# Privacy-Preserved Neural Graph Databases

Yangqiu Song

Department of CSE, HKUST

Slides Credit: Qi Hu

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

THE DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING 計算機科學及工程學系

KnowComp Group

Understanding the World by Computational Knowledge

# Our Research in the Era of LLMs

- LLMs have “killed” many research directions
- What do we do? IMHO,
  - The challenges that LLMs still face
  - The existing/new applications that LLMs enable

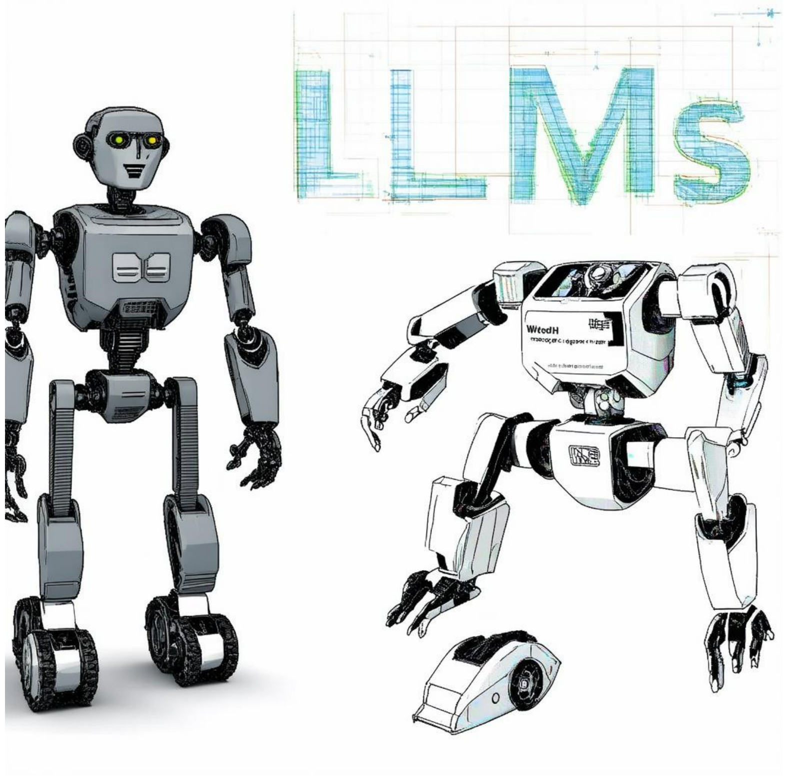



Image generated by Stable Diffusion v3 Medium - by fal.ai

# Challenges

- Factuality hallucination emphasizes the discrepancy between generated content and verifiable real-world facts, typically manifesting as factual inconsistency or fabrication

- Specific domain knowledge/Long-tail knowledge

Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232. 2023 Nov 9.
Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, Xin Luna Dong. Head-to-Tail: How knowledgeable are Large Language Models? A.K.A. Will LLMs replace knowledge graphs? In arXiv2023.

# New Applications

- LLMs provides interactive natural language interface to many things
  - Self-driving cars
  - Excel spread sheets
  - Local databases
  - Etc.

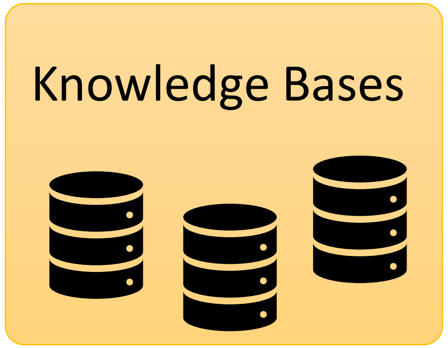
- LLMs provides much better representation for free texts to enable
  - Semantic search in text-rich databases
  - Search engines
  - Etc.

# Retrieval Augmented Generation (RAG)

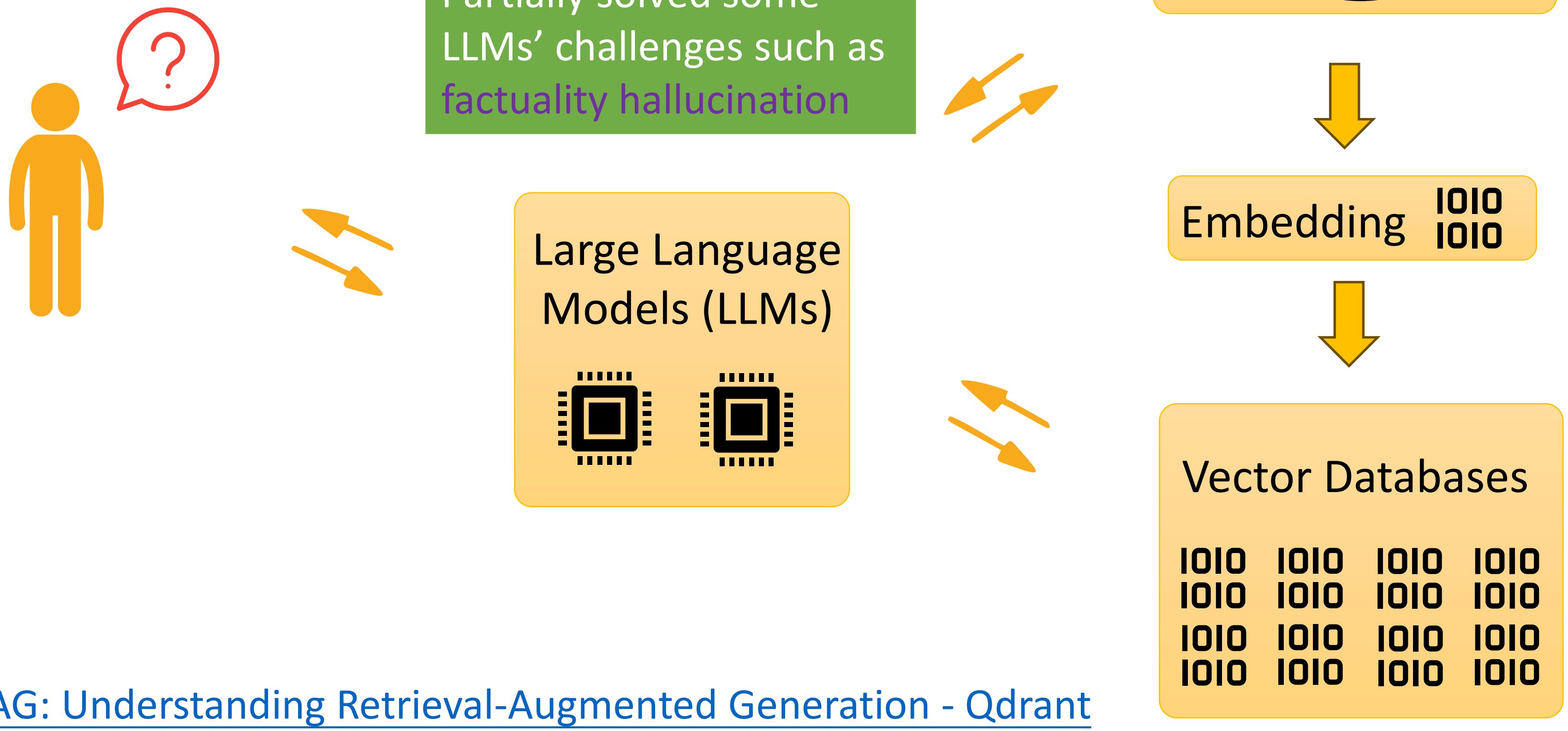
**1. Retrieval**: Fetches relevant documents from a large dataset.

**2. Augmentation**: Uses retrieved documents to provide context.

**3. Generation**: Generates responses based on both the input and retrieved context.




Partially solved some LLMs' challenges such as factuality hallucination




Enabled by LLMs to have a better fuzzy semantic search when there is an open-world assumption

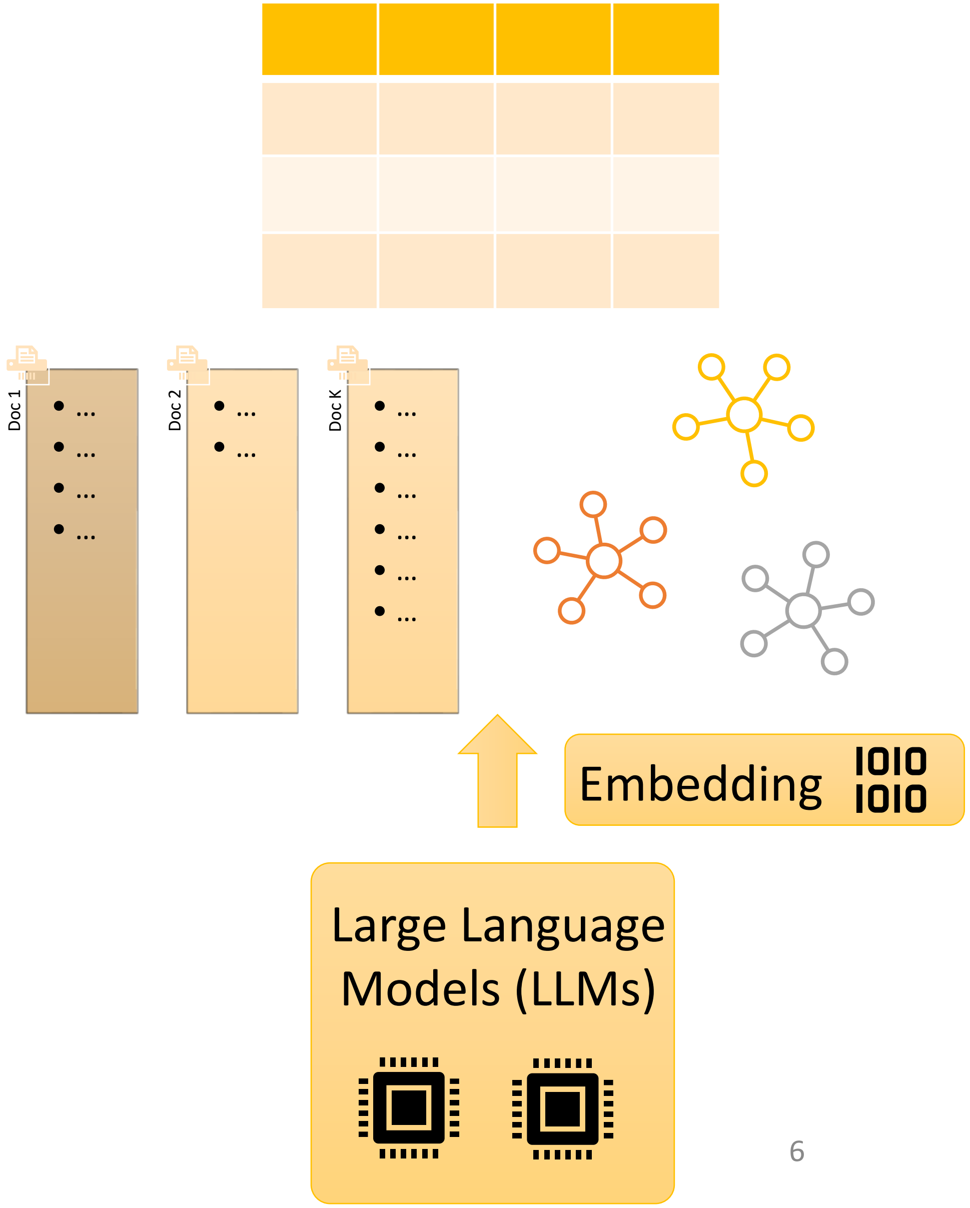
- Retrieved information may not be accurate

[What is RAG: Understanding Retrieval-Augmented Generation - Qdrant]

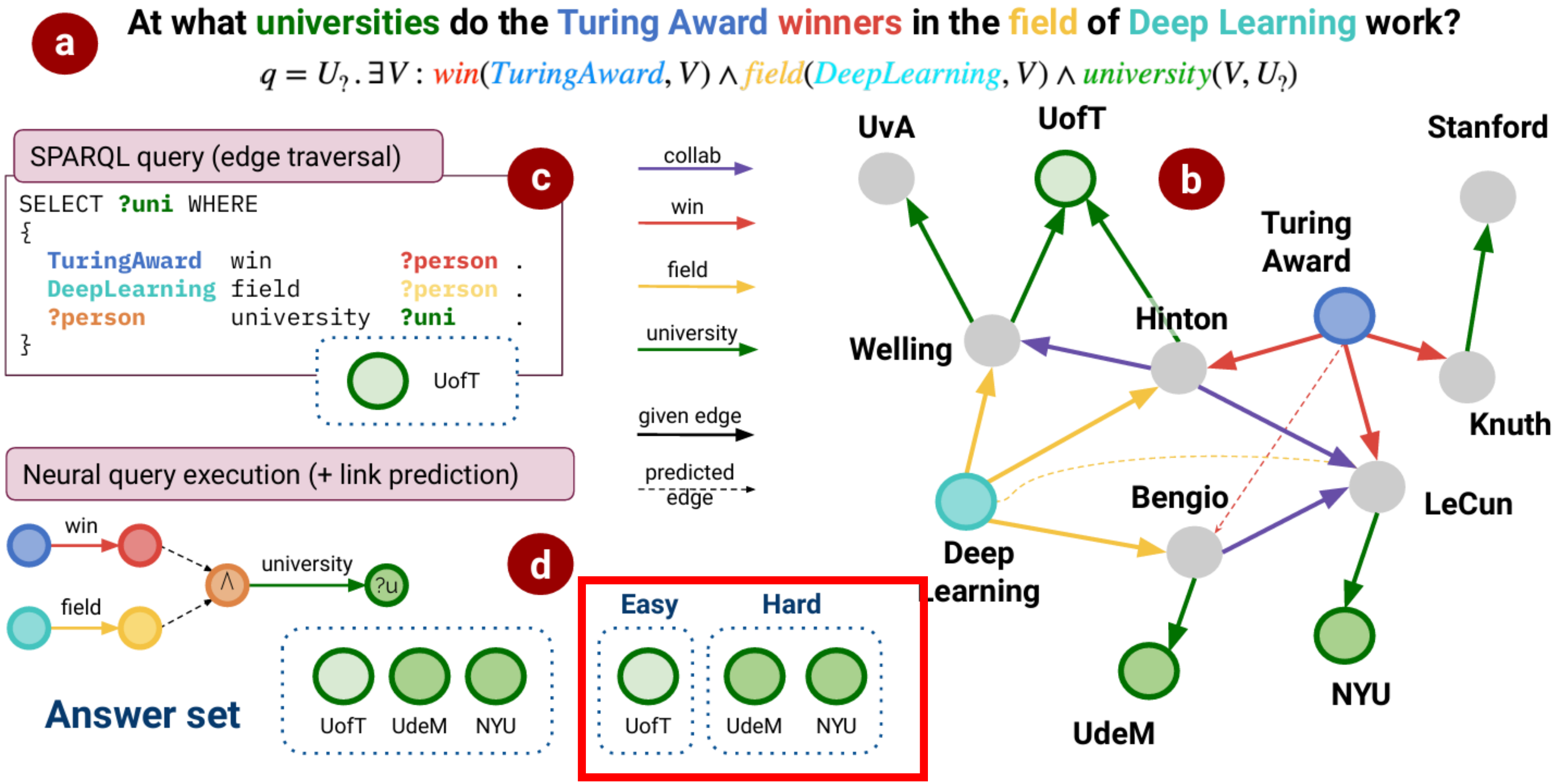
# From Vector DBs to Neural Graph DBs

- Why Graphs?
  - Sometimes we need globally and structural referenced knowledge
  - Ability of reasoning with high complexity
    - NP-complete problems, e.g., Max-Sat (Chalier et al., 2022) , subgraph matching or counting, subset sum, etc.
  - The trade-offs between scalability and computational complexity

- Leverage both neural and symbolic reasoning power



# Graph Query

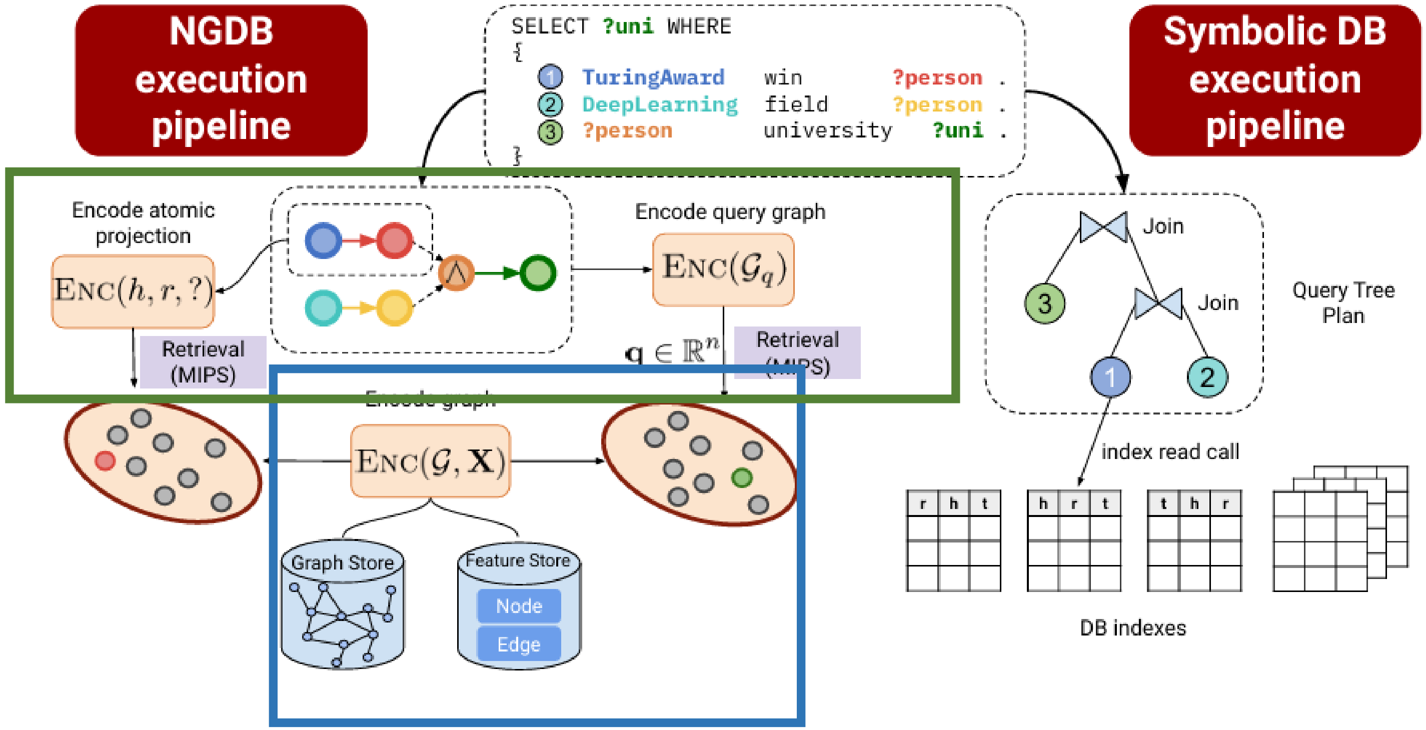



**Limitation**: Missing knowledge results in incomplete answer set.

Complex Graph Queries (Figure taken from Ren et al)


Ren H, Galkin M, Cochez M, Zhu Z, Leskovec J. Neural graph reasoning: Complex logical query answering meets graph databases. arXiv preprint arXiv:2303.14617. 2023 Mar 26.

# Neural Graph Databases (NGDBs)




Neural Graph Databases (Figure taken from Ren et al)

**Neural Graph Storage**: employ graph store and feature store to obtain latent representations in the embedding store.

**Neural Query Engine**: derive the computation graph of the query and execute in the latent space.

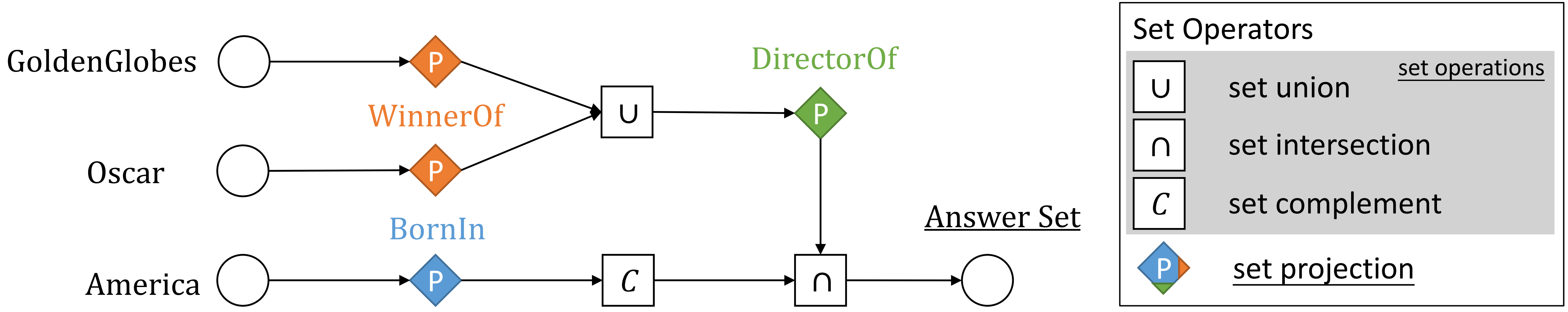


Ren H, Galkin M, Cochez M, Zhu Z, Leskovec J. Neural graph reasoning: Complex logical query answering meets graph databases. arXiv preprint arXiv:2303.14617. 2023 Mar 26.

# Complex Queries on Neuralized Knowledge Graphs

- A working example: Tree-Formed Queries (TFQ):
  - Tree-form query family contains the queries that can be converted into the computational tree

**Natural Language:** Find non-American directors whose movie won Golden Globes or Oscar?
**Logical Formula:** $q = V_? \exists V_1 . (\mathrm{Won}(V_1, \mathrm{GoldenGlobes}) \vee \mathrm{Won}(V_1, \mathrm{Oscar})) \wedge \neg \mathrm{BornIn}(V_?, \mathrm{America}) \wedge \mathrm{Direct}(V_?, V_1)$
**Set Operator Tree:** $\mathrm{DirectorOf}(\mathrm{WinnerOf}(\mathrm{GoldenGlobes}) \cup \mathrm{WinnerOf}(\mathrm{Oscar})) \cap \mathrm{BornIn}(\mathrm{America})^C$

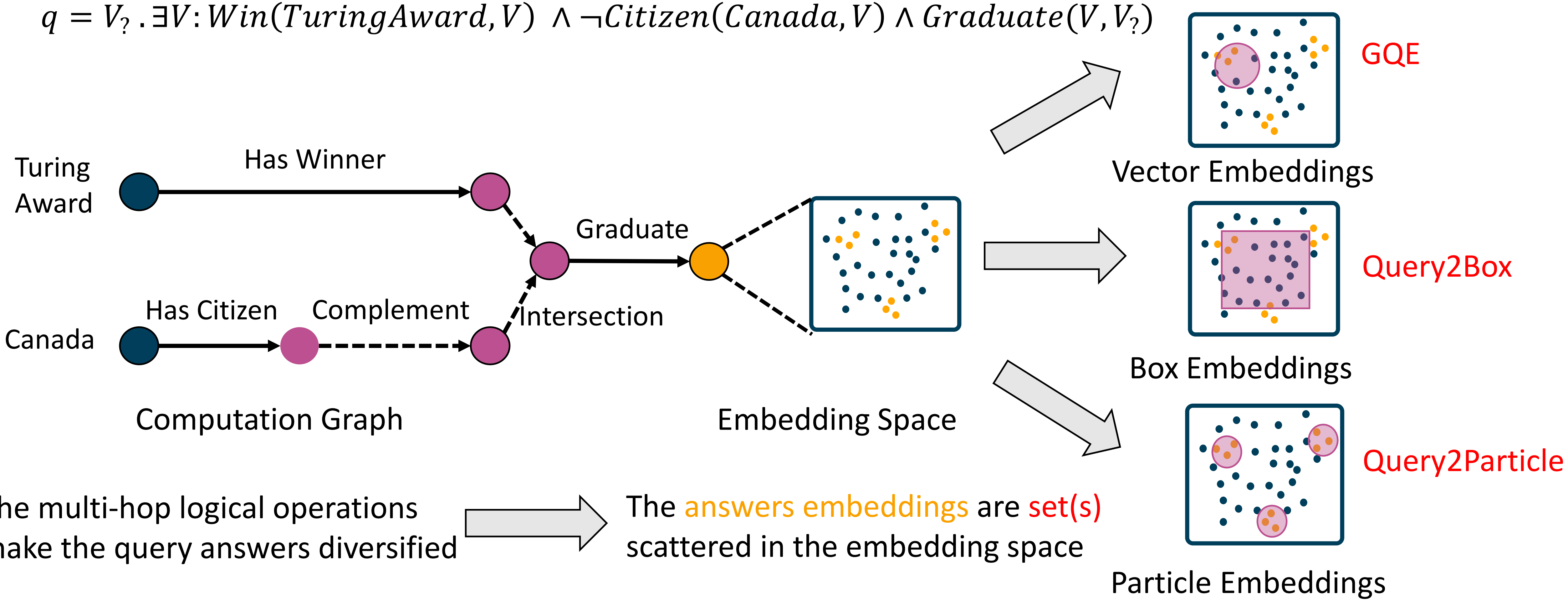





Example from: Zihao Wang, Weizhi Fei, Hang Yin, Yangqiu Song, Ginny Y Wong, and Simon See . Wasserstein-Fisher-Rao Embedding: Logical Query Embeddings with Local Comparison and Global Transport In Findings of ACL 2023

# Embedding Space and Set Representations

$$q = V_? \,.\, \exists V : Win(TuringAward, V) \land \neg Citizen(Canada, V) \land Graduate(V, V_?)$$



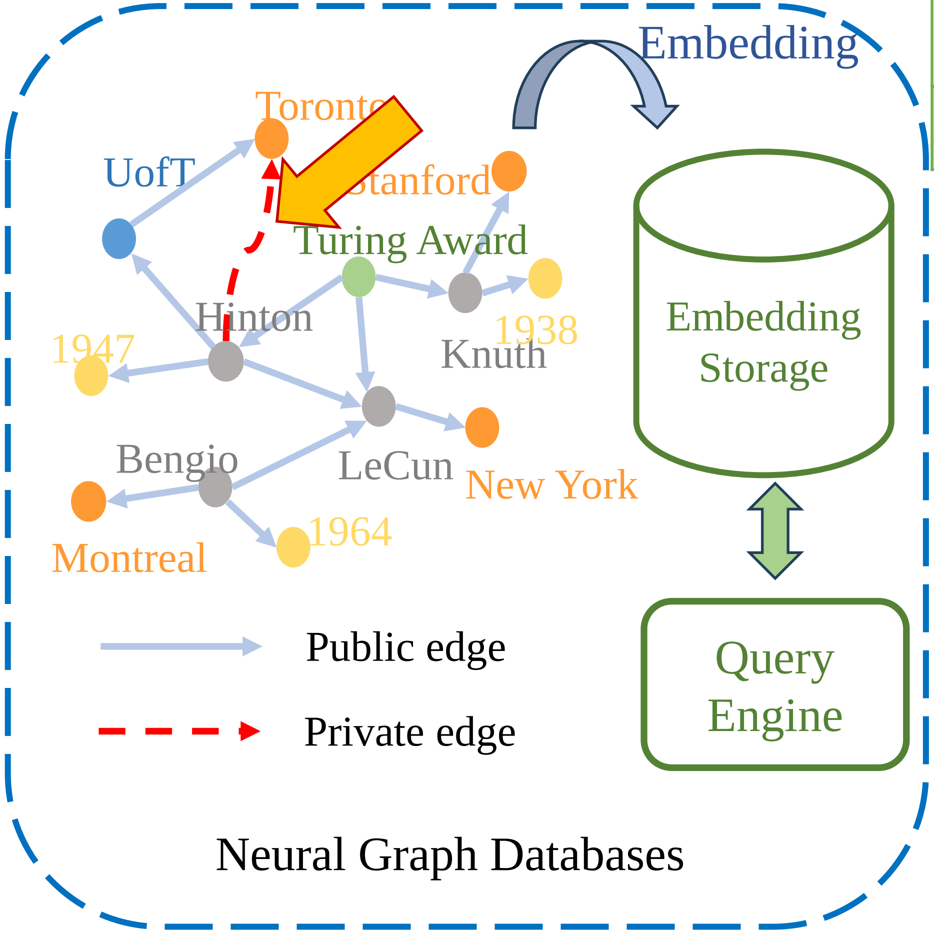

The multi-hop logical operations make the query answers diversified

The answers embeddings are set(s) scattered in the embedding space

William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, Jure Leskovec. Embedding Logical Queries on Knowledge Graphs. NeurIPS 2018.
Hongyu Ren, Weihua Hu, Jure Leskovec. Query2box: Reasoning over Knowledge Graphs in Vector Space using Box Embeddings. ICLR 2020.
Example from: Jiaxin Bai, Zihao Wang, Hongming Zhang, Yangqiu Song: Query2Particles: Knowledge Graph Reasoning with Particle Embeddings. NAACL-HLT (Findings) 2022.

# Privacy Issues in NGDBs

An attacker attempts to infer private information about Hinton's living place in the NGDBs. Attackers can leverage well-designed queries to retrieve desired privacy. The intersection of these queries can make a fair guess.



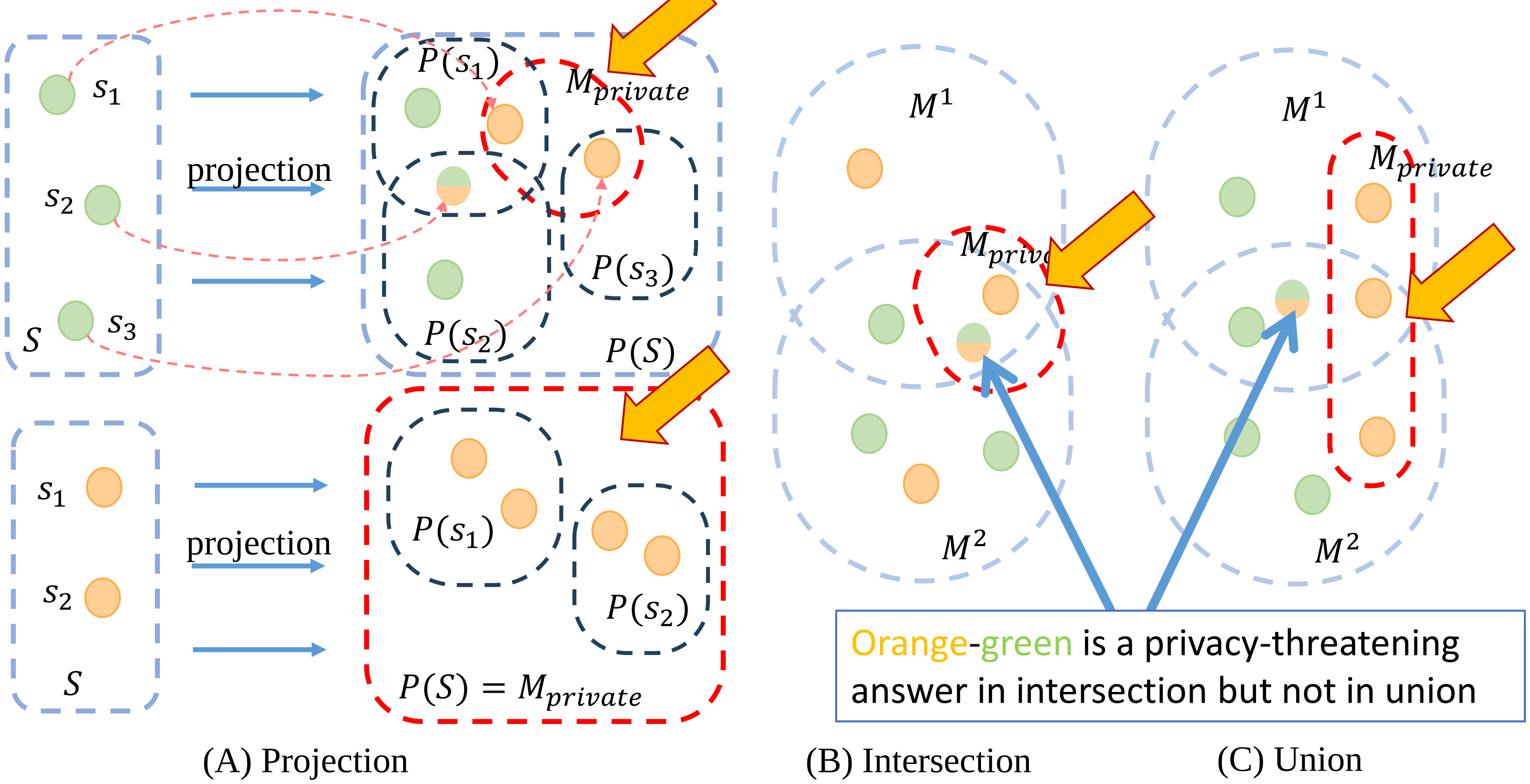

| | |
|---|---|
| **Query** | $q = V_?.\exists V: Win(V, Turing\ Award) \wedge BornIn(V, 1938) \wedge LiveIn(V, V_?)$ |
| **Interpretation** | Find where the Turing Award winner who was born in 1938 lived. |

Complex Queries

| Query | Answer |
|---|---|
| $q_1 = V_?.LiveIn(Hinton, V_?)$ | Privacy risk query detection |
| $q_2 = V_?.\exists X_1, X_2: Win(X_1, Turing\ Award) \wedge GreaterThan(X_2, 1940) \wedge BornIn(X_1, X_2) \wedge Livein(X_1, V_?)$ | Montreal, Toronto… |
| $q_3 = V_?.\exists X_1: CollabWith(LeCun, X_1) \wedge LiveIn(X_1, V_?)$ | Montreal, Toronto… |
| $q_4 = V_?.\exists X_1, X_2: Win(X_1, Turing\ Award) \wedge SmallerThan(X_2, 1950) \wedge BornIn(X_1, X_2) \wedge LiveIn(X_1, V_?)$ | Toronto, Stanford… |

Privacy Risk Queries

# Privacy-preserved NGDBs: Adversarial Training Examples




Green nodes denote non-private answers, orange nodes denote privacy-threatening answers, and orange-green nodes denote different privacy risks in subsets. Red dashed arrows denote privacy projection. The answers circled in red dashed line are at risk to leak privacy.



Privacy-Preserved Neural Graph Databases. Qi Hu, Haoran Li, Jiaxin Bai, Zihao Wang, Yangqiu Song. KDD 2024.

# Privacy-preserved NGDBs: Adversarial Training Examples

Query Encoding:

$$q_{i+1} = f_P(q_i, r), \quad r \in \mathcal{R} \cup \mathcal{A},$$
$$q_{i+1} = f_I(q_i^1, ..., q_i^n),$$
$$q_{i+1} = f_U(q_i^1, ..., q_i^n),$$

The query encoding procedure can be decomposed to sub-queries and finally to atomic queries.

Learning Objective: $L = L_u + \beta L_p$

$$L_u = -\frac{1}{N} \sum_{v \in \mathcal{M}^q_{\text{public}}} \log p(q, v),$$

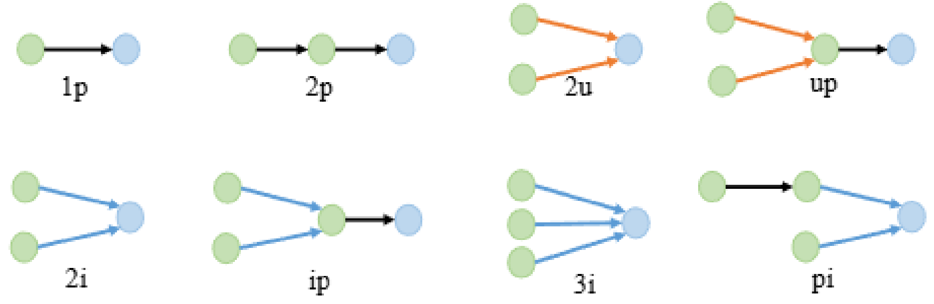
The original objective for public queries; increase the likelihood

$$L_p = \frac{1}{|\mathcal{A}_{\text{private}}|} \sum_{r(u,v) \in \mathcal{A}_{\text{private}}} \log p(f_p(e_v, r), u).$$

The privacy protection objective is to obfuscate private atomic queries; decrease the likelihood

# Privacy-preserved NGDBs: Experiments

- Multi-relational knowledge graphs with numerical attributes
  - Attribute value projections can be the same as traditional relation projection if the values themselves are entities, e.g., locations
  - Attributes and their values are more aligned with real-world privacy considerations
  - Attribute values are vulnerable to be attacked as we can use group queries to attack individual's information, which has been widely used as an illustration in differential privacy




Query Type

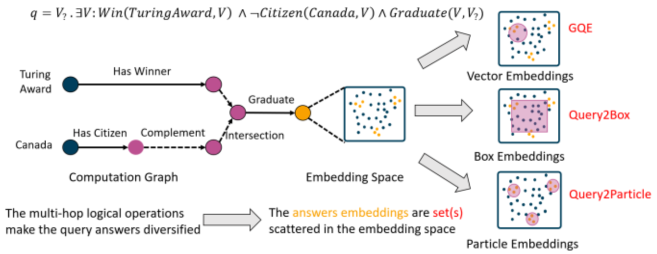
$$V_?.\exists X_1, X_2: Win(X_1, Turing\ Award) \wedge GreaterThan(X_2, 1940) \wedge BornIn(X_1, X_2) \wedge Livein(X_1, V_?)$$

| Graphs | Data Split | #Nodes | #Edges | #Pri. Edges |
|---|---|---|---|---|
| FB15k-N | Training | 22,964 | 1,037,480 | 8,000 |
| | Validation | 24,021 | 1,087,296 | |
| | Testing | 27,144 | 1,144,506 | |
| DB15k-N | Training | 27,639 | 340,936 | 6,000 |
| | Validation | 29,859 | 381,090 | |
| | Testing | 36,358 | 452,348 | |
| YAGO15k-N | Training | 30,351 | 383,772 | 1,600 |
| | Validation | 31,543 | 417,356 | |
| | Testing | 33,610 | 453,688 | |

# Privacy-preserved NGDBs: Experiments

| Dataset | Encoding | Model | Public HR@3 | Public MRR | Private HR@3 | Private MRR |
|---|---|---|---|---|---|---|
| FB15k-N | GQE | Baseline | **21.99** | **20.26** | 28.99 | 27.82 |
| | | Noise | 15.89 | 14.67 | 21.54 | 21.37 |
| | | P-NGDB | 15.92 | 14.73 | **10.77** | **10.21** |
| | Q2B | Baseline | **18.70** | **16.88** | 30.28 | 28.98 |
| | | Noise | 12.34 | 12.19 | 20.01 | 19.71 |
| | | P-NGDB | 12.28 | 11.18 | **10.17** | **9.38** |
| | Q2P | Baseline | **26.45** | **24.48** | 29.08 | 31.85 |
| | | Noise | 20.13 | 19.77 | 22.35 | 23.17 |
| | | P-NGDB | 19.48 | 18.19 | **14.15** | **14.93** |
| DB15k-N | GQE | Baseline | **24.16** | **22.37** | 39.26 | 37.25 |
| | | Noise | 18.01 | 16.35 | 28.59 | 28.37 |
| | | P-NGDB | 17.58 | 16.29 | **10.52** | **10.79** |
| | Q2B | Baseline | **15.94** | **14.98** | 42.19 | 39.78 |
| | | Noise | 10.76 | 10.28 | 26.49 | 25.93 |
| | | P-NGDB | 10.19 | 9.49 | **8.92** | **7.99** |
| | Q2P | Baseline | **25.72** | **24.12** | 46.18 | 43.48 |
| | | Noise | 19.89 | 19.32 | 33.56 | 33.17 |
| | | P-NGDB | 20.26 | 19.00 | **19.38** | **18.45** |
| YAGO15k-N | GQE | Baseline | **26.06** | **24.37** | 43.55 | 40.81 |
| | | Noise | 20.32 | 20.27 | 38.52 | 38.29 |
| | | P-NGDB | 19.58 | 19.82 | **7.56** | **7.33** |
| | Q2B | Baseline | **23.39** | **22.53** | 42.73 | 40.55 |
| | | Noise | 16.85 | 15.37 | 28.23 | 28.54 |
| | | P-NGDB | 17.07 | 16.03 | **6.26** | **5.79** |
| | Q2P | Baseline | **29.41** | **27.87** | 42.56 | 45.79 |
| | | Noise | 22.85 | 21.21 | 34.26 | 33.68 |
| | | P-NGDB | 23.27 | 22.59 | **7.34** | **7.17** |

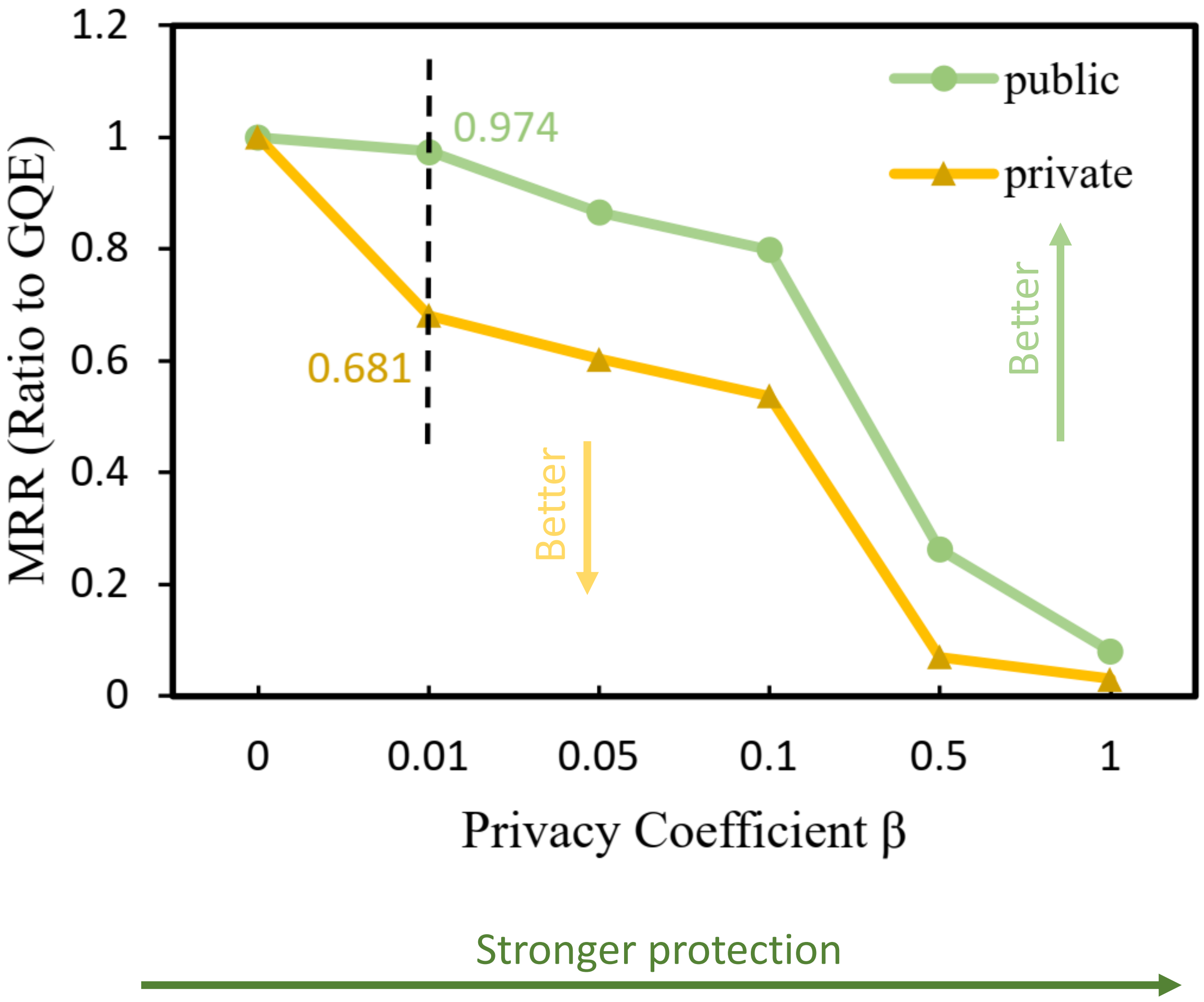
Three commonly used query encoding methods




The protection methods hurt the retrieval quality on public sets, but to make fair comparison, we tune the parameter to get similar performance

P-NGDB's retrieval performance on private sets drops more significantly denotes better privacy protection

# Privacy-preserved NGDBs: Experiments



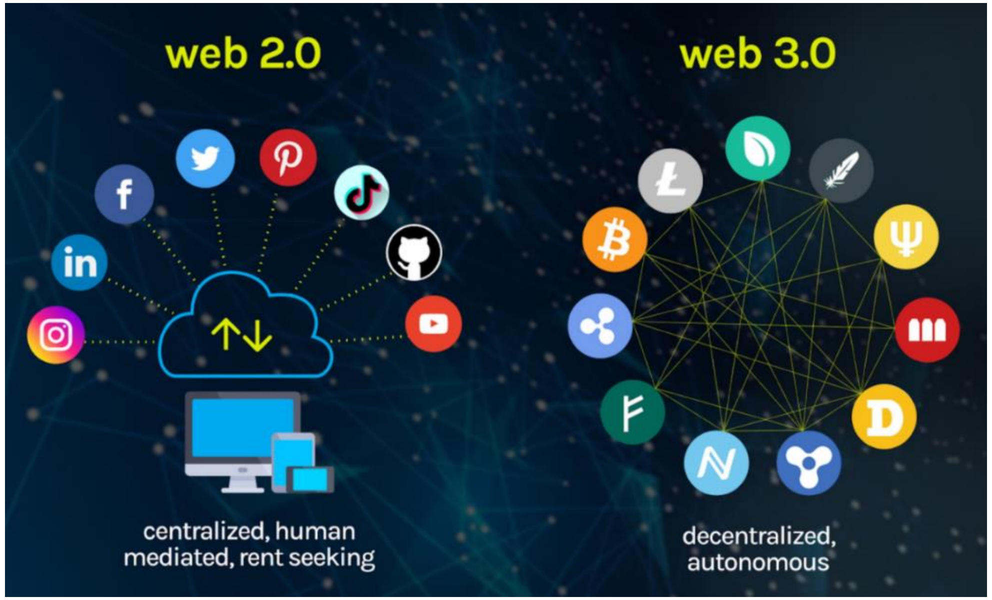

$$L = L_u + \beta L_p$$

There is a tradeoff between retrieval performance and privacy protection.

We can select suitable privacy coefficients $\beta$ according to the task.

# An Outlook

- From Web2.0 to Web3.0
  - Decentralized data: users own their (neural) knowledge bases/graphs
    - Monetarize by users' data and time
  - Permissionless, trustless, but accessible to users' owned knowledge or data

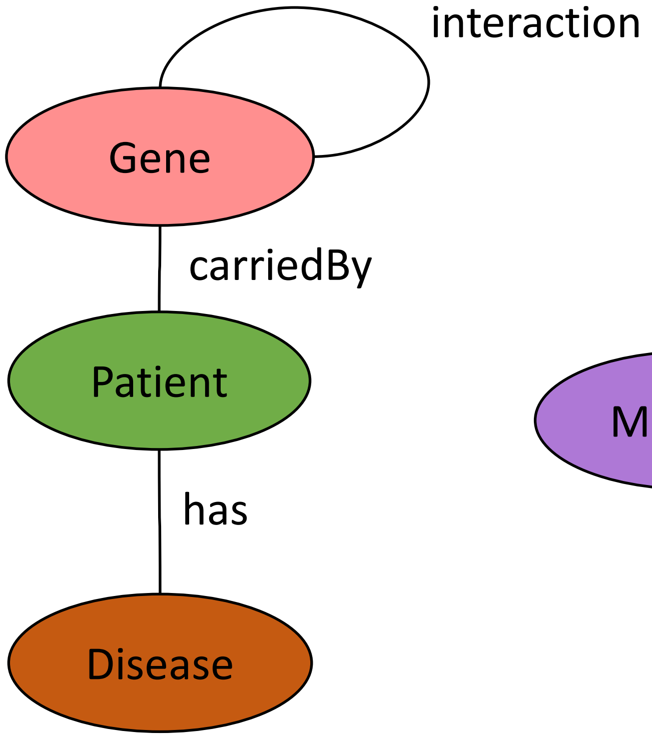



- Security and privacy of data and knowledge is the key!

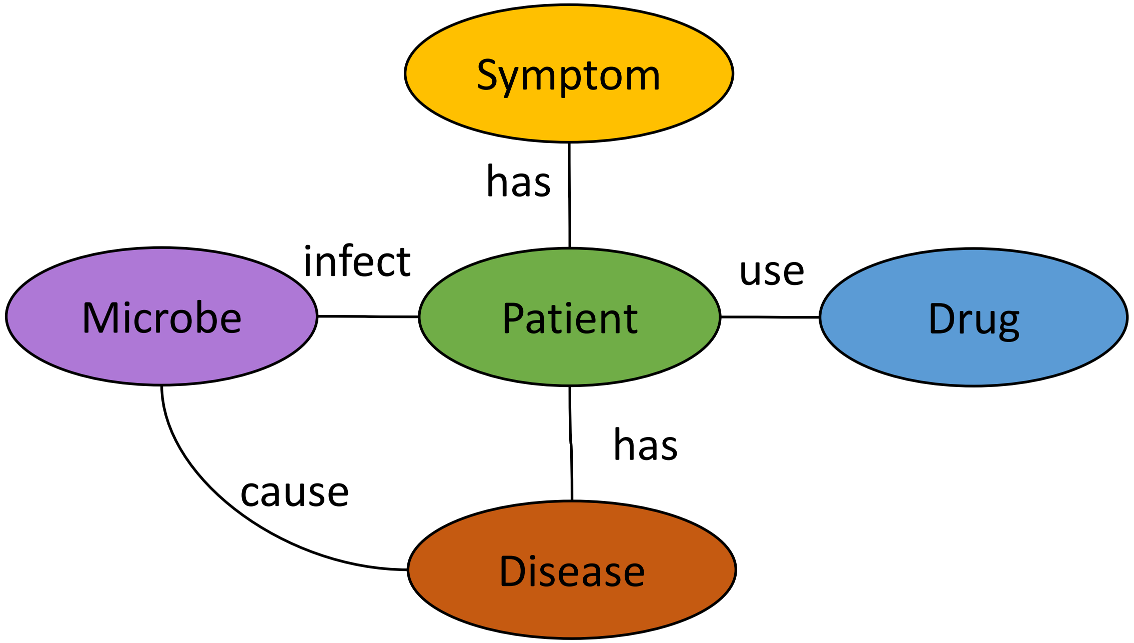


Figure from: https://vitalflux.com/what-is-web3-0-features-design-skills-nfts/

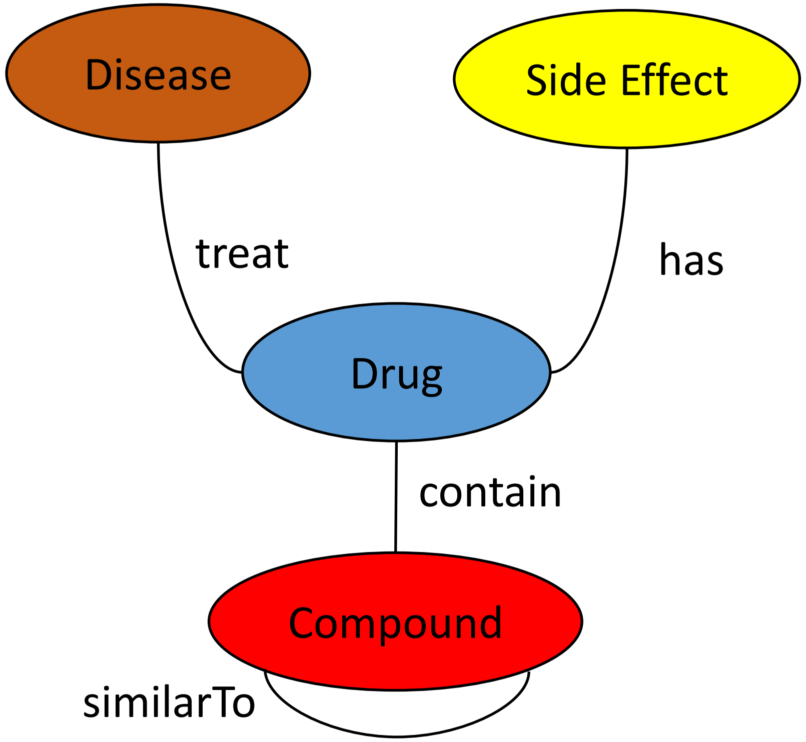
# Knowledge Sharing




KG 1 from a gene engineering company




KG 2 from a hospital

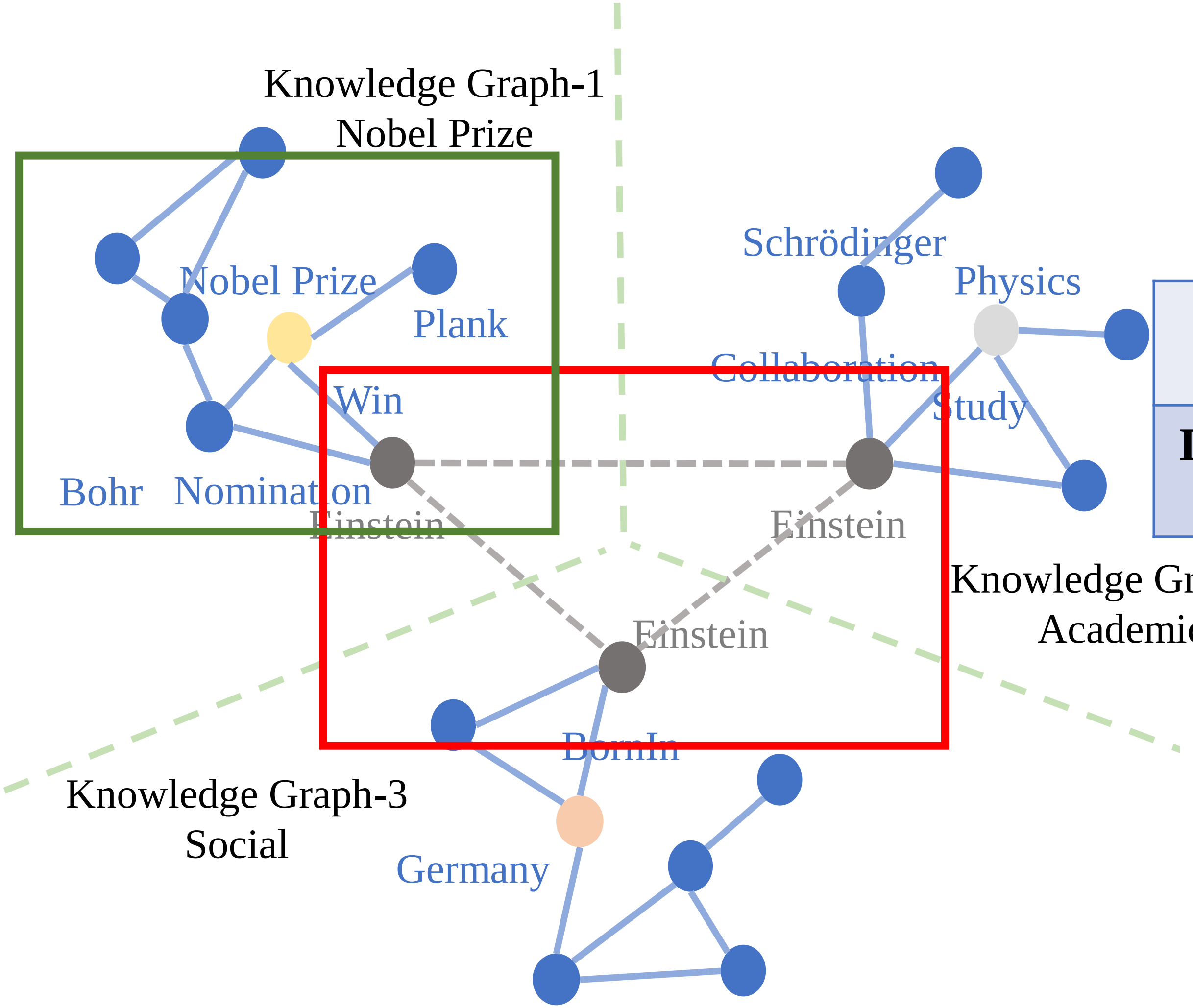



KG 3 from a pharmaceutical company

# Knowledge Sharing

- Each party has its private part of data, which cannot be disclosed to others
  - Patient information
  - Drag chemical compound
  - Personal gene expressions
- Even if privacy is not a concern, they would not expose their knowledge to other companies except they can also benefit from others
  - Existing drug repurposing failure cases

# Types of Queries for Knowledge Sharing




**Definition 3.1** (Cross-graph Query). A complex query $q$ is a cross-graph query if there exists query answers $V_? \in \mathcal{V}$, such that there are $V_1, \cdots, V_k \in \mathcal{V}$ in the knowledge graph that can satisfy the given logical expressions and the atomic expressions in the query can not be found in a single knowledge graph.

| | |
|---|---|
| **Query** | $q = V_?.\exists V: Win(V, Nobel\ Prize) \wedge BornIn(V, Germany) \wedge Study(V, V_?)$ |
| **Interpretation** | Find what research topics which Nobel Prize winner who was born in Germany studied. |

**Definition 3.2** (In-graph Query). A complex query $q$ is an in-graph query if for all answers $V_? \in \mathcal{V}$ to the query, such that there are $V_1, \cdots, V_k \in \mathcal{V}$ in the knowledge graph that can satisfy the given logical expressions and the atomic expressions in the query are from a single knowledge graph.

May be solved by previous work



FedCQA: Answering Complex Queries on Multi-Source Knowledge Graphs via Federated Learning Qi Hu, Weifeng Jiang, Haoran Li, Zihao Wang, Jiaxin Bai, Qianren Mao, Yangqiu Song, Lixin Fan, Jianxin Li. Arxiv, 2024.

# Federated Graph Machine Learning

- Horizontal federated learning
  - Node embeddings should be aligned
    - Very unlikely
- Vertical federated learning
  - Nodes should be partially aligned
    - Possible but sometimes unlikely
  - Aligned nodes are in different embedding space but features are not complementary
- Federated transfer learning
  - Nodes and their embeddings are aligned
    - Possible
  - Nodes and their embeddings are not aligned
    - Likely

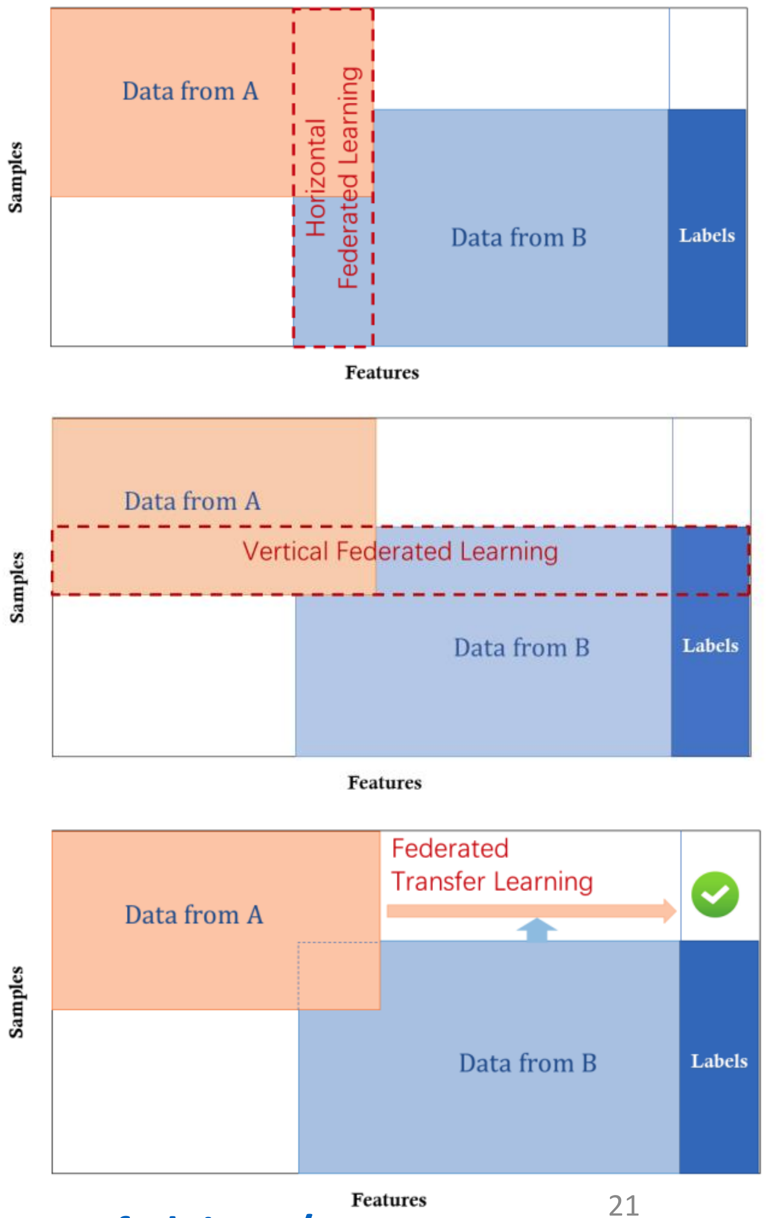
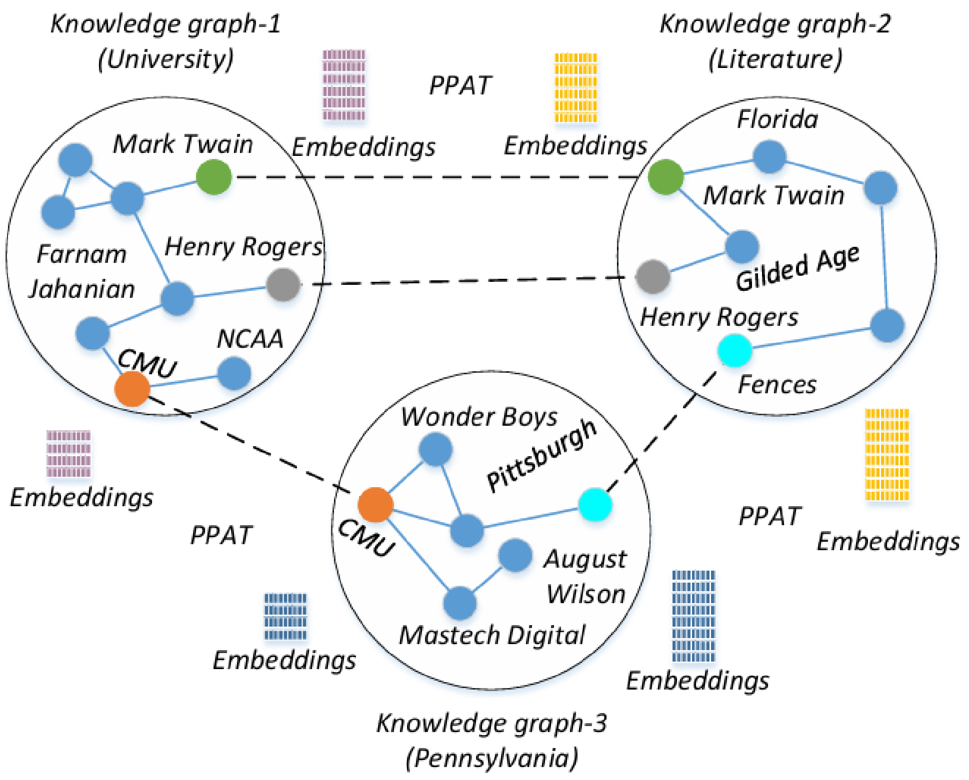



Figure credit: WeBank Tutorial, Chapter 1 - Introduction to Federated Learning. [https://www.fedai.org/](https://www.fedai.org/)

# Existing methods: Federated Knowledge Graph Embedding

- Learning a low-dimensional representation of a knowledge graph's entities and relations while preserving their semantic meaning.

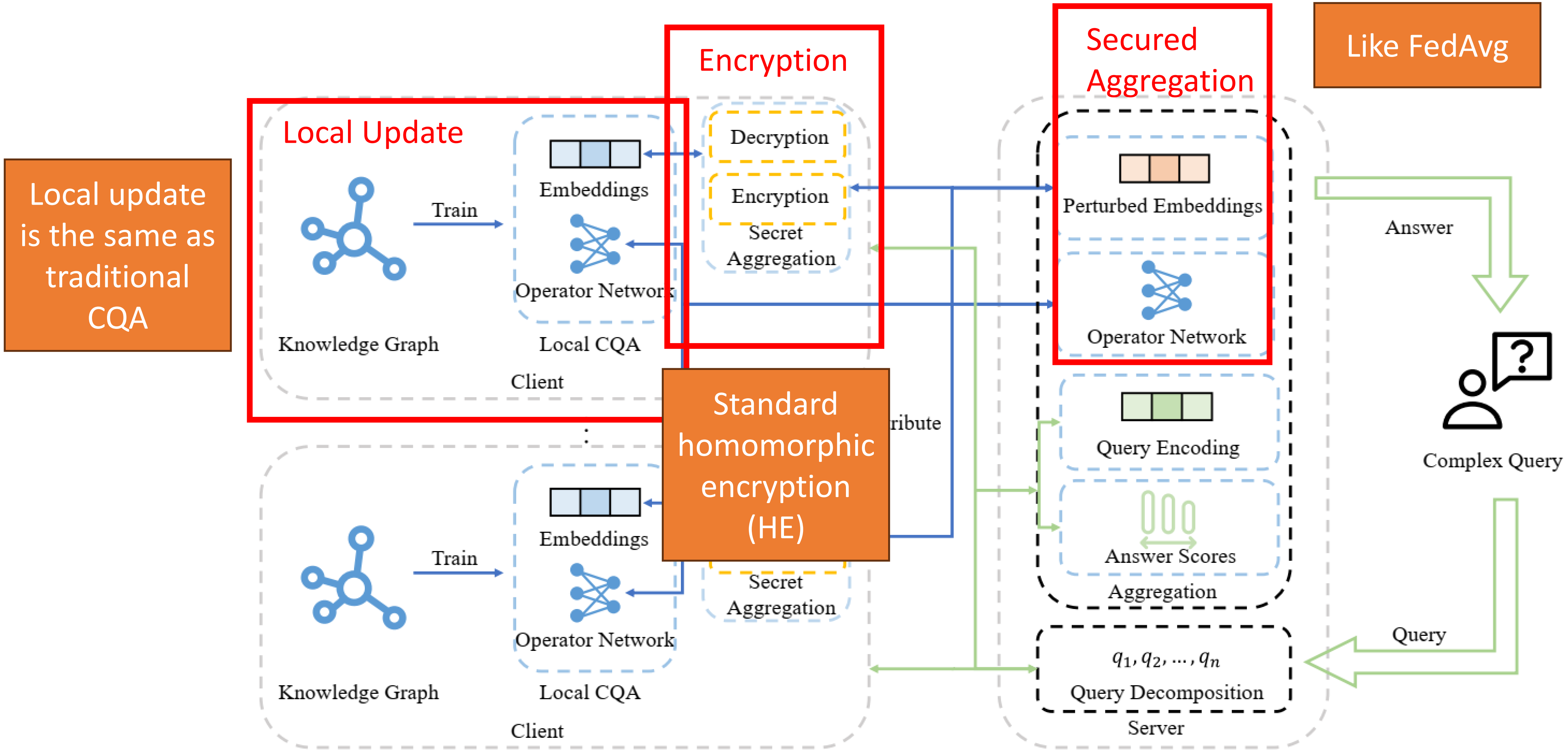



Limitation: Only focus on one-hop relations and cannot support complex queries on the learned graph systems.

Federated Knowledge Graph Embedding (Figure taken from Peng et al)

Hao Peng, Haoran Li, Yangqiu Song, Vincent Zheng, and Jianxin Li. 2021. Differentially private federated knowledge graphs embedding. In CIKM 2021.

# Federated NGDBs – Training

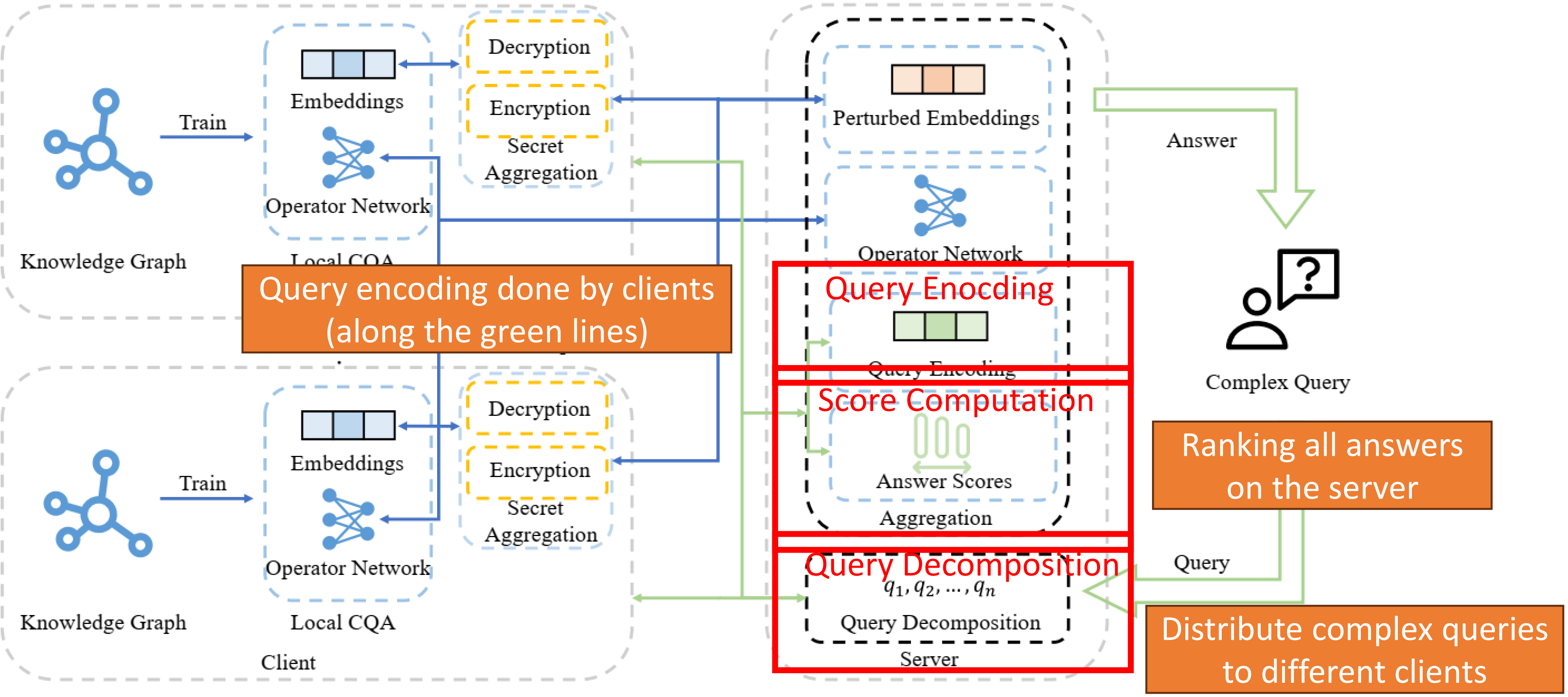



The blue line denotes the training process, and the green line denotes the retrieval process.



FedCQA: Answering Complex Queries on Multi-Source Knowledge Graphs via Federated Learning. Qi Hu, Weifeng Jiang, Haoran Li, Zihao Wang, Jiaxin Bai, Qianren Mao, Yangqiu Song, Lixin Fan, Jianxin Li. Arxiv 2024.

# Federated NGDBs – Inference (Queries)




The blue line denotes the training process, and the green line denotes the retrieval process.



FedCQA: Answering Complex Queries on Multi-Source Knowledge Graphs via Federated Learning. Qi Hu, Weifeng Jiang, Haoran Li, Zihao Wang, Jiaxin Bai, Qianren Mao, Yangqiu Song, Lixin Fan, Jianxin Li. Arxiv 2024.

# Federated NGDBs - Experiments

KnowComp

Understanding the World by Computational Knowledge

Split according relations

| Graphs | #Clients | #Nodes | #Relations | #Edges |
|---|---|---|---|---|
| FB15k-237 | 3 | 13,651 | 79 | 103,359 |
| | 5 | 12,639 | 47.4 | 62,015 |
| FB15k | 3 | 14,690 | 448.3 | 197,404 |
| | 5 | 14,279 | 269 | 118,442 |
| NELL995 | 3 | 40,204 | 66.7 | 47,601 |
| | 5 | 28,879 | 40 | 28,560 |

Statistics of Knowledge Graphs

Sampled in-graph queries Evaluation

| Graphs | #C | In-graph Train. | In-graph Valid. | In-graph Test. | Cross-graph Test. |
|---|---|---|---|---|---|
| FB15k-237 | 3 | 317,226 | 11,528 | 11,539 | 32,573 |
| | 5 | 180,552 | 6,619 | 6,673 | 31,469 |
| FB15k | 3 | 592,573 | 19,206 | 19,267 | 53,660 |
| | 5 | 344,418 | 11,409 | 11,437 | 53,154 |
| NELL995 | 3 | 208,070 | 8,810 | 8,750 | 24,954 |
| | 5 | 117,231 | 5,177 | 5,118 | 24,237 |

Statistics of Sampled Queries

# Federated NGDBs - Experiments

We evaluate the performance change on in- & cross- graph queries

Our FedCQA perform well on all datasets well maintaining good properties of both FedE and FedR

FedE performs well but has to share embeddings to the server

FedR secured entities for local clients but cannot support cross-graph queries

| Graph | Setting | In-graph | | Cross-graph | | In-graph | | Cross-graph | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | HR@3 | MRR | HR@3 | MRR | HR@3 | MRR | HR@3 | MRR | | | | |
| | Local | 12.64 | 12.03 | - | - | 14.55 | 13.63 | - | - | 13.32 | 12.73 | - | - |
| | Central | 13.13 | 12.39 | 13.03 | 12.28 | 14.93 | 14.66 | 15.02 | 14.81 | 13.28 | 12.61 | 13.36 | 12.91 |
| | FedE | **13.72** | **13.23** | **12.74** | **11.63** | 14.82 | 14.27 | 14.79 | 13.93 | 13.12 | 12.23 | **12.62** | **12.08** |
| | FedR | 12.89 | 11.98 | - | - | 14.32 | 14.23 | - | - | **13.92** | **12.92** | - | - |
| | FedCQA | 13.54 | 12.43 | 12.63 | 11.32 | **15.32** | **14.32** | **14.83** | **14.11** | 12.93 | 12.11 | 12.55 | 11.96 |
| FB15k | Local | 22.05 | 18.21 | - | - | 24.32 | 22.64 | - | - | 22.87 | 20.51 | - | - |
| | Central | 29.53 | 25.65 | 30.21 | 25.33 | 38.62 | 34.14 | 38.03 | 34.36 | 38.87 | 35.86 | 37.97 | 36.13 |
| | FedE | 24.31 | 26.74 | **27.95** | **25.21** | 43.68 | 39.62 | 39.72 | 35.95 | 34.27 | 30.18 | **31.19** | 26.03 |
| | FedR | 20.29 | 18.61 | - | - | 25.32 | 22.71 | - | - | 23.64 | 20.97 | - | - |
| | FedCQA | **25.63** | **26.87** | 24.77 | 25.17 | **44.02** | **39.27** | **40.27** | **36.31** | **34.85** | **33.83** | 31.80 | **28.99** |
| | Local | 11.85 | 11.03 | - | - | 15.86 | 13.02 | - | - | 13.85 | 13.85 | 12.94 | - |
| | Central | 12.87 | 11.95 | 13.06 | 12.46 | 16.74 | 14.82 | 16.42 | 15.63 | 15.41 | 14.23 | 16.27 | 15.83 |
| | FedE | 13.29 | 12.72 | 12.46 | 11.82 | **17.23** | 14.12 | **16.28** | 14.01 | 14.27 | 13.81 | 14.18 | 13.71 |
| | FedR | 12.01 | 11.23 | - | - | 16.04 | 13.26 | - | - | 12.48 | 11.67 | - | - |
| | FedCQA | **14.21** | **13.27** | **13.76** | **12.67** | 16.62 | **15.28** | 16.27 | **16.23** | **16.28** | **15.38** | **16.09** | **15.27** |

Table: The retrieval performance of distributed knowledge graph complex query answering models when there are 3 clients

# Federated NGDBs – More Clients

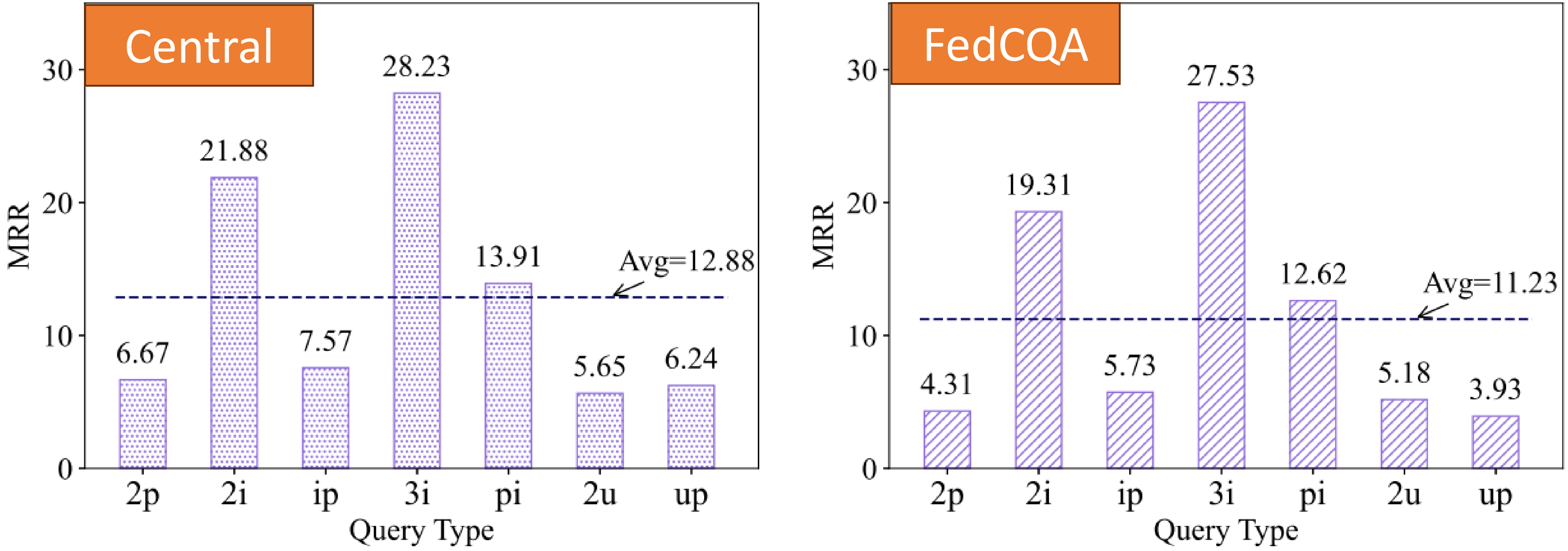
Improve performance on both in- & cross-graph queries.

| Graph | Setting | FB15k-237 | | | | FB15k | | | | NELL995 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | In-graph | | Cross-graph | | In-graph | | Cross-graph | | In-graph | | Cross-graph | |
| | | HR@3 | MRR | HR@3 | MRR | HR@3 | MRR | HR@3 | MRR | HR@3 | MRR | HR@3 | MRR |
| GQE | Local | 11.44 | 10.65 | - | - | 14.65 | 13.8 | - | - | 11.23 | 10.37 | - | - |
| | FedCQA | **12.42** | **11.60** | **11.20** | **10.79** | **16.13** | **15.78** | **15.28** | **14.91** | **12.48** | **11.91** | **11.49** | **11.02** |
| Q2P | Local | 19.83 | 17.51 | - | - | 36.10 | 35.04 | - | - | 20.03 | 18.62 | - | - |
| | FedCQA | **21.40** | **20.83** | **20.71** | **19.94** | **40.81** | **37.96** | **38.56** | **35.73** | **24.59** | **23.75** | **23.85** | **22.90** |
| Tree-LSTM | Local | 10.48 | 10.09 | - | - | 15.26 | 14.37 | - | - | 14.52 | 13.89 | - | - |
| | FedCQA | **13.79** | **13.27** | **12.74** | **12.18** | **15.44** | **15.81** | **15.28** | **14.24** | **15.68** | **14.28** | **14.57** | **12.89** |

Table: The retrieval performance of distributed knowledge graph complex query answering models when there are 5 clients

# Federated NGDBs – Compared with Central Training

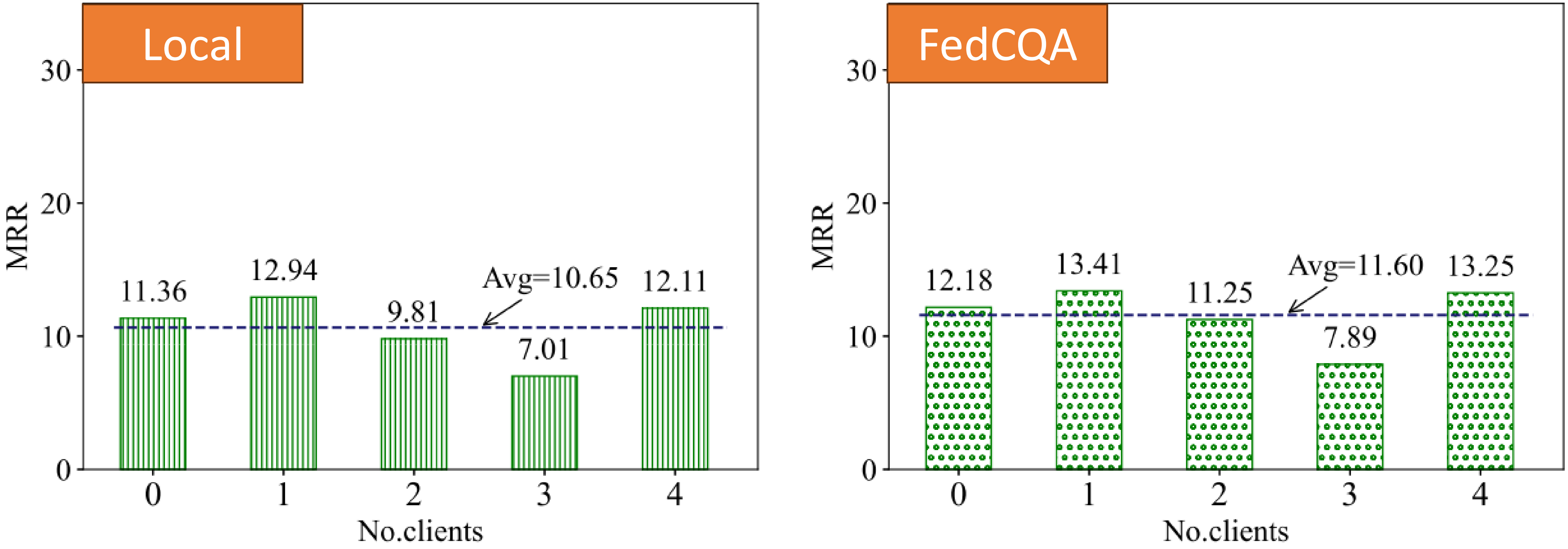
Different query types, the retrieval performance close to central training.




Only evaluated on cross-graph queries

# Federated NGDBs – Compared with Local Training

For clients, all participants can benefit from FedCQA training.




Only evaluated on in-graph queries

# Federated NGDBs – More Results

| Setting | FedE | FedR | FedCQA |
|---|---|---|---|
| Relative Rounds to FedE | 1.00 | 1.32 | 1.09 |

Table: Communication Rounds.

For convergence speed, FedCQA is slower than FedE but faster than FedR

| Setting | FB15k-237 | FB15k | NELL995 |
|---|---|---|---|
| Local | 8.46 | 13.04 | 7.83 |
| FedCQA | 10.17 | 14.98 | 9.17 |

Table: More Clients (10), in MRR

For more clients, our FedCQA is still useful

| Setting | FB15k-237 | FB15k | NELL995 |
|---|---|---|---|
| Local | 10.22 | 20.21 | 9.64 |
| FedCQA | 11.42 | 22.47 | 11.36 |

Table: Overlapped relations, in MRR

When there are relations overlapped, our FedCQA is still useful

# Conclusions

- The combination of LLMs and KGs (or NGDBs) is a promising direction
  - Retrieval augmented generation
  - Co-training
- NGDBs brings better retrieval performance (for open-world assumptions) while introducing novel privacy risks
- Privacy in NGDBs needs further explored
  - Inherent Privacy: we proposed privacy preserved NGDBs
  - Distributed Learning: we proposed federated NGDBs

# Thank you for your attention ☺

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

THE DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

KnowComp Group
Understanding the World by Computational Knowledge