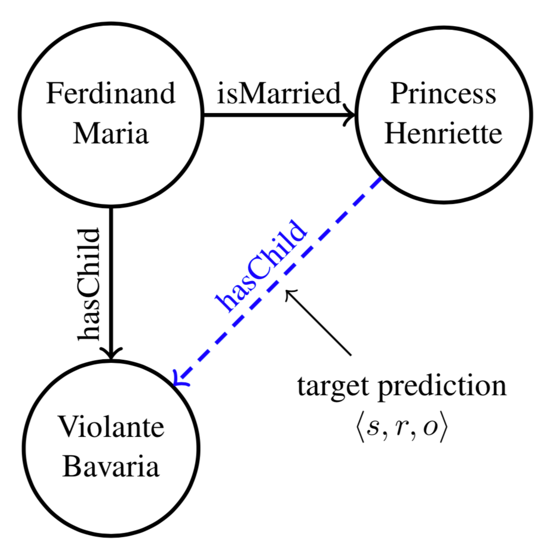
Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications

Pouya Pezeshkpour Yifan Tian Sameer Singh

Graph Embeddings for Link Prediction

In this work, we propose efficient adversarial modifications for link prediction models to evaluate robustness, and study interpretability and error correcting.

Completing Knowledge Graphs: Predicting a missing link from observed graph structure.



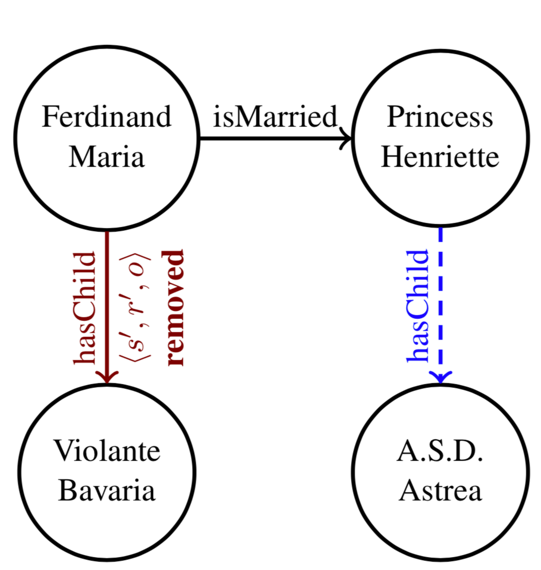
- Existing models:
  - Embed *s*, *r*, and *o*
  - Maximize score $\psi(s, r, o)$ for observed facts
    - DistMult: $e_s R e_o$
    - ConvE: $f(vec(f([\overline{e_s}; \overline{r_r} * w]))W)e_o$
- Embeddings are inscrutable…
  - Are these embeddings robust to small changes?
  - Can we explain why a fact/link was predicted?

Adversarial Modifications (CRIAGE)

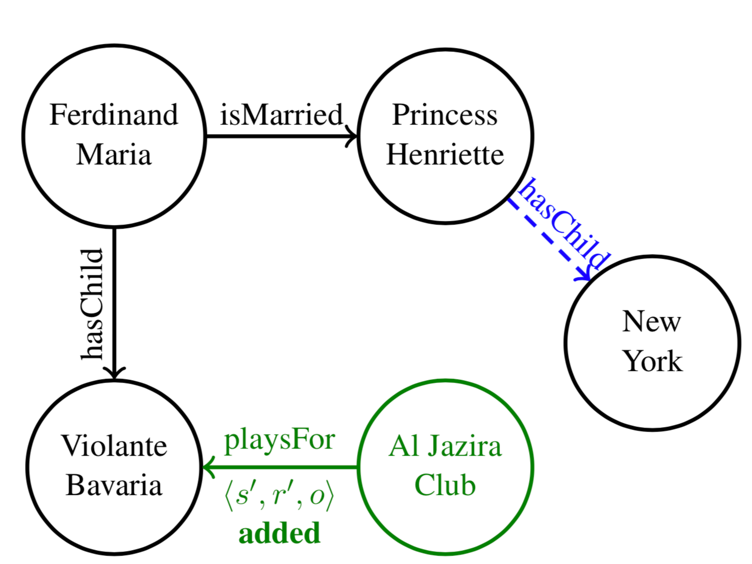
- Completion Robustness and Interpretability via Adversarial Graph Edits (CRIAGE)

Minimally change the graph so that target fact prediction changes the most after embeddings are relearned.

Removing an existing link



Adding a fake link



Efficiently Identifying the Modification

For target triple $< s, r, o >$ and graph G, we identify:

- Removing / Adding : Find (s', r', o) such that score $\psi(s, r, o)$ trained on G is maximally different from score $\overline{\psi}(s, r, o)$ trained after removing or adding (s', r', o) :

$$\underset{(s',r')}{\operatorname{argmax}}\, \psi(s, r, o) - \overline{\psi}(s, r, o)$$

Two Primary Challenges:

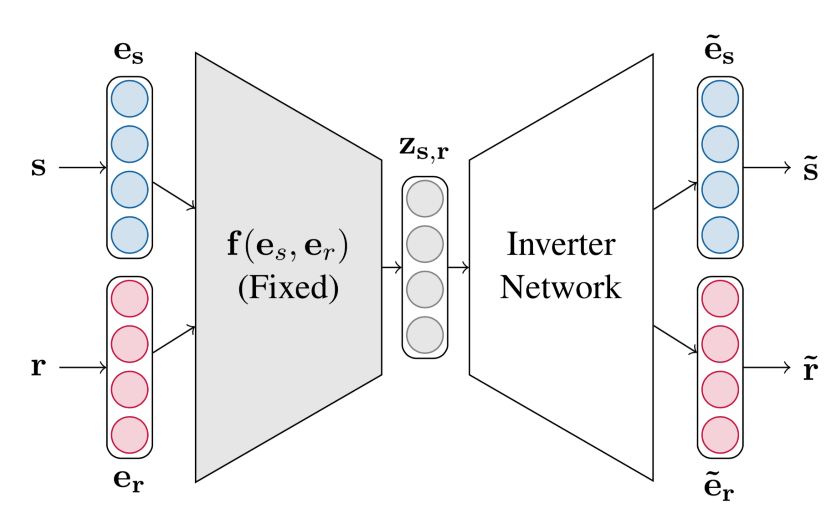
1. **Retraining is too expensive:** Taylor approximation on gradient of loss and utilizing graph structure.

$$\Delta(\nabla_e loss(e)) = H_e(loss) \times (e - \overline{e})$$

$e, \overline{e}$ = optimum embedding & H = Hessian

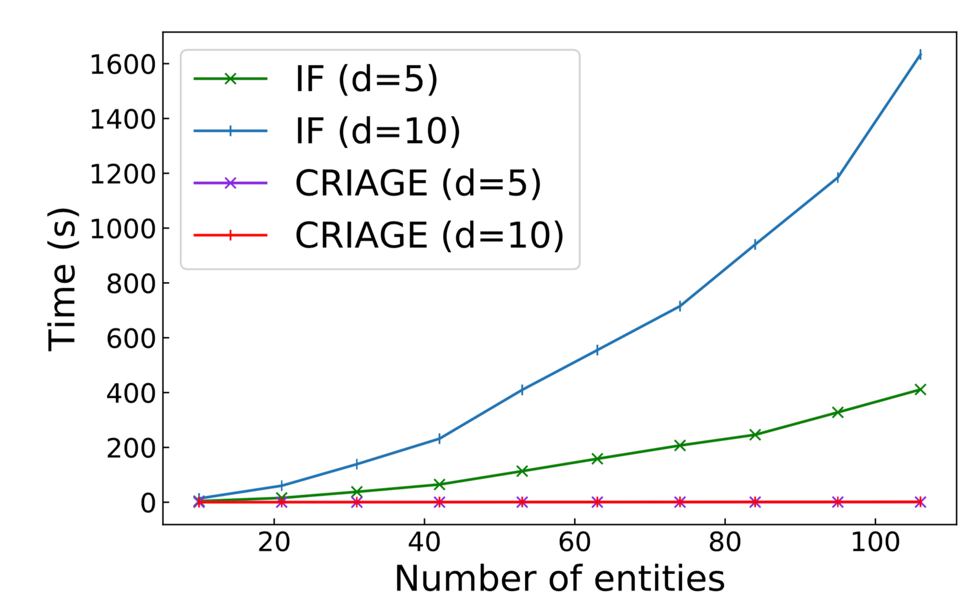
$$\Rightarrow \overline{e} = e - H_e(loss)^{-1} \times \Delta(\nabla_e loss(e))$$

2. **Too many links to search:** Learn a continuous space of links using an inverter, and use gradient descent.



CRIAGE vs Influence Functions

- Influence Functions (IF)*:
  - Similar motivation, but doesn't exploit graph structure

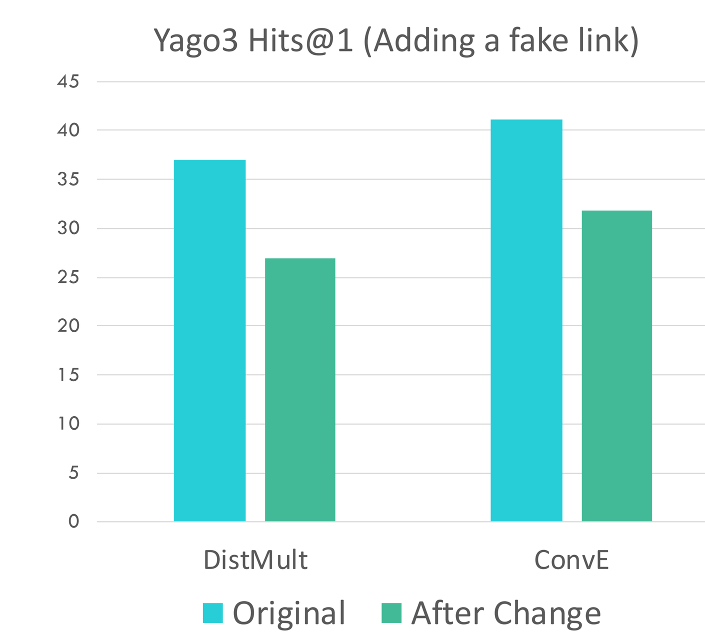


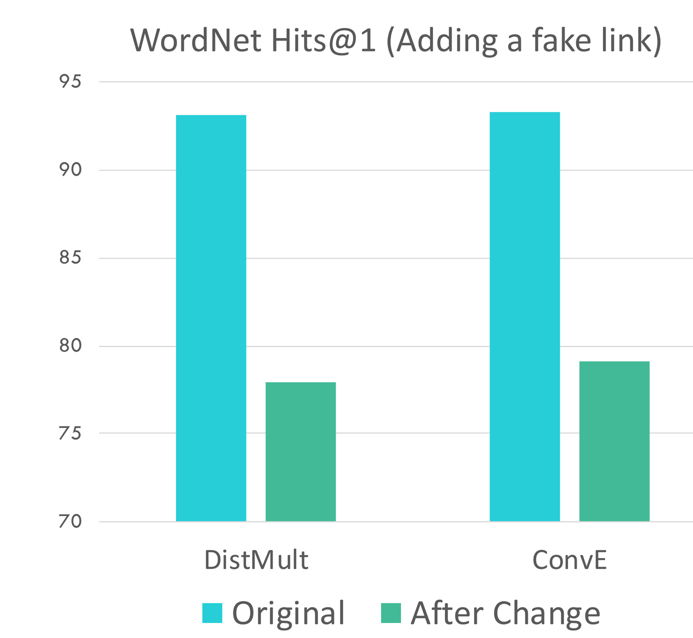
* Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions."

Robustness and Interpretability

Robustness

Does adding a fake link affect performance?





Interpretability

Which link, when removed, changes the prediction?
Find common patterns in removed link $R(a, b)$

DistMult and ConvE: isMarriedTo(a,c) ∧ hasChild(c,b) ⇒ hasChild(a,b)

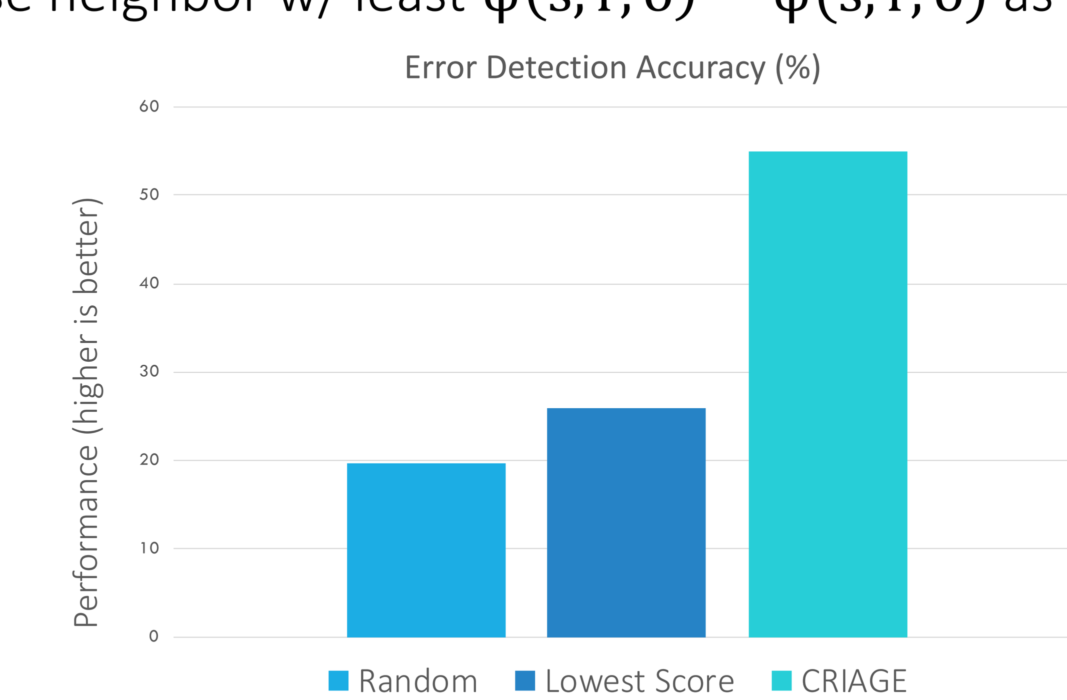
Only in DistMult:
playsFor(a,c) ∧ isLocatedIn(c,b) ⇒ wasBornIn(a,b)*
isAffiliatedTo(a,c) ∧ isLocatedIn(c,b) ⇒ diedIn(a,b)*

Only in ConvE:
hasAdvisor(a,c) ∧ graduatedFrom(c,b) ⇒ graduatedFrom(a,b)
influences(a,c) ∧ influences(c,b) ⇒ influences(a,b)

* Identified as rules by [Yang et. al. 2015]

Error Correcting

Introduce errors and see if we can detect it.
Choose neighbor w/ least $\psi(s, r, o) - \overline{\psi}(s, r, o)$ as incorrect.



For any questions, email: **pezeshkp@uci.edu**

Website: **https://pouyapez.github.io/criage**