GQBE: Querying Knowledge Graphs by Example Entity Tuples

Nandish Jayaram [1], Mahesh Gupta [1], Arijit Khan [2], Chengkai Li [1], Xifeng Yan [2], Ramez Elmasri [1]

University of Texas at Arlington [1]   University of California, Santa Barbara [2]

Knowledge Graphs
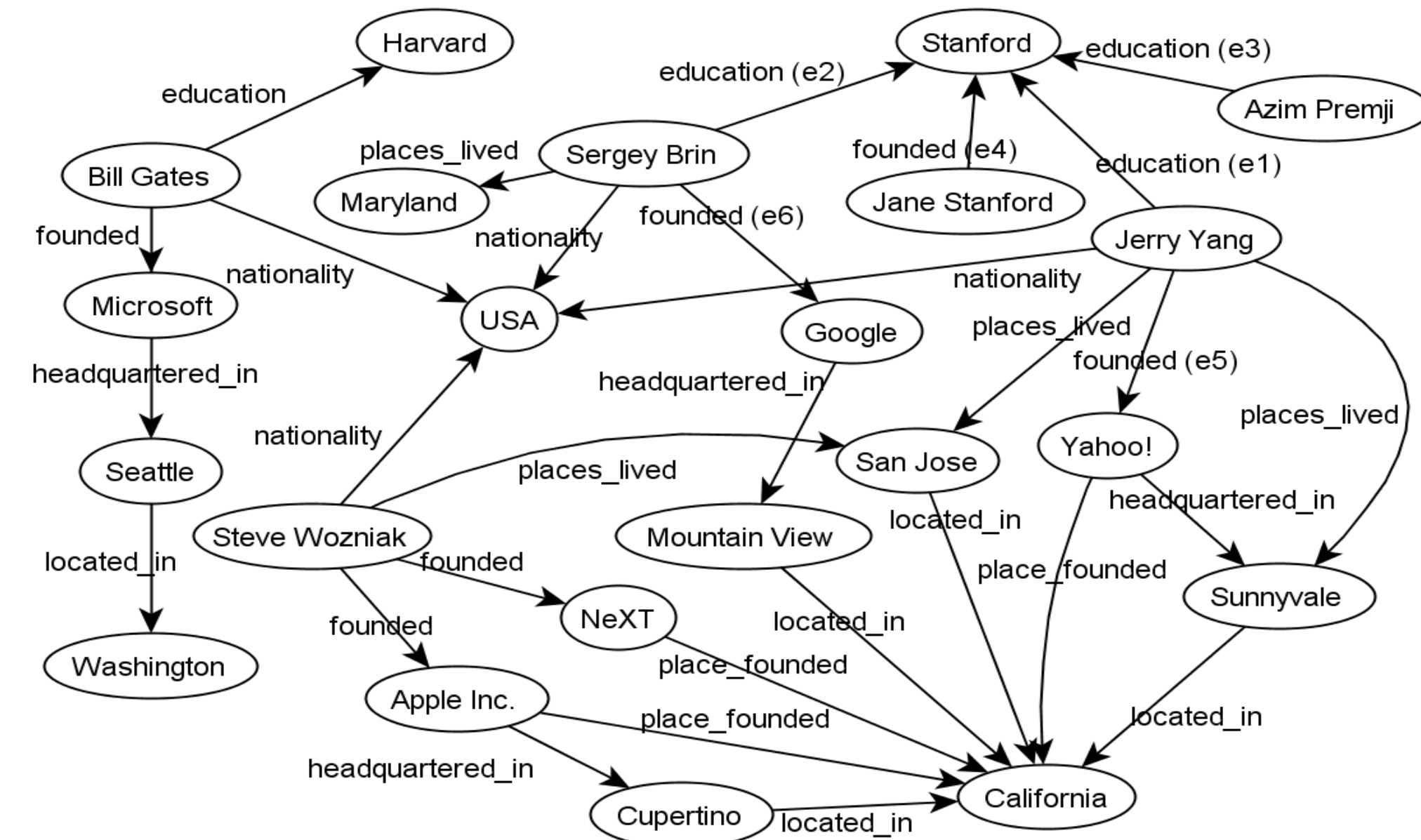
Linking Open Data:
    52 billion RDF triples

Dbpedia:
    470 million facts

Freebase:
    1.8 billion facts

Yago:
    120 million facts

Usability Challenges

- Big and complex data.
    - Lack of schema.
    - Challenging to users (both novice and experienced) and developers.
- How to query the graph, and understand the results.
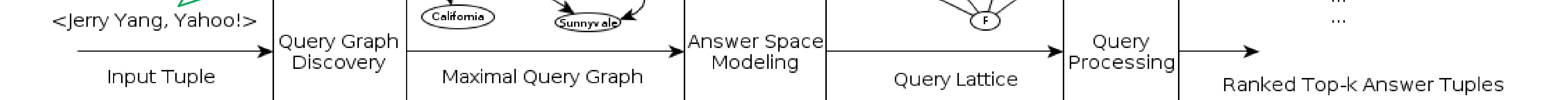
Related Work

- Query-by-example in relational databases [Zloof, 1975].
- Keyword search and keyword-based query formulation [Chang et al., 2011].
- Set expansion [Wang et al., 2007].
- Semantic query by example [Lim et al., 2013].
- XML query relaxation [Amer-Yahia et al., 2005].

Search Interface
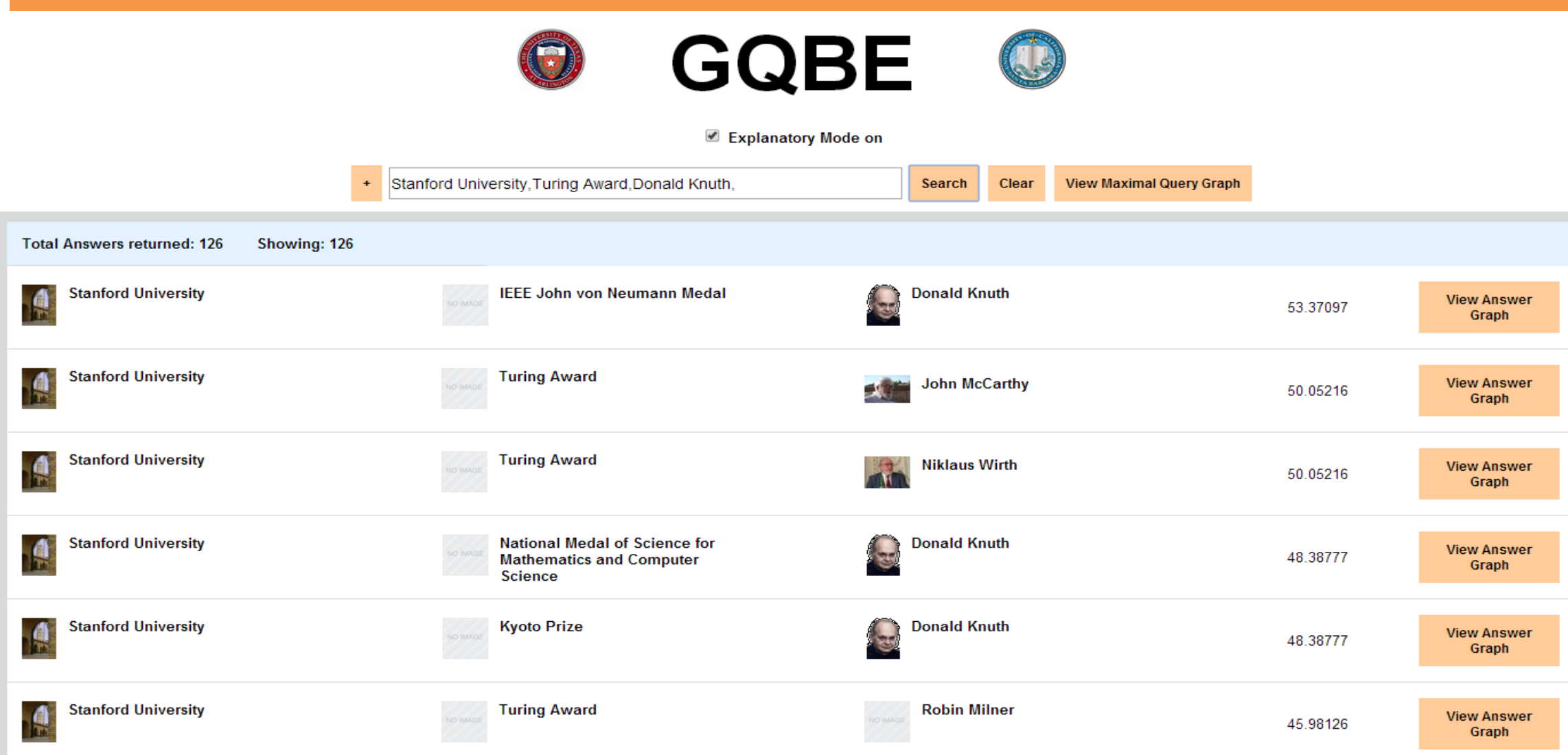
GQBE Architecture

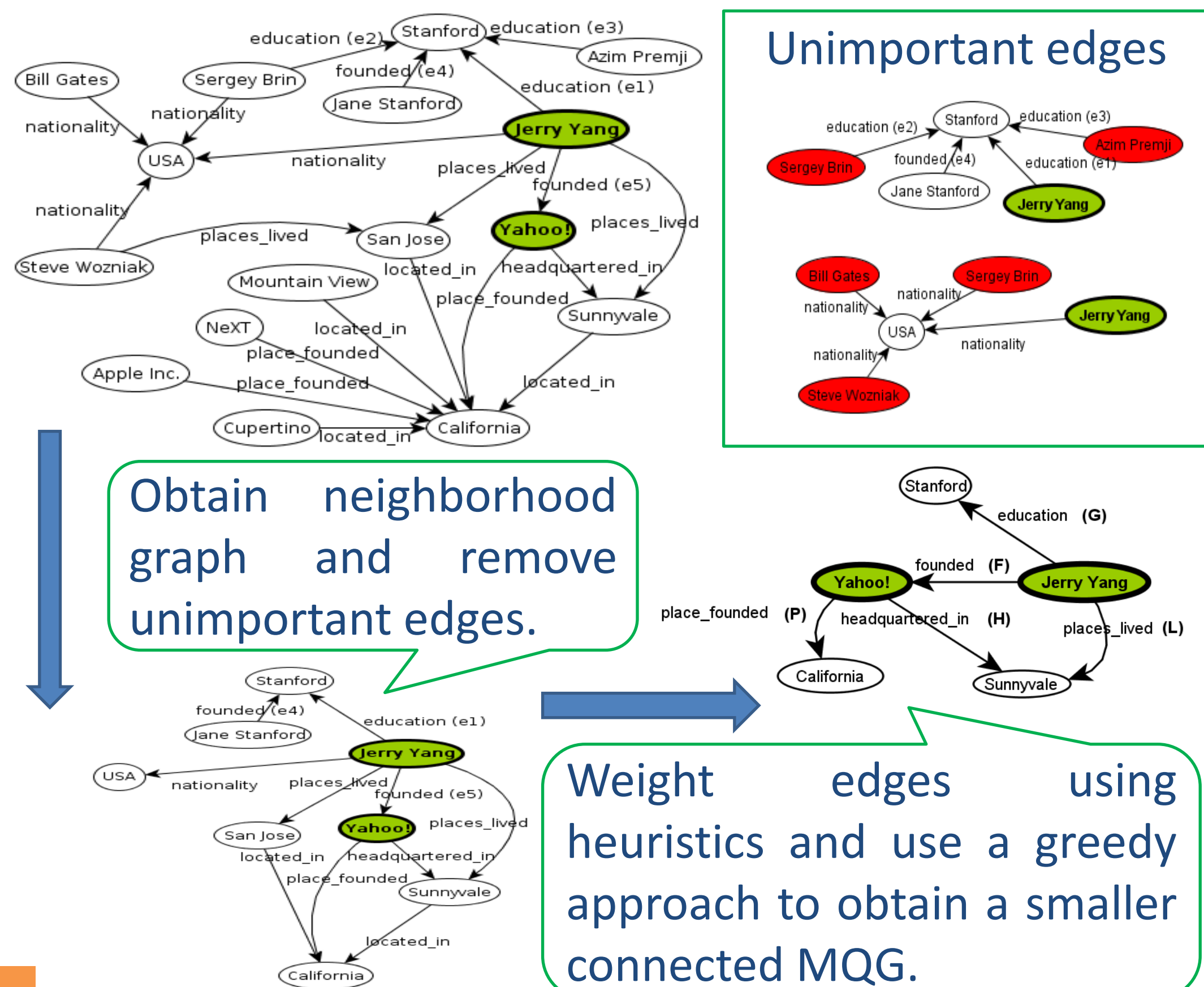Input is an example tuple: an example of what the user wants to find.

Output is a set of similar answer tuples.

<David Fillo, Yahoo!>
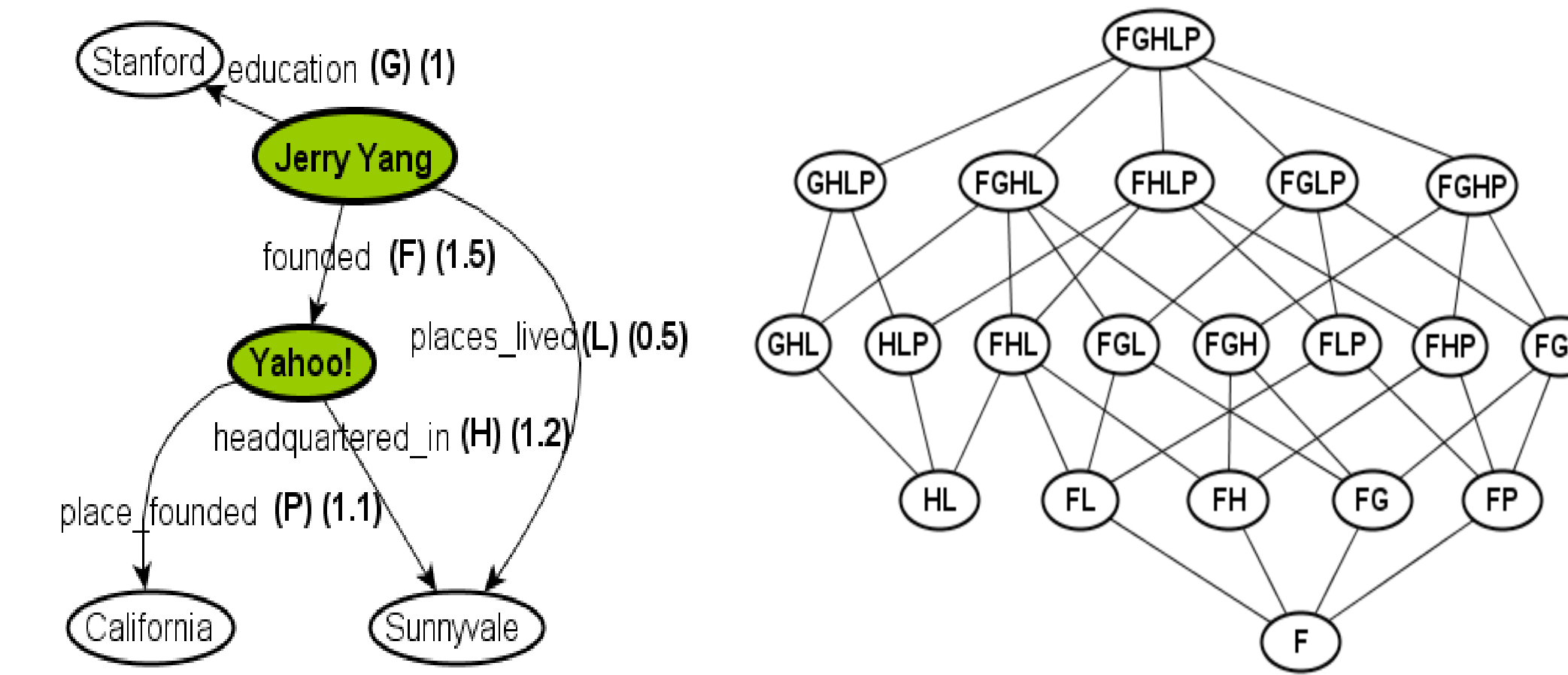<Bill Gates, Microsoft>
<Steve Wozniak, Apple Inc.>
...

<Jerry Yang, Yahoo!> → Query Graph Discovery → Answer Space Modeling → Query Processing → Ranked Top-k Answer Tuples

Input Tuple   Maximal Query Graph   Query Lattice

Query Graph Discovery

Unimportant edges

Obtain neighborhood graph and remove unimportant edges.

Weight edges using heuristics and use a greedy approach to obtain a smaller connected MQG.
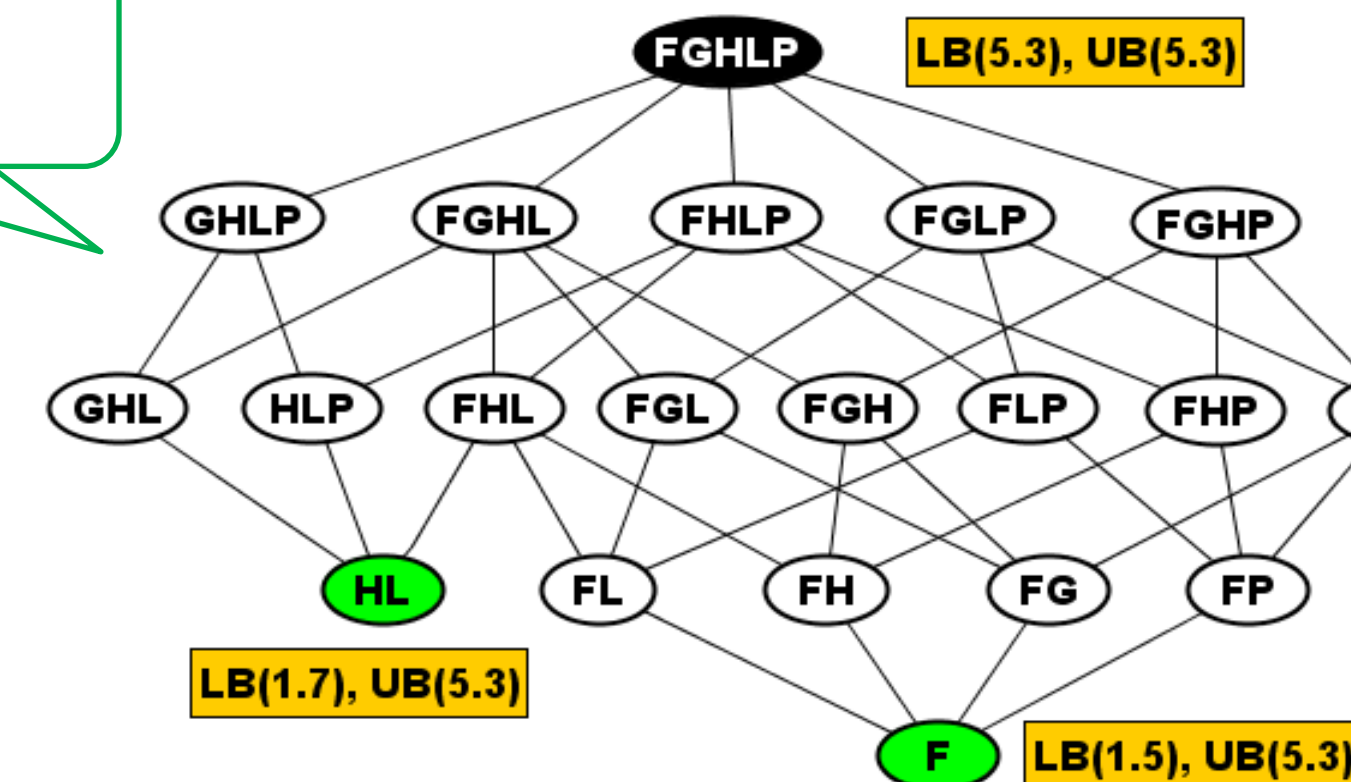
Answer Space Modeling

- Nodes (F) and (HL) are two minimal query trees.
- Node (F) corresponds to the sub-graph that connects Jerry Yang and Yahoo through edge *founded*.
- Node (FGHLP) is the MQG, and it corresponds to the entire query graph on the left.
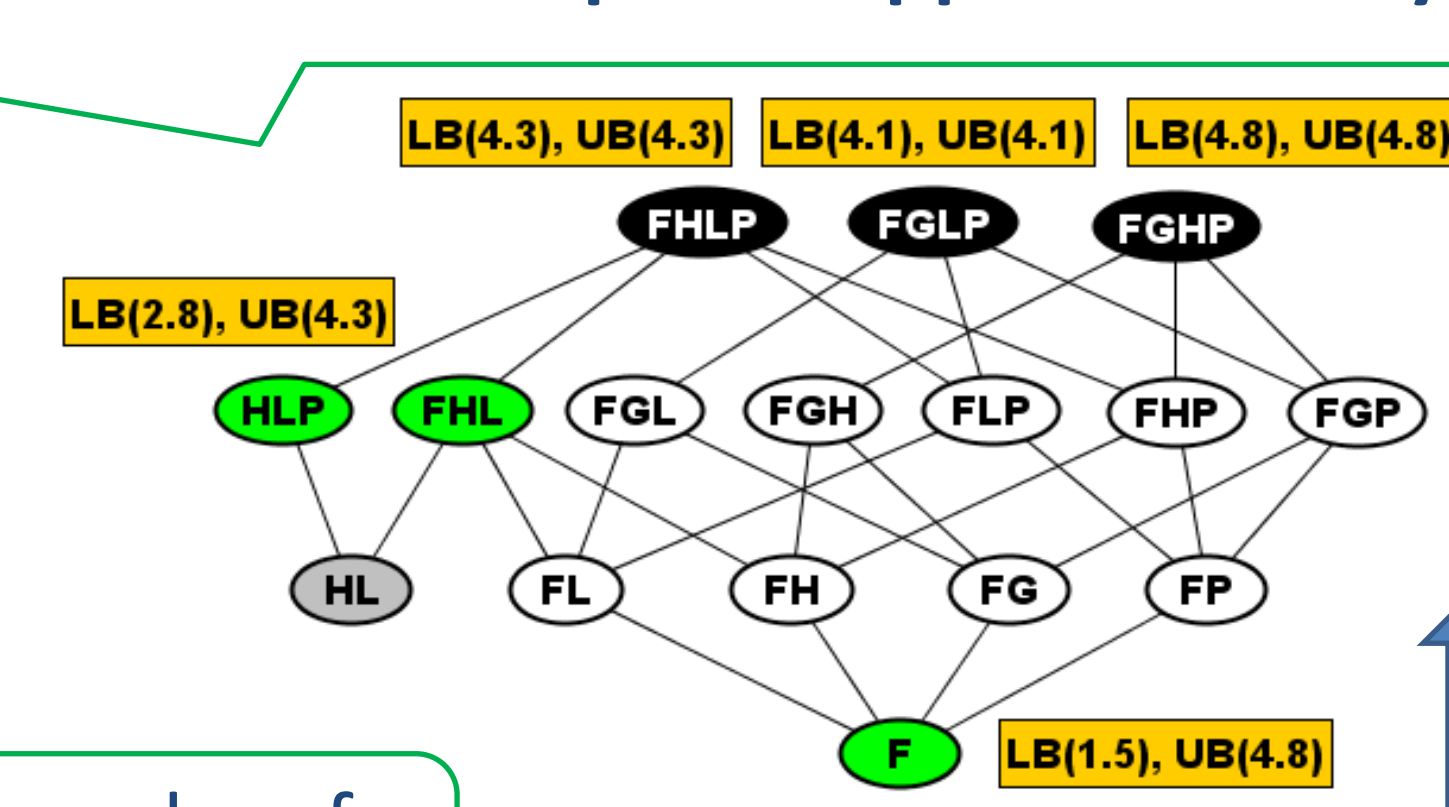- Each internal node in the lattice is a sub-graph of the MQG.

Query Processing

Initial Lattice

LB(5.3), UB(5.3)

Modified Lattice with Recomputed Upper Boundary

LB(4.3), UB(4.3)   LB(4.1), UB(4.1)   LB(4.8), UB(4.8)

LB(2.8), UB(4.3)

LB(1.7), UB(5.3)

LB(1.5), UB(5.3)

LB(1.5), UB(4.8)

Upper Bound (UB) of a node is the score of its highest-scored super-graph in the lattice.

All the super-graphs of a null node are pruned.

LB(5.3), UB(5.3)

LB(2.7), UB(5.3)

LB(1.5), UB(5.3)

LB(1.7), UB(5.3)

LB(1.5), UB(5.3)

Lower Bound (LB) is the edge matching score of the corresponding graph.

(GHL), a node which does not have any answers, is a null node.

Multiple-Tuple Query Graphs

Multiple MQGs are merged based on edge labels and vertex label matches.
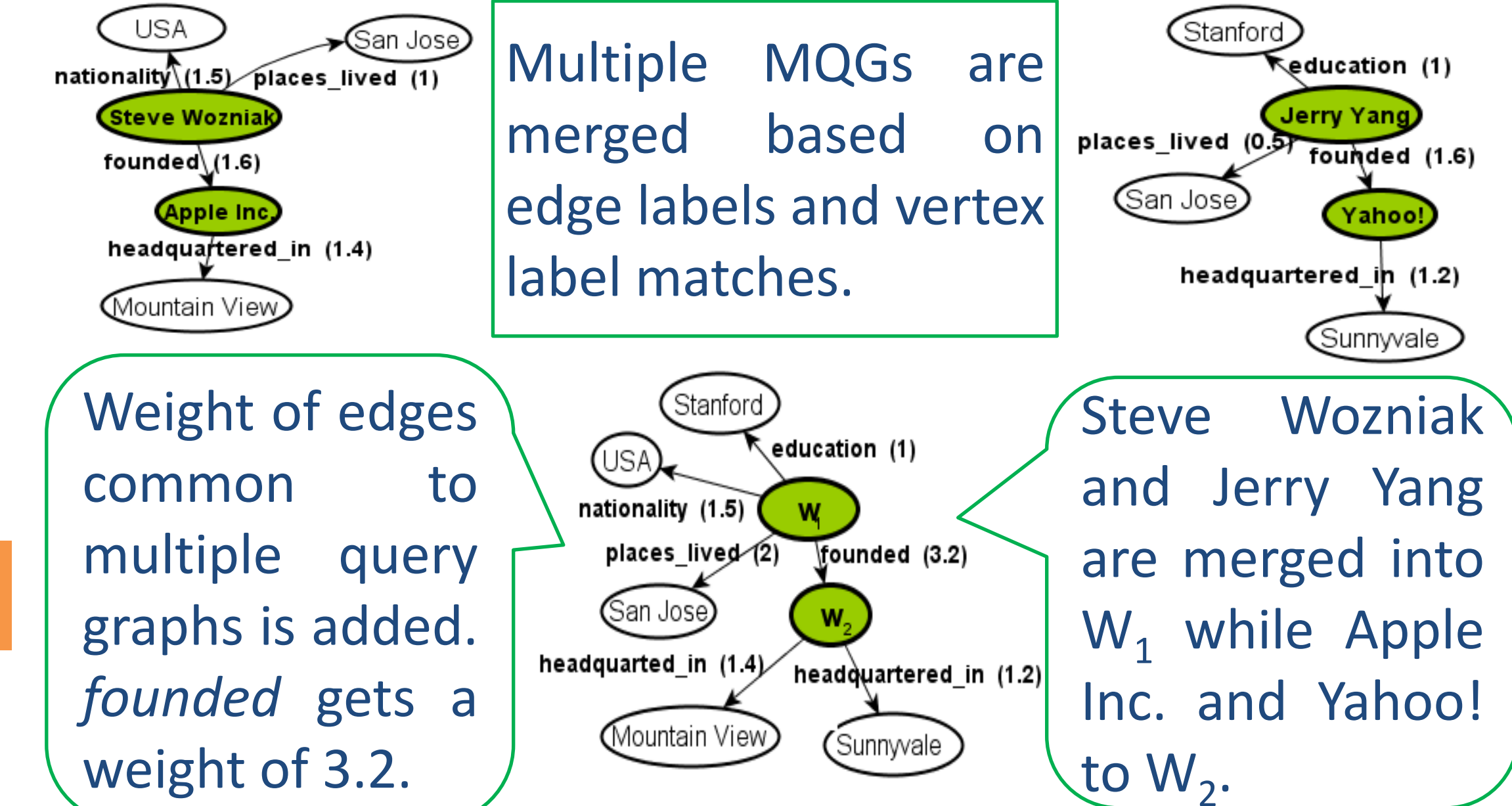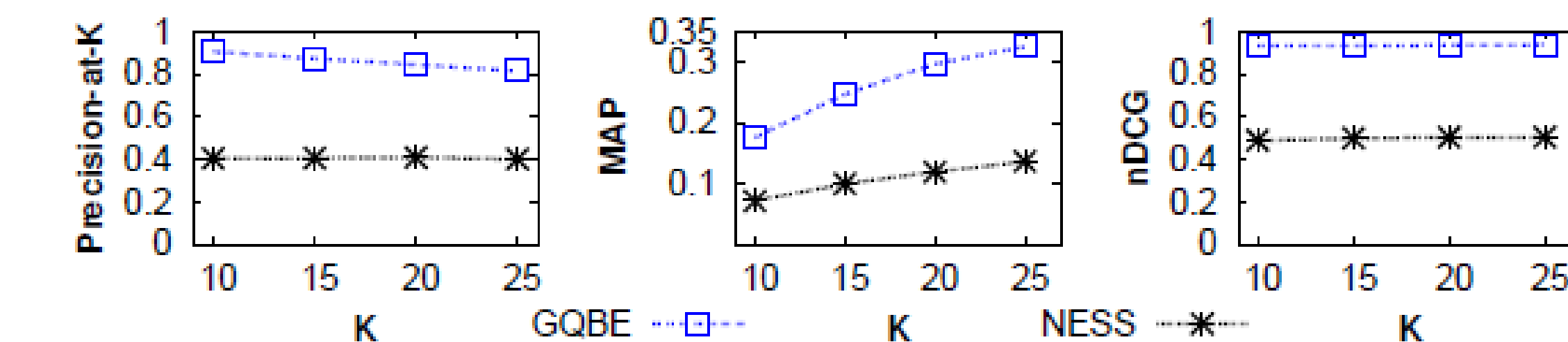
Weight of edges common to multiple query graphs is added. *founded* gets a weight of 3.2.

Steve Wozniak and Jerry Yang are merged into W_1 while Apple Inc. and Yahoo! to W_2.

Experiments

Ground truth based accuracy comparison of GQBE and NESS. The measured parameters are precision-at-k, Mean Average Precision and normalized Discounted Cumulative Gain.

Query	PCC	Query	PCC	Query	PCC	Query	PCC
F_1	0.79	F_2	0.78	F_3	0.60	F_4	0.80
F_5	0.34	F_6	0.27	F_7	0.06	F_8	0.26
F_9	0.33	F_{10}	0.77	F_{11}	0.58	F_{12}	undefined
F_{13}	undefined	F_{14}	0.62	F_{15}	0.43	F_{16}	0.29
F_{17}	0.64	F_{18}	0.30	F_{19}	0.40	F_{20}	0.65

Pearson Correlation Coefficient (PCC) between GQBE and Amazon Mechanical Turk Workers, for k=30. MTurk workers were presented with answer pairs and asked for their preference between the two answers in each pair. The user preferred ranking was compared with GQBE's ranking list. 20000 such opinions collected.
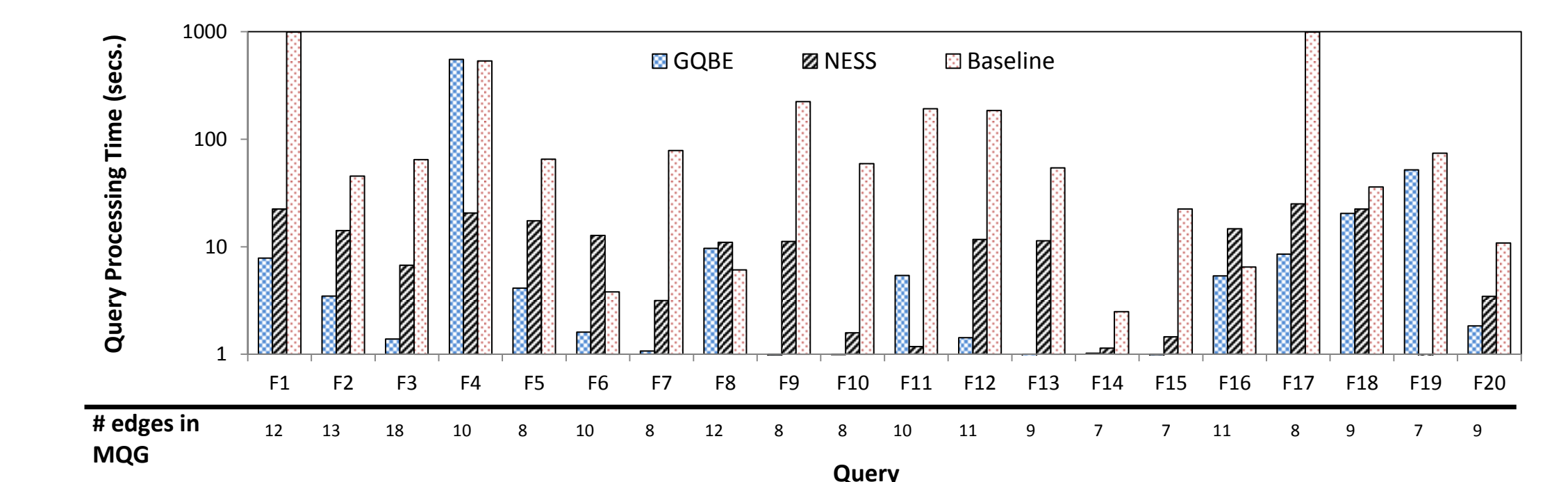
Query processing times of GQBE, NESS and Baseline.

Technical Details and Demo

[1] N. Jayaram, A. Khan, C. Li, X. Yan and R. Elmasri. Querying knowledge graphs by example entity tuples, in arxiv.org/abs/1311.2100

[2] Demo URL: http://idir.uta.edu/gqbe