

# WEBCOP: LOCATING NEIGHBORHOODS OF MALWARE ON THE WEB

- Jay Stokes  
*Microsoft Research*
- Reid Andersen
- Christian Seifert
- Kumar Chellapilla  
*Microsoft Search*

# Detecting Malicious Web Pages

The screenshot shows a Windows Internet Explorer browser window displaying a Bing search results page. The address bar shows the search query: `http://www.bing.com/search?q=site%3A+similarminds.com&src=IE-SearchBox`. The search results are for the query `site: similarminds.com`. The page displays several search results, including:

- similarminds.com - Site Info from Alexa**: `similarminds.com` is one of the top 100,000 sites in the world and is in the Directories category. [www.alexac.com/siteinfo/similarminds.com](http://www.alexac.com/siteinfo/similarminds.com) - [Cached page](#) - [Mark as spam](#)
- SimilarMinds.com**: Free Personality Test Site ... \*Jung tests are similar in underlying theory to Myers-Briggs typology (of which this site has no affiliation) [similarminds.com](http://similarminds.com)
- Peronality Type Descriptions**: `similarminds.com` ... Global 5 Primary Type Descriptions [similarminds.com/software.html](http://similarminds.com/software.html) - [Cached page](#) - [Mark as spam](#)
- similarminds.com - Quantcast Audience Profile**: This site reaches approximately 56,59 U.S. monthly people. The site is popular among a rather female, teen group. The typical visitor visits [elearners.com](http://elearners.com) and reads Washington Post. [www.quantcast.com/similarminds.com](http://www.quantcast.com/similarminds.com) - [Cached page](#) - [Mark as spam](#)
- Site Profile for similarminds.com (rank #31,681) | Compete**: Website profile of `similarminds.com` - traffic charts, graphs, popularity rank, ratings, trust safety profiles, scores, deals, promotional codes, and more. [siteanalytics.compete.com/similarminds.com](http://siteanalytics.compete.com/similarminds.com) - [Cached page](#) - [Mark as spam](#)
- Personality Tests and Tools**: `similarminds.com` ... tests are similar in underlying theory to Myers-Briggs typology (of which this site ... [sminds.com](http://sminds.com) - [Cached page](#) - [Mark as spam](#)

The browser status bar at the bottom shows "Done", "Internet | Protected Mode: Off", and "100%".

# Detecting Malicious Web Pages

site: similarminds.com - Bing - Windows Internet Explorer

http://www.bing.com/search?q=site%3A+similarminds.com&src=IE-SearchBox

site: similarminds.com

Web Images Videos Shopping News Maps More MSN Hotmail Sign in | United States | Preferences

bing MS Beta 9962

site: similarminds.com

Make Bing your decision engine

ALL RESULTS 1-10 of 25,000 results - [Advanced](#)

**similarminds.com - Site Info from Alexa**  
similarminds.com is one of the top 100,000 sites in the world and is in the Directories category  
[www.alexametrics.com/siteinfo/similarminds.com](http://www.alexametrics.com/siteinfo/similarminds.com) - [Cached page](#) - [Mark as spam](#)

**SimilarMinds.com**  
Free Personality Test Site ... \*Jung tests are similar in underlying theory to Myers-Briggs typology (of which this site has no affiliation)  
[similarminds.com](http://similarminds.com)

**Peronality Type Descriptions**  
similarminds.com ... Global 5 Primary Type Descriptions  
[similarminds.com/software.html](http://similarminds.com/software.html) - [Cached page](#) - [Mark as spam](#)

**similarminds.com - Quantcast Audience Profile**  
This site reaches approximately 56,59 U.S. monthly people. The site is popular among a rather female, teen group. The typical visitor visits [elearners.com](http://elearners.com) and reads Washington Post.  
[www.quantcast.com/similarminds.com](http://www.quantcast.com/similarminds.com) - [Cached page](#) - [Mark as spam](#)

**Site Profile for similarminds.com (rank #31,681) | Compete**  
Website profile of similarminds.com - traffic charts, graphs, popularity rank, ratings, trust safety profiles, scores, deals, promotional codes, and more.  
[siteanalytics.compete.com/similarminds.com](http://siteanalytics.compete.com/similarminds.com) - [Cached page](#) - [Mark as spam](#)

**Personality Tests and Tools**  
similarminds.com ... tests are similar in underlying theory to Myers-Briggs typology (of which this site ...  
[sminds.com](http://sminds.com) - [Cached page](#) - [Mark as spam](#)

**CAREFUL!**  
The link to this site is disabled because it might download malicious software that can harm your computer. [Learn More](#)

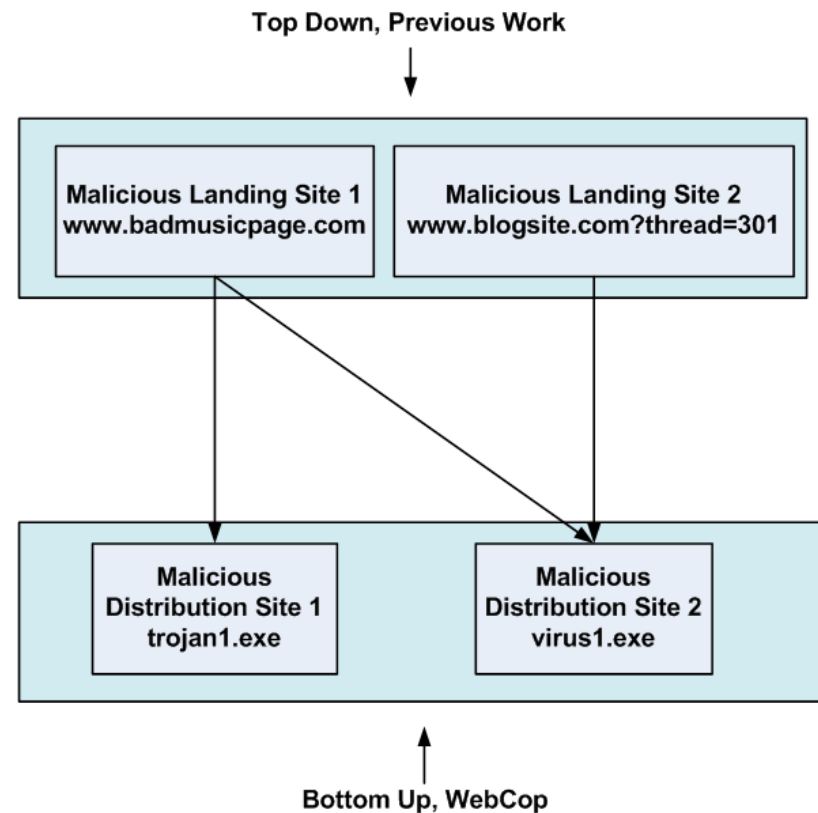
We suggest you choose another result, but if you want to risk it, [visit the website](#).

Internal preview Help improve Bing

http://similarminds.com/ Internet | Protected Mode: Off 100%

# Production System

- Drive-By Download
  - ▣ Malware is automatically downloaded
  - ▣ No user interaction
  - ▣ Strider HoneyMonkey (Wang 2006)
- Top-Down Approach
- Obfuscated JavaScript redirections
- Other notable work (Moshchuk 2006, Provos 2007, 2008)



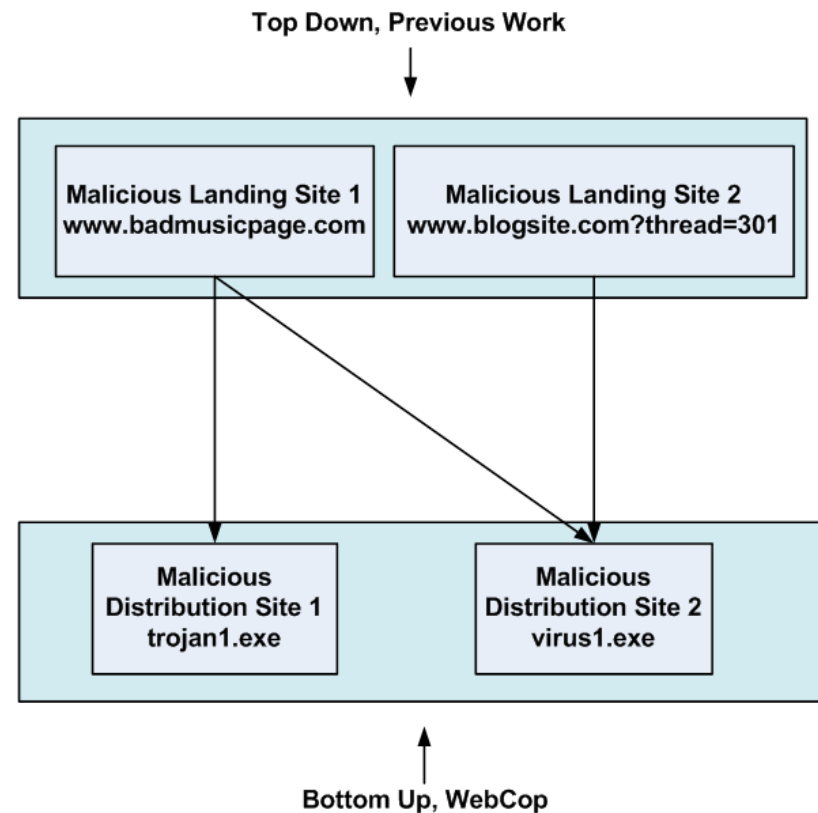
# Drive-by Detection Limitations



- Difficult to identify suspicious pages to scan
- Production system looks for changes after running malware in a virtual machine
  - ▣ Attackers adapt and learn to avoid detection
  - ▣ Malware will often detect it is running in a VM
  - ▣ Halt execution
- Centrally Located Service

# Top-Down with Crawler

- ❑ Moshchuk 2006,  
Stamminger 2009
- ❑ Crawl the web
- ❑ Direct Links
- ❑ Download and test  
executables
- ❑ AM Scan



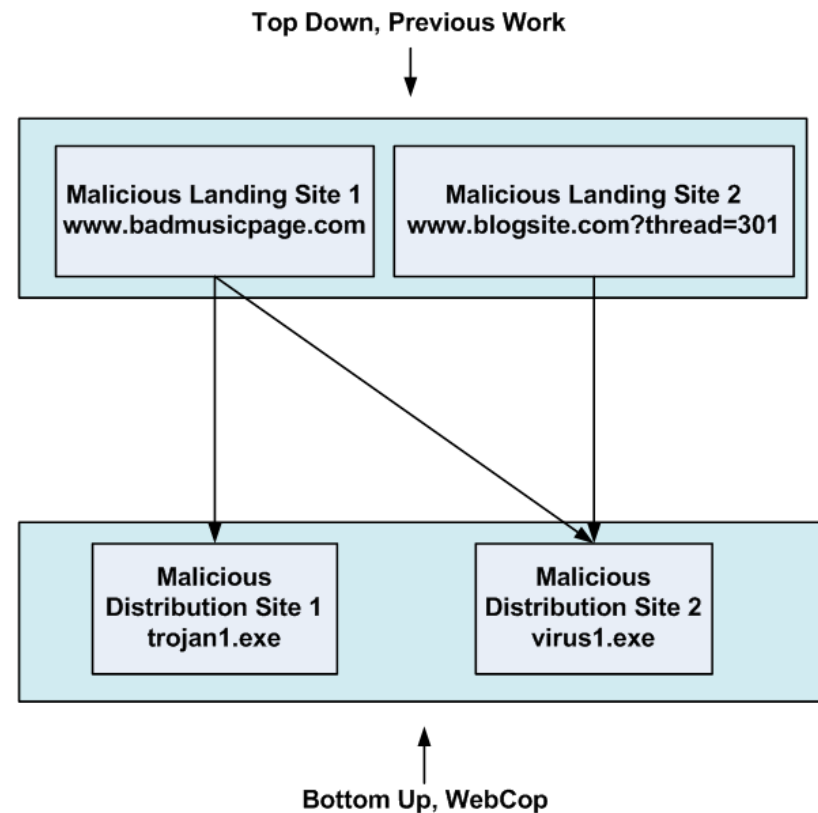
# Top-Down Crawling Limitations



- ❑ Downloading all executables from the internet is problematic
- ❑ Need to simulate user input
  - ▣ Installation, web surfing
- ❑ Scanning with an AM engine
  - ▣ May require full system scan (Stamminger 2009)
- ❑ To avoid reimaging, test in a VM
  - ▣ Again, malware can detect VM and hide
- ❑ Centrally located service

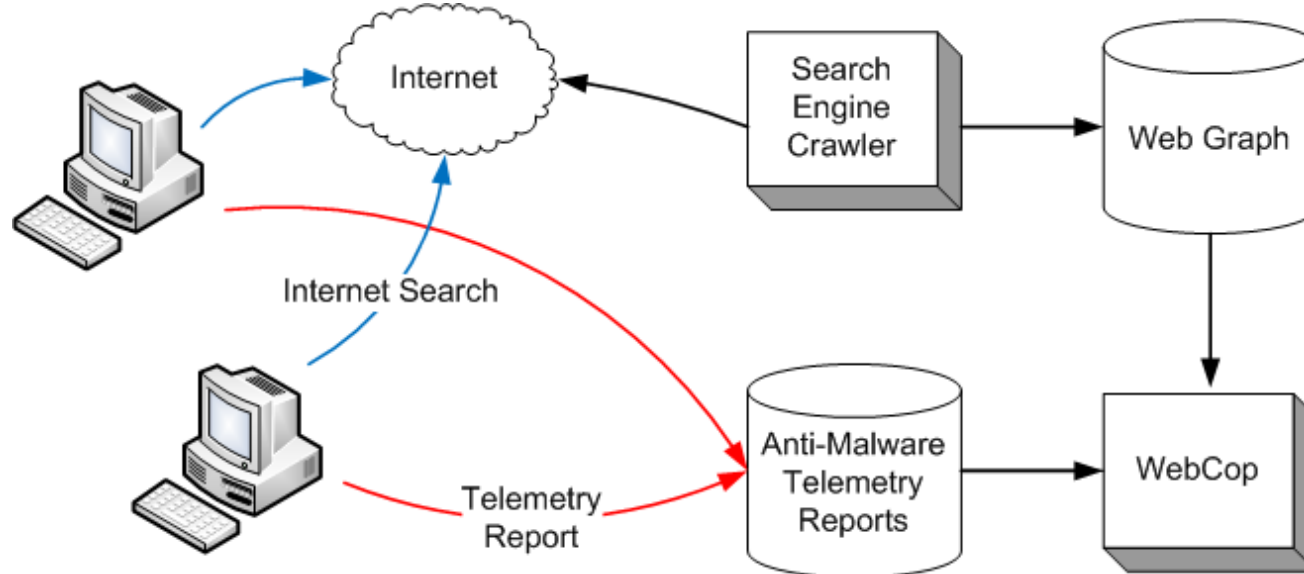
# WebCop Solution

- ❑ Bottom-Up Approach
- ❑ Anti-Malware reports indicate malware distribution pages
- ❑ Crawler discovers all web pages linking to the malware
- ❑ Direct Links
- ❑ Additional Goal:
  - ❑ Identify neighborhoods of malware on the web





# WebCop System



# WebCop Advantages

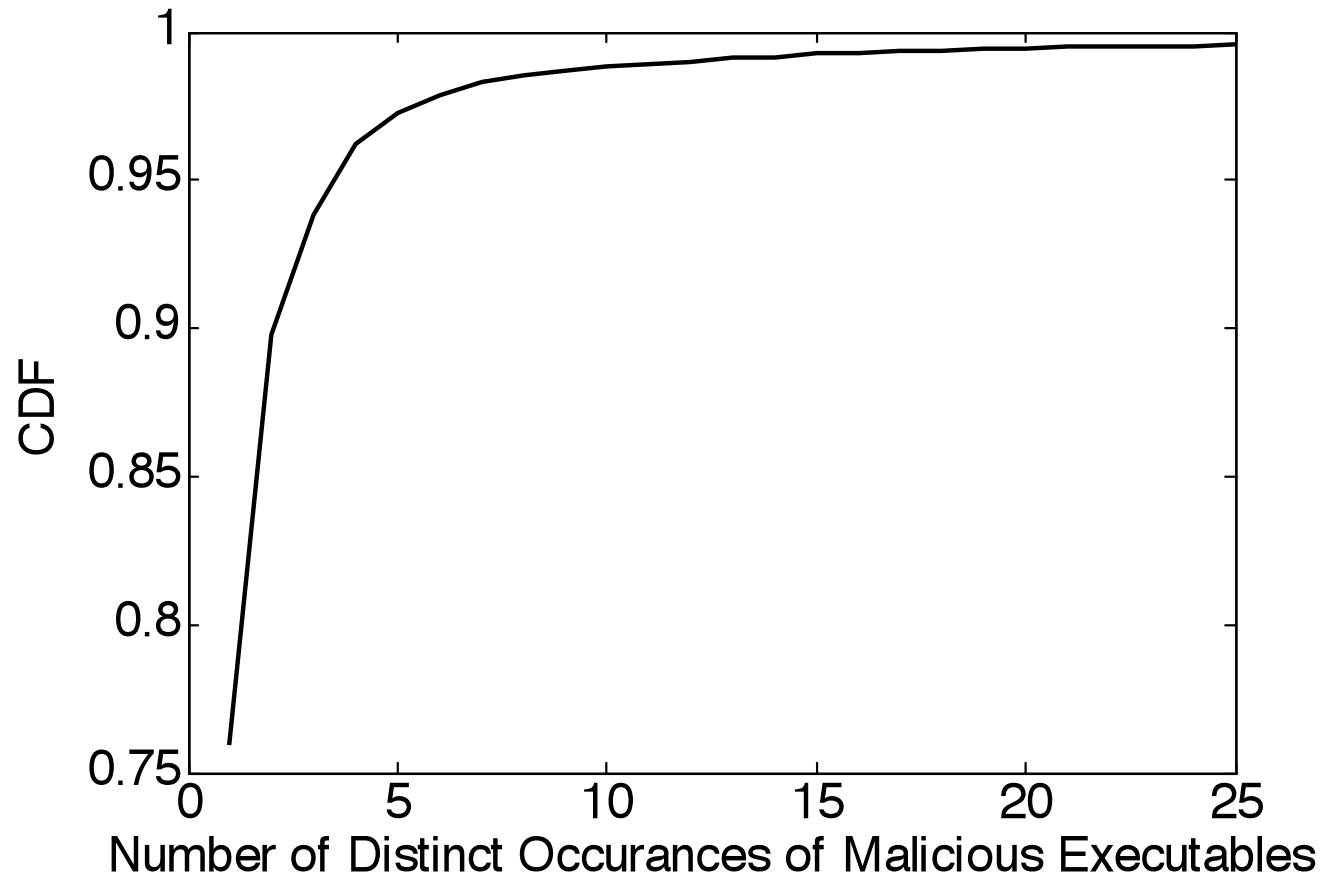


- ❑ WebCop only deals with hard classifications
- ❑ Distributed worldwide sensor network
  - ▣ Millions of clients
- ❑ Targeted detection
- ❑ AM service detects malware running on native OS
  - ▣ Not in a VM
  - ▣ Malware will not try to hide
- ❑ Users input all UI interactions

# Telemetry Reports

- Automatically submitted to backend
  - ▣ File is downloaded from internet
  - ▣ Malware detection
  - ▣ Unknown file was not signed by a trusted entity
- Reports include
  - ▣ Distribution page URL
  - ▣ File Hash
- Most recent 1 million distinct labeled URLs through end of May 2009
  - ▣ 837,882 Malware URLs
  - ▣ 162,118 Benign URLs
- Telemetry reports from a URL are usually only seen during a one month period
  - ▣ Only 8.7% overlap of malicious distribution URLs between April and May, 2009

# Occurrences of Executables

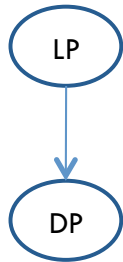


# Link Analysis

- Web graph from June 1, 2009
- Intersecting distribution pages
  - ▣ Occurs in both AM reports and web graph

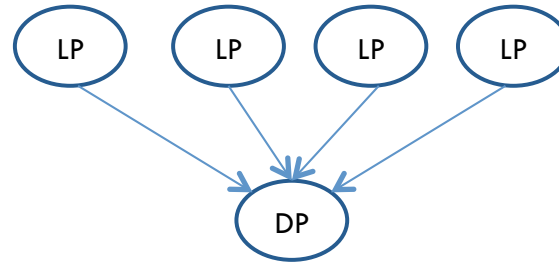
Measure	Count
Number of intersecting malware distribution pages	10,853
Number of malware landing pages	391,893

# Median Malware Topologies



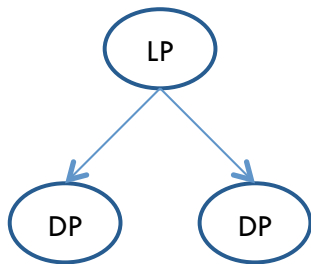
2984

Single Edge



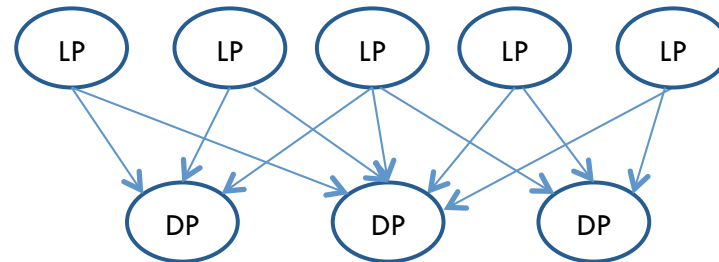
2498

Fan-In



388

Fan-Out



547

Complex

# Malware Subgraph Statistics

Measure	Topology	Median	Average
Number Landing Pages	Fan-In	4	31.3
	Complex	5	33.7
Number Distribution Pages	Fan-Out	2	3.5
	Complex	3	4.9
Number Edges	Fan-In	4	31.3
	Fan-Out	2	2.9
	Complex	11	72.2

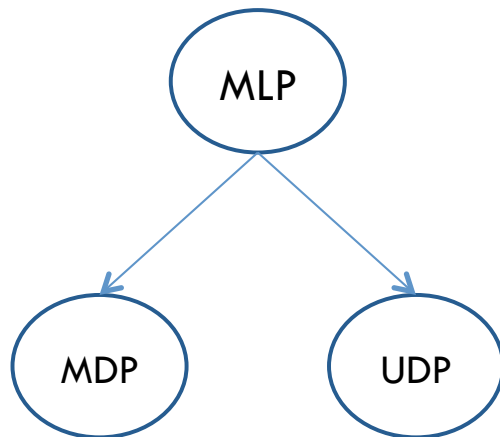
# Comparison with Production System



- Drive-by detections from April 6 – June 1, 2009
- Little overlap
  - ▣ 2 matching distribution pages
  - ▣ 0 matching landing pages
- Complementary to current production system
- Lists can be combined



# Locating Potential New Malware

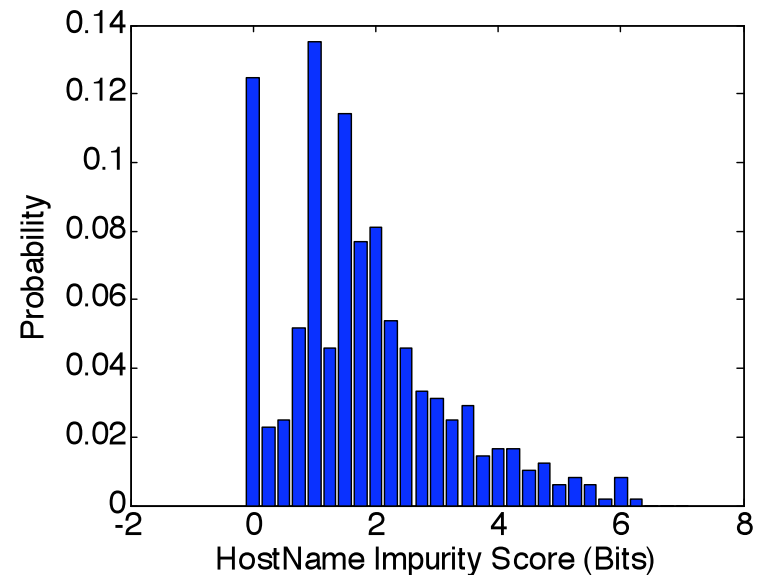


Unknown Executable Two-Hops  
Away from Malware

- Neighborhood graph
  - ▣ Unknown distribution pages (UDP)
- Identified **346,084** unknown distribution pages
- 32 suspicious pages for each labeled malware pages
- Suspicious Executables
  - ▣ Download and scan
  - ▣ More sophisticated automated analysis
  - ▣ Rank for analysts

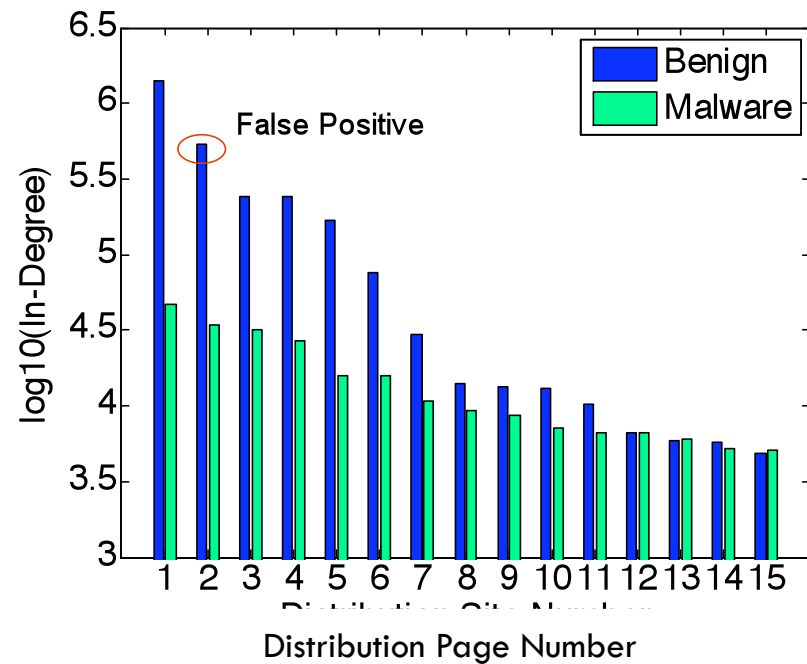
# HostName Impurity

- How often do landing and distribution pages share same hostname?
- HostName impurity score
$$hi(n) = -\sum_j P(\omega_j) \log_2 P(\omega_j)$$
- $W_i$  - fraction of nodes sharing same hostname
- Low score, most nodes in neighborhood share same hostname



# Discover AM False Positives

- Use graph topology
- In-Degree
  - ▣ Total number of edges where node is the head
- Malware distribution page with 540K links



# Will WebCop Work in Production?

Telemetry Reports	Malicious Intersecting Distribution Pages	Malicious Landing Pages
May 2009 Only	2,763	158,333
March – May, 2009	4,633	212,688
Most Recent One Million Reports	10,853	391,893

- ❑ Queues of distribution pages (e.g. 2 or 3 months)
- ❑ Telemetry reports only seen for a short time
- ❑ Find large number of new landing pages each month

# Conclusions



- WebCop provides
  - ▣ Targeted, bottom-up approach for detecting malware landing pages on the internet
  - ▣ Large scale evaluation of malicious internet neighborhoods composed of direct links
  - ▣ New way to detect false positives in an AM service using the internet web graph
  - ▣ New method to discover potential malware

# WEBCOP: LOCATING NEIGHBORHOODS OF MALWARE ON THE WEB

- Jay Stokes  
*Microsoft Research*
- Reid Andersen
- Christian Seifert
- Kumar Chellapilla  
*Microsoft Search*

# Microsoft Security Essentials



## □ Privacy Statement

- “... , by accepting this privacy statement, you agree to send reports to Microsoft”
- “... reports include information about ... cryptographic hash, ...”
- “... might collect full URLs ...”