Adversarial Attack against Cross-lingual Knowledge Graph Alignment

Zeru Zhang[1], **Zijie Zhang**[1], **Yang Zhou**[1], **Lingfei Wu**[2], **Sixing Wu**[3], **Xiaoying Han**[1], **Dejing Dou**[4,5], **Tianshi Che**[1], **Da Yan**[6]
[1]Auburn University, [2]JD.COM Silicon Valley Research Center, [3]Peking University,
[4]University of Oregon, [5]Baidu Research, [6]University of Alabama at Birmingham
{zeruzhang, zijiezhang, yangzhou, xzh0003, tianshiche}@auburn.edu,
lwu@email.wm.edu, wusixing@pku.edu.cn, dou@cs.uoregon.edu, yanda@uab.edu

Abstract

Recent literatures have shown that knowledge graph (KG) learning models are highly vulnerable to adversarial attacks. However, there is still a paucity of vulnerability analyses of cross-lingual entity alignment under adversarial attacks. This paper proposes an adversarial attack model with two novel attack techniques to perturb the KG structure and degrade the quality of deep cross-lingual entity alignment. First, an entity density maximization method is employed to hide the attacked entities in dense regions in two KGs, such that the derived perturbations are unnoticeable. Second, an attack signal amplification method is developed to reduce the gradient vanishing issues in the process of adversarial attacks for further improving the attack effectiveness.

1 Introduction

Today, multilingual knowledge graphs (KGs), such as WordNet (Miller, 1992), DBpedia (Auer et al., 2007), YAGO (Hoffart et al., 2011), and ConceptNet (Speer et al., 2017), are becoming essential sources of knowledge for various AI-related applications, e.g., personal assistants, medical diagnosis, and online question answering. Cross-lingual entity alignment between multilingual KGs is a powerful tool that align the same entities in different monolingual KGs together, automatically synchronize different language-specific KGs and revolutionize the understanding of these ubiquitous multilingual KGs in a transformative manner (Xu et al., 2020b; Sun et al., 2020a; Berrendorf et al., 2021b,a).

Unfortunately, real-world KGs are typically noisy due to two main reasons: (1) massive fake information injected by malicious parties and users on online encyclopedia websites (e.g., Wikipedia (Wik) and Answers.com (Ans)), social networks (e.g., Twitter and Facebook), online communities (e.g., Reddit and Yahoo Answers), news websites, and search engines that usually serve as data sources of the KGs; and (2) direct adversarial attacks on the KGs. Google Knowledge Graph has been criticized for providing answers without source attribution or citation, and thus undermines people's ability to verify information and to develop well-informed opinions (Dewey, 2016).

Recent studies have shown that KG learning models remain highly sensitive to adversarial attacks, i.e., carefully designed small perturbations in KGs can cause the models to produce wrong prediction results, including knowledge graph embedding (Minervini et al., 2017; Pujara et al., 2017; Pezeshkpour et al., 2019; Zhang et al., 2019; Banerjee et al., 2021) and knowledge graph-based dialogue generation (Xu et al., 2020a). However, existing techniques focus on the adversarial attacks on single KG learning tasks. These techniques cannot be directly utilized to attack the cross-lingual entity alignment models, as they have to analyze relations within and across KGs. Two critical questions still keep unsolved: (1) Can small perturbations on KGs defeat cross-lingual entity alignment models? (2) How to design effective and unnoticeable perturbations against cross-lingual entity alignment?

The majority of cross-lingual entity alignment techniques aim to train the model by minimizing the distance between pre-aligned entity pairs in training data, such that the corresponding entity embeddings across KGs are close to each other, and the entity pairs with the smallest distance in test data are output as alignment results (Mao et al., 2020a; Wu et al., 2020b; Mao et al., 2020b; Tang et al., 2020; Yan et al., 2021; Zhu et al., 2021; Mao et al., 2021; Pei et al., 2020).

In terms of the distribution of entities in a KG, one idea of perturbing an entity unobtrusively is to move the entity to a dense region in the KG with many similar entities by adding/deleting relations to/from it is able to move it to a dense region in the KG with many similar entities, such that it is non-trivial to recognize the modified entity in the

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5320–5337
November 7–11, 2021.

dense region with many similar entities.

Existing gradient-based adversarial attack methods (Goodfellow et al., 2015; Madry et al., 2018) search for the weakest input features to attack by calculating the loss gradient. However, the vanishing gradient problem is often encountered when training neural networks with poor backward signal propagation and thus leads to the attack failures (Athalye et al., 2018). Can we enhance the attack signal propagation for improving the attack effectiveness?

In this work, an entity density estimation and maximization method is employed to first estimate the distribution of entities in KGs. Based on the estimated KG distributions, the entities to be attacked are then moved to dense regions in two KGs by maximizing their densities. The attacked entities are hidden in dense regions in two KGs, such that they are surrounded by many neighbors in dense regions as well as indistinguishable from these neighbors. In addition, the surrounding of many neighbors makes it difficult to identify the correctly aligned entity pairs among many similar candidate entities.

We comprehensively study how poor signal propagation on neural networks leads to vanishing gradients in adversarial attacks over cross-lingual entity alignment. An attack signal amplification method is developed to secure informative attack signals with both well-conditioned Jacobian and competent signal propagation from the alignment loss. This reduces the gradient vanishing issues in the process of adversarial attacks for further improving the attack effectiveness.

Extensive experiments over real-world KG datasets validate the superior attack performance of the EAA model against several state-of-the-art cross-lingual entity alignment models. To our best knowledge, this work is the first to study adversarial attacks on cross-lingual entity alignment.

2 Problem Formulation

Given two input knowledge graphs G^1 and G^2. Each is denoted as $G^k = (E^k, R^k, T^k)$ $(1 \leq k \leq 2)$, where $E^k = \{e_1^k, \cdots, e_{N^k}^k\}$ is the set of N^k entities, $R^k = \{r_{ij}^k = (e_i^k, e_j^k) : 1 \leq i, j \leq N^k, i \neq j\}$ is the set of relations, and $T^k = E^k \times R^k \times E^k$ is the set of triples. Each triple $t_l^k = (e_i^k, r_{ij}^k, e_j^k) \in T^k$ in G^K denotes head entity e_i^k connected to tail entity e_j^k through relation r_{ij}^k. $\mathbf{A}^k$ is an $N^k \times N^k$ adjacency matrix that denotes the structure information of G^K. By using knowledge graph embedding (KGE), each triple can be presented as $(\mathbf{e}_i^k, \mathbf{r}_{ij}^k, \mathbf{e}_j^k)$, where boldfaced $\mathbf{e}_i^k$, $\mathbf{r}_{ij}^k$, and $\mathbf{e}_j^k$ represent the embedding vectors of head e_i^k, relation r_{ij}^k, and tail e_j^k respectively.

D contains a set of pre-aligned entity pairs $D = \{(e_i^1, e_j^2) | e_i^1 \leftrightarrow e_j^2, e_i^1 \in E^1, e_j^2 \in E^2\}$, where $e_i^1 \leftrightarrow e_j^2$ indicates that two entities e_i^1 and e_j^2 are the equivalent ones in different language-specific KGs. The cross-lingual entity alignment aims to utilize D as the training data to identify the one-to-one entity alignments between entities e_i^1 and e_j^2 in two cross-lingual KGs G^1 and G^2 in the test data.

Most of existing cross-lingual entity alignment models are supervised learning methods with minimizing the distances (or maximizing the similarities) between the embeddings of pre-aligned entity pairs e_i^1 and e_j^2 in D (Wang et al., 2018; Sun et al., 2020d; Wu et al., 2020b; Pei et al., 2020; Tang et al., 2020; Yan et al., 2021). The entity pairs e_i^1 and e_j^2 in the test data with the largest similarities are selected as the alignment results. The following loss function is minimized to learn a KGE model $h : e_i^k \in E^k \mapsto \mathbf{e}_i^k$. h is often implemented as a graph convolutional network (GCN) for deep KGE.

$$\begin{aligned} \min_h \mathcal{L} = &- \sum_{(e_i^1, e_j^2) \in D} \log \sigma\big((\mathbf{e}_i^1)^T \cdot \mathbf{e}_j^2\big) \\ &+ \sum_{(e_{i'}^1, e_{j'}^2) \notin D} \log \sigma\big((\mathbf{e}_{i'}^1)^T \cdot \mathbf{v}_{j'}^2\big) \end{aligned} \tag{1}$$

where (e_i^1, e_j^2) and $(e_{i'}^1, e_{j'}^2)$ are positive and negative entity pairs. $(\mathbf{e}_i^1)^T$ is the transpose of $\mathbf{e}_i^1$. $\sigma(\cdot)$ is the sigmoid function. The inner product $\cdot$ denotes the similarity between two embedding vectors.

Given a trained deep KGE model $\mathbf{e}_i^k = h(e_i^k)$, an adversarial attacker aims to maximally degrade the alignment performance of h by injecting effective and unnoticeable relation perturbations (including relation addition and deletion) into two clean KGs G^k $(1 \leq k \leq 2)$, leading to two perturbed KGs $\hat{G}^k = (\hat{E}^k, \hat{R}^k, \hat{T}^k)$.

$$\max_{\hat{\mathbf{A}}^k} \mathcal{L} \; s.t. \; |\hat{\mathbf{A}}^k - \mathbf{A}^k| \leq \Delta, 1 \leq k \leq 2 \tag{2}$$

where $\mathbf{A}^k$ and $\hat{\mathbf{A}}^k$ are clean and perturbed adjacency matrices respectively. Δ is the allowed attack budget, i.e., allowed relation modifications.

3 Unnoticeable Adversarial Attacks

Existing GCN-based entity alignment methods often initialize entity features with random initialization or pre-trained word embeddings of entity names and utilize adjacency matrix of KGs to learn the entity embeddings (Wang et al., 2018; Sun et al., 2020d; Wu et al., 2020b; Yan et al., 2021). Thus, the embedding of an entity mainly depends on the embeddings of its neighbor entities. In order to modify the embedding of a target entity for the purpose of adversarial attacks, we need to remove some positive (i.e., existing) relations and add some negative (i.e., non-existing) relations between the target entity and its neighbors in adjacency matrix, and thus degrade the accuracy of entity embedding and alignment. We use the i^{th} row of adjacency matrix $\mathbf{A}^k$ (i.e., $\mathbf{A}^k_i$) to represent structure features of each entity e^k_i and analyze the impact of each structure feature (i.e., positive or negative relation) on the alignment accuracy.

As shown in Figure 1, assuming that $\mathbf{e}^1_i$ and $\mathbf{e}^2_j$ are pre-aligned entity embeddings, if we hide an entity e^1_i in a dense region with many similar e^1_ks by modifying its associated relations, then the surrounding of many e^1_ks makes it difficult to differentiate e^1_i from many similar e^1_ks and identify the correctly aligned entity pairs $\mathbf{e}^1_i$ and $\mathbf{e}^2_j$ among many similar candidate entities e^1_ks. In addition, if another pair of entity embeddings $\mathbf{e}^1_k$ and $\mathbf{e}^2_j$ are more similar than the pre-aligned entity embeddings $\mathbf{e}^1_i$ and $\mathbf{e}^2_j$, i.e., $(\mathbf{e}^1_k)^T \cdot \mathbf{e}^2_j > (\mathbf{e}^1_i)^T \cdot \mathbf{e}^2_j$, then we will obtain an incorrect alignment result (e^1_k, e^2_j).

In this work, we will leverage our proposed kernel density estimation method (Zhang et al., 2020b) to estimate the distribution of perturbed KGs and maximize the distance between pre-aligned entity pairs for degrading the performance of entity alignment as well as for hiding the attacked entities in dense regions in two KGs. The kernel density estimation method is essentially to estimate a probability density function (PDF) $f(x)$ of a random variable x for revealing the intrinsic distribution of x (Parzen, 1962). Let $\mathbf{x}^k$ be a N^k-dimensional random variable to denote the structure features of all entities $\{\mathbf{A}^k_i, \cdots, \mathbf{A}^K_{N^K}\}$ in KG G^k for estimating a PDF $f(\mathbf{x}^k)$.

$$f(\mathbf{x}^k) = \frac{1}{N^k \det(\mathbf{B})} \sum_{i=1}^{N^k} \mathcal{K}\left(\mathbf{B}^{-1}\left(\mathbf{x}^k - \mathbf{A}^k_i\right)\right) \tag{3}$$

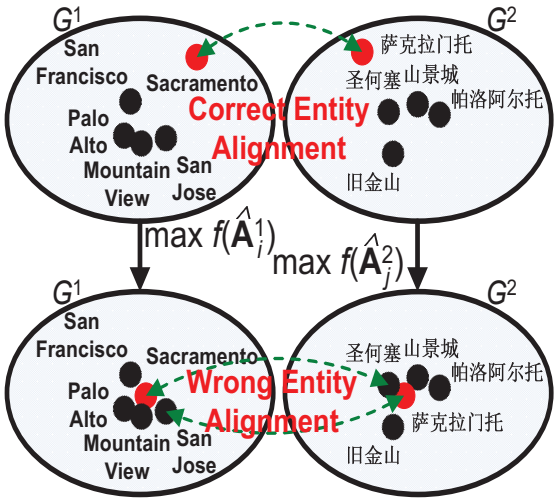


Figure 1: Unnoticeable Adversarial Attacks

where $\det(\cdot)$ denotes the determinant operation. $\mathbf{B} > 0$ is a bandwidth to be estimated. It is an $N^k \times N^k$ diagonal matrix $\mathbf{B} = diag(b_1, \cdots, b_{N^k})$, which has strong influence on the density estimation $f(\mathbf{x}^k)$. A good $\mathbf{B}$ should be as small as the data can allow. $\mathcal{K}$ is a product symmetric kernel that satisfies $\int \mathcal{K}(x)dx = 1$ and $\int x\mathcal{K}(x)dx = 0$. The vector form $f(\mathbf{x}^k)$ can be rewritten as an element form, where $\mathbf{x}^k_j$ denotes the j^{th} dimension in $\mathbf{x}^k$.

$$f(\mathbf{x}^k) = \frac{1}{N^k} \sum_{i=1}^{N^k} \prod_{j=1}^{N^k} \frac{1}{b_j} \mathcal{K}\Big(\frac{\mathbf{x}^k_j - \mathbf{A}^k_{ij}}{b_j}\Big) \tag{4}$$

We then calculate the derivative $\frac{\partial f(\mathbf{x}^k)}{\partial b_j}$ about each b_j in $\mathbf{B}$.

$$\frac{\partial f(\mathbf{x}^k)}{\partial b_j} = \frac{1}{N^k} \sum_{i=1}^{N^k} \frac{\partial\Big[\prod_{l=1}^{N^k} \frac{1}{b_l} \mathcal{K}\big(\frac{\mathbf{x}^k_l - \mathbf{A}^k_{il}}{b_l}\big)\Big]}{\partial b_j} =$$
$$-\frac{1}{N^k} \sum_{i=1}^{N^k} \Big(\frac{1}{b_j} + \frac{\mathbf{x}^k_l - \mathbf{A}^k_{il}}{b_j^2} K\big(\frac{\mathbf{x}^k_l - \mathbf{A}^k_{il}}{b_j}\big)\Big)$$
$$\prod_{l=1}^{N^k} \frac{1}{b_l} \mathcal{K}\big(\frac{\mathbf{x}^k_l - \mathbf{A}^k_{il}}{b_l}\big) \tag{5}$$

We make use of a greedy search method to determine bandwidths in the kernel density estimation method. For a non-trivial/trivial dimension j, updating the bandwidth b_j will have a

strong/weak influence over $f(\mathbf{x}^k)$. We greedily reduce b_j with a sequence $b_0, b_0 s, b_0 s^2, \cdots$ for a parameter $0 < s < 1$, until b_j is smaller than a certain threshold τ_j, to validate whether a small update in b_j is able to lead to a large update in $f(\mathbf{x}^k)$.

We use an initial $\mathbf{B} = diag(b_0, \cdots, b_0)$ for a large b_0 to estimate $\frac{\partial f(\mathbf{x}^k)}{\partial b_j}$, and reduce b_j when $\frac{\partial f(\mathbf{x}^k)}{\partial b_j}$ is larger than a certain threshold.

$$\begin{aligned}\frac{\partial f(\mathbf{x}^k)}{\partial b_j} &= \frac{1}{N^k}\sum_{i=1}^{N^k}\frac{\partial\Big[\prod_{l=1}^{N^1}\frac{1}{b_l}\mathcal{K}\big(\frac{\mathbf{x}_l^1-\mathbf{A}_{il}^1}{b_l}\big)\Big]}{\partial b_j} \\ &= \frac{1}{N^k}\sum_{i=1}^{N^k}\frac{\mathcal{K}\big(\frac{\mathbf{x}_j^1-\mathbf{A}_{ij}^1}{b_j}\big)}{\mathcal{K}\big(\frac{\mathbf{x}_j^1-\mathbf{A}_{ij}^1}{b_j}\big)}\prod_{l=1}^{N^k}\mathcal{K}\big(\frac{\mathbf{x}_l^1-\mathbf{A}_{il}^1}{b_l}\big) \\ &= \frac{1}{N^k}\sum_{i=1}^{N^k}\frac{\partial f(\mathbf{x}_i^k)}{\partial b_j}\end{aligned} \tag{6}$$

We derive the corresponding variance $\mathrm{Var}\Big(\frac{\partial f(\mathbf{x}^k)}{\partial b_j}\Big)$ as follows.

$$\mathrm{Var}\Big(\frac{\partial f(\mathbf{x}^k)}{\partial b_j}\Big) = \mathrm{Var}\Big(\frac{1}{N^k}\sum_{i=1}^{N^k}\frac{\partial f(\mathbf{x}_i^k)}{\partial b_j}\Big) \tag{7}$$

According to the estimated bandwidth $\mathbf{B}$ by Algorithm 1, we can calculate density $f(\mathbf{x}^k)$ of $\mathbf{x}^k$ in Eq.(3). The perturbation process is to maximize the following attack loss $\mathcal{L}_A$ for producing unnoticeable perturbations, in terms of the estimations $f(\mathbf{x}^1)$ and $f(\mathbf{x}^2)$ in two KGs G^1 and G^2.

$$\begin{aligned}\max_{\hat{\mathbf{A}}^k}\mathcal{L}_A = \Big[&\sum_{(e_i^1,e_j^2)\in D} -\log\sigma\big((\hat{\mathbf{e}}_i^1)^T\cdot\hat{\mathbf{e}}_j^2\big) \\ &+ f(\hat{\mathbf{A}}_i^1) + f(\hat{\mathbf{A}}_j^2)\Big] \\ &+ \sum_{(e_{i'}^1,e_{j'}^2)\notin D}\log\sigma\big((\mathbf{e}_{i'}^1)^T\cdot\mathbf{v}_{j'}^2\big) \\ s.t.\ &|\hat{\mathbf{A}}_i^k - \mathbf{A}_i^k| \le \Delta, 1\le k\le 2\end{aligned} \tag{8}$$

where $\hat{\mathbf{A}}_i^1 = \mathbf{A}_i^1 + \delta_i^1$ (and $\hat{\mathbf{A}}_j^2 = \mathbf{A}_j^2 + \delta_j^2$) denote perturbations of clean structure features $\mathbf{A}_i^1$ (and $\mathbf{A}_j^2$) in G^1 (and G^2) by adding a small amount of relation perturbations δ_i^1 (and δ_j^2), such that $\hat{\mathbf{e}}_i^1$ is far away from $\hat{\mathbf{e}}_j^2$ and thus the alignment accuracy is decreased. In addition, we push e_i^1 and e_j^2 to dense regions to generate $\hat{e}_i^1$ and $\hat{e}_j^2$, by maximizing $f(\hat{\mathbf{A}}_i^1)$ and $f(\hat{\mathbf{A}}_j^2)$, such that $\hat{e}_i^1$ and $\hat{e}_j^2$ are indistinguishable from their neighbors in perturbed KGs. This reduces the possibility of perturbation detection by humans or defender programs.

Algorithm 1 Kernel Density Estimation

Input: KG $G^k = (E^k, R^k, T^k)$, parameter $0 < s < 1$, initial bandwidth b_0, and parameter c.
Output: Bandwidth matrix $\mathbf{B}$.

```
Initialize all b_1, ..., b_{N^k} with b_0;
for each j = 1 to N^k
  do
    Estimate derivative ∂f(x^k)/∂b_j and variance Var(∂f(x^k)/∂b_j);
    Compute τ_j = sqrt(2 · Var(∂f(x^k)/∂b_j) · log(cN^k));
    if |∂f(x^k)/∂b_j| > τ_j, then Update b_j = b_j s;
  while |∂f(x^k)/∂b_j| > τ_j
Return B.
```

We leverage the Projected Gradient Descent (PGD) technique (Madry et al., 2018) to produce perturbed adjacency matrices $\hat{\mathbf{A}}^1$ and $\hat{\mathbf{A}}^2$ of two KGs G^1 and G^2.

$$\begin{aligned}(\mathbf{A}_i^1)^{(t+1)} &= \Pi_{\triangle^1}\mathrm{sgn}\big[\mathrm{ReLU}\big(\nabla_{(\mathbf{A}_i^1)^t}\mathcal{L}_A\big] \\ (\mathbf{A}_j^2)^{(t+1)} &= \Pi_{\triangle^2}\mathrm{sgn}\big[\mathrm{ReLU}\big(\nabla_{(\mathbf{A}_j^2)^t}\mathcal{L}_A\big], \\ &t = 1, \cdots, T\end{aligned} \tag{9}$$

where $(\mathbf{A}_i^1)^{(t+1)}$ and $(\mathbf{A}_j^2)^{(t+1)}$ denotes the perturbations of $\mathbf{A}_i^1$ and $\mathbf{A}_j^2$ derived at step t. ϵ specifies the budget of allowed perturbed relations for each attacked entity. $\triangle^k = \{(\delta^k)^t | \mathbf{1}^T(\delta^k)^t \le \epsilon, (\delta^k)^t \in \{0,1\}^{N^k}\}$, where $(\delta^k)^t = \|(\mathbf{A}_i^1)^t - \mathbf{A}_i^1\|_2^2$, represents the constraint set of the projection operator Π, i.e., it encodes whether a relation in $\mathbf{A}_i^1$ is modified or not. The composition of the ReLU and sign operators guarantees $(\mathbf{A}_i^1)^t \in \{0,1\}^{N^1}$ and $(\mathbf{A}_j^2)^t \in \{0,1\}^{N^2}$, as it adds (or removes) an relation or keeps it unchanged when an derivate in the gradient is positive (or negative). The outputs $(\mathbf{A}_i^1)^T$ and $(\mathbf{A}_j^2)^T$ at final step T are used as the perturbed adjacency matrices $\hat{\mathbf{A}}_i^1$ and $\hat{\mathbf{A}}_j^2$.

4 Effective Adversarial Attacks

Unfortunately, the above PGD-based unnoticeable attack method needs to iteratively calculate the gradient $\nabla_{(\mathbf{A}_i^1)}\mathcal{L}_A$, which mainly depends on

$\frac{\partial\left(\log\sigma((\mathbf{e}_i^1)^T\cdot\mathbf{e}_j^2)\right)}{\partial\mathbf{A}_i^1}$ in the GCN-based entity alignment models.

Given an alignment signal $\phi\big((\mathbf{e}_i^1)^T,\mathbf{e}_j^2\big) = \frac{\partial\left(\log\sigma((\mathbf{e}_i^1)^T\cdot\mathbf{e}_j^2)\right)}{\partial(\mathbf{e}_i^1)^T}$ and a Jacobian matrix $\mathbf{J}_i = \frac{\partial(\mathbf{e}_i^1)^T}{\partial\mathbf{A}_i^1}$, the gradient of $\log\sigma((\mathbf{e}_i^1)^T\cdot\mathbf{e}_j^2)$ is calculated as follows.

$$\begin{aligned}&\frac{\partial\left(\log\sigma((\mathbf{e}_i^1)^T\cdot\mathbf{e}_j^2)\right)}{\partial\mathbf{A}_i^1}\\ =&\frac{\partial\left(\log\sigma((\mathbf{e}_i^1)^T\cdot\mathbf{e}_j^2)\right)}{\partial(\mathbf{e}_i^1)^T}\frac{\partial(\mathbf{e}_i^1)^T}{\partial\mathbf{A}_i^1}\\ =&\phi\big((\mathbf{e}_i^1)^T,\mathbf{e}_j^2\big)\mathbf{J}_i\end{aligned}\tag{10}$$

It is obvious that the gradient is determined with both the signal and the Jacobian together. The situation that either the signal has saturating gradient or the Jacobian is insignificant is able to result in vanishing gradients in $\frac{\partial\left(\log\sigma((\mathbf{e}_i^1)^T\cdot\mathbf{e}_j^2)\right)}{\partial\mathbf{A}_i^1}$ and thus the attack failures.

All singular values of a neural network's input-output Jacobian matrix concentrate near 1 is a property known as dynamical isometry (Pennington et al., 2017). Ensuring the mean squared singular value of a network's input-output Jacobian is $O(1)$ is essential for avoiding the exponential vanishing or explosion of gradients. We leverage the dynamical isometry theory for improving the effectiveness of the PGD adversarial attacks. Concretely, a neural network is dynamical isometry if all singular values λ_{ir} of the Jacobian $\mathbf{J}_i$ are close to 1, i.e., $1-\lambda_{ir}\leq\xi$ for $\forall r, r\in\{1,\cdots,\min\{N^1,N^2\}\}$ and a small positive number $\xi\approx 0$. In our problem, when the Jacobian matrix $\mathbf{J}_i$ is dynamical isometry, the signal $\phi\big((\mathbf{e}_i^1)^T,\mathbf{e}_j^2\big)$ backpropagates isometrically over the neural network and maintains the norm and all angles between vectors.

Intuitively, if we select a good attack signal amplification factor α to amplify $\mathbf{e}_i^1$ and $\mathbf{e}_j^2$ as follows, then this can improve the diffusion of attack signals. In addition, a good α should guarantee the relative order of the network's output logits invariant, to ensure the decision boundary of entity alignment unchanged.

$$\tilde{\mathbf{e}}_i^1=\alpha\mathbf{e}_i^1,\ \tilde{\mathbf{e}}_j^2=\alpha\mathbf{e}_j^2\tag{11}$$

We rewrite the gradients with α as follows.

$$\begin{aligned}&\frac{\partial\left(\log\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)\right)}{\partial\mathbf{A}_i^1}\\ =&\frac{\partial\left(\log\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)\right)}{\partial(\tilde{\mathbf{e}}_i^1)^T}\frac{\partial(\tilde{\mathbf{e}}_i^1)^T}{\partial(\mathbf{e}_i^1)^T}\frac{\partial(\mathbf{e}_i^1)^T}{\partial\mathbf{A}_i^1}\\ =&\phi\big((\tilde{\mathbf{e}}_i^1)^T,\tilde{\mathbf{e}}_j^2\big)\alpha\mathbf{J}_i\end{aligned}\tag{12}$$

Notice that $\phi\big((\tilde{\mathbf{e}}_i^1)^T,\tilde{\mathbf{e}}_j^2\big) = \frac{\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)\left(1-\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)\right)\tilde{\mathbf{e}}_j^2}{\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)} = \big(1-\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)\big)\tilde{\mathbf{e}}_j^2$. When α is close to ∞, the alignment signal $\phi\big((\tilde{\mathbf{e}}_i^1)^T,\tilde{\mathbf{e}}_j^2\big)$ approaches zero and thus the vanishing gradient problem is encountered in adversarial attacks. In addition, all singular values of $\alpha\mathbf{J}_i$ are equal to zeros if $\alpha=0$. $\frac{\partial\left(\log\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)\right)}{\partial\mathbf{A}_i^1}$ is equal to zero, which leads to the vanishing gradient problem too.

Therefore, a desired α for avoiding the exponential vanishing of gradients should stand in between 0 and ∞, in order to guarantee the signal $\phi\big((\tilde{\mathbf{e}}_i^1)^T,\tilde{\mathbf{e}}_j^2\big)$ large enough, i.e., $\|\phi\big((\tilde{\mathbf{e}}_i^1)^T,\tilde{\mathbf{e}}_j^2\big)\|_2>\eta$ for a positive threshold η, as well as make all singular values of $\alpha\mathbf{J}_i$ close to 1, such that the signal $\phi\big((\tilde{\mathbf{e}}_i^1)^T,\tilde{\mathbf{e}}_j^2\big)$ can be well backpropagated from the output layer to the input layer.

In order to make the mean of singular values of $\alpha\mathbf{J}_i$ close to 1, the first option of α is the inverse of the mean of singular values of $\mathbf{J}_i$.

$$\alpha=\frac{|D|N}{\sum_{i=1}^{|D|}\sum_{r=1}^{N}\lambda_{ir}}\tag{13}$$

where λ_{ir} is the r^{th} singular value of $\mathbf{J}_i$. $|D|$ is the size of the set D of pre-aligned entity pairs and $N=\min\{N^1,N^2\}$.

For the purpose of ensuring $\|\phi\big((\tilde{\mathbf{e}}_i^1)^T,\tilde{\mathbf{e}}_j^2\big)\|_2>\eta$, the second option of α should be satisfied with $1-\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)>\eta/\|\tilde{\mathbf{e}}_j^2\|_2$. The feasible α can be obtained through the following theorem.

Theorem 1. *Let entity embedding vectors $\tilde{\mathbf{e}}_k^2$ and $\tilde{\mathbf{e}}_l^2$ be the most similar and least similar to $(\tilde{\mathbf{e}}_i^1)^T$ ($1\leq k,l\leq N^2$), i.e., $\tilde{\mathbf{e}}_k^2=\mathrm{argmax}_{\tilde{\mathbf{e}}_k^2}(\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_k^2$ and $\tilde{\mathbf{e}}_l^2=\mathrm{argmin}_{\tilde{\mathbf{e}}_l^2}(\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_l^2$, and $c=(\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_k^2$. Also, suppose that d is the minimal norm of entity embedding vectors in G^2, i.e., $d=\min_{\tilde{\mathbf{e}}_m^2}\|\tilde{\mathbf{e}}_m^2\|_2$ for $\forall e_m^2\in E^2$. For a given $0<\eta<d/2$, if $\alpha<$*

Algorithm 2 Effective Adversarial Attacks

Input: KG $G^k = (E^k, R^k, T^k)$, set of pre-aligned entity pairs $D = \{(e_i^1, e_j^2)|e_i^1 \leftrightarrow e_j^2\}$, trained entity embedding model h, noise budget ϵ, and signal threshold η.
Output: Perturbed adjacency matrices $\{\hat{\mathbf{A}}_i^1, \hat{\mathbf{A}}_j^2|(e_i^1, e_j^2) \in D\}$.

1: **for** each pair (e_i^1, e_j^2) in D
2: Set $\hat{\mathbf{e}}_i^1 = \mathbf{e}_i^1 = h(e_i^1)$, $\hat{\mathbf{e}}_j^2 = \mathbf{e}_j^2 = h(e_j^2)$;
3: Compute $\alpha_1 = \frac{|D|N}{\sum_{i=1}^{|D|}\sum_{r=1}^{N}\lambda_{ir}}$ in Eq.(13);
4: **for** $t = 1, \cdots, T$
5: Initialize $\alpha_2 = 1.0$;
6: **if** $1 - \sigma((\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_j^2) \leq \eta / \|\tilde{\mathbf{e}}_j^2\|_2$
7: Update $\alpha_2 = \sqrt{\frac{1}{c}\log\frac{d-\eta}{\eta}}$ in Theorem 1;
8: Amplify $\tilde{\mathbf{e}}_i^1 = \alpha_1\alpha_2\mathbf{e}_i^1$, $\tilde{\mathbf{e}}_j^2 = \alpha_1\alpha_2\mathbf{e}_j^2$;
9: Calculate $\frac{\partial\left(\log\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)\right)}{\partial\tilde{\mathbf{A}}_i^1}$ and $\frac{\partial\left(\log\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2)\right)}{\partial\tilde{\mathbf{A}}_j^2}$;
10: Use the PGD to update $\hat{\mathbf{A}}_i^1$, $\hat{\mathbf{A}}_j^2$ in Eq.(9);
11: **Return** $\{\hat{\mathbf{A}}_i^1, \hat{\mathbf{A}}_j^2|(e_i^1, e_j^2) \in D\}$.

$\sqrt{\frac{1}{c}\log\frac{d-\eta}{\eta}}$, *then* $1-\sigma((\tilde{\mathbf{e}}_i^1)^T\cdot\tilde{\mathbf{e}}_j^2) > \eta/\|\tilde{\mathbf{e}}_j^2\|_2$ *for* $\forall e_j^2 \in E^2$.

Proof. Please see Appendix A for proof.

Algorithm 2 combines the above two kinds of α to produce effective adversarial attacks with attack signal amplification. The perturbed entity embeddings $\hat{\mathbf{e}}_i^1$ and $\hat{\mathbf{e}}_j^2$ are initialized with clean ones $\mathbf{e}_i^1$ and $\mathbf{e}_j^2$ in step 2. The first amplification factor α_1 is calculated in step 3. The second factor α_2 is computed in steps 5-7. α_1 and α_2 are integrated together for enhancing the attack signal propagation of neural networks in steps 8-9. The PGD attack method with attack signal amplification is utilized to perturb the KGs. The algorithm repeats the above iterative procedure until convergence.

5 Experimental Evaluation

Table 1 presents the statistics of the DBP15K datasets (Sun et al., 2017). They consist of three different cross-lingual datasets which are $DBP15K_{ZH-EN}$, $DBP15K_{JA-EN}$, and $DBP15K_{FR-EN}$. Each cross-lingual dataset contains two monolingual KGs in different languages and 15,000 pre-aligned entity pairs between two KGs. In the experiment, 30% pre-aligned entity pairs are used for training data and the remaining are used for test data.

Dataset		#Entities	#Relations	#Triples	#Alignments
ZH-EN	ZH	66,469	2,830	153,929	15,000
	EN	98,125	2,317	237,674	
JA-EN	JA	65,744	2,043	164,373	15,000
	EN	95,680	2,096	233,319	
FR-EN	FA	66,858	1,379	192,191	15,000
	EN	105,889	2,209	278,590	

Table 1: Statistics of Datasets

We compare the EAA model with seven state-of-the-art attack models. Sememe-based Word Substitution (**SWS**) incorporates the sememe-based word substitution and swarm optimization-based search to conduct word-level attacks (Zang et al., 2020). Inflection Word Swap (**IWS**) perturb the inflectional morphology of words to craft plausible and semantically similar adversarial examples (Tan et al., 2020; Morris et al., 2020). We utilize the above two word-level attack models to replace associated entities of a relation based on semantics. **GF-Attack** attacks graph embedding methods by devising new loss and approximating the spectrum (Chang et al., 2020). **LowBlow** is a general low-rank adversarial attack model which is able to affect the performance of various graph learning tasks (Entezari et al., 2020). We use the above two graph attack models to directly add/remove relations in terms of graph topology. **CRIAGE** aims to add/remove the facts to/from the KG that degrades the performance of link prediction (Pezeshkpour et al., 2019). **DPA** contains a collection of data poisoning attack strategies against knowledge graph embedding (Zhang et al., 2019). **RL-RR** uses reinforcement learning policy to produce deceptively perturbed KGs while keeping the downstream quality of the original KG (Raman et al., 2021). To our best knowledge, this work is the first to study adversarial attacks on cross-lingual entity alignment.

We evaluate four versions of EAA to show the strengths of different components. **EAA-P** uses the basic PGD (Madry et al., 2018) to produce adversarial attacks. **EAA-D** only utilizes the KDE and density maximization to generate effective and unnoticeable attacks. **EAA-A** employs only our attack signal amplification strategy to improve the performance of the basic PGD attack. **EAA** operates with the full support of both KDE and signal amplification components.

We validate the effectiveness of the above attack models with three representative cross-lingual entity alignment algorithms. **AttrGNN** integrates

	AttrGNN		RNM		REA	
Attacks	**Hits@1**	**MRR**	**Hits@1**	**MRR**	**Hits@1**	**MRR**
Clean	0.796	0.845	0.841	0.875	0.792	0.818
SWS	0.726	0.839	0.745	0.862	0.764	0.848
IWS	0.708	0.761	0.729	0.823	0.759	0.804
GF-Attack	0.709	0.815	0.724	0.833	0.733	0.844
LowBlow	0.677	0.773	0.678	0.776	0.697	0.797
CRIAGE	0.646	0.704	0.655	0.719	0.662	0.715
DPA	0.603	0.712	0.636	0.751	0.635	0.733
RL-RR	0.562	0.684	0.628	0.713	0.637	0.722
EAA	**0.497**	**0.538**	**0.525**	**0.636**	**0.538**	**0.641**

Table 2: Results on $DBP15K_{ZH-EN}$ with 5% perturbed relations

	AttrGNN		RNM		REA	
Attacks	**Hits@1**	**MRR**	**Hits@1**	**MRR**	**Hits@1**	**MRR**
Clean	0.783	0.834	0.872	0.899	0.799	0.823
SWS	0.724	0.839	0.774	0.854	0.788	0.843
IWS	0.718	0.787	0.755	0.804	0.745	0.796
GF-Attack	0.715	0.824	0.747	0.826	0.767	0.845
LowBlow	0.737	0.783	0.728	0.802	0.723	0.821
CRIAGE	0.705	0.756	0.699	0.769	0.707	0.769
DPA	0.643	0.725	0.723	0.753	0.669	0.766
RL-RR	0.689	0.716	0.691	0.765	0.706	0.768
EAA	**0.579**	**0.612**	**0.618**	**0.642**	**0.621**	**0.652**

Table 3: Results on $DBP15K_{JA-EN}$ with 5% perturbed relations

both attribute and relation triples for better performance of cross-lingual entity alignment (Liu et al., 2020). **RNM** is a novel relation-aware neighborhood matching model for entity alignment (Zhu et al., 2021). To our best knowledge, **REA** is the only robust cross-lingual entity alignment solution against adversarial attacks by detecting noise in the perturbed inter-KG entity links (Pei et al., 2020).

We use two popular metrics in entity alignment to verify the attack effectiveness: $Hits@k$ (i.e., the ratio of correctly aligned entities ranked in the top k candidates) and MRR (i.e., mean reciprocal rank). A smaller $Hits@k$ or MRR indicates a worse entity alignment but a better attack. K is fixed to 1 in all tests.

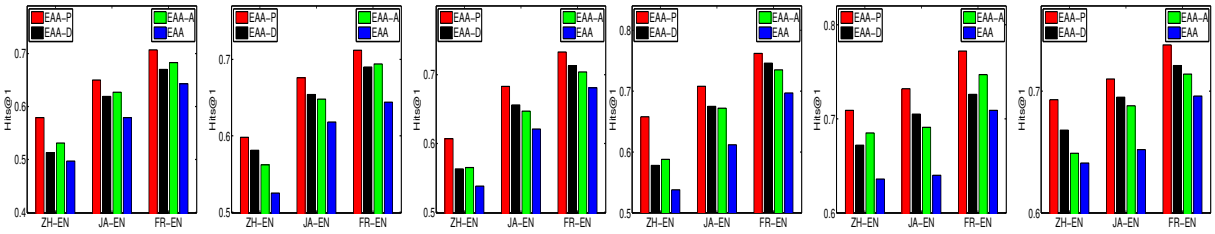
Attack performance on various datasets with different entity alignment algorithms. Table 2-4 exhibit the Hits@1 and MRR scores of three GCN-based entity alignment algorithms on test data by nine attack models over three groups of cross-lingual datasets. Clean represents that the experiments run on the original KGs without any perturbations. For all other attack models, the number of perturbed relations is fixed to 5% in these experiments. It is observed that among nine attack methods, no matter how strong the attacks are, the EAA method achieve the lowest Hits@1 and MRR scores on perturbed KGs in most experiments, showing the effectiveness of EAA for the adversarial attacks. Compared to the entity alignment results under other attack models, EAA, on average, achieves 17.7%, 12.8%, and 12.8% improvement of Hits@1 and 17.6%, 16.9%, and 13.7% boost of MRR on $DBP15K_{ZH-EN}$, $DBP15K_{JA-EN}$, and $DBP15K_{FR-EN}$ respectively. In addition, the promising performance of EAA with all three entity alignment models implies that EAA has great potential as a general attack solution to other entity alignment methods, which is desirable in practice.

	AttrGNN		RNM		REA	
Attacks	**Hits@1**	**MRR**	**Hits@1**	**MRR**	**Hits@1**	**MRR**
Clean	0.919	0.91	0.938	0.954	0.812	0.855
SWS	0.782	0.873	0.814	0.886	0.807	0.846
IWS	0.755	0.801	0.803	0.836	0.802	0.806
GF-Attack	0.715	0.828	0.779	0.848	0.792	0.848
LowBlow	0.792	0.841	0.799	0.826	0.793	0.852
CRIAGE	0.733	0.864	0.744	0.873	0.781	0.831
DPA	0.704	0.757	0.796	0.817	0.695	0.791
RL-RR	0.754	0.792	0.745	0.823	0.754	0.784
EAA	**0.643**	**0.697**	**0.644**	**0.709**	**0.681**	**0.696**

Table 4: Results on $DBP15K_{FR-EN}$ with 5% perturbed relations

Ablation study. Figure 2 and 3 present the Hits@1 and MRR scores achieved by three entity alignment methods under adversarial attacks with four variants of our EAA attack model. We have observed the complete EAA achieves the lowest Hits@1 (< 0.681) and the smallest MRR scores (< 0.709) respectively, which are obviously better than other versions. Notice that EAA-A achieves the better attack performance than EAA-P in most tests. A reasonable explanation is that our attack signal amplification technique is able to alleviate the vanishing gradient issue, which effectively helps maintain the utility of adversarial attacks in GCN-based entity alignment models. In addition, EAA-D also performs well in most experiments, compared with EAA-P. A rational guess is that it is difficult to correctly match the entities in two KGs when they lie in dense regions with many similar entities. These results illustrate both KDE and signal amplification methods are important in producing effective and unnoticeable attacks in entity alignment.

Attack performance with varying perturbed relations. Figure 4 presents the performance of entity alignment under nine attack models by varying the ratios of perturbed edges from 5% to 30%. It is obvious that the attacking performance improves for each attacker with an increase in the number of perturbed edges. This phenomenon indicates that current GCN-based entity alignment methods are very sensitive to adversarial attacks. EAA achieves

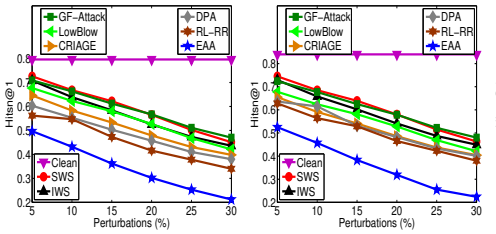


(a) AttrGNN (b) RNM (c) REA

Figure 2: $Hits@1$ of EAA variants

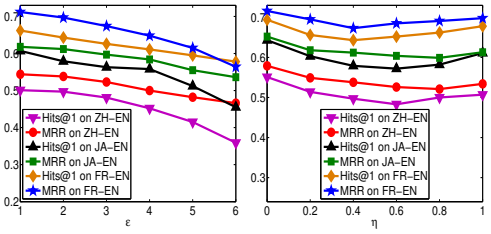
(a) AttrGNN (b) RNM (c) REA

Figure 3: MRR of EAA variants



(a) AttrGNN on ZH-EN (b) RNM on ZH-EN (c) REA on ZH-EN

Figure 4: $Hits@1$ with varying perturbed relations



(a) Perturbation budget ϵ (b) Signal threshold η

Figure 5: Results with varying parameters

the lowest Hits@1 values (< 0.538), which are still better than the other eight methods in most tests. Especially, when the perturbation ratio is large than 10%, the Hits@1 values drop quickly.

Impact of perturbation budget ϵ. Figure 5 (a) measures the performance effect of ϵ in the EAA model for the entity alignment by varying ϵ from 1 to 6. It is observed that when increasing ϵ, both Hits@1 and MRR scores of the EAA model decreases substantially. This demonstrates it is difficult to train a robust entity alignment model under large ϵ constraint. However, a large ϵ can be easily detected by humans or by defender programs. Notice that the average number of associated relations of each entity in three datasets is between 2.3 and 2.9. Thus we suggest generating both effective and unnoticeable attacks for the entity alignment task under ϵ between 2 and 3, such that ϵ is smaller than the average number of associated relations.

Impact of signal threshold η. Figure 5 (b) shows the impact of η in our EAA model over three groups of datasets. The performance curves initially drop when η increases. Intuitively, this can help alleviate the vanishing gradient issue in the PGD adversarial attacks. Later on, the performance curves keep relatively stable or even increasing when η continuously increases. A reasonable explanation is that the too large η makes the upper bound of α too small. This results in poor-conditioned Jacobian and thus leads to the vanishing gradient issue again. Thus, it is important to determine the optimal η for the EAA model.

6 Related Work

Knowledge graph alignment. Knowledge graph alignment techniques have attracted active research in the last decade (Xu et al., 2020b; Sun et al., 2020a; Berrendorf et al., 2021b,a) and can be broadly classified into two categories: (1) Translation-based techniques, which denote entities by computing the plausibility of relational facts measured by a specific fact plausibility scoring function, including MtransE (Chen et al., 2017a), IPTransE (Zhu et al., 2017), JAPE (Sun et al., 2017), BootEA (Sun et al., 2018), RSNs (Guo et al., 2019), NAEA (Zhu et al., 2019), OTEA (Pei et al., 2019b), TransEdge (Sun et al., 2019), HyperKA (Sun et al., 2020c). The idea of this kind of methods are originated from cross-lingual word embedding techniques. Thus, they are able to capture fine-grained fact semantics. However, they fail to preserve the global topological structure of knowledge graphs; (2) GCN-based methods, which utilize GCN models to model global structure information of knowledge graphs by recursively aggregating the features of neighbors of each entity, such as GCN-Align (Wang et al., 2018), SEA (Pei et al., 2019a), MuGNN (Cao et al., 2019), HopGCN (Xu

et al., 2019c), NAEA (Zhu et al., 2019), AVR-GCN (Ye et al., 2019), RDGCN (Wu et al., 2019a), HGCN-JE (Wu et al., 2019b), KECG (Li et al., 2019), MRAEA (Mao et al., 2020a), AliNet (Sun et al., 2020d), CG-MuAlign (Zhu et al., 2020), NMN (Wu et al., 2020b), DAT (Zeng et al., 2020), SSP (Nie et al., 2020), RREA (Mao et al., 2020b), DINGAL (Yan et al., 2021), RNM (Zhu et al., 2021), JEANS (Chen et al., 2021), Dual-AMN (Mao et al., 2021), KE-GCN (Yu et al., 2021). These methods can fully utilize the topological and neighborhood information to learn better representations of entities. However, it is difficult to model fine-grained fact semantics.

Adversarial attacks on text and graph data. Recent studies have presented that NLP and graph models, especially DNN models, are highly sensitive to adversarial attacks, i.e., carefully designed small deliberate perturbations in input intended to result in analysis failures (Song et al., 2018; Chen et al., 2020; Xu et al., 2019a; Wang et al., 2019; Zhang et al., 2020a; Huq and Pervin, 2020).

In the NLP area, the majority of research efforts focus on attacking the corpus in different models, including dialogue generation (Niu and Bansal, 2018), machine translation (Belinkov and Bisk, 2018; Tan et al., 2020; Niu et al., 2020), model-agnostic attacks (Wallace et al., 2019; Zang et al., 2020; Morris et al., 2020), natural language inference (Abdou et al., 2020; Chan et al., 2020; Li et al., 2020b), reading comprehension (Jia and Liang, 2017; Blohm et al., 2018; Tan et al., 2020), and sentiment classification (Wu et al., 2020c; Kurita et al., 2020; Wang et al., 2020).

Graph data analysis have attracted active research in the last decade (Cheng et al., 2009; Zhou et al., 2009, 2010; Cheng et al., 2011; Zhou and Liu, 2012; Cheng et al., 2012; Lee et al., 2013; Su et al., 2013; Zhou et al., 2013; Zhou and Liu, 2013; Palanisamy et al., 2014; Zhou et al., 2014; Zhou and Liu, 2014; Su et al., 2015; Zhou et al., 2015b; Bao et al., 2015; Zhou et al., 2015d; Zhou and Liu, 2015; Zhou et al., 2015a,c; Lee et al., 2015; Zhou et al., 2016; Zhou, 2017; Palanisamy et al., 2018; Zhou et al., 2018b,a; Ren et al., 2019; Zhou et al., 2019c,b,d; Zhou and Liu, 2019; Wu et al., 2020a, 2021a; Zhou et al., 2020b; Zhang et al., 2020b; Zhou et al., 2020c,a; Goswami et al., 2020; Zhou et al., 2021b; Zhao et al., 2021; Ren et al., 2021; Jin et al., 2021; Wu et al., 2021b; Zhou et al., 2021a; Zhang et al., 2021; Liu et al., 2021). Various adversarial attack models have been developed to show the vulnerability of graph learning models in node classification (Dai et al., 2018; Zügner et al., 2018; Wang and Gong, 2019; Xu et al., 2019b; Zügner and Günnemann, 2019; Takahashi, 2019; Entezari et al., 2020; Sun et al., 2020b; Ma et al., 2020; Zügner et al., 2020; Xi et al., 2021; He et al., 2021), community detection (Chen et al., 2017b; Waniek et al., 2018; Chen et al., 2019; Li et al., 2020a), network embedding (Chen et al., 2018; Bojchevski and Günnemann, 2019; Chang et al., 2020), graph classification (Dai et al., 2018; Xi et al., 2021), link prediction (Zhou et al., 2019a), similarity search (Dey and Medya, 2020), malware detection (Hou et al., 2019), and graph matching (Zhang et al., 2020b).

Only recently, researchers have started to develop adversarial attack techniques to maximally degrade the performance of knowledge graph learning in knowledge graph embedding (Minervini et al., 2017; Pujara et al., 2017; Pezeshkpour et al., 2019; Zhang et al., 2019; Banerjee et al., 2021) and knowledge graph-based dialogue generation (Xu et al., 2020a). REA detects noise in the perturbed inter-graph links for robust cross-lingual entity alignment (Pei et al., 2020). RL-RR aims to produce deceptively perturbed knowledge graphs, which maintain the downstream performance of the original knowledge graph while significantly deviating from the original knowledge graph's semantics and structure (Raman et al., 2021).

7 Conclusions

We have studied the problem of adversarial attacks against cross-lingual entity alignment. First, we proposed to utilize kernel density estimation technique to estimate and maximize the densities of attacked entities and generate effective and unnoticeable perturbations, by pushing attacked entities to dense regions in two KGs. Second, we analyze how gradient vanishing causes failures of gradient-based adversarial attacks. We design an attack signal amplification method to ensure informative signal propagation. The EAA model achieves superior performance against representative attack models.

8 Ethical Considerations

In this work, all the three knowledge graph datasets are open-released by previous works for research (Sun et al., 2017). All the three datasets are widely used in training/evaluating the cross-lingual entity alignment, for example, (Liu et al.,

2020; Zhu et al., 2021; Pei et al., 2020; Yan et al., 2021; Mao et al., 2021). All the three datasets are open-accessed resources that everyone can see and no privacy-related data (such as gender, nickname, birthday, etc.) are included. All the three knowledge graph datasets are originally collected and filtered from Wikipedia (under the license CC BY-SA 3.0). It is allowed to reuse them in research. But if it needs commercial use, it may need to ask for additional permission from the original author/copyright owner (Wik; Sun et al., 2017). To summary, as research work, this work has no concerns on the dataset and other aspects. But if someone wants to use the same/similar data as us in commercial, they have to further check the licenses.

References

Answers.com. https://www.answers.com.

Wikipedia. http://www.wikipedia.org/.

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. The sensitivity of language models and humans to winograd schema perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7590–7604.

Anish Athalye, Nicholas Carlini, and David A. Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 274–283.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007.*, pages 722–735.

Prithu Banerjee, Lingyang Chu, Yong Zhang, Laks V.S. Lakshmanan, and Lanjun Wang. 2021. Stealthy targeted data poisoning attack on knowledge graphs. In *Proceedings of the 37th IEEE International Conference on Data Engineering, IEEE 2021*.

Xianqiang Bao, Ling Liu, Nong Xiao, Yang Zhou, and Qi Zhang. 2015. Policy-driven autonomic configuration management for nosql. In *Proceedings of the 2015 IEEE International Conference on Cloud Computing (CLOUD'15)*, pages 245–252, New York, NY.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Max Berrendorf, Evgeniy Faerman, and Volker Tresp. 2021a. Active learning for entity alignment. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, pages 48–62.

Max Berrendorf, Ludwig Wacker, and Evgeniy Faerman. 2021b. A critical assessment of state-of-the-art in entity alignment. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, pages 18–32.

Matthias Blohm, Glorianna Jagfeld, Ekta Sood, Xiang Yu, and Ngoc Thang Vu. 2018. Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 108–118.

Aleksandar Bojchevski and Stephan Günnemann. 2019. Adversarial attacks on node embeddings via graph poisoning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 695–704.

Yixin Cao, Zhiyuan Liu, Chengjiang Li, Zhiyuan Liu, Juanzi Li, and Tat-Seng Chua. 2019. Multi-channel graph neural network for entity alignment. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1452–1461.

Alvin Chan, Yi Tay, Yew-Soon Ong, and Aston Zhang. 2020. Poison attacks against text datasets with conditional adversarially regularized autoencoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 4175–4189.

Heng Chang, Yu Rong, Tingyang Xu, Wenbing Huang, Honglei Zhang, Peng Cui, Wenwu Zhu, and Junzhou Huang. 2020. A restricted black-box adversarial framework towards attacking graph embedding models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New Yrok, NY, USA, February 7 - 12, 2020*.

Jinyin Chen, Lihong Chen, Yixian Chen, Minghao Zhao, Shanqing Yu, Qi Xuan, and Xiaoniu Yang. 2019. Ga-based q-attack on community detection. *IEEE Trans. Comput. Social Systems*, 6(3):491–503.

Jinyin Chen, Yangyang Wu, Xuanheng Xu, Yixian Chen, Haibin Zheng, and Qi Xuan. 2018. Fast gradient attack on network embedding. *CoRR*, abs/1809.02797.

Liang Chen, Jintang Li, Jiaying Peng, Tao Xie, Zengxu Cao, Kun Xu, Xiangnan He, and Zibin Zheng. 2020. A survey of adversarial learning on graphs. *CoRR*, abs/2003.05730.

Muhao Chen, Weijia Shi, Ben Zhou, and Dan Roth. 2021. Cross-lingual entity alignment with incidental supervision. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 645–658.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017a. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1511–1517.

Yizheng Chen, Yacin Nadji, Athanasios Kountouras, Fabian Monrose, Roberto Perdisci, Manos Antonakakis, and Nikolaos Vasiloglou. 2017b. Practical attacks against graph-based clustering. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30-November 03, 2017*, pages 1125–1142.

Hong Cheng, David Lo, Yang Zhou, Xiaoyin Wang, and Xifeng Yan. 2009. Identifying bug signatures using discriminative graph mining. In *Proceedings of the 18th International Symposium on Software Testing and Analysis (ISSTA'09)*, pages 141–152, Chicago, IL.

Hong Cheng, Yang Zhou, Xin Huang, and Jeffrey Xu Yu. 2012. Clustering large attributed information networks: An efficient incremental computing approach. *Data Mining and Knowledge Discovery (DMKD)*, 25(3):450–477.

Hong Cheng, Yang Zhou, and Jeffrey Xu Yu. 2011. Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(2):1–33.

Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. 2018. Adversarial attack on graph structured data. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1123–1132.

Caitlin Dewey. 2016. You probably haven't even noticed google's sketchy quest to control the world's knowledge. https://www.washingtonpost.com/news/the-intersect/wp/2016/05/11/you-probably-havent-even-noticed-googles-sketchy-quest-to-control-the-worlds-knowledge/.

Palash Dey and Sourav Medya. 2020. Manipulating node similarity measures in networks. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*.

Negin Entezari, Saba Al-Sayouri, Amirali Darvishzadeh, and Evangelos Papalexakis. 2020. All you need is low (rank): Defending against adversarial attacks on graphs. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining, WSDM 2020, Houston, TX, February 3-7, 2020*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sayan Goswami, Ayam Pokhrel, Kisung Lee, Ling Liu, Qi Zhang, and Yang Zhou. 2020. Graphmap: Scalable iterative graph processing using nosql. *The Journal of Supercomputing (TJSC)*, 76(9):6619–6647.

Lingbing Guo, Zequn Sun, and Wei Hu. 2019. Learning to exploit long-term relational dependencies in knowledge graphs. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2505–2514.

Xinlei He, Jinyuan Jia, Michael Backes, Neil Zhenqiang Gong, and Yang Zhang. 2021. Stealing links from graph neural networks. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. 2011. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011 (Companion Volume)*, pages 229–232.

Shifu Hou, Yujie Fan, Yiming Zhang, Yanfang Ye, Jingwei Lei, Wenqiang Wan, Jiabin Wang, Qi Xiong, and Fudong Shao. 2019. α*Cyber*: Enhancing robustness of android malware detection system against adversarial attacks on heterogeneous graph based model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 609–618.

Aminul Huq and Mst. Tasnim Pervin. 2020. Adversarial attacks and defense on texts: A survey. *CoRR*, abs/2005.14108.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 2021–2031.

Ruoming Jin, Dong Li, Jing Gao, Zhi Liu, Li Chen, and Yang Zhou. 2021. Towards a better understanding of linear models for recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'21)*, Virtual Event.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2793–2806.

Kisung Lee, Ling Liu, Karsten Schwan, Calton Pu, Qi Zhang, Yang Zhou, Emre Yigitoglu, and Pingpeng Yuan. 2015. Scaling iterative graph computations with graphmap. In *Proceedings of the 27th IEEE international conference for High Performance Computing, Networking, Storage and Analysis (SC'15)*, pages 57:1–57:12, Austin, TX.

Kisung Lee, Ling Liu, Yuzhe Tang, Qi Zhang, and Yang Zhou. 2013. Efficient and customizable data partitioning framework for distributed big rdf data processing in the cloud. In *Proceedings of the 2013 IEEE International Conference on Cloud Computing (CLOUD'13)*, pages 327–334, Santa Clara, CA.

Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. 2019. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2723–2732.

Jia Li, Honglei Zhang, Zhichao Han, Yu Rong, Hong Cheng, and Junzhou Huang. 2020a. Adversarial attack on community detection by hiding individuals. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 917–927.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. BERT-ATTACK: adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6193–6202.

Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. 2021. From distributed machine learning to federated learning: A survey. *CoRR*, abs/2104.14362.

Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, and Tat-Seng Chua. 2020. Exploring and evaluating attributes, values, and structures for entity alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6355–6364.

Jiaqi Ma, Shuangrui Ding, and Qiaozhu Mei. 2020. Towards more practical adversarial attacks on graph neural networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Online*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021. Boosting the speed of entity alignment 10x: Dual attention matching network with normalized hard sample mining. In *WWW '21: The Web Conference 2021, April 19-23, 2021*.

Xin Mao, Wenting Wang, Huimin Xu, Man Lan, and Yuanbin Wu. 2020a. MRAEA: an efficient and robust entity alignment approach for cross-lingual knowledge graph. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 420–428.

Xin Mao, Wenting Wang, Huimin Xu, Yuanbin Wu, and Man Lan. 2020b. Relational reflection entity alignment. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1095–1104.

George A. Miller. 1992. WORDNET: a lexical database for english. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992*.

Pasquale Minervini, Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2017. Adversarial sets for regularising neural link predictors. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 119–126.

Hao Nie, Xianpei Han, Le Sun, Chi Man Wong, Qiang Chen, Suhui Wu, and Wei Zhang. 2020. Global structure and local semantics-preserved embeddings for entity alignment. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3658–3664.

Tong Niu and Mohit Bansal. 2018. Adversarial over-sensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 486–496.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8538–8544.

Balaji Palanisamy, Ling Liu, Kisung Lee, Shicong Meng, Yuzhe Tang, and Yang Zhou. 2014. Anonymizing continuous queries with delay-tolerant mix-zones over road networks. *Distributed and Parallel Databases (DAPD)*, 32(1):91–118.

Balaji Palanisamy, Ling Liu, Yang Zhou, and Qingyang Wang. 2018. Privacy-preserving publishing of multilevel utility-controlled graph datasets. *ACM Transactions on Internet Technology (TOIT)*, 18(2):24:1–24:21.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Shichao Pei, Lu Yu, Robert Hoehndorf, and Xiangliang Zhang. 2019a. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3130–3136.

Shichao Pei, Lu Yu, Guoxian Yu, and Xiangliang Zhang. 2020. REA: robust cross-lingual entity alignment between knowledge graphs. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 2175–2184.

Shichao Pei, Lu Yu, and Xiangliang Zhang. 2019b. Improving cross-lingual entity alignment via optimal transport. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3231–3237.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. 2017. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS) 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4785–4795.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3336–3347.

Jay Pujara, Eriq Augustine, and Lise Getoor. 2017. Sparsity and noise: Where knowledge graph embeddings fall short. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1751–1756.

Mrigank Raman, Siddhant Agarwal, Peifeng Wang, Aaron Chan, Hansen Wang, Sungchul Kim, Ryan Rossi, Handong Zhao, Nedim Lipka, and Xiang Ren. 2021. Learning to deceive knowledge graph augmented models via targeted perturbation. In *9th International Conference on Learning Representations, ICLR 2021, Online, May 4-7, 2021, Conference Track Proceedings*.

Jiaxiang Ren, Zijie Zhang, Jiayin Jin, Xin Zhao, Sixing Wu, Yang Zhou, Yelong Shen, Tianshi Che, Ruoming Jin, and Dejing Dou. 2021. Integrated defense for resilient graph matching. In *Proceedings of the 38th International Conference on Machine Learning, (ICML'21)*, pages 8982–8997, Virtual Event.

Jiaxiang Ren, Yang Zhou, Ruoming Jin, Zijie Zhang, Dejing Dou, and Pengwei Wang. 2019. Dual adversarial learning based network alignment. In *Proceedings of the 19th IEEE International Conference on Data Mining (ICDM'19)*, pages 1288–1293, Beijing, China.

Wenzhuo Song, Shengsheng Wang, Bo Yang, You Lu, Xuehua Zhao, and Xueyan Liu. 2018. Learning node and edge embeddings for signed networks. *Neurocomputing*, 319:42–54.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4444–4451.

Zhiyuan Su, Ling Liu, Mingchu Li, Xinxin Fan, and Yang Zhou. 2013. Servicetrust: Trust management in service provision networks. In *Proceedings of the 10th IEEE International Conference on Services Computing (SCC'13)*, pages 272–279, Santa Clara, CA.

Zhiyuan Su, Ling Liu, Mingchu Li, Xinxin Fan, and Yang Zhou. 2015. Reliable and resilient trust management in distributed service provision networks. *ACM Transactions on the Web (TWEB)*, 9(3):1–37.

Jian Sun, Yu Zhou, and Chengqing Zong. 2020a. Dual attention network for cross-lingual entity alignment. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3190–3201.

Yiwei Sun, Suhang Wang, Xianfeng Tang, Tsung-Yu Hsieh, and Vasant G. Honavar. 2020b. Adversarial attacks on graph neural networks via node injections: A hierarchical reinforcement learning approach. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 673–683.

Zequn Sun, Muhao Chen, Wei Hu, Chengming Wang, Jian Dai, and Wei Zhang. 2020c. Knowledge association with hyperbolic knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5704–5716.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I*, pages 628–644.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4396–4402.

Zequn Sun, JiaCheng Huang, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Transedge: Translating relation-contextualized embeddings for knowledge graphs. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I*, pages 612–629.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020d. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 222–229.

Tsubasa Takahashi. 2019. Indirect adversarial attacks via poisoning neighbors for graph convolutional networks. In *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, December 9-12, 2019*, pages 1395–1400.

Samson Tan, Shafiq R. Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2920–2935.

Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang, Hong Chen, and Cuiping Li. 2020. BERT-INT: A bert-based interaction model for knowledge graph alignment. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3174–3180.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2153–2162.

Binghui Wang and Neil Zhenqiang Gong. 2019. Attacking graph-based classification via manipulating the graph structure. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 2023–2040.

Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. 2020. T3: tree-autoencoder constrained adversarial text generation for targeted attack. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6134–6150.

Wenqi Wang, Lina Wang, Run Wang, Zhibo Wang, and Aoshuang Ye. 2019. Towards a robust deep neural network in texts: A survey. *CoRR*, abs/1902.07285.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 349–357.

Marcin Waniek, Tomasz P. Michalak, Michael J. Wooldridge, and Talal Rahwan. 2018. Hiding individuals and communities in a social network. *Nature Human Behaviour*, 2:139–147.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020a. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (ACL'20)*, pages 5811–5820, Online.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2021a. Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, (IJCAI'20)*, pages 3766–3772, Online.

Sixing Wu, Minghui Wang, Dawei Zhang, Yang Zhou, Ying Li, and Zhonghai Wu. 2021b. Knowledge-aware dialogue generation via hierarchical infobox accessing and infobox-dialogue interaction graph

network. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence, (IJCAI'21)*, Online.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019a. Relation-aware entity alignment for heterogeneous knowledge graphs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5278–5284.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2019b. Jointly learning entity and relation representations for entity alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 240–249.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2020b. Neighborhood matching network for entity alignment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6477–6487.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020c. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4166–4176.

Zhaohan Xi, Ren Pang, Shouling Ji, and Ting Wang. 2021. Graph backdoor. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*.

Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. 2019a. Adversarial attacks and defenses in images, graphs and text: A review. *CoRR*, abs/1909.08072.

Hongcai Xu, Junpeng Bao, and Gaojie Zhang. 2020a. Dynamic knowledge graph-based dialogue generation with improved adversarial meta-learning. *CoRR*, abs/2004.08833.

Kaidi Xu, Hongge Chen, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Mingyi Hong, and Xue Lin. 2019b. Topology attack and defense for graph neural networks: An optimization perspective. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3961–3967.

Kun Xu, Linfeng Song, Yansong Feng, Yan Song, and Dong Yu. 2020b. Coordinated reasoning for cross-lingual knowledge graph alignment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9354–9361.

Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. 2019c. Cross-lingual knowledge graph alignment via graph matching neural network. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3156–3161.

Yuchen Yan, Lihui Liu, Yikun Ban, Baoyu Jing, and Hanghang Tong. 2021. Dynamic knowledge graph alignment. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*.

Rui Ye, Xin Li, Yujie Fang, Hongyu Zang, and Mingzhong Wang. 2019. A vectorized relational graph convolutional network for multi-relational network alignment. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4135–4141.

Donghan Yu, Yiming Yang, Ruohong Zhang, and Yuexin Wu. 2021. Knowledge embedding based graph convolutional network. In *WWW '21: The Web Conference 2021, April 19-23, 2021*.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6066–6080.

Weixin Zeng, Xiang Zhao, Wei Wang, Jiuyang Tang, and Zhen Tan. 2020. Degree-aware alignment for entities in tail. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 811–820.

Gong Zhang, Yang Zhou, Sixing Wu, Zeru Zhang, and Dejing Dou. 2021. Cross-lingual entity alignment with adversarial kernel embedding and adversarial knowledge translation. *CoRR*, abs/2104.07837.

Hengtong Zhang, Tianhang Zheng, Jing Gao, Chenglin Miao, Lu Su, Yaliang Li, and Kui Ren. 2019. Data poisoning attack against knowledge graph embedding. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4853–4859.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Abdulrahmn F. Alhazmi, and Chenliang Li. 2020a. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3):24:1–24:41.

Zijie Zhang, Zeru Zhang, Yang Zhou, Yelong Shen, Ruoming Jin, and Dejing Dou. 2020b. Adversarial attacks on deep graph matching. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020 (NeurIPS'20)*, Virtual.

Xin Zhao, Zeru Zhang, Zijie Zhang, Lingfei Wu, Jiayin Jin, Yang Zhou, Ruoming Jin, Dejing Dou, and Da Yan. 2021. Expressive 1-lipschitz neural networks for robust multiple graph learning against adversarial attacks. In *Proceedings of the 38th International Conference on Machine Learning, (ICML'21)*, pages 12719–12735, Virtual Event.

Kai Zhou, Tomasz P. Michalak, Marcin Waniek, Talal Rahwan, and Yevgeniy Vorobeychik. 2019a. Attacking similarity-based link prediction in social networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 305–313.

Yang Zhou. 2017. *Innovative Mining, Processing, and Application of Big Graphs*. Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, USA.

Yang Zhou, Amnay Amimeur, Chao Jiang, Dejing Dou, Ruoming Jin, and Pengwei Wang. 2018a. Density-aware local siamese autoencoder network embedding with autoencoder graph clustering. In *Proceedings of the 2018 IEEE International Conference on Big Data (BigData'18)*, pages 1162–1167, Seattle, WA.

Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. 2009. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment (PVLDB)*, 2(1):718–729.

Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. 2010. Clustering large attributed graphs: An efficient incremental approach. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM'10)*, pages 689–698, Sydney, Australia.

Yang Zhou, Chao Jiang, Zijie Zhang, Dejing Dou, Ruoming Jin, and Pengwei Wang. 2019b. Integrating local vertex/edge embedding via deep matrix fusion and siamese multi-label classification. In *Proceedings of the 2019 IEEE International Conference on Big Data (BigData'19)*, pages 1018–1027, Los Angeles, CA.

Yang Zhou, Kisung Lee Ling Liu, Qi Zhang, and Balaji Palanisamy. 2019c. Enhancing collaborative filtering with multi-label classification. In *Proceedings of the 2019 International Conference on Computational Data and Social Networks (CSoNet'19)*, pages 323–338, Ho Chi Minh City, Vietnam.

Yang Zhou and Ling Liu. 2012. Clustering analysis in large graphs with rich attributes. In Dawn E. Holmes and Lakhmi C. Jain, editors, *Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification*. Springer.

Yang Zhou and Ling Liu. 2013. Social influence based clustering of heterogeneous information networks. In *Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'13)*, pages 338–346, Chicago, IL.

Yang Zhou and Ling Liu. 2014. Activity-edge centric multi-label classification for mining heterogeneous information networks. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'14)*, pages 1276–1285, New York, NY.

Yang Zhou and Ling Liu. 2015. Social influence based clustering and optimization over heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):1–53.

Yang Zhou and Ling Liu. 2019. Approximate deep network embedding for mining large-scale graphs. In *Proceedings of the 2019 IEEE International Conference on Cognitive Machine Intelligence (CogMI'19)*, pages 53–60, Los Angeles, CA.

Yang Zhou, Ling Liu, and David Buttler. 2015a. Integrating vertex-centric clustering with edge-centric clustering for meta path graph analysis. In *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'15)*, pages 1563–1572, Sydney, Australia.

Yang Zhou, Ling Liu, Kisung Lee, Balaji Palanisamy, and Qi Zhang. 2020a. Improving collaborative filtering with social influence over heterogeneous information networks. *ACM Transactions on Internet Technology (TOIT)*, 20(4):36:1–36:29.

Yang Zhou, Ling Liu, Kisung Lee, Calton Pu, and Qi Zhang. 2015b. Fast iterative graph computation with resource aware graph parallel abstractions. In *Proceedings of the 24th ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC'15)*, pages 179–190, Portland, OR.

Yang Zhou, Ling Liu, Kisung Lee, and Qi Zhang. 2015c. Graphtwist: Fast iterative graph computation with two-tier optimizations. *Proceedings of the VLDB Endowment (PVLDB)*, 8(11):1262–1273.

Yang Zhou, Ling Liu, Chang-Shing Perng, Anca Sailer, Ignacio Silva-Lepe, and Zhiyuan Su. 2013. Ranking services by service network structure and service attributes. In *Proceedings of the 20th International Conference on Web Service (ICWS'13)*, pages 26–33, Santa Clara, CA.

Yang Zhou, Ling Liu, Calton Pu, Xianqiang Bao, Kisung Lee, Balaji Palanisamy, Emre Yigitoglu, and Qi Zhang. 2015d. Clustering service networks with entity, attribute and link heterogeneity. In *Proceedings of the 22nd International Conference on Web Service (ICWS'15)*, pages 257–264, New York, NY.

Yang Zhou, Ling Liu, Sangeetha Seshadri, and Lawrence Chiu. 2016. Analyzing enterprise storage workloads with graph modeling and clustering.

IEEE Journal on Selected Areas in Communications (JSAC), 34(3):551–574.

Yang Zhou, Jiaxiang Ren, Dejing Dou, Ruoming Jin, Jingyi Zheng, and Kisung Lee. 2020b. Robust meta network embedding against adversarial attacks. In *Proceedings of the 20th IEEE International Conference on Data Mining (ICDM'20)*, pages 1448–1453, Sorrento, Italy.

Yang Zhou, Jiaxiang Ren, Ruoming Jin, Zijie Zhang, Dejing Dou, and Da Yan. 2020c. Unsupervised multiple network alignment with multinominal gan and variational inference. In *Proceedings of the 2020 IEEE International Conference on Big Data (BigData'20)*, pages 868–877, Atlanta, GA.

Yang Zhou, Jiaxiang Ren, Ruoming Jin, Zijie Zhang, Jingyi Zheng, Zhe Jiang, Da Yan, and Dejing Dou. 2021a. Unsupervised adversarial network alignment with reinforcement learning. *To appear in ACM Transactions on Knowledge Discovery from Data (TKDD)*.

Yang Zhou, Jiaxiang Ren, Sixing Wu, Dejing Dou, Ruoming Jin, Zijie Zhang, and Pengwei Wang. 2019d. Semi-supervised classification-based local vertex ranking via dual generative adversarial nets. In *Proceedings of the 2019 IEEE International Conference on Big Data (BigData'19)*, pages 1267–1273, Los Angeles, CA.

Yang Zhou, Sangeetha Seshadri, Lawrence Chiu, and Ling Liu. 2014. Graphlens: Mining enterprise storage workloads using graph analytics. In *Proceedings of the 2014 IEEE International Congress on Big Data (BigData'14)*, pages 1–8, Anchorage, AK.

Yang Zhou, Sixing Wu, Chao Jiang, Zijie Zhang, Dejing Dou, Ruoming Jin, and Pengwei Wang. 2018b. Density-adaptive local edge representation learning with generative adversarial network multi-label edge classification. In *Proceedings of the 18th IEEE International Conference on Data Mining (ICDM'18)*, pages 1464–1469, Singapore.

Yang Zhou, Zeru Zhang, Sixing Wu, Victor Sheng, Xiaoying Han, Zijie Zhang, and Ruoming Jin. 2021b. Robust network alignment via attack signal scaling and adversarial perturbation elimination. In *Proceedings of the 30th Web Conference (WWW'21)*, pages 3884–3895, Virtual Event / Ljubljana, Slovenia.

Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative entity alignment via joint knowledge embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4258–4264.

Qi Zhu, Hao Wei, Bunyamin Sisman, Da Zheng, Christos Faloutsos, Xin Luna Dong, and Jiawei Han. 2020. Collective multi-type entity alignment between knowledge graphs. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2241–2252.

Qiannan Zhu, Xiaofei Zhou, Jia Wu, Jianlong Tan, and Li Guo. 2019. Neighborhood-aware attentional representation for multilingual knowledge graphs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1943–1949.

Yao Zhu, Hongzhi Liu, Zhonghai Wu, and Yingpeng Du. 2021. Relation-aware neighborhood matching model for entity alignment. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. 2018. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 2847–2856.

Daniel Zügner, Oliver Borchert, Amir Akbarnejad, and Stephan Günnemann. 2020. Adversarial attacks on graph neural networks: Perturbations and their patterns. *ACM Trans. Knowl. Discov. Data*, 14(5):57:1–57:31.

Daniel Zügner and Stephan Günnemann. 2019. Adversarial attacks on graph neural networks via meta learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Appendix

A Theoretical Analysis

Theorem 1. *Let entity embedding vectors* $\tilde{\mathbf{e}}_k^2$ *and* $\tilde{\mathbf{e}}_l^2$ *be the most similar and least similar to* $(\tilde{\mathbf{e}}_i^1)^T$ *(*$1 \leq k, l \leq N^2$*), i.e.,* $\tilde{\mathbf{e}}_k^2 = \text{argmax}_{\tilde{\mathbf{e}}_k^2}(\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_k^2$ *and* $\tilde{\mathbf{e}}_l^2 = \text{argmin}_{\tilde{\mathbf{e}}_l^2}(\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_l^2$*, and* $c = (\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_k^2$*. Also, suppose that* d *is the minimal norm of entity embedding vectors in* G^2*, i.e.,* $d = \min_{\tilde{\mathbf{e}}_m^2} \|\tilde{\mathbf{e}}_m^2\|_2$ *for* $\forall e_m^2 \in E^2$*. For a given* $0 < \eta < d/2$*, if* $\alpha < \sqrt{\frac{1}{c}\log\frac{d-\eta}{\eta}}$*, then* $1-\sigma((\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_j^2) > \eta/\|\tilde{\mathbf{e}}_j^2\|_2$ *for* $\forall e_j^2 \in E^2$*.*

Proof. $1-\sigma((\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_j^2) > \eta/\|\tilde{\mathbf{e}}_j^2\|_2$ *is equivalent to* $\sigma((\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_j^2) < 1 - \eta/\|\tilde{\mathbf{e}}_j^2\|_2$*. We convert it to* $\frac{1}{1+\exp\left(-(\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_j^2\right)} < 1 - \eta/\|\tilde{\mathbf{e}}_j^2\|_2$*. As* $(\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_j^2 \leq c$*, we have* $\frac{1}{1+\exp\left(-\alpha^2(\mathbf{e}_i^1)^T \cdot \mathbf{e}_j^2\right)} \leq \frac{1}{1+\exp(-\alpha^2 c)}$*. If we can prove* $\frac{1}{1+\exp(-\alpha^2 c)} < 1 - \eta/\|\tilde{\mathbf{e}}_j^2$*, then we can testify* $\frac{1}{1+\exp\left(-\alpha^2(\mathbf{e}_i^1)^T \cdot \mathbf{e}_j^2\right)} < 1 - \eta/\|\tilde{\mathbf{e}}_j^2$*. Thus, we need to solve* $\exp\left(\alpha^2 c\right) < \frac{\|\tilde{\mathbf{e}}_j^2\|_2 - \eta}{\eta}$*.*

As $\|\tilde{\mathbf{e}}_j^2\|_2 \geq d$*, feasible* α *for* $\exp\left(\alpha^2 c\right) < \frac{d-\eta}{\eta}$ *is also feasible for* $\exp\left(\alpha^2 c\right) < \frac{\|\tilde{\mathbf{e}}_j^2\|_2 - \eta}{\eta}$*. Since* $\exp$ *is a monotonic increasing function, by solving the above inequality, we have feasible* $\alpha < \sqrt{\frac{1}{c}\log\frac{d-\eta}{\eta}}$*.*

Notice that $0 < \eta < d/2$*. This makes* $\frac{d-\eta}{\eta} > 1$ *and the upper bound of* α *be positive. Therefore, for any* $\alpha < \frac{1}{c}\log\frac{d-\eta}{\eta}$*,* $1 - \sigma((\tilde{\mathbf{e}}_i^1)^T \cdot \tilde{\mathbf{e}}_j^2) > \eta/\|\tilde{\mathbf{e}}_j^2\|_2$ *is satisfied.*