

A Crawler-based Study of Spyware in the Web

Alex Moshchuk, Tanya Bragin,
Steve Gribble, Hank Levy

Department of Computer Science and Engineering
University of Washington
Seattle, WA

What do we mean by spyware?

- Difficult to define spyware precisely
 - No clean line between good and bad behavior
- Spyware is a *software parasite* that:
 - Collects information of value and relays it to a 3rd party
 - Hijacks resources or functions of PC
 - Installs surreptitiously, without user consent
 - Resist detection and de-installation
- Spyware provides value to others, but not to you

Spyware today

- Most Internet PCs have, or have had, spyware
- Harsh consequences for victims
- Explosion of anti-spyware software market
- We have very little quantitative data on spyware

The goal of this work

- Quantify the nature and extent of the spyware problem from the Internet point of view
- Example questions:
 - How prevalent is spyware on the Web?
 - What Web categories are most infected?
 - What are the spyware trends over time?

Talk overview

- We studied the two methods by which spyware infects victims
 - Spyware piggy-backed on executables
 - E.g., Kazaa ships bundled with multiple spyware programs
 - Drive-by download installation
 - Malicious web content exploits browser flaws to install spyware
- We repeated each study to understand the trends
 - May 2005, October 2005
 - We present data for October

Popularity of sites in our study

- Does anyone visit any of the sites we've examined?
 - Popularity ratings (using Alexa) confirm that we have crawled sites across all popularity rankings
 - A few very popular sites
 - Many popular sites
 - Intuition
 - Companies will put adware in popular, easy-to-reach places

Outline

- Introduction
- Executable file study
- Drive-by download study
- Related work and conclusions

Crawling for executables

- Measure spyware prevalence in sites people tend to visit
- We defined 10 interesting Web categories
 - E.g., games, news, celebrities, pirate, wallpaper
- For each category, we:
 - Used Google to identify several hundred domains
 - Crawled each domain (to depth 3) to find executables
 - Downloaded executables for offline analysis
- Crawled about 20 million URLs over 2,500 domains
- Collected 20,000 executables
 - 19% of domains had downloadable executables

Analyzing executables

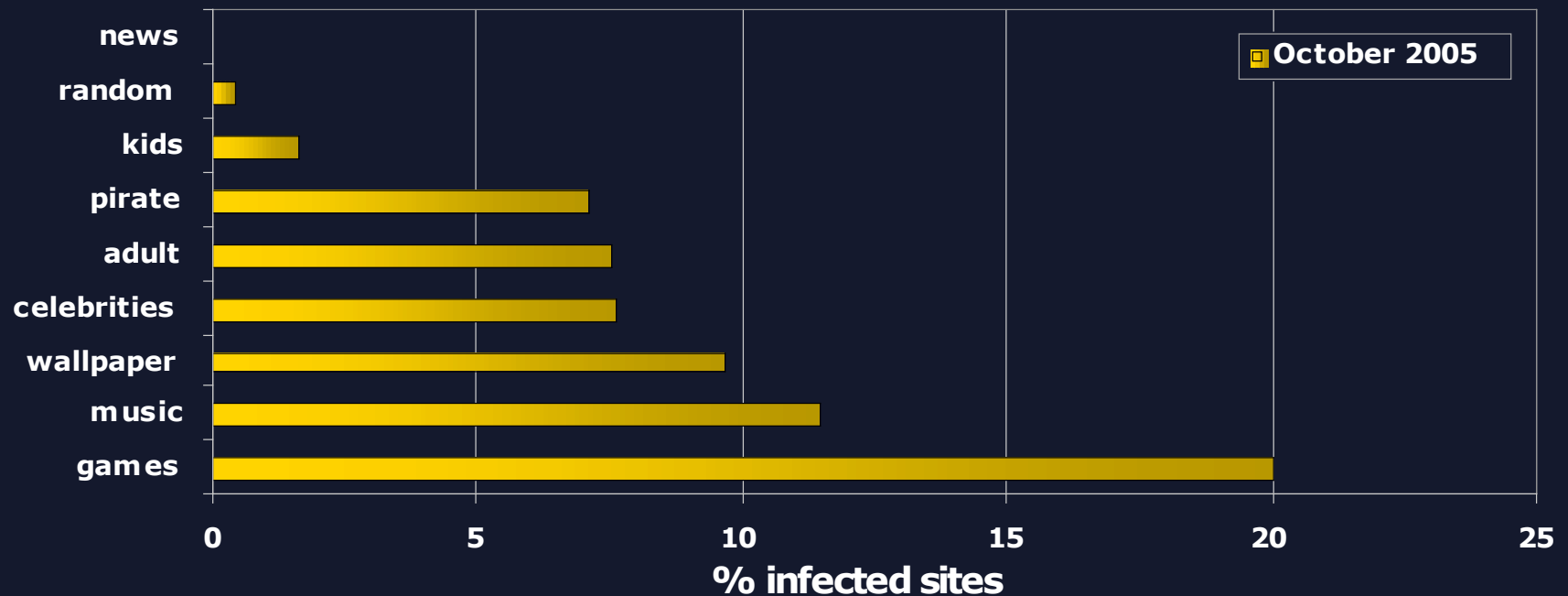
- For each executable, we:
 - Cloned a clean WinXP virtual machine (VMware)
 - Automatically installed the executable into the VM
 - Ran an anti-spyware tool to look for infections
 - We used Lavasoft Ad-Aware
- Automating installation required some heuristics
 - E.g., pressing "Next," agreeing to EULAs, ...
- An executable is *infected* if Ad-Aware finds spyware
 - Limited to what Ad-Aware can detect
 - We found choice of the tool rarely matters

High-level results

- We found a lot of piggy-backed spyware
 - 1 in 20 executables contained spyware
 - 1 in 25 domains were infectious
- We observed few spyware variants
 - We encountered 1,294 infected executables but only 89 spyware programs
- No significant change in amount of piggy-backed spyware from May 2005 to October 2005

Where is the spyware found?

- Spyware is concentrated on specific popular Web zones
 - High-profile organizations tend to have spyware-free sites
 - Downloads from unknown sources are risky



Spyware on c|net

- We examined 2,000 executables on download.com
 - In May, we found spyware in 110 programs (4.6%)
 - In October, we found spyware in only 6 programs
- c|net implemented a no-spyware policy between our crawls
 - Mostly effective
 - Some programs can still fool the filters

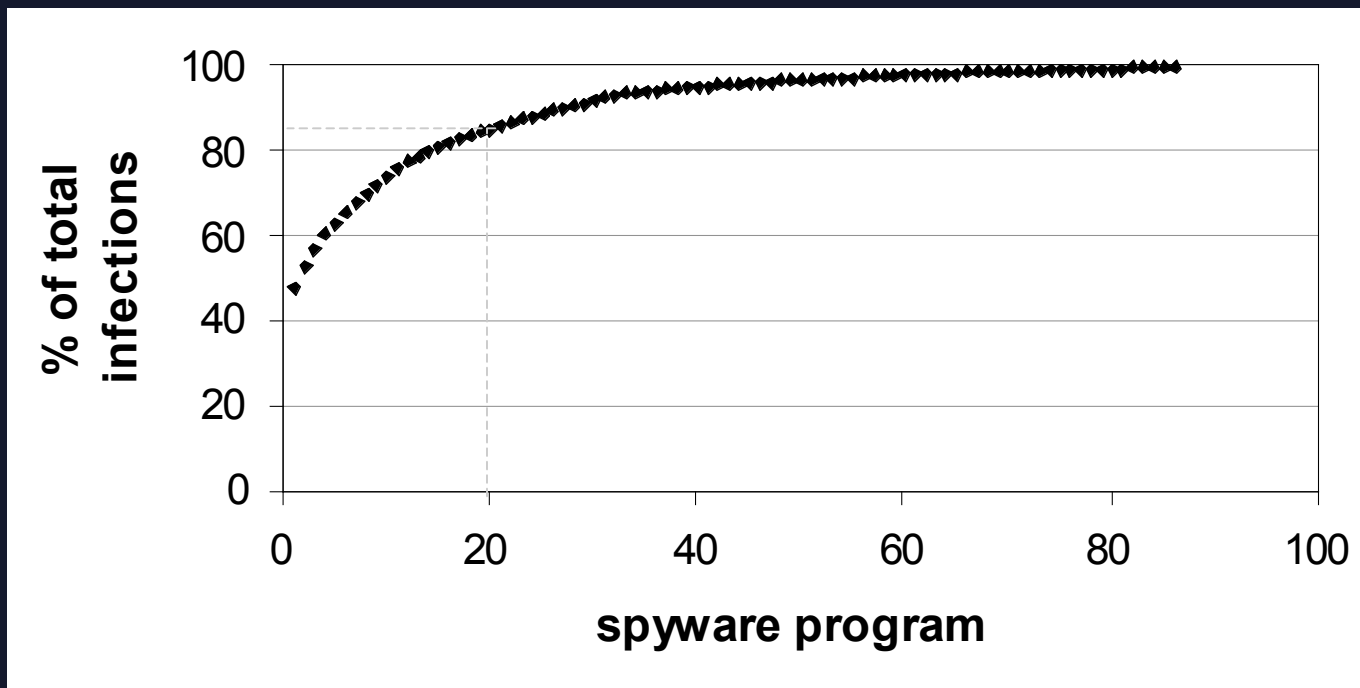
How is spyware distributed across sites?

- A small # of sites have a large # of infected executables
 - Easy to detect and blacklist, given our tool

Top spyware sites	# infected executables
scenicreflections.com	503
gamehouse.com	164
screensavershot.com	137
screensaver.com	107
hidownload.com	50
games.aol.com	30
appzplanet.com	27
dailymp3.com	27
free-games.to	27

Distribution of spyware programs

- A few offenders are responsible for most infected executables
- Top offenders are well-known (e.g., WhenU)
- Many spyware programs are rare
- Signature-based detection should be effective



What kinds of spyware do we find?

- We measured the prevalence of five spyware functions:
 - Keyloggers
 - Dialers
 - Trojan downloaders
 - Browser hijackers
 - Adware
- Adware and browser hijackers are most common (86%)
- Trojan downloaders pose a risk (13%)
- Keyloggers and dialers are more rare (1%)

Piggy-backed spyware summary

- A large number of executables are infected (1 in 20)
- Spyware is focused on a small number of popular sites
- Most of it is benign
- Only a few variants matter
- Implications:
 - Easy to identify and defend against the main culprits
 - Signature-based techniques should be effective

Outline

- Introduction
- Executable file study
- Drive-by download study
- Related work and conclusions

Drive-by download study

- First study examined downloadable executables
- Next, we look at Web pages with drive-by downloads
 - Web content exploits browser flaws to install spyware
 - Victims are infected just by visiting a malicious page

Methodology

- Goal: find malicious Web pages automatically
- Detect attacks as they happen in practice
 - Crawl our Web categories
 - Render each page in an unmodified Web browser inside a clean VM
 - Internet Explorer (6.0, unpatched)
 - Mozilla Firefox (1.0.6)
 - Run anti-spyware check to look for spyware

Using Event Triggers

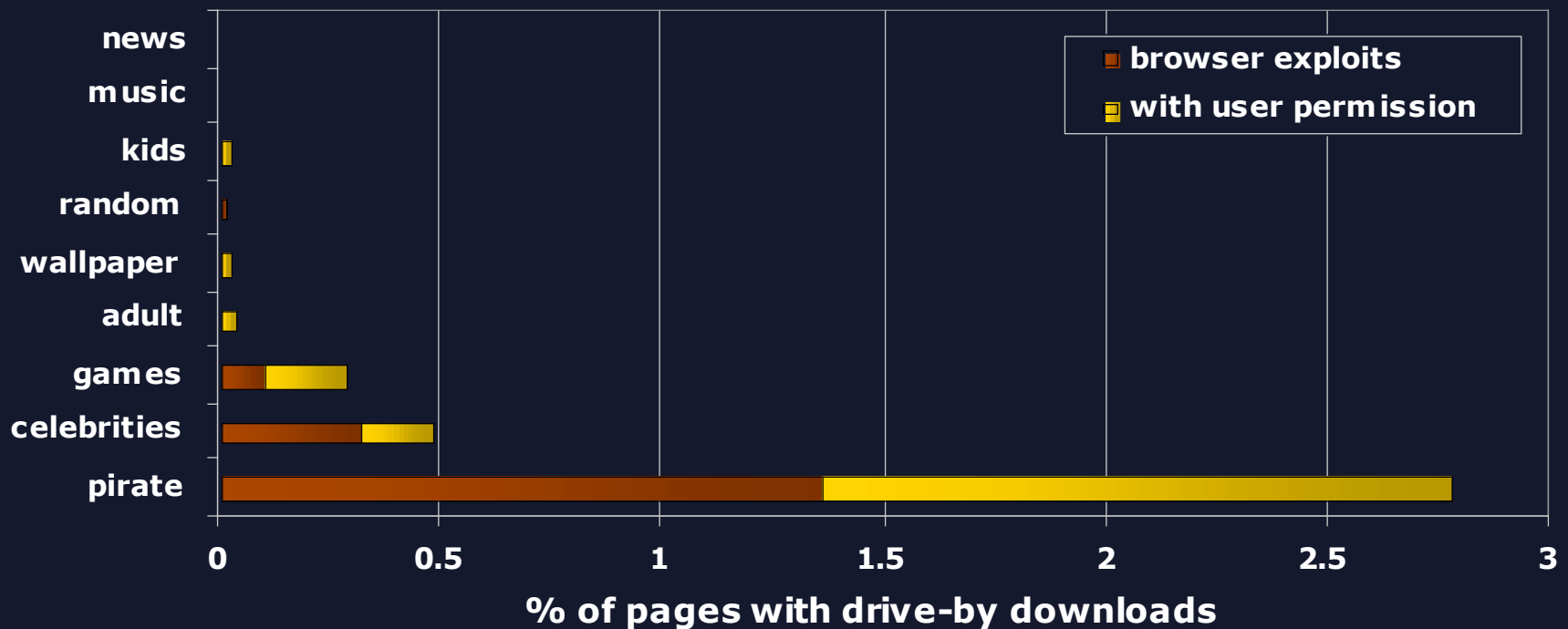
- Event triggers are a performance optimization
- Triggers detect suspicious activity
 - Process creation
 - Suspicious registry modifications
 - Files written outside browser temp. folders
- Run Ad-Aware check *only* when a trigger fires
 - No false negatives
 - 41% false positives
 - Benign software installations
 - Background noise
 - Spyware not detected by Ad-Aware

High-level results

- There are many Web pages with drive-by downloads
 - 0.4% of Web pages are infectious
- 50% of attacks exploited browser flaws
 - These bypass the browser security framework
- Little variation
 - Only 36 spyware programs responsible for 186 attacks
- Different threats than piggy-backed spyware programs

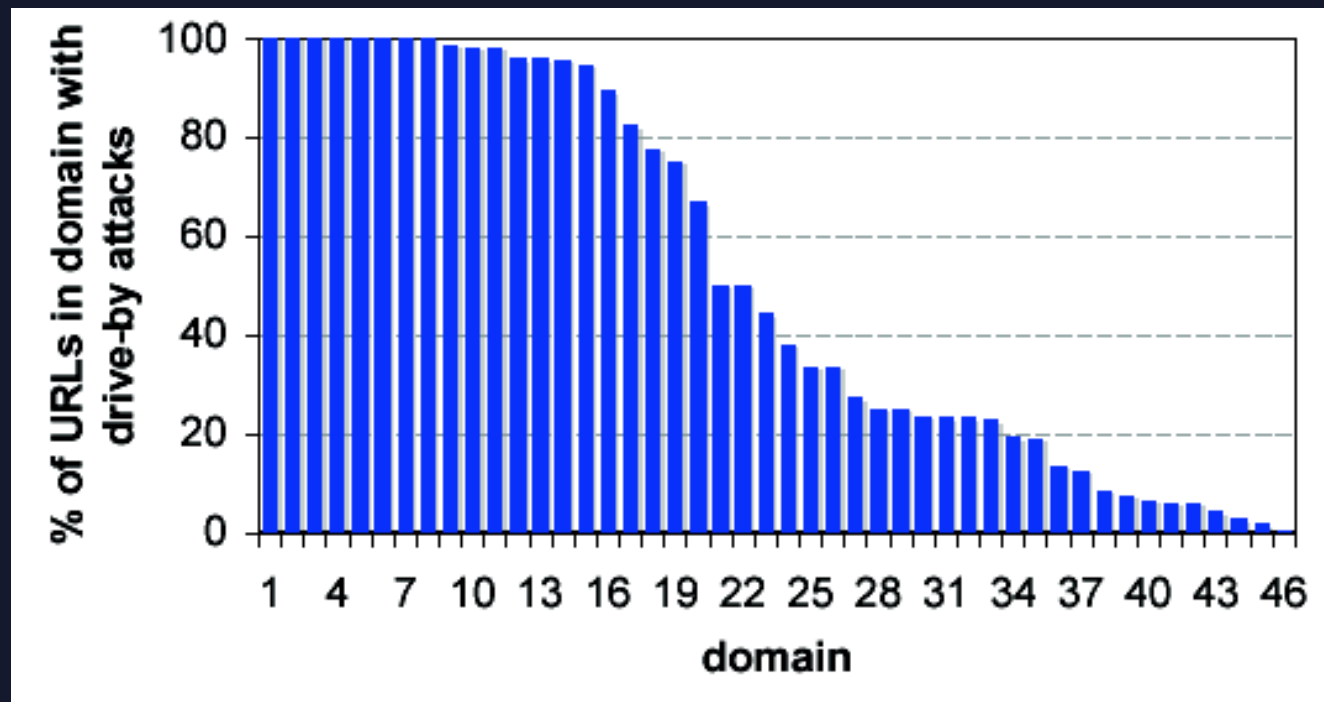
Where are drive-bys found?

- Non-uniform distribution
- Surprisingly many browser exploits!



Spyware prevalence in infectious domains

- Infectious sites often attempt attacks on a large number of their Web pages
 - Sufficient to identify bad sites, rather than bad pages



Is the Firefox browser susceptible?

- Successful drive-by downloads appeared on 0.08% of pages
 - All require user consent
 - All are based on Java
- Firefox is *not* 100% safe, but it is *safer to use* than IE
 - Firefox flaws are not yet being exploited
 - We found 13 times more attacks for IE than for Firefox

Drive-by download trends

- The number of pages with drive-by downloads is decreasing
 - All categories experienced a decrease from May to October
 - Overall, Web page infection decreased 93%
- Our results suggest spyware is past its prime
- Possible reasons:
 - Success rate of attacks is declining
 - Widespread adoption of anti-spyware tools
 - Recent lawsuits discouraging attackers

Drive-by download summary

- Despite the decline, there are still many infectious pages
- 50% of these pages infect *without* user consent
- Malicious content is focused on a small number of sites
- Only a few variants matter
- Firefox is also susceptible
- Implications:
 - Patching security holes is important
 - Automated crawler-based tools are effective at finding sites with malicious content

How big is our Ad-Aware limitation?

- We relied on Ad-Aware to identify known spyware
 - How much spyware are we missing by not using other tools?
- For drive-by downloads, triggers limit how much we miss
 - Upper bound: 41% false positives when a trigger fires
- For piggy-backed spyware, we compared Ad-Aware to Webroot Spy Sweeper
 - Of 100 random executables, only 1 was missed by Ad-Aware

		Spy Sweeper	
		clean	infected
Ad-Aware	clean	90	1
	infected	1	8

Outline

- Introduction
- Executable file study
- Drive-by download study
- Related work and conclusions

Related Work

- Honeypots
- Strider HoneyMonkey
 - Tool to find Web pages with browser exploits
 - Method similar to our trigger-based VM approach
 - We focus more on analysis
- Webroot Phileas, Sunbelt
 - Automated web crawling for new spyware variants
- SiteAdviser
 - Upcoming commercial service to rate safety of Web sites

Conclusions

- We addressed key questions about spyware:
 - Prevalence
 - Location
 - Trends
- Takeaway lessons:
 - Despite the decreasing trend, spyware is still a big problem
 - Spyware is usually not as dangerous as people claim
 - Signature-based defenses should be effective
 - Need automated tools to identify what matters in practice
 - Opt-in schemes for browser security are not effective

Questions?