Querying Cloud Data with Confidentiality and Integrity Assurance

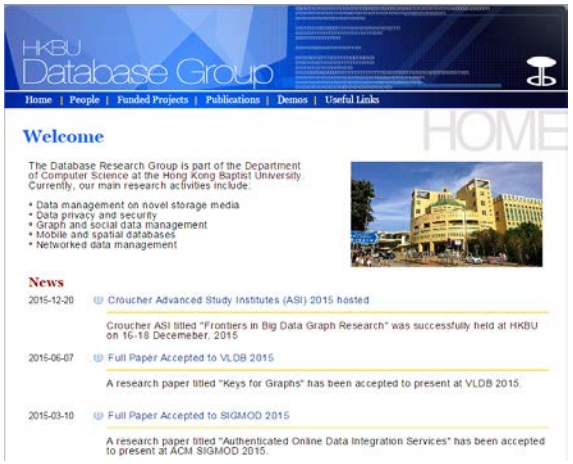
Jianliang Xu

Database Research Group
Hong Kong Baptist University (HKBU)

http://www.comp.hkbu.edu.hk/~db

Group Profile

- People
  - 3+1 faculty members
  - 9 PhD students and research/project assistants



- Focused Research Areas
  - Data management on new hardware
  - Data security and privacy
  - Graph and social data management
  - Mobile and spatial databases

- Funding
  - RGC, ITF, NSFC
  - HK$10+M grants secured in 2012-2016

Outsourcing in the Cloud

- Scalability
- Elasticity
- Self-manageability
- Pay-per-use pricing

Cloud Service Provider (SP)

Data & Algorithm

Data Owner

Client

Security Challenge #1: Integrity

- Scalability
- Elasticity
- Self-manageability
- Pay-per-use pricing

Incorrect results

- Hacking attack
- Incomplete search
- Program bug
- In favor of sponsor

Cloud Service Provider (SP)

Data & Algorithm

Data Owner

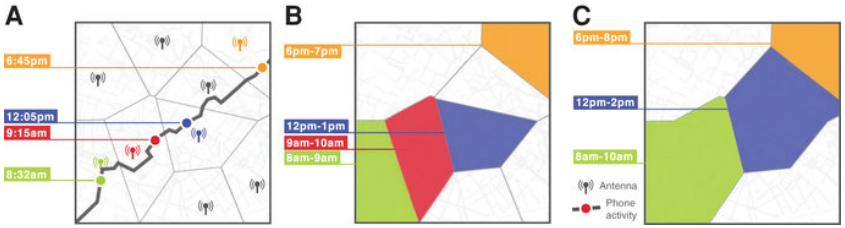
Client

Security Challenge #2: Confidentiality

- Data Privacy
  - Private asset of the data owner
  - Containing commercial intelligence information or sensitive personal information
  - Protected against the cloud server and/or the query client

Hey, wanna buy some data? Only slightly used! Heap of clicks left in it. Its last owner was a little old lady who only used it for shopping ...

- Query Privacy
  - 2006: AOL search data leak
  - 2013: 4 spatio-temporal locations could identify people!



[Nature SRep, 2013]

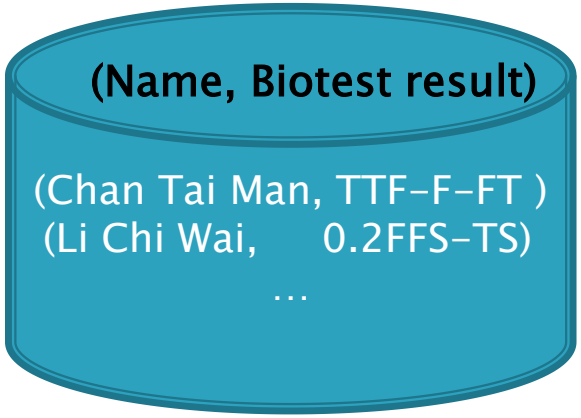
Roadmap

- Motivation and Challenges
- Secure Data Search
  - Private Search on Key-Value Stores [ICDE14]
  - Structure-Preserving Subgraph Queries [ICDE15]
- Query Integrity Assurance
  - Privacy-Preserving Query Authentication [SIGMOD12, PVLDB14]
  - Multi-Source Query Authentication [SIGMOD15]
- Summary

Private Search on Key-Value Store

- Example: eHR (Electronic Health Record)
  - A doctor accesses an eHR database for the test result of a patient through a patient id

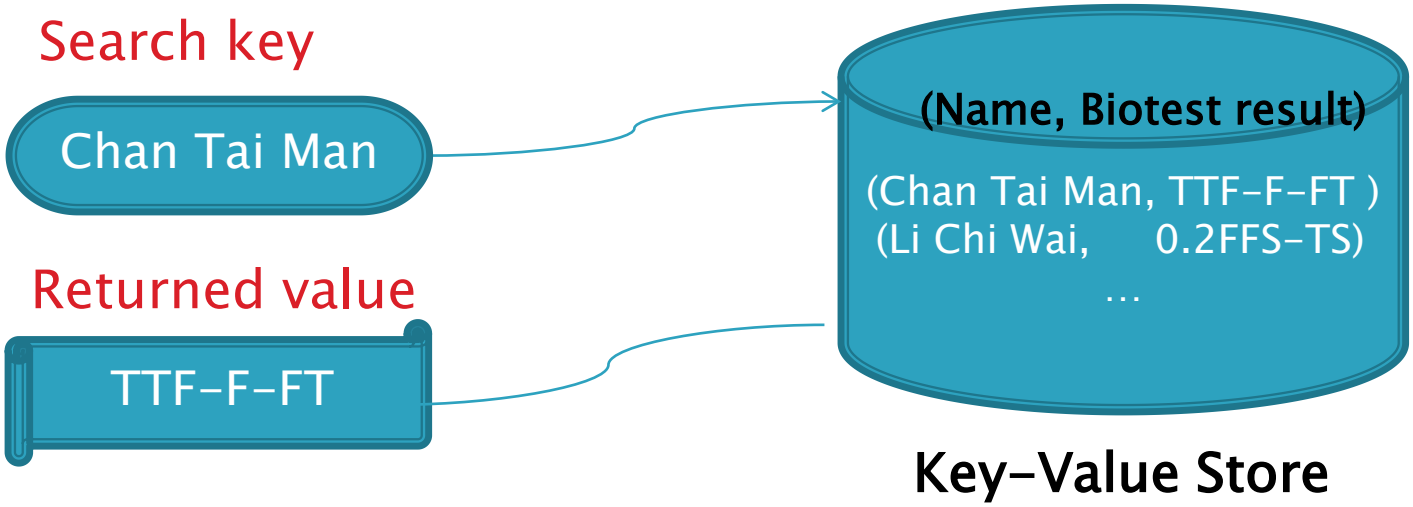
- Security goal: mutual privacy
  - The doctor should receive **ONLY** this patient's result, but no any other's result
  - The SP should NOT know which test result is accessed, as it was protected by the patient-doctor confidentiality



Key-Value Store

Problem Definition

- A mutual privacy model:
  - Search key or returned value is **NOT** learned by the server
  - **ONLY** the value that matches the search key is returned



Preliminary 1: Homomorphic Encryption

- Plaintext space M and Cipher text space C
  - $m_1, m_2 \in M$ and their ciphertexts $c_1, c_2 \in C$
- It holds that

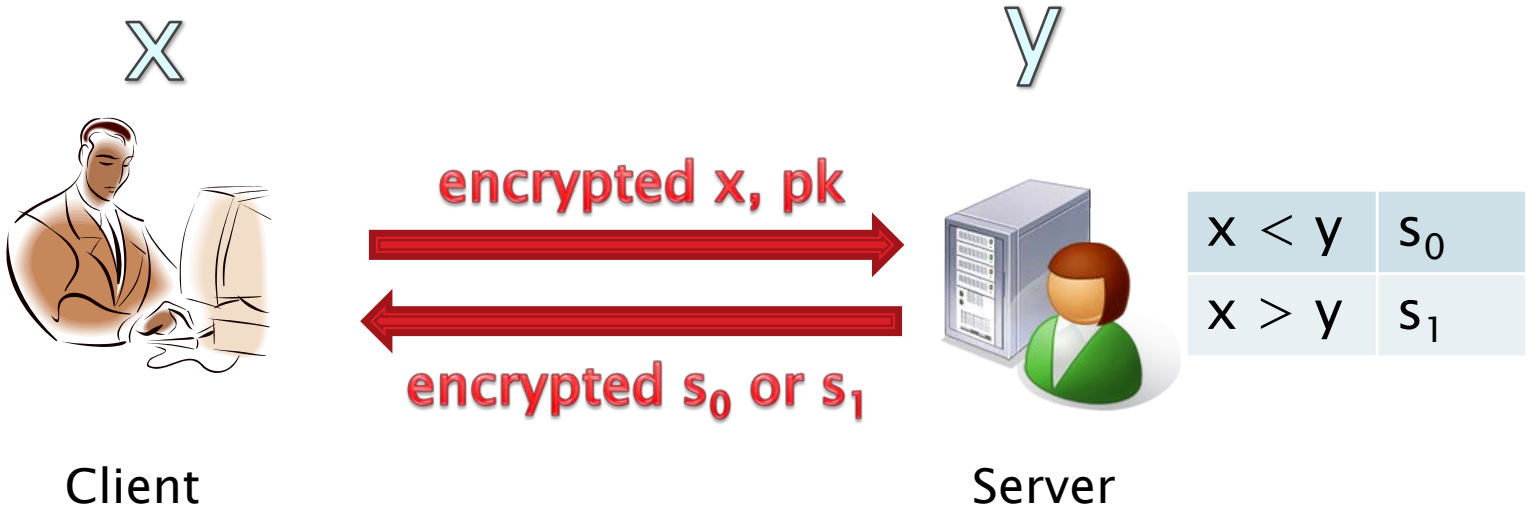
$$E^{-1}(c_1 \odot c_2) = m_1 \oplus m_2$$

- Examples: Paillier, Goldwasser-Micali (GM), RSA, El Gamal

Credit: Craig Gentry, inventor of the first fully homomorphic encryption

Preliminary 2: Conditional Oblivious Transfer

- GT-COT: Conditional Oblivious Transfer for "Greater Than"
  - The value to receiver is determined by the result of "greater-than" predicate on private inputs x, y from both parties
    - The client gets s_0 if $x < y$, or gets s_1 if $x > y$
    - The predicate result is only known to the client
    - GT-COT can be implemented based on Paillier encryption

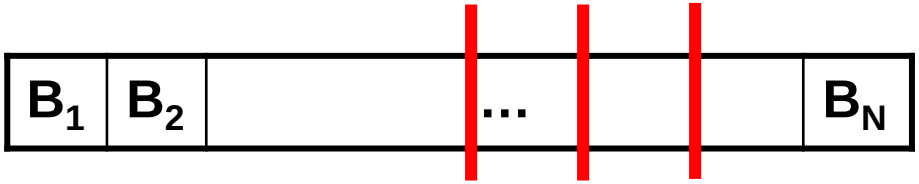


$x < y$	s_0
$x > y$	s_1

[Blake and Kolesnikov, 2004]

Applying GT-COT to Key-Search Problem

- A basic approach
  - Sort keys in ascending order
  - Invoke GT-COT for each key
  - Cost: O(N) calls

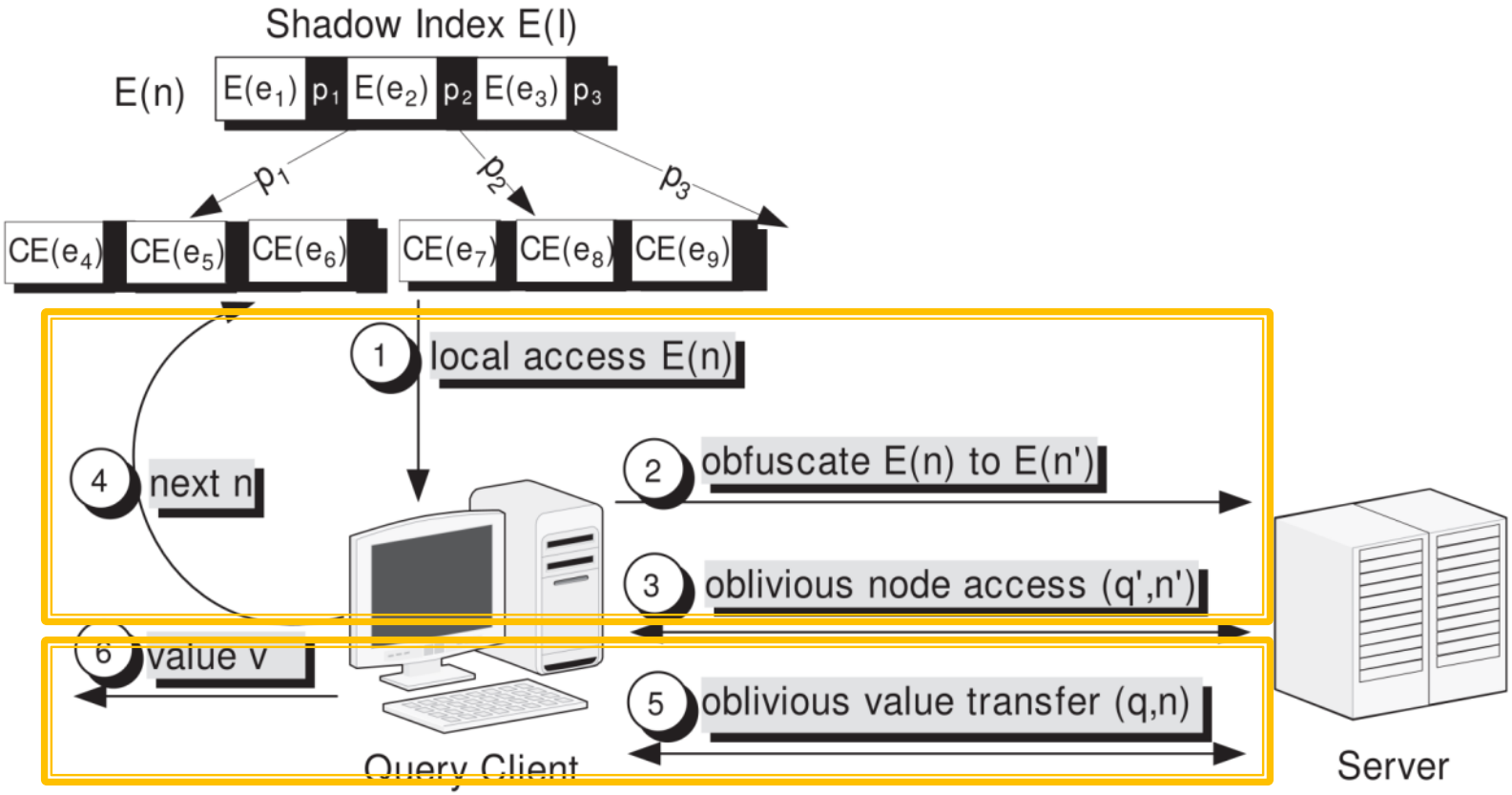


- Improved version
  - Apply binary search
  - Only need O(log N) calls to GT-COT

This doesn't work as the server needs to know the next y to compare with x

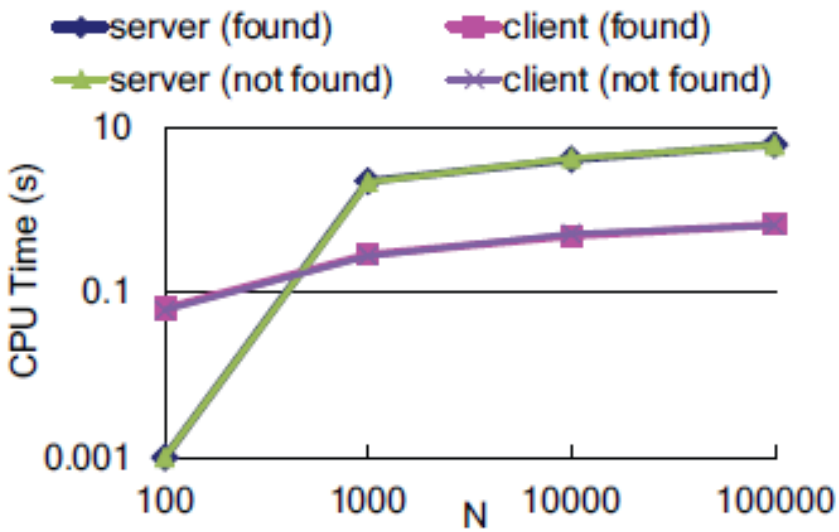
Our Proposal: Oblivious Index Traversal

- Reducing complexity from O(N) to O(log N)
- Hide the y value from the server by another layer of homomorphic encryption

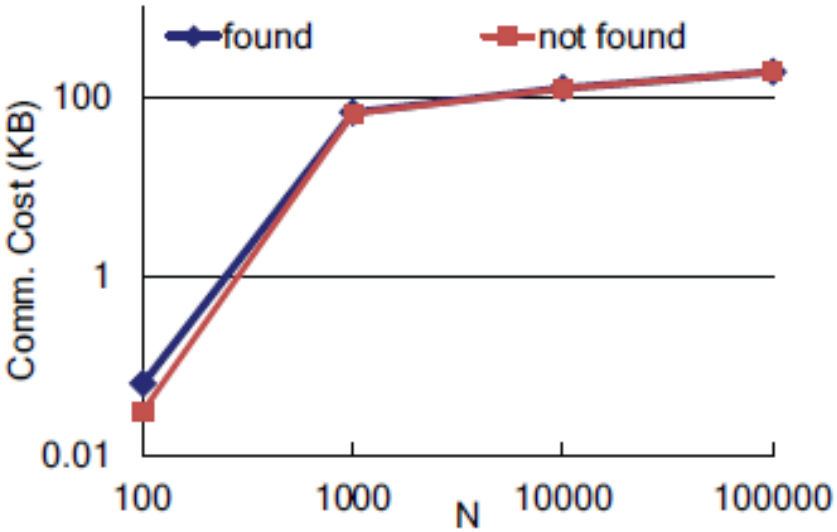


[IEEE ICDE 2014]

Performance Results



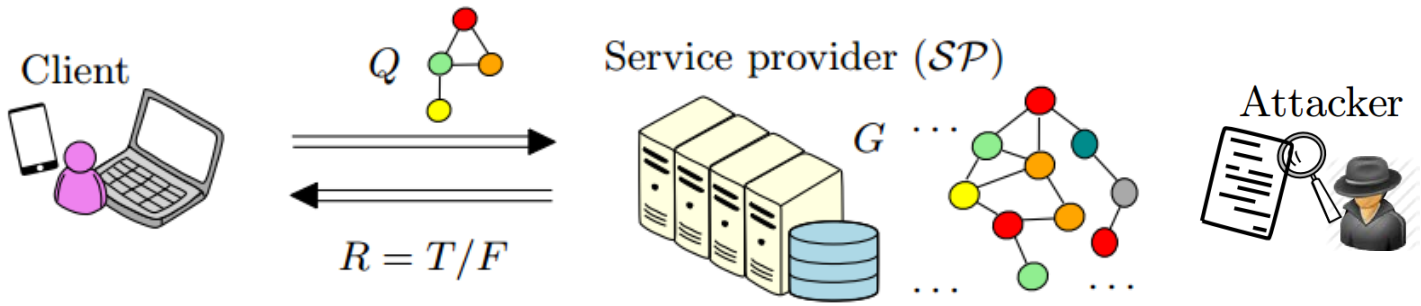
(a) CPU Time



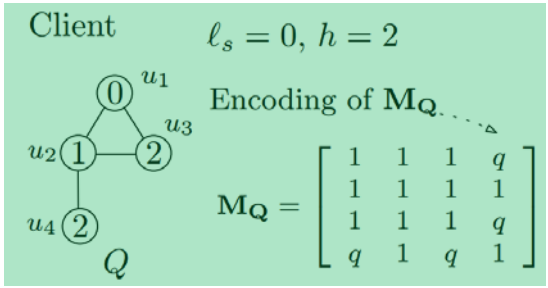
(b) Communication Bandwidth

Structure-Preserving Subgraph Queries

- **Subgraph queries** in social media search, bioinformatics, web topology…
- **Challenge**: evaluate subgraph queries at the SP while protecting the structure of query graph against SP



- **Methodology**
  - Encode query graph with cyclic group-based (CBG) encryption
  - Reduce and verify candidate mappings in encrypted domain



[IEEE ICDE 2015]

Roadmap

- Motivation and Challenges
- Secure Data Search
  - Private Search on Key-Value Stores [ICDE14]
  - Structure-Preserving Subgraph Queries [ICDE15]
- Query Integrity Assurance
  - Privacy-Preserving Query Authentication [SIGMOD12, PVLDB14]
  - Multi-Source Query Authentication [SIGMOD15]
- Summary

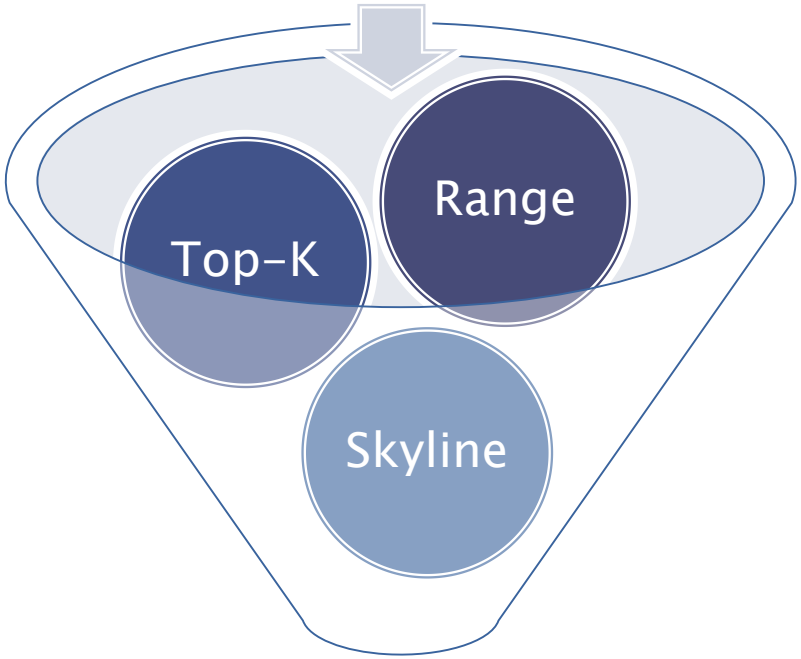
Authenticated Query Processing

- Empower the query client to verify the integrity of query results

- What to verify?
  - Soundness: all results are not tampered with
  - Completeness: no missing results

Query Flow vs Authentication Flow

Raw Data



Query Result

Raw Data

Crypto, Indexing

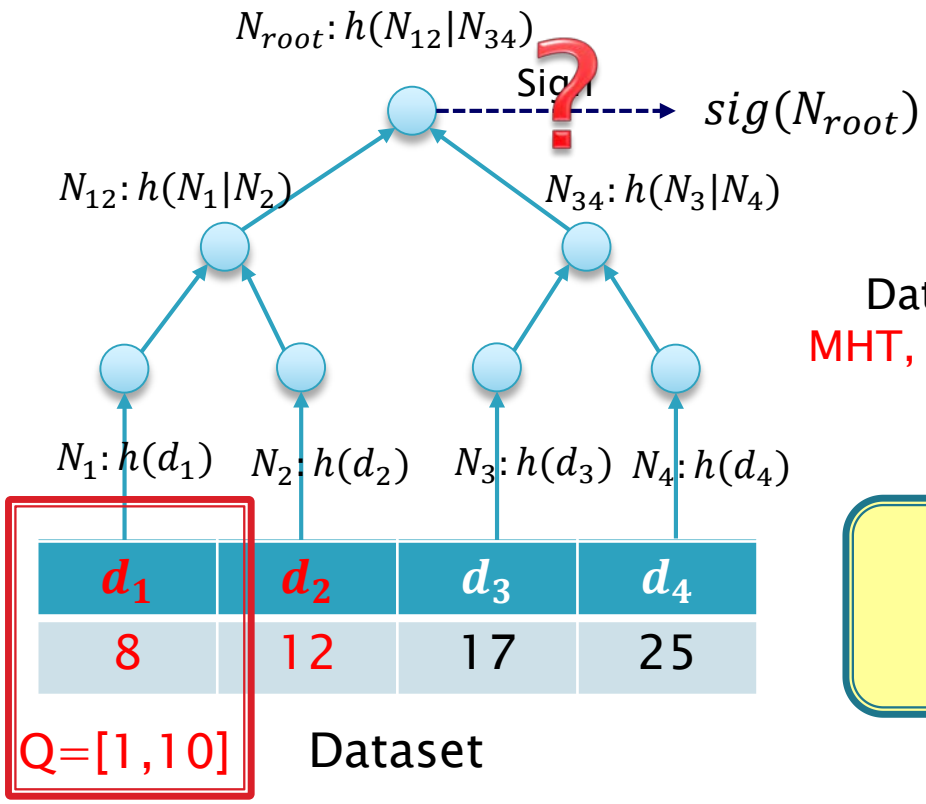
Authenticated Data Structure

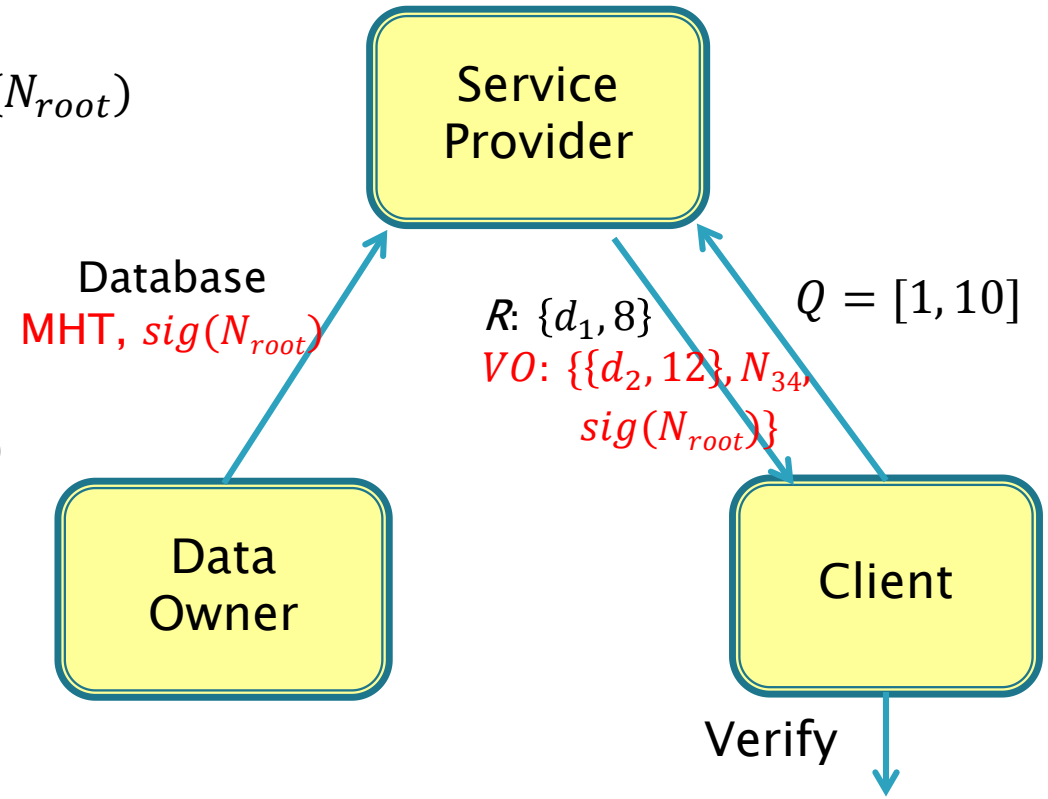
Prune, Optimization

VO: Proof of Query Result

Merkle Hash Tree

Merkle Hash Tree (MHT)



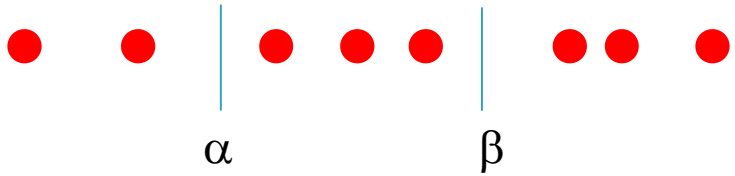


- Soundness: 8 ∈ [1,10]; root
- Completeness: 12,N_{34} ∉ [1,10]

Privacy-Aware Query Authentication

- **Problem**: To verify query results without letting query client learn anything beyond

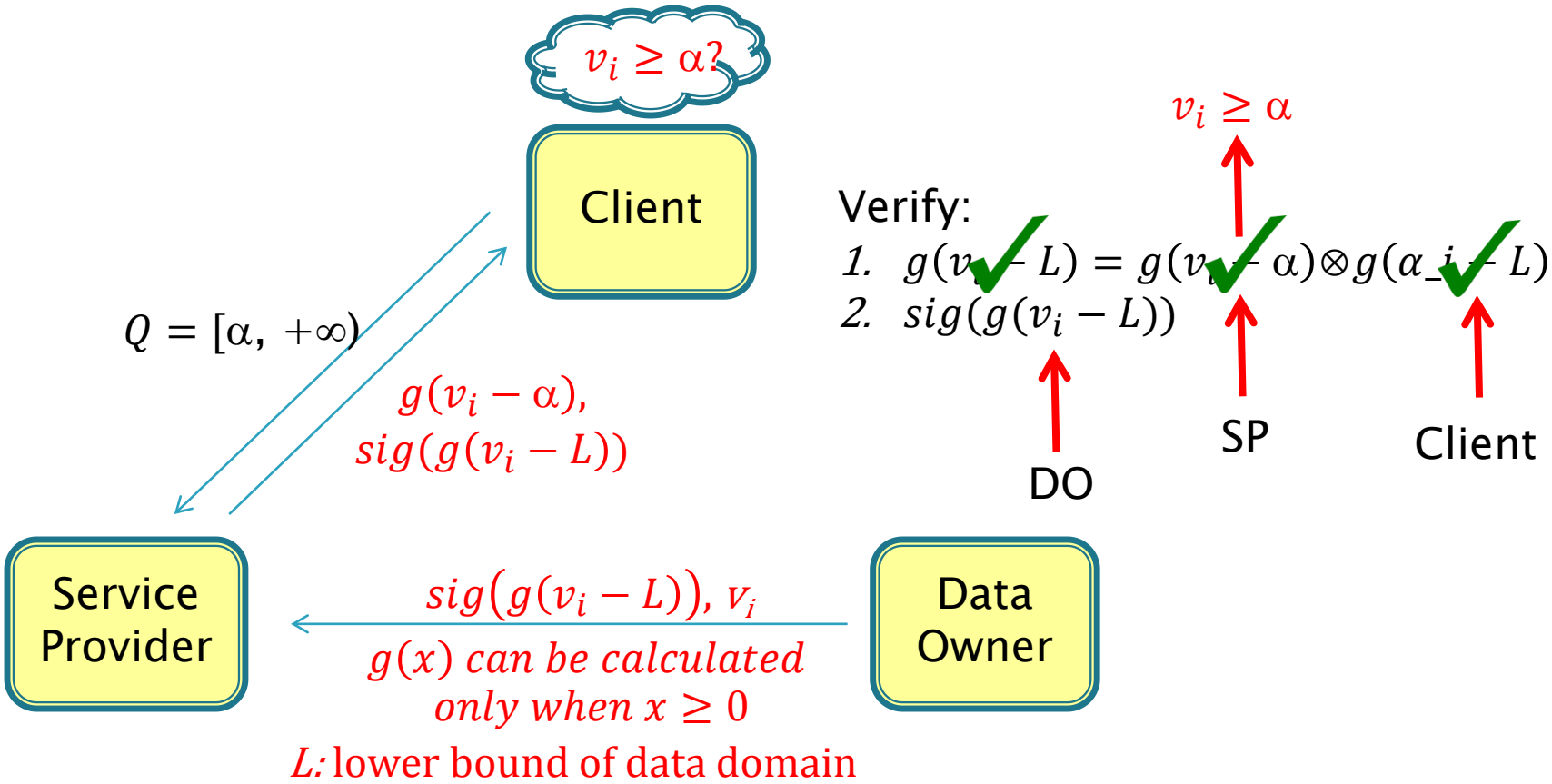
- Motivating scenarios
  - **Range Query**: Count the number of employees in a certain salary range [α, β] without knowing specific salaries
  - **Top-k Query**: Find new nearby friends without learning detailed locations and profiles



[ACM SIGMOD 2012, PVLDB 2014]

Building Block for Private Comparison

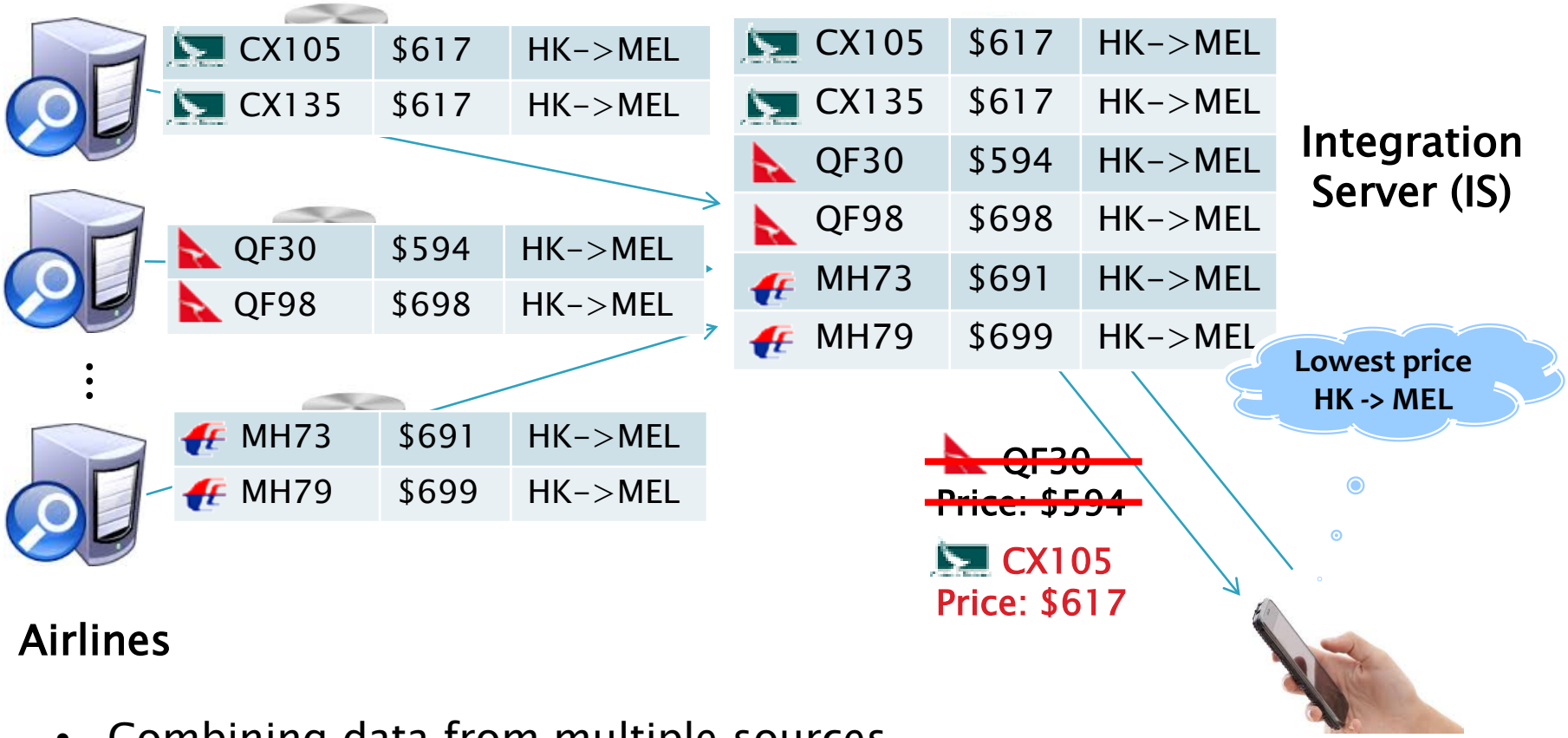
- To verify $v_i \geq \alpha$ without knowing v_i (v_i is *private value*)
- Basic idea: joint computing by SP and client



Roadmap

- Motivation and Challenges
- Secure Data Search
  - Private Search on Key-Value Stores [ICDE14]
  - Structure-preserving Subgraph Queries [ICDE15]
- Query Integrity Assurance
  - Privacy-Preserving Query Authentication [SIGMOD12, PVLDB14]
  - Multi-Source Query Authentication [SIGMOD15]
- Summary

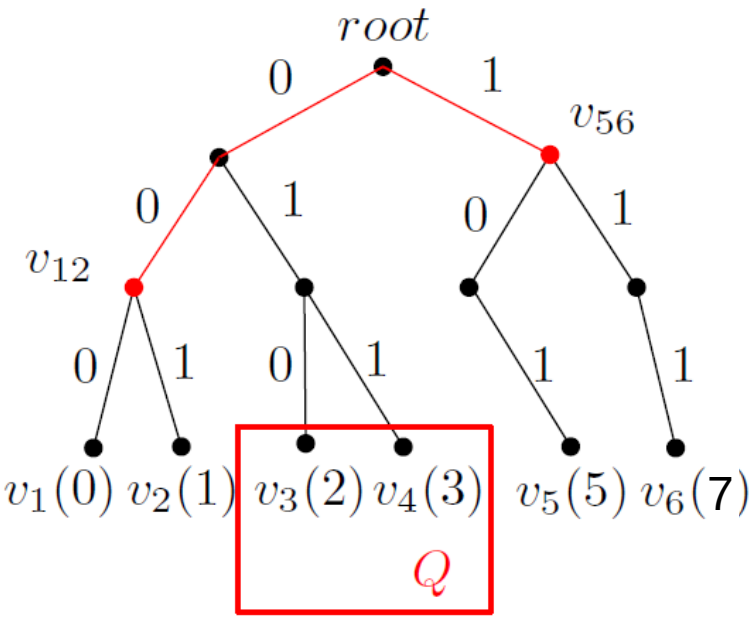
Example: Metasearch Engines



- Combining data from multiple sources
- Providing users with a unified query interface

Design Intuition – Prefix Tree

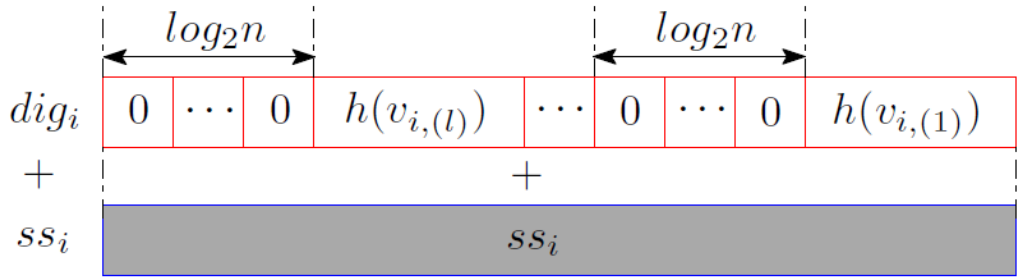
- Verify far-away non-result values without using the whole dataset
  - v_{12} proves v_1, v_2
  - v_{56} proves v_5, v_6
- Issue: $v_1, \ldots, v_6$ have sigs; v_{12}, v_{56} don't
- Challenge: aggregatable signature needed



Prefix Tree based on v_1 to v_6

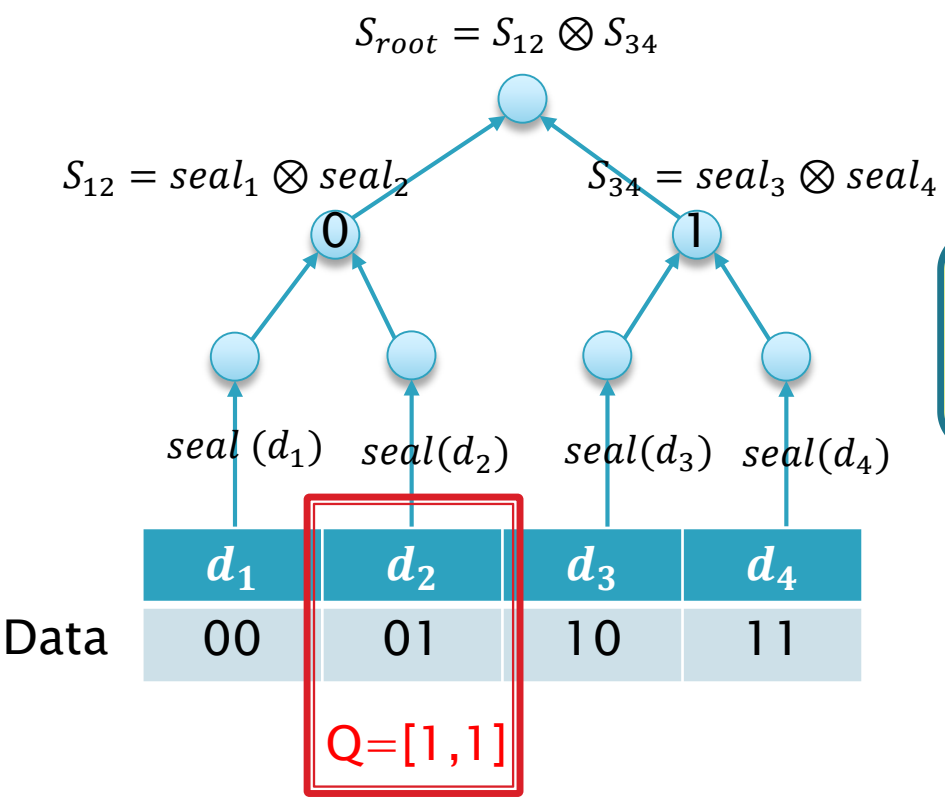
Homomorphic Secret Sharing Seal

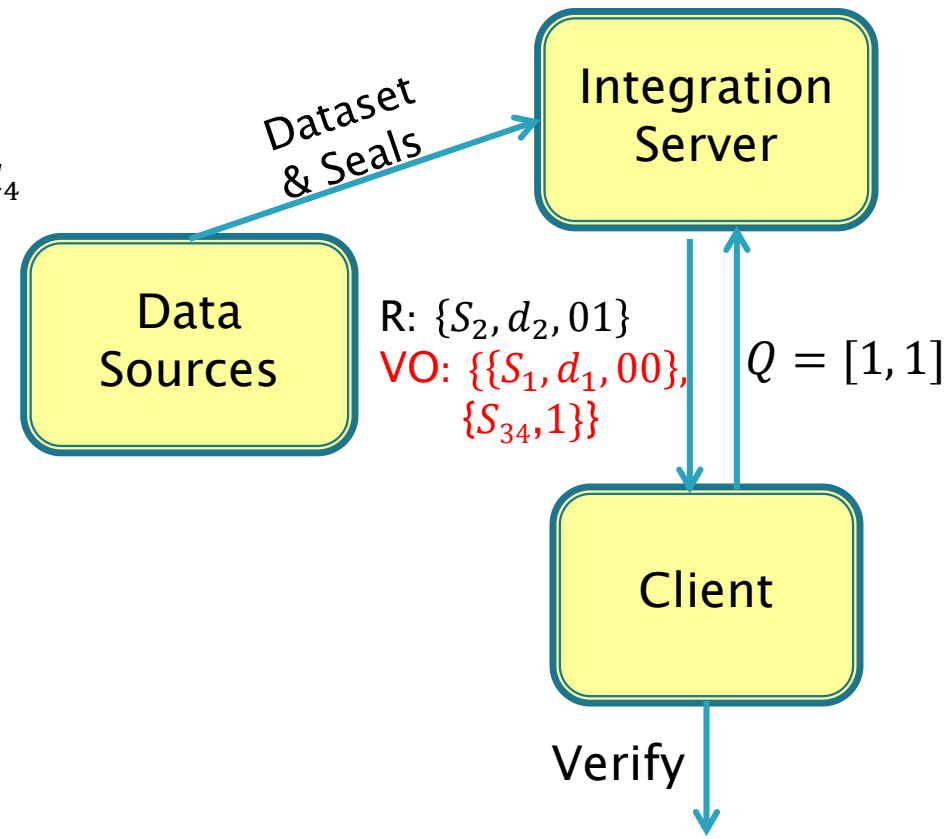
- Content:
  - $h(\cdot)$
  - $dig_i + ss_i$
  - $seal(v) = \mathcal{G}(v, ss_i) = \left(g^{ss_i} \cdot g^{h(v_{(3)})|h(v_{(2)})|h(v_{(1)})}\right)^a \ mod\ n$



- Seal design
  - Seals are "additively" homomorphic
  - Seals can be folded by the integration server
    - $v_1 = 000 \Rightarrow \boldsymbol{S_1 = \mathcal{G}(000, ss_1)}$ from o_1
    - $v_2 = 001 \Rightarrow \boldsymbol{S_2 = \mathcal{G}(001, ss_2)}$ from o_2
    - $\boldsymbol{seals(prefix(v_1, v_2)) = \mathcal{G}(00, ss_1 + ss_2) = S_1 \otimes S_2}$

Authenticated Prefix Tree





- Soundness: $d_2 \in [1,1]$; S_1, S_2, S_{34}
- Completeness: $d_1, 1 \notin [1,1]$; secret

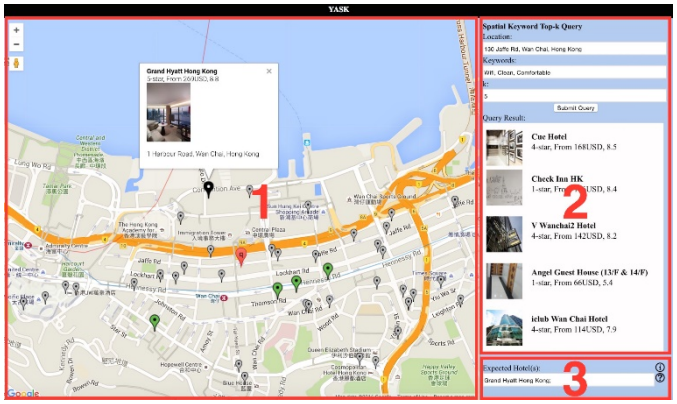
Summary

- Protection of data access is crucial in cloud settings
- It is possible and yet meaningful to support secure query processing with confidentiality and integrity assurance
  - Cryptographic-only approach cannot scale well
  - Integration with **database techniques** is crucial

Querying Spatial Data

- **Why-Not Questions [ICDE15, ICDE16a]**
  - Some object(s) unexpectedly missing from query results
  - How to **minimally** modify the initial query to revive missing object(s)?

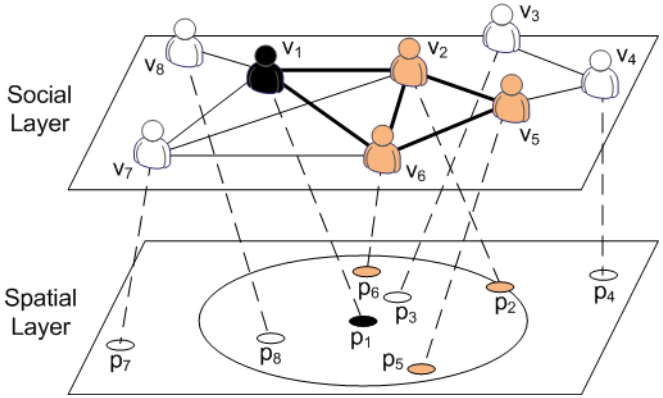
Research Session 6A, Wednesday



[PVLDB 2016 Demo]

- **Geo-Social Group Queries [TKDE15, ICDE16b]**
  - Group-based activity planning and marketing, friend gathering…
  - Challenge: **efficient** processing while considering both spatial and social constraints

TKDE Poster Session, Tuesday

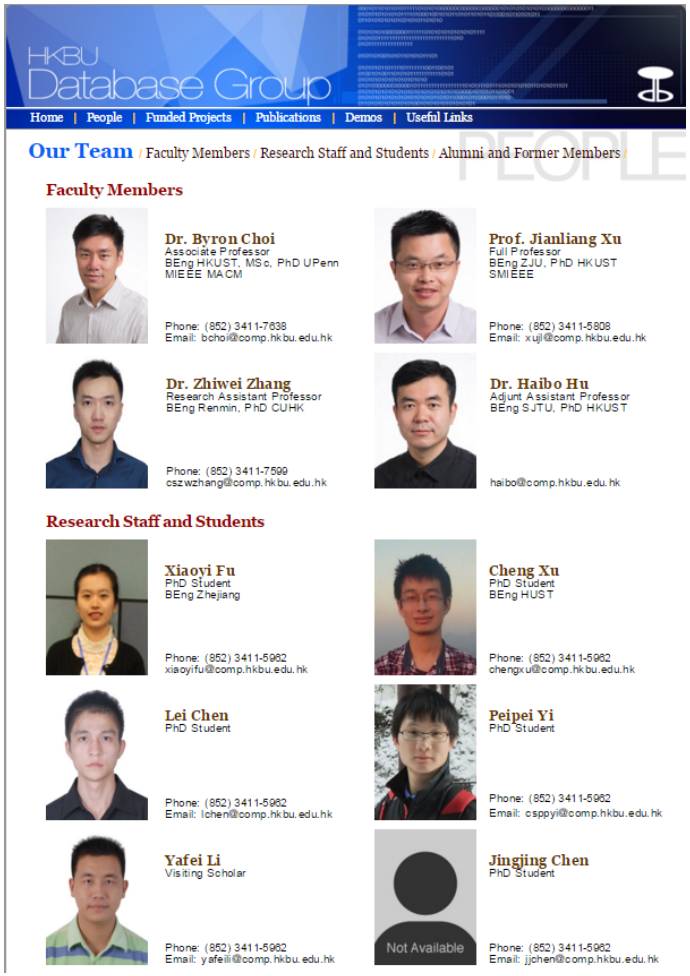


Acknowledgments

- Members of HKBU DB Group

- External Collaborators
  - Christian S. Jensen (Aalborg University)
  - Prof. Sourav S. Bowmick (Nanyang Technological University)
  - Prof. Wang-Chien Lee (Penn State University)
  - Dr. Rui Chen (Samsung Research America)

- Funding Agencies
  - Research Grants Council of Hong Kong
  - Innovative Technology Fund
  - Hong Kong Scholars Program



Thank You!

References

- **[PVLDB16]** L. Chen, J. Xu, C. S. Jensen, and Y. Li. "YASK: A Why-not Question Answering Engine for Spatial-Keyword Query Services." PVLDB, 2016. (Demo)
- **[ICDE16a]** L. Chen, J. Xu, X. Lin, C. S. Jensen, and H. Hu. "Answering Why-Not Spatial Keyword Top-k Queries via Keyword Adaption." *ICDE*, 2016.
- **[ICDE16b]** Y. Li, R. Chen, J. Xu, Q. Huang, H. Hu, and B. Choi. "Geo-Social K-Cover Group Queries for Collaborative Spatial Computing." *ICDE*, 2016 (Poster)
- **[SIGKDD15]** R. Chen, Q. Xiao, Y. Zhang, and J. Xu. "Differentially Private High-Dimensional Data Publishing via Sampling-Based Inference." *ACM KDD*, August 2015.
- **[SIGMOD15]** Q. Chen, H. Hu, and J. Xu. "Authenticated Online Data Integration Services." *ACM SIGMOD*, May/June 2015.
- **[ICDE15]** Z. Fan, B. Choi, J. Xu, and S. S. Bhowmick. "Asymmetric Structure-Preserving Subgraph Query for Large Graphs." *IEEE ICDE*, April 2015.
- **[TKDE15a]** D. Wu, B. Choi, J. Xu, and Christian S. Jensen. "Authentication of Moving Top-k Spatial Keyword Queries." IEEE Transactions on Knowledge and Data Engineering (TKDE), 2015.
- **[TKDE15b]** Y. Peng, Z. Fan, B. Choi, J. Xu, and S. S. Bhowmick. "Authenticated Subgraph Similarity Search in Outsourced Graph Databases." IEEE Transactions on Knowledge and Data Engineering (TKDE), 2015.
- **[TKDE15c]** Z. Fan, B. Choi, Q. Chen, J. Xu, H. Hu, and S. S. Bhowmick. "Structure-Preserving Subgraph Query Services." IEEE Transactions on Knowledge and Data Engineering (TKDE), 2015.
- **[PVLDB14]** Q. Chen, H. Hu, and J. Xu. "Authenticating Top-k Queries in Location-based Services with Confidentiality." PVLDB, 2014.
- **[TKDE14]** X. Lin, J. Xu, H. Hu, and W.-C. Lee. "Authenticating Location-Based Skyline Queries in Arbitrary Subspaces." IEEE TKDE, 2014.
- **[ICDE14]** H. Hu, J. Xu, X. Xu, K. Pei, B. Choi, and S. Zhou. "Private Search on Key-Value Stores with Hierarchical Indexes." IEEE ICDE, 2014.
- **[SIGMOD12]** H. Hu, J. Xu, Q. Chen, Z. Yang. "Authenticating Location-based Services without Compromising Location Privacy." ACM SIGMOD, 2012.