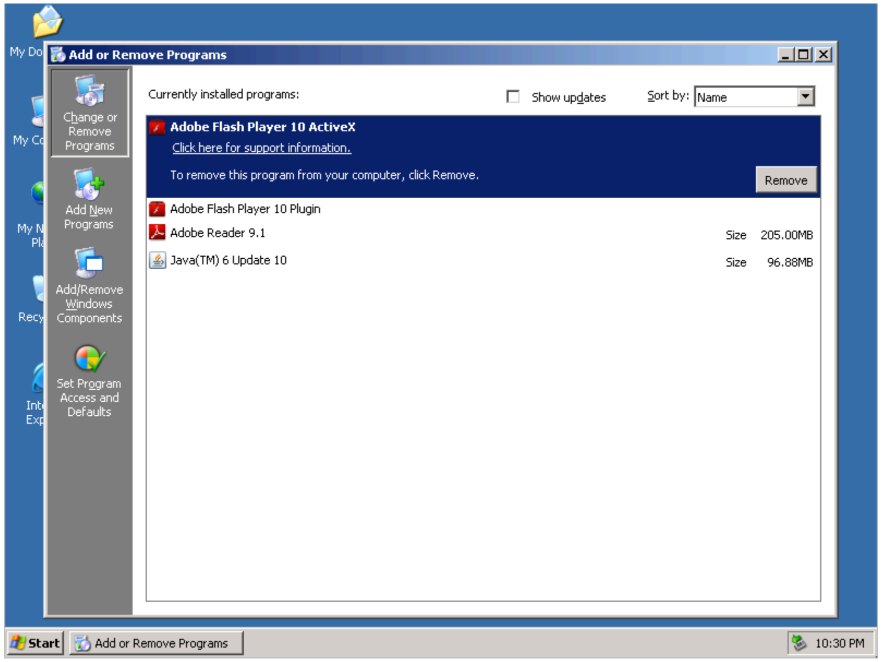
BARRACUDA LABS

Detection System Specification

- Input Source
  - Daily list of Alexa top 25,000 websites
    - Domains only (no path elements)
- Database Node (2U)
  - Houses system database and session artifacts (e.g., DDL packet capture files)
  - Runs Debian Linux and PostgreSQL
- URL Processing Node (1U)
  - Server that will process URLs by running many virtual machines at once
  - Runs Debian Linux and uses KVM virtualization container

Virtual Machine Configuration

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



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. 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 multi-threaded process is instantiated that spawns a series of threads
- Each thread continuously does the following
  - Obtains an unprocessed URL
  - Starts a sterile, isolated VM that will be used to process the URL
    - Network traffic recording begins just before VM invocation
    - A script inside the VM directs the browser to visit the URL
  - Permits 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 determine whether a drive-by download occurred

DDL Identification

- Employed simple detection heuristic
  - For a given network session, attempted to determine whether an executable was pushed to the VM
  - For example, looked for MZ header and PE header within a given ethernet frame
- For arbitrary HTTP traffic, would produce many false positives
  - Context of the detection (index of top ranked domains) essential to its utility
  - February 2012 Case Study
    - Two false positives
    - Both FPs served malware, but via social vectors
  - May 2012 Case Study
    - No false positives

Estimating Impact

- For each DDL site, need to conservatively estimate affected users
- Alexa published the popularity of a site as a percentage of all views
  - – Leveraged a popular website's visitor statistics to derive the actual number of all views
    - For example, in February 2012, Wikipedia self-reported 15.756 billion views
      - – Alexa indicated Wikipedia comprised 0.5416% of views
    - Working backward, Alexa based that percentage off of (15,756 * 1,000,000)/(29 * (0.5416/100)) = ~100.31 billion views each day
- To convert views to users, used Alexa-provided views per user estimation

Estimating Impact Cont’d

- For a set of affected users, need to conservatively estimate the subset that were successfully compromised
  - Used visitor statistics of popular websites, vendor studies
- For example, over 50% of users run a seldom-targeted or exploit-resistant platform (e.g., those using Mac OS X)
- Users with exploit-compatible software must be vulnerable to a given exploit – consider Java as an example
  - 73% of users have the Java web plugin installed (Adobe)
  - 42% of those use a version of Java vulnerable to exploitation (Qualys)
- After applying above filters, estimated that ~15.5% of users served malicious content are likely to be successfully compromised
- Hard to validate portions of this estimate, but overall result consistent with exploit kit control panel load percentages (12%-17%)

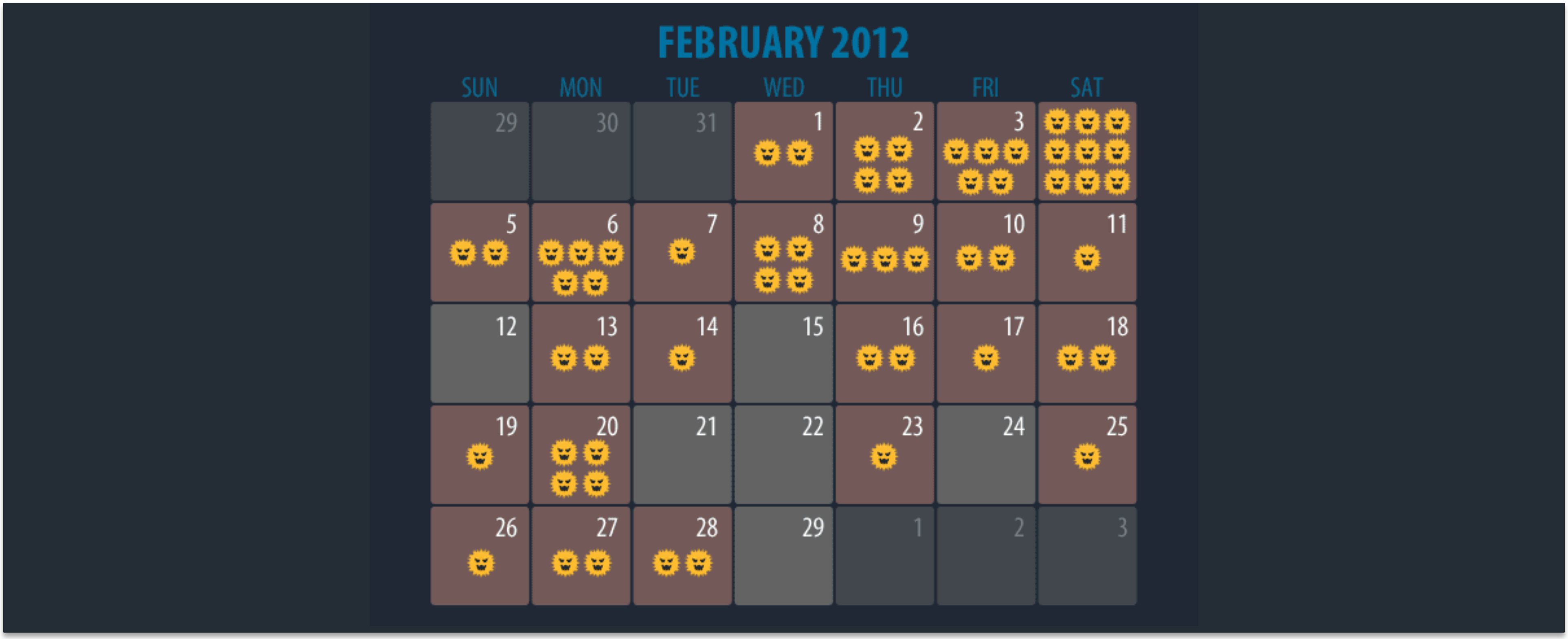
Analysis

BARRACUDA LABS

Case Study: February 2012

- When visited, 58 of the Alexa top 25,000 domains resulted in a drive-by download
  - Malicious content served by at least one top-ranked site 73% of the days in February
- Employ previously-described estimations
  - ~10.5 million users served malicious content
  - ~1.6 million users likely successfully compromised

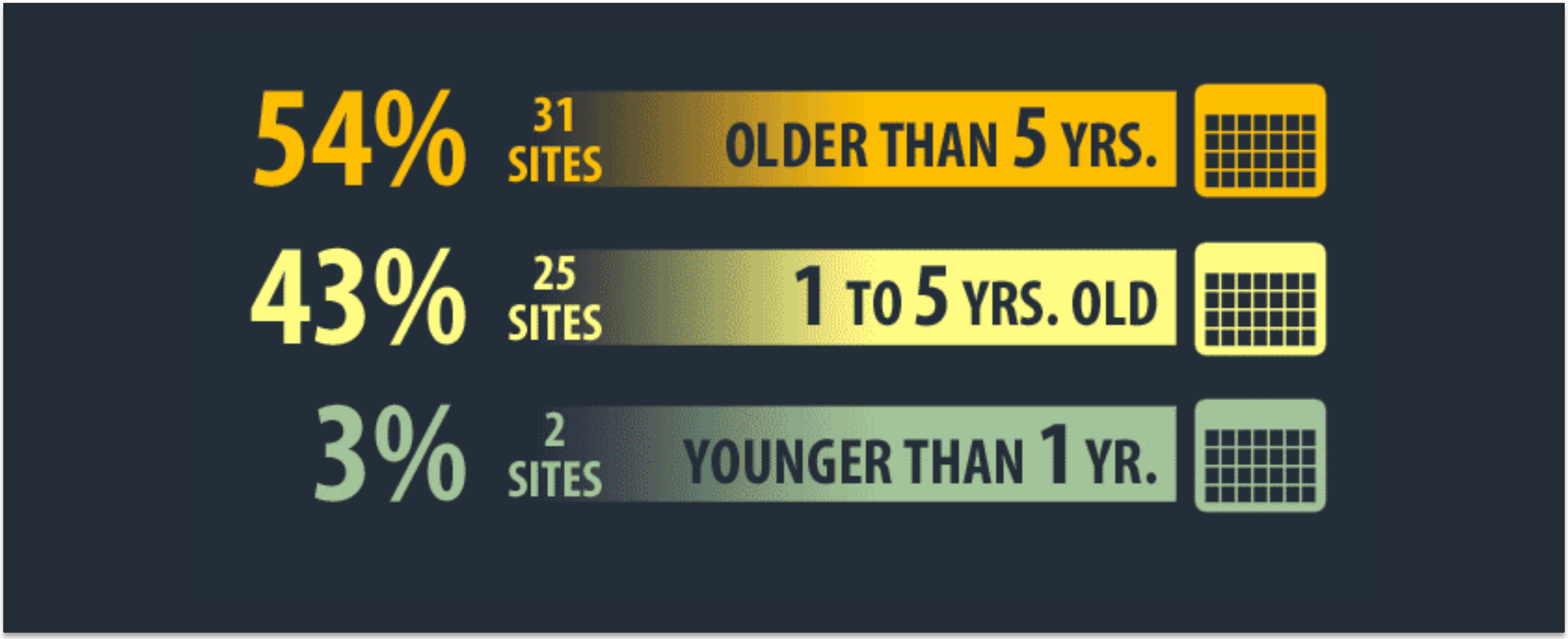
Top-Ranked Site DDL Calendar



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



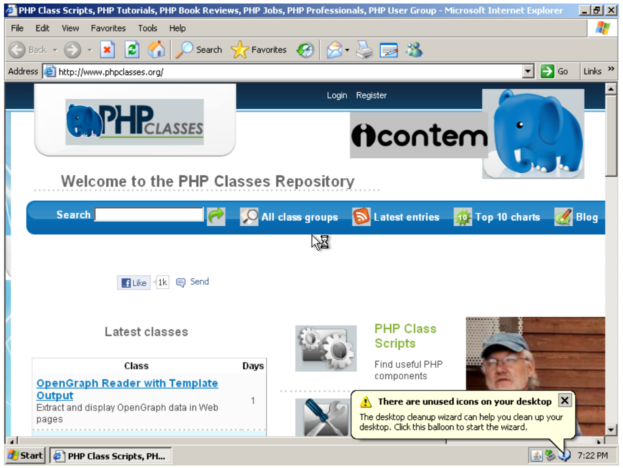
54% 31 SITES OLDER THAN 5 YRS.

43% 25 SITES 1 TO 5 YRS. OLD

3% 2 SITES YOUNGER THAN 1 YR.

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

Address http://www.phpclasses.org/

Welcome to the PHP Classes Repository

Latest classes

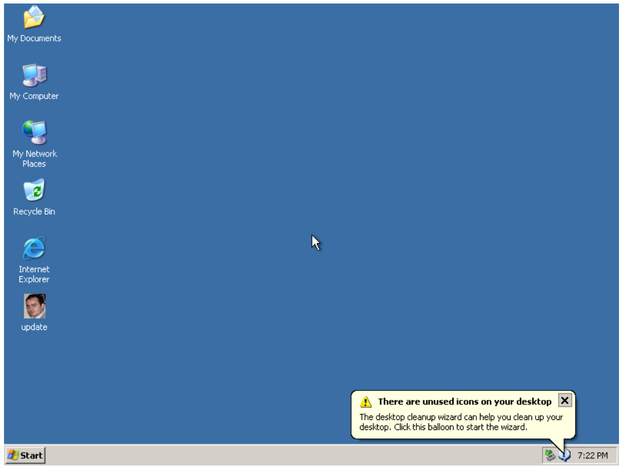
Class	Days
OpenGraph Reader with Template Output Extract and display OpenGraph data in Web pages	1

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

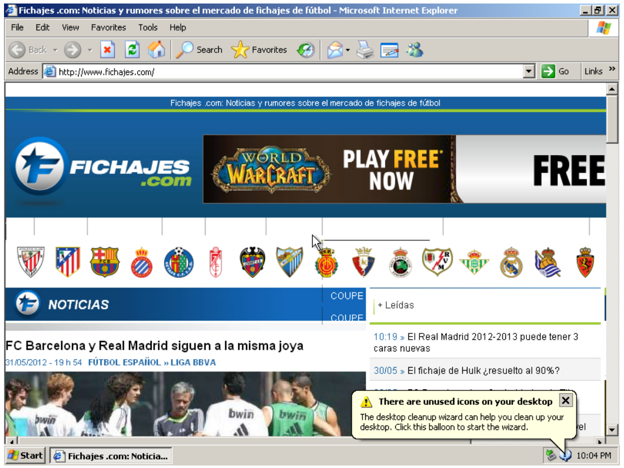
The desktop cleanup wizard can help you clean up your desktop. Click this balloon to start the wizard.

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
  - ~7.8 million users served malicious content
  - ~1.2 million users likely successfully compromised
- For the May 2012 study, functionality was added to examine recurring maliciousness
  - Most sites (72%) compromised for a single day, others for a week or more
  - Average period of compromise ~36 hours

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

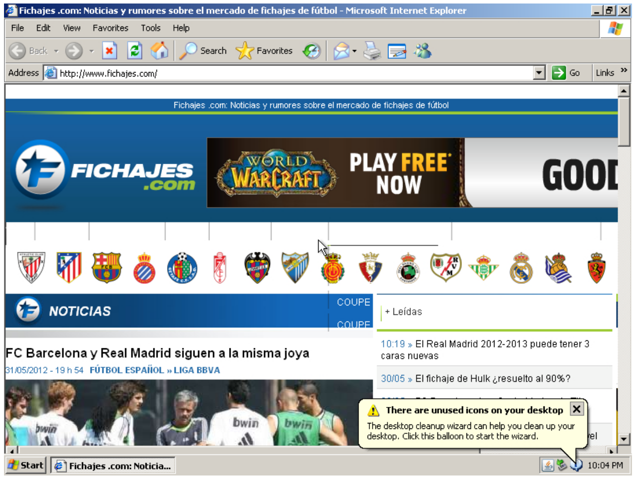
31/05/2012 - 19 h 54 FÚTBOL ESPAÑOL » LIGA BBVA

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.



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.

May 2012 DDL Properties

- Performed 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 resulted from direct website compromise
  - Hypothesized that Java was an overwhelmingly popular target in DDLs
    - Results matched expectation
      - 87.1% of DDLs included one or more exploits for Java

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 the mainstream, popular web is not a safe place

Threatglass

- A free-to-use web frontend for Barracuda Labs' URL analysis system
  - Designed for both casual end-users and researchers
- Provides VM screenshots and network activity visualizations for each drive-by download
  - Full PCAP of DDL session also available
- Encourages community participation via comment system, website submission support

Questions?

DDL Site Details, Source Data
threatglass.com

BARRACUDA LABS