

Towards a Unified Evaluation Protocol for Link Prediction Interpretation over Knowledge Graphs

Beshani Weralupitiya
bw4052@rit.edu
Rochester Institute of Technology
Rochester, NY, USA

Abstract

Link prediction is a critical task in knowledge graph applications, enabling the inference of missing information within graph-structured data. However, existing evaluation protocols for link prediction are computationally expensive, lack statistical soundness, and suffer from various challenges, such as limited interpretability, inconsistent graph splitting methods, leading to unreliable model assessments. This research aims to establish new link prediction interpretation methods for knowledge graphs that are robust and efficient with statistical guarantees. It also aims to create a unified evaluation protocol that integrates variations like subgraph selection, graph splitting, and redundancy reduction. As the initial step, we have developed a knowledge graph splitting mechanism, which we present in this paper. Our research focuses on designing a robust evaluation framework, scalable rule-mining techniques, and reliable statistical estimators to enhance the interpretability of link prediction models.

CCS Concepts

• **Computing methodologies** → **Semantic Networks**; • **Information Systems** → *Data mining*.

Keywords

Knowledge Graphs, Graph Splitting

ACM Reference Format:

Beshani Weralupitiya. 2025. Towards a Unified Evaluation Protocol for Link Prediction Interpretation over Knowledge Graphs. In *Proceedings of* . ACM, New York, NY, USA, 3 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

A knowledge graph is a structured representation of information from the real world, comprising entities of interest and their relationships [12]. Knowledge graphs are at the core of many services, including search engines, social networks, product catalogs, and large language models [7, 15, 18, 26]. A knowledge graph comprises a set of triples of the form (s, p, o) , where s and o are entities connected via a predicate p . This can be seen as a directed edge from s to o with label p . Despite their wide adoption, most knowledge

graphs remain incomplete, since it is hard to capture all the relationships among entities [5]. For example, there is no birthplace information for 71% out of around 3 million people in Freebase [25]. To address this incompleteness, link prediction has emerged as a crucial task to complete knowledge graphs. It infers missing triples using models trained on existing data [3]. These models utilize knowledge graph embeddings to encode entities and predicates into a continuous vector space, which enables the prediction of new triples. [10]. However, link prediction interpretation is challenging due to the sub-symbolic nature of the knowledge graph embeddings that link prediction relies on [2–4, 22].

One approach to enhance the interpretability of link prediction over knowledge graphs is through rule-based methods, where Horn rules are mined from model predictions. Krishnan and Rivero [13] proposed a model-agnostic method that extracts rules from triples deemed plausible during link prediction evaluation. The method focuses on rule quality measurements like support and confidence to accomplish the interpretation. Although this approach helps interpret diverse knowledge graph embedding architectures, it still relies on rule mining and rule quality measurements. Rule mining is computationally expensive, as models may generate a large number of deemed-plausible triples. Existing rule mining techniques like AMIE [9] use an exact and exhaustive process that employs a breadth-first search strategy, making it computationally expensive. Although AMIE mitigates this issue with heuristic approximations for rule quality estimation, these approximations are limited to path-based rules and tied to the concept of predicate functionality, limiting their generalizability. Alternatively, AnyBURL [17] adopts a random exploratory mining, where it generates candidate rules via random walks over the knowledge graph at hand. However, this method lacks a clear termination criterion, making it challenging to ensure the completeness of mined rules. Furthermore, the approximate measurements used for rule quality lack statistical soundness. As a result, more efficient link prediction interpretation methods with statistical guarantees are needed.

Another challenge lies in the link prediction evaluation protocol. Variations in the protocol, such as subgraph selection, graph splitting, or redundancy reduction, have significant impacts on reported accuracy and introduce biases in experiments [20, 23]. Focusing on graph partitioning, random splitting fails to preserve the topological and semantic properties of the graph at hand, introducing biases that affect model performance and interpretability [1, 6]. Limited research has addressed this, with Tiwari et al. [23] proposing a degree-based splitting approach, leveraging statistical similarity between in-degrees and out-degrees of the graph to ensure topological consistency. However, this method is sensitive to the processing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

, Toronto, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

order of the triples as there is no measurement indicating the triples that should be moved to the test split.

The main goal of my PhD research is to enhance the entire link prediction evaluation protocol for knowledge graphs, addressing the above critical challenges. Despite the substantial impact of these factors on model evaluation, they remain underexplored in the context of knowledge graphs, unlike other domains like computer vision and natural language processing [19, 24]. Ambiguous evaluation protocols can lead to misleading conclusions, obscuring true model performance and undermining interpretability. To bridge this gap, we propose a unified evaluation protocol for link prediction interpretability. These research goals are further elaborated in Sections 2, 3, and 4.

2 Research Agenda

To address our overarching research goal, the following research questions are proposed:

- **RQ1:** Can we design a knowledge graph splitting strategy that preserves topological and semantic properties across the original graph and the training and test splits?
- **RQ2:** Can we develop a scalable rule-mining technique that efficiently identifies meaningful rules without excessive computational costs and expected termination criteria?
- **RQ3:** Can we establish robust statistical estimators to reliably measure rule quality while maintaining computational efficiency?
- **RQ4:** Can we integrate the above outcomes to enhance both predictive performance and interpretability of link prediction models over knowledge graphs?

In Section 3, we present our current research efforts that aim to address RQ1.

3 Knowledge Graph Splitting Mechanism

Given a knowledge graph G , we represent it as (E, R, T_G) , where E is the set of entities, R is the set of predicates, and $T_G \subseteq E \times R \times E$ is the set of triples. We aim to generate training and test splits T_x with $x \in \{tr, te\}$ such that $T_{tr} \cup T_{te} = T_G$. Initially, $T_{tr} = T_G$ and $T_{te} = \emptyset$. At each step, we transfer one triple $(s, p, o) \in T_{tr}$ to T_{te} , modeled as a multi-armed bandit problem [21]. The triple (s, p, o) is chosen as follows:

$$\begin{cases} (s, p, o) \in T_{tr} & \text{with probability } \rho \\ \arg \min_{(s,p,o) \in T_{tr}} f & \text{with probability } 1 - \rho \end{cases} \quad (1)$$

where ρ is the probability of selecting (s, p, o) randomly and f is the reward function defined below. Assuming that (s, p, o) is selected, we define $T'_{tr} = T_{tr} \setminus \{(s, p, o)\}$ and $T'_{te} = T_{te} \cup \{(s, p, o)\}$. After (s, p, o) is selected, we update the sets of triples as follows: $T_{tr} = T'_{tr}$ and $T_{te} = T'_{te}$. Thus, step $i + 1$ starts with the training and test splits computed in step i . We continue the process until certain stopping criteria are met. One of the possible stopping criteria is the size of the test split of the datasets that are publicly available and commonly used to evaluate link prediction [11].

To assess the structural properties, we define a degree tensor $\mathbf{D}(T_y) \in \mathbb{N}^{\delta \times m \times n}$, where $T_y \in \{T_G, T_x\}$, $\delta \in \{v^+, v^-\}$ encodes in-degree and outdegree directions respectively, $m = |\mathcal{P}|$ corresponds to the size of the extended predicate index set $\mathcal{P} := R \cup \{*\}$, and $n = |E|$. The special symbol $*$ represents a non-predicate dimension used to aggregate over all predicates.

The individual measures are provided below. We use $\#$ to denote the size of a set.

- $D(T_y)[v^+, p, e] = \#\{(_, p, e) \in T_y\}$
- $D(T_y)[v^-, p, e] = \#\{(e, p, _) \in T_y\}$
- $D(T_y)[v^+, *, e] = \#\{(_, _, e) \in T_y\}$
- $D(T_y)[v^-, *, e] = \#\{(e, _, _) \in T_y\}$

The reward function penalizes structural inconsistencies via constraint satisfaction framework:

$$f = \begin{cases} H(\mathcal{E}) & \text{if } KS(d_{v,p}^{tr}, d_{v,p}^G) \\ \infty & \text{Otherwise} \end{cases} \quad (2)$$

where KS is the two-sample Kolmogorov–Smirnov test [14, 16], used to assess whether the degree distributions $d_{v,p}^y = D(T_y)[v, p, \cdot]$ for $v \in \delta, p \in \mathcal{P}$ significantly diverge between T_x and T_G . If the test fails, then $f = \infty$. Otherwise, f is computed using the harmonic mean [8], H over an error tensor $\mathcal{E} \in \mathbb{R}^{\delta \times |x| \times m}$. This aggregates all the deviations captured in the error tensor \mathcal{E} . This is defined as:

$$\mathcal{E}[v, t, p] = |D(T_G)[v, p, \cdot] - D(T_x)[v, p, \cdot]|_1 \quad (3)$$

where each entry of \mathcal{E} represents the ℓ_1 -norm difference in structure between T_G and T_x across direction v , predicate p , and split t .

The Kolmogorov–Smirnov test is defined as follows:

$$KS(d_{v,p}^{tr}, d_{v,p}^G) = \sup_{\theta} |\Delta(d_{v,p}^{tr}, \theta) - \Delta(d_{v,p}^G, \theta)| \leq c(\alpha) \quad (4)$$

where $\Delta(d, \theta)$ is the empirical CDF of d evaluated at θ , and $c(\alpha)$ is the critical value at confidence level α :

$$\Delta(d, \theta) = \frac{1}{|E|} \sum_{e \in E} \mathbb{I}[D(T_y)[v, p, \cdot] \leq \theta] \quad (5)$$

4 Conclusion and Future Work

In this paper, we introduce a statistically grounded graph splitting mechanism that preserves both topological and semantic properties, making significant progress toward addressing RQ1 and overcoming a key limitation in existing link prediction evaluation protocols. Moving forward, we will focus on further refining this work and expanding it to tackle RQ2 and RQ3, including developing scalable rule-mining techniques and robust statistical estimators for rule quality while maintaining efficient computational costs and termination criteria. Ultimately, these efforts will culminate in a unified protocol for more reliable and interpretable link prediction evaluation (RQ4).

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. 2346959.

References

- [1] Dan Archdeacon. 1996. Topological graph theory. *A survey. Congressus Numerantium* 115, 5-54 (1996), 18.
- [2] Federico Bianchi, Gaetano Rossiello, Luca Costabello, Matteo Palmonari, and Pasquale Minervini. 2020. Knowledge Graph Embeddings and Explainable AI. *CoRR* abs/2004.14843 (2020). arXiv:2004.14843 <https://arxiv.org/abs/2004.14843>
- [3] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.), 2787–2795. <https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html>
- [4] Guillaume Bouchard, Sameer Singh, and Théo Trouillon. 2015. On Approximate Reasoning Capabilities of Low-Rank Vector Spaces. In *2015 AAAI Spring Symposium, Stanford University, Palo Alto, California, USA, March 22-25, 2015*. AAAI Press. <http://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10257>
- [5] Zhe Chen, Yuehan Wang, Bin Zhao, Jing Cheng, Xin Zhao, and Zongtao Duan. 2020. Knowledge Graph Completion: A Review. *IEEE Access* 8 (2020), 192435–192456. doi:10.1109/ACCESS.2020.3030076
- [6] Gerrit Jan de Bruin, Cor J. Veenman, H. Jaap van den Herik, and Frank W. Takes. 2020. Experimental Evaluation of Train and Test Split Strategies in Link Prediction. In *Complex Networks & Their Applications IX - Volume 2, Proceedings of the Ninth International Conference on Complex Networks and Their Applications, COMPLEX NETWORKS 2020, 1-3 December 2020, Madrid, Spain (Studies in Computational Intelligence, Vol. 944)*. Springer, 79–91. doi:10.1007/978-3-030-65351-4_7
- [7] Xin Luna Dong. 2019. Building a Broad Knowledge Graph for Products. In *ICDE*. 25.
- [8] Wirth F Ferger. 1931. The nature and use of the harmonic mean. *J. Amer. Statist. Assoc.* 26, 173 (1931), 36–40.
- [9] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M. Suchanek. 2015. Fast rule mining in ontological knowledge bases with AMIE+. *VLDB J.* 24, 6 (2015), 707–730.
- [10] Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. 2016. Jointly Embedding Knowledge Graphs and Logical Rules. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 192–202. doi:10.18653/v1/D16-1019
- [11] Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. OpenKE: An Open Toolkit for Knowledge Embedding. In *Proceedings of EMNLP*.
- [12] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2022. Knowledge Graphs. *ACM Comput. Surv.* 54, 4 (2022), 71:1–71:37. doi:10.1145/3447772
- [13] Narayanan Asuri Krishnan and Carlos R. Rivero. 2024. A Method for Assessing Inference Patterns Captured by Embedding Models in Knowledge Graphs. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee (Eds.). ACM, 2030–2041. doi:10.1145/3589334.3645505
- [14] Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven, and Dimitrios Gunopulos (Eds.). ACM, 631–636. doi:10.1145/1150402.1150479
- [15] Lihui Liu and Hanghang Tong. 2023. Knowledge Graph Reasoning and Its Applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Long Beach, CA, USA) (KDD ’23)*. Association for Computing Machinery, New York, NY, USA, 5813–5814. doi:10.1145/3580305.3599564
- [16] Frank J. Massey. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Amer. Statist. Assoc.* 46, 253 (1951), 68–78. <http://www.jstor.org/stable/2280095>
- [17] Christian Meilicke, Melisachew Wudage Chekol, Patrick Betz, Manuel Fink, and Heiner Stuckenschmidt. 2024. Anytime bottom-up rule learning for large-scale knowledge graph completion. *VLDB J.* 33, 1 (2024), 131–161.
- [18] Natalya Fridman Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-scale knowledge graphs: lessons and challenges. *Comm. ACM* 62, 8 (2019), 36–43.
- [19] ChaeHun Park, Eugene Jang, Wonsuk Yang, and Jong Park. 2021. Generating Negative Samples by Manipulating Golden Responses for Unsupervised Learning of a Response Evaluation Model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 1525–1534. doi:10.18653/v1/2021.naacl-main.120
- [20] Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. 2021. Knowledge Graph Embedding for Link Prediction: A Comparative Analysis. *ACM Trans. Knowl. Discov. Data* 15, 2 (2021), 14:1–14:49. doi:10.1145/3424672
- [21] Aleksandrs Slivkins. 2019. Introduction to Multi-Armed Bandits. *CoRR* abs/1904.07272 (2019). arXiv:1904.07272 <http://arxiv.org/abs/1904.07272>
- [22] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. <https://openreview.net/forum?id=HkgEQnRqYQ>
- [23] Sudhanshu Tiwari, Iti Bansal, and Carlos R. Rivero. 2021. Revisiting the Evaluation Protocol of Knowledge Graph Completion Methods for Link Prediction. In *WWW ’21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 809–820. doi:10.1145/3442381.3449856
- [24] Bharadwaj Veeravalli, Xiaolin Li, and Chi Chung Ko. 2000. Efficient partitioning and scheduling of computer vision and image processing data on bus networks using divisible load analysis. *Image Vis. Comput.* 18, 11 (2000), 919–938. doi:10.1016/S0262-8856(99)00085-2
- [25] Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge base completion via search-based question answering. In *23rd International World Wide Web Conference, WWW ’14, Seoul, Republic of Korea, April 7-11, 2014*, Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel (Eds.). ACM, 515–526. doi:10.1145/2566486.2568032
- [26] Zhenhao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024. Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. In *SIGIR*. 2905–2909.