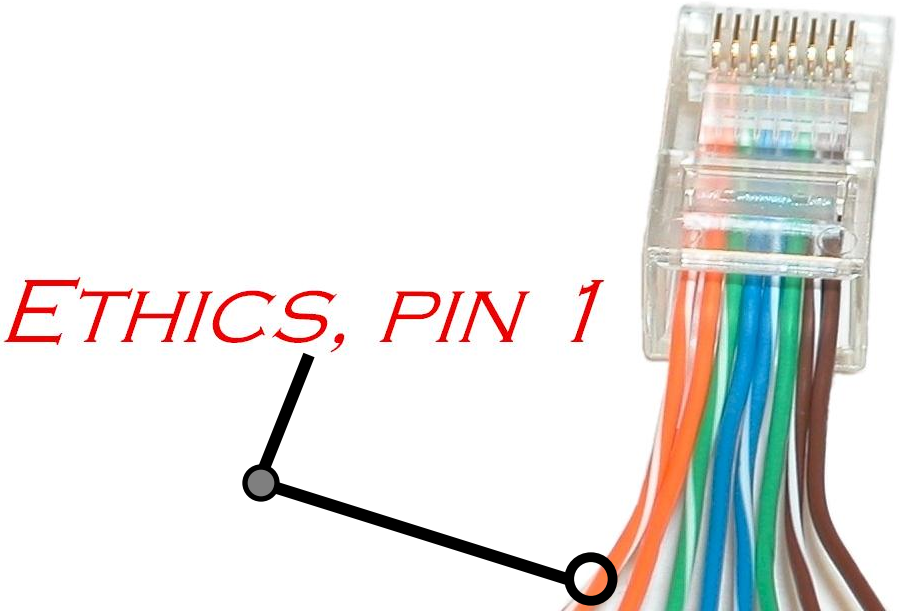
Conducting an Ethical Study of Web Traffic



John F. Duncan
L. Jean Camp

Outline

- What and Why of the Study
- Successes and Failures of Ethical Design Goals
- Surprise Data Use and IRB Implications
- Subjects and Their Determination to be Complicated
- Common Rule ANPR
- Menlo Report
- Comments, Implications for Our Work
- Assorted Calls for Involvement

Conclusions Highlights

- Researchers have little incentive to “follow the rules”
- IRBs are often unequipped to perform their duties
- Collaboration needed between ethics boards, researchers

Study Goals

- **Original Motivation:**
  - Homophily in web browsing

- **Why?**
  - Wisdom of crowds in phishing detection
  - Is this site new? To your friends?

- **Results:**
  - With 10 friends, 99% of clicks to previously visited sites
  - 95% of web sites are previously visited

Technical Implementation

- **Where:** Campus dormitory
- **Collection Point:** Dedicated FreeBSD server
- **Data Collection:**
  - Mirror packets destined for TCP port 80 with Berkeley packet filter
  - No content fetched, no HTTPS traffic observed
  - Session information removed from URL when found
  - Automated requests removed when found

Overview: Data Set

- **Timeframe:** March 5, 2008 through May 3, 2008

- **What:** HTTP GET requests to TCP port 80, URL requested*, referring URL*, identity of the user agent* (* - filtered)

- **Recorded:**
  - 408 million HTTP requests
  - 1,083 unique MAC addresses
  - 29.8 million page requests from 967 unique users
  - 630,000 distinct Web servers
  - 110,000 distinct servers provided referrers

Study Details: After

- **This Paper:**
  - Release instruments & note difficulties
  - Aid other researchers & improve dialog
  - Discuss human-subjects protocols & study design

Meiss et al, HyperText 2009

- **Additional Research Use:**
  - There is no typical http session
  - The concept of session can mislead researchers
  - Pure networking
    - No approval or review process

- **Governing Frameworks Matter!**
  - Administrative bodies must be involved

IRB Interactions

- **Original Language:**
  - "Non-public browser requests (sites that use encryption such as your email, financial institutions, academic records, and Oncourse) will not be viewed or recorded in this experiment."

IRB Interactions

- **Final Language:**
  - "Any web traffic that is encrypted – i.e., traffic that is sent via https:// or through a VPN connection – cannot be captured or recorded as part of this study. So if you decide you don't want any of your web browsing information to be recorded as part of this study, you can encrypt your web traffic in one of two ways:
    - Ensure that the website you are using has a URL starting with https:// instead of http://
    - Use IU's Virtual Private Network (VPN) service: [link removed]

Successes With Ethical Design

- **Data Minimization:**
  - User data was de-identified as it was captured
  - Session data removed from URLs when identified
  - All URL data was de-identified before data was processed
- **Discussion With Stakeholders**
- **Opt-out Increased Security**

Ethical Design Failures

- **Opt-out vs. Opt-in**
- **Gauging Student Understanding**
- **Student Government vs. IRB / HSC**
- **Additional Use of Data**

Tension Between Security, Privacy and Trust

- **Poorly-written Anonymization Service:**
  - Attempting to obscure traffic to an adult chat site
  - Spoofing requests from uninvolved clients
  - Name of site communicated sexual orientation

- **Our Options:**
  1. Adjust collection to identify machine, then individual, notify individual & direct them to better resources
  2. Do nothing

Our Decision: Do Nothing

- **Why?**
  - Individual identification too problematic
  - Social loss of privacy vs. technical increase in privacy
  - Identification could cause harm!

Other Situations

- **Criminal Activity?**
  - Child endangerment and abuse must be reported
  - No other case merits this abrogation of study parameters

AOL Search Data, User 17556639

17556639 how to kill your wife
17556639 how to kill your wife
17556639 wife killer
17556639 how to kill a wife
17556639 poop
17556639 dead people
17556639 pictures of dead people
17556639 killed people
17556639 dead pictures
17556639 dead pictures
17556639 dead pictures
17556639 murder photo
17556639 steak and cheese
17556639 photo of death
17556639 photo of death
17556639 death
17556639 dead people photos
17556639 photo of dead people
17556639 www.murderdpeople.com
17556639 decapatated photos
17556639 decapatated photos
17556639 car crashes3
17556639 car crashes3
17556639 car crash photo

Human Subjects Motivation and Proposals

- **HSC Material Motivation:**
  - Face-to-face interviews
  - Biomedical research
  - Sociological experiments
  - Etc.

- **Study Benefit to Society:**
  - Privacy systems design
  - Exploration of anonymization techniques
  - Improving standards for network research

Human Subjects Proposals

- **Belmont / Common Rule Update:**
  - Advance Notice of Proposed Rule-Making
  - July 26 - September 26, 2011
  - All research, focus on medical
- **Menlo Report:**
  - Notice and Request for Comments
  - September 15 - December 28, 2011
  - Networking research specifically
- **Various Comments on Menlo & Belmont**

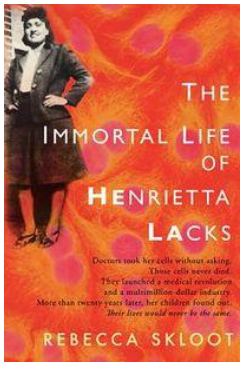
Belmont Report (1979)

- Driven by disclosure of Tuskegee Experiment in 1972
- A balance of societal risk and individual benefit

- **Respect for Persons**
- **Beneficence**
- **Justice**

Common Rule/Belmont ANPR (2011)

- **Bioinformatics**
  - New categories of risk
- **Control over Inherently Identifiable Materials**
  - DNA
  - Medical records
  - Tissues
  - Patented genes and cures from patients' illness
- **A Balance of Societal Risk and Individual Benefit**
  - Driven in part by *The Immortal Life of Henrietta Lacks*



Harvard Comments on Common Rule ANPR

- **Who:**
  - 10 colleagues
- **HIPAA Standards:**
  - Not adequate to protect privacy OR allow research
- **Proposal:**
  - Consistent *set of standards*, not consistent set of reviewers
  - Create a Safe Harbor for research that meets specific criteria

Harvard Comment Implications

- **For Us:**
  - Experimental information sharing
  - Additional use acceptable

- **Proposed "Safe Harbor" List :**
  - Research designed to limit risk / maximize benefit
  - Remove review of low-risk research
  - Function of: class of data; source of data; method of sharing data; informed consent mechanisms; and class of subjects

Professional Society Comments on Common Rule ANPR

- **Represent 310,000 Professionals**
  - Human Subjects and computing research
  - Data security for Human Subjects research information

Recommendations for Common Rule

- Regulations might **unintentionally** restrict research
- Uniform application of data security rules
- A means for updating regulations

Professional Society Comments on Common Rule ANPR

- "*Question 7:* **What research activities, if any, should be added to the published list of activities that can be used in a study that qualifies for expedited review?**

- Networking research that uses only the information **normally and openly transmitted** over the network(s) should be subject to expedited review. Depending on the particulars of the project, such research could fall into the Excused category."

Professional Society Comment Implications

- **For Us:**
  - Data may meet proposed Safe Harbor
  - Medical standards review not appropriate

- **Consequences:**
  - Much networking research may be prohibited
  - Additional use would require notification
  - Human subjects training for network researchers

Menlo Report (2011)

- **Critical Challenges:**
  - Identification of stakeholder interests
  - Appropriate communication with stakeholder

- **Menlo Report:**
  - Consult with HSC / IRB
  - Identifiable stakeholders
  - Driven in part by Common Rule update
  - Problems remain with anonymity research
  - Adds to Belmont

Implications of Menlo to Our Research

- **Our Experiment Used IRB:**
  - Very professional, ready to learn
  - Wanted to support research
  - Identifiable stakeholders
  - Method for communication was documented
- **Additional Use:**
  - May have been Excused or Exempt
  - Uniform process
- **Problems Remain!**

USACM Comments on Menlo

- **First:**
  - Collect and analyze data on current practices
  - Evaluate the advantages and disadvantages of research ethics board models
- **Proposed:** IRBs OR national / regional bodies
- **Security Research:**
  - Consider related work and guidance world-wide
  - Bodies must include specialists in research ethics
  - Offer to help

Camp's Comments on Menlo

- **IRBs:**
  - Lacking technical expertise, under-resourced
  - Structural incentives to deny research
    - low-risk vs. high-risk
  - Ensuring compliance beyond many IRBs
  - Risks are an incentive to avoid the IRB process

Camp's Comments on Menlo

- **Alternative:**
  - Draw from the IEEE and ACM!
    - National review body of Ethicists and Technologists
  - Consistent review & uniform standards
  - Enable self-regulation via structural changes
  - Societies in IEEE, ACM, for USA

- **Data Re-use**
  - Inherent access to data
  - Publisher bears responsibility as well!

Implications of Camp's Comments on Menlo

- Additional use would have been likely excused but within a structure
- Strengths and weaknesses of IRB shown by this work
- Ethics boards require technological expertise, more predictability in providing research guidelines
- Technical bodies require ethical and philosophical expertise

Isolated Efforts Inadequate

- Many labor hours expended
- Results only now disseminated
- Were these efforts duplicated?
- From whom should we have learned?

Conclusions - Researchers Must Take The Initiative!

- **Ethical Systems in Network Research:**
  - Reduce data footprint
  - Address stakeholder concerns, even if by proxy
- **Challenges:**
  - Research boards and researchers need a common language
  - Identifying subjects in networking research
  - Subjects may be technological naïve or malicious
  - Participants must understand experiments to provide meaningful informed consent
- **Menlo and Common Rule ANPR**

Meta Conclusion

- **GET INVOLVED:**
  - No IEEE- USA or SIGCOM Comments on Menlo because of a lack of volunteers
    - Frustrating but EQUALLY IMPORTANT
    - No big-data or large-scale networking researchers on either committee

Now Is The Time –
An Inflection Point For Our Communities

Questions / Discussion