Web Monitoring & Content Filtering

CSH6 Chapter 31

“Web Monitoring and Content Filtering”

Steven Lovaas

Topics

- Introduction
- Terminology
- Motivation
- General Techniques
- Implementation
- Enforcement
- Vulnerabilities
- The Future

Introduction

- Internet = cesspool?
- Objectionable content for some people
- Choice to monitor and/or control access
- Controversial
  - Privacy
  - Free speech
  - "Censorware" vs "content filtering"
- Chapter avoids taking sides in arguments

Terminology

- Proxy: intermediary computer in access path
- Anonymizing Proxy: proxy that conceals identity of original user
- Privacy-enhancing technologies (PET): maintain privacy of network access
- Encryption: concealing content
- Mixing networks: encrypted content sent to intermediary server to extract origin
- Onion routing (TOR): strip ID info from originating router at multiple steps

Motivation

- Preventing Dissent
- Protecting Children
- Supporting HR Policies
- Enforcing Laws
- National Security

Preventing Dissent

- Repressive governments (e.g., PRC, N. Korea)
  - Deprive citizens of information
  - Stifle questions about doctrine
  - Repress challenges to authority
- Constitutional rule of law
  - US 1st Amendment
  - Cultural bias against repression
  - Resistance to public libraries' use of filtering

Protecting Children

- US government limits access by children to certain types of information
  - "Adult" materials
- Public schools & libraries
  - *In loco parentis*
  - Segment of population opposes exposing children to material to which the parents object
  - SCOTUS ruled many implementations unconstitutional
  - Child Internet Protection Act (CIPA) applies to schools receiving certain government funding
  - Also try to shield children against violating laws

Supporting HR Policies

- Avoid waste of time by employees
  - Although some studies show freedom to be positive influence on productivity
- Prevent *hostile work-environment*
  - Equal Opportunity Employment laws
  - But many types of material potentially offensive
  - Complicates task of determining rules
- Avoid legal liability for filtering
  - Pop-up warnings instead of blocks
  - AKA “Monitor & Notify”

Enforcing Laws

- Law-enforcement organizations (LEOs) rarely block access
- May monitor (surveillance)
- PET a serious problem
- 2014: FBI Director warns that phone encryption without back doors = serious problem for LEOs

FBI Director Calls On Congress To 'Fix' Phone Encryption By Apple, Google

Posted: 10/16/2014 2:46 pm EDT | Updated: 10/16/2014 2:59 pm EDT

0 Share Tweet LinkedIn Email Comment

WASHINGTON -- FBI Director James Comey called Thursday for "a regulatory or legislative fix" for technology companies' expanding use of encryption to protect user privacy, arguing that without such a fix, "homicide cases could be stalled, suspects could walk free, and child exploitation victims might not be identified or recovered."

Comey said he understood the "justifiable surprise" many Americans felt after former National Security Agency contractor Edward Snowden's disclosures about mass government surveillance, but he contended that recent shifts by companies like Apple and Google to make data stored on cell phones inaccessible to law enforcement went too far.

"Perhaps it's time to suggest that the post-Snowden pendulum has swung too far in one direction -- in a direction of fear and mistrust," said Comey, speaking at the Brookings Institution in Washington in his first major policy speech since taking over the FBI 13 months ago.

http://www.huffingtonpost.com/2014/10/16/james-comey-phone-encryption_n_5996808.html

National Security

- NSA surveillance debate
  - PRISM & Edward Snowden
  - Terrorism
  - Constitutional law issues: 4th Amendment

https://www.eff.org/nsa-spying

ELECTRONIC FRONTIER FOUNDATION
DEFENDING YOUR RIGHTS IN THE DIGITAL WORLD

HOME | ABOUT | OUR WORK | DEEPLINKS BLOG | PRESS ROOM

NSA Spying on Americans

FAQ | How It Works | Key Officials | NSA Primary Sources | State Secrets Privilege | Timeline | Word Games

The US government, with assistance from major telecommunications carriers including AT&T, has engaged in a massive illegal dragnet surveillance of domestic communications and communications records of millions of ordinary Americans since at least 2001. Since this was first reported on by the press and discovered by the public in late 2005, EFF has been at the forefront of the effort to stop it and bring government surveillance programs back within the law and the Constitution.

TAKE ACTION NOW

Oppose NSA Mass Spying!

History of NSA Spying Information since 2005 (See EFF's full timeline of events here)

News reports in December 2005 first revealed that the National Security Agency (NSA) has been intercepting Americans' phone calls and Internet communications. Those news reports, combined with a USA Today story in May 2006 and the statements of several members of Congress, revealed that the NSA is also receiving wholesale copies of American's telephone and other communications records. All of these surveillance activities are in violation of the privacy safeguards established by Congress and the US Constitution.

General Techniques

- Matching the Request
- Matching the Host
- Matching the Domain
- Matching the Content

Matching the Request

- **Web request format:**

  Protocol://server.organization.tld/path-to-file.file-format

- **Lists of keywords to examine**
  - **Protocol**
  - **Server**
  - **Organization**
  - **Top-level-domain (tld)**
  - **Path-to-file**
  - **File-format**
- **Whitelist: to allow**
- **Blacklist: to reject**

Matching the Host: Servers

- Block list of servers
  - By IP address
    - ✓ May not be permanent
    - ✓ Block could affect innocent sites later
    - ✓ IP address may be shared by innocent sites
  - By name
    - ✓ Must update Domain Name System (DNS) mappings regularly
    - ✓ But organizations may use multiple names
    - ✓ Again, multiple organizations can share server name

Matching the Host: Intermediaries

- Many ways of locating site or page
  - Search tools (Google, Bing, etc.)
  - Portals (organized listings of sites for specific purpose – e.g., IA)
- Few organizations block search engines and portals
  - Too useful to block
  - Dictatorships have blocked them
  - Some countries block specific engines that violate local laws (e.g., Nazi stuff in France)

Matching the Domain

- Internet Corporation for Assigned Names and Numbers (ICANN)
  - Governs definition of new TLDs
- Long-standing requests for .xxx pornography domain
  - Make filtering easier
  - Comply with laws against access by children
  - Conservative religious groups feared legitimizing pornography
  - Porn providers disagreed on benefits
- Accepted in 2011

Matching the Content (1)

- String matching easy
  - But false +s & false –s are problems
- Exact matches (e.g., word "sex") can miss related terms ("sexy," "sexual"…)
- Generic matches (e.g., string "sex")
  - Can block harmless content (e.g., "Essex," "Sussex," "asexual"…)
  - Easily block legitimate use in context; e.g.,
    - ✓Survey asking for sex of respondent
    - ✓Scientific or technical text (biology, law…)
- Misleading URLs (phishing, pharming)
  - Whitehouse.com was porn site
  - Jumbled random text (2sfh.com/7hioh)

Matching the Content (2)

- Metastudy (2010)
  - Web-filtering effectiveness
  - 2001-2008 studies
  - Average accuracy 78%
  - Some improvement in 2007-2008: 83%
- Some products analyze content of Web pages
  - But HTTPS causes difficulties
  - Textual analysis doesn't work well
- Graphics analysis
  - Reference images
  - Anti-sex: skin tones, manual review, matching to known-bad images

Implementation

- Manual "Bad URL" Lists
  - Firewalls usually allow lists for blocking
  - Does not scale well
- Third-Party Block Lists
  - Huge number of Web pages
  - Proprietary databases
  - Some blockers try to conceal their rules
  - But some have been found to impose political or religious views (e.g., blocking birth-control or abortion sites)
  - Parental controls developed to allow finer control of content blocking

Enforcement

- Proxies
- Firewalls
- Parental Tools

Proxies

- Proxy intercepts request and redirects to destination if necessary
- Often used to share commonly accessed sites or pages to reduce bandwidth utilization
- Browsers can be configured to check proxy first
- Ideal locus for filtering requests

Firewalls

- More sophisticated firewalls implementing Unified Threat Management (UTM)
- Being called
  - Service gateways
  - Security gateways
- Integrate multiple security functions
  - Antivirus
  - Intrusion detection
  - Filtering Web and email content
- Effective control requires funneling traffic through device
  - Block alternatives such as WAPs

Parental Tools

- Some large ISPs provide filtering
  - AOL, MSN
- Stand-alone products for parental controls
  - Net Nanny
  - CYBERsitter
  - CyberPatrol
- Installed on individual (family, child) computers
  - Multiple security functions available

Vulnerabilities

- Spoofing
- Tunneling
- Encryption
- Anonymity
- Translation Sites
- Caching Services

Spoofing

- Modifying incoming packets' origination IP address
  - Consequence of poor security policies on routers
  - Fail to check legitimacy of inbound packets' origination addresses
- Defeat spoofing through *reverse-path filtering*
  - Configure routers to check their routing tables
  - Drop packets when source address differs from actual origin

Tunneling

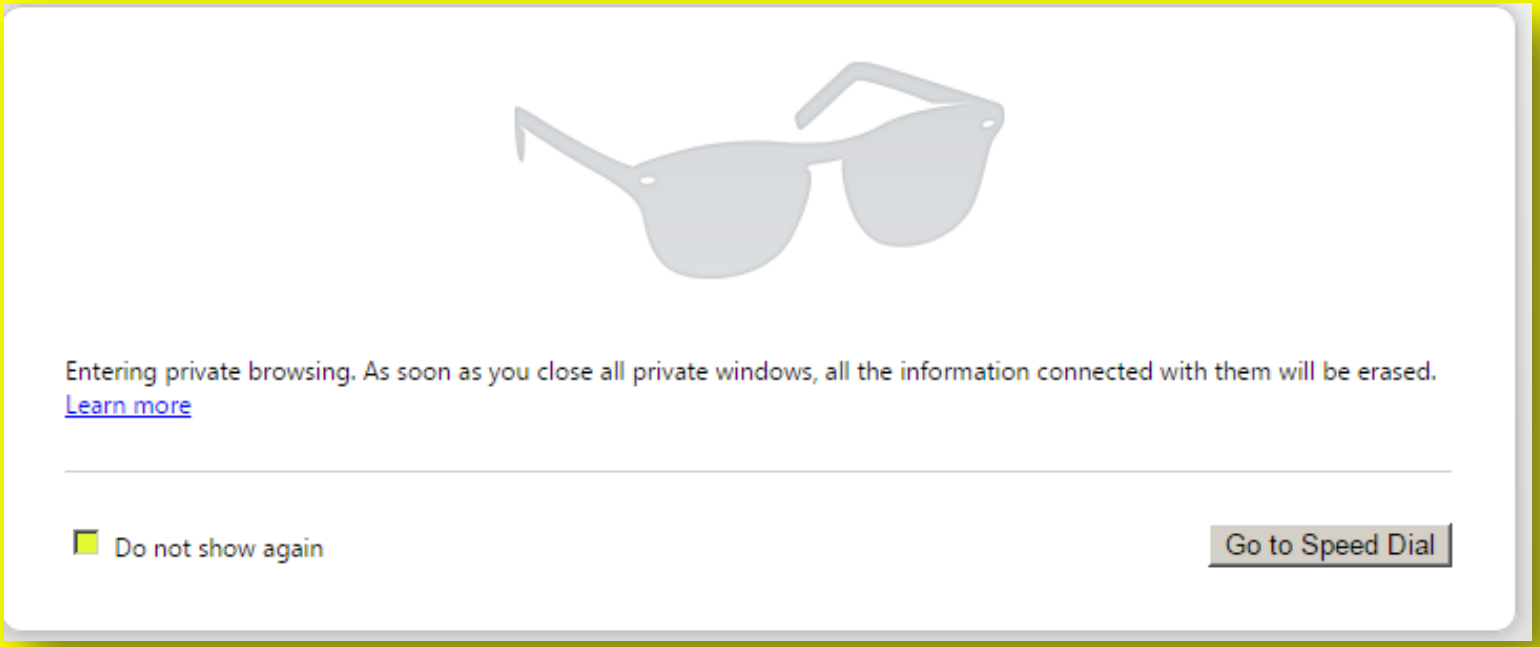
- ***Protocol tunneling = dynamic application tunneling***
  - **Applications send data on specific, commonly allowed ports**
    - ✓**E.g., encrypt data through SSH**
    - ✓**TCP port 22**
  - **Difficult to detect encrypted data from unauthorized source**
  - **Some sites allow tunneling IPv6 traffic through IPv4 connections**
- ***Application tunneling = static application tunneling***
  - **Reconfigure client application program**
  - **Redirect requests through different port**

Encryption

- Encrypted traffic from forbidden sites cannot be analyzed
  - But origination and destination IP addresses still available
  - Can therefore still block access to forbidden IP address
- Even VPN connections still show IP addresses
- Steganography (see Ch 7)
  - Far more difficult to detect
  - Store forbidden/secret content inside files
  - E.g., in images or documents

Anonymity

- Some organizations block access to forbidden sites for subsets of users
  - E.g., schools → students
- But anonymity interferes with control
  - Users may access *anonymizing services* to conceal traffic
  - E.g., The Onion Router (TOR)
  - Or *private browsing* settings on browsers



Translation Sites

- BabelFish < http://www.babelfish.com/ >
- Google Translate < https://translate.google.com/ >
- May allow escape from site blocking
  - Translation page acts as proxy
- Can evade content blocking
  - If data converted to non-English, content filters may fail
  - "Translating" English-language site from foreign language to English passes original content without alteration
  - But evades filtering

Caching Services

- Proxy servers typically cache much-used content
- So accessing data through Google may show wrong origination IP address
- Historical records (archives) on search engines also have different source addresses
- Google's "safe search" feature not a control
  - Cannot be password protected
  - Simply method for individual user to filter sexually objectionable material

The Future

- Major home-filtering products already improving
  - Net Nanny & McAfee Parental Controls
  - Allow secure limitation to safe-search options
  - Object recognition for data typical of porn sites
- Strong political / ideological conflicts
  - "Protection of the innocent" vs
  - "Rights of free speech"

Now go and study