Adversarial Robustness of Representation Learning for Knowledge Graphs

Peru Bhardwaj

Contact: bhardwap@tcd.ie
Knowledge Graph

Example - Financial Details of a Bank’s Customer
Knowledge Graph

Example - Financial Details of a Bank’s Customer

<table>
<thead>
<tr>
<th>s</th>
<th>r</th>
<th>o</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karl</td>
<td>credit_card</td>
<td>Card_K</td>
</tr>
<tr>
<td>Karl</td>
<td>SSN_info</td>
<td>SSN_K</td>
</tr>
<tr>
<td>Karl</td>
<td>contact_num</td>
<td>Mobile_K</td>
</tr>
<tr>
<td>Karl</td>
<td>lives_in</td>
<td>Country_K</td>
</tr>
</tbody>
</table>
Knowledge Graph

Example - Financial Details of a Bank’s Customer

Card_K

credit_card

Karl

contact_num

lives_in

Mobile_K

SSN_info

SSN_K

Country_K

<table>
<thead>
<tr>
<th>s</th>
<th>r</th>
<th>o</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karl</td>
<td>credit_card</td>
<td>Card_K</td>
</tr>
<tr>
<td>Karl</td>
<td>SSN_info</td>
<td>SSN_K</td>
</tr>
<tr>
<td>Karl</td>
<td>contact_num</td>
<td>Mobile_K</td>
</tr>
<tr>
<td>Karl</td>
<td>lives_in</td>
<td>Country_K</td>
</tr>
</tbody>
</table>
Missing Link Prediction

Card_K → credit_card
SSN_K → SSN_info → contact_num
Mobile_K

Karl
lives_in
deposits

Country_K

Acc

owned_by

Joe

Missing Link Prediction

Card_K
credit_card
SSN_K
contact_num
Mobile_K
lives_in
Country_K
deposits
Acc
owned_by
Country_J
involved_in
Poaching
Illegal_Drugs

Karl

Joe
Missing Link Prediction

Money Laundering

Card_K
credit_card
SSN_info
contact_num
lives_in
Country_K
Mobile_K
lives_in
Country_J
Acc
owned_by
deposits

Joe
involved_in
involved_in
Poaching
Illegal_Drugs

Karl
Missing Link Prediction

Use case – Anti Money Laundering
Missing Link Prediction

Knowledge Graph Embeddings for entities and relations
Missing Link Prediction with KGE

Score the missing triple using embeddings

\[ P(Karl, \text{works\_with}, Joe) \propto f(e_{Karl}, e_{\text{works\_with}}, e_{Joe}) \]

\[ P(Karl, \text{works\_with}, Joe) \propto f(\text{[embeddings]}\) \]
Missing Link Prediction with KGE

Rank the missing triple against synthetic negatives

\[ \mathcal{P}(Karl, \text{works with}, Joe) \]

\[ \mathcal{P}(Karl, \text{works with}, \text{Person}_1) \]

\[ \mathcal{P}(Karl, \text{works with}, \text{Person}_2) \]
Where to find KGE in practice?

Security Sensitive

High Stakes
KGE in High-Stakes Applications

Incentives for bad actors!
Adversarial Robustness of Representation Learning for Knowledge Graphs
Adversarial Attacks on KGE

Aim – Degrade the rank of target triple
Adversarial Attacks on KGE

Adversarial Edits to Training Data

Karl

Predicted Link

works_with

Joe

Card_K

SSN_K

Mobile_K

contact_num

lives_in

Owned_by

Country_K

Acc

Country_J

lives_in

Illegal_Drugs

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

Adversarial Edits to Training Data

How to search through the combinatorial space of candidate adversarial additions?
Adversarial Robustness of Representation Learning for Knowledge Graphs

☑️ Instance Attribution Methods for Adversarial Deletions
[Bhardwaj et.al. EMNLP 2021]
Instance Attribution Methods

Identify the most influential training triple
Instance Attribution Methods

Identify the most influential training triple

1. Instance Similarity
2. Gradient Similarity
3. Influence Functions [Koh and Liang, 2017]
Adversarial Robustness of Representation Learning for Knowledge Graphs

☑️ Instance Attribution Methods for Adversarial Deletions  
[Bhardwaj et.al. EMNLP 2021]

☑️ Relation Inference Patterns for Adversarial Additions  
[Bhardwaj et.al. ACL 2021]
Adversarial Additions for KGE

Aim – Degrade the rank of TargetTriple

\[ \mathcal{P}(Karl, works\_with, Joe) \]

\[ \mathcal{P}(Karl, works\_with, Person\_1) \]

\[ \mathcal{P}(Karl, works\_with, Person\_2) \]
Adversarial Additions for KGE

Exploit inductive abilities to improve the rank of a decoy triple

\[ \mathcal{P}(Karl, works\_with, Joe) \]

\[ \mathcal{P}(Karl, works\_with, Person\_1) \]

\[ \mathcal{P}(Karl, works\_with, Person\_2) \]
Adversarial Additions for KGE

Symmetry
Adversarial Additions for KGE

Inversion

Karl

Predicted Link

works_with

Joe

involved_in

Poaching

involved_in

Illegal_Drugs

Alice

has_employee

has_employee

deposits

owned_by

deposits

Bob

lives_in

Country_J

Acc
Adversarial Additions for KGE

Composition

Alice

Karl

Joe

owned_by

deposits

works_with

lives_in

Poaching

Illegal_Drugs

Carol

Acc

Country_J

Owned_by

Deposits

Works_with

Lives_in
Adversarial Additions for KGE

Exploit inductive abilities to improve the rank of a decoy triple

Steps

- Adversarial Relations
- Decoy Entity
- Remaining Entity
Evaluation Setup

Evaluation Setup

Original graph

Poisoned graph

Original KGE Model

Poisoned KGE Model

Original ranks

Poisoned ranks
Evaluation Summary

4 KGE models, 2 benchmark datasets
Evaluation Summary

4 KGE models, 2 benchmark datasets

- Proposed attacks are more effective than baselines
Evaluation Summary

4 KGE models, 2 benchmark datasets

- Proposed attacks are more effective than baselines
- Proposed attacks based on simpler methods are more or as effective as more complex attacks
Adversarial Robustness of KGE

Impact of the Contributions
Adversarial Robustness of KGE

Impact of the Contributions

Improve Security

Safeguard the KGE models against potential harm from adversaries in user facing applications
Adversarial Robustness of KGE

Impact of the Contributions

**Improve Security**

Safeguard the KGE models against potential harm from adversaries in user facing applications

**Improve Understanding**

Understand the black-box model predictions through post-hoc explanations, or inductive abilities
Adversarial Robustness of Representation Learning for Knowledge Graphs

- Instance Attribution Methods for Adversarial Deletions
  [Bhardwaj et.al. EMNLP 2021]

- Relation Inference Patterns for Adversarial Additions
  [Bhardwaj et.al. ACL 2021]