



# Mining Summaries for Knowledge Graph Search

Qi Song<sup>1</sup>   Yinghui Wu<sup>1</sup>   Xin Luna Dong<sup>2</sup>

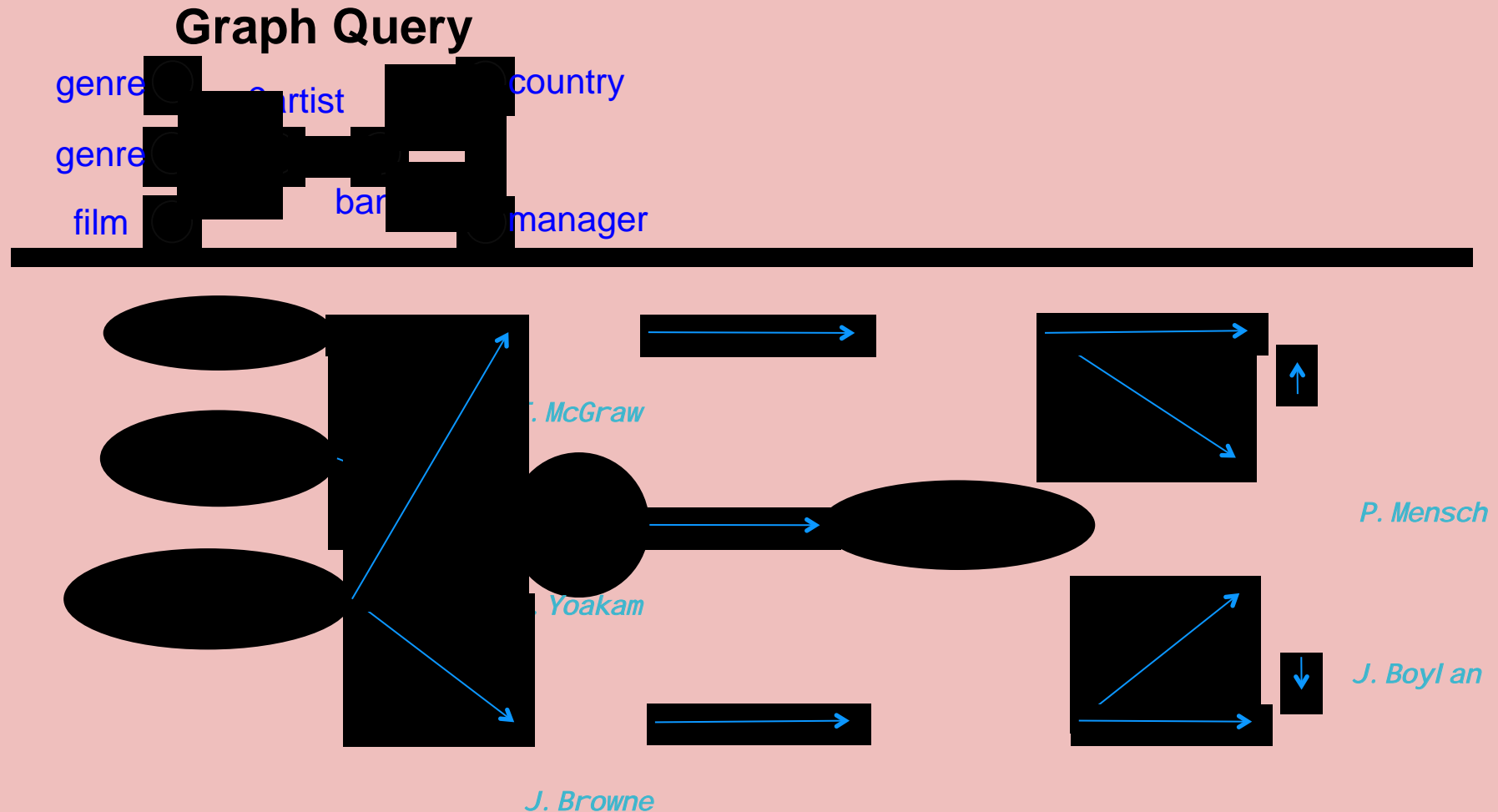
1

2

Sponsored by:

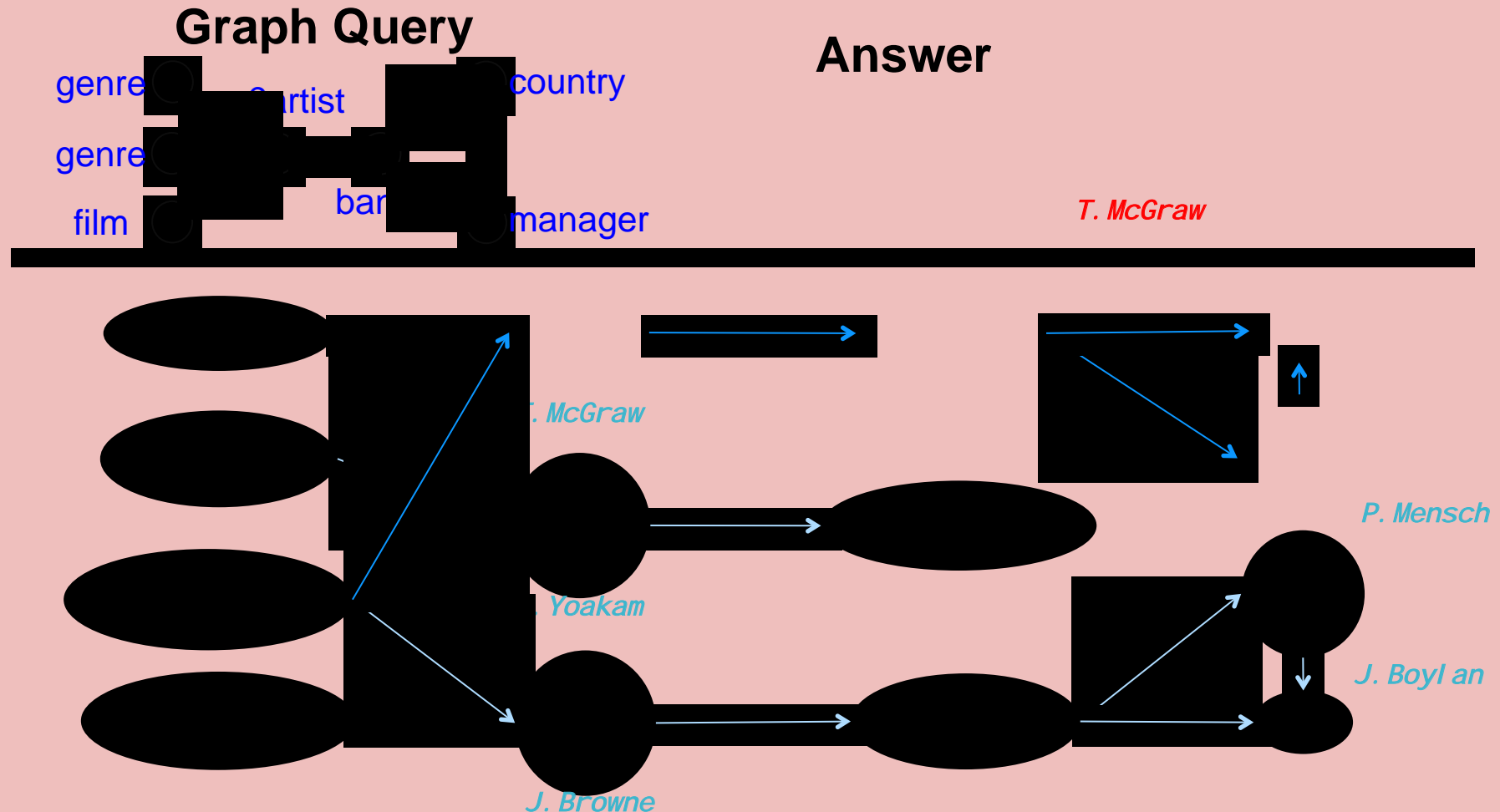
# Searching real world graph data

- A Knowledge Graph  $G$ : used to represent knowledge bases
- A Graph query  $Q$ : graph with types on each node



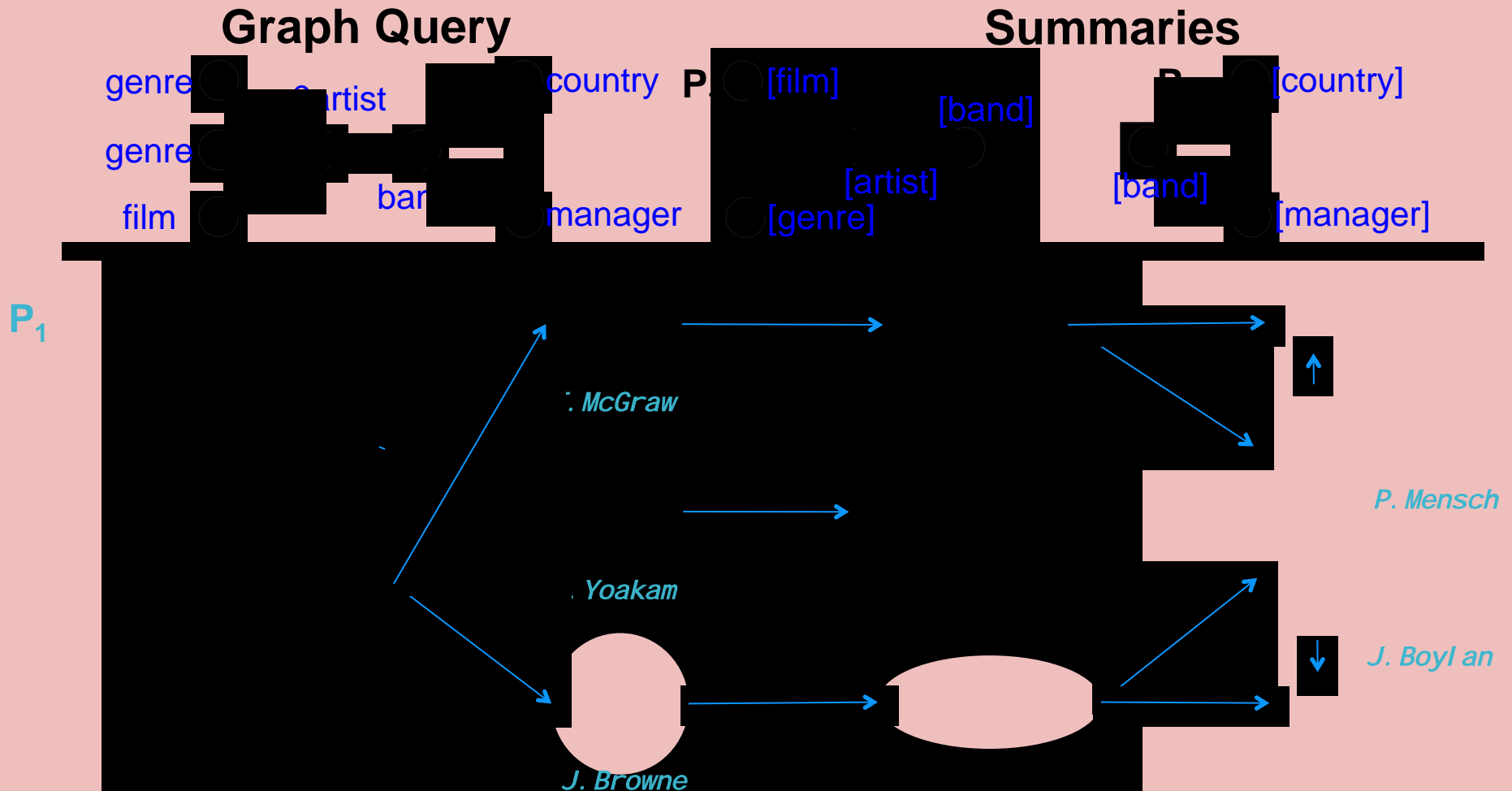
# Searching real world graph data

- Knowledge Graph  $G$ : used to represent knowledge bases
- Graph query  $Q$ : graph with types on each node
- Answer  $Q(G)$ : the set of entities with certain type in the subgraphs of  $G$  that are **isomorphic** to  $Q$ .
- Challenges: **usability & scalability**



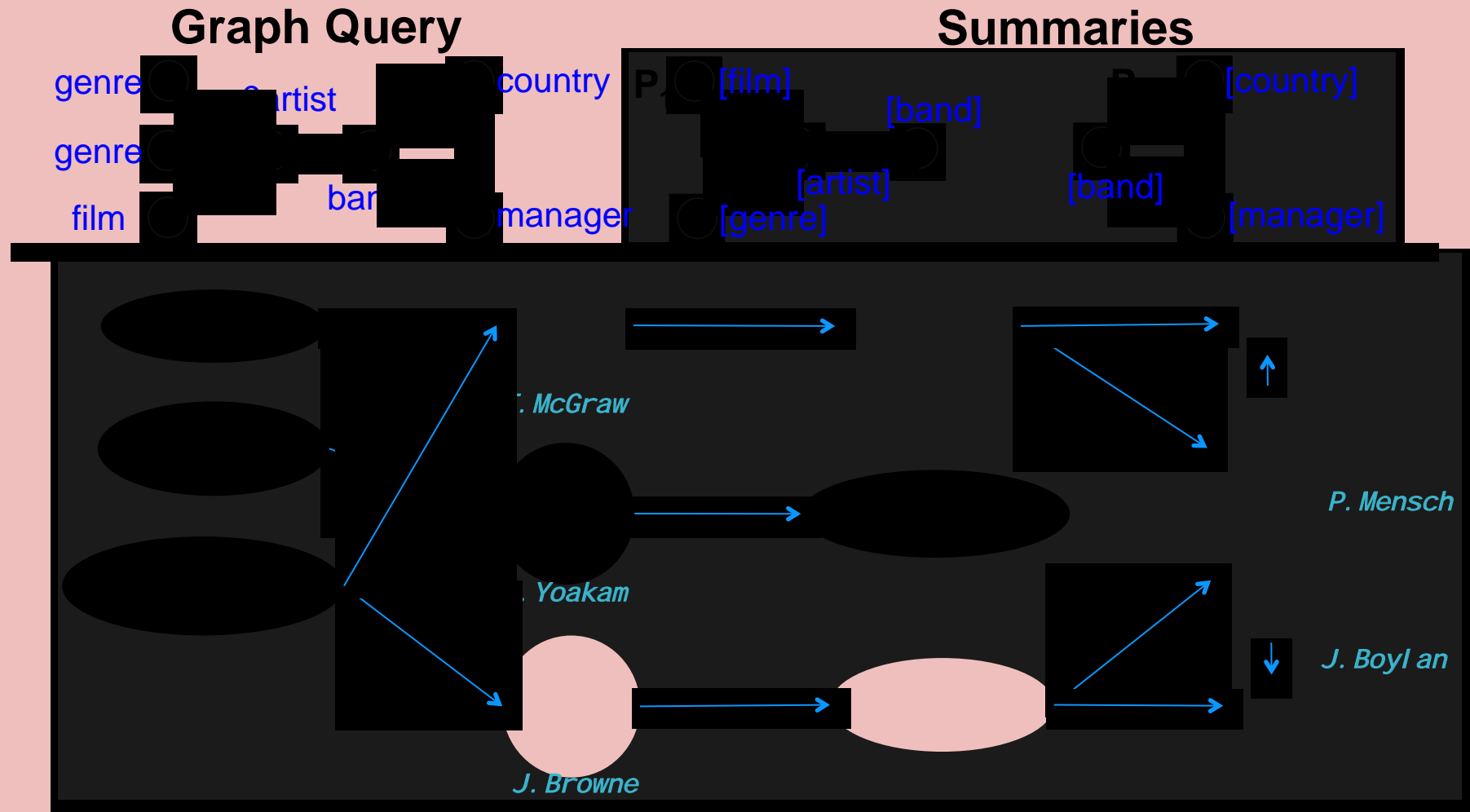
# Use summarization to facilitate query evaluation

- Graph summarization: describe the data graph with a small amount of information



# Use summarization to facilitate query evaluation

- A Graph summarization: describe the data graph with a small amount of information
- A Summary based query evaluation: Query Q can be answered by accessing only the entities summarized by “relevant” patterns



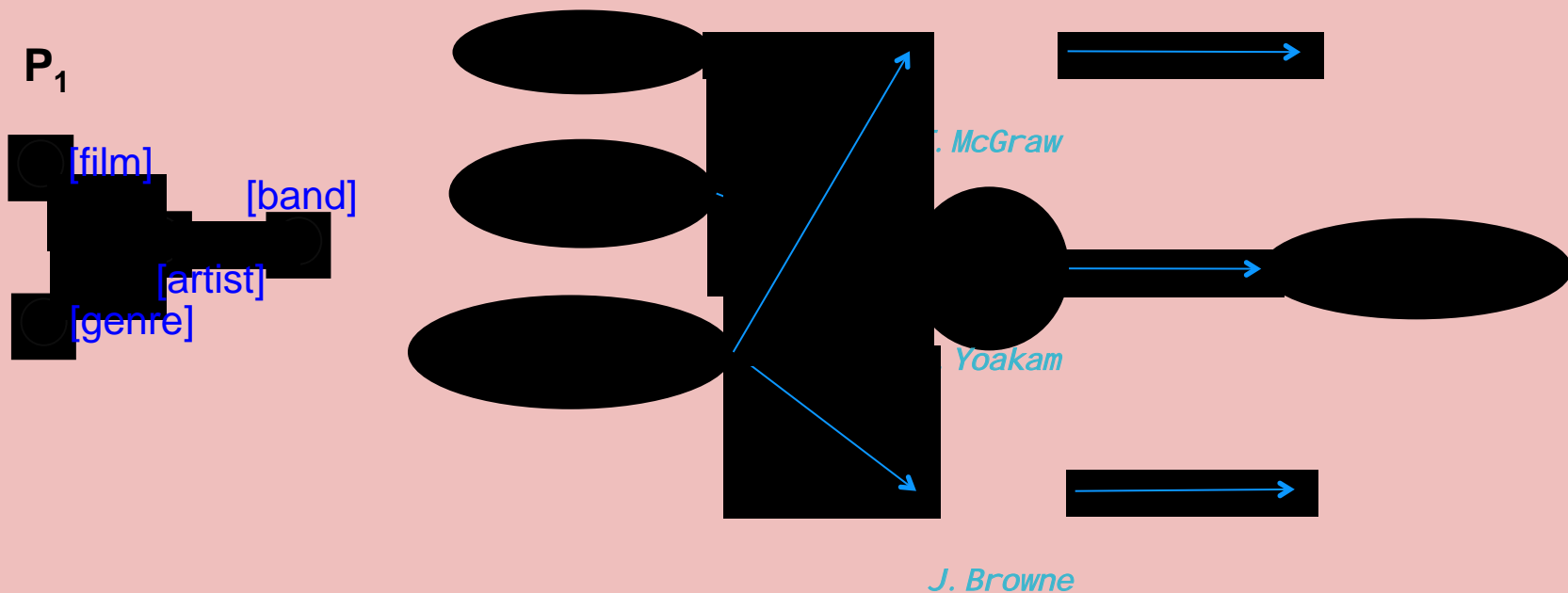
# Use summarization to facilitate query evaluation

- ⌌ How to construct summaries in a schema-less KG?
  - ⌌ Traditional isomorphism based frequent pattern mining may not work
  - ⌌ **D-summaries: summarize similar entities up to a bounded hop d**
- ⌌ How to leverage the summaries to support KG search?
  - ⌌ How to measure the quality of KG summarization
  - ⌌ **Diversified graph summarization problem and approximate algorithms**

## D-summaries

### Subgraph isomorphism VS **d-hop dual simulation**

- Relax 1-1 to many-many relation
- Bounded match with hop d
- Dual-simulation: parent-children matching
- Quadratic time solvable



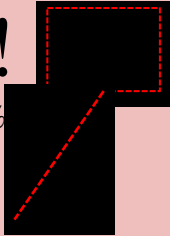
# Diversified knowledge graph summarization

## Problem definition:

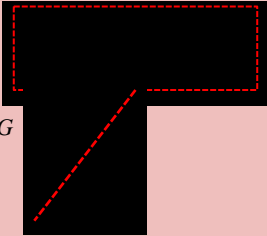
- Given: knowledge graph  $G$ , integers  $k$  and  $d$
- Output: a set of  $k$   $d$ -summaries that maximizes the bi-criteria quality function.

## Objective function

$$F(S_G) = (1 - \frac{\sum_{P_i \in S_G} |P_i \cap S_G|}{card(S_G)}) + \frac{\sum_{P_i \in S_G} |P_i \setminus S_G|}{card(S_G)}$$



Informativeness



Difference

# Diversified knowledge graph summarization

- A 2-approximation algorithm *approxDis*:
  - ^ Mining frequent patterns based on d-similarity
  - ^ Calculate pair-wise score and select top score pairs
  - ^ Have to wait until all frequent patterns are generated
- A Anytime algorithm *streamDis*:
  - ^ Maintain a cache during pattern mining
  - ^  $O(N_t \cdot b_p (b_p + |V|)(b_p + |E|) + \frac{k}{2} N_t^2)$
  - ^ Can be interrupted at any time
  - ^ Maintain 2-approximation (better than pure heuristic)

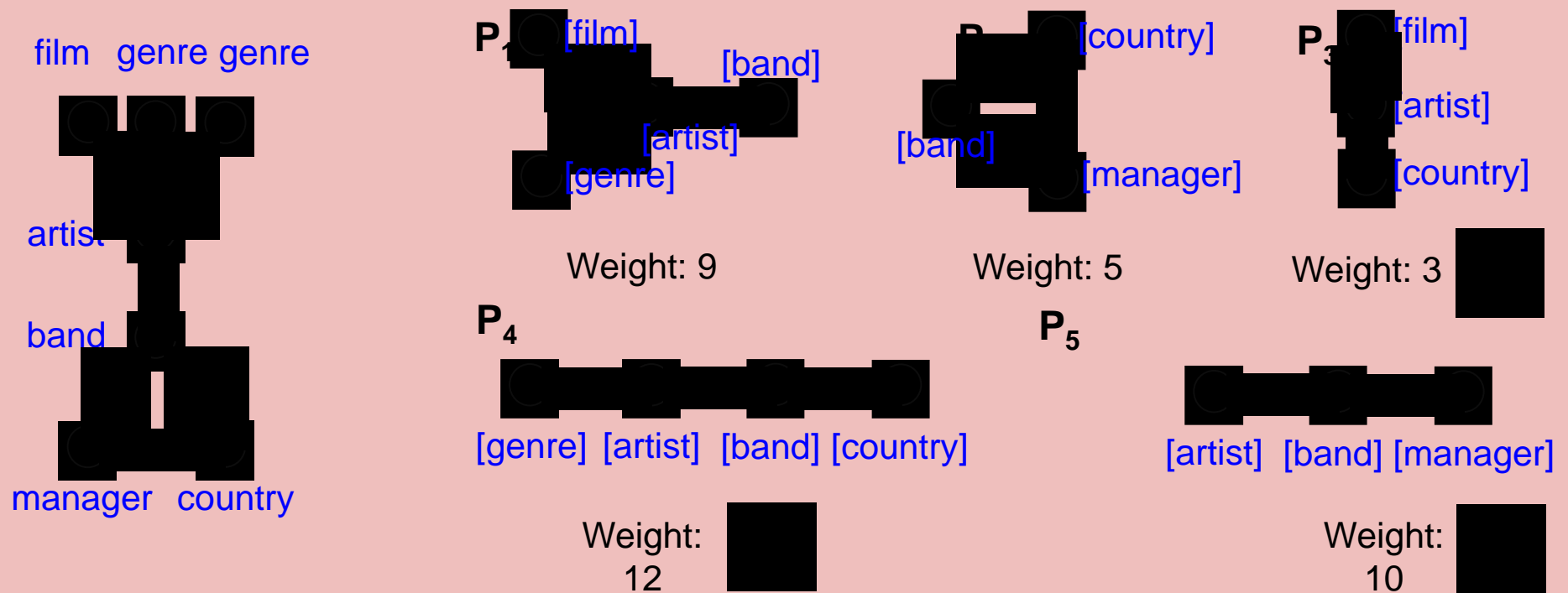
*approxDis*

*streamDis*

# “Summaries + A” scheme for query evaluation

A Pattern selection

^ Iteratively selects a view with minimum weight



A Query answering *evalSum*: “Summaries + A”

## Experimental study

### A Datasets: real-world and synthetic knowledge graphs

- ^A Yago: 1.54M nodes, 2.37M edges, 324k labels
- ^A DBPedia: 4.86M nodes, 15M edges, 676 labels
- ^A Freebase: 40M nodes, 63M edges, 9630 labels
- ^A BSBM: up to 60M nodes, 152M edges and 3080 labels

### A Algorithms:

- ^A Summarization: *approxDis*, *streamDis* and its counterpart *heuDis*, *GRAMI*\*
- ^A Query evaluation: *evalSum*, *evalRnd* (performs random selection), *evalGRAMI* (employs FPGs mined by GRAMI), *evalNo* (directly employ subgraph isomorphism algorithm)

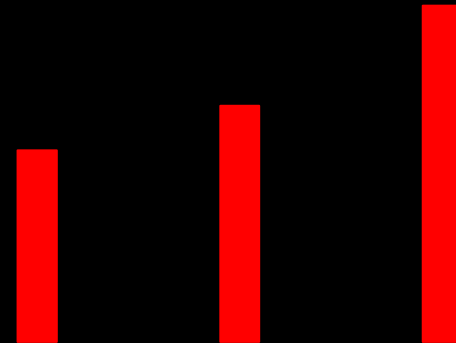
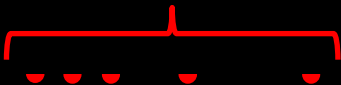
\* M. Elseidy, E. Abdelhamid, S. Skiadopoulos, and P. Kalnis. GRAMI: frequent subgraph and pattern mining in a single large graph. *PVLDB*, 7(7):517–528, 2014.

Source code: <https://github.com/songqi1990/KnowGraphSum>

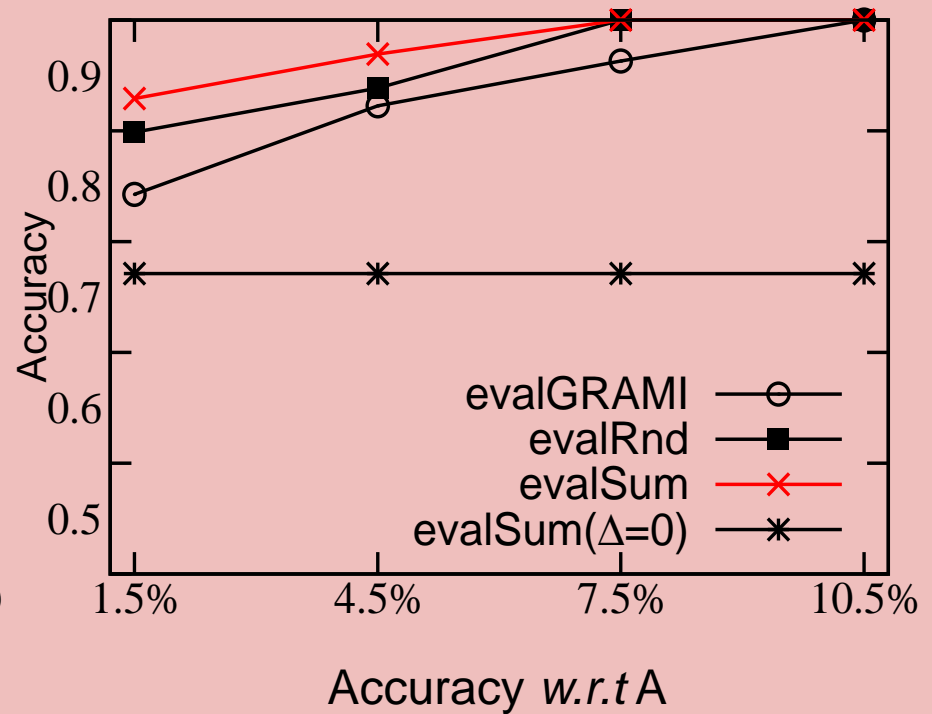
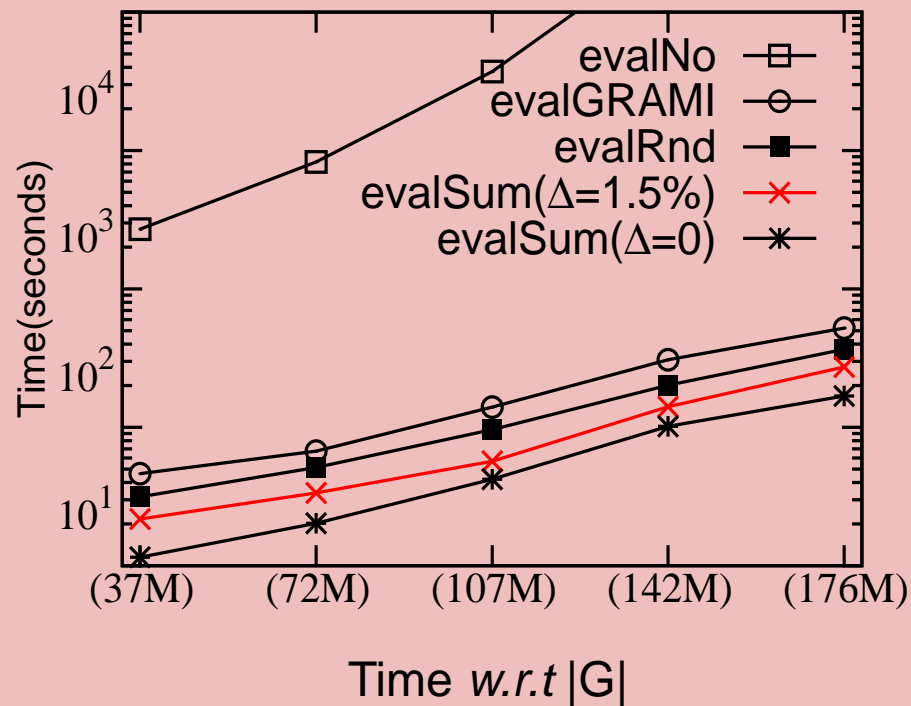
# Effectiveness of summary discovery

^ fi()t' f&%t'žt %ft . ')! #ižt' ffflt ("t  
^ fift ("t "%žt %'f#(\$ f#)&ž, fift)t ži()  
f&%t'žt %ft 1

^ i 't' (&ž\$ fž%), t ži()t' )! fP%"/~ ~



# Effectiveness of *evalSum*



32\*\$!(Źfi)!' )! fP%~' °' ž  
~t&t~ °. P/žŽ' ž#\$-4 3/2 ž! fi"°! fl#\*~. \$  
/Ž. ~)° "\$ Ž#( ~"#,

## Conclusion and future work

### A Mining Summaries for Knowledge Graph Search:

- ^A We proposed a class of d-summaries
- ^A We developed feasible summary mining algorithms and efficient query evaluation algorithm
- ^A We show that our algorithms efficiently generate concise summaries that significantly reduces query evaluation cost

### A Future work

- ^A Distributed query evaluation over different information source
- ^A Query suggestion, data integration, knowledge fusion using views



**Thanks!**