Additional Experimental Results

1 Setup

We present additional experimental results using Stratified TWCS (STWCS) [1, 2], a sampling strategy that first partitions entity clusters into non-overlapping strata and then applies TWCS within each stratum. YAGO and NELL have two strata, while DBPEDIA and FACTBENCH have four. For the second-stage size m, we used the same values as in the TWCS setup: $m = 3$ for YAGO, NELL, DBPEDIA, and FACTBENCH; and $m = 5$ for SYN 100M.

Although these results are consistent with those presented in the main text, they are less central and are therefore provided online due to space constraints.

2 Results

We provide results for the efficiency, scalability, and robustness analyses.

2.1 Efficiency

Table 1 compares the performance of *a*HPD with Wald and Wilson baselines under STWCS.

Table 1. Performance on YAGO, NELL, DBPEDIA, and FACTBENCH. Best performance are in **bold**.

		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
STWCS	Wald	33±6	0.43±0.08	106±57	1.34±0.71	240±82	2.77±0.94	148±52	1.80±0.62
	Wilson	39±6	0.52±0.08	101±56	1.26±0.70	214±88	2.47±1.01	**132±57**	**1.60±0.68**
	*a*HPD	**31±2**	**0.41±0.03**	**87±54**	**1.10±0.68**	**203±93**	**2.34±1.07**	**132±57**	**1.60±0.68**

Consistent with results from SRS and TWCS, *a*HPD outperforms both Wald and Wilson on YAGO, NELL, and DBPEDIA, where KG accuracy is skewed, and remains competitive on FACTBENCH, which serves as a controlled scenario to investigate quasi-symmetric cases.

2.2 Scalability

Table 2 showcases the performance of *a*HPD on SYN 100M datasets, compared to Wald and Wilson methods.

Table 2. Performance on SYN 100M with accuracy values $\mu \in \{0.9, 0.5, 0.1\}$. Best performance are in **bold**.

		SYN 100M					
		$\mu = 0.9$		$\mu = 0.5$		$\mu = 0.1$	
Sampling	Interval	Triples	Cost	Triples	Cost	Triples	Cost
STWCS	Wald	118±59	1.11±0.56	377±84	3.57±0.79	113±56	1.07±0.53
	Wilson	124±58	1.18±0.55	**376±78**	**3.56±0.74**	119±54	1.13±0.51
	*a*HPD	**110±60**	**1.04±0.57**	**376±78**	**3.56±0.74**	**107±57**	**1.01±0.54**

The trend observed in Table 1 holds even as the dataset size scales up, as shown in Table 2, confirming the consistency of *a*HPD performance across different sampling strategies, regardless of the complexity of the sampling scheme.

SIGMOD '25, June 22-27, 2025, Berlin, Germany.

2.3 Robustness

Figure 1 illustrates the annotation costs of *a*HPD at different precision levels, comparing them with Wilson costs under STWCS. The reduction ratio of *a*HPD relative to Wilson is also displayed for each case.

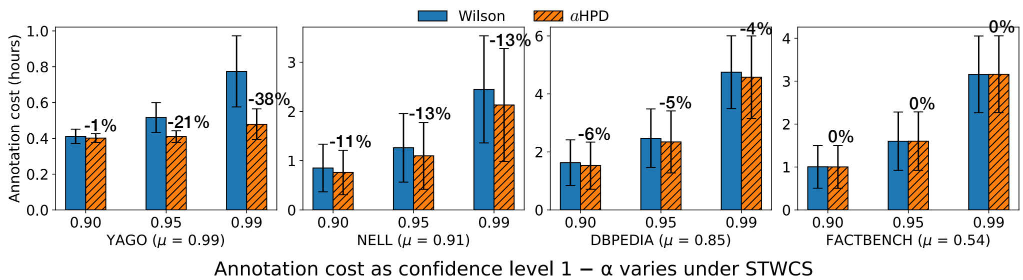

Fig. 1. Annotation cost comparison between *a*HPD and Wilson at different confidence levels $1 - \alpha$ under STWCS on YAGO, NELL, DBPEDIA, and FACTBENCH KGs. We also report the reduction ratio (in %) of *a*HPD over Wilson.

The results in Figure 1 align with those under SRS and TWCS, reinforcing the all-around superiority of *a*HPD compared to the considered baselines. These outcomes provide further evidence to support the adoption of *a*HPD with any sampling strategy and in any scenario where there is a need to evaluate KG accuracy with limited annotations.

References

[1] 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. https://doi.org/10.14778/3342263.3342642

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