CPSC156: The Internet
Co-Evolution of Technology and Society

Lecture 24: April 24, 2007

Review for Second Exam

In-class exam: Thurs., 4/26/07

- Test on material in Lectures 12 (Feb. 22) through 22 (Apr. 17).
- Lecture notes
- Homework assignments and solution sets
- Exams and HWs from earlier version of 156 and 155.
- Reading assignments

Topics

- **Open source**
- Public records
- Identity, anonymity, and accountability
- Privacy
- Cryptographic primitives
- Spam and viruses
- Browser-based security

Open Source

- Lecture 12 (February 22, 2007)
- Review the basics of copyright law that you studied for Exam 1, because copyright law is one of the foundations of the open-source movement.
- Review the beginning of Lecture 12 so that you understand the model to which open source is an alternative.

Open Source Basics

- Technical definition:
  Anyone can look at the source code.
- Benefits:
  - Interoperability
  - Education
  - Cross-platform compatibility (if code can be compiled by users)
- Still protects intellectual property:
  - Uses of code are still limited by a license.
  - Developers maintain rights to code and official releases of the product.

Consequences of Open Source

- Software can be closely scrutinized; performance can be analyzed and attributed to parts of code.
- Ideas behind code can become standards.
- **Distributors** can specialize in building more features on top of open-source software, offering customized packages with support options.
- "Open source" is a more business-friendly term than "free."

Free Software

- Technical definition (from the *Free Software Foundation*): Users have the freedom to:
  (1) run the software, for any purpose;
  (2) study how the program works and adapt it to their needs;
  (3) redistribute copies;
  (4) improve the program and release improvements to the public.
- Access to source code is necessary for (2) and (4); so, "Free" can include "Open Source."

Software Licenses

- The software license indicates what users can do with software and code.
- **Traditional licenses** strictly govern use of software based on purchase.
- **Open-source and free licenses** indicate how code can be used, reused, and distributed, usually asserting user rights like the "four freedoms."

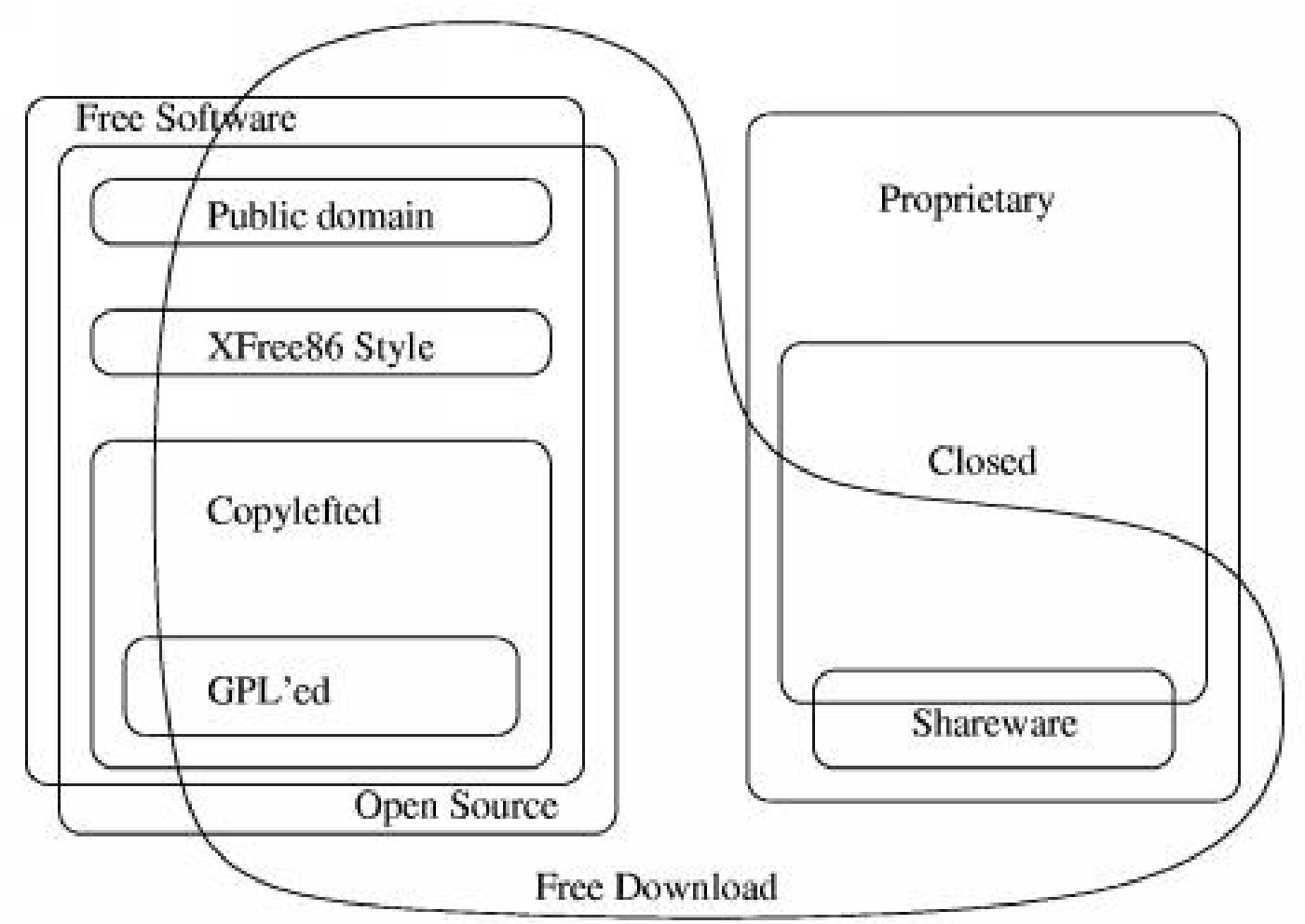
GNU General Public License (GPL)

- Formalization of ideal software-distribution model by the Free Software Foundation and the GNU Project.
- Developers can choose to release their "free" software under the GPL license.
- Requires that users maintain the original copyright on the code and clearly mark any changes when distributing it.
- Source code is included, and users can modify and compile it where and how they see fit.
- **Copyleft:** Redistributed versions must give users the same rights (must include source code that can be modified, *etc.*).

Other Popular Licenses

- Many licenses exist that come from organizations that develop "free" software.
- **GPL-compatible** licenses are those that allow software covered to be combined with GPL-covered software to produce larger "free" products.
  - Examples (non-copyleft): MIT, BSD (Berkeley), X11 (windowing system)
- Other public licenses: Netscape-JavaScript, Artistic, W3C Software

Relationships Among Models



Source: Free Software Foundation

Topics

- Open source
- **Public records**
- Identity, anonymity, and accountability
- Privacy
- Cryptographic primitives
- Spam and viruses
- Browser-based security

Public Records

- Lecture 14 (March 6, 2007)
- More in-depth explanations of the issues raised in these notes can be found on the EPIC website.

"Public Records" in the Internet Age

Depending on State and Federal law, "public records" can include:

- Birth, death, marriage, and divorce records
- Court documents and arrest warrants (including those of people who were acquitted)
- Property ownership and tax-compliance records
- Driver's license information
- Occupational certification

They are, by definition, "open to inspection by any person."

How "Public" are They?

Traditionally: Many public records were "practically obscure."

- Stored at the local level on hard-to-search media, *e.g.*, paper, microfiche, or offline computer disks.
- Not often accurately and usefully indexed.

Now: More and more public records, especially Federal records, are being put on public web pages in standard, searchable formats.

What are "Public Records" Used For?

In addition to straightforward, known uses (such as credential checks by employers and title searches by home buyers), they're used for:

- Commercial profiling and marketing
- Dossier compilation
- Identity theft and "pretexting"
- Private investigation

Discussion point: Will "reinventing oneself" and "social forgiveness" be things of the past?

Do We Need a More Nuanced Approach?

Can we distinguish among

- Private information
  - Only the "data subject" has a right to it.
  - ? Example: Legal activity in a private home.
- Public information
  - Everyone has a right to it.
  - ? Example: Government contracts with businesses
- Nonpublic personal information
  - Only parties with a legitimate reason to use it have a right to it.
  - Example: Certain financial information (see, *e.g.*, the Graham-Leach-Bliley Act)

Discussion point: Should some Internet-accessible "public records" be only conditionally accessible? Should data subjects have more control?

Topics

- Open source
- Public records
- **Identity, anonymity, and accountability**
- Privacy
- Cryptographic primitives
- Spam and viruses
- Browser-based security

Identity, Anonymity, and Accountability

- Lecture 15 (March 8, 2007)
- Reading assignment "Identity and Anonymity: Some Conceptual Distinctions and Issues for Research," by G. Marx

1. A Person's Legal Name

- Usually an answer to the question "Who are you?"
- Involves connection to biological and social lineage.
- Many people may have the same name, but the assumption is often made that there is at most one person of each name born to particular parents at a given time and place.

2. A Person's Address

- Usually an answer to the question "Where are you?"
- Involves location and reachability in actual space or cyberspace.
- Need not involve knowing the person's name or even a pseudonym.
- Note that a person may be unreachable even if his name and address are known; this was true of, *e.g.*, Robert Vesco when he was a fugitive in Cuba.

3. Unique ID: Linkable

- Unique alphanumeric strings, biometric patterns, or pseudonyms that *can* be linked back to actual people *but need not be*.
- Involves trusted intermediaries and conditions under which they should link IDs to people.
- Social security numbers could be used in this way if we had widespread agreement on how they should be used and when they should be linked to names and addresses.

4. Unique ID: Unlinkable

- Unique alphanumeric strings, biometric patterns, or pseudonyms that *cannot* be linked back to actual people.
- Provides a means of discerning information about people without identifying them; someone tested for AIDS may be given a number that he can use to call for results but never have to reveal his name or address.
- Spies, undercover operatives, and con artists may use fraudulent IDs and never reveal their real names to those they deal with.

5. Distinct Appearance or Behavior Patterns

- Some information is necessarily revealed when one interacts with others.
- "Being unnamed is not necessarily the same as being unknown." To a limited extent, you "know" the person you see at 8:15 a.m. on the M23 bus every day.
- Leakage of identifying information is a condition of social existence and has been greatly expanded by new technologies.

6. Social Categorization

- Forms of identification that do not distinguish among members of a group; the group may be defined by gender, ethnicity, religion, age, economic class, *etc.*
- Number of categories has exploded with new technology and expanded bureaucracy.
- New categories (credit scores, IQs, life-style categories used in marketing, *etc.*) may or may not be known by the people in them; this was not true of traditional social categories.

7. Certification: Proof of the possession of knowledge or skill

- Knowing a secret password, being able to swim, *etc.* are ways to prove that one is entitled to certain privileges or is a member of a certain group.
- These proofs may be linkable to individual people (as passwords often are) but need not be.
- Provides essential balance between the need to control sensitive personal information and the need to restrict access to and prevent abuse of systems.

Topics

- Open source
- Public records
- Identity, anonymity, and accountability
- **Privacy**
- Cryptographic primitives
- Spam and viruses
- Browser-based security

Privacy

- Lectures 16 and 17 (March 27 and 29)
- Reading assignment "A Taxonomy of Privacy," by D. Solove
- Consider the implications of this taxonomy of privacy in the Internet environment.

Motivation for this Work

"Under the **secrecy paradigm**, privacy is tantamount to complete secrecy, and a privacy violation occurs when concealed data [are] revealed to others. If the information is not previously hidden, then no privacy interest is implicated by the collection or dissemination of the information."

Solove's thesis in this article is that the secrecy paradigm has strongly influenced court decisions but is a thoroughly inadequate organizing principle for privacy law.

A. Information Collection

1. Surveillance
2. Interrogation

B. Information Processing

1. Aggregation
2. Identification
3. Insecurity
4. Secondary Use
5. Exclusion

C. Information Dissemination

1. Breach of Confidentiality
2. Disclosure
3. Exposure
4. Increased Accessibility
5. Blackmail
6. Appropriation
7. Distortion

D. Invasion

1. Intrusion
2. Decisional Interference

Question: General Approach

Is this a useful taxonomy? Are these 16 categories truly distinct, and are they collectively exhaustive of all privacy violations?

Is this highly particularized, incremental approach the right one, or would a broad, simply stated "right to privacy" be more effective?

Question: Relationship to Digital Identity

Compare Solove's privacy taxonomy to Marx's identity taxonomy. (See lecture 15 and the March 8 reading assignment.)

Are the authors' conceptual frameworks consistent, inconsistent, complementary, or overlapping? Would widely available pseudonymous or anonymous communication make Solove's goal simpler?

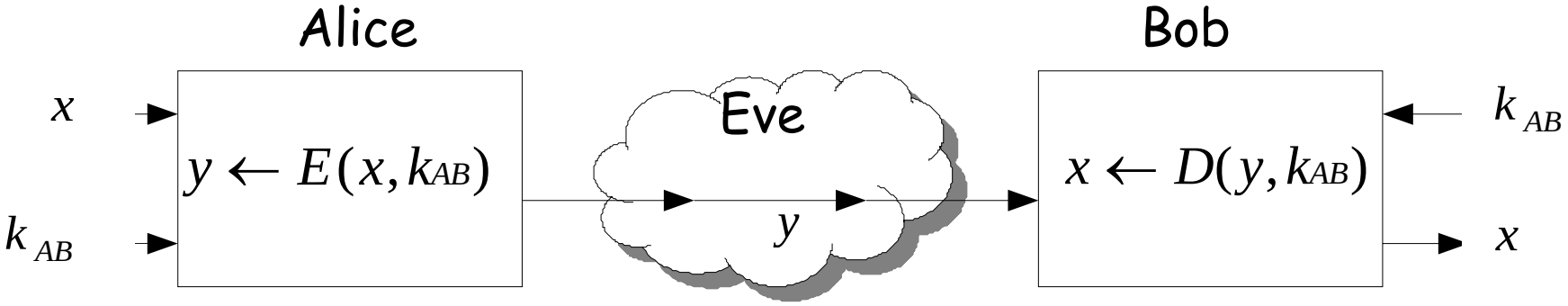
Topics

- Open source
- Public records
- Identity, anonymity, and accountability
- Privacy
- **Cryptographic primitives**
- Spam and viruses
- Browser-based security

Cryptographic Primitives

- Lectures 19, 20 and 21 (April 5, 10, and 12, 2007)
- Excerpt from Schneier's "Applied Cryptography" (distributed in paper form)
- Consider how these primitives are used in various Internet technologies that we have discussed this semester, including the ones that you used in HW4 and HW5.

Symmetric-key Encryption (1)



x: plaintext

y: ciphertext

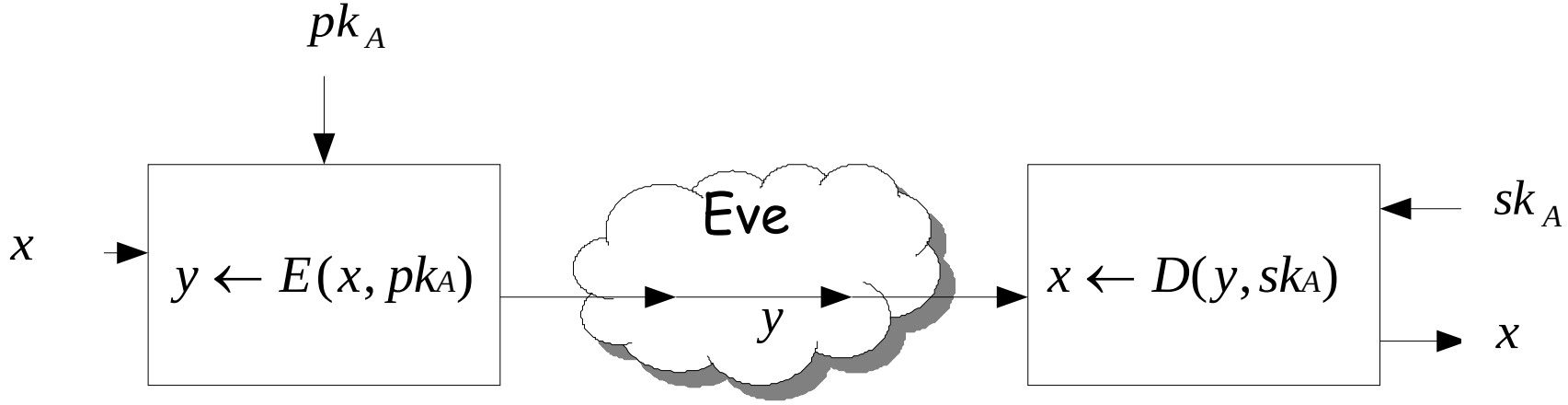
E: encryption function

D: decryption function

k_{AB}: Alice's and Bob's shared secret key

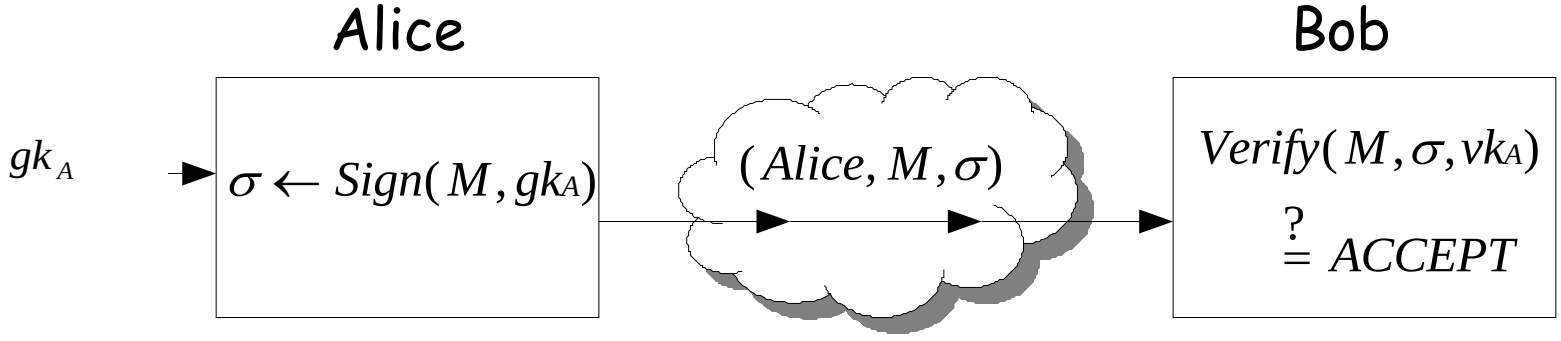
Public-key Encryption (1)

- Alice runs the *key generator* to obtain a *public-key, secret-key pair* (pk_A, sk_A).
- She publishes "Alice: pk_A" and keeps sk_A secret.
- To communicate securely, Bob first looks up pk_A.



Digital Signatures (1)

- Alice runs the *key generator* to obtain a *verification-key, signing-key pair* (vk_A, gk_A).
- She publishes "Alice: vk_A" and keeps gk_A secret.
- To verify her signature, Bob first looks up vk_A.



Discussion Point

Note that Alice's digital signature on digital document M_1 will be different from her digital signature on digital document M_2. Thus, digital signatures are not analogous to handwritten signatures.

Why is this necessary?

Certificates

- We can leverage our confidence in the integrity of a single verification key.
- Let *CA* (for "certifying authority") be an entity with a valid, highly available verification key vk_{CA}. Verisign is an example of a *CA*.
- The *CA* can verify that Alice, Bob, Eve, *etc.*, have appropriate IDs and have not presented keys that are already owned by someone else. It can then publish "certified" name-key pairs:

  $(Alice, pk_A, Sign((Alice, pk_A), gk_{CA}))$,

  $(Bob, pk_B, Sign((Bob, pk_B), gk_{CA}))$,

  $(Eve, pk_E, Sign((Eve, pk_E), gk_{CA}))$, …

Cryptographic Hash Functions

- A "hash function" h maps an arbitrary-length input to a fixed-length (*e.g.*, 256-bit) output. Note that, of necessity, there are many inputs x_1, x_2, x_3, … that are mapped to the same hash value y.
- h is a (keyed) "cryptographic hash function" if

  - each user has a secret key k that he uses to compute $y = h_k(x)$;

  - it is computationally infeasible for someone who does not know k to find *any* x* such that $h_k(x^*) = y$, even if he knows h and y.
- Note that this is different from encryption, in which each ciphertext produced with a given key corresponds to a *unique* plaintext that must be recoverable.

Topics

- Open source
- Public records
- Identity, anonymity, and accountability
- Privacy
- Cryptographic primitives
- **Spam and viruses**
- Browser-based security

Spam and Viruses

- Lecture 18 (April 3, 2007)
- Go over the technical slides carefully and think about how the cryptographic primitives we discussed play a role.
- HW4

What is Spam?

Source: Mail Abuse Prevention System, LLC

- Spam is unsolicited bulk e-mail (primarily used for advertising).
- An electronic message is spam IF:

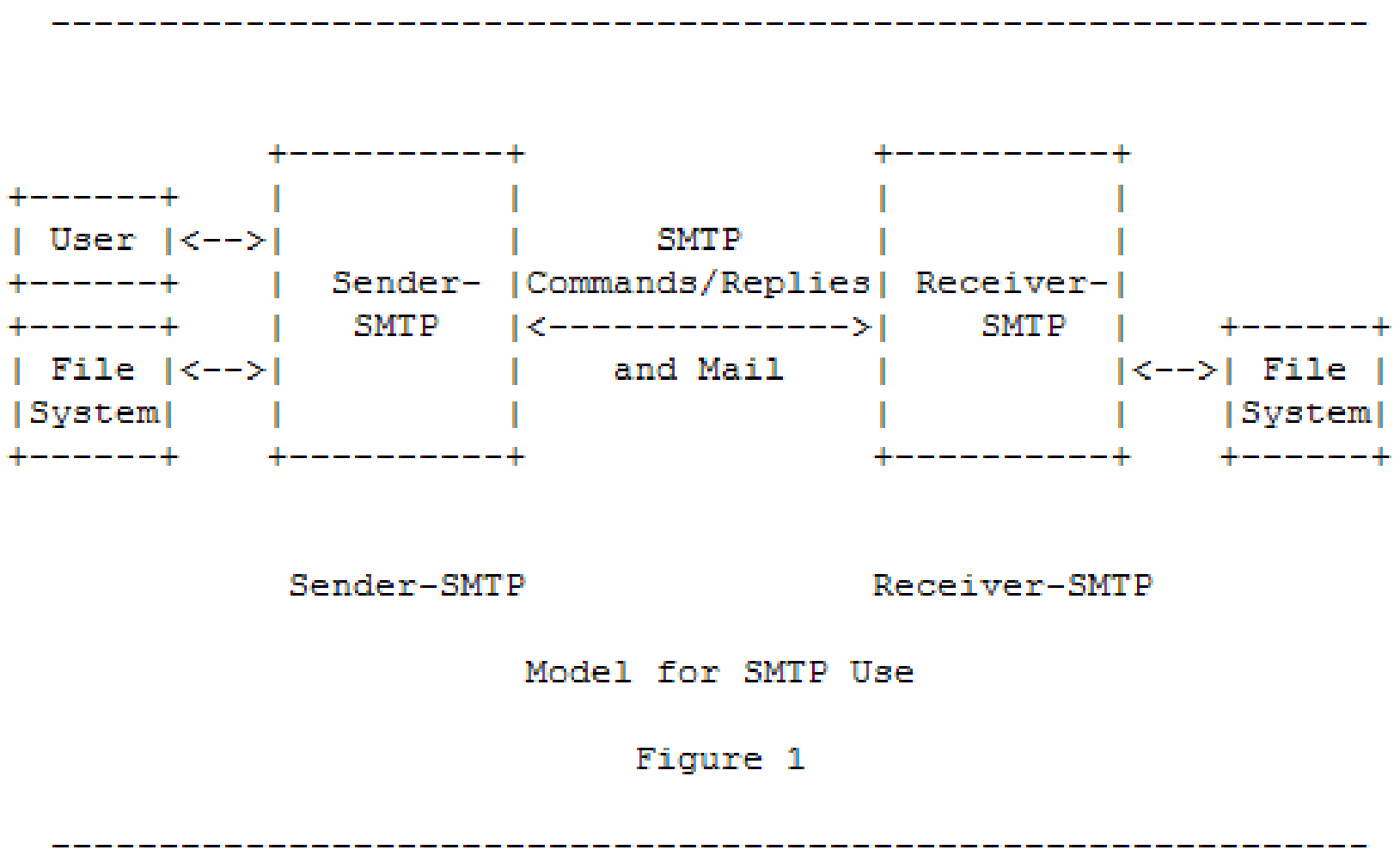
(1) the recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients; AND

(2) the recipient has not verifiably granted deliberate, explicit, and still-revocable permission for it to be sent; AND

(3) the transmission and reception of the message appears to the recipient to give a disproportionate benefit to the sender.

How is E-mail Sent?

Source: RFC 821 (SMTP)



Model for SMTP Use

Figure 1

Tracking Spam

- SMTP runs on top of TCP.
  - Packets are acknowledged.
  - **Source** of packets is known in any successful mail session.
- SMTP servers add the IP address and hostname of every mail server or host involved in the sending process to the e-mail's message header.
- **But,** dynamic IP addresses and large ISPs can make it difficult to identify senders.

Spoofing E-mail Headers

- Most e-mail programs use (and most people see) only the standard "To," "Cc," "From," "Subject," and "Date" headers.
- All of these are provided as part of the mail data by the mail sender's client.
- Any of this information can be falsified.
- The only headers you can always believe are message-path headers from trusted SMTP servers.

Viruses

A **computer virus** is a piece of code, often malicious, that is intended to transmit itself between computers and replicate itself and/or execute instructions without the user's knowledge or intent.

Examples: Michelangelo, I-Love-You, Melissa, Slammer, Code Red

How Does One Get Infected?

Simple answer:

Run malicious code on your computer.

Simple reaction:

Then I won't.

Problem:

What if you are tricked into doing it?
Or don't know it's happening?

Types of Viruses

- **Trojan Horses:** disguised to do one thing, but do another when run
- **Boot Sector Viruses:** reside in system sectors; run in the background while resident in memory; copy themselves to other disks
- **File Infectors:** modify portions of executable files on disk so that virus code is unknowingly executed
- **Macro Viruses:** take advantage of the programmability of documents; run when infected files are accessed
- **Worms:** replicate across networks, possibly through proprietary software protocols
- **E-mail Viruses:** transmitted through e-mail, often through attachments

Topics

- Open source
- Public records
- Identity, anonymity, and accountability
- Privacy
- Cryptographic primitives
- Spam and viruses
- **Browser-based security**

Browser-based Security

- Lecture 22 (April 17, 2007)
- http://crypto.stanford.edu/antiphishing
- HW5