Lecture 1 Introduction & Overview

Dr. Chen Zhang
Department of Computer Science
The Hang Seng University of Hong Kong

What is Big data

Big data refers to large, diversified sets of data sourcing from multiple channels such as social media platforms, websites, electronic check-ins, sensors, product purchases, and call logs.

What’s the characteristics of Big Data?

Characteristics of Big data

- Volume: the size and amounts of big data that companies manage and analyze.
- Variety: the diversity and range of different data types, including unstructured data, semi-structured data and structured data.
- Value: the benefits that the organization derives from the data.

Characteristics of Big data

- Velocity: the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time.
- Veracity: the “truth” or accuracy of data and information assets.

Data breaches happen all the time

Yahoo

Date: 2013

Impact: 3 billion accounts

Over 3 billion customers' information (including account information such as security questions and answers) has been accessed by a hacking group.

Data breaches happen all the time

Facebook

Date: 2019

Impact: 533 million users

Two datasets from Facebook apps were exposed to the public internet. The information related to more than 530 million Facebook users and included phone numbers, account names, and Facebook IDs.

Data breaches happen all the time

Sina Weibo

Date: 2020

Impact: 538 million accounts

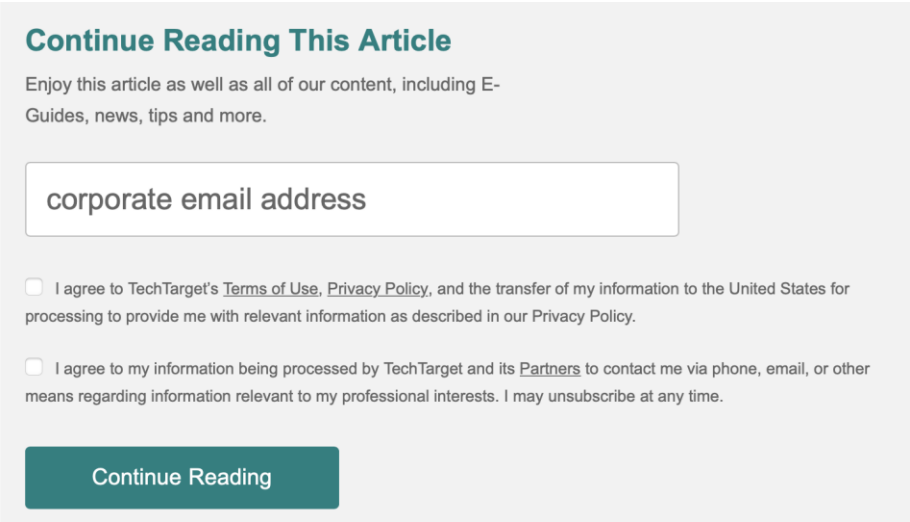
An attacker obtained part of its database, impacting 538 million Weibo users and their personal details including real names, site usernames, gender, location, and phone numbers.

There is a great need to protect big data security and privacy!

What's the difference between security and privacy?

Data Security vs. Privacy

- Privacy: the right to have some control over how your personal information is collected and used. Privacy focuses on individuals' **autonomy** and **decision-making** rights regarding their personal information.



Personal Data Often Collected When You Download Apps
https://www.youtube.com/watch?v=m4W-PLsSUhE

Data Security vs. Privacy

- Privacy: the right to have some control over how your personal information is collected and used.
- Security: protecting data from getting into the wrong hands, through a breach, leak, or cyber attack.

Can you have security without privacy?

Can You Have Security Without Privacy?

- Yes, but they go better together.
- For example: a company may write into their privacy policy that they can share or sell a user's data.
- privacy is less protected, but the organization's systems and the systems of those they sell the data to can be secure.
- The more that information is shared, the more likely their identity information will be leaked.

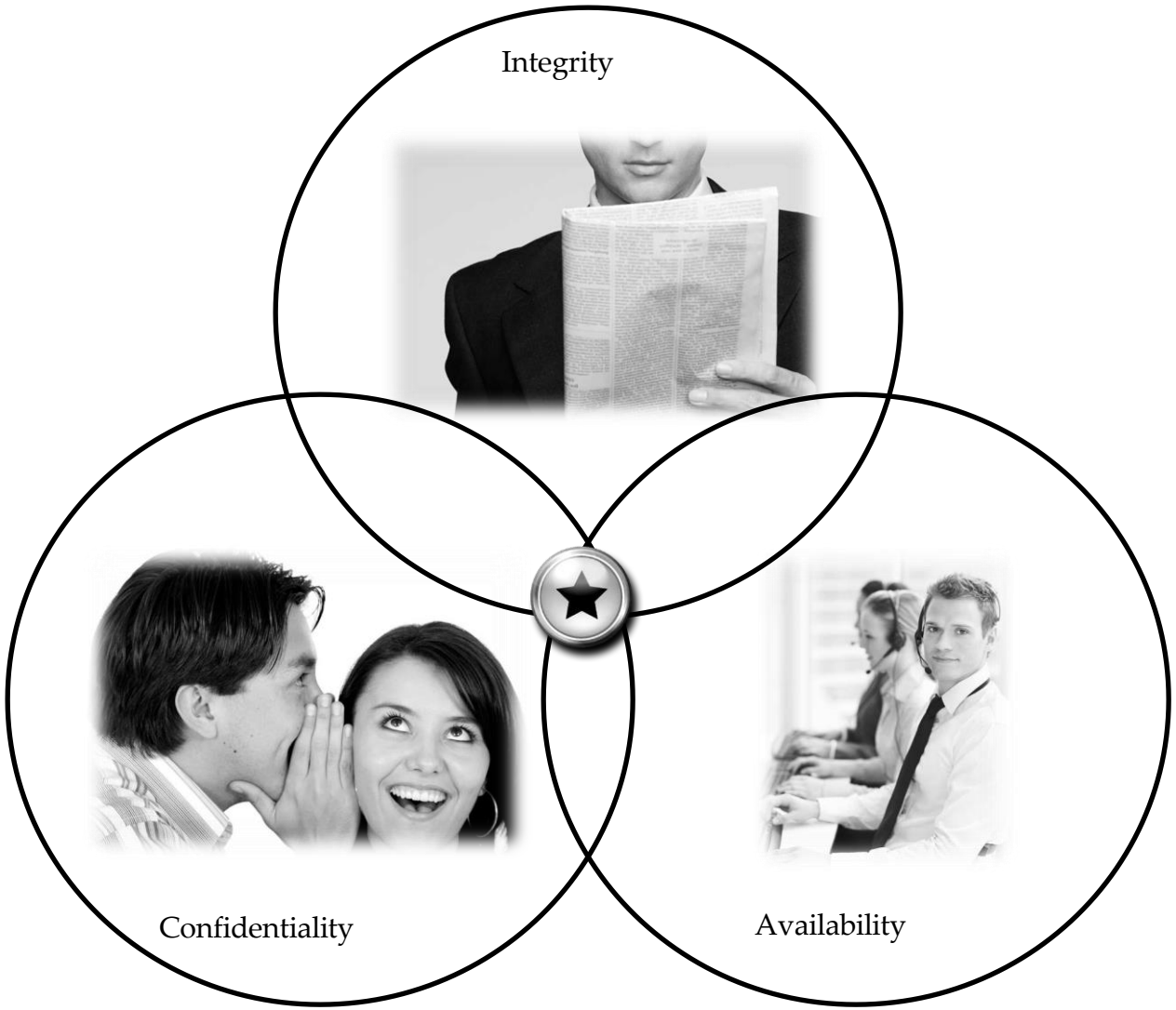
How to define the security of systems?

Defining Security

- The security of a system, application, or protocol is always relative to
  - An adversary with specific capabilities
  - A set of desired properties
- For example, standard file access permissions in Linux and Windows are not effective against an adversary who can boot from a CD

Security Goals

- C.I.A.



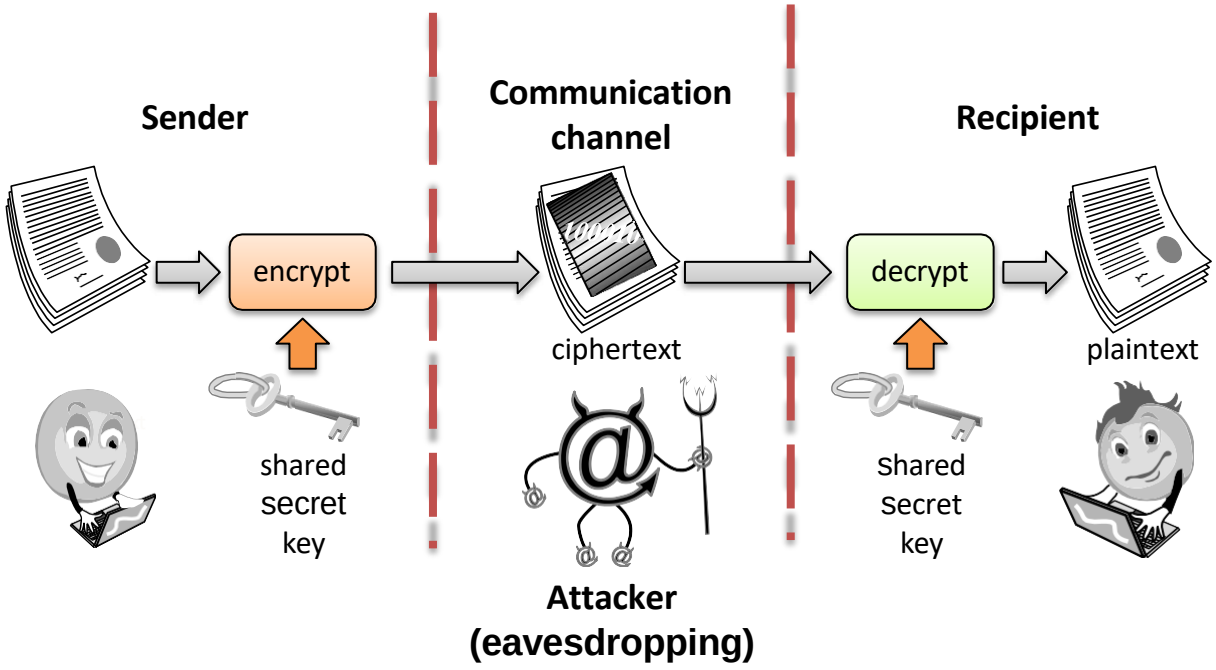
Confidentiality

- **Confidentiality** is the avoidance of the unauthorized disclosure of information.
  - confidentiality involves the protection of data, providing access for those who are allowed to see it while disallowing others from learning anything about its content.

How to ensure data confidentiality?

Tools for Confidentiality

- **Encryption:** the transformation of information using a secret, called an encryption key, so that the transformed information can only be read using another secret, called the decryption key (which may, in some cases, be the same as the encryption key).



Tools for Confidentiality

- **Access control:** rules and policies that limit access to confidential information to those people and/or systems with a "need to know."
  - This need to know may be determined by identity, such as a person's name or a computer's serial number, or by a role that a person has, such as being a manager or a computer security specialist.

Tools for Confidentiality

- **Authorization:** the determination if a person or system is allowed access to resources, based on an access control policy.
  - Such authorizations should prevent an attacker from tricking the system into letting him have access to protected resources.
- **Physical security:** the establishment of physical barriers to limit access to protected computational resources.
  - Such barriers include locks on doors, the placement of computers in windowless rooms, the use of sound dampening materials, and even the construction of buildings or rooms with walls incorporating copper meshes (called **Faraday cages)** so that electromagnetic signals cannot enter or exit the enclosure.

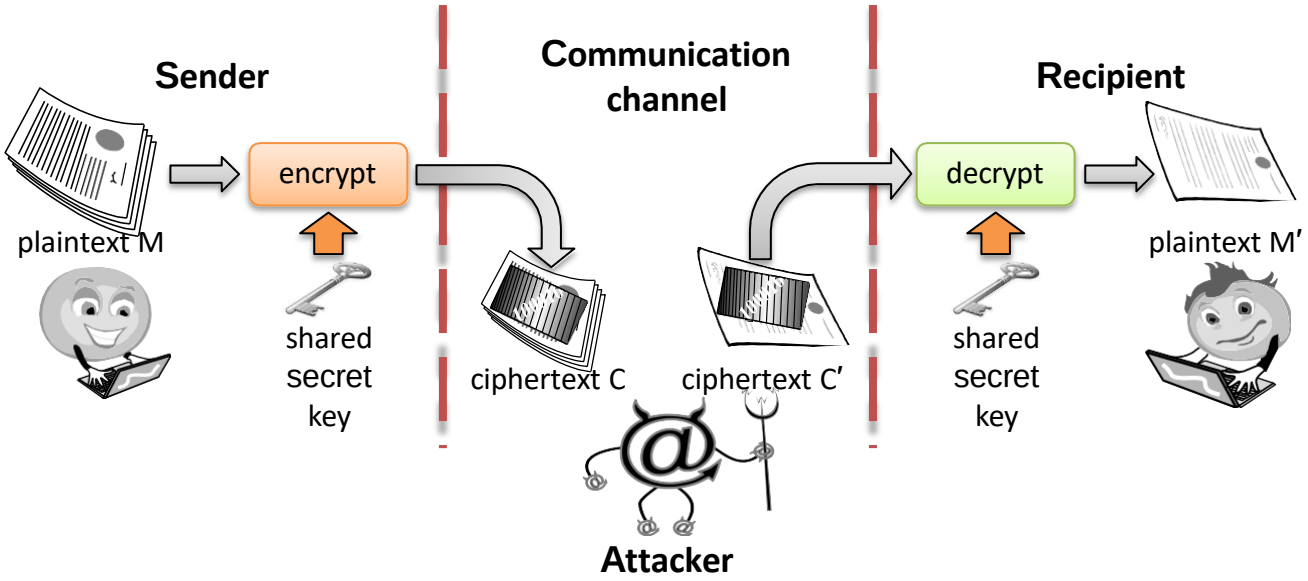
Integrity

- **Integrity:** the property that information has not been **altered** in an unauthorized way.

Can Encryption ensure data integrity?

Integrity

- **Integrity:** the property that information has not been **altered** in an unauthorized way.
- **Encryption cannot address integrity issues**



Even the transmitted data is encrypted,
an **eavesdropping** attacker can still modify data

Tools for Integrity

- **Backups:** the periodic archiving of data.
- **Checksums:** the computation of a function that maps the contents of a file to a numerical value. A checksum function depends on the entire contents of a file and is designed in a way that even a small change to the input file (such as flipping a single bit) is highly likely to result in a different output value.
- **Data correcting codes:** methods for storing data in such a way that small changes can be easily detected and automatically corrected. E.g., hamming code

Availability

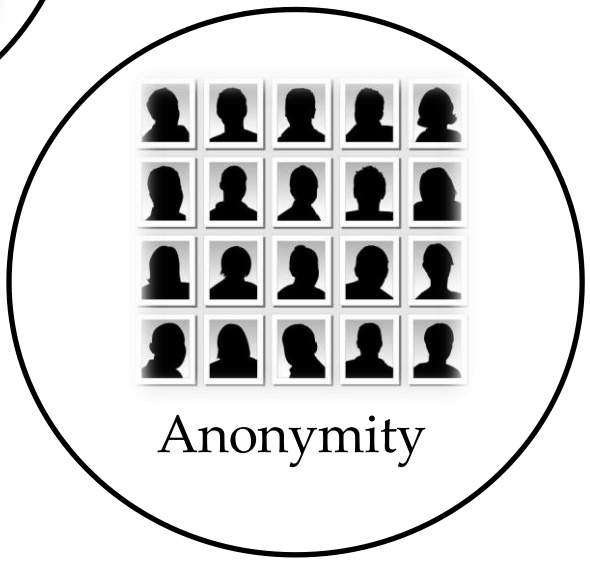
- **Availability:** the property that information is accessible and modifiable in a timely fashion by those authorized to do so.
- **Tools:**
  - **Physical protections:** infrastructure meant to keep information available even in the event of physical challenges (e.g., use laptop bag to protect laptop).
  - **Computational redundancies:** computers and storage devices that serve as fallbacks in the case of failures.

Other Security Concepts

- A.A.A.

Authenticity

Assurance



Anonymity

Anonymity

- **Anonymity:** the property that certain records or transactions not to be attributable to any individual.
- **Tools:**
  - **Aggregation:** the combining of data from many individuals so that disclosed sums or averages cannot be tied to any individual.
  - **Mixing:** the intertwining of transactions, information, or communications in a way that cannot be traced to any individual.
  - **Proxies:** trusted agents that are willing to engage in actions for an individual in a way that cannot be traced back to that person.
  - **Pseudonyms:** fictional identities that can fill in for real identities in communications and transactions, but are otherwise known only to a trusted entity.

Authenticity

- **Authenticity** is the ability to determine that statements, policies, and permissions issued by individuals or systems are real.
- **Primary tool:**
  - **digital signatures.** These are cryptographic computations that allow a person or system to commit to the authenticity of their documents in a unique way that achieves **nonrepudiation**, which is the property that authentic statements issued by a person or system cannot be denied.

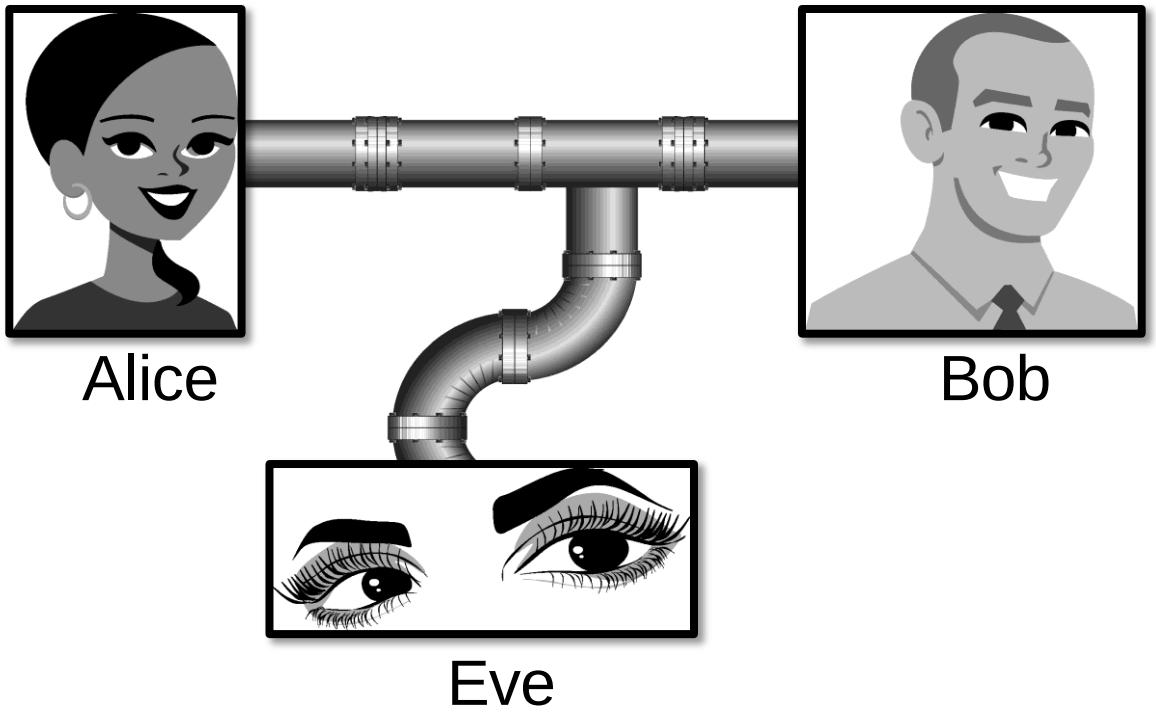
Assurance

- **Assurance** refers to how trust is provided and managed in computer systems.
- **Trust management** depends on:
  - **Policies,** which specify behavioral expectations that people or systems have for themselves and others.
    - For example, the designers of an online music system may specify policies that describe how users can access and copy songs.
  - **Permissions,** which describe the behaviors that are allowed by the agents that interact with a person or system.
    - For instance, an online music store may provide permissions for limited access and copying to people who have purchased certain songs.
  - **Protections,** which describe mechanisms put in place to enforce permissions and polices.
    - We could imagine that an online music store would build in protections to prevent people from unauthorized access and copying of its songs.

What kind of threats and attacks may we encounter?

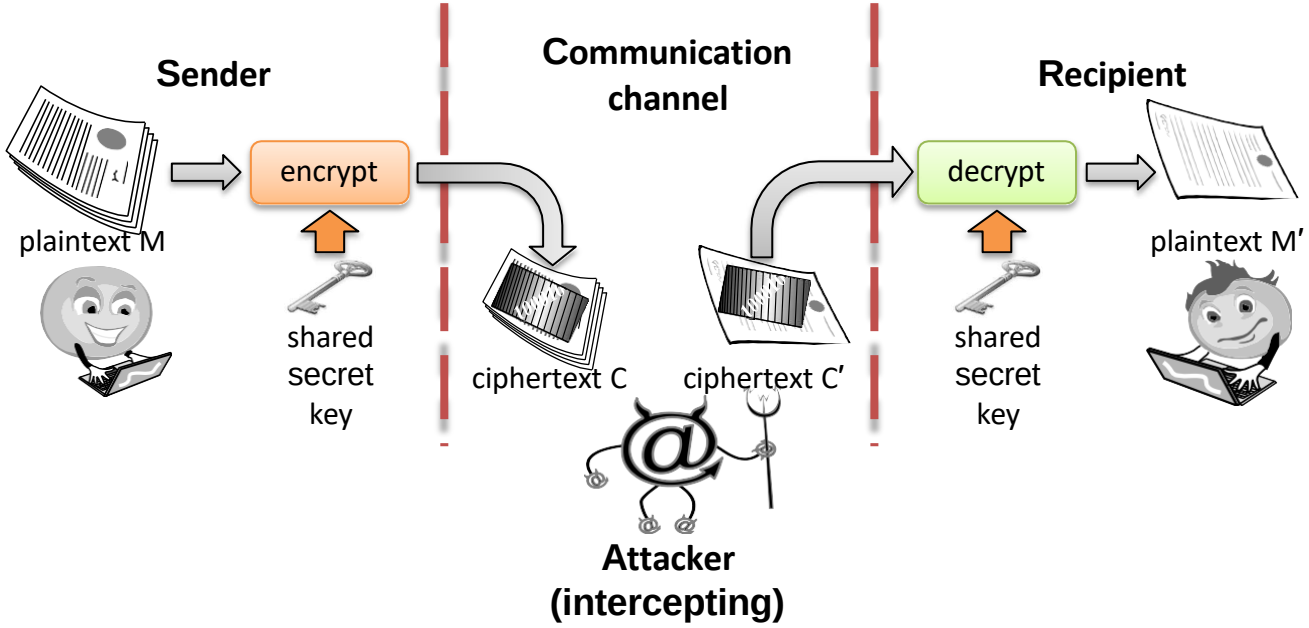
Threats and Attacks

- **Eavesdropping:** the interception of information intended for someone else during its transmission over a communication channel.



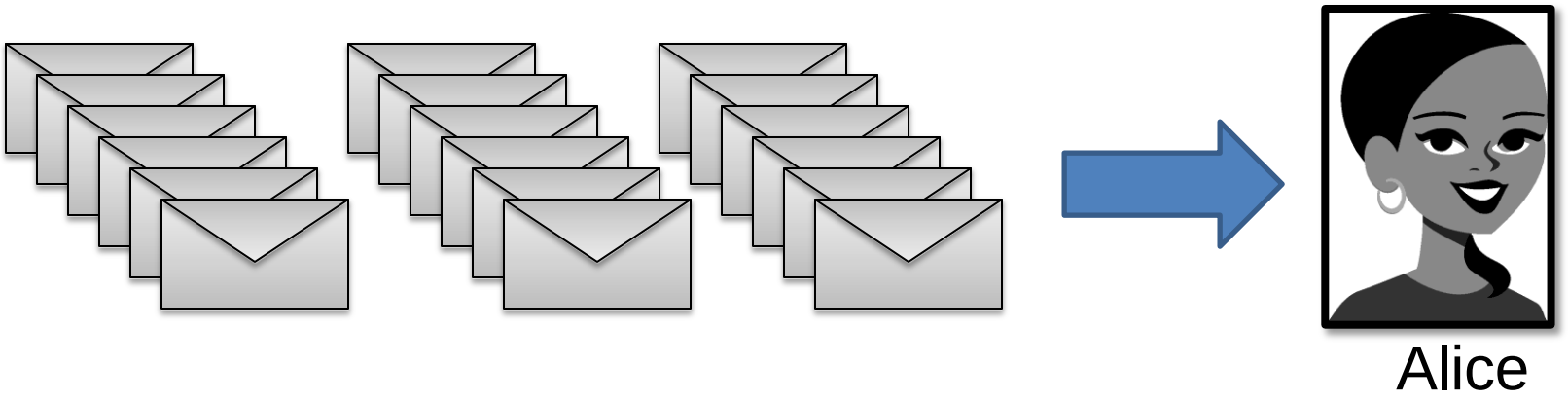
Threats and Attacks

- **Alteration:** unauthorized modification of information.
  - **Example:** the **man-in-the-middle attack,** where a network stream is intercepted, modified, and retransmitted.



Threats and Attacks

- **Denial-of-service:** the interruption or degradation of a data service or information access.
  - **Example:** email **spam,** to the degree that it is meant to simply fill up a mail queue and slow down an email server.



Threats and Attacks

- **Masquerading:** the fabrication of information that is purported to be from someone who is not actually the author.

"From: Alice"
(really is from Eve)

Threats and Attacks

- **Repudiation :** the denial of a commitment or data receipt.
  - Happens when an application or system does not adopt controls to properly track and log users' actions, thus permitting malicious manipulation or forging the identification of new actions.

Threats and Attacks

- **Correlation** and **traceback:** the integration of multiple data sources and information flows to determine the source of a particular data stream or piece of information.

Introduction of Cryptography

Cryptography

- Is
  - A useful tool
  - The basis for many security mechanisms
- Is not
  - The solution to all security problems
  - Reliable unless implemented properly
  - Reliable unless used properly
  - Something you should try to invent yourself unless
    - you spend a lot of time becoming an expert
    - you subject your design to outside review

Auguste Kerckhoffs

- A cryptosystem should be secure even if everything about the system, except the **key**, is **public knowledge**.

baptised as **Jean-Guillaume-Hubert-Victor-François-Alexandre-Auguste Kerckhoffs von Nieuwenhof**

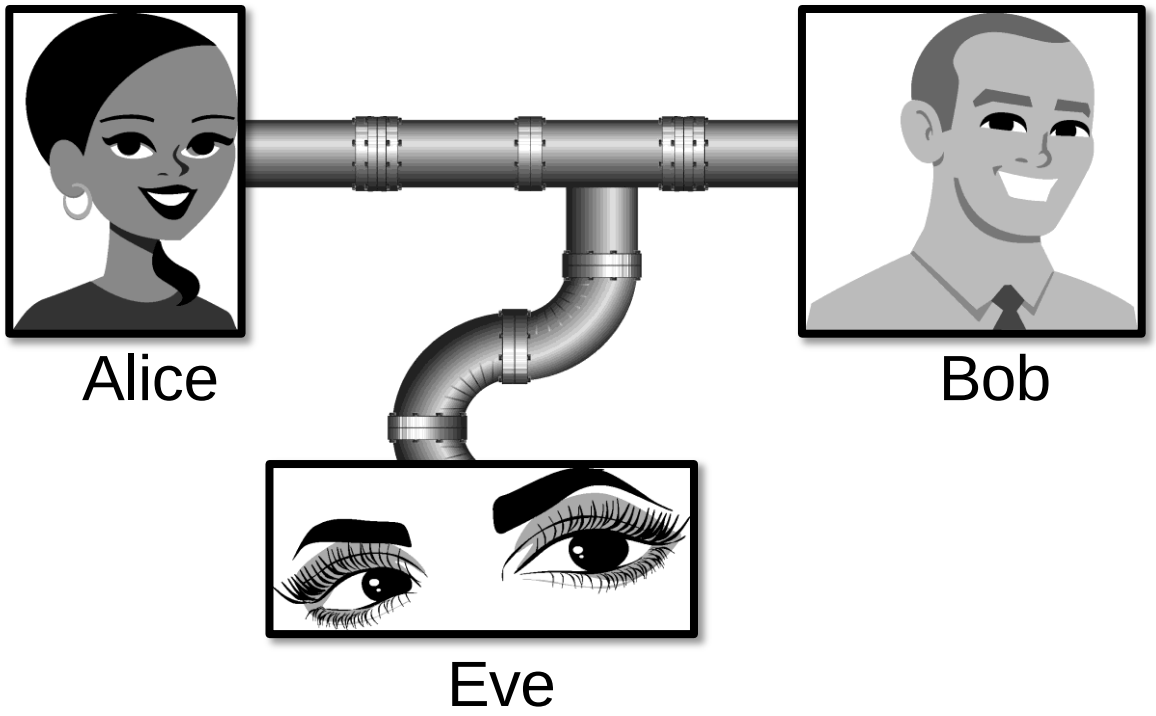
Crypto threat model

- Assume attacker knows the cryptosystem
- Attacker does not know random numbers
  - Generated as systems run, not in advance

- Easy lessons
  - Use good random number generators
  - No harm in public review of cryptography
    - This prevents silly and not-so-silly mistakes
    - Benefit from community of experts

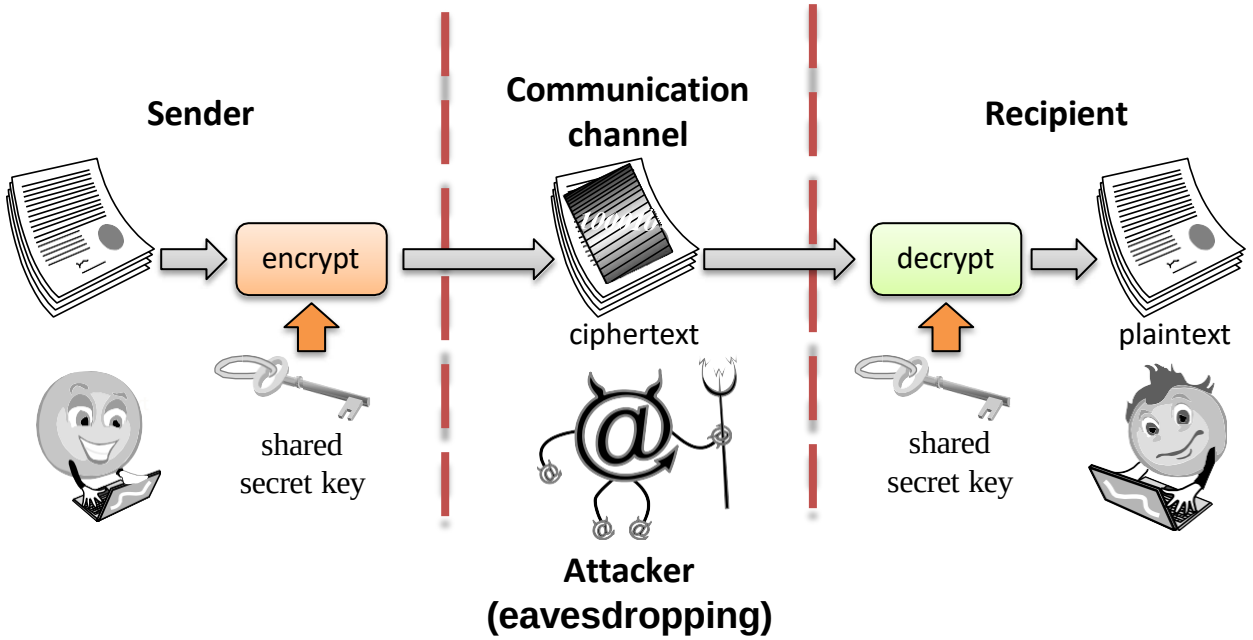
Cryptographic Concepts

- **Encryption**: a means to allow two parties, customarily called Alice and Bob, to establish confidential communication over an insecure channel that is subject to eavesdropping.



Encryption and Decryption

- The message P is called the **plaintext.**
- Alice will convert plaintext P to an encrypted form using an encryption algorithm E that outputs a **ciphertext C for P.**



Encryption and Decryption

- As equations:

$$C = E(P)$$

$$P = D(C)$$

- The encryption and decryption algorithms are chosen so that it is infeasible for someone other than Alice and Bob to determine plaintext M from ciphertext C. Thus, ciphertext C can be transmitted over an insecure channel that can be eavesdropped by an adversary.

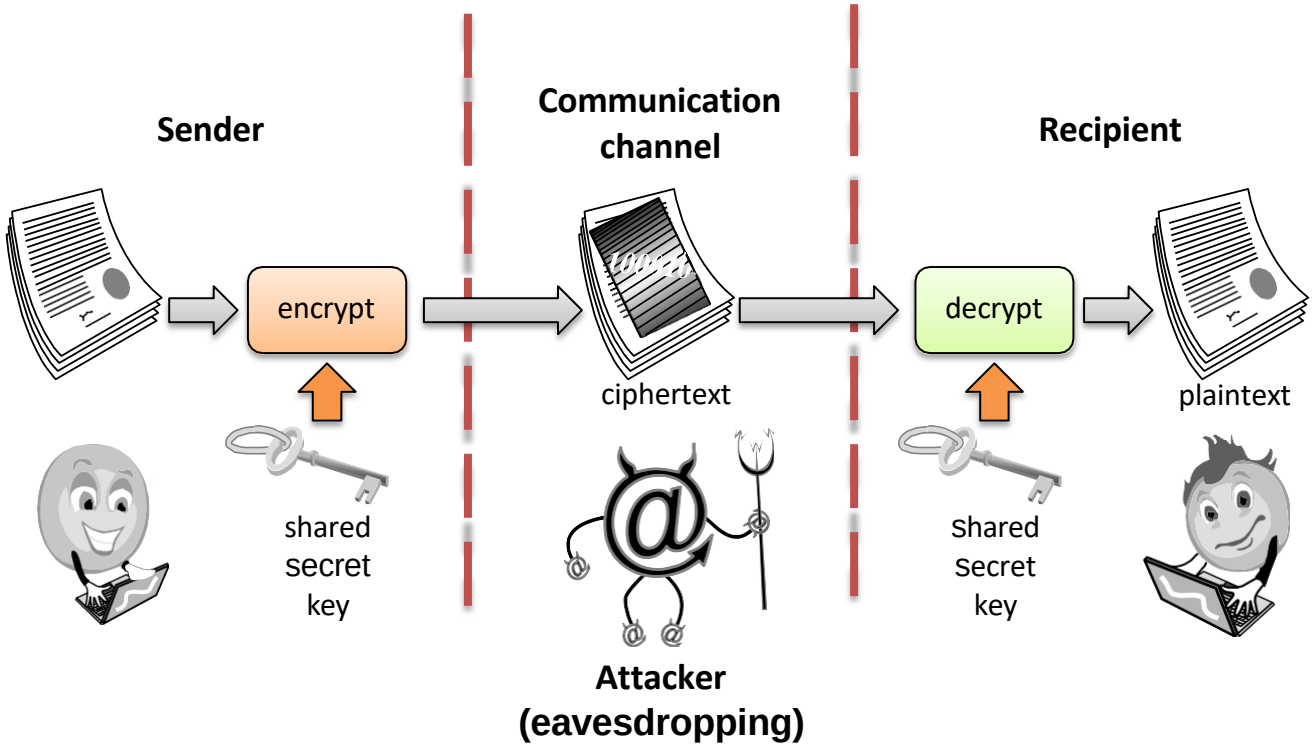
Cryptosystem

1. The set of possible plaintexts
2. The set of possible ciphertexts
3. The set of encryption keys
4. The set of decryption keys
5. The correspondence between encryption keys and decryption keys
6. The encryption algorithm to use
7. The decryption algorithm to use

Symmetric encryption & asymmetric encryption

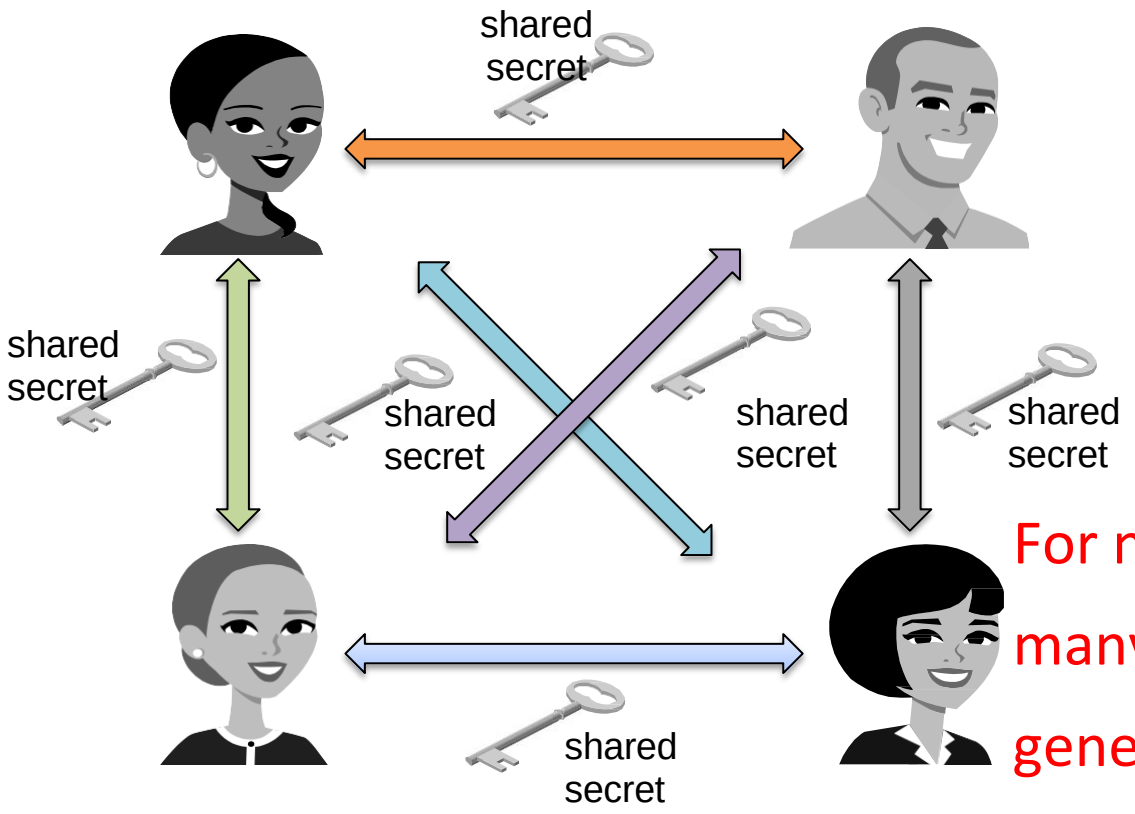
Symmetric Cryptosystems

- Alice and Bob share a secret key, which is used for both encryption and decryption.



Symmetric Key Distribution

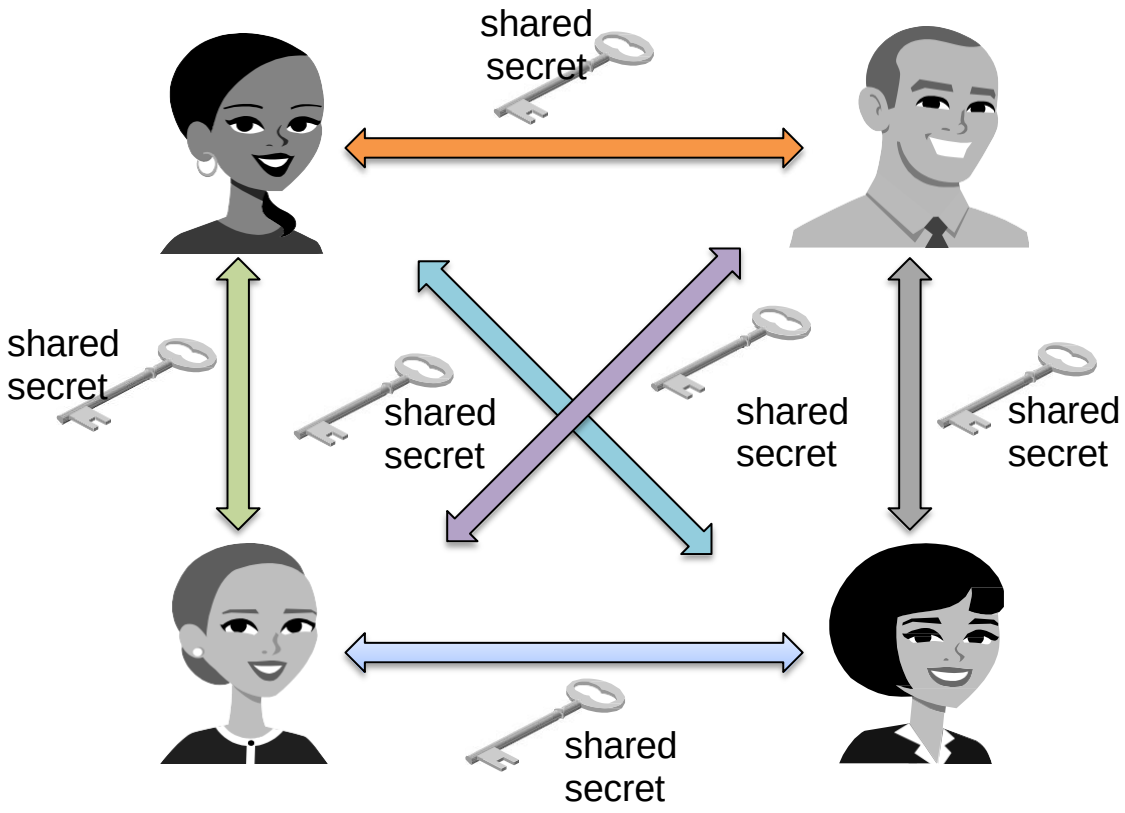
- Requires each pair of communicating parties to share a (separate) secret key.



For n users, how many keys are generated in total?

Symmetric Key Distribution

- Requires each pair of communicating parties to share a (separate) secret key.



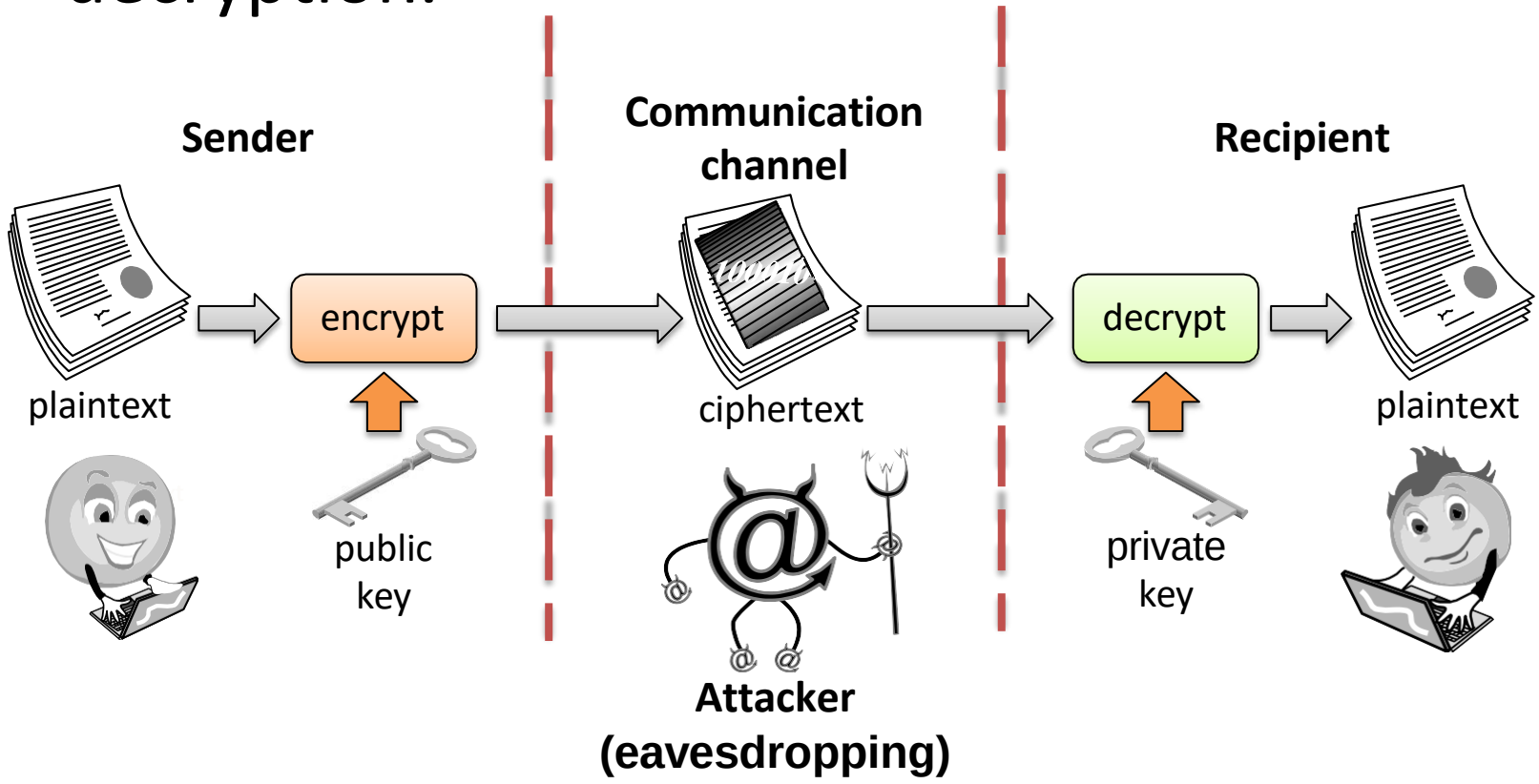
n (n–1)/2 keys

Public-Key Cryptography

- Bob has two keys: a **private key,** SB, which Bob keeps secret, and a **public key,** PB, which Bob broadcasts widely.
  - In order for Alice to send an encrypted message to Bob, she need only obtain his public key, PB, use that to encrypt her message, P, and send the result, $C = E_{PB}(P)$, to Bob. Bob then uses his secret key to decrypt the message as $P = D_{SB}(C)$.

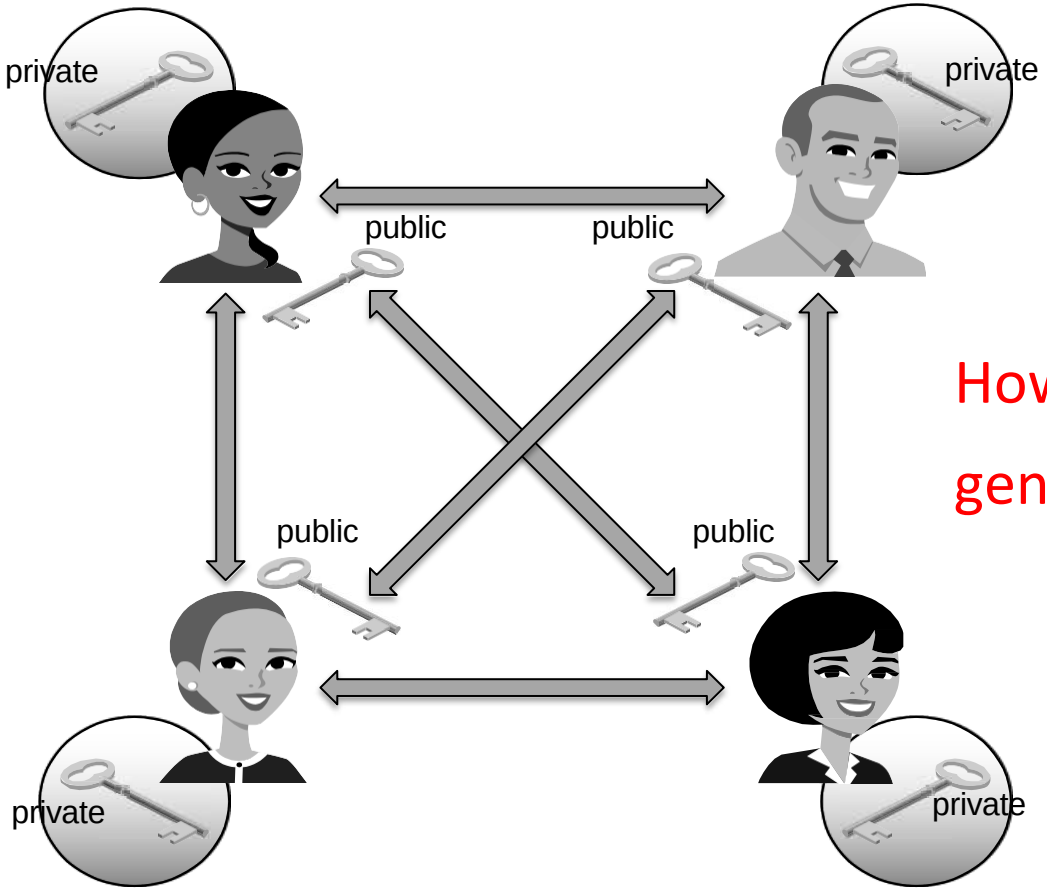
Public-Key Cryptography

- Separate keys are used for encryption and decryption.



Public Key Distribution

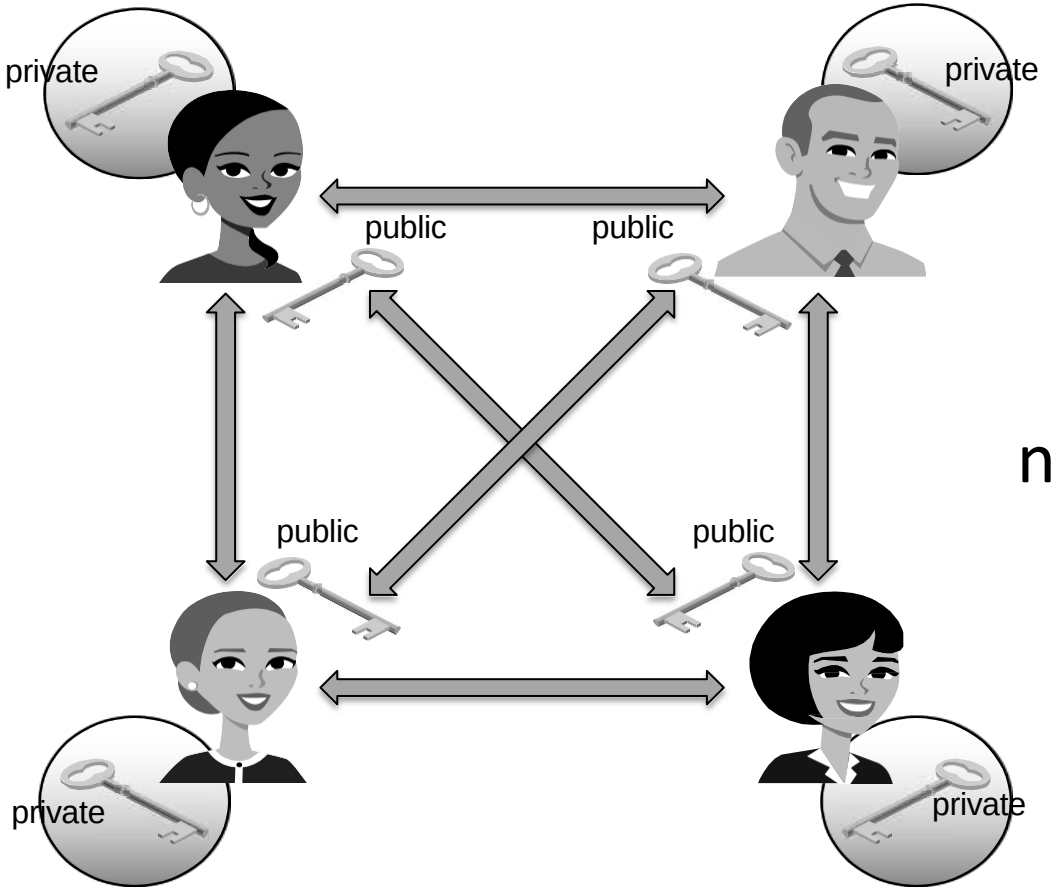
- Only one key is needed for each recipient



How many keys are generated in total?

Public Key Distribution

- Only one key is needed for each recipient



n key pairs

Sharing keys for symmetric encryption using public key encryption

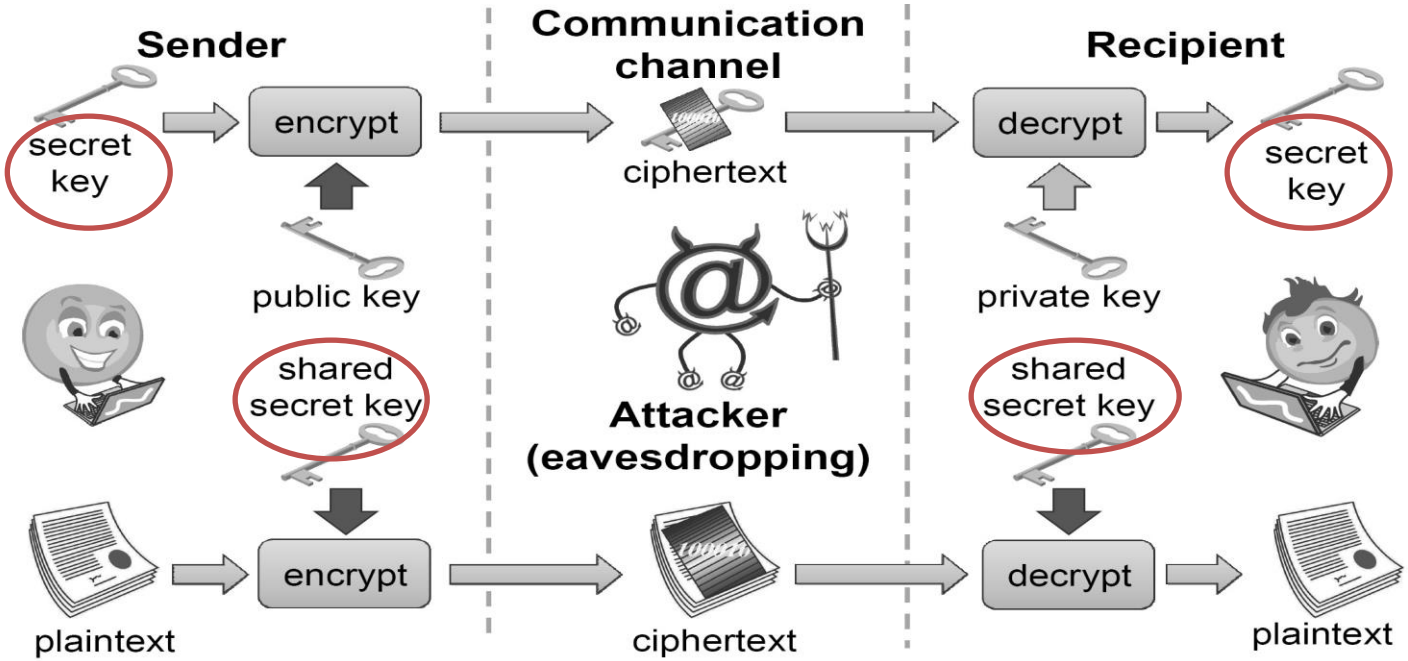


Figure 1.12: Use of a public-key cryptosystem to exchange a shared secret key, which is subsequently employed for communicating with a symmetric encryption scheme. The secret key is the "plaintext" message sent from the sender to the recipient.

Digital Signatures

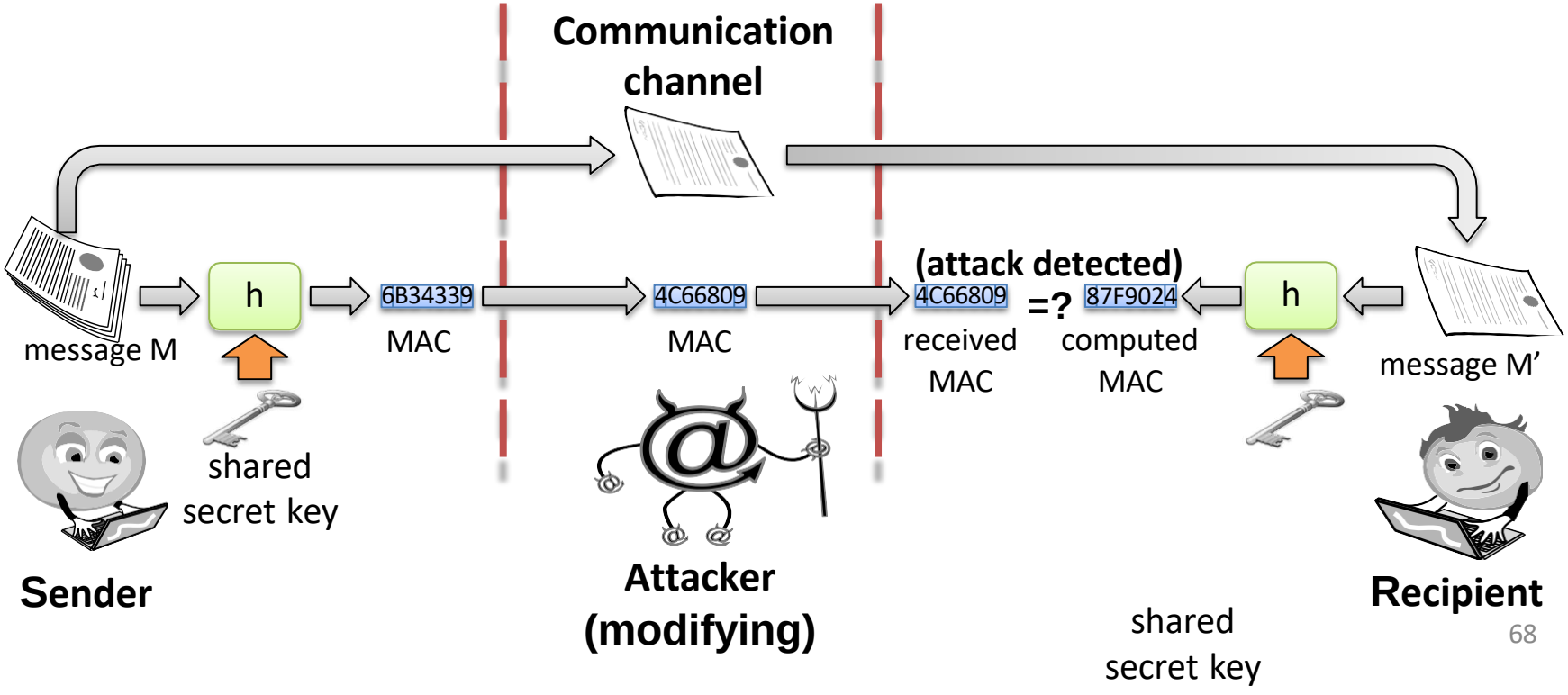
- Public-key encryption provides a method for doing digital signatures
- To sign a message, M, Alice just encrypts it with her private key, SA, creating $C = E_{SA}(M)$.
- Anyone can decrypt this message using Alice's public key, as $M' = D_{PA}(C)$, and compare that to the message M.

Cryptographic Hash Functions

- A checksum on a message, M, that is:
- **One-way**: it should be easy to compute Y=H(M), but hard to find M given only Y
- **Collision-resistant:** it should be hard to find two messages, M and N, such that H(M)=H(N).
- **Examples:** SHA-1, SHA-256.

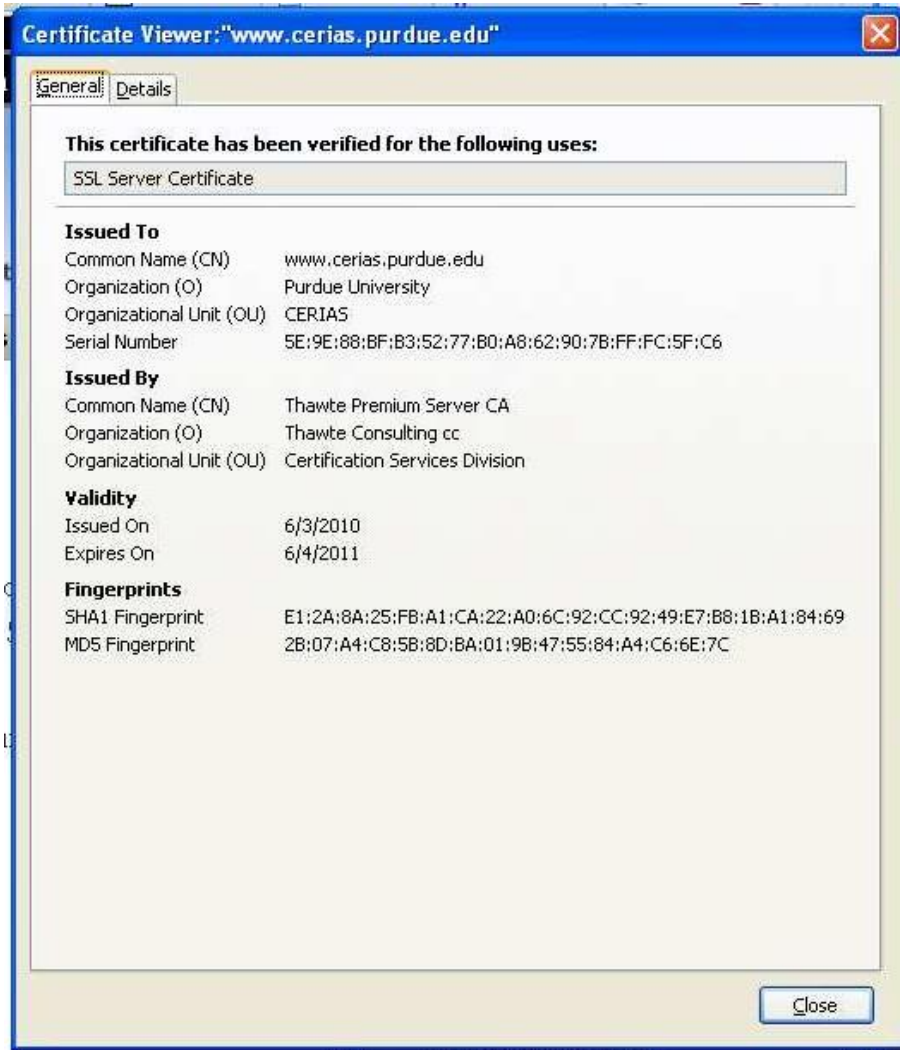
Message Authentication Codes

- Allows for Alice and Bob to have data **integrity**, if they share a secret key.
- Given a message M, Alice computes H(K||M) and sends M and this hash to Bob.



Digital Certificates

- **certificate authority (CA)** digitally signs a binding between an identity and the public key for that identity.

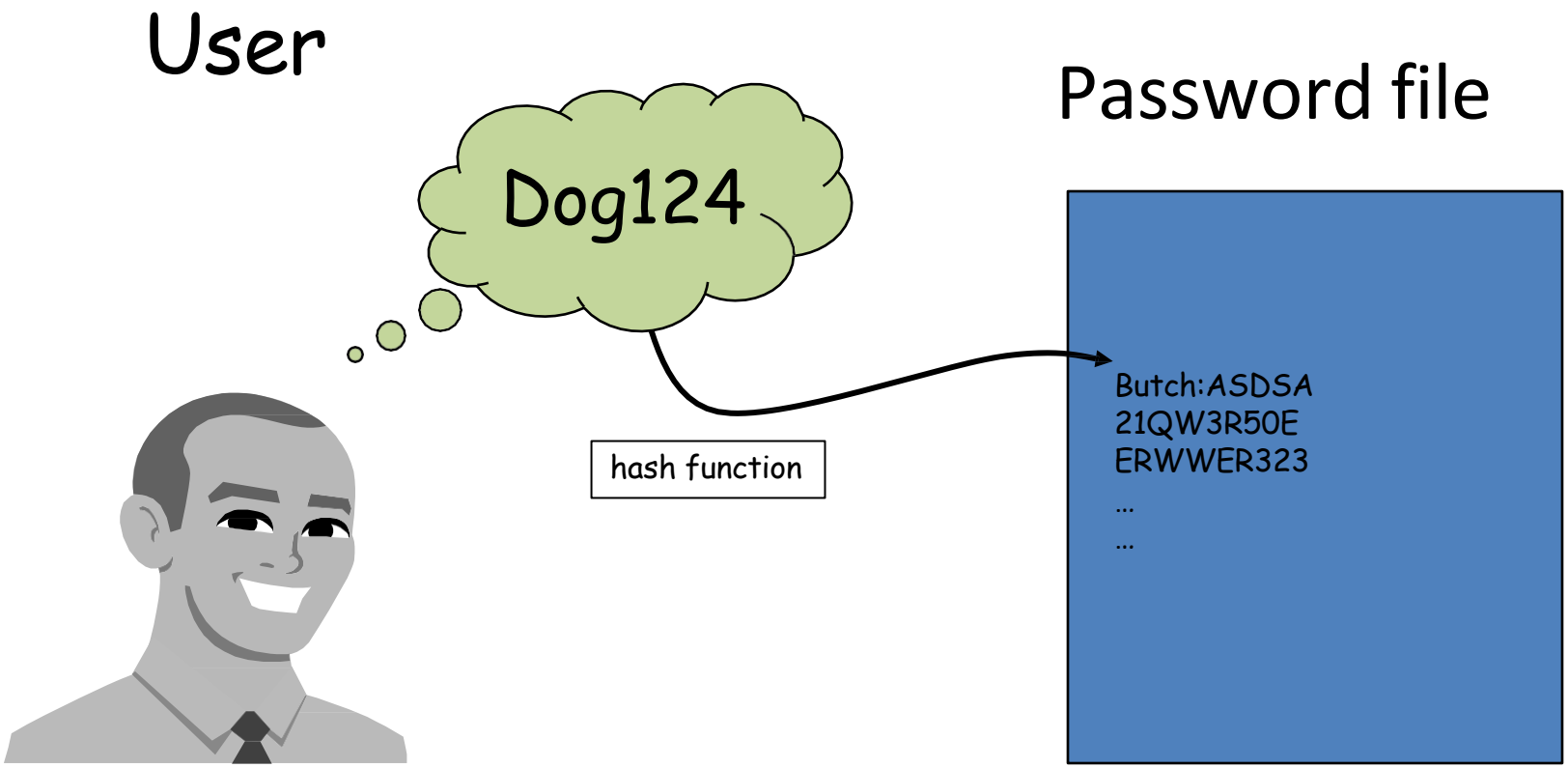


Passwords

- A short sequence of characters used as a means to authenticate someone via a secret that they know.

- Userid: _____
- Password: _____

How a password is stored?



Strong Passwords

What is a strong password

- UPPER/lower case characters
- Special characters
- Numbers

When is a password strong?

- Seattle1
- M1ke03
- P@$$w0rd
- TD2k5secV

Password Complexity

A fixed 6 symbols password:

Numbers

$10^6 = 1{,}000{,}000$

UPPER or lower case characters

$26^6 = 308{,}915{,}776$

UPPER and lower case characters

$52^6 = 19{,}770{,}609{,}664$

32 special characters (&, %, $, £, “, |, ^, §, etc.)

$32^6 = 1{,}073{,}741{,}824$

ASCII standard (character encoding standard for electronic communication)

7 bit 2^7 =128 symbols

$128^6 = 4{,}398{,}046{,}511{,}104$

Password Length

26 UPPER/lower case characters = 52 characters
10 numbers
32 special characters
=> 94 characters available

5 characters: 94^5 =	7,339,040,224
6 characters: 94^6 =	689,869,781,056
7 characters: 94^7 =	64,847,759,419,264
8 characters: 94^8 =	6,095,689,385,410,816
9 characters: 94^9 =	572,994,802,228,616,704

Longer passwords are better

Password Validity: Brute Force Test

Password does not change for 60 days

how many passwords should I try for each second?

5 characters:	1,415 PW /sec
6 characters:	133,076 PW /sec
7 characters:	12,509,214 PW /sec
8 characters:	1,175,866,008 PW /sec
9 characters:	110,531,404,750 PW /sec

Secure Passwords

A strong password includes characters from at least three of the following groups:

Group	Example
Lowercase letters	a, b, c, ...
Uppercase letters	A, B, C, ...
Numerals	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Non-alphanumeric (symbols)	( ) ` ~ ! @ # $ % ^ & * - + = | \ { } [ ] : ; " ' < > , . ? /
Unicode characters	€, Γ, ƒ, and λ

Use pass phrases eg. "I re@lly want to buy 11 Dogs!"

Summary of Lecture 1

- Security Concepts:
  - Confidentiality; Integrity; Availability
  - Authenticity; Assurance; Anonymity
- Overview on the crypto tools
  - Symmetric/public crypto., cryptographic hash, digital signature, digital certificate.
- Secure Password
  - Common means for authentication
  - Usually stored via hash values

Discussion

Think about the way you use your computer in your personal life. Which is most important: confidentiality, integrity, availability? Justify your answer.