Credible Intervals for Knowledge Graph Accuracy Estimation

STEFANO MARCHESIN, University of Padua, Italy
GIANMARIA SILVELLO, University of Padua, Italy

Knowledge Graphs (KGs) are widely used in data-driven applications and downstream tasks, such as virtual assistants, recommendation systems, and semantic search. The accuracy of KGs directly impacts the reliability of the inferred knowledge and outcomes. Therefore, assessing the accuracy of a KG is essential for ensuring the quality of facts used in these tasks. However, the large size of real-world KGs makes manual triple-by-triple annotation impractical, thereby requiring sampling strategies to provide accuracy estimates with statistical guarantees. The current state-of-the-art approaches rely on Confidence Intervals (CIs), derived from frequentist statistics. While efficient, CIs have notable limitations and can lead to interpretation fallacies. In this paper, we propose to overcome the limitations of CIs by using *Credible Intervals* (CrIs), which are grounded in Bayesian statistics. These intervals are more suitable for reliable post-data inference, particularly in KG accuracy evaluation. We prove that CrIs offer greater reliability and stronger guarantees than frequentist approaches in this context. Additionally, we introduce *a*HPD, an adaptive algorithm that is more efficient for real-world KGs and statistically robust, addressing the interpretive challenges of CIs.

CCS Concepts: • **Information systems → Information integration**.

Additional Key Words and Phrases: Confidence intervals, Knowledge graphs, Bayesian statistics, Accuracy estimation

ACM Reference Format:
Stefano Marchesin and Gianmaria Silvello. 2025. Credible Intervals for Knowledge Graph Accuracy Estimation. *Proc. ACM Manag. Data* 3, 3 (SIGMOD), Article 142 (June 2025), 26 pages. https://doi.org/10.1145/3725279

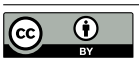
1 Introduction

Over the past decade, the emergence of large-scale Knowledge Graphs (KGs) like Wikidata [50], DBpedia [1], YAGO [19], and NELL [35] has opened up new avenues in database research [21] and empowered a range of machine learning applications, including virtual assistants [20, 36], search engines, recommender, and question answering systems [48, 49]. In this context, ensuring the accuracy of KG content is pivotal to delivering a high-quality user experience, as emphasized by applications like Saga [21], where the on-demand evaluation of the KG accuracy is a key feature.

The importance of assessing KG accuracy is further accentuated by the limitations of current KG construction processes, which often result in large volumes of incorrect facts, inconsistencies, and high sparsity [11, 45]. These issues not only undermine entity-oriented search applications by introducing potential misinformation [17, 33, 38], but they also hinder the performance of downstream tasks such as KG embeddings, which are integral to many machine learning applications [20, 36]. KG embeddings are sensitive to unreliable data and perform poorly on KGs extracted from noisy sources [45]; thus, assessing the KG accuracy is essential for making informed decisions about its use in both direct and downstream applications.

Auditing KG accuracy involves annotating facts with correctness labels. However, two major challenges arise when dealing with real-life KGs. First, obtaining high-quality labels is expensive [14,

Authors' Contact Information: Stefano Marchesin, stefano.marchesin@unipd.it, University of Padua, Padua, Italy; Gianmaria Silvello, gianmaria.silvello@unipd.it, University of Padua, Padua, Italy.



ACM 2836-6573/2025/6-ART142
https://doi.org/10.1145/3725279

41]. Secondly, annotating every fact in a large-scale KG is unfeasible [46]. Real-life KGs encompass hundreds of millions, or even billions, of relational facts, typically represented as (s, p, o) triples. Therefore, cost-effective and scalable solutions are crucial. Given the importance of efficiently evaluating KG accuracy, several studies have begun addressing these challenges [54], framing the KG accuracy evaluation as a constrained minimization problem [14, 31, 46]. The state-of-the-art methods employ iterative procedures involving sampling strategies for efficient data collection, point estimators to determine accuracy, and $1 - \alpha$ intervals to quantify the uncertainties associated with the sampling procedure.

For $1 - \alpha$ intervals, where α represents the risk of concluding that a difference exists between the estimated and true KG accuracy, current approaches to KG accuracy estimation rely on Confidence Intervals (CIs) [14, 31, 46]. CIs are rooted in frequentist statistics, where probability is interpreted as the long-term frequency of events over repeated sampling. Thus, the $1 - \alpha$ probability pertains to the reliability of the interval estimation procedure (stochastic process) rather than to the specific interval computed given a sample (single event). Once a CI is built from the sample, it either includes the true (but unknown) KG accuracy or it does not – it is no longer a matter of probability. However, in real-life scenarios, where KG accuracy is typically assessed via a single, non-repeated process, there is instead a need for efficient $1 - \alpha$ intervals that can be reliably interpreted in a *one-shot* setting. To further reduce the appeal of CIs for the considered task, recent studies have also highlighted that their use can lead to interpretation fallacies regarding confidence, precision, and likelihood [37].

Contributions. The core contributions of this work are:

- We highlight the limitations of existing KG evaluation methods relying on CIs and pioneer the use of Bayesian Credible Intervals (CrIs) for KG accuracy evaluation. By framing the task in Bayesian terms for the first time, we show that CrIs offer interpretable results at minimal costs, avoiding CI interpretation fallacies and providing reliable solutions in a one-shot setting.
- We consider two main families of CrIs: Equal-Tailed (ET) and Highest Posterior Density (HPD) intervals. We formally prove that HPD CrIs provide the optimal representation of parameter values that are most consistent with the data across all annotation scenarios in KG accuracy evaluation. This ensures comprehensive applicability across diverse KG settings, using both informative and uninformative priors.
- To address the challenge of choosing the right prior, one of the most well-known resistances to using Bayesian methods in statistics, we introduce the *adaptive* HPD (*a*HPD) algorithm. *a*HPD exploits multiple priors concurrently to generate different $1 - \alpha$ HPD intervals, competing to achieve convergence in the minimization problem. This ensures that the most efficient outcome is always obtained from competing solutions without choosing a specific prior "a priori".
- Finally, we show that *a*HPD outperforms the state-of-the-art approaches with real and synthetic data at both small and large scales, proving robust regardless of the precision level required by the evaluation process and reducing annotation costs by up to 47% in high-precision scenarios.

Although CI fallacies and Bayesian CrIs are well-established concepts in statistical domains, they remain largely unfamiliar to the database community. This work bridges this gap by emphasizing their relevance for data quality management and providing a solution that addresses the interpretative fallacies currently affecting KG quality estimation. Moreover, by removing the need for prior selection, our newly proposed *a*HPD algorithm enables analysts to evaluate KG accuracy with minimal statistical expertise. This makes the approach not only interpretable and efficient but also

user-friendly. *a*HPD is flexible and general enough to accommodate scenarios where (a) analysts have no prior knowledge about KG accuracy, or (b) they possess reliable prior information, such as insights derived from similar KGs. This versatility ensures broad applicability while maintaining ease of use across diverse settings.

Outline. Section 2 introduces the problem, its optimization, and the state-of-the-art sampling techniques and their corresponding estimators. Section 3 discusses CIs and their limitations for KG accuracy evaluation. Section 4 delves into CrIs, highlights the challenges of prior selection, and presents the *a*HPD algorithm as an effective solution. Sections 5 and 6 detail the experimental setup and results, respectively. Section 7 reviews related work. Section 8 concludes the paper and outlines future research.

2 Background

We present the problem, outline its optimization via an iterative evaluation framework, and report the state-of-the-art sampling techniques adopted in this work and the respective estimators.

2.1 Preliminaries

Following the notation introduced by Bonifati et al. [6], we consider KGs as ground RDF graphs defined as $G = (V, R, \eta)$, where $V = \{E \cup A\}$ is the set of nodes, E the set of entities and A of attributes. R denotes the set of relationships between nodes and $\eta : R \rightarrow E \times \{E \cup A\}$ is the function that assigns an ordered pair of nodes to each relationship. The η function generates the ternary relation T of G, which consists of the set of (s, p, o) triples where $s \in E$, $p \in R$, and $o \in \{E \cup A\}$, with a size $M = |T|$. Given T, we define an entity cluster $C_e = \{(s, p, o) \in T \mid s = e\}$ as the set of triples that share the same subject $e \in E$.

We treat triples as first-class citizens of a KG; hence, we redefine a KG as $G = (V, R, T, \eta)$. A triple is also called a fact; we use the two terms interchangeably.

2.2 Problem: Constrained Minimization

Let the correctness of a triple $t \in T$ be represented by an indicator function $\mathbb{1}(t)$, where $\mathbb{1}(t) = 1$ indicates that the triple is correct, and $\mathbb{1}(t) = 0$ indicates it is incorrect. The accuracy of a KG can then be defined as the proportion of correct triples τ within G:

$$\mu = \frac{\sum_{t \in T} \mathbb{1}(t)}{M} = \frac{\tau}{M} \tag{1}$$

where the value of $\mathbb{1}(t)$ is determined through manual annotation. Following [14, 31], we consider correctness a binary problem, similar to triple validation [13], since an atomic fact is correct or incorrect.

However, manually evaluating every triple in a large-scale KG to determine its accuracy is impractical. Therefore, the standard approach is to estimate μ using an estimator $\hat{\mu}$, calculated from a relatively small sample drawn according to a sampling strategy $\mathcal{S}$, which selects a subset $T_{\mathcal{S}} \subset T$ of triples to annotate. This results in a sample $G_{\mathcal{S}} = (V_{\mathcal{S}}, R_{\mathcal{S}}, T_{\mathcal{S}}, \eta)$, where $V_{\mathcal{S}} \subset V$ and $R_{\mathcal{S}} \subset R$.

To ensure the accuracy of G, the estimator $\hat{\mu}$ must be unbiased, meaning $E[\hat{\mu}] = \mu$. Additionally, since $\hat{\mu}$ is a point estimator, it is necessary to provide a $1 - \alpha$ interval to quantify the uncertainties associated with the sampling procedure. The larger the sample $G_{\mathcal{S}}$, the narrower the interval until it reaches zero width when the sample is equivalent to G. A key measure related to intervals is the Margin of Error (MoE), which is half the width of the interval.

Let $G_{\mathcal{S}} = \mathcal{S}(G)$ be a sample drawn using the sampling strategy $\mathcal{S}$, and $\hat{\mu}$ an estimator of μ based on that sample. Let $\text{cost}(G_{\mathcal{S}})$ denote the cost of manually evaluating the correctness of elements in

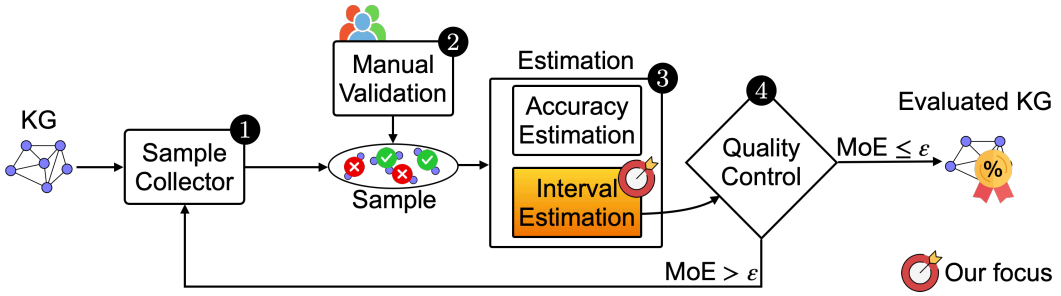


Fig. 1. Efficient KG accuracy evaluation framework [31].

the sample. Accordingly to [14, 31], we define the KG accuracy evaluation task as a constrained minimization problem:

Problem. Given a KG G and an upper bound ε for the MoE of a $1-\alpha$ interval:

$$\begin{aligned} \underset{\mathcal{S}}{\text{minimize}} \quad & \text{cost}(G_{\mathcal{S}}) \\ \text{subject to} \quad & E[\hat{\mu}] = \mu, \text{MoE} \leq \varepsilon \end{aligned}$$

2.3 Optimization: Iterative Procedure

The minimization problem is optimized with an evaluation framework that operates as the iterative procedure shown in Figure 1.

In phase ❶, a small batch of triples is drawn from the KG according to a chosen sampling strategy $\mathcal{S}$. This strategy should be selected to minimize the cost function, representing the optimization objective. In phase ❷, manual annotations are obtained for these newly sampled triples and combined with any existing annotations. In phase ❸, using the annotations accumulated via the chosen sampling design, an unbiased estimator computes $\hat{\mu}$, the estimate of the KG accuracy. The annotated sample also constructs the corresponding $1-\alpha$ interval, quantifying the uncertainties in the sampling procedure. In phase ❹, a quality control step checks whether the generated interval meets the predefined upper bound, ensuring that MoE $\leq \varepsilon$. If the criterion is satisfied, the process concludes, and the accuracy estimate and the corresponding $1-\alpha$ interval are reported. If not, the process returns to phase ❶. Thus, the framework iteratively samples and estimates, stopping when the interval meets the specified threshold ε. Doing so avoids oversampling and unnecessary manual annotations, ensuring accurate and cost-effective estimates.

The fact that the problem is solved once the interval is sufficiently narrow highlights the pivotal role of **$1-\alpha$ intervals** in the optimization process. Given a sampling strategy, faster-shrinking intervals allow the problem to be solved with smaller samples, thereby reducing the number of required annotations compared to other, slower-to-converge intervals. At the same time, it is also essential that $1-\alpha$ intervals are reliable and guarantees the statistical properties they (nominally) entail, as first outlined in [31]. However, to ensure reliability in the evaluation process, the approach by Marchesin and Silvello [31] required a trade-off, thus sacrificing some efficiency. In this work, by shifting the interval estimation framework from frequentist to Bayesian statistics, we demonstrate that it is possible to obtain intervals that are guaranteed to be the shortest – and thus the most efficient – while also providing probabilistic solid guarantees on their reliability.

2.4 Sampling and Accuracy Estimation

The approaches proposed to efficiently evaluate KG accuracy [14, 31, 46] adopt simple and cluster random sampling techniques, paired with well-known point estimators [10]. Below, we detail both sampling strategies and the corresponding unbiased estimators.

Simple Random Sampling. Simple Random Sampling (SRS) draws a sample of n_S triples from G without replacement. In large-scale KGs, the probability of choosing the same triple more than once is low, making sampling with replacement a good approximation to sampling without replacement [9] and a practical solution.

Under SRS, the estimator for μ is the sample proportion:

$$\hat{\mu}_{\text{SRS}} = \frac{\sum_{i=1}^{n_S} \mathbb{1}(t_i)}{n_S} = \frac{\tau_S}{n_S} \tag{2}$$

with estimation variance given by $V(\hat{\mu}_{\text{SRS}}) = \frac{\hat{\mu}_{\text{SRS}}(1-\hat{\mu}_{\text{SRS}})}{n_S}$.

When SRS is applied, $\hat{\mu}_{\text{SRS}}$ is an unbiased estimator of μ [10].

Cluster Random Sampling. Cluster sampling is a cost-effective alternative to SRS when estimating the accuracy of large-scale KGs [14, 31, 46]. Among the different cluster sampling strategies that can be employed, Two-stage Weighted Cluster Sampling (TWCS) represents the state-of-the-art for KG accuracy evaluation [14, 31]. TWCS is divided into two stages. In the first stage, TWCS samples n_C entity clusters with probabilities π_i proportional to their sizes. That is, by denoting the size of the ith entity cluster as $M_i = |C_{e_i}|$, the cluster selection probability can be defined as $\pi_i = M_i/M$. In the second stage, TWCS draws $\min\{M_i, m\}$ triples from each sampled cluster using SRS without replacement.

Let $\hat{\mu}_i$ be the (estimated) accuracy of the ith entity cluster. The estimator for μ under TWCS is:

$$\hat{\mu}_{\text{TWCS}} = \frac{\sum_{i=1}^{n_C} \hat{\mu}_i}{n_C} \tag{3}$$

with estimation variance $V(\hat{\mu}_{\text{TWCS}}) = \frac{1}{n_C(n_C-1)} \sum_{i=1}^{n_C} (\hat{\mu}_i - \hat{\mu}_{\text{TWCS}})^2$.

When TWCS is applied with a second-stage sample of approximately size m, $\hat{\mu}_{\text{TWCS}}$ is an unbiased estimator of μ [10].

3 State of the Art: Confidence Intervals (CIs)

Previous research on efficient KG accuracy evaluation has utilized CIs to estimate $1-\alpha$ intervals [14, 31, 46]. However, no universal formula exists for constructing CIs under arbitrary sampling strategies (and point estimators). Instead, multiple approaches have been developed [8], each offering CIs with distinct properties [37]. Among these, we review only the methods employed by state-of-the-art KG accuracy estimation techniques — specifically, the Wald and Wilson approaches [9, 53].

This section outlines Wald and Wilson methods for constructing $1-\alpha$ CIs. Next, we discuss the common misinterpretations and the resulting fallacies regarding CIs concerning their confidence, precision, and likelihood properties [37], which are desirable characteristics for tasks where analysts seek reasonable post-data inference, such as in KG accuracy estimation.

3.1 The Wald Method

The Wald method relies on normal approximation to build $1-\alpha$ CIs and involves inverting the acceptance region of the Wald large-sample normal test [9]:

$$\left| \frac{\hat{\mu}_{\mathcal{S}} - \mu}{\sqrt{V(\hat{\mu}_{\mathcal{S}})}} \right| \leq z_{\alpha/2} \tag{4}$$

where μ represents the true KG accuracy, $\hat{\mu}_{\mathcal{S}}$ and $V(\hat{\mu}_{\mathcal{S}})$ the estimated KG accuracy and its variance under a given sampling strategy $\mathcal{S}$, and $z_{\alpha/2}$ the critical value of the standard normal distribution for a confidence level $1-\alpha$.

By solving Equation (4) in μ, we obtain the $1-\alpha$ Wald interval:

$$\hat{\mu}_{\mathcal{S}} \pm z_{\alpha/2}\sqrt{V(\hat{\mu}_{\mathcal{S}})} \tag{5}$$

The Wald interval is an appealing CI due to its simplicity. However, when applied to binomial proportions, it is known to incur overshooting and zero-width intervals [8], also presenting an erratic behavior when used to gauge KG accuracy [31]. These phenomena aggravate in KGs with accuracy values deviating from the center of the (accuracy) range – where normal approximation works best – towards its boundaries, that is, one or zero. Since many real-life KGs, such as YAGO, DBpedia, and Wikidata, involve a certain degree of manual curation, they rarely have accuracy values as low as 0.5. Conversely, fully automated KGs, which are noisier and more error-prone, frequently present accuracy values well below 0.5 [45]. Thus, this tendency of real-life KGs towards the boundaries of the accuracy range makes Wald an often unreliable choice for KG accuracy estimation.

3.2 The Wilson Method

To overcome Wald limitations, Marchesin and Silvello [31] relied on the Wilson method [53], which can be used to build $1-\alpha$ binomial CIs. Similarly to Wald, the method involves inverting the acceptance region of the significance test in Equation 4, but requires two changes. First, it requires assuming that the sampling strategy used to collect the sample is SRS. Secondly, it requires replacing the estimated standard error with the null standard error. Based on these changes, the test can thus be rewritten as:

$$\left| \sqrt{\frac{n_{\mathcal{S}}}{\mu(1-\mu)}} \cdot (\hat{\mu}_{\text{SRS}} - \mu) \right| \leq z_{\alpha/2} \tag{6}$$

By solving Equation (6) in μ, we get the $1-\alpha$ Wilson interval:

$$\frac{\hat{\mu}_{\text{SRS}} + \frac{z^2_{\alpha/2}}{2n_{\mathcal{S}}}}{1 + \frac{z^2_{\alpha/2}}{n_{\mathcal{S}}}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n_{\mathcal{S}}}} \cdot \sqrt{\frac{\hat{\mu}_{\text{SRS}}(1-\hat{\mu}_{\text{SRS}})}{n_{\mathcal{S}}} + \frac{z^2_{\alpha/2}}{4n^2_{\mathcal{S}}}} \tag{7}$$

which, compared to the Wald interval, consists of a relocated center estimate (left) and a corrected standard deviation (right).

Wilson intervals correct the erratic behavior of Wald ones, resulting in higher reliability, although at the expense of a drop in efficiency as accuracy approaches its boundaries [31]. Note that design effect adjustments are also required to adapt Wilson to complex sampling strategies [25, 26].

3.3 Confidence Intervals (CIs) Limitations

CIs fall under frequentist statistics, where probability is interpreted as the long-term frequency of events across repeated sampling. This means that, when building a CI, we are not directly estimating the probability that the parameter of interest – the KG accuracy in our case – lies within

the interval. Instead, we estimate how frequently such intervals, across repeated trials, will catch the true parameter. This is often referred to as *frequency probability*.

Hence, CIs interpretation can be counterintuitive and, for this reason, is often misunderstood by practitioners [18]. Three major fallacies exist in interpreting CIs, as shown by Morey et al. [37]. Although these fallacies do not affect the quantitative outcomes of KG accuracy evaluation, they pose a significant risk to the proper use of CIs. Misinterpretations stemming from them can lead to conclusions as flawed as those caused by quantitative inaccuracies. In the following, we first outline these fallacies and then provide a practical example where they occur.

Fallacy 1. The probability that a $1-\alpha$ CI contains the true parameter is $1-\alpha$ for a particular observed interval.

This is a fallacious interpretation because once a CI is constructed, the probability that it contains the true parameter is either 1 (if it does) or 0 (if it does not). The $1-\alpha$ confidence level represents the reliability of the interval estimation procedure across repeated samples rather than the probability that any specific interval captures the true parameter. This misconception stems from the belief that a $1-\alpha$ CI provides the likelihood that the unknown parameter falls within the observed interval [18]. However, the $1-\alpha$ confidence level does not assess whether the specific interval computed from the data includes the true parameter. CIs rely on pre-data assumptions not intended for post-data inference [40].

Fallacy 2. The width of a CI reflects the precision of our knowledge about the parameter.

This interpretation is flawed because there is no inherent relationship between the width of a CI and the precision of the parameter estimate. The $1-\alpha$ confidence level reflects the reliability of the interval estimation procedure, not the precision of a particular observed interval. Hence, it is impossible to know whether a specific narrow or wide interval belongs to the $1-\alpha$ proportion that reliably captures the true parameter or the α proportion that does not. Consequently, interpreting CI width as a measure of precision can lead to incorrect post-data conclusions [37].

Fallacy 3. A CI includes the likely values of the parameter.

While a CI is constructed to have a fixed $1-\alpha$ probability of capturing the true parameter across repeated samples, this does not imply that any particular observed interval contains the most likely values for the parameter. Once the interval is calculated, it either includes the true parameter or does not [39]. Indeed, CIs can sometimes exclude many reasonable values or, in extreme cases, be infinitesimally narrow or even empty (e.g., Wald intervals [8]), thereby excluding all possible values [37].

Example 1. Let us consider an analyst who wants to estimate the accuracy of NELL ($\mu = 0.91$, see Table 1) using the evaluation framework described in Section 2.3. The analyst adopts the most straightforward approach with SRS for sampling, the Wald method with a significance level $\alpha = 0.05$ for interval estimation, and a MoE threshold of $\varepsilon = 0.05$ for convergence. Under this setup, the evaluation procedure halts (i.e., MoE $\leq \varepsilon$) with a sample size of $n_{\mathcal{S}} = 30$ and an estimated accuracy of $\hat{\mu}_{\text{SRS}} = 1.00$ – reasonable outcomes given the underlying KG accuracy.[1] With these values, the estimation variance is $V(\hat{\mu}_{\text{SRS}}) = 0.00$. Substituting this into Equation (5) produces a zero-width interval: CI $= [1.00, 1.00]$, meaning that there is an absolute certainty that the estimated accuracy $\hat{\mu}_{\text{SRS}}$ is correct. However, this also implies that the probability that the obtained CI contains the true accuracy cannot be $1-\alpha$. Therefore, placing $1-\alpha$ confidence in an interval that assumes absolute certainty would clearly be a mistake for the analyst (**Fallacy 1**). At the same time, a

[1] In our experiments, these results occurred in 7% of the 1,000 iterations run on NELL.

zero-width CI does not imply that effect size is determined with perfect precision, as that would require annotating the entire population, nor can it imply that there is a 0.95 ($1 - \alpha$) probability that μ is exactly 1.00, thus leading to **Fallacy 2**. Finally, a zero-width interval excludes all the possible values the accuracy can take, resulting in **Fallacy 3**. Hence, these fallacies highlight that the post-data properties required to make valid inferences from the interval are absent, potentially leading to arbitrary conclusions.

Task-Specific Deterrents. The frequentist assumption underlying CIs – and the corresponding interpretation fallacies this can generate – makes these intervals not ideal for KG accuracy evaluation. In fact, CIs offer guarantees over repeated trials rather than for a specific instance. However, in real-life scenarios, KG accuracy evaluation is typically conducted as a single, non-repeated process. Hence, CIs are inconvenient and $1 - \alpha$ intervals providing a reliable interpretation in a one-shot setting are very much desired.

Moreover, the lack of a unified framework for constructing CIs contributes to their complexity. Different CIs are generated using distinct construction methods, often relying on intricate procedures that involve assumptions and approximations that are rarely tested [30]. Additionally, CIs constructed through different methods can vary in efficiency and reliability across different regions of the accuracy space [31]. Therefore, establishing a unified framework that enables the seamless computation of multiple $1 - \alpha$ intervals would be crucial for improving consistency and comparability.

Finally, the long-term properties of CIs necessitate additional metrics to assess their reliability, such as coverage probability [31]. However, calculating this probability requires repeated iterations of the entire KG accuracy evaluation procedure, making it impractical for real-world applications.

4 Proposed Solution: Credible Intervals (CrIs)

To overcome the limitations of CIs for KG accuracy evaluation, we can resort to Credible Intervals (CrIs) [12]. These intervals are grounded in Bayesian statistics and operate on the principle that estimated parameters are random variables following a distribution.

Bayesian statistics differs from the frequentist approach as it is a post-data theory [37]. It uses the information from the data, combined with model assumptions and prior information, to determine what is reasonable to believe [15, 52].

In the Bayesian framework, probability represents a degree of belief in an event, captured by a distribution. Before observing data, this belief is expressed through a prior distribution, reflecting initial assumptions or knowledge about the parameter of interest. For instance, when estimating KG accuracy, the prior distribution encodes beliefs about likely accuracy values based on prior knowledge or judgment. Once data is collected, this initial belief is updated through the likelihood, representing the posterior probability – the focus of Bayesian statistics – from the observed data. This probability is obtained by conditioning the prior probability with the information summarized by the likelihood via Bayes' rule [7].

A CrI is an interval derived from the posterior distribution, indicating the range within which an unobserved parameter value will likely fall with a certain probability. Since CrIs are defined on the posterior distribution, a $1 - \alpha$ CrI has the following properties:

(1) It contains the parameter of interest with $1 - \alpha$ probability (addressing Fallacy 1).
(2) Its width reflects the precision of the estimate (addressing Fallacy 2).
(3) It covers the plausible values of the parameter (addressing Fallacy 3). Thus, Bayesian interval estimation avoids CIs interpretation fallacies by providing $1 - \alpha$, interpretable and reliable intervals in a one-shot setting.

We introduce two families of CrIs: ET and HPD intervals [7]. We demonstrate that HPD intervals are the shortest possible, providing the best representation of parameter values most consistent with the data. Additionally, we show that ET and HPD intervals are equivalent when the posterior distribution is symmetric. Finally, we address the challenges of selecting an appropriate prior by proposing a new adaptive solution that eliminates the need for manual prior selection while ensuring optimal performance.

4.1 Bayesian Interval Estimation

We model the process of annotating triples for correctness as a binomial process $\text{Bin}(n_S, \mu)$, where $n_S \geq 1$ is the number of annotated triples and μ represents the binomial proportion (cf. Eq. (1)). Since beta distributions are the standard conjugate priors for binomial distributions, using them for inference on μ is suitable [3].

Let $\tau_S \sim \text{Bin}(n_S, \mu)$, and assume that μ has a prior distribution $\text{Beta}(a, b)$, where $a, b > 0$ represent our prior belief regarding correct (a) and incorrect (b) triples in the KG. Due to the conjugate relationship between beta and binomial distributions, the posterior distribution of μ can be derived as $\text{Beta}(a + \tau_S, b + n_S - \tau_S)$ [47]. This means that to obtain the beta posterior, we simply update the parameters of the beta prior by adding the observed number of correct (τ_S) and incorrect ($n_S - \tau_S$) triples.

Let $F(\mu \mid G_S)$ denote the posterior Cumulative Distribution Function (CDF) of μ. For a given α, we can identify an interval (l, u) such that:

$$1 - \alpha = \Pr(l \leq \mu \leq u \mid G_S) = F(u) - F(l) \tag{8}$$

Therefore, a $1 - \alpha$ CrI is an interval (l, u) that satisfies the condition $1 - \alpha = F(u) - F(l)$ [44]. However,it remains to decide how to compute such an interval.

4.2 Equal-Tailed (ET) Intervals

A common solution is that of Equal-Tailed (ET) intervals [34]. These are $1 - \alpha$ CrIs that allocate equal probability to the tails of the posterior distribution – i.e., below the lower bound l and above the upper bound u. In other words, a $1 - \alpha$ ET CrI is obtained by taking the central $1 - \alpha$ region of the posterior distribution, leaving $\alpha/2$ probability in each tail. By denoting $\text{qBeta}(i; \cdot, \cdot)$ as the ith quantile of a beta distribution, we define the $1 - \alpha$ ET CrI as:

$$\begin{aligned} l &= \text{qBeta}(\alpha/2; a + \tau_S, b + n_S - \tau_S) \\ u &= \text{qBeta}(1 - \alpha/2; a + \tau_S, b + n_S - \tau_S) \end{aligned} \tag{9}$$

While ET CrIs are intuitive and effective for symmetric posteriors, they become suboptimal for skewed distributions because, by design, they allocate $\alpha/2$ of the distribution in each tail. To illustrate this, consider Figure 2, which shows three scenarios with increasing skewness in the posterior distribution for KG accuracy. In the symmetric case (Figure 2(a)), the ET CrI captures the most probable values (purple region), providing a good summary of the distribution. However, as skewness increases (Figures 2(b,c)), the ET CrI includes lower-probability parameter values (red region) while excluding parts of the HPD region (blue region). This results in a longer interval than necessary to satisfy the condition $1 - \alpha = F(u) - F(l)$.

To further support this, we compared the CDF of the HPD region excluded by the ET CrI with the CDF of any equally wide region covered by ET but falling outside the HPD region. In the moderately skewed case (Figure 2(b)), the CDF of these regions falling outside the HPD region, but covered by the ET CrI, is always less than 75% of the CDF of the HPD region not covered by the ET CrI. In the highly skewed case (Figure 2(c)), the CDF of these non-HPD regions is not even the 20% of the CDF of the excluded HPD region. We can see that, in both cases, the exclusion of portions

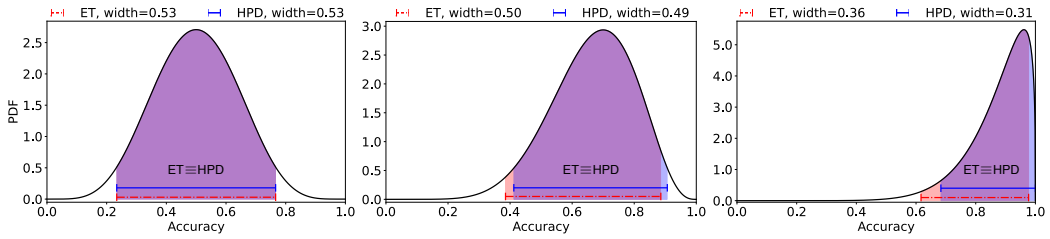


Fig. 2. Comparison of ET and HPD CrIs across three posterior distributions for KG accuracy with increasing skewness (from left to right). In panel (a), where the posterior is symmetric, both ET and HPD intervals capture the most probable values (purple region). However, as skewness increases in panels (b) and (c), the ET interval deviates from the highest posterior density region, including less likely parameter values (red region), and resulting in unnecessarily longer intervals to satisfy the $1-\alpha = F(u) - F(l)$ condition. In contrast, the HPD intervals remain optimal, covering only the highest posterior density region (purple and blue regions), thereby highlighting the superior performance of HPD in skewed distributions compared to ET.

of the HPD region forces the ET interval to expand in width to satisfy the $1-\alpha = F(u) - F(l)$ constraint, making it a suboptimal choice.

Moreover, for skewed posteriors, the inability of ET CrIs to fully cover the HPD region makes them vulnerable to Fallacy 3. Thus, for KG accuracy evaluation, where skewed distributions are common, ET limitations are of significance.

4.3 Highest Posterior Density (HPD) Intervals

The $1-\alpha$ CrI that best describes the posterior distribution is the interval containing the most probable posterior values [34], i.e., the Highest Posterior Density (HPD) interval.

Definition. Let $f(\mu|G_{\mathcal{S}})$ represent the Probability Density Function (PDF) of the posterior distribution. A $1-\alpha$ CrI is an HPD interval if it satisfies the following two conditions:

(1) $F(u) - F(l) = 1-\alpha$;
(2) For $l \leq \mu \leq u$, $f(\mu|G_{\mathcal{S}})$ is greater than for any other interval where condition (1) holds.

The first condition corresponds to the general definition of a $1-\alpha$ CrI (see Eq. (8)), while the second ensures that every point inside the HPD interval has higher PDF value than points outside. Given these conditions, within the Bayesian framework designed in Section 4.1, the HPD interval is the smallest possible and unique. As we will see in Section 6, these properties translate into more cost-effective solutions for KG accuracy evaluation.

First, we prove both properties for the **standard case**, where the number of correct ($\tau_{\mathcal{S}}$) and incorrect ($n_{\mathcal{S}} - \tau_{\mathcal{S}}$) triples resulting from the annotation process are nonzero. Due to space limits, we report sketch proofs. The complete proofs are in the online repository.[2]

Theorem 1. Given a prior distribution $\text{Beta}(a, b)$ for μ and a KG correctness annotation process, where $0 < \tau_{\mathcal{S}} < n_{\mathcal{S}}$, the $1-\alpha$ HPD interval is the smallest (l, u) interval satisfying the condition $F(u) - F(l) = 1-\alpha$.

Sketch Proof. When $0 < \tau_{\mathcal{S}} < n_{\mathcal{S}}$, the beta posterior $\text{Beta}(a+\tau_{\mathcal{S}}, b+n_{\mathcal{S}}-\tau_{\mathcal{S}})$ is unimodal and continuous over $[0,1]$, ensuring the existence of a CrI. To find the smallest interval satisfying $F(u) - F(l) = 1-\alpha$, we apply the method of Lagrange multipliers, which minimizes the interval

[2]https://github.com/KGAccuracyEval/credible-intervals4kg-estimation

length $(u - l)$ under the constraint that the posterior probability over $[l, u]$ equals $1 - \alpha$. That is, we minimize the Lagrangian function $\mathcal{L} = (u - l) + \lambda\Big(\int_l^u f(\mu \mid G_S)\, d\mu - (1 - \alpha)\Big)$, where λ is the Lagrange multiplier.

From the first-order conditions $\partial\mathcal{L}/\partial l = 0$ and $\partial\mathcal{L}/\partial u = 0$, we derive $f(l \mid G_S) = f(u \mid G_S) = -\frac{1}{\lambda}$. We compute the second-order derivatives to verify that (l, u) minimizes $\mathcal{L}$. The unimodal nature of the posterior ensures that the derived Hessian matrix is positive definite, confirming that (l, u) is the smallest interval satisfying $F(u) - F(l) = 1 - \alpha$. □

Theorem 2. Under the assumptions of Theorem 1, the $1 - \alpha$ HPD interval is unique.

Sketch Proof. Let (l^*, u^*) represent the HPD interval. Assume by contradiction that there exists another interval $(l', u') \neq (l^*, u^*)$ such that $(u' - l') = (u^* - l^*)$ and $\int_{l'}^{u'} f(\mu \mid G_S)\, d\mu = 1 - \alpha$. Since the posterior is unimodal, Th. 1 ensures that the HPD interval contains the mode. Similarly, the alternative interval (l', u') must also contain the mode. Hence, given that $(l', u') \neq (l^*, u^*)$ but $(u' - l') = (u^* - l^*)$, two cases arise: (i) $l' < l^*$ and $u' < u^*$; or (ii) $l' > l^*$ and $u' > u^*$. We prove case (i); case (ii) is analogous.

Since both intervals satisfy the condition $F(u) - F(l) = 1 - \alpha$, then it must hold that $\int_{l'}^{u'} f(\mu \mid G_S)\, d\mu = \int_{l^*}^{u^*} f(\mu \mid G_S)\, d\mu$. Removing the common terms yields: $\int_{l'}^{l^*} f(\mu \mid G_S)\, d\mu = \int_{u'}^{u^*} f(\mu \mid G_S)\, d\mu$. However, since l' lies outside the HPD region while u' lies inside, it follows that $\int_{l'}^{l^*} f(\mu \mid G_S)\, d\mu < \int_{u'}^{u^*} f(\mu \mid G_S)\, d\mu$, which contradicts the equality assumption. □

To compute the $1 - \alpha$ HPD CrI, we need to find the (l, u) interval that minimizes the Lagrangian function $\mathcal{L}$ defined in Th. 1. We resort to sequential quadratic programming by adopting the Sequential Least SQuares Programming (SLSQP) optimizer [27], that minimizes a Lagrangian of several variables with any combination of bounds, equality, and inequality constraints. SLSQP reformulates the optimization problem into a sequence of quadratic subproblems whose objectives are second-order approximations of the original Lagrangian and whose constraints are linearized versions of the original ones. The method then applies globalization techniques to guarantee convergence.

In our case, the objective of the Lagrangian is minimizing the (l, u) interval width, with l and u bounded to the $[0, 1]$ probability domain. The enforced (equality) constraint is $\int_l^u f(\mu \mid G_S)\, d\mu = 1 - \alpha$. To accelerate convergence, we use the corresponding $1 - \alpha$ ET CrI as an initial guess for the SLSQP optimizer.

Limiting Cases. The minimality and uniqueness properties of the $1 - \alpha$ HPD interval can be extended to limiting cases where the beta posterior exhibits exponentially increasing or decreasing behavior. These cases arise when the beta prior is uninformative – i.e., $a = b \leq 1$ – and the annotation process yields either all correct ($\tau_S = n_S$) or all incorrect ($\tau_S = 0$) triples.

When $\tau_S = n_S$, resulting in an exponentially increasing posterior, the $1 - \alpha$ HPD interval is computed as:

$$l = \text{qBeta}(\alpha; a + \tau_S, b);\ u = 1 \tag{10}$$

Likewise, when $\tau_S = 0$, producing an exponentially decreasing posterior, the $1 - \alpha$ HPD interval becomes:

$$l = 0;\ u = \text{qBeta}(1 - \alpha; a, b + n_S) \tag{11}$$

As a corollary to Theorem 1, we prove that the $1 - \alpha$ HPD interval is minimal in both limiting cases.

Corollary 1. Given an uninformative Beta(a, b) prior for μ and a KG annotation process resulting in either all correct ($\tau_{\mathcal{S}} = n_{\mathcal{S}}$) or all incorrect ($\tau_{\mathcal{S}} = 0$) triples, the $1 - \alpha$ HPD interval is the smallest (l, u) interval satisfying $F(u) - F(l) = 1 - \alpha$.

Proof. We prove $\tau_{\mathcal{S}} = 0$, the proof for $\tau_{\mathcal{S}} = n_{\mathcal{S}}$ is symmetrical. With an uninformative prior (i.e., $a = b \leq 1$) and an annotation outcome $\tau_{\mathcal{S}} = 0$, the resulting beta posterior has parameters $a \leq 1$ and $b + n_{\mathcal{S}} > 1$, exhibiting an exponentially decreasing behavior with the highest density in $x = 0$. Due to the monotonicity of the beta posterior, for any $x_1 < x_2$, we have $f(x_1 \mid G_{\mathcal{S}}) > f(x_2 \mid G_{\mathcal{S}})$. Thus, the HPD interval, with a lower bound at 0 and an upper bound at the $1 - \alpha$ quantile, is by construction the shortest interval containing μ with probability $1 - \alpha$. □

Similarly, we prove its uniqueness as a corollary to Theorem 2.

Corollary 2. Under the assumptions of Corollary 1, the $1 - \alpha$ HPD interval is unique.

Sketch Proof. The proof follows the same rationale as the proof of Theorem 2. We prove for both limiting cases – $\tau_{\mathcal{S}} = 0$ and $\tau_{\mathcal{S}} = n_{\mathcal{S}}$ – by contradiction. Suppose another interval (l', u') exists with the same properties as the $1 - \alpha$ HPD interval. This interval could either overlap with the HPD or lie entirely outside its range. In the overlapping case, the proof mirrors that of Theorem 2. For the external case, due to the monotonicity of the posterior distribution, it is trivial to show that the (l', u') interval cannot have the same width as the HPD while covering μ with probability $1 - \alpha$, thus proving the uniqueness of the HPD interval. □

Posterior Skewness. The properties of $1-\alpha$ HPD CrIs make them suited for representing skewed posterior distributions. To verify this, we revisit Figure 2 with a focus on HPD intervals. When the distribution is skewed, as shown in Figures 2(b,c), the $1 - \alpha$ HPD interval exhibits two expected properties. First, it is smaller than the corresponding $1 - \alpha$ ET interval. Secondly, it only covers the most probable parameter values, resulting in a shift compared to the ET interval. Thus, for skewed posteriors, HPD CrIs provide the most precise and reliable summaries for post-data inference.

On the other hand, when the posterior is unimodal and symmetric, as in Figure 2(a), the HPD interval coincides with the ET interval. This relationship is formalized in the following theorem.

Theorem 3. When the beta posterior for μ is unimodal and symmetric, the $1 - \alpha$ HPD interval is equivalent to the $1 - \alpha$ ET interval.

Sketch Proof. Let ω represent the mode of a unimodal and symmetric posterior. By the symmetry property, for any $\delta > 0$, we have $f(\omega - \delta \mid G_{\mathcal{S}}) = f(\omega + \delta \mid G_{\mathcal{S}})$. Therefore, the $1 - \alpha$ HPD interval is symmetric around ω – i.e., the lower and upper bounds, l and u, are equidistant from ω. Consequently, it follows that $\int_0^l f(\mu \mid G_{\mathcal{S}})\, d\mu = \int_u^1 f(\mu \mid G_{\mathcal{S}})\, d\mu = \frac{\alpha}{2}$, which correspond to the CDFs of the $\alpha/2$ and $1 - \alpha/2$ quantiles, respectively. That is, the $1 - \alpha$ HPD and ET intervals coincide. □

This scenario occurs when, given a Beta(a, b) prior, the annotation process yields a Beta$(a + \tau_{\mathcal{S}}, b + n_{\mathcal{S}} - \tau_{\mathcal{S}})$ posterior where $a + \tau_{\mathcal{S}} = b + n_{\mathcal{S}} - \tau_{\mathcal{S}}$, which implies a post-data balance between correct and incorrect triples and hence a KG accuracy of 0.5.

Implications. These theoretical results encompass all the scenarios that can be encountered in the annotation process for evaluating KG accuracy. These results stem from modeling the task, for the first time, through a beta-binomial process and account for configurations involving informative and uninformative priors. This led to new proofs tailored to the specific requirements of the considered task.

These findings have notable implications for data quality management because they show that HPD CrIs consistently optimize the convergence rate of the KG accuracy minimization problem across all possible scenarios. This translates into minimizing the number of annotations required to audit KG accuracy, thereby reducing temporal and monetary costs that are critical factors for data curation [42].

Furthermore, these theoretical insights extend beyond KG accuracy evaluation to any task involving a similar annotation process. As such, they hold broader applicability and significance, offering valuable contributions also to related domains.

4.4 The Challenge of Choosing the Right Prior

The Bayesian framework we designed for KG accuracy evaluation offers a seamless approach to build an infinite number of $1-\alpha$ CrIs by varying the a, b parameters of the beta prior. This contrasts with frequentist methods, which involve different and complex procedures to generate $1-\alpha$ CIs that are challenging to interpret [30, 37]. However, adopting the unified and flexible Bayesian approach requires choosing appropriate priors that optimize efficiency while ensuring reliability.

We focus on the most general and challenging scenario, where no prior knowledge or belief about the KG accuracy is available. In this case, using informative (or subjective) priors can be detrimental, leading to deceptive interpretations or inefficient solutions [29]. To mitigate the bias introduced by these priors, we can resort to uninformative (or objective) priors [44], which reflect a lack of available information about the underlying process. In other words, uninformative priors do not encode the analyst's particular belief (or bias), making them suited to provide an objective analytical foundation. This allows the framework to be used by anyone – layman or expert – regardless of their specific expectations about the outcomes of the KG evaluation process.

For beta distributions, uninformative priors are characterized by parameters $a = b \leq 1$. In the literature, three proper (i.e., $a, b > 0$) uninformative beta priors are widely used: Kerman $\text{Beta}(\frac{1}{3}, \frac{1}{3})$ [24], Jeffreys $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ [22], and Uniform $\text{Beta}(1, 1)$ [2]. The Jeffreys prior is commonly adopted as the default choice for binomial proportion problems [8]. However, it represents a trade-off between Kerman and Uniform priors [23], and as a result, it is never the most efficient option.

As an example, Figure 3 presents the expected width of $1-\alpha$ HPD intervals under Kerman, Jeffreys, and Uniform priors for $n_{\mathcal{S}} = 30$ and $\alpha = 0.05$. We can see that the HPD interval based on Jeffreys prior is never the shortest of the three. In contrast, Kerman- and Uniform-based HPD intervals achieve minimal widths depending on the specific accuracy space regions. Kerman-based intervals are optimal in the extreme regions, while Uniform-based intervals are optimal in the central region.

Despite being minor, these differences in interval widths lead to appreciable variations in annotation costs for KG accuracy evaluation. As we will see experimentally in Section 6.2, the annotation costs obtained under different priors are statistically different in most cases. This further underscores the importance of addressing the prior selection problem.

Therefore, choosing the right uninformative prior is a nontrivial decision. Since the outcome of the annotation process – and thus the region of the accuracy space where the estimate will fall – cannot be known in advance, it is impossible to choose, *a priori*, one prior that consistently optimizes efficiency.

4.5 The *adaptive* HPD (*a*HPD) algorithm

To address the challenge of prior selection, we propose the *adaptive* HPD (*a*HPD) algorithm (see Algorithm 1), where multiple uninformative priors are concurrently employed to generate distinct $1-\alpha$ HPD intervals.

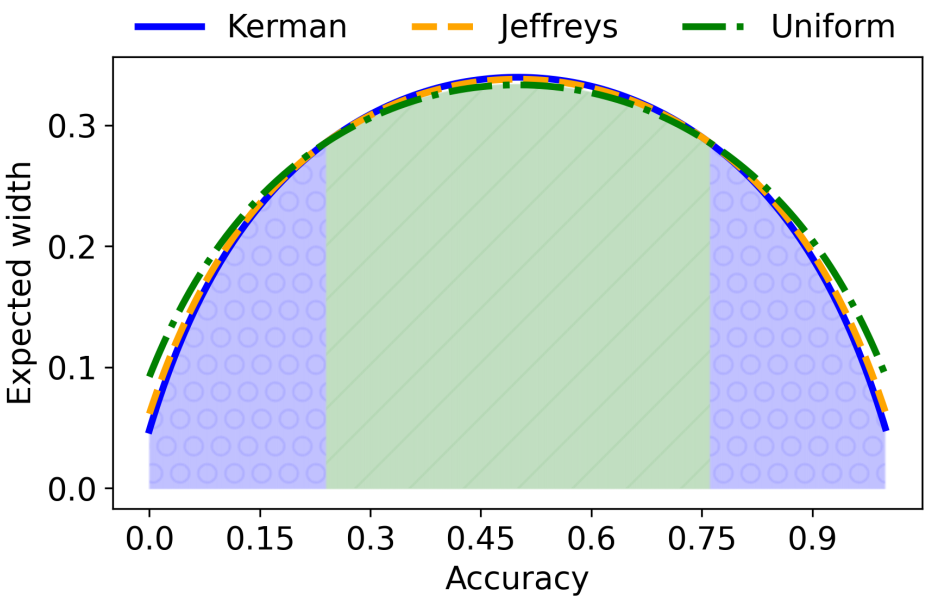


Fig. 3. Expected width of HPD credible intervals under Kerman, Jeffreys, and Uniform priors for $n_{\mathcal{S}} = 30$ and $\alpha = 0.05$. The circle pattern (○) under the curves represents the set of accuracy values where Kerman prior provides the shortest expected width, whereas the line pattern (//) the set of accuracy values where Uniform performs best.

At each stage of the KG evaluation process, the algorithm samples and annotates a batch of facts (lines 6-7), which combines with previous annotations to update the accuracy estimate (lines 8-9). Based on the number of correct facts and the sample size, *a*HPD first applies design effect adjustments in case of complex sampling strategies [26, 31] and then calculates the posterior parameters for each uninformative prior provided as input (lines 10-14); there is no limit to the number of priors that can be fed to the algorithm. It then builds the corresponding $1 - \alpha$ HPD intervals. When the number of correct facts equals the sample size (lines 15-16), in the first limiting case, the posterior distribution takes an exponentially increasing shape, and the interval bounds are computed using Equation (10). When the number of correct facts is zero (lines 17-18), in the second limiting case, the posterior distribution follows an exponentially decreasing pattern, and the intervals are derived via Equation (11). For the standard case, having defined the objective function, constraints, and bounds (lines 1-3), the algorithm adopts the SLSQP optimizer to generate the $1 - \alpha$ HPD intervals, using the corresponding ET intervals as initial guesses (lines 20-21). It selects the smallest one among the generated intervals and computes its MoE (lines 23-24). The process iterates until MoE $\leq \varepsilon$ (line 5), ensuring the selected $1 - \alpha$ HPD interval meets the desired precision. Once the criterion is satisfied, the algorithm returns the estimated accuracy and the chosen interval. Note that the computation of posterior parameters (line 14) and the construction of $1 - \alpha$ intervals (lines 16, 18, 20-21) can be parallelized for each input prior, ensuring that the algorithm remains efficient regardless of the number of considered priors.

The *a*HPD algorithm dynamically adjusts to the region of the accuracy space where the estimate lies after the annotation process. This adaptive approach ensures that the first interval achieving

Algorithm 1 *adaptive* HPD

Input:
The KG G;
The significance level α;
The sampling strategy $\mathcal{S}$;
The upper bound ε for the MoE;
The set of uninformative beta prior parameters $\{(a_1, b_1), \ldots, (a_k, b_k)\}$;

Output: The estimated accuracy and the smallest $1-\alpha$ HPD interval.

Define objective function (`obj`) to minimize interval width $(u - l)$
Define constraint function (`constr`) to satisfy $F(u) - F(l) = 1 - \alpha$
`bounds` $\leftarrow [[0, 1], [0, 1]]$ ▷ Set l and u bounds
`mu` $\leftarrow 0$, `moe` $\leftarrow 1$, `sample` $\leftarrow \emptyset$ ▷ Initialization
while `moe` $> \varepsilon$ **do**
  `batch` $\leftarrow \mathcal{S}(G)$ ▷ Sample batch of facts from G via $\mathcal{S}$
  `annot` $\leftarrow \mathbb{1}$(`batch`) ▷ Annotate batch and store annotations
  `sample` $\leftarrow$ `sample` $\cup$ `annot` ▷ Update sample with new annotations
  `mu` $\leftarrow \hat{\mu}_{\mathcal{S}}$(`sample`) ▷ Estimate G accuracy from sample
  `tau` $\leftarrow$ sum(`annot`), `n` $\leftarrow$ |`sample`| ▷ Correct triples and sample size
  if $\mathcal{S}$!= SRS **then** ▷ Complex sampling strategy
    Adjust for complex sampling effects using design effect corrections [26, 31]
  end if
  Compute posterior parameters from each prior and given annotation outcome:
  $(\mathtt{a\_post}_i, \mathtt{b\_post}_i) \leftarrow (a_i + \mathtt{tau}, b_i + \mathtt{n} - \mathtt{tau})$
  if `tau` == `n` **then** ▷ Limiting case (1): all facts correct
    $\mathtt{l}_i \leftarrow \text{qBeta}(\alpha; \mathtt{a\_post}_i, \mathtt{b\_post}_i)$, $\mathtt{u}_i \leftarrow 1 \; \forall i \in \{1, \ldots, k\}$
  else if `tau` == 0 **then** ▷ Limiting case (2): no correct facts
    $\mathtt{l}_i \leftarrow 0$, $\mathtt{u}_i \leftarrow \text{qBeta}(1 - \alpha; \mathtt{a\_post}_i, \mathtt{b\_post}_i) \; \forall i \in \{1, \ldots, k\}$
  else ▷ Standard case: unimodal posterior
    Compute, for each posterior, the $1 - \alpha$ ET CrI as initial guess:
    $\mathtt{guess}_i \leftarrow \text{ET}(1 - \alpha; \mathtt{a\_post}_i, \mathtt{b\_post}_i)$
    Compute, for each posterior, the $1 - \alpha$ HPD CrI using SLSQP optimizer:
    $(\mathtt{l}_i, \mathtt{u}_i) \leftarrow \text{SLSQP}(\mathtt{obj}, \mathtt{guess}_i, \mathtt{constr}, \mathtt{bounds})$
  end if
  $(\mathtt{l}, \mathtt{u}) \leftarrow \arg\min_i(\mathtt{u}_i - \mathtt{l}_i)$ ▷ Select smallest $1 - \alpha$ HPD CrI
  `moe` $\leftarrow (\mathtt{u} - \mathtt{l})/2$ ▷ Update MoE
end while
return `mu`, $(\mathtt{l}, \mathtt{u})$

the desired precision halts the evaluation process, guaranteeing the most efficient outcome from the set of competing solutions.

aHPD with Informative Priors. Although in this work we focus on the most general and challenging scenario, where no prior knowledge or belief about the KG accuracy is available, if we have (reliable) prior knowledge, then *a*HPD can be seamlessly used with such informative priors to optimize performance further.

Example 2. Let us assume that an analyst who wants to estimate the accuracy of DBpedia ($\mu = 0.85$, see Table 1) already knows the accuracy of two similar KGs – which have accuracy values $\mu = 0.80$

Table 1. Statistics for YAGO, NELL, DBPEDIA, FACTBENCH, and SYN 100M datasets.

	YAGO	NELL	DBPEDIA	FACTBENCH	SYN 100M
Num. of facts	1,386	1,860	9,344	2,800	101,415,011
Num. of clusters	822	817	2,936	1,157	5,000,000
Avg. cluster size	1.69	2.28	3.18	2.42	20.28
Accuracy (μ)	0.99	0.91	0.85	0.54	{0.9,0.5,0.1}

and $\mu = 0.90$, respectively. Based on this knowledge, the analyst can set up two informative priors, e.g., ($a_1 = 80$, $b_1 = 20$) and ($a_2 = 90$, $b_2 = 10$), and plug them into *a*HPD. The performance obtained with these informative priors under TWCS, averaged over 1,000 repetitions, are 63 ± 36 triples and a cost of 0.72 ± 0.41, instead of 222 ± 83 triples and a cost of 2.55 ± 0.95 when using *a*HPD equipped with Kerman, Jeffreys, and Uniform uninformative priors.

Hence, *a*HPD can lead to even more efficient solutions if prior knowledge is available. Meanwhile, poor informative priors can be detrimental, leading to deceptive interpretations or inefficient solutions. In these cases, relying on uninformative priors is better.

5 Experimental Setup

In the following, we report the considered datasets, the adopted cost function, the experimental configuration, and the evaluation procedure. We release data and code used in this work.[3] Moreover, we provide more experimental results presenting additional sampling strategies in the online repository. While these results are consistent with those given in the main text, they are less critical and thus are made available online due to space limits.

Datasets. We consider data from various real-life KGs, whose statistics are reported in Table 1. Note that dataset sizes are rather small due to the manual labeling costs.

YAGO: a dataset sampled by Ojha and Talukdar [41] from the YAGO2 KG [19], with facts referring to people, organizations, countries, and movies. The dataset was manually annotated by crowdsourcing workers, resulting in a ground-truth accuracy of $\mu = 0.99$.

NELL: drawn by Ojha and Talukdar [41] from the NELL KG [35], this dataset focuses on sports-related facts. It includes manual annotations by crowdsourcing workers for each fact, yielding a ground-truth accuracy of $\mu = 0.91$.

DBPEDIA: a dataset sampled by Marchesin et al. [32] from DBpedia [1],[4] covering a broad range of topics such as entertainment, news, history, sports, business, and science. Each fact in the dataset was manually annotated by at least three layman workers. The correctness label for every fact was then derived via quality-weighted majority voting, where (layman) worker quality was evaluated on an extra pool of facts supervised by expert workers. This annotation process resulted in a ground-truth accuracy of $\mu = 0.85$.

FACTBENCH: a benchmark created by Gerber et al. [16] to assess fact validation algorithms. It contains correct facts from DBpedia [1] and Freebase [4, 5], and generates incorrect facts by modifying correct ones while adhering to domain and range constraints. We consider a configuration where incorrect facts are a mix produced using different negative example generation strategies, resulting in a ground-truth accuracy of $\mu = 0.54$.

[3] https://github.com/KGAccuracyEval/credible-intervals4kg-estimation
[4] https://downloads.dbpedia.org/wiki-archive/dbpedia-dataset-version-2015-10.html

YAGO and NELL are the reference datasets used in the literature to evaluate KG accuracy estimation approaches [14, 31, 41, 46]. The DBPEDIA dataset is a proxy for one of the largest and widely used real-life KGs. Conversely, FACTBENCH represents a controlled scenario designed to assess the ability of automatic fact validation methods to distinguish between real (correct) and synthetic (incorrect) facts. Although unrealistic, FACTBENCH can be used to evaluate the behavior of KG accuracy estimation solutions in the quasi-symmetric case. Note that other datasets used by prior work exist, such as MOVIE [14] and OPIEC [46], but they are either not publicly (MOVIE) or only partially (OPIEC) available.

We assess the scalability of the methods by using **SYN 100M** [31], a synthetic dataset with over 100 million triples, where correctness labels are generated by setting the probability of a triple being true to a fixed rate. We consider ground-truth accuracy values of $\mu \in \{0.9, 0.5, 0.1\}$.

Annotation Cost. We use the cost function proposed by [14] and later adopted in [31, 46] to measure the manual effort required to assess the correctness of facts in a given sample. The function is based on the assumption that annotating a fact for an entity already identified incurs a lower cost than annotating a fact from an unseen entity. Specifically, it divides the annotation process into entity identification, which involves linking the entity to the corresponding real-world concept, and fact verification, which entails collecting evidence and auditing the fact stated by the triple.

The cost function is expressed as:

$$\text{cost}(G_S) = |E_S| \cdot c_1 + |T_S| \cdot c_2 \tag{12}$$

Here, c_1 and c_2 represent the average time cost (in seconds) for entity identification and fact verification, respectively. In line with prior research [14, 31, 46], we set $c_1 = 45$ and $c_2 = 25$ (seconds).

Sampling Strategies. We adopt SRS and TWCS as sampling strategies (Section 2.4). Based on the recommendation by Gao et al. [14] to choose the second-stage size m of TWCS in the range $\{3, 5\}$, we set $m = 3$ for YAGO, NELL, DBPEDIA, and FACTBENCH, which have small clusters, and $m = 5$ for SYN 100M, which presents larger clusters (see Table 1).

Interval Estimation. We consider Wald (Section 3.1) and Wilson (Section 3.2) CIs as baselines and compare them with the *a*HPD (Section 4.5). Wald represents an efficient yet unreliable baseline, which we consider only to emphasize the efficiency of our approach further. Conversely, Wilson represents the current state-of-the-art, providing a trade-off between efficiency and (frequentist) reliability. For *a*HPD, we employ three uninformative priors: Kerman Beta$(\frac{1}{3}, \frac{1}{3})$, Jeffreys Beta$(\frac{1}{2}, \frac{1}{2})$, and Uniform Beta$(1, 1)$. We also evaluate the performance of ET (Section 4.2) and HPD (Section 4.3) CrIs under these priors. We apply the design effect adjustments from [31] when the considered intervals are used with TWCS.

Evaluation Procedure. We set the significance level at $\alpha \in \{0.10, 0.05, 0.01\}$ and define the upper bound for the MoE as $\varepsilon = 0.05$. When not specified, the significance level defaults to $\alpha = 0.05$. We use two metrics to compare methods: the number of annotated triples and the annotation cost (in hours). The number of triples represents a simple and plain measure, while the annotation cost reflects a more realistic measure of temporal (and monetary) costs.

Results are reported when MoE ≤ 0.05, ensuring the methods are compared once they satisfy the optimization constraint. For this reason, we do not report the width of CIs, as all solutions have MoE ≤ 0.05. Similarly, accuracy estimates are omitted since all methods provide unbiased estimates with minimal deviation from ground-truth accuracy (less than 0.02).

Table 2. Performance of ET and HPD CrIs on YAGO, NELL, DBPEDIA, and FACTBENCH under different priors using SRS. For each interval, the best performances are in bold. We also report the performance obtained by *a*HPD equipped with the considered priors.

		YAGO	NELL	DBPEDIA	FACTBENCH
		$\mu = 0.99$	$\mu = 0.91$	$\mu = 0.85$	$\mu = 0.54$
Interval	Prior	Triples	Triples	Triples	Triples
ET	Kerman	**38±6**	**104±40**	**187±36**	380±3
	Jeffreys	39±7	107±39	188±34	379±3
	Uniform	41±9	113±36	190±32	**378±3**
HPD	Kerman	**32±5**	**96±44**	**182±42**	380±3
	Jeffreys	33±6	99±42	184±39	379±3
	Uniform	34±9	106±40	187±36	**378±3**
*a*HPD	{K, J, U}	**32±5**	**96±44**	**182±42**	**378±3**

To account for the inherent variability of sampling strategies, we repeat the evaluation procedure 1,000 times per method, reporting the mean and standard deviation for both the number of triples and the annotation cost.

6 Experimental Results

The objectives of the experiments are to (1) assess the impact of prior selection on KG accuracy estimation methods; (2) evaluate the efficiency of the *a*HPD compared to the state-of-the-art; (3) test whether *a*HPD is still efficient with large-scale datasets; and, (4) verify the robustness of *a*HPD across different levels of precision.

6.1 Summary of the Experimental Findings

(F1) *There is no one-prior-fits-all solution.* Section 6.2 shows that there is no single prior that we can apply to obtain top performance consistently, thus motivating the need for the *a*HPD.

(F2) *aHPD outperforms state-of-the-art solutions in all real-life cases.* Section 6.3 shows that *a*HPD reduces convergence costs compared to Wald and Wilson in common cases with a skewed accuracy distribution.

(F3) *aHPD scales efficiently to large datasets.* Section 6.4 shows that the dataset size does not affect the performance of *a*HPD, which remains the most efficient solution.

(F4) *aHPD is robust across varying levels of precision.* Section 6.5 shows that *a*HPD preserves its optimal performance regardless of the precision level required by the evaluation process.

6.2 Prior Selection

The performance of ET and HPD CrIs under Kerman, Jeffreys, and Uniform priors are reported in Table 2, together with those of *a*HPD equipped with the three considered priors. For sampling, we resorted to SRS, which allows for a clean comparison between CrIs built from different priors without introducing clustering. Such effects can hamper the interpretation of the results, making it challenging to determine whether the observed outcomes are due to the specific prior under consideration or to clustering effects.

The results in Table 2 confirm that Kerman and Uniform priors lead to the most efficient solutions in the extreme and central regions of the accuracy space, respectively (cf. Section 4.4). Conversely,

Table 3. Performance on YAGO, NELL, DBPEDIA, and FACTBENCH. For TWCS, we set $m = 3$. For each sampling strategy, best performance are in bold. † denotes a statistically significant difference between *a*HPD and Wald, while ‡ indicates a significant difference between *a*HPD and Wilson, as determined by standard independent t-tests with $p < 0.01$.

		YAGO		NELL		DBPEDIA		FACTBENCH	
		$\mu = 0.99$		$\mu = 0.91$		$\mu = 0.85$		$\mu = 0.54$	
Sampling	Interval	Triples	Cost	Triples	Cost	Triples	Cost	Triples	Cost
SRS	Wald	33±7	0.62±0.12	103±43	1.89±0.76	188±38	3.55±0.71	382±3	6.37±0.10
	Wilson	41±10	0.76±0.18	114±36	2.08±0.63	190±32	3.60±0.60	**378±3**	**6.32±0.10**
	*a*HPD	**32±5**	**0.60±0.09**†,‡	**96±44**	**1.76±0.79**†,‡	**182±42**	**3.45±0.78**†,‡	**378±3**	**6.32±0.10**†
TWCS	Wald	32±4	0.42±0.05	126±70	1.58±0.87	243±75	2.80±0.87	**254±37**	**3.08±0.45**
	Wilson	35±4	0.47±0.06	129±67	1.62±0.84	234±73	2.69±0.84	257±39	3.11±0.48
	*a*HPD	**31±2**	**0.41±0.03**†,‡	**112±68**	**1.40±0.85**†,‡	**222±83**	**2.55±0.95**†,‡	257±39	3.11±0.48

the Jeffreys prior never achieves top performance. This pattern holds for both ET and HPD CrIs, highlighting the generality of the problem of choosing an appropriate prior.

Focusing on HPD intervals, the annotation costs under different priors are statistically different in 8 out of 12 cases, as determined by standard independent t-tests with $p < 0.01$. This emphasizes that the choice of the prior significantly affects the outcomes. In this regard, *a*HPD, which automatically selects the optimal prior (cf. Table 2), proves to be statistically better than HPD solutions relying on suboptimal priors in 75% of cases. Therefore, *a*HPD emerges as a robust approach for consistently selecting the optimal prior for a given annotation process, removing the need to choose *a priori*.

It is worth noting that other priors could be explored, which may perform better than the considered three in specific regions of the accuracy space. This further stresses the problem of prior selection when optimizing efficiency. For this work, we focused on three of the most popular priors, showing that even with a relatively small set, it is impossible to guarantee optimal performance across the entire accuracy space without knowing where the accuracy estimate will fall after the annotation process. This strengthens the need for an adaptive solution as the *a*HPD.

As a side note, Table 2 also shows that HPD intervals are better than ET ones when the accuracy is skewed (YAGO, NELL, DBPEDIA) and equal when the accuracy is quasi-symmetric (FACTBENCH), as proved by Theorems 1-3.

6.3 Efficiency

The performance of *a*HPD against Wald and Wilson baselines is reported in Table 3. We compared methods considering both SRS and TWCS as sampling strategies. We also performed statistical analyses to assess the significance of the performance differences between *a*HPD and baselines. Specifically, we conducted standard independent t-tests between *a*HPD and Wald, and between *a*HPD and Wilson, reporting statistical results when $p < 0.01$.

Table 3 provides the following insights. *a*HPD statistically outperforms both Wald and Wilson methods across all datasets where the KG accuracy is skewed – i.e., YAGO, NELL, and DBPEDIA. This is a significant outcome, as it shows that *a*HPD surpasses not only the Wilson method, which sacrifices efficiency to ensure high (frequentist) reliability [31] but also the unreliable, yet very efficient Wald method. Thus, *a*HPD, grounded in Bayesian statistics, provides interpretable and reliable results while improving efficiency.

For datasets with quasi-symmetric accuracy distributions, such as FACTBENCH, the performance of *a*HPD matches those of the Wilson method. This result is expected. Since FACTBENCH has an accuracy of $\mu = 0.54$, we know that, among the considered priors, the Uniform one achieves optimal

Table 4. Performance on SYN 100M with accuracy values $\mu \in \{0.9, 0.5, 0.1\}$. For TWCS we set $m = 5$. For each sampling strategy, best performance are in bold. † denotes a statistically significant difference between *a*HPD and Wald, while ‡ indicates a significant difference between *a*HPD and Wilson, as determined by standard independent t-tests with $p < 0.01$.

		SYN 100M					
		$\mu = 0.9$		$\mu = 0.5$		$\mu = 0.1$	
Sampling	Interval	Triples	Cost	Triples	Cost	Triples	Cost
SRS	Wald	122±43	2.37±0.83	384±1	7.46±0.03	124±43	2.42±0.83
	Wilson	131±34	2.54±0.66	**380±1**	**7.39±0.03**	133±35	2.58±0.68
	*a*HPD	**114±46**	**2.22±0.89**†,‡	**380±1**	**7.39±0.03**†	**117±45**	**2.28±0.88**†,‡
TWCS	Wald	120±50	1.13±0.48	384±63	3.64±0.60	121±53	1.14±0.50
	Wilson	121±47	1.14±0.45	**374±65**	**3.54±0.62**	121±49	1.15±0.47
	*a*HPD	**106±52**	**1.01±0.49**†,‡	**374±65**	**3.54±0.62**†	**108±54**	**1.02±0.51**†,‡

performance (see Table 2). We also know that, in this scenario, HPD and ET intervals are the same (Theorem 3). Then, since the Wilson CI can be seen as an approximation of the ET CrI based on the Uniform prior [23], it follows that *a*HPD and Wilson-based solutions yield equal performance. Hence, these findings indicate that *a*HPD offers an efficient and reliable solution suited to all the considered scenarios. It consistently provides statistically superior performance in cases of skewed KG accuracy while remaining competitive in (quasi-)symmetric situations.

6.4 Scalability

Table 4 reports the *a*HPD performance on SYN 100M datasets compared to the Wald and Wilson methods using SRS and TWCS for sampling. As before, we conducted statistical analyses to assess the significance of *a*HPD performance with respect to baselines.

We can see that the performance of all methods remain within the same order of magnitude as those observed in small-scale datasets (cf. Table 3). This suggests that KG accuracy estimation methods are not affected by the size of the datasets. Such a finding is unsurprising, as both the accuracy estimators and the $1 - \alpha$ intervals are independent of the underlying (dataset) population. In the case of CIs, they only rely on the sample and its characteristics. Similarly, for CrIs, using uninformative priors makes the interval estimation primarily driven by the annotation process – that is, by the sample and its characteristics. Instead, the main factor influencing the performance of the methods is the distribution of correctness labels across the dataset, or in other words, the accuracy level of the dataset. Consequently, the performance observed in large-scale datasets mirror those of small ones, motivating the consistency in the results of SYN 100M (large-scale) and YAGO, NELL, DBPEDIA, and FACTBENCH (small-scale)

As such, *a*HPD proves again to be the most statistically efficient solution when the KG accuracy is skewed (i.e., $\mu \in \{0.9, 0.1\}$) and a competitive approach in the symmetric case (i.e., $\mu = 0.5$), where it achieves top performance alongside Wilson. This further consolidates *a*HPD as the preferred solution in real-world applications.

It is worth mentioning that, with unbiased sampling strategies, if two datasets have correctness labels distributed in opposite proportions – i.e., presenting symmetric accuracy values of μ and $1 - \mu$, respectively – the evaluation process incurs the same costs. In other words, the iterative procedure requires (approximately) the same number of triples to converge, regardless of whether

the dataset accuracy is μ or $1-\mu$. This property explains the consistent performance of the methods on SYN 100M with $\mu = 0.9$ and $\mu = 0.1$.

6.5 Robustness

Figure 4 presents the annotation costs of *a*HPD at various precision levels, comparing them to Wilson costs under SRS and TWCS sampling strategies. The reduction ratio of *a*HPD relative to Wilson is also shown for each case. Since we have already demonstrated that *a*HPD is a superior alternative to Wald, this comparison focuses solely on Wilson, the current state-of-the-art for KG accuracy estimation [31].

The results in Figure 4 confirm the superior performance of *a*HPD not only at the standard $\alpha = 0.05$ level but also at less ($\alpha = 0.10$) and more ($\alpha = 0.01$) stringent precision levels. Overall, *a*HPD exhibits reduction ratios across all skewed datasets and precision levels, underscoring its versatility across diverse application scenarios. For instance, on YAGO, the cost reduction achieved by *a*HPD reaches up to 47% under SRS and 39% under TWCS when $\alpha = 0.01$. To reinforce these findings, we conducted statistical analyses for $1-\alpha$ intervals with precision levels $\alpha \in \{0.1, 0.01\}$, obtaining results consistent with those found for $\alpha = 0.05$ (cf. Table 3). These outcomes further substantiate the higher efficiency of *a*HPD across precision levels.

The $\alpha = 0.01$ setting is particularly noteworthy, as it reflects a high-precision scenario where a substantial degree of confidence in the estimates is essential. Such scenarios are common in fields like healthcare, where ensuring data quality is critical to prevent introducing inference errors into downstream applications that rely on the KG. In this context, where the number of required annotations increases due to the stringent precision demands, the cost savings achieved by *a*HPD become even more impactful. Since manual annotations in these cases often require expert knowledge, the ability to reduce costs while maintaining reliability is precious.

Furthermore, it is also important to consider that, in real-world scenarios, the evaluation process typically involves multiple annotators (often 3-5) per fact, whose annotations are aggregated to determine the final correctness label. As the number of annotators increases, so do the corresponding costs. Depending on the available annotation budget, the cost reduction introduced by *a*HPD can make the difference between an evaluation process that concludes successfully (due to convergence) and one that terminates prematurely (due to budget exhaustion).

As expected, there are no benefits in using *a*HPD over Wilson on FACTBENCH (quasi-symmetric) – but no drawbacks either. This favors *a*HPD, as it reduces annotation costs in skewed cases without introducing additional costs in quasi-symmetric ones. Thus, the robustness analysis across different precision levels confirms the all-around superiority of *a*HPD compared to the considered baselines, providing strong evidence to support its adoption with any sampling strategy and in any scenario where there is a need to evaluate KG accuracy with limited annotations.

7 Related Work

The efficient evaluation of KGs accuracy has been largely overlooked. The first approach, KGEval [41], is an iterative algorithm that alternates between control and inference stages. In the control stage, KGEval uses crowdsourcing to select facts for evaluation. In the inference stage, it applies type consistency and Horn-clause coupling constraints [28, 35] to the evaluated facts to automatically estimate the correctness of additional facts. This process repeats until no more facts are evaluated. KGEval does not scale to real-life KGs [14] and is susceptible to error propagation due to its probabilistic inference mechanism, making it unsuitable for our work.

To overcome KGEval limitations, Gao et al. [14] resorted to sampling strategies and estimators that gauge KG accuracy with statistical guarantees. To minimize annotation costs, [14] proposed the use of cluster sampling techniques (i.e., TWCS) over standard sampling (i.e., SRS). On the other

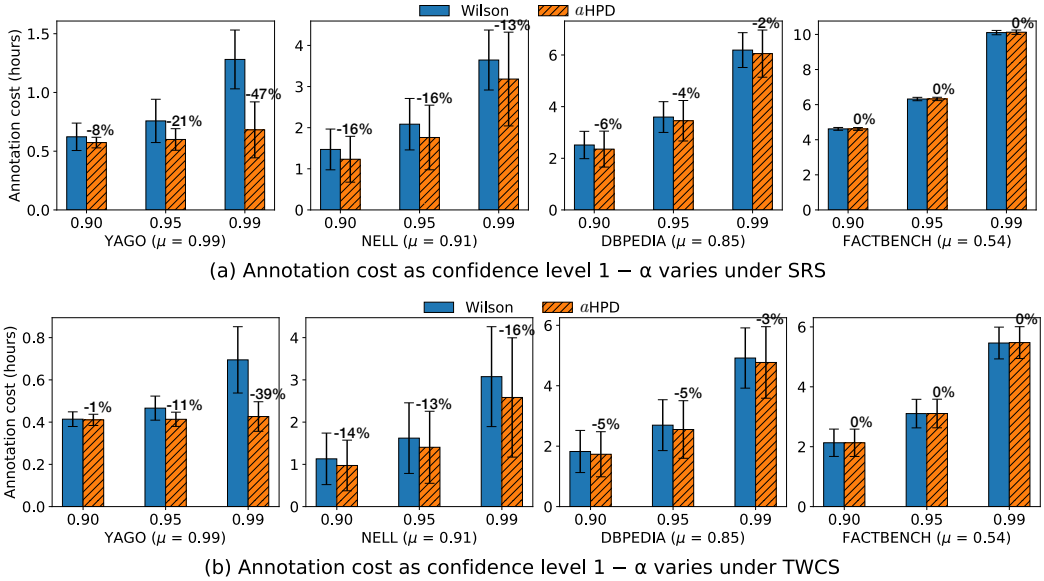


Fig. 4. Annotation cost comparison between *a*HPD and Wilson at different confidence levels $1-\alpha$ under SRS (a) and TWCS (b) on YAGO, NELL, DBPEDIA, and FACTBENCH KGs. We also report the reduction ratio (in %) of *a*HPD over Wilson.

hand, the authors relied on the Wald interval [9], known to have reliability issues when used on binomial proportions [8, 51]. To address Wald issues for KG accuracy evaluation, Marchesin and Silvello [31] proposed to use the Wilson interval [53], being it better suited for binomial proportions.

Although reliable, the Wilson interval shares Wald's frequentist interpretation, preventing a one-shot probabilistic understanding of its reliability. Additionally, the interval balances efficiency and reliability, sometimes requiring more annotations than Wald for greater reliability. In contrast, our solution offers a one-shot interpretation of interval confidence, making it better suited for the task while remaining the most efficient option. Moreover, the proposed approach, *a*HPD, eliminates the need for analysts to select a specific prior, addressing a key barrier to using Bayesian methods.

Qi et al. [46] developed an efficient human-machine collaborative framework to minimize annotation costs through inference. This framework leverages statistical techniques similar to those in [14] and combines inference mechanisms akin to those in [41]. Their work focuses on the interplay between human annotations and inference mechanisms rather than on CIs and their impact on the minimization problem. Therefore, we do not include their method in our experimental analysis, but the *a*HPD method can be integrated into Qi et al. [46]'s framework to enhance efficiency.

8 Conclusions

In this paper, we exposed the limitations of Confidence Intervals (CIs), used in current state-of-the-art solutions for evaluating KG accuracy. Rooted in frequentist statistics, these efficient intervals often lead to interpretation fallacies regarding confidence, precision, and likelihood, making them inconvenient for the task. We defined the interval estimation problem in Bayesian terms to address these limitations, introducing Bayesian Credible Intervals (CrIs). Using a prior distribution reflecting the initial belief about the KG accuracy and updating it with observed data through the likelihood, CrIs define the range of probable parameters in the posterior distribution. In other words, these

intervals represent where the KG accuracy is likely to fall with $1 - \alpha$ probability. Since they are built directly on the posterior distribution, CrIs avoid the interpretative pitfalls of CIs, offering reliable and efficient solutions in a *one-shot* setting.

We considered two families of CrIs: Equal-Tailed (ET) and Highest Posterior Density (HPD) intervals. Through theoretical and empirical analyses, we demonstrated that HPD CrIs represent the most efficient solution in real-life cases – where the KG accuracy is skewed – while remaining competitive with state-of-the-art in controlled scenarios – where the KG accuracy is (quasi-)symmetric.

We also addressed one of the major resistances to using Bayesian methods: the challenge of choosing the right prior. Prior selection is a critical aspect of Bayesian approaches, as selecting a poor prior can lead to deceptive interpretations or inefficient solutions. We proposed the *adaptive* HPD (*a*HPD) algorithm, where multiple objective priors are concurrently used to generate different $1 - \alpha$ HPD intervals. These intervals compete to achieve convergence in the minimization problem, ensuring the most efficient outcome is always obtained from the set of competing solutions. In this way, *a*HPD removes the need to choose a specific prior in advance.

Through extensive experiments on both real and synthetic data, we showed that *a*HPD outperforms the state-of-the-art approaches across different scales, providing robust results regardless of the precision required for the evaluation process. Notably, *a*HPD reduced annotation costs by up to 47% in high-precision scenarios. Based on this thorough evaluation, which involved both SRS and TWCS sampling strategies, we recommend that practitioners adopt *a*HPD with TWCS as the most efficient and reliable solution for auditing KG accuracy in a one-shot setting.

Future Work and Limitations. We plan to explore the use of the *a*HPD algorithm in dynamic environments, where KG contents evolve over time [43]. Consider a typical scenario for KGs, where batches of new content arrive at irregular intervals. Assume an initial evaluation has been conducted, resulting in a given KG accuracy estimate. When the volume of new content reaches a certain, predefined mass, we can perform a new evaluation. Leveraging the Bayesian framework we introduced to define the problem, we can consider the initial estimate as prior knowledge and use it as an informative prior for the new evaluation process. Equipped with reliable informative priors, *a*HPD can converge significantly faster than solutions based on uninformative priors (see Example 2).

Despite promising, this approach also hides potential limitations. Massive updates (e.g., half the size of the KG) with significantly different accuracy levels can render the informative prior less effective and, potentially, even deceptive. Exploring these challenges represents a critical direction for future research.

Acknowledgments

The work was supported by the HEREDITARY project, as part of the EU Horizon Europe program under Grant Agreement 101137074. ChatGPT and Grammarly were used as writing assistants to help rephrase certain sentences in this manuscript.

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007 (LNCS, Vol. 4825)*. Springer, 722–735. doi:10.1007/978-3-540-76298-0_52

[2] T. Bayes. 1763. An Essay Towards Solving a Problem in the Doctrine of Chances. *Phil. Trans. R. Soc.* 53 (1763), 370–418. doi:10.1098/rstl.1763.0053

[3] J. O. Berger. 1985. *Statistical Decision Theory and Bayesian Analysis, 2nd Edition*. Springer. doi:10.1007/978-1-4757-4286-2

[4] K. D. Bollacker, R. P. Cook, and P. Tufts. 2007. Freebase: A Shared Database of Structured General Human Knowledge. In *Proc. of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*. AAAI, 1962–1963. doi:10.5555/1619797.1619981

[5] K. D. Bollacker, C. Evans, P. K. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proc. of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*. ACM, 1247–1250. doi:10.1145/1376616.1376746

[6] A. Bonifati, G. H. L. Fletcher, H. Voigt, and N. Yakovets. 2018. *Querying Graphs*. Morgan & Claypool Publishers. doi:10.2200/S00873ED1V01Y201808DTM051

[7] G. E. P. Box and G. C. Tiao. 2011. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons.

[8] L. D. Brown, T. T. Cai, and A. DasGupta. 2001. Interval Estimation for a Binomial Proportion. *Statist. Sci.* 16, 2 (2001), 101–117. http://www.jstor.org/stable/2676784

[9] G. Casella and R. L. Berger. 2002. *Statistical Inference*. Thomson Learning. https://books.google.it/books?id=0x_vAAAAMAAJ

[10] W. G. Cochran. 1977. *Sampling Techniques, 3rd Edition*. John Wiley. doi:10.1017/S0013091500025724

[11] O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. 2013. Building, maintaining, and using knowledge bases: a report from the trenches. In *Proc. of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*. ACM, 1209–1220. doi:10.1145/2463676.2465297

[12] W. Edwards, H. Lindman, and L. J. Savage. 1963. Bayesian Statistical Inference for Psychological Research. *Psychol. Rev.* 70, 3 (1963), 193–242. doi:10.1037/h0044139

[13] D. Esteves, A. Rula, A. J. Reddy, and J. Lehmann. 2018. Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis. *ACM J. Data Inf. Qual.* 9, 3 (2018), 16:1–16:26. doi:10.1145/3177873

[14] J. Gao, X. Li, Y. E. Xu, B. Sisman, X. L. Dong, and J. Yang. 2019. Efficient Knowledge Graph Accuracy Evaluation. *Proc. VLDB Endow.* 12, 11 (2019), 1679–1691. doi:10.14778/3342263.3342642

[15] A. Gelman. 2008. Rejoinder. *Bayesian Anal.* 3, 3 (2008), 467 – 477. doi:10.1214/08-BA318REJ

[16] D. Gerber, D. Esteves, J. Lehmann, L. Bühmann, R. Usbeck, A. C. Ngonga Ngomo, and R. Speck. 2015. DeFacto - Temporal and multilingual Deep Fact Validation. *J. Web Semant.* 35 (2015), 85–101. doi:10.1016/j.websem.2015.08.001

[17] F. Hasibi, K. Balog, and S. E. Bratsberg. 2017. Dynamic Factual Summaries for Entity Cards. In *Proc. of SIGIR*. ACM, 773–782. doi:10.1145/3077136.3080810

[18] R. Hoekstra, R. D. Morey, J. N. Rouder, and E. J. Wagenmakers. 2014. Robust Misinterpretation of Confidence Intervals. *Psychon. Bull. Rev.* 21 (2014), 1157–1164. doi:10.3758/s13423-013-0572-3

[19] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artif. Intell.* 194 (2013), 28–61. doi:10.1016/j.artint.2012.06.001

[20] I. F. Ilyas, J. Lacerda, Y. Li, U. F. Minhas, A. Mousavi, J. Pound, T. Rekatsinas, and C. Sumanth. 2023. Growing and Serving Large Open-domain Knowledge Graphs. In *Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023*. ACM, 253–259. doi:10.1145/3555041.3589672

[21] I. F. Ilyas, T. Rekatsinas, V. Konda, J. Pound, X. Qi, and M. A. Soliman. 2022. Saga: A Platform for Continuous Construction and Serving of Knowledge at Scale. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*. ACM, 2259–2272. doi:10.1145/3514221.3526049

[22] H. Jeffreys. 1946. An Invariant Form for the Prior Probability in Estimation Problems. *Proc. R. Soc. A* 186, 1007 (1946), 453–461. doi:10.1098/rspa.1946.0056

[23] S. Jin, M. Thulin, and R. Larsson. 2017. Approximate Bayesianity of Frequentist Confidence Intervals for a Binomial Proportion. *Am. Stat.* 71, 2 (2017), 106–111. doi:10.1080/00031305.2016.1208630

[24] J. Kerman. 2011. Neutral Noninformative and Informative Conjugate Beta and Gamma Prior Distributions. *Electron. J. Stat.* 5 (2011), 1450 – 1470. doi:10.1214/11-EJS648

[25] L. Kish. 1965. *Survey Sampling*. Wiley. https://books.google.it/books?id=xiZmAAAAIAAJ

[26] L. Kish. 1995. Methods for Design Effects. *J. Off. Stat.* 11, 1 (1995), 55. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/methods-for-design-effects.pdf

[27] D. Kraft. 1988. *A Software Package for Sequential Quadratic Programming*. Technical Report. DLR German Aerospace Center - Institute for Flight Mechanics, Koln, Germany. 1 – 33 pages.

[28] N. Lao, T. M. Mitchell, and W. W. Cohen. 2011. Random Walk Inference and Learning in A Large Scale Knowledge Base. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK*. ACL, 529–539. https://aclanthology.org/D11-1049/

[29] E. Lesaffre and A. B. Lawson. 2012. *Bayesian Biostatistics*. John Wiley & Sons. doi:10.1002/9781119942412

[30] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke. 2019. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *J. Open Source Softw.* 4, 40 (2019), 1541. doi:10.21105/joss.01541

[31] S. Marchesin and G. Silvello. 2024. Efficient and Reliable Estimation of Knowledge Graph Accuracy. *Proc. VLDB Endow.* 17, 9 (2024), 2392–2404. doi:10.14778/3665844.3665865

[32] S. Marchesin, G. Silvello, and O. Alonso. 2024. Utility-Oriented Knowledge Graph Accuracy Estimation with Limited Annotations: A Case Study on DBpedia. *Proc. of the 12th AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2024, Pittsburgh, Pennsylvania, USA, October 16–19, 2024* 12, 1 (2024), 105–114. doi:10.1609/hcomp.v12i1.31605

[33] S. Marchesin, G. Silvello, and O. Alonso. 2024. Veracity Estimation for Entity-Oriented Search with Knowledge Graphs. In *Proc. of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*. ACM, 1649–1659. doi:10.1145/3627673.3679561

[34] R. McElreath. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. Chapman & Hall. doi:10.1201/9780429029608

[35] T. M. Mitchell, W. W. Cohen, E. R. Hruschka Jr., P. P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. D. Mishra, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. A. Platanios, A. Ritter, M. Samadi, B. Settles, R. C. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2018. Never-ending learning. *Commun. ACM* 61, 5 (2018), 103–115. doi:10.1145/3191513

[36] J. Mohoney, A. Pacaci, S. R. Chowdhury, A. Mousavi, I. F. Ilyas, U. F.q Minhas, J. Pound, and T. Rekatsinas. 2023. High-Throughput Vector Similarity Search in Knowledge Graphs. *Proc. ACM Manag. Data* 1, 2 (2023), 197:1–197:25. doi:10.1145/3589777

[37] R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E. J. Wagenmakers. 2016. The Fallacy of Placing Confidence in Confidence Intervals. *Psychon. Bull. Rev.* 23 (2016), 103–123. doi:10.3758/s13423-015-0947-8

[38] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. J. Smola. 2013. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proc. of WWW*. International World Wide Web Conferences Steering Committee / ACM, 953–964. doi:10.1145/2488388.2488471

[39] J. Neyman. 1941. Fiducial Argument and the Theory of Confidence Intervals. *Biometrika* 32, 2 (1941), 128–150. http://www.jstor.org/stable/2332207

[40] J. Neyman and H. Jeffreys. 1937. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philos. Trans. R. Soc. A* 236, 767 (1937), 333–380. doi:10.1098/rsta.1937.0005

[41] P. Ojha and P. Talukdar. 2017. KGEval: Accuracy Estimation of Automatically Constructed Knowledge Graphs. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. ACL, 1741–1750. doi:10.18653/v1/d17-1183

[42] H. Paulheim. 2018. How much is a Triple? Estimating the Cost of Knowledge Graph Creation. In *Proc. of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8-12, 2018 (CEUR, Vol. 2180)*. CEUR-WS.org. https://ceur-ws.org/Vol-2180/ISWC_2018_Outrageous_Ideas_paper_10.pdf

[43] A. Polleres, R. Pernisch, A. Bonifati, D. Dell'Aglio, D. Dobriy, S. Dumbrava, L. Etcheverry, N. Ferranti, K. Hose, E. Jiménez-Ruiz, M. Lissandrini, A. Scherp, R. Tommasini, and J Wachs. 2023. How Does Knowledge Evolve in Open Knowledge Graphs? *TGDK* 1, 1 (2023), 11:1–11:59. doi:10.4230/TGDK.1.1.11

[44] S. J. Press. 2002. *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons. doi:10.1002/9780470317105

[45] J. Pujara, E. Augustine, and L. Getoor. 2017. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. ACL, 1751–1756. doi:10.18653/v1/d17-1184

[46] Y. Qi, W. Zheng, L. Hong, and L. Zou. 2022. Evaluating Knowledge Graph Accuracy Powered by Optimized Human-Machine Collaboration. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*. ACM, 1368–1378. doi:10.1145/3534678.3539233

[47] H. Raiffa and R. Schlaifer. 1961. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Adminitration, Harvard University. https://books.google.it/books?id=wPBLAAAAMAAJ

[48] R. Reinanda, E. Meij, and M. de Rijke. 2020. Knowledge Graphs: An Information Retrieval Perspective. *Found. Trends Inf. Retr.* 14, 4 (2020), 289–444. doi:10.1561/1500000063

[49] M. Samadi, P. P. Talukdar, M. M. Veloso, and T. M. Mitchell. 2015. AskWorld: Budget-Sensitive Query Evaluation for Knowledge-on-Demand. In *Proc. of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. AAAI Press, 837–843. http://ijcai.org/Abstract/15/123

[50] D. Vrandecic and M. Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. doi:10.1145/2629489

[51] S. Wallis. 2013. Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods. *J. Quant. Linguistics* 20, 3 (2013), 178–208. doi:10.1080/09296174.2013.799918

[52] L. Wasserman. 2008. Comment on Article by Gelman. *Bayesian Anal.* 3, 3 (2008), 463 – 465. doi:10.1214/08-BA318D

[53] E. B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212. doi:10.1080/01621459.1927.10502953

[54] B. Xue and L. Zou. 2023. Knowledge Graph Quality Management: A Comprehensive Survey. *IEEE Trans. Knowl. Data Eng.* 35, 5 (2023), 4969–4988. doi:10.1109/TKDE.2022.3150080

Received October 2024; revised January 2025; accepted February 2025