Conformal Prediction Sets for Graph Neural Networks

Soroush H. Zargarbashi[1], Simone Antonelli[1], Aleksandar Bojchevski[2]

(1) CISPA Helmholtz Center for Information Security, (2) University of Cologne

TL;DR

- Conformal prediction for semi-supervised node classification returns *sets* that provably contain the true label.
- Extensive study of graphs + exchangeability.
- Neighborhood diffusion leads to more efficient prediction sets, and significantly improves correct singletons.

1. Motivation

Quantification of the model's uncertainty is critical

Models are usually uncalibrated with no uncertainty guarantees

Distribution-free coverage guarantees via Conformal Prediction

2. Conformal Prediction

Given: *holdout labeled set* $D_{cal} \subset D$ (exchangeable),

black-box model M trained with $D_{tr} \cup D_{val}$,

user-specified α, test input $\mathbf{x}_{n+1}$

Compute: $q = \mathrm{Quantile}(\alpha, \{s_i = s(x_i, y_i)\})$

Returns: $\mathbb{P}[\underbrace{\mathbf{y}_{n+1}}_{\text{true label}} \in \underbrace{\{y : s(\mathbf{x}_{n+1}, y) \geq q\}}_{\text{prediction set}}] > 1 - \alpha.$

Any conformity score $s(\cdot, \cdot)$ works! Goal: Design a good s!

3. Scores, and Metrics

APS: $s(\mathbf{x}, y) = -\left(\sum_{c=1}^{K} \pi(\mathbf{x})_c \mathbb{I}[\pi(\mathbf{x})_c > \pi(\mathbf{x})_y] + u \cdot \pi(\mathbf{x}, y)\right)$

RAPS: Reducing size by *penalizing low rank* scores.

Metrics:

- Coverage: guaranteed and user-specified!
- Average Set Size: lower is better
- (Correct) Singleton Hits: higher is better

Github Repository

ICML 2023

4. CP for Node Classification

Settings:

- **Transductive:** Exchangeable

Train: entire graph | Calibrate: entire graph | Train: entire graph

- **Simultaneous Inductive:** Exchangeable

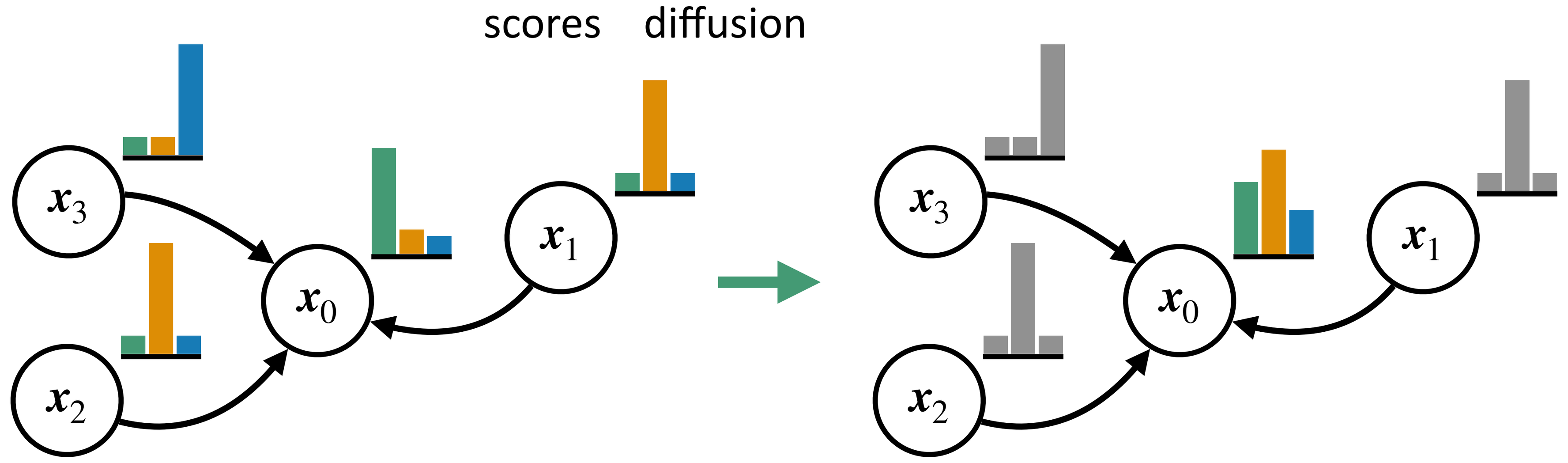
Train: sub-graph | Calibrate: entire graph | Train: entire graph

- **Inductive:** Not Exchangeable

Train: sub-graph | Calibrate: sub-graph | Train: entire graph

5. DAPS: Diffused Adaptive Prediction Sets

1-hop DAPS : $\hat{S} = (1-\lambda)\underbrace{S}_{\text{scores}} + \underbrace{\lambda D^{-1}AS}_{\text{diffusion}}$



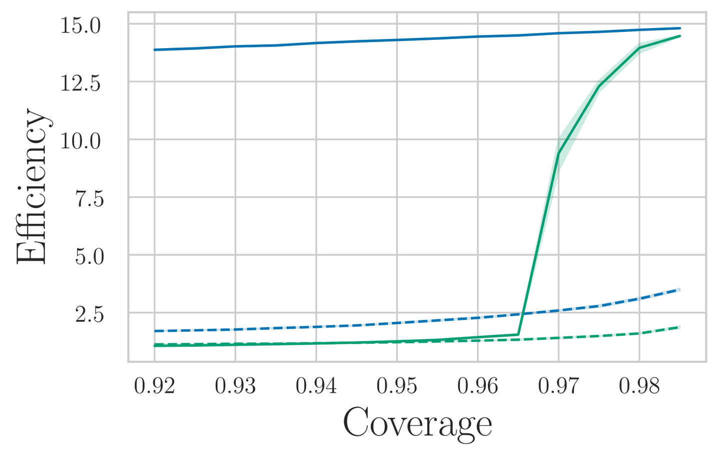
K-hop DAPS: Apply on k-hop neighborhood

Score Propagation: $\hat{S} = (1-\lambda)S + (I - \lambda D^{-1}A)^{-1}S$

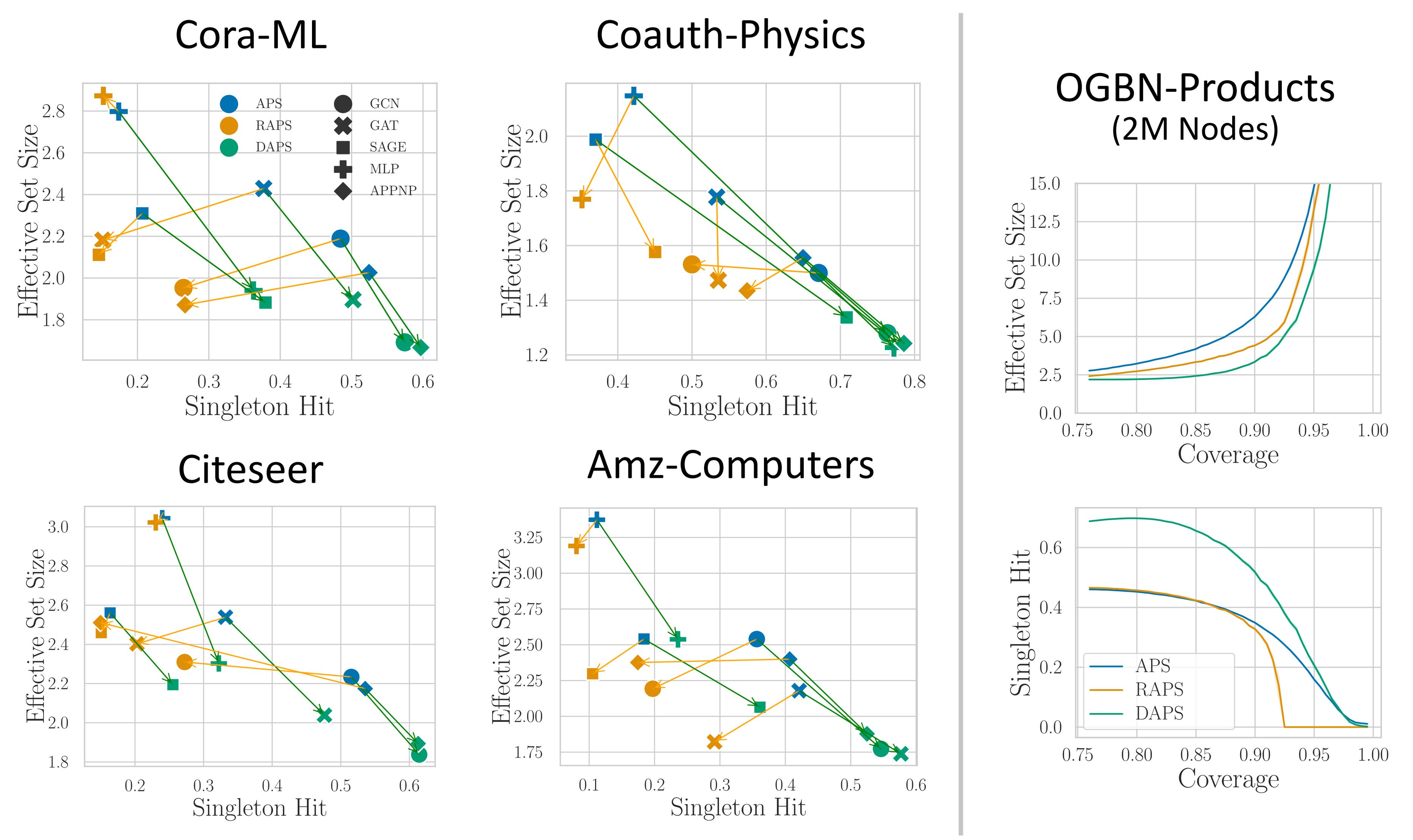
Intuition: Diffusion makes $\underbrace{\pi(y\,|\,x)}_{\text{predicted}}$ closer to $\underbrace{p(y\,|\,x)}_{\text{true}}$.

6.1. DAPS Handles Hard Predictions

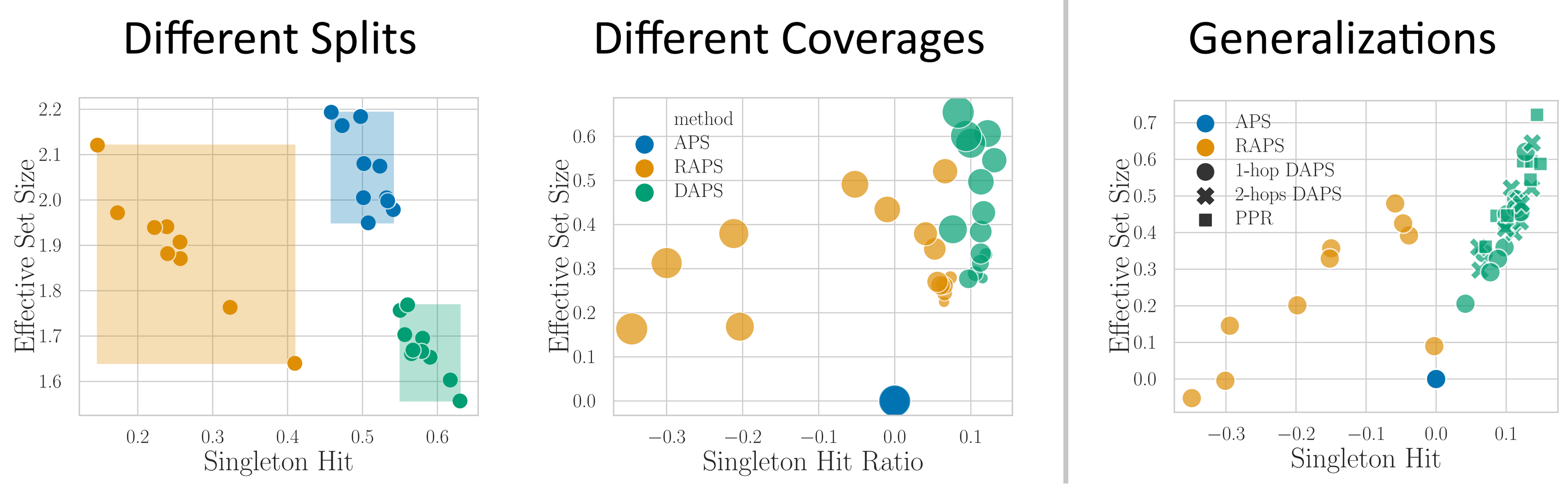
We can define DAPS over black-box prediction machines, which return only top class labels.



6.2. DAPS is Pareto Dominating and Scalable

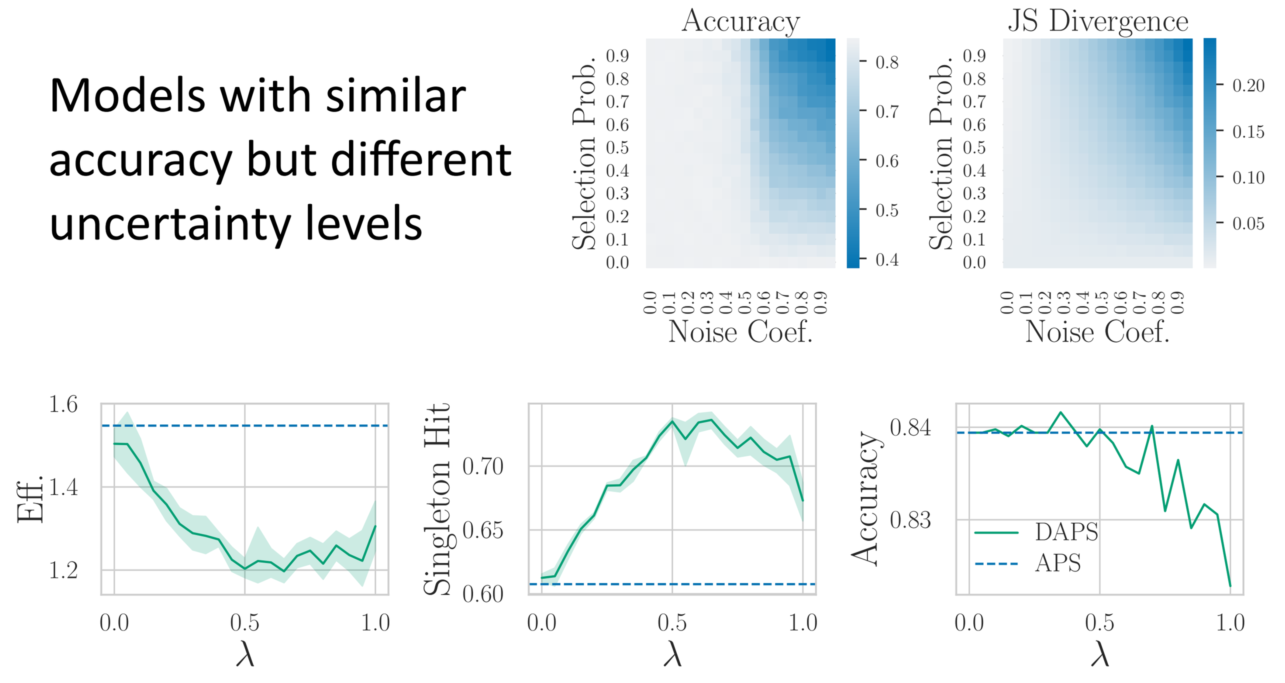


6.3. DAPS (and its Generalizations) are More Stable



6.4. DAPS Enhances Uncertainty Awareness

Models with similar accuracy but different uncertainty levels



Applicable to all Scores