CIKM 2017 SINGAPORE

Relaxing Graph Pattern Matching With Explanations

Jia Li[1], Yang Cao[2], Shuai Ma[1]

[1]Beihang University, China

[2]University of Edinburgh, UK

北京航空航天大學 BEIHANG UNIVERSITY

University of Edinburgh

Background

Graph pattern matching

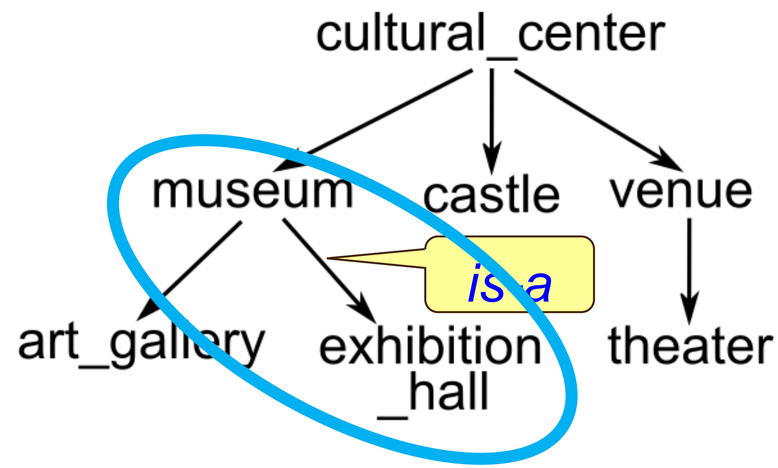
Bijective function

Subgraph isomorphism

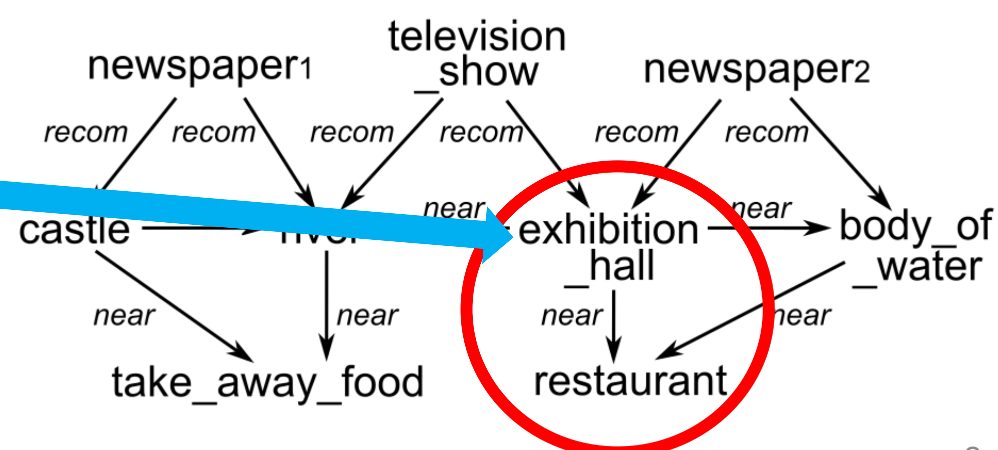
Too Restrictive to find matches

Taxonomy subgraph isomorphism

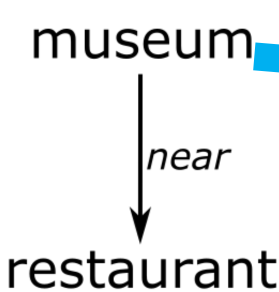
(Partial) Taxonomy T



Data graph G



Pattern graph Q



Background

Graph pattern matching

↓

Bijective function

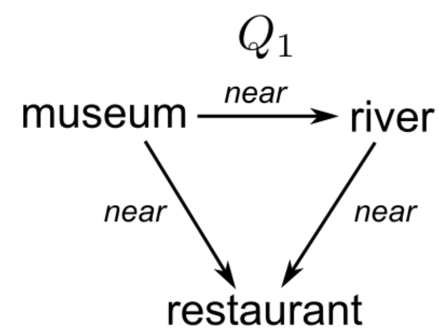
Subgraph isomorphism

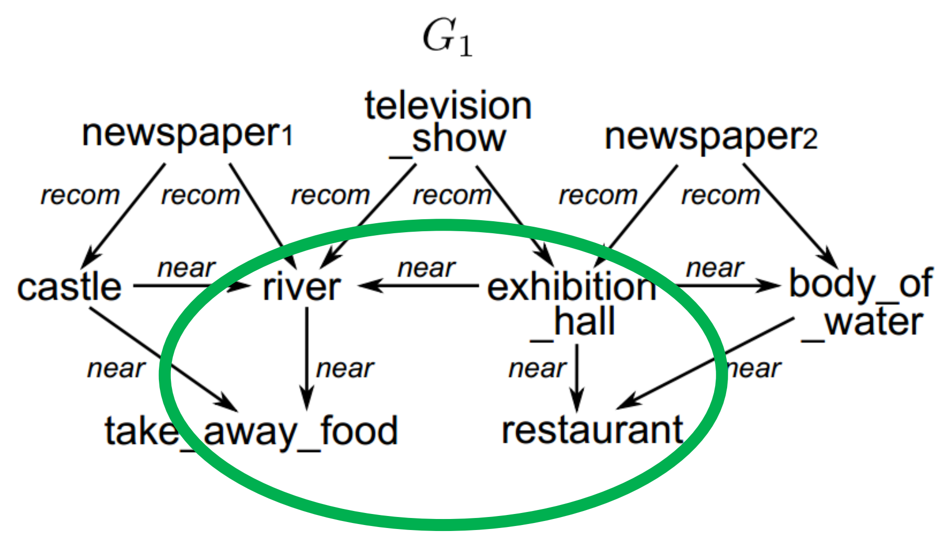
Too Restrictive to find matches

↓

Still too restrictive

Taxonomy subgraph isomorphism





Relax the topological constraints of taxonomy isomorphism

Taxonomy simulation

- Taxonomy simulation

Given a data graph $G(V$ … omy $T(V_T, E_T, f_T)$,

Relation instead of bijective function

G matches Q w.r.t. T via *taxo*… *simulation*, denoted by $Q \trianglelefteq_T G$, if there exits

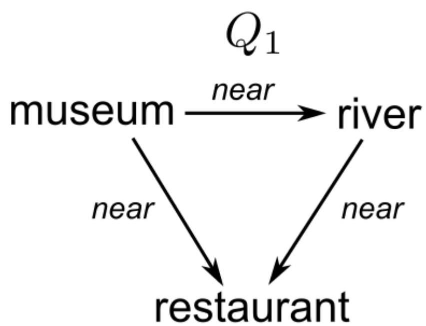
Relaxed label matching

a *left-total* binary match *relation* $R_T \subseteq$ … :

(1) for each $(u, v) \in R^T$, $f(v) \in desc_T\,(f_Q(u))$; and

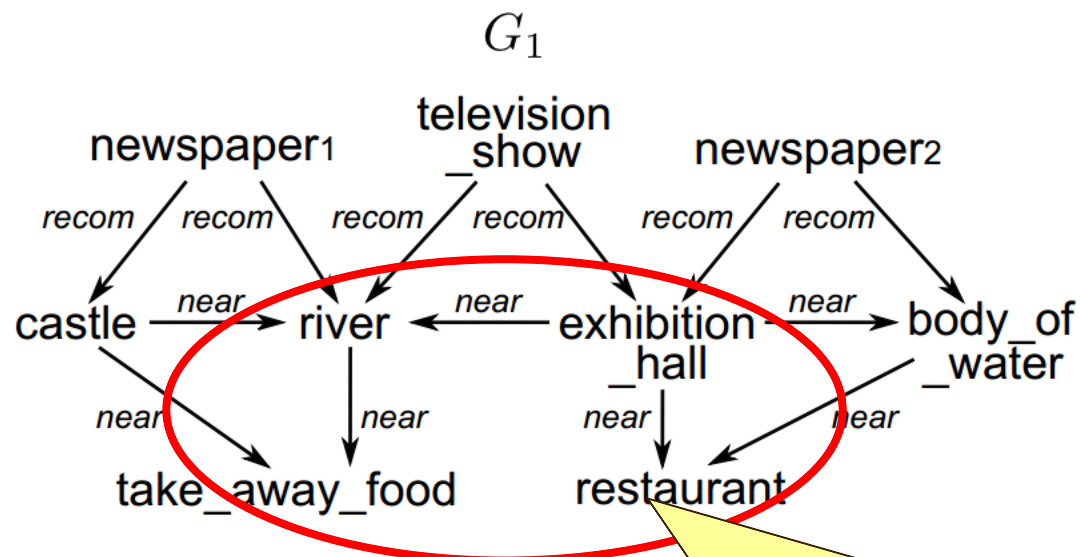
(2) for each edge $e = (u, u')\in E_Q$, there exists an edge $e' = (v, v') \in E$ such that $(u', v') \in R$ and $f_Q\,(e) = f\,(e')$.

Taxonomy simulation



Match results for Q1 in G1

- museum: exhibition_hall
- river: river
- restaurant: { take_away_food, restaurant }



Relation-based structural mapping
Taxonomy-based label matching

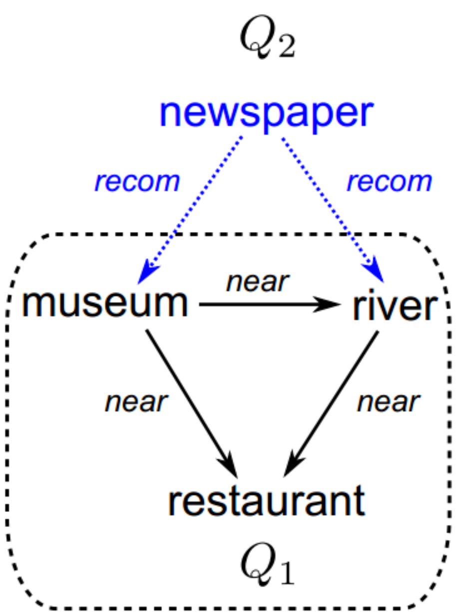
It is in O(|Q||G|) time to compute taxonomy simulation

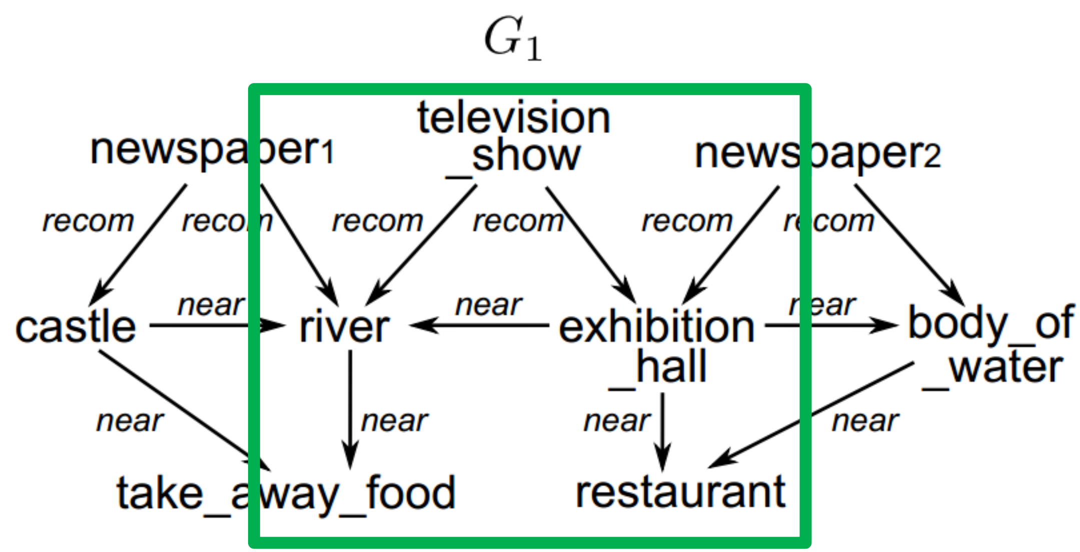
comes with no price w.r.t graph simulation!

Taxonomy simulation

- An experiment (*percentage of patterns with non-empty match results*)

$\lvert V_Q \rvert$	**2**	**4**	**6**	**8**	**10**
DBpedia	90%	18%	0%	0%	0%
YAGO	54%	2%	0%	0%	0%





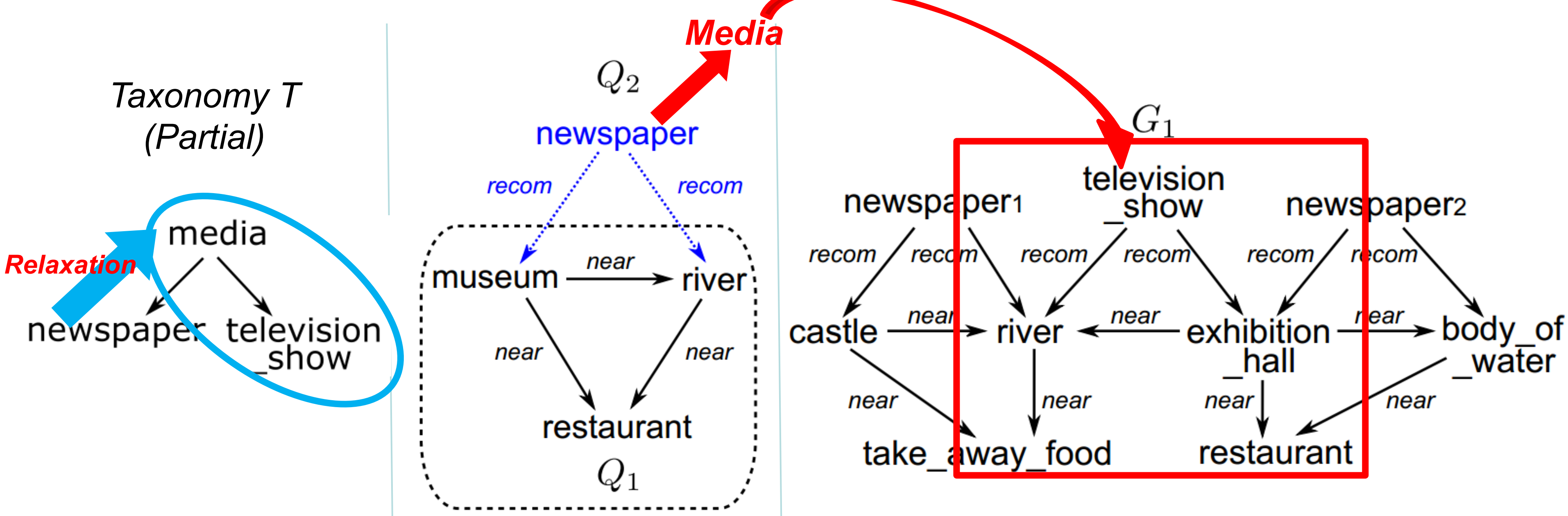
We need to further relax taxonomy simulation for larger patterns

Taxonomy simulation relaxation

- Label relaxation

A *label relaxation δ w.r.t.* a taxonomy *T* is of form $l \rightarrow l'$ such that l' is an ancestor label of l in *T*.

- Pattern relaxation
  - A *pattern relaxation* Δ for Q *w.r.t. T* is a set of label relaxations for Q.
  - Q $\oplus$ Δ is the *relaxed pattern* derived from Q by applying Δ.



A relaxation framework

- **Ranking top-k relaxations.**
- **Evaluating top-k relaxations.**
- **Relaxation explanation.**

Ranking top-k relaxations

Topological Ranking Function

Diversified Topological Ranking Function

Relaxation Ratio: similarity distance to *Q*

Information Ratio: ability to capture answers

Diversification Ratio: dissimilarity distance among top-k relaxations

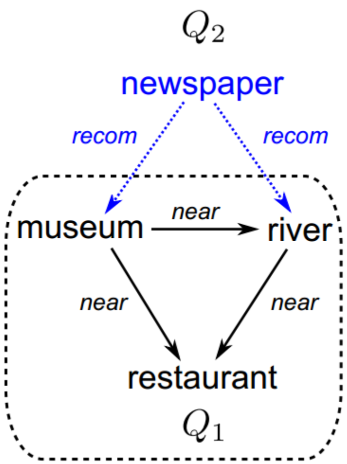
- Problems:
  - Top-k pattern relaxation problem (kPR): topological ranking
  - Diversified top-k relaxation problem(kPR_{DF}): diversified topological ranking

- Results:
  - kPR problem is in PTIME: in quadratic time, adopt *Lawler's procedure* for computing top-k results
  - kPR_{DF} problem is NP-hard and APX-hard: reduction to well-solved *maximum dispersion problem* (maxDP)

Evaluating top-k relaxations

- Problem:

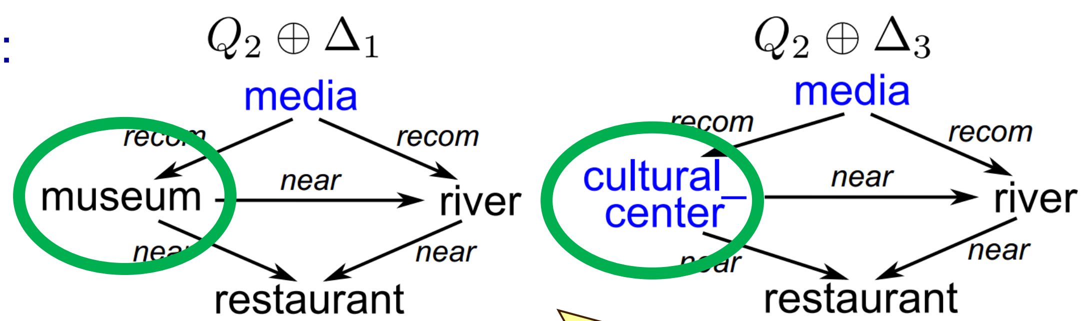
Given Q, G, T and k pattern relaxations $\Delta_1, \ldots, \Delta_k$, we aim to compute answers to the relaxed patterns $Q \oplus \Delta_1, \ldots, Q \oplus \Delta_k$ in G *w.r.t.* T.



Pattern relaxations:

$$\Delta_1 = \{ \delta_1 \}$$

$$\Delta_3 = \{ \delta_1, \delta_2 \}$$



Almost the same

Label relaxations:

δ_1 = newspaper → media

δ_2 = museum → cultural_center

δ_3 = river → natural_place

δ_4 = river → body_of_water

$Q_2 \oplus \Delta_1(G) \subseteq Q_2 \oplus \Delta_3(G)$

$Q_2 \oplus \Delta_1(G)$ can be derived from $Q_2 \oplus \Delta_3(G)$ via ***bounded decremental taxonomy simulation***

One pass of evaluation to compute both!

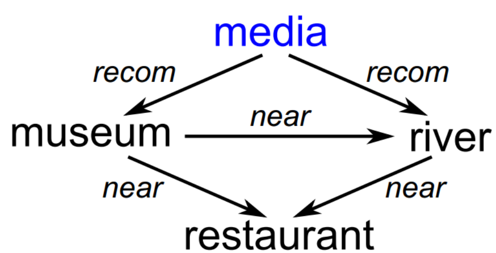
Evaluating top-k relaxations

- Problem:

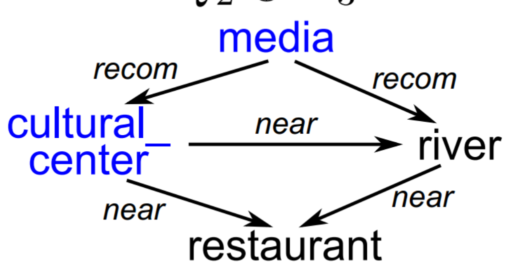
Given Q, G, T and k pattern relaxations $\Delta_1, \ldots, \Delta_k$, we aim to compute answers to the relaxed patterns $Q \oplus \Delta_1, \ldots, Q \oplus \Delta_k$ in G *w.r.t.* T.

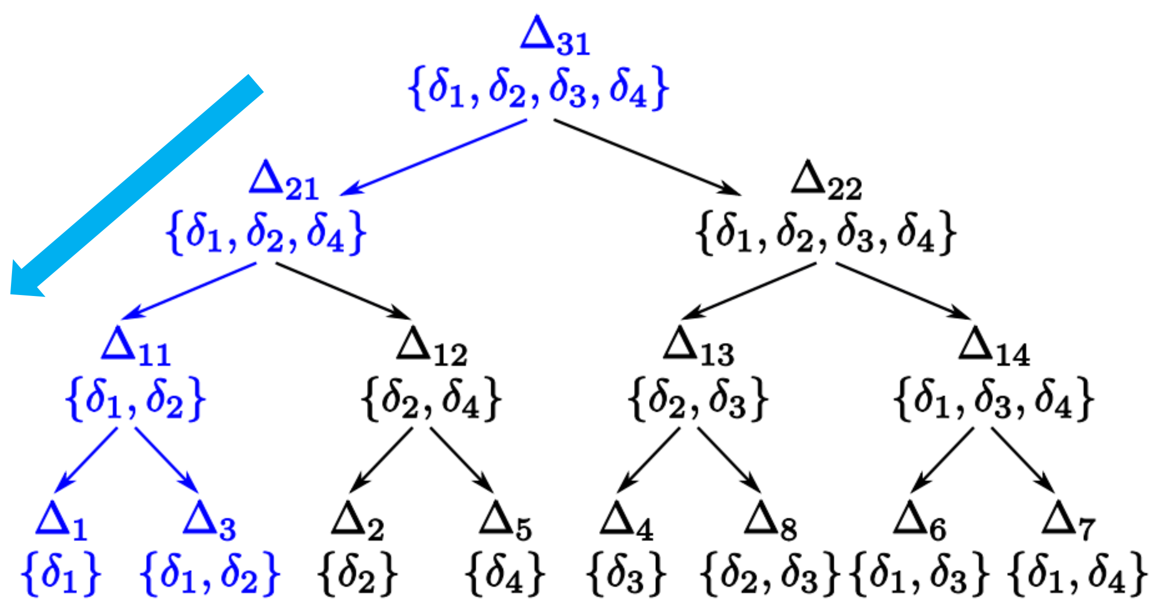
- An algorithm to maximize computation sharing
    - Minimum pairing tree construction
    - Bounded decremental evaluation

$Q_2 \oplus \Delta_1$



$Q_2 \oplus \Delta_3$





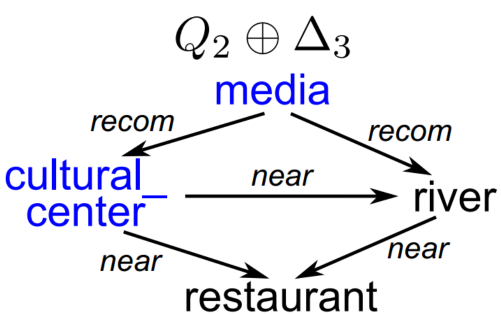
Relaxation Explanation

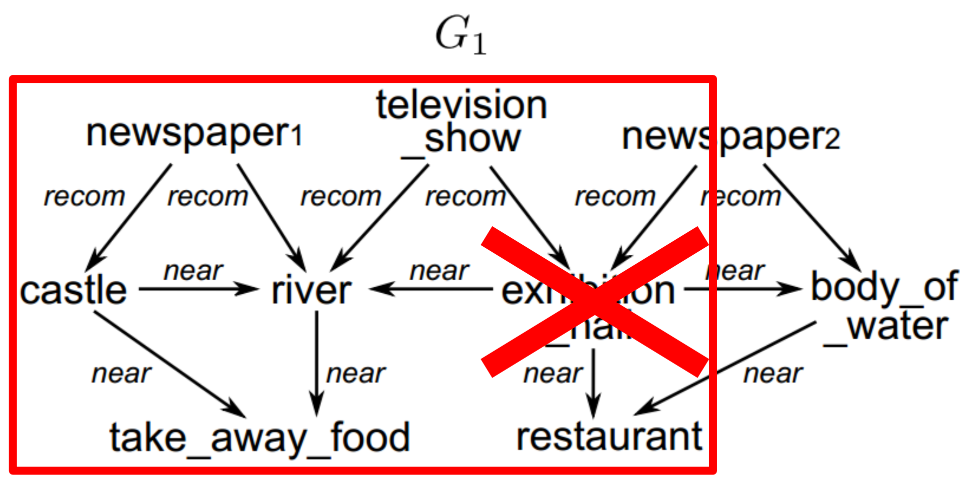
Can we explain why we return a match by relaxation?

- Explanation:

Given data graph *G*, pattern *Q*, taxonomy *T* , pattern relaxation Δ, and a node *v* in *G* that is in the match result $(Q \oplus \Delta)(G)$ to the relaxed pattern $Q \oplus \Delta$,

an explanation for *v w.r.t.* Δ, denoted by $E\Delta(v)$, is a subset of Δ such that *v* is in $(Q \oplus E\Delta(v))(G)$.





$\Delta_3 = \{ \cancel{\delta_1 = \text{newspaper} \rightarrow \text{media}},$

$\delta_2 = \text{museum} \rightarrow \text{cultural\_center} \}$

$\mathbf{E\Delta_3}$ **(** ***exhibition_hall*** **) = {** δ_1 **}**

Relaxation Explanation

Can we explain why we return a match by relaxation?

- Explanation:

Given data graph G, pattern Q, taxonomy T , pattern relaxation Δ, and a node v in G that is in the match result $(Q \oplus \Delta)(G)$ to the relaxed pattern $Q \oplus \Delta$,

an *explanation* for v *w.r.t.* Δ, denoted by $E\Delta(v)$, is a subset of Δ such that v is in $(Q \oplus E\Delta(v))(G)$.

- Problem:

Input: G, Q, T, Δ, v.

Output: minimum explanation for v in Δ.

Instances: MRE_{TF} , MRE_{DF}

- Results:
  - MRE_{TF}: optimal linear algorithm
  - MRE_{DF}: NP-hard, parameterized algorithm by M

Experimental setting

- Real-life graphs:

(1) YAGO:

data graph: (5.13M, 5.39M),

taxonomy graph: a forest with 6488 nodes, average height 3.27 (maximum height 13)

(2) DBpedia:

data graph: (4.43M, 8.43M),

taxonomy graph: a forest with 735 nodes, average height 2.29 (maximum height 6)

- Pattern graphs:

implement a generator for producing random pattern graphs $Q(V_Q, E_Q, f_Q)$, controlled by 3 parameters: $|V_Q|$ varying from 2 to 10, $|E_Q| = \lfloor \alpha |V_Q| \rfloor$, and the number $\lfloor \beta |V_Q| \rfloor$ of labels

Effectiveness of taxonomy simulation and relaxation

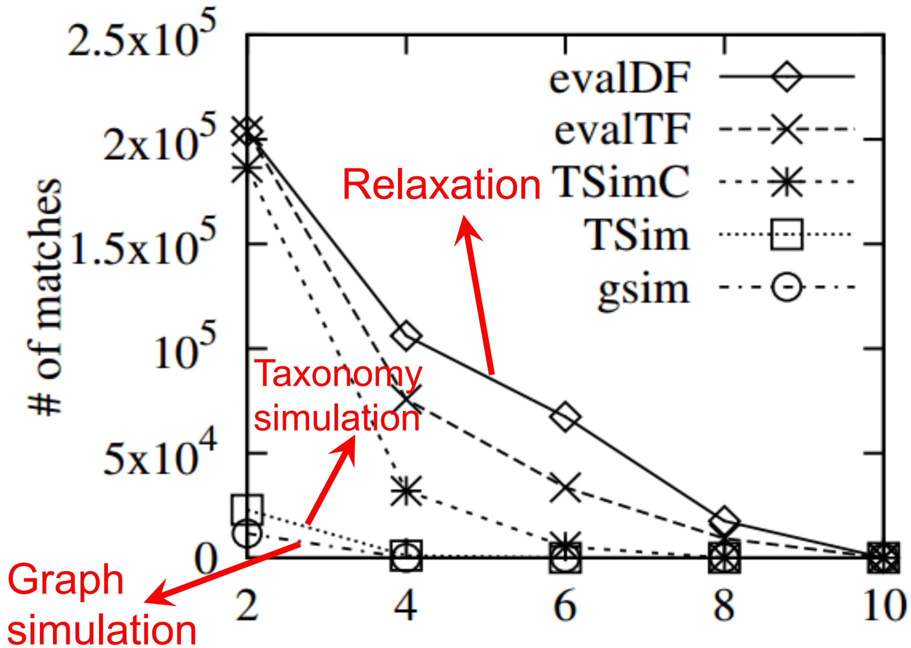
- Quality

$$\mathrm{acc}(S, Q, G) = \sum_{(u,v)\in S} \mathrm{valid}(u, v) / |S|$$

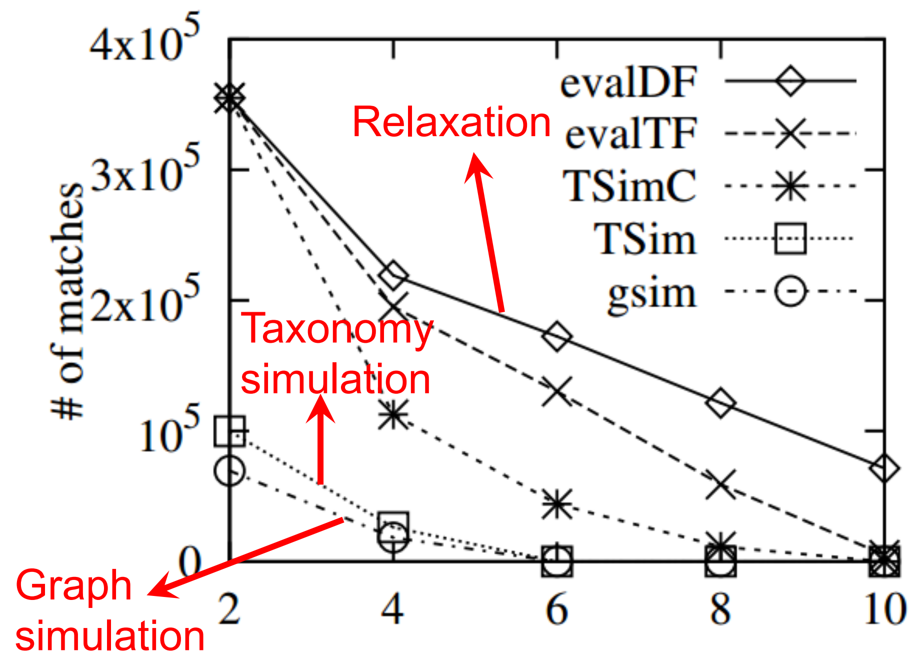
- *DBpedia*
  - *Taxonomy simulation: 98%*
  - *Relaxations: 77%*
- *YAGO*
  - *Taxonomy simulation: 94%*
  - *Relaxations: 71%*

Effectiveness of taxonomy simulation and relaxation

➢ Quantity (number of matches vs. $|V_Q|$)



Varying $|V_Q|$ (YAGO)



Varying $|V_Q|$ (DBpedia)

Taxonomy simulation vs. graph simulation

- 1,116 vs 0 ($|V_Q|$=4)

Relaxation vs. taxonomy simulation

Taxonomy simulation vs. graph simulation

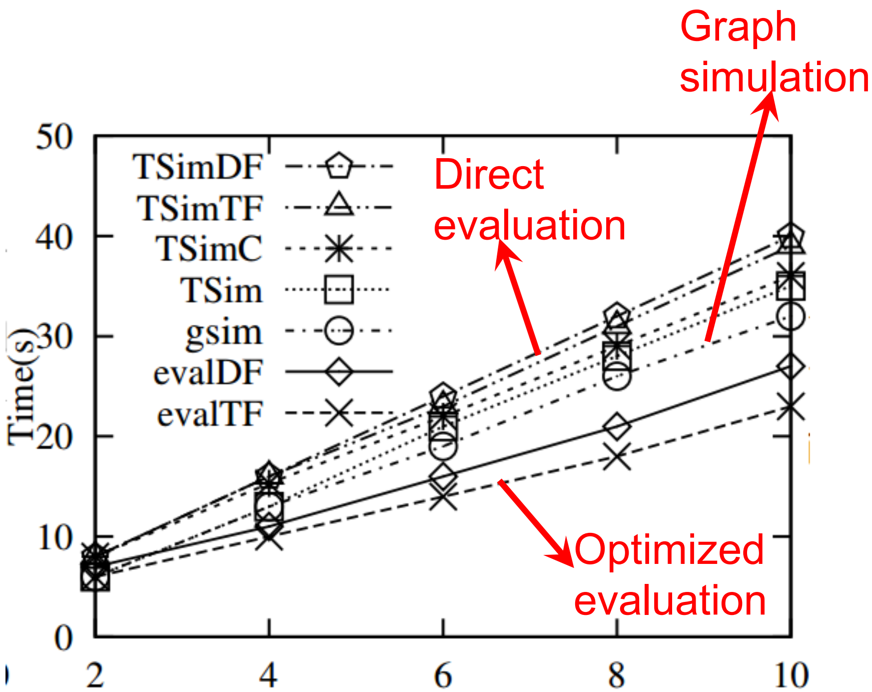
- 26,242 vs 18,384 ($|V_Q|$=4)

Relaxation vs. taxonomy simulation

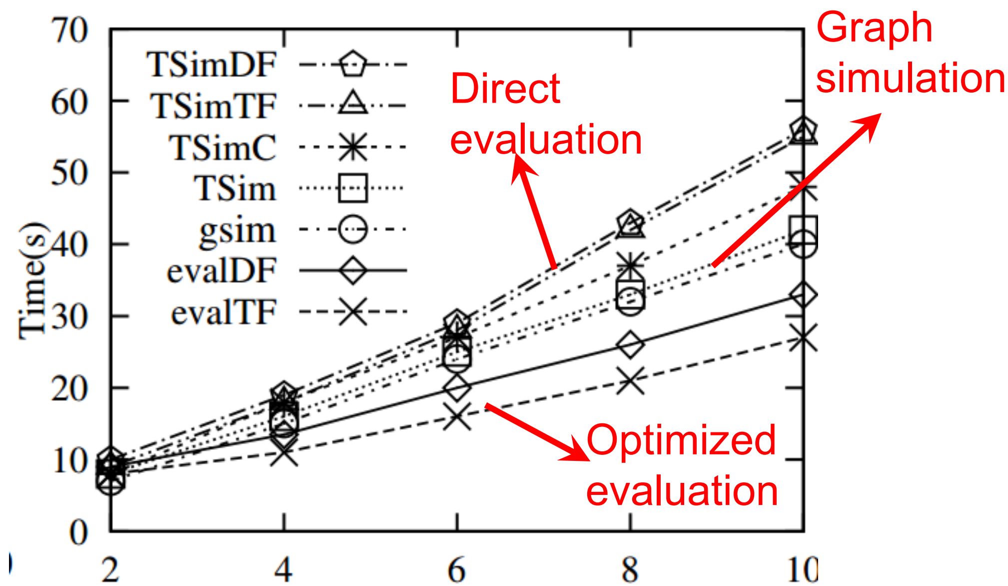
•

$|V_Q| \leq 4$: Taxonomy simulation; $|V_Q| > 4$: Relaxation

Efficiency of relaxation



(d) Varying $|V_Q|$ (YAGO)



(c) Varying $|V_Q|$ (DBpedia)

Direct evaluation / optimized evaluation

- 1.57 times faster ($|V_Q|$=6, YAGO)
- 1.62 times faster ($|V_Q|$=6, DBpedia)

Summary

A framework for relaxing graph pattern matching queries

- Taxonomy simulation by combining taxonomy with graph simulation
- Relaxation framework for taxonomy simulation
    - Ranking functions for taxonomy simulation patterns
    - Computing top-k relaxed patterns
    - Evaluating top-k relaxed patterns
    - Relaxation explanation

Thanks!

Subgraph isomorphism and graph simulation

- ✓ Subgraph isomorphism: Graph G matches pattern Q via subgraph isomorphism denoted by $Q \lhd G$, if there exists a subgraph G_s of G that is isomorphic t NP-hard there exists a bijection *h* from V_Q to V_s, such that

  (a) edge $(u,u') \in E_Q$ if and only if $(h(u),h(u')) \in E_s$; (b) for each $u \in V_Q$, $l_Q(u)=l(h(u))$.

- ✓ Graph simulation: Graph G matches pattern Q via graph simulation, denoted by $Q \prec G$, if there exists a binary match relation $R \subseteq V_Q \times V$ such that

  Quadratic time

  (a) for each $(u,v) \in R$, $l_Q(u)=l(v)$;

  (b) for each $u \in V_Q$, there exists $v \in V$, such that (i) $(u,v) \in R$, and (ii) for any edge (u,u') in Q, there exists an edge (v,v') in G such that $(u',v') \in R$.