CSE543- Computer Security

Module: Privacy

Prof. Syed Rafiul Hussain
Department of Computer Science and Engineering
The Pennsylvania State University

Data Privacy

- From Slashdot (11/24/2013)
  - An anonymous reader writes "*The NSA snoops traffic and has backdoors in encryption algorithms. Law enforcement agencies are operating surveillance drones domestically (not to mention traffic cameras and satellites). Commercial entities like Google, Facebook and Amazon have vast data on your internet behavior. The average Joe has sophisticated video-shooting and sharing technology in his pocket, meaning your image can be spread anywhere anytime. Your private health, financial, etc. data is protected by under-funded IT organizations which are not under your control. Is privacy even a valid consideration anymore, or is it simply obsolete? If you think you can maintain your privacy, how do you go about it?*"

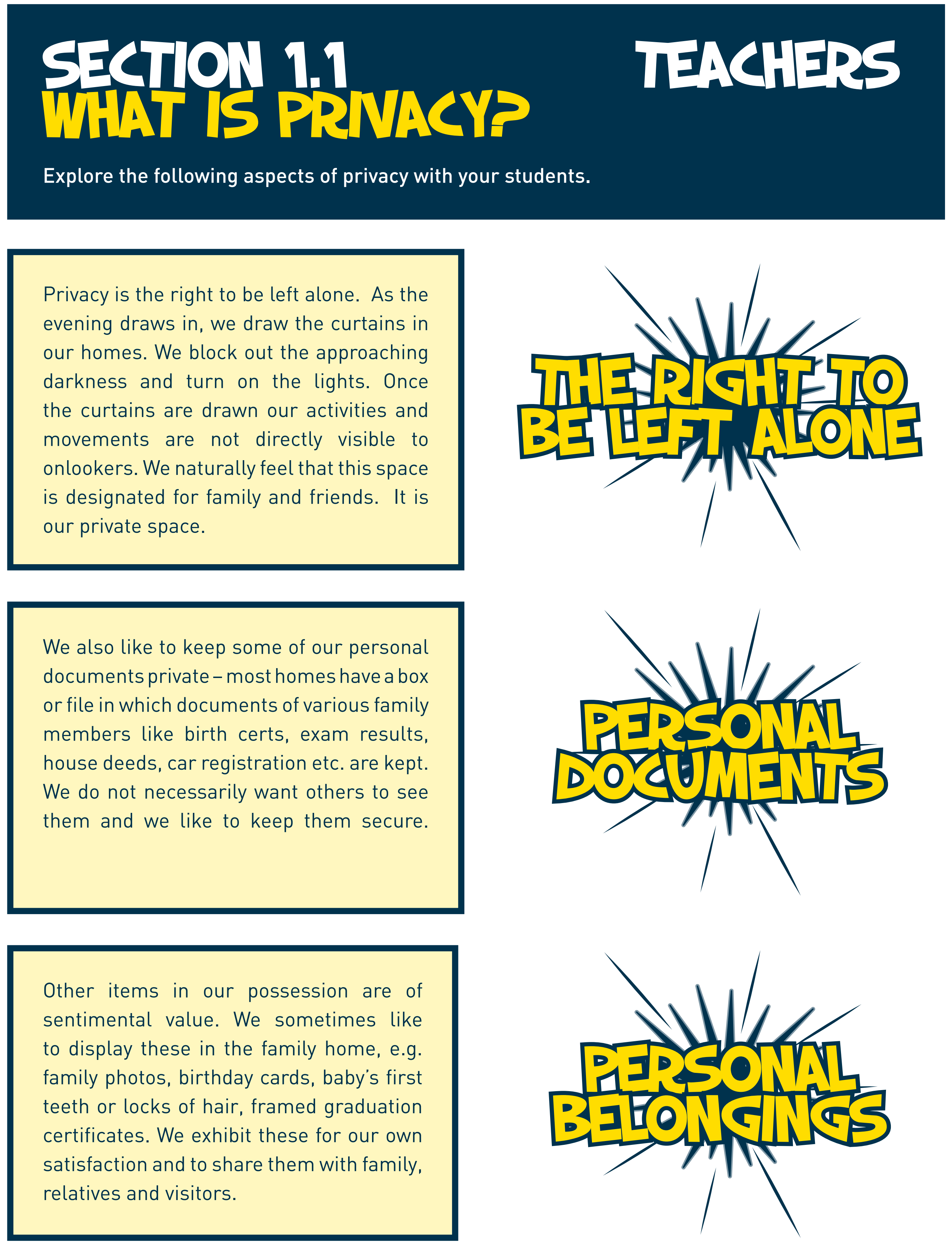
What is Privacy?

- What is a reasonable expectation of privacy today?
- How do you maintain your privacy to this level?

What is Data Privacy?

- Australia (Info & Privacy Commission)
- from the right to be left alone to the right to have some control over how your **personal** or **health** information is properly collected, stored, used or released
  - **information privacy** – the way in which government agencies or organizations handle personal information such as age, address, physical or mental health records
  - **freedom from excessive surveillance** – the right to go about our daily lives without being surveilled or have all our actions caught on camera.

Ireland (Data Protection Commissioner)



SECTION 1.1 TEACHERS

WHAT IS PRIVACY?

Explore the following aspects of privacy with your students.

Privacy is the right to be left alone. As the evening draws in, we draw the curtains in our homes. We block out the approaching darkness and turn on the lights. Once the curtains are drawn our activities and movements are not directly visible to onlookers. We naturally feel that this space is designated for family and friends. It is our private space.

THE RIGHT TO BE LEFT ALONE

We also like to keep some of our personal documents private – most homes have a box or file in which documents of various family members like birth certs, exam results, house deeds, car registration etc. are kept. We do not necessarily want others to see them and we like to keep them secure.

PERSONAL DOCUMENTS

Other items in our possession are of sentimental value. We sometimes like to display these in the family home, e.g. family photos, birthday cards, baby's first teeth or locks of hair, framed graduation certificates. We exhibit these for our own satisfaction and to share them with family, relatives and visitors.

PERSONAL BELONGINGS

Privacy is the right to be left alone and to live one's life with the minimum of interference.

Privacy “Statements”

Australia

http://www.ipc.nsw.gov.au/privacy/privacy_forgovernment/govt_privacy/privacy_faqprivacy.html

The *Privacy Act 1988* (Privacy Act) regulates how personal information is handled. The Privacy Act defines personal information as:

…information or an opinion (including information or an opinion forming part of a database), whether true or not, and whether recorded in a material form or not, about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion.

Personal information includes information such as:

your name or address
bank account details and credit card information
photos
information about your opinions and what you like.

EU - Data Protection Directive

http://epic.org/privacy/intl/eu_data_protection_directive.html

The EU Commission's strategy sets out proposals on how to modernize the EU framework for data protection rules through a series of the following key goals:

- **Strengthening the Rights of Individuals** so that the collection and use of personal data is limited to the minimum necessary. Individuals should also be clearly informed in a transparent way on how, why, by whom, and for how long their data is collected and used. People should be able to give their informed consent to the processing of their personal data, for example when surfing online, and should have the "right to be forgotten" when their data is no longer needed or they want their data to be deleted.
- **Enhancing the Free Flow of Information in the Single Market Dimension** by reducing the administrative burden on companies and ensuring a true level-playing field. Current differences in implementing EU data protection rules and a lack of clarity about which country's rules apply harm the free flow of personal data within the EU and raise costs.
- **…**
- **More Effective Enforcement of Privacy Rules** by strengthening and further harmonizing the role and powers of Data Protection Authorities. Improved cooperation and coordination is also strongly needed to ensure a more consistent application of data protection rules across the Single Market.

What is Privacy?

- US
  - This broad concept of privacy has been given a more precise definition in the law. Since the Warren-Brandeis article, according to William Prosser, American common law has recognized four types of actions for which one can be sued in civil court for invasion of privacy.
  - They are, to quote Prosser:
    - Intrusion upon the plaintiff's seclusion or solitude, or into his private affairs.
    - Public disclosure of embarrassing private facts about the plaintiff.
    - Publicity which places the plaintiff in a false light in the public eye.
    - Appropriation, for the defendant's advantage, of the plaintiff's name or likeness.
  - HIPAA (Health Insurance Portability and Accountability Act of 1996)
    - The HIPAA Privacy Rule establishes national standards to protect individuals' medical records and other personal health information and applies to health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically. The Rule requires appropriate safeguards to protect the privacy of personal health information, and sets limits and conditions on the uses and disclosures that may be made of such information without patient authorization. The Rule also gives patients rights over their health information, including rights to examine and obtain a copy of their health records, and to request corrections.

Protection Privacy

- How do you protect your privacy in practice?
- Slashdot responses (11/24/2013)
  - not respond truthfully (may not be practical or be checked)
  - change your browser (be careful about compatibility)
    - use multiple browser profiles or control use of cookies
  - encryption (beware of traffic analysis)
  - don't use social networks
  - assume that you are not interesting (is your head in sand?)
  - give up (assume all electronic communication is public)
- Others?

Can We Do Something?

- Suppose a research agency wants to evaluate medical data
  - Can we give them medical data that cannot be tracked to a specific identity?
- Suppose medical records have fields
  - Name
  - Address
  - Visit Date
  - Doctor
  - Diagnosis
  - …

- Can we just remove identifying information (name, address)…?

Inference Attack

- An **Inference Attack** uses data analysis in order to illegitimately gain knowledge about a subject or database. A subject's sensitive information can be considered as leaked if an adversary can infer its real value with a high confidence.
  - Assume that the adversary can choose the query
  - Could query by doctor and date
  - Could cross-reference with external knowledge about doctor or date or condition or ...
  - To find a particular subject's sensitive information with high confidence
- How do we know whether removing some identifying information from records (anonymization of data) will prevent inference attacks?

Netflix De-Anonymization

- Narayanan and Shmatikov de-anonymization technique
  - Adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset
- Overview
  - Model: Database N records of M attributes (NxM)
  - Adversary Goal: de-anonymize an anonymous record r from the public database
  - Compute score for each record r from *auxiliary info*
  - Claim: For *sparse* datasets, like Netflix reviews, much less auxiliary info is necessary to distinguish records

Netflix De-Anonymization

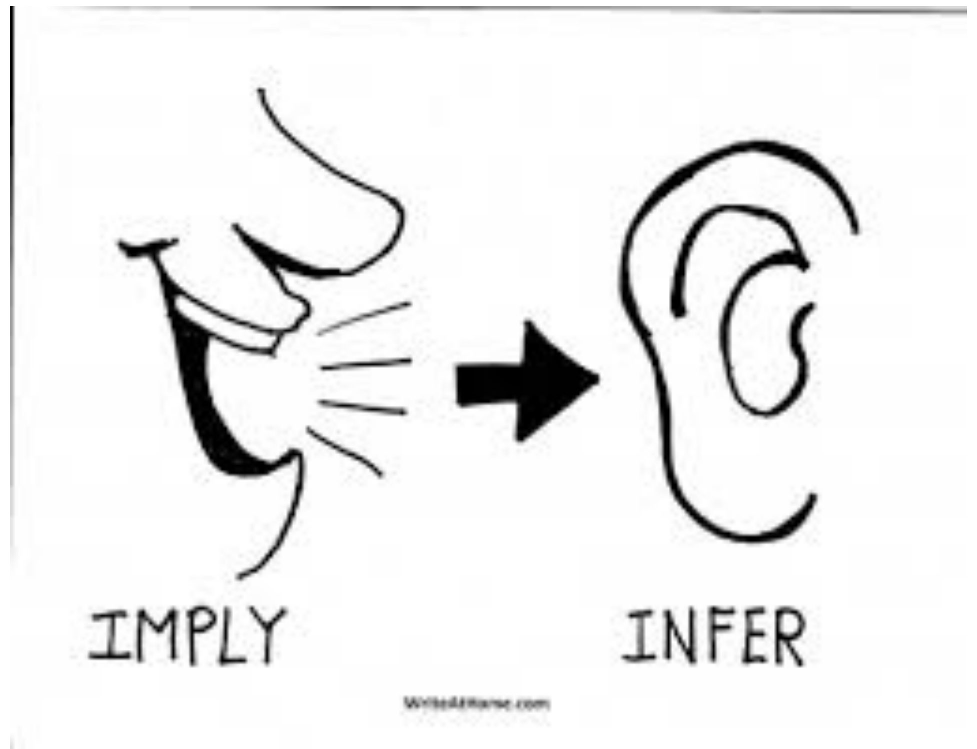
- Applied to Netflix Prize dataset
  - Anonymized dataset of 500,000 Netflix subscribers
  - Finding: simply removing identifying information is insufficient for anonymity
- How much does an adversary need to know about a Netflix subscriber to identify if her record is in the DB?
  - Auxilary info: Individual ratings of a movie and the dates of ratings
  - Result: If adversary knows 8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error, 99% of records be uniquely identified

Netflix De-Anonymization

- Approach
  - Auxiliary info: IMDb reviews - other movie reviews
    - Obtained Netflix info for some acquaintances - very few records were perturbed in Netflix dataset
  - Given this info, compute *similarity* between non-anonymous records and those in data set - for two attributes: *rating* and *date*
  - Find *best match* - and test if much better than next match (e.g., compare difference to standard deviation)
  - Bias toward more *unusual attribute* values

Preventing Inference

- Is there a method that prevents detection of identifying information in records in databases?
  - While still returning accurate answers to queries?
- Maximizing the accuracy of query results while minimizing the chances of identifying records



Differential Privacy

- Consider a trusted party that holds a dataset of sensitive information (e.g. medical records, voter registration information, email usage) with the goal of providing global, statistical information about the data publicly available, while preserving the privacy of the users whose information the data set contains.
- "Epsilon" is a "privacy-budget" used to define "Differential Privacy"
  - A randomized algorithm A (for providing global, statistical info) is epsilon-differentially private if for all data sets D1 and D2 that differ in only a single element (data about one person):
  - Probability that output of A for D1 (with person's data) contains user data is no greater than $e^{epsilon}$ * *probability* of any output of A for D2
  - When epsilon is small, then probabilities would be very close
- That is, algorithm A should behave essentially the same on the two data sets

Differential Privacy Systems

- What does it mean in practice? Privacy is composable
  - Database and Algorithm A
  - Adversary requests queries on a database using A
    - Untrusted queries
  - Data owner can specify a "privacy budget" regarding an individual
  - The system computes a "privacy cost" for each query
  - Only allows the query if the cost does not exceed the budget
- Example systems: PINQ and Airavat
  - Fuzz: restrict budget for covert information as well

Cell Phones

- A target of data collection are cell phones
  - Have them with you all the time
  - Track useful information (GPS)
  - Download nearly arbitrary code to phones
- Is your cellular information private?
  - Short answer: no
  - Long answer: different parties have (or want) access to your data for different purposes
- Who should be allowed to access cellular info? Providers? Law enforcement? App developers?

TaintDroid

- Runtime taint tracking in Android
  - Identify security-critical data (manual)
  - Track its propagation throughout program at runtime
    - Each instruction's impact on tainting must be defined
    - Keep metadata about memory locations regarding taint
  - See if tainted data is output by the program

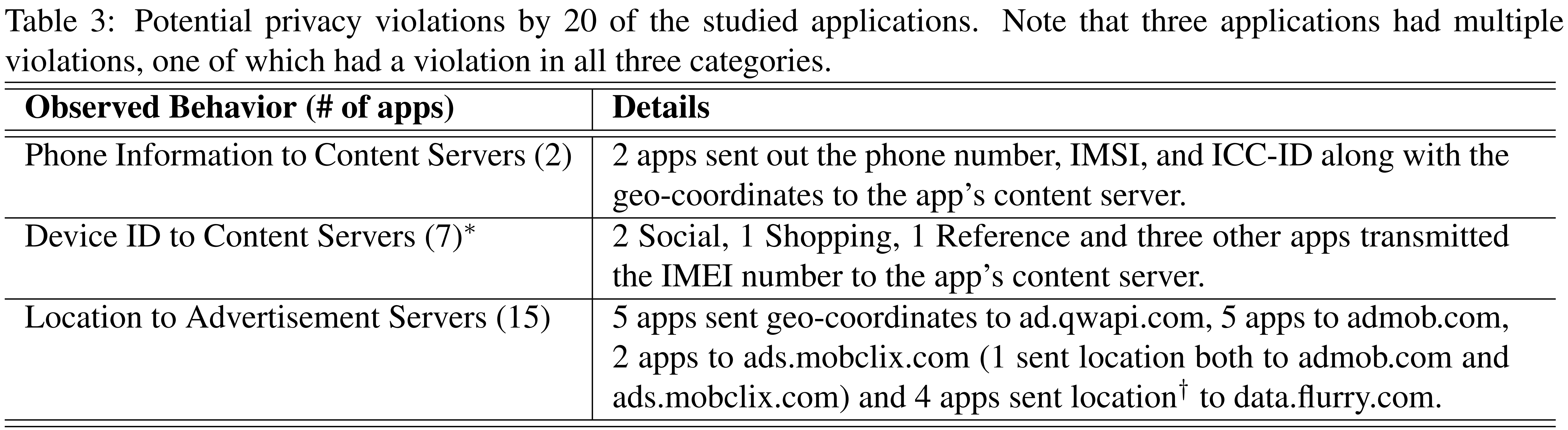
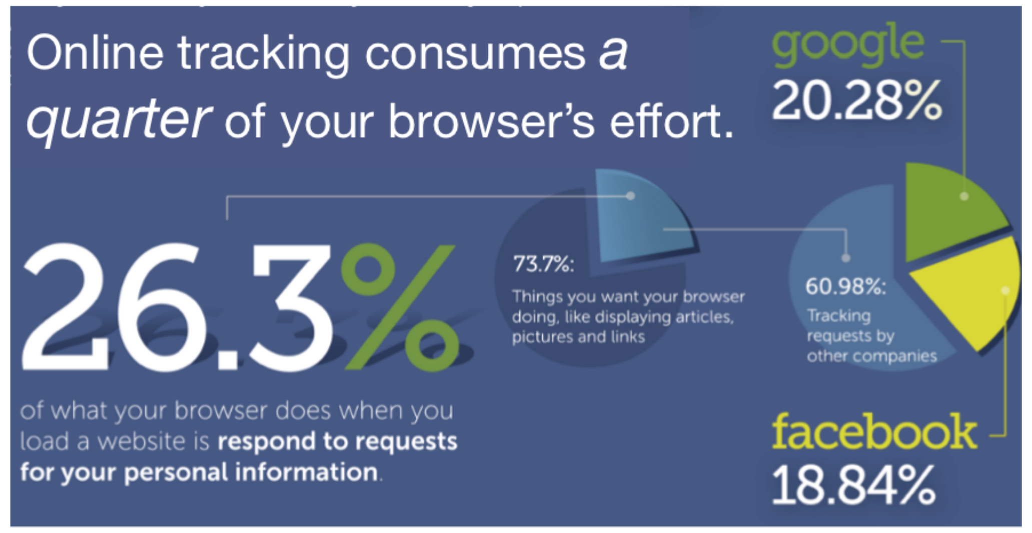


Table 3: Potential privacy violations by 20 of the studied applications. Note that three applications had multiple violations, one of which had a violation in all three categories.

Observed Behavior (# of apps)	**Details**
Phone Information to Content Servers (2)	2 apps sent out the phone number, IMSI, and ICC-ID along with the geo-coordinates to the app's content server.
Device ID to Content Servers (7)*	2 Social, 1 Shopping, 1 Reference and three other apps transmitted the IMEI number to the app's content server.
Location to Advertisement Servers (15)	5 apps sent geo-coordinates to ad.qwapi.com, 5 apps to admob.com, 2 apps to ads.mobclix.com (1 sent location both to admob.com and ads.mobclix.com) and 4 apps sent location† to data.flurry.com.

Web Privacy

- Have you ever …
  - Searched for a product on some website
  - …Advertisement for the same product shows up on another website?
  - Reason: Tracking! Profile users for targeted advertisement
- Study by WSJ found
  - 75% of top 1000 sites feature social networking plugins
    - Match users' identities with their browsing activities
- abine and UC Berkeley found
  - Online tracking is 25% of browser traffic
    - 20.28% Google analytics
    - 18.84% Facebook

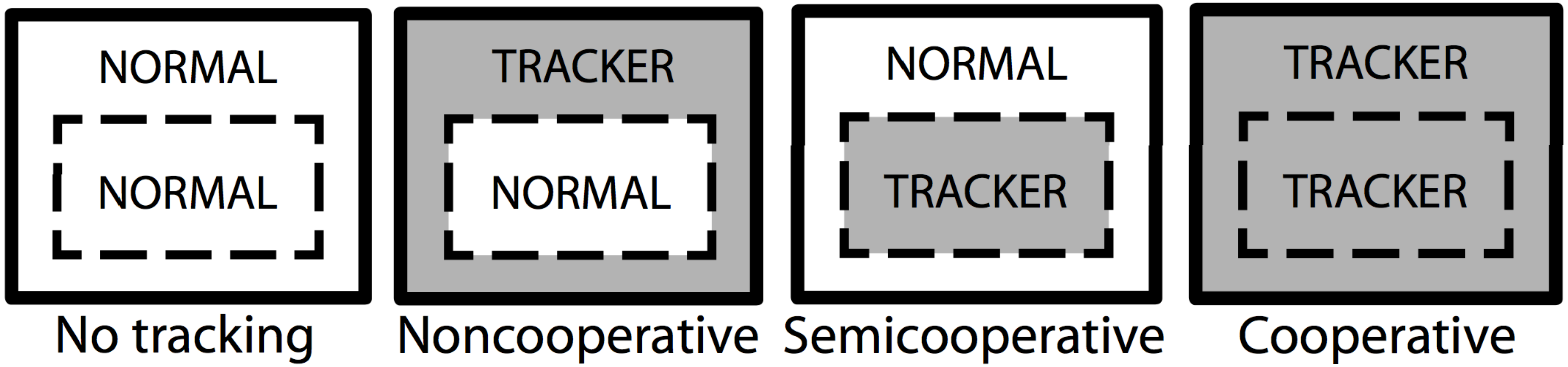


http://www.abine.com/

Web Privacy

- Tracking is done when one site embeds content in another

Protecting Browser State from Web Privacy Attacks : Jackson et al.



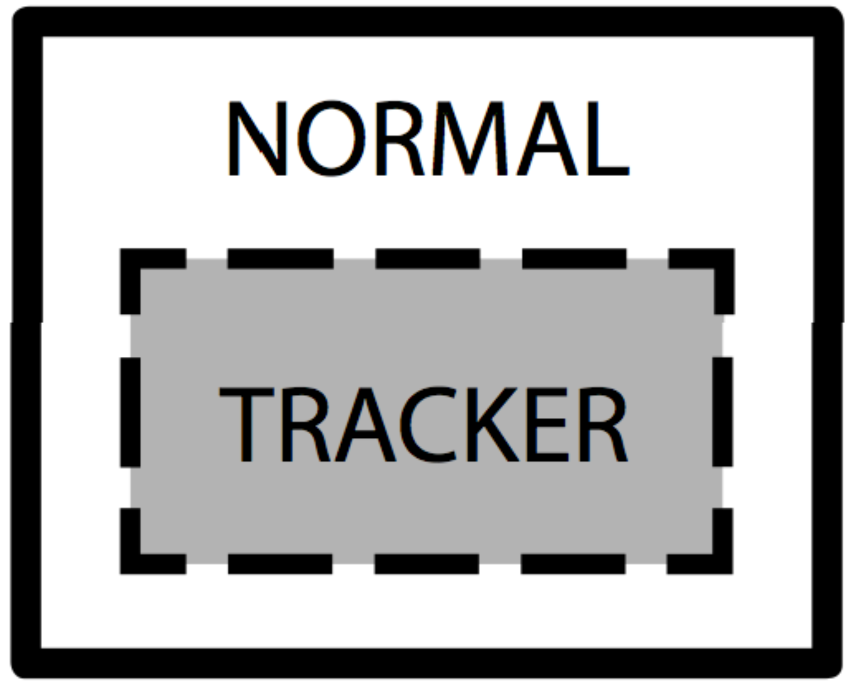
- “Tracker
  - Social networking sites
  - Analytics
  - Advertisement agencies
  - …

Web Privacy

- Objective of tracking code is to maintain state of users across multiple sites
  - Build profile of sites visited
- Semi-cooperative tracking done by
  - Javascript
    - e.g., Cached redirect URLs
  - Web bugs
    - 1x1 images
    - Ever wondered why email clients have "Display images"?
  - IFrames
  - Cookies
    - Traditional, flash, HTML5 LocalStorage, ...
    - Defense: Disable third-party cookies

Third-Party Cookies

- A third-party cookie is a cookie from a website different from the website being viewed
- Browsers can block third-party cookies
  - Different browsers have different variations
    - Some have different origin for (hosted, embedded)
    - Some completely block
- Limitation
  - Other ways exist to store state
    - HTML5 LocalStorage
    - Redirect caching
    - ETags



Semicooperative

Web Privacy

- What should the web privacy policy be?
  - What is a reasonable default?
    - *Tracking* or *no tracking*
  - What choices should users be able to make?
    - Control *collection* and/or *use* of data
  - Who should develop/manage such policy enforcement?
    - Third-parties trusted to administer policies - like OS distributors for MAC policies on hosts
- Multiple perspectives on privacy
  - Fundamental human right
  - Maximize welfare (of whom?)

Web Privacy Technologies

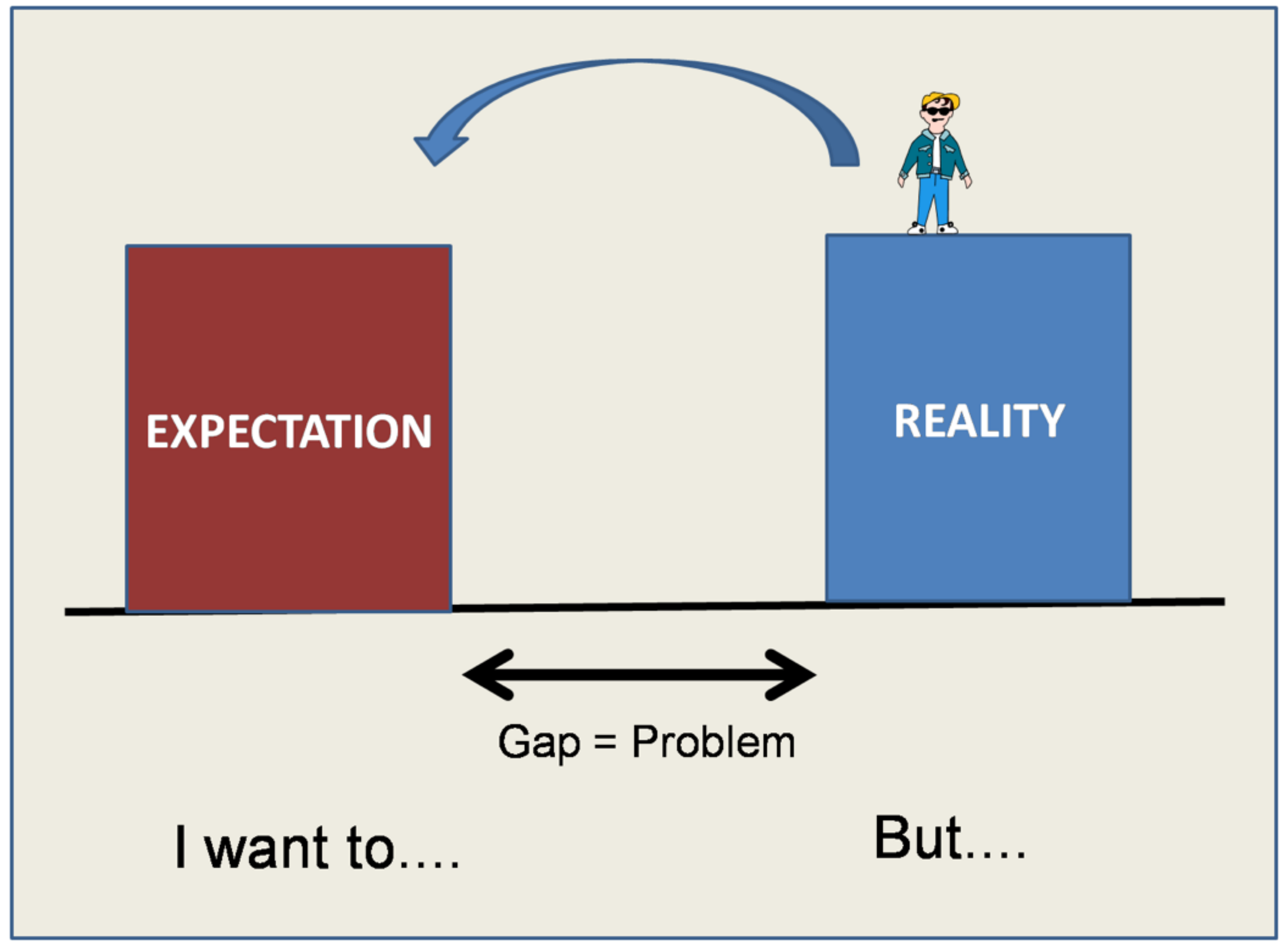
- What technologies are available to users to protect their privacy?
- (1) Opt-out Cookies
  - Tells the website not to install third-party advertiser or other cookies on your browser
  - Problems: Install such cookies manually and may get removed
- (2) Blocking Third-Party web content
  - A "block list" implemented by some browser extension
  - Problems: Variable quality, block content and tracking
- (3) Do Not Track
  - HTTP header, *DNT*, that signals a user's preference
  - Problems: Websites may not honor it

Communicating Anonymously

- What if you want to access a website anonymously?
    - Avoid government or adversarial tracking
- Is this possible on the Internet?
    - Traffic analysis: the process of intercepting and examining messages in order to **deduce information from patterns** - even encrypted communications
    - Someone has access to one or more Internet routers, they can intercept messages and determine information, such as the source and destination

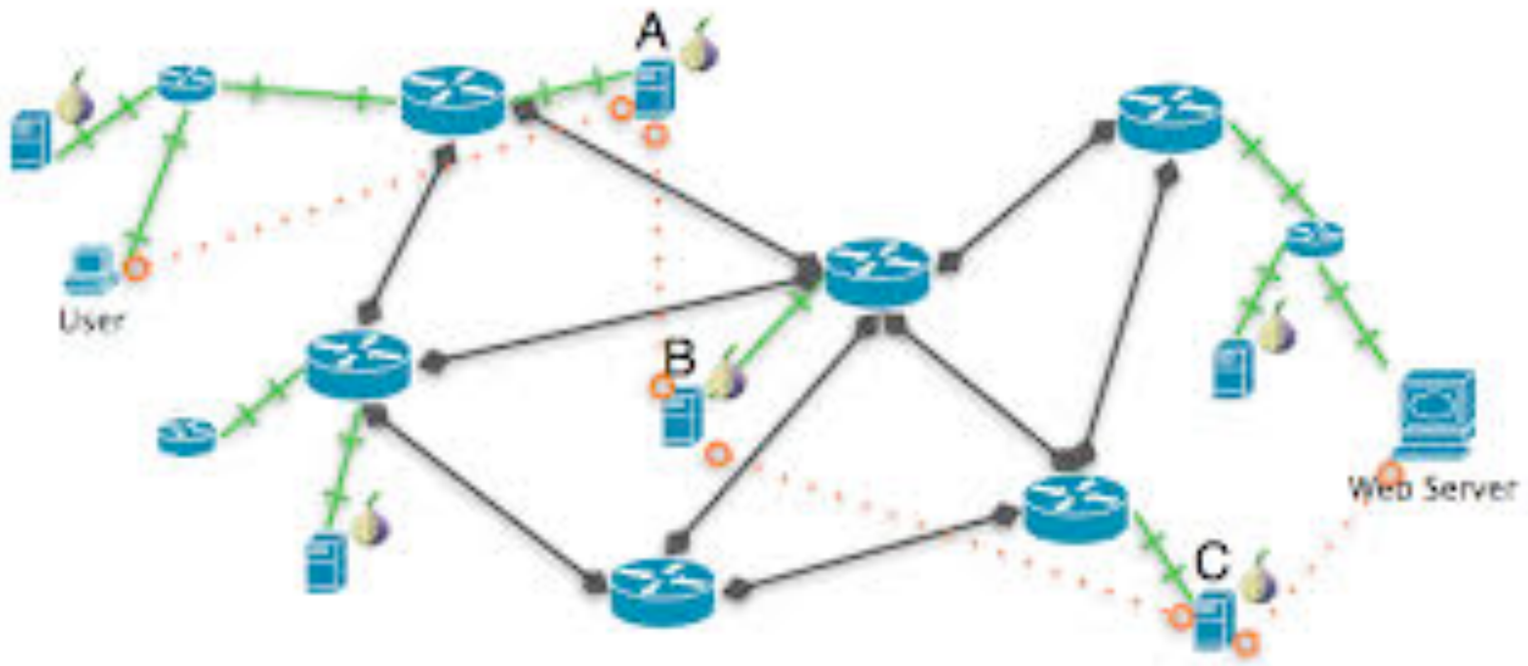
Reasonable Expectation

- Your communication traffic is public
- Traffic analysis is practical
- Some parties may want to block communications with some websites
- So what can you do?



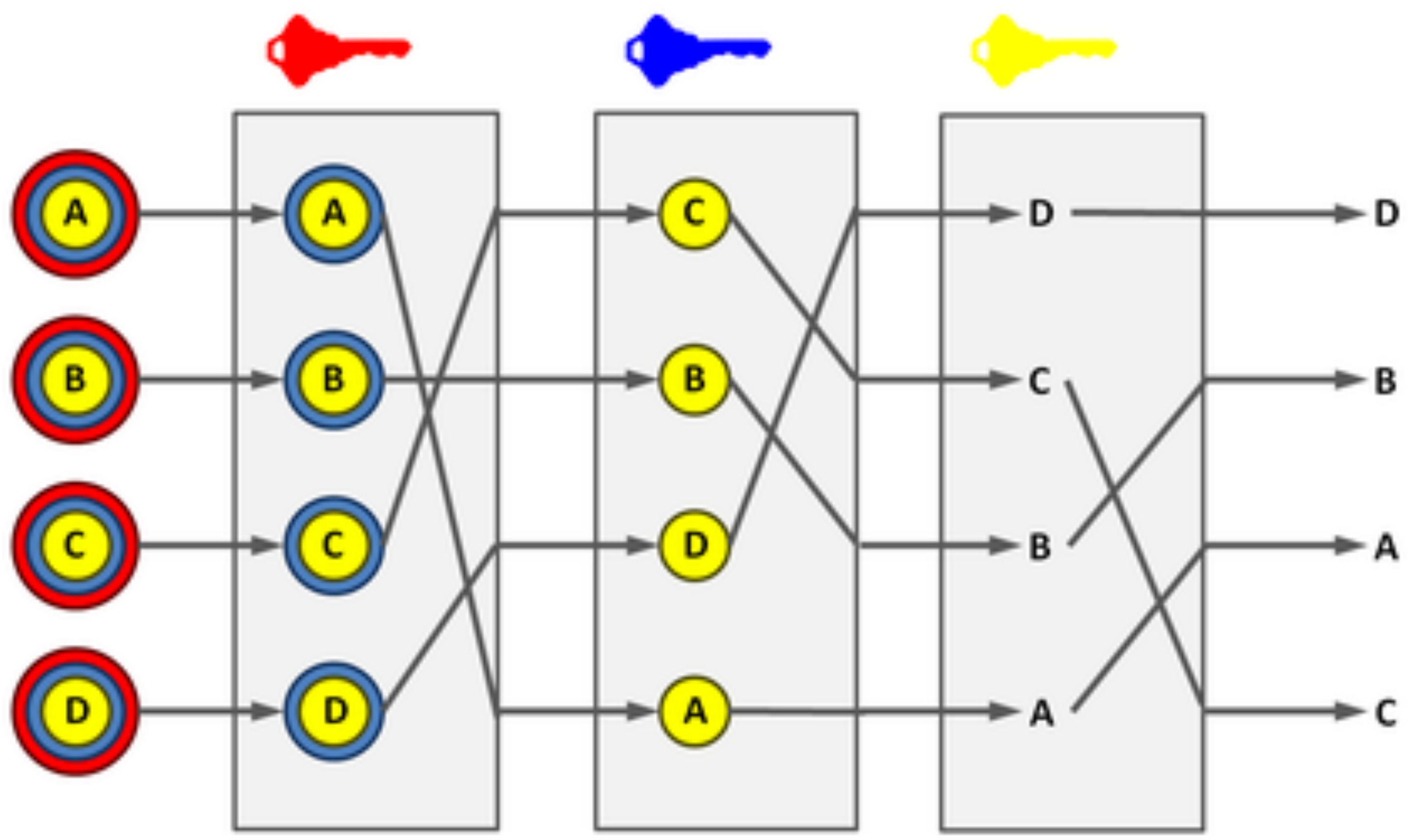
Anonymous Routing

- Prevent adversary in the network from deducing the source and destination of communications
- Goals
  - Complicate traffic analysis
  - Separate identification from routing
    - Anonymous connections: hop-to-hop
  - Support many applications



Onion Routing

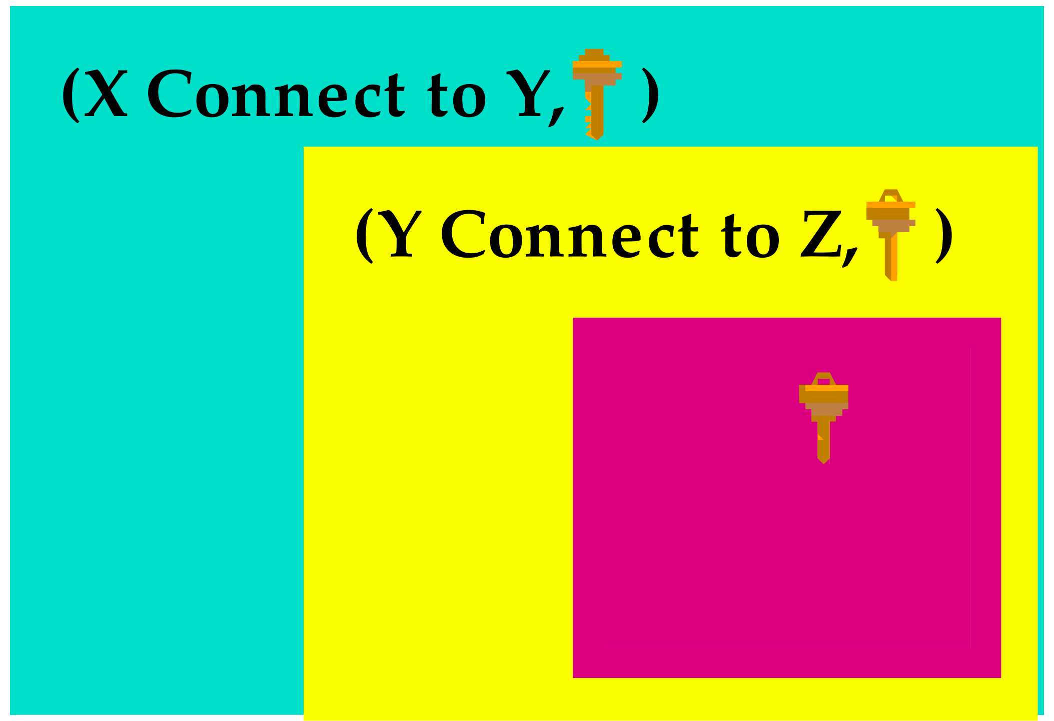
- A combination of techniques to encapsulate communications to make traffic analysis more difficult
  - Mixes: intermediaries that may pad, reorder, delay communications to complicate traffic analysis
  - Onion Routers: Communication infrastructure that act as mixes
  - Connections: Point-to-point between pairs of onion routers
  - Communications: changed on each link
- Idea: create end-to-end connections through a sequence of onion routers that change communications on each hop
  - Key to changing data - the "onion"



Simple decryption mix net. Messages are encrypted under a sequence of public keys. Each mix node removes a layer of encryption using its own private key. The node shuffles the message order, and transmits the result to the next node.

Onion

- Initiator's proxy (W) chooses an anonymous connection
  - W-X-Y-Z, then destination
- Public key crypto is used to limit each onion router to only "peel" the layer intended for it
  - How would W create a public key message that only X could read?
  - How would W create messages for Y and Z inside the message for X?
- For efficiency, only encrypt a header using public key
  - Rest via symmetric key crypto



Onion

- Onion Routing Process

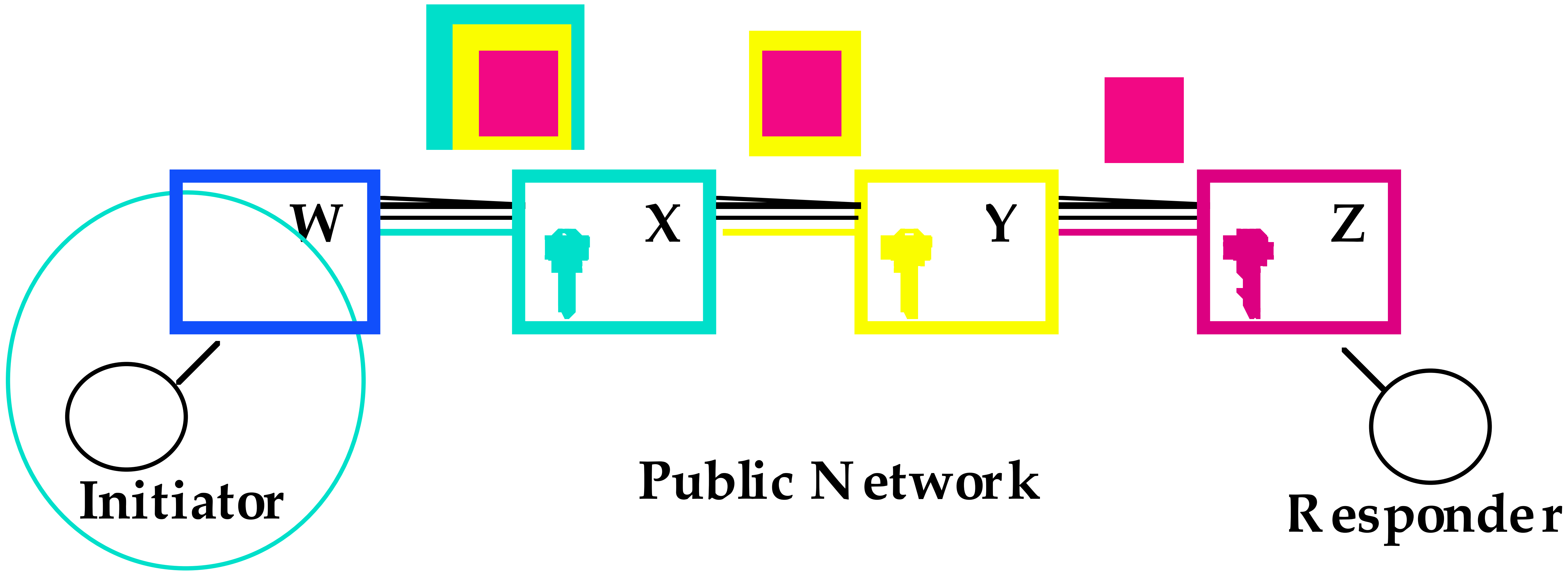


Figure 5: Use of an Onion

Limitations of Onion Routing

- Performance-Anonymity Trade-off
  - How many onion routers are necessary?
- Traffic analysis is still possible
  - Does not completely eliminate analysis
- Web traffic may be distinct
  - May be difficult to hide
- Onion routers may be compromised
  - Broken if initiator's proxy is compromised
- Denial of service is possible

Tor - The Onion Router

- Second-generation Onion Router
- Significant improvements
  - Perfect forward secrecy: Instead of using public keys that could eventually be compromised, use per-hop keys that are deleted when no longer in use
  - Performance improvements: Shared TCP streams, congestion control
  - Integrity checking: None before, end-to-end now
- Subsequent improvements include
  - Guard nodes
  - Improved path selection algorithms
- Used by Edward Snowden to send information about PRISM to the Guardian and Washington Post

Guard Nodes

- Prevent de-anonymization by traffic analysis
- From Tor documentation
  - if an attacker controls or monitors the first hop and last hop of a circuit, then the attacker can de-anonymize the user by correlating timing and volume information.
- Approach
  - Tor clients pick a few Tor nodes as its "guards", and uses one of them as the first hop for all circuits (as long as those nodes remain operational).
- If the guard nodes chosen by a user are not attacker-controlled all their future circuits will be safe

Using Tor

- Tor Browser
  - Configured to browse using Tor network
- But that alone is not enough - need to change your habits
  - Don't torrent over Tor - sends your IP address
  - Don't enable or install browser plugins - reveal your IP address
  - Use HTTPS versions of websites - Tor only encrypts in the Tor network
  - Don't open documents downloaded through Tor while online - they might contain internet resources (pdf and doc)
  - Use a bridge - to hide that you are using Tor - get friends to also

Take Away

- Maintaining private use of digital services is difficult
  - Ease of broad access to data is often a goal
  - Difficult to know what privacy means to users and privacy can be broken using external data
- Databases
  - Queries of private databases may reveal secrets
  - Even "anonymized" release of data may insufficiently protect anonymity (Netflix)
- Communication privacy
  - Prevent traffic analysis during secure communication
  - Onion routing - available in Tor