How to address this issue?

- Enforce privacy-preserving data mining
  - E.g., government can use it to alleviate people’s worry in excessive data collection
- Naïve approach: removing identifying information from data
  - E.g., removing full name from collected data before analysis or release
- The approach doesn’t work when statistical inference attack is performed

Statistical Inference Attack

- Uses data analysis 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 data can be queried by any user publicly
  - Assume that the adversary can choose the query (e.g., full name and date)
  - Could cross-reference with external knowledge about full name or date
  - To find a particular subject’s sensitive information with high confidence

Netflix De-Anonymization

- Narayanan and Shmatikov de-anonymization technique (2008)
  - 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

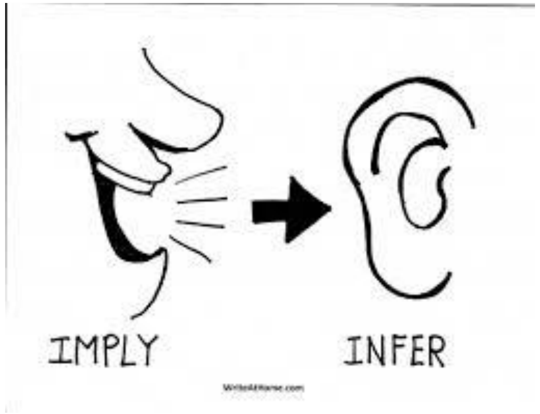
- Applied to Netflix Prize dataset
  - Anonymized dataset of 500,000 Netflix subscribers
  - Finding: simply removing identifying information is insufficient for anonymity
- How much does an adversary need to know about a Netflix subscriber to identify if her record is in the DB?
  - Auxiliary info: 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)

Preventing Data 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



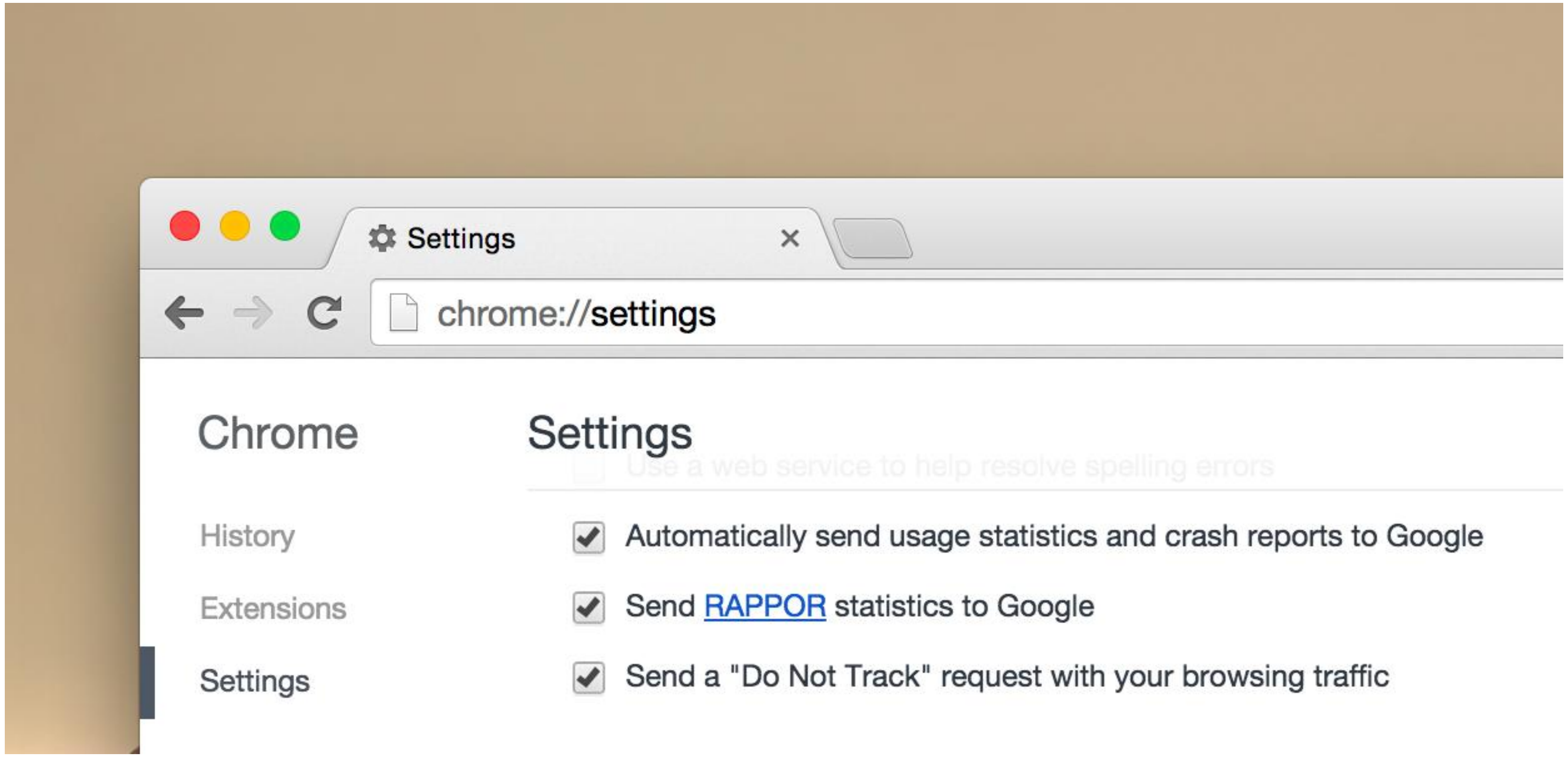
Differential Privacy

- Consider a party that holds a dataset of sensitive information (*e.g.* medical records, voter registration information, email usage)
  - Its goal is to provide global, statistical information about the data publicly available, while preserving the privacy of the included users.
- "Epsilon"-Differential Privacy
  - Assume two datasets D1 and D2 only differs a single element (data about one person)
  - A randomized algorithm A (for providing global, statistical info) is epsilon-differentially private if $\Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \times \Pr[\mathcal{A}(D_2) \in S]$.
  - 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
  - Add noises to ensure the cost does not exceed the budget or allows the query if so
- Example systems:
  - Google RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response)
  - Apple's differential privacy deployment

Google RAPPOR

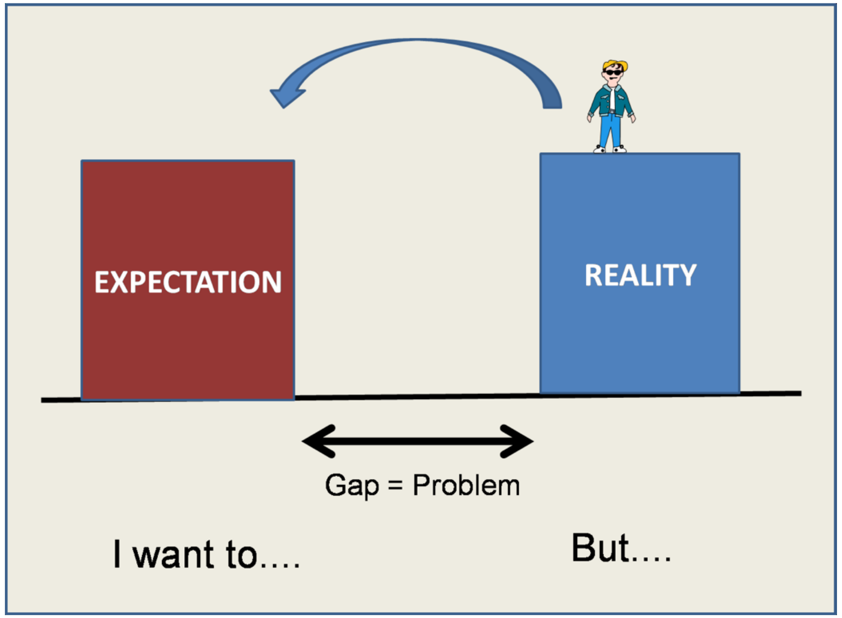


Preventing Communication Inference

- 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

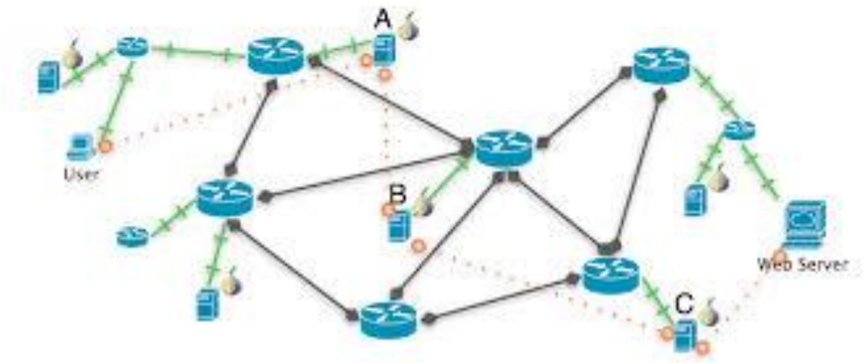
Reasonable Expectation

- 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

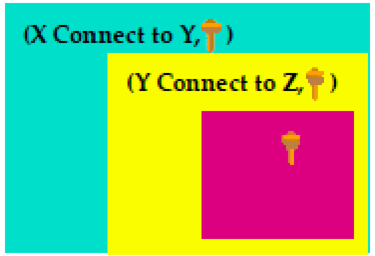


Onion Routing

- 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”

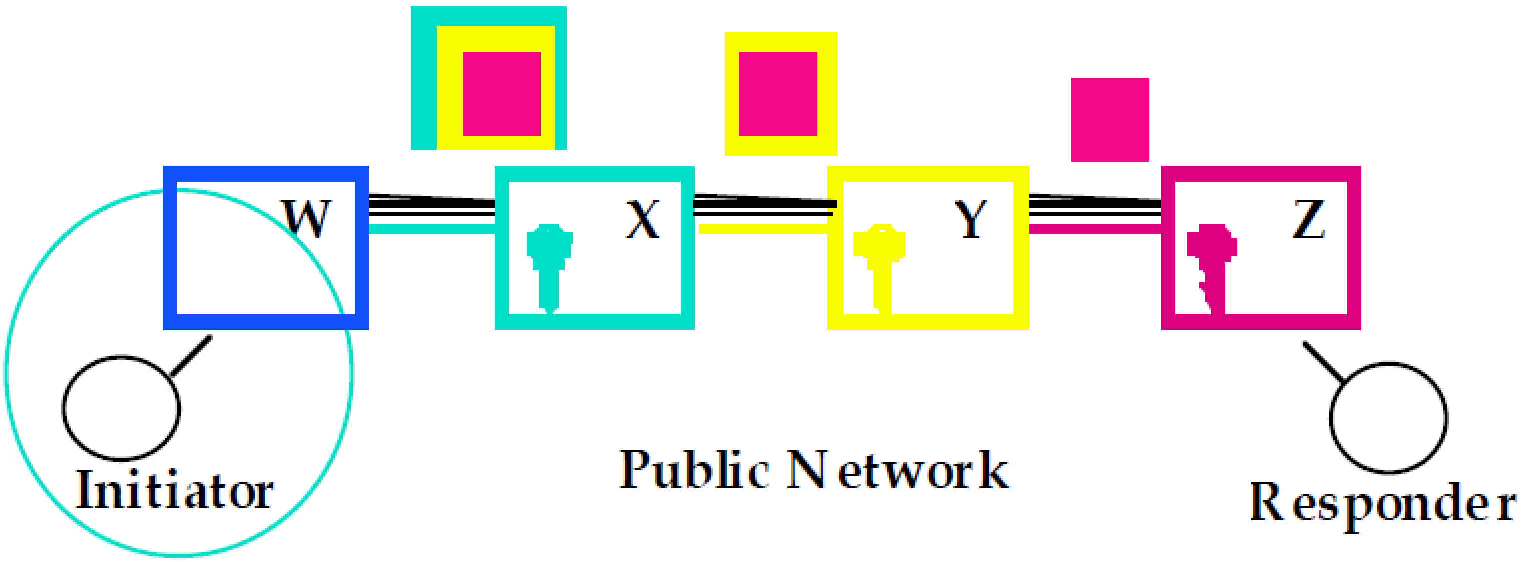
Onion

- 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



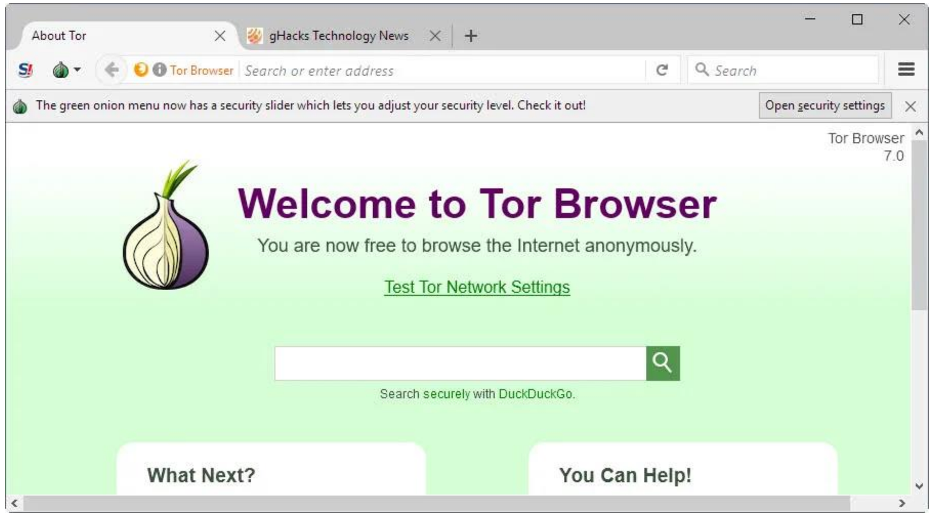
Limitations of Onion Routing

- 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

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

Privacy Impacts of Emerging Technologies

- RFID
- Electronic Voting
- VoIP
- Cloud Computing

Radio Frequency Identification (RFID)

- RFID tags are small, low-power, small-distance wireless radio transmitters
  - 5 centimeters to several meters
- When a tag receives a signal from a RFID reader on the correct frequency, it responds with its unique ID number
- Deployment: implanted under skin, embedded in credit card or identity badge, placed in shipping or inventory label

RFID for entrance management

RFID Privacy Concerns

- RFID tracking could be pervasive
  - As RFID tags become more prevalent, and RFID readers are installed in more places, it becomes possible to track individuals wherever they go
  - As RFID tags are put on more items, it will become increasingly possible to discern personal information by reading those tags

RFID tracking

Mitigation: RFID sleeve/shield

Electronic Voting

- Paper ballots are inefficient
- Electronic voting automatically collects ballots and is far more efficient
- How to protect voter privacy?
  - Voter info can be learnt by machine
- Solution needs to ensure accountability as well
  - Encrypting a vote using public key protects confidentiality, but unauthorized people can vote then

Electronic Voting Machine

Voice over IP (VoIP)

- VoIP transmits voice traffic over Internet, instead of traditional PSTN network
  - Your analog voice is converted to digital signals sent over Internet
- Major VoIP carriers: Skype, Google Talk and Vonage
- While VoIP provides encryption to voice calls, it also allows service providers to track calls' sources and destinations
  - VoIP provider has end-to-end visibility

VoIP network

Cloud Computing

- Cloud computing is on-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user
- Because resources are managed by cloud provider, a third-party, there are privacy issues related to regional laws
  - Physical location of information in the cloud may have significant effects on privacy and confidentiality protections
  - Cloud data may have more than one legal location at a time, with different legal consequences
  - Laws could oblige cloud providers to examine user data for criminal activity
  - Legal uncertainties make it difficult to assess the status of cloud data

Summary

- What data is considered private is subjective
- Privacy laws vary widely by jurisdiction
- New privacy issues: Inference attacks
  - Protection: differential privacy
- New privacy enhancement technologies: anonymous communication
- Emerging technologies are fraught with privacy uncertainties, including both technological and legal issues

Slides credit

- Module: Privacy, Trent Jaeger
- Online Tracking, Amir Houmansadr