Effect-Level Validation for Causal Discovery in Interactive Telemetry

Hoang Dang · Luan Pham · Minh Nguyen

Independent Researcher, Vietnam · RMIT University, Australia · Florida Atlantic University, USA

NeurIPS 2026 · Evaluations & Datasets Track

Which Graphs Give Reliable Answers?

The problem

Key question

Without a ground-truth graph, how do we know a discovered graph can be trusted for the decision it will feed?

Example decision

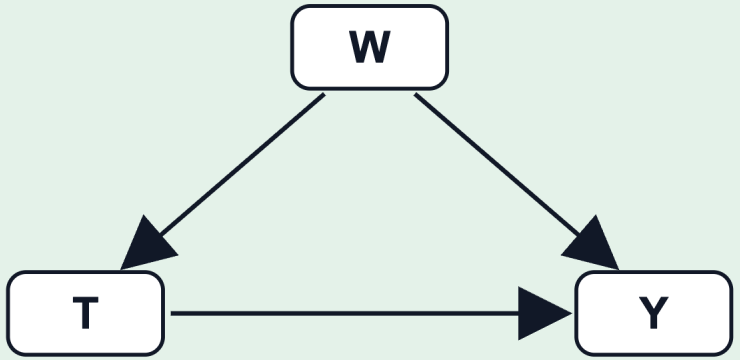
Does early player-versus-player (PvP) play make players come back the next day? If yes, invest in PvP onboarding.

Why accuracy scores fail

The true graph is unknown, so SHD, F1 and AUROC cannot be computed.

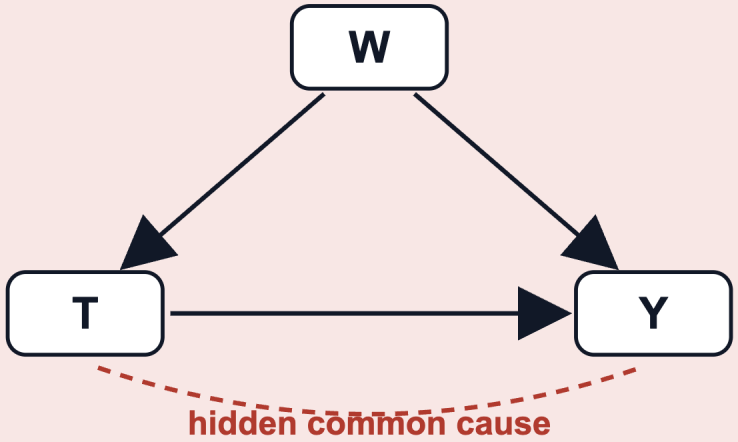
Same skeleton, opposite answers

A · effect identifiable



Adjust for W and the effect of T on Y is clean.

B · effect not identifiable



A hidden common cause cannot be adjusted away.

Both graphs score almost the same on graph-recovery metrics. Only one can answer the question.

A real game, and a tempting number

Telemetry from a deployed Web3 role-playing game: 1,280 paying players. Four example rows (0 = no, 1 = yes):

Web3 wallet	Sessions	Level-1 play time	PvE battles	Early PvP	Day-1 return
1	7	9.4	23	0	1
1	8	3.7	22	1	1
0	7	20.5	18	1	1
0	2	5.4	20	0	0

Day-1 retention

89% with early PvP

65% without

24 pp gap

Why the gap is misleading

PvP players already had more Web3 wallets (73% vs 42%), more PvE battles (18.9 vs 8.7) and more sessions (5.3 vs 3.5). The raw gap is heavily confounded.

1. Causal Discovery

Run several suitable algorithms: PC, GRaSP, BOSS, FCI…

↓

2. Domain Constraints

Apply time and system-knowledge constraints, level by level

↓

3. Admissibility Gate

Check the identifiability assumptions

✓ Pass ✗ Reject

↓

4. Causal Effect Estimation

Estimate the causal effect with suitable estimators

↓

5. Reliability Validation

Cross-algorithm stability, placebo, subsampling, E-value, …

How many graphs pass the gate?

C0 · minimal rules

10/21

PvP left with no parents, so the run returns the raw 24 pp gap.

C1 · + rule on Web3

0/21

Undirected edges remain, so the effect is not identified.

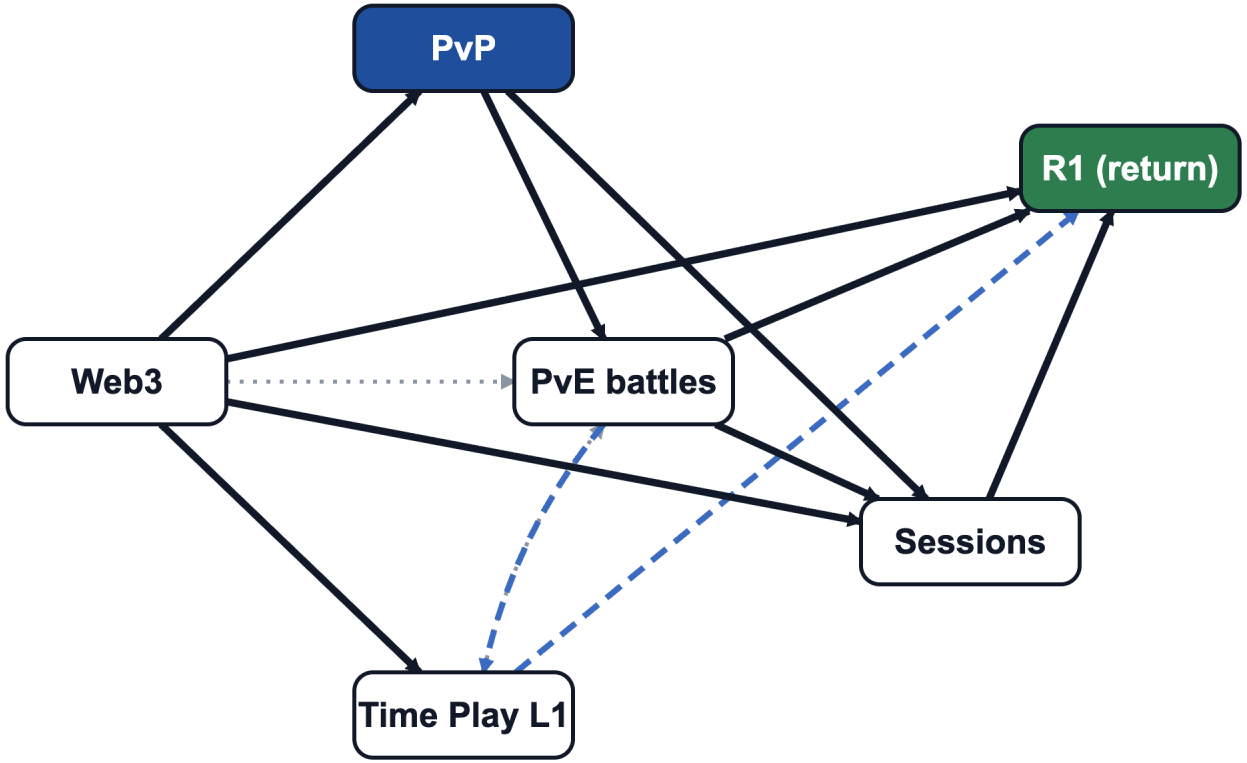
C2–C3 · full rule set

21/21

Every run is admitted and returns the same effect.

21 runs per level = 3 algorithms (PC, GRaSP, BOSS) × 7 significance levels.

Algorithms agree on the graph and on the effect



— 6/6 runs -- 4/6 ····· 2/6

Effect of PvP on R1 (ATE), 6 runs = 3 algorithms × 2 rule sets:

Algorithm	C2	C3
PC	0.1739	0.1739
GRaSP	0.1739	0.1739
BOSS	0.1739	0.1739

Total effect, not direct

No PvP → R1 edge. The 0.174 flows through Sessions and PvE battles, adjusting for Web3 only.

It holds only if the C2–C3 structure is right.

Contributions

1. **A 3-check gate**
   A graph is used only if it follows the rules, identifies the effect and compares like with like.

2. **Evidence that agreement is not proof**
   The effect drops from 13.9 to 4.3 pp when the adjustment set changes.

3. **A real-game case study**
   With a known-truth simulation: the checks cannot rule out hidden confounders.

Paper · Code · Data

PAPER CODE DATA