Knowing Your Enemy: Understanding and Detecting Malicious Web Advertising

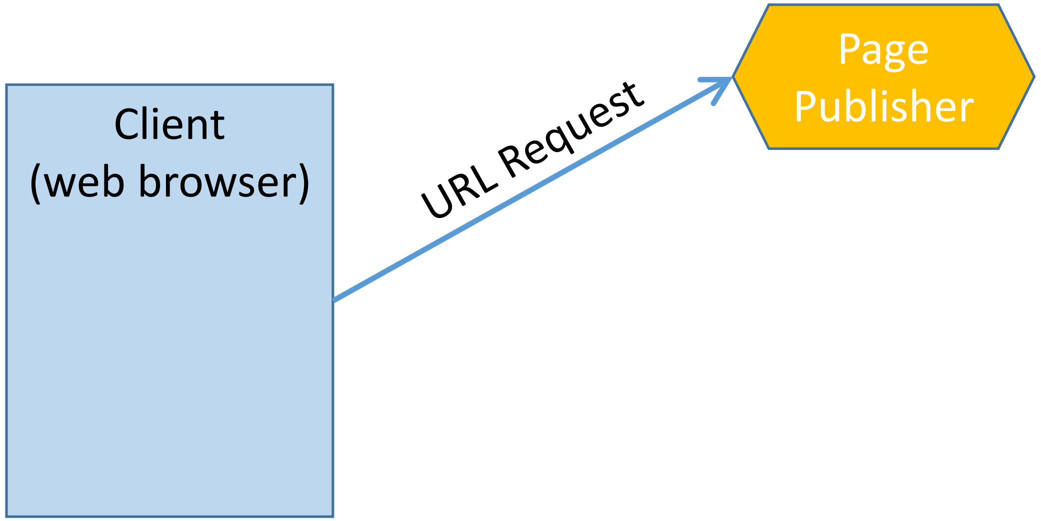
Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang

In *Proc. ACM CCS*, 2012

Dr. Kevin W. Hamlen

Language-based Security

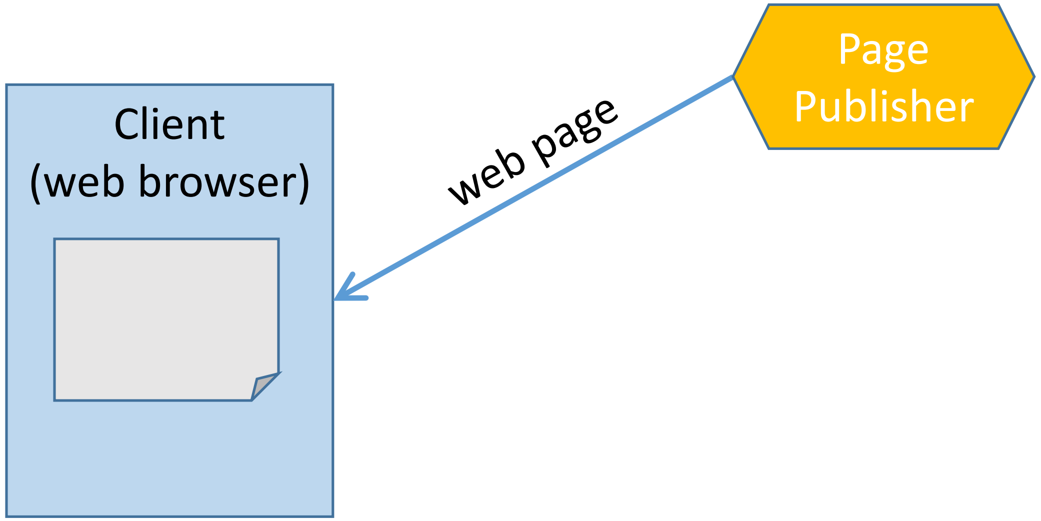
Loading a Web Advertisement


Client
(web browser)

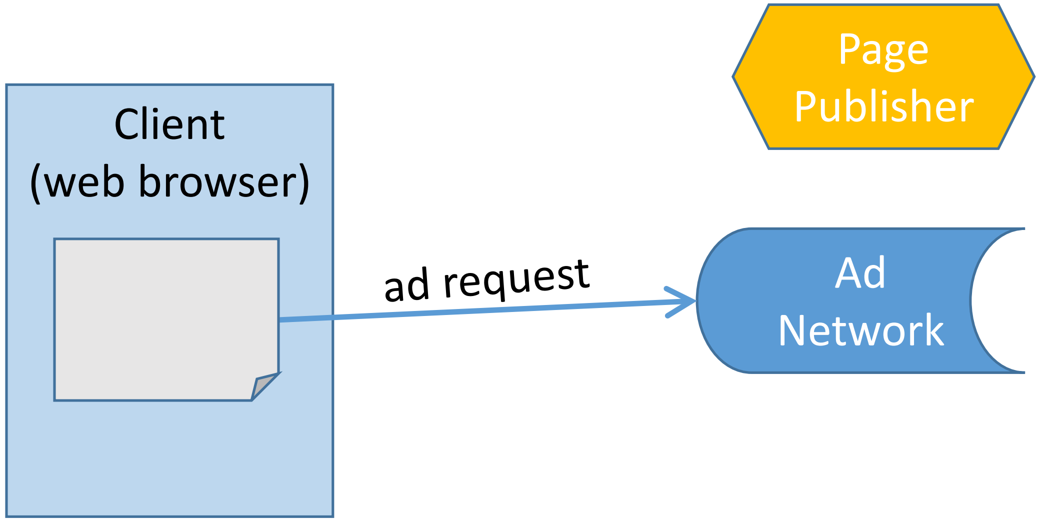
URL Request

Page
Publisher

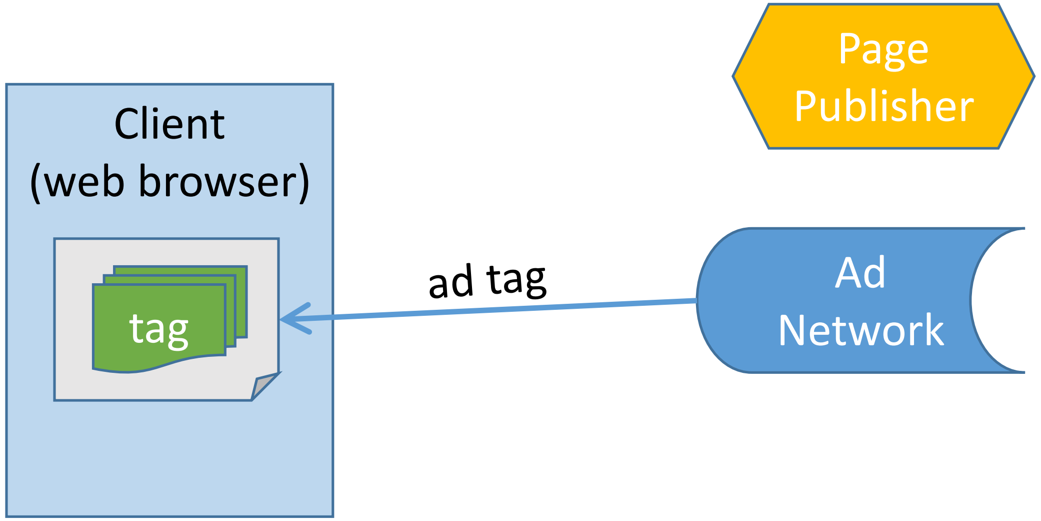
Loading a Web Advertisement



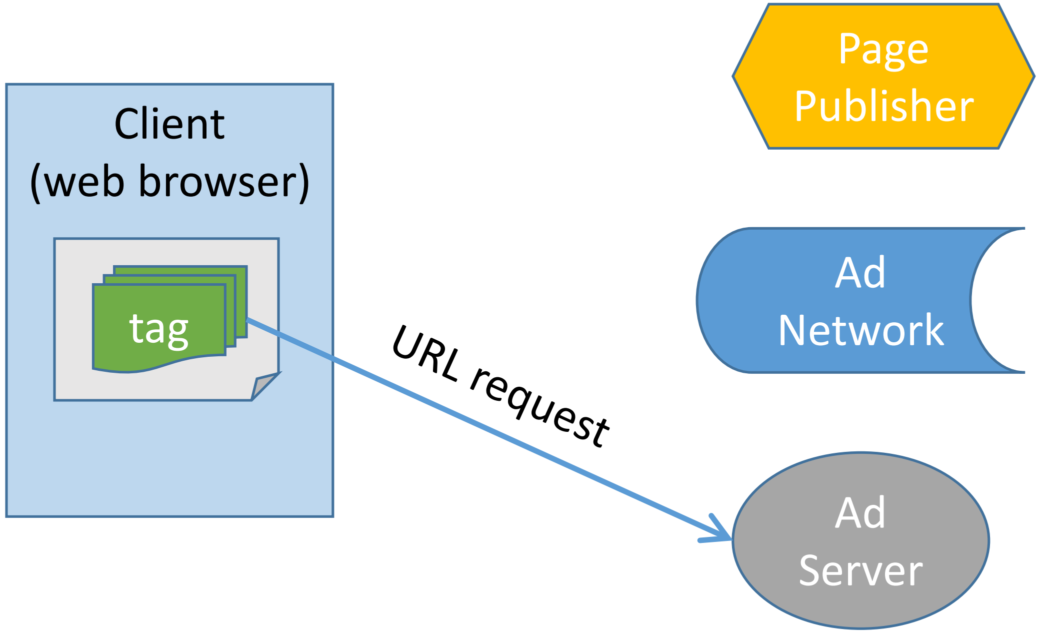
Loading a Web Advertisement



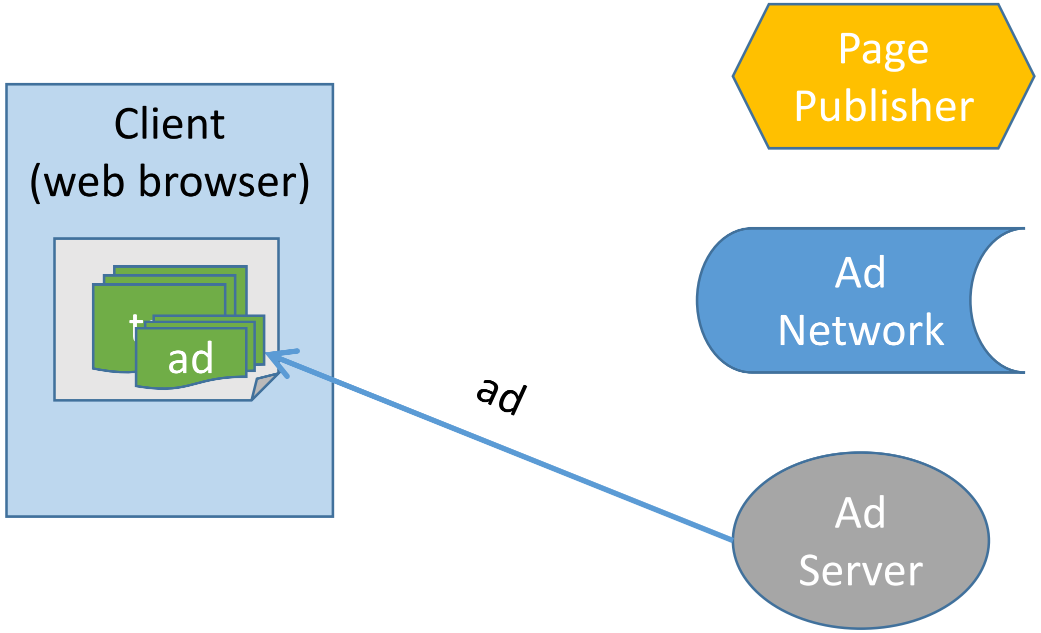
Loading a Web Advertisement



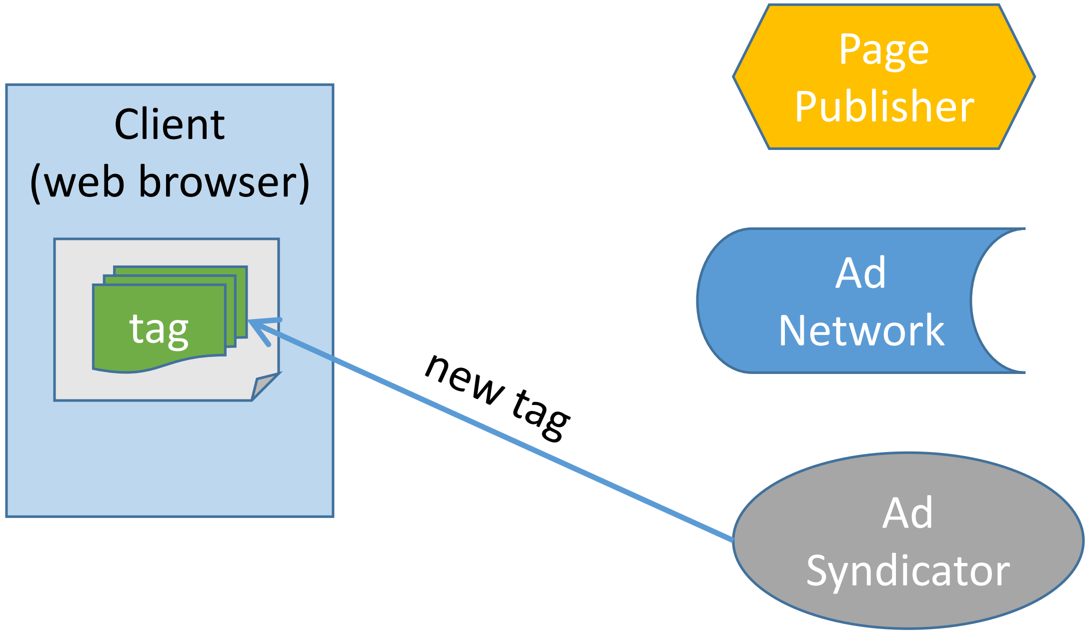
Loading a Web Advertisement



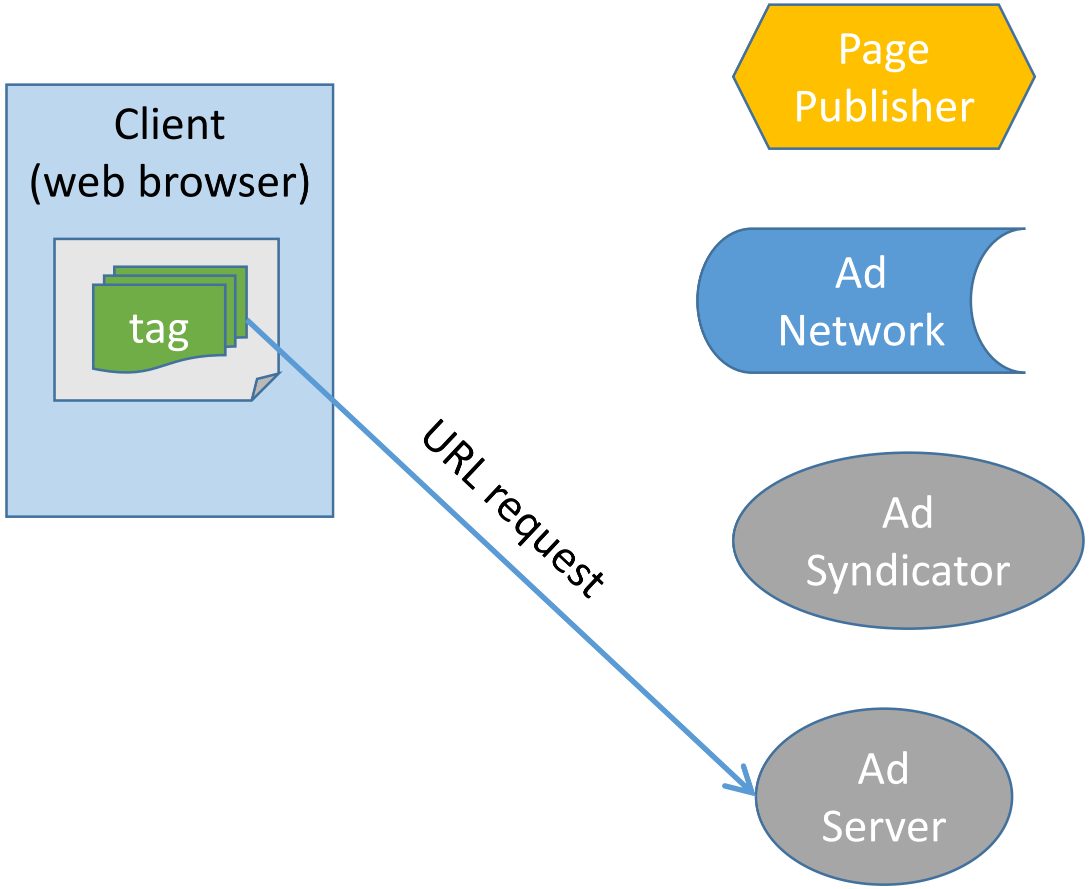
Loading a Web Advertisement



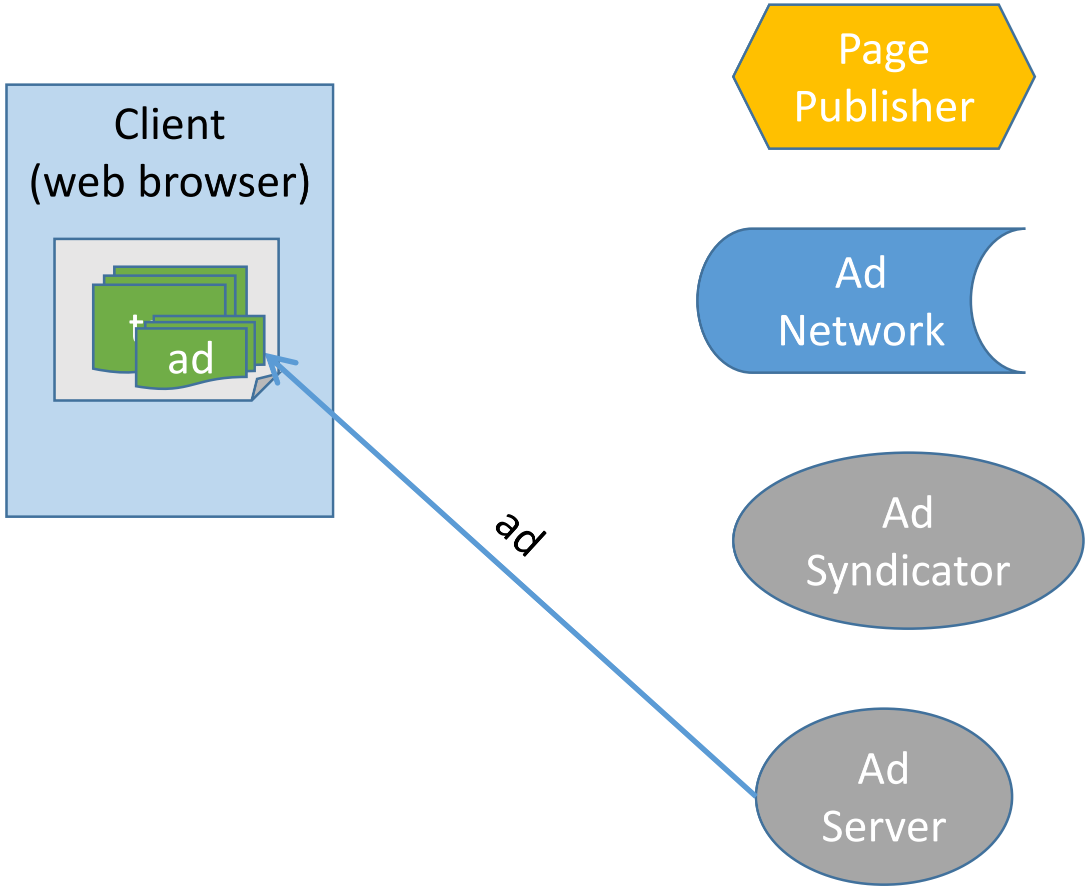
Ad Syndication



Ad Syndication



Ad Syndication



Client
(web browser)

ad

ad

Page
Publisher

Ad
Network

Ad
Syndicator

Ad
Server

Malicious Advertisements

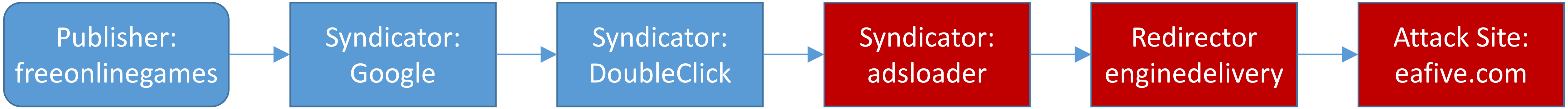
- Various goals
  - Click fraud
    - Accrue unmerited ad revenue
    - *pay-per-impression* – advertisers pay by number of URL requests for their ads
    - *pay-per-click* – advertisers pay by number of people who click on their ads
    - malvertisements trick browsers into sending URL requests that are never displayed
    - malvertisements redirect clicks to ads, generating false clicks
  - Scams / Phishing
    - Collect private user information (credit card info, usernames/passwords, etc.)
    - Impersonate legitimate sites (e.g., your bank)
    - Harvested info used in other criminal activities (identity theft, spam, etc.)
  - Drive-by-download
    - Infect client machine with malware
    - Exploit browser vulnerabilities
    - Infections facilitate other attacks (botnet zombies, ransomware, all of the above attacks)

Two Steps of Malvertising

- Enablers
  - ad syndicators
  - malicious ad tags
  - malicious ad networks
  - malicious redirectors
  - malicious ad servers
- Payloads
  - the actual malicious code that gets delivered
  - the actual malicious sites to which the client is ultimately redirected
- This paper: Measure and detect the *enabler* half of this picture.
  - Payload detection based on stock products
  - Google Safe-Browsing and Microsoft Forefront

Example Fake-AV Malvertisement Campaign

- Drive-by-download attack
  - victim browsers redirected to fake-AV site
  - fake-AV malware pretends to detect viruses and sells fake fixes
- Impact
  - infected at least 65 publisher pages in summer 2011
  - infected pages include top Alexa sites (e.g., freeonlinegames.com)
- Delivery included five levels of indirection:



Publisher: freeonlinegames → Syndicator: Google → Syndicator: DoubleClick → Syndicator: adsloader → Redirector enginedelivery → Attack Site: eafive.com

Attacker Gambits

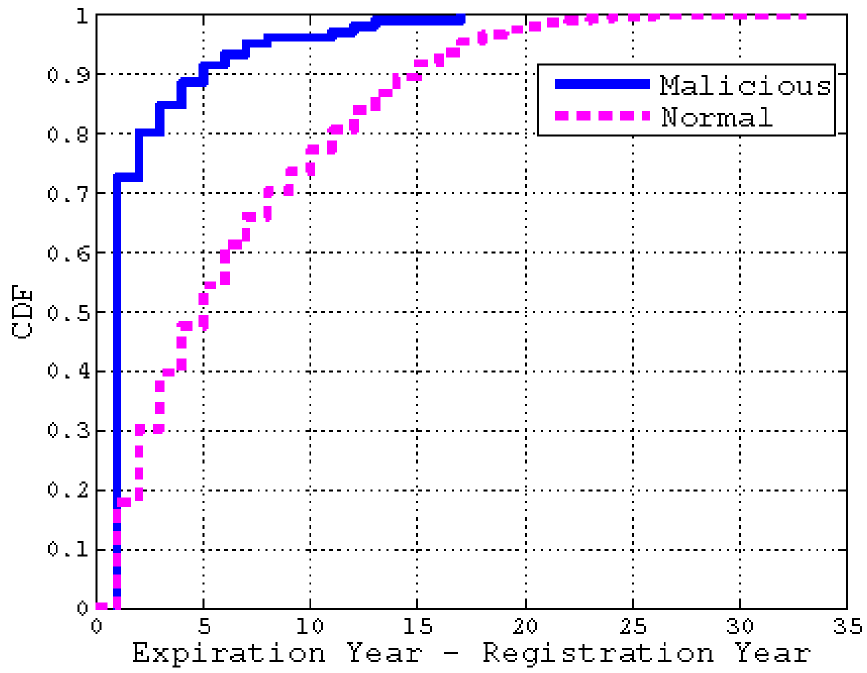
- Domain name impersonation
    - adsloader.com ≠ adloader.com
- Subversion of legitimate (often trusted) ad networks
    - GoogleServices, DoubleClick
    - over 24 ad networks total (!)
- Conditional redirection (cloaking)
    - adsloader.com redirects visitors at most once (per IP)
    - only IE agents redirected
    - empty referrers not redirected
- Honeynet evasion
    - enginedelivery withholds malicious content from Amazon EC2 IPs
- Conditional payload delivery
    - only IE6 received Fake-AV solicitation from eafive.com
- Domain and payload rotation
    - 16 different redirectors
    - 84 different fake-AV scanners

Measurement Study

- Crawl 90,000 web sites continuously for ~3 months (summer ’11)
- Infer redirection chains
  - HTML code (attributes containing URLs)
  - HTTP redirection (302-responses)
  - JavaScript net accesses (mine script texts for domain names of requests)
  - 24.8M chains and 21.9M URLs collected
- Identify malicious nodes
  - detection based on stock products (Google Safe Browsing, Micosoft Forefront)
  - Paths containing malicious nodes are malicious paths.
  - Descendants of malicious nodes might not be malicious!

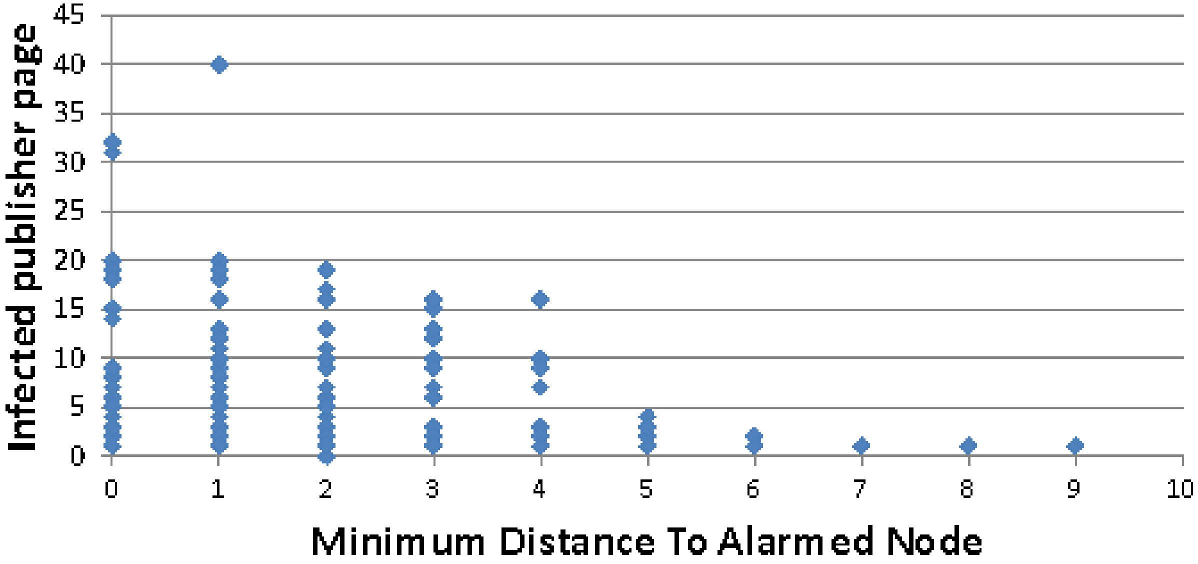
Distinguishing Features of Malicious Nodes

- Node roles: known publisher / known ad-node / unknown
  - non-malicious paths: 93.1% known
  - malicious paths: 8.4% known
- URL patterns (Example: /showthread.php?t=12345678)
- Short domain name life expectancies
- Short, diverse associations w/publishers



Syndication and Redirection Cloaking

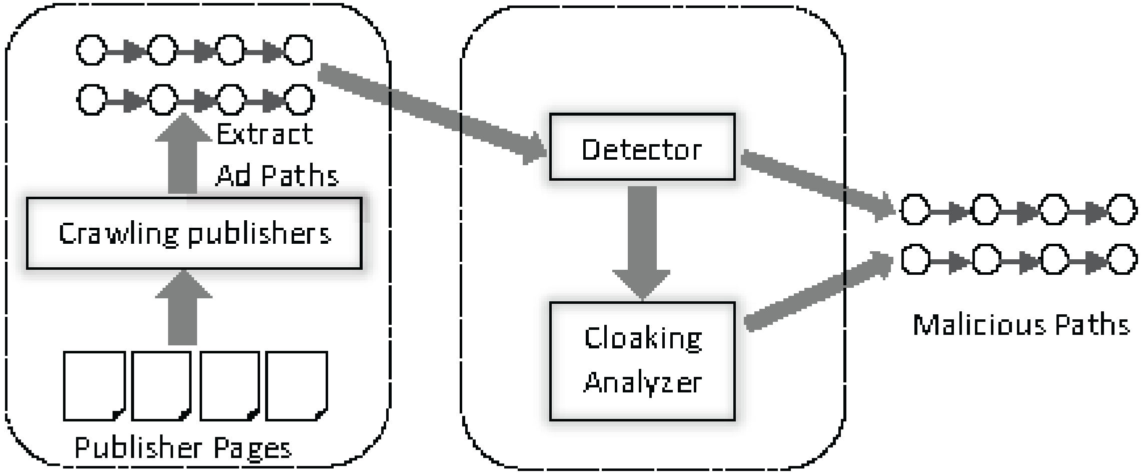
- Syndication Rates
  - 64% of all paths involve syndication (multiple ad networks)
  - 86 well-known networks compromised
  - 92% of DoubleClick-facilitated attacks are via syndicated paths
- Redirection cloaking
  - Malvertisement paths tend to be longer due to redirection cloaking
  - Early malicious redirectors tend to be involved in many attacks

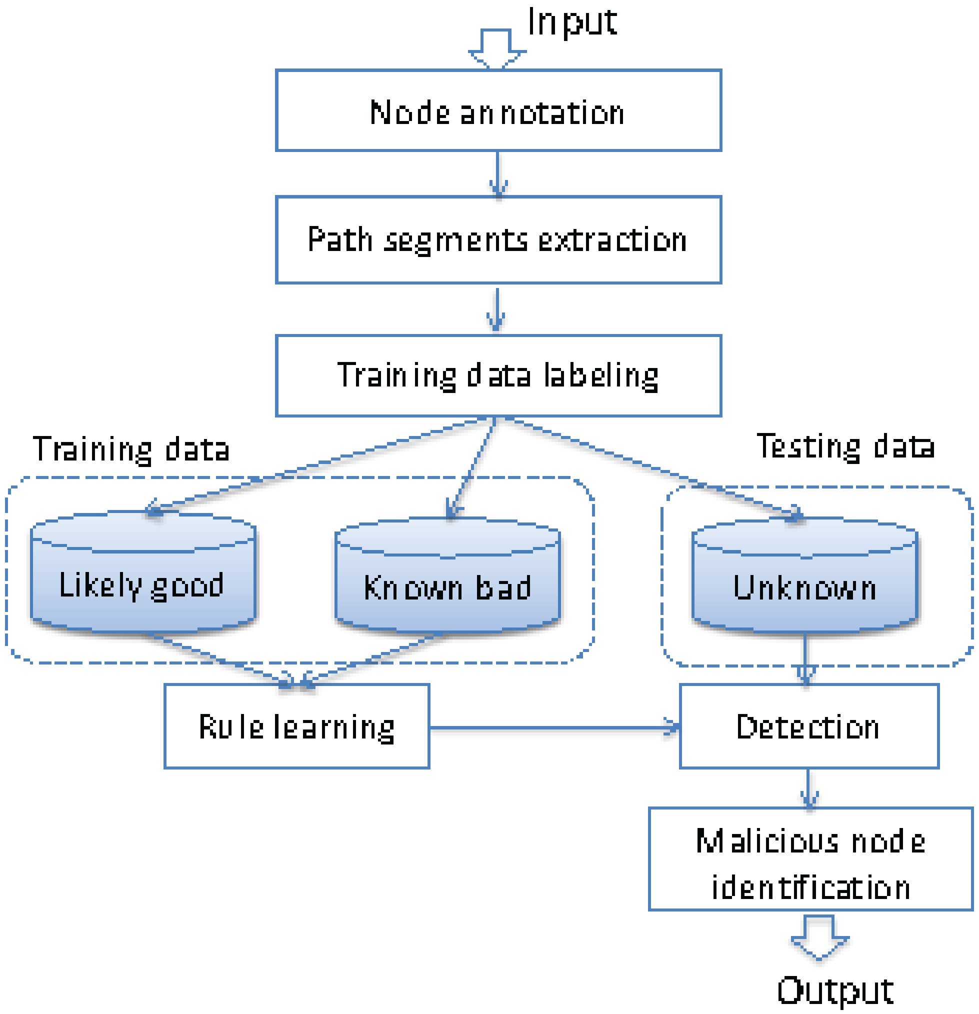


From Measurement to Detection

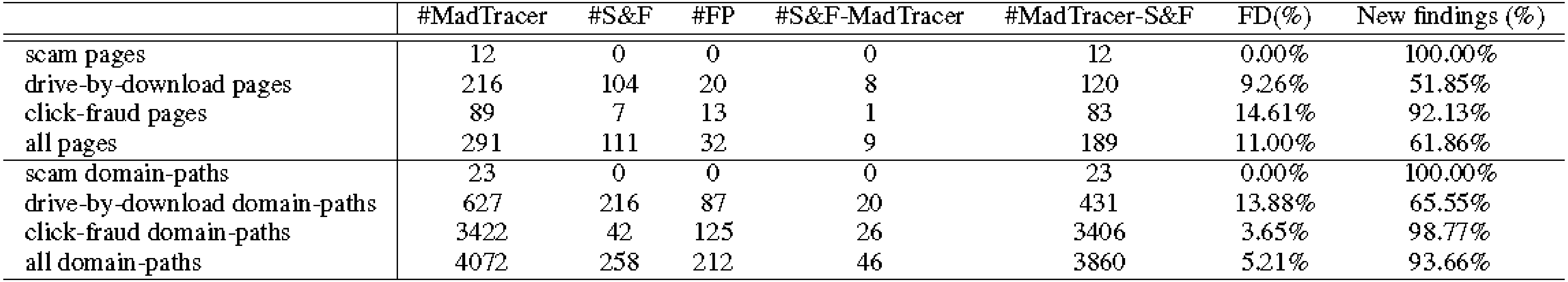
- Goal: Use path statistics to reliably detect malvertisements
- Major finding:
    - Blindly applying heuristics to full redirection paths doesn’t work well.
        - too slow, difficult to implement
        - too many false positives
    - But heuristically identifying short, suspicious path segments works very well.
        - faster, easier to implement
        - malicious nodes tend to be clustered along the path
        - node roles in the segments are key

MadTracer Architecture





MadTracer Detection Results



	#MadTracer	#S&F	#FP	#S&F-MadTracer	#MadTracer-S&F	FD(%)	New findings (%)
scam pages	12	0	0	0	12	0.00%	100.00%
drive-by-download pages	216	104	20	8	120	9.26%	51.85%
click-fraud pages	89	7	13	1	83	14.61%	92.13%
all pages	291	111	32	9	189	11.00%	61.86%
scam domain-paths	23	0	0	0	23	0.00%	100.00%
drive-by-download domain-paths	627	216	87	20	431	13.88%	65.55%
click-fraud domain-paths	3422	42	125	26	3406	3.65%	98.77%
all domain-paths	4072	258	212	46	3860	5.21%	93.66%

Conclusions

- Malvertising is a significant threat to the internet revenue model
  - much of the internet funded by advertising (billion-dollar industry)
  - at least 1% of top sites fell victim to malvertising campaigns in 2011
- Simple detection approaches don’t work
  - IP black-listing fails because malicious campaigns rotate servers too quickly.
  - Honeypotting is frustrated by highly selective attacks.
  - Full referrer paths of many legitimate ads display “suspicious” characteristics (long path lengths, unknown nodes, short domain lifetimes, etc.). This can result in high false positive rates.
- But detecting short, malicious sub-paths works well
  - Malicious nodes operate in close proximity on a malicious path.
  - Possible to identify node roles in these sub-paths.
- Open problem: It’s still an arms race.
  - As these heuristics catch on, malvertisers will adopt new topologies to counter them.
  - The race will continue as defenders compensate with new heuristics.

Discussion Questions

- Is there a principled answer to the malvertising problem?
  - language-based security?
  - formal methods?
  - browser security?
  - script analysis?
- What about economic/financial solutions?
  - better revenue models?
  - incentive schemes?