Broadening DNS Research: beyond just DNS anonymization

(work in progress)

John Heidemann
joint work with Liang Zhu
USC/ISI and USC/CS Dept.

22 October 2012

how can students do research on DNS?

instrument a small, local server?

data not necessarily representative

intern at (large company or operator)?

challenging to continue work when summer’s over; difficult for others to build on results

talk to the right folks?

perhaps in 1990s, but much tougher today

Our Goal

- broaden field of DNS researchers
- with sharable DNS data
  - combine technical and legal methods
  - address privacy questions
  - support IRB (Institutional Review Board) oversight => clean for academic use
- ultimately, accelerate DNS evolution

Challenge: Privacy Concerns

- what if data shows (important figure) is browsing (embarassing site)
  - Sergey Brin … Google for dummies
  - Larry Ellison … 99only.com
  - Felix Baumgartner … Jolt Cola
  - (your example goes here)
- general privacy concerns
  - given enough data and effort, often something pops out
  - ex: 2006 AOL search data and searcher #4417749
- DNS-specific concerns
  - database-like use of DNS, ex: RBHL

Context: Growing Interest in Careful Sharing

- data sharing efforts
  - CRAWDAD.cs.dartmouth.edu: wireless datasets, NSF-supported
  - www.PREDICT.org: Protected Repository for the Defense of Infrastructure Against Cyber Threats, DHS-supported
  - SIE.isc.org: Security Information Exchange
  - ISC, CAIDA, USC, U. Mich, Ga. Tech., ICSI, and others…
- scrutiny of and guidelines for sharing
  - interest in sharing guidelines *and* more open data in academia (ACM Internet Measurements Conference)
  - role of IRB oversight in network research
  - The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research (Dittrich and Kenneally, eds.)
- can we bring these together?

Our Approach: Combined Technical and Policy

- technical
  - aggregation
  - anonymization
  - separation
- policy
  - legal agreements
  - researcher-to-data
- best practices

- builds on existing work
- work-in-progress

(but hope that combination provides some new insight)

Aggregation for Anonymity

- built-in aggregation via recursive resolvers
  - replace end-user IP addresses
  - aggregate data from many users

  ⇒ part of anonymization
- effects depend on observer's place in hierarchy
- open questions
  - can we estimate degree of aggregation?
  - can we identify (and filter when necessary) streams with insufficient aggregation?
  - what is the hierarchy, in practice?

Anonymization

- lots of collection tools
  - tcpdump, dnscap+dnsqr, nmsg, LANDER, etc.
- fewer anonymization
  - tcpmkpub (ISCI), U. Md. extensions for DNS
- our approach
  - building on ISCI/U. Md. approach
    - anonymize each DNS label (+salt) via hash
    - prefix-preserving anonymization of IPs (cryptopan)
    - hash ID field
  - hashes don't fit in pcap => output to simple text format
  - applies to queries and replies (examine each reply)

Attacks on Anonymity

statistical attacks

- stream with mix of frequent and infrequent labels
- adversary can identify frequent labels
  - www.
  - .com
- very powerful attack, *but* probably doesn't show much that is a suprise

injection attacks

- assume an adversary
  - can inject arbitrary queries
  - can observe anonymized results
- very powerful attack if part of injection is not anonymized
  - unusual query, special time, etc.
  - effectively creates a side-channel

Controlling Access

- control access to traces to manage side-channel attacks
- legal agreement to access data
  - cannot attempt to de-anonymize
  - cannot redistribute data
- researcher-to-data
  - have researcher do analysis on provider's computers
  - provider has better control over local security and can audit analysis

Separating Access

- risk comes from saying “A asked for B”
- much less sensitive
  - “A asked for something”
  - and “someone asked for B”
  - and “reply for B is C”
- idea: separate streams
  - separate request and reply streams
  - remove linkage information (timing and IDs)
  - prohibit external linkage
- separate streams answer some research questions
- (work-in-progress)

Benefits

- enable new research
  - broader set of groups
  - new questions
- supported by publically available datasets
- perhaps sharing between commercial groups?
- open question: what questions can be done…
  - …with anonymized data only?
  - …*started* with anonymized, then moved?
  - what can definitely not be done

Alternatives

- many existing tools do DNS capture
  - our anonymization as optional back-end?
- some existing anonymization tools
  - tcpmkpub + U. Md. extensions
- regardless of choice of tool,
  sharing policy and IRB approaches benefit all

Broadening DNS Research

- work-in-progress
- combining
  - complete anonymization
  - stream separation
  - policy and access control
- …to enable access to DNS data
- http://www.isi.edu/ant/