CCKS 2020 • 知识图谱前沿趋势

知识图谱数据管理研究新进展

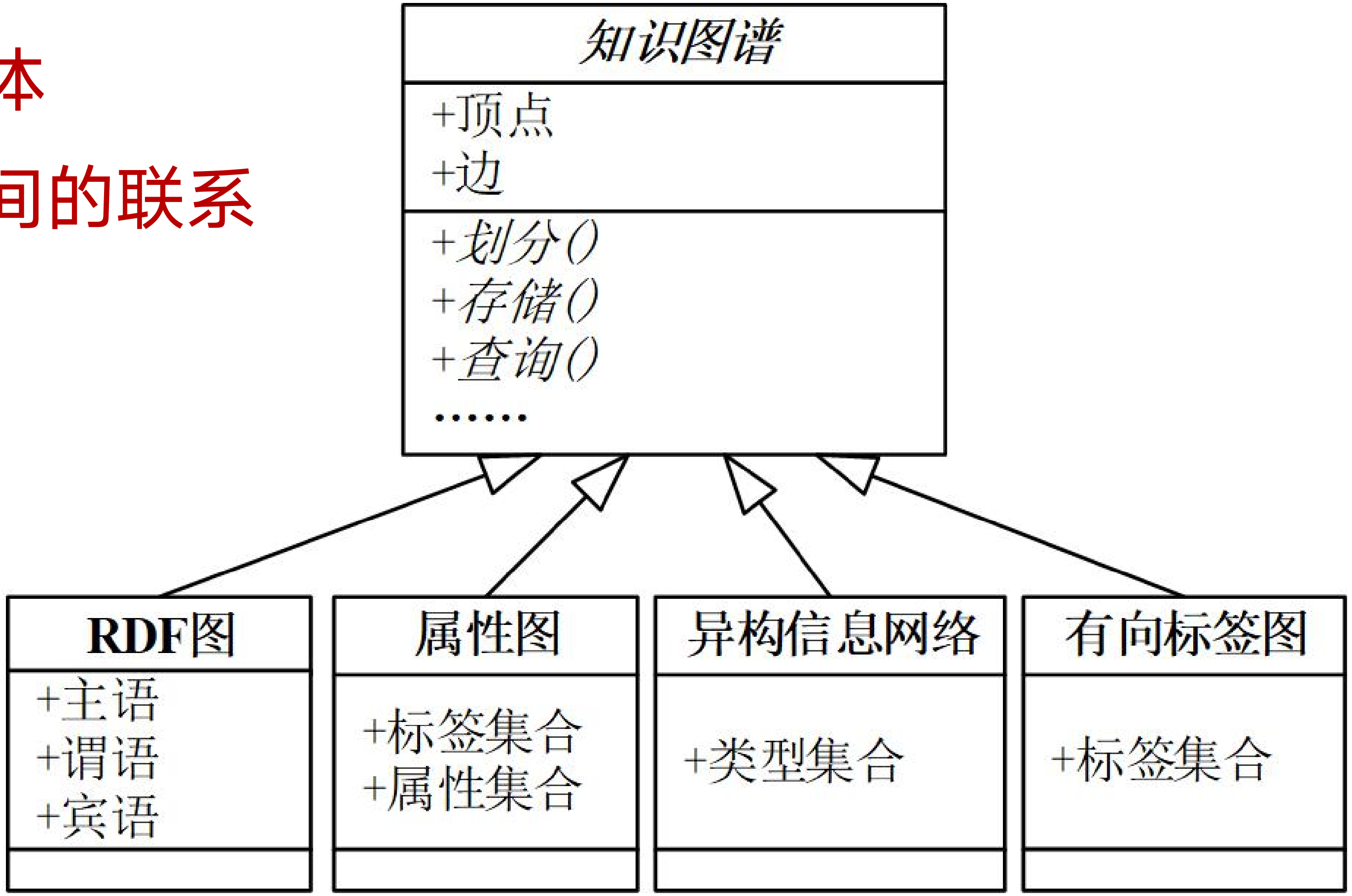
天津大学 智能与计算学部 人工智能学院

王　鑫

wangx@tju.edu.cn

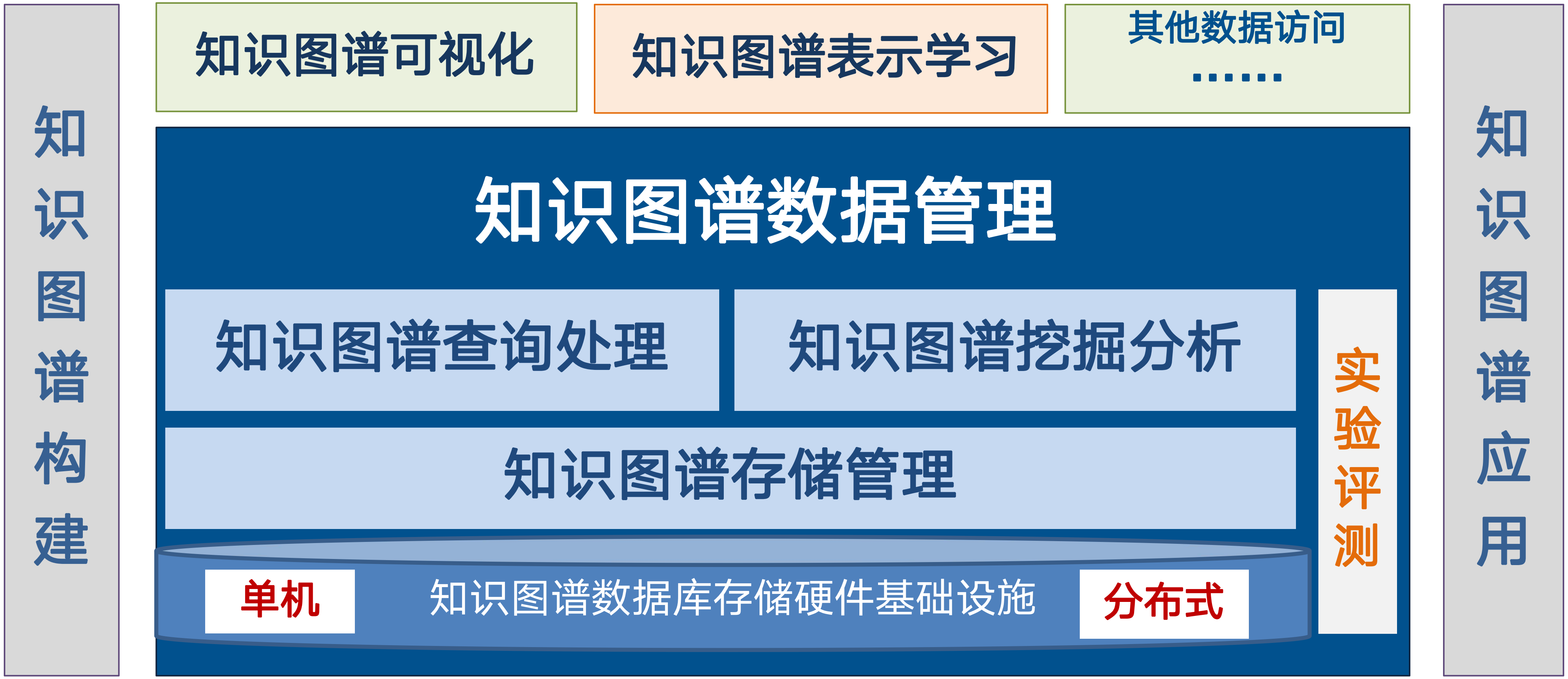
2020年11月14日 • 南昌

- **知识图谱**目前并没有统一的严格定义
  - 图$G = (V, E)$的某种扩展形式
    - V顶点集合，表示实体
    - E边集合，表示实体间的联系
  - RDF图
  - 属性图
  - 异构信息网络
  - 有向标签图



天津大学 Tianjin University

■ 知识图谱数据管理系统的构成



- 王鑫, 邹磊, 王朝坤, 等. 知识图谱数据管理研究综述. 软件学报. 2019.

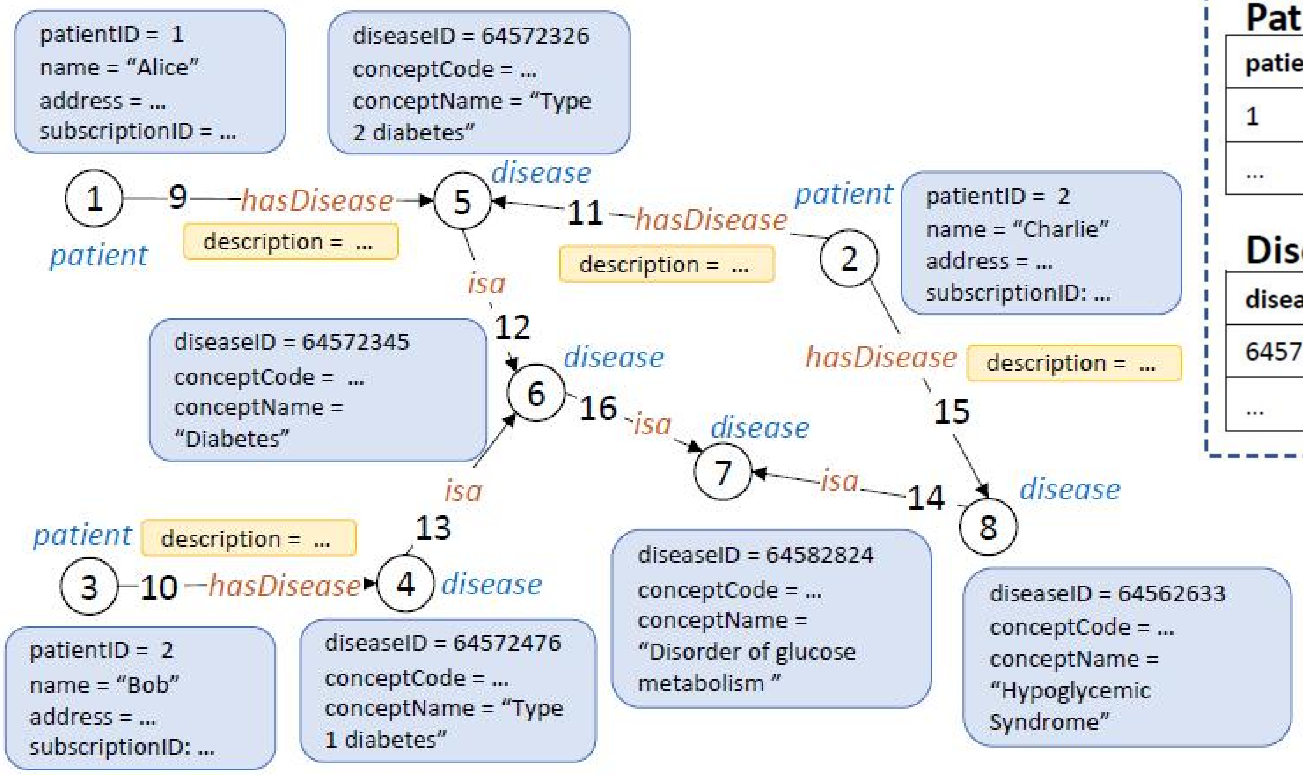
Tianjin University 天津大学

IBM Db2 Graph

Synergistic with other analytics
Retrofittable to existing data

In-DBMS graph query approach

Graph overlay approach to expose a graph view of the relational data



Patient Table

patientID	name	address	subscriptionID
1	Alice	...	115
...			...

HasDisease Table

patientID	diseaseID	description
1	64572326	...
...	...	...

Disease Table

diseaseID	conceptCode	conceptName
64572326	44054006	“Type 2 diabetes”
...	...	

DiseaseOntology Table

sourceID	targetID	type
64572326	73211009	“isa”
...	...	

```
"v_tables": [
{
    "table_name": "Patient",
    "prefixed_id": true,
    "id": "'patient'::patientID",
    "fix_label": true,
    "label": "'patient'",
    "properties": ["patientID", "name", "address", "
    subscriptionID"]
},
```

• Tian, et al. IBM Db2 Graph: Supporting Synergistic and Retrofittable Graph Queries Inside IBM Db2. SIGMOD 2020.

天津大学 Tianjin University

Microsoft A1

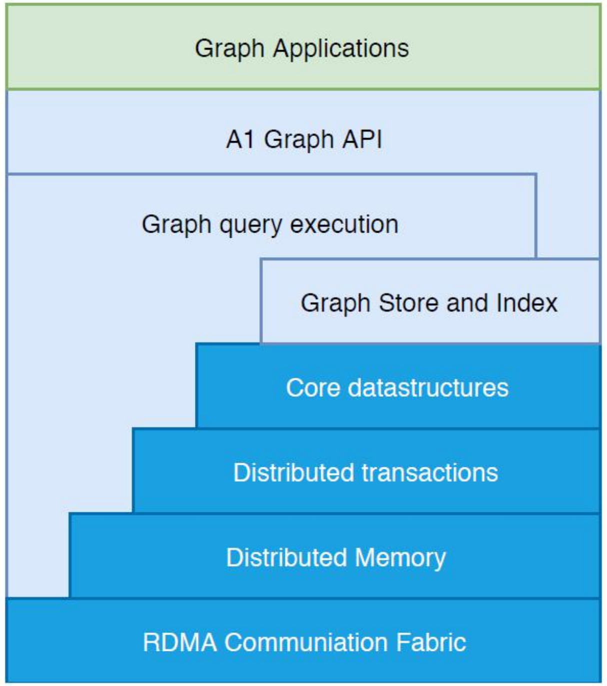
- Distributed In-Memory Graph Database used by the Bing search engine

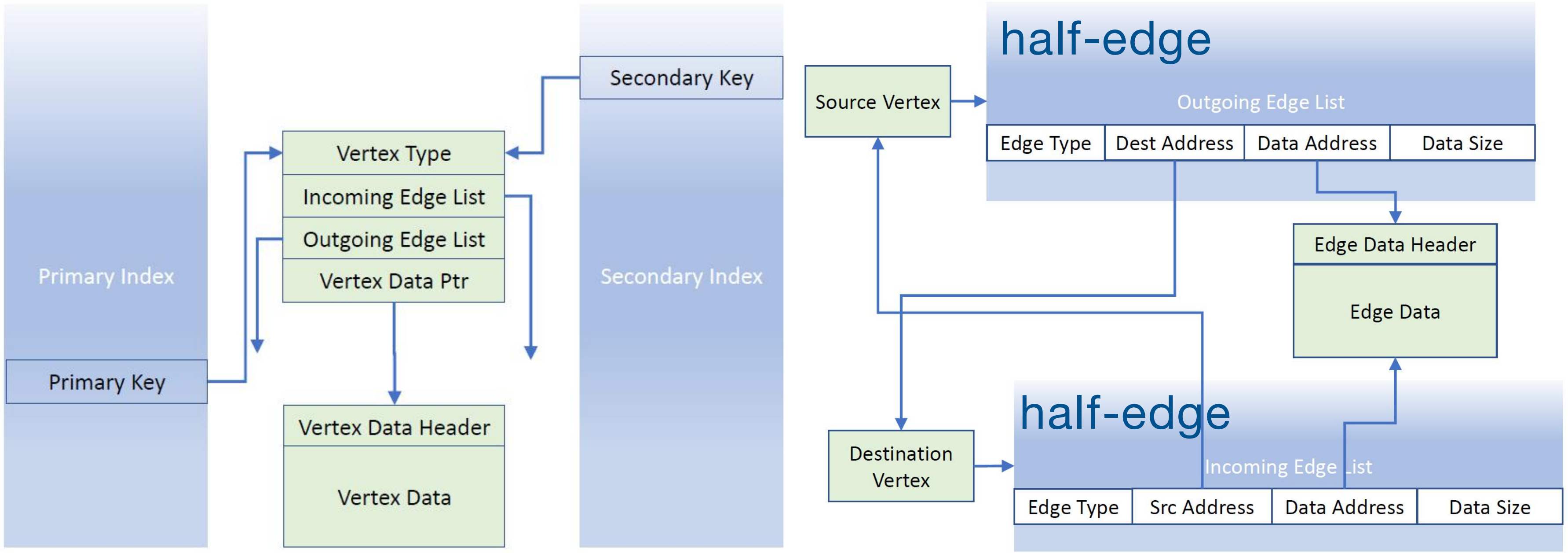
Availability of cheap DRAM

High speed RDMA (Remote Direct Memory Access)

- Store tens of billions of vertices and edges
- Throughput of 350+ million of vertex reads per second

Query latency in single digit milliseconds





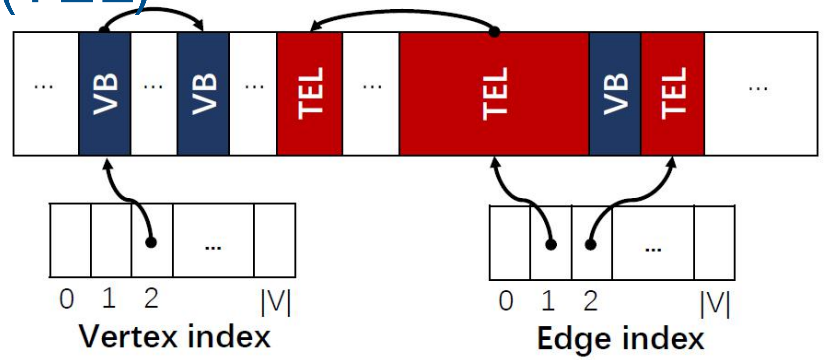
- Buragohain, et al. A1: A Distributed In-Memory Graph Database. SIGMOD 2020.

LiveGraph

- Graph transactional system
- Real-time graph analytics

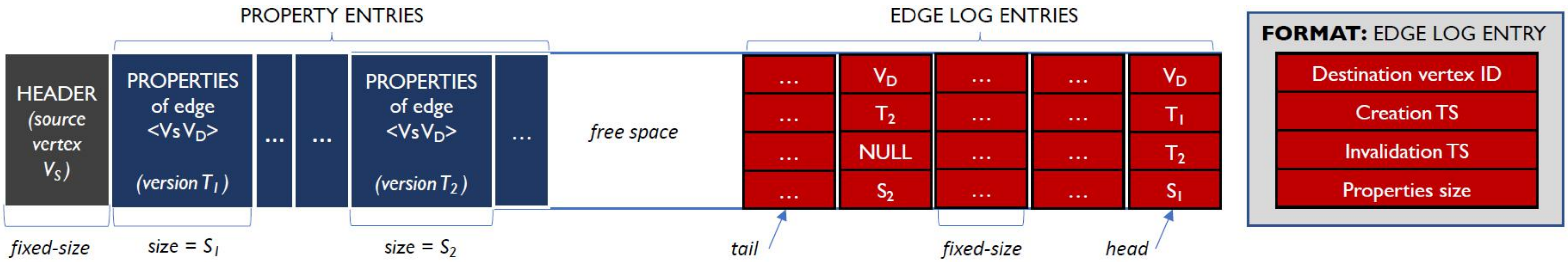
- A Transactional Graph Storage System with Purely Sequential Adjacency List Scans

- Transactional Edge Log (TEL)



- Purely sequential, yet mutable, edge storage

Cost	Seek		Scan (per edge)
B+ Tree	$O(\log N)$	random	sequential w. random
LSMT	$O(\log N)$	random	sequential w. random
Linked List	$O(1)$	random	random
CSR	$O(1)$	random	sequential
TEL	$O(1)$	random	sequential

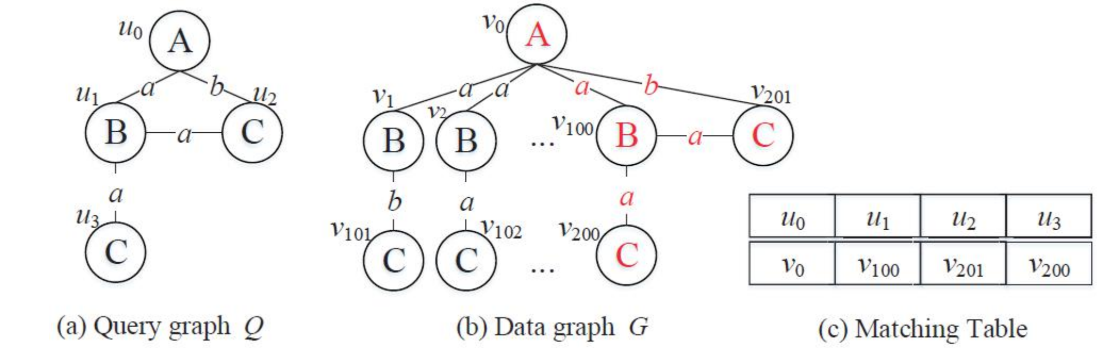


- Zhu, et al. LiveGraph: A Transactional Graph Storage System with Purely Sequential Adjacency List Scans. VLDB 2020.

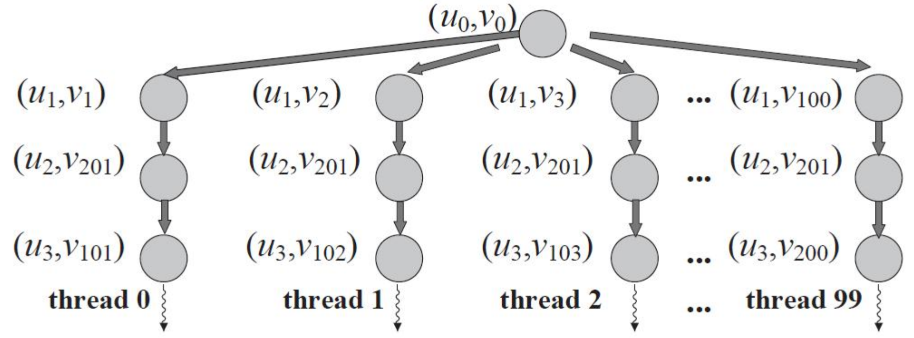
天津大学 Tianjin University

GSI：GPU-friendly Subgraph Isomorphism

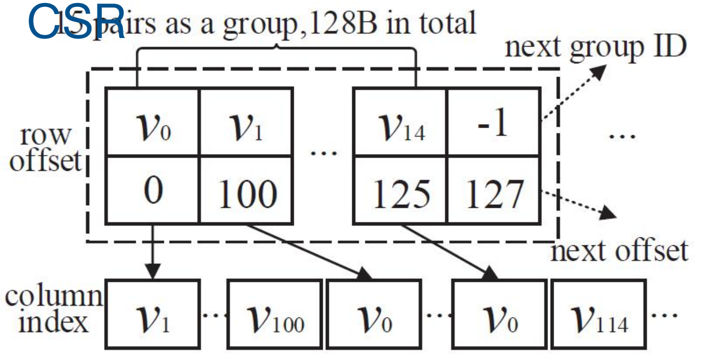
- Existing GPU-based solutions adopt two-step output scheme performing the same join twice in order to write intermediate results concurrently
- Subgraph isomorphism NP-
- Prealloc-Combine strategy, vertex-oriented

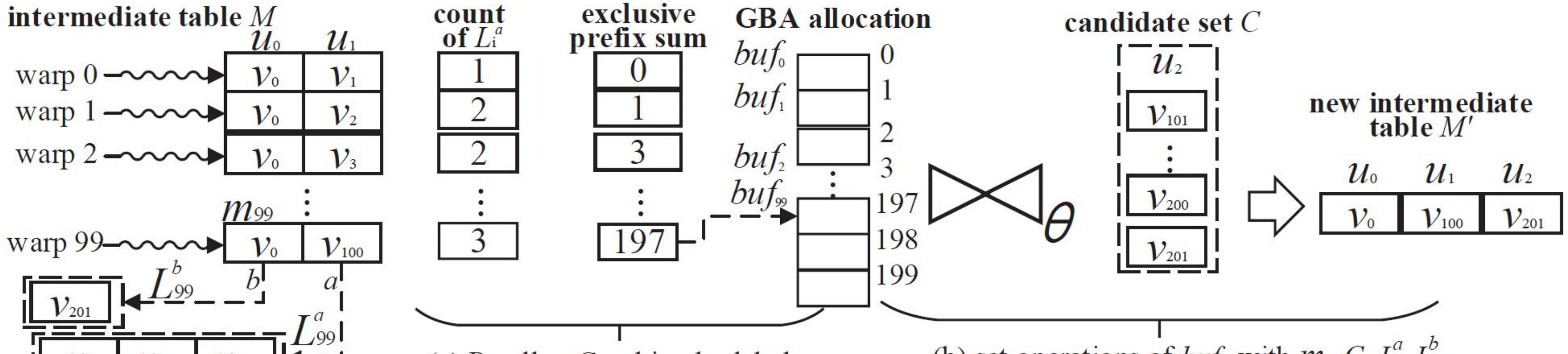


(a) Query graph Q (b) Data graph G (c) Matching Table



Partitioned CSR



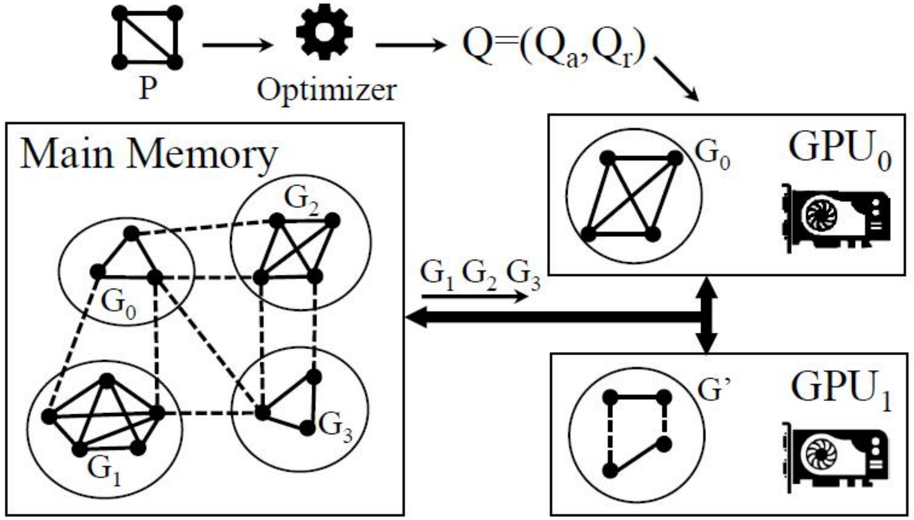


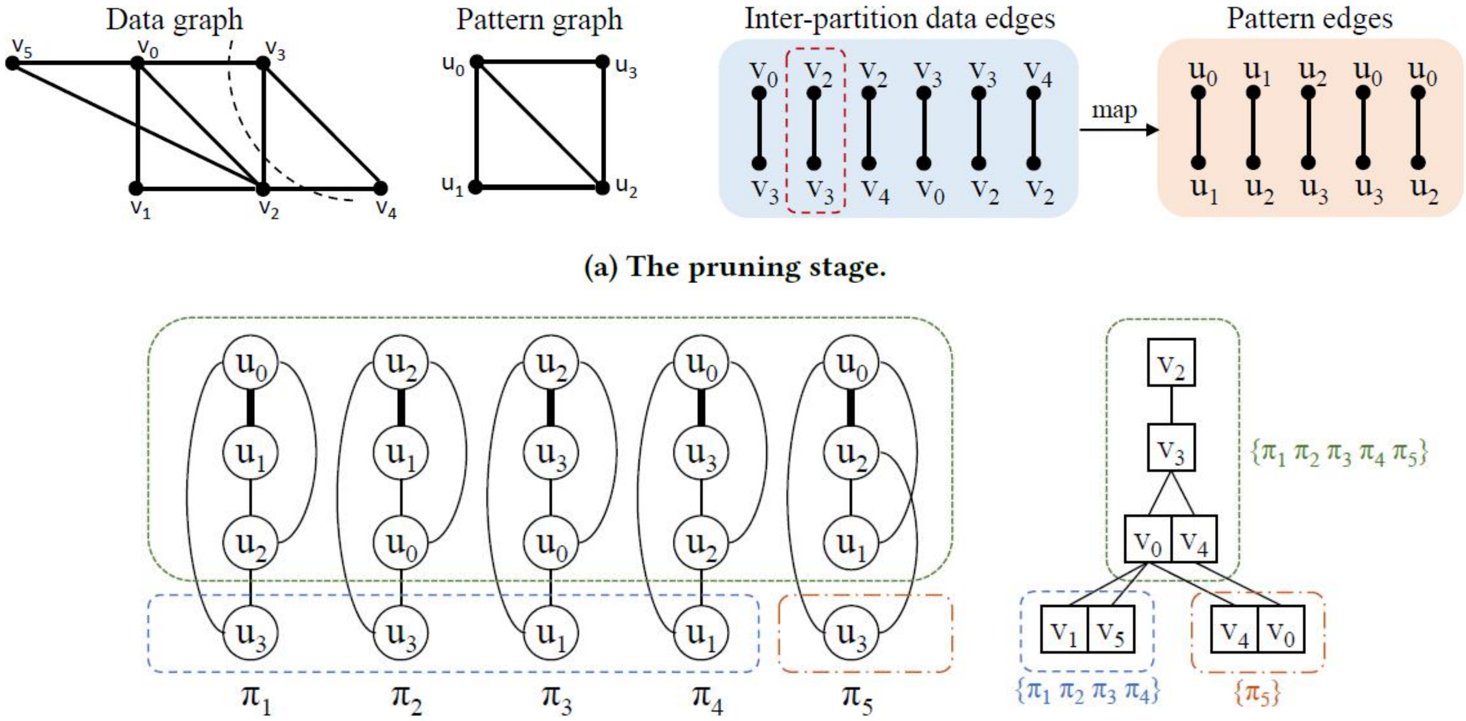
(a) Prealloc-Combine by label a (b) set operations of buf_i with m_i, C, L_i^a, L_i^b

- Zeng, et al. GSI: GPU-friendly Subgraph Isomorphism. ICDE 2020.

GPU-Accelerated Subgraph Enumeration on Partitioned Graphs

- Existing methods can only handle graphs that fit into the GPU memory
- This approach can scale to large graphs beyond the GPU memory
- Divides the graph into partitions
- enumerating the instances across different partitions





(a) The pruning stage.

- Guo, et al. GPU-Accelerated Subgraph Enumeration on Partitioned Graphs. SIGMOD 2020.

2 知识图谱查询处理：近似查询

Tianjin University 天津大学

Approximate Pattern Matching

- in Massive Graphs with Precision and Recall Guarantees

real-world (257 billion edges)
synthetic (1.1 trillion edges)
massive cluster
(256 nodes/9,216 cores)

- Combines edit-distance based matching with systematic graph pruning
- Identifying all exact matches for up to k edit-distance subgraphs

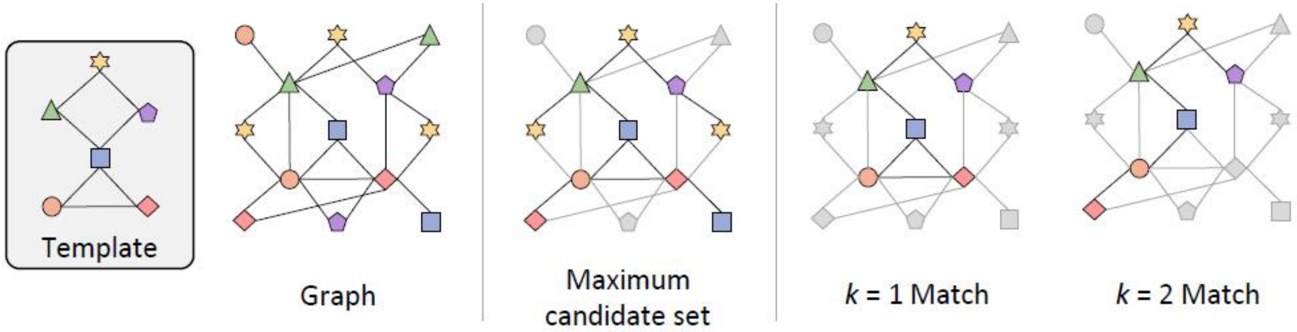
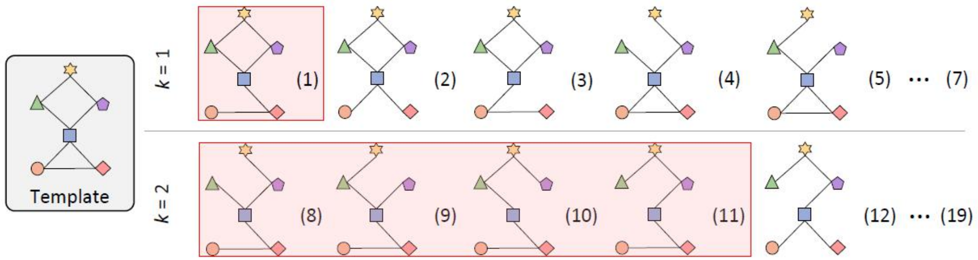
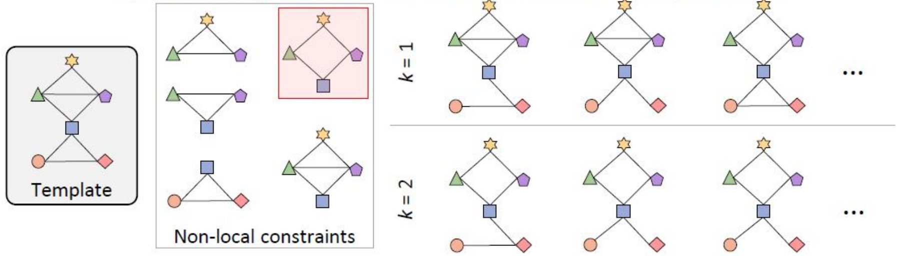


Figure 1: Edit-distance based approximate matching: (left) a search template $\mathcal{H}_0$ and background graph $\mathcal{G}$, and (right) example matches at $k = 1$ and $k = 2$ edit-distance. (Center) the *maximum candidate set* for the search template - the (approximate) match superset.



(a) Example of the containment rule (Def. 1).

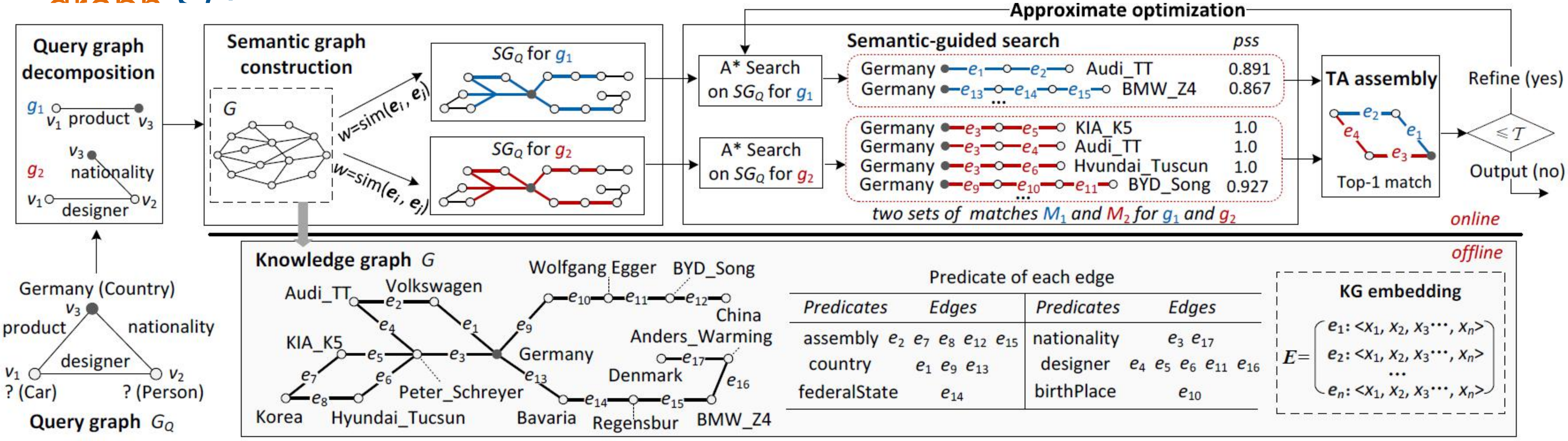


(b) Work reuse, eliminating redundant constraints checks.

- Reza, et al. Approximate Pattern Matching in Massive Graphs with Precision and Recall Guarantees. SIGMOD 2020.

Top-k Similarity Search Semantic Guided and Response Times Bounde

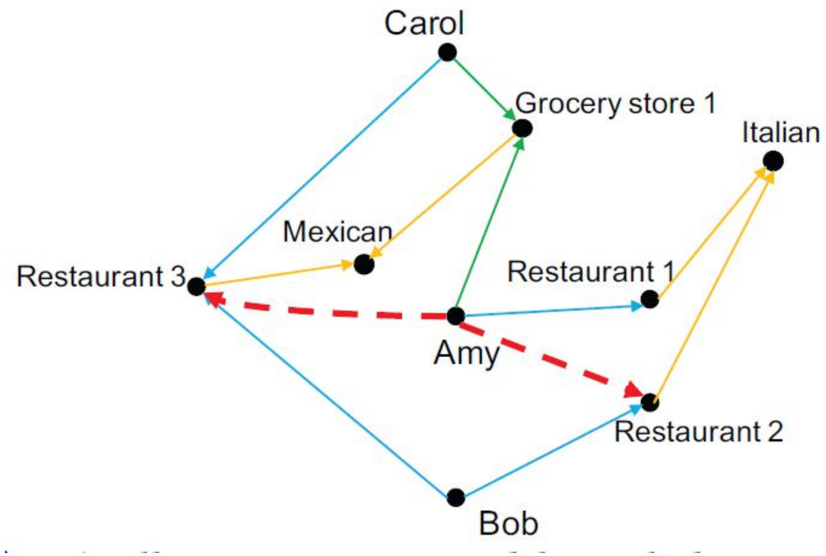
- Structural gap between G_Q and the predefined schema in G causes mismatch
- Users cannot view the answers until the graph query terminates
- Leverage a knowledge graph embedding model to build the semantic graph SG



- Wang, et al. Semantic Guided and Response Times Bounded Top-k Similarity Search over Knowledge Graphs. SIGMOD 2020.

Online Indices for Predictive Top-k Entity and Aggregate Queries on Knowledge Graphs

- Top-k entity queries and aggregate queries
- An incremental index on top of low dimensional entity vectors transformed from network embedding vectors
- Provide theoretical guarantees of accuracy



A virtual knowledge graph

- Spatial indexing of the embedding vectors

(Q1) What are the top-5 most likely restaurants Amy would rate high but has not been to yet?

$\mathbf{h} + \mathbf{r}$

(Q2) What is the average age of all the people who would like Restaurant 2?

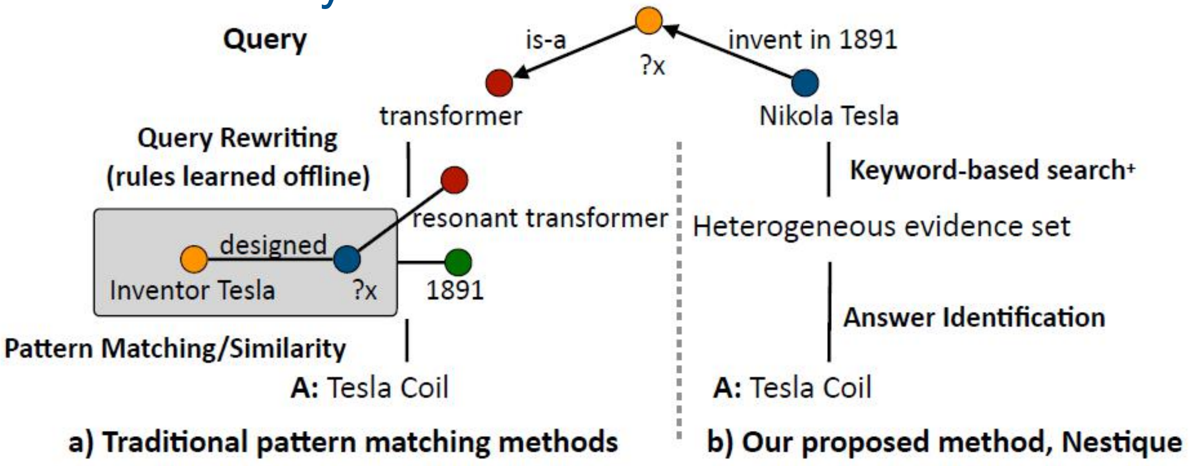
$\mathbf{t} - \mathbf{r}$

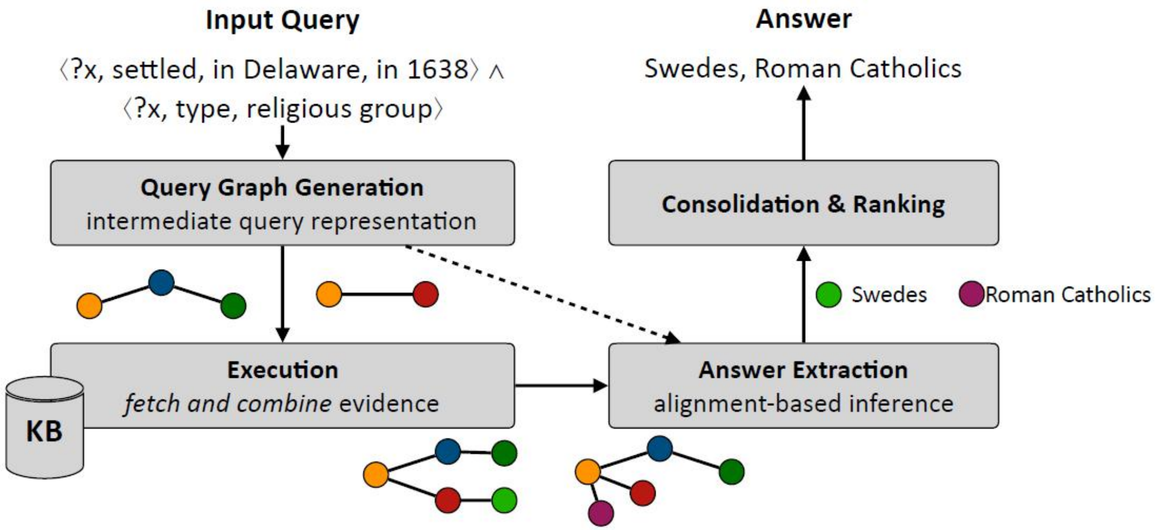
- Use a Johnson-Lindenstrauss (JL) type random projection
- Performs A* search with top-k best choices in the index

- Li, et al. Online Indices for Predictive Top-k Entity and Aggregate Queries on Knowledge Graphs. ICDE 2020.

Online Schemaless Querying of Heterogeneous Open Knowledge Bases

- Open KB derived automatically from unstructured text without any pre-specified ontology
- Finds matches for individual query components
- Identifies an answer by reasoning over the collective evidence
- An alignment-based algorithm for extracting answers based on textual and semantic similarity





Online and does not rely on any offline process to learn transformation functions

- Bhutani, et al. Online Schemaless Querying of Heterogeneous Open Knowledge Bases. CIKM 2019.

天津大学 Tianjin University

SPARQL Rewriting Towards Desired Results

- A Gap between the user's real desire and the actual meaning of a SPARQL query
- Query-restricting: Generate a new query Q′ by applying a set of modifiers

(1) Q′ is similar to Q (2) Q′ returns a result containing as fewer entity tuples in E^- as possible

- Query-relaxing: Generate a similar query Q′ whose resul overlaps with E^+ as much as poss

Query:

```
SELECT ?person ?award
WHERE {
  ?person given_name "Michael".
  ?person family_name "Jordan".
  ?person award_received ?award.
}
```

Result:

?person	?award
Michael Jordan	All-NBA Team
⋮	
Michael Jordan	NBA MPV Award
Michael I. Jordan	ACM Fellow
⋮	
Michael I. Jordan	Rumelhart Prize

(a)

Query:

```
SELECT ?company ?person
WHERE {
  ?company locationCity California.
  ?company industry Software.
  ?company founders ?person.
}
```

Result:

?company	?person
Google Inc.	Larry Page
Google Inc.	Sergey Brin
⋮	

(b)

NP-Hard no *polynomial-time approximation scheme*

propose a $(1-1/e)$-approximation method for *query-restricting*

2 heuristics for *query-relaxing*

Query Modifiers

- AddE(t): adding an edge triplet t in T_Q;
- ModE(t, t'): replacing an edge triplet t with t';
- DelE(t): deleting an edge triplet t from T_Q;
- AddF(f): adding a filter f in F_Q;
- ModF(f, f'): replacing a filter f with a f' ;
- DelF(f): deleting a filter f from F_Q.

- Jian, et al. SPARQL Rewriting: Towards Desired Results. SIGMOD 2020.

Tianjin University 天津大学

Aggregation Support for Modern Graph Analytics in TigerGraph

- GSQL: the specification of aggregation in graph analytics
- PageRank Query: Cross-Iteration Composition via Accumulators

```
CREATE QUERY PageRank (float maxChange, int maxIteration, float dampingFactor) {

  MaxAccum<float> @@maxDifference;       // max score change in an iteration
  SumAccum<float> @received_score;       // sum of scores received from neighbors
  SumAccum<float> @score = 1;            // initial score for every vertex is 1.

  AllV = {Page.*};                       // start with all vertices of type Page

  WHILE @@maxDifference > maxChange LIMIT maxIteration DO
     @@maxDifference = 0;

     S = SELECT      v
         FROM        AllV:v -(LinkTo>)- Page:n
         ACCUM       n.@received_score += v.@score/v.outdegree()
         POST-ACCUM  v.@score = 1-dampingFactor + dampingFactor * v.@received_score,
                     v.@received_score = 0,
                     @@maxDifference += abs(v.@score - v.@score');
  END;
}
```

- Deutsch, et al. Aggregation Support for Modern Graph Analytics in TigerGraph. SIGMOD 2020.

2 知识图谱查询处理：其他工作

天津大学 Tianjin University

分析挖掘

- IDAR: Fast Supergraph Search Using DAG Integration [VLDB 2020] 超图搜索
- Distributed Subgraph Counting: A General Approach [VLDB2020] 通用框架
- G-thinker: A Distributed Framework for Mining Subgraphs in a Big Graph [ICDE 2020]
- Mining an "Anti-Knowledge Base" from Wikipedia Updates with Applications to Fact Checking and Beyond [VLDB2020] 挖掘反三元组
- MultiImport: Inferring Node Importance in a Knowledge Graph from Multiple Input signals [KDD 2020] 节点重要性 latent variable model, attentive GNN
- Neural Subgraph Isomorphism Counting [KDD2020] 基于神经网络

实验评测

- In-Memory Subgraph Matching: An In-depth Study [SIGMOD 2020]
- Realistic Re-evaluation of Knowledge Graph Completion Methods: An Experimental Study [SIGMOD 2020]
- A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs [VLDB 2020]

谢谢大家