A Crawler-based Study of Spyware in the Web

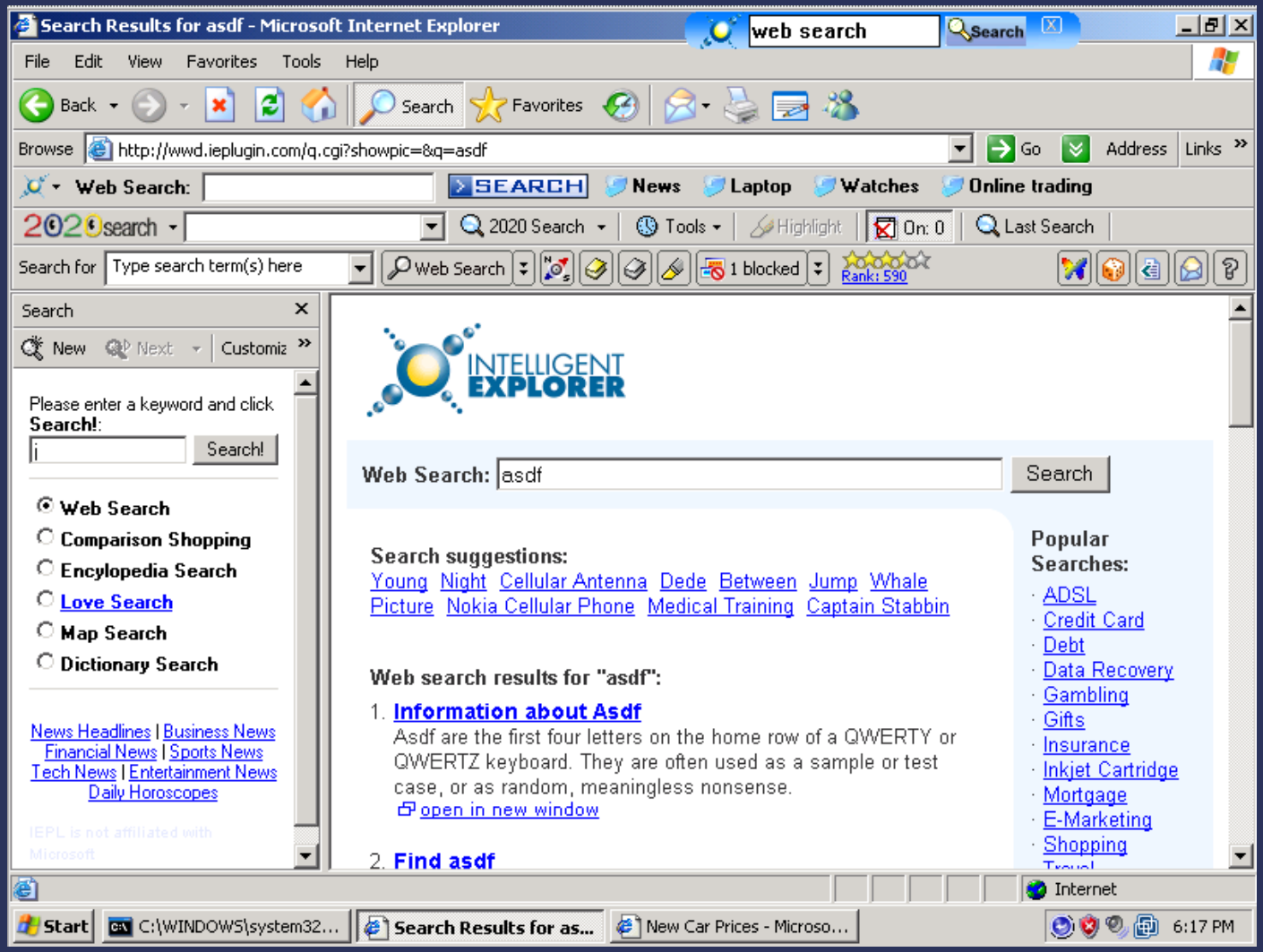
Alex Moshchuk, Tanya Bragin,
Steve Gribble, Hank Levy

What is spyware?

- Broad class of malicious and unwanted software
- Steal control of a PC for the benefit of a 3rd party

- Characteristics:
  - Installs without user knowledge or consent
  - Hijacks computer's resources or functions
  - Collects valuable information and relays to a 3rd party
  - Resists detection and uninstallation

You know it when you see it



How do people get spyware?

- Spyware piggybacked on popular software
  - Kazaa, eDonkey
- Drive-by downloads
  - Web page installs spyware through browser
  - With or without user consent
- Trojan downloaders
  - Spyware downloads/installs more spyware

Why measure spyware?

- Understand the problem before defending against it
- Many unanswered questions
  - What's the spyware density on the web?
  - Where do people get spyware?
  - How many spyware variants are out there?
  - What kinds of threats does spyware pose?
- New ideas and tools for:
  - Detection
  - Prevention

Approach

- Large-scale study of spyware:
  - Crawl “interesting” portions of the Web
  - Download content
  - Determine if it is malicious
- Two strategies:
  - Executable study
    - Find executables with known spyware
  - Drive-by download study
    - Find Web pages with drive-by downloads

Outline

- Introduction
- Executable file study
- Drive-by download study
- Summary
- Conclusions

Analyzing executables

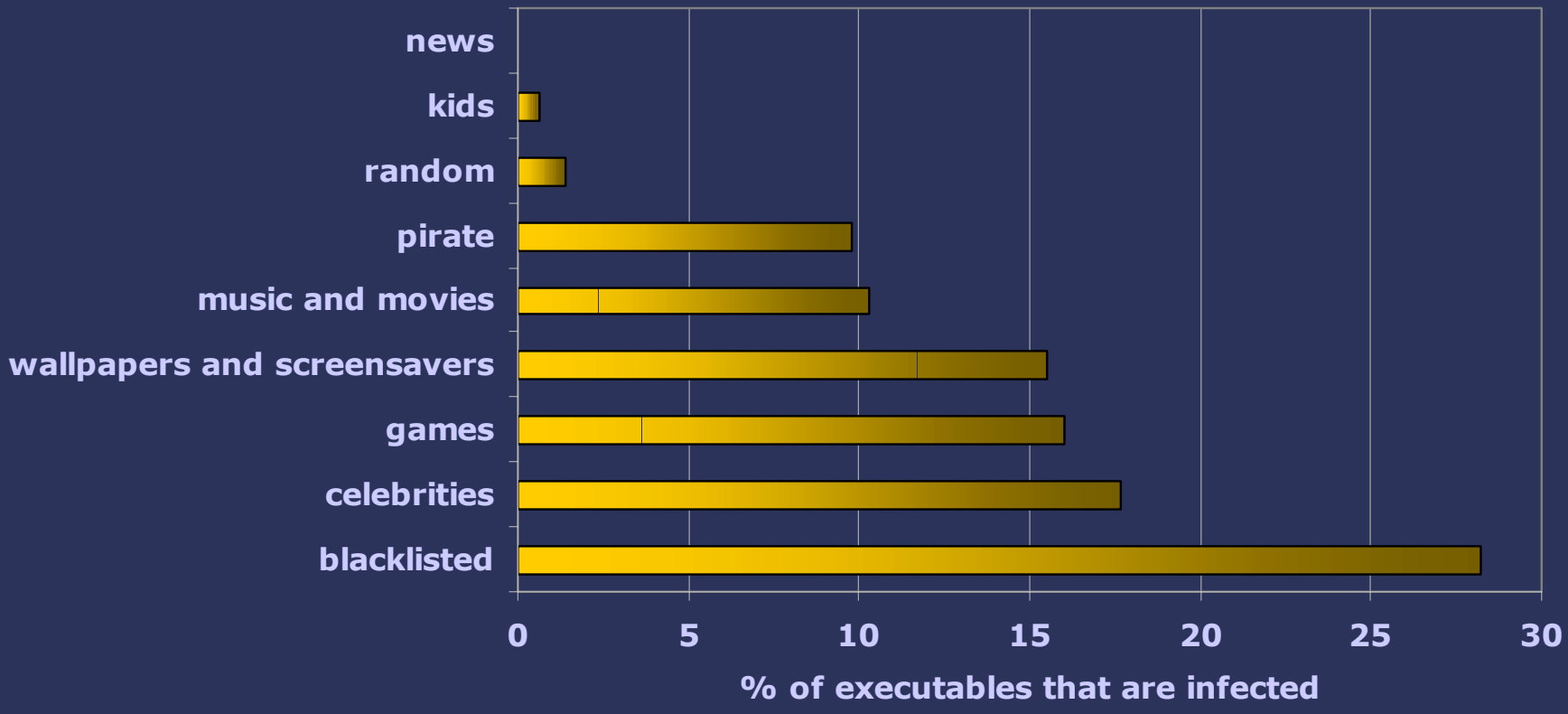
- Web crawler collects a pool of executabes
- Analyze each in a virtual machine:
  - Clone a clean WinXP VM
  - Automatically install executable
  - Run analysis to see what changed
    - Currently, an anti-spyware tool (Ad-Aware)
- Average analysis time – 90 sec. per executable

Executable study results

- Crawled 32 million pages in 9,000 domains
- Downloaded 26,000 executables
- Found spyware in 12.3% of them
  - Most installed just one spyware program
    - Only 6% installed three or more spyware variants
  - Few spyware variants encountered in practice
    - 142 unique spyware threats

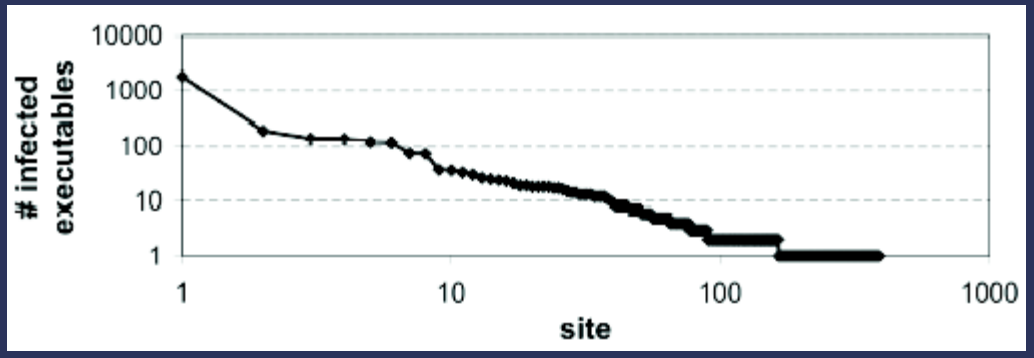
Main targets

- Visit a site and download a program
- What's the chance that you got spyware?

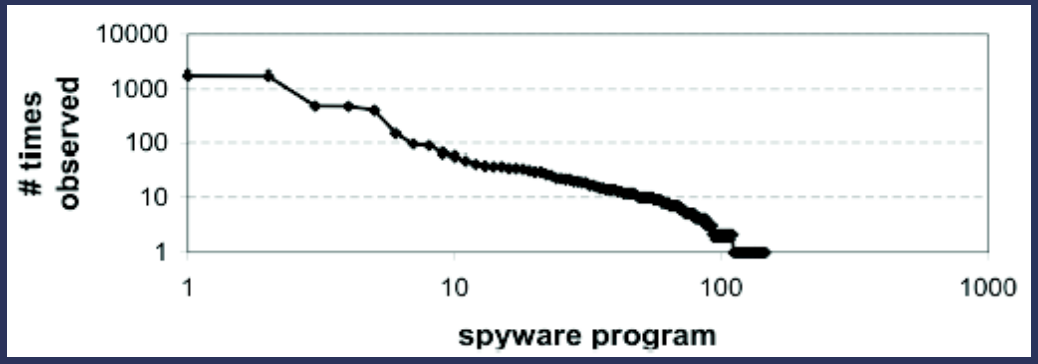


Popularity

- A small # of sites have large #of spyware executables:



- A small # of spyware variants are responsible for the majority of infections:



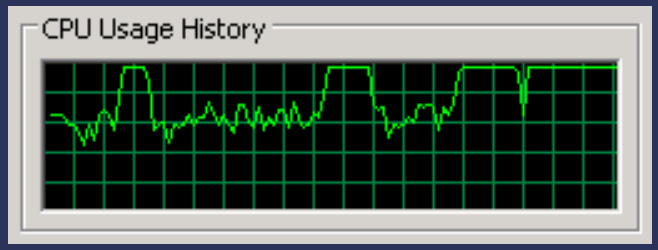
Types of spyware

- Quantify the kinds of threats posed by spyware
- Consider five spyware functions
  - What's the chance an infected executable contains each function?

Keylogger	0.05%
Dialer	1.2%
Trojan downloader	12%
Browser hijacker	62%
Adware	88%

Example of a Nasty Executable

- http://aaa1screensavers.com/
  - "Let all your worries melt away into this collection of clouds in the sky – 100% free!"
  - http://aaa1screensavers.com/free/clouds.exe
- Installs 11 spyware programs initially
  - Includes a trojan downloader; continually installs more spyware
    - 10 more within first 20 minutes
- 12 new items on desktop, 3 browser toolbars
- Shows an ad for every 1.5 pages you visit
- CPU usage is constantly 100%
- No uninstallers
- Ad-Aware can't clean
- System stops responding in 30 mins
  - Restarting doesn't help
- Unusable system and no screensaver!



Outline

- Introduction
- Executable file study
- Drive-by download study
- Summary
- Conclusions

Finding drive-by downloads

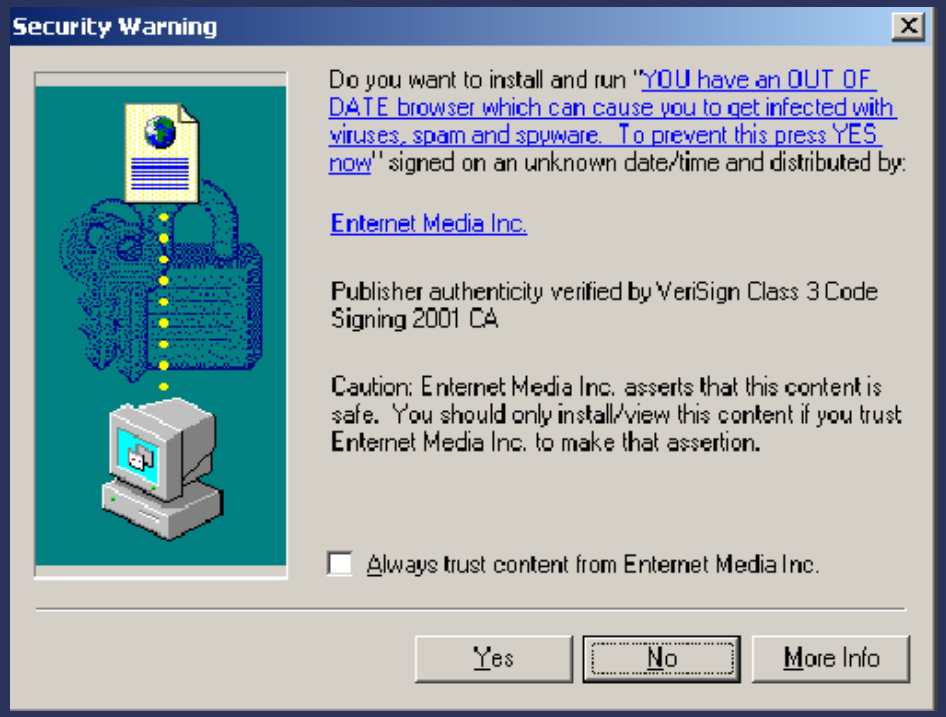
- Evaluate the safety of browsing the Web
- Approach: automatic virtual browsing
  - Render pages in a real browser inside a clean VM
    - Internet Explorer
    - Mozilla Firefox
  - Identify malicious pages
    - Define triggers for suspicious browsing activity
    - Run anti-spyware check only when trigger fires

Event triggers

- Real-time monitoring for non-normal behavior:
  - Process creation
  - File events
    - Example: foo.exe written outside IE folders.
  - Registry events
    - Example: new auto-start entry for foo.exe
- No false negatives (theoretically)
- 41% false positives:
  - Legitimate software installations
  - Background noise
  - Spyware missed by our anti-spyware tool

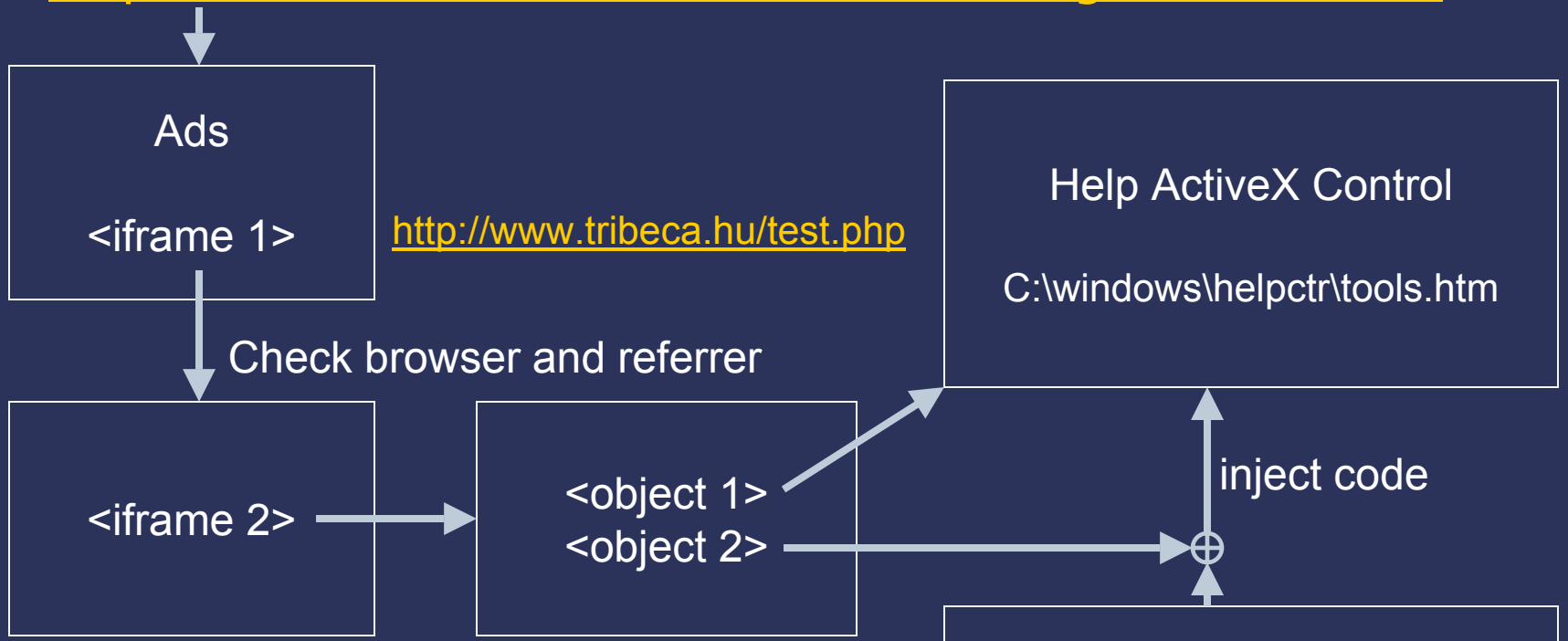
More on automatic browsing

- Caveats and tricks
  - Restore clean state before navigating to next page
  - Speed up virtual time
  - Monitor for crashes and freezes
- Deciding what to say to security prompts:
  - "yes"
    - Emulate user consent
  - "no" (or no prompt)
    - Find security exploits



Example of a security exploit

- http://www.1000dictionaries.com/free_games_1.html

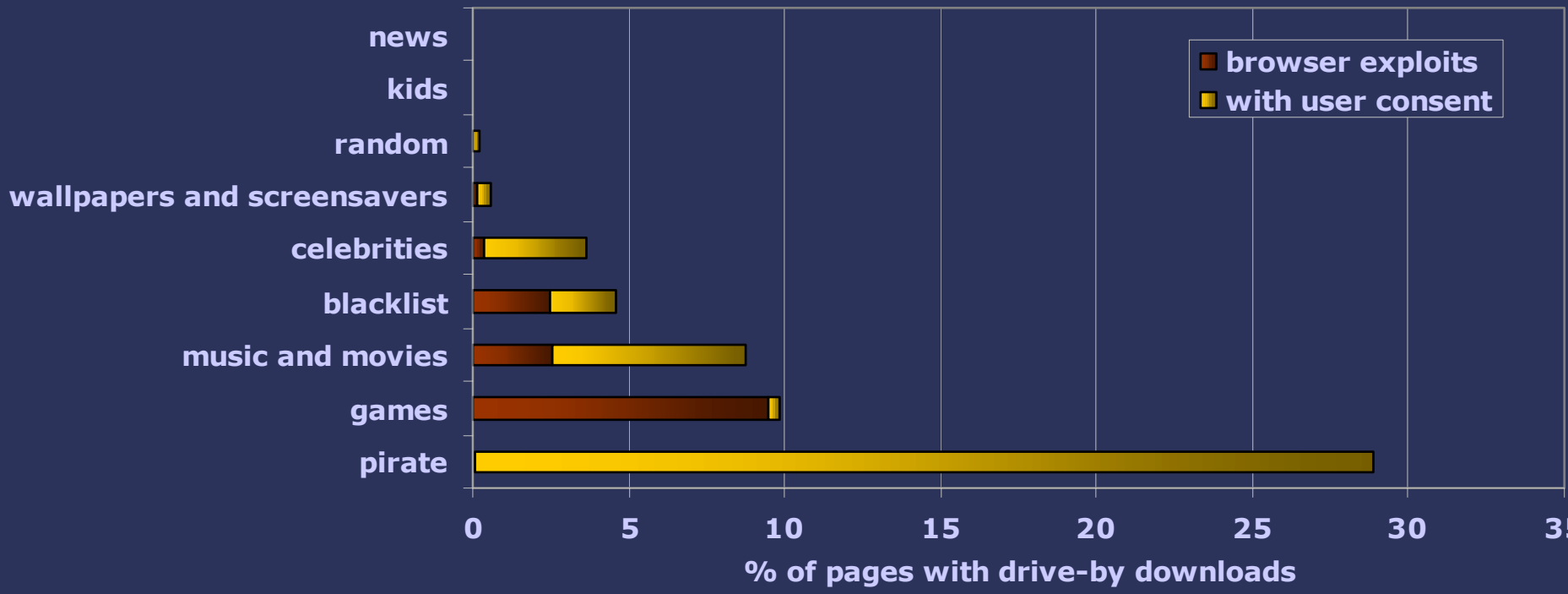


- Local help objects bypass security restrictions; unsecured "local zone"
- Cross-zone scripting vulnerability in ActiveX Help allows JavaScript to inject code into a local help control

Drive-by download results

(unpatched Internet Explorer, unpatched WinXP)

- Examined 50,000 pages
- 5.5% carried drive-by downloads
  - 1.4% exploited browser vulnerabilities



Types of spyware

- Is drive-by download spyware more dangerous?

	Executables	Drive-by Downloads
Keylogger	0.05%	0%
Dialer	1.2%	0.2%
Trojan Downloader	12%	50%
Browser hijacker	62%	84%
Adware	88%	75%

Is Firefox better than IE?

- Repeat drive-by download study with Mozilla Firefox
- Found 189 (0.4%) pages with drive-by downloads
  - All require user consent
  - All are based on Java
    - Work in other browsers
- Firefox is not 100% safe
  - However, much safer than IE

adult	0
celebrity	**33**
games	0
kids	0
music	**1**
news	0
pirate	**132**
random	0
wallpaper	0
blacklist	**23**
Total:	**189**

Summary

- Lots of spyware on the Web
  - 1 in 8 programs is infected with spyware
  - 1 in 18 Web pages has a spyware drive-by download
  - 1 in 70 Web pages exploits browser vulnerabilities
- Most of it is just annoying (adware)
  - But a significant fraction poses a big risk
- Spyware companies target specific popular content
  - Most piggy-backed spyware in games & celebrity sites
  - Most drive-by downloads in pirate sites
- Few spyware variants are encountered in practice

Conclusion and Future Work

- Addressed key questions about spyware
- Built useful tools and infrastructure

- More details:
  A Crawler-based Study of Spyware in the Web
  NDSS06

- Looking forward:
  - Real-time protection with a trigger-based Web proxy
  - Automatically detect new spyware
    - Use triggers as truth
  - Increase the scale of the study
  - Study change of spyware over time (see paper!)

Questions?