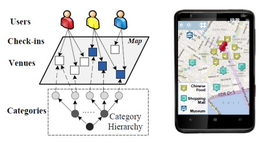
Answering Why-Questions for Subgraph Queries in Multi-Attributed Graphs

Qi Song∗, Mohammad Hossein Namaki∗, Yinghui Wu∗§
Washington State University∗ Pacific Northwest National Laboratory§
{qi.song, m.namaki, yinghui.wu}@wsu.edu∗

ICDE MACAU 2019

Introduction

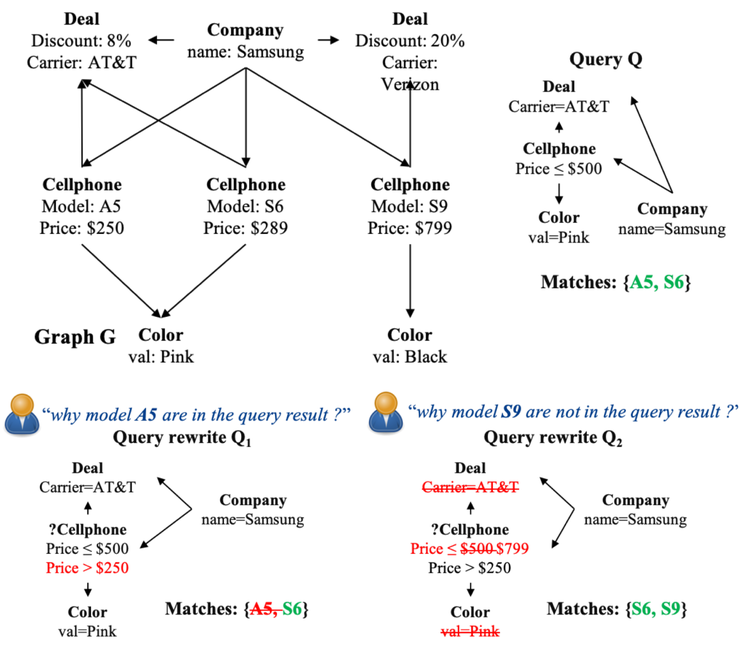
- **Subgraph queries have been applied to access and understand complex networks.**



POI Recommendation

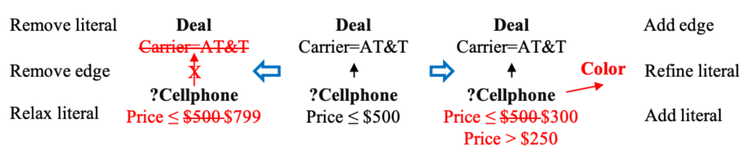
Knowledge Extraction

- **Writing queries is nevertheless a nontrivial task for end users:**
  - ➢ The graph is large and heterogeneous.
  - ➢ Users often need to revise the queries multiple times to find desirable answers.
- **Why-questions.**
  - ➢ Why question: "why some (unexpected) entities are in the query answer?"; and
  - ➢ Why-not question: "why certain entities are missing from the query result?"



Problem Formulation

- **Categorization of Why-Questions.**
  - ➢ Why: why the nodes in V_{N_u} are included as matches for u_o in G
  - ➢ Why-not: why the nodes in V_{Cu}, are not matches of u of Q?
- **Answers for Why-Questions**
  - ➢ Query rewrites: query editing operators;



  - ➢ Answer closeness $cl(O, V_u)$:

- Why: the fraction of V_{N_u} that are excluded from $Q'(u_o, G)$.
- Why-not: the fraction of new matches in V_{C_u} that are introduced in $Q'(u_o, G)$.

$$\mathrm{cl}(O, V_{N_u}) = \frac{|(Q(u_o,G) \setminus Q'(u_o,G)) \cap V_{N_u}|}{|V_{N_u}|} \qquad \mathrm{cl}(O, V_{C_u}) = \frac{|(Q'(u_o,G)) \cap V_{C_u}|}{|V_{C_u}|}$$

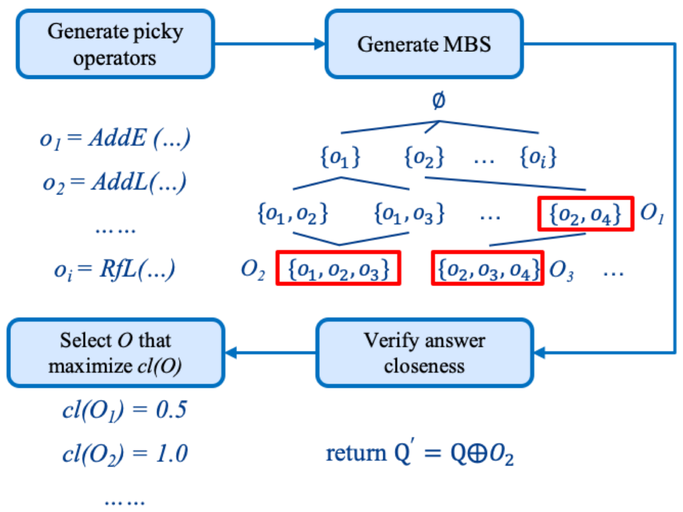
Guard condition: avoid over-refinement or over-relaxation

  - ➢ Problem statement
    - ❖ Given a query Q, answer $Q(u_o, G)$, graph G, a Why-question W, editing budget B,
    - ❖ Compute a query rewrite $Q' = Q \oplus O^*$, such that

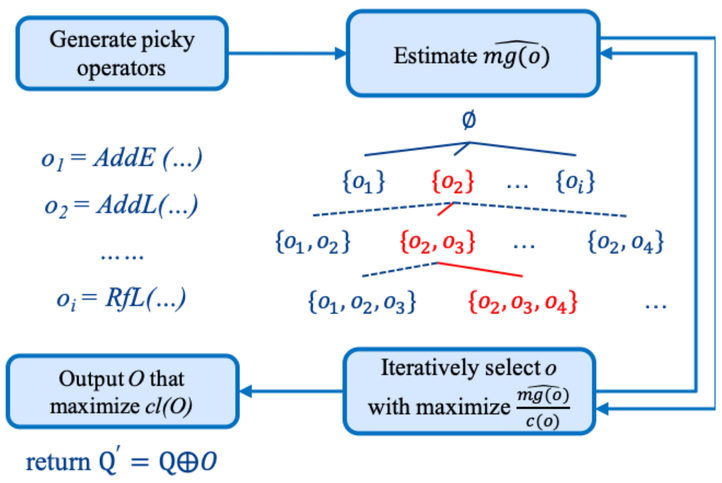
$$O^* = \underset{O: c(O) \le B}{\mathrm{argmax}}\ cl(O, V_u)$$

Answering Why Questions

- **Computing optimal query rewrites**
  - ➢ Maximum bounded set (MBS): with $c(O) \le B$ and all of its superset has cost exceeds B.
  - ➢ An exact algorithm (ExactWhy):



    - ❖ Time cost: $O(|Q|\left|N_{d_Q+1}(Q(u_o,G))\right| + |O_s|^{3B}|N_{d_Q+1}(V_{N_u})|^{|Q|})$.
- **Approximating optimal query rewrites**
  - ➢ Given refinement operator set O, the marginal gain of an operator o to O: $mg(O,o) = cl(O \oplus \{o\}) - cl(O)$;
  - ➢ Function $cl(\cdot)$ is submodular;
  - ➢ An approximation algorithm ApproxWhy:



    - ❖ Approximation : $\frac{1}{2} \cdot \left(1 - \frac{1}{e}\right) \cdot cl(O^*, V_{N_u}) - 6B\varepsilon$;
    - ❖ Time cost: $O(|Q|\left|N_{d_Q+1}(Q(u_o,G))\right| + |O_s|\left|N_{d_Q+1}(V_{N_u})\right|^{|Q|} + |O_s|^2\left|N_{d_Q+1}(V_{N_u})\right|)$.

Answering Why-not Questions

- **Computing optimal query rewrites**
  - ➢ An exact algorithm (ExactWhyNot):
    - ❖ Following the similar manner with ExactWhy but considers only relaxation operators;
    - ❖ Time cost:$O(|Q||O_s|^{2B}|N_{d_Q}(V_{C_u})|^{|Q|})$.
- **A faster heuristic**
  - ➢ A heuristic algorithm HeuWhyNot:
    - ❖ Following the similar manner with ApproxWhy;
    - ❖ Time cost: $O(|Q|\left|N_{d_Q}(V_{C_u})\right| + |O_s|^2\left|N_{d_Q}(V_{C_u})\right|)$.

Extensions

- **Q that contains multiple output nodes;**
- **Why-empty: answer set is empty;**
- **Why-so-many: too many answers.**

Evaluation

Datasets

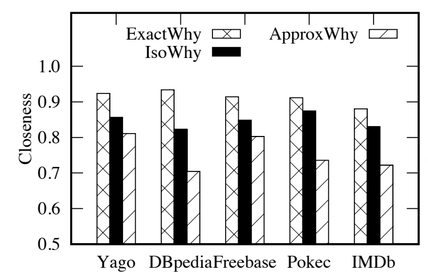
Name	#Nodes	#Edges	#Labels	Attributes per node
DBPedia	4.86M	15M	676	9
Yago	1.54M	2.37M	324K	5
Freebase	40.32M	63.2M	8630	8
Pokec	1.6M	30.6M	10	60
IMDb	1.7M	5.2M	8	6
BSBM	up to 50M	up to 126M	up to 3080	up to 20

Algorithms

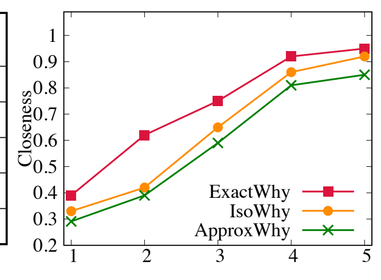
- **ExactWhy, ApproxWhy, IsoWhy**
- **ExactWhyNot, FastWhyNot, IsoWhyNot**

Results

- **Answering Why questions: Effectiveness**



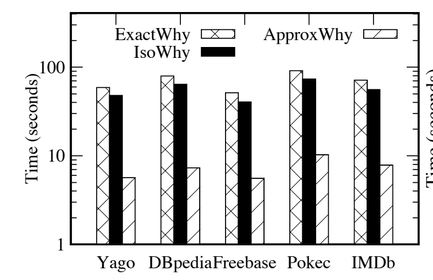
Answer Closeness



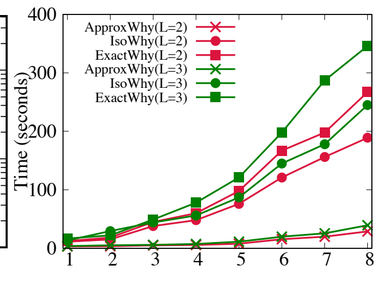
Varying B

ApproxWhy computes good query rewrites that have closeness at least 85% to their optimal counterpart

- **Answering why questions: Efficiency**



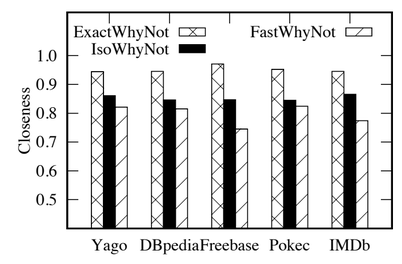
Efficiency



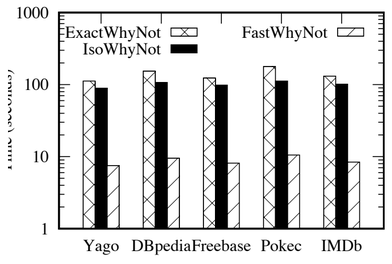
Vayring $|E_Q|$

It is feasible to answer why questions for large graphs.

- **Answering why-not questions**



Answer Closeness

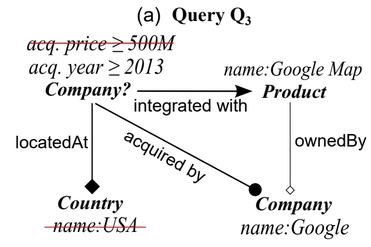


Efficiency

The result is consistent with the results for Why questions.

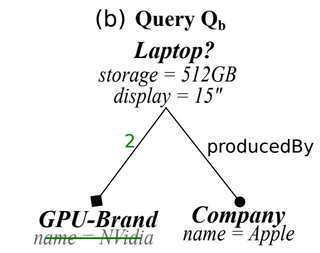
- **Case study**

Business inquiry



Why-not *Urban-Engines* and *Waze*?

Product recommendation



Why empty?

Following Up work and Acknowledgement

- Answering Why-Questions by Exemplars (SIGMOD'2019);
- NAVIGATE: Explainable Visual Graph Exploration by Examples (SIGMOD'19 Demo);
- This work is supported in part by NSF IIS-1633629, USDA/NIFA 2018-67007-28797, Siemens and Huawei HIRP.