Binomial Confidence Intervals for Knowledge Graph Accuracy Estimation (Extended Abstract)*

Stefano Marchesin[1], Gianmaria Silvello[1]

[1]Department of Information Engineering, University of Padua, Padua, Italy

Abstract

Data accuracy is a critical aspect of data quality, particularly in the context of Knowledge Graphs (KGs). Accurately auditing KGs is essential for informed decision-making in entity-centric services and applications. However, manual accuracy evaluation of large-scale KGs is prohibitively costly, prompting research into efficient sampling techniques for KG accuracy estimation. In this extended abstract, we report our endeavours in tackling the shortcomings of existing KG accuracy estimation methods, which predominantly rely on the Wald method for constructing Confidence Intervals (CIs). When used to gauge binomial proportions, such as KG accuracy, Wald intervals suffer from reliability issues such as zero-width and overshooting. We introduce a solution based on the Wilson method, which addresses these challenges and ensures broad applicability across diverse evaluation scenarios. The results demonstrate that the proposed solution enhances the reliability of accuracy estimates by up to two times compared to the state-of-the-art, without compromising efficiency. Moreover, this improvement remains consistent regardless of KG size or topology.

Keywords

Data Quality, Knowledge Graph, Cluster Sampling, Confidence Interval

1. Introduction

In recent years, large-scale Knowledge Graphs (KGs) containing millions of relational facts, represented as subject-predicate-object triples (s, p, o), have gained significant prominence. Notable examples include Wikidata [2], DBpedia [3], YAGO [4], and NELL [5]. However, existing construction processes for KGs are prone to errors, resulting in sparse graphs that contain inaccurate facts [6, 7]. Therefore, precise evaluation of KG accuracy is essential for refining construction processes, understanding data reliability, and supporting downstream applications [8, 9].

KG accuracy plays a critical role not only in database management [10, 11, 12], but also in search, recommendation, and question-answering systems [13, 14]. Industrial applications like Saga [10] further emphasize the importance of accurate KGs for delivering high-quality user experiences in entity-centric services, making reliable, on-demand accuracy assessments vital for knowledge platforms.

However, evaluating KG accuracy requires labeling facts for correctness, a process that is both expensive and labor-intensive [15, 16]. Given the impracticality of annotating every fact in large-scale KGs [17], recent works have approached the problem as a constrained minimization task [16, 17], combining sampling strategies, accuracy estimators, and Confidence Intervals (CIs) to ensure estimation robustness. These methods rely on Wald CIs [18], which assume normal approximation [19]. However, since KG accuracy estimation concerns binomial proportions, Wald intervals suffer from zero-width and overshooting issues, especially when proportions (i.e., accuracies) approach 0 or 1 [20, 19] – a common scenario in real-world KGs [6, 7].

Contributions. Thus, efficient and cost-effective evaluation approaches must ensure reliable CIs that account for the binomial nature of KG accuracy. To this end, in [1], we expose the limitations of state-of-the-art KG accuracy estimation methods that rely on Wald intervals, validating these shortcomings

SEBD 2025: 33rd Symposium On Advanced Database Systems, June 16-19, 2025, Ischia, Italy

*This work represents an extended abstract of [1], presented at VLDB 2024.

stefano.marchesin@unipd.it (S. Marchesin); gianmaria.silvello@unipd.it (G. Silvello)
https://www.dei.unipd.it/~marches1/ (S. Marchesin); https://www.dei.unipd.it/~silvello/ (G. Silvello)
0000-0003-0362-5893 (S. Marchesin); 0000-0003-4970-4554 (G. Silvello)

through experiments on real and synthetic KGs. Then, we introduce and evaluate a set of binomial CIs that overcome Wald's drawbacks, identifying the Wilson interval [21] as offering the best balance between efficiency and reliability. We also extend Wilson and other binomial intervals to accommodate complex sampling designs, such as clustering and stratification. Finally, we benchmark our Wilson-based solutions against the state-of-the-art, which relies on Wald intervals, finding that Wilson intervals are up to twice as reliable and more efficient than state-of-the-art in common, real-world scenarios. At the same time, we also demonstrate the scalability of our methods on a synthetic KG exceeding 100 million triples, showing consistent reliability and efficiency regardless of KG size or structure.

Outline. The rest of this paper is as follows. In Section 2, we describe the problem and evaluation framework, also reporting the considered sampling techniques and estimators. In Section 3, we detail Wald and Wilson CIs, mentioning Wilson's theoretical advantages. In Section 4, we cover experimental setup and results. In Section 5, we conclude the paper.

2. Background

In this section, we first introduce the required notation and concepts, then we present the problem and its optimization, and finally we report the considered sampling strategies and estimators.

2.1. Preliminaries

Following [22], we define a KG as a directed, edge-labeled multi-graph $G = (V, R, \eta)$, where $V = \{E \cup A\}$ represents entities (E) and attributes (A), R is the set of relationships, and $\eta : R \to E \times (E \cup A)$ maps each relationship to an ordered pair of nodes. The ternary relation T comprises (s, p, o) triples, where $s \in E$, $p \in R$, and $o \in E \cup A$, with $M = |T|$ denoting its size. We also define an entity cluster $G[e] = \{(s, p, o) \in T \mid s = e\}$ as the set of triples sharing the same subject e.

2.2. Problem Formulation

The accuracy of a KG G is defined as the mean accuracy of its triples:

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

where $\mathbb{1}(t)$ is an indicator function returning 1 if a triple t is correct and 0 otherwise. Manual annotation is used to determine correctness, making large-scale evaluations costly.

To address this, we estimate $\mu(G)$ using an unbiased estimator $\hat{\mu}$ over a sample $T_{\mathcal{S}} \subset T$ drawn via a sampling strategy $\mathcal{S}$. The goal is to minimize annotation costs while ensuring a CI with a Margin of Error (MoE) below a threshold ε. Formally:

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

The problem requires a sampling strategy $\mathcal{S}$ that minimizes the cost of manually evaluating triples. Simultaneously, it must satisfy a constraint on the MoE, ensuring it remains below a specified upper bound ε. The problem remains unsolved until the CI reaches the desired width, making the CI a central component of the optimization process. CIs that shrink rapidly accelerate the process convergence, reducing sample sizes and thus annotation costs. However, they must remain reliable, encompassing the true KG accuracy approximately $1 - \alpha$ times. Despite its importance, previous research has largely overlooked the impact of CI selection, focusing instead on sampling strategies [16, 17].

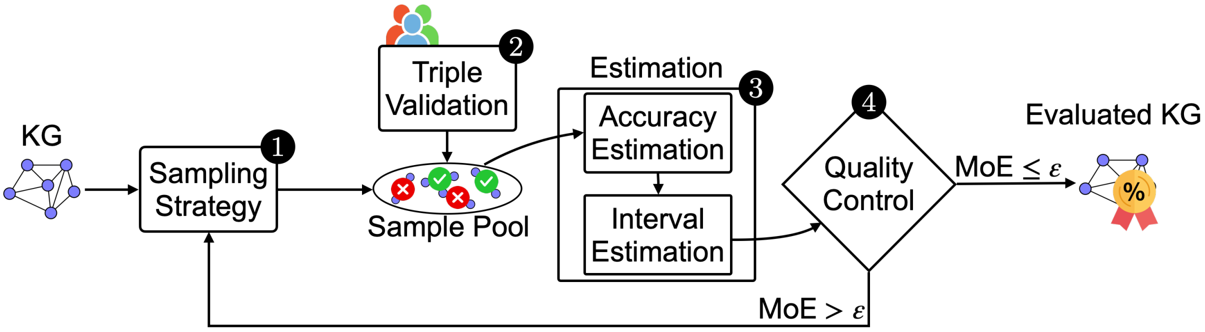

Figure 1: Efficient KG Accuracy Evaluation Framework.

2.3. Evaluation Framework

The evaluation framework addresses the constrained minimization problem through an iterative procedure, structured into four key phases, as illustrated in Figure 1.

Phase (1): Sampling. A small batch of triples is sampled from the KG using a specified sampling strategy $\mathcal{S}$. The chosen strategy aims to minimize annotation costs while maintaining the representativeness of the sample.

Phase (2): Validation. Each sampled triple undergoes manual annotation to validate its correctness. Annotations from each iteration are aggregated with previous ones to build a progressively larger annotated pool.

Phase (3): Estimation. An unbiased estimator $\hat{\mu}$ computes the accuracy of the KG based on the accumulated annotations and the sampling design $\mathcal{S}$. The corresponding $1 - \alpha$ CI is then constructed, with the objective of ensuring fast convergence and high reliability.

Phase (4): Quality Control. A quality control mechanism assesses whether the generated CI satisfies the predefined MoE threshold ε. Specifically, the framework checks if:

$$\mathrm{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$$

If this criterion is met, the process terminates, returning the final accuracy estimate and its associated CI. Otherwise, the procedure loops back to the sampling phase for additional data collection.

This iterative framework efficiently balances cost and precision by:

- **Preventing oversampling:** the iterative nature ensures that sampling stops as soon as the desired CI width is achieved, avoiding unnecessary annotations.
- **Ensuring reliability:** by focusing on unbiased point estimators and robust CIs, the framework provides reliable estimates that accurately represent the true KG accuracy.

Overall, the framework guarantees efficient, cost-effective, and reliable KG accuracy estimation by seamlessly integrating sampling strategies, unbiased estimation, and rigorous quality control.

2.4. Sampling and Accuracy Estimation

Recent approaches for efficient KG accuracy evaluation [16, 17] leverage well-established sampling methods and estimators [23]. We outline these strategies and their unbiased estimators below.

Simple Random Sampling. Simple Random Sampling (SRS) selects a sample of n_T triples from G without replacement. For large KGs, the probability of selecting the same triple twice is negligible, allowing the use of sampling with replacement [18], which is computationally more efficient.

The unbiased estimator for $\mu(G)$ under SRS is the sample proportion:

$$\hat{\mu}_{\text{SRS}} = \frac{1}{n_T} \sum_{i=1}^{n_T} \mathbb{1}(t_i)$$

with estimation variance:

$$V(\hat{\mu}_{\text{SRS}}) = \frac{\hat{\mu}_{\text{SRS}}(1 - \hat{\mu}_{\text{SRS}})}{n_T}$$

Cluster Sampling. Cluster sampling is an efficient alternative for evaluating large KGs [16, 17]. We first introduce Weighted Cluster Sampling (WCS), and then present Two-stage Weighted Cluster Sampling (TWCS) together with its estimator.

In WCS, n entity clusters are sampled with probabilities π_i proportional to their sizes, where $M_i = |G[e_i]|$ is the size of the ith cluster, and $\pi_i = M_i/M$. Since WCS requires manual evaluation of all triples in sampled clusters, it may become costly when clusters are large.

To address this, TWCS adopts a two-stage approach:

Stage 1: sample entity clusters using WCS;

Stage 2: from each sampled cluster i, select $\min\{M_i, m\}$ triples using SRS without replacement.

Let $\hat{\mu}_i$ denote the (estimated) mean accuracy of the sampled triples in the ith cluster. The unbiased estimator of $\mu(G)$ under TWCS is:

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

with estimation variance:

$$V(\hat{\mu}_{\text{TWCS}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\mu}_i - \hat{\mu}_{\text{TWCS}})^2$$

Stratified Sampling. Stratified Sampling (SS) partitions entity clusters into Q non-overlapping strata [23, 18]. Samples must be drawn from each stratum, raising the question of sample allocation. We employ proportional allocation [23], ensuring that the sample size in each stratum matches the proportion of units it contains.

Combining stratification with TWCS leads to the Stratified TWCS (STWCS) design. In each stratum q, TWCS is applied with a second-stage sample size m, yielding an unbiased estimator $\hat{\mu}_q$. Formally, let E_q be the set of N_q entities in stratum q, $\mathcal{C}_q = \{G[e] \mid e \in Eq\}$ the corresponding cluster family, and $C_q = \sum_{i=1}^{N_q} M_i$ its total size. The stratum weight W_q is given by $W_q = C_q/M$. The unbiased estimator for $\mu(G)$ under STWCS is:

$$\hat{\mu}_{\text{STWCS}} = \sum_{q=1}^{Q} W_q \cdot \hat{\mu}_q$$

with estimation variance:

$$V(\hat{\mu}_{\text{STWCS}}) = \sum_{q=1}^{Q} W_q^2 V(\hat{\mu}_q)$$

3. Interval Estimation

To quantify uncertainty in the sampling procedure, we estimate CIs for the sample. A CI identifies a range where the true population value is "likely" to lie, with a given confidence level $1 - \alpha$. Wider intervals indicate higher uncertainty. While multiple methods exist for computing CIs, binomial CIs are appropriate when defining KG accuracy as the proportion of correct triples (τ) over total triples (M) [19]. Although various binomial CIs exist [24, 25], state-of-the-art methods for KG accuracy estimation [16, 17] commonly use the Wald interval [18]. We first outline the limitations of the Wald interval and then introduce the Wilson interval [21], a binomial method that addresses these issues. In [1], we also provide a theoretical comparison between Wald, Wilson, and other binomial CIs that identifies the Wilson interval as the best trade-off between efficiency and reliability. However, due to space reasons, we omit such analysis in this extended abstract. We refer the interested reader to the original paper [1].

3.1. The Wald Interval

The Wald interval, based on normal approximation, is derived by inverting the acceptance region of the Wald test [18]:

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

where μ and $\hat{\mu}$ represent the true and estimated KG accuracies, $V(\hat{\mu})$ the estimation variance, and $z_{\alpha/2}$ the critical value for significance level α. Assuming a sufficiently large sample size ($n_{\mathcal{S}} \geq 30$ [26]), the $1 - \alpha$ CI becomes:

$$\hat{\mu} \pm z_{\alpha/2}\sqrt{V(\hat{\mu})}$$

While the Wald interval is simple, it is known to be flawed [19, 20]. When $\hat{\mu}$ approaches 0 or 1, $V(\hat{\mu})$ tends to zero, producing a zero-width interval that falsely implies certainty. Additionally, the interval may extend beyond the valid $[0, 1]$ range, violating the binomial nature of $\hat{\mu}$. These issues contribute to its erratic coverage probability, defined as the likelihood that the CI contains the true parameter. In this regard, the Wald interval frequently underperforms, offering coverage significantly below the nominal level $1 - \alpha$, especially for small samples or skewed data [19]. To ensure more reliable CIs, alternative methods are required.

3.2. The Wilson Interval

The Wilson interval [21] improves upon the Wald interval by using the null standard error in its test inversion:

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

Solving for μ yields:

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

This interval features a shifted center estimate and a corrected standard deviation. Unlike the Wald interval, the Wilson interval never collapses to zero width at the boundaries, preventing misleading certainty. Its asymmetry, pushing the center estimate toward the range midpoint, further enhances robustness. The Wilson interval thus offers more reliable coverage, especially for small samples and skewed observations, making it a superior choice for KG accuracy estimation.

However, we point out that the Wilson interval assumes the use of SRS as sampling strategy. Therefore, to apply it to more complex sampling strategies, such as TWCS and STWCS, design effect adjustments are necessary. Due to space constraints, we refer interested readers to the original paper [1].

Table 1
Data statistics for YAGO, NELL, and SYN 100M.

	YAGO	NELL	SYN 100M
Number of facts	1,386	1,860	101,415,011
Number of clusters	822	817	5,000,000
Average cluster size	1.69	2.28	20.28
Accuracy (μ)	0.99	0.91	n/a

4. Experimental Evaluation

In this section, we first outline the experimental setup and then present the experimental results.

4.1. Experimental Setup

The experimental setup covers: (i) dataset selection and rationale; (ii) annotation cost modeling for manual fact evaluation; (iii) implementation details; and (iv) evaluation procedure and chosen metrics.

Datasets. Table 1 summarizes the key statistics of the considered datasets. We select YAGO and NELL [15], as they are frequently used datasets in the literature for KG accuracy estimation[16, 17], and then we introduce SYN 100M, a large-scale synthetic KG used to test scalability. More datasets are considered in the original work [1].

YAGO is sampled from YAGO2 [4], a large KG of general knowledge (e.g., people, cities, movies). Each fact in the sample is manually annotated, resulting in a ground-truth accuracy of $\mu = 0.99$.

NELL is drawn from the NELL KG [5], focusing on sports-related facts (e.g., athletes, teams, stadiums). As with YAGO, manually annotated labels are provided, with a ground-truth accuracy of $\mu = 0.91$.

To assess scalability, we generate **SYN 100M**, a synthetic KG containing over 100 million triples. Clusters were generated with a mean size of 20 and a standard deviation of 15. Correctness labels were generated by setting the probability of a triple being true to a fixed rate. Specifically, we set probability rates to $\{0.1, 0.5, 0.9\}$, leading to ground-truth accuracies $\mu \in \{0.9, 0.5, 0.1\}$.

Cost Function. To quantify the cost of manually evaluating fact correctness within the sampled subset $G_{\mathcal{S}}$, we adopt the cost function from [16]. The function assumes that annotating additional facts for an already identified entity is less costly than evaluating facts from previously unseen entities. It is defined as:

$$\text{cost}(G_{\mathcal{S}}) = |E_{\mathcal{S}}| \cdot c_1 + |T_{\mathcal{S}}| \cdot c_2 \tag{1}$$

Here, $|E_{\mathcal{S}}|$ and $|T_{\mathcal{S}}|$ denote the number of unique entities and facts in $G_{\mathcal{S}}$, respectively. The parameters c_1 and c_2 represent the average annotation cost (in seconds) for entity identification and fact verification. Following [16], we set $c_1 = 45$ seconds and $c_2 = 25$ seconds.

Implementation. We evaluate three **sampling strategies**: SRS, TWCS, and STWCS. Following Gao et al. [16], we set the second-stage sample size m to 3 for YAGO and NELL (due to smaller average cluster sizes) and to 5 for SYN 100M (where clusters are larger). For STWCS, entity clusters are stratified by the subject entity's degree centrality using the Cumulative Square Root of Frequency (Cumulative $\sqrt{F}$) method [27], with $Q = 2$ strata per KG.

To construct **confidence intervals**, we use the Wald interval as a baseline and compare it with our Wilson-based solution. Configurations using the Wald interval represent state-of-the-art baselines [16, 17] and are labeled as {method} (Wald). Our newly proposed Wilson-based approaches are labeled as {method} (Wilson).

Table 2
Performance on YAGO and NELL. Methods achieving nominal coverage are in **bold**.

	YAGO			NELL		
	$\mu = 0.99$			$\mu = 0.91$		
Method	Triples	Cost	Coverage	Triples	Cost	Coverage
SRS (Wald)	33±6	0.61±0.11	0.20±0.03	107±40	1.96±0.71	0.82±0.02
SRS (Wilson)	40±10	0.75±0.17	**0.99±0.01**	116±33	2.13±0.59	**0.98±0.01**
TWCS (Wald)	31±3	0.42±0.05	0.31±0.03	124±68	1.56±0.85	0.75±0.03
TWCS (Wilson)	35±4	0.46±0.05	**0.95±0.01**	128±65	1.60±0.82	**0.98±0.01**
STWCS (Wald)	34±8	0.45±0.10	0.31±0.03	105±59	1.32±0.74	0.73±0.03
STWCS (Wilson)	40±7	0.54±0.10	**0.94±0.02**	100±57	1.26±0.72	0.92±0.02

Evaluation Procedure. We set the significance level to $\alpha = 0.05$ and the MoE upper bound to $\varepsilon = 0.05$. A minimum of 30 annotated triples is required, and the evaluation procedure is repeated 1,000 times for each method. Performance comparison relies on three key metrics: number of annotated triples, annotation cost (measured in hours), and empirical coverage. Empirical coverage is defined as the proportion of evaluation runs in which the constructed CIs contain the ground-truth accuracy. This metric assesses how close the CIs are to the nominal coverage probability – i.e., 95% when $\alpha = 0.05$.

Results are presented only when MoE ≤ 0.05, ensuring all methods are evaluated when they meet the optimization objective. Consequently, we omit CI widths, as all reported solutions have MoE ≤ 0.05. Similarly, accuracy estimates are excluded since all methods provide unbiased estimates with negligible differences (≤ 0.02) from the ground truth.

4.2. Experimental Results

We present the results of the comparison between the proposed Wilson-based methods and the Wald-based baselines, adopting SRS, TWCS, and STWCS as sampling strategies. The evaluation spans two KGs: YAGO and NELL. The results, summarized in Table 2, detail the number of annotated triples, the annotation cost, and the empirical coverage.[1]

The analysis in Table 2 reveals the following trends. On YAGO and NELL, Wilson-based methods demonstrate lower efficiency but higher reliability than Wald counterparts – improving reliability on YAGO by up to a factor of two. Notably, the relative increase in annotation costs for Wilson solutions is less pronounced when using TWCS or STWCS compared to the SRS scenario. On NELL, STWCS (Wilson) even reduces annotation costs by 5% relative to STWCS (Wald) while achieving empirical coverage close to the nominal 0.95 level, thus boosting reliability by 26%.

Overall, TWCS and STWCS significantly reduce evaluation costs compared to SRS, albeit with a slight drop in coverage. This reduction stems from the complexity introduced by clustering and stratification in TWCS and STWCS. Nevertheless, when paired with Wilson-based methods, both strategies consistently achieve coverage probabilities near the nominal value across all evaluated KGs, outperforming Wald solutions. Consequently, Wilson-based approaches again emerge as the optimal balance between efficiency and reliability.

We also assess the scalability of the proposed methods by examining whether the findings from previous experiments hold for SYN 100M. Results are summarized in Table 3.

The results reveal two key insights. First, despite SYN 100M being orders of magnitude larger than NELL and YAGO, the number of annotations and annotation cost remained comparable across all methods. This indicates that the annotation scheme primarily influences evaluation cost, while the size and topological structure of the KG have negligible impact. This outcome is particularly noteworthy, as

[1] Note that further analyses are available in the original paper [1], highlighting how the methods perform across KGs with varying accuracy levels, sizes, and topologies.

Table 3
Performance on SYN 100M with $\mu \in \{0.9, 0.5, 0.1\}$. Methods achieving nominal coverage are in **bold**.

	SYN 100M								
	$\mu = 0.9$			$\mu = 0.5$			$\mu = 0.1$		
Method	Triples	Cost	Coverage	Triples	Cost	Coverage	Triples	Cost	Coverage
SRS (Wald)	122±43	2.38±0.83	0.83±0.02	385±2	7.50±0.03	**0.95±0.02**	126±43	2.44±0.84	0.84±0.02
SRS (Wilson)	131±34	2.56±0.66	**0.93±0.02**	380±1	7.39±0.03	**0.95±0.02**	133±34	2.59±0.66	**0.93±0.02**
TWCS (Wald)	122±52	1.16±0.50	0.83±0.02	386±62	3.65±0.59	**0.95±0.02**	119±52	1.12±0.49	0.81±0.02
TWCS (Wilson)	123±51	1.16±0.48	**0.93±0.02**	379±59	3.58±0.56	**0.94±0.02**	124±47	1.17±0.45	**0.93±0.02**
STWCS (Wald)	117±60	1.11±0.57	0.77±0.02	385±76	3.64±0.72	0.92±0.02	117±56	1.11±0.53	0.82±0.02
STWCS (Wilson)	122±57	1.15±0.54	0.88±0.02	382±76	3.61±0.72	**0.94±0.02**	125±57	1.18±0.56	0.88±0.02

it not only confirms that the procedure is insensitive to KG size – consistent with previous findings by Gao et al. [16] – but also demonstrates its robustness to structural differences between KGs (cf. Table 1). Secondly, the performance patterns observed for YAGO and NELL persist in SYN 100M. Specifically, the proposed methods achieve significant coverage improvements when the KG accuracy is near 0 or 1. Additionally, they offer lower annotation costs when the accuracy deviates from these boundaries.

5. Conclusions

In this extended abstract, we reported the key limitations of current state-of-the-art approaches for KG accuracy evaluation. Specifically, their reliance on Wald-based CIs leads to zero-width and overshooting intervals, undermining estimation reliability. To address these issues, we introduced the Wilson interval, a binomial interval ensuring higher statistical guarantees. Through various analyses we demonstrated that Wilson offers the best balance between efficiency and reliability.

Building on Wilson intervals, we developed solutions that push the state-of-the-art forward. Extensive experiments on diverse real-world and synthetic KGs – varying in accuracy levels, sizes, and topologies – show that our Wilson-based methods (i) can be up to twice as reliable as existing solutions when KG accuracy approaches its boundaries and (ii) achieve greater efficiency than Wald-based methods when KG accuracy gets close to 0.5.

After a comprehensive comparison of advanced sampling strategies, including TWCS and STWCS, we recommend practitioners adopt TWCS paired with Wilson intervals to achieve efficient and reliable KG accuracy estimation.

Acknowledgments

The work was supported by the HEREDITARY project, as part of the EU Horizon Europe program under Grant Agreement 101137074.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] S. Marchesin, G. Silvello, Efficient and Reliable Estimation of Knowledge Graph Accuracy, Proc. VLDB Endow. 17 (2024) 2392–2404. URL: https://www.vldb.org/pvldb/vol17/p2392-marchesin.pdf. doi:10.14778/3665844.3665865.

[2] D. Vrandecic, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. G. Ives, 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, volume 4825 of *LNCS*, Springer, 2007, pp. 722–735. URL: https://doi.org/10.1007/978-3-540-76298-0_52. doi:10.1007/978-3-540-76298-0_52.

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

[5] T. M. Mitchell, W. W. Cohen, E. R. H. 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, J. Welling, Never-ending learning, Commun. ACM 61 (2018) 103–115. URL: https://doi.org/10.1145/3191513. doi:10.1145/3191513.

[6] O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, A. Doan, 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, 2013, pp. 1209–1220. URL: https://doi.org/10.1145/2463676.2465297. doi:10.1145/2463676.2465297.

[7] J. Pujara, E. Augustine, L. Getoor, 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, 2017, pp. 1751–1756. URL: https://doi.org/10.18653/v1/d17-1184. doi:10.18653/v1/d17-1184.

[8] S. Marchesin, G. Silvello, O. Alonso, 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, 2024, pp. 1649–1659. URL: https://doi.org/10.1145/3627673.3679561. doi:10.1145/3627673.3679561.

[9] S. Marchesin, G. Silvello, O. Alonso, 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 (2024) 105–114. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/31605. doi:10.1609/hcomp.v12i1.31605.

[10] I. F. Ilyas, T. Rekatsinas, V. Konda, J. Pound, X. Qi, M. A. Soliman, 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, 2022, pp. 2259–2272. URL: https://doi.org/10.1145/3514221.3526049. doi:10.1145/3514221.3526049.

[11] I. F. Ilyas, J. Lacerda, Y. Li, U. F. Minhas, A. Mousavi, J. Pound, T. Rekatsinas, C. Sumanth, 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, 2023, pp. 253–259. URL: https://doi.org/10.1145/3555041.3589672. doi:10.1145/3555041.3589672.

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

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

[14] M. Samadi, P. P. Talukdar, M. M. Veloso, T. M. Mitchell, 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, 2015, pp. 837–843. URL: http://ijcai.org/Abstract/15/123.

[15] P. Ojha, P. P. Talukdar, 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, 2017, pp. 1741–1750. URL: https://doi.org/10.18653/v1/d17-1183. doi:10.18653/v1/d17-1183.

[16] J. Gao, X. Li, Y. E. Xu, B. Sisman, X. L. Dong, J. Yang, Efficient Knowledge Graph Accuracy Evaluation, Proc. VLDB Endow. 12 (2019) 1679–1691. URL: http://www.vldb.org/pvldb/vol12/p1679-gao.pdf. doi:10.14778/3342263.3342642.

[17] Y. Qi, W. Zheng, L. Hong, L. Zou, 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, 2022, pp. 1368–1378. URL: https://doi.org/10.1145/3534678.3539233. doi:10.1145/3534678.3539233.

[18] G. Casella, R. L. Berger, Statistical Inference, Duxbury advanced series in statistics and decision sciences, Thomson Learning, 2002. URL: https://books.google.it/books?id=0x_vAAAAMAAJ.

[19] L. D. Brown, T. T. Cai, A. DasGupta, Interval Estimation for a Binomial Proportion, Statistical Science 16 (2001) 101–117. URL: http://www.jstor.org/stable/2676784.

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

[21] E. B. Wilson, Probable Inference, the Law of Succession, and Statistical Inference, Journal of the American Statistical Association 22 (1927) 209–212. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502953. doi:10.1080/01621459.1927.10502953.

[22] A. Bonifati, G. H. L. Fletcher, H. Voigt, N. Yakovets, Querying Graphs, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2018. URL: https://doi.org/10.2200/S00873ED1V01Y201808DTM051. doi:10.2200/S00873ED1V01Y201808DTM051.

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

[24] S. E. Vollset, Confidence Intervals for a Binomial Proportion, Statistics in Medicine 12 (1993) 809–824. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780120902. doi:https://doi.org/10.1002/sim.4780120902.

[25] R. G. Newcombe, Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods, Statistics in Medicine 17 (1998) 857–872. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819980430%2917%3A8%3C857%3A%3AAID-SIM777%3E3.0.CO%3B2-E. doi:https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.

[26] R. V. Hogg, E. A. Tanis, D. L. Zimmerman, Probability and Statistical Inference, Pearson, 2013. URL: https://books.google.it/books?id=I_7tnQEACAAJ.

[27] T. Dalenius, J. L. Hodges, Minimum Variance Stratification, Journal of the American Statistical Association 54 (1959) 88–101. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1959.10501501. doi:10.1080/01621459.1959.10501501.