1 EXTENDED APPENDIX

We have a page limit of 2 page of appendix. Hence, as mentioned in the call for papers of Web Conference https://www2023.thewebconf.org/calls/research-tracks/, this is an extended document for further analysis such as detailed proof of our theoretical results. This document is solely for the completion of full proofs. Also, additional empirical results have been provided with detailed hyperparameters table to assist the reviewing process and clarify doubts(if any). As mentioned in the call for papers, such extended appendix will not be part of the proceedings. Hence, after acceptance of our work, we will release this document on public Github along with the code.

1.1 Theoretical Motivation

Let us consider that there are two distributions: One for the positive graph's edge weights(scores) and the other for the negative graph. With this intuition in mind we now define a proxy metric for the ranking metrics. This is needed as the ranking metrics being discrete we would not be able to define integrals in the continuous space of the domain. The proxy should have a one to one correspondence with the ranking metric(i.e. it should be monotone increasing with ranking metric) along with being continuous. As a convention we perform ranking in ascending order i.e. the positive triples' score should be lesser than the negative triples' score for ranking. Note that is not a limiting assumption since we could always invert the scores in the opposite case. Given a probability distribution of the scores(positive and negative), we define the expected ranking metric for a positive triple with score $x = a$ as the integral of the negative distribution starting from $x = a$ till $x = \infty$ i.e. it is the area under the curve of the negative probability distribution from a till ∞. Intuitively one could imagine that in the ideal scenario the two distributions would have no overlap with each other and the positive triples could be easily distinguished from the negative ones. Formally,

Definition 1.1 (Expected Ranking(ER)). Consider the positive triples to have the distribution D^+ and the negative triples to have the distribution D^-. For a positive triple with score a its expected ranking(ER) is defined as, $ER(a) = \int_{x=a}^{x=\infty} D^-(x)dx$

the proxy to the overall ranking metric is then defined as the expectation of the above expected rank, under the distribution D^+. Since this is supposed to be the proxy of the ranking metric in expectation, we term it as PERM(Proxy of the Expected Ranking Metric). Formally we can define PERM as below,

Definition 1.2 (PERM). Consider the positive triples to have the distribution D^+ and the negative triples to have the distribution D^-. The PERM metric is then defined as, $PERM = \int_{x=-\infty}^{x=\infty} D^+(x)ER(x)dx$

It is easy to see that PERM has a monotone increasing correspondence with the actual ranking metrics. That is, as many of the negative triples get a higher score than the positive triples the distribution of the negative triples will shift further right of the positive distribution and the area under the curve would increase for a given triple(x=a). The intuition of the proxy metric is visualized in the below diagram.

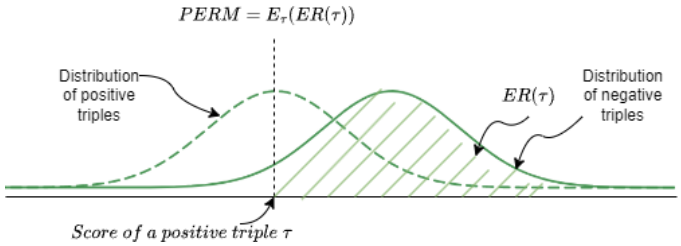

Figure 1: Figure gives an intuition of the metric PERM which is designed to be a proxy to the ranking metrics for ease of theoretical analysis. For a given positive triple τ with score x_τ the expected rank($ER(\tau)$) is defined as the area under the curve of the negative distribution from x_τ to ∞(shown in the shaded area above). PERM is then defined as the expectation of the expected rank under the positive distribution.

Now that we have established a monotone increasing correspondence of PERM with the ranking metrics, we are only left to show that there exists a one-one correspondence between PERM and $\mathcal{KP}$. We make the following assumptions: i) The scores of the KGE method follows a normal distribution, each for the positive(μ) and negative(ν) scores with means m_μ and m_ν respectively (ii) As the KGE method converges the sufficient statistic of the scores of the positive triples consistently lies on one side of the half plane formed by the sufficient statistic of the negative triples, irrespective of the data distribution.

Lemma 1.1. *$\mathcal{KP}$ has a monotone increasing correspondence with the Proxy of the Expected Ranking Metrics(PERM) under the above stated assumptions and as m_ν deviates from m_μ*

Proof. Considering the 0-dimensional PD as used by $\mathcal{KP}$ and a normal distribution for the edge weights(can be extended to other distributions using techniques like [7]) of the graph(scores of the triples), we have univariate gaussian measures μ and ν for the positive and

negative distributions respectively. Denote by m_μ and m_ν the means of the distributions μ and ν respectively and by Σ_μ, Σ_ν the respective covariance matrices. Thus we have a closed form solution to the 2-wasserstein distance between the measures μ and ν:

$$dW_2^2(\mu,\nu) = \|\mu-\nu\|^2 + B(\Sigma_\mu,\Sigma_\nu)^2 \tag{1}$$

where $B(\Sigma_\mu,\Sigma_\nu)^2 = tr(\Sigma_\mu+\Sigma_\nu-2(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}})^{\frac{1}{2}})$. Note that while $\mathcal{KP}$ uses the 0-dimensional PD for computational reasons, however the above equation holds for the higher dimensional PDs(after appropriate normalizations) using sliced wasserstein distance if the distribution along the slices are gaussian and we leave this analysis for future works.

Now consider that as the learning progresses the generating process of the scores change such that the means of the two distributions are separated out, variance remaining the same. Even if the variance changes it changes similarly for both distributions, with the initial variance being equal. We now find how each of PERM and the wasserstein distance($\mathcal{KP}$) changes with change in the distributions.

PERM: We first find the change in PERM with change in the means. Without loss of generality we assume m_μ is fixed and vary m_ν. Note since we have univariate measures we write the variances as $\Sigma_\mu = \sigma_\mu^2$ and $\Sigma_\nu = \sigma_\nu^2$. Writing the expression for PERM(P) and taking derivatives with respect to the mean we have

$$\begin{aligned}
P &= \int_{x=-\infty}^{x=\infty} D^+(x)\Big(\int_{y=x}^{y=\infty} D^-(x)dy\Big)dx \\
\frac{\partial P}{\partial m_\nu} &= \int_{x=-\infty}^{x=\infty} D^+(x)\left(\int_{y=x}^{y=\infty}\frac{\partial D^-(x)}{\partial m_\nu}dy\right)dx \\
&= \int_{x=-\infty}^{x=\infty} D^+(x)\left(\int_{y=x}^{y=\infty}\frac{y-m_\nu}{\sigma_{nu}^2}D^-(y)dy\right)dx
\end{aligned}$$

Note in the above expression $\int_{y=a}^{y=\infty}\frac{y-m_\nu}{\sigma_{nu}^2}D^-(y)dy > 0$ if $a > m_\nu$ and if $a < m_\nu$, say $m_\nu - d = a$ then

$$\begin{aligned}
&\int_{y=a}^{y=\infty}\frac{y-m_\nu}{\sigma_{nu}^2}D^-(y)dy \\
&= \int_{y=m_\nu-d}^{y=m_\nu+d}\frac{y-m_\nu}{\sigma_{nu}^2}D^-(y)dy + \int_{y=m_\nu+d}^{y=\infty}\frac{y-m_\nu}{\sigma_{nu}^2}D^-(y)dy \\
&= 0 + \int_{y=m_\nu+d}^{y=\infty}\frac{y-m_\nu}{\sigma_{nu}^2}D^-(y)dy \\
&= \int_{y=m_\nu+d}^{y=\infty}\frac{y-m_\nu}{\sigma_{nu}^2}D^-(y)dy \\
&> 0
\end{aligned}$$

Thus we have $\frac{\partial P}{\partial m_\nu} > 0$. Taking the derivative with respect to the variance Σ_ν, we get

$$\begin{aligned}
P &= \int_{x=-\infty}^{x=\infty} D^+(x)\left(\int_{y=a}^{y=\infty} D^-(y)dy\right)dx \\
\frac{\partial P}{\partial \Sigma_\nu} &= \int_{x=-\infty}^{x=\infty} D^+(x)\left(\int_{y=x}^{y=\infty}\frac{\partial D^-(y)}{\partial \Sigma_\nu}dy\right)dx \\
&= \int_{x=-\infty}^{x=\infty} D^+(x)\left(\int_{y=x}^{y=\infty}\frac{(y-m_\nu)^2-\sigma^2}{\sigma_{nu}^4}D^-(y)dy\right)dx \\
&= \frac{1}{\sigma^4}\int_{x=-\infty}^{x=\infty} D^+(x)\left(\int_{y=x}^{y=\infty}(y-m_\nu)^2D^-dy - \sigma^2\int_{y=a}^{y=\infty}D^-(y)dy\right)dx \\
&= \frac{1}{\sigma^4}\int_{x=-\infty}^{x=\infty} D^+(x)\left(\int_{y=x}^{y=\infty}(y-m_\nu)^2D^-dy - \sigma^2\right)dx \\
&= \frac{1}{\sigma^4}\int_{x=-\infty}^{x=\infty} D^+(x)\left(-\int_{y=-\infty}^{y=x}(y-m_\nu)^2D^-dy\right)dx \\
&< 0
\end{aligned}$$

$\therefore \frac{\partial P}{\partial \Sigma_\mu} > 0, \frac{\partial P}{\partial \Sigma_\nu} > 0$ Similarly taking the derivative with respect to the variance Σ_μ, we get

$$
\begin{aligned}
P &= \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=x}^{y=\infty} D^-(y) dy \right) dx \\
\frac{\partial P}{\partial \Sigma_\mu} &= \int_{x=-\infty}^{x=\infty} \frac{\partial D^+(x)}{\partial \Sigma_\nu} \left(\int_{y=x}^{y=\infty} D^-(y) dy \right) dx \\
&= \int_{x=-\infty}^{x=\infty} \frac{(x-\mu)^2 - \sigma_\mu^2}{\sigma_\mu^4} D^+(x) ER(x) dx \\
&= \int_{x=-\infty}^{x=\infty} \frac{(x-\mu)^2}{\sigma_\mu^4} D^+(x) ER(x) dx + \frac{PERM}{\sigma_\mu^2}
\end{aligned}
$$

$\mathcal{KP}$: We now find the change in $\mathcal{KP}$ with change in the means. Writing the expression for $\mathcal{KP}$ and taking derivatives with respect to the mean(m_ν) we have

$$
\begin{aligned}
\frac{\partial W_2^2(\mu, \nu)}{\partial m_\nu} &= 2|m_\mu - m_\nu| \\
&> 0
\end{aligned}
$$

Taking the derivative with respect to Σ_ν we get

$$
\begin{aligned}
\frac{\partial W_2^2(\mu, \nu)}{\partial \Sigma_\nu} &= I - \Sigma_\mu^{\frac{1}{2}} (\Sigma_\mu^{\frac{1}{2}} \Sigma_\nu \Sigma_\mu^{\frac{1}{2}})^{\frac{-1}{2}} \Sigma_\mu^{\frac{1}{2}} \\
&= 1 - \frac{\sigma_\mu}{\sigma_\nu} \\
&= \frac{\sigma_\nu - \sigma_\mu}{\sigma_\nu} \\
&\begin{cases} \geq 0, & \text{if } \sigma_\nu \geq \sigma_\mu \\ < 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

Similarly taking the derivative with respect to Σ_μ we get

$$
\begin{aligned}
\frac{\partial W_2^2(\mu, \nu)}{\partial \Sigma_\nu} &= 1 - \frac{\sigma_\nu}{\sigma_\mu} \\
&= \frac{\sigma_\mu - \sigma_\nu}{\sigma_\mu} \\
&\begin{cases} \geq 0, & \text{if } \sigma_\mu \geq \sigma_\nu \\ < 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

Now under the assumptions above consider a change dm_ν in the mean of the negative scores and $d\sigma$ in the standard deviations of both the distributions. Also the variances of the initial distributions are same $\Sigma_\mu = \Sigma_\nu$. This would be a fair assumption say at the start of the

traiing process. As the generating process of the scores changes the the gradient of PERM along the direction $(dm_\nu, d\sigma_\mu, d\sigma_\nu)$ is

$$
\begin{aligned}
&\left\langle (dm_\nu, d\sigma, d\sigma), \left(\frac{\partial PERM}{\partial m_\nu}, \frac{\partial PERM}{\partial \Sigma_\mu}, \frac{\partial PERM}{\partial \Sigma_\nu} \right) \right\rangle \\
&= \frac{\partial PERM}{\partial m_\nu} dm_\nu + \left(\frac{\partial PERM}{\partial \Sigma_\mu} + \frac{\partial PERM}{\partial \Sigma_\nu} \right) d\sigma \\
&= \frac{\partial PERM}{\partial m_\nu} dm_\nu + \\
&\left(\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) - \frac{PERM}{\sigma_\mu^2} \right) d\sigma + \\
&\int_{x=-\infty}^{x=\infty} D^+ \left(- \int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y) \right) d\sigma \\
&\approx \frac{\partial PERM}{\partial m_\nu} (1 - d\sigma^2) + \\
&\left(\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) - \frac{PERM}{\sigma_\mu^2} \right) d\sigma \\
&+ \int_{x=-\infty}^{x=\infty} D^+(x) \left(- \int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y) \right) d\sigma \\
&\geq \frac{\partial PERM}{\partial m_\nu} \left(\frac{1}{2} \right) + \left(\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) - \frac{PERM}{\sigma_\mu^2} \right) d\sigma \\
&- \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y) \right) d\sigma \\
&= \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=x}^{y=\infty} \frac{y - m_\nu}{2\sigma_{nu}^2} D^-(y) dy \right) dx - \frac{PERM}{\sigma_\mu^2} d\sigma + \\
&\quad \left(\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) \right) d\sigma \\
&\quad - \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y) \right) d\sigma \\
&= \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=x}^{y=\infty} \frac{y - m_\nu}{2\sigma_{nu}^2} D^-(y) dy \right) dx - \\
&\quad \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=x}^{y=\infty} \frac{d\sigma}{\sigma_{mu}^2} D^-(y) dy \right) dx + \\
&\quad \left(\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) \right) d\sigma \\
&\quad - \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y) \right) d\sigma \\
&= \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=x}^{y=\infty} \left(\frac{y - m_\nu}{2\sigma^2} - \frac{d\sigma}{\sigma^2} \right) D^-(y) dy \right) dx + \\
&\quad \left(\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) \right) d\sigma \\
&\quad - \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y) \right) d\sigma
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=x}^{y=\infty} (\frac{y - m_\nu - 2d\sigma}{2\sigma^2}) D^-(y) dy \right) dx + \\
&\quad \left(\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) \right) d\sigma \\
&\quad - \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y) \right) d\sigma \\
&\approx \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=x}^{y=\infty} (\frac{y - m_\nu}{2\sigma^2}) D^-(y) dy \right) dx + \\
&\quad \left(\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) \right) d\sigma \\
&\quad - \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y) \right) d\sigma \\
&= \frac{1}{2} \frac{\partial PERM}{\partial m_\nu} + \left(\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) \right) d\sigma \\
&\quad - \int_{x=-\infty}^{x=\infty} D^+(x) \left(\int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y) \right) d\sigma
\end{aligned}
$$

The approximation in the 4th equation above is due to $dm^2 + 2d\sigma^2 = 1$ and by using the binomial expansion eliminating higher order terms. Also, as $2d\sigma^2 < 1$ the next equation follows.The final approximation is as we can arbitrarily have $\mu_2 > d\sigma (< 1)$, for example by adding a constant to the scores. Now we have

$$
\begin{aligned}
&\int_{x=-\infty}^{x=\infty} \frac{x - m_\mu^2}{\sigma_\mu^4} D^+(x) \int_{y=x}^{\infty} D^-(y) \geq \\
&\int_{x=-\infty}^{x=\infty} \frac{D^+}{\sigma_4} (- \int_{y=x}^{\infty} (y - m_\nu)^2 D^-(y)) \\
&\geq \int_{x=-\infty}^{x=\infty} D^+ (- \int_{y=-\infty}^{x} \frac{(y - m_\nu)^2}{\sigma_2^4} D^-(y))
\end{aligned}
$$

Also we have shown that $\frac{\partial PERM}{\partial m_\nu} \geq 0$. Thus we have

$$
\begin{aligned}
&\left\langle (dm_\nu, d\sigma, d\sigma), (\frac{\partial PERM}{\partial m_\nu}, \frac{\partial PERM}{\partial \Sigma_\mu}, \frac{\partial PERM}{\partial \Sigma_\nu}) \right\rangle \\
&\geq 0
\end{aligned}
$$

Similarly the gradient of $\mathcal{KP}$ along the direction $(dm_\nu, d\sigma_\mu, d\sigma_\nu)$ is

$$
\begin{aligned}
&\left\langle (dm_\nu, d\sigma, d\sigma), (\frac{\partial W_2^2(\mu,\nu)}{\partial m_\nu}, \frac{\partial W_2^2(\mu,\nu)}{\partial \Sigma_\mu}, \frac{\partial W_2^2(\mu,\nu)}{\partial \Sigma_\nu}) \right\rangle \\
&= \frac{\partial W_2^2(\mu,\nu)}{\partial m_\nu} dm_\nu + \frac{\sigma_\nu - \sigma_\mu}{\sigma_\nu} d\sigma + \frac{\sigma_\mu - \sigma_{nu}}{\sigma_\mu} d\sigma \\
&= \frac{\partial W_2^2(\mu,\nu)}{\partial m_\nu} dm_\nu + (\frac{\sigma_\nu - \sigma_\mu}{\sigma_\nu} + \frac{\sigma_\mu - \sigma_{nu}}{\sigma_\mu}) d\sigma \\
&\approx \frac{\partial W_2^2(\mu,\nu)}{\partial m_\nu} dm_\nu + (\frac{\sigma_\nu - \sigma_\mu + \sigma_\mu - \sigma_{nu}}{\sigma_\nu}) d\sigma \\
&> 0
\end{aligned}
$$

Since both PERM and and $\mathcal{KP}$ vary in the same manner as the distribution changes, the two have a one-one correspondence [8]. □

The above lemma shows that there is a one-one correspondence between $\mathcal{KP}$ and PERM and by definition PERM has a one-one correspondence with the ranking metrics. Therefore, the next lemma follows as a natural consequence

Theorem 1.3. *$\mathcal{KP}$ has a one-one correspondence with the Ranking Metrics under the above stated considerations.*

The above theorem states that, with high probability, there exists a correlation between $\mathcal{KP}$ and the ranking metrics under the stated assumption of having a normal distribution. Note that we do not require this assumption but use it for theoretical convenience as it gives a closed-form solution to the equations. Even if a distribution is not normal we could convert it to one by using for example the Box-Cox transform [7]. The theorem may hold even for non-normal distributions, which we leave for future works to investigate and is out of scope of this work. The other assumption states that the trained KGE method has a consistent separation between the scores of positive and negative triples, i.e., for any dataset, the positive scores must be lesser(or greater) than the negative scores in expectation. Otherwise, the method may not be reliable and may not have good performance. In this sense, the assumptions are mild and are satisfied by many existing KGE methods.

Theorem 1.4. *Considerate to theorem 1.3, the relative change in $\mathcal{KP}$ on addition of random noise to the scores is bounded by a function of the original and noise-induced covariance matrix as $\frac{\Delta\mathcal{KP}}{\mathcal{KP}} \leq max((1-|\Sigma_{\mu_1}^{+1}\Sigma_{\mu_2}^{-1}|^{\frac{3}{2}}), (1-|\Sigma_{\nu_1}^{+1}\Sigma_{\nu_2}^{-1}|^{\frac{3}{2}}))$, where Σ_{μ_1} and Σ_{ν_1} are the covariance matrices of the positive and negative triples' scores respectively and Σ_{μ_2} and Σ_{ν_2} are that of the corrupted scores.*

Proof. Consider a zero mean random noise to simulate the process of varying the distribution of the scores of the KGE method. Let m_{μ_1} and m_{ν_1} be the means of the positive and negative triples' scores of the original method and $\Sigma_{\mu_1}, \Sigma_{\nu_1}$ be the respective covariance matrices. Let m_{μ_2} and m_{ν_2} be the means of the positive and negative triples' scores of the corrupted method and $\Sigma_{\mu_2}, \Sigma_{\nu_2}$ be the respective covariance matrices. Since the corruption process is random the means do not change and $m_{\mu_1} = m_{\mu_2} = m_{\mu}$ and $m_{\nu_1} = m_{\nu_2} = m_{\nu}$. Denote by $\mathcal{KP}_1, \mathcal{KP}_2$ the knowledge persistence scores for the respective methods. Then by definition of $\mathcal{KP}$ we have

$$\mathcal{KP}_1 = \inf_{\gamma_1 \in \Pi(x,y)} \left(\int_{\gamma_1} D^p(x,y) d\gamma_1(x,y) \right)^{\frac{1}{p}}$$
$$\mathcal{KP}_2 = \inf_{\gamma_2 \in \Pi(x,y)} \left(\int_{\gamma_2} D^p(x,y) d\gamma_2(x,y) \right)^{\frac{1}{p}}$$

Here $x, y \in R^n$ refers to the PD with n=2, $\Pi(x,y) \in R^{n \times n}$ is the transportation plan(or the set of coupling matrices), $D(x,y)$ refers to the distance(transportation cost) between points x and y. Note that by writing the kantorovich dual [12] of the above we get

$$\mathcal{KP}_1 = \sup_{\substack{\Phi(x),\Psi(y) \\ \Phi(x)+\Psi(y) \leq c(x,y)}} \int_x \Phi(x) d\mu_1(x) + \int_y \Psi(y) d\nu_1(y)$$
$$\mathcal{KP}_2 = \sup_{\substack{\Phi(x),\Psi(y) \\ \Phi(x)+\Psi(y) \leq c(x,y)}} \int_x \Phi(x) d\mu_2(x) + \int_y \Psi(y) d\nu_2(y)$$

Taking the difference between the two measures we have

$$
\begin{aligned}
&\mathcal{KP}_1 - \mathcal{KP}_2 \\
&= \inf_{\gamma_1 \in \Pi(x,y)} \int_{\gamma_1} Distance(x,y) d\gamma_1(x,y) \\
&- \inf_{\gamma_2 \in \Pi(x,y)} \int_{\gamma_2} Distance(x,y) d\gamma_2(x,y) \\
&= \sup_{\Phi,\Psi} \int_x \Phi(x) d\mu_1(x) + \int_y \Psi(y) d\nu_1(y) \\
&- (\sup_{\Phi,\Psi} \int_x \Phi(x) d\mu_2(x) + \int_y \Psi(y) d\nu_2(y)) \\
&= \sup_{\Phi,\Psi} \int_x \Phi(x) d\mu_1(x) + \int_y \Psi(y) d\nu_1(y) \\
&+ \inf_{\Phi,\Psi} -1 \times (\int_x \Phi(x) d\mu_2(x) + \int_y \Psi(y) d\nu_2(y)) \\
&\leq \sup_{\Phi,\Psi} \int_x \Phi(x) d\mu_1(x) + \int_y \Psi(y) d\nu_1(y) \\
&- \int_x \Phi(x) d\mu_2(x) - \int_y \Psi(y) d\nu_2(y) \\
&\leq \sup_{\Phi,\Psi} \int_x \Phi(x)(d\mu_1(x) - d\mu_2(x)) + \int_y \Psi(y)(d\nu_1(y) - d\nu_2(y))
\end{aligned}
$$

Now by definition of the measure μ_1 we have

$$
\begin{aligned}
\frac{\partial \mu_1}{\partial x} &= -\mu_1 \Sigma_{\mu_1}^{-1}(x - m_{\mu_1}) \\
d\mu_1(x_i) &= -(\mu_1 \Sigma_{\mu_1}^{-1}(x - m_{\mu_1}))[i] dx_i \\
\therefore d\mu_1(x) &= det(diag(-\mu_1 \Sigma_{\mu_1}^{-1}(x - m_{\mu_1}))) dx
\end{aligned}
$$

Similarly $d\mu_2(x) = det(diag(-\mu_2 \Sigma_{\mu_2}^{-1}(x - m_{\mu_2})))dx$. Thus we have,

$$
\begin{aligned}
&\int_x \Phi(x)(d\mu_1(x) - d\mu_2(x)) \\
&= \int_x \Phi(x)(det(diag(-\mu_1 \Sigma_{\mu_1}^{-1}(x - m_{\mu_1}))) \\
&\quad - det(diag(-\mu_2 \Sigma_{\mu_2}^{-1}(x - m_{\mu_2})))) dx \\
&= \int_x \Phi(x)(1 - \frac{det(diag(\mu_2 \Sigma_{\mu_2}^{-1}(x - m_{\mu_2})))}{det(diag(-\mu_1 \Sigma_{\mu_1}^{-1}(x - m_{\mu_1})))}) \\
&\quad det(diag(\mu_1 \Sigma_{\mu_1}^{-1}(x - m_{\mu_1}))) dx \\
&= \int_x \Phi(x)(1 - \frac{det(diag(\mu_2 \Sigma_{\mu_2}^{-1}(x - m_{\mu_2})))}{det(diag(\mu_1 \Sigma_{\mu_1}^{-1}(x - m_{\mu_1})))}) \\
&\quad det(diag(\mu_1 \Sigma_{\mu_1}^{-1}(x - m_{\mu_1}))) dx \\
&= \int_x \Phi(x)(1 - \frac{\mu_2}{\mu_1}^n \frac{det(diag(\Sigma_{\mu_2}^{-1}(x - m_{\mu_2})))}{det(diag(\Sigma_{\mu_1}^{-1}(x - m_{\mu_1})))}) d\mu_1
\end{aligned}
$$

Now we know that

$$\frac{\mu_2}{\mu_1} = \sqrt{\Sigma_{\mu_2}^{-1}\Sigma_{\mu_1}}\exp^{-(x-\mu_1)^T(\Sigma_{\mu_2}^{-1}-\Sigma_{\mu_1}^{-1})(x-\mu)}$$
$$\geq \sqrt{\Sigma_{\mu_2}^{-1}\Sigma_{\mu_1}}$$

The last equation above is as after the addition of noise the variance would increase and $\sigma_{\mu_2}^2 > \sigma_{\mu_1}^2 \rightarrow \Sigma_{\mu_2} > \Sigma_{\mu_1} \rightarrow \Sigma_{\mu_2}^{-1} < \Sigma_{\mu_1}^{-1}$ for a PSD covariance (which is true in our univariate case) and so the exponent would be > 0. Thus from the above equations we have

$$\int_x \Phi(x)(d\mu_1(x) - d\mu_2(x))$$
$$\leq \int_x \Phi(x)(1 - \sqrt{\Sigma_{\mu_2}^{-1}\Sigma_{\mu_1}}^n \frac{det(diag(\Sigma_{\mu_2}^{-1}(x - m_{\mu_2})))}{det(diag(\Sigma_{\mu_1}^{-1}(x - m_{\mu_1})))})d\mu_1$$
$$= \int_x \Phi(x)(1 - det(\Sigma_{\mu_1}\Sigma_{\mu_2}^{-1})^{\frac{n}{2}+1})d\mu_1$$

Similarly we can show that $\int_x \Psi(x)(dv_1(y) - dv_2(y)) \leq \int_y \Psi(y)(1 - det(\Sigma_{v_1}\Sigma_{v_2}^{-1})^{\frac{n}{2}+1})dv_1$. Thus we can infer that

$$\mathcal{KP}_1 - \mathcal{KP}_2$$
$$\leq \sup_{\Phi,\Psi} \int_x \Phi(x)(d\mu_1(x) - d\mu_2(x)) + \int_y \Psi(y)(dv_1(y) - dv_2(y))$$
$$\leq max((1 - det(\Sigma_{\mu_1}\Sigma_{\mu_2}^{-1})^{\frac{n}{2}+1}), (1 - det(\Sigma_{v_1}\Sigma_{v_2}^{-1})^{\frac{n}{2}+1}))\times$$
$$\sup_{\Phi,\Psi} \int_x \Phi(x)d\mu_1 + \int_y \Psi(y)dv_1$$
$$= max((1 - det(\Sigma_{\mu_1}\Sigma_{\mu_2}^{-1})^{\frac{n}{2}+1}), (1 - det(\Sigma_{v_1}\Sigma_{v_2}^{-1})^{\frac{n}{2}+1}))\mathcal{KP}_1$$
$$\therefore \frac{\Delta\mathcal{KP}}{\mathcal{KP}} \leq max\left(\left(1 - det(\Sigma_{\mu_1}\Sigma_{\mu_2}^{-1})^{\frac{n}{2}+1}\right), \left(1 - det(\Sigma_{v_1}\Sigma_{v_2}^{-1})^{\frac{n}{2}+1}\right)\right)$$

In our case as we work in the univariate setting $n = 1$ and thus we have $\frac{\Delta\mathcal{KP}}{\mathcal{KP}} \leq max\left(\left(1 - det(\Sigma_{\mu_1}\Sigma_{\mu_2}^{-1})^{\frac{3}{2}}\right), \left(1 - det(\Sigma_{v_1}\Sigma_{v_2}^{-1})^{\frac{3}{2}}\right)\right)$, as required. □

Theorem 1.4 gives a bound on the change in $\mathcal{KP}$ while inducing noise in the KGE predictions. In the ideal case, we see the error/change would be 0, and as the noise is increased(and the ranking changed), gradually, the $\mathcal{KP}$ value also changes in a bounded manner as desired.

1.2 Concepts for Persistent Homology

Simplicial Complex: Consider a set V of $(n+1)$ affinely independent points $V = v_0, v_1, \ldots, v_n$. An n dimensional simplex(n-simplex) $\sigma^{(n)}$ is the convex hull of these $n+1$ points. Formally,

$$\sigma^{(n)} = \{v \mid v = \sum_{i=0}^{n} k_i v_i\ ,\ \sum_{i=0}^{n} k_i = 1\ ,\ k_i \geq 0 \forall\ i\}$$

For example, a point is a 0-simplex, an edge is a 1-simplex, a triangle is a 2-simplex etc. A facet of an n-simplex is an $(n-1)$-simplex, having all vertices but one of the original simplex. Thus there are n such facets. For example a triangle (with three vertices) has three edges. Define a boundary operator ∂_n over the n-simplex which gives the union of the n facets i.e. $\bigcup_{i=0}^{n} \sigma_i^{n-1}$.A simplicial complex C has the properties: (i) Any face of a simplex in C is a simplex and (ii) the intersection of two distinct simplices gives common face of both. Finally, C_d is the set of all $d-$dimensional simplices in C and the dimension of a simplicial complex is the maximum dimension of all simplices in it. Thus we have sets of simplices $C_0, C_1, \ldots, C_n$ in an n-dimensional simplicial complex. In our case, graphs containing entities and its relations can be considered as 1-dimensional simplicial complexes with nodes and edges as the simplices.

Homology: Homology theory studies objects by assigning groups $C_0, C_1, C_2 \ldots$ to them [4]. These groups are connected by a homomorphic boundary operator $\partial_k : C_{k+1} \rightarrow C_k$. The map of the boundary operator from C_{k+1} to C_k is called the image space $im(\partial_{k+1})$ and the map of the operator from C_k to the null element of C_{k-1} is called the kernel space $ker(\partial_k)$. The homomorphic boundary operator thus creates an image space which is a subset of the kernel space i.e. $im(\partial_{k+1}) \subseteq ker(\partial_k)$. This kind of a structure is called a chain complex and it induces the property that $H_k = ker(\partial_k)/im(\partial_k)$,

$$H_k = ker(\partial_k)/im(\partial_k)$$

where H_k is the $k-th$ homology group. The rank of H_k, also called Betti numbers [3], can be used to study the homological features of the space. Specifically, $rank(H_0)$ gives the number of connected components, $rank(H_1)$ gives the topological features such as number of tunnels/holes and so on.

Simplicial homology concerns with the application of homology theory to simplicial complexes. Consider C_k to be the vector field generated by the $k-$dimensional simplces over $\mathbb{Z}/2\mathbb{Z}$. Then the $k-th$ boundary operator would map

$$\partial_k : \sum_i \sigma_i \rightarrow \sum_i \sum_{\substack{dim(\tau)=k-1 \\ \tau \subset \sigma_i}} \tau$$

which gives the sum of the boundaries/facets of all the k-simplices in C_k.

Persistent Homology (PH): PH studies the topological features [4] such as components in 0-dimension (e.g., a node), holes in 1-dimension (e.g., a void area bounded by triangle edges) and so on, spread over a scale. Thus, one need not choose a scale beforehand. The number(rank) of these topological features(homology group) in every dimension at a particular scale is given by Betti numbers [3]. However, these Betti numbers are unstable [4] creating a need for Persistent Homology. Consider the simplicial complex C with weights $a_0 \leq a_1 \leq a_2 \ldots a_{m-1}$, which could represent the edge weights, for example, the triple score from the KG embedding method in our case. We can then define a Filtration process, which refers to generating a nested sequence of complexes $\phi \subseteq C_1 \subseteq C_2 \subseteq \ldots C_m = C$ in time/scale as the simplices below the threshold weights are added in the complex. This is thought to represent the evolution of C as the weights/scale are changed in increasing order. The filtration process results in the creation(birth) and destruction(death) of components, holes, etc. Thus each structure is associated with a birth-death pair $(a_i, a_j) \in R^2$ with $i \leq j$. The persistence or lifetime of each component can then be given by $a_j - a_i$. A **persistence diagram (PD)** summarizes the (birth,death) pair of each object on a 2D plot, with birth times on the x axis and death times on the y axis. The points near the diagonal are shortlived components and generally are considered noise (local geometry), whereas the persistent objects (global topology) are treated as features. We consider local and global geometry/topology to compare two PDs (i.e., positive and negative triple graphs). Figure **??** shows the filtration process for a sample graph with the associated persistence diagram. Note here we show the filtration on the edge weights only, however it could be on done on any quantity such as the node values etc.

Distance Computation: In order to compare two persistence diagrams we could use the $p-th$ Wasserstein distance between them as defined below (with $\Omega = R^2$ in this case)

Definition 1.5. Let $p \in [1, \infty)$ and D: $\Omega \times \Omega \rightarrow [0, \infty)$ be the cost of transporting the measure μ to ν, then the $p-th$ Wasserstein distance [11] between the measures is given by

$$W_p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\Omega \times \Omega} D(x,y)^p \partial\gamma \right)^{\frac{1}{p}} \tag{2}$$

where Π is the set of all the possible transport plans with the marginals μ and ν.

Computation of Wasserstein distance is expensive of the order $O(N^3 log(N))$ [11]. Sliced Wasserstein distance is generally an alternative with computational and statistical benefits. It involves finding 1-dimensional projections along with all the directions of S^{d-1} and then computing the integral of the 1-d Wasserstein distance between these projections. The Sliced Wasserstein distance(SW) between measures μ and ν is:

$$SW_p(\mu, \nu) = \left(\int_{S^{d-1}} W_p^p(R_\mu(., \theta), R_\nu(., \theta)) \right)^{\frac{1}{p}}$$

where $R_\mu(., \theta)$ is the projection of μ along θ. In practice a Monte Carlo average over L samples is done in place of the integral. Computing the Sliced Wasserstein distance takes $O(LNd + LNlog(N))$ which can be improved to linear time $O(Nd)$ for SW_2 as it has been shown to have a closed form solution [6]

1.3 Implementation Details

In order to train and test the KG embedding methods we make use of the pykg2vec [14] library. We use the standard/best hyperparameters for these datasets that the previous works have reported [1, 2, 5, 9, 10, 13, 14]. The details of the hyperparameters for each method and dataset combination are given in tables 2, 3, 4, 5. As standard practice, the validation runs are executed 20 times on average. All of the methods are run on a single P100 GPU machine for a maximum of 100 epochs each and evaluated every 5 epochs. The dataset details are in the Table 1.

1.4 Additional Results

Regression plots between $\mathcal{KP}$ and ranking metrics in intra-method setting: Figures 2 and 3 show the linear regression plots(along with the error bands) between $\mathcal{KP}$ and the Hits@10 ranking metric for the translation(TransE, TransR) and bilinear(Complex, RotatE) methods. Here each point represents the $\mathcal{KP}$ value and the corresponding Hits@10 value at the given epoch. Here, evaluation is done every 2 epochs and training proceeds for 100 epochs. We can see in all the plots that $\mathcal{KP}$ behaves as an approximately monotonically increasing

Table 1: Existing Benchmark Datasets.

Dataset	Triples	Entities	Relations
FB15K	592,213	14.951	1,345
FB15K-237	272,115	14,541	237
WN18	151,442	40,943	18
WN18RR	93,003	40,943	11
Yago3-10	1 Million	123,182	37

function of Hits@10. The plot has a good linear resemblance in some methods(eg: translation methods), however in the other methods we see some non-linear behaviour indicating the prediction power of $\mathcal{KP}$ may change with the class of KG embedding methods.

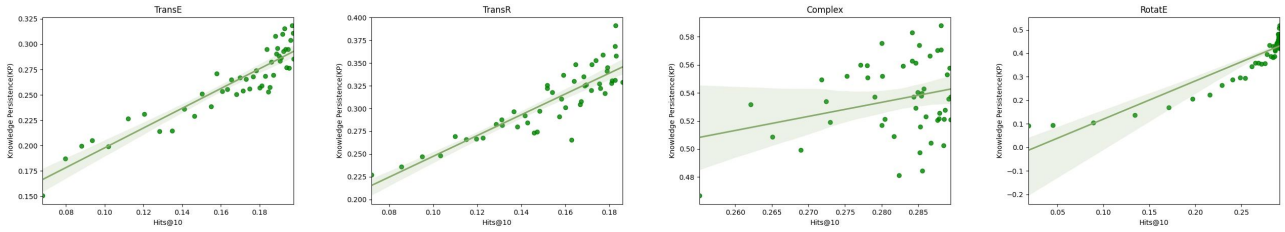

Figure 2: Figure shows the regression plots between Hits@10(on x axis) and $\mathcal{KP}$(on y axis) for the KG embedding methods on the FB15K237 dataset.

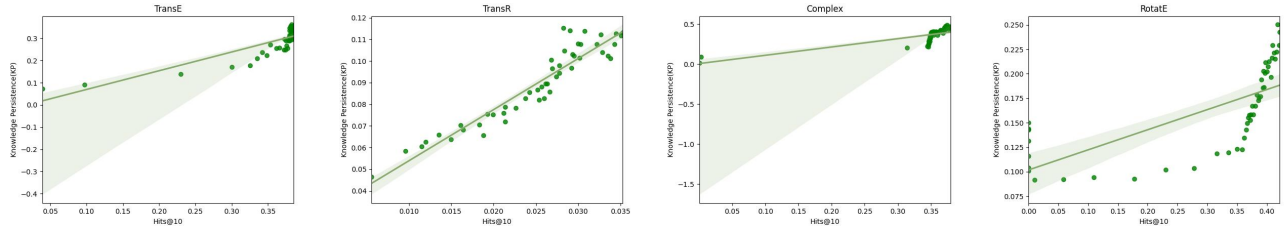

Figure 3: Figure shows the regression plots between Hits@10(on x axis) and $\mathcal{KP}$(on y axis) for the KG embedding methods on the WN18RR dataset.

REFERENCES

[1] Ivana Balažević, Carl Allen, and Timothy Hospedales. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. 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)*. 5185–5194.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS*. 1–9.

[3] Cecil Jose A Delfinado and Herbert Edelsbrunner. 1993. An incremental algorithm for Betti numbers of simplicial complexes. In *Proceedings of the ninth annual symposium on Computational geometry*. 232–239.

[4] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. 2000. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*. IEEE, 454–463.

[5] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[6] Kimia Nadjahi, Alain Durmus, Pierre E Jacob, Roland Badeau, and Umut Simsekli. 2021. Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Random Projections. *Advances in Neural Information Processing Systems* 34 (2021).

Datasets	**FB15K**						
Methods	learning rate	hidden size	batch size	margin	optimizer	sampling	neg rate
TransE	1	50	50	5	adam	adversarial negative sampling	100
TransH	0.1	50	50	0.5	adam	adversarial negative sampling	100
TransR	0.1	50	50	4	adam	adversarial negative sampling	100
Complex	0.1	50	50	-	adam	adversarial negative sampling	100
RotatE	0.1	50	50	9	adam	adversarial negative sampling	100
TuckER	0.1	200	50	-	adam	adversarial negative sampling	100
ConvKB	0.1	50	50	-	adam	adversarial negative sampling	100

Table 2: Table showing the hyperparameters used in the KG embedding methods trained on the FB15K dataset.

Datasets	**FB15K237**						
Methods	learning rate	hidden size	batch size	margin	optimizer	sampling	neg rate
TransE	1	200	2048	5	sgd	bernoulli	25
TransH	0.005	50	1200	0.5	sgd	uniform	1
TransR	0.5	50	2048	4	sgd	bernoulli	1
Complex	0.0695	200	256	-	adagrad	bernoulli	1
RotatE	0.0001	1000	1024	9	adam	adversarial negative sampling	128
TuckER	0.0005	200	128	-	adam	uniform	0
ConvKB	0.0001	100	128	-	adam	adversarial negative sampling	1

Table 3: Table showing the hyperparameters used in the KG embedding methods trained on the FB15K237 dataset.

Datasets	**WN18**						
Methods	learning rate	hidden size	batch size	margin	optimizer	sampling	neg rate
TransE	0.1	50	50	6	adam	adversarial negative sampling	100
TransH	0.1	50	50	0.5	adam	adversarial negative sampling	100
TransR	0.1	50	50	4	adam	adversarial negative sampling	100
Complex	0.5	50	50	-	adam	adversarial negative sampling	100
RotatE	0.1	50	50	6	adam	adversarial negative sampling	100
TuckER	0.1	50	50	-	adam	adversarial negative sampling	100
ConvKB	0.1	50	500	-	adam	adversarial negative sampling	100

Table 4: Table showing the hyperparameters used in the KG embedding methods trained on the WN18 dataset.

Datasets	**WN18RR**						
Methods	learning rate	hidden size	batch size	margin	optimizer	sampling	neg rate
TransE	0.00002	1024	128	6	adam	uniform	64
TransH	0.005	50	1200	0.5	sgd	uniform	1
TransR	0.5	50	2048	4	sgd	bernoulli	1
Complex	0.5	200	2048	-	adagrad	bernoulli	25
RotatE	0.00002	1024	128	6	adam	adversarial negative sampling	64
TuckER	0.0005	200	128	-	adam	uniform	0
ConvKB	0.0001	100	128	-	adam	adversarial negative sampling	1

Table 5: Table showing the hyperparameters used in the KG embedding methods trained on the WN18RR dataset.

[7] Remi M Sakia. 1992. The Box-Cox transformation technique: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)* 41, 2 (1992), 169–178.
[8] C. Spearman. 1907. Demonstration of Formulæ for True Measurement of Correlation. *The American Journal of Psychology* 18, 2 (1907), 161–169. http://www.jstor.org/stable/1412408
[9] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2018. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*.
[10] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*. PMLR, 2071–2080.
[11] Cédric Villani. 2008. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.
[12] C. Villani. 2009. Optimal transport. *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* 338 (2009).
[13] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.
[14] Shih-Yuan Yu, Sujit Rokka Chhetri, Arquimedes Canedo, Palash Goyal, and Mohammad Abdullah Al Faruque. 2021. Pykg2vec: A Python Library for Knowledge Graph Embedding. *J. Mach. Learn. Res.* 22 (2021), 16–1.