Towards Adversarially Robust Knowledge Graph Embeddings

Peru Bhardwaj
ADAPT Centre, School of Computer Science and Statistics
Trinity College Dublin
Dublin 2, Ireland
peru.bhardwaj@adaptcentre.ie

Abstract

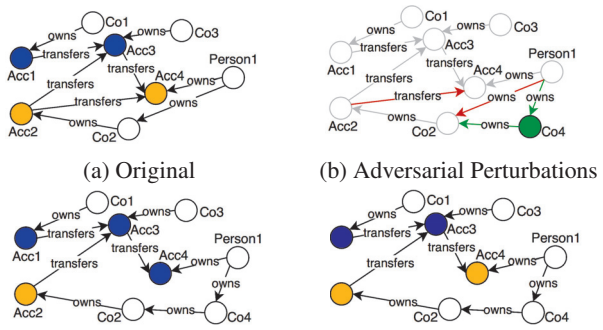
Knowledge graph embedding models enable representation learning on multi-relational graphs and are used in security sensitive domains. But, their security analysis has received little attention. I will research security of these models by designing adversarial attacks against them, improving their adversarial robustness and evaluating the effect of proposed improvement on their interpretability.

Introduction

Graph representation learning encodes structural information in graphs into low dimensional feature vectors or embeddings (Cai, Zheng, and Chang 2018). The graph structure to be encoded can have different characteristics – (un)directed, (un)labelled, (non-)attributed. This leads to different families of graph representation learning like network embeddings, graph neural networks (GNN) and knowledge graph embeddings (KGE). They differ from each other in their neural network architectures and underlying theoretical assumptions to encode structural information. Embeddings learned from graphs can be used to classify nodes, to predict links or as background knowledge to other machine learning (ML) tasks.

Research shows that interconnections in social networks can be rewired to mislead humans and skew their decisions (Stewart et al. 2019). Similar strategies can also be used by adversaries to fool graph representation learning, especially in security sensitive applications like anti-money laundering[1]. Here, financial data is represented as a multi-relational graph and representation learning is used to identify suspicious bank accounts. This is done by using the learned embeddings for node classification as shown in Figure 1. Malicious parties (like drug cartels, politicians, criminals or rogue industrialists) that want to launder income from illicit activities can fool this learning algorithm. Based on the predicted results, they can change the flow of transactions and add or remove bank accounts to evade detection. Hence, there is a need to identify and fix security vulnerabilities of graph representation learning.

[1]https://www.markrweber.com/graph-deep-learning



(a) Original (b) Adversarial Perturbations

(c) Adversarial Attack (d) Adversarial Robustness

Figure 1: Example scenario for adversarial attack on anti-money laundering. The financial relational graph consists of bank accounts (Acc), companies (Co) and people (Person). (a) Original KGE model is used to classify the accounts as suspicious (orange) or not (blue). (b) Adversarial graph perturbations – triples to be removed (red) and added (green). (c) Node classification by original KGE model under the adversarial attack. (d) Node classification by adversarially robust KGE model under adversarial attack. Note that the robust model has same results as baseline in (a).

In adversarial machine learning, the intentionally perturbed inputs that can fool a learning system to output incorrect results with high confidence are called *adversarial examples* (Tsipras et al. 2019). These inputs can be used to generate an *adversarial attack*, which aims to reduce predictive performance of the ML model. Adversarial inputs for graphs (or *adversarial graph perturbations*) can take many forms like addition or removal of nodes, edges, triples or node features. Since graph representation learning is based on interconnections in the graph, these adversarial perturbations can propagate through the entire graph. A proactive approach to security advocates that the system designer should anticipate such adversarial attacks, simulate them and incorporate defenses against them during system design. Recent research efforts have mostly focused on adversarial attacks and defenses for GNN and network embedding models. But the security of KGE models has received little attention.

Research Problem

KGE models learn representations from multi-relational graphs (or knowledge graphs) using ranking based objective functions. Embeddings are initialized with random values and updated iteratively such that actual facts in the graph are ranked higher than false facts. Generating attacks against KGE models is more challenging than other machine learning models, because their loss functions depend on latent embeddings rather than input data points, i.e. triples. Thus, state-of-art attacks (that use the gradient of loss function w.r.t. inputs) can only be used to perturb the latent space of KGE models. Determining input perturbations from the perturbed latent space requires novel approaches.

Recently, two initial attempts have been made to solve this problem. (Zhang et al. 2019) generates input perturbations from the perturbed latent space by scoring all possible perturbations and (Pezeshkpour, Tian, and Singh 2019) uses an encoder-decoder based inverter neural network. But the generated attacks have been successful in reducing the predictive performance by 50% only (Pezeshkpour, Tian, and Singh 2019). This attack success is low, compared to the adversarial attack success on GNNs, where the node classification accuracy is reduced to 1% (Zügner, Akbarnejad, and Günnemann 2018). For proper security evaluation, there is a need to generate more effective adversarial attacks and compute bounds on the predictive performance degradation that they cause. Hence, the first research question for my doctoral research is – **RQ1:** *Can we design adversarial attacks that degrade the predictive performance of KGE models more than the state-of-art adversarial attacks?*

ML systems are defended against adversarial attacks by designing algorithms that are intrinsically less sensitive to adversarial inputs, i.e. they are adversarially robust (Tsipras et al. 2019). But there is no study so far to improve adversarial robustness of KGE models. Hence, the second research question for my doctoral research is – **RQ2:** *Can we improve the adversarial robustness of KGE models to defend them against existing and proposed adversarial attacks?*

Interpretability of ML models is the ability to explain their predictions in human understandable terms. Black-box models like neural networks can be interpreted by providing explanations for their predictions. These explanations for image classifiers have shown that adversarially robust neural networks are more interpretable than standard ones (Tsipras et al. 2019). However, no similar analysis exists for KGE models. Hence, the third research question for my research is – **RQ3:** *Does the proposed improvement in adversarial robustness of KGE models effect their interpretability?*

Research Proposal

To answer the three research questions, my doctoral research has three research objectives. A timeline to complete the research objectives is given in Table 1. The first research objective is – **RO1:** *Propose and evaluate novel adversarial attacks on KGE models.* I will evaluate the proposed attacks using the work in (Pezeshkpour, Tian, and Singh 2019) and (Zhang et al. 2019) as baseline. The attack success will be measured by computing the predictive performance degradation of downstream ML tasks like node classification or link prediction. For proper risk assessment of KGE models, I will also select usecase scenarios (like anti-money laundering) and evaluate the proposed attack under different threat models. I will further compute theoretical bounds on the proposed attack's success for effective evaluation of defenses.

The second research objective of my research is – **RO2:** *Improve adversarial robustness of KGE models against adversarial attacks.* On images, adversarial robustness is improved by training the model with l-norm perturbations of input data (adversarial training). I will adapt this method to graph data by generating the perturbations in latent space instead of input space. I will also add the adversarial examples generated for RO1 to input data and introduce an adversarial loss function that penalizes high scores for these inputs to improve robustness. I will evaluate the proposed methods by comparing the predictive performance of original and robust KGE models under adversarial attacks.

My third research objective is – **RO3:** *Evaluate the effect of proposed improvement in adversarial robustness on interpretability of KGE models.* I will provide explanations for KGE model predictions by identifying facts in the knowledge graph that have high impact on the predictive performance of downstream ML task. These explanations will be aggregated for each relation to extract rules from the knowledge graph. I will use the extracted rules for original and robust KGE models to compare their interpretability.

Table 1: PhD Completion Plan

Research Objectives	Time Period
RO1	July 2019 - December 2019
RO2	January 2020 - July 2020
RO3	August 2020 - December 2020
Thesis Writing	January 2021 - August 2021

Acknowledgments

This research is funded by Science Foundation Ireland's ADAPT Centre (Grant No. 13/RC/2106) and Accenture Labs, Ireland.

References

Cai, H.; Zheng, V. W.; and Chang, K. C.-C. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*.

Pezeshkpour, P.; Tian, Y.; and Singh, S. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *NAACL*.

Stewart, A. J.; Mosleh, M.; Diakonova, M.; Arechar, A. A.; Rand, D. G.; and Plotkin, J. B. 2019. Information gerrymandering and undemocratic decisions. *Nature* 573:117–121.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness may be at odds with accuracy. In *ICLR*.

Zhang, H.; Zheng, T.; Gao, J.; Miao, C.; Su, L.; Li, Y.; and Ren, K. 2019. Data poisoning attack against knowledge graph embedding. In *International Joint Conference on Artificial Intelligence*.

Zügner, D.; Akbarnejad, A.; and Günnemann, S. 2018. Adversarial attacks on neural networks for graph data. In *International Conference on Knowledge Discovery & Data Mining*, 2847–2856.