**Towards Data Poisoning Attack**

**against Knowledge Graph Embedding**

**一、Abstract**

知识图嵌入（KGE）是种学习知识图中实体与关系的连续嵌入的技术。KGE应用广泛，但其对对抗攻击的鲁棒性没有得到很好研究。本文提出了一系列投毒攻击，通过在知识图中添加/删除实体，可有效改变任意目标fact的可信性。

**二、Introduction**

知识图谱是很多现实应用中的一个重要资源。在知识图谱中，一个知识会被表示为三元组（头实体，关系，尾实体），例如：（爱因斯然，职业，科学家）表示的是“爱因斯然的职业是科学家”。三元组虽然可以有效地记录大量知识，但其本身的符号特性使其难以直接作为模型的输入。

为解决这一问题，知识图嵌入（KGE）可将三元组投影到连续的向量空间中。KGE既保留了实体与关系的特性，又扩大了这些知识的应用领域，如：问答，推荐任务。

KGE数据来源很open，易受到攻击，KGE的不可靠会造成最终应用带来损失。因此对KGE进行鲁棒性分析非常必要。

现有的基于图像的对抗攻击并不能直接应用到KGE上。

**Contributions：**

·第一次对KGE的vulnerability进行了分析。

·针对KGE，提出了一系列有效的数据投毒攻击（分为直接和非直接攻击）。

·实验表明，该策略只对图进行少量改变，就可显著恶化目标的预测结果。

**三、Related Work**

针对图学习任务的对抗攻击：包括了节点分类、图分类、link预测、节点嵌入。此前方法都将投毒攻击表示为双层优化问题。

就本文作者所知，目前还没有针对异类图的对抗攻击的研究，像知识图谱这样的异类图，其节点和link是不同类型的。这篇文章是首先揭示这个问题的。

**四、Data Poisoning Attack against KGE Methods**

敌手目标：操纵已学习的embedding，降低（或提高）目标fact的可信度。本文专注于降低可信度的情况。

由于实体数量大于关系的数量，所以关系更稳定，不易操作。本文对实体进行操作。

**Defifinition 1 (Problem Defifinition)：**

定义目标fact为，表示的embedding，其余两者同理。攻击目标是制造扰动（添加或删除facts）以最小化的可信度，即。预算为M，即只能制造M个扰动。

**1. Direct Attack：**

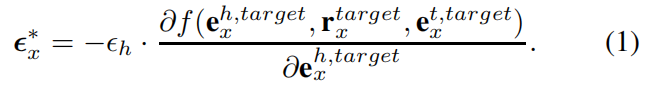
直接攻击的目标是确定一个扰动的集合，这些扰动会改变实体的embedding，并最小化目标fact的可信度。

首先，确定实体embedding的最佳移动方向；然后，分析KGE模型训练方式与scoring function，对可能的扰动进行排序；最后，实施分数最高的M个扰动。

【名词解释：

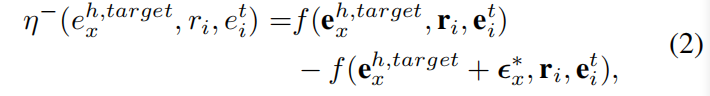
（1 、scoring function用于评估一个扰动的优劣，即扰动可使embedding在目标方向上移动多少.

（2 、是embedding shifting vector，表示扰动步长，则最优转换向量为：

】

具体可分为直接删除/添加攻击。

**直接删除攻击：**

为将embedding从 转换成 ，需要找到并删除一个或多个直接涉及实体 的facts。假设一个包含的fact，它的可信度越高，对最终 的 embedding的贡献越大，那么删除该fact的效益要与相应的分数函数成正比。反之，当转化为后，若fact的可信度（分数）变高，说明该fact对嵌套转化有积极影响，不应该删除，也即删除该fact的效益分数要与成反比。因此，定义删除某个fact的效益分数为两者之差：

其中的分别表示三者的embedding。

**直接添加攻击：**

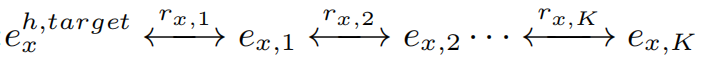
添加包含的fact，使得可信。添加某个fact的效益分数定义为：



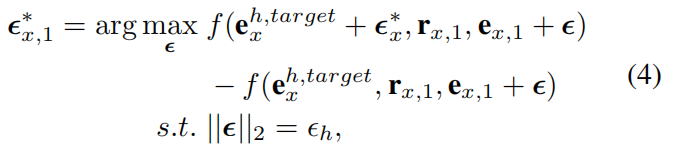
**2. Indirect Attack：**

直接攻击易被数据完整性检查检测出来。非直接攻击更复杂也更隐蔽。

非直接攻击会扰动包含其他实体的facts，并将扰动的影响传播到目标的fact。

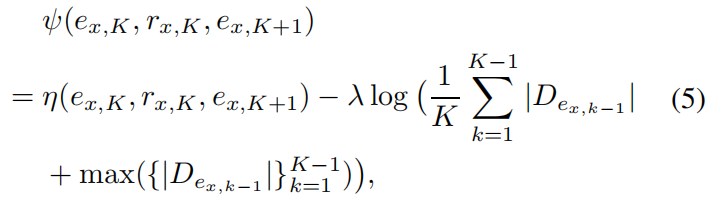
先找到的K-hop邻居实体，这些被称作代理实体，对包含这些实体的facts进行扰动。在K-hop邻居实体与之间的实体叫中间实体。代理实体会将影响传播到，传播路径为。

直接影响的是邻居，其embedding shifting vector 应满足：



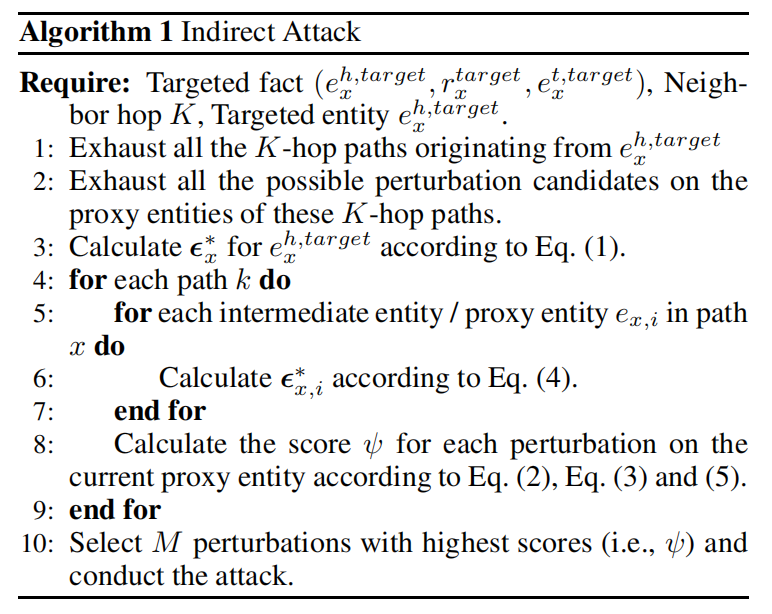
确定了后，就可以计算出。

定义间接扰动的效益分数为：



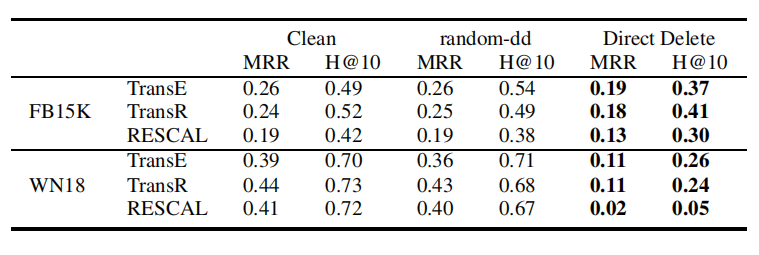
这个式子中右边第一项估计了扰动的直接效益，第二、三项评估的是中间实体在传播路径上对目标实体的影响能力。（5）式的结果越大，扰动效果越好。

非直接攻击流程描述如下：



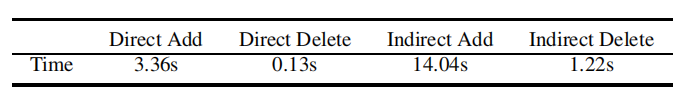
**五、Experiments**

1. 以直接删除攻击效果为例：



（FB15K和WN18是数据集，TransE、TransR、RESCAL是嵌套方式，random-dd是基线方法，MRR和H@10是评估指标，其值越低，说明攻击效果越好。）

可见，直接删除攻击效果最好。

1. 攻击效率：

均攻击时间都不超过15s，其中直接删除攻击平均时间不超过1s。攻击效率很高。