

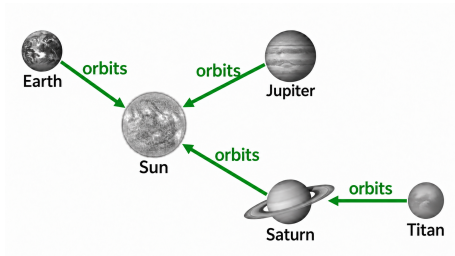
When Confidence Turns Against You: Exploiting Link Prediction Scores for Data Poisoning

Adel Memariani, Arnab Sharma, Michael Röder, Axel-Cyrille Ngonga Ngomo
Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University

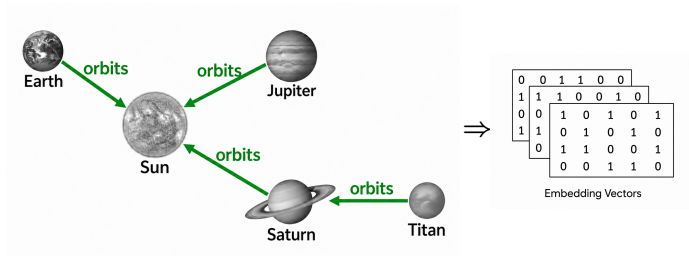


The 25th International Conference on Knowledge Engineering and Knowledge Management
September 29th - October 1st, 2026 - University of Torino, Italy

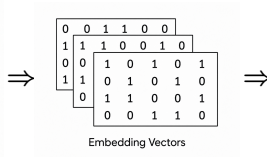
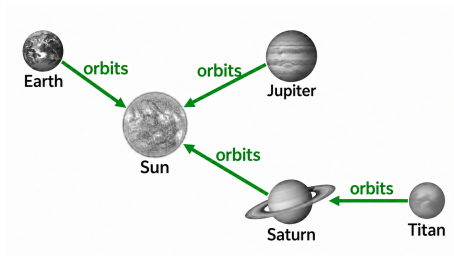
Motivation: poisoning can change link prediction



Motivation: poisoning can change link prediction



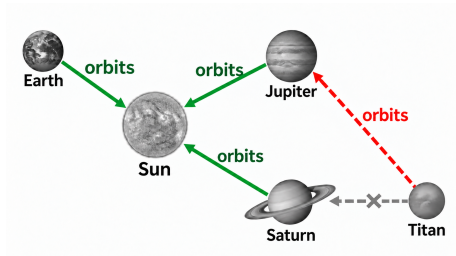
Motivation: poisoning can change link prediction



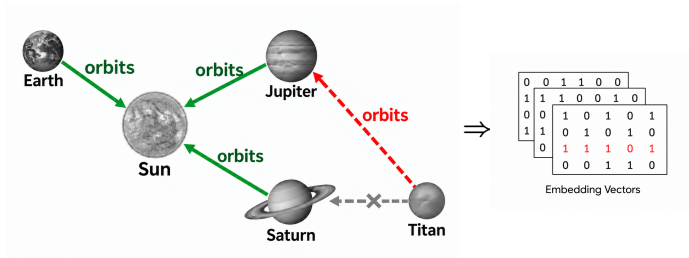
Link Prediction
(Titan, orbits, ?)

1	Saturn	0.93
2	Jupiter	0.12
3	Sun	0.07

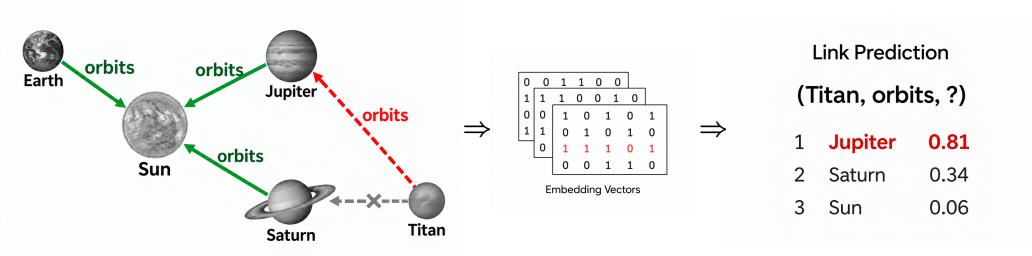
Motivation: poisoning can change link prediction



Motivation: poisoning can change link prediction



Motivation: poisoning can change link prediction



- ▶ Link prediction models learn a scoring function for triples.
- ▶ At prediction phase, **plausibility/confidence scores** are used to rank candidate facts.

- ▶ Link prediction models learn a scoring function for triples.
- ▶ At prediction phase, **plausibility/confidence scores** are used to rank candidate facts.

$\mathcal{E}$: entities $\mathcal{R}$: relations $\mathcal{G}$: knowledge graph

$$\mathcal{G} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$$

- ▶ Link prediction models learn a scoring function for triples.
- ▶ At prediction phase, **plausibility/confidence scores** are used to rank candidate facts.

$\mathcal{E}$: entities $\mathcal{R}$: relations $\mathcal{G}$: knowledge graph

$$\mathcal{G} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$$

$$\forall (h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E} : \text{Score}(h, r, t) \in \mathbb{R}$$

For example, $\text{Score}(\text{Titan}, \text{orbits}, \text{Saturn}) = 0.93$

- ▶ Link prediction models learn a scoring function for triples.
- ▶ At prediction phase, **plausibility/confidence scores** are used to rank candidate facts.

$\mathcal{E}$: entities $\mathcal{R}$: relations $\mathcal{G}$: knowledge graph

$$\mathcal{G} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$$

$$\forall (h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E} : \text{Score}(h, r, t) \in \mathbb{R}$$

For example, $\text{Score}(\text{Titan}, \text{orbits}, \text{Saturn}) = 0.93$

Score $\uparrow \Rightarrow$ more plausible

Score $\downarrow \Rightarrow$ less plausible



Can link prediction scores reveal effective targets for poisoning attacks?

Attacker capabilities

- ▶ Obtain plausibility scores for triples.
- ▶ Edit a limited number of training triples.

In our experiments

A surrogate KGE model provides scores as a proxy for predictions.

Quantile-Based Triple Selection

$T = \mathcal{G}_{\text{train}}$, Each triple $t \in T$ receives a plausibility score s_t

We then **sort** the triples by increasing score:

$$s_{t_{(1)}} \leq s_{t_{(2)}} \leq \dots \leq s_{t_{(|T|)}}$$

Quantile-Based Triple Selection

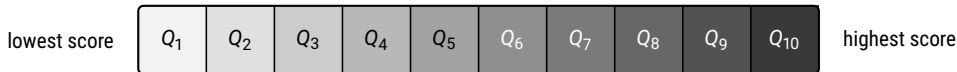
$T = \mathcal{G}_{\text{train}}$, Each triple $t \in T$ receives a plausibility score s_t

We then **sort** the triples by increasing score:

$$s_{t_{(1)}} \leq s_{t_{(2)}} \leq \dots \leq s_{t_{(|T|)}}$$

A triple's position in this ordered list determines its score quantile.

$$\xi(T^\uparrow, a, u) = \{t_{(i)} \mid \lfloor a|T| \rfloor < i \leq \lfloor u|T| \rfloor\}, \quad 0 \leq a < u \leq 1.$$



Random vs. Quantile-Based Poisoning

Triple selection:

Random baseline: $\tau \sim \text{Uniform}(\mathcal{G}_{\text{train}})$ random training triple

Quantile-based: $\tau \sim \text{Uniform}(Q_k)$ triple from score quantile Q_k

Triple selection:

Random baseline: $\tau \sim \text{Uniform}(\mathcal{G}_{\text{train}})$ random training triple

Quantile-based: $\tau \sim \text{Uniform}(Q_k)$ triple from score quantile Q_k

Perturbation [same for both]:

$$z \sim \text{Uniform}\{\text{head}, \text{relation}, \text{tail}\}$$

$$\tau' = \begin{cases} (h', r, t), & z = \text{head}, \quad h' \sim \text{Uniform}(\mathcal{E} \setminus \{h\}) \\ (h, r', t), & z = \text{relation}, \quad r' \sim \text{Uniform}(\mathcal{R} \setminus \{r\}) \\ (h, r, t'), & z = \text{tail}, \quad t' \sim \text{Uniform}(\mathcal{E} \setminus \{t\}) \end{cases}$$

Perturbation budget: $\mathcal{B} \subseteq \mathcal{G}_{\text{train}}, |\mathcal{B}| = \rho |\mathcal{G}_{\text{train}}|, \rho \in \{0.02, 0.04, 0.06\}$

¹MRR: Mean Reciprocal Rank.

Perturbation budget: $\mathcal{B} \subseteq \mathcal{G}_{\text{train}}, |\mathcal{B}| = \rho |\mathcal{G}_{\text{train}}|, \rho \in \{0.02, 0.04, 0.06\}$

Relative MRR¹ drop: $\Delta_{\text{MRR}} = \frac{\text{MRR}_{\text{clean}} - \text{MRR}_{\text{poisoned}}}{\text{MRR}_{\text{clean}}}$

¹MRR: Mean Reciprocal Rank.

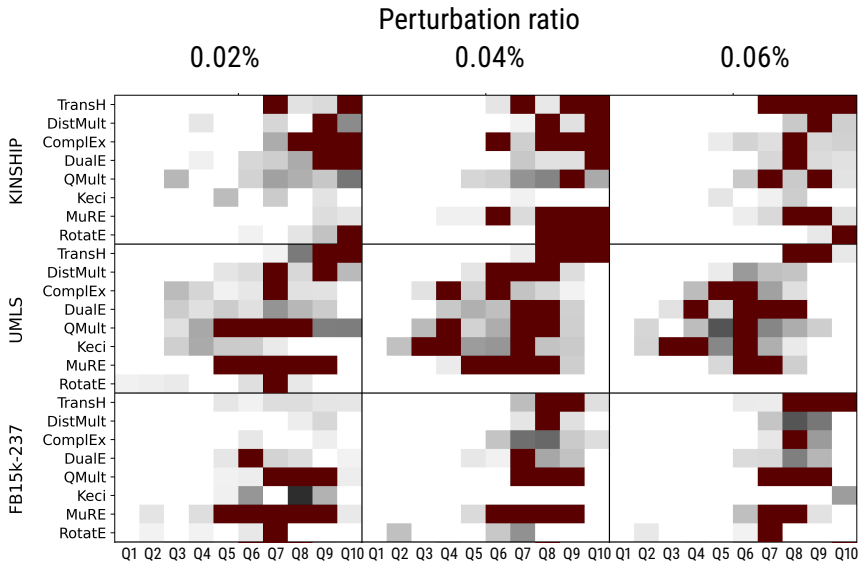
Perturbation budget: $\mathcal{B} \subseteq \mathcal{G}_{\text{train}}, |\mathcal{B}| = \rho |\mathcal{G}_{\text{train}}|, \rho \in \{0.02, 0.04, 0.06\}$

Relative MRR¹ drop: $\Delta_{\text{MRR}} = \frac{\text{MRR}_{\text{clean}} - \text{MRR}_{\text{poisoned}}}{\text{MRR}_{\text{clean}}}$

- ▶ 5 datasets
- ▶ 8 link prediction models
- ▶ 6 random seeds
- ▶ 3 noise ratios
- ▶ Statistical tests (paired t-test)
- ▶ 9,792 experiments

¹MRR: Mean Reciprocal Rank.

Statistical Significance



- ▶ Score-guided poisoning can be more damaging than random poisoning.
- ▶ Mid-high scoring triples (Q7-Q9) are often more influential in denser graphs.

Takeaway: Model confidence can itself become an attack signal.

Thank you

Data Science Group at Paderborn University

Web: dice-research.org

Code: github.com/dice-group

Twitter: [@DiceResearch](https://twitter.com/DiceResearch)



This work has been supported by the Ministry of Culture and Science of North Rhine-Westphalia (MKW NRW) within the project SAIL (grant no NW21-059D) and the project WHALE (LFN 1-04) funded under the Lamarr Fellow Network programme.

Evaluation

DB	Ratio	Att.	TransH	DistMult	ComplEx	DualE	QMult	Keci	RotatE	MuRE
KINSHIP	0.02	Base	-3.2%	-2.7%	-2.8%	-3.5%	-2.9%	-4.9%	-1.7%	-4.7%
		Surr	 -10.5%	 -5.0%	 -4.7%	 -4.9%	 -5.1%	 -5.9%	 -4.9%	 -5.2%
	0.04	Base	-5.4%	-5.5%	-6.2%	-7.8%	-6.7%	-9.9%	-3.9%	-6.2%
		Surr	 -15.9%	 -8.2%	 -8.3%	 -9.8%	 -11.6%	 -10.0%	 -6.5%	 -9.0%
	0.06	Base	-5.9%	-8.4%	-9.8%	-11.1%	-10.1%	-11.7%	-6.5%	-9.3%
		Surr	 -20.6%	 -13.3%	 -12.5%	 -13.1%	 -14.1%	 -13.3%	 -8.5%	 -11.8%
UMLS	0.02	Base	-3.5%	-5.1%	-4.9%	-6.2%	-5.2%	-4.9%	-2.7%	-5.2%
		Surr	 -6.3%	 -6.9%	 -6.2%	 -7.6%	 -7.8%	 -6.1%	 -4.7%	 -7.7%
	0.04	Base	-6.0%	-10.4%	-10.3%	-11.4%	-10.5%	-8.0%	-6.1%	-9.8%
		Surr	 -9.2%	 -12.8%	 -12.7%	 -13.2%	 -13.2%	 -11.9%	 -5.0%	 -13.4%
	0.06	Base	-9.1%	-14.7%	-14.9%	-15.9%	-15.3%	-13.1%	-10.1%	-15.1%
		Surr	 -12.4%	 -16.5%	 -16.9%	 -19.0%	 -18.6%	 -16.0%	 -6.3%	 -18.0%
FB15k-237	0.02	Base	-1.9%	-4.1%	-3.2%	-2.3%	-1.0%	-4.4%	-1.2%	-2.3%
		Surr	 -3.4%	 -5.8%	 -4.0%	 -5.7%	 -3.6%	 -12.9%	 -2.9%	 -4.8%
	0.04	Base	-3.3%	-6.0%	-4.9%	-3.3%	-3.8%	-11.9%	-2.4%	-4.3%
		Surr	 -5.4%	 -8.4%	 -7.3%	 -5.1%	 -5.8%	 -10.8%	 -4.0%	 -7.0%
	0.06	Base	-4.4%	-8.2%	-6.6%	-5.5%	-5.1%	-9.3%	-4.5%	-7.0%
		Surr	 -6.3%	 -12.4%	 -11.8%	 -8.8%	 -7.7%	 -11.8%	 -4.9%	 -9.9%
NELL-995 h100	0.02	Base	-4.9%	-0.2%	-0.1%	-2.8%	-3.4%	-2.2%	-2.1%	-1.4%
		Surr	 -14.4%	 -6.3%	 -8.4%	 -6.3%	 -5.6%	 -8.6%	 -3.2%	 -4.0%
	0.04	Base	-7.9%	-5.3%	-5.4%	-2.3%	-4.4%	-2.3%	-5.0%	-4.1%
		Surr	 -11.8%	 -9.9%	 -11.2%	 -6.8%	 -12.9%	 -13.3%	 -5.2%	 -7.5%
	0.06	Base	-11.5%	-2.9%	-6.3%	-4.3%	-8.1%	-7.0%	-5.2%	-9.6%
		Surr	 -15.0%	 -11.9%	 -13.9%	 -11.7%	 -12.9%	 -13.5%	 -9.0%	 -9.3%
WN18RR	0.02	Base	-4.3%	-4.3%	-5.5%	-1.3%	-1.4%	-3.6%	-3.8%	-5.7%
		Surr	 -5.6%	 -5.4%	 -12.5%	 -37.7%	 -0.5%	 -9.2%	 -5.2%	 -6.0%
	0.04	Base	-8.6%	-7.9%	-11.7%	-18.2%	-5.8%	-6.2%	-5.0%	-9.6%
		Surr	 -8.9%	 -9.1%	 -16.8%	 -32.4%	 -9.2%	 -10.4%	 -10.6%	 -12.3%
	0.06	Base	-10.2%	-12.2%	-17.2%	-30.9%	-11.1%	-7.9%	-6.9%	-13.5%
		Surr	 -13.8%	 -13.5%	 -25.8%	 -44.5%	 -13.8%	 -13.0%	 -14.6%	 -16.5%
Quantiles:  Q1  Q2  Q3  Q4  Q5  Q6  Q7  Q8  Q9  Q10										

Statistical Significance

