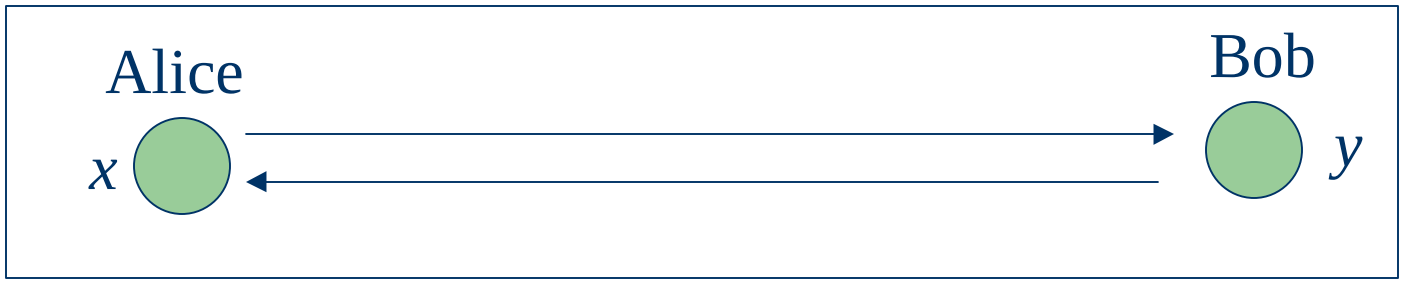
Secure Multi-party Protocols for Approximate Pattern Matching

Prof. Mikhail Atallah

Wenliang Du (Kevin)

CERIAS

General Framework of Secure Multiparty Computation



- Alice has private data *x*,
- Bob has private data *y*,
- They want to jointly compute *f(x,y)*,
- Only Alice (or Bob, or both) knows the result.

Objectives: Privacy in E-commerce

- Extend the general framework to fit into various e-commerce models.
- Design protocols to achieve privacy in e-commerce.
- Focus on *approximate* pattern matching, which is very important to some e-commerce applications.

Motivation

- Current client/server model: server is trusted.
- How to maintain privacy if the server is not trusted, and the client's query contains private information?
- Examples:
  - DNA sequence matching.
  - Patent searching.

Why Approximate Pattern Matching?

- Real-life requirement:
  - In certain area, we are interested in knowing “is *q* similar to *s?*”
  - Exact pattern matching is unrealistic in many situations
- Examples:
  - Matching fingerprint, voice, signature
  - Matching DNA sequence
  - Image template matching

Approximate Pattern Matching

- How to conduct approximate pattern matching?
  - Sum-of-squares: $\Sigma (a[k] - b[k])^2$.
  - Sum-of-absolute-values: $\Sigma |a[k] - b[k]|$.
  - String Edit Distance: the cost of transforming one string to another through insertion, deletion, or substitution.

Why Not Use Encryption or Hash Function?

- They work in exact pattern matching.
- They won't work in approximate pattern matching:
  - If *q* is close to *d*, after the encryption or hash, they will not be close to each other any more.
  - To know if they are close to each other, one has to decrypt them, then compare.
  - Privacy is lost if original information *q* or *d* is disclosed.

Why Not Use Anonymous Communication?

- Anonymity hides the sender's (of the information) identity, not the privacy of the information.
  - **Patent, new discovery (represented by images or words).**
- Some information automatically discloses the owner's identity.
  - **Face image.**
- Disclosing the information makes it easier for others to find the identity of the owner.
  - **DNA sequence, fingerprint.**
- Anonymous communication assumes trusted third parties.

Related Work

- The problem is called secure multiparty computation (SMC) in general.
- General secure multiparty computation research is still in theoretical stages of investigation.
- Exact pattern matching has been studied in the SMC framework, but approximate pattern matching has not.

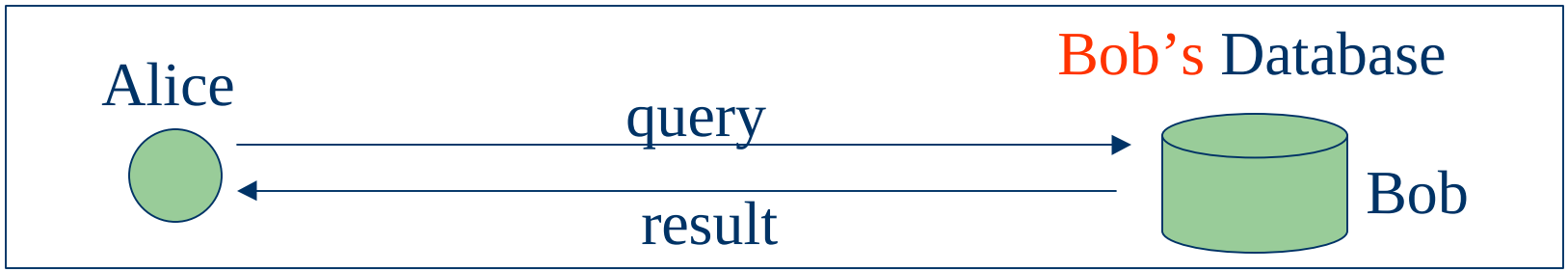
Models

From the current e-commerce models and their different requirements for privacy, the following models suggest themselves:

- PIR Model
- PIRPD Model
- SSO Model
- SSCO Model

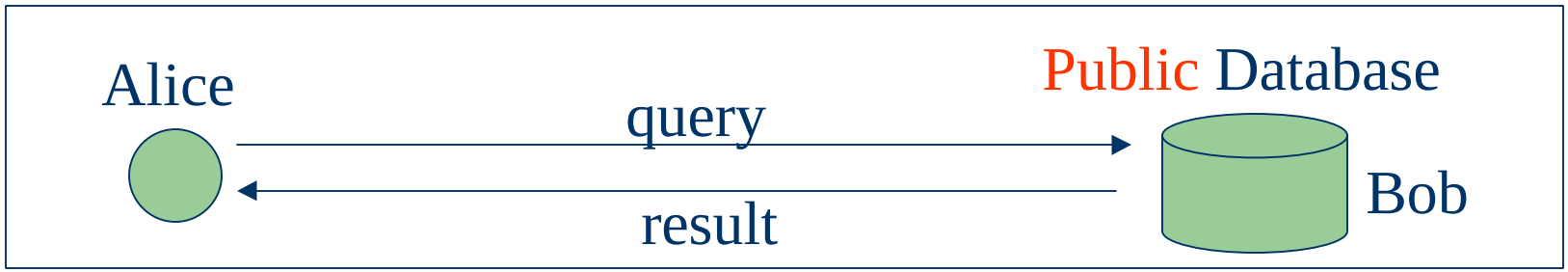
Private Information Retrieval (PIR Model)

- Alice's requirement.
  - I don't want Bob to know my query.
  - I don't want Bob to know my result.
- Bob's requirement.
  - I don't want to disclose any information to Alice other than the response to her query.



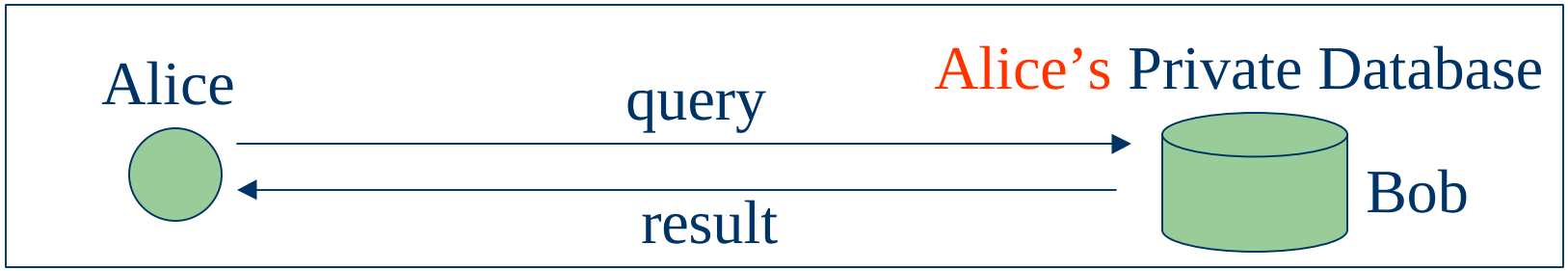
Private Information Retrieval From a Public Database (PIRPD Model)

- Alice's requirement.
  - I don't want Bob to know my query.
  - I don't want Bob to know my result.



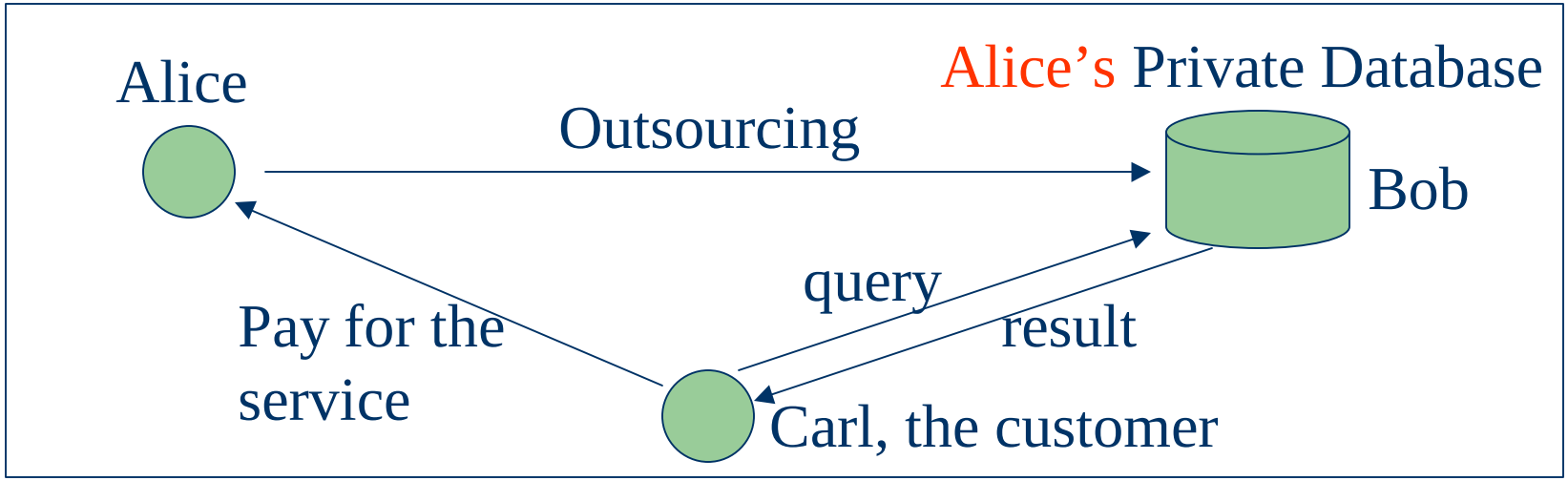
Secure Storage Outsourcing (SSO Model)

- Alice does not have enough storage, she has to store her database at Bob's place.
- Alice's requirement.
  - I don't want Bob to know my database.
  - I don't want Bob to know my query.



Secure Storage and Computation Outsourcing (SSCO Model)

- Alice has a database, but she does not have enough resources to support the storage and the database operations.



SSCO Model (Continued)

- Alice's requirement:
  - I don't want either Bob or the client to know my database.
  - I want to charge Carl for each of his queries.
- Carl's (the customer) requirement:
  - I don't want either Alice or Bob to know my query or the result.

Our Solutions

- PIR/APPROX protocols for different metrics.
  - *sum-of-squares* metrics,
  - *sum-of-absolute-values* metrics,
  - *string-edit* metrics.
- SSO/APPROX protocol.
  - *sum-of-squares* metrics.
- SSCO/APPROX protocol.
  - *sum-of-squares* metrics.

Research in Progress

- Improve current solutions for PIR
- Find a practical solution for PIRPD model
- Identify new models
- Extend to other applications
  - Proximity queries
  - Genetic database queries
  - GIS