Data Sharing:
Querying and Linking Distributed and Autonomous Data

Yang Cao, Wenfei Fan, Peter Buneman

University of Edinburgh

Background

Distributed databases: healthcare, business intelligence, e-government, ...

- *Tasks*: querying, linking and sharing
- *Data*: distributed, heterogeneous, large scale

Challenges:

- **Privacy & security**: data owners ("private") vs. users ("open")
- **Heterogeneity**: relations, key-value, graphs
- **Scalability**: limited resources vs. big data analytics
- **Accountability**: ownership and accountability of shared data collection

Querying Shared Data with Security Heterogeneity [4]

Challenges

- heterogeneous security requirements
- centralized evaluation not possible

Solution:

1. Modeling data sharing protocols
2. Query evaluation under protocols
   - heterogeneous distributed query plans
   - optimal security-efficiency trade-offs
   - data movement and security allocation
   - leverage security heterogeneity in plans

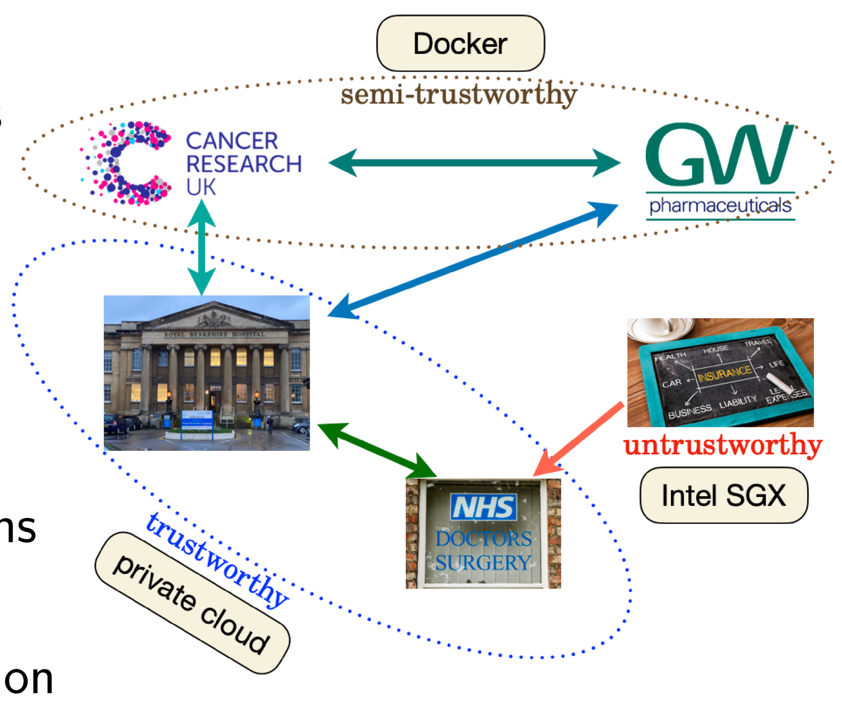

Figure: Heterogeneous data sharing

Data sharing protocols ρ specifying:

- capsules: logic units for computations over shared data
- hosts: data owners that host capsules
- pair-wise privacy requirements:
  - permitted capsule specifications
  - secure communication overheads

Heterogeneous distributed query plan: A DAG of

$$\textit{atomic operations}: \delta = (\mathbf{op}, \mathbf{t_c}, X_1, \ldots, X_n, j)$$

- **op**: an **RA** operator
- $\mathbf{t_c}$: type of the security facility for operation δ
- X_i: data from site i for δ
- j: the host site of δ.

Cost model: $\mathbf{toll}(\delta)$:

- upfront cost, secure communication cost, computation cost
- submodular set function: X_i to capsule at j

Planning under protocols: find optimal (minimum toll) plan

[Complexity]:

1. decidable in NEXPTIME;
2. PSPACE-hard even when ρ is linear;
3. Σ_3^p-hard even when Q is in **SPC** and ρ is linear.
   Moreover, 2 and 3 hold even when $\mathcal{S}$ has two sites only.

Intractable to make the best (optimal) use of heterogeneity in data sharing.

[Algorithm]: Nonetheless, a two phase approach with guarantees:
Step (1): generating toll-minimized distributed plan ξ_Q

- toll-minimized $\xi_{\mathbf{op}}$ for each **op** of Q
- an $O(\log n)$-approximation algorithm for $\bowtie$

Step (2): optimizing ξ_Q within the toll budget

- via an automic operaton κ for "rebalancing" ξ_Q
- a near-optimal design of κ (2-approximation of the optimal for $\bowtie$)

[Effectiveness]: it speeds up state-of-the-art secure database system (SMCQL) by 18+ times over 1GB of **TPCH** data.

Heterogeneous Entity Resolution [5]

Main task: link entities across a relational database D and a graph G

Challenge: traditional ER methods work for relations only (homogeneous).

HER: decide whether tuple $t \in D$ matches vertex v of G:

1. convert relations D into a graph G_D
2. whether t matches $v \Rightarrow$ whether v_t in G_D matches v in G
3. parametric simulation between G_D and G
   - graph simulation extended with label matching functions
   - remains in quadratic-time
   - learned threshold for score functions
   - parallelizable to scale over large graphs and relations

Scaling via Consistent Data Caching [1]

Background:

- large communication overhead and limited resources per node
- transactional accesses (workloads) become prevalent

Main task: scale out via data cache with consistency guarantees

Results:

- prove traditional policies (e.g., LRU) are not competitive for transactions
- formulate consistent cache scheme; show it's NP-complete even for uni-size accesses, in contrast to trivially PTIME for conventional caching
- develop a consistent cache policy that is theoretically competitive and optimal when transactions access data items of unit size
- transaction batching and reordering for caching
- implemented and tested with Memcached@HBase: 126% improvement on average for transaction throughput, while guaranteeing consistency

Accountability of Shared Curated Data [3, 2]

Background:

- datasets are often *shared/copied* $\rightarrow$ *modified* $\rightarrow$ *published* $\rightarrow \cdots$
- a data collection contains contributions from multiple users
- contributions form a dependency hierarchy (copy-modify-contribute)
- however, entire dataset is often treated as one single "article"

Main task: how to account the ownership of pieces of data in a dataset

Results:

- a model of citation graph for databases
- method for generating data summaries of "optimal" granularity
  - stress measures of data summaries
  - compute accountability by measuring stress
- application to a collectively curated pharmacology database (GtoPdb)

Publications

[1] S. An, Y. Cao, and W. Zhao. Competitive consistent caching for transactions. In *ICDE*, 2022.

[2] P. Buneman, D. Dosso, M. Lissandrini, and G. Silvello. Data citation and the citation graph. *Quant. Sci. Stud.*, 2(4), 2021.

[3] P. Buneman, D. Dosso, M. Lissandrini, and G. Silvello. Expanding the citation graph for data citations. In *SEBD*, pages 276–283, 2022.

[4] Y. Cao, W. Fan, Y. Wang, and K. Yi. Querying shared data with security heterogeneity. In *SIGMOD*, 2020.

[5] W. Fan, L. Geng, R. Jin, P. Lu, R. Tugay, and W. Yu. Linking entities across relations and graphs. In *ICDE*, 2022.