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

Yuli Liu[3] · **Zhiheng Zhang**[1,2*]

[1] School of Statistics and Data Science, Shanghai University of Finance and Economics
[2] Institute of Big Data Research, Shanghai University of Finance and Economics
[3] Department of Ocean Science and Technology , University of Macau, Macao SAR, China

Paper | Project Page | Code

CENTRAL MESSAGE Exploratory rewards may misrepresent no-exploration steady-state performance. CAUSAL-UCB addresses this mismatch through interference-aware doubly robust evaluation and calibrated optimistic selection.

Background: why the target is different

A network operator chooses a global configuration while interacting, black-box local controllers adapt and settle. The operator ultimately cares about the no-exploration steady-state utility—not the reward seen while deliberately perturbing those controllers to measure them.

The observational trap

Measurement-level exploration changes local actions and potentially the post-burn-in context distribution. Even abundant logged data can misrank deployment configurations when the logging and target regimes differ.

EXPLORATORY LOGGING
$\eta > 0$: perturbed actions, shifted contexts, observable rewards

TARGET DEPLOYMENT
$\eta = 0$: unperturbed steady-state utility to optimize

Causal estimand and decision unit

$$W_\infty(z;0) = \mathbb{E}_{C^- \sim \nu_{z,0}}\left[\int_{\widetilde{\mathcal{A}}} m_z(\widetilde{a}, C^-)\, g_{z,0}(d\widetilde{a} \mid C^-)\right]$$

Here z is the global configuration; $\widetilde{A} = (A, E)$combines a local action with its neighborhood exposure. The optimization benchmark is $z^\star \in \arg\max_{z \in \mathcal{Z}} W_\infty(z;0)$, and target regret accumulates the loss from selecting configurations other than z*.

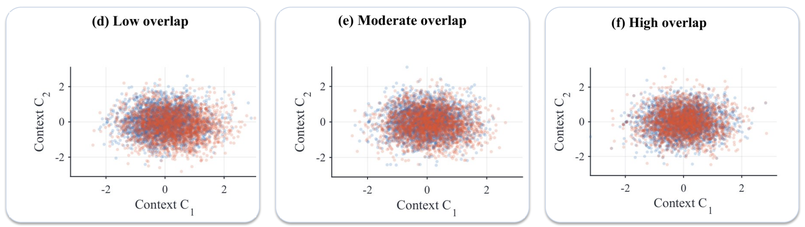
Two obstacles: interference and context shift

01 NETWORK INTERFERENCE
A focal-action propensity is insufficient when neighboring actions jointly affect outcomes. Use the action–exposure law.

02 CONTEXT TRANSPORT
Correcting logged actions does not identify the post-burn-in context law under zero exploration.

PROPOSITION 3 Identical exploratory logs can conceal different no-exploration target values.

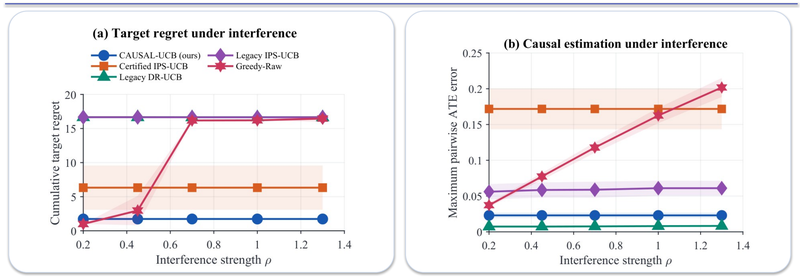
Visual evidence: better overlap is not context transport



Paper Fig. 5(d–f) · Blue: target, coral: logging. Better action support does not certify the target context distribution.

IDENTIFICATION CONDITIONS
Joint action–exposure overlap, conditional identification, and independently calibrated exposure, transport, nuisance, clipping, and temporal-dependence budgets are required.

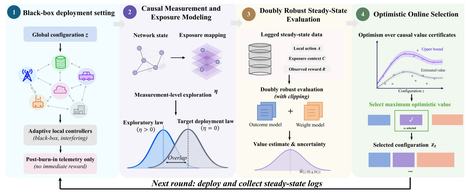
How interference affects decisions and causal estimation



Paper Fig. 4(a–b) · Five interference levels, 24 paired seeds per condition. Target regret and pairwise ATE error are distinct outcomes.

Method: the CAUSAL-UCB decision loop

Each slow round deploys one global configuration, waits for local controllers to settle, collects post-burn-in windows, and updates predictable nuisance models. The next configuration is chosen using an optimistic bound on its causal deployment value.



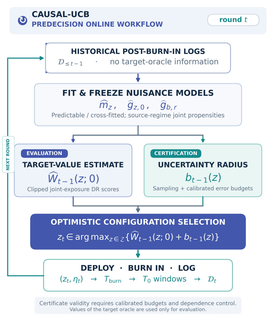


Fig. 1. CAUSAL-UCB framework and online workflow. Left: Interference-aware causal evaluation and optimistic selection. Right: Predecision value estimation and uncertainty certification from historical data.

Exposure-aware doubly robust evaluation

$$\widehat{\psi}_{z,r}(\widetilde{O}) = \widehat{\mu}_z(C^-) + \widehat{w}_{z,r}(\widetilde{A}, C^-)\left\{R - \widehat{m}_z(\widetilde{A}, C^-)\right\}$$

The direct term integrates the fitted response over the target joint action–exposure law. The residual correction uses clipped target/behavior density ratios from the actual source regime; historical scores are frozen rather than refit using future observations.

Why DR is not a complete solution
Orthogonal nuisance correction addresses estimation sensitivity, but cannot by itself identify the target context law or remove exposure approximation and cross-regime response differences.

A uniform predecision value certificate

$$\left|\widehat{W}_t(z;0) - W_\infty(z;0)\right| \le 2B_\psi\sqrt{\frac{\mathrm{comp}(\mathcal{Z}) + \log(2/\delta)}{n_{\mathrm{eff},t}(z)}} + \epsilon_{\mathrm{tot},t}(z) + \widehat{\kappa}_t(z)$$

Theorem 1 separates effective-sample fluctuation from nuisance products, target-law estimation, clipping, exposure/response transport, context shift, configuration mismatch, and residual temporal dependence. Plug-in ESS and exposure-enrichment checks are diagnostics: they count as certificate inputs only when separately calibrated.

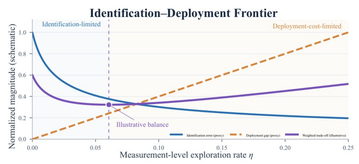
Optimistic selection and three guarantees

THEOREM 1 · VALUE CERTIFICATE
Joint-exposure DR evaluation with calibrated sampling, nuisance, transport, and dependence allowances.

THEOREM 2 · IDENTIFICATION FRONTIER
In a stated slow-mixing family, more measurement improves inference but also incurs deployment loss.

THEOREM 3 · ONLINE REGRET
Valid predecision certificates and selected-information growth control target regret.

PROPOSITION 1 · RESTRICTED CASE
Known local interference yields an anytime, executable radius shared by estimation, selection, and reporting.



Paper Fig. 2 · Illustrative identification–deployment frontier.
The diagram is schematic; it is not an empirical performance guarantee.

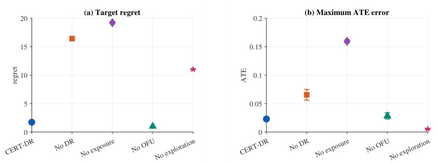
Independent certificate audit: what was verified

RESTRICTED CERTIFICATE AUDIT

300 / 300

Full-path target coverage and OFU path inequality

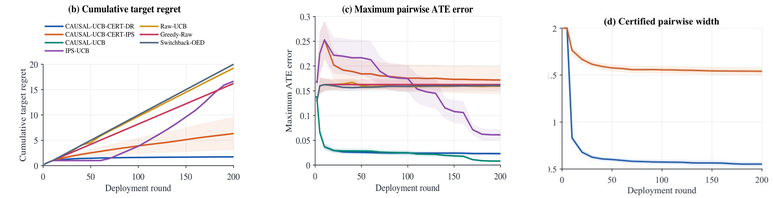
300 independent seeds × 160 rounds. Applies to the known-local-interference certificate, not unrestricted black-box networks.



Paper Fig. 6 · Mechanism ablations: removing DR or exposure changes regret and causal-estimation error.

Experimental protocol and decision quality

Controlled study: a 24-node ring, five global configurations, 200 rounds, and 24 paired random seeds. The audit uses 300 independent seeds; METR-LA and FlockLab supply topology/context traces for separate 180-round, eight-seed semi-synthetic replays.



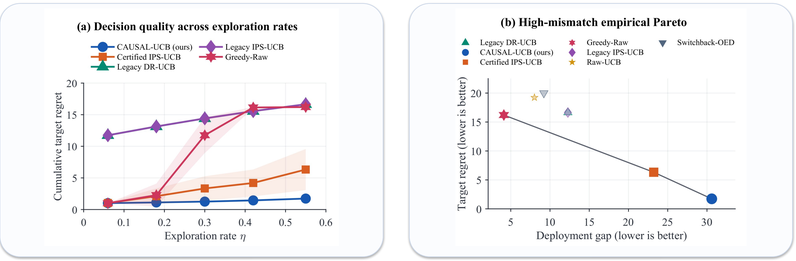
Paper Fig. 3(b–d): target regret, pairwise ATE error, and certified interval width across rounds.

High mismatch (200 rounds)	Target regret ↓	ATE error ↓	Deployment gap ↓
CAUSAL-UCB (ours)	**1.732**	**0.0230**	**30.607**
Certified IPS-UCB	6.326	0.1718	23.217
Legacy DR-UCB	16.647	0.0081	12.279
Greedy-Raw	16.207	0.1626	4.120

Main finding. CAUSAL-UCB attains the smallest high-mismatch target regret and zero terminal gap in this benchmark. Yet its deployment gap is larger than that of Certified IPS-UCB and Greedy-Raw: better target decisions are not free measurements.

Deployment cost: a distinct objective

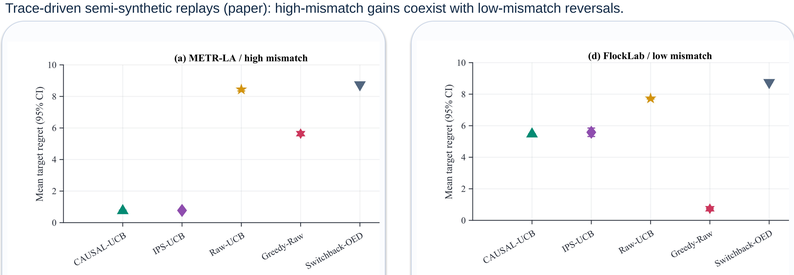
Better target decisions can require more measurement; Fig. 7 shows both sides of this trade-off.



Paper Fig. 7(a–b) · Curves and points describe a regime-specific trade-off, not a Pareto-optimality guarantee.

Transfer is regime-dependent

Trace-driven semi-synthetic replays (paper): high-mismatch gains coexist with low-mismatch reversals.



METR-LA: CAUSAL-UCB 0.754 vs IPS-UCB 0.774

FlockLab: Greedy-Raw 0.735 vs CAUSAL-UCB 5.472

SCOPE · Real network-derived topology/context traces, but model-generated counterfactual outcomes. These are not live deployments. Benefits depend on mismatch; deployment cost is reported separately.