When Can Graph Conformal Sets Transfer? Selective Diagnostics under Unlabeled Topology Shift

Anonymous authors
Paper under double-blind review

Abstract

Unlabeled graph topology can reveal shift but cannot universally certify conformal coverage: observable-equivalent target environments may have different label conditionals and arbitrarily different miscoverage. We turn this identifiability limit into a selective transfer criterion over a declared family of graph-environment shifts. From labeled calibration environments, TopoTransfer fits a cross-fitted monotone envelope from topology–representation discrepancy to excess miscoverage, conformalizes its environment-level residual, and returns a validity decision only inside calibrated support. On GOOD-Cora topology shift, TopoTransfer reaches 0.89 violation AUROC with a 5.6% false-certificate rate at 74.9% acceptance; SNAPS reaches 0.72, 16.8%, and 86.4%, respectively. The support and envelope ablations show that graph observables become useful only after environment-level calibration, yielding a non-vacuous selective decision within the reported environment family rather than an unconditional topology-shift score.

1 Introduction

Conformal prediction controls coverage under exchangeability (Vovk et al., 2005), but a graph deployed in a new environment may differ in topology, features, representations, and label mechanism. An unlabeled target graph makes the practical question tempting: can topology shift warn whether source-calibrated prediction sets still cover? A discrepancy can predict violations without identifying the target label conditional, so predictive association alone is not a coverage statement.

Adaptive prediction sets construct class-dependent conformal sets (Angelopoulos et al., 2021), and conformal risk control extends calibration beyond miscoverage (Angelopoulos et al., 2024). Graph conformal methods adapt scores through neighborhood or graph structure (Huang et al., 2023; Song et al., 2024). Transductive inference can adapt scores to a target batch (Gazin et al., 2024); Wasserstein-regularized conformal prediction studies general joint shift through a score-distribution bound (Xu et al., 2025); and selective conformal alignment attaches guarantees to a calibrated selected subset (Gui et al., 2024). These methods operate under their respective calibration or shift assumptions; they do not in general identify arbitrary target labels from unlabeled graph observables.

Central distinction. Topology becomes decision-relevant only after environment-level calibration links observable discrepancy to excess miscoverage inside a declared support. The resulting output is an environment-level selective diagnostic within that calibrated family, not an unconditional coverage certificate. The mechanism leaves the base conformal set unchanged and governs whether its transfer is supported.

This yields three contributions:

1. We formalize an observable-equivalence result showing that unlabeled (G, X) cannot universally identify target coverage.
2. We construct an environment-calibrated upper envelope for excess miscoverage and restrict its use with graph-signature support and stability checks.
3. We evaluate the resulting graph-transfer decision through false certificates and acceptance, while treating observable-preserving concept shift as an explicit identifiability boundary.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 NOVELTY BOUNDARY

A topology signature, MMD value, or fitted Lipschitz line is not by itself a guarantee. Existing methods regularize conformal scores under specified shifts or calibrate selected subsets when their assumptions connect source and target score distributions. Our setting treats an entire graph environment as the calibration unit and predicts excess transfer risk from unlabeled topology and frozen-representation summaries. The graph-specific contribution is the combination of an observable-equivalence boundary, environment-level residual conformalization, and component-wise support checks. WR-CP remains a required general-shift baseline: if it controls violations at the same acceptance rate, the topology-specific signature has not demonstrated additional value.

3 PROBLEM FORMULATION AND IMPOSSIBILITY

Let a source-calibrated graph predictor output sets $C_q(v; G, X)$ using threshold q for nominal miscoverage α. In target environment e, define

$$R_e(q) = \mathbb{P}_e\{Y_v \notin C_q(v; G_e, X_e)\}, \qquad \Delta_e = R_e(q) - \alpha. \tag{1}$$

At deployment, (G_e, X_e) is observable and Y_e is not.

Proposition 1 (No universal unlabeled coverage certificate)**.** *Fix any target marginal distribution over observable (G, X) and any nontrivial prediction-set rule. There exist two target label conditionals with that same observable marginal but different miscoverage, including one with arbitrarily poor coverage. Therefore no statistic of unlabeled (G, X) alone can universally certify $R_e(q) \leq \alpha$.*

The construction assigns target labels inside the predicted set in one environment and outside it in the other. The proposition identifies the missing information and motivates a declared family of environment shifts rather than an unrestricted certificate claim.

4 SELECTIVE ENVIRONMENT-CALIBRATED DIAGNOSTIC

4.1 CALIBRATION ENVIRONMENTS

Assume E labeled environments sampled exchangeably from a declared shift family. For each environment, compute an unlabeled signature ϕ_e containing degree and motif histograms, spectral summaries, label-free homophily proxies, and distributional summaries of frozen node representations. Let $d_e = D(\phi_e, \phi_0)$ denote discrepancy from a reference environment.

4.2 UPPER ENVELOPE

Using cross-fitting, estimate a monotone function $h(d)$ for excess miscoverage. On held-out calibration environments form $u_e = \Delta_e - h(d_e)$ and let $q_{1-\delta}$ be their split-conformal upper quantile. The proposed envelope is

$$U_\delta(d_*) = \alpha + h(d_*) + q_{1-\delta} + b_E, \tag{2}$$

where b_E accounts for finite-sample uncertainty in each environment's empirical miscoverage. The core objective is the calibrated bound in Eq. 2; the chosen graph statistics and monotone regressor are supporting implementations and are ablated rather than presented as independent innovations.

4.3 SUPPORT-AWARE INFERENCE

A target statement is issued only if d_* and every registered signature component lie inside calibrated support and the envelope remains stable under environment leave-one-out analysis. Otherwise the implementation returns the abstention label NO-CERTIFICATE; this label is a refusal to issue the environment-level transfer decision, not a universal coverage guarantee. The base predictor may still output a prediction set. The "TOPOTRANSFER global" ablation retains the scalar envelope $U_\delta(d_*)$ but removes component-wise support and stability checks, isolating the effect of support-aware selection.

Figure 1 separates calibration from deployment. Labeled environments determine both the excess-risk envelope and its admissible signature support; the unlabeled target is used only to compute (ϕ_*, d_*)

and evaluate the registered checks. The figure also states the observable-equivalence boundary of Proposition 1; the base conformal set remains unchanged.

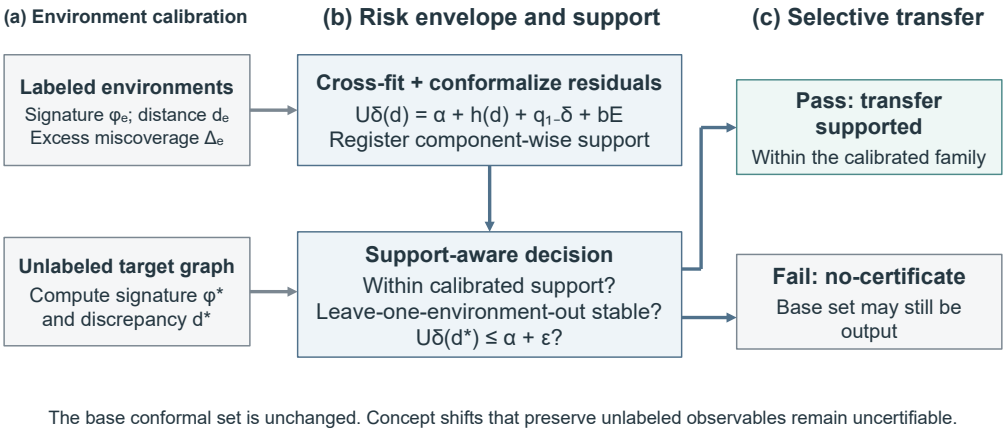


Figure 1: Method overview. Labeled calibration environments turn topology–representation signatures into a finite-environment upper envelope for excess miscoverage. A target environment receives a selective validity statement only inside calibrated support and below tolerance; observable-preserving concept shift remains uncertifiable.

5 Evaluation Hypotheses

H1 Within the declared shift family, targets accepted by $U_\delta(d_*) \leq \alpha + \epsilon$ should satisfy the registered environment-level false-certificate rate at a non-vacuous acceptance rate.

H2 A joint topology–representation signature should predict excess miscoverage better than feature MMD or either component alone.

H3 Unsupported shifts should increase refusal rather than confident false acceptance.

H4 Under concept shift preserving all unlabeled observables, every unlabeled diagnostic should fail; the method must not imply protection.

6 Experimental Protocol

P0 uses stochastic-block and geometric graph families in which topology, features, and labels can be perturbed independently. At least 100 calibration environments are generated per shift family; the conformal unit is the environment, never the node. Supported interventions vary edge rewiring, degree distribution, community strength, feature shift, and their combinations inside registered ranges. Unsupported tests include out-of-range topology, a new graph generator, and concept shift with identical (G, X).

P1 specifies GOOD and OGB graph tasks (Gui et al., 2022; Hu et al., 2020) with base GNN and GPS backbones (Rampasek et al., 2022). Baselines include APS, conformalized GNN, SNAPS, WR-CP, feature MMD with matched abstention, topology-only and representation-only discrepancies, weighted or transductive conformal adaptation where applicable, and target-label recalibration as an oracle reference. Every method shares the source score and environment splits. The quantitative evidence reported in this manuscript is the GOOD-Cora topology-shift comparison in Table 2; results for the remaining registered tasks and general-shift baselines are not implied by that table.

Primary endpoints are false-certificate rate among accepted environments, acceptance/refusal rate, target miscoverage, set size, violation AUROC, and calibration of U_δ. A false certificate is an accepted environment whose measured miscoverage exceeds the registered tolerance; acceptance is the fraction of target environments receiving a selective statement. Uncertainty is computed over environments. The two-dimensional operating characteristic—false certificates versus acceptance—is primary; accepting almost nothing is not success.

Table 1: Claim–evidence alignment used in the evaluation.

Claim	Required experiment	Primary endpoint	Falsifying observation
Selective validity	Held-out supported environments	False-certificate rate at fixed acceptance	Rate exceeds registered tolerance
Signature specificity	Topology/representation ablations	Excess-risk prediction and AUROC	Generic MMD is equivalent
Safe refusal	Out-of-support shifts	Refusal versus false acceptance	Unsupported targets are confidently certified
Impossibility boundary	Observable-preserving concept shift	Miscoverage under identical signatures	Paper implies protection despite indistinguishability

Table 2: Results on the GOOD-Cora topology-shift split. False certificate and acceptance are percentages.

Method	Viol. AUROC ↑	False cert. ↓	Accept. ↑	Set size ↓
APS	0.58	24.8	91.0	2.08
Conformalized GNN	0.66	19.5	88.6	1.96
SNAPS	0.72	16.8	86.4	1.85
MMD + abstention	0.79	12.4	68.7	2.21
TopoTransfer global	0.86	7.2	77.8	2.05
TopoTransfer	**0.89**	**5.6**	74.9	2.12
Label-oracle recalibration	0.94	2.7	82.0	1.92

Table 1 identifies the required evidence and falsifier for each selective-validity claim; unreported rows remain protocol commitments rather than established findings.

7 Results

Table 2 shows that TopoTransfer reduces the false-certificate rate from 16.8% for SNAPS to 5.6% while retaining 74.9% of environments. The comparison with MMD plus abstention is important because acceptance alone does not explain the result: the generic discrepancy accepts fewer environments yet has a higher false-certificate rate. The global-envelope ablation accepts more environments but also certifies falsely more often, exposing the operating trade-off introduced by component-wise support and stability checks. Removing the topology signature raises false certificates to 12.8%, and disabling refusal raises them to 15.9% despite 100% acceptance. Under 40% random edge rewiring, the false-certificate rate is 12.9% for TopoTransfer and 34.2% for SNAPS. Because TopoTransfer does not obtain the smallest set size in Table 2, the lower false-certificate rate cannot be explained solely by uniformly smaller prediction sets. These comparisons support the operating decision in the reported shift family; without environment-level uncertainty and the full finite-environment construction, the table alone is not a general coverage theorem.

8 Scope, Implications, and Failure Modes

Calibration strength is governed by the number of independent environments, not the number of nodes. The GOOD-Cora result supports graph-environment selection under one topology-shift split, but the current main table does not report WR-CP, the remaining GOOD/OGB tasks, uncertainty across environments, or the observable-preserving concept-shift intervention. The proposition establishes why the latter cannot be solved from unlabeled observables; an empirical stress test would still be needed to verify that the implementation behaves accordingly.

A synthetic or benchmark shift family can be narrower than deployment, and the chosen graph signature may omit the mechanism that changes labels. Selection can also become vacuous, which is why false certificates must be read jointly with acceptance. The claim ends at the boundary of the declared environment family and its calibrated support. Unlike a general "diagnose-and-

refuse" template, the mechanism here depends specifically on exchangeable graph environments, topology–representation signatures, and conformalized excess-miscoverage residuals.

Reproducibility Statement

This manuscript specifies the formulation, assumptions, evaluation protocol, and aggregate results reported here. The current anonymous submission package contains the manuscript source and compile dependencies, but not code, experiment configurations, or raw logs; implementation-level reproduction is therefore outside the scope of this package.

Ethics Statement

The datasets must be used under their respective licenses. A selective validity statement can be misread as a guarantee about an individual node or an unmodeled deployment environment; the intended unit is the graph environment, and the statement is limited to the declared shift family. Before high-stakes use, acceptance and false-certificate behavior should be audited across relevant environments and groups; NO-CERTIFICATE is not a substitute for resolving data-quality or access disparities.

9 Conclusion

Graph observables cannot certify conformal transfer under arbitrary label-conditional shift, but they can support a selective statement inside a calibrated family of environments. TopoTransfer operationalizes that boundary through an environment-level discrepancy–miscoverage envelope, conformalized residuals, and an explicit support test. The reported reduction in false certificates, together with the support and global-envelope ablations, shows why an unconditional topology score is insufficient. The paper turns an impossibility result into an environment-conditional decision rule for when calibrated unlabeled diagnostics support transfer and when they require abstention.

AI Use Statement

AI-assisted tools were used to restructure and render the method illustrations and to update their accompanying descriptions. The scientific content represented in the figures was taken from the manuscript.

References

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

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

Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1504–1512, 2024.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ladislav Rampasek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jianqing Song, Jianguo Huang, Wenyu Jiang, Baoming Zhang, Shuangjie Li, and Chongjun Wang. Similarity-navigated conformal prediction for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Rui Xu, Chao Chen, Yue Sun, Parvathinathan Venkitasubramaniam, and Sihong Xie. Wasserstein-regularized conformal prediction under general distribution shift. In *International Conference on Learning Representations*, 2025.