
Rethinking Knowledge Graph Evaluation Under the Open-World Assumption

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Most knowledge graphs (KGs) are incomplete, which motivates one important
2 research topic on automatically complementing knowledge graphs. However,
3 evaluation of knowledge graph completion (KGC) models often ignores the
4 incompleteness—facts in the test set are ranked against all unknown triplets which
5 may contain a large number of missing facts not included in the KG yet. Treating all
6 unknown triplets as false is called the closed-world assumption. This closed-world
7 assumption might negatively affect the fairness and consistency of the evaluation
8 metrics. In this paper, we study KGC evaluation under a more realistic setting,
9 namely the *open-world assumption*, where unknown triplets are considered to
10 include many missing facts not included in the training or test sets. For the currently
11 most used metrics such as mean reciprocal rank (MRR) and Hits@K, we
12 point out that their behavior may be unexpected under the open-world assumption.
13 Specifically, with not many missing facts, their numbers show a logarithmic trend
14 with respect to the true strength of the model, and thus, the metric increase could be
15 insignificant in terms of reflecting the true model improvement. Further, considering
16 the variance, we show that the degradation in the reported numbers may result
17 in incorrect comparisons between different models, where stronger models may
18 have lower metric numbers. We validate the phenomenon both theoretically and
19 experimentally. Finally, we suggest possible causes and solutions for this problem.

20 1 Introduction

21 Knowledge graph (KG) is a structural method to store facts about some field or the world. Because
22 most KGs are incomplete, the knowledge graph completion (KGC) task is proposed to automatically
23 complement the existing KG with missing facts. However, when we do not know the missing facts
24 in advance, we must manually evaluate whether each predicted completion is correct, which is an
25 impossible task for modern KGs. This problem is called the *open-world problem* and the assumption
26 that KGs are incomplete is called the *open-world assumption*. A general solution is to extract the
27 training, validation and test sets from the existing incomplete KG and then evaluate the trained models
28 on the test set. Then, a natural question is whether the conclusion drawn from the incomplete test set
29 is consistent with the true strength of the model, which should be measured on the complete KG.

30 To answer this question, we need to investigate the metrics used to evaluate KGC models. KGC
31 models are often evaluated by ranking-based metrics, such as mean reciprocal rank (MRR) and
32 Hits@K. Under the open-world assumption, when a missing fact that should have been included in
33 the test answers is predicted by the model, **its ranking could be higher than some test answers,**
34 **which makes the rankings of these test answers drop.** In this situation, despite actually recognizing
35 more right answers, the metrics drop instead.

Table 1: The filtered ranking as well as the metric MRR. “w/o c”: without correction and “with c”: with correction. Query: “*What sports were included in the 1956 Summer Olympics?*”

	Test Answers		Missing Answers					
	swimming	sailing	water polo	boxing	dressage	show jumping	canoe sprint	cycling
ranking w/o c	5	5	1	2	3	4	7	9
ranking with c	1	1	1	1	1	1	3	4

36 To intuitively show the problem, we train BetaE [Ren and Leskovec, 2020], one state-of-the-art
 37 multi-hop KGC model, on the FB15k-237 dataset [Toutanova and Chen, 2015]. One of the test
 38 queries is “*What sports were included in the 1956 Summer Olympics?*”. The two test answers are
 39 swimming and sailing, both with the filtered rankings (refer to Section 2) of 5, so the MRR on this
 40 query is 20%. However, when we manually check the first 30 predictions, we found that many of
 41 them are in fact sports included in the 1956 Summer Olympics but not included in the answer set.¹
 42 We present these missing answers in Table 1. We can see that all the four sports previously ranking
 43 higher than the two test answers turn out to be missing true answers. Thus, if we correct the answer
 44 set by adding these missing answers to the test set, the actual filtered rankings of the two test answers
 45 are both 1, and the corrected MRR on the new test set becomes 82% which is much higher than the
 46 reported 20%, indicating that the model strength on this query is significantly underestimated.

47 In this paper, we study the odd behavior of the ranking-based metrics under the open-world assump-
 48 tion, and summarize two problems affecting the KGC evaluation: 1) **Metric Degradation**. It means
 49 that with the increasing of the actual model strength, the increasing of the reported metric becomes
 50 slower and slower. Thus, the reported metric might not be able to reflect the true model improvement.
 51 2) **Metric Inconsistency**. It means that when comparing two models, the model with lower reported
 52 metric may actually have better performance if we evaluate them on the complete KG.

53 Our main contributions include that: For the first time, we theoretically analyze the evaluation of
 54 KGC under the open-world assumption and point out the degradation and inconsistency problems.
 55 Furthermore, we suggest that the degradation and inconsistency may be related to the focus-on-top
 56 behavior of the metrics, and provide a solution to relieve the two problems. Finally, we verify the
 57 theoretical analysis through experiments on an artificial closed-world KG.

58 2 Background and related work

59 **Knowledge graph completion** Current KGC models can be mainly categorized into three classes:
 60 logic-based, embedding-based, and neural-based. **Logic-based models** [Joseph and Riley, 1998,
 61 Richardson and Domingos, 2006] use some explicit rules for KGC, which are manually provided or
 62 mined by some rule-mining methods, such as [Galárraga et al., 2013, Yang et al., 2017, Sadeghian
 63 et al., 2019]. These models search through the existing KG and deduce missing facts according to the
 64 given rules. However, this process can be time-consuming and noise-sensitive. At the same time, if
 65 the KGs are highly incomplete, the performance could be poor. **Embedding-based models** [Bordes
 66 et al., 2013, Yang et al., 2015, Trouillon et al., 2016, Sun et al., 2019] represent entities and relations
 67 by learned vectors or tensors, where the possibility of a fact is measured by a *score function*. These
 68 models have good scalability and can be applied to large and sparse KGs. Some works aim to
 69 generalize embedding-based models to more patterns [Trouillon et al., 2016, Abboud et al., 2020]
 70 and more assumptions (such as multiple answers) [Vilnis et al., 2018, Ren et al., 2020, Abboud et al.,
 71 2020]. One of the interesting directions is to consider multi-hop reasoning [Ren et al., 2020, Ren and
 72 Leskovec, 2020, Zhang et al., 2021], where a query can be composed by several conditions, such
 73 as “*Who is the Canadian and won the Turing Award ?*” Note that the open-world problem could be
 74 more severe in the multi-hop setting, because missing in any condition leads to missing in the final
 75 results. **Neural-based models** combine neural networks with embeddings. Dettmers et al. [2018]
 76 and Nguyen et al. [2018] use a convolution networks as the score function to enlarge the capacity of
 77 the models. Nathani et al. [2019], Vashishth et al. [2020] and Wang et al. [2021] use graph neural
 78 networks on the KGs to learn the embeddings or directly predict the links.

79 **KG evaluation** Current KGC evaluation resorts to manually split training, validation and test sets
 80 from the incomplete KG. Given a test query $r(e_h; ?)$ (which entities have the relation r with the

¹All sports held at this Olympics are in: https://en.wikipedia.org/wiki/1956_Summer_Olympics.

81 head entity e_h ?), a typical method is to predict a score for all entities as the tail entity, rank all the
 82 entities, and then measure the average of a ranking-based function $h(r)$ on the test answers. Here, the
 83 most-used metrics are MRR $h(r) = 1/r$ and the Hits@K $h(r) = \mathbb{1}(r \leq K)$. Because there could be
 84 multiple answers for a query, the metrics should be **filtered**, which means the answers in the training
 85 and test sets do not occupy a position so that the number of training and test answers does not affect
 86 the metrics. The details of the filtering can be found in [Bordes et al., 2013]. Due to the nonlinearity
 87 of most ranking-based metrics, some works have theoretically investigated their behavior. Wang
 88 et al. [2013] point out some ranking-based metrics always converge to 1 on different models as the
 89 number of objects to rank goes to infinity, so that the performance of models is indistinguishable.
 90 Krichene and Rendle [2020] analyze the behavior of ranking-based metrics under negative sampling.
 91 They point out the sampled metrics can be inconsistent with exact metrics and all metrics lose their
 92 focus-on-top feature and collapse to a linear one, AUC-ROC, in the small sample limit. Sun et al.
 93 [2020] focus on the unfair tie-breaking methods. Akrami et al. [2020] find some data argumentation
 94 such as adding inverse relations could be a kind of excessive data leakage during evaluation.

95 3 Open-world problem

96 In this section, we formally define the open-world problem that will be analyzed in our paper.

97 **Definition 3.1** (Knowledge Graph). A *knowledge graph* is a relational graph $G = (E; F; R)$ where
 98 E is the vertex set containing *entities*, F is the edge set containing *facts*, and R is the *relation* set.
 99 Each edge $f \in F$ is labeled by a relation. If an edge f between entities e_h and e_t is labeled by
 100 relation $r \in R$, we denote the edge f as $r(e_h; e_t)$ where e_h is the head entity and e_t is the tail entity.

101 In this paper, we assume E and R are fixed. Therefore, we sometimes directly use G to denote the
 102 fact set F , and $r(e_h; e_t) \in G$ means there is relation r between e_h and e_t in the KG G .

103 A set of KGs with the same entities E and relations R but different facts F is called a *world* and
 104 denoted by $W(E; R)$. An (open-world) KG can be considered as an observation or understanding of
 105 the world where there could be unobserved or unknown facts, while the closed-world KG contains all
 106 the true facts of the world. Formally, the closed-world and open-world KGs are defined as follows:

107 **Definition 3.2** (Closed-World KG and Open-World KG). For a world $W(E; R)$, the closed-world
 108 KG G is the closure of the world.

$$G = \bigcup_{G^j \in W} G^j;$$

109 where the union is defined on the fact set. And for a KG $G^j \in W$, if $G^j \notin G$, G^j is open-world.²

110 Some property of closed-world KGs: 1) There is a one-to-one correspondence between closed-world
 111 KGs and worlds $W(E; R)$. 2) All the KGs in a world are subgraphs of the closed-world one. 3) If G
 112 is the closed-world KG of the world W , we have $f \in G \iff \exists G^j \in W; f \in G^j$.

113 The third property is critical. It means with the closed-world G , we **know** there is **no such a relation**
 114 r between e_h and e_t in the world when $r(e_h; e_t) \notin G$. Given the closed-world KG, we have all
 115 knowledge of the world, including both the *positive* and *negative* one. Conversely, if a KG is open-
 116 world, we **do not know** whether the triplet is *false* or *unknown* when $r(e_h; e_t) \notin G$. In other words,
 117 an open-world KG only contains *positive* knowledge.

118 In the rest of the paper, we denote the closed-world KG as G_{full} . Because we want to study the
 119 evaluation, we denote the existing open-world dataset as G_{test} , and extract the training set G_{train}
 120 from G_{test} . Here, $G_{train} \subseteq G_{test} \subseteq G_{full}$ and the facts in G_{train} , $G_{test} \setminus G_{train}$, $G_{full} \setminus G_{test}$
 121 are **training facts**, **test facts** and **missing facts** respectively. In addition, we also call the facts in
 122 $G_{full} \setminus G_{train}$ **full test facts** and $G_{test} \setminus G_{train}$ **sparse test facts**.

123 Now, we can formally define the *open-world problem*. We believe the actual strength of a model
 124 should be evaluated on the full test facts $G_{full} \setminus G_{train}$. However, because the closed-world KG
 125 G_{full} is unavailable, the evaluation is often performed over $G_{test} \setminus G_{train}$. The question is:

126 *Whether the conclusions from evaluation on the sparse test facts $G_{test} \setminus G_{train}$*
 127 *lead to consistent conclusions from evaluation on the full test facts $G_{full} \setminus G_{train}$.*

²Some works use open-world to refer to not only facts but also entities may be incomplete.

128 4 Theoretical analysis on metric degradation and inconsistency

129 To study the open-world problem, we theoretically analyze the behavior of ranking-based metrics
 130 with missing facts. All the proofs are in Appendix A.1. The randomness comes from two sources:
 131 the missing of facts and the predictions of the model. We model them as two random events.

- 132 • **Missing Fact Model:** For a full test fact $r(e_h; e_t) \in G_{full} \cap G_{train}$, X means it is a missing
 133 fact with $P(X) = \frac{1}{N}$ while \bar{X} means it is included in the sparse test set $G_{test} \cap G_{train}$ with
 134 $P(\bar{X}) = 1 - \frac{1}{N}$. $\frac{1}{N}$ is called the sparsity of the KG.
- 135 • **Prediction Model:** For simplicity of analysis, we model KGC as a classification task. In fact,
 136 an ideal (oracle) KGC model is exactly a classification model, which identifies all the correct
 137 facts. Here, for a full test fact $r(e_h; e_t) \in G_{full} \cap G_{train}$, Y means the answer e_t is correctly
 138 classified as positive with $P(Y) = l$. l is called the strength of a model. We break ties uniformly
 139 at random for entities classified into the same class.

140 4.1 Expectation degradation

141 We first assume the independence of the random events X and Y . We show that the expectation
 142 of the metrics will degrade with missing facts. Specifically, the increasing of the metrics shows a
 143 logarithmic trend, so that it could be too flat to reflect the true increasing of the model strength.

144 Assume the number of entities is N_{entity} in the KG. For a given query $r(e_h; ?)$, let N be the number
 145 of full test answers. The random variant m is the number of missing answers $G_{full} \cap G_{test}$, and it
 146 follows the binomial distribution $B(N; \frac{1}{N})$. The other $N - m$ answers are test answers. We denote
 147 the filtered ranking of the entity e as $r(e)$. Then we have the lemma.

148 **Lemma 4.1** (Expectation of ranking-based metrics). *With the modeling of missing fact and prediction*
 149 *as above, the expectation of ranking-based metric $M = \frac{1}{N - m} \sum_{i=1}^m \frac{1}{f(r(e_i))}$ can be expressed as*

$$E(M) = \frac{1}{(N + 1)} \sum_{k=0}^N \frac{1}{f(k + 1)} (1 - \frac{1}{N})^k (\frac{1}{N})^1 \quad (1)$$

150 where $\hat{\cdot}$ is the cumulative distribution function (cdf) of binomial distribution $B(N + 1; \frac{1}{N})$ and
 151 $0 < \frac{1}{N} < 1$.

152 Generally, the N_{entity} is a large number and the answer rate $\frac{N}{N_{entity}} < 10\%$ in almost all queries,
 153 so the item $\frac{1}{N}$ is negligible. In the rest of the paper, we denote $\hat{E} = \frac{1}{(N + 1)} \sum_{k=0}^N (1 - \frac{1}{N})^k (\frac{1}{N})^1 = f(k + 1)$
 154 which is a good approximation of E .

155 With this lemma, we get a closed-form expression of the expectation of the ranking-based metric M .
 156 However, this expression cannot explain why the metrics will degrade. Next, we derive the form of
 157 derivative of the expectation \hat{E} w.r.t the model strength l to account for its degradation.

158 **Corollary 4.1** (Derivative of Expectation w.r.t Model Strength). *Let $g(r) = r = f(r); r \in \mathbb{N}_+$ and*
 159 *$g(0) = 0$. Under the condition of 4.1, the derivative of the expectation w.r.t the model strength l is*

$$\frac{d\hat{E}(M)}{dl} = \frac{1}{l(N + 1)} E_{k \sim B(N + 1; \frac{1}{N})} g(k) \quad (2)$$

160 And for the most-used metrics MRR and Hits@K, their derivative are expressed as follows.

161 **Corollary 4.2** (Derivative of MRR w.r.t strength l). *For MRR where $f(r) = \frac{1}{r}; r \in \mathbb{N}_+$, its*
 162 *derivative w.r.t. l is*

$$\frac{d\hat{E}(\text{MRR})}{dl} = \frac{1}{l(N + 1)} \sum_{k=0}^N \frac{1}{(k + 1)^2} \quad (3)$$

163 where $\sum_{k=0}^N \frac{1}{(k + 1)^2} = \frac{1}{N + 1}$.

164 When l and N are not too small, the term $\frac{1}{(k + 1)^2}$ is negligible. In this situation, the derivative of the
 165 metric MRR w.r.t the model strength is approximately of $O(1 = Nl)$, which will result in insignificant
 166 changes in the metric with the increasing of the model strength.

167 **Corollary 4.3** (Derivative of Hits@K w.r.t strength l). For Hits@K where $f = 1$ for $r = k$ and
 168 $f = +1$ otherwise, the derivative is

$$\frac{d\hat{E}(\text{Hits@K})}{dl} = (K - 1); \quad (4)$$

169 where B is the cdf of the binomial distribution $B(N; l)$.

170 The behavior of Hits@K is similar to MRR when $K = N$. When l is not too small and K is not
 171 too large, the derivative $(K - 1)$ is so small that the increase could be insignificant.

172 Finally, we further approximate Equation (1) by a more intuitive expression with a tolerable error.

173 **Theorem 4.1** (Expectation of MRR). For MRR, we can further approximate its expectation by

$$\hat{E}(\text{MRR}) = \frac{\ln(l) + \ln(\frac{1}{1-l}) + \ln(N+2) + \gamma}{(N+1)} := \bar{E};$$

174 where $\gamma = 0.577$ is the Euler's constant and the error³ $e = j\bar{E} \max\{F_{\frac{1}{2(N+1)^2}}; \frac{(1-l)^{N+1}}{(1-l)^{N+1}}\}$
 175 $\frac{\ln(1-l)}{(N+1)} g$.

176 Firstly, we explain the rationales of the approximation \bar{E} as follows.

- 177 • N is large enough. For many KGs, especially those commonsense KGs which are not limited to
 178 a certain field, the number of answers is often quite large. In the experiment conducted by Ren
 179 and Leskovec [2020], there are many tests queries with dozens or hundreds of answers.
- 180 • Sparsity α is not too small. For most real-world KGs, although we do not exactly know their
 181 sparsity, we expect many of them could have a rather high sparsity due to the incompleteness
 182 of knowledge extraction and the long-tail distribution of commonsense knowledge.
- 183 • Model strength l is not too small. Here we are more concerned with how to select and evaluate
 184 those models that perform well.

185 Under the above conditions, the relative error e is negli-
 186 gible. To further show that our approximation is reliable,
 187 we do numerical simulations as shown in Figure 1. For
 188 $l = 0.2$ and $l > 0.3$, the analytical and numerical
 189 curves almost overlap, which means the \log approximation
 190 is accurate. At the same time, we point out the \log trend
 191 is just the reason why the curve becomes flatter and flatter.
 192 The details of the simulation is in Appendix A.2.

193 Theorem 4.1 illustrates the metric degradation intuitively.
 194 There are some conclusions about the MRR under the
 195 open-world assumption: 1) Although the theoretical maxi-
 196 mum of MRR is 1, the expectation of the MRR of a perfect
 197 model $l = 1$ is still much lower than 1 and depends on the
 198 sparsity of the KG. 2) With the sparsity α not very small,
 199 the MRR will be a \log function of the strength l times the
 200 answer number N which means that as the model gets stronger, the increase of the metric MRR will
 201 be less and less significant. The sparser the KG is, the more severe the degradation problem is. Note
 202 that in the closed-world KG, the curve should be very closed to $y = x$.

203 4.2 Inconsistency due to high variance

204 In Figure 1, another notable phenomenon is the vibrated curves, which suggests instability of the
 205 metric and relatively high variance. This phenomenon combined with the flattening of the expectation
 206 can lead to *inconsistency*, which means higher MRR might not mean stronger models unless the
 207 difference of the metric is large enough, because the increasing of expectation could be easily
 208 overwhelmed by the variance.

209 One trivial method to solve the problem is to use more test queries. Here we show the number of
 210 queries required to ensure the reliability of conclusions can be very large.

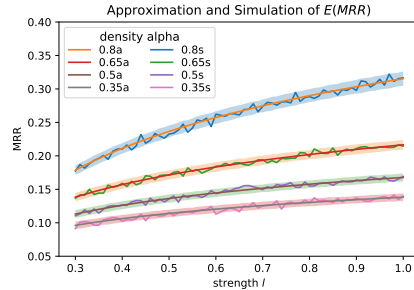


Figure 1: Theoretical \log approximation (a) and numerical simulation (s). The shadow shows the $[-2; 2]$ interval, where σ is the numerical std.

³The error bound has changes compared to the submitted version.

211 **Theorem 4.2** (Consistence with High Probability). Assuming the number of test queries N_q is large
 212 enough ($N_q > 50$), we can approximate the average MRR $M = \frac{1}{N_q} \sum_{i=1}^{N_q} \text{MRR}(q_i)$ to follow a
 213 normal distribution. Given two independent models \mathcal{M}_1 and \mathcal{M}_2 whose strength is l and $l + \delta$
 214 respectively. The probability of inconsistency between two models can be approximated as follows.

$$P[M(\mathcal{M}_1) > M(\mathcal{M}_2)] = \frac{1}{(N+1)} \Phi\left(\frac{\sqrt{N_q} \ln(1 + \frac{\delta}{l})}{\sqrt{V(\delta; l) + V(\delta; l + \delta)}}\right); \quad (5)$$

215 where Φ is the cdf of the standard normal distribution.

216 Note for a given KG, the sparsity δ and the answer number N are fixed. Assuming $\frac{\delta}{l} \ll 1$, we have
 217 $V(\delta; l) \approx V(\delta; l + \delta) := V$ and $\ln(1 + \frac{\delta}{l}) \approx \frac{\delta}{l}$. Then we have the following corollary.

218 **Corollary 4.4** (Lower Bound of the Number of Queries). Under the above assumption of $\frac{\delta}{l} \ll 1$,
 219 with the upper bound of inconsistency probability p , the number of test queries required N_q has a
 220 lower-bound as follows.

$$N_q \geq \frac{c(\delta; l; N; p)}{(\frac{\delta}{l})^2}; \quad (6)$$

221 where $c(\delta; l; N; p) = 2(1 - l(N + 1))^{-1} (p)^2 V$.

222 Note that the required number is of the second order $O((1 - l)^{-2})$, which means one should be
 223 particularly careful when comparing two models with close strength. For example, when we set
 224 $\delta = 0.35; l = 0.7; N = 43$ and $p = 5\%$ we have $c \approx 2.85$ where we use the numerical variance
 225 $V = 7.4 \cdot 10^{-3}$. In this situation, when $l = 0.05$, $N_q \approx 1140$, while when $l = 0.01$,
 226 $N_q \approx 28500$ which cannot be easily satisfied.

227 4.3 Correlation between missing and misclassification

228 In the previous two subsections, we analyze the degradation and inconsistency with independence
 229 assumption between missing facts and model predictions. In some conditions, there could be
 230 correlation between the missing data and the trained models. Let us give some examples:

- 231 • The missing facts in closed-world KG G_{Full} follow some non-uniform distribution. For example,
 232 in some KGs, the missing facts are more frequently related to some certain entities. The model
 233 could be under-trained on these entities because there are more missing facts related to them
 234 when training. And when the model is tested, the queries with more missing test answers
 235 correspond to the lower predicting capacity.
- 236 • The target KG has been preliminarily complemented by some models. In this situation, the
 237 missing facts show a negative correlation with the predictive power of this type of models.

238 Next, we will extend the previous analysis to the situation without the independence assumption. For
 239 this goal, we use the correlation coefficient to model the correlation between random event X : a fact is
 240 missing and Y : this fact is predicted by the model.

241 **Definition 4.1** (Correlation between fact missing and model prediction). The correlation coefficient
 242 r between two random events X and Y can be defined as follows.

$$r = \frac{P(XY) - P(X)P(Y)}{\sqrt{P(X)P(X)P(Y)P(Y)}}; \quad (7)$$

243 With the correlation coefficient r , the prediction accuracy on missing answers $e \in G_{Full} \cap G_{test}$ and
 244 on test answers $e \in G_{test} \cap G_{train}$ can be calculated as the conditional probability as $P(Y|X) =$
 245 $\frac{l + \delta}{l(1 - \delta)} = \frac{1}{1 - \frac{\delta}{l}}$ and $P(Y|\bar{X}) = \frac{l}{l(1 - \delta)} = \frac{1}{1 - \frac{\delta}{l}}$. We denote them as h_1 and h_2 respectively.

246 Theorem 4.1 can then be generalized as follows:

247 **Theorem 4.3.** We have the approximation for MRR

$$E(MRR) \approx \frac{h_2}{h_1} \frac{\ln(h_1) + \ln(\frac{1}{1 - \frac{\delta}{l}}) + \ln(N + 2)}{(N + 1)} := \bar{E}; \quad (8)$$

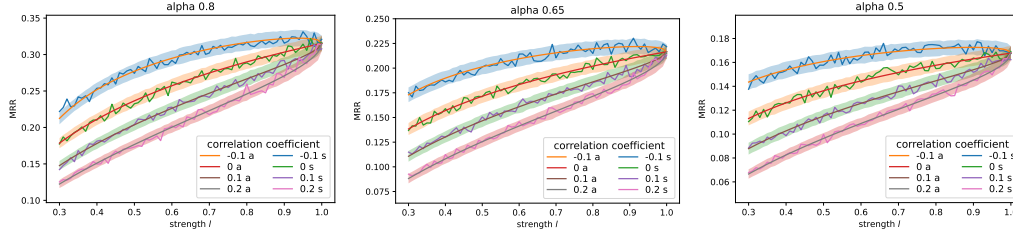


Figure 2: Theoretical approximation (a) and numerical simulation (s) of the expectation with correlation coefficient ρ . Details of the figures are the same as in Figure 1.

248 The theoretical error analysis is in Appendix A.1. We also evaluate the approximation by numerical
 249 simulation and the results are shown in Figure 2. The approximation fits the numerical simulation
 250 well. Comparing different models with the same model strength l but different correlation coefficient
 251 ρ , the models with smaller ρ have the higher MRR. The phenomenon is consistent with Corollary 4.5.
 252

253 **Corollary 4.5** (The derivative w.r.t l with correlation). *If we have the inequality $l_1 > (N + 2)$*
 254 $\exp\left(\rho + \frac{(1-l)}{l}\right)$, then the derivative $\frac{\partial E(\text{MRR})}{\partial l} < 0$.

255 The condition in Corollary 4.5 has requirements for lower bound l_1 and ρ . If l_1 is close to 0, the
 256 condition can be violated. This corollary suggests more severe inconsistency. In a reasonable range,
 257 the expectation of MRR is monotonically decreasing w.r.t l . This conclusion suggests that the metric
 258 MRR may favor the models with smaller l instead of larger l . Note that the inconsistency is of
 259 expectation, which cannot be solved by more test queries.

260 5 Relationship between focus-on-top and degradation

261 We have pointed out the degradation and inconsistency under the open-world assumption for some
 262 most-used ranking-based metrics. Specifically, the derivative of the metrics w.r.t l can be too small
 263 to reflect the increasing of the true improvement of the model strength (degradation). According
 264 to Corollary 4.1, the derivative is related to the expectation of $g(r) = r=f(r)$ where r follows a
 265 binomial distribution. We point out the degradation is due to the too small expectation of g relative to
 266 the denominator N , which is inherently caused by a property of the metrics called *focus-on-top*.

267 The focus-on-top property means that the metrics are more sensitive to ranking change in top places.
 268 For example, MRR changes from 1 to 0.5 when the ranking changes from 1 to 2, but only changes
 269 from $1e-2$ to $0.99e-2$ when the ranking changes from 100 to 101. This property can simulate the
 270 human behavior that people pay more attention to the top answers. However, under the open-world
 271 assumption, the focus-on-top property causes negative impacts by making the function $1=f(r)$
 272 decrease too fast so that the expectation of g is too small relative to N . For example, according to
 273 Corollary 4.3, a smaller K means more focus-on-top and smaller derivative.

274 We can also understand the relationship intuitively. Focusing-on-top means that a few missing
 275 answers can have a large impact on the metric, especially when the model performance is already good
 276 and the rankings of the rest answers fall into the sensitive range. It is also consistent with our
 277 observation that the flattening problem is more severe when the strength l increases.

278 There is a trade-off between focus-on-top and consistency. Therefore, one solution to the degradation
 279 and inconsistency is to add in some less focus-on-top metrics as a verification when evaluating, which
 280 have a relatively slower descending rate. For example, the \log -MRR where $f(r) = \log_2(r + 1)$ and
 281 ρ -MRR where $f(r) = r^p$; $0 < p < 1$ are less focus-on-top than the standard MRR. If the conclusions
 282 of these less focus-on-top metrics are consistent with the MRR or Hits@K, the credibility of the
 283 conclusions will be greatly enhanced.

284 6 Experiments on an artificial KG

285 In this section, we aim to conduct experiments with practical KGC models on a meaningful closed-
 286 world KG to further verify our conclusions. The reason we want a **closed-world KG** is for comparing

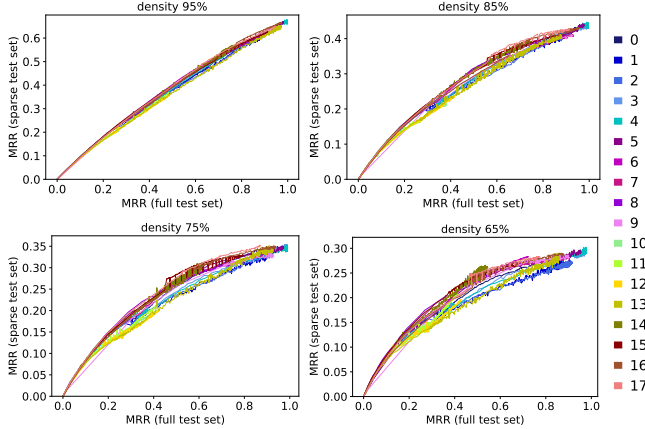


Figure 3: Full test and sparse test MRR on the artificial family tree KG. Note the ranges of y-axis are different.

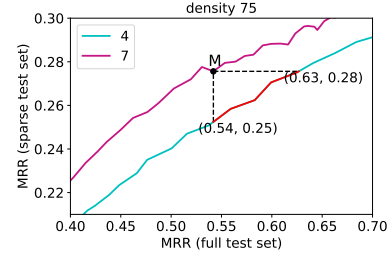


Figure 4: A zoom-in of two curves 4 and 7 under density $d = 0.75$. The checkpoints of model 4 lying on the red segment all have better full test MRR than the checkpoint M of model 7, while reporting worse sparse test MRR under the open-world setting.

287 the reported MRR with the true model strength which should be measured on the full test set. To find
 288 a closed-world KG, however, it is impractical to resort to existing real-world ones since we have no
 289 guarantee that the KG has no missing facts. Therefore, we must resort to some artificial KGs.

290 For this purpose, we generate an artificial **family tree KG**, which contains 6,004 entities, 23 relations,
 291 and 192,532 facts. The relation set contains all common family relations such as parent, child,
 292 husband, wife, sister, and brother. The details are included in Appendix A.3. The generated KG is
 293 **closed-world** since all the facts can be deduced by a symbolic reasoning tool called DLV system
 294 [Leone et al., 2006] (free for academic use). With the closed-world KG, we can simulate the practical
 295 open-world setting by artificially controlling the degree of random fact missing, which is measured
 296 by density $d = |G_{test}|/|G_{full}|$. The relation between d and l is explained in Appendix A.3.

297 With the closed-world KG, we aim to verify our previous conclusions restated below:

- 298 1. There is metric degradation which means the curves of metric increasing become flatter
 299 and flatter with the increasing of model strength l . Further, the degradation may result in
 300 inconsistency, where stronger models report lower metric numbers.
- 301 2. Considering the correlation between fact missing and model prediction, if the correlation
 302 degrees vary among different models, the inconsistency problem may become more severe.
- 303 3. The degrees of degradation and inconsistency are related to the focus-on-top property of the
 304 metrics. With less focus-on-top metrics, these problems could be relieved.

305 All codes are provided in the supplement and will be released after publication. The experiments
 306 were run on two clusters with four NVIDIA A40 and six NVIDIA GeForce 3090 GPUs respectively.

307 6.1 Degradation and inconsistency under independence assumption

308 We train four KGC models with different hyperparameter settings (which results in 18 different models
 309 in total) and test them on full test set $G_{full} \cap G_{train}$ and sparse test set $G_{test} \cap G_{train}$ respectively.
 310 The full test metric can be considered as a measurement of model strength l which is what we
 311 really want to measure, while the sparse test metric is what we can observe in practice. We plot
 312 the sparse-full test curve under different densities in Figure 3, where each curve represents a model
 313 whose label is shown in the right legend, and the details of the models are given in Appendix A.4.

314 From the figure, we first observe that these curves are indeed shaped like *log* curves. The increasing
 315 of the sparse test MRR is slower and slower with the increasing of the full test MRR. Due to the
 316 flattening of the curves, the same sparse MRR has a rather broad interval of the corresponding full
 317 metric. This phenomenon indicates the degradation of the metric MRR. And as the sparsity increases,
 318 the range of the y-axis shrinks (i.e., the curves become flatter), which means the degradation is more
 319 severe. Further, these results demonstrate the inconsistency problem of MRR. To illustrate this point
 320 more clearly, we zoom in a part of the full figure with two curves as shown in Figure 4. For the model
 321 checkpoint corresponding to point M on curve 7, any model checkpoint corresponding to a point on
 322 the red segment of curve 4 is **actually stronger than that model**, but reports a **lower sparse MRR**.

(a)

(b)

Figure 5: (a) Full test and sparse test MRR on independent (above) and correlated (bottom) family tree KG. (b) MRR, and less focus-on-top metrics. Both are under density 75%

323 6.2 Correlation between fact missing and model prediction

324 In this part, we will simulate the third example we provided in Section 4.3 to check our theory
325 considering the correlation between fact missing and model prediction. Here we use one of the trained
326 ComplEx model (labeled as 16) [Trouillon et al., 2016] to predict on the full test set n_{train}
327 and use its predictions to choose the test set n_{train} and missing facts n_{test} . The
328 missing facts are highly correlated with this ComplEx model and therefore could be correlated with
329 other models according to the correlation between different frameworks and model settings. Then we
330 test the other models except for this ComplEx model on the correlated test set. The results of density
331 $d = 75\%$ are shown in Figure 5a and others are shown in Appendix A.5. Though the correlation
332 coefficient is not available for these different models, we indeed observe the gaps between different
333 models become larger than the independent setting, which suggests the inconsistency is more severe.

334 6.3 Less focus-on-top metrics

335 The next conclusion is that with less focus-on-top metrics, the degradation and inconsistency can be
336 relieved. In Figure 5b, we show the sparse-full curves with some less focus-on-top metrics (MRR
337 and p-MRR) for the experiments from Section 6.1. Their curves are more close to instead of
338 the log function. The flattening is less significant due to the wider range of y-axis. We also observe that
339 the gaps between different models become smaller, which indicates inconsistency is also relieved.
340 Additional results with more metrics, density and correlation are shown in Appendix A.6.

341 7 Conclusion and future work

342 In this paper, we study KGC evaluation under the open-world assumption. Theoretically, we model
343 KGC as a positive-negative classification and then deduce an approximation of the expectation of the
344 ranking-based metrics with or without the independence assumption. According to the approximation,
345 we illustrate the degradation and inconsistency of these metrics under the open-world assumption.
346 Furthermore, we point out the focus-on-top property of ranking-based metrics worsen the degradation
347 and inconsistency. Finally, we generate a closed-world family tree KG and do experiments to verify
348 our theoretical conclusions. There is still some future work. First, our analysis is based on the
349 positive-negative classification model, which may be too idealistic. In practice, the ranking in positive
350 and negative parts may not be uniformly at random. A more realistic modeling of the KGC task is a
351 direction for our future research. In addition, the correlation between missing facts and prediction
352 could be more complex than our analysis. Finally, we are curious about the possibility to find a more
353 fundamental solution to the open-world problem, which we leave for future work.

354 **References**

355 Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. Boxe: A box embed-
356 ding model for knowledge base completion. *Neural Information Processing Systems* 2020.

357 Farahnaz Akrami, Mohammed Samiul Saeef, Qingheng Zhang, Wei Hu, and Chengkai Li. Realistic
358 Re-evaluation of Knowledge Graph Completion Methods: An Experimental Study. *Proceedings*
359 *of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020.

360 Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko.
361 Translating embeddings for modeling multi-relational data. *Neural Information Processing*
362 *Systems*, 2013.

363 Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2D
364 Knowledge Graph Embeddings. *Association for the Advancement of Artificial Intelligence*
365 *Conference* 2018.

366 Luis Antonio Galárraga, Christina Teioudi, Katja Hose, and Fabian Suchanek. Amie: Association
367 rule mining under incomplete evidence in ontological knowledge bases. *Proceedings of the*
368 *22nd International Conference on World Wide Web*, 2013.

369 Patrick Hohenecker and Thomas Lukasiewicz. Ontology reasoning with deep neural networks. In
370 *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* 2020.

371 Giarratano Joseph and Gary Riley. *Expert systems: principles and programming*. 1998.

372 Walid Krichene and Steffen Rendle. On Sampled Metrics for Item Recommendation. *Proceedings*
373 *of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*
374 *2020*.

375 Nicola Leone, Gerald Pfeifer, Wolfgang Faber, Thomas Eiter, Georg Gottlob, Simona Perri, and
376 Francesco Scarcello. The dlv system for knowledge representation and reasoning. *ACM Trans.*
377 *Comput. Logic*7(3):499562, jul 2006. ISSN 1529-3785.

378 Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. Learning attention-based
379 embeddings for relation prediction in knowledge graphs. *Proceedings of the 57th Annual*
380 *Meeting of the Association for Computational Linguistics* 2019.

381 Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A novel embedding model
382 for knowledge base completion based on convolutional neural networks. *Proceedings of the*
383 *2018 Conference of the North American Chapter of the Association for Computational Linguistics:*
384 *Human Language Technologies* 2018.

385 Hongyu Ren and Jure Leskovec. Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge
386 Graphs. In *Advances in Neural Information Processing Systems* 2020.

387 Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in
388 vector space using box embeddings. *International Conference on Learning Representations*
389 *2020*.

390 Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning* 62(1):
391 107–136, Feb 2006.

392 Ali Sadeghian, Mohammadreza Armandpour, Patrick Ding, and Daisy Zhe Wang. DRGM: End-to-
393 End Differentiable Rule Mining on Knowledge Graphs. 2019.

394 Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by
395 relational rotation in complex space. *International Conference on Learning Representations*
396 *2019*.

397 Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha Talukdar, and Yiming Yang. A re-evaluation
398 of knowledge graph completion methods. *Proceedings of the 58th Annual Meeting of the*
399 *Association for Computational Linguistics* 2020.

- 400 Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text
401 inference. In the 3rd workshop on continuous vector space models and their compositionality
402 2015.
- 403 Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. Complex
404 embeddings for simple link prediction. International Conference on Machine Learning, 2016.
- 405 Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-
406 relational graph convolutional networks. International Conference on Learning Representations
407 2020.
- 408 Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge
409 graphs with box lattice measures. Annual Meeting of the Association for Computational
410 Linguistics 2018.
- 411 Hongwei Wang, Hongyu Ren, and Jure Leskovec. Relational message passing for knowledge graph
412 completion. In the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining, 2021.
- 413 Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type
414 ranking measures. Proceedings of the 26th Annual Conference on Learning Theory, 2013.
- 415 Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and
416 relations for learning and inference in knowledge bases. International Conference on Learning
417 Representation, 2015.
- 418 Fan Yang, Zhilin Yang, and William W Cohen. Differentiable learning of logical rules for knowledge
419 base reasoning. Advances in Neural Information Processing Systems, 2017.
- 420 Zhanqiu Zhang, Jie Wang, Jiajun Chen, Shuiwang Ji, and Feng Wu. Cone: Cone embeddings for
421 multi-hop reasoning over knowledge graphs. Neural Information Processing Systems, 2021.

422 Checklist

- 423 1. For all authors...
- 424 (a) Do the main claims made in the abstract and introduction accurately reflect the paper's
425 contributions and scope? [Yes]
- 426 (b) Did you describe the limitations of your work? [Yes] We use some assumptions and
427 approximations, for which we have stated in Section 7.
- 428 (c) Did you discuss any potential negative societal impacts of your work? [?] Our work
429 does not involve specific application scenarios and has no negative social impact.
- 430 (d) Have you read the ethics review guidelines and ensured that your paper conforms to
431 them? [Yes] Our work conform to the ethics review guidelines. Our work does not
432 involve ethical risks. All datasets are commonly used and public, except the ones which
433 we generated ourselves. And all people are virtual in the artificial KG.
- 434 2. If you are including theoretical results...
- 435 (a) Did you state the full set of assumptions of all theoretical results? [Yes]
- 436 (b) Did you include complete proofs of all theoretical results? [Yes] The complete proofs
437 of all theoretical results are shown in Appendix.
- 438 3. If you ran experiments...
- 439 (a) Did you include the code, data, and instructions needed to reproduce the main experi-
440 mental results (either in the supplemental material or as a URL)? [Yes] We include the
441 code in the supplemental material.
- 442 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they
443 were chosen)? [Yes] All the details of our experiments, including data generation and
444 splits, hyperparameters are specified in Section 6 and Appendix A.3 and A.4.
- 445 (c) Did you report error bars (e.g., with respect to the random seed after running experi-
446 ments multiple times)? [Yes] The simulation and the experiments on artificial family
447 tree, we both report the results after multiple times.

- 448 (d) Did you include the total amount of compute and the type of resources used (e.g., type
 449 of GPUs, internal cluster, or cloud provider)? [Yes] See Section 6.
- 450 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- 451 (a) If your work uses existing assets, did you cite the creator? [Yes]
- 452 (b) Did you mention the license of the assets? [Yes] See Section 6 and A.3.
- 453 (c) Did you include any new assets either in the supplemental material or as a [Yes]?
 454 We generate an artificial family tree KG and we include it in the supplemental material.
- 455 (d) Did you discuss whether and how consent was obtained from people whose data you're
 456 using/curating? [N/A] The data is generated ourselves.
- 457 (e) Did you discuss whether the data you are using/curating contains personally identi-
 458 fiable information or offensive content? [N/A] The data does not contain personally
 459 identifiable information or offensive content. It is a virtual KG.
- 460 5. If you used crowdsourcing or conducted research with human subjects...
- 461 (a) Did you include the full text of instructions given to participants and screenshots, if
 462 applicable? [N/A]
- 463 (b) Did you describe any potential participant risks, with links to Institutional Review
 464 Board (IRB) approvals, if applicable? [N/A]
- 465 (c) Did you include the estimated hourly wage paid to participants and the total amount
 466 spent on participant compensation? [N/A]

467 A Appendix

468 A.1 Proof

469 A.1.1 Lemma 4.1

470 Proof. First, according to the linearity of expectation, we have

$$E(M) = E_m E \frac{1}{N} \sum_{i=1}^N \frac{1}{f(r(e_i))} = E_m E \frac{1}{f(r(e))} ;$$

471 where $E(\frac{1}{f(r(e))}) = E(\frac{1}{f(r(e_1))}) = \dots = E(\frac{1}{f(r(e_{N-m}))})$, and here $\{e_1, e_2, \dots, e_{N-m}\}$ is the test
 472 answer set. We denote the minimal item as $e_{(1)}$. Then, using the conditional expectation, we
 473 have

$$\begin{aligned} E \frac{1}{f(r(e))} &= P(Y) E \frac{1}{f(r(e))} | Y + (1 - P(Y)) E \frac{1}{f(r(e))} | \bar{Y} \\ &= l E \frac{1}{f(r(e))} | Y + (1 - l) E \frac{1}{f(r(e))} | \bar{Y} ; \end{aligned}$$

474 For the first item, because the ranking of each positive entity is uniformly at random and note the
 475 ranking is iterated, we have $P(r = k | Y) = \frac{1}{m+1}$; $k = 1, 2, \dots, m+1$. Note m is a random
 476 variant following the binomial distribution $B(N; l)$, we have

$$\begin{aligned} E \frac{1}{f(r(e))} | Y &= E_m \frac{1}{m+1} \sum_{k=1}^{m+1} \frac{1}{f(k)} \\ &= \sum_{m=0}^N \binom{N}{m} (l)^m (1-l)^{N-m} \frac{1}{m+1} \left(\sum_{k=1}^{m+1} \frac{1}{f(k)} \right) \\ &= \sum_{k=1}^{N+1} \frac{1}{f(k)} \sum_{m=k-1}^N \binom{N}{m} (l)^m (1-l)^{N-m} \frac{1}{m+1} \\ &= \frac{1}{l(N+1)} \sum_{k=1}^{N+1} \frac{1}{f(k)} \sum_{m=k-1}^N \binom{N}{m} (l)^{m+1} (1-l)^{N-m} \\ &= \frac{1}{l(N+1)} \sum_{k=1}^{N+1} \frac{1}{f(k)} P(m \geq k) \\ &= \frac{1}{l(N+1)} \sum_{k=0}^N \frac{1}{f(k+1)} (1 - \hat{F}(k)); \end{aligned}$$

477 where \hat{F} is the cdf of binomial distribution $B(N+1; l)$. To prove this lemma, we only need to prove
 478 $0 < (1-l) E(\frac{1}{f(r(e))} | \bar{Y}) < (1-l) \frac{\ln(N_{entity} - N)}{N_{entity} - N}$. The left side is obvious, and the right side can be
 479 proved as follow. Here we use N_e to denote the number of entity instead of N_{entity} . Condition on
 480 m , note that the minimal ranking of negative entities is $m+1$ because there have been missing
 481 answers with higher rankings, and the maximal ranking of them is $(N-m)$ which is the
 482 number of entities except for the iterated ones. So we have

$$E \frac{1}{f(r(e))} | \bar{Y} = \frac{1}{N_e - N} E_m \sum_{k=m+1}^{N_e - N + m} \frac{1}{k} = E_m \frac{\ln(N_e - N + m) - \ln(m)}{N_e - N} = \frac{\ln(N_e - N)}{N_e - N} ;$$

483 The first inequality is because $\frac{1}{k} > \ln(k) - \ln(k-1)$. This proves the lemma. □

484 A.1.2 Corollary 4.1

485 Proof. We need the derivative of the cdf of binomial distribution. Assuming the cdf of $B(N; p)$,
 486 we have

$$\begin{aligned} \frac{d(\hat{K})}{dp} &= \sum_{k=0}^X \frac{d}{dp} \binom{N}{k} p^k (1-p)^{N-k} \\ &= \sum_{k=0}^X \binom{N}{k} k p^{k-1} (1-p)^{N-k} - \sum_{k=0}^X \binom{N}{k} (N-k) p^k (1-p)^{N-k-1} \\ &= \sum_{k=0}^X \binom{N}{k} k p^{k-1} (1-p)^{N-k} - \sum_{k=0}^X \binom{N}{k+1} (k+1) p^k (1-p)^{N-k-1} \\ &= \sum_{k=0}^X \binom{N}{k} k p^{k-1} (1-p)^{N-k} - \sum_{k=1}^{X+1} \binom{N}{k} k p^{k-1} (1-p)^{N-k} \\ &= \binom{N}{K+1} (K+1) p^K (1-p)^{N-K-1} \end{aligned}$$

487 So for \hat{K} is the cdf of $B(N+1; l)$, we have

$$\frac{d\hat{K}}{dl} = \frac{N+1}{k+1} (k+1) l^k (1-l)^{N-k}$$

488 and

$$\begin{aligned} \frac{d\hat{E}}{dl} &= \frac{1}{l(N+1)} \sum_{k=0}^X \frac{k+1}{f(k+1)} \binom{N+1}{k+1} l^{k+1} (1-l)^{N-k} \\ &= \frac{1}{l(N+1)} \sum_{k=1}^{X+1} \frac{k}{f(k)} \binom{N+1}{k} l^k (1-l)^{N+1-k} \\ &= \frac{1}{l(N+1)} E_{k \sim B(N+1;l)} g(k): \end{aligned}$$

489 The final equation is because $g(0) = 0$. □

490 A.1.3 Corollary 4.2

491 Proof. For MRR, $g(r) = 1; 8r \geq 2N_+$ and $g(0) = 0$. Just replace g into Corollary 4.1, we can get
 492 this corollary. □

493 A.1.4 Corollary 4.3

494 Proof. According to the Corollary 4.1, we have

$$\begin{aligned} \frac{d\hat{E}(\text{Hits}@K)}{dl} &= \frac{1}{N+1} \sum_{k=1}^X \binom{N+1}{k} l^k (1-l)^{N+1-k} \\ &= \sum_{k=1}^X \binom{N}{k-1} l^k (1-l)^{N-(k-1)} \\ &= (K-1); \end{aligned}$$

495 where \hat{K} is the cdf of the binomial distribution $B(N; l)$. □

496 A.1.5 Theorem 4.1

497 Proof. Let $E^0 = \hat{E} = \frac{1}{N+1} \sum_{k=0}^N \frac{1}{k+1} (1 - \hat{K})$ and $t = l$. In the same way in Corollary 4.1,
 498 we have

$$\frac{dE^0}{dt} = \frac{1}{t(N+1)} \sum_{k=0}^N \frac{N+1}{k+1} t^{k+1} (1-t)^{N-k} = \frac{1}{t(N+1)} (1-t)^{N+1}:$$

499 For $0 < t_0 < t < 1$, we have

$$\frac{1 - (1 - t_0)^{N+1}}{t(N+1)} \frac{dE^0}{dt} = \frac{1}{t(N+1)}:$$

500 Then we integrate them from t_0 to 1.

$$\frac{1 - (1 - t_0)^{N+1}}{N+1} \ln(t_0) = E^0_{j=t=1} - E^0_{j=t=t_0} = \frac{1}{N+1} \ln(t_0):$$

501 Because t_0 is arbitrary, we replace t_0 as general t .

$$\frac{1 - (1 - t)^{N+1}}{N+1} (\ln(t) + \ln(\frac{1}{t})) = \hat{E}_{j=1} - E^0 = \frac{1}{N+1} \ln(t) + \ln(\frac{1}{t}):$$

Note that

$$\hat{E}_{j=1} - E^0 = \hat{E}_{j=1} - E^0 = \frac{1}{N+1} \sum_{k=1}^{N+1} \frac{1}{k} = \frac{\ln(N+2) + \frac{1}{N+1}}{N+1}:$$

502 We denote it as E_1 , where $\frac{1}{N+1} \sum_{k=1}^{N+1} \frac{1}{k}$ is the residual of the sum of harmonic series and $0 < \frac{1}{N+1} < \frac{1}{2(N+1)}$. Then, we have

$$1 - (1 - t)^{N+1} (E_1 + \frac{1}{N+1} \ln(t)) = E_1 - E^0 (E_1 + \frac{1}{N+1} \ln(t)):$$

For the second inequality, it is equivalent to

$$\hat{E} - E = \frac{1}{N+1} - \frac{1}{2(N+1)^2}:$$

504 For the first inequality, we have

$$\begin{aligned} \hat{E} - E &= \frac{E_1}{N+1} - \hat{E} = 1 - \frac{1}{(1 - (1 - t)^{N+1})} - \frac{1}{N+1} \\ &= \frac{E_1}{N+1} + \frac{1}{N+1} - \frac{(1 - t)^{N+1}}{1 - (1 - t)^{N+1}} - \frac{1}{N+1} \\ &= \frac{(1 - t)^{N+1}}{1 - (1 - t)^{N+1}} - \frac{\ln(1 - (1 - t)^{N+1})}{(N+1)} - \frac{1}{N+1} \\ &= \frac{(1 - t)^{N+1}}{1 - (1 - t)^{N+1}} - \frac{\ln(1 - (1 - t)^{N+1})}{(N+1)}: \end{aligned}$$

505 Therefore, we have the error bound:

$$j\hat{E} - E_j \leq \max \left\{ \frac{1}{2(N+1)^2}, \frac{(1 - t)^{N+1}}{1 - (1 - t)^{N+1}} - \frac{\ln(1 - (1 - t)^{N+1})}{(N+1)} \right\}$$

506

□

507 A.1.6 Theorem 4.2

Proof. Given all the independence assumptions, $M(M_2) \sim M(M_1)$ follows normal distribution $N(\frac{\ln(1 + \frac{1}{N+1})}{(N+1)}, \frac{V(\cdot; 1) + V(\cdot; 1+1)}{N_q})$. So

$$Z = \frac{M(M_2) - M(M_1) \frac{\ln(1 + \frac{1}{N+1})}{(N+1)}}{\sqrt{\frac{V(\cdot; 1) + V(\cdot; 1+1)}{N_q}}} \sim N(0; 1):$$

508 Then $M(M_2) \sim M(M_1)$ is equivalent to

$$Z = \frac{M(M_2) - M(M_1) \frac{\ln(1 + \frac{1}{N+1})}{(N+1)}}{\sqrt{\frac{V(\cdot; 1) + V(\cdot; 1+1)}{N_q}}} = \frac{P \sqrt{N_q} \ln(1 + \frac{1}{N+1})}{(N+1) \sqrt{V(\cdot; 1) + V(\cdot; 1+1)}}:$$

509 So the probability is as shown in the theorem.

□

510 A.1.7 Corollary 4.4

511 Proof. Just solve N_q from the Theorem 4.2. □

512 A.1.8 Theorem 4.3

513 Proof. We can generalize the Lemma 4.1 as follows.

514 Lemma A.1 (Expectation with Correlation) Under the same assumptions as the lemma 4.1 and the
 515 correlation coefficient is ρ , the expectation of the metric M :

$$E(M) = \frac{l_2}{l_1} \frac{1}{(N+1)} \sum_{k=0}^N \frac{1}{f(k+1)} (1 - \rho)^k + \rho; \quad (9)$$

516 where f is the cdf of binomial distribution $B(N+1; l_1)$ and $0 < \rho < (1 - l_2) \frac{\ln(N_{entity} - N)}{N_{entity}}$.

517 The proof of the lemma is similar to what we have shown in A.1.1. Given the lemma, it can be
 518 similarly expressed as $\frac{1}{l_1} \frac{1}{(N+1)} \sum_{k=0}^N \frac{1}{f(k+1)} (1 - \rho)^k$.

519 In the similar way in A.1.5, let $E^0 = l_1 \hat{E}$ and $t = l_1$ we have

$$1 - (1 - l_1)^{N+1} (E_1 + \frac{l_2^{N+1}}{N+1} - l_1 E) = E_1 - E^0 (E_1 + \frac{l_2^{N+1}}{N+1} - l_1 E):$$

520 where $E_1 = \frac{l_2(\ln(N+2) + \frac{1}{N+1})}{N+1} = E_{j|l_1=1} = \hat{E}_{j|l_1=1}$. Also using the same technique, the error
 521 bound is

$$\hat{E} - E = \frac{l_2}{l_1 (N+1)^2}$$

522 and

$$\begin{aligned} \hat{E} - E &= \frac{E_1 - \hat{E}}{l_1} = \frac{1}{l_1} \frac{1}{(1 - (1 - l_1)^{N+1})} \\ &= \frac{E_1 + l_2^{N+1} - l_1 E}{l_1} = \frac{(1 - l_1)^{N+1}}{1 - (1 - l_1)^{N+1}} \\ &= \frac{(1 - l_1)^{N+1}}{1 - (1 - l_1)^{N+1}} \frac{l_2 \ln(1 - (1 - l_1))}{l_1 (N+1)}. \end{aligned}$$

523 The error bound is that

$$j \hat{E} - E_j \leq \max \left\{ \frac{l_2}{l_1 (N+1)^2}, \frac{(1 - l_1)^{N+1}}{1 - (1 - l_1)^{N+1}} \frac{l_2 \ln(1 - (1 - l_1))}{l_1 (N+1)} \right\}$$

524 □

525 A.1.9 Corollary 4.5

526 Proof. Let $(N+1) = c$ and $\ln(N+2) + \frac{1}{N+1} = d$, we have

$$\frac{\partial E}{\partial l_1} = \frac{l_2(1 - \ln l_1 - d)}{c l_1^2}; \quad \frac{\partial E}{\partial l_2} = \frac{\ln(l_1) + d}{c l_1};$$

527 and

$$\frac{\partial \hat{E}}{\partial l_1} = \frac{s}{l_1(1 - l_1)}; \quad \frac{\partial \hat{E}}{\partial l_2} = \frac{r}{l_1(1 - l_1)}.$$

528 So the derivative w.r.t

$$\begin{aligned} \frac{\partial E}{\partial l_1} &= \frac{p}{l_1(1 - l_1)} l_2(1 - \ln l_1 - d) - \frac{r}{l_1(\ln(l_1) + d)} \\ &= \frac{p}{c l_1^2} l_2(1 - \ln(l_1) + d) - \frac{r}{l_1} \end{aligned}$$

529 Because of the conditions $(N + 2) \exp\left(\frac{q}{1 - l}\right)$, we have

$$\ln(l_1) + d + \frac{r}{1 - l}$$

530 and then

$$\frac{1}{1 - (\ln(l_1) + d)} \left(1 + \frac{r}{1 - l}\right) > 1 + \frac{r}{1 - l} = l_2:$$

531 Combining this inequality with the derivative expression, we have $\frac{\partial F}{\partial l} < 0$. □

532 A.2 Details of the simulation

533 In the Figures 1 and 2, we choose $N_{\text{ent}} = 14505 \times 30\% = 4351$, where we assume $N_{\text{entity}} = 14505$
 534 as the same as FB15k-237, the total answers accounts for one percent of all entities and the test set
 535 accounts for thirty percent of the total answers. For each d we repeat the simulation with 500
 536 times to calculate the average MRR and the standard derivation.

537 A.3 Details of artificial family tree KG

538 Our codes are modified from [Hohenecker and Lukasiewicz, 2020] (BSD license) to generate the
 539 KG. Firstly, it generates all the parent-child relations and then deduced other relations by a symbolic
 540 reasoning systems called DLV system [Leone et al., 2006]. We generate 20 family trees then merge
 541 them into a whole. Each family tree has three layer depth and 300 entities, with maximal branching
 542 width 20 at each internal node. The final artificial KG has 6,004 entities, 23 relations and 192,532
 543 facts. The relations are listed as follow:

- | | | | |
|-----------------|---------------------|--------------------|-----------------------|
| 544 • parentOf | 550 • wifeOf | 556 • girlCousinOf | 562 • granddaughterOf |
| 545 • sisterOf | 551 • husbandOf | 557 • boyCousinOf | 563 • grandsonOf |
| 546 • brotherOf | 552 • grandmotherOf | 558 • cousinOf | 564 • grandchildOf |
| 547 • siblingOf | 553 • grandfatherOf | 559 • daughterOf | 565 • nieceOf |
| 548 • motherOf | 554 • auntOf | 560 • sonOf | 566 • nephewOf |
| 549 • fatherOf | 555 • uncleOf | 561 • childOf | |

567 Note that $G_{\text{full}} = G_{\text{test}} \cup G_{\text{train}}$. We use density to denote the ratio $\rho_j = \frac{|G_{\text{test}} \cap j|}{|G_{\text{full}} \cap j|}$ and then
 568 in the open-world KG G_{test} we split the training set and test set with ratio $\rho_j = \frac{|G_{\text{train}} \cap j|}{|G_{\text{test}} \cap j|}$.
 569 For each facts, it is a missing fact with probability d , a test fact with probability $(1 - d)\rho_j$ and a
 570 training fact with probability $(1 - d)(1 - \rho_j)$. We set $\rho_j = 0.7$ and $d = 95\%, 85\%, 75\%, 65\%$ which corresponds
 571 to $\rho_j = \frac{|G_{\text{test}} \cap j|}{|G_{\text{full}} \cap j|} = \frac{d(1 - \rho_j)}{1 - d} = 85\%; 63\%; 47\%; 35\%$

572 As the same as [Ren et al., 2020, Ren and Leskovec, 2020], we organize the test by queries which
 573 means we firstly randomly sample the test queries $(s, ?)$ and then search answers $G_{\text{full}} \cap G_{\text{test}}$
 574 as missing answers and $G_{\text{test}} \cap G_{\text{train}}$ as test answers. In order to simulate the real situation of
 575 common KGs, we filter out the queries with less than 10 answers in the closed-world graph
 576 Finally, for each sparsity d we choose 500 test queries. For training, we use all facts in G_{train} .

577 A.4 Details of models

578 Different models are trained on artificial family tree KG. During training, we test them on full test
 579 set and sparse test set to plot the sparse-full curve to show the inconsistency. We choose different
 580 framework, including RotatE, pRotatE [Sun et al., 2019], ComplEx [Trouillon et al., 2016] and BetaE
 581 [Ren and Leskovec, 2020]. We also test Q2B [Ren et al., 2020] and TransE [Bordes et al., 2013]
 582 models, both of which cannot fit the KG well. For each framework, we use several settings, where
 583 their label in Figure 3 and the hyper-parameters of the models are shown in Table 2. Here, we have
 584 filtered some models which maximal strength ≤ 0.1 .

Table 2: Detail of the models trained on family tree KG.

label	model	dimension	gamma	step	batchsize	negative sampling
0	RotatE	1000	24	100000	1024	128
1	RotatE	500	12	100000	256	128
2	RotatE	500	12	100000	1024	512
3	RotatE	500	24	100000	1024	128
4	RotatE	1000	24	100000	1024	128
5	pRotatE	1000	24	12000	1024	128
6	pRotatE	250	24	12000	1024	128
7	pRotatE	500	24	12000	1024	128
8	pRotatE	500	24	12000	128	512
9	pRotatE	500	6	12000	1024	128
10	BetaE	1000	60	400000	1024	128
11	BetaE	500	240	400000	1024	128
12	BetaE	500	60	400000	1024	128
13	BetaE	500	15	400000	1024	128
14	BetaE	100	60	400000	1024	128
15	ComplEx	1000	500	100000	1024	128
16	ComplEx	1000	200	100000	512	256
17	ComplEx	2000	500	100000	1024	128

Figure 6: Full test and sparse test MRR on independent (above) and correlated (bottom) family tree KG. Density = 95%; 85%; 65% from left to right.

585 A.5 Experiments with correlation

586 Here we show more results of the experiments on the correlated family tree KG in Figure 6.

587 A.6 Family tree experiments with MRR, Hits@K and more less focus-on-top metrics

588 The results for other density and more metrics in the independent situation are shown in Figures 7-10.

589 And the results in the correlated situation are shown in Figures 11-14.

Figure 7:d = 95% independent

Figure 8:d = 85% independent

Figure 9:d = 75% independent

Figure 10:d = 65% independent

Figure 11:d = 95% correlated

Figure 12:d = 85% correlated

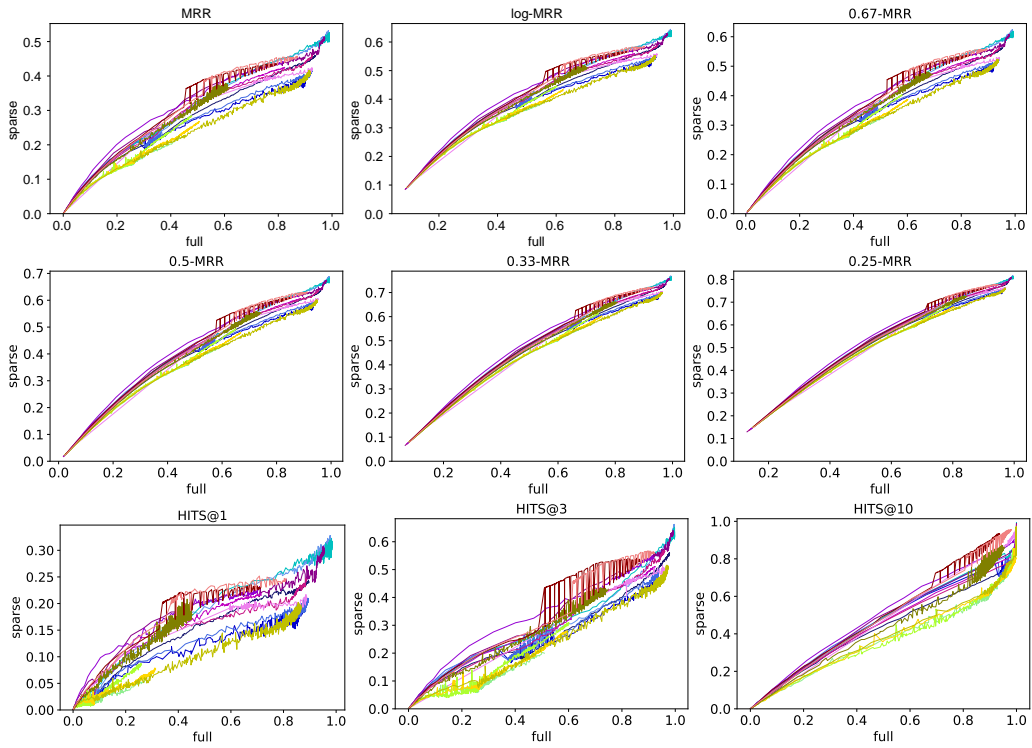


Figure 13: $d = 75\%$ correlated

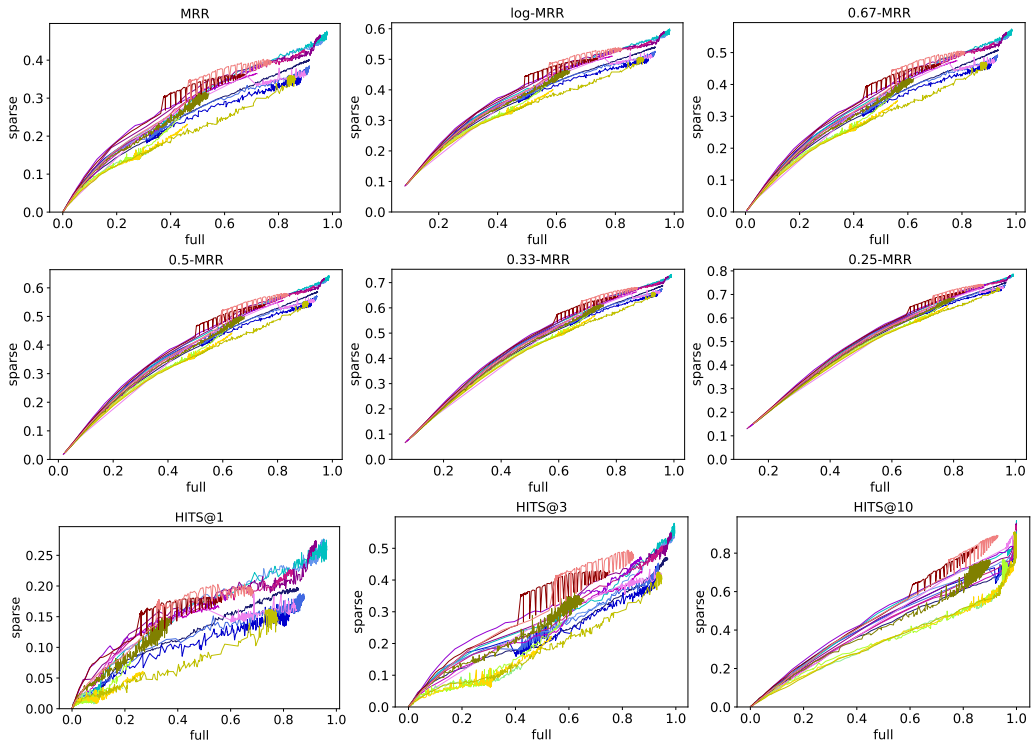


Figure 14: $d = 65\%$ correlated