Proceedings of Machine Learning Research 266:1–20, 2025 Conformal and Probabilistic Prediction with Applications

Conformal multi-hop relation detection and classification in knowledge graphs

Frederick Law LAW16@LLNL.GOV
Lawrence Livermore National Laboratory
Livermore, CA 94550, USA

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

Knowledge graphs (KGs) have seen an increasing use in application domains where information may be deemed proprietary, protected, or sensitive, such as enterprise, medical, or security applications. For such systems, incorporating uncertainty quantification (UQ) is critically necessary when KG information is passed to others for any downstream usage. Moreover, such systems often have constraints on data availability due to safety or legal restrictions, and as such full access to well-labeled training data may be unavailable. Conformal prediction is a distribution-free UQ strategy which is well-equipped to handle both of these concerns, as it produces prediction sets with statistically valid guarantees and is highly compatible with black-box models, which may be shared more easily than training data. In this work, we develop a novel conformal framework for simultaneously detecting and classifying multi-hop relations between entities in a KG, which only assumes access to a pre-trained KG model over triples and does not require multi-hop training data. Our framework utilizes a greedy approach, wherein we use successive conformal predictors to build a sparsely-supported scoring function in the high-dimensional multi-hop relation space. In numerical experiments on publicly available benchmark KGs with variable size and multi-hop length, our conformal multi-hop relation sets offer substantial reduction relative to the multi-hop relation space.

Keywords: Conformal prediction, uncertainty quantification, knowledge graph, multi-hop

1. Introduction

Knowledge graphs (KGs) represent large-scale, heterogeneous graph structured data, comprised of semantic information of entities and the relationships between them. Such facts are represented by triples (h, r, t) where h and t are the respective head and tail entities, and r is the relation type between them. While originally developed for linguistic and data-science focused applications such as question-answering, in recent years KGs have found application in a wide variety of fields including biomedical engineering, cybersecurity, and education (Zou, 2020). A defining characteristic of KGs is that they are typically incomplete, as not all information can be readily verified or included due to scale and complexity. Consequently, a primary goal with KGs has been KG completion (Shen et al., 2022), where the goal is to infer new triples (h, r, t) based on existing data. This is commonly done by embedding the KGs using a graph neural network (GNN) to capture the relevant features of the KG for inference, see (Wang et al., 2017) for an overview on KG embedding techniques.

For KG systems in which reliability and trustworthiness are paramount, the incorporation of uncertainty quantification (UQ) is necessary for any further downstream analysis.

Moreover, due to the complex nature of KGs, in scale, heterogeneity, and incompleteness, we seek a distribution-free approach using conformal prediction, which offers valid statistical guarantees and is flexible for a variety of KG queries. One approach to incorporate UQ into KGs has been through the GNN embedding itself, an area for which the literature is rich, see (Wang et al., 2024) for a comprehensive review. However, the use of conformal prediction in GNN is still an emerging area. One work that has incorporated conformal prediction with GNNs is (Huang et al., 2023) in which provided a base GNN, they train an additional correction GNN by simulating the conformal prediction process and minimizing a conformal-aware loss. A different direction has been taken in (H. Zargarbashi et al., 2023) by introducing diffused adaptive prediction sets, which leverages the graph structure to smooth nonconformity values.

Another approach to incorporating UQ for KGs has been to directly address the KG context. This is typically done by augmenting the existing KG with verified factual information (Bahaj and Ghogho, 2025), or modifying the choice of embedding (Chen et al., 2019). Work has been done incorporating conformal prediction for KG as in (Ni, 2024) which leverages the Learn-Then-Test framework for multi-hop reasoning on systems which combine KGs with large language models (LLMs). A recent work which leverages conformal prediction outside of the GNN embedding is (Zhu et al., 2025) which utilizes a split conformal prediction approach for link prediction in KGs, with various nonconformity measures.

In this work, we focus on incorporating conformal prediction into the KG task of simultaneously detecting and classifying multi-hop relationships between existing KG entities. For instance, in a KG comprised of people we may be interested in inferring long-range connections between seemingly disparate entities, e.g. is person A a friend of a friend of person B. We are interested in the non-intrusive case where we do not have direct access to the training data, but only a pre-trained base model f which scores KG triples (h, r, t), as well as limited calibration and testing data. This approach is heavily focused on portability, for use in cases where training data is potentially proprietary, sensitive, or otherwise inaccessible. Our contribution in this work is two-fold:

1. Introduce a multi-label regularized adaptive prediction sets (MLRAPS) method which slightly modifies the existing regularized adaptive prediction sets (RAPS).

2. Develop a greedily-constructed scoring function, only requiring a base model $f(h, r, t)$ and calibration data, for use with MLRAPS to produce conformal multi-hop relation sets to both detect and classify multi-hop relations.

In Section 2 we first outline our simultaneous detection and classification framework with conformal prediction using RAPS on the simpler single-hop case. We then extend this to the multi-hop case in Section 3 which requires both multi-label conformal classification as well as the introduction of our greedily-constructed scoring function. In Section 4 we demonstrate our method on benchmark KGs, with experimental results showing smaller conformal sets and greater efficiency than a more naive choice of scoring function.

2. Single-hop relation detection and classification

We will first detail the single-hop relation detection and classification case before extending to the multi-hop case. We consider a KG to be a set of triples (or triplets) $(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$

where $\mathcal{E} = \{e^{(1)}, \ldots, e^{(N_e)}\}$ is a set of N_e entities and $\mathcal{R} = \{r^{(1)}, \ldots, r^{(N_r)}\}$ is a set of N_r relation types. We assume that we have access to pre-trained, base scoring model $f : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$, most commonly a KG embedding trained on a dataset $\mathcal{D}_{\text{train}}$. However, we do not assume that we have direct access to $\mathcal{D}_{\text{train}}$.

For a given head, tail pair (h, t), we are interested in detecting whether there exists a latent relationship between h and t, as well as classifying the relation type if one exists. In order to simultaneously detect and classify relation types, we expand the relation space by introducing a unique "no relationship" or "NoRel" class r_{NoRel}. Triples of the form $(h, r_{\text{NoRel}}, t) \in \mathcal{E} \times \mathcal{R}' \times \mathcal{E}$, where $\mathcal{R}' := \mathcal{R} \cup \{r_{\text{NoRel}}\}$, indicate that (h, r, t) is false for all $r \in \mathcal{R}$. We will refer to triples (h, r, t) for $r \in \mathcal{R}$ which contain real relations as true triples.

Our objective is to produce a conformal prediction set of relation types $C_{1-\alpha}(h, t) \subset \mathcal{R}'$ where α is the significance level. For the single hop case, we will use RAPS, as introduced in (Angelopoulos et al., 2021), as our conformal prediction method, which computes a nonconformity value as a regularized cumulative likelihood over scores, see Algorithm 1. This will require a scoring function $\boldsymbol{G_0}(h, t) : \mathcal{E} \times \mathcal{E} \to [0, 1]^{N_r+1}$ which provides a score for each $r \in \mathcal{R}$ as well as r_{NoRel}.

Algorithm 1: Regularized adaptive prediction sets (RAPS)

Input: Significance level $\alpha \in [0, 1]$, scoring function $\tilde{\boldsymbol{g}} : \mathcal{X} \to [0, 1]^K$, calibration data $D_{\text{calib}} = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \{1, \ldots, K\}$, test point $(x, y) \in \mathcal{X} \times \{1, \ldots, K\}$ exchangeable with D_{calib}, regularization hyperparameters $\beta > 0$ and $\lambda > 0$

Output: Conformal set $C_{1-\alpha}(x) \subset \{1, \ldots, K\}$ with $\mathbb{P}[y \in C_{1-\alpha}(x)] \geq 1 - \alpha$

for $i \in \{1, \ldots, N\}$ **do**

  $\boldsymbol{S}^i \leftarrow \pi_i(\tilde{\boldsymbol{g}}(x_i))$; ▷ π_i permutes $g(x_i)$ descending

  $R_i \leftarrow \sum_{j=1}^{z_i} \boldsymbol{S}^i_j + \lambda \cdot \max(0, z_i - \beta)$; ▷ $z_i = \pi_i(y_i)$, same π_i as above

end

$\hat{q} \leftarrow \lceil (1 - \alpha)(1 + N) \rceil$-th smallest R_i;

$\boldsymbol{S} \leftarrow \pi_x(\tilde{\boldsymbol{g}}(x))$; ▷ π_x permutes $g(x)$ descending

$C_{1-\alpha}(x) \leftarrow \{y \mid \sum_{j=1}^{\pi_x(y)} \boldsymbol{S}_j + \lambda \cdot \max(0, \pi_x(y) - \beta) \leq \hat{q}\}$

We assume that we have access to dataset $\mathcal{D}_{\text{calib-real}} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ which contains a set of true calibration triples, as well as $\mathcal{D}_{\text{NoRel}} \subset \mathcal{E} \times \{r_{\text{NoRel}}\} \times \mathcal{E}$ which contains a set of known "NoRel" triples. Furthermore, we assume we have calibration and testing sets $\mathcal{D}_{\text{calib-NR}}, \mathcal{D}_{\text{test-NR}} \subset \mathcal{E} \times \mathcal{R}' \times \mathcal{E}$ which contain a mixture of true and "NoRel" triples, and whose elements are exchangeable. If such datasets are unavailable, they can be constructed with sufficient 1-hop data, see Remark 1.

To first score relations for true triples, we define $\boldsymbol{g_0} : \mathcal{E} \times \mathcal{E} \to \mathbb{R}^{N_r}$ by

$$\boldsymbol{g_0}(h, t) = \begin{pmatrix} f(h, r^{(1)}, t) \\ \vdots \\ f(h, r^{(N_r)}, t) \end{pmatrix} \tag{1}$$

We then evaluate $\boldsymbol{g_0}$ on elements from $\mathcal{D}_{\text{calib-real}}$ and $\mathcal{D}_{\text{NoRel}}$. Using the labeled training-target pairs $\{(\boldsymbol{g_0}(h, t), 1) \mid (h, r_{\text{NoRel}}, t) \in \mathcal{D}_{\text{NoRel}}\} \cup \{(\boldsymbol{g_0}(h, t), 0) \mid (h, r, t) \in \mathcal{D}_{\text{calib-real}}\}$, we

fit a probabilistic binary classifier $f_{\text{class}} : \mathbb{R}^{N_r} \to [0,1]$, where small values of $f_{\text{class}}(\boldsymbol{g_0}(h,t))$ imply a latent relationship exists between h and t, and large values imply a lack of relation.

Ideally, if there is no latent relation between h and t, then the base model $f(h,r,t)$ should be small over all r and $\boldsymbol{g_0}(h,t)$ will be flat. Conversely, if a latent relationship r exists, then $f(h,r,t)$ should have a strong signal for this r and $\boldsymbol{g_0}(h,t)$ will be peaked.

To assign a score to the "NoRel" class, we scale the outputs of f_{class} based on the feature vector $\boldsymbol{g_0}(h,t)$ itself to define a final scoring function $\boldsymbol{G_0} : \mathcal{E} \times \mathcal{E} \to [0,1]^{N_r+1}$ as

$$\boldsymbol{G_0}(h,t) = \boldsymbol{\sigma}\left(\begin{array}{c} \boldsymbol{g_0}(h,t) \\ \frac{f_{\text{class}}(\boldsymbol{g_0}(h,t))}{1-f_{\text{class}}(\boldsymbol{g_0}(h,t))}||\boldsymbol{g_0}(h,t)||_\infty \end{array}\right) \tag{2}$$

where $\boldsymbol{\sigma}$ denotes the softmax function. If there is no latent relationship between h and t, then we expect $f_{\text{class}}(\boldsymbol{g_0}(h,t))$ to be large and thus the last coordinate in Equation (2), corresponding to r_{NoRel}, will be the largest value. If there is latent relationship, we expect $f_{\text{class}}(\boldsymbol{g_0}(h,t))$ to be small and thus the "NoRel" score will not be the dominant coordinate in Equation (2). We then leverage RAPS with level α, hyperparameters β and λ, scoring function $\boldsymbol{G_0}(h,t)$, and calibration data $\mathcal{D}_{\text{calib-NR}}$ (with $x = (h,t)$ and $y \in \mathcal{R}'$) to produce our conformal relation set $C_{1-\alpha}(h,t) \subset \mathcal{R}'$ which can be evaluated on $\mathcal{D}_{\text{test-NR}}$.

A major motivation for using conformal prediction to produce relation sets is the context of downstream analysis. The inclusion of r_{NoRel} in the relation space allows for flexible interpretability using $C_{1-\alpha}(h,t)$. For instance, if $r_{\text{NoRel}} \notin C_{1-\alpha}(h,t)$, then an analyst can be confident that a relationship is likely to exist between h and t, and that $C_{1-\alpha}(h,t)$ contains that true relationship with probability at least $1-\alpha$. Conversely, if $r_{\text{NoRel}} \in C_{1-\alpha}(h,t)$, then an analyst can carefully scrutinize other elements in $C_{1-\alpha}(h,t)$, and if none of the other $r \in C_{1-\alpha}(h,t)$ seem satisfactory, then the analyst can be confident no relationship exists between h and t.

Additionally, depending on the specific domain of information captured in the KG, there are likely "nonsense" relations that do not make sense between certain entities. For instance consider a KG comprised of historical figures and locations. For entities "Winston Churchill" and "United Kingdom", relations such as "BornIn" or "LivedIn" may make sense. In contrast, relations typically between humans such as "MarriedTo" or "SiblingOf" do not make sense, but may reasonably be in the relation space. As a result, conformal relation sets not only offer valid statistical guarantees, but can also be further pared down by leveraging domain expertise to remove "nonsense" answers in the conformal set.

Remark 1 *In practice, "NoRel" data may not always be provided, as many KGs are focused on known relationships between entities, not on known, lack of relationships between on entities. However, "NoRel" data can easily be generated provided $\mathcal{D}_{train}$, $\mathcal{D}_{calib}$, and $\mathcal{D}_{test}$, as is common for many publicly available KGs, by parsing through all triples and sampling (h, r_{NoRel}, t) for $h,t \in \mathcal{E}$ and $(h,r,t) \notin \mathcal{D}_{train} \cup \mathcal{D}_{calib} \cup \mathcal{D}_{test}$ for any $r \in \mathcal{R}$. Datasets $\mathcal{D}_{calib\text{-}real}$, $\mathcal{D}_{NoRel}$, $\mathcal{D}_{calib\text{-}NR}$, and $\mathcal{D}_{test\text{-}NR}$ can then be formed by sampling "NoRel" triples, partitioning $\mathcal{D}_{calib}$, and blending "NoRel" triples into $\mathcal{D}_{test}$ and a partition of $\mathcal{D}_{calib}$, using the other partition as $\mathcal{D}_{calib\text{-}real}$. Generating "NoRel" triples this way requires parsing $\mathcal{D}_{train}$, as "NoRel" triples represents known, definitive lack of relation. However, if a set of "NoRel" triples can be provided a priori then access to $\mathcal{D}_{test}$ is not required.*

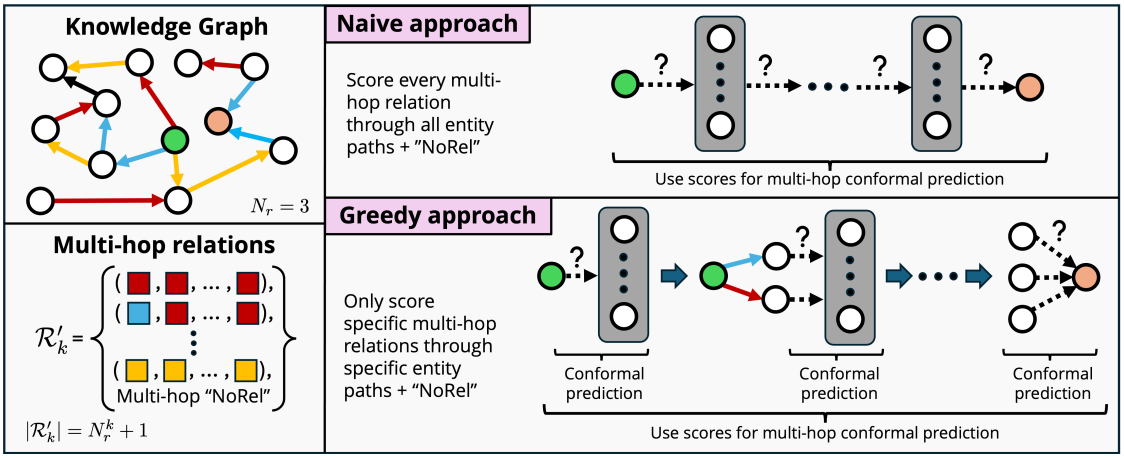

Figure 1: Overview of conformal multi-hop relation detection and classification.

3. Multi-hop relation detection and classification

We now detail the extension of our conformal single-hop relation detection and classification to the multi-hop case. For the remainder of the article, let $k \in \mathbb{N}$, $k \geq 2$ denote multi-hop length. For a head $h = e_0$, a tail $t = e_k$, and relations $r_1, \ldots, r_k \in \mathcal{R}$, we say a k-hop relation $\boldsymbol{r} = (r_1, \ldots, r_k)$ exists between h and t, denoted as $(h, \boldsymbol{r}, t)$, if there exists entities $e_1, \ldots, e_{k-1} \in \mathcal{E}$ such that (e_{s-1}, r_s, e_s) is true for $s = 1, \ldots, k$. Let $\mathcal{R}_k = \bigotimes_{s=1}^{k} \mathcal{R}$ denote the space of all k-hop relations. Let $r_{k\text{-hop,NoRel}}$ denote a k-hop "NoRel" class such that $(h, r_{k\text{-hop,NoRel}}, t)$ indicates that no k-hop relation $\boldsymbol{r}$ exists between h and t. As in the single-hop case, let $\mathcal{R}'_k := \mathcal{R}_k \cup \{r_{k\text{-hop,NoRel}}\}$.

Similarly to the single-hop case, given an (h, t) pair, our goal is to simultaneously detect and classify if a k-hop relation exists between h and t using conformal prediction sets $C_{k\text{-hop},1-\alpha}(h, t) \subset \mathcal{R}'_k$. Following the steps in the single-hop case, if we have a scoring function over all k-hop relation classes $\boldsymbol{g} : \mathcal{E} \times \mathcal{E} \to [0,1]^{N_r^k}$, analogous to Equation (1) in the single-hop case, we can train a probabilistic, binary classifier $f_{k\text{-hop,class}}$ on appropriately partitioned data and use a final scoring function $\boldsymbol{G} : \mathcal{E} \times \mathcal{E} \to [0,1]^{N_r^k+1}$

$$\boldsymbol{G}(h,t) \propto \begin{pmatrix} \boldsymbol{g}(h,t) \\ \frac{f_{k\text{-hop,class}}(\boldsymbol{g}(h,t))}{1-f_{k\text{-hop,class}}(\boldsymbol{g}(h,t))} ||\boldsymbol{g}(h,t)||_\infty \end{pmatrix} \tag{3}$$

where $||\boldsymbol{G}(h,t)||_1 = 1$.

3.1. Comparison to single-hop case

Unlike the single-hop case, where we anticipate a single primary relation r between head h and t, in the multi-hop case it is possible for there to exist multiple k-hop relations between h and t. Moreover, if the KG is well-connected, then for larger k we anticipate the number of k-hop relations to grow, as there will be more paths of length k connecting h and t. As a result, classifying k-hop relation types is a multi-label classification problem, which will require multi-label conformal prediction.

There are numerous works in the area of multi-label conformal classification, including label power set (LP) (Papadopoulos, 2014) which works directly in the power set of all labels, instance reproduction (IR) (Wang et al., 2014) which creates a single-label instance for each multi-label data point, and binary relevance (BR) (Wang et al., 2015) which combines a binary classification task on each individual class. More recent work in the field includes p-norm (Maltoudoglou et al., 2022) and (Katsios and Papadopulos, 2024) which addresses the dimensionality challenges of LP by assigning a nonconformity value based on the p-norm and Mahalanobis distance of multi-hot encoded error vector. Other works incorporate conformalized quantile techniques (Cauchois et al., 2021) which additionally develop inner and outer conformal sets, tree-based approaches (Tyagi and Guo, 2023) which leverage a hierarchical structure with multiple-testing, and (Angelopoulos et al., 2024) which handles multi-label classification as a conformal risk control task.

Algorithm 2: Multi-label RAPS (MLRAPS)

Input: Significance level $\alpha \in [0,1]$, scoring function $\tilde{\boldsymbol{g}} : \mathcal{X} \to [0,1]^K$, calibration data $D_{\text{calib}} = \{(x_i, Y_i)\}_{i=1}^N \subset \mathcal{X} \times 2^{\{1,\dots,K\}}$, test point $(x, Y) \in \mathcal{X} \times 2^{\{1,\dots,K\}}$ exchangeable with D_{calib}, regularization hyperparameters $\beta > 0$ and $\lambda > 0$

Output: Conformal set $C_{1-\alpha}(x) \subset \{1,\dots,K\}$ with $\mathbb{P}[Y \subseteq C_{1-\alpha}(x)] \geq 1-\alpha$

for $i \in \{1,\dots,N\}$ **do**

$\boldsymbol{S}^i \leftarrow \pi_i(\tilde{\boldsymbol{g}}(x_i))$; ▷ π_i `permutes` $g(x_i)$ `descending`

$R_i \leftarrow \sum_{j=1}^{z_i} \boldsymbol{S}_j^i + \lambda \cdot \max(0, z_i - \beta)$; ▷ $z_i = \max_{y_i \in Y_i} \pi_i(y_i)$`, same` π_i `as above`

end

$\hat{q} \leftarrow \lceil (1-\alpha)(1+N) \rceil$-th smallest R_i;

$\boldsymbol{S} \leftarrow \pi_x(\tilde{\boldsymbol{g}}(x))$; ▷ π_x `permutes` $g(x)$ `descending`

$C_{1-\alpha}(x) \leftarrow \{y \mid \sum_{j=1}^{\pi_x(y)} \boldsymbol{S}_j + \lambda \cdot \max(0, \pi_x(y) - \beta) \leq \hat{q}\}$

In this work, we consider a simple multi-label modification to RAPS (MLRAPS). For the task of classifying inputs $x \in \mathcal{X}$ into the classes $\{1,\dots,K\}$ using a scoring function $\tilde{\boldsymbol{g}} : \mathcal{X} \to [0,1]^K$, the RAPS nonconformity value for a calibration data point $(x, y) \in \mathcal{X} \times \{1,\dots,K\}$ is the sum over the coordinates of $\tilde{\boldsymbol{g}}(x)$ such that $\tilde{\boldsymbol{g}}_i(x) \geq \tilde{\boldsymbol{g}}_y(x)$, i.e. over all classes i which are at least as likely as the true label y according to $\tilde{\boldsymbol{g}}$. For MLRAPS (Algorithm 2), the label y is replaced with the multi-label $Y \subset \{1,\dots,K\}$, and so we now sum over classes i such that $\tilde{\boldsymbol{g}}_i(x) \geq \min_{y \in Y} \tilde{\boldsymbol{g}}_y(x)$, i.e. over all classes i which are at least as likely as the least likely label in Y. Under the same exchangeability assumptions as RAPS, but now on data with multi-labels, this modified MLRAPS achieves the same statistical guarantees.

Proposition 2 *Let $\tilde{\boldsymbol{g}} : \mathcal{X} \to [0,1]^K$ be a scoring function, let $D_{calib} = \{(x_i, Y_i)\}_{i=1}^N \subset \mathcal{X} \times 2^{\{1,\dots,K\}}$ be a set of calibration data, and let $(x, Y) \in \mathcal{X} \times 2^{\{1,\dots,K\}}$ be exchangeable with elements of D_{calib}. Then for $\alpha \in [0,1]$, the conformal set $C_{1-\alpha}(x) \subset \{1,\dots,K\}$ produced by MLRAPS in Algorithm 2 satisfies*

$$\mathbb{P}[Y \subseteq C_{1-\alpha}(x)] \geq 1-\alpha$$

Proof The event $Y \subseteq C_{1-\alpha}(x)$ is equivalent to $y \in C_{1-\alpha}(x)$ for all $y \in Y$, and $y \in C_{1-\alpha}(x)$ if and only if $R(x,y) \leq \hat{q}$ for $R(x,y) = \sum_{j=1}^{\pi_x(y)} \boldsymbol{S}_j + \lambda \cdot \max(\pi_x(y) - \beta)$. Let us define $R(x,Y) = R(x,y^*)$ where $y^* \in Y$ and $\pi_x(y^*) \geq \pi_x(y)$ for $y \in Y$. Since R is monotonic in y under π_x, i.e. $R(x,y) > R(x,y')$ when $\pi_x(y) > \pi_x(y')$, then $R(x,y) \leq \hat{q}$ for all $y \in Y$ if and only if $R(x,Y) \leq \hat{q}$. Recall $\hat{q}$ is the $\lceil (1-\alpha)(1+N) \rceil$-th smallest $R(x_i, Y_i)$ for $(x_i, Y_i) \in D_{\text{calib}}$. Since (x,Y) is assumed exchangeable with D_{calib}, we have our result

$$\mathbb{P}[Y \subseteq C_{1-\alpha}(x)] = \mathbb{P}[R(x,Y) \leq \hat{q}] \geq 1 - \alpha$$

■

We assume we have access to the multi-label k-hop dataset $\mathcal{D}^{k\text{-hop}}_{\text{calib-real}} \subset \mathcal{E} \times 2^{\mathcal{R}_k} \times \mathcal{E}$ which contains true multi-label calibration k-hops, as well as $\mathcal{D}^{k\text{-hop}}_{\text{NoRel}} \subset \mathcal{E} \times \{r_{k\text{-hop,NoRel}}\} \times \mathcal{E}$ which contains known "NoRel" k-hops. Additionally, we assume we have access to $\mathcal{D}^{k\text{-hop}}_{\text{calib-NR}}, \mathcal{D}^{k\text{-hop}}_{\text{test-NR}} \subset \mathcal{E} \times \left(2^{\mathcal{R}_k} \cup \{r_{k\text{-hop,NoRel}}\}\right) \times \mathcal{E}$ which contain true multi-label k-hops and "NoRel" k-hops, and whose elements are exchangeable. Note that a "triple" in $\mathcal{E} \times 2^{\mathcal{R}_k} \times E$ would be (h, R_k, t), where $R_k \subset \mathcal{R}_k$ is a subset of k-hop relations. If these datasets are unavailable, they can be constructed with sufficient 1-hop data, see Remark 4.

We use the datasets $\mathcal{D}^{k\text{-hop}}_{\text{calib-real}}$ and $\mathcal{D}^{k\text{-hop}}_{\text{NoRel}}$ along with a scoring function $\boldsymbol{g}(h,t)$ to generate the training-target pairs for "NoRel" data $\{(\boldsymbol{g}(h,t), 1) \,|\, (h, r_{k\text{-hop,NoRel}}, t) \in \mathcal{D}^{k\text{-hop}}_{\text{NoRel}}\}$ as well as $\{(\boldsymbol{g}(h,t), 0) \,|\, (h, R_k, t) \in \mathcal{D}^{k\text{-hop}}_{\text{calib-real}}\}$ for true data to fit our binary classifier $f_{k\text{-hop,class}}$. We then use $f_{k\text{-hop,class}}$ alongside $\boldsymbol{g}(h,t)$ to define our final scoring function $\boldsymbol{G}(h,t)$ as in Equation (3). Lastly, we use MLRAPS with level α, hyperparameters β and λ, scoring function $\boldsymbol{G}(h,t)$, and calibration data $\mathcal{D}^{k\text{-hop}}_{\text{calib-NR}}$ (with $x = (h,t)$ and $Y \subset \mathcal{R}'_k$) to form our conformal sets $C_{k\text{-hop},1-\alpha}(h,t) \subset \mathcal{R}'_k$ which can be evaluated on $\mathcal{D}^{k\text{-hop}}_{\text{test-NR}}$. We note that regardless of the choice of $\boldsymbol{G}(h,t)$, our $C_{k\text{-hop},1-\alpha}(h,t)$ is valid by the validity of MLRAPS.

3.2. Naive approach

In order to construct $\boldsymbol{G}(h,t)$ we need an appropriate scoring function $\boldsymbol{g} : \mathcal{E} \times \mathcal{E} \to [0,1]^{N_r^k}$ which provides a score for every $\boldsymbol{r} \in \mathcal{R}_k$. This step of the process is the most challenging, as in order to score every k-hop relation, we may need to analyze every possible path from h to t of length k, through all possible entities, and over all possible relations. Indeed, a naive choice would be the following:

$$\boldsymbol{g}_{\text{naive}}(h,t)_{\boldsymbol{r}} \propto \max_{e_1,\ldots,e_{k-1} \in \mathcal{E}} \sum_{s=1}^{k} \tilde{f}(e_{s-1}, r_s, e_s) \tag{4}$$

where $h = e_0$ and $t = e_k$ and $\tilde{f}(e, r, e') = \mathbf{A}(e)_{r,e'}$ with $\mathbf{A}(e)$ is defined as

$$\mathbf{A}(e) = \boldsymbol{\sigma} \begin{pmatrix} f(e, r^{(1)}, e^{(1)}) & \cdots & f(e, r^{(1)}, e^{(N_e)}) \\ \vdots & \ddots & \vdots \\ f(e, r^{(N_r)}, e^{(1)}) & \cdots & f(e, r^{(N_r)}, e^{(N_e)}) \end{pmatrix} \tag{5}$$

The matrix $\mathbf{A}(e)$ in Equation (5) corresponds to normalizing all of the base scores $f(e, r, e')$ for all hops (r, e') away from e. This construction gives a relative score to the path of (r, e')

away from e compared to all other paths, as well as normalizing the scores to $[0, 1]$ to avoid cancellation in the sum. We note that for the "last" hop, $f(e, r, t)$ where we have a known tail t, we set $\tilde{f}(e, r, t) = \boldsymbol{\sigma}(\boldsymbol{g_0}(e, t))_r$, i.e. the r-th component of the softmax of Equation (1).

This choice of naive scoring assigns to a k-hop relation $\boldsymbol{r}$ the maximum aggregate (normalized) base model score over all possible paths of length k with these relation types. Doing so requires a search over N_e^{k-1} many intermediate entity paths for a single k-hop relation $\boldsymbol{r}$, a computational hurdle worsened as this must be done for N_r^k many k-hop relations. For large-scale KG, this approach may be computationally intractable even for moderate k.

Moreover, this naive approach is particularly wasteful in the context of KGs, as some relation classes may only make sense for certain entities. For instance, suppose that for a given head h, only a single relation r^* makes sense for a single hop away, that is (h, r, e) is not true for $r \neq r^*$ and all $e \in \mathcal{E}$. Then assigning scores to k-hops $\boldsymbol{r} = (r_1, \ldots, r_k)$ where $r_1 \neq r^*$ is unnecessary. This is analogous to there being more "nonsense" answers in the k-hop relation space as k grows.

3.3. Greedy approach

To design a more efficient scoring function, we propose a greedily-constructed $\boldsymbol{g}(h, t)$ which iteratively expands a multi-hop neighborhood originating from h, which seeks to both reduce the number of intermediate entity paths considered for a given $\boldsymbol{r}$ as well as reduce the number of k-hop relations $\boldsymbol{r}$ that are assigned meaningful scores.

Suppose that we have access to two set-valued functions, $C_{\text{1-hop}} : \mathcal{E} \to 2^{\mathcal{R} \times \mathcal{E}}$ as well as $C_{\text{end}} : \mathcal{E} \to 2^{\mathcal{R}}$. The set $C_{\text{1-hop}}(e)$ is a 1-hop neighborhood of relation, tail pairs (r, e') away from e, whereas the set $C_{\text{end}}(e, e')$ is a set of relations r between e and e'.

For a given pair (h, t), we compute $\boldsymbol{g}(h, t)$ as follows. First, we compute the 1-hop set $C_1(h) = C_{\text{1-hop}}(h)$, and let $\boldsymbol{\xi_1} = (r_1, e_1) \in C_{\text{1-hop}}(h)$. Then for $s = 2, \ldots, k-1$, we iteratively expand to an s-hop neighborhood by

$$C_s(h) = \bigcup_{(\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_{s-1}}) \in C_{s-1}(h)} \{(\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_s}) \,|\, \boldsymbol{\xi_s} \in C_{\text{1-hop}}(e_{s-1})\}$$

where $\boldsymbol{\xi_s} = (r_s, e_s) \in C_{\text{1-hop}}(e_{s-1})$. Thus, at iteration s, we have an s-hop neighborhood away from h, as well as the paths through relations and intermediate entities to get there. After $k-1$ iterations, we then use C_{end} to connect to the target tail t,

$$C_k(h, t) = \bigcup_{(\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_{k-1}}) \in C_{k-1}(h)} \{(\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_k}) \,|\, r_k \in C_{\text{end}}(e_{k-1}, t)\}$$

where $\boldsymbol{\xi_k} = (r_k, t)$. We then define our scoring function $\boldsymbol{g}(h, t) : \mathcal{E} \times \mathcal{E} \to [0, 1]^{N_r^k}$ as

$$\boldsymbol{g}(h, t)_{\boldsymbol{r}} \propto \begin{cases} \max_{e_1, \ldots, e_{k-1}} \sum_{s=1}^{k} \tilde{f}(e_{s-1}, r_s, e_s), & \text{if } (\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_k}) \in C_k(h, t) \\ \varepsilon, & \text{otherwise} \end{cases} \tag{6}$$

where $e_0 = h$ and $e_k = t$ and $||\boldsymbol{g}(h, t)||_1 = 1$ Here $\varepsilon > 0$ is a small number chosen for numerical stability. We will refer to the set of $\boldsymbol{r}$ whose $\boldsymbol{g}(h, t)$ values are not assigned proportional to ε as the *effective support* of $\boldsymbol{g}(h, t)$, denoted as $\text{supp}_{\text{eff}}(\boldsymbol{g})$, and likewise for $\boldsymbol{G}(h, t)$.

This greedily-constructed $\boldsymbol{g}(h,t)$ places probability mass on k-hop relations $\boldsymbol{r}$ which are traversed from a path $(h, r_1, e_1), (e_1, r_2, e_2), \ldots, (e_{k-1}, r_k, t)$ using the set-valued functions $C_{\text{1-hop}}$ and C_{end}. Comparatively, the naive approach using Equation (4) may seek to place probability mass on all k-hop relations. Moreover, for a k-hop relation $\boldsymbol{r}$, our greedy $\boldsymbol{g}(h,t)$ only computes the maximum aggregate (normalized) base model score over the paths traced out with $C_{\text{1-hop}}$, not over all N_e^{k-1} paths as in the naive approach. Note we recover the naive scoring function if our set-valued functions $C_{\text{1-hop}}$ and C_{end} are as large as possible. That is, if $C_{\text{1-hop}}(e) = \mathcal{R} \times \mathcal{E}$ for all $e \in \mathcal{E}$ and $C_{\text{end}}(e, e') = \mathcal{R}$ for all $e, e' \in \mathcal{E}$, then $\boldsymbol{g}(h,t)$ as in Equation (6) is exactly $\boldsymbol{g}_{\text{naive}}(h,t)$.

It remains to determine our choice of set-valued functions $C_{\text{1-hop}}$ and C_{end}. If they are too large, then our greedy method will lose efficiency, as we will have to scan over too many k-hop relations and paths per k-hop relation. However, care must be taken to ensure that $C_{\text{1-hop}}$ is not too small, as $\boldsymbol{g}(h,t)$ is used to form $\boldsymbol{G}(h,t)$ which will is used in the k-hop MLRAPS to produce $C_{k\text{-hop},1-\alpha}(h,t)$.

If $C_{\text{1-hop}}$ is too small, our greedy approach may undercover the (r, e') neighborhood away from e, and the effective support $\text{supp}_{\text{eff}}(\boldsymbol{G})$ may be too small. Consequently, during MLRAPS calibration, if a k-hop relation in the label set is outside $\text{supp}_{\text{eff}}(\boldsymbol{G})$, then the nonconformity value for that data point will be large, requiring summing through all of $\text{supp}_{\text{eff}}(\boldsymbol{G})$, and summing noise outside $\text{supp}_{\text{eff}}(\boldsymbol{G})$. If this happens too frequently, then the MLRAPS calibration will be poor and $C_{k\text{-hop},1-\alpha}(h,t)$ may be large. Thus, we want to choose $C_{\text{1-hop}}$ and C_{end} in a manner where we can cover the correct set of (r, e') and r with confidence.

3.4. Conformal set-valued functions

We consider constructing $C_{\text{1-hop}}$ and C_{end} using conformal prediction sets. The choice for C_{end} is natural, as it is almost exactly the single-hop case. We assume we also have access to a dataset of 1-hop calibration triples $\mathcal{D}_{\text{calib}} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$. Let $\boldsymbol{g_2}(e, e') = \boldsymbol{\sigma}(\boldsymbol{g_0}(e, e'))$ where $\boldsymbol{g_0}$ is the single-hop scoring function in Equation (1). We use RAPS with level α_{int}, hyperparameters β_{int} and λ_{int}, scoring function $\boldsymbol{g_2}(e, e')$ and calibration data $\mathcal{D}_{\text{calib}}$ (with $x = (e, e')$ and $Y \subset \mathcal{R}$), to produce our conformal set-valued function $C_{\text{end},1-\alpha_{\text{int}}} : \mathcal{E} \times \mathcal{E} \to 2^{\mathcal{R}}$. We shall refer to $\alpha_{\text{int}}, \beta_{\text{int}}, \lambda_{\text{int}}$ as interior parameters, used to construct $C_{\text{1-hop}}$ and C_{end}, to distinguish them from the α, β, λ parameters used to construct $C_{k\text{-hop},1-\alpha}(h,t)$.

For $C_{\text{1-hop}}$ we consider two options: a direct approach which scans all possible hops (r, e') from e simultaneously, and a sequential approach which first grabs the best relations r from e and then subsequently the best 1-hop tails e'.

3.4.1. Direct approach

For the direct approach, we shall consider all possible hops (r, e') away from e together. We treat this as a multi-label classification problem where the input is the current entity e, and the output is a subset of $\mathcal{R} \times \mathcal{E}$. Let $\boldsymbol{g_1}_{,\text{direct}} : \mathcal{E} \to [0,1]^{N_r N_e}$ be defined as

$$\boldsymbol{g_1}_{,\text{direct}}(e) = \text{vec}(\mathbf{A}(e))$$

where $\mathbf{A}(e)$ is defined in Equation (5).

We assume we have access to a multi-label calibration dataset $\mathcal{D}_{\text{direct}} \subset \mathcal{E} \times 2^{\mathcal{R}\times\mathcal{E}}$. Elements of $\mathcal{D}_{\text{direct}}$ correspond to a head entity e paired with a set of (r, e') such that (e, r, e') is true. Note that if such a dataset is unavailable, it can be constructed by parsing through $\mathcal{D}_{\text{calib}}$. We then use MLRAPS with interior level α_{int}, hyperparameters β_{int} and λ_{int}, scoring function $\boldsymbol{g_{1,\text{direct}}}(e)$, and calibration data $\mathcal{D}_{\text{direct}}$ (with $x = e$ and $Y \subset \mathcal{R} \times \mathcal{E}$) to produce conformal relation sets as our set-valued function $C^{\text{direct}}_{\text{1-hop},1-\alpha_{\text{int}}} : \mathcal{E} \to 2^{\mathcal{R}\times\mathcal{E}}$.

Since the direct approach uses MLRAPS with $\mathcal{R} \times E$ treated as the full space classification space, we expect $C^{\text{direct}}_{\text{1-hop},1-\alpha_{\text{int}}}$ to be potentially large as there are $N_r N_e$ total possible elements to consider. This is both a benefit and a weakness. On the plus side, we can be confident the direct approach will not undercover the (r, e') neighborhood away from e, and thus avoid problems with $\text{supp}_{\text{eff}}(\boldsymbol{G})$ being too small. However, if $C^{\text{direct}}_{\text{1-hop},1-\alpha_{\text{int}}}$ is too large then it may not be computationally tractable for larger k as at each iteration we build a 1-hop set off every element in the previous 1-hop set.

3.4.2. Sequential approach

In the sequential approach, we instead first find a set of relations r to "hop away" from e with, and once those are found we then find a set of tails e' for each (e, r). As in the direct approach, this will be done by treating this as two multi-label classification problems. In the first, the input will be an entity e with the output as a subset of $\mathcal{R}$. In the second, the input will be an entity relation pair (e, r) and the output will be a subset of $\mathcal{E}$.

We define $\boldsymbol{g_{1,r\text{-seq}}} : \mathcal{E} \to [0, 1]^{N_r}$ and $\boldsymbol{g_{1,t\text{-seq}}} : \mathcal{E} \times \mathcal{R} \to [0, 1]^{N_e}$ by:

$$\boldsymbol{g_{1,r\text{-seq}}}(e)_r \propto \max \mathbf{A}(e)_{r,*}$$
$$\boldsymbol{g_{1,t\text{-seq}}}(e, r) \propto \mathbf{A}(e)_{r,*}$$

where $||\boldsymbol{g_{1,r\text{-seq}}}(e)||_1 = 1$ and $||\boldsymbol{g_{1,t\text{-seq}}}(r, e)||_1 = 1$. Here $\boldsymbol{g_{1,r\text{-seq}}}(e)$ is the normalized column-wise maximum of $\mathbf{A}(e)$, and $\boldsymbol{g_{1,t\text{-seq}}}(e, r)$ is the normalized r-th row of $\mathbf{A}(e)$. The intuition behind this is as follows. To find only the best relations r which (e, r, e') may be true, we can scan all (r, e') pairs, and for each r, just take the maximum score over all the possible e'. Large scores will correspond to relations r for which there is at least one entity e' where (e, r, e') is true, and small scores will correspond to relations r where (e, r, e') is not true for all entities e'. And to find the tails e' for a given (e, r), we can just take all the scores from the r-th column of $\mathbf{A}(e)$.

We assume that we have two multi-label calibration sets $\mathcal{D}_{r\text{-seq}} \subset \mathcal{E} \times 2^{\mathcal{R}}$ and $\mathcal{D}_{t\text{-seq}} \subset \mathcal{E} \times \mathcal{R} \times 2^{\mathcal{E}}$. As in the direct case, if such datasets are not available they can be constructed by parsing $\mathcal{D}_{\text{calib}}$. Elements of $\mathcal{D}_{r\text{-seq}}$ correspond to a head entity e paired with a set of relations r such that there exists an e' where (e, r, e') is true. Elements of $\mathcal{D}_{t\text{-seq}}$ correspond to head relation pairs (e, r) paired with a set of tails e' such that (e, r, e') is true.

We first use MLRAPS with interior level α_{int}, hyperparameters β_{int} and λ_{int}, scoring function $\boldsymbol{g_{1,r\text{-seq}}}(e)$, and calibration data $\mathcal{D}_{r\text{-seq}}$ (with $x = e$ and $Y \subset \mathcal{R}$) to produce conformal prediction sets $C^{r\text{-seq}}_{1-\alpha_{\text{int}}} : \mathcal{E} \to 2^{\mathcal{R}}$. We then use MLRAPS with interior level α_{int}, hyperparameters β_{int} and λ_{int}, scoring function $\boldsymbol{g_{1,t\text{-seq}}}(e, r)$, and calibration data $\mathcal{D}_{t\text{-seq}}$ (with $x = (e, r)$ and $Y \subset \mathcal{E}$) to produce conformal prediction sets $C^{t\text{-seq}}_{1-\alpha_{\text{int}}} : \mathcal{E} \times \mathcal{R} \to 2^{\mathcal{E}}$. Finally, we define our sequential 1-hop set-valued function as $C^{\text{seq}}_{\text{1-hop},1-\alpha_{\text{int}}} : \mathcal{E} \to 2^{\mathcal{R}\times\mathcal{E}}$ by

$$C^{\text{seq}}_{\text{1-hop},1-\alpha_{\text{int}}}(e) = \{(r, e') \,|\, r \in C^{r\text{-seq}}_{1-\alpha_{\text{int}}}(e), e' \in C^{t\text{-seq}}_{1-\alpha_{\text{int}}}(e, r)\}$$

Unlike the direct approach which considers all (r, e') from e at once, we expect the sequential approach to offer smaller conformal 1-hop sets, as it breaks up its search first over $\mathcal{R}$ and then over $\mathcal{E}$. While this implies more efficient sets and potentially better scalability to large k, the sequential approach may undercover more frequently, and thus be less stable than the direct approach.

Since our sequential 1-hop set-valued function requires performing two conformal classifications, one layered into the next, with the same interior level α_{int}, there is no guarantee that $C^{\text{seq}}_{\text{1-hop},1-\alpha_{\text{int}}}(e)$ achieves at least $1-\alpha_{\text{int}}$ coverage. This is in contrast to the direct approach which performs only one conformal classification, and thus $C^{\text{direct}}_{\text{1-hop},1-\alpha_{\text{int}}}$ has the guarantee of $1-\alpha_{\text{int}}$ coverage. However, since the conformal choice of $C_{\text{1-hop}}$ (direct or sequential) must run iteratively and then eventually feed into C_{end}, neither approach will lead to guaranteed $1-\alpha$ coverage on the final k-hop neighborhood set $C_k(h,t)$.

We note that exact coverage is not necessarily the primary goal in the construction of a conformal $C_{\text{1-hop}}$ and C_{end}, nor in $C_k(h,t)$. Rather, the goal is to have an oft-reliable set-valued function which can then be used to expand a neighborhood around the target head entity h until we reach the target tail entity t. Since $C_k(h,t)$ is only used to construct $\boldsymbol{g}(h,t)$, and subsequently $\boldsymbol{G}(h,t)$, which is then used in a final MLRAPS calibration on k-hop relations, we will maintain valid statistical coverage on the k-hop relation sets. And so long as $|\operatorname{supp}_{\text{eff}}(\boldsymbol{G})|$ does not get too small, we can expect our approach to be stable.

Remark 3 *For the sequential approach, different choices of α_{int} can be made for $C^{r\text{-}seq}_{1-\alpha_{int}}$ and $C^{t\text{-}seq}_{1-\alpha_{int}}$, however for ease of presentation we have kept these the same. Additionally, since α_{int} is the primary source of the size of $C_{1\text{-}hop}$, for both the direct and sequential approaches, in practice it may be beneficial to use additional calibration data to tune α_{int}.*

Remark 4 *In practice, many KGs may not naturally come with k-hop calibration or test data, much less multi-label versions. Likewise, many datasets may not have "NoRel" k-hop data either. However, provided 1-hop calibration and testing datasets $\mathcal{D}_{calib}$ and $\mathcal{D}_{test}$, multi-label k-hop data can be generated by parsing through these datasets. In fact, one could mix $\mathcal{D}_{calib}$ and $\mathcal{D}_{test}$ together, parse $\mathcal{D}_{calib} \cup \mathcal{D}_{test}$ to create a larger set of multi-label k-hop data, and then repartition out to calibration and testing sets. Additionally, "NoRel" k-hop data can be generated as in the single-hop case, see Remark 1, but now parsing for multi-hops.*

4. Numerical experiments

We demonstrate our conformal multi-hop relation detection and classification on the benchmark CoDEx KG datasets (Safavi and Koutra, 2020), which comprise information extracted from WikiData. We test on both CoDEx Small (CoDEx-S) and CoDEx Medium (CoDEx-M) which are provided as 1-hop datasets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{calib}}$, $\mathcal{D}_{\text{test}}$, whose sizes are shown in Table 1. We consider CoDEx-S with $k = 2, 3$ and CoDEx-M with $k = 2$. We guarantee exchangeability in all our benchmark experiments by always blending calibration and test data and randomly repartitioning, as then calibration and test data are drawn uniformly from the combined original data. For real-application deployment, additional caution should be used to ensure exchangeability.

In each case, we use the $\mathcal{D}_{\text{train}}$ dataset to train a base model $f(h,r,t)$ using DistMult (Yang et al., 2015) with 100 epochs, implemented through PyKEEN (Ali et al., 2021).

Table 1: Sizes of benchmark CoDEx datasets.

	$N_e = \lvert\mathcal{E}\rvert$	$N_r = \lvert\mathcal{R}\rvert$	$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{calib}}$	$\mathcal{D}_{\text{test}}$
CoDEx-S	2034	42	32,888	1827	1828
CoDEx-M	17,050	51	185,584	10,310	10,311

We choose DistMult for its simplicity and efficiency, as preliminary testing with more sophisticated KG embedding architectures such as RGCN did not demonstrate a significant improvement in accuracy for f. In handling relatively small KG and k, we perform all our tests on an 8-core Macbook Pro. Since the CoDEx datasets do not come with k-hop data, we use the 1-hop datasets $\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}}$ to generate $\mathcal{D}_{\text{calib}}^{k\text{-hop}}$ and $\mathcal{D}_{\text{test}}^{k\text{-hop}}$ through mixing and parsing as discussed in Remark 4. We additionally leverage the $\mathcal{D}_{\text{train}}$ to generate k-hop "NoRel" triples as also discussed in Remark 4. We blend and partition $\mathcal{D}_{\text{calib}}^{k\text{-hop}}$, $\mathcal{D}_{\text{test}}^{k\text{-hop}}$, and $\mathcal{D}_{\text{test}}^{k\text{-hop}}$ to produce our datasets $\mathcal{D}_{\text{calib-real}}^{k\text{-hop}}$, $\mathcal{D}_{\text{NoRel}}^{k\text{-hop}}$, $\mathcal{D}_{\text{calib-NR}}^{k\text{-hop}}$, and $\mathcal{D}_{\text{test-NR}}^{k\text{-hop}}$. Although the CoDEx datasets do contain hard negatives (e.g. 1-hop "NoRels"), these cannot easily be used to generate k-hop "NoRels", as this would require at least 1-hop along every k-hop path from h to t to have be a hard negative. Lastly, we generate the datasets $\mathcal{D}_{\text{direct}}$ for the direct approach and $\mathcal{D}_{r\text{-seq}}$, $\mathcal{D}_{t\text{-seq}}$ for the sequential approach by parsing through $\mathcal{D}_{\text{calib}}$. Sizes of these datasets for CoDEx-S with $k = 2, 3$ and CoDEx-M with $k = 2$ are provided in Table 2, where we have under-sampled the number of 2-hops in CoDEx-M and 3-hops in CoDEx-S for computational considerations. For each case, we use a support vector machine with radial basis function kernel for $f_{k\text{-hop,class}}$, and use $\varepsilon = 10^{-10}$ for Equation (6).

Table 2: Dataset sizes for conformal k-hop detection and classification.

	$\mathcal{D}_{\text{direct}}$	$\mathcal{D}_{r\text{-seq}}$	$\mathcal{D}_{t\text{-seq}}$	$\mathcal{D}_{\text{calib-real}}^{k\text{-hop}}$	$\mathcal{D}_{\text{NoRel}}^{k\text{-hop}}$	$\mathcal{D}_{\text{calib-NR}}^{k\text{-hop}}$	$\mathcal{D}_{\text{test-NR}}^{k\text{-hop}}$
CoDEx-S (k=2)	644	644	775	210	146	2523	2804
CoDEx-S (k=3)	644	664	775	96	96	144	72
CoDEx-M (k=2)	4288	4288	4896	150	100	1776	1974

We consider five primary diagnostics for each test case performed: the average 1-hop set size $|C_{1\text{-hop}}|$, the average effective support size $|\operatorname{supp}_{\text{eff}}(\boldsymbol{G})|$, the average conformal k-hop relation set size $|C_{k\text{-hop},0.9}(h,t)|$ (using MLRAPS with $\alpha = 0.1$), the "NoRel" false positive rate (FPR), and the "NoRel" false negative rate (FNR). We also report empirical coverage at the $\alpha = 0.1$ level. Both the average $|C_{1\text{-hop}}|$ and average $|\operatorname{supp}_{\text{eff}}(\boldsymbol{G})|$ will inform how efficient the test case is, as smaller values mean less paths need to be searched. However, average $|C_{1\text{-hop}}|$ and $|\operatorname{supp}_{\text{eff}}(\boldsymbol{G})|$ will also inform how stable the case is, as if they are too small they will tend to undercover and lead to large $|C_{k\text{-hop},0.9}|$ is. Ideally the average $|C_{k\text{-hop},0.9}|$ is small relative to $\mathcal{R}_k'$ while also maintaining a small $|C_{1\text{-hop}}|$. The "NoRel" FPR is the proportion of (h,t) for which there does exist a nonempty set of k-hop relations, but $r_{k\text{-hop,NoRel}} \in C_{k\text{-hop},1-\alpha}(h,t)$. Conversely, the "NoRel" FNR is the proportion of (h,t) with known k-hop "NoRel", but $r_{k\text{-hop,NoRel}} \notin C_{k\text{-hop},1-\alpha}(h,t)$. Unless otherwise specified, all test cases are performed with hyperparameters $\beta = \beta_{\text{int}} = 1$ and $\lambda = \lambda_{\text{int}} = 10$ for both the

interior MLRAPS used for $C_{\text{1-hop},1-\alpha_{\text{int}}}(e, e')$ and for the k-hop MLRAPS $C_{k\text{-hop},1-\alpha}(h, t)$, to regularize conformal sets to be as small as possible.

Table 3: k-hop statistics using $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{calib}} \cup \mathcal{D}_{\text{test}}$.

	Total k-hops	Mean $M_k(h,t)$	Max $M_k(h,t)$	$\hat{\tau}_{0.90}(M_k(h,t))$
CoDEx-S (k=2)	847,065	2.70	192	6
CoDEx-S (k=3)	42,007,992	80.95	5669	152
CoDEx-M (k=2)	3,629,061	1.34	194	3

Since, for benchmarking purposes, we have access to the training data $\mathcal{D}_{\text{train}}$ in addition to $\mathcal{D}_{\text{calib}}$ and $\mathcal{D}_{\text{test}}$, we can compute how many k-hops exist between entities (h, t). For $h, t \in \mathcal{E}$ such that there exists at least one k-hop relation, let $M_k(h, t) : \mathcal{E} \times \mathcal{E} \to \{1, 2, \dots\}$ denote the number of k-hop relations between h and t, as found by parsing $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{calib}} \cup \mathcal{D}_{\text{test}}$. In Table 3 we show the total number of k-hop relations found, the mean $M_k(h, t)$, the maximum $M_k(h, t)$, as well as the 90th quantile $\hat{\tau}_{0.90}(M_k(h, t))$. We observe that in going from $k = 2$ to $k = 3$ for CoDEx-S, we see a significant increase in the total number of k-hops, the mean $M_k(h, t)$, and the maximum $M_k(h, t)$. This underscores the challenge behind k-hop relation prediction, as for well-connected KGs we expect the number of paths, and thus multi-hop relations, between h and t to grow dramatically as k increase.

Ideally we would like our conformal set at level $\alpha = 0.1$ to be just slightly larger than $\hat{\tau}_{0.90}(M_k(h, t))$. If 90% of the time an (h, t) pair has less than $\hat{\tau}_{0.90}(M_k(h, t))$ many k-hop relations, then it stands to reason we would want a $C_{k\text{-hop},0.90}(h, t)$ to be of a similar size. However, we stress that in practice, when training data is not accessible, *such information cannot be derived*. Moreover, a defining characteristic of KGs is that they are incomplete, so even if training data is accessible, the information displayed in Table 3 must be taken with a grain of salt. We include these statistics here for completeness, but emphasize that by the incomplete nature of KGs, they *cannot* be used as a true barometer for calibrating conformal k-hop relation sets.

4.1. CoDEx-S

For CoDEx-S, we first consider $k = 2$ case, testing both the direct and sequential approaches, as well as using $\lambda_{\text{int}} = 0, 10$ for $C^{\text{direct}}_{\text{1-hop},1-\alpha_{\text{int}}}(e)$ and $C^{\text{seq}}_{\text{1-hop},1-\alpha_{\text{int}}}(e)$. All four of these tests will be performed with internal level $\alpha_{\text{int}} = 0.10$, and for each case we additionally test using $\lambda = 0, 10$ for $C_{k\text{-hop},1-\alpha}(h, t)$. We also test the naive approach for comparison. For this test case, 2-hop relation space has size $|\mathcal{R}'_2| = 1765$.

In Table 4 we see that the naive, greedy, and sequential approaches all yield very small $|C_{\text{2-hop},0.9}|$ compared to $|\mathcal{R}'_2| = 1765$. In particular, we observe that our greedy approaches outperforms the naive approach for all tests in producing smaller conformal 2-hop relation sets $|C_{\text{2-hop},0.9}|$, with the best case (sequential, $\lambda_{\text{int}} = 10$) having $|C_{\text{2-hop},0.9}|/|\mathcal{R}'_2 = 0.0053$, a massive reduction in 2-hop relation space.

Comparing greedy approaches, we see that the sequential approach not only has smaller 1-hop sets $|C_{\text{1-hop}}|$ and smaller $|\operatorname{supp}_{\text{eff}}(\boldsymbol{G})|$ than the direct approach, and is thus more efficient, but also yields smaller conformal 2-hop relation sets. Comparing $\lambda_{\text{int}} = 0, 10$, we

see that $\lambda_{\text{int}} = 10$ is better for both the sequential and direct approach. All tests for this case have roughly the same "NoRel" FPR and "NoREL" FNR. We note that $C_{\text{1-hop}} \subset \mathcal{R} \times \mathcal{E}$, and thus the reported $|C_{\text{1-hop}}|$ sizes are very small compared to $|\mathcal{R} \times \mathcal{E}| = N_r N_e = 85,428$.

Table 4: Diagnostics for CoDEx-S conformal 2-hop relation prediction, $\alpha_{\text{int}} = 0.10$.

	Naive	Direct		Sequential	
		$\lambda_{\text{int}} = 0$	$\lambda_{\text{int}} = 10$	$\lambda_{\text{int}} = 0$	$\lambda_{\text{int}} = 10$
Average $\lvert C_{\text{1-hop}} \rvert$	N/A	2493.90	2030.96	1050.62	662.17
Average $\lvert \text{supp}_{\text{eff}}(\boldsymbol{G}) \rvert$	N/A	295.94	155.03	221.49	124.49
Average $\lvert C_{\text{2-hop},0.9} \rvert$	43.58	25.06	12.31	13.06	9.45
"NoRel" FPR	0.06	0.06	0.07	0.06	0.06
"NoRel" FNR	0.09	0.06	0.05	0.07	0.07
Empirical coverage, $\alpha = 0.1$	0.899	0.898	0.900	0.900	0.896

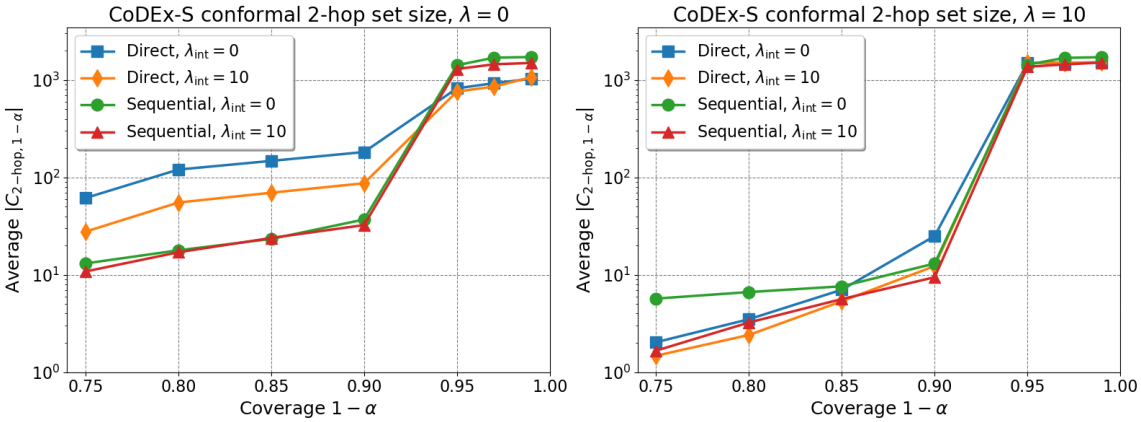

Figure 2: Conformal 2-hop relation set sizes for CoDEx-S.

In Figure 2 we plot the conformal 2-hop relation set size $|C_{\text{2-hop},1-\alpha}|$ for various choices of α with $\lambda = 0, 10$. We observe that in the $\lambda = 0$ case, the sequential approach with both $\lambda_{\text{int}} = 0, 10$ noticeably outperforms the direct approach. At $\lambda = 10$ however, all four test cases are much closer together, with the notable difference being between $\lambda_{\text{int}} = 0$ and $\lambda_{\text{int}} = 10$ rather than the direct versus sequential. Overall, the best results are obtained when choosing $\lambda_{\text{int}} = 10$ and $\lambda = 10$, with the difference between sequential and direct being minor. This suggests that regularization, both for the interior 1-hop conformal prediction, as well as for the conformal k-hop relation prediction, is preferred to no regularization.

We next consider CoDEx-S with $k = 3$, comparing the direct and sequential approach with $\alpha_{\text{int}} = 0.10, 0.05$, to demonstrate the capability of scaling to larger choices of k. We use a fixed $\lambda_{\text{int}} = 10$ and consider $\lambda = 0, 10$. With $N_r = 42$, the 3-hop relation space has size $|\mathcal{R}'_3| = 74,089$. Since the only change is in the multi-hop length, the $C_{\text{1-hop}}$ and G

functions are the same as in the $k = 2$ case. We do not test the naive approach for $k = 3$ as it proved computationally intractable. In Table 5 we display our five diagnostics.

Table 5: Diagnostics for CoDEx-S conformal 3-hop relation prediction.

	Direct		Sequential	
	$\alpha_{\text{int}} = 0.10$	$\alpha_{\text{int}} = 0.05$	$\alpha_{\text{int}} = 0.10$	$\alpha_{\text{int}} = 0.05$
Average $\lvert C_{\text{1-hop}}\rvert$	2030.93	3101.89	663.94	1664.25
Average $\lvert \text{supp}_{\text{eff}}(\boldsymbol{G})\rvert$	6969.38	16371.42	6357.97	17644.52
Average $\lvert C_{\text{3-hop},0.9}\rvert$	63,989.01	1760.04	63,434.36	833.88
"NoRel" FPR	1.00	0.04	1.00	0.06
"NoRel" FNR	0.00	0.04	0.00	0.08
Empirical coverage, $\alpha = 0.1$	0.889	0.917	0.875	0.889

We observe that for both the sequential and direct approaches at $\alpha_{\text{int}} = 0.1$, our method is quite unstable, with enormous conformal 3-hop relation sets, along with a "NoRel" FPR of 1. Examining Figure 3, we see that for $\alpha = 0.10$, the nonconformity values are massive for a large proportion of calibration points.

This can be explained by the relatively small $|C_{\text{1-hop}}|$ at $\alpha_{\text{int}} = 0.1$. Both methods with $\alpha_{\text{int}} = 0.10$ produce too small $C_{\text{1-hop}}$ sets, undercovering the 1-hop neighborhood away in $\mathcal{R} \times \mathcal{E}$. As a result, $\text{supp}_{\text{eff}}(\boldsymbol{G})$ is too small, with too many labels outside $\text{supp}_{\text{eff}}(\boldsymbol{G})$ during calibration. This leads to summing through noise outside of $\text{supp}_{\text{eff}}(\boldsymbol{G})$ until all true labels are covered which leads to a sharp jump in nonconformity values and thus a large quantile in MLRAPS, which leads to large $C_{\text{3-hop},1-\alpha}$. We see a "NoRel" FPR of 1 due to the scaling of the "NoRel" score in Equation (3), which should always place "NoRel" within $\text{supp}_{\text{eff}}(\boldsymbol{G})$, and thus all conformal sets will contain "NoRel" when unstable.

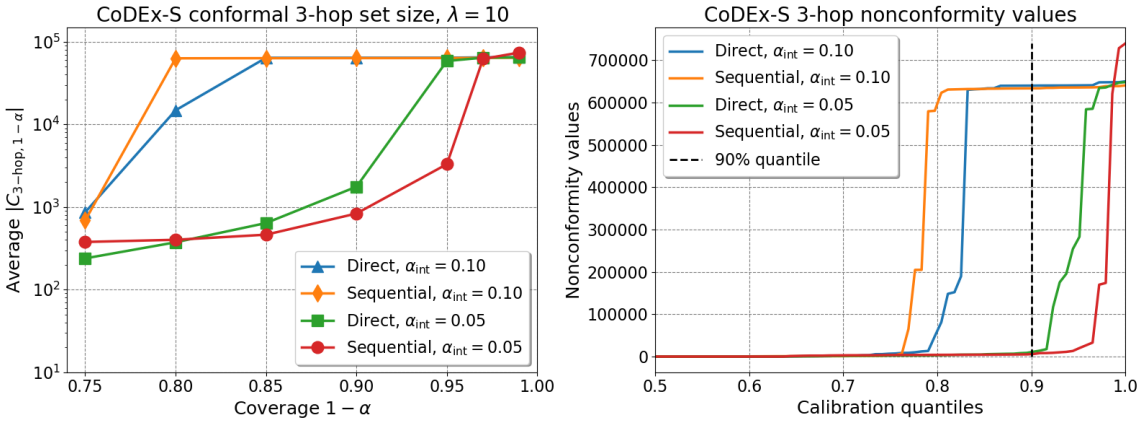

Figure 3: Conformal 3-hop relation set sizes for CoDEx-S and nonconformity values.

However, once we decrease the interior level to $\alpha_{\text{int}} = 0.05$, we see in Table 5 that the $C_{\text{1-hop}}$ size grows for both the direct and sequential approach, leading to larger $|\text{supp}_{\text{eff}}(\boldsymbol{G})|$

and thus more stable calibration. Consequently, we see smaller $|C_{\text{3-hop},0.9}|$, and improved "NoRel" FPR and FNR. We can additionally see this improvement in the nonconformity values in Figure 3, where nonconformity values have decreased sufficiently so the inevitable jump does not occur until $\alpha > 0.1$.

Although $\alpha = 0.05$ shrinks the conformal 3-hop relation sets, they are still on average of moderate size at 1760 for the direct approach and 834 for the sequential approach. However, since $|\mathcal{R}_3'| = 74,089$, these sets provide a substantial reduction in the 3-hop relation space, with $|C_{\text{3-hop},0.9}|/|\mathcal{R}_3'| = 0.024$ for the direct approach and 0.011 for the sequential approach. Moreover, due to the increase in potential "nonsense" answers for longer multi-hops, there is still utility behind these moderately sized conformal 3-hop relation sets for downstream analysis. We note that for $k = 3$ we only have 144 calibration and 72 test data points, which explains some empirical coverage deviating from the expected 0.9.

4.2. CoDEx-M

Lastly we consider the CoDEx-M dataset with $k = 2$, with the goal of demonstrating the capability of scaling to larger KGs. Compared to CoDEx-S, CoDEX-m has over 8 times more entities and slightly more relations. For this dataset $|\mathcal{R} \times \mathcal{E}| = N_r N_e = 869,550$, significantly larger than in CoDEx-S where the dimensionality was 85,428. As a result, we expect the task of estimating a 1-hop set with $C_{\text{1-hop}}$ to be more difficult. The 2-hop relation space has size $|\mathcal{R}_2'| = 2602$, slightly larger than CoDEx-S with $k = 2$, but much smaller than CoDEX-S with $k = 3$.

Table 6: Diagnostics for CoDEx-M conformal 2-hop relation prediction.

	Naive	Direct		Sequential	
		$\alpha_{\text{int}} = 0.10$	$\alpha_{\text{int}} = 0.05$	$\alpha_{\text{int}} = 0.10$	$\alpha_{\text{int}} = 0.05$
Average $\lvert C_{\text{1-hop}}\rvert$	N/A	86,855.31	97,715.20	6237.77	12,302.73
Average $\lvert \text{supp}_{\text{eff}}(\boldsymbol{G})\rvert$	N/A	534.42	730.63	604.65	954.85
Average $\lvert C_{\text{2-hop},0.9}\rvert$	255.52	53.14	50.03	2142.38	102.14
"NoRel" FPR	0.02	0.03	0.03	1.00	0.04
"NoRel" FNR	0.07	0.04	0.05	0.00	0.04
Empirical coverage, $\alpha = 0.1$	0.915	0.928	0.914	0.920	0.894

We test our greedy approaches using $\lambda_{\text{int}} = 10$ and $\lambda = 0$ with $\alpha_{\text{int}} = 0.05, 0.10$. Similarly to CoDEx-S with $k = 2$, we additionally test the naive approach. In Table 6 we see that our greedy approaches outperform the naive approach in producing smaller $|C_{\text{2-hop},1-\alpha}|$, except for the sequential approach with $\alpha_{\text{int}} = 0.10$.

When $\alpha_{\text{int}} = 0.10$ with the sequential approach, we see unstable calibration, similar to the results for CoDEx-S when $\alpha_{\text{int}} = 0.10$. As before, this is due to $C_{\text{1-hop}}$ being too small, leading to $\boldsymbol{G}(h, t)$ undercovering and producing too many large nonconformity values as seen in Figure 4. Unlike our previous experiments, for this KG we see that the direct approach noticeably outperforms the sequential approach in producing smaller conformal 2-hop sets, at the cost of much large $C_{\text{1-hop}}$. Not only is the direct approach stable with $\alpha_{\text{int}} = 0.10$, but it has smaller conformal 2-hop relation sets of size 50 whereas the sequential approach

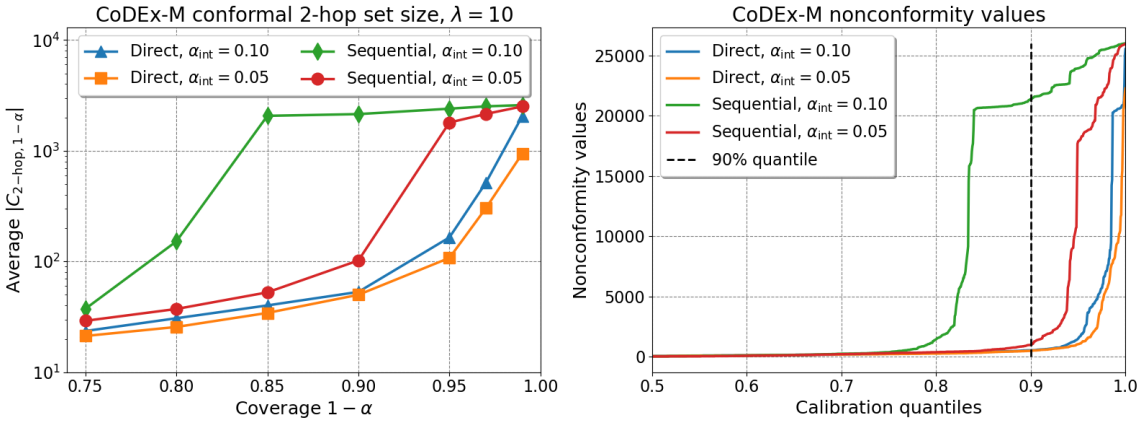

Figure 4: Conformal 2-hop relation set sizes for CoDEx-M and nonconformity values.

only goes down to 102 with $\alpha_{\text{int}} = 0.05$. This corresponds to $|C_{2\text{-hop},0.9}|/|\mathcal{R}'_2| = 0.019$ for the direct approach and 0.039 for the sequential approach, which is a substantial reduction in the 2-hop relation space.

Stability at $\alpha_{\text{int}} = 0.10$ for a larger KG corresponds to increased size of the 1-hop space of $\mathcal{R} \times \mathcal{E}$ for $C_{1\text{-hop}}$. As this space grows, it becomes easier to undercover with $C_{1\text{-hop}}$, and thus it is consistent that the direct approach which produces much large $C_{1\text{-hop}}$ than the sequential approach is more stable over α_{int} for a larger KG.

In Table 7 we show the proportion of 1-hop space $\mathcal{R} \times \mathcal{E}$ covered by $C_{1\text{-hop}}$ for CoDEx-S and CoDEx-M, for the direct and sequential approach, and $\alpha_{\text{int}} = 0.05, 0.10$. In Table 8 we show the proportion of k-hop relation space $\mathcal{R}'_k$ covered by $\text{supp}_{\text{eff}}(\boldsymbol{G})$, for the direct and sequential approach, for $\alpha_{\text{int}} = 0.05, 0.10$, and for CoDEx-S with $k = 2, 3$ and CoDEx-M with $k = 2$. We mark the unstable cases in Table 8 with an asterisk. Looking first at the proportion of 1-hop space covered, we see that decreasing α_{int} leads to more coverage as expected. But moving from CoDEx-S to CoDEX-M, the direct approach covers a larger percentage of the 1-hop space, whereas the sequential approach covers roughly the same percentage, despite there being many more entities in CoDEx-M.

Table 7: Proportion of 1-hop space covered, $|C_{1\text{-hop}}|/|\mathcal{R} \times \mathcal{E}|$, with $\lambda_{\text{int}} = 10$

	Direct		Sequential	
	$\alpha_{\text{int}} = 0.10$	$\alpha_{\text{int}} = 0.05$	$\alpha_{\text{int}} = 0.10$	$\alpha_{\text{int}} = 0.05$
CoDEx-S	0.024	0.036	0.0078	0.019
CoDEx-M	0.10	0.11	0.0073	0.014

However, we see that for CoDEx-M, the sequential approach covers more of the 2-hop relation space $\mathcal{R}'_k$. This suggests that the sequential approach is parsing more relations than the direct approach, but combined with the proportion of 1-hop space covered in Table 7, implies the sequential approach parses far fewer mid-path entities than the direct approach.

Table 8: Proportion of k-hop relation space covered, $|\operatorname{supp}_{\text{eff}}(\boldsymbol{G})|/|\mathcal{R}'_k|$, with $\lambda_{\text{int}} = 10$

	Direct		Sequential	
	$\alpha_{\text{int}} = 0.10$	$\alpha_{\text{int}} = 0.05$	$\alpha_{\text{int}} = 0.10$	$\alpha_{\text{int}} = 0.05$
CoDEx-S (k=2)	0.088	N/A	0.070	N/A
CoDEx-S (k=3)	0.094*	0.22	0.0865*	0.24
CoDEx-M (k=2)	0.21	0.28	0.23*	0.37

Since CoDEx-M and CoDEx-S have a similar number of relations (51 vs. 42), when adding more entities the direct approach adapts by increasing its 1-hop space coverage, while the sequential approach covers well in the relations with $C^{r\text{-seq}}_{\text{1-hop},1-\alpha_{\text{int}}}(e)$ but poorly covering with $C^{t\text{-seq}}_{\text{1-hop},1-\alpha_{\text{int}}}(e,r)$. As a consequence, the sequential approach seems to "miss" correct k-hop paths more often, and so even though $\operatorname{supp}_{\text{eff}}(\boldsymbol{G})$ is larger than the direct approach, it is placing probability mass on the wrong k-hop relations. Thus, although the sequential approach is typically "cheaper" than the direct approach as less of $\mathcal{R} \times \mathcal{E}$ must be covered by $C_{\text{1-hop}}$, it is more sensitive to instability, and may require more fine-tuning of α_{int}.

Remark 5 *In Section 2 and Section 3.4, we use RAPS for the single-hop case and for C_{end} in the multi-hop case, assuming there is at most one 1-hop relation for a given (h,t). In practice this may not be true, depending on the KG. For our benchmarks, less than 1% of (h,t) pairs in CoDEx-S dataset have more than one 1-hop relation between them. In the CoDEx-M dataset, this is less than 2%. However, if a KG has many 1-hop relations between h and t, it is straightforward to substitute RAPS with MLRAPS.*

5. Conclusion and future work

We have developed a greedily-constructed scoring function for conformal multi-hop relation detection and prediction on KGs which only requires a pre-trained scoring function on triples. Our scoring function is built by iteratively expanding a neighborhood from a target head h until the target tail t is reached, and only assigns relevant probability mass to parsed multi-hop relations. We introduced two approaches for expanding a neighborhood, direct and sequential, both of which use interior conformal prediction methods to build a set of 1-hop neighbors in relation and entity space. Numerical experiments on benchmark CoDEx KGs yield positive results, allowing for efficient parsing of multi-hops to produce reasonably sized conformal multi-hop relation sets for downstream analysis.

Future work includes further fine-tuning of our approach, including the introduction of variable interior parameter α_{int} which can be adjusted based on the choice of direct or sequential, as well as the current neighborhood size. Other work includes optimizing implementation for high-performance computing on much larger KG, investigating the computational complexity with respect to KG size and hop length k, as well as leveraging other conformal classification methods. While we have focused on using our modified MLRAPS algorithm, the framework established can be use with any multi-label conformal classification algorithm with nonconformity scoring $s(x, Y)$ where $x = (h,t)$ and $Y \subset \mathcal{R}'_k$ which can leverage $\boldsymbol{G}(h,t)$ as a heuristic approximation of uncertainty.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This document has been reviewed for release (LLNL-CONF-2005227).

References

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 22 (82):1–6, 2021.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.

Adil Bahaj and Mounir Ghogho. A step towards quantifying, modelling and exploring uncertainty in biomedical knowledge graphs. *Computers in Biology and Medicine*, 184: 109355, 2025.

Maxime Cauchois, Suyash Gupta, and John C. Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.*, 22(1), 2021.

Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. Embedding uncertain knowledge graphs. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

S. H. Zargarbashi, S. Antonelli, and A. Bojchevski. Conformal prediction sets for graph neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202. PMLR, 2023.

K. Huang, Y. Jin, and J. Candès, E.and Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2023.

Kostas Katsios and Harris Papadopulos. Multi-label conformal prediction with a mahalanobis distance nonconformity measure. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*, pages 522–535. PMLR, 2024.

Lysimachos Maltoudoglou, Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. Well-calibrated confidence measures for multi-label text classification with a large number of labels. *Pattern Recognition*, 122:108271, 2022.

Bo Ni. Reliable knowledge graph reasoning with uncertainty quantification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5463–5466. Association for Computing Machinery, 2024.

Harris Papadopoulos. A cross-conformal predictor for multi-label classification. In *Artificial Intelligence Applications and Innovations*, pages 241–250, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

Tara Safavi and Danai Koutra. CoDEx: A Comprehensive Knowledge Graph Completion Benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8328–8350. Association for Computational Linguistics, 2020.

Tong Shen, Fu Zhang, and Jingwei Cheng. A comprehensive overview of knowledge graph completion. *Knowledge-Based Systems*, 255:109597, 2022.

Chhavi Tyagi and Wenge Guo. Multi-label classification under uncertainty: A tree-based conformal prediction approach. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 488–512. PMLR, 2023.

Fangxin Wang, Yuqing Liu, Kay Liu, Yibo Wang, Sourav Medya, and Philip S. Yu. Uncertainty in graph neural networks: A survey. *Transactions on Machine Learning Research*, 2024.

Huazhen Wang, Xin Liu, Bing Lv, Fan Yang, and Yanzhu Hong. Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional chinese medicine. *PLOS ONE*, 9(6):1–14, 2014.

Huazhen Wang, Xin Liu, Ilia Nouretdinov, and Zhiyuan Luo. A comparison of three implementations of multi-label conformal prediction. In *Statistical Learning and Data Sciences*, pages 241–250, Cham, 2015. Springer International Publishing.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Y. Zhu, N. Potyka, J. Pan, B. Xiong, Y. He, E. Kharlamov, and S. Staab. Conformalized answer set prediction for knowledge graph embedding, 2025. arXiv:1410.5093.

Xiaohan Zou. A survey on application of knowledge graph. *Journal of Physics: Conference Series*, 1487(1):012016, 2020.