Generating a Privacy Footprint on the Internet

Balachander Krishnamurthy, AT&T Labs – Research

Craig E. Wills, Worcester Polytechnic Institute

To be presented at the ACM Internet Measurement Conference, October 2006.

Generating a privacy footprint on the Internet

You don't have any privacy on the Internet. Get over it.

(Scott McNealy, CEO, Sun Microsystems)

Er, *former* CEO

Privacy footprint

- Various daily interactions done on the Web (commerce, banking, email, search etc.)
- Some require supply of private information
- Web sites use many techniques to track users (1x1 pixel Web bugs, cookies)
- Aggregators (doubleclick, googlesyndication: tracking across sites)
- Is there diffusion of user information across unrelated sites?
- Measure of dissemination of user-related information: privacy footprint

Does privacy matter?

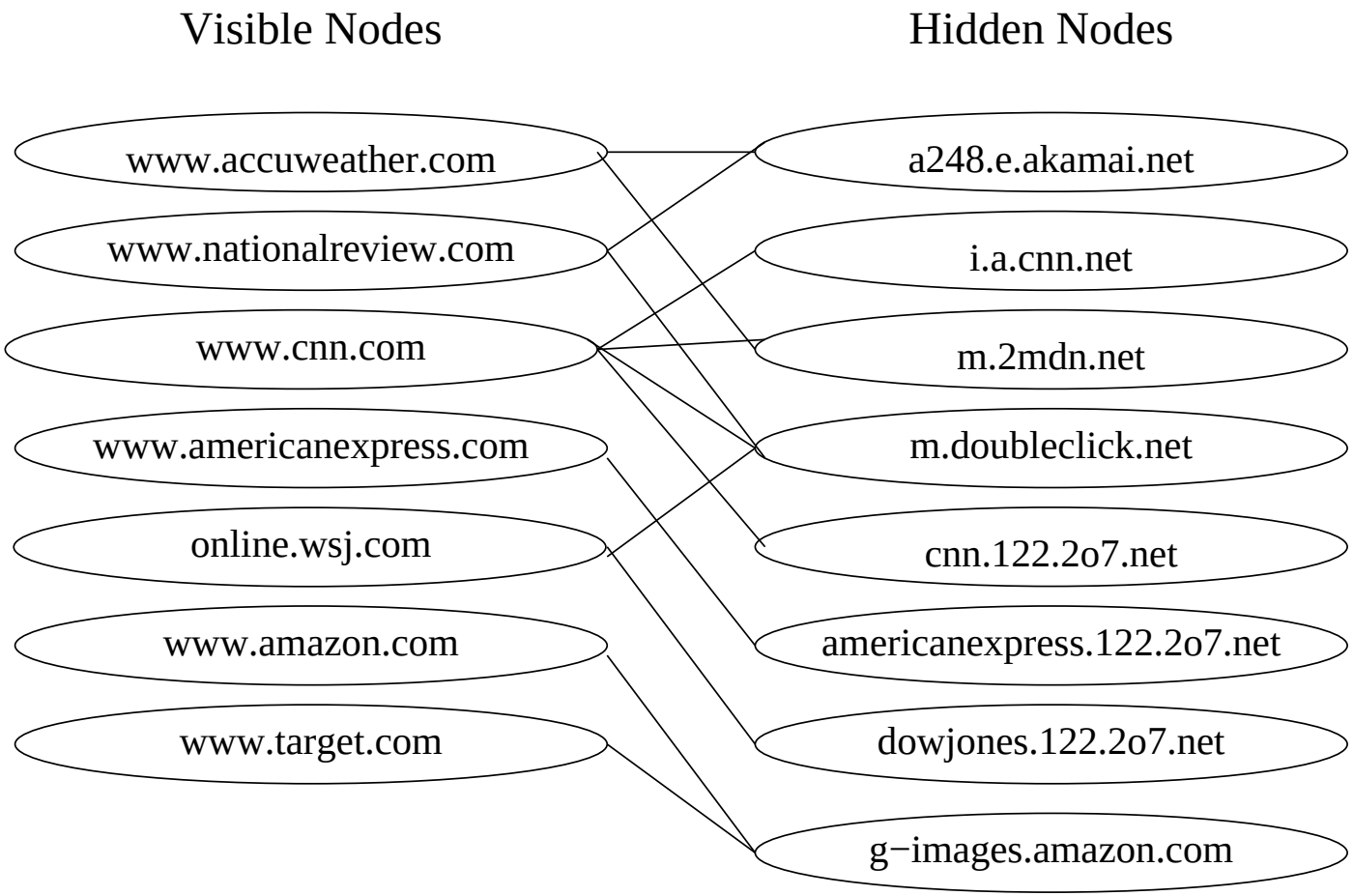
Maybe ask AOL customers?

- Depends on the information disseminated
- Depends what data collectors might do with it
- What can we do about it?

But first we need to know *what* information is being diffused and could be tracked, by whom, how, and then see if it is blockable.

Study

Examine connections between visible nodes (servers explicitly visited) and hidden (visited as by-product)



Mechanics of data collection

- Use similar technique to earlier Cat and Mouse work (WWW'06)
- *Pagestats* – Firefox extension records request/response, objects fetched
- Interface allows list of URLs in file to be fetched serially
- 1075 Web sites in over a dozen Alexa categories chosen as visible nodes
- Extract hidden nodes corresponding to each visible node
- Examine how cookies are used
- Examine broad set of sites and also narrow examination to interesting subset of Web sites that raise more privacy concerns (fiduciary sites)

Node association

Two visible nodes are *associated* if accessing them results in accessing the same hidden node.

Association can be due to several reasons:

1. server: Identical server name (`www.google-analytics.com`)
2. domain: Aggregated by merging hidden nodes with same 2nd-level domain names. E.g. `timecom.112.2o7.net` and `msnbcom.112.2o7.net`
3. adns: Aggregated by merging hidden nodes that share same ADNS. e.g. `google-analytics.com` and `googlesyndication.com` have the same ADNS, but different domains.

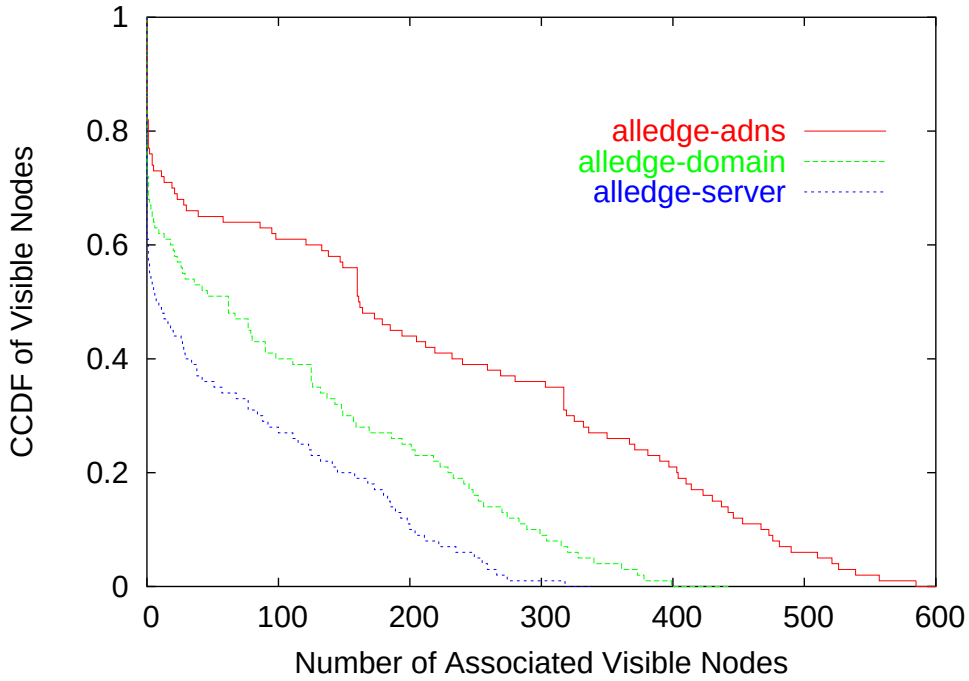
Grouping by ADNS: empirical examination of false positive

Maybe many organizations just outsource DNS?

- Examined servers in top-15 frequently occuring ADNSs (56% of ADNS that handle multiple servers)
- Used *dig*, WHOIS, IP address clustering, and *traceroute* as spot checks
- Error rate: around 5% (servers from different organizations using the same ADNS)
- Many cases where what appeared to be a server in one organization (e.g. `lads.myspace.com`) was actually a DNS CNAME alias to a server (e.g. `lads.myspace.com.edgesuite.net`) in another organization (e.g. Akamai).

Results: Node association

CCDF Number of other visible nodes associated with each visible node



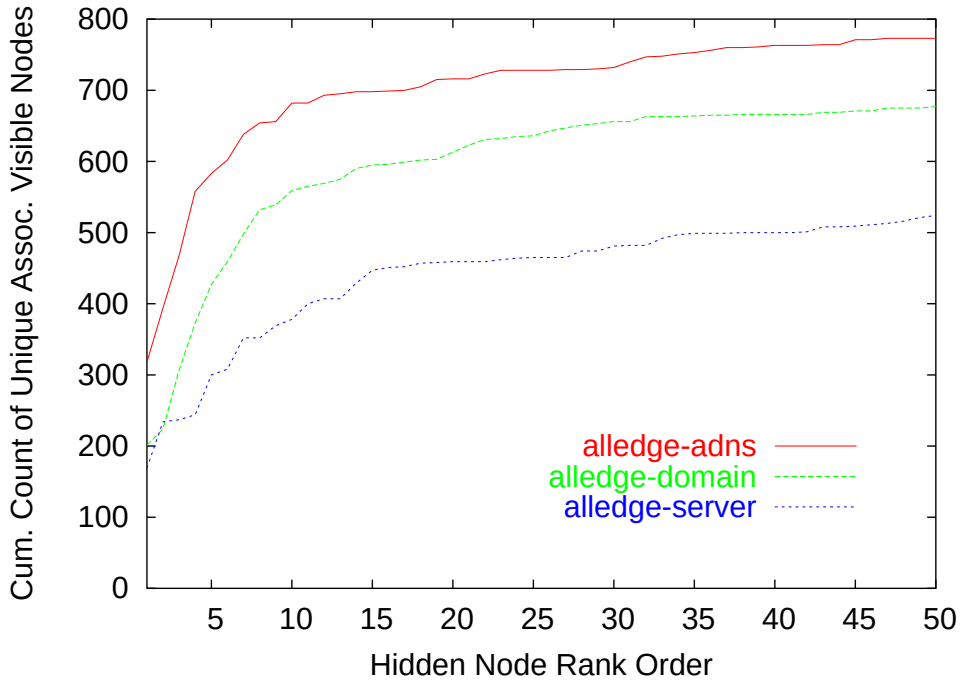
Y-axis: Degree of association: 61% server, 72% domain, 82% ADNS

X-axis: Single visible node's maximal association: 338 (31%) server, 443 (41%) domain, 609 (57%) ADNS

Strong relationship among popular sites visited—behind the scenes

Association among visible nodes

Cumulative count of unique associated visible nodes based on hidden node rank order



Some visible nodes are associated via more than one hidden node.

Top-10 ADNS nodes are connected to 682 (63%) visible nodes.

Privacy footprint

Summarize interconnections for a given set of sites

Metrics of interest:

1. Number, percentage of visible nodes associated with at least one other visible node
2. Mean/median/max of number of visible node associations for any given visible node
3. Top-n rank ordered hidden nodes contribution to associations

1075 Alexa Web sites privacy footprint Apr06 vs. Oct05

Timeframe/ Approach	Visible Nodes w/ Assoc's (%)	Number of Assoc's for a Visible Node			Assoc's Via Top-10 Hidden Nodes
		Med.	Mean	Max	
apr06/adns	879 (82)	225	247	609	682
apr06/domain	779 (72)	125	144	443	559
oct05/adns	853 (78)	121	170	527	585
oct05/domain	718 (66)	80	98	347	456

# of visible nodes with associations up by 5% but # of mean associations for visible nodes up by 50%. Seemingly unrelated websites traversals can be more readily correlated.

Fiduciary sites

Some sites raise more serious privacy concerns: ones that users have personal fiduciary connections. e.g. credit, financial, insurance, mortgage, travel etc.

Selected 81 such sites

Examined cookie information by narrowing focus to hidden nodes fetched when cookies are required.

Privacy Footprint of 81 Fiduciary-Related Sites

Edges/ Approach	Visible Nodes w/ Assoc's (%)	Number of Assoc's for a Visible Node			Assoc's Via Top-10 Hidden Nodes
		Med.	Mean	Max	
alledges/adns	52 (64)	11	11	32	40
alledges/domain	41 (51)	6	7	25	32
cookie/adns	47 (58)	10	10	32	38
cookie/domain	37 (46)	7	7	20	30

Fooprint smaller than larger Alexa set but most common hidden node domains are similar: `doubleclick.net`, `atdmt.com`, `2o7.net`.

How to block privacy leakage?

- AdBlock: can reduce mean number of associations by half in ADNS and by two thirds in domain approach; larger drop when narrowed to cookies.
- Concern not eliminated as AdBlock was designed to block ads, not tracking
- Blacklisting top hidden nodes: similar to AdBlock results as associations are with multiple hidden nodes

Conclusions

- Privacy is an increasingly serious concern
- We have scratched the surface of the extent of largely hidden data gathering
- We illuminate the extent of privacy sharing
- Concerns vary with set of sites visited
- Goal is to prevent leakage for important subsets of sites visited