Privacy

What is Privacy?

- "The right of individuals to control or influence what information related to them may be collected and stored and by whom and to whom that information may be disclosed." (OSI Reference Model)
- "Privacy is the interest that individuals have in sustaining a 'personal space', free from interference by other people and organisations."
- "[T]he house of every one is to him as his castle and fortress." (Semayne's Case, 1603)
- "The right to be let alone." (Future U.S. Supreme Court Justice Louis Brandeis, 1890)

Why Protect Privacy?

- "You have zero privacy anyway. Get over it". (Scott McNealy, CEO, Sun Microsystems)
- (Also see David Brin's *The Transparent Society*)
- That said, people do care
- From a purely pragmatic perspective, organizations that get caught in privacy violations can suffer
- Real risks: blackmail, job-hunting problems, relationship problems, insurance problems, identity theft

Why Protect Privacy?

"Privacy is a fundamental tenet of legal systems and political philosophies that value individual freedom, autonomy, and political participation. . . The underlying values that privacy protects include individuality and autonomy; intimacy; fairness; and limited, tolerant government." (National Research Council)

Kinds of Privacy

Bodily integrity Protects the individual from intrusive searches and seizures;

Decisional privacy Protects the individual from interference with decisions about self and family;

Information privacy Protects the individuals interest in controlling the flow of information about the self to others;

Communications privacy A subset of information privacy that protects the confidentiality of individuals communications.

Privacy is not Confidentiality

- *Privacy* is a reason for confidentiality
- More than confidentiality is needed to protect privacy
- Confidentiality protects more than just privacy

Abuses of Privacy

- Reading traffic
- Learning identity
- Tracking identity
- Tracking behavior

Reading Traffic

- Reading traffic is easy
- Easy way to collect passwords, too
- Especially easy on wireless nets. . .

Eavesdropping Example


```
$ telnet example.com 110
+OK Cubic Circle's v1.31 1998/05/13 POP3 ready <56ec
user smb
+OK smb selected
pass secret
-ERR cucipop: Invalid password or username (check ca
quit
+OK Not really your day, is it?
Connection closed by foreign host.
```

Eavesdropping


```
# dsniff
dsniff: listening on bge0
-----------------
04/26/05 01:17:15 tcp gg1.cs.columbia.edu.63471 -> e

user smb

pass secret
```

But recovering the password isn't the point

You Can Learn a Lot That Way—

- What is the content of the email?
- Who are the correspondents?

  $\Rightarrow$ Traffic analysis
- What web pages does the target visit?

Traffic Analysis

Traffic Analysis

- Who talks to whom
- How often, for how long?
- Often much more useful than actual content

Why is it Useful?

- Very hard to hide
- Even encryption doesn't block traffic analysis
- Can show chain of responsibility
- More amenable to machine processing (no need to parse speech or text)

Example

- The (false) alert is an example of traffic analysis: a CS machine was trying to talk to invalid addresses
- Pick out the botnet controller
- Find out who else the botnet controller is talking to

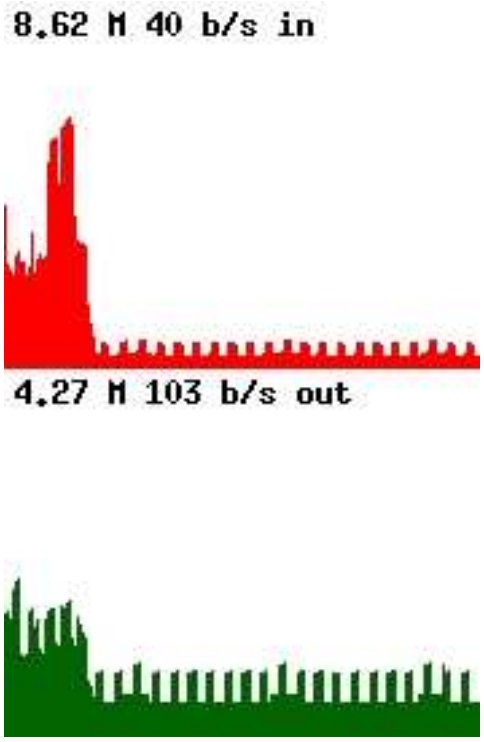
Web Data

- What web sites or URLs does the target visit?
- Note: image sizes can be quite distinctive
- Can be combined with other analyses

Mail Left in Draft Folders

- Allegedly, al Qaeda members compose messages, but leave them in draft folders on Web mailers
- That way, they're never sent and monitored, but someone else logs in and picks them up
- Look for connections that upload/download a lot of data
- Correlate with logins to accounts that don't send or receive email

Application Identification



- The low, spikey pattern at the right is an IM client sending keep-alives
- The larger peak at the left is email retrieval
- Note how the IM pattern is identifiable even when super-imposed on the email pattern

Mail Logs

- Who's talking to whom?
- Can be sensitive within an organization
- (Complex) interpersonal relationships
- Who's leaking information?

From the SAGE Code of Ethics

"I will access private information on computer systems only when it is necessary in the course of my technical duties. I will maintain and protect the confidentiality of any information to which I may have access, regardless of the method by which I came into knowledge of it."

From the ACM Code of Ethics

"It is the responsibility of professionals to maintain the privacy and integrity of data describing individuals. This includes taking precautions to ensure the accuracy of data, as well as protecting it from unauthorized access or accidental disclosure to inappropriate individuals. . .
"User data observed during the normal duties of system operation and maintenance must be treated with strictest confidentiality, except in cases where it is evidence or the violation of law, organizational regulations, or this Code. . ."

Web Bugs

- Embed unique image URL in email or web page
- See who retrieves that URL
- Note: most HTML mailers ignore IMG tags, for just that reason
- But it works well for 3rd-party web ads

Personally Identifiable Information

- Variety of daily activities that leave trails – cookies, third party cookies, website registration
- Cooperation with third party advertising sites – ISPs watching traffic for sale to advertisers, inserting ads in traffic (Rogers in Canada)
- top 10 third party cookies in 70% of first party servers (top 2: Doubleclick, Goolge Analytics)
- top 10 third party families in over 60% fiduciary sites (top 2: Google, Omniture)

Log Files

- System log, Auth.log, Sudo.log, History log, Apache log, Search log
- Trace back malicious behavior on the system
- Anonymization if sensitive information – IPs, times tamps, port numbers, file/directory names, passwords, packets sizes, search queries; **hard problem**
- AOL "anonymized" queries database release

Authentication

Authentication

- Authentication schemes can impact privacy
- Logins leak information
  - Common usernames
  - Common passwords
  - Common biometrics, such as fingerprints
- Who has access to the records?

Biometrics

- Hard to change a biometric
- Easy to correlate biometrics across sites
- (Many other problems)

Secondary Uses

Linkages and Secondary Uses

- Often, the primary use of gathered data is innocuous
- But too much data is sometimes collected
- *Secondary uses*, such as using drivers' licenses as an airplane boarding card and a liquor authorization card, create much more trouble
- Example: some bars use swipe readers, not just to verify the authenticity of the license, but also to collect names, addresses, and demographic data

Example: Drivers' License Verifiers

- Some bars use a swipe reader to verify drivers' licenses
- Easier to fake picture and text than mag stripe
- (Actually, writing a mag stripe isn't hard. . .)
- But — the readers record name, address, gender, etc., and build up databases

Databases

- Corporations — and sometimes the government — collect massive databases on personal behavior
- Credit records are the obvious example
- In the U.S., *all* medical insurance claims are tracked by the Medical Information Bureau (MIB).

Example: Digital Content and Digital Rights Management

- Nominal purpose is to ensure that you've paid for content
- But the content owner then knows exactly what you watch or listen to
- What does TiVo know about your viewing habits?

Fair Information Practices

- First "code of fair information practices" developed in 1973 at HEW
- Basic rules for minimizing information collection, ensuring due process, protection against secret collection, provide security, ensure accountability
- Emphasize individual knowledge and consent
- Principles are broadly accepted, but individual principles not implemented uniformly

Fair Information Principles and Practices

- Collection limitation
- Data quality
- Purpose specification
- Use limitation
- Security
- Openness/notice
- Individual participation
- Accountability

Legal Protecions

- U.S.: Patchwork of laws, i.e., FERPA, Video Privacy Protection Act
- Limited U.S. constitutional protection inferred by Supreme Court
- Few limits in the U.S. on private sector behavior
- EU: Strong, mandatory privacy protections

Defenses

- Encryption
- Mixnets
- Authorization-based credentials
- Minimization

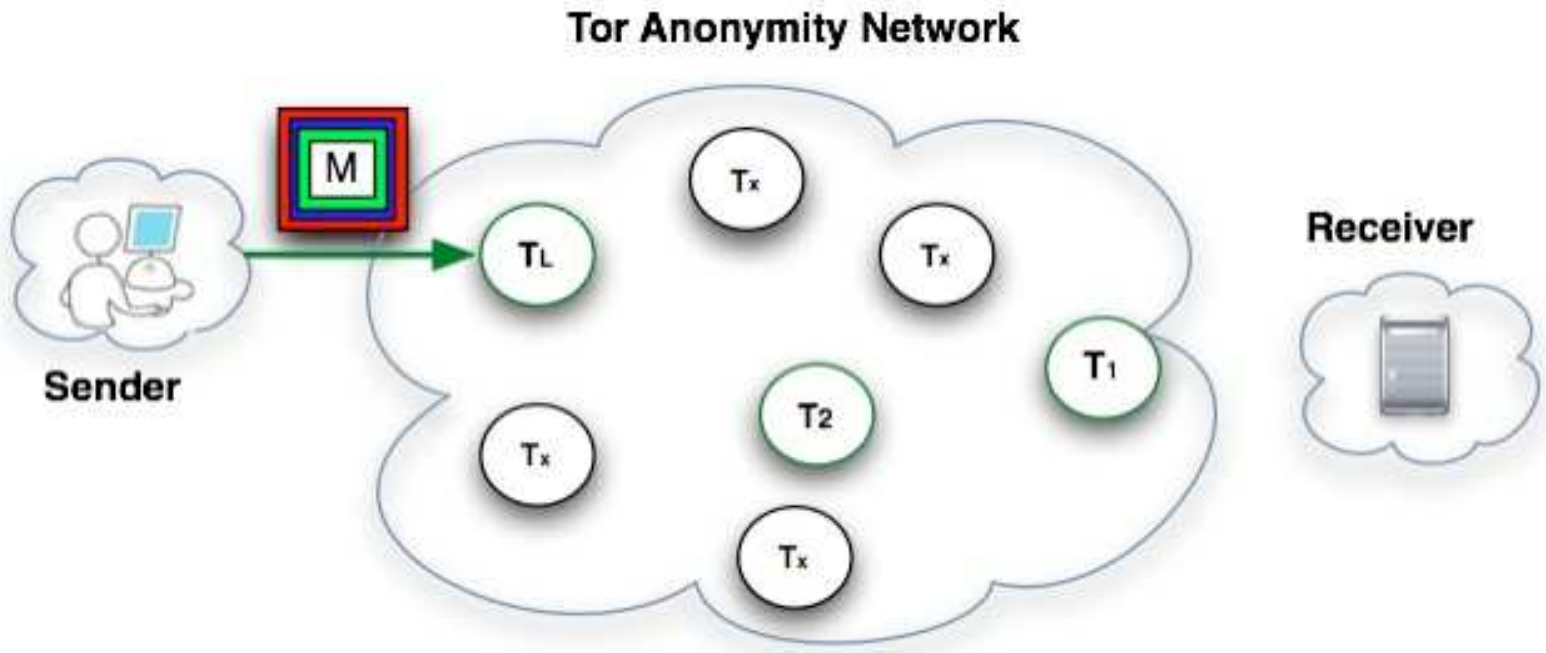
Encryption

- The obvious solution
- Very hard to guard against traffic analysis
- Doesn't guard against misuse by authorized parties
- Difficult to deploy in large-scale systems

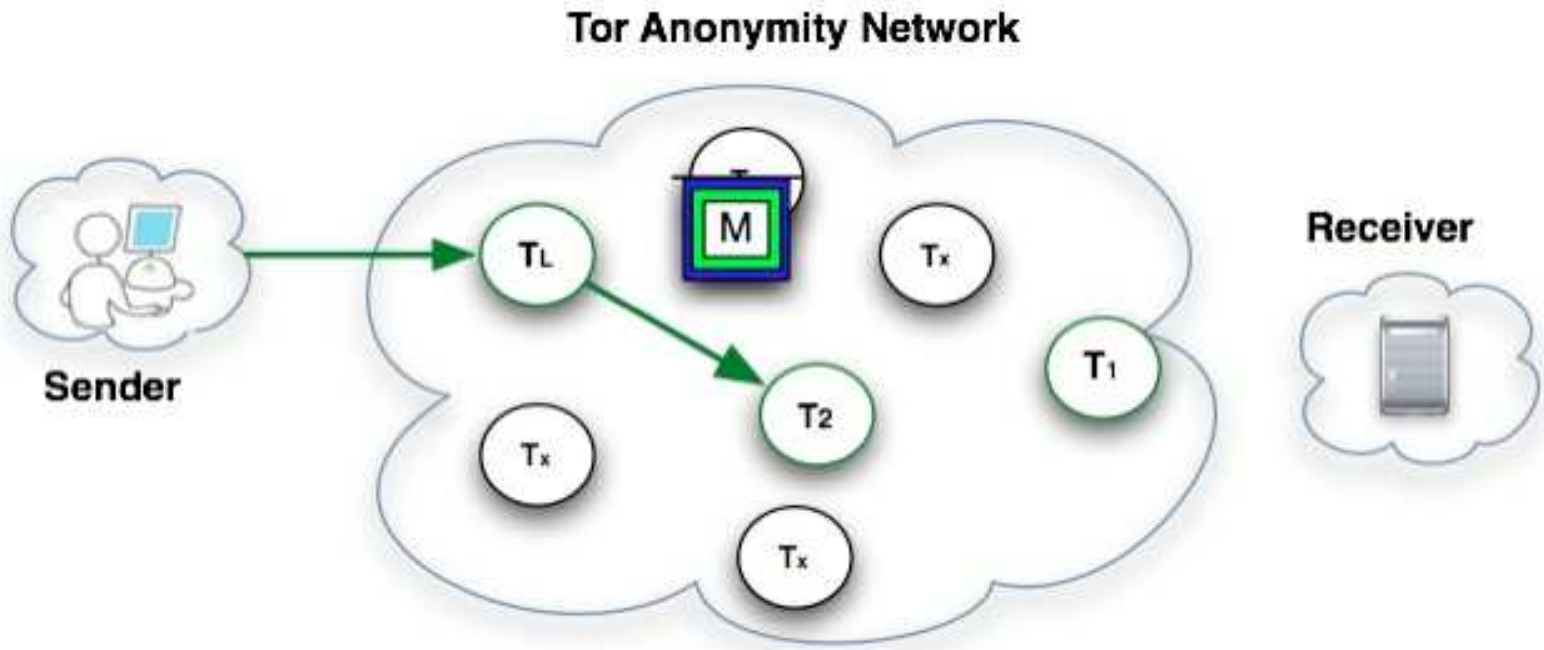
Mixnets

- Aggregate traffic
- Insert dummy traffic
- Delay traffic
- Chain through multiple mix nodes
- Goal is to prevent traffic analysis
- Real-world systems, such as Tor, do this

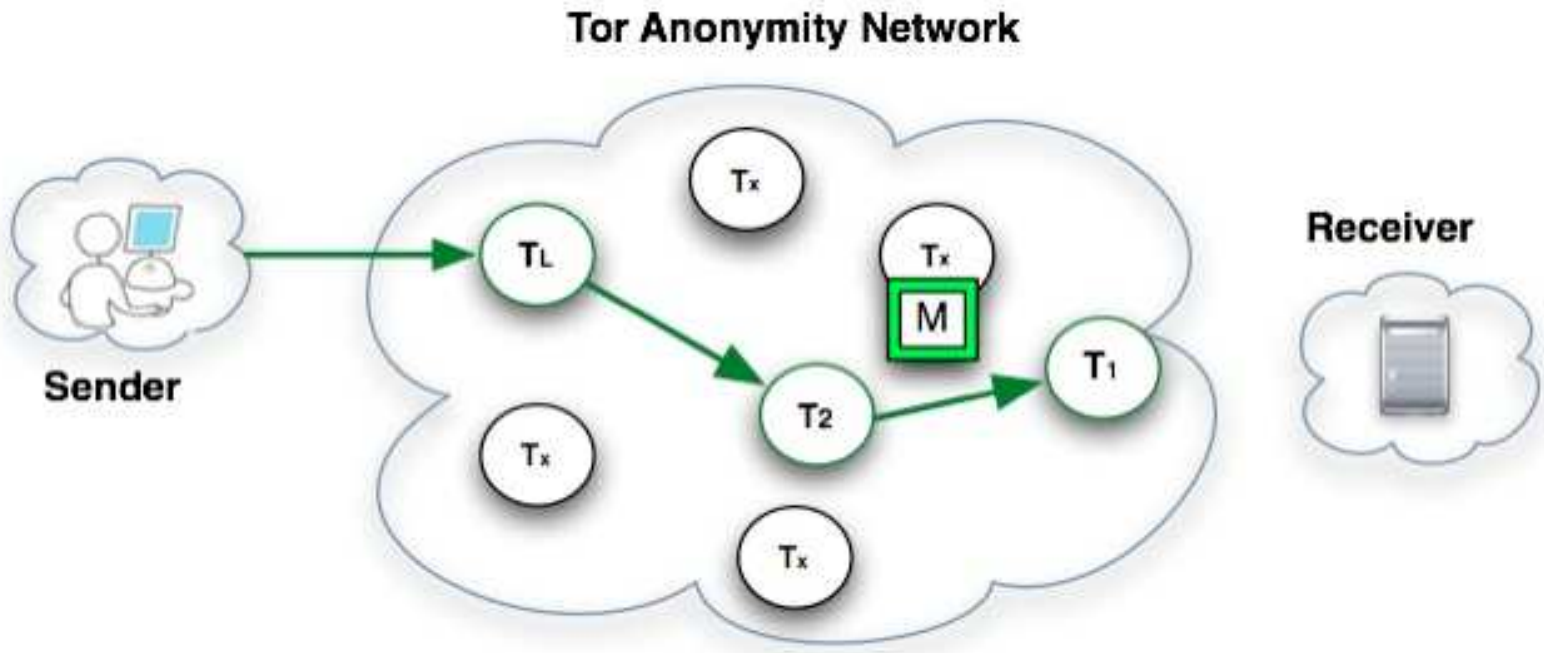
Tor Anonymity Network



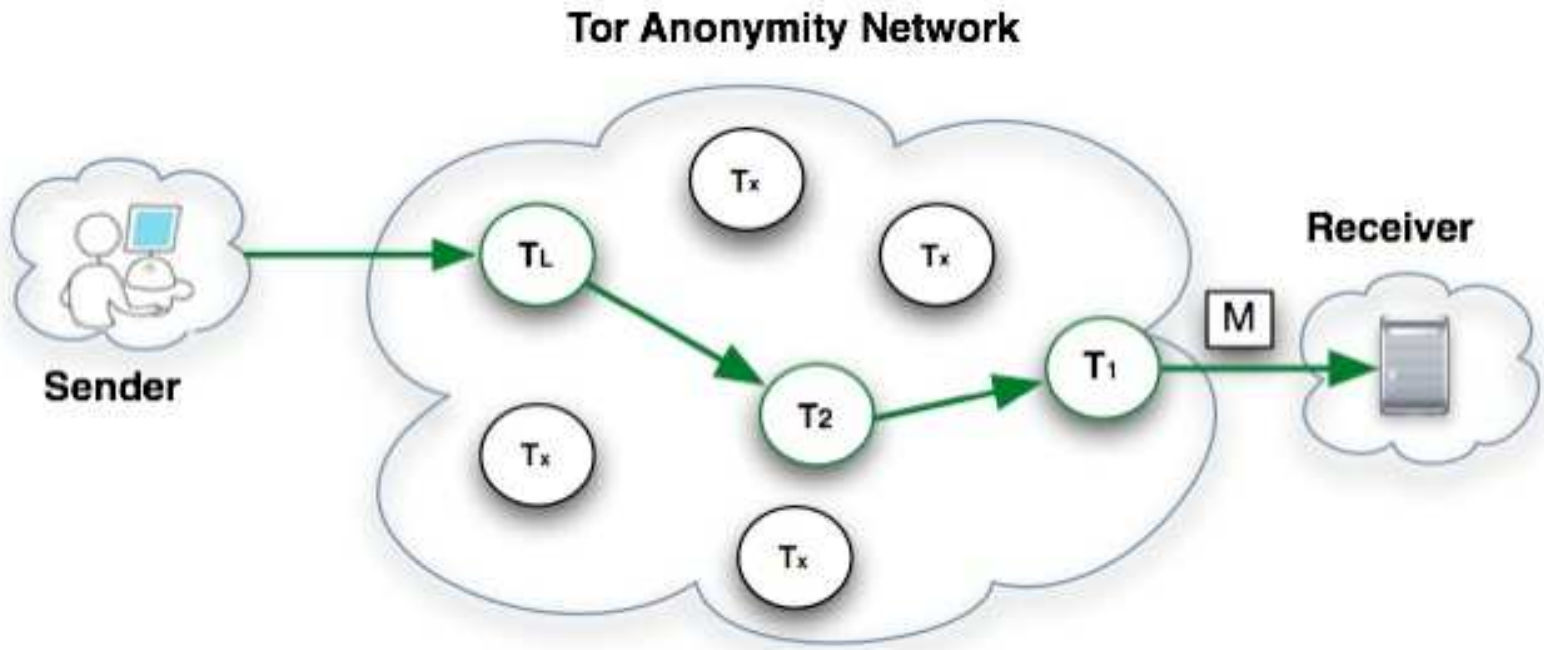
Tor Anonymity Network



Tor Anonymity Network



Tor Anonymity Network



Authorization-based Credentials

- Give users some sort of anonymous token that grants access
- Example: Cash versus credit cards (yes, merchants track you by credit card number)
- Rarely used — people don't think that way

Digital cash

- Electronic money that provides anonymity of purchases similarly to real cash
- Often heavy crypto
- No infrastructure exisiting to facilitate transition to digital cash, credit cards work OK, privacy concerns not high enough
- Issues: might create hurdles for proper taxation, money laundering

Minimization

- Don't collect data unless you need it
- Data that doesn't exist can't be misused
- Data that doesn't exist can't be compromised

Preserving Privacy

- Plan for it from the beginning
- Minimize collection
- Use security mechanisms to protect data
- Make sure management buys in