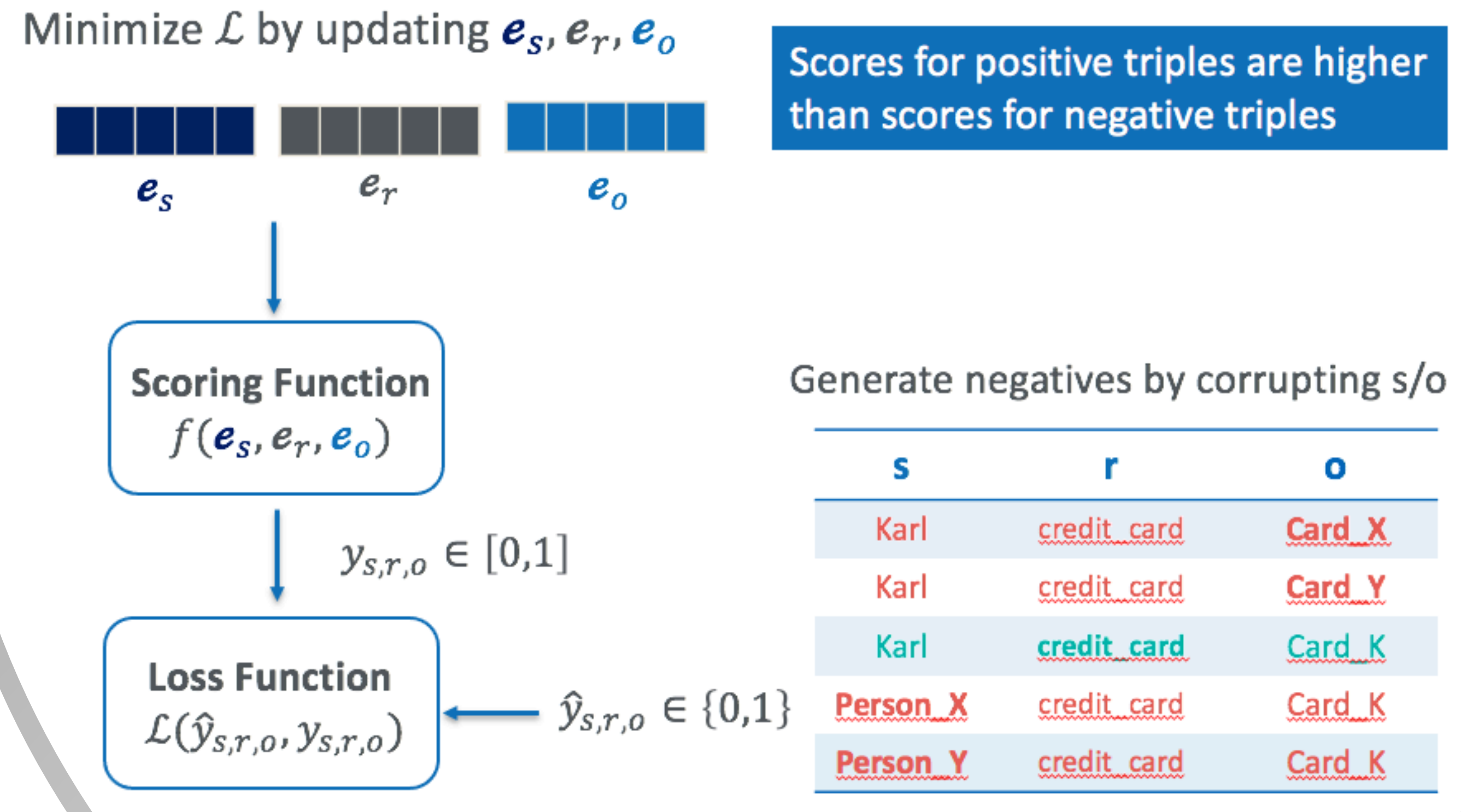
Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods

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

What are Knowledge Graph Embeddings?


Minimize $\mathcal{L}$ by updating $\boldsymbol{e}_s, \boldsymbol{e}_r, \boldsymbol{e}_o$
$\boldsymbol{e}_s$
$\boldsymbol{e}_r$
$\boldsymbol{e}_o$
Scores for positive triples are higher than scores for negative triples
Scoring Function
$f(\boldsymbol{e}_s, \boldsymbol{e}_r, \boldsymbol{e}_o)$
$y_{s,r,o} \in [0,1]$
Loss Function
$\mathcal{L}(\hat{y}_{s,r,o}, y_{s,r,o})$
$\hat{y}_{s,r,o} \in \{0,1\}$
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

Motivation

Knowledge graph embedding models enable representation learning on multi-relational graphs and are used in security sensitive domains. But their security vulnerabilities have received little attention.

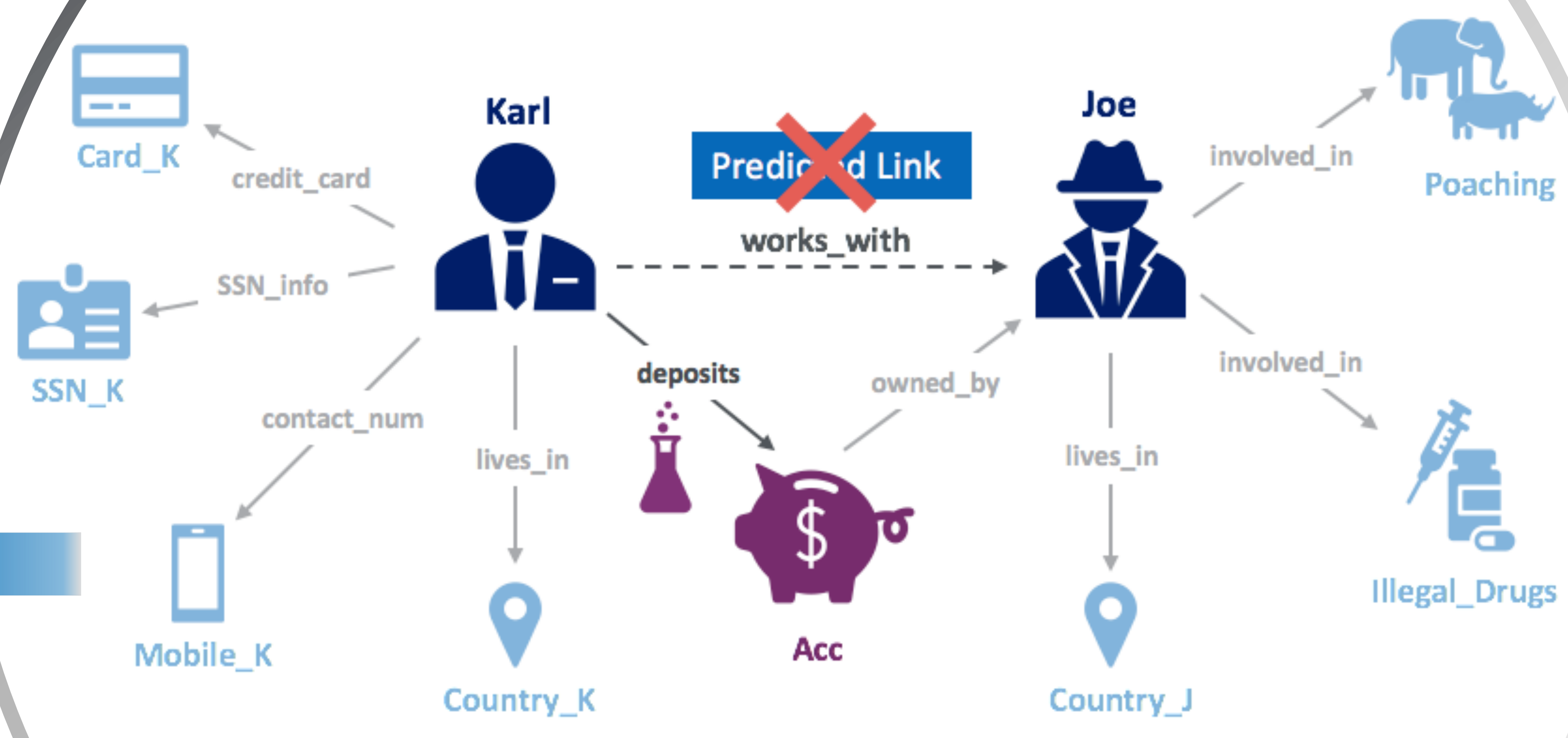
What are Instance Attribution Methods?

Methods to dentify the training triple that is most influential to model's prediction on target triple

1. Similarity between feature representations of instances (Instance Similarity)
2. Similarity between gradients due to instances (Gradient Similarity)
3. Influence Functions

Adversarial Deletions against Knowledge Graph Embeddings

Target triple becomes False by deleting training triple



Example scenario for adversarial attacks against KGE models for fraud detection. The missing triple to predict is (Karl, works_with, Joe). Original KGE model predicts this triple as True.

But a malicious attacker uses the instance attribution methods to identify the influential triple and deletes it.

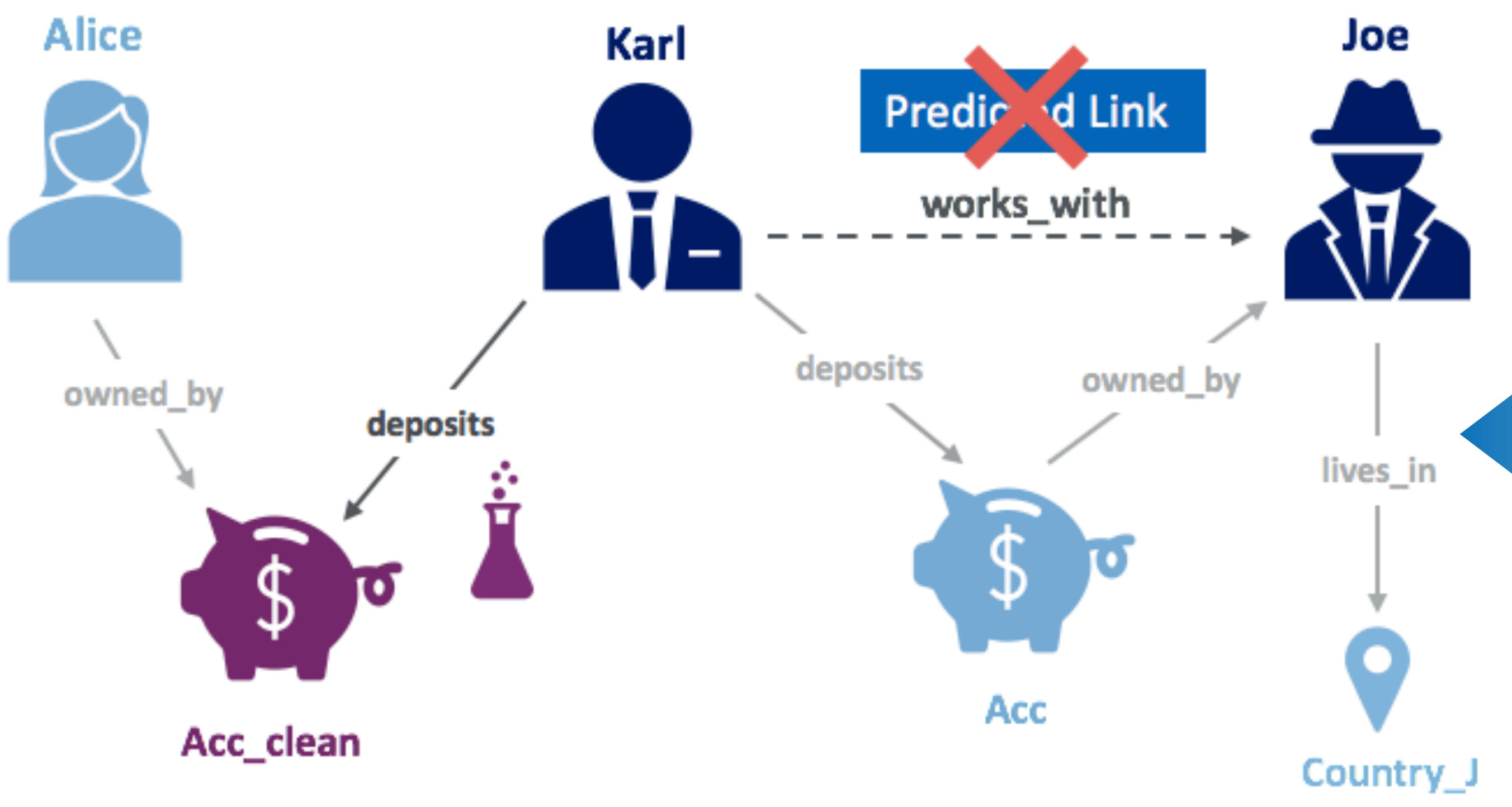
Now the KGE model predicts the target triple as False.

Future Work

1. Influence of training sub-graph instead of individual triples
2. Adversarial robustness of KGE models

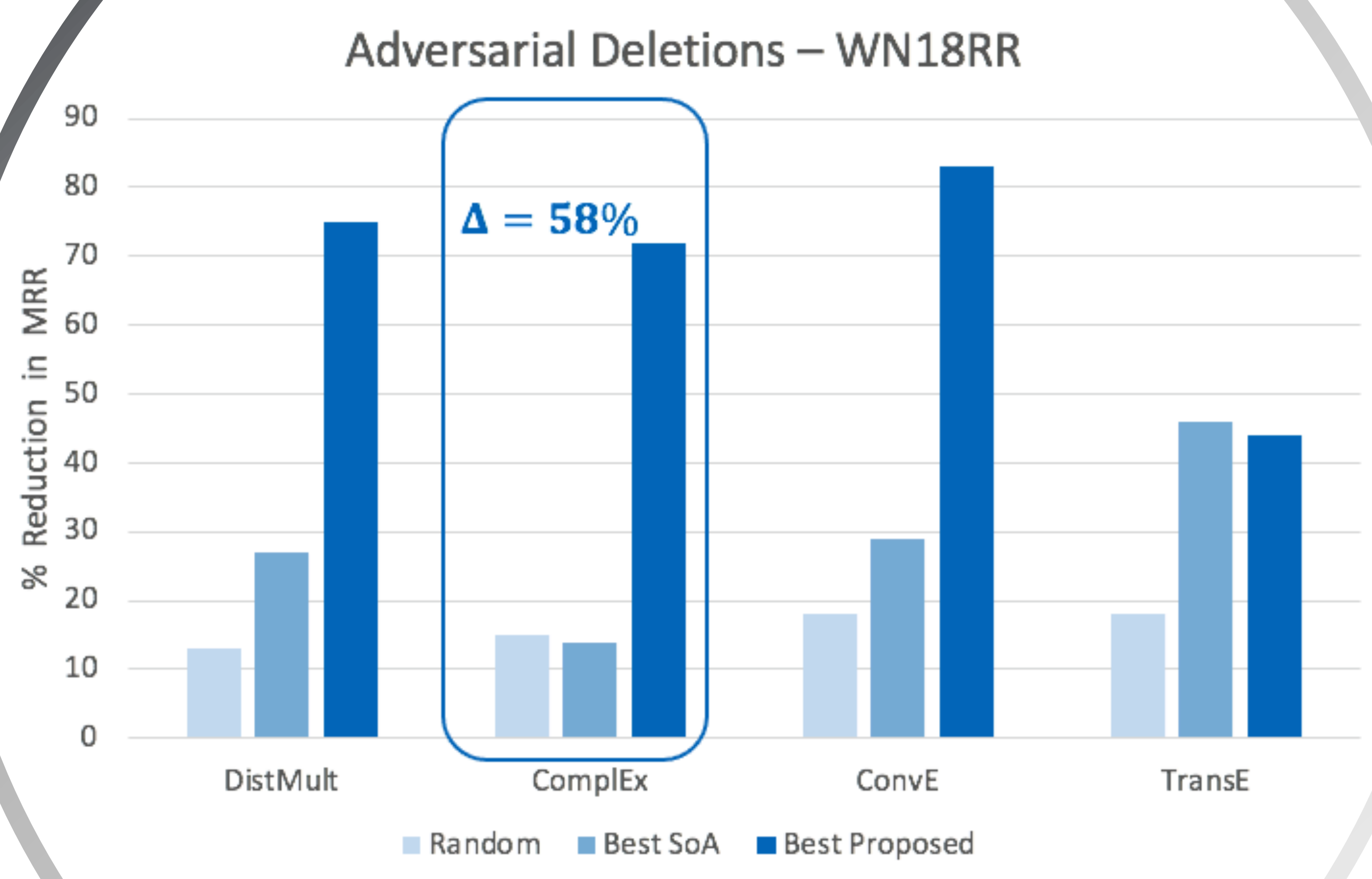
Adversarial Additions

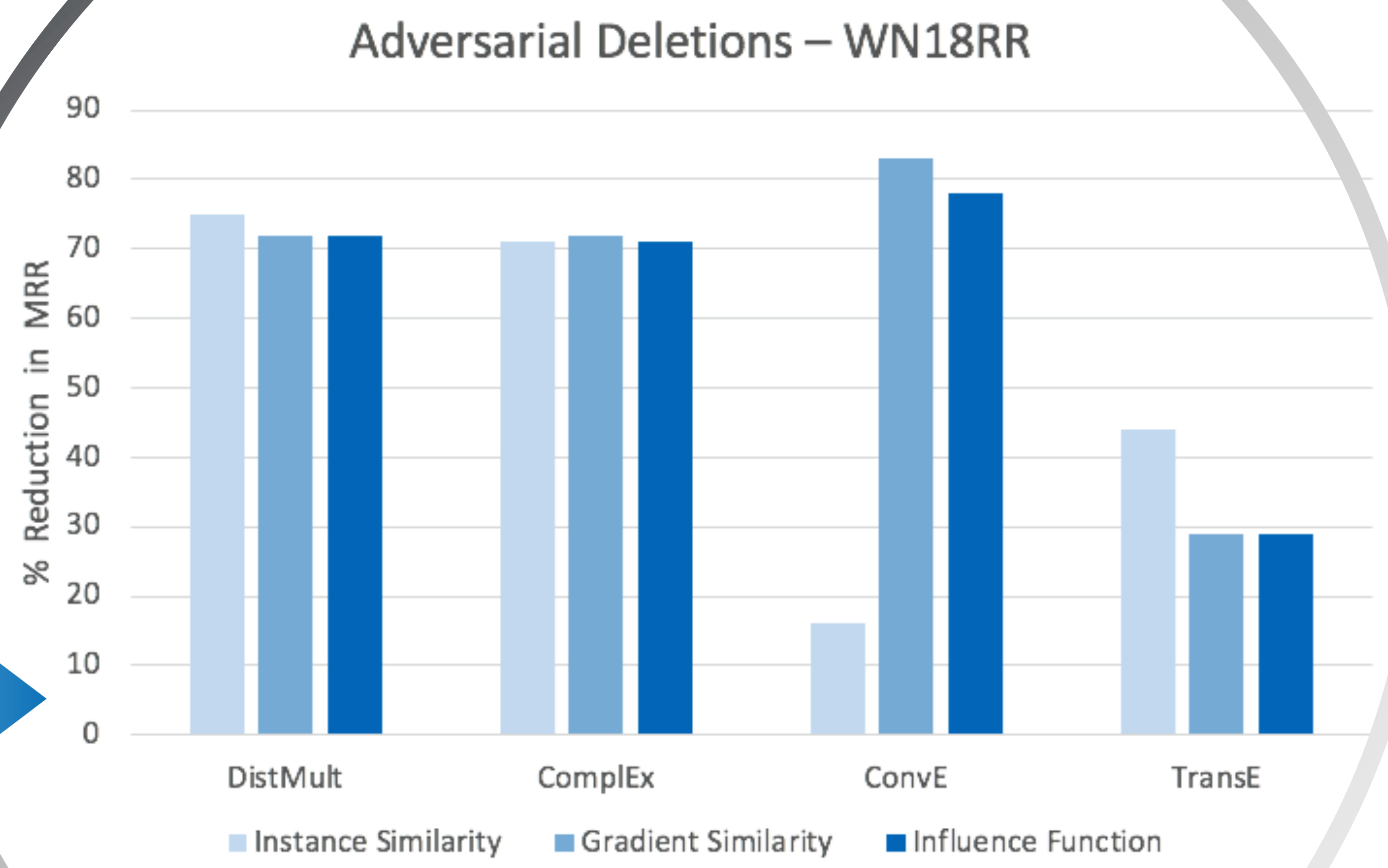
Target triple becomes False by adding triples to training set



The malicious attacker selects adversarial addition by replacing an entity of the influential triple with the most dissimilar entity in embedding space.

Proposed Vs State-of-Art





Instance Attribution Methods

Challenges

How to measure impact of candidate edit on target prediction?

✓ Use instance attribution methods

How to search through combinatorial space of candidate adversarial additions?

✓ Replace entity in influential triple

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

ADAPT
Engaging Content
Engaging People

accenture

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN