Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

Peru Bhardwaj, John Kelleher, Luca Costabello, Declan O'Sullivan

Contact: peru.bhardwaj@adaptcentre.ie

EMNLP 2021

Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

Contact: peru.bhardwaj@adaptcentre.ie

PeruBhardwaj/AttributionAttack

Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

- Adversarial Deletions + Additions

Contact: peru.bhardwaj@adaptcentre.ie

PeruBhardwaj/AttributionAttack

Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

Instance Similarity | Gradient Similarity | Influence Function

- ✓ Adversarial Deletions + Additions
- ✓ Identify influential training examples

Contact: peru.bhardwaj@adaptcentre.ie

PeruBhardwaj/AttributionAttack

Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

Instance Similarity | Gradient Similarity | Influence Function

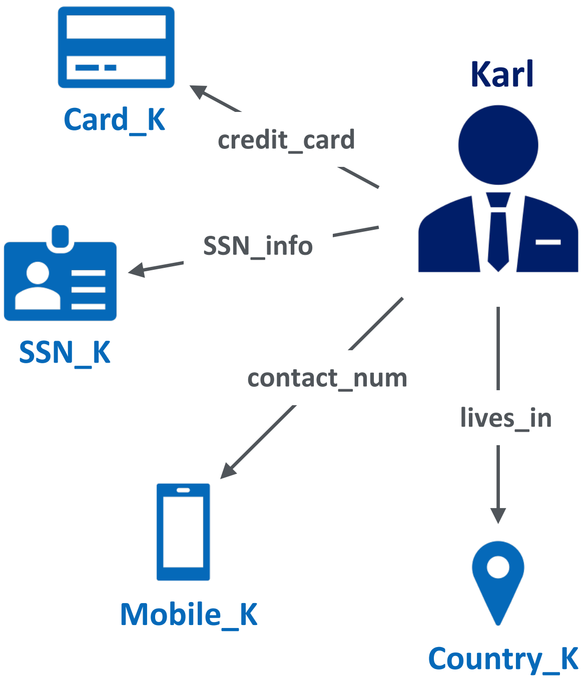
- Adversarial Deletions + Additions
- Identify influential training examples
- Outperform state-of-art attacks

Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

Adversarial Attacks on **Knowledge Graph** Embeddings via Instance Attribution Methods

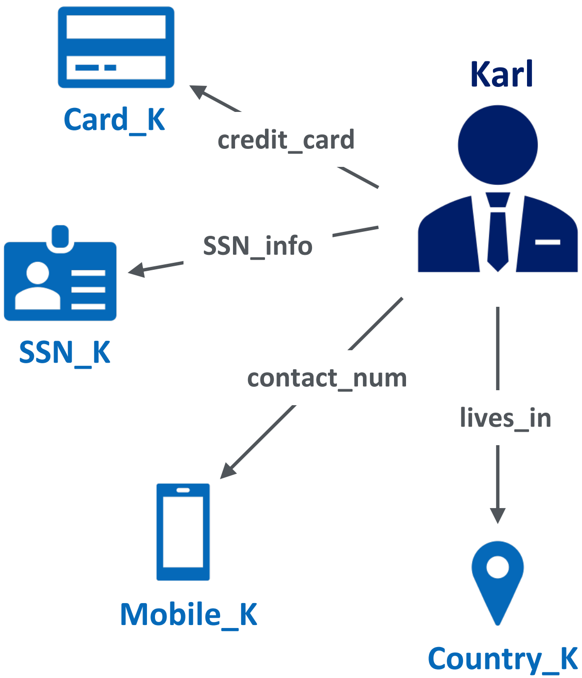
Knowledge Graph

Example - Financial Details of a Bank's Customer



Knowledge Graph

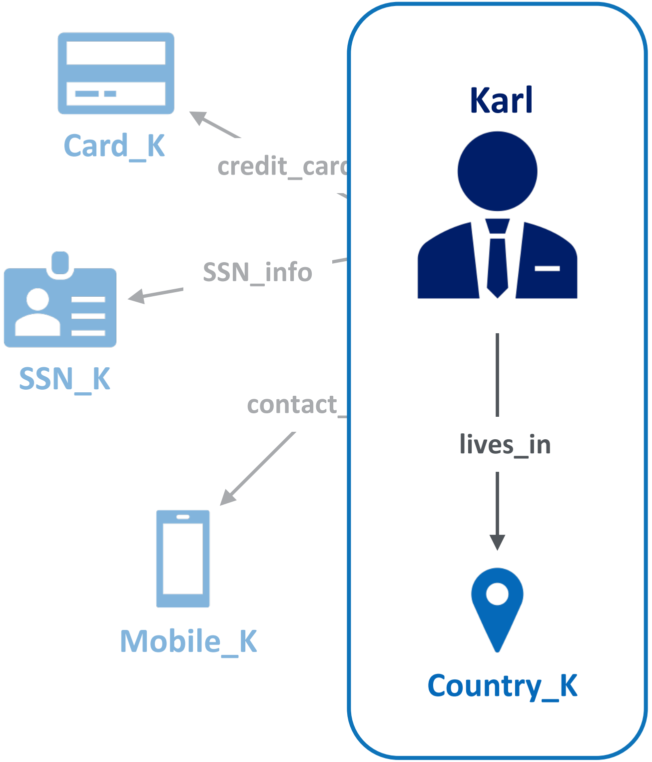
Example - Financial Details of a Bank's Customer



s	r	o
Karl	credit_card	Card_K
Karl	SSN_info	SSN_K
Karl	contact_num	Mobile_K
Karl	lives_in	Country_K

Knowledge Graph

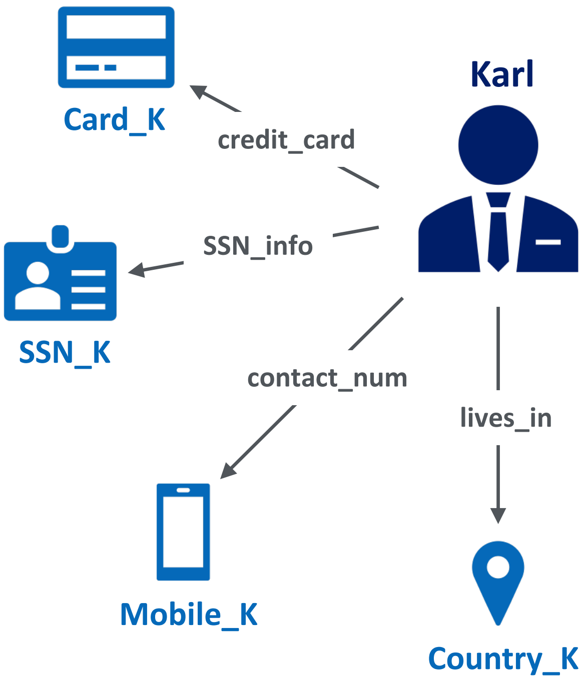
Example - Financial Details of a Bank's Customer



s	r	o
Karl	credit_card	Card_K
Karl	SSN_info	SSN_K
Karl	contact_num	Mobile_K
Karl	lives_in	Country_K

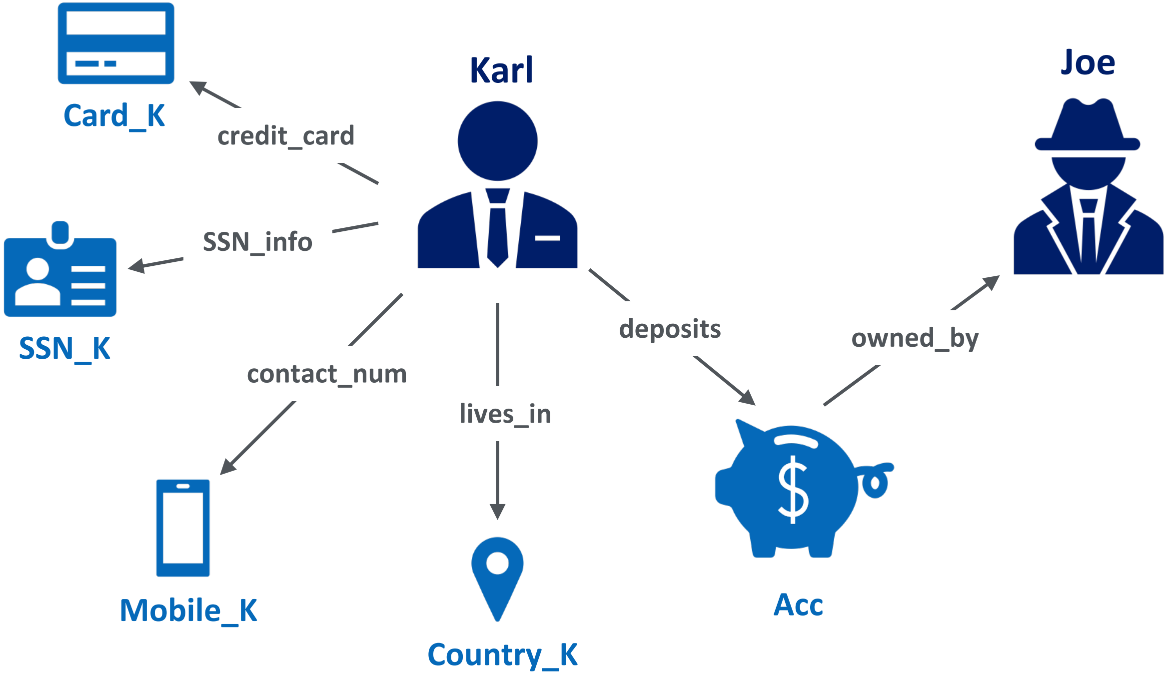
Knowledge Graph

Example - Financial Details of a Bank's Customer

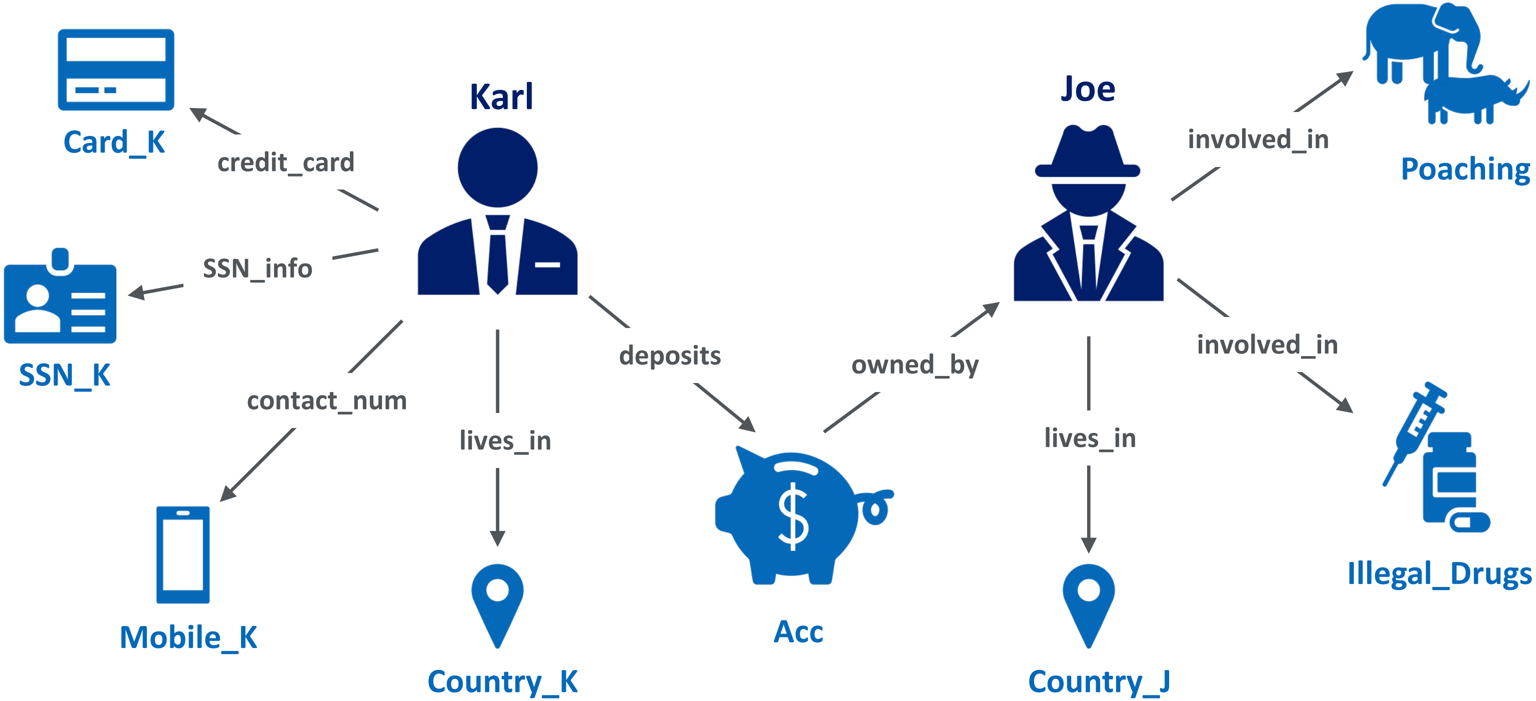


s	r	o
Karl	credit_card	Card_K
Karl	SSN_info	SSN_K
Karl	contact_num	Mobile_K
Karl	lives_in	Country_K

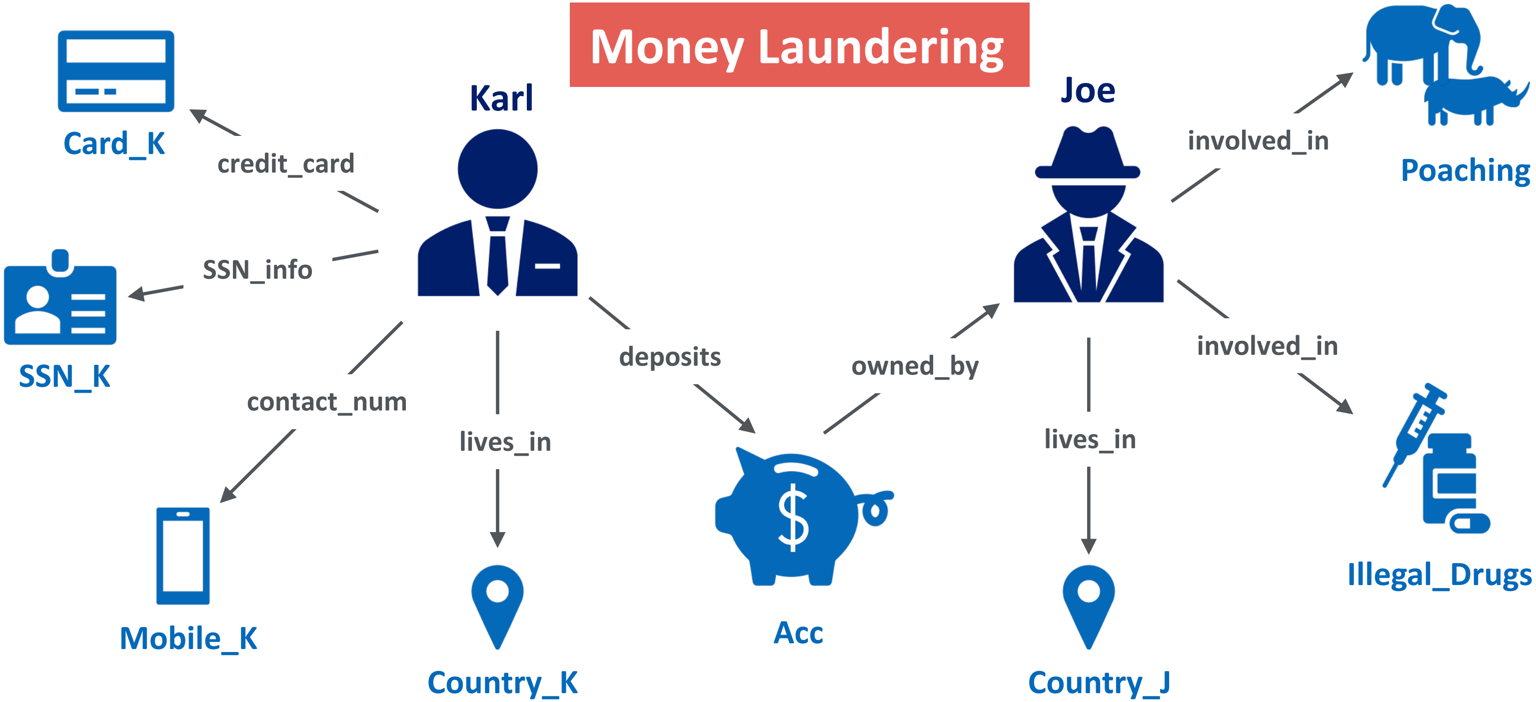
Missing Link Prediction



Missing Link Prediction

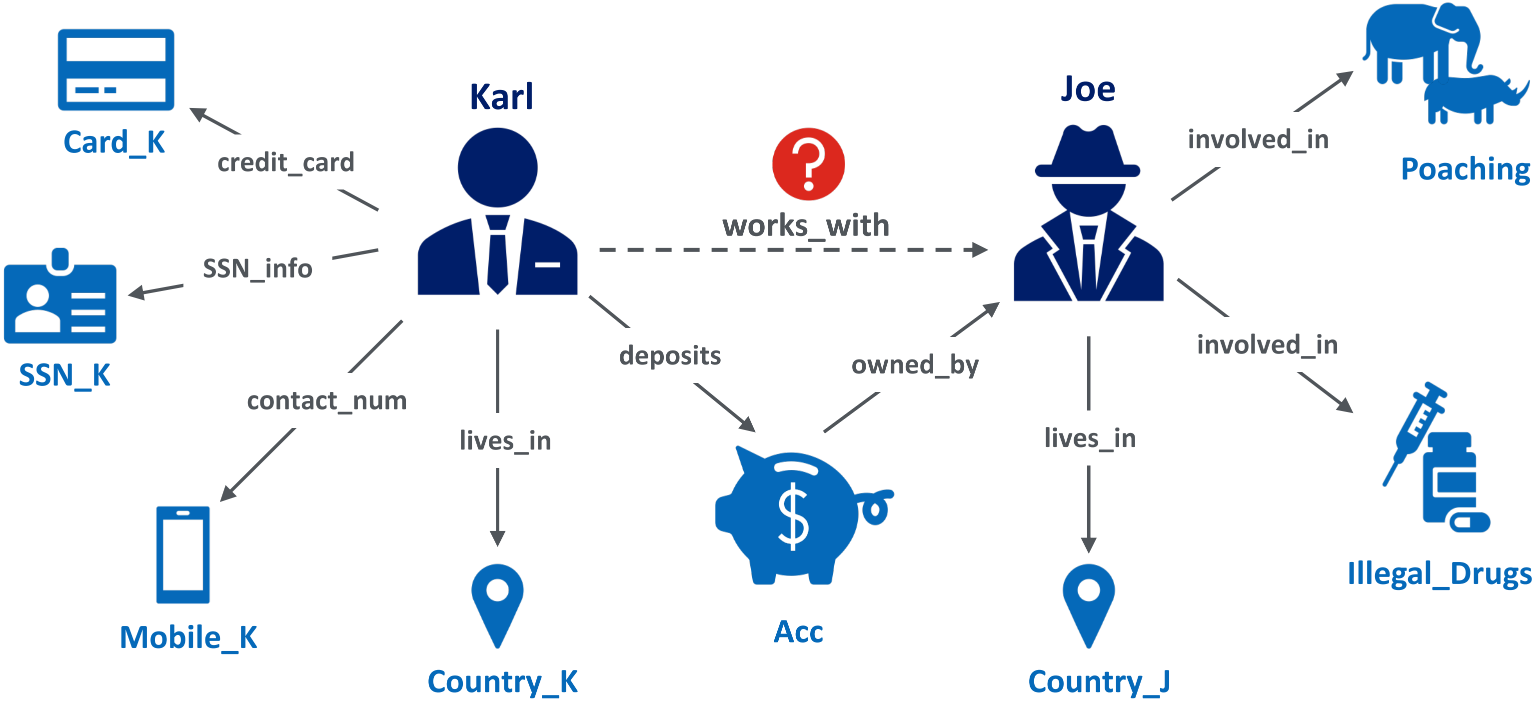


Missing Link Prediction



Missing Link Prediction

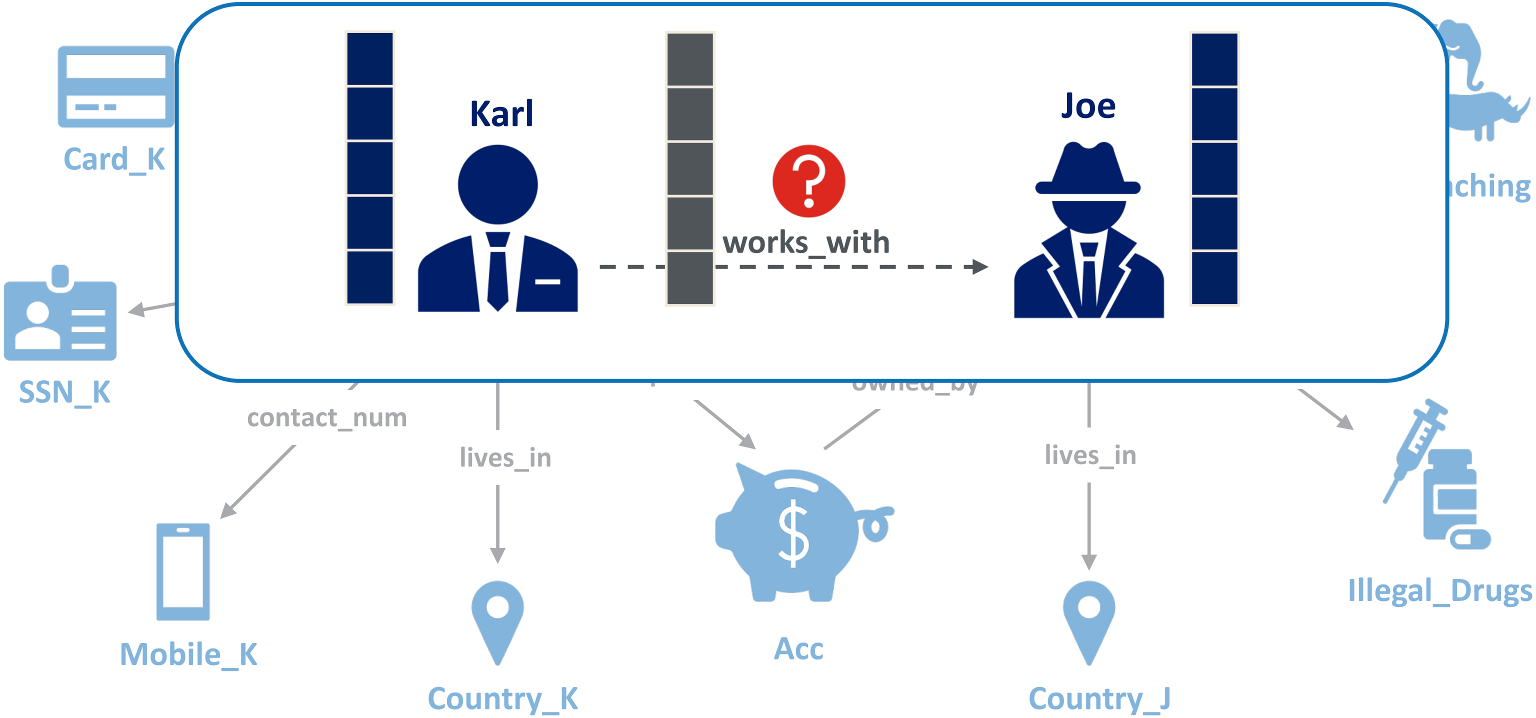
Use case – Anti Money Laundering



Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

Missing Link Prediction

Use case – Anti Money Laundering

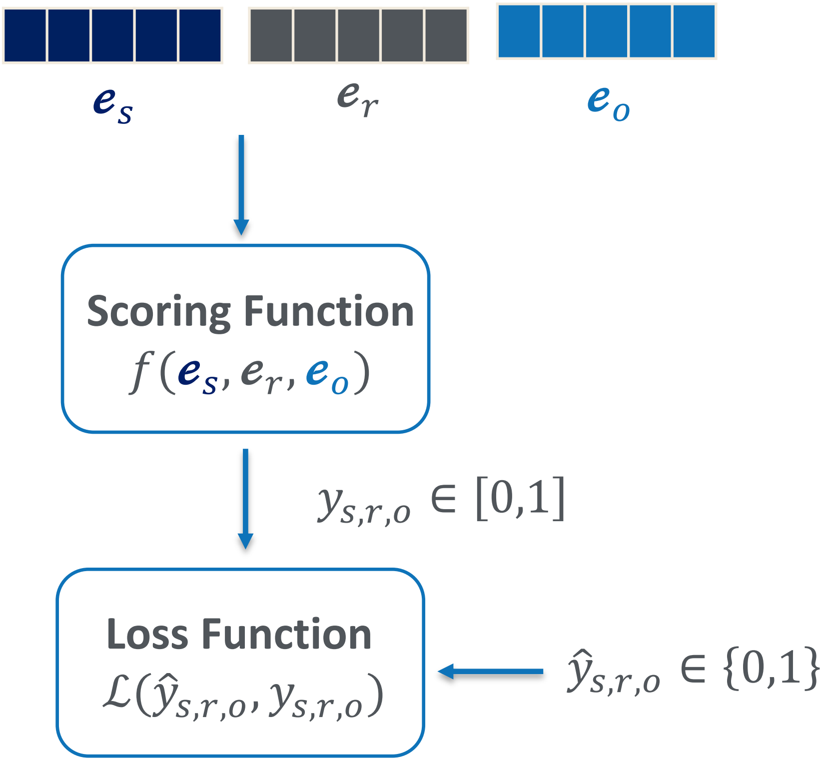


Knowledge Graph Embeddings (KGE)

Generate negatives by corrupting s/o

s	r	o
Karl	credit_card	**Card_X**
Karl	credit_card	**Card_Y**
Karl	**credit_card**	Card_K
Person_X	credit_card	Card_K
Person_Y	credit_card	Card_K

Knowledge Graph Embeddings (KGE)



Generate negatives by corrupting s/o

s	**r**	**o**
Karl	credit_card	**Card_X**
Karl	credit_card	**Card_Y**
Karl	**credit_card**	Card_K
Person_X	credit_card	Card_K
Person_Y	credit_card	Card_K

Knowledge Graph Embeddings (KGE)

Minimize $\mathcal{L}$ by updating $\boldsymbol{e}_s, \boldsymbol{e}_r, \boldsymbol{e}_o$

$\boldsymbol{e}_s$ $\boldsymbol{e}_r$ $\boldsymbol{e}_o$

Scoring Function

$f(\boldsymbol{e}_s, \boldsymbol{e}_r, \boldsymbol{e}_o)$

$y_{s,r,o} \in [0,1]$

Loss Function

$\mathcal{L}(\hat{y}_{s,r,o}, y_{s,r,o})$

$\hat{y}_{s,r,o} \in \{0,1\}$

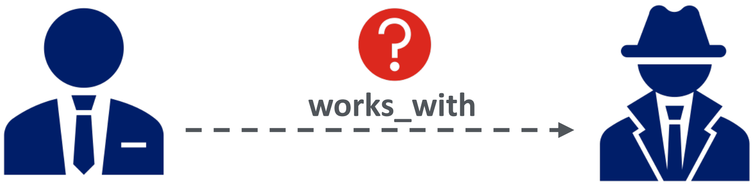
Scores for positive triples are higher than scores for negative triples

Generate negatives by corrupting s/o

s	r	o
Karl	credit_card	**Card_X**
Karl	credit_card	**Card_Y**
Karl	**credit_card**	Card_K
Person_X	credit_card	Card_K
Person_Y	credit_card	Card_K

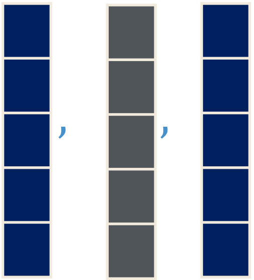
Missing Link Prediction with KGE

Use case – Anti Money Laundering



$$\mathcal{P}(Karl, works\_with, Joe) \propto f(\boldsymbol{e}_{Karl}, \boldsymbol{e}_{works\_with}, \boldsymbol{e}_{Joe})$$

$$\mathcal{P}(Karl, works\_with, Joe) \propto f(\quad , \quad , \quad)$$

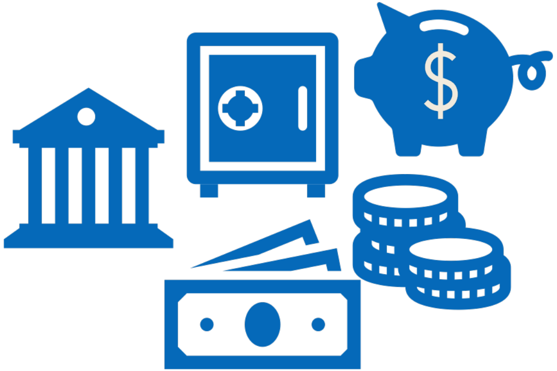


Bhardwaj et al. 2021. Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Metrics. In EMNLP

Where to find KGE in practice?

Security Sensitive

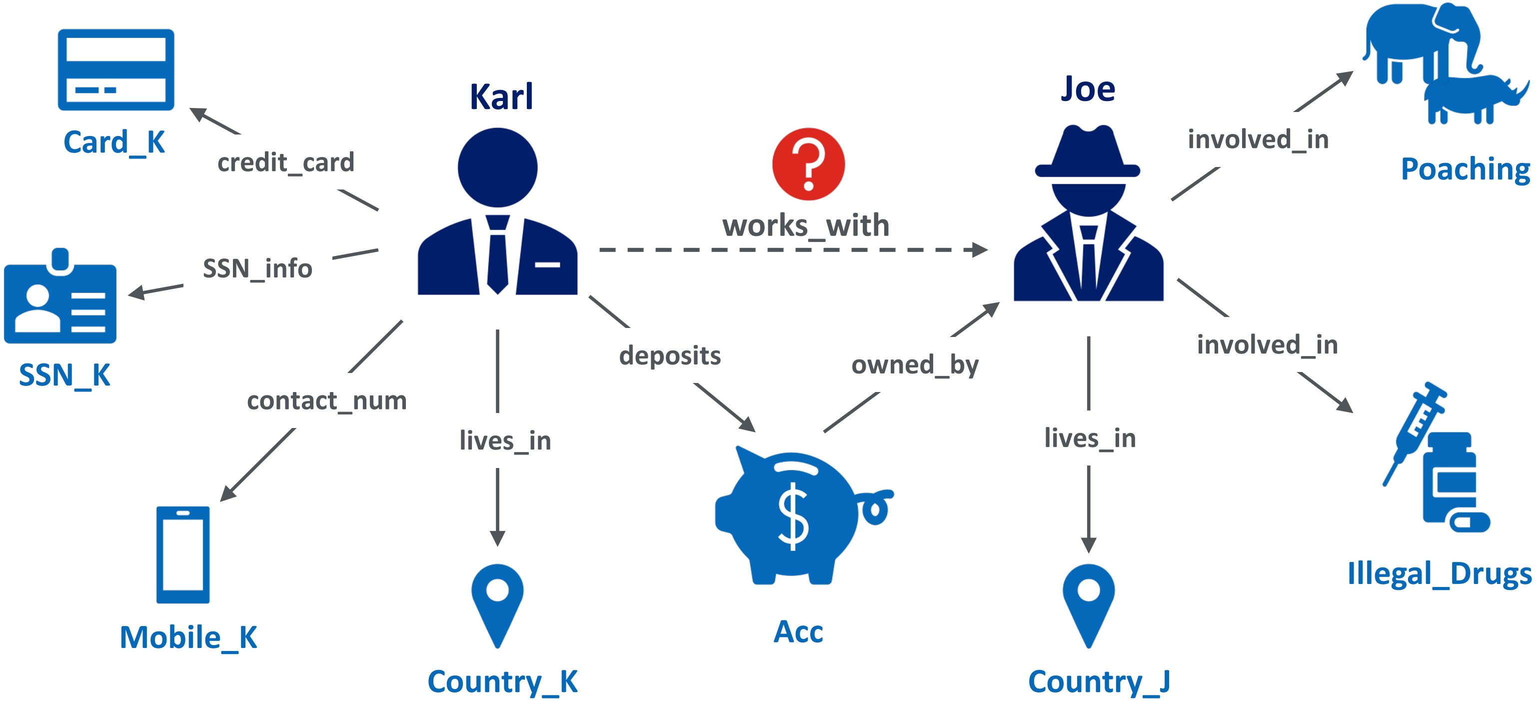
High Stakes



Bhardwaj et al. 2021. Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Metrics. In EMNLP

KGE in High-Stakes Applications

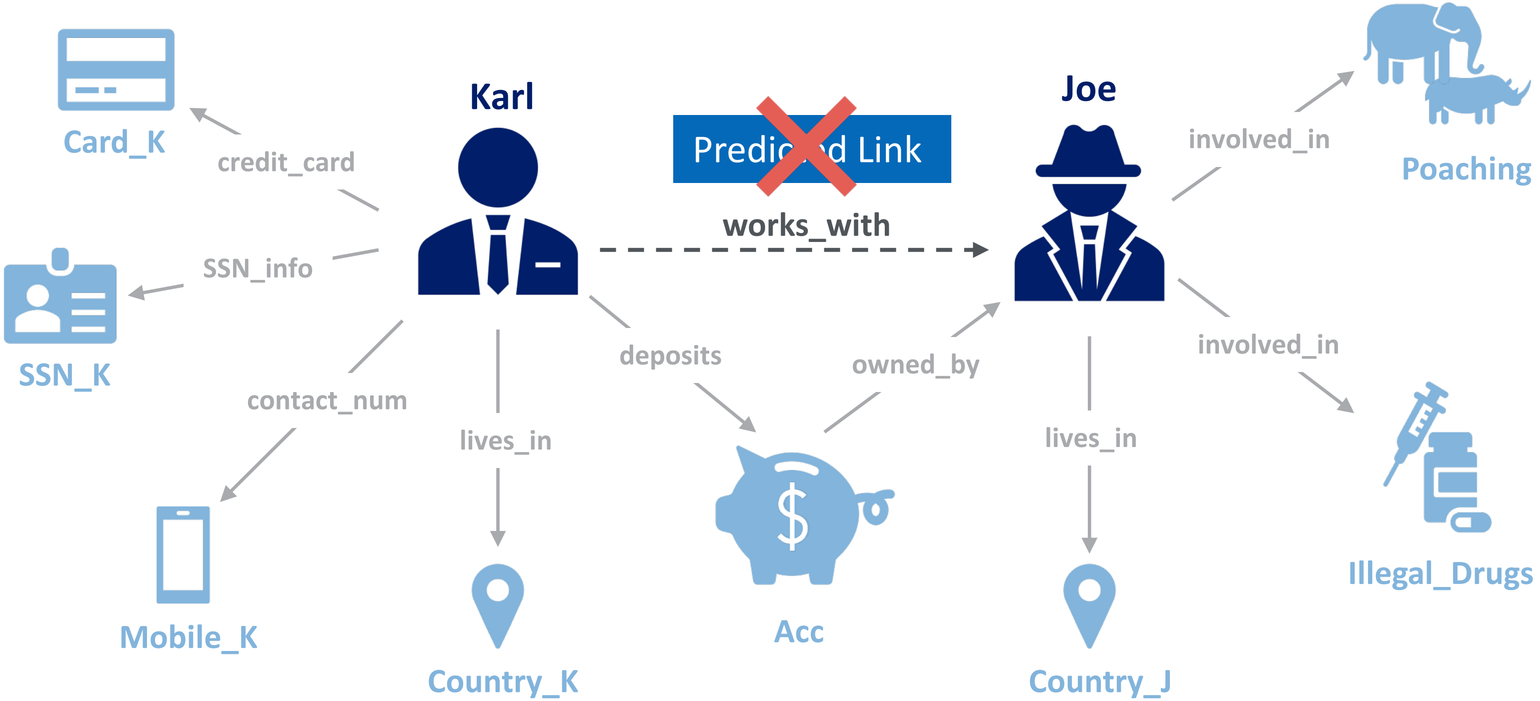
Use case – Anti Money Laundering



Bhardwaj et al. 2021. Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Metrics. In EMNLP

KGE in High-Stakes Applications

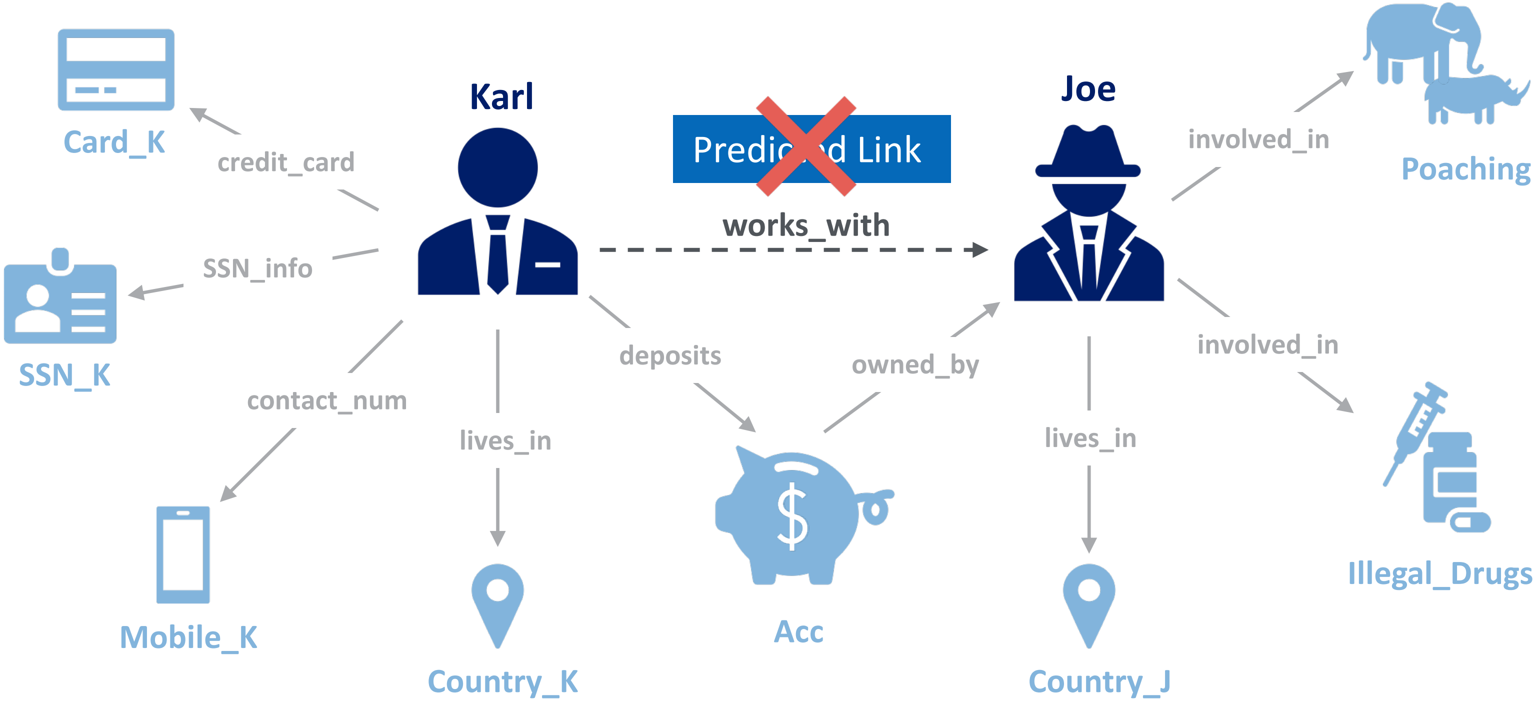
Incentives for bad actors!



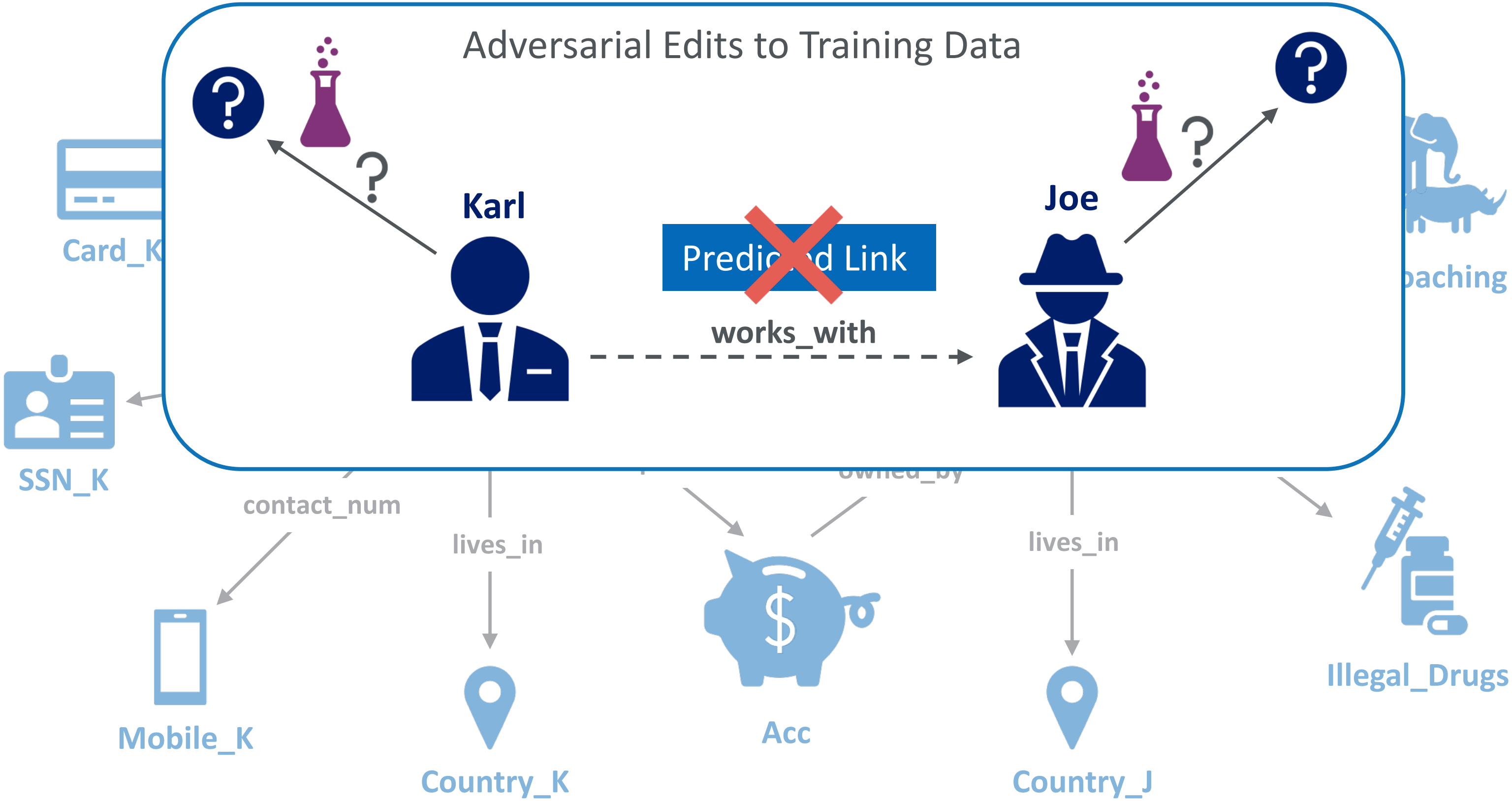
Adversarial Attacks on Knowledge Graph Embeddings
via Instance Attribution Methods

Adversarial Attacks on KGE

Aim – Degrade the prediction on target triple

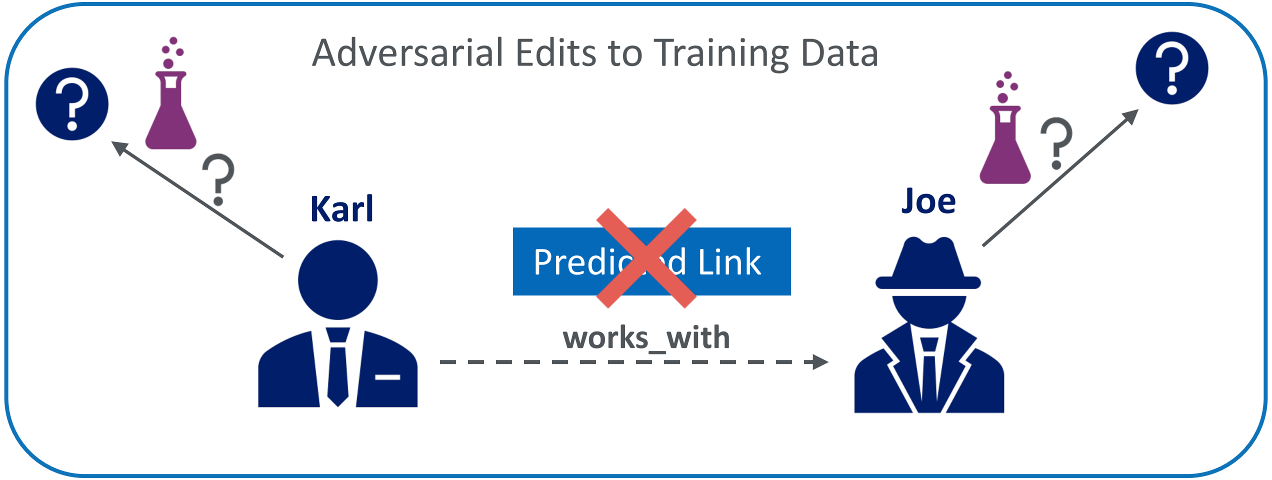


Adversarial Attacks on KGE



Adversarial Attacks on KGE

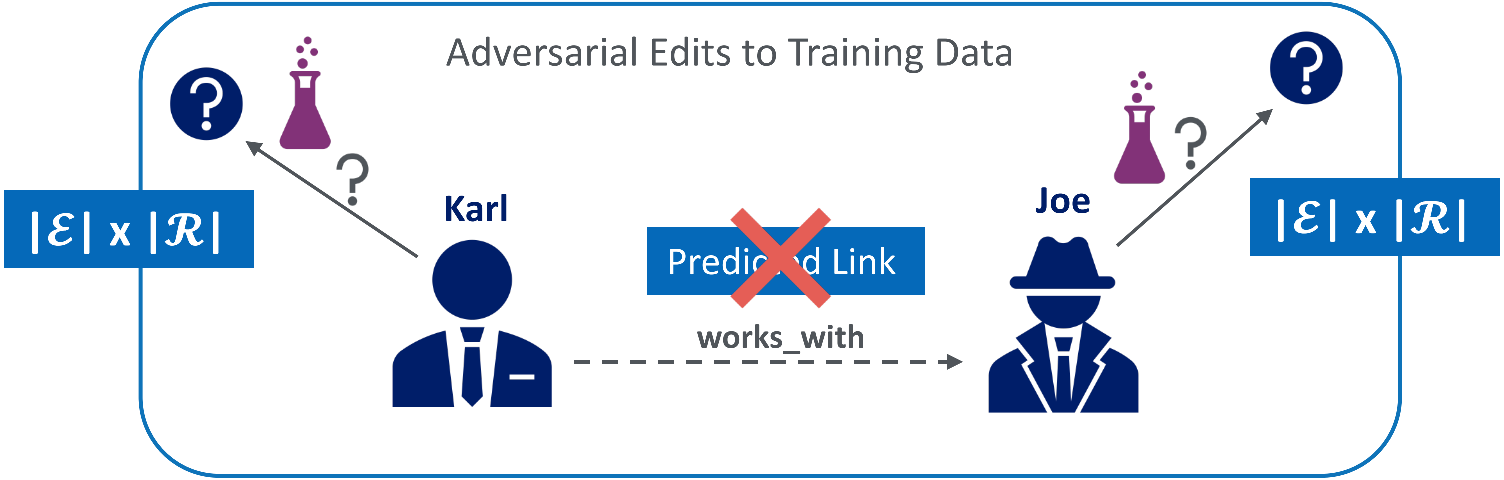
Challenge – Metric for adversarial impact



How to measure the impact of a candidate adversarial perturbation on the prediction of target triple?

Adversarial Attacks on KGE

Challenge – Large Search Space

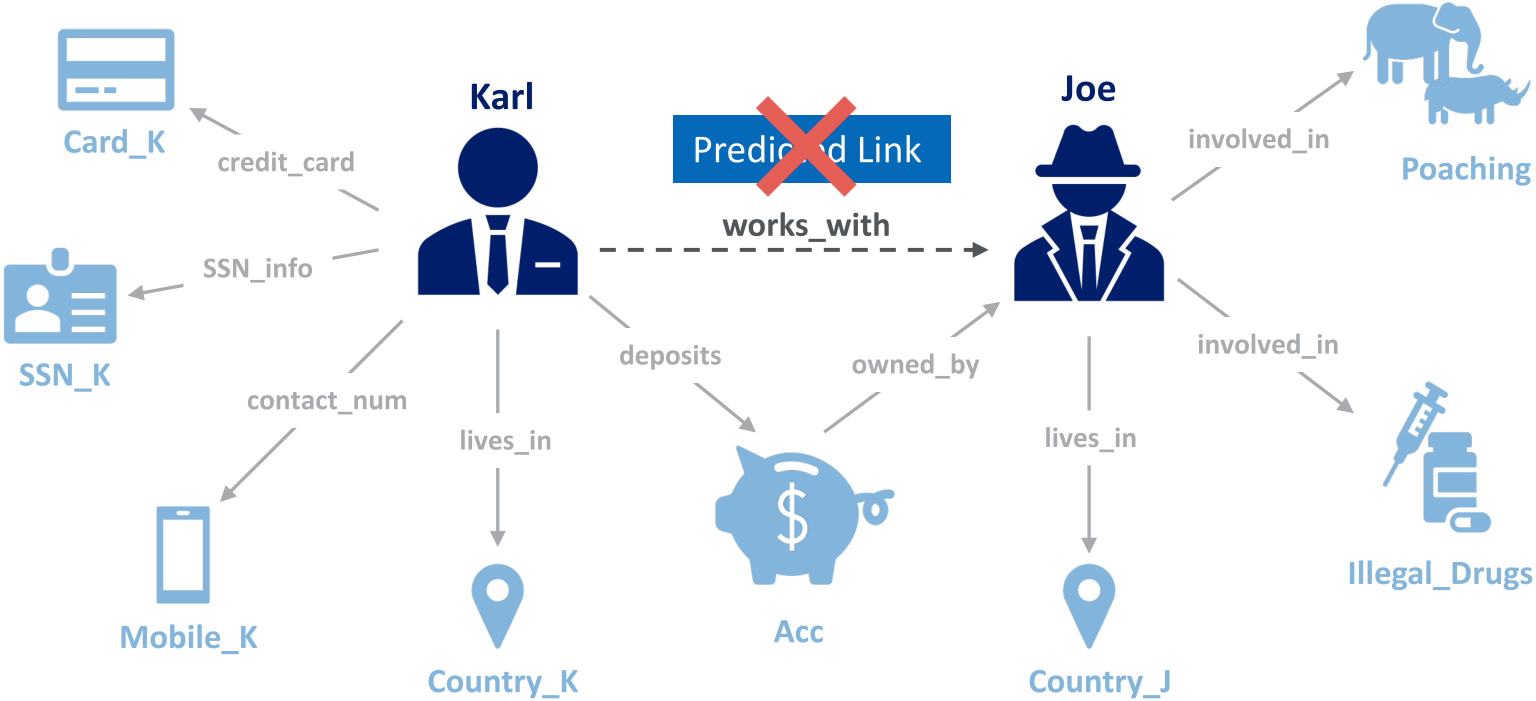


How to search through the combinatorial space of candidate adversarial additions?

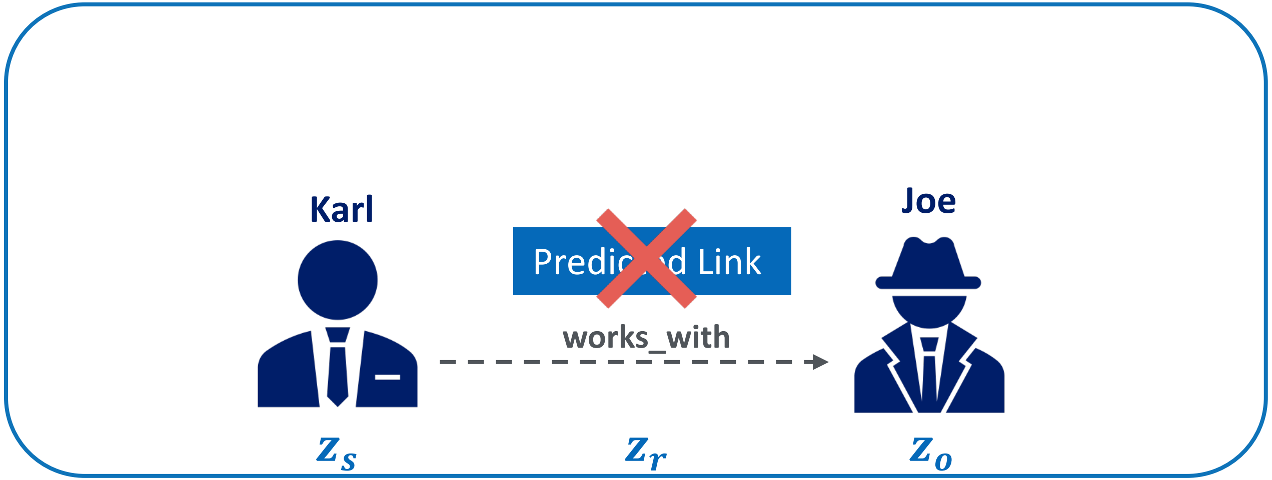
Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

Instance Attribution Metrics

Identify the most influential training triple



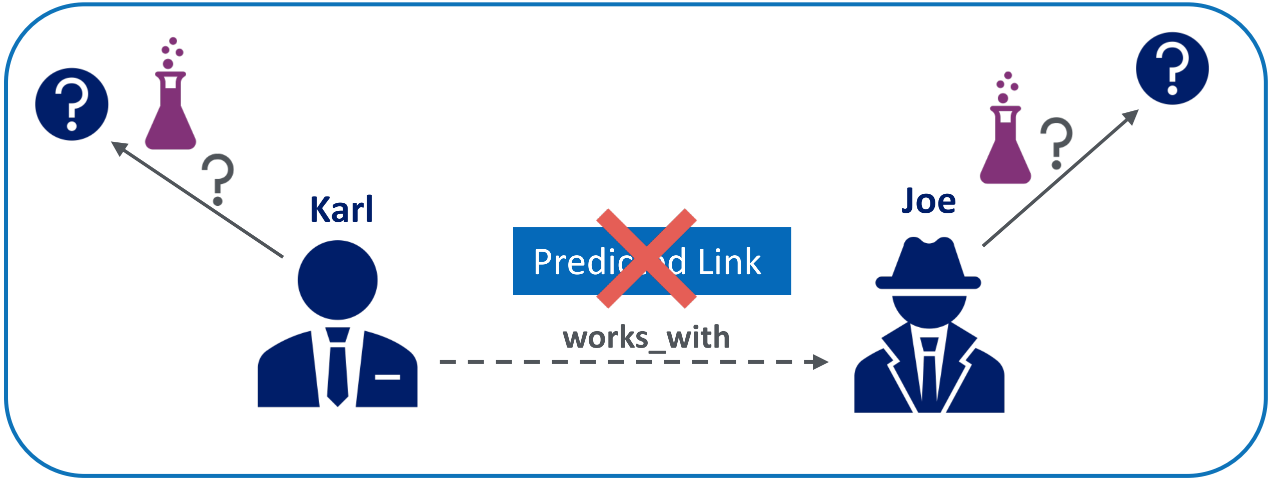
Instance Attribution Metrics



$$z := (z_s, z_r, z_o)$$

Target Triple

Instance Attribution Metrics

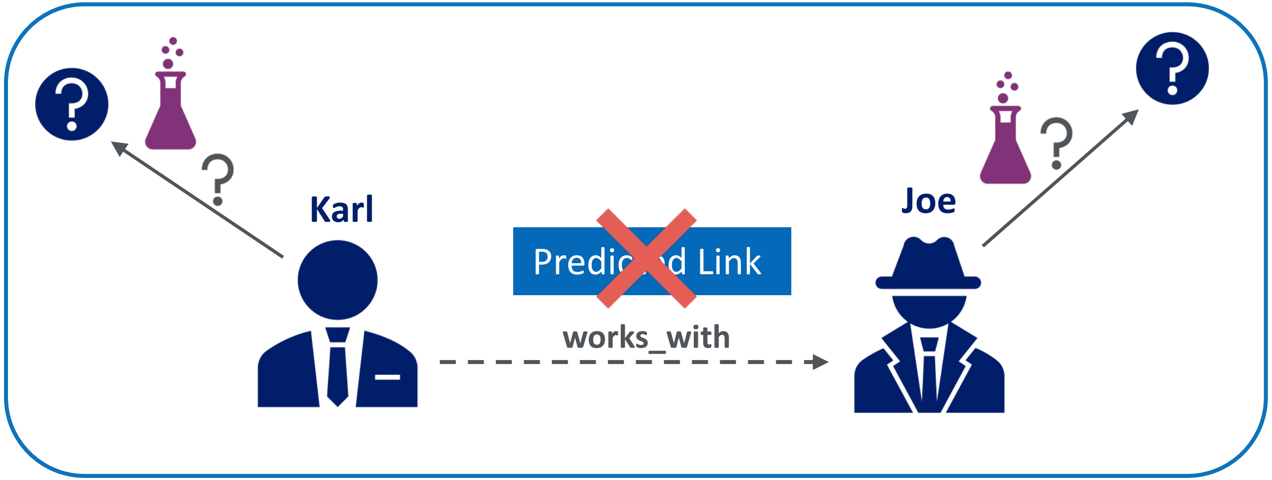


$$x \coloneqq (x_s, x_r, x_o)$$

Candidate Influential Triple

Instance Attribution Metrics

1. Instance Similarity

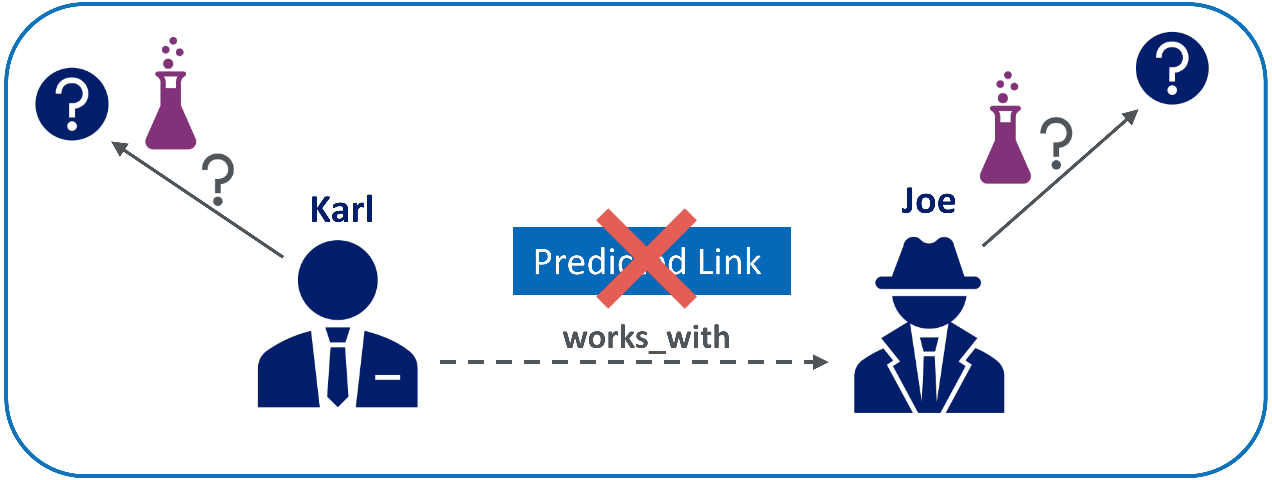


Similarity between $f(e_{z_s}, e_{z_r}, e_{z_o})$ and $f(e_{x_s}, e_{x_r}, e_{x_o})$

where
z – Target triple, x – Candidate triple

Instance Attribution Metrics

2. Gradient Similarity



Similarity between $\boldsymbol{g}(\boldsymbol{z}, \widehat{\boldsymbol{\theta}})$ and $\boldsymbol{g}(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})$

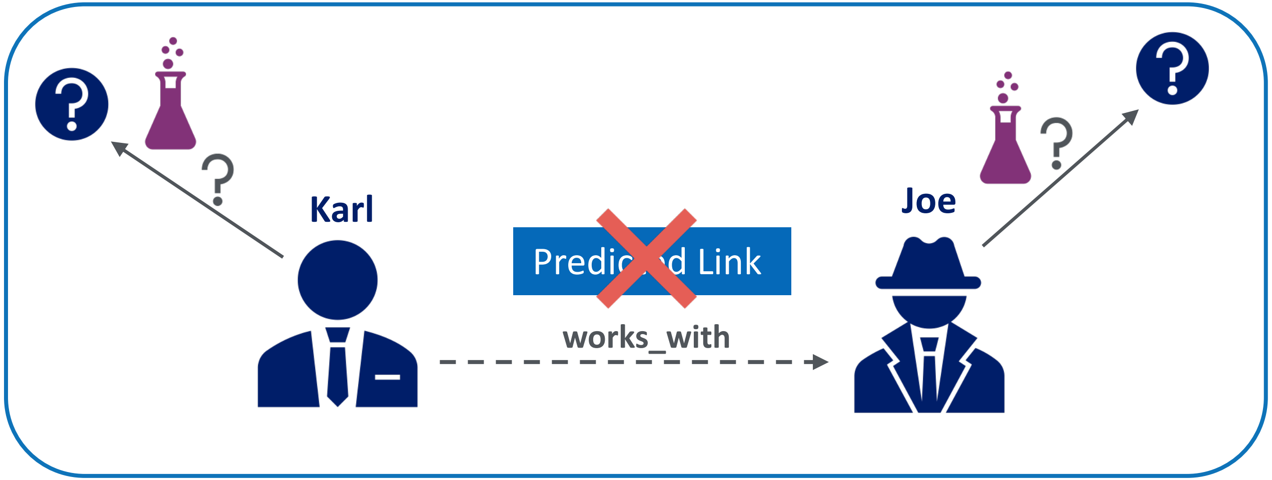
where

z – Target triple, x – Candidate triple

and $\boldsymbol{g}(\boldsymbol{z}, \widehat{\boldsymbol{\theta}}) = \nabla_{\theta}\mathcal{L}(\boldsymbol{z}, \widehat{\boldsymbol{\theta}})$

Instance Attribution Metrics

3. Influence Functions [Koh and Liang, 2017]



Dot product between $\boldsymbol{g}(\boldsymbol{z}, \widehat{\boldsymbol{\theta}})$ and $\boldsymbol{H}_{\widehat{\boldsymbol{\theta}}}^{-1}\boldsymbol{g}(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})$

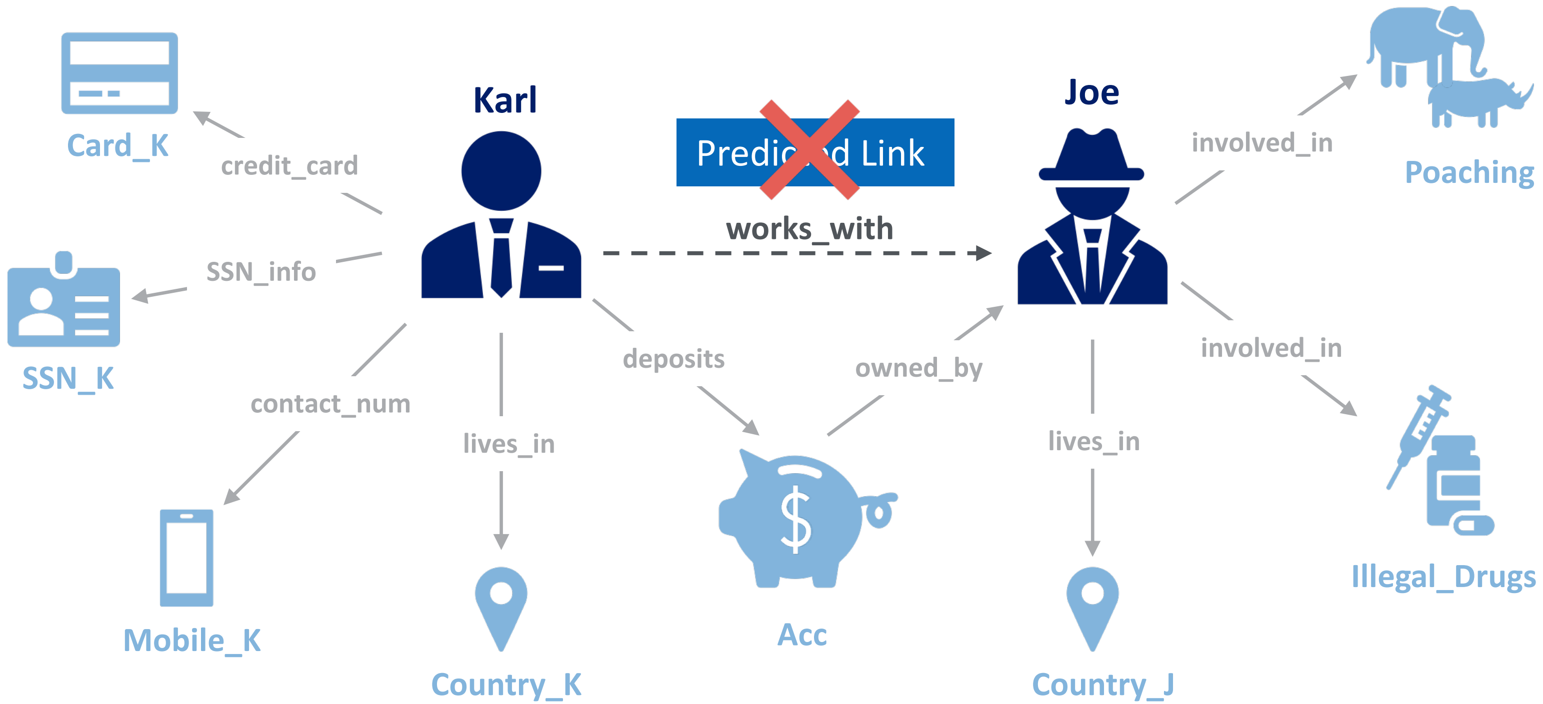
where

z – Target triple, x – Candidate triple

and $\boldsymbol{g}(\boldsymbol{z}, \widehat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{\mathcal{L}}(\boldsymbol{z}, \widehat{\boldsymbol{\theta}})$

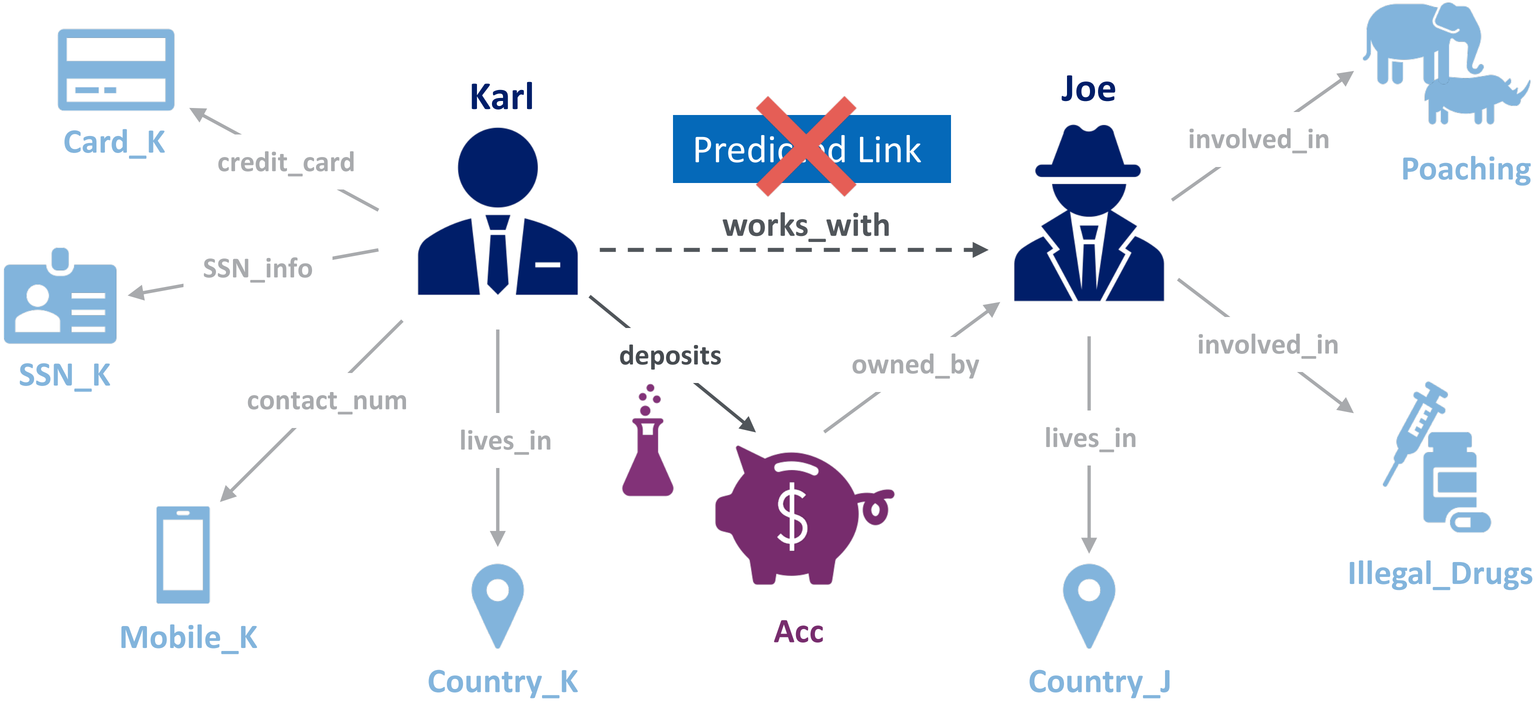
Adversarial Deletions

Identify the most influential training triple



Adversarial Deletions

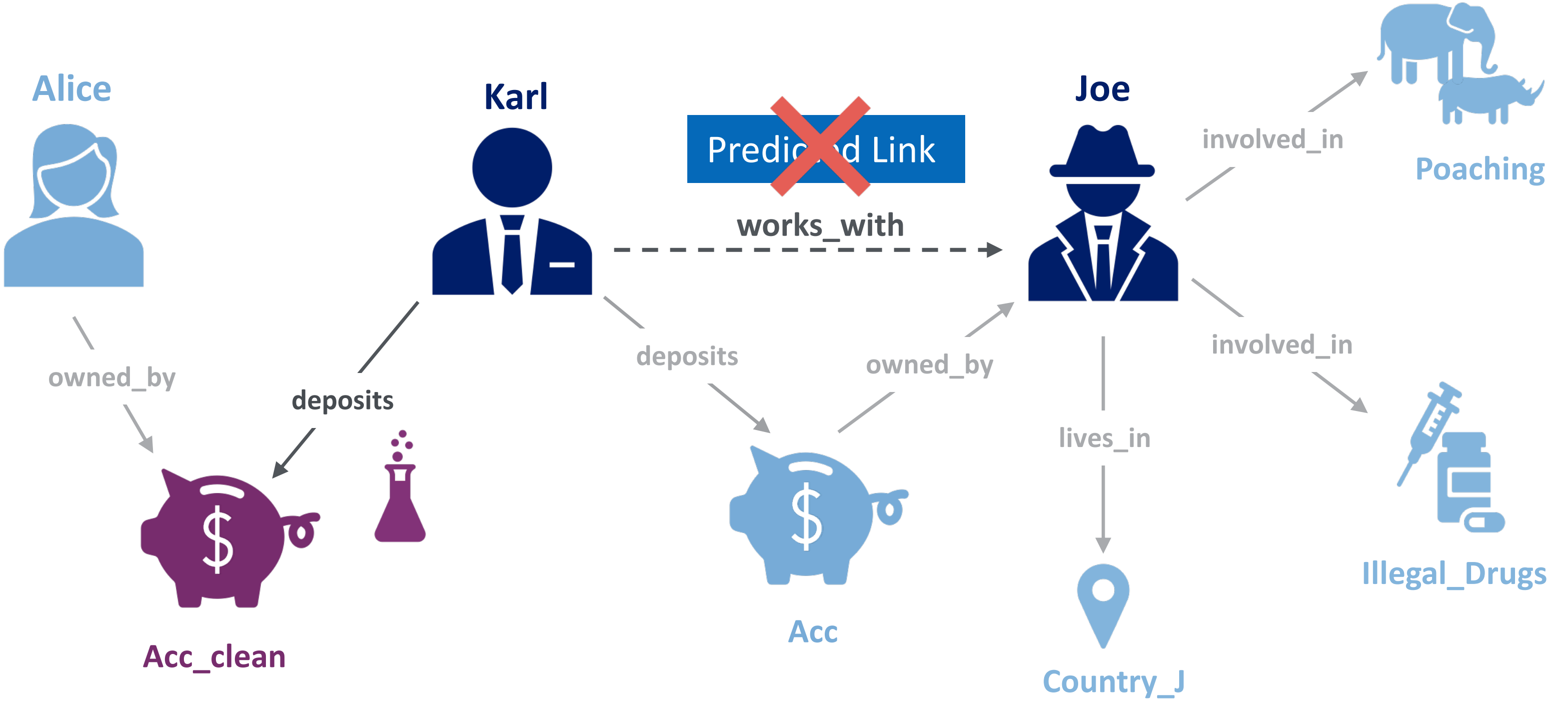
Identify the most influential training triple



Bhardwaj et al. 2021. Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Metrics. In EMNLP

Adversarial Additions

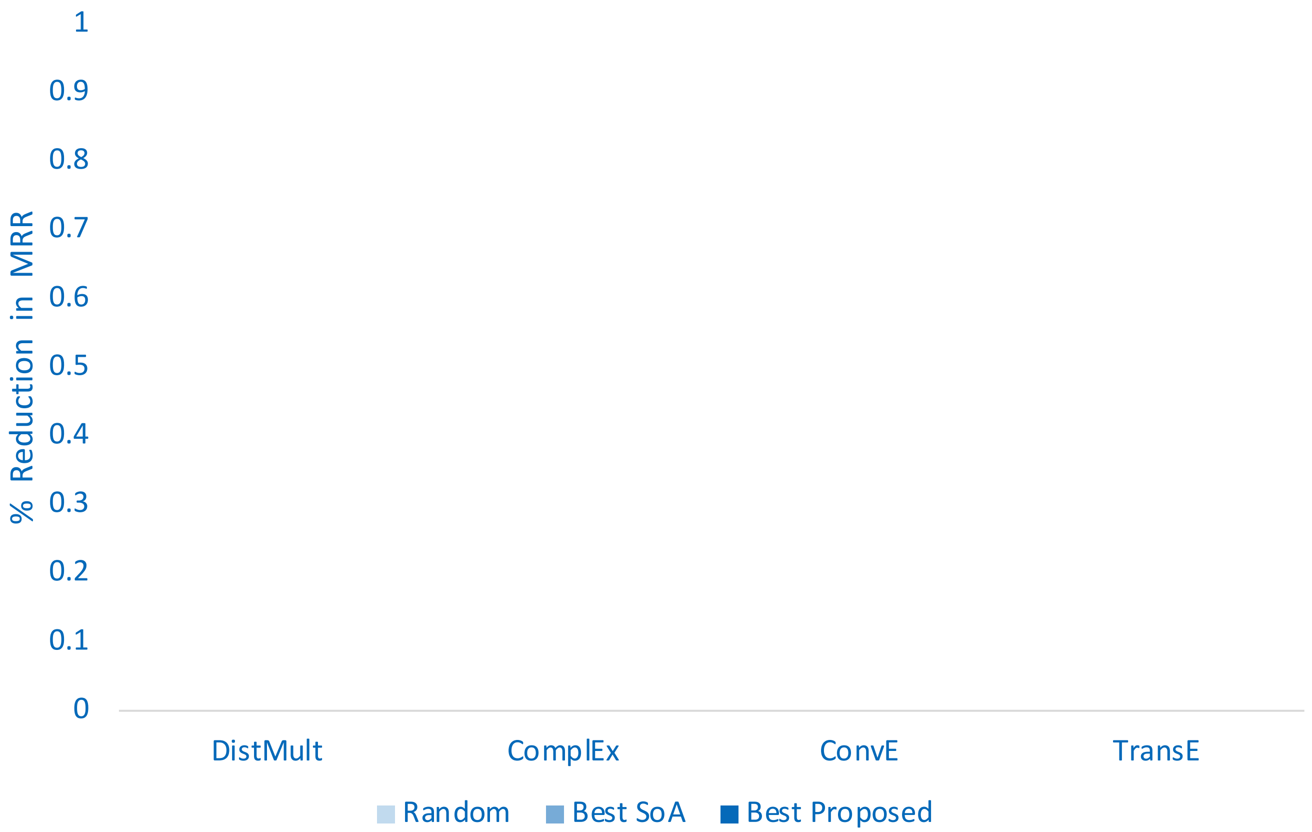
Replace with dissimilar entity



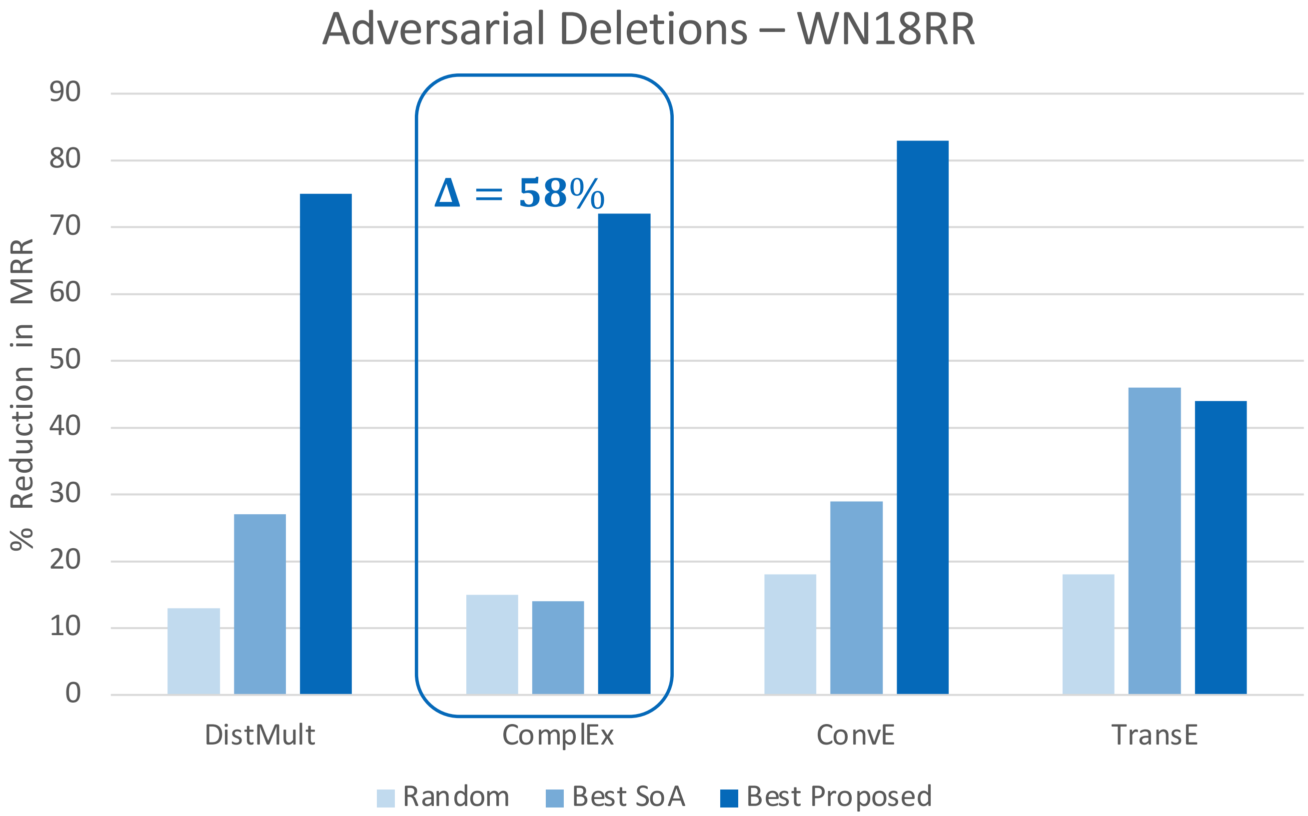
Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

Evaluation

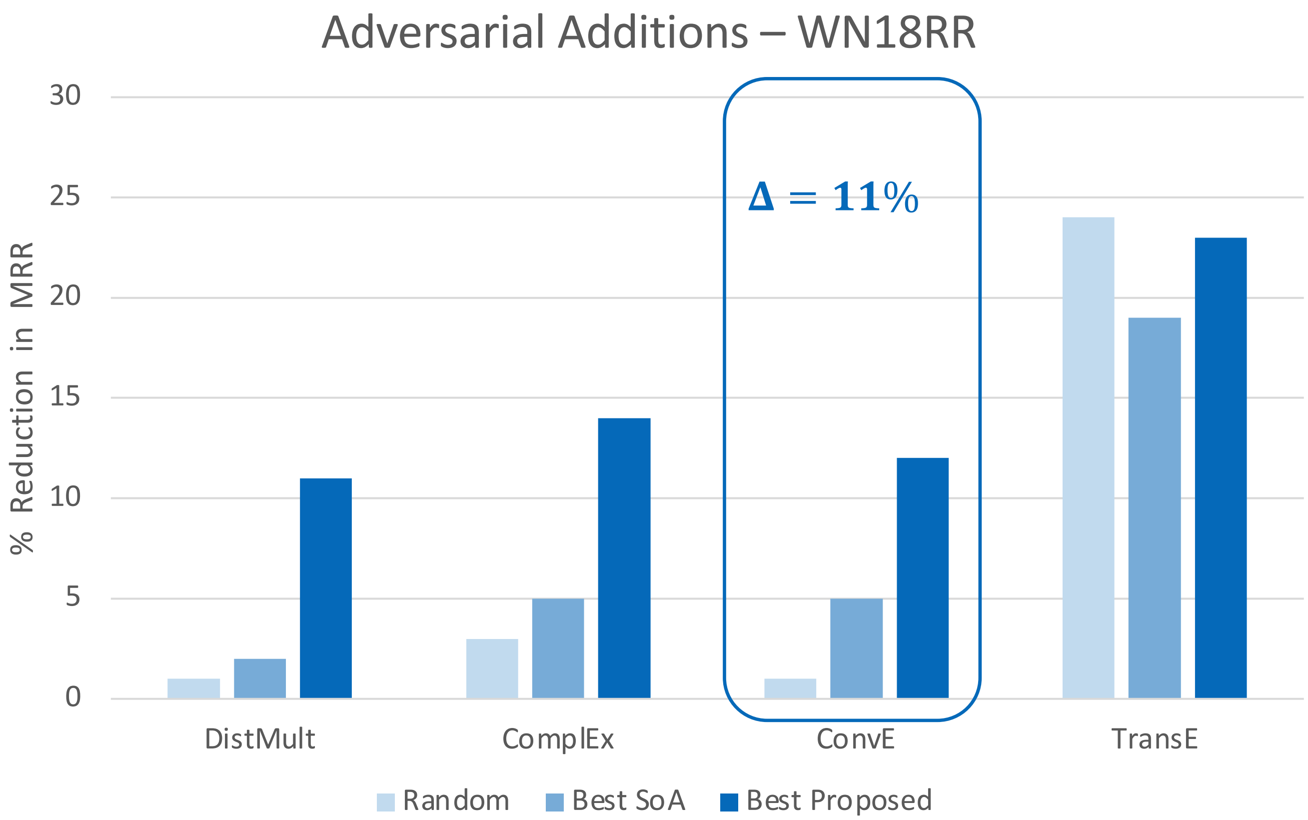
Proposed Vs State-of-Art



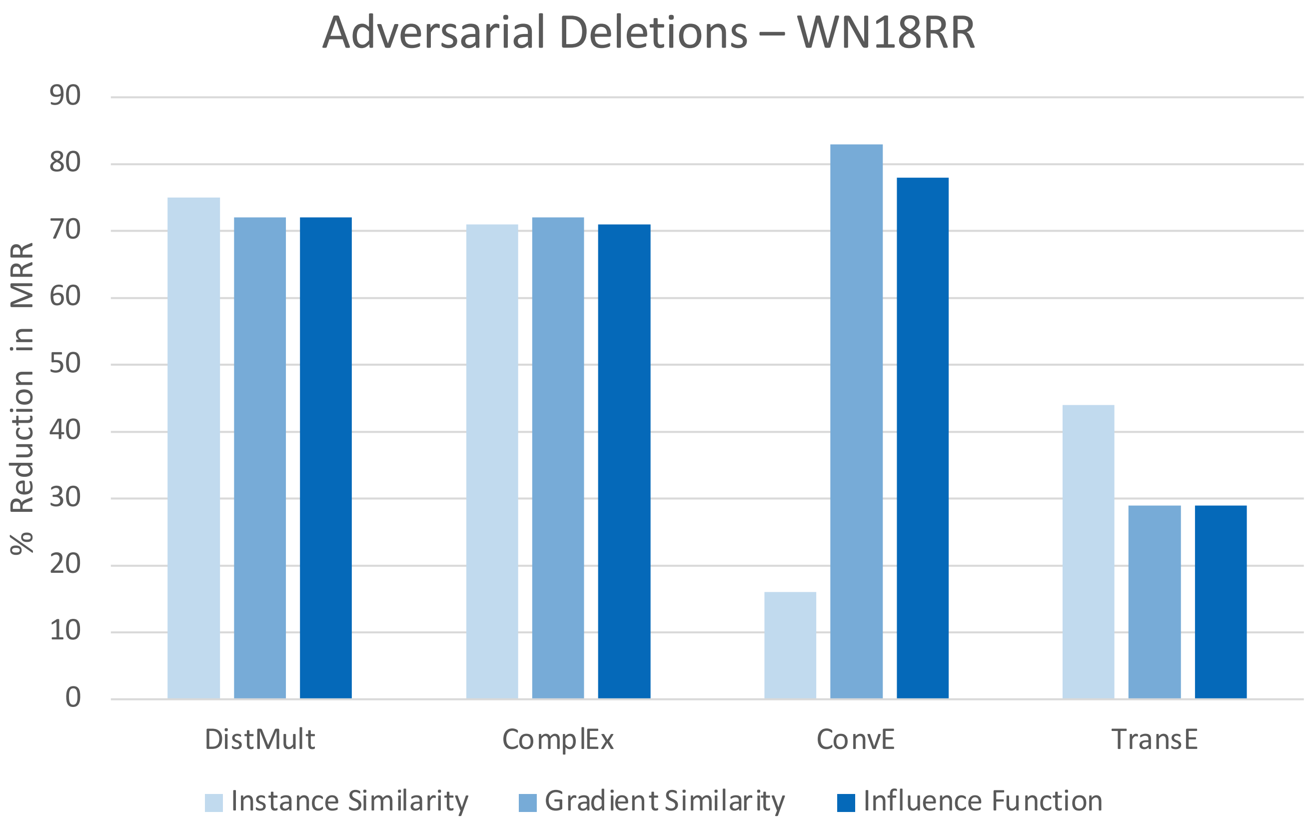
Proposed Vs State-of-Art



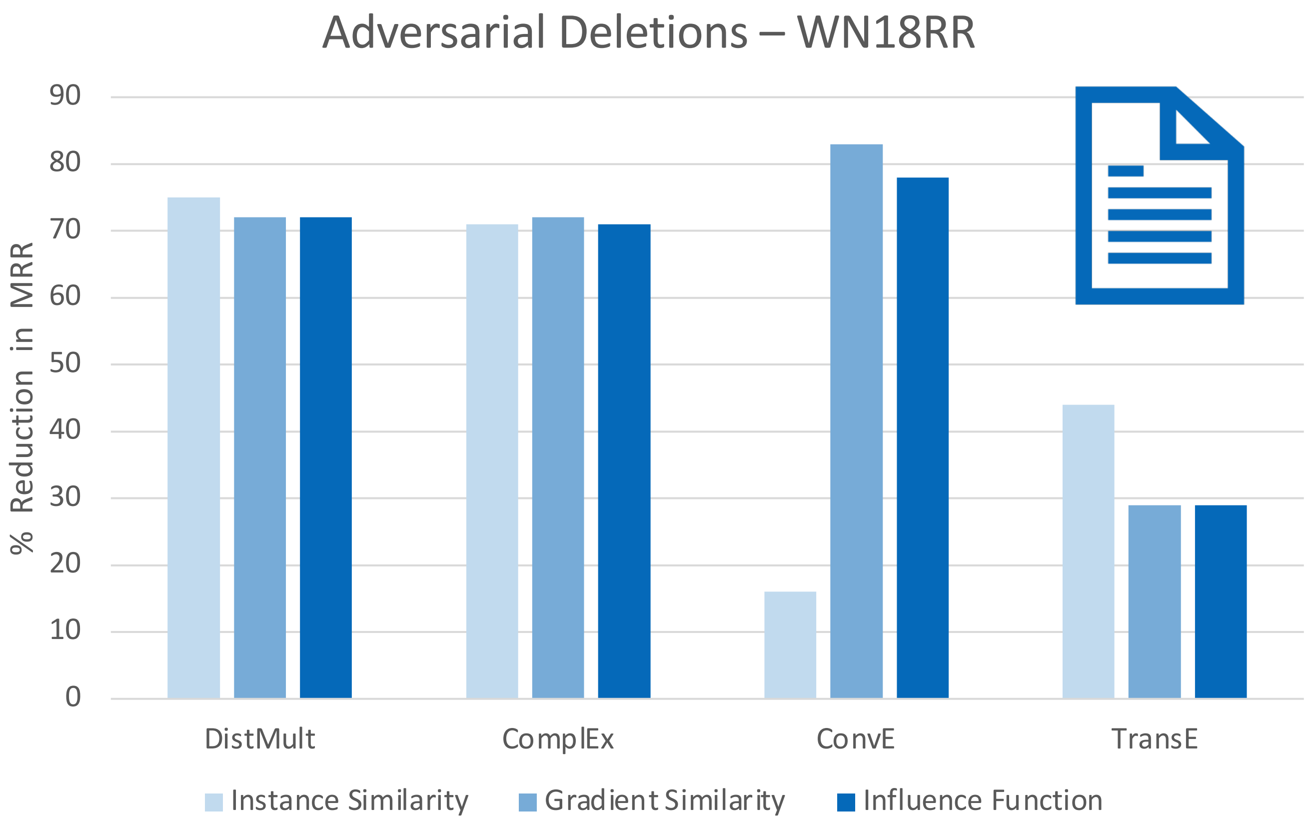
Proposed Vs State-of-Art



Instance Attribution Methods



Instance Attribution Methods



Adversarial Attacks on KGE

Future Directions

Sub-graph Influence

Can we measure the influence of a training sub-graph on the model’s prediction for target triple?

Bhardwaj et al. 2021. Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Metrics. In EMNLP

Adversarial Attacks on KGE

Future Directions

Sub-graph Influence

Can we measure the influence of a training sub-graph on the model’s prediction for target triple?

Adversarial Robustness

Can we improve the adversarial robustness of KGE models to defend them against adversarial attacks?

Bhardwaj et al. 2021. Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Metrics. In EMNLP

Contact: peru.bhardwaj@adaptcentre.ie

PeruBhardwaj/AttributionAttack

Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

- Adversarial Deletions + Additions

Contact: peru.bhardwaj@adaptcentre.ie

PeruBhardwaj/AttributionAttack

Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

Instance Similarity | Gradient Similarity | Influence Function

- ✓ Adversarial Deletions + Additions
- ✓ Identify influential training examples
- ✓ Outperform state-of-art attacks