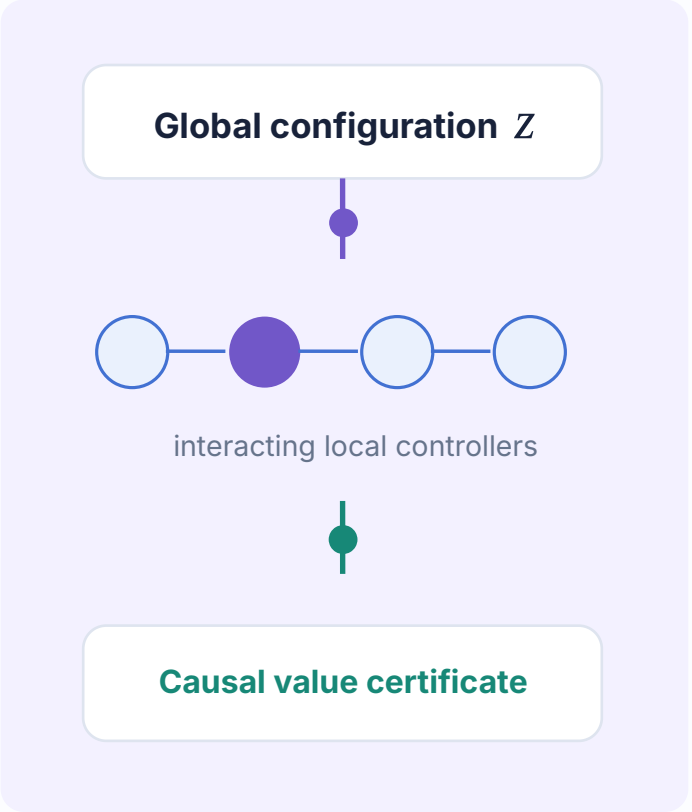
NEURIPS 2026 • MAIN TRACK POSTER

Online Causal Configuration for Networked Systems via Doubly Robust Steady-State Learning

Yuli Liu[1] · Zhiheng Zhang[2,3]

[1] University of Macau [2,3] Shanghai University of Finance and Economics



CAUSAL-UCB | Poster ID 149162

The measurement regime is not the deployment regime

Exploration provides information — but also changes the operating system.

LOGGING • $\eta > 0$

Explore to learn

Local actions and post-burn-in contexts are observed under a measurement policy.

$\neq$

DEPLOYMENT • $\eta = 0$

Optimize to deploy

The target is steady-state utility after local controllers adapt without measurement exploration.

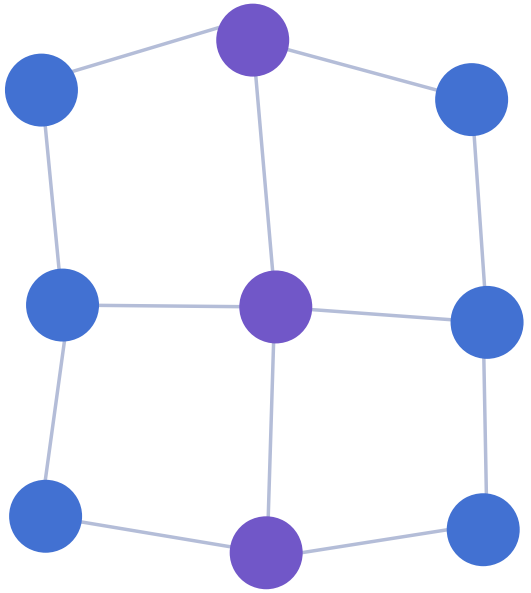
Identification obstruction

Identical exploratory logs can conceal different no-exploration values.

One global choice, many interfering local decisions

The operator selects a network configuration; the local system determines its eventual outcome.

Global Z_t



Interference through neighbors

Target value

$$W_\infty(z;\, 0)$$

- **Post-burn-in feedback**

  No immediate reward for the global decision.

- **Causal exposure**

  Local actions and neighboring actions jointly matter.

METHOD

CAUSAL-UCB: a closed loop from logs to decisions

The same target value and uncertainty budget guide evaluation, action choice, and reporting.

01

Deploy & observe

Choose global z; wait for burn-in; log local telemetry.

02

Model interference

Summarize neighborhood exposure and measurement policy.

03

Correct the value

Clipped doubly robust estimation of $W_\infty(z; 0)$.

04

Act optimistically

Select the configuration with highest causal UCB.

Value estimate + predecision radius → certified optimistic choice → next round

A causal value certificate, not just a point estimate

The radius separates stochastic fluctuation from identifiable, transport, and temporal errors.

Doubly robust score

$$\hat{\psi} = \hat{\mu}(C) + \hat{w} \cdot (R - \hat{m}(A, E, C))$$

Theorem 1 · high-probability deviation

$$|W(\hat{z}; 0) - W_\infty(z; 0)| \leq b_t(z)$$

Valid under calibrated assumptions; diagnostic fits alone are not certificates.

What contributes to $b_t(z)$?

- Sampling variation
- Action & exposure mismatch
- Context transport & clipping
- Temporal dependence / drift

Why better identification can cost more during deployment

The guarantees clarify the tension instead of pretending exploration is free.

THEOREM 2 • WORST-CASE FRONTIER

Inference error

$$\Omega\left((1+\sum n_t)^{-1}\Omega^2\right)$$

More measurement can improve identification.

THEOREM 2 • MEASUREMENT COST

Deployment loss

$$\Omega\left(\sum n_t\right)$$

But the same measurement perturbs deployment.

Theorem 3 • Conditional regret guarantee

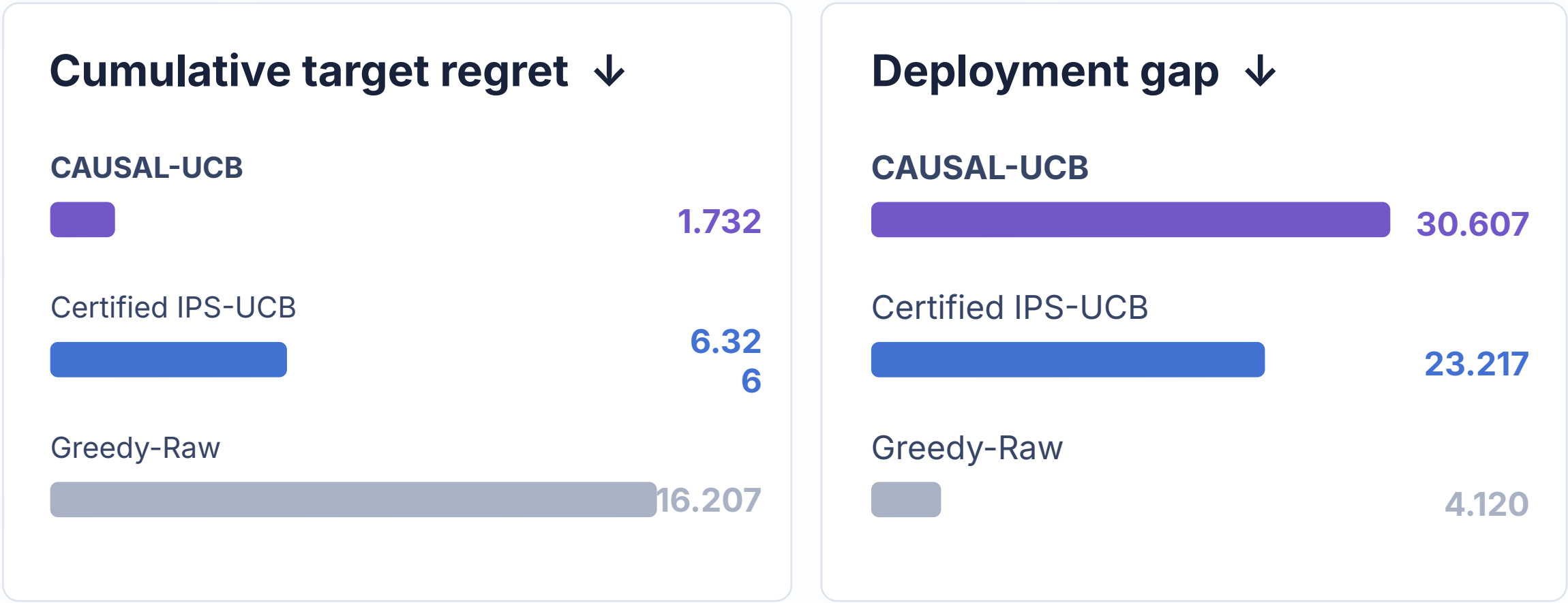
Valid certificates + selected-information growth → controlled online regret.

Scope: finite horizon, predetermined exploration schedule, stated stationary family.

CONTROLLED EVIDENCE

High mismatch: better target decisions, higher measurement cost

24 paired seeds · 200 rounds · 5 candidate configurations · source: Table 2



What transfers, and what the results do not claim

Trace-driven semi-synthetic replay uses real network traces with model-generated counterfactual outcomes.

METR-LA

0.754 vs 0.774 (IPS-UCB)

High-mismatch cumulative regret

FLOCKLAB

0.735 vs 1.612 (IPS-UCB)

High-mismatch cumulative regret

Three takeaways

1. **Optimize no-exploration value**
2. **Use calibrated causal certificates**
3. **Report gains alongside deployment costs**

Scope: benefits are regime-dependent; these trace replays are not live network deployments.