CS3319 Foundations of Data Science

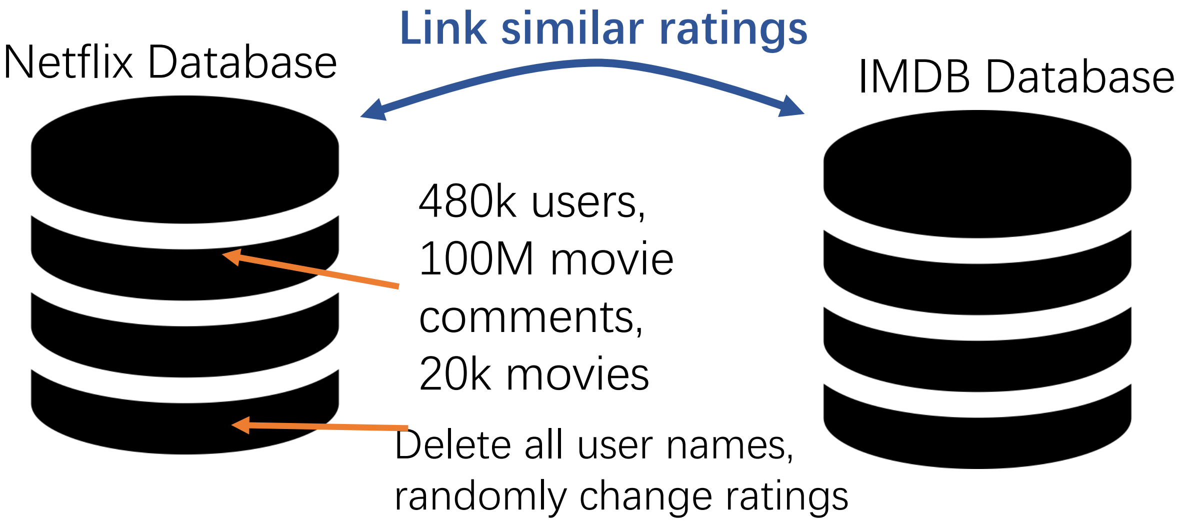
# 9. Privacy

Jiaxin Ding
John Hopcroft Center

上海交通大学
约翰·霍普克罗夫特计算机科学中心
John Hopcroft Center for Computer Science

# Privacy Concerns from Users

- Netflix prize **De-anonymization** attack[1]





[1] Narayanan, Arvind, and Vitaly Shmatikov. “Robust de-anonymization of large sparse datasets.” *2008, IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008.

# Privacy

- **Privacy** is the ability of an individual or group to seclude themselves or information about themselves, and thereby express themselves selectively.

- When something is private to a person, it usually means that something is inherently special or sensitive to them.

# What do we mean by privacy?

- **Anonymization**
  - Hide the information that can be used to infer the **identity**

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | **Zip code** | **Age** | **Nationality** | **Condition** |
| 1 | 130** | $<30$ | * | AIDS |
| 2 | 130** | $<30$ | * | Heart Disease |
| 3 | 130** | $<30$ | * | Viral Infection |
| 4 | 130** | $<30$ | * | Viral Infection |
| 5 | 130** | $\geq 40$ | * | Cancer |
| 6 | 130** | $\geq 40$ | * | Heart Disease |
| 7 | 130** | $\geq 40$ | * | Viral Infection |
| 8 | 130** | $\geq 40$ | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

# Example Attack 1

- **K-anonymity**: at least k records share the same **quasi-identifier** (e.g. zip code, age, nationality)

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | **Zip code** | **Age** | **Nationality** | **Condition** |
| 1 | 130** | $<30$ | * | AIDS |
| 2 | 130** | $<30$ | * | Heart Disease |
| 3 | 130** | $<30$ | * | Viral Infection |
| 4 | 130** | $<30$ | * | Viral Infection |
| 5 | 130** | $\geq 40$ | * | Cancer |
| 6 | 130** | $\geq 40$ | * | Heart Disease |
| 7 | 130** | $\geq 40$ | * | Viral Infection |
| 8 | 130** | $\geq 40$ | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

# Example Attack 1

- We have records from 2 hospitals
- If we know someone visited both hospital, what can we know?
- If we know her/his age is 28, what can we know?

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip code | Age | Nationality | Condition |
| 1 | 130** | <30 | * | AIDS |
| 2 | 130** | <30 | * | Heart Disease |
| 3 | 130** | <30 | * | Viral Infection |
| 4 | 130** | <30 | * | Viral Infection |
| 5 | 130** | ≥40 | * | Cancer |
| 6 | 130** | ≥40 | * | Heart Disease |
| 7 | 130** | ≥40 | * | Viral Infection |
| 8 | 130** | ≥40 | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

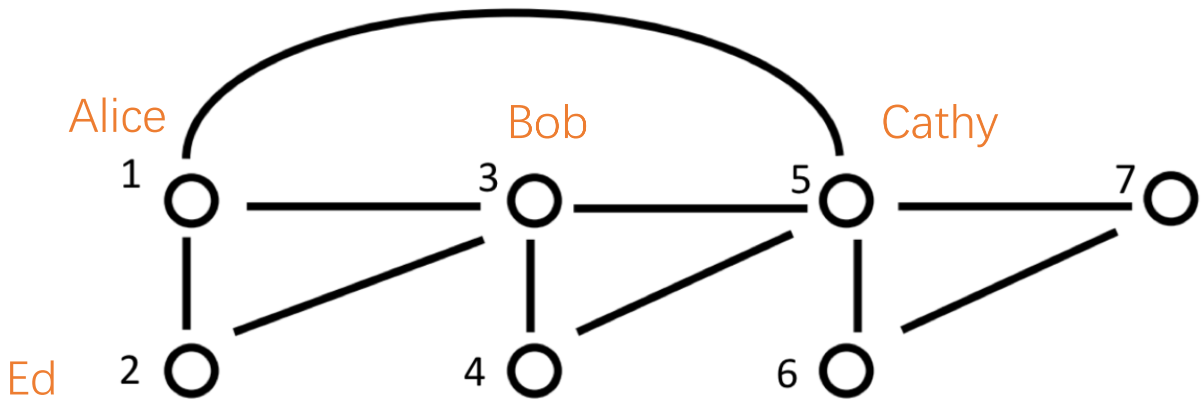
Hospital A (4-anonymous)

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip code | Age | Nationality | Condition |
| 1 | 130** | <35 | * | AIDS |
| 2 | 130** | <35 | * | Tuberculosis |
| 3 | 130** | <35 | * | Flu |
| 4 | 130** | <35 | * | Tuberculosis |
| 5 | 130** | <35 | * | Cancer |
| 6 | 130** | <35 | * | Cancer |
| 7 | 130** | ≥35 | * | Cancer |
| 8 | 130** | ≥35 | * | Cancer |
| 9 | 130** | ≥35 | * | Cancer |
| 10 | 130** | ≥35 | * | Tuberculosis |
| 11 | 130** | ≥35 | * | Viral Infection |
| 12 | 130** | ≥35 | * | Viral Infection |

Hospital B (6-anonymous)

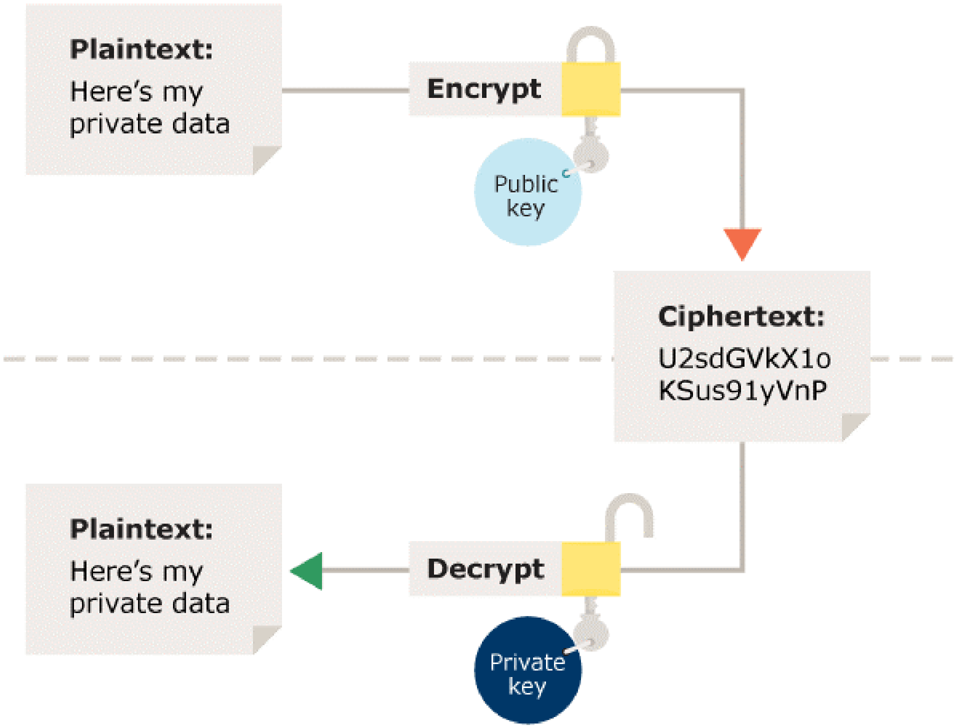
# Example Attack 2

- Anonymous communication graph
- **Auxiliary knowledge**
  - Alice has communicated to Bob, Cathy, and Ed
  - Cathy has communicated to everyone, except Ed



# What do we mean by privacy?

- Encryption
  - Alice sends a message to Bob such that any other does not learn the message without the **key**. Bob gets the correct message.



# What do we mean by privacy?

- Encryption
  - E.g. **RSA algorithm**.

$$Theorem\colon\ (m^e)^d = m\ (mod\ n)$$

**RSA**

**RSA: The first and most popular asymmetric encryption**

$$E(m) = m^e\ (\mathrm{mod}\, n)$$

$$\mathrm{D}(c) = c^d\ (\mathrm{mod}\, n)$$

**Example:**

Choose two primes: $p = 11, q = 13,$

$n = p \times q = 143.$

Choose public key $e = 7.$

Extended Euclidean algorithm find

private key $d = 103, \mathrm{s.t.\,ed} = 1(\mathrm{mod(p-1)(q-1)}).$

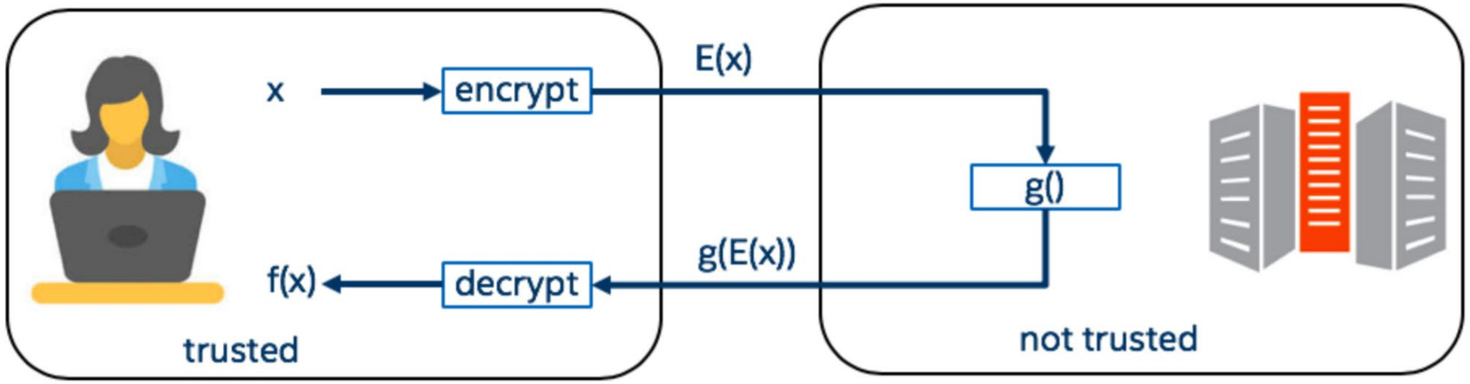
We want to encrypt $m = 9$,

$$E(m) = 9^7\ (\mathrm{mod}\ 143) = 48 = c$$

$$D(c) = 48^{103}(\mathrm{mod}\ 143) = 9 = m$$

# What do we mean by privacy?

- **Computing with a not-trusted third party**
  - Alice stores encrypted data on a server controlled by Bob. Server returns correct query answers to Alice, without Bob learning anything about the data.



# What do we mean by privacy?

- **Computing with a not-trusted third party**
  - Homomorphic encryption (e.g. RSA)

$$E(m) = m^e \pmod n$$

$$\mathrm{D}(c) = c^d \pmod n$$

$E(m_1) = m_1{}^e \qquad E(m_2) = m_2{}^e$

Ergo … $E(m_1) \times E(m_2)$

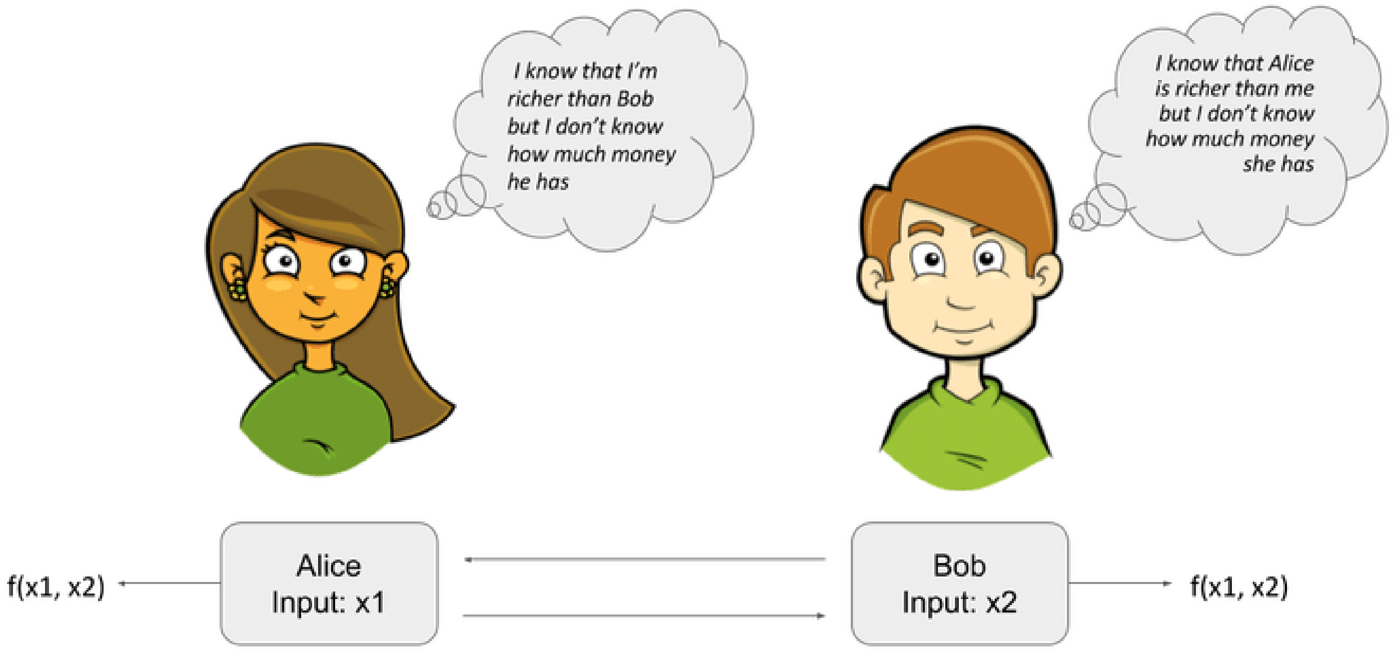
$$= m_1{}^e \times m_2{}^e$$
$$= (m_1 \times m_2)^e$$
$$= E(m_1 \times m_2)$$

**Multiplicative Homomorphism**

$$\boxed{E(m_1) \times E(m_2) = E(m_1 \times m_2)}$$

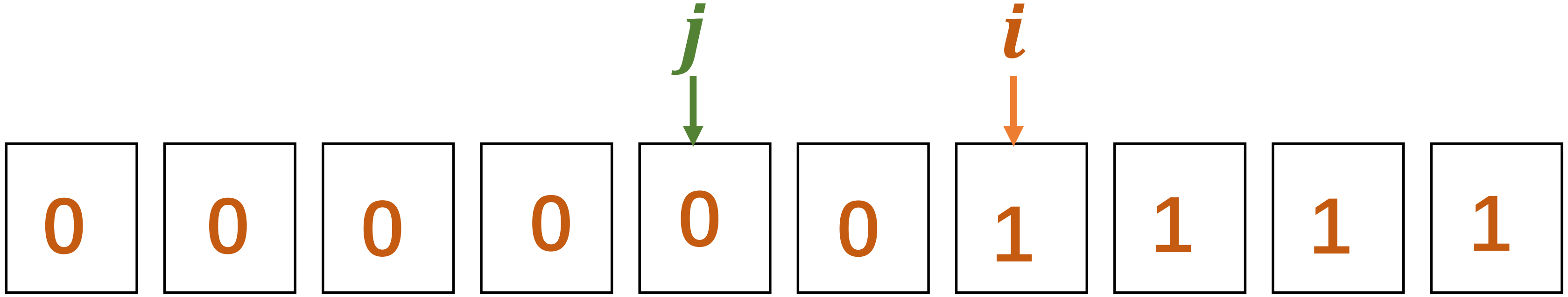
# What do we mean by privacy?

- Secure Multiparty Computation
  - **The millionaire's problem**: Alice and Bob want to know **who** of them has **more money** without letting the other know the exact amount of money one owns.



# What do we mean by privacy?

- **The millionaire's problem** protocol
  - Assume Alice has $\boldsymbol{i}$ million, and Bob has $\boldsymbol{j}$ million, $i, j$ are integers in [1, 10]
  - **Intuition:**
    - We have 10 boxes, Alice has the keys, while Bob does not have the keys
    - Alice opens all boxes. For box k, if **k<i**, Alice puts **0** in it; else Alice puts **1**. Afterwards, Alice closes all boxes
    - Bob picks up the ***j*th** box, and destroys all the other boxes
    - Alice opens the box and knows who is richer



# What do we mean by privacy?

- **Access Control**
  - A set of agents want to access a set of resources (could be files or records in a database).
  - Access control rules specify who is allowed to access certain resources.