A DIAGNOSTIC RADIUS FOR KNOWLEDGE-INDUCED CAUSAL EFFECT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Observational causal discovery usually identifies an equivalence class of causal graphs rather than a unique graph. In applied settings, background or expert knowledge can further constrain uncertain edge orientations, after which sound orientation rules propagate their logical consequences. We give a query-specific certificate: the fewest closure orientations that must be contradicted before a fixed adjustment set can fail, conditional on a correct equivalence class, computed exactly by retractions without enumerating the space of knowledge states. On synthetic graphs whose generator spreads instances across radius levels by targeting the treatment's separation from the adjustment set, this breakdown radius tracks survival to corrupted knowledge at a rank correlation of 0.75 to 0.81 under three- and four-tier corruption, against at most 0.14 for structural Hamming distance to the true graph. When the query is instead drawn without targeting separation, the radius ranks survival at 0.29 and the structural proxy separation at 0.34, both well above SHD, with the radius still leading SHD by about 0.3. On published networks, with knowledge scenarios constructed by large language model elicitation, it ranks survival within fixed knowledge states, where knowledge counts are constant. We release code and benchmarks so that a breakdown radius can be reported alongside any causal estimate that relies on an adjustment set.

1 INTRODUCTION

From observational data, causal discovery recovers an equivalence class of graphs, represented by a *completed partially directed acyclic graph* (CPDAG). Some of its edges are directed; the others stay undirected because the data cannot decide them. In applied work, the remaining edges are oriented by a person. The analyst asserts orientations they believe to hold, Meek's rules (Meek, 1995) propagate those assertions to everything they entail, and the result is a maximally oriented partially directed acyclic graph (MPDAG) from which a covariate adjustment set Z is read off with the generalized adjustment criterion (Perković et al., 2018; Perković, 2020), often as the optimal set of Henckel et al. (2022), and the reported estimate is adjusted for it.

The assertions that produced Z are checked for consistency with the data and never against the process that generated it. Meek's algorithm either returns a valid MPDAG or reports a contradiction; it cannot ask whether an asserted orientation is the one that holds in the world, because the data are silent on those edges. Expert knowledge is also variable: two analysts rarely supply the same claims, and an elicited language model supplies yet another set. The data cannot flag a wrong claim on these edges, but they do fix the space of claims an analyst could make, and within that space the dependence of the conclusion on the claims actually made can be measured.

We measure it,[1] in the spirit of sensitivity analysis (Rosenbaum, 2002; Cinelli & Hazlett, 2020) and of the breakdown point of robust statistics. Rosenbaum's Γ states how much hidden bias a matched comparison tolerates; we state how many orientation errors, with the skeleton fixed by the data, an adjustment set tolerates. Fix the analyst's graph G_0 and the adjustment set Z read from it, and consider the knowledge states they could have reached by withholding a claim, reversing one, or adding one they had declined to make. Each consistent revision is an MPDAG for the same CPDAG, and each either admits Z or does not. Following Taeb et al. (2026), we order these states by

[1] Code and benchmarks: https://anonymous.4open.science/r/bkrobust/.

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.

the models they represent and measure distance by elementary revisions. The *breakdown radius* r_{val} is then the distance from G_0 to the nearest state in which Z is no longer a valid adjustment set: the fewest elementary revisions of the analyst's closure orientations that can invalidate the conclusion. It is computed from the analyst's own inputs, with no appeal to the data-generating graph.

The robustness of an adjustment set is a separate property from the accuracy of the knowledge behind it. Benchmarks of elicited knowledge score its agreement with a published graph (Zhou et al., 2026; Kitson & Constantinou, 2026), a score that needs the graph the analyst lacks. The radius scores the conclusion itself and, like a sensitivity parameter, is reported beside the estimate.

Contributions.

- We equip the space of knowledge states that a CPDAG admits with a model-oriented metric and define the breakdown radius of a committed adjustment set on it (Section 4).
- We prove that failure is upward-closed on a join-semilattice whose closure operator has the anti-exchange property, so the nearest failure is reached by retracting orientations, and that distance on this space is an orientation Hamming distance. Infinite robustness is decided by a single adjustment-criterion call on $\widehat{\mathcal{C}}$, and a finite radius is found exactly by retractions, with a SAT certificate when the failure is not nearby (Section 5).
- On published networks with knowledge elicited from language models, r_{val} ranks survival within a fixed knowledge state, where the number of claims is constant (Section 6.3). On synthetic graphs it ranks survival far better than the structural Hamming distance (SHD) to the data-generating graph, and keeps ranking it when separation, a structural stand-in that is often undefined on published networks, is held fixed (Section 6.2).

Section 7 carries the same geometry to the effect scale. Under a linear-Gaussian model, inside the region r_{val} certifies, the reported effect carries no identification bias, and a mean-bias profile illustrates how analyses sharing a radius can differ.

2 Related work

CPDAGs with background knowledge. Meek (1995) gave the orientation rules and the consistency test for asserted orientations, and Perković et al. (2017); Perković (2020); Perković et al. (2018) characterise the resulting MPDAGs, effect identification in them and their valid adjustment sets. MPDAGs also arise from tiered or interventional knowledge (Scheines et al., 1998; Hauser & Bühlmann, 2012), and discovered graphs are marginalised after discovery (Bang et al., 2026). Methods that make discovery robust to unreliable knowledge (Hyttinen et al., 2014; Wang et al., 2026; Long et al., 2023; Li et al., 2024; Ejaz & Bareinboim, 2025) change what discovery returns, whereas the radius takes the discovered graph as given (Appendix A). Taeb et al. (2026) already define covering-graph distances, including for MPDAGs, and apply metric neighbourhoods to the robustness of covariate adjustment under graph misspecification, the closest predecessor of this work. We apply their model-oriented distance construction to the poset of orientation-knowledge states compatible with a fixed CPDAG, ordered as sets of compatible causal DAG hypotheses rather than observational distributions. Unlike their unrestricted MPDAG poset, which is not graded, this one is graded; this gives the distance in closed form and an exact computation of the distance to a query's failure boundary (Section 5). Padh et al. (2025) bound a causal query over a collection of plausible graphs, a broader graph-uncertainty optimisation where we instead give an exact orientation-error-budget certificate for one fixed set; Henckel et al. (2024) give an adjustment-based distance between two graphs for evaluating a learned graph, not the distance from a knowledge state to the failure of a fixed adjustment set.

Knowledge elicited from language models. Benchmarks score the orientations language models give against published graphs (Zhou et al., 2026; Kitson & Constantinou, 2026; Kiciman et al., 2024; Vashishtha et al., 2025); prior-data fitted networks estimate effects in a forward pass, one given a partial graph (Balazadeh Meresht et al., 2025; Robertson et al., 2025; Reuter et al., 2026).

Robustness of adjustment and sensitivity analysis. Existing work tests disagreement between valid adjustment sets (Su & Henckel, 2022), selects sets under graph uncertainty (Hu & van der

Pas, 2025), or enumerates the effects an equivalence class admits (Maathuis et al., 2009; Guo & Perković, 2021), all for fixed knowledge; the radius varies it. Sensitivity analysis asks how strong an unmeasured confounder must be to overturn a conclusion (Rosenbaum, 2002; Cinelli & Hazlett, 2020; Javurek et al., 2026) ; the radius asks it of background knowledge.

3 PRELIMINARIES

We follow Pearl (2009) and Spirtes et al. (2000); the notation is collected in App. Table 3. A causal directed acyclic graph (DAG) over $\mathcal{V} = \{X_1, \ldots, X_d\}$ encodes direct causal influences, with parents, ancestors and descendants pa, an, de. Under the Markov and faithfulness assumptions, DAGs encoding the same conditional independencies form a Markov equivalence class, represented by a CPDAG sharing their skeleton and v-structures (Andersson et al., 1997). Background knowledge that constrains further orientations refines the class to an MPDAG (Meek, 1995; Perković et al., 2017). We write $[G]$ for the DAGs an MPDAG G represents and $\mathrm{dir}(G)$ for its directed edges.

Meek closure and closure orientations. Meek's rules R1–R4 orient an undirected edge whenever the alternative would create a new v-structure or a directed cycle. We write $\mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K})$ for the MPDAG obtained by imposing orientations $\mathcal{K}$ on a CPDAG $\widehat{\mathcal{C}}$ and closing under R1–R4, and call $\mathcal{K}$ *consistent* when this does not fail. Closure propagates consequences, so the MPDAG generally carries more orientations than were asserted. We call

$$\mathcal{K}_G \;:=\; \mathrm{dir}(G) \setminus \mathrm{dir}(\widehat{\mathcal{C}}) \tag{1}$$

the *closure orientations* of G, so that $\mathcal{K} \subseteq \mathcal{K}_{\mathrm{Meek}(\widehat{\mathcal{C}},\mathcal{K})}$.

Adjustment. For the effect of X on Y, a set Z is a valid adjustment set in an MPDAG G when it satisfies the generalized adjustment criterion (Perković et al., 2018; Perković, 2020), equivalently when it is valid in every DAG of $[G]$; the effect is identified by adjustment in G if and only if such a set exists. Among valid sets we use the optimal set $\mathbf{O}$, whose efficiency interpretation depends on the estimation regime (Henckel et al., 2022; Rotnitzky & Smucler, 2020). Because validity is universal over $[G]$, withdrawing an orientation enlarges $[G]$ and can only make validity harder to satisfy.

Standing assumptions. We assume causal sufficiency and correct conditional independence information, so the skeleton and v-structures of $\widehat{\mathcal{C}}$ are given and are not themselves perturbed. Background knowledge is orientation knowledge about the edges the data left undirected.

4 THE GEOMETRY OF KNOWLEDGE STATES AND THE BREAKDOWN RADIUS

An analyst holds an estimated CPDAG $\widehat{\mathcal{C}}$, asserts orientations $\mathcal{K}$ for edges the data left undirected, obtains $G_0 = \mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K})$, chooses a valid adjustment set Z for a query (X, Y) in G_0 (in what follows the optimal adjustment set), and holds Z fixed for estimation. This section builds the space in which G_0 sits and measures how far Z is from failing.

A running example. Consider the effect of statin therapy on cardiovascular events, with eight variables `Age`, `Smoke`, `BMI`, `Chol` (cholesterol), `CRP` (an inflammation marker), `Statin` (treatment X), `BP` and `CVD` (outcome Y). The CPDAG (Figure 1a) has six compelled edges and five undirected ones forming a single component on $\{\texttt{Smoke}, \texttt{BMI}, \texttt{Chol}, \texttt{CRP}, \texttt{Statin}\}$, so the data do not say whether inflammation precedes or follows treatment. The analyst asserts $\texttt{Smoke} \to \texttt{BMI}$, $\texttt{BMI} \to \texttt{Chol}$, $\texttt{Chol} \to \texttt{CRP}$ and $\texttt{Statin} \to \texttt{CRP}$ (statins lower CRP). Meek closure orients all five edges, adding $\texttt{Chol} \to \texttt{Statin}$ on the analyst's behalf (Figure 1b), so $|\mathcal{K}| = 4$ while $|\mathcal{K}_{G_0}| = 5$. The optimal adjustment set is $Z = \{\texttt{Age}, \texttt{Smoke}\}$.

Elementary moves. The skeleton and v-structures are fixed by the data, so knowledge can change only the orientation of undirected edges; an assertion that contradicts a compelled edge is inconsistent with the data, and Meek's algorithm already rejects it. Two *atomic moves* remain: orient an undirected edge, or retract (un-orient) a closure orientation. A reversal costs two moves. Moves act on the closure $\mathcal{K}_{G_0}$: an orientation that R1–R4 entailed is as revisable as one the analyst stated.

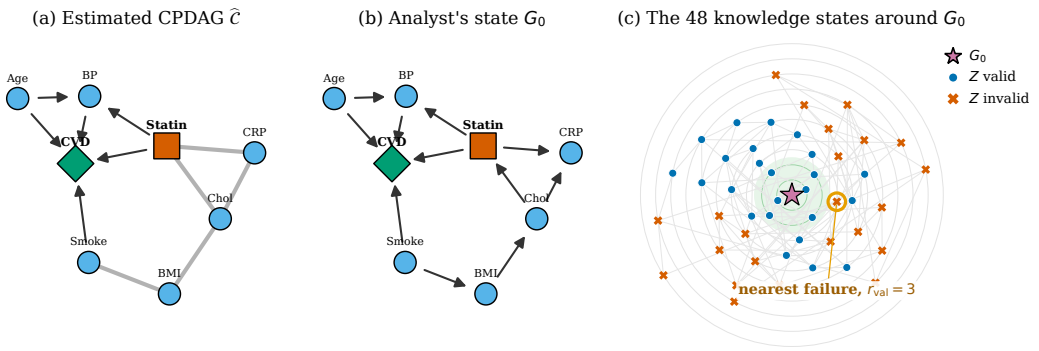

Figure 1: The committed set survives every knowledge state within two revisions of G_0; the nearest failure lies at $r_{\text{val}} = 3$. (a) Estimated CPDAG; grey edges are undirected. (b) G_0 after Meek closure of four asserted orientations. (c) The 48 knowledge states in shells by distance from G_0; edges are covering pairs, and the shaded region, where $Z = \{\texttt{Age}, \texttt{Smoke}\}$ is valid throughout, is the certified region.

Definition 1 (Knowledge states)**.** The *space of knowledge states* induced by $\widehat{\mathcal{C}}$ is $\mathfrak{G}_{\widehat{\mathcal{C}}} = \{\,\mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K}) : \mathcal{K} \text{ consistent with } \widehat{\mathcal{C}}\,\}$; equivalently, $G \in \mathfrak{G}_{\widehat{\mathcal{C}}}$ if and only if $\mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K}_G) = G$.

The fixpoint test enforces $[G] \subseteq [\widehat{\mathcal{C}}]$. Weaker tests, that G is an MPDAG in its own right or that its undirected part is chordal, admit graphs outside the class (Appendix C). Following Taeb et al. (2026), we order states by the models they represent,

$$G_1 \preceq G_2 \iff [G_1] \subseteq [G_2], \tag{2}$$

so more knowledge means a lower element, with $\widehat{\mathcal{C}}$ at the top and the DAGs of the class at the bottom. We write $G_1 \lessdot G_2$ when G_2 covers G_1; covering pairs form the *neighbour graph*.

Definition 2 (Distance)**.** $d(G, H)$ is the length of a shortest path between G and H in the neighbour graph: the least number of elementary revisions carrying one knowledge state to the other.

Every state is joined to $\widehat{\mathcal{C}}$ by a chain of covering pairs, so the neighbour graph is connected and d is a metric on $\mathfrak{G}_{\widehat{\mathcal{C}}}$ (Corollary 1 gives it in closed form). The running example has 48 states and 96 covering pairs.

Definition 3 (Breakdown radius)**.** Let $G_0 = \mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K})$ be the analyst's knowledge state, and let $Z \subseteq \mathcal{V} \setminus \{X, Y\}$ be a valid adjustment set for (X, Y) in G_0; in this paper we take Z to be the optimal adjustment set of G_0. Holding Z fixed, its breakdown radius is

$$r_{\text{val}} = \min\big\{\, d(G_0, G) \;:\; G \in \mathfrak{G}_{\widehat{\mathcal{C}}},\ Z \text{ is not a valid adjustment set for } (X, Y) \text{ in } G \,\big\}, \tag{3}$$

with $r_{\text{val}} = \infty$ when no state invalidates Z. Since Z is valid in G_0, $r_{\text{val}} \geq 1$.

The claims the radius counts are the orientations of the closure $\mathcal{K}_{G_0}$, whether the analyst stated them or Meek's rules entailed them from what was stated.[2]

Remark 1 (Reading the radius). $r_{\text{val}} = k$ asserts that *some* revision of k closure orientations invalidates Z and that no revision of fewer does. The states within distance $r_{\text{val}} - 1$ are certified, so $r_{\text{val}} - 1$ revisions are safe and a radius of one means that no revision is guaranteed safe. The radius locates the nearest failure; it does not say which orientation the analyst got wrong, and it counts closure orientations, not elicited claims. Nor does failure of the committed Z mean failure of identification: another adjustment set may remain valid at that state.

In the running example $r_{\text{val}} = 3$ (Figure 1c). A nearest failing state retracts $\texttt{BMI} \to \texttt{Chol}$, $\texttt{Smoke} \to \texttt{BMI}$ and $\texttt{Chol} \to \texttt{Statin}$, the last supplied by closure rather than asserted: without all three, the data cannot rule out that $\texttt{Statin}$ is an ancestor of the confounder block, so $\texttt{Smoke}$ may be downstream of treatment and adjusting for it invalid. Withdrawing any one or two of them leaves Z valid.

[2] Elicited claims, $|\mathcal{K}| \leq |\mathcal{K}_{G_0}|$, are a different unit (Appendix H.2).

Algorithm 1 Breakdown radius $r_{\mathrm{val}}(\widehat{\mathcal{C}}, \mathcal{K}, X, Y, Z)$

$G_0 \leftarrow \mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K})$; if Z invalid in G_0, **return** INVALIDINPUT ▷ status, not a radius
if Z valid in $\widehat{\mathcal{C}}$, **return** ∞ ▷ Corollary 1(b)
$\mathcal{F} \leftarrow \{\mathcal{K}_{G_0}\}$ ▷ frontier of closed orientation sets
for $k = 1, 2, \ldots, b$ **do** ▷ upward search, depth budget b
  $\mathcal{F} \leftarrow \{S \setminus \{e\} : S \in \mathcal{F},\ e \in S,\ \mathrm{cl}(S \setminus \{e\}) = S \setminus \{e\}\}$ ▷ covers (Theorem 1)
  if some $S \in \mathcal{F}$ has Z invalid in $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$ **then return** k ▷ with $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$ as witness
for $k = b+1, \ldots, |\mathcal{K}_{G_0}| - 1$ **do** ▷ SAT certificate
  if SAT(closed $S \subseteq \mathcal{K}_{G_0}$, $|\mathcal{K}_{G_0} \setminus S| \le k$, Z invalid in $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$) **then return** k
return $|\mathcal{K}_{G_0}|$ ▷ retracting to $\widehat{\mathcal{C}}$, which fails

5 EXACT COMPUTATION BY RETRACTION SEARCH

Evaluating Definition 3 by enumerating $\mathfrak{G}_{\widehat{\mathcal{C}}}$ is feasible only for small examples: the space grows exponentially in the number of undirected edges, and building its covering relation is quadratic in its size. We show that the search can be confined to states obtained from G_0 by *withdrawing* orientations. All proofs are in Appendices C.4 and C.5, with a map of how the results depend on one another in Appendix C.2.

Lemma 1 (Failure is upward-closed)**.** *If Z is invalid in G and $G \preceq G'$, then Z is invalid in G'.*

Lemma 2 (Join-semilattice)**.** *For $G, H \in \mathfrak{G}_{\widehat{\mathcal{C}}}$, $G \vee H = \mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K}_G \cap \mathcal{K}_H)$ is the least upper bound of G and H, so $\mathfrak{G}_{\widehat{\mathcal{C}}}$ is a join-semilattice with maximum $\widehat{\mathcal{C}}$.*

Write $\mathrm{cl}(S) = \mathcal{K}_{\mathrm{Meek}(\widehat{\mathcal{C}}, S)}$ for the closure operator that Meek's rules induce on sets of orientations. The order is well behaved because this operator never lets two new orientations force each other.

Proposition 1 (Anti-exchange)**.** *Fix $H \in \mathfrak{G}_{\widehat{\mathcal{C}}}$. On the ground set $\mathcal{K}_H$, $S \mapsto \mathrm{cl}(S) = \mathcal{K}_{\mathrm{Meek}(\widehat{\mathcal{C}}, S)}$ for $S \subseteq \mathcal{K}_H$ is an anti-exchange closure operator: for closed $S \subseteq \mathcal{K}_H$ and distinct $x, y \in \mathcal{K}_H \setminus S$, if $y \in \mathrm{cl}(S \cup \{x\})$ then $x \notin \mathrm{cl}(S \cup \{y\})$.*

What the search uses is its consequence: every covering step adds or removes exactly one orientation, so the order is graded by the corank $|\mathcal{K}_G|$ (decreasing upward); the rank is $\rho(G) = m - |\mathcal{K}_G|$, for m the number of undirected edges of $\widehat{\mathcal{C}}$.

Theorem 1 (Monotone reachability)**.** *If $r_{\mathrm{val}} < \infty$, a nearest failing state is reachable from G_0 by a path that only retracts orientations. Hence r_{val} equals the least number of retractions from G_0 that invalidates Z.*

The reason is that the join of G_0 with a nearest failing state also fails (Lemma 1) and is no farther from G_0, and gradedness reaches it by single retractions.

Corollary 1 (Distance and the radius, in closed form)**.** *For $G, H \in \mathfrak{G}_{\widehat{\mathcal{C}}}$: (a) $d(G, H) = |\mathcal{K}_G \triangle \mathcal{K}_H|$, the number of closure orientations that one of the two states holds and the other does not ($\triangle$ is the symmetric difference), so orienting or retracting one edge costs one and reversing one costs two. (b) $r_{\mathrm{val}} = \infty$ if and only if Z is a valid adjustment set in $\widehat{\mathcal{C}}$; otherwise $r_{\mathrm{val}} \le |\mathcal{K}_{G_0}|$. (c) $r_{\mathrm{val}} = \min\{\, |\mathcal{K}_{G_0} \setminus \mathrm{dir}(D)| : D \in [\widehat{\mathcal{C}}],\ Z$ invalid for (X, Y) in $D \,\}$.*

Part (c) reads as an orientation-error budget: if the true DAG lies in $[\widehat{\mathcal{C}}]$ and contradicts fewer than r_{val} closure orientations of G_0, Z is a valid adjustment set in it.

Theorem 1 uses only that failure is upward-closed, so the same search also finds the nearest state whose worst-case bias exceeds a tolerance (Section 7).

Lemma 3 (Order identification)**.** *For $G, H \in \mathfrak{G}_{\widehat{\mathcal{C}}}$, $[G] \subseteq [H]$ if and only if $\mathrm{dir}(H) \subseteq \mathrm{dir}(G)$.*

By Lemma 3, covers are generated by set arithmetic on $\mathcal{K}_G$, and validity in an MPDAG is decided in polynomial time by the adjustment criterion, without listing $[G]$.

The algorithm. A single adjustment-criterion call on $\widehat{\mathcal{C}}$ settles whether r_{val} is finite (Corollary 1(b)). Otherwise Algorithm 1 runs a breadth-first search over retractions and stops at the first shell that

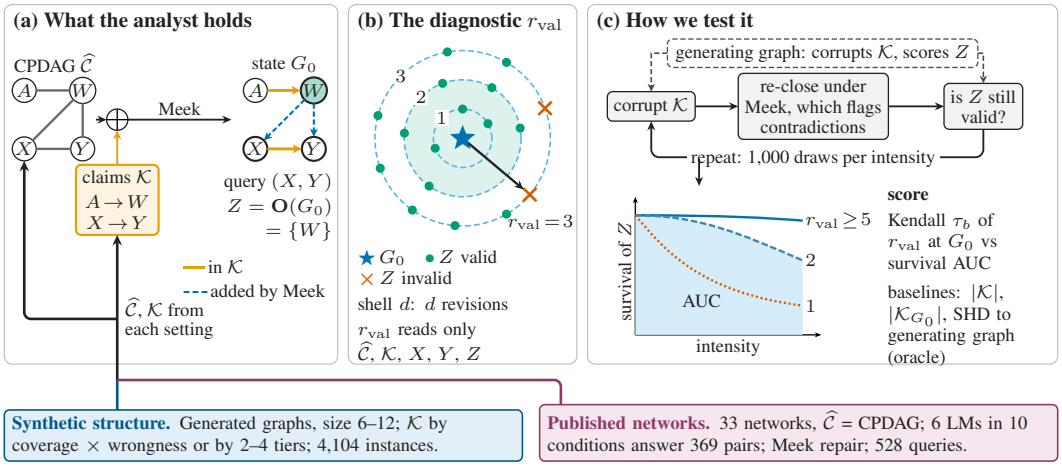

Figure 2: One protocol for both experiments. (a) What the analyst holds. (b) The radius, which reads only $\widehat{\mathcal{C}}$, $\mathcal{K}$, X, Y and Z; the shells are schematic, not the space of the toy in (a). (c) Each corrupted draw is re-closed and Z scored there; the survival AUC is ranked against r_{val} and the baselines by Kendall's τ_b.

contains a failure. Beyond a small depth budget it asks a SAT solver whether a failure lies within k retractions; an unsatisfiable answer certifies that none does (Appendix D.1). Both stages are exact, so the budget affects running time only. The search avoids building the whole space, though its worst case remains exponential; the timings of Appendix D come from an earlier back-door implementation, not the final one.

6 The radius ranks survival under corrupted knowledge

We test whether a larger radius means a set survives wrong knowledge longer, in two settings sharing one protocol (Figure 2): synthetic structure, where every corruption is known (Section 6.2), and published structures with knowledge elicited from language models (Section 6.3). The radius reads only $(\widehat{\mathcal{C}}, \mathcal{K}, X, Y, Z)$; the data-generating DAG only builds the CPDAG and corrupts knowledge.

6.1 Protocol

Survival. We commit Z at G_0 and corrupt the claims at a grid of intensities, 1,000 draws per point, re-closing each draw under Meek's rules. Survival is scored in that corrupted, re-closed state, with the same adjustment criterion as the radius: Z survives a draw when it is still a valid adjustment set there, and the survival AUC is the surviving fraction, averaged over a ten-point intensity grid. Survival measures whether the adjustment guarantee holds, conditional on the corruption being consistent – not the probability that Z is valid in the data-generating DAG. Draws with no consistent MPDAG are set aside, as r_{val} is defined on consistent states. Two error classes are simulated:

- *flip* errors reverse claims independently at random;
- *tiered* errors misplace variables within an ordering the analyst holds, so errors are block-correlated, as when an analyst imports a wrong causal story.

Predictors. Each predictor, taken at the uncorrupted state, is scored by Kendall's τ_b with the survival AUC. Two predictors are the analyst's counts $|\mathcal{K}|$ and $k_{G_0} = |\mathcal{K}_{G_0}|$; two grade accuracy, as benchmarks of elicited knowledge do (Zhou et al., 2026; Kitson & Constantinou, 2026): the SHD from G_0 to the data-generating graph (raw sign; survival tracks the guarantee, not correctness, so accuracy need not predict it) and the claim accuracy k_{acc}, and neither is available to an analyst. The fifth is a structural stand-in for the radius that an analyst could read off the graph without any search: the *separation* s, the length of a shortest path from X to a member of Z inside X's undirected component of $\widehat{\mathcal{C}}$. Few orientations lie between X and Z when s is small, so small s suggests a nearby

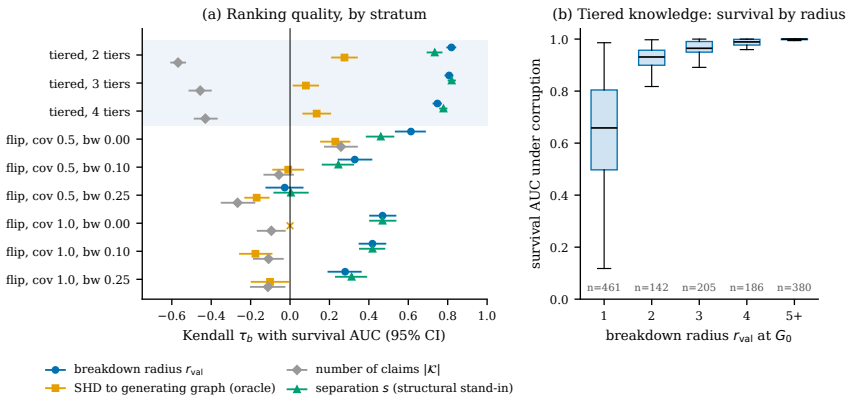

Figure 3: The radius ranks survival at τ_b of 0.75–0.81 at three and four tiers, against at most 0.14 for SHD to the data-generating graph; the two-tier value reflects corruptions that change nothing (Section 6.2). (a) Kendall τ_b between each predictor and survival AUC, per stratum, 95% instance-bootstrap intervals (2,000 resamples); tiered strata shaded, cov is coverage, bw base wrongness; the cross marks a stratum where SHD is constant (G_0 is the data-generating DAG). (b) Survival AUC under tiered corruption by the radius at G_0 (1,374 instances).

failure; but s certifies nothing, names no orientation, and is undefined when no member of Z lies in that component. It tests whether the exact search buys anything over reading the graph. Head-to-head comparisons use a paired bootstrap of the difference in τ_b, and App. Table 9 gives every setting of both experiments.

6.2 Verification on synthetic structure

Among admissible queries drawn without targeting separation, 66% have $r_{\text{val}} = 1$, which leaves little to rank, so the generator places the query at the minimum, median or maximum achievable separation to spread instances across radius levels. It spans 4,104 instances (32.8M draws; Table 9).

Tiered knowledge. At three and four tiers r_{val} ranks survival at $\tau_b = 0.81$ and 0.75 (480 instances each, intervals within ± 0.02), against 0.08 and 0.14 for SHD (Figure 3a), and still at 0.76 and 0.75 on corruptions that change the closure. At two tiers the whole-sample 0.82 (SHD 0.28) falls to -0.03 on such corruptions: there it measures how often a relocation changes nothing (Appendix H). The claim count ranks survival negatively (-0.57 to -0.43; App. Table 10), and median survival AUC rises from 0.66 at $r_{\text{val}} = 1$ to 1.00 at $r_{\text{val}} \geq 5$ (Figure 3b).

Independent flips. Flips are harder: 84% of draws are contradictions Meek closure flags. r_{val} still ranks survival positively in five of six flip strata, τ_b from 0.28 to 0.61 (Figure 3a), leading SHD under either sign wherever it varies, and the claim count by $+0.24$ to $+0.56$ (App. Table 10). The exception is half coverage with 25% base wrongness, where knowledge is sparse and already wrong and the negated single-claim fragility $-\varphi_1$ beats r_{val} by 0.37 (App. Table 10).

The radius against its structural stand-in. Separation ranks survival at least as well as the radius: slightly but significantly better at three and four tiers on this generator (Figure 3a, App. Table 10) and when the query is drawn uniformly (480 instances, Appendix H.1: pooled $\tau_b = 0.34$ vs. 0.29, both leading SHD), so where s is defined reading the graph is a good heuristic. The radius is exact ($r_{\text{val}} \leq s$ on every tiered instance of the main sweep), names orientations whose retraction breaks Z, and ranks survival at fixed s (0.49 tiered, 0.38 flips) while s barely does at fixed r_{val} (Appendix H).

6.3 Validation on published networks with elicited knowledge

Elicitation and release. On published Bayesian networks and DAGs from applied studies, one fixed questionnaire asks, for each chain component of at most twenty nodes, for a causal order, read as an orientation or a decline on each of 369 undirected pairs. Six language models answer

Table 1: On published networks the radius ranks survival, including inside a fixed knowledge state. Kendall τ_b [95% CI] with the survival AUC, network-cluster bootstrap, on the original cohort restricted to nonempty committed adjustment sets (see text). Pooled: eight real-name conditions (2,422 units, 24 networks); balanced: 31 queries scorable in all eight (248 units, 7 networks); within state: inside each network and condition (1,973 units, 92 states, 18 networks). Separation is scored only where defined (1,389, 96 and 1,213 units); there, within state, r_{val} scores 0.68 and s 0.49. Bold: interval excludes zero.

predictor	pooled	balanced	within state
r_{val}	**+0.319 [0.122, 0.485]**	**+0.499 [0.240, 0.624]**	**+0.419 [0.053, 0.679]**
k_{G_0}	**+0.214 [0.079, 0.339]**	+0.380 [−0.074, 0.562]	constant
$\lvert\mathcal{K}\rvert$	**+0.185 [0.071, 0.327]**	+0.333 [−0.125, 0.484]	constant
SHD, generating graph (oracle)	+0.054 [−0.077, 0.200]	+0.285 [−0.167, 0.456]	constant
separation s, where defined	+0.284 [−0.029, 0.492]	**+0.387 [0.221, 0.511]**	+0.491 [−0.135, 0.598]
claim accuracy k_{acc} (oracle)	+0.001 [−0.132, 0.135]	–	constant

it under ten conditions, two with scrambled names (App. Table 14), with Qwen2.5 7B as primary. Inconsistent answers are repaired by dropping the least confident orientations; states, questionnaire and answers are released with a datasheet (Appendix I). Survival is scored as on synthetic structure. The scrambled-name control of Kitson & Constantinou (2026) detects no memorisation at 7B or 72B, and the elicited orientations are not significantly more accurate than matched random ones, which the diagnostic does not require (App. Table 15).

Ranking. Within a single knowledge state, $|\mathcal{K}|$, k_{G_0}, SHD and k_{acc} are constant; the radius still ranks survival there (Table 1, within state; permutation $p < 0.001$). Its interval stays above zero with any single network removed, and scrambled names give a similar value, so we find no evidence that the ranking rests on recall of the networks (Appendix H.3). On the balanced panel, which fixes the query set across conditions, the counts' intervals include zero, though paired differences are not significant (Appendix H). Pooled over the eight real-name conditions the radius beats SHD to the generating graph by $+0.27$ $[0.04, 0.46]$; its lead over the counts is not significant. A pipeline error excluded 94 eligible records with valid empty sets; including them gives pooled $\tau_b = 0.32$ $[0.13, 0.49]$, within-state 0.43 $[0.07, 0.69]$ (Appendix H). Separation is the only other predictor that varies within a state, and it is undefined on 38% of the within-state units; where defined, the radius's paired lead over it is $+0.19$ $[0.07, 0.90]$ (15 network clusters).

7 From breakdown to magnitude: the radius on the effect scale

How far can an estimate move once the committed set is no longer guaranteed? Of the first five queries on each of 32 published networks, 49 commit a valid set with finite radius and non-zero bias in an enumerable component, and 44 have $r_{\mathrm{val}} = 1$ (Appendix E). Sets sharing a radius can still differ in how far; this section, the only one needing a parametric model, separates them.

Assumption 1 (Linear-Gaussian model)**.** The observational law is generated by a linear-Gaussian structural equation model whose DAG lies in $[\widehat{\mathcal{C}}]$, with population covariance $\Sigma \succ 0$.

Write $\beta(S;\Sigma)$ for the coefficient on X in the population regression of Y on $\{X\} \cup S$. The analyst reports $\theta_Z = \beta(Z;\Sigma)$, one number whichever DAG generated the data. Each candidate DAG D assigns the effect $\tau_D = 0$ if $Y \notin \mathrm{de}_D(X)$ and $\tau_D = \beta(\mathrm{pa}_D(X);\Sigma)$ otherwise, as in IDA (Maathuis et al., 2009), so a state G leaves the finite set $T(G) = \{\tau_D : D \in [G]\}$ open, and the worst-case bias it allows is

$$B(G) \;=\; \max\{\,|\theta_Z - \tau| \;:\; \tau \in T(G)\,\}. \tag{4}$$

Theorem 2 (Zero bias on the certified region)**.** *Under Assumption 1, every state within distance $r_{\mathrm{val}} - 1$ of G_0 has $B(G) = 0$: inside the region the radius certifies, the reported effect carries no identification bias. If moreover the true DAG lies in $[\widehat{\mathcal{C}}]$ and contradicts fewer than r_{val} closure orientations of G_0 (Cor. 1(c)), θ_Z equals the true effect; this is a population identity and does not by itself validate confidence intervals after data-dependent selection of the graph and adjustment set.*

Table 2: Mean-bias profiles beside the radius, five of 49 queries (Table 6). D: largest distance within X's component; $B(\widehat{\mathcal{C}})/|\theta_Z|$: largest possible bias, relative to the estimate; UNR: never exceeds ε.

					r_μ at tolerance ε					
network	query	r_{val}	D	$B(\widehat{\mathcal{C}})/\lvert\theta_Z\rvert$	2%	10%	25%	40%	60%	90%
barley	dg25 → s2225	3	6	1.53	3	3	3	3	3	5
barley	frspdag → dg25	2	6	1.00	2	2	2	2	3	3
magic-irri	G3212 → FT	2	7	1.23	2	2	2	4	4	6
Schipf_2010	A → U	1	8	1.00	1	1	3	4	5	7
Kampen_2014	AFF → CDR	1	13	1.00	1	2	5	7	9	12

Appendix E gives the proof and a retraction search for the nearest state exceeding a bias tolerance, but the worst case usually jumps to its maximum as soon as validity fails and rarely separates queries sharing a radius, so we report the mean instead.

An illustrative mean-bias profile. B depends only on the orientations in X's undirected component of $\widehat{\mathcal{C}}$, since τ_D depends only on $\text{pa}_D(X)$ and chain components are oriented independently. We average B uniformly (not as an expectation over true DAGs or an error model) over the states $\mathfrak{G}^X_{\widehat{\mathcal{C}}}$ that agree with G_0 outside that component, at each distance from G_0,

$$\mu(d) = \text{mean}\{B(G) : G \in \mathfrak{G}^X_{\widehat{\mathcal{C}}},\ d(G_0, G) = d\}, \qquad r_\mu(\varepsilon) = \min\{d : \mu(d) > \varepsilon|\theta_Z|\}. \tag{5}$$

$r_\mu(\varepsilon)$ is the smallest number of revisions at which the average bias exceeds ε of the reported effect – descriptive, not a guarantee (Theorem 2 and the worst-case ceiling $B(\widehat{\mathcal{C}})$ are). Queries with $\theta_Z = 0$ are excluded, since a relative tolerance is then undefined.

What the profile shows. In Table 2, the radius alone ranks barley dg25 → s2225 safest, yet its mean bias crosses every tolerance but 90% at once and keeps rising, while the two queries sharing $r_{\text{val}} = 1$ and $B(\widehat{\mathcal{C}})/|\theta_Z| = 1.00$ reach 90% only after seven and twelve revisions – illustrations, not corpus findings, since deeper components (Appendix E.5) grade more finely.

8 Limitations and Conclusion

Limitations. *Discovery error.* The radius takes the skeleton and v-structures of $\widehat{\mathcal{C}}$ as correct: a missed weak edge can open a back-door path no revision reaches, so the radius should be reported beside a skeleton-stability check. *Stand-in.* Where defined, separation ranks average-case survival at least as well, better at three and four tiers (Section 6.2); the radius adds exactness and named orientations. *Worst case.* The radius locates the nearest failure, so it is conservative (44% of the states beyond the certified region of the running example still admit Z; Appendix F); as a ranking it concerns consistent corruptions, since Meek closure flags contradictions (Section 6). *Units.* The radius counts closure orientations, not elicited claims; the claim radius $r_{\text{claim}} \leq r_{\text{val}}$ is positively associated with it ($\tau_b = 0.66$) but can rank queries differently, and where one claim entails many orientations both should be reported (Appendix H.2 and Remark 4). *Model.* Only Section 7 needs Assumption 1 (Appendix E.6 treats general models). *Coverage.* The SAT stage is reached on 6 of 256 timing instances (legacy implementation), so its scaling is untested; the effect-scale results use one coefficient draw per network.

Conclusion. The breakdown radius is an exact, query-specific certificate: if the true DAG lies in the equivalence class of $\widehat{\mathcal{C}}$ and contradicts fewer than r_{val} of the analyst's closure orientations, the committed adjustment set remains valid. It is a worst-case guarantee, not a universal predictor of average-case survival: it tracks survival closely on radius-spanning stress tests, more modestly on uniformly drawn queries, where separation does at least as well, and within fixed elicited knowledge states. It certifies one committed set rather than identification itself, and the elicited scenarios test the diagnostic, not the accuracy of the knowledge. On the effect scale the certified region carries no population identification bias, which does not by itself validate intervals after the graph and set are chosen from data.

REPRODUCIBILITY STATEMENT

The space of knowledge states, the distance and the breakdown radius are defined in Section 4, and the order-theoretic results they rest on are stated and proved in Appendix C, with the anti-exchange property proved in Appendix C.5. The exact search is given in Section 5 and its constraint encodings in Appendix D.1. Table 9 lists every setting of both experiments, and the datasheet of Appendix I documents the elicitation from the questionnaire and prompts to the repaired knowledge states. The code, the released knowledge states and the committed result files needed to reproduce every table are available at `https://anonymous.4open.science/r/bkrobust/`.

ETHICS STATEMENT

This work is a diagnostic for causal effect estimation and involves no human subjects and no personal data. The graphs used in Section 6.3 are published Bayesian networks and DAGs drawn from applied papers, and no data were collected for this study. The language models we elicit from are open-weight models, and their prompts contain only the variable names of published networks or scrambled labels. The diagnostic is meant to quantify a dependence that usually goes unstated: that of a reported causal estimate on expert assertions the data cannot check. We see one risk of misuse. Reporting a large radius as evidence that the analyst's knowledge agrees with the process that generated the data would invert its meaning, and Remark 1 states the distinction. In the running example, the valid adjustment sets under fixed knowledge happen to share one radius (Appendix F), so using it to choose among them is left to future work; the radius can be reported beside each candidate set's efficiency. Asserting more orientations to raise it makes none of them more accurate.

AI USE STATEMENT

Large language models enter this work in two roles, which we keep apart. As objects of study, six open-weight models were queried for causal orientations (Section 6.3 and Appendix I); their answers are data, and they took no part in analysing them. As tools, we used an AI coding assistant (Claude, from Anthropic) to write the experiment code and to revise text. We have reviewed all AI-assisted work, and we take responsibility for the final content of this paper, including text, claims, code and released data produced with the aid of generative AI.

REFERENCES

Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Vahid Balazadeh Meresht, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse Cresswell, and Rahul Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 154945–154984, 2025.

Christine W. Bang and Vanessa Didelez. Do we become wiser with time? On causal equivalence with tiered background knowledge. In *39th Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 119–129, 2023.

Christine W. Bang and Vanessa Didelez. Constraint-based causal discovery with tiered background knowledge and latent variables in single or overlapping datasets. In *4th Conference on Causal Learning and Reasoning*, volume 275, pp. 1116–1146, 2025.

Christine W Bang, Tobias Ellegaard Larsen, Anne Helby Petersen, and Alex Markham. Marginalising and interpreting graphs after causal discovery. In *UAI 2026 Workshop on Causality for Decision Making*, 2026.

Bertrand Charpentier, Simon Kibler, and Stephan Günnemann. Differentiable DAG sampling. In *International Conference on Learning Representations*, 2022.

David Maxwell Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 87–98, 1995.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B*, 82(1):39–67, 2020.

Chris Cundy, Aditya Grover, and Stefano Ermon. BCD nets: Scalable variational approaches for Bayesian causal discovery. In *Advances in Neural Information Processing Systems*, 2021.

Paul H. Edelman and Robert E. Jamison. The theory of convex geometries. *Geometriae Dedicata*, 19(3):247–270, 1985.

Adiba Ejaz and Elias Bareinboim. Less greedy equivalence search. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 123068–123109, 2025.

Richard Guo and Emilija Perković. Minimal enumeration of all possible total effects in a Markov equivalence class. In *24th International Conference on Artificial Intelligence and Statistics*, 2021.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society: Series B*, 84(2):579–599, 2022.

Leonard Henckel, Theo Würtzen, and Sebastian Weichwald. Adjustment identification distance: A gadjid for causal structure learning. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1569–1598, 2024.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

Zhongyi Hu and Stéphanie van der Pas. Selecting valid adjustment sets with uncertain causal graphs. *arXiv:2511.01662*, 2025.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *30th Conference on Uncertainty in Artificial Intelligence*, 2014.

Emil Javurek, Dennis Frauen, Marie Brockschmidt, Jonas Schweisthal, and Stefan Feuerriegel. Amortizing causal sensitivity analysis via prior data-fitted networks. *arXiv:2605.10590*, 2026.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024.

Neville K. Kitson and Anthony C. Constantinou. LLM-augmented causal discovery: Probabilistic fusion of edge existence and orientation. In *International Conference on Probabilistic Graphical Models*, 2026.

Wenrui Li, Wei Zhang, Qinghao Zhang, Xuegong Zhang, and Xiaowo Wang. Weakly supervised causal discovery based on fuzzy knowledge and complex data complementarity. *IEEE Transactions on Fuzzy Systems*, 32(12):7002–7014, 2024.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. DiBS: Differentiable Bayesian structure learning. In *Advances in Neural Information Processing Systems*, 2021.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *Advances in Neural Information Processing Systems*, 2022.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, 2022.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 79065–79098, 2026.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *11th Conference on Uncertainty in Artificial Intelligence*, pp. 403–410, 1995.

Valentyn Melnychuk, Vahid Balazadeh, Stefan Feuerriegel, and Rahul G Krishnan. Frequentist consistency of prior-data fitted networks for causal inference. In *43rd International Conference on Machine Learning*, 2026.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Kirtan Padh, Zhufeng Li, Cecilia Casolo, and Niki Kilbertus. Your assumed DAG is wrong and here's how to deal with it. In *Fourth Conference on Causal Learning and Reasoning*, volume 275, pp. 1239–1267, 2025.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Emilija Perković. Identifying causal effects in maximally oriented partially directed acyclic graphs. *36th Conference on Uncertainty in Artificial Intelligence*, 2020.

Emilija Perković, Markus Kalisch, and Marloes H. Maathuis. Interpreting and using CPDAGs with background knowledge. In *33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

Arik Reuter, Anish Dhir, Codrin Diaconu, Jake Robertson, Sebastian Ossen, Frank Hutter, Adrian Weller, Mark van der Wilk, and Bernhard Schölkopf. Use what you know: Causal foundation models with partial graphs. *arXiv:2602.14972*, 2026.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-context learning for causal effect estimation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 174811–174848, 2025.

Paul R. Rosenbaum. *Observational Studies*. Springer, 2nd edition, 2002.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188): 1–86, 2020.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The TETRAD project: Constraint based aids to causal model specification. In *Multivariate Behavioral Research*, volume 33, pp. 65–117, 1998.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Zehao Su and Leonard Henckel. A robustness test for estimating total effects with covariate adjustment. In *38th Conference on Uncertainty in Artificial Intelligence*, volume 180, pp. 1886–1895, 2022.

Armeen Taeb, F Richard Guo, and Leonard Henckel. Model-oriented graph distances via partially ordered sets. *arXiv:2511.10625*, 2026. v3.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 95330–95371, 2025.

Zidong Wang, Xi Lin, Chuchao He, and Xiaoguang Gao. Robust causal discovery under imperfect structural constraints. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 36757–36765, 2026.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

Yu Zhou, Xingyu Wu, Jibin Wu, Liang Feng, and Kay Chen Tan. CausalBN-Bench: A comprehensive benchmark for causal learning capability of LLMs. *IEEE Transactions on Artificial Intelligence*, 2026.

A Extended related work

Two lines of work sit next to those discussed in Section 2: causal discovery whose background knowledge may itself be wrong, and amortized and foundation-model approaches, which supply or consume such knowledge.

A.1 Causal discovery under imperfect background knowledge

Beyond the conflict-resolution and soft-knowledge work already cited in Section 2 (Hyttinen et al., 2014; Wang et al., 2026; Li et al., 2024), Bang & Didelez (2023) and Bang & Didelez (2025) study a complementary source of structure in background knowledge: a partial temporal or causal tiering of the variables, rather than edge-level orientation assertions, and characterise the equivalence classes and identification results such tiering licenses, including under latent confounding and across overlapping datasets. Ejaz & Bareinboim (2025) take background knowledge as a search heuristic rather than a hard constraint, using it to prune the space explored by greedy equivalence search. In all of these, a consistent constraint, once admitted, is taken as given. The breakdown radius quantifies how much the downstream conclusion depends on such constraints: it measures how many of them must be withdrawn before the adjustment set loses validity, whichever consistent state agrees with the data-generating graph.

Language models are a second source of imperfect background knowledge. Kiciman et al. (2024) give early evidence that LLMs can supply causal judgements competitive with, and sometimes better calibrated than, purely data-driven discovery; Vashishtha et al. (2025) elicit causal orderings from an LLM instead of full graphs, since orderings are easier to elicit reliably, and show that even partial, noisy orderings improve downstream discovery. Long et al. (2023) is the closest in spirit to our motivation. They treat an LLM as an imperfect expert whose suggestions are checked for consistency with the data before being incorporated, which is the pipeline whose downstream sensitivity our radius measures. Section 6.3 computes the radius on published networks for background knowledge elicited in this way.

A.2 Amortized and foundation-model approaches to causal structure and effects

A separate line of work learns to produce a causal graph, or a causal effect, directly from data via a model trained (or conditioned) once and reused across problem instances, rather than by running a discovery algorithm from scratch on each dataset. Its starting point is continuous structure learning on a single dataset: NOTEARS relaxes the acyclicity constraint (Zheng et al., 2018), and DiBS (Lorch et al., 2021), BCD Nets (Cundy et al., 2021) and differentiable DAG sampling (Charpentier et al., 2022) place variational or differentiable inference over graphs in place of combinatorial search. Lorch et al. (2022) and Löwe et al. (2022) amortize this inference: a network trained once, across many synthetic datasets, maps a new dataset directly to a posterior over graphs with no retraining. The same idea, applied to Bayesian inference generally by Müller et al. (2022) and to tabular prediction by TabPFN (Hollmann et al., 2023), now extends to causal effect estimation. CausalPFN (Balazadeh Meresht et al., 2025), CausalFM (Ma et al., 2026) and Do-PFN (Robertson et al., 2025) train a prior-data fitted network once on simulated causal data and return an effect estimate for a new problem in a single forward pass, with no graph discovered or asserted. Reuter et al. (2026) come closest to our setting: they condition such a network on a *partial* graph, so that whatever background knowledge is available still constrains the amortized estimate. Auditing this pipeline with our radius is a possible extension, with a learned model taking the place of Meek's rules between the partial graph and the estimate, though the present results are established only for the classical elicit-then-adjust pipeline with a fixed adjustment set. Melnychuk et al. (2026) show that the bias of a prior-data fitted network need not vanish with the data, so misspecification in the training prior, the model-based analogue of a wrong orientation, leaves a residual bias. This paper asks, for the classical elicit-then-adjust pipeline, how many structural commitments must be wrong before the reported effect changes; extending the same question to an amortized estimator conditioned on a partial graph, with that partial graph taken as the knowledge state, is left open.

B Notation

A ball of radius k about G_0 is $\{G : d(G_0, G) \le k\}$ and the shell at distance k is $\{G : d(G_0, G) = k\}$. Every guarantee in the paper is stated on a ball; shell-wise quantities are descriptive (Remark 5).

C Order-theoretic results and proofs

Throughout, $\widehat{\mathcal{C}}$ is a CPDAG, $\mathfrak{G}_{\widehat{\mathcal{C}}}$ the space of Definition 1, and for $G \in \mathfrak{G}_{\widehat{\mathcal{C}}}$ we write $\mathcal{K}_G = \mathrm{dir}(G) \setminus \mathrm{dir}(\widehat{\mathcal{C}})$, so that by Meek's theorem $[G] = \{D \in [\widehat{\mathcal{C}}] : D \models \mathcal{K}_G\}$ and G is *maximally oriented*: every edge left undirected in G takes both orientations among the DAGs in $[G]$. We write cl for the induced closure operator on orientation sets, $\mathrm{cl}(S) = \mathcal{K}_{\mathrm{Meek}(\widehat{\mathcal{C}},S)}$.

C.1 Membership

Definition 1 characterises membership by a fixpoint condition, and two weaker conditions show why. Asking only that a graph be a valid MPDAG *in its own right* admits graphs from a different Markov equivalence class: on a chain $A - B - C$ the collider $A \to B \leftarrow C$ is vacuously Meek-closed and extends itself. Asking that each component of the undirected subgraph be chordal is a necessary structural property of CPDAGs, not by itself a characterisation, and it misses MPDAGs that carry background knowledge, because knowledge can place a directed edge inside an otherwise undirected component so that the chord which would resolve it is present but directed. The minimal instance is $\widehat{\mathcal{C}} = K_4$ minus one edge, with $\mathcal{K} = \{V_0 \to V_1\}$. The resulting state is Meek-closed, has five DAG extensions and is maximally oriented, yet its undirected part is not chordal. Such states are absent below five undirected edges and make up 10.2% of space elements at eight, so the fixpoint condition matters in practice.

C.2 How the results fit together

Figure 4 is the dependency graph of the formal results stated in Sections 5 and 7 and proved in this appendix and in Appendix E. The order-theoretic spine on the left terminates in Theorem 1, and

Table 3: Symbols used throughout. Every symbol is defined at first use in the text.

Symbol	Description
GRAPHS AND KNOWLEDGE	
$\widehat{\mathcal{C}}$	the estimated CPDAG; $[\widehat{\mathcal{C}}]$ its Markov equivalence class
$\mathcal{K}$	the orientations the analyst elicits and asserts
$\mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K})$	the MPDAG obtained by imposing $\mathcal{K}$ and closing under R1–R4
G_0	the analyst's knowledge state, $G_0 = \mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K})$
$\mathcal{K}_G$	the closure orientations of G, $\mathrm{dir}(G) \setminus \mathrm{dir}(\widehat{\mathcal{C}})$; $\mathcal{K} \subseteq \mathcal{K}_{G_0}$
$\lvert\mathcal{K}_G\rvert$	the corank: how many closure orientations G carries
$\rho(G)$	the rank, $\rho(G) = m - \lvert\mathcal{K}_G\rvert$, for m the number of undirected edges of $\widehat{\mathcal{C}}$
$[G]$	the DAGs the state G represents
Z	the adjustment set read off G_0 and then held fixed
THE SPACE AND ITS GEOMETRY	
$\mathfrak{G}_{\widehat{\mathcal{C}}}$	the space of knowledge states induced by $\widehat{\mathcal{C}}$
$G \preceq H$	model inclusion, $[G] \subseteq [H]$; lower means more knowledge
$G \vee H$	the join, i.e., the least state weaker than both
$d(G, H)$	hop count on the covering graph of $\preceq$
r_{val}	breakdown radius: distance to the nearest state where Z is invalid
EFFECT SCALE (requires the linear-Gaussian law of Assumption 1, Section 7)	
Σ	the population covariance of the observational law
$\beta(S; \Sigma)$	coefficient on X in the population regression of Y on $\{X\} \cup S$
θ_Z	$\beta(Z; \Sigma)$: the single number the pipeline reports
τ_D	0 if $Y \notin \mathrm{de}_D(X)$, else $\beta(\mathrm{pa}_D(X); \Sigma)$: the effect the candidate DAG D assigns
$T(G)$	$\{\tau_D : D \in [G]\}$: the effect ambiguity G leaves open
$B(G)$	worst-case bias, $\max\{\lvert\theta_Z - \tau\rvert : \tau \in T(G)\}$
$\beta^{\uparrow}(d)$	the largest B over the ball of radius d about G_0
ε	the bias tolerance the analyst is willing to carry
r_ε	ε-bias radius: least d with $\beta^{\uparrow}(d) > \varepsilon\lvert\theta_Z\rvert$ (defined only for $\theta_Z \neq 0$)
$\mu(d)$	mean of B over the states at distance d from G_0 that agree with it outside X's undirected component
r_μ	mean-bias profile: at each ε, the least d with $\mu(d) > \varepsilon\lvert\theta_Z\rvert$ (defined only for $\theta_Z \neq 0$)

every lemma on it serves that theorem. Corollary 1 is a direct corollary of Theorem 1 together with Lemmas 2 to 4 and is omitted from the figure for space; its proof is below. Assumption 1, the one parametric restriction in the paper, enters only on the right-hand branch. The results of Sections 4 and 5 therefore hold for any observational law compatible with $\widehat{\mathcal{C}}$, and only those of Section 7 require the linear-Gaussian model.

C.3 THE DOMAIN OF THE ORDER-THEORETIC ARGUMENT

Anti-exchange, and the convex-geometry characterisation it feeds, are applied to a fixed finite ground set and to no larger one. Fix $G \in \mathfrak{G}_{\widehat{\mathcal{C}}}$ and take as ground set its closure orientations $\mathcal{K}_G$, with the operator $S \mapsto \mathrm{cl}(S)$ restricted to subsets $S \subseteq \mathcal{K}_G$.

Remark 2 (Ground set and consistency). The argument uses three properties of this domain. (i) *Every subset is consistent*: any $D \in [G]$ contains every orientation of $S \subseteq \mathcal{K}_G$ and lies in $[\widehat{\mathcal{C}}]$, so it witnesses the consistency of S. (ii) *Closure stays inside the ground set*: $\mathrm{Meek}(\widehat{\mathcal{C}}, S) \succeq G$ by monotonicity of closure, so $\mathrm{cl}(S) \subseteq \mathcal{K}_G$ by Lemma 3, and cl is a closure operator on the Boolean lattice of subsets of $\mathcal{K}_G$. (iii) Given (i) and (ii), anti-exchange yields that every covering pair of closed sets inside $\mathcal{K}_G$ differs by one element, so a shortest route to a failing state decomposes into single retractions.

Remark 3 (Rank and chain length). Gradedness is stated through the corank $\lvert\mathcal{K}_G\rvert$ (decreasing upward); equivalently the rank $\rho(G) = m - \lvert\mathcal{K}_G\rvert$, for m the number of undirected edges of $\widehat{\mathcal{C}}$. A retraction moves *upward* in $\preceq$ while *removing* an orientation, so $\lvert\mathcal{K}_G\rvert$ decreases along an upward chain: if $G_0 \preceq J$ then $\mathcal{K}_J \subseteq \mathcal{K}_{G_0}$ and the length of a saturated chain from G_0 to J is the nonnegative quantity $\lvert\mathcal{K}_{G_0}\rvert - \lvert\mathcal{K}_J\rvert$, which equals $d(G_0, J)$ along such a chain.

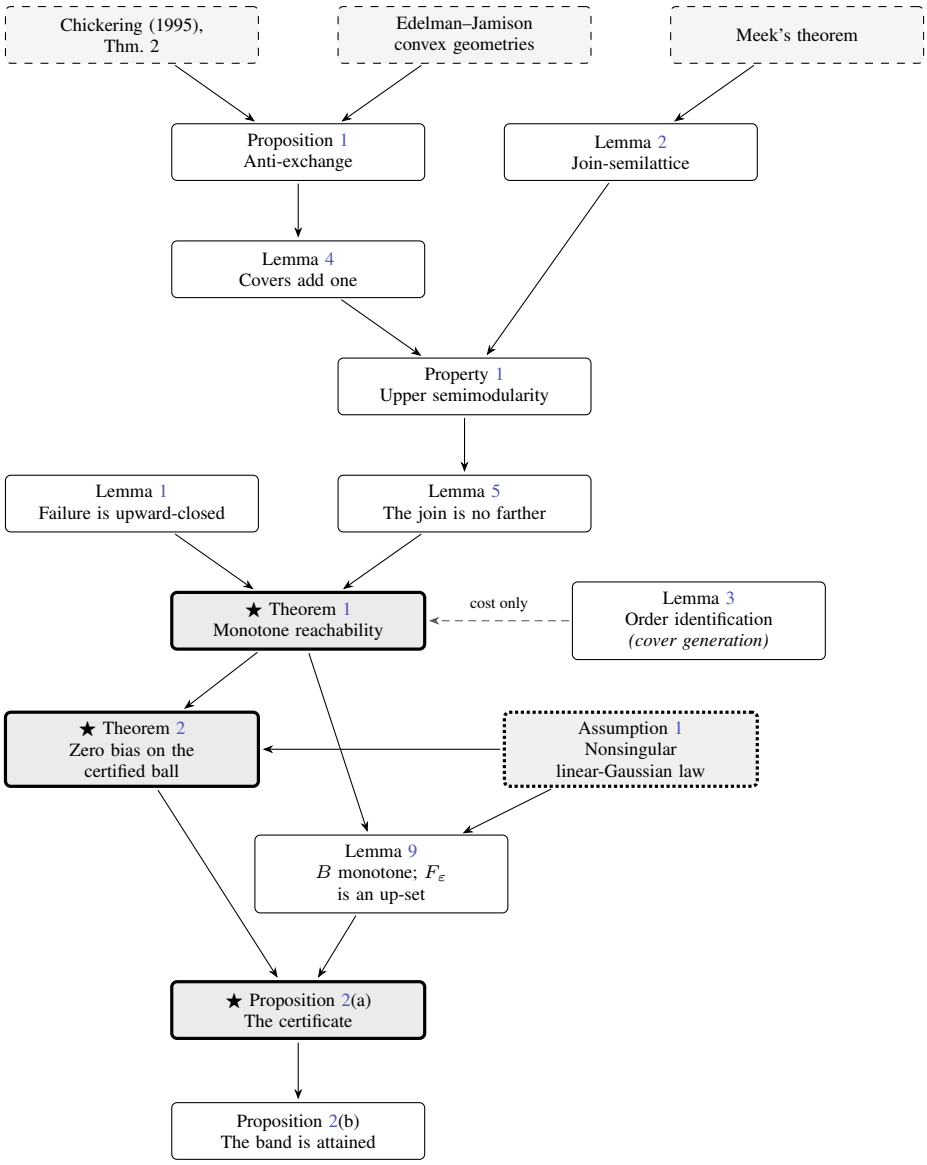

Figure 4: Dependency graph of the formal results. An arrow means "is used in the proof of". Dashed boxes are external results taken as input; the heavy boxes marked ★ are the main results; the dotted box is the paper's only parametric assumption. Lemma 3 makes cover generation cheap and carries no proof weight, hence its dashed arrow. Assumption 1 feeds only the right-hand branch, so the validity radius of Sections 4 and 5 is model-free.

C.4 PROOFS

Proof of Lemma 1. Validity of Z in a state is a universal statement over the DAGs it represents. If Z is invalid in G, some $D \in [G]$ has Z invalid; since $[G] \subseteq [G']$, that same D lies in $[G']$ and witnesses failure there. □

Proof of Lemma 2. Write $J = \mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K}_G \cap \mathcal{K}_H)$. *Upper bound*: any $D \in [G]$ satisfies $\mathcal{K}_G$, hence $\mathcal{K}_G \cap \mathcal{K}_H$, hence lies in $[J]$; symmetrically for H. *Least*: let $W \in \mathfrak{G}_{\widehat{\mathcal{C}}}$ with $[G] \cup [H] \subseteq [W]$ and take $e = (a \to b) \in \mathcal{K}_W$. Every $D \in [G]$ satisfies e, so the edge $a \sim b$ is oriented identically across $[G]$; since G is maximally oriented, G contains $a \to b$ as a directed edge, and as $e \notin \mathrm{dir}(\widehat{\mathcal{C}})$ we get $e \in \mathcal{K}_G$, and symmetrically $e \in \mathcal{K}_H$. Hence $\mathcal{K}_W \subseteq \mathcal{K}_G \cap \mathcal{K}_H$ and $[J] \subseteq [W]$. □

Meets exist for any pair with a common lower bound. Since $\mathcal{K}_G \cup \mathcal{K}_H$ may be inconsistent and the empty model set is not an element, $\mathfrak{G}_{\widehat{\mathcal{C}}}$ is a join-semilattice with a top.

Lemma 4 (Covers add one orientation)**.** *If $G \lessdot H$ in $\mathfrak{G}_{\widehat{\mathcal{C}}}$ then $|\mathcal{K}_G \setminus \mathcal{K}_H| = 1$.*

Proof. By Edelman & Jamison (1985), a closure operator satisfies anti-exchange (Proposition 1) if and only if its closed sets form a convex geometry, and in a convex geometry every cover in the lattice of closed sets adds exactly one element. □

In particular $|\mathcal{K}_G| = |\mathrm{dir}(G)| - |\mathrm{dir}(\widehat{\mathcal{C}})|$ is a corank function (rank $\rho(G) = m - |\mathcal{K}_G|$, for m the number of undirected edges of $\widehat{\mathcal{C}}$), and the order is graded.

Property 1 (Upper semimodularity)**.** For $G \lessdot H$ and any $W \in \mathfrak{G}_{\widehat{\mathcal{C}}}$, either $G \vee W = H \vee W$ or $G \vee W \lessdot H \vee W$.

Proof. Work on the orientation side, where by Lemma 2 the join is intersection. By Lemma 4, $\mathcal{K}_G = \mathcal{K}_H \sqcup \{e\}$. If $e \notin \mathcal{K}_W$ then $\mathcal{K}_G \cap \mathcal{K}_W = \mathcal{K}_H \cap \mathcal{K}_W$ and the joins coincide. If $e \in \mathcal{K}_W$ then $\mathcal{K}_G \cap \mathcal{K}_W = (\mathcal{K}_H \cap \mathcal{K}_W) \sqcup \{e\}$; intersections of closed sets are closed, so both are elements, and no closed set lies strictly between two sets differing by one element. □

Lemma 5 (The join is no farther)**.** *For all $G_0, G \in \mathfrak{G}_{\widehat{\mathcal{C}}}$, $d(G_0, G \vee G_0) \leq d(G_0, G)$.*

Proof. Let $G_0 = P_0, \ldots, P_d = G$ be a shortest path in the neighbour graph and put $Q_i = P_i \vee G_0$, so $Q_0 = G_0$ and $Q_d = G \vee G_0$. Each consecutive pair $\{P_i, P_{i+1}\}$ is a covering pair in one direction, so by Property 1 Q_i and Q_{i+1} are equal or adjacent. Deleting repetitions leaves a walk of length at most d. □

Proof of Theorem 1. Let G be a nearest failing state, $d(G_0, G) = r_{\mathrm{val}}$, and put $J = G \vee G_0$. Then $G \preceq J$, so by Lemma 1 Z fails at J; by Lemma 5, $d(G_0, J) \leq r_{\mathrm{val}}$. Semimodularity gives the Jordan–Dedekind chain condition, hence gradedness, so for the comparable pair $G_0 \preceq J$ every path changes rank by one at each step and a saturated chain of length $|\mathcal{K}_{G_0}| - |\mathcal{K}_J|$ exists (Remark 3): the distance to J is realised by a monotone upward path. A retraction-only search therefore reaches a failing state in at most r_{val} steps, and cannot do better, since r_{val} is the minimum over all of $\mathfrak{G}_{\widehat{\mathcal{C}}}$. □

Proof of Corollary 1. (a) By Lemma 4 every covering step adds or removes exactly one orientation, so every path between G and H has length at least $|\mathcal{K}_G \triangle \mathcal{K}_H|$. For the converse, let $J = G \vee H$ and $S = \mathcal{K}_G \cap \mathcal{K}_H$. Both $\mathcal{K}_G$ and $\mathcal{K}_H$ are closed ($G, H \in \mathfrak{G}_{\widehat{\mathcal{C}}}$), and cl is monotone and extensive, so $S \subseteq \mathrm{cl}(S) \subseteq \mathrm{cl}(\mathcal{K}_G) \cap \mathrm{cl}(\mathcal{K}_H) = S$: the intersection of closed sets is closed, so $\mathcal{K}_J = S$ by Lemma 2. By Remark 3, $\mathcal{K}_J \subseteq \mathcal{K}_G$ gives a saturated chain of retractions from G to J of length $|\mathcal{K}_G| - |\mathcal{K}_J| = |\mathcal{K}_G \setminus \mathcal{K}_H|$, and symmetrically from H to J of length $|\mathcal{K}_H \setminus \mathcal{K}_G|$. Concatenating gives a walk from G to H of length $|\mathcal{K}_G \triangle \mathcal{K}_H|$, so $d(G, H) \leq |\mathcal{K}_G \triangle \mathcal{K}_H|$; with the lower bound, equality.

(b) $[G] \subseteq [\widehat{\mathcal{C}}]$ for every $G \in \mathfrak{G}_{\widehat{\mathcal{C}}}$, and validity of an adjustment set is universal over the DAGs represented, so if Z is valid in $\widehat{\mathcal{C}}$ it is valid in every $D \in [\widehat{\mathcal{C}}]$, hence in every state: $r_{\mathrm{val}} = \infty$. Conversely, if Z is invalid in $\widehat{\mathcal{C}}$ then $\widehat{\mathcal{C}}$ itself is a failing state, and by (a), $d(G_0, \widehat{\mathcal{C}}) = |\mathcal{K}_{G_0} \triangle \mathcal{K}_{\widehat{\mathcal{C}}}| = |\mathcal{K}_{G_0}|$ since $\mathcal{K}_{\widehat{\mathcal{C}}} = \emptyset$; so $r_{\mathrm{val}} \leq |\mathcal{K}_{G_0}|$.

(c) Let $D \in [\widehat{\mathcal{C}}]$ with Z invalid for (X, Y) in D, and put $S = \mathcal{K}_{G_0} \cap \mathrm{dir}(D)$. As in (a), S is the intersection of the closed set $\mathcal{K}_{G_0}$ with $\mathrm{dir}(D)$, and D witnesses that every orientation of S is consistent, so, arguing as in Remark 2(i)–(ii), $\mathrm{cl}(S) \subseteq \mathcal{K}_{G_0}$ and $\mathrm{cl}(S) = S$. Let $G' = \mathrm{Meek}(\widehat{\mathcal{C}}, S)$; then D satisfies S, so $\mathrm{dir}(G') \subseteq \mathrm{dir}(D)$, giving $D \in [G']$ by Lemma 3, so Z is invalid in G'. By (a), $d(G_0, G') = |\mathcal{K}_{G_0} \setminus S| = |\mathcal{K}_{G_0} \setminus \mathrm{dir}(D)|$, so the right-hand side is an upper bound on r_{val}.

Conversely, by Theorem 1 take a nearest failing state G with $\mathcal{K}_G \subseteq \mathcal{K}_{G_0}$ and $|\mathcal{K}_{G_0} \setminus \mathcal{K}_G| = r_{\mathrm{val}}$. Since Z is invalid in G, some $D \in [G] \subseteq [\widehat{\mathcal{C}}]$ has Z invalid, and $\mathrm{dir}(G) \subseteq \mathrm{dir}(D)$ gives $\mathcal{K}_G \subseteq \mathrm{dir}(D)$, so $|\mathcal{K}_{G_0} \setminus \mathrm{dir}(D)| \leq |\mathcal{K}_{G_0} \setminus \mathcal{K}_G| = r_{\mathrm{val}}$. The two bounds meet. □

Remark 4 (The claim radius, in the same form). The argument of Corollary 1(c) applies verbatim to the elicited set $\mathcal{K}$ of Appendix H.2 in place of $\mathcal{K}_{G_0}$:

$$r_{\mathrm{claim}} = \min\{\, |\mathcal{K} \setminus \mathrm{dir}(D)| : D \in [\widehat{\mathcal{C}}],\ Z \text{ invalid for } (X, Y) \text{ in } D \,\},$$

hence $r_{\mathrm{claim}} \leq r_{\mathrm{val}}$, since $\mathcal{K} \subseteq \mathcal{K}_{G_0}$. For a failing D, retracting $R = \mathcal{K} \setminus \mathrm{dir}(D)$ leaves claims contained in $\mathrm{dir}(D)$, so by soundness of the Meek closure – the extensions of $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$ are exactly the DAGs of $[\widehat{\mathcal{C}}]$ containing $S - D \in [\mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K} \setminus R)]$ and Z fails there; conversely if retracting R makes Z invalid, some failing $D \in [\mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K} \setminus R)]$ contains $\mathcal{K} \setminus R$, so $\mathcal{K} \setminus \mathrm{dir}(D) \subseteq R$. This gives a direct exact optimisation encoding of r_{claim}: minimise violated commitments over the DAG extensions of $\widehat{\mathcal{C}}$ subject to an adjustment-failure witness. The released encodings reach the same optimum through an explicit closed intermediate state rather than this bare form (Appendix D.1).

Proof of Lemma 3. (⇐) Weaker orientation constraints admit more extensions. (⇒) Let $a \to b \in \mathrm{dir}(H)$. Every $D \in [H]$ orients it that way, and $[G] \subseteq [H]$, so every $D \in [G]$ does too; since G is maximally oriented, $a \to b \in \mathrm{dir}(G)$. □

Proof of Lemma 9. Each τ_D, defined as 0 if $Y \notin \mathrm{de}_D(X)$ and as $\beta(\mathrm{pa}_D(X); \Sigma)$ otherwise, is a function of the graph D and the law Σ alone: it does not depend on the knowledge state through which D was reached. Hence $[G] \subseteq [H]$ gives

$$T(G) = \{\tau_D : D \in [G]\} \ \subseteq\ \{\tau_D : D \in [H]\} = T(H).$$

By equation 4, B is the maximum of the fixed function $\tau \mapsto |\theta_Z - \tau|$ over that set, and a maximum over a subset is no larger, so $B(G) \leq B(H)$.

For the consequences, let $G \in F_\varepsilon$ and $G \preceq H$. Then $B(H) \geq B(G) > \varepsilon$, so $H \in F_\varepsilon$ and F_ε is upward-closed. If $\varepsilon \leq \varepsilon'$ and $B(G) > \varepsilon'$ then $B(G) > \varepsilon$, so $F_{\varepsilon'} \subseteq F_\varepsilon$. □

The argument uses no property of adjustment sets, of d-separation or of the particular Σ: only that B is a maximum over a set indexed by $[\cdot]$, and that $[\cdot]$ grows as knowledge is withdrawn.

Remark 5 (Order is not distance). Lemma 9 orders only comparable states, and neither consequence is a statement about $d(G_0, \cdot)$. States at distance 1 and 2 are typically incomparable. Appendix E.4 exhibits a four-node instance in which the nearer carries bias 1.34 and the further none, and along a path that is not a chain of retractions nothing forces B to rise. What is monotone in the distance is $\beta^{\uparrow}$, and for a different and more elementary reason than Lemma 9: $\beta^{\uparrow}(d)$ maximises B over the ball of radius d, balls grow with d, and a maximum over a superset is no smaller. The distinction is practical as well as formal, because a maximum tabulated over the *shells* need not rise with d: distinct shells are disjoint, so no inclusion holds between them and a shell maximum may fall while the running maximum over the ball does not. Lemma 9 makes F_ε an up-set and so confines the search to retractions; ball inclusion makes $\beta^{\uparrow}$ a staircase.

C.5 The anti-exchange property

We prove Proposition 1. Fix $H \in \mathfrak{G}_{\widehat{\mathcal{C}}}$ as in Remark 2 and work inside its domain, ground set $\mathcal{K}_H$; for $S \subseteq \mathcal{K}_H$ write $G_S = \mathrm{Meek}(\widehat{\mathcal{C}}, S)$ for the associated state, so $\mathrm{cl}(S) = \mathcal{K}_{G_S}$. Distinct $x, y \in \mathcal{K}_H$ can never orient the same edge of $\widehat{\mathcal{C}}$'s skeleton, since $\mathcal{K}_H$ (being $\mathrm{dir}(H) \setminus \mathrm{dir}(\widehat{\mathcal{C}})$, a set of directed edges of the single DAG-like state H) contains at most one orientation of each edge; the case of the axiom in which x and y orient the same edge therefore cannot arise and needs no

argument. The remaining case, x and y orienting distinct undirected edges of G_S, needs more care. An edge $a \to b$ is *covered* in D if $\mathrm{pa}(b) = \mathrm{pa}(a) \cup \{a\}$, and reversing a covered edge preserves the equivalence class. Chickering's theorem connects any two DAGs of $[\widehat{\mathcal{C}}]$ by a sequence of covered-edge reversals with every intermediate DAG in $[\widehat{\mathcal{C}}]$. The orientations of $\mathcal{K}_{G_S}$ are uncompelled in $[\widehat{\mathcal{C}}]$, so the theorem is free to reverse them, and the existence of a path alone does not keep it inside $[G_S]$. Reorienting within a chordal chain component, the classical device, does not apply either, because an MPDAG carrying background knowledge can have a non-chordal undirected part. The instance of the Membership discussion shows this: for $\widehat{\mathcal{C}} = K_4$ minus the edge $\{V_2, V_3\}$ and $\mathcal{K} = \{V_0 \to V_1\}$, the undirected part $V_0 - V_2 - V_1 - V_3 - V_0$ of the resulting state is a chordless four-cycle.

The argument instead uses the length of such a path, which the theorem also supplies. We take the following as an established input, as the opening paragraph of this appendix takes Meek's theorem.

Chickering (1995), Thm. 2 (Chickering, 1995). If $D, D' \in [\widehat{\mathcal{C}}]$ differ on m edges, there is a sequence of exactly m distinct covered-edge reversals in D, after each of which the graph remains in $[\widehat{\mathcal{C}}]$, and after all of which $D = D'$.

Lemma 6 (No agreeing edge is ever touched)**.** *In the sequence supplied by the theorem above for $D, D' \in [\widehat{\mathcal{C}}]$ differing on m edges, the m reversed edges are exactly the m edges on which D and D' differ, each reversed once; no edge on which D and D' agree is ever reversed.*

Proof. Let $D = D_0, D_1, \ldots, D_m = D'$ be the supplied sequence and let δ_k be the number of edges on which D_k and D' differ, so $\delta_0 = m$ and $\delta_m = 0$. Each step reverses exactly one edge, so $|\delta_k - \delta_{k-1}| = 1$; a walk of m unit steps from m to 0 must decrease at every step, so every step reverses an edge on which the current graph and D' still differ. The reversals are of distinct edges and there are m of them, so they are exactly the initially differing edges, one each. □

Lemma 7 (The sequence stays inside $[G]$)**.** *For $D_A, D_B \in [G]$ there is a sequence of covered-edge reversals from D_A to D_B with every intermediate DAG in $[G]$.*

Proof. D_A and D_B both extend G, so they agree on all of $\mathrm{dir}(G)$. Apply the theorem above to D_A, D_B: by Lemma 6 no edge of $\mathrm{dir}(G)$ is ever reversed, so every intermediate D_k retains $\mathrm{dir}(G)$; each D_k lies in $[\widehat{\mathcal{C}}]$ by the theorem, and a DAG of $[\widehat{\mathcal{C}}]$ that contains $\mathrm{dir}(G)$ is by definition an element of $[G]$. □

Completing the proof. Assume for contradiction that $y \in \mathrm{cl}(S \cup \{x\})$ and $x \in \mathrm{cl}(S \cup \{y\})$ for distinct undirected edges $e_1 = u_1 - v_1$ and $e_2 = u_2 - v_2$ of G_S. Monotonicity and idempotence of cl force $\mathrm{cl}(S \cup \{x\}) = \mathrm{cl}(S \cup \{x, y\}) = \mathrm{cl}(S \cup \{y\})$, so every $D \in [G_S]$ satisfies $x \iff y$: no extension exhibits a mixed configuration $x \wedge \neg y$ or $\neg x \wedge y$.

G_S is maximally oriented and e_1 is undirected in G_S, so some $D_A \in [G_S]$ satisfies x and some $D_B \in [G_S]$ satisfies $\neg x$; by the equivalence just established, $D_A \models x \wedge y$ and $D_B \models \neg x \wedge \neg y$, so D_A and D_B differ on both e_1 and e_2. Apply Lemma 7 (with $G = G_S$) to D_A, D_B: every intermediate D_k lies in $[G_S]$, and by Lemma 6 e_1 is reversed at exactly one step i and e_2 at exactly one step j, with $i \neq j$ since the reversed edges are distinct. If $i < j$ then D_i has e_1 already reversed and e_2 not yet, so $D_i \models \neg x \wedge y$; if $j < i$ then $D_j \models x \wedge \neg y$. Either way an element of $[G_S]$ exhibits a mixed configuration, contradicting the paragraph above. □

C.6 Exhaustive corroboration

The statements proved above were also checked exhaustively on small graphs (Table 4). The two lemmas behind the anti-exchange proof correspond to rows L2 and L3. Row AE checks anti-exchange in its semantic form, by a route that does not pass through Chickering's theorem. The remaining results were checked in the same way. The fixpoint condition of Definition 1 coincides with reachability on all 1,588 states over every CPDAG on three and four nodes with an undirected edge (133 classes). Lemma 4 holds on 5,348 covering pairs and Lemma 3 on 80,480 ordered pairs. Retraction-only radii equal brute-force distances in 150 comparisons at three nodes, 13,008 at four and 1,977,820 across all 8,782 CPDAGs on five nodes, as Theorem 1 states.

Table 4: Exhaustive check of the lemmas behind the anti-exchange proof over every CPDAG on at most five nodes, with one sampled row at $n = 6$. Each statement holds on every one of the 3,584,369 instances checked.

check	statement	instances
L1	a covered-edge reversal preserves the equivalence class, exhaustive at $n = 3, 4, 5$ (all 29,281 DAGs at $n = 5$)	50,758
L1, $n = 6$	same, 4,000 sampled DAGs	7,034
L2	a difference-reducing sequence of exactly m reversals exists, exhaustive over every equivalence class and ordered pair at $n = 3, 4, 5$	206,878
L3	every intermediate DAG stays inside $[G]$, exhaustive over all 8,782 CPDAGs and 116,992 knowledge states at $n = 5$	2,907,242
L4	$[G]$ is connected under reversals that remain in $[G]$, same scope as L3	118,643
AE	no two distinct undirected edges induce the same bipartition of $[G]$, same scope as L3	293,814

D COMPUTATION

Two complementary searches. Corollary 1(b) settles infinite robustness with a single adjustment-criterion call on $\widehat{\mathcal{C}}$, so what remains is to locate a finite radius. The upward search of Algorithm 1 finds a nearby failure quickly. When the nearest failure lies beyond the depth budget, certifying that Z survives every state within distance k needs more than the upward search: the SAT stage of Algorithm 1 asks, at each radius k, whether there is a closed state obtained by retracting at most k orientations in which Z is invalid; one Boolean variable encodes each possible orientation, and clauses encode Meek-closedness, acyclicity, the absence of new v-structures, and a witness DAG in which Z violates the adjustment criterion. An unsatisfiable answer at k is a proof that no failure lies within k retractions. Algorithm 1 runs the upward search under a small depth budget and moves to the SAT stage when no failure lies within it; we call this combination the hybrid. The switch affects only running time, since both stages return the same radius.

Instances and timing. We time every method on the same instances and the same query. The family has 256 Erdős–Rényi instances: $n \in \{8, 10, 12, 14, 16, 18, 20, 24\}$ variables, edge probability $p \in \{0.3, 0.5, 0.7, 0.85\}$ and eight draws per cell, with G_0 the data-generating DAG under full knowledge. An instance whose CPDAG has no undirected edge is redrawn (14 of 256), so k, the number of undirected edges of $\widehat{\mathcal{C}}$, ranges from 1 to 26. The query is the first pair (X, Y) whose optimal adjustment set is non-empty and valid. Three methods answer it: building the whole space and searching it breadth-first, the retraction search alone, and the hybrid of Algorithm 1 with depth budget 3. Each run has a time limit of 120 s. All timings in this appendix were measured with an earlier implementation using a back-door validity oracle, which the generalized criterion replaced (Appendix D.1). They do not benchmark the final generalized-adjustment implementation; they are reported only as legacy measurements of the search strategy.

Results. On all 256 instances, every method that answers within 120 s returns the same radius; Table 5 gives the per-k medians. The retraction search removes the enumeration bottleneck. The median cost of building the space rises from 0.6 ms at $k = 1$ to 54 s at $k = 10$, and enumeration exceeds 120 s on 2 of the 8 instances at $k = 10$ and on all 20 with $k \geq 11$. The median retraction search stays below 0.1 s at every $k \leq 10$, and on 12 of the 20 instances with $k \geq 11$ it answers in under 0.1 s.

The top-state check (Corollary 1(b)) settles 164 of the 256 instances – 64% of the family, with G_0 the fully informed data-generating DAG – in one adjustment-criterion call on $\widehat{\mathcal{C}}$, median 0.14 ms, before either search runs. The hybrid returns the radius on all 256 instances, with median 0.25 ms and maximum 2.0 s ($n = 18$, $k = 20$). Of the remaining 92 instances, 86 are answered by the bounded search within its depth-3 budget; the SAT stage is reached only on the 6 whose radius

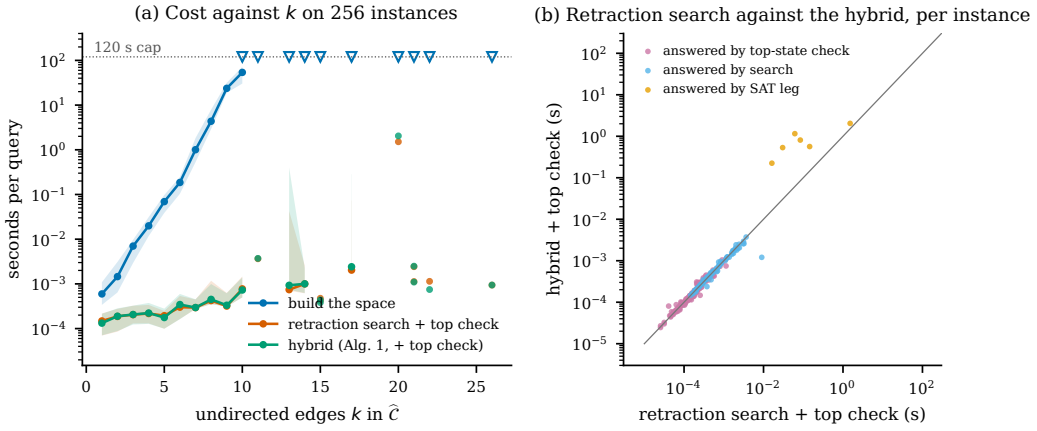

Figure 5: Legacy timings, measured with an earlier implementation using the back-door criterion; they do not benchmark the final generalized-adjustment implementation. Cost per query on one family of 256 instances, with the top-state check (Corollary 1(b)) run first on every method that reaches Algorithm 1. (a) Median time against the number k of undirected edges, for building the space, the retraction search, and the hybrid, on the same instances; bands are interquartile ranges over at least three completed runs, dots are individual runs elsewhere, and triangles mark the k at which at least one run exceeds 120 s. The cost of building the space grows exponentially in k and exceeds 120 s on every instance beyond $k = 10$. (b) Retraction search against the hybrid, per instance, coloured by which stage answers: the top-state check, the bounded search, or the SAT stage.

is finite but lies beyond that budget (radii $4, 4, 4, 5, 5, 6$, between 0.22 and 2.04 s each). These timings therefore exercise the SAT stage lightly; its scaling on harder, finite-radius instances is not established here. For transparency: without the top-state check, the same family gives hybrid median 1.1 ms, maximum 45.6 s, and 85 SAT dispatches.

Runtime on the evaluated states. The timings above use fully informed states, while Section 6 evaluates corrupted ones. With the same earlier implementation, on all 4,104 of those the radius is computed in under 0.2 s, with a median of 3.2 ms under flips (2,730 instances) and 1.9 ms under tiered corruption (1,374 instances); the 95th percentiles are 79 ms and 57 ms. On the elicited states of the published networks, the median over the 99 queries within the scope stated in Appendix E.4 is 27 ms.

D.1 The constraint encodings

Three encodings of the radius query differ in what they search and in what they assume. Each returns an *upper* bound on the exact radius, so none can report a radius that is too small.

	searches	assumes	guarantee
E1	retractions of $\mathcal{K}_{G_0}$	Theorem 1	$r_{\text{val}} \le r_{E1}$
E2	every closed state, distance via the join	Lemma 2, Lemma 4	$r_{\text{val}} \le r_{E2} \le r_{E1}$
E3	walks of single-orientation covering steps	nothing	$r_{\text{val}} \le r_{E3}$

E1 is the encoding used by Algorithm 1. E2 exploits Lemma 2: since the join is intersection and $\mathcal{K}_{G_0}$ is a constant, the distance objective collapses to the symmetric difference $|\mathcal{K}_{G_0} \triangle \mathcal{K}_G|$, a single linear constraint over one copy of the orientation variables – exactly the closed form of Corollary 1(a).

E3 assumes no theorem of this paper. Consecutive states in an E3 walk are closed and differ by one orientation, and two sets differing by a single element admit nothing strictly between them, so each such pair is a covering pair without appeal to Lemma 4 or Proposition 1. An E3 walk of length k to a failing state therefore proves $r_{\text{val}} \le k$ directly. On 1,007 instances $r_{E3} = r_{E1}$ on every one, as

Theorem 1 predicts. They comprise every four-node instance and generated graphs of up to twenty nodes, dense graphs with high knowledge among them. E3 unrolls one copy of the closure encoding per step, so its size grows with the radius as well as with the graph.

The E1 encoding, in detail. The SAT stage of Algorithm 1 is a ladder of solver calls implemented in `bkrobust.sat.e1`, built on the shared closure encoding of `bkrobust.sat.closure` and the failure encoding of `bkrobust.sat.failure`, and solved with Google OR-Tools' CP-SAT solver.

Variables. One Boolean s_e per ordered adjacent pair $e = (u, v)$ of $\widehat{\mathcal{C}}$'s skeleton ($O(n^2)$ many), pinned to $\mathrm{dir}(\widehat{\mathcal{C}})$ on the edges $\widehat{\mathcal{C}}$ already directs and, for the retraction-only search, forced to 0 on every $e \notin \mathcal{K}_{G_0} \cup \mathrm{dir}(\widehat{\mathcal{C}})$; for $e \in \mathcal{K}_{G_0}$ a paired Boolean $\overline{s}_e = 1 - s_e$ marks retraction. A Boolean und_e per skeleton edge records "present and left undirected," and an integer $\mathrm{pos}_v \in \{0, \dots, n-1\}$ per node orders the state graph. The witness DAG reuses one Boolean $e_{u,v}$ per ordered adjacent pair ($O(n^2)$), an integer epos_v per node, a Boolean rx_v marking descent from X in D, Booleans ancy_v, cn_v and forb_v marking that v is an ancestor of Y, a causal node (rx_v and $v = Y$ or ancy_v) and in the forbidden set (a causal node or a descendant of one), and a Boolean ancz_v marking that $v \in Z$ or v has a descendant in Z within D', the proper back-door graph (D with each edge $X \to v$ for which cn_v deleted) – each the target of a topological fixpoint with $O(n^2)$ auxiliary support literals – plus an explicit bounded simple path: position variables $q_{i,v}$ for $i = 0, \dots, n-1$ ($O(n^2)$) and activity flags act_i ($O(n)$), and mode flags $\mathrm{mode}_a, \mathrm{mode}_b$ combined into a top-level literal fails. In total $O(n^2)$ variables. The R3/R4 Meek-rule clauses and the collider/non-collider clauses at each path position are each indexed by four vertices, giving $O(n^4)$ constraints overall; the remaining families are $O(n^2)$ or $O(n^3)$.

Constraints. (i) $F = \{e : s_e\}$ *is Meek-closed, where* $F = \mathrm{dir}(\widehat{\mathcal{C}}) \cup S$ *for* $S = F \setminus \mathrm{dir}(\widehat{\mathcal{C}}) \subseteq \mathcal{K}_{G_0}$. Each rule instance is a clause forbidding its firing configuration: R1 forbids $s_{a,b} \wedge \mathrm{und}_{b,c}$ for a, c non-adjacent; R2 forbids $s_{a,c} \wedge s_{c,b} \wedge \mathrm{und}_{a,b}$; R3 forbids $\mathrm{und}_{a,b} \wedge \mathrm{und}_{a,c} \wedge \mathrm{und}_{a,e} \wedge s_{c,b} \wedge s_{e,b}$ for non-adjacent c, e; R4 forbids $\mathrm{und}_{a,b} \wedge \mathrm{und}_{a,c} \wedge s_{c,w} \wedge s_{w,b}$ for non-adjacent c, b, exactly as encoded (this clause does not additionally require a, w adjacent, as Meek's original R4 does. This is not a weakening: in a consistent state, block (ii) supplies an extension D; if some extension had $b \to a$, then $a \to c$ would give the cycle $a \to c \to w \to b \to a$, and $c \to a$ would give the v-structure $c \to a \leftarrow b$ (c, b non-adjacent) absent from $\widehat{\mathcal{C}}$; so $a \to b$ holds in every extension, is compelled given S, and by completeness of Meek's rules (Meek, 1995) $a - b$ cannot remain undirected in a closed consistent state. The clause as encoded therefore excludes nothing legitimate, even without the a, w-adjacency condition of Meek's original R4). A further clause forbids any v-structure $a \to b \leftarrow c$ (a, c non-adjacent) absent from $\widehat{\mathcal{C}}$, and $\mathrm{pos}_u < \mathrm{pos}_v$ is enforced whenever $s_{u,v}$, ruling out directed cycles; together these force every satisfying F to be the closure orientations of a genuine knowledge state, hence consistent. (ii) D *is a consistent extension of* $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$: every skeleton edge of D is oriented exactly one way; $s_{u,v} \Rightarrow e_{u,v}$ pins D to contain $\mathrm{dir}(\widehat{\mathcal{C}}) \cup S$; $\mathrm{epos}_u < \mathrm{epos}_v$ whenever $e_{u,v}$ rules out cycles; and the v-structure clause of (i), applied to e, rules out v-structures absent from $\widehat{\mathcal{C}}$. (iii) Z *is invalid in* D under the generalized adjustment criterion, matching `bkrobust.gac.dag_level.is_gac_valid_dag`: mode_a is the disjunction over $z \in Z$ of forb_z, where $\mathrm{rx}_v = \bigvee_{u \to v \in D} (u{=}X \vee \mathrm{rx}_u)$ and the other indicators are evaluated in topological order (safe because D is already forced acyclic); mode_b asserts a simple X-to-Y path of length $\leq n$ in D', position-indexed by q, act, that is d-connecting given Z, with collider/non-collider conditions at each internal position keyed on ancz; the model asserts $\mathrm{mode}_a \vee \mathrm{mode}_b$ and fixes $\mathrm{fails} = 1$. Given $Z \cap \mathrm{forb} = \varnothing$, Z blocks every proper non-causal path in D if and only if it d-separates X and Y in D' (Perković et al., 2018), so (iii) is exactly the failure of the criterion. (Definition 3 takes $Z \subseteq \mathcal{V} \setminus \{X, Y\}$, so the reference definition's "$X \in Z$ or $Y \in Z$" disjunct is vacuous and is not encoded.) The d-connecting path is encoded explicitly, not as a Boolean fixpoint, because the Bayes-ball state graph has cycles: an unguarded fixpoint over-reports failure and a level-guarded one under-reports it – silently weakening the ladder's UNSAT certificates – whereas every d-connecting path can be taken simple, so bounding its length by n loses nothing. (iv) *Cardinality.* $|\mathcal{K}_{G_0} \setminus S| \leq k$ is posted, once per rung, as a single native linear constraint $\sum_{e \in \mathcal{K}_{G_0}} \overline{s}_e \leq k$ on a clone of the base model – there is no hand-built sequential-counter, totalizer, or PySAT cardinality network. The whole model is solved by Google OR-Tools' CP-SAT (a lazy-clause-generation CP/SAT solver), single-threaded and seeded for determinism.

Validation of the criterion and of the reported radii. The survival oracle and the radius use one implementation of the generalized adjustment criterion on MPDAGs (Perković et al., 2018), i.e. the b-adjustment criterion for MPDAGs, stated with b-possibly causal paths (Perković et al., 2017, Thm. 4.4), validated semantically – by agreement with enumeration of DAG extensions, not by trusting the implementation – on every case of the exhaustive three-to-four node scope (74,568 cases, 28,464 non-amenable) and on a 2.4M-case five-node sample (`results/axisa2/gac_agreement.json`). The E1 encoding agrees with the brute-force radius on all 17,976 queries on three and four nodes, for optimal and arbitrary valid sets (`results/axisa2/sat_gac_validation.json`). An earlier version of the pipeline decided validity with Pearl's back-door criterion (Pearl, 2009), which is sufficient but not necessary: on the CPDAG $W - X - Y$ with $G_0 : W \to X \to Y$ and the valid, non-optimal $Z = \{W\}$, it gives radius 1 where the generalized criterion gives 2. For the committed optimal set the two coincide on every small graph we checked (5,304 queries on three and four nodes, 58,414 on a five-node sample; `results/axisa2/oset_radius_agreement.json`), and every reported radius was recomputed under the generalized criterion: all 4,104 synthetic, 480 uniform-query, 3,027 elicited and 99 effect-scale units agree exactly, as does the running example (`results/radius_gac_recompute/`). Only the radius oracle changed: every corrupted-state validity label and hence every survival AUC was computed with the generalized criterion from the start (`bkrobust.robustness.survival`, `real_survival`, `run_unconfounded_v2`), so no label, AUC or table entry changes. The back-door criterion survives only in the instance-admission gates of the two synthetic generators. The uniform-query gate applies it in the fully oriented true DAG to the optimal and the empty set, neither of which contains a descendant of X, so there the two criteria coincide; its perturbation check already uses the generalized criterion. The gate of the main synthetic generator requires that some atomic perturbation of the true knowledge invalidate some valid set, and decides this with the back-door criterion over all back-door-valid sets. Both are selection rules: they determine which instances enter a sweep, not any radius, label or AUC reported on them.

Lemma 8 (Soundness and completeness of the E1 rung)**.** *Fix k. The rung-k model is satisfiable if and only if there is a Meek-closed $S \subseteq \mathcal{K}_{G_0}$ with $|\mathcal{K}_{G_0} \setminus S| \leq k$ and Z invalid in* $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$.

Proof idea. Because validity is universal over $[G]$ (Section 3), Z is invalid in $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$ iff some $D \in [\mathrm{Meek}(\widehat{\mathcal{C}}, S)]$ has Z invalid in D, and $[\mathrm{Meek}(\widehat{\mathcal{C}}, S)]$ is exactly the DAGs sharing $\widehat{\mathcal{C}}$'s skeleton, containing $\mathrm{dir}(\widehat{\mathcal{C}}) \cup S$, acyclic, with no v-structure beyond $\widehat{\mathcal{C}}$'s (Meek, 1995; Perković et al., 2017) – precisely constraint block (ii). Given such S, D (so $F = \mathrm{dir}(\widehat{\mathcal{C}}) \cup S$ is the directed-edge set of $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$), setting $s_e = \mathbb{1}[e \in F]$ satisfies (i) because F is Meek-closed, being the closure set of a real MPDAG, setting $e_{u,v} = \mathbb{1}[u \to v \in D]$ satisfies (ii), D's invalidity witnesses (iii), and the cardinality bound is the hypothesis. Conversely, a satisfying assignment gives $F = \{e : s_e\}$ and, by the domain restriction on s_e, $S = F \setminus \mathrm{dir}(\widehat{\mathcal{C}}) \subseteq \mathcal{K}_{G_0}$; with $D = \{e_{u,v}\}$, D is acyclic, shares $\widehat{\mathcal{C}}$'s skeleton, contains $F = \mathrm{dir}(\widehat{\mathcal{C}}) \cup S$, and has no v-structure beyond $\widehat{\mathcal{C}}$'s – block (ii) – so D is a consistent extension of $\widehat{\mathcal{C}}$ and F is consistent. Block (i) makes F closed under R1–R4, so S is the corresponding closure set relative to $\widehat{\mathcal{C}}$. By soundness and completeness of Meek's rules for consistent background knowledge (Meek, 1995), $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$ orients exactly S beyond $\widehat{\mathcal{C}}$, so $S = \mathcal{K}_{\mathrm{Meek}(\widehat{\mathcal{C}}, S)}$ and $\mathrm{Meek}(\widehat{\mathcal{C}}, S)$ is a genuine knowledge state (the exhaustive check of the closure clauses against the closed-knowledge-state space, `bkrobust.sat.closure.solutions`, corroborates this). Then $D \in [\mathrm{Meek}(\widehat{\mathcal{C}}, S)]$ satisfies the reference invalidity predicate by construction of (iii), and $|\mathcal{K}_{G_0} \setminus S| \leq k$ by (iv). □

Taking the least SAT rung over $k = 0, 1, \ldots$ therefore returns $r_{E1} = \min\{|\mathcal{K}_{G_0} \setminus S| : S \subseteq \mathcal{K}_{G_0} \text{ closed}, \ Z \text{ invalid in } \mathrm{Meek}(\widehat{\mathcal{C}}, S)\}$, which equals r_{val} by Theorem 1. This realises Corollary 1(c) through an explicit closed intermediate state S rather than a bare search over $D \in [\widehat{\mathcal{C}}]$ with $|\mathcal{K}_{G_0} \setminus \mathrm{dir}(D)| \leq k$; E2 (`bkrobust.sat.e2`) keeps the same decomposition – it reuses the structural constraints (i), (ii) and the failure block (iii) unchanged, but not E1's retraction-only domain restriction (*Variables*, above: s_e forced to 0 outside $\mathcal{K}_{G_0} \cup \mathrm{dir}(\widehat{\mathcal{C}})$); that restriction is lifted, so S ranges over every subset of the full skeleton consistent with block (i) rather than only over $S \subseteq \mathcal{K}_{G_0}$, i.e. E2 searches every closed state F, not only retractions of $\mathcal{K}_{G_0}$, and only the objective

changes, from (iv) to the join-distance sum of Lemma 2 – so neither released encoding is literally the bare-D formulation of Corollary 1(c), though both are equivalent to it by the argument above.

Statuses. The released `bkrobust.hybrid.breakdown_radius` reports each outcome of Algorithm 1 as a named method rather than a bare number: `degenerate` for INVALIDINPUT (Z invalid in G_0), `top_state` for $r_{\mathrm{val}} = \infty$, `local_up_fast` and `e1_ladder` for a finite radius found by the search or the SAT stage, and `exact=False` for a timeout. The experiments additionally record, before the radius is computed, whether the knowledge was inconsistent and whether an adjustment set exists (Appendix H).

The ladder. Algorithm 1 calls the top-state check of Corollary 1(b) once, then the bounded retraction search for $k = 1, \dots, b$, and only then dispatches one rung-k solver call for each $k = b+1, \dots, |\mathcal{K}_{G_0}| - 1$ in increasing order, stopping at the first SAT; `bkrobust.sat.e1` implements the same ladder standalone from $k = 0$, with a per-rung time limit (default 60 s). A run is *exact* when every rung up to the returned radius resolves to SAT or UNSAT within its limit (the rung $|\mathcal{K}_{G_0}|$ is always satisfiable once Z fails in $\widehat{\mathcal{C}}$); a rung that times out (CP-SAT returns UNKNOWN) instead reports $r_{\mathrm{val}} \geq k$ with `exact=False`, a partial bound rather than a certified value.

The encodings agree with exhaustive enumeration wherever enumeration is feasible. The closure clauses reproduce the enumerated space on all 133 CPDAGs with an undirected edge on three and four nodes (1,588 elements), and the radii equal the brute-force radius on all 1,044 exhaustive instances on four nodes and on all 346 generated instances up to twenty nodes for which a brute-force radius is computable.

Table 5 gives the per-k medians behind Figure 5.

Table 5: Legacy timings (earlier back-door implementation; not a benchmark of the final implementation). Median seconds per query and median speed-up of the retraction search (with the top-state check) over building the space, by the number k of undirected edges, on the family of Figure 5. Search and hybrid both run the top-state check first. Ratios are medians over instances on which both methods answer within 120 s. For $k \geq 11$ building the space exceeds 120 s on every instance, so no ratio is given.

k	instances	space	search	hybrid	space/search
1	32	0.0006	0.00015	0.00013	5.2
2	43	0.0015	0.00019	0.00019	12.5
3	35	0.0070	0.00020	0.00021	38.6
4	35	0.020	0.00022	0.00022	137
5	24	0.069	0.00019	0.00018	447
6	21	0.186	0.00030	0.00035	1,196
7	12	1.00	0.00029	0.00030	3,344
8	19	4.35	0.00042	0.00045	9,559
9	7	23.7	0.00032	0.00033	81,870
10	8	53.9[†]	0.00077	0.00073	62,190
11–26	20	> 120	0.00045–1.51	0.00041–2.04	–

[†]Median over the 6 of 8 instances on which building the space takes under 120 s.

E THE BIAS RADIUS IN FULL

The worst-case bias radius r_ε of Section 7 is developed here with its proofs and experiments, followed by the coverage and the full table of the mean-bias profile r_μ of Section 7. Assumption 1 is in force throughout this appendix and nowhere before it: every statement in Appendix C.4 is order-theoretic and holds for any observational law compatible with $\widehat{\mathcal{C}}$.

E.1 ZERO BIAS ON THE CERTIFIED BALL, AND MONOTONICITY

The bias vanishes identically on the region that r_{val} already certifies. Recall from Appendix B that the *closed ball* of radius k about G_0 is $\{G \in \mathfrak{G}_{\widehat{\mathcal{C}}} : d(G_0, G) \leq k\}$. The certified region is the ball of

radius $r_{\mathrm{val}} - 1$. It contains G_0 and every state within distance $r_{\mathrm{val}} - 1$ of it, and since r_{val} is defined as a minimum, no state in it makes Z invalid.

Theorem 2 states that, under Assumption 1, if $d(G_0, G) \leq r_{\mathrm{val}} - 1$ then $T(G) = \{\theta_Z\}$, and hence $B(G) = 0$; equivalently, $\sup\{B(G) : d(G_0, G) \leq r_{\mathrm{val}} - 1\} = 0$.

The proof is in Appendix E.3. The bias on this ball is algebraically zero at the population level, for every law that satisfies Assumption 1 and is compatible with the equivalence class. If moreover the true DAG lies in $[\widehat{\mathcal{C}}]$ and contradicts fewer than r_{val} closure orientations of G_0, θ_Z equals the true effect (Theorem 2); this is a population identity and does not by itself validate confidence intervals after data-dependent selection of the graph and the adjustment set. Every state in the ball yields the same estimand. The converse can fail only in the conservative direction. At distance r_{val} some state admits a graph in which Z is invalid, yet $\beta(Z; \Sigma)$ may still equal τ_D for the particular Σ at hand. Hence r_{val} is a lower bound on every bias radius defined below.

Lemma 9 (Monotonicity of worst-case bias)**.** *If $[G] \subseteq [H]$ then $T(G) \subseteq T(H)$ and $B(G) \leq B(H)$. Consequently $F_\varepsilon = \{G : B(G) > \varepsilon\}$ is upward-closed, and $F_{\varepsilon'} \subseteq F_\varepsilon$ whenever $\varepsilon \leq \varepsilon'$.*

The proof is in Appendix C.4; it uses only the fact that $T(G)$ is indexed by $[G]$.

Lemma 9 is monotonicity along the order. Monotonicity along the distance is a different and weaker property, and it does not hold for B: two states at the same distance are typically incomparable, and a state one shell further out can carry strictly less bias than a nearer one (Appendix E.4 gives a minimal instance). The worst case over the ball is monotone in the distance, and a certificate must be built from it, because a guarantee has to cover whichever wrong claims the analyst actually holds.

E.2 The radius and its computation

Writing $\beta^{\uparrow}(d)$ for the largest B over the ball of radius d about G_0, define

$$r_\varepsilon \;=\; \min\{\, d(G_0, G) \;:\; B(G) > \varepsilon|\theta_Z| \,\} \;=\; \min\{\, d : \beta^{\uparrow}(d) > \varepsilon|\theta_Z| \,\}. \tag{6}$$

As in equation 5, queries with $\theta_Z = 0$ are excluded, since the relative tolerance is then undefined and unstable when $|\theta_Z|$ is small; we report the ceiling $B(\widehat{\mathcal{C}})$ alongside r_ε for this reason, as $B_{\mathrm{rel}}(\widehat{\mathcal{C}}) = B(\widehat{\mathcal{C}})/|\theta_Z|$ in the tables. The reduction that makes r_{val} computable applies verbatim. The proof of Theorem 1 uses one property of failure, that it is upward-closed; adjustment sets and d-separation do not enter. It is therefore a statement about arbitrary up-sets, and Lemma 9 places $F_{\varepsilon|\theta_Z|}$ among them. The nearest ε-violating state is again reachable by retraction alone, and the search is again confined to withdrawals of the analyst's own commitments. The up-set of G_0 is parameterised by the subsets of $\mathcal{K}_{G_0}$, and shell d is generated directly from the d-subsets, so a shell can be built without traversing any shell beneath it.

By Theorem 2 no state in the ball of radius $r_{\mathrm{val}} - 1$ needs to be evaluated, and since, restricted to the retraction up-set ($0 \leq k \leq |\mathcal{K}_{G_0}|$), the maximum over a ball is attained on its outermost shell, interior states need no evaluation either: every state at retraction depth $j < k$ lies below some state at depth k (it can be extended by further retractions), and B is monotone (upward) in the order (Lemma 9), so the maximum over depths at most k is attained at depth k. A single traversal answers every ε, including a threshold chosen afterwards, because each radius is then a lookup on a monotone staircase, and monotonicity licenses bisection, which locates a distant first crossing in a logarithmic number of shells. This shortcut is licensed by Theorem 2, a population identity: $B(G) = 0$ on the certified ball holds for the population covariance Σ, but not in general for an unrestricted empirical covariance $\hat{\Sigma}$, which need not satisfy the conditional independences of G_0. The *Finite samples* paragraph below applies the same shortcut to $\hat{\Sigma}$ and states what that does and does not certify. Evaluating B requires no graph enumeration. The ambiguity set is determined by the possible parent sets of the treatment, which are obtained by orienting the undirected edges at X each way and keeping the assignments whose Meek closure is consistent. This costs $2^{\deg(X)}$ polynomial closures, where $\deg(X)$ counts undirected edges at the treatment alone. Two-sided bounds come without an exhaustive pass. r_{val} and any radius at a smaller ε bound r_ε from below at no cost; a single greedy retraction chain that reaches a violating state bounds it from above with an explicit witness state; and one evaluation at $\widehat{\mathcal{C}}$, the maximum of the order and hence of B, settles $r_\varepsilon = \infty$ for every threshold above it. These costs hold in the linear-Gaussian regime of Assumption 1, where each τ_D is a linear solve.

Maximising B over the ball gives a certificate in effect units.

Proposition 2 (Effect-scale certificate)**.** *Under Assumption 1: (a) if at most k of the orientations of $\mathcal{K}_{G_0}$ are false of the data-generating DAG, then $|\theta_Z - \tau^*| \le \beta^\uparrow(k)$; in particular $\theta_Z = \tau^*$ whenever $k < r_{\mathrm{val}}$, the error is at most $\varepsilon|\theta_Z|$ when $k < r_\varepsilon$, and when $r_{\varepsilon_x} \le k < r_{\varepsilon_y}$ the worst case lies in $(\varepsilon_x|\theta_Z|, \varepsilon_y|\theta_Z|]$; (b) for every k there is a graph $D \in [\widehat{\mathcal{C}}]$ that reproduces the observed covariance matrix Σ, contradicts at most k of the analyst's orientations, and realises $|\theta_Z - \tau_D| = \beta^\uparrow(k)$.*

The proofs are in Appendix E.3. For part (a), withdrawing exactly the false orientations gives a state that still represents the data-generating DAG, so that DAG lies in the ambiguity set of this state and the bound applies to it. The budget is counted in orientations of the closure, the same unit as r_{val}. Meek's rules cascade, so a few wrong assertions can produce a larger number of wrong closure orientations, and the budget k must be read in closure orientations.

By part (b) the bound is tight: it is attained by a graph the analyst has no evidence against.

For part (b), take a state at distance at most k attaining $\beta^\uparrow(k)$ and, inside it, a graph attaining the maximum. That graph is Markov equivalent to the data-generating DAG, so some linear-Gaussian parameterisation of it reproduces Σ, and it withholds at most k of $\mathcal{K}_{G_0}$ by the parameterisation of the up-set. Under Assumption 1, Markov equivalent graphs impose the same covariance constraints, so the witness cannot be ruled out on the evidence the analyst holds. The band is therefore as tight as the CPDAG, the data and the error budget allow, and refining the ε grid refines it further.

E.3 Proofs for the bias radius

Proof of Theorem 2. Let $d(G_0, G) \le r_{\mathrm{val}} - 1$. By the definition of r_{val} as a minimum (Definition 3), no state at distance at most $r_{\mathrm{val}} - 1$ fails, so Z is a valid adjustment set in G. Validity in an MPDAG is a universal statement over the DAGs it represents, so Z is a valid adjustment set in every $D \in [G]$. Fix such a D. Under Assumption 1 the law is generated by a linear-Gaussian SCM and $\Sigma \succ 0$, so $\beta(\cdot; \Sigma)$ is well defined and a valid adjustment set identifies the total effect through the population least-squares coefficient: $\beta(Z; \Sigma) = \tau_D$. This is the only step of the argument that uses linearity, and positive definiteness makes the coefficient unique. Hence $\tau_D = \theta_Z$ for every $D \in [G]$, so $T(G) = \{\theta_Z\}$ and $B(G) = \max |\theta_Z - \theta_Z| = 0$. As this holds at every such G, including G_0 itself, the supremum over the closed ball is 0. □

Proof of Proposition 2(a). Let D^* be the data-generating DAG, $\tau^* = \tau_{D^*}$, and let $S \subseteq \mathcal{K}_{G_0}$ be the set of orientations of $\mathcal{K}_{G_0}$ that D^* violates, so $|S| \le k$ by hypothesis. Put $H = \mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K}_{G_0} \setminus S)$. Then $D^* \in [\widehat{\mathcal{C}}]$ satisfies every orientation in $\mathcal{K}_{G_0} \setminus S$, so $D^* \in [H]$ by Meek's theorem, and therefore $\tau^* \in T(H)$. By the parameterisation of the up-set of G_0 (Theorem 1 and Lemma 4), $H \succeq G_0$ with $d(G_0, H) = |S| \le k$. Combining,

$$|\theta_Z - \tau^*| \;\le\; \max\{|\theta_Z - \tau| : \tau \in T(H)\} \;=\; B(H) \;\le\; \beta^\uparrow(k),$$

the last step because H lies in the ball of radius k over which $\beta^\uparrow(k)$ is a maximum. The three specialisations follow. If $k < r_{\mathrm{val}}$ then H lies in the ball of radius $r_{\mathrm{val}} - 1$, so $B(H) = 0$ by Theorem 2 and $\theta_Z = \tau^*$. If $k < r_\varepsilon$ then $\beta^\uparrow(k) \le \varepsilon|\theta_Z|$ by equation 6, since r_ε is the least d at which $\beta^\uparrow$ exceeds $\varepsilon|\theta_Z|$. If $r_{\varepsilon_x} \le k < r_{\varepsilon_y}$ then $\beta^\uparrow(k) \le \varepsilon_y|\theta_Z|$ by the same argument, while $\beta^\uparrow(r_{\varepsilon_x}) > \varepsilon_x|\theta_Z|$ and $\beta^\uparrow$ is non-decreasing, so $\beta^\uparrow(k) > \varepsilon_x|\theta_Z|$. □

Proof of Proposition 2(b). Fix k. The maximum defining $\beta^\uparrow(k)$ is over a finite set, so it is attained: choose H with $d(G_0, H) \le k$ and $B(H) = \beta^\uparrow(k)$, and by Theorem 1 H may be taken comparable to G_0, so $\mathcal{K}_H \subseteq \mathcal{K}_{G_0}$ with $|\mathcal{K}_{G_0}| - |\mathcal{K}_H| = d(G_0, H) \le k$. Choose $D \in [H]$ attaining $|\theta_Z - \tau_D| = B(H)$, which again exists because $T(H)$ is finite. Then $D \in [H] \subseteq [\widehat{\mathcal{C}}]$, so D is Markov equivalent to the data-generating DAG; under Assumption 1 Markov equivalent DAGs impose the same constraints on a jointly Gaussian law, so some linear-Gaussian parameterisation of D has population covariance Σ. Finally D satisfies every orientation of $\mathcal{K}_H$, so the orientations of $\mathcal{K}_{G_0}$ it can contradict lie in $\mathcal{K}_{G_0} \setminus \mathcal{K}_H$, of which there are at most k. □

Remark 6 (What the witness is). Part (b) asserts that the witness D reproduces the observed covariance matrix and respects all but at most k of the analyst's claims. It makes no claim that D is the

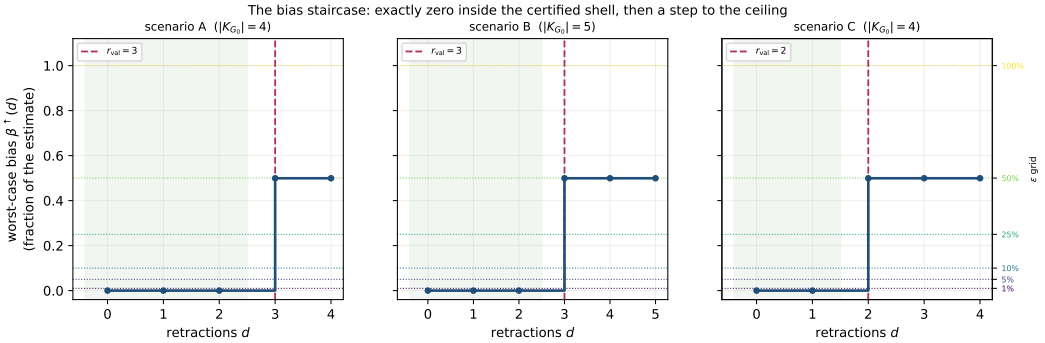

Figure 6: The bias staircase on the running example, one panel per knowledge state. The worst-case bias is zero for $d < r_{\mathrm{val}}$ (shaded), then steps to the ceiling; the dotted rules are the ε grid, so each radius can be read off the plot. Table 7 reports a sampled mean over drawn coefficients, which rises gradually; the curve here is the supremum.

data-generating DAG. This is the sense in which the analyst cannot rule D out on the evidence they hold, and the band is tight relative to the CPDAG, the data and the error budget.

E.4 Evaluation of the worst-case radius

The construction under direct computation. Seven consequences of Appendices E.1 and E.2 can be checked by direct computation, and we checked them on 400 instances, each with three independent parameter draws. All 436,798 comparisons below agree at an absolute tolerance of 10^{-9}.

consequence	comparisons
$B = 0$ on every state of the certified ball	1,758
B monotone over every ordered pair of the enumerated space	372,099
staircase non-decreasing and attained on its outermost shell (retraction up-set, $0 \le k \le \lvert\mathcal{K}_{G_0}\rvert$)	17,160
incremental, bisection and brute-force radii coincide	5,844
r_ε non-decreasing in ε and never below r_{val}	5,844
realised error within the bound of Proposition 2(a), budget in closure orientations	22,680
ambiguity set from the parent sets of X equals the enumerated one	11,413

The staircase on the running example. Figure 6 gives the exact worst case in place of the sampled-mean column of Appendix F. The mean rises shell by shell. The worst case, at the reported parameter draw, is flat for $d < r_{\mathrm{val}}$, then jumps to $B(\widehat{\mathcal{C}})$, the largest bias any perturbation can produce, and stays there. What rises gradually with depth is the proportion of states at which Z fails, while the size of the worst failure is fixed from r_{val} on (Appendix F reports the shape over further coefficient draws). The radii computed over the retraction up-set agree with a breadth-first search over the whole enumerated space in all three knowledge states.

Depths beyond $\lvert\mathcal{K}_{G_0}\rvert$ add no information: $d(G_0, \widehat{\mathcal{C}}) = \lvert\mathcal{K}_{G_0}\rvert$ and $\widehat{\mathcal{C}}$ maximises B, so $\beta^{\uparrow}(k) = B(\widehat{\mathcal{C}})$ for every $k \ge \lvert\mathcal{K}_{G_0}\rvert$. When $r_\varepsilon = r_{\mathrm{val}}$, the bias radius still adds information, since it supplies the magnitude that r_{val} leaves open. The validity radius says that the third revision breaks the set, and the band says what breaking it costs.

Why a supremum over the ball. The construction maximises over a ball and uses a supremum over the ambiguity set. The two instances below show that both choices are needed.

First, bias is non-monotone in distance. At four nodes, the smallest size at which the question is non-trivial and where 355 witnesses exist, one instance has a state at distance one carrying bias 1.34 and a state at distance two carrying none. The two states are incomparable in the order, so Lemma 9 is untouched, and a certificate read off a single perturbation path would be wrong on this instance.

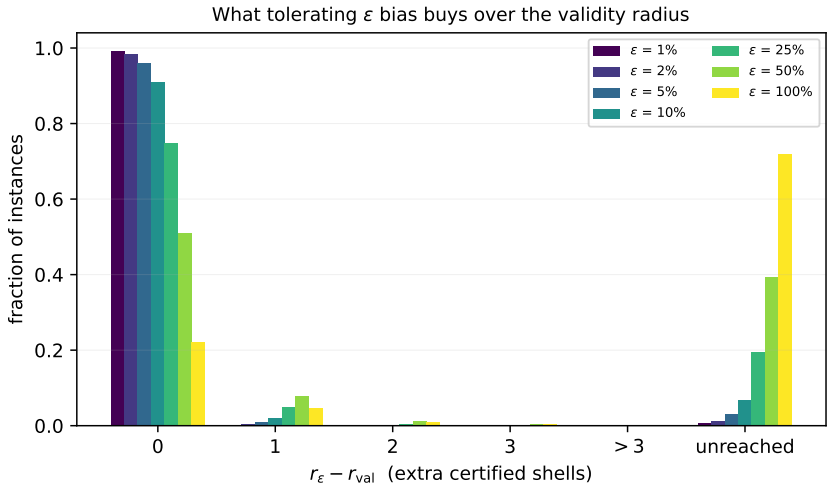

Figure 7: Distribution of $r_\varepsilon - r_{\mathrm{val}}$ over 900 instances. As ε grows, the bias radius extends past r_{val} by about one shell on a growing share of instances, and at $\varepsilon = 100\%$ the threshold is never reached on 70.1% of them, a ceiling that r_{val} cannot express.

For this reason $\beta^{\uparrow}$ maximises over the ball, and a maximum tabulated shell by shell need not rise with d (Remark 5).

Second, a sampled mean absolute bias in place of the supremum would break the search. On identical pairs, the mean is non-monotone under the order on at least 34.7% of comparable pairs (278 of 801). A pair counts only when the deficit exceeds three combined Monte-Carlo standard errors in each of two independent repetitions, so this rate is a lower bound, and the median margin is 9.6 standard errors. The worst case defined here is monotone on every one of these pairs. Because $\{\text{mean} > \varepsilon\}$ is not an up-set, the retraction search does not apply to it. A mean-based distance disagrees with brute force over the whole space on 269 of 644 thresholds, in the inspected cases by missing a crossing at distance two. On a separate matched subsample of 50 instances, the mean-based and worst-case distances differ on at least one threshold on 42, in both directions: 23 larger, 14 smaller and 5 mixed, with 8 equal. There the mean is non-monotone under the order on 49 of the 50, while the supremum is monotone by Lemma 9.

How this bears on r_μ. The mean-bias profile r_μ of Section 7 is also built from a mean, and neither difficulty arises for it. It averages the exact worst case B over each shell, with no coefficient sampling, and it is computed on the full space $\mathfrak{G}_{\widehat{\mathcal{C}}}$, or, when that space is too large, on the states restricted to X's undirected component, without the retraction search; together these cover 62 of the 66 queries of Appendix E.5 (48 by whole-space enumeration, 14 by component-restricted enumeration).

What the refinement adds, over 900 instances. The experiment covers 900 instances over five generators, node counts 6 to 12, three knowledge coverages and four corruption regimes. The main contribution of the bias radius is a ceiling. The fraction of instances on which the threshold is never reached grows from 1.4% at $\varepsilon = 1\%$ to 70.1% at $\varepsilon = 100\%$ (Figure 7). On seven instances in ten, no perturbation of the analyst's knowledge, however large, can move the estimate by its own magnitude. r_{val} cannot make this statement, and it comes from the same traversal. The radius itself also extends: r_ε exceeds r_{val} on 1.0% of instances at $\varepsilon = 1\%$, on 24.5% at 25% and on 75.6% at 100%, typically by one shell (mean gain 1.00 to 1.14).

Three further observations hold on all 900 instances. Empirically, $r_0 = r_{\mathrm{val}}$ on every one; the converse to Theorem 2 allows a coincidental cancellation at the breakdown radius for non-generic parameters, and none is observed. The staircase is a step in 84.1% of instances and a ramp in 15.1%, so the worst case usually saturates as soon as validity fails, and the gradual rise of a sampled mean mostly reflects a growing fraction of failing states within each shell. The bound of Proposition 2(a) holds on all 900 instances. Among instances with a non-degenerate bound, the ratio of realised error to the bound has median 1.000 and maximum 1.000: Proposition 2(b) guarantees only that

some DAG attains the bound, not that the actual data-generating DAG typically does, so this near-universal attainment is an empirical observation rather than a consequence of the proposition.

Incremental traversal and bisection return the same radius on every instance; incremental traversal evaluates a median of one state (mean 2.09, maximum 65), and the greedy retraction chain alone returns the exact radius on 98.4% of instances.

Finite samples. r_{val} takes no numerical input and has no sampling distribution given its graphical inputs; when $\widehat{\mathcal{C}}$ is instead estimated from data, r_{val} inherits the sampling variability of $\widehat{\mathcal{C}}$. r_ε depends on a covariance and has one. The plug-in radius (reported below in graphically truncated form) substitutes, in equation 6, the unbiased sample covariance $\hat{\Sigma}$ for Σ, without projecting it onto the Gaussian model of $\widehat{\mathcal{C}}$, and reads every downstream quantity, including $\hat{\theta}_Z = \beta(Z;\hat{\Sigma})$, off it. The identity $B = 0$ on the certified ball (Theorem 2) is a population statement and need not hold for $\hat{\Sigma}$: for $G_0\colon W \to X \to Y$ with $Z = \varnothing$ (so $r_{\mathrm{val}} = 2$) and, in node order (W, X, Y), $\hat{\Sigma} = \left(\begin{smallmatrix} 1 & .5 & .4 \\ .5 & 1 & .5 \\ .4 & .5 & 1 \end{smallmatrix}\right)$, $\hat{\theta}_Z = 0.5$ while $\hat{\tau}_{G_0} = \beta(\{W\};\hat{\Sigma}) = 0.4$, so $\hat{B}(G_0) = 0.1$ already exceeds a 10% relative tolerance and the raw plug-in radius there is 0. The implementation nonetheless skips the certified ball, as in the population case, so the quantity we report is a *graphically truncated plug-in radius*: the population construction – zero inside the ball – evaluated with $\hat{\Sigma}$ outside it. Because B is monotone in the order for any covariance (Lemma 9), skipping changes no value of the staircase at depths $\geq r_{\mathrm{val}}$; it only removes crossings below r_{val}, so the reported radius is never smaller than the raw plug-in radius, that is, it is less conservative than the raw plug-in.

We evaluated the plug-in construction on 120 instances at sample sizes from 100 to 10,000. It exceeds the population radius in at most 5.0% of cases at $n = 100$ and in 0 to 0.8% at $n = 10{,}000$. Both radii are at least r_{val} and the shortcut acts only below r_{val}, so these exceedances arise from sampling noise at depths $\geq r_{\mathrm{val}}$. A nonparametric bootstrap resamples the observations with replacement and recomputes $\hat{\Sigma}$ and the whole staircase (200 replicates), holding r_{val} fixed and judging every replicate against the absolute threshold $\varepsilon|\hat{\theta}_Z|$ fixed once from the observed data. Reporting the 5% quantile of the resampled radius brings the exceedance rate to 0.0% in every cell, while understating the radius on 1 to 17% of instances. This is a single quantile of the radius, not a pointwise or simultaneous band over states or depths, and we claim no finite-sample coverage guarantee: the rates above are empirical, on this corpus. This bootstrap variant was empirically conservative on the evaluated instances; it is not a certified lower confidence bound for the population radius.[3]

On real graphs. The same computation runs on 32 published networks, each with a background knowledge set elicited from a language model, on the first five (X, Y) queries per network, 160 instances in all. We report the 66 instances, from 21 of the 32 networks, on which G_0 commits a valid adjustment set and the computation completes within 120 seconds per query. Of the other 94, 27 admit no identified adjustment set at G_0, on 33 the optimal-set construction does not apply, so no set is committed, and 34 lie in the largest networks (the `munin` family, `link`, `diabetes`, `arth150` and `andes`). Of the 66, 13 have no failure anywhere in the enumerated space, including every query on `win95pts`, and 53 have a finite radius.

On 46 of these 53 the radius is 1: a single retraction of a closure orientation already invalidates the committed adjustment set (Remark 1), and the witness returned with the radius names an orientation whose retraction breaks Z, a candidate for the analyst to check first; other single orientations may do the same. The ceiling of the bias radius then grades these analyses. On 31 of the 53, the tolerance $\varepsilon = 100\%$ is never crossed, so an opposite-sign effect is excluded against every retraction, though zero may remain compatible. On the other 22, $r_\varepsilon = r_{\mathrm{val}}$ at every tolerance. r_{val} is purely graph-theoretic and uses only the CPDAG and the asserted knowledge. The r_ε values are in addition conditional on a linear-Gaussian structural equation model attached to the network's DAG, with one fixed set of coefficients (Assumption 1).

[3] Implementation: `bkrobust.epsilon` (modules `finite_sample` and `profile`) in the released code.

E.5 THE MEAN-BIAS PROFILE: COVERAGE AND THE FULL TABLE

How to read r_μ. Three properties go with the profile of Section 7. It is reported beside r_{val} and certifies nothing: μ is an average and bounds nothing, since $\mu(d) \le \beta^{\uparrow}(d)$ and every guarantee is stated against $\beta^{\uparrow}$. It marks an onset, the first distance at which the average exceeds ε. μ need not be monotone in d, because far shells contain heavily oriented states that pin the effect down and pull the average back. On `mediator` $I \to Y$, μ rises to 4.57 at $d = 4$ and falls to 1.00 by $d = 7$; over the 49 instances, μ is non-monotone on 9, and 16 of the 271 finite cells record a crossing that does not persist to the outermost shell. Finally it is conditional on Assumption 1: every r_μ and $B(\widehat{\mathcal{C}})$ is computed against a linear-Gaussian model attached to the network's DAG. UNR marks a tolerance the mean profile never crosses, ε above the peak of μ; it is a statement about the mean, not about the worst case $B(\widehat{\mathcal{C}})$, which μ does not bound. We report r_μ per query rather than as a median, since a median over a denominator that shrinks as ε rises would track the denominator.

Which states enter the average. $B(G)$ depends only on the orientations inside the undirected component of X in $\widehat{\mathcal{C}}$: τ_D depends on $\mathrm{pa}_D(X)$ alone, and the chain components of a CPDAG are oriented independently, since no Meek rule has premises in one component and its conclusion in another. Averaging B over every state of a shell would therefore mix in states that differ from G_0 only on components irrelevant to the query, each with the bias of its restriction to X's component, and dilute μ without moving r_{val} or $B(\widehat{\mathcal{C}})$: starting from a single relevant orientation, q disconnected, oriented pairs divide the whole-space $\mu(1)$ by $q + 1$. We therefore average over the states that agree with G_0 outside X's component, the only ones that can change B. On the 39 instances whose whole space is enumerable we checked that B is constant across states sharing the same restriction to X's component. The restriction matters: 16 of the 39 spaces contain other undirected components, and on those the whole-space average delays every onset (on `hepar2` `RHepatitis` $\to$ `THepatitis` it gives onsets from 1 to 9 across the tolerances, where the restricted profile crosses every tolerance at $d = 1$).

Component size as a confound. The graded behaviour follows D, the largest distance from G_0 within X's component (Pearson $+0.75$ between D and the number of distinct finite values of r_μ). A deeper component offers more distances at which the mean can cross a tolerance, so differences between profiles partly reflect component size rather than fragility, and we report them descriptively. That number is 1.00 on average on the 21 instances with $D \le 3$, 1.25 on the 12 with $D \in [4, 6]$, 2.82 on the 11 with $D \in [7, 9]$ and 3.60 on the 5 with $D \ge 10$. Across the 49 instances the r_μ profile alone takes 22 distinct signatures and 32 together with the ceiling, against 3 for r_{val}. Much of this comes from the ceiling itself: $B(\widehat{\mathcal{C}})$ alone takes 21 distinct values and $(r_{\mathrm{val}}, B(\widehat{\mathcal{C}}))$ together 22, so the profile adds 10 signatures beyond the radius and the ceiling, some of which the component size alone may explain. The profile is descriptive, not a guarantee; the guarantees are Theorem 2 and the worst-case ceiling.

Coverage. r_μ is computed from the enumerated space, which has up to 3^k elements in the k undirected edges of $\widehat{\mathcal{C}}$ and a covering relation quadratic in that size. On 48 of the 66 instances the whole space has at most $2{,}997$ states, the largest being `hepar2` at $k = 9$. The other 18 belong to eight networks with larger undirected-edge counts (`Kampen_2014`, `andes`, `child`, `ecoli70`, `magic-irri`, `magic-niab`, `paths`, `win95pts`); for these we enumerate only the states that agree with G_0 outside X's component, which is all μ needs, with distances from Corollary 1(a). On the 39 queries both ways apply, the two computations agree exactly. This covers 14 of the 18; the other 4 (`ecoli70` and three `paths` queries, with 13 or 14 undirected edges in X's component) exceed 900 s. This is a property of the implementation behind these mean-profile experiments, which evaluates B by enumerating DAG extensions state by state, not of the method: the parent-set route of Appendix E.2, which needs no graph enumeration, was not used for this table (scripts `experiments/mean_table*.py`). Every r_μ is exact, with no sampling or estimation. The worst-case radius covers all 66, at $49{,}427$ subsets in total, because Theorem 1 confines its search to the retraction up-set. Of the 62 computed instances, 13 have no failure anywhere and therefore zero bias everywhere, so their r_μ is UNR at every tolerance; Table 6 lists the remaining 49.

Table 6: All 49 instances with a finite r_{val} and non-zero bias, one row per query, with knowledge elicited from the primary language model. $|\mathcal{K}|$ is the size of the elicited knowledge and D the largest distance from G_0 among the states that differ from it only inside X's undirected component, over which μ averages. r_μ is the smallest perturbation size at which the mean bias exceeds ε; μ is non-monotone on 9 of the 49, and * marks a crossing that does not persist to the outermost shell (16 of the 271 finite cells, on 5 instances). UNR: the tolerance exceeds the peak of μ, so the mean profile crosses it at no distance; this bounds the mean, not the worst case, which peak and r_μ need not reach (e.g. `Schipf_2010` A → PA: peak 0.56, $B_{\mathrm{rel}}(\widehat{\mathcal{C}}) = 1.000$, UNR at 60% and 90%). "peak" is $\max_d \mu(d)/|\theta_Z|$ with the d attaining it, and $B_{\mathrm{rel}}(\widehat{\mathcal{C}}) = B(\widehat{\mathcal{C}})/|\theta_Z|$. Every r_μ, peak and $B(\widehat{\mathcal{C}})$ entry is conditional on a linear-Gaussian SEM with one fixed draw of coefficients on the network's DAG (Assumption 1); r_{val} is purely graph-theoretic, and $r_{\mathrm{val}} = 1$ means that a single retraction reaches a failure.

					r_μ at tolerance ε							
Network	Query	$\lvert\mathcal{K}\rvert$	r_{val}	D	2%	10%	25%	40%	60%	90%	peak (d)	$B_{\mathrm{rel}}(\widehat{\mathcal{C}})$
barley	dg25 → aks_m2	7	3	6	3	5	UNR	UNR	UNR	UNR	0.18 (5)	0.184
barley	dg25 → s2225	7	3	6	3	3	3	3	3	5	1.52 (5)	1.525
water	CKNI_12_15 → CBODN_12_45	6	3	6	3	3	3	3	3	3	2.25 (5)	2.245
barley	frspdag → dg25	7	2	6	2	2	2	2	3	3	1.00 (3)	1.000
magic-irri	G3212 → FT	8	2	7	2	2	2	4	4	6	1.22 (6)	1.225
Acid_1996	x1 → x10	1	1	2	1	1	1	1	1	1	1.93 (1)	1.927
Acid_1996	x1 → x11	1	1	2	1	1	1	1	1	1	1.63 (1)	1.629
Acid_1996	x1 → x12	1	1	2	1	1	1	UNR	UNR	UNR	0.38 (1)	0.384
Acid_1996	x1 → x13	1	1	2	1	1	UNR	UNR	UNR	UNR	0.13 (1)	0.126
Didelez_2010	Age → HRT	3	1	4	1	1	1	1	1	1	3.45 (1)	3.453
Didelez_2010	Age → Occ	3	1	4	1	1	1	1	1	1	1.00 (1)	1.000
Didelez_2010	Age → S	3	1	4	1	1	1	1	1	1	1.22 (1)	1.223
Didelez_2010	Age → Smo	3	1	4	1	1	1	1	1	1	1.00 (1)	1.000
Didelez_2010	Age → TCI	3	1	4	1	1	1	1	1	1	1.47 (1)	1.473
Kampen_2014	AFF → AIS	10	1	13	1	1	2	5	8	12	1.00 (12)	1.000
Kampen_2014	AFF → CDR	10	1	13	1	2	5	7	9	12	1.00 (12)	1.000
Kampen_2014	AFF → FTW	10	1	13	1	1	1	2	2*	4*	1.02 (8)	1.619
Kampen_2014	AFF → SAN	10	1	13	1	1	1	1	1	1	28.34 (7)	58.153
Polzer_2012	Caries → Diabetes	5	1	8	1	1	1	2	2	3	1.00 (8)	1.000
Schipf_2010	A → PA	6	1	8	1	1*	1*	2*	UNR	UNR	0.56 (5)	1.000
Schipf_2010	A → S	6	1	8	1*	1*	1*	3*	UNR	UNR	0.53 (5)	1.000
Schipf_2010	A → T2DM	6	1	8	1	1*	2*	3*	5*	UNR	0.64 (5)	0.791
Schipf_2010	A → TT	6	1	8	1	1	1	1	1	5	1.29 (7)	1.422
Schipf_2010	A → U	6	1	8	1	1	3	4	5	7	1.00 (7)	1.000
Sebastiani_2005	ANXA2.5 → ANXA2.11	4	1	2	1	1	1	1	1	1	1.00 (1)	1.000
Shrier_2008	Coach → IntraGameProprioception	3	1	2	1	1	1	UNR	UNR	UNR	0.30 (1)	0.296
Shrier_2008	Coach → TeamMotivation	3	1	2	1	1	1	1	1	1	1.00 (1)	1.000
Thoemmes_2013	e0 → s1	1	1	2	1	1	1	1	1	1	1.00 (1)	1.000
Thoemmes_2013	e0 → s2	1	1	2	1	1	1	1	1	1	1.00 (1)	1.000
Thoemmes_2013	e0 → s3	1	1	2	1	1	1	1	1	1	1.00 (1)	1.000
Thoemmes_2013	e0 → x	1	1	2	1	1	1	1	1	1	1.00 (1)	1.000
Thoemmes_2013	e0 → y	1	1	2	1	1	1	1	1	1	1.00 (1)	1.000
alarm	ANAPHYLAXIS → BP	1	1	2	1	1	1	1	1	1	1.00 (1)	1.000
alarm	ANAPHYLAXIS → HRBP	1	1	2	1	1	1	1	1	1	1.00 (1)	1.000
andes	GIVEN_1 → SNode_100	10	1	2	1	1	1	1	1	1	1.00 (1)	1.000
andes	HORIZ53 → SNode_118	10	1	4	1	1	1	1	1	1	1.00 (1)	1.000
asia	asia → dysp	3	1	2	1	1	1	1	1	1	1.00 (1)	1.000
asia	asia → either	3	1	2	1	1	1	1	1	1	1.00 (1)	1.000
asia	asia → tub	3	1	2	1	1	1	1	1	1	1.00 (1)	1.000
asia	asia → xray	3	1	2	1	1	1	1	1	1	1.00 (1)	1.000
asia	bronc → dysp	3	1	2	1	1	1	1	1	UNR	0.84 (2)	1.256
child	CO2 → DuctFlow	10	1	14	1	1	2	2	3	3	1.00 (3)	1.000
hepar2	RHepatitis → THepatitis	5	1	2	1	1	1	1	1	1	1.00 (1)	1.000
magic-niab	G1217 → FT	5	1	4	1	1	1	1	UNR	UNR	0.44 (1)	0.441
magic-niab	G1217 → YLD	5	1	4	1	1	1	1	1	1	1.13 (1)	1.127
mediator	I → X	4	1	7	1	1	1	2	2	4	1.00 (6)	1.000
mediator	I → Y	4	1	7	1	1	1	1	1	1	4.57 (4)	4.967
mediator	I → Z	4	1	7	1	1	2*	2*	3*	UNR	0.75 (4)	1.000
mediator	X → Y	4	1	7	1	1	2	2	2	6	1.00 (6)	1.000

E.6 Beyond the linear-Gaussian model

The validity radius r_{val} is graphical and model-free: it is defined by the validity of Z and holds for any observational law compatible with $\widehat{\mathcal{C}}$. The effect-scale results use Assumption 1 in one place. We define both θ_Z and each τ_D as population least-squares coefficients, and linearity makes such a coefficient equal to the causal effect when one adjusts for a valid set. In a general structural causal model, a valid adjustment set identifies the effect through the g-formula, which is a regression coefficient only when the outcome regression is linear in X with no X–Z interaction. Nonsingularity of Σ makes the coefficient unique. Redefining τ_D and θ_Z as g-formula contrasts adjusting for $\mathrm{pa}_D(X)$ and Z carries Proposition 2 and Lemma 9 to any structural causal model with positivity, because their order-theoretic content uses only that $B(G)$ is a maximum over a set indexed by $[G]$ and that $[\cdot]$ grows as knowledge is withdrawn. The estimand then becomes a g-formula contrast, and each τ_D is computed by a nonparametric regression in place of the $2^{\deg(X)}$ linear solves of the linear-Gaussian case.

F The worked example in full

Table 7 gives the shell-by-shell profile of the running example under the knowledge state of Section 4. Its bias column is a sampled mean over drawn linear-Gaussian coefficients. The exact worst case, shown in Figure 6 and Table 8, is flat for $d < r_{\mathrm{val}}$ and then steps directly to the ceiling. The gradual rise of the mean past shell 3 therefore reflects the growing fraction of states in each shell at which Z is invalid. Over the 900 instances of Appendix E.4, the worst-case staircase is a step in 84.1% of instances and a ramp in 15.1%.

Table 7: Shell profile of the running example: number of states at each distance from G_0, how many invalidate $Z = \{\texttt{Age}, \texttt{Smoke}\}$, and the sampled mean absolute bias of the Z-adjusted estimator there; Figure 6 gives the exact worst-case staircase for the same instance. On shells 0 to 2 the bias is zero by Theorem 2, and the computed means are below 10^{-15}. The mean absolute total effect across the space is 1.15 and the asymptotic standard error at $n = 1000$ is 0.024.

shell	states	Z invalid	mean abs. bias
0	1	0	0
1	2	0	0
2	4	0	0
3	**5**	**1**	**0.033**
4	8	3	0.077
5	8	4	0.127
6	8	5	0.214
7	4	3	0.314
8	5	4	0.466
9	2	2	0.551
10	1	1	0.619

The two tables share the states and Z-invalid columns and differ in the last. What rises shell by shell in Table 7 is the proportion of states at which Z is invalid, while at this draw the worst case is reached at r_{val} and grows no further. The shape also depends on the coefficients. Over 400 further coefficient draws the worst case is flat past r_{val} on 221 and rises once more on 179, with a median relative rise of zero. At $\varepsilon = 50\%$ the threshold is never reached on 334 of the 400 draws, which is the ceiling of Section 7 on a single graph.

Three knowledge states, compared. Alongside the state used in the main text, in which the analyst asserts $\texttt{Statin} \to \texttt{CRP}$ (statins lower CRP), we computed two others on the same CPDAG. The illustrative data-generating DAG is built to reverse that edge, so the assertion, clinically plausible and consistent with the data, is the wrong one in this example. A mild state declines to say whether inflammation precedes treatment. A compound state holds the reversed assertion and also omits a claim that agrees with the data-generating DAG. The mild state and the main-text state both have

Table 8: The same shells, with the exact worst-case bias $\beta^{\uparrow}(d)$ at one fixed coefficient draw, replacing the sampled mean of Table 7. The reported estimate is $\theta_Z = 0.3262$ and here coincides with the effect in the data-generating DAG, so the realised error is zero. The worst case is flat for $d < r_{\mathrm{val}} = 3$ and then steps to $B(\widehat{\mathcal{C}})$, the largest bias any perturbation can produce. Since $\beta^{\uparrow}$ maximises over the whole ball and $d(G_0, \widehat{\mathcal{C}}) = |\mathcal{K}_{G_0}| = 5$, it equals $B(\widehat{\mathcal{C}})$ at every $d \geq 5$; shells past 5 add no information. Radii computed over the retraction up-set agree with a search over the whole space.

shell	states	Z invalid	worst-case bias	relative
0	1	0	0.0000	0.0%
1	2	0	0.0000	0.0%
2	4	0	0.0000	0.0%
3	**5**	**1**	**0.1629**	**49.9%**
4	8	3	0.1629	49.9%
5	8	4	0.1629	49.9%
6	8	5	0.1629	49.9%
7	4	3	0.1629	49.9%
8	5	4	0.1629	49.9%
9	2	2	0.1629	49.9%
10	1	1	0.1629	49.9%

$r_{\mathrm{val}} = 3$, and the compound state has $r_{\mathrm{val}} = 2$: a wrong assertion combined with a gap brings the first invalidating state one shell closer, and the radius registers the difference.

Against a naive count. A simpler distance is the Hamming distance on the knowledge set, which counts the assertions that differ. The two correlate with Spearman coefficient 0.73 to 0.77 across the three states, and about 29% of elements differ by two or more. The difference is systematic. The naive count understates how far a state is, because one assertion edit can cascade through Meek's rules into several covering steps. One element of this space sits eight covering steps from G_0 and only four assertion edits away, because those four edits reverse the orientation of the entire confounder block.

Calibration. Because every DAG of the class is known here, each element of the space can be treated in turn as a hypothetical data-generating state. Of the states lying outside the certified radius, a fraction of 0.43 to 0.47 across the three knowledge states still admit Z. The certificate is therefore conservative: a state outside the radius is one where validity is no longer certified, which is weaker than a prediction that Z is invalid there.

Other valid sets. Computing the radius for every valid adjustment set of G_0, rather than only the optimal one, gives a flat profile here: all twenty-eight valid sets have $r_{\mathrm{val}} = 3$, and the optimal set is the most efficient in causal linear models under least-squares adjustment (Henckel et al., 2022) and as robust as every alternative. The reason is structural and specific to this instance. At the first invalidating state one DAG extension makes the treatment an ancestor of the entire confounder block, so every valid set loses validity at the same distance. When the confounders are separated in the graph, the valid sets can reach different radii, and the radius of each can be reported beside its efficiency.

G EXPERIMENTAL PROTOCOL

Table 9 lists the settings of both experiments stage by stage. Rows that span both columns are shared by the two settings, which differ only in where the structure and the claims come from and in the unit of the bootstrap.

Table 9: Experimental protocol of Sections 6.2 and 6.3. The data-generating graph sets up only the CPDAG, the corruption and the oracle baselines; the radius uses the analyst's inputs alone.

stage	synthetic structure (Section 6.2)	published networks (Section 6.3)
structure	Generated DAGs; undirected-component size $\{6, 8, 10, 12\}$, crossed with the minimum, median and maximum achievable separation between treatment and adjustment set	33 published networks, Bayesian networks and DAGs from applied studies; chain components of at most 20 nodes, so answers exist on 32 networks, of which 27 enter the per-query records and 24 the survival ranking (Appendix I)
CPDAG $\widehat{\mathcal{C}}$	The CPDAG of the data-generating graph, so that the skeleton and the v-structures are correct, as the theory assumes	
claims $\mathcal{K}$	Flipped knowledge: coverage $\{0.5, 1.0\}$ $\times$ base wrongness $\{0, 0.10, 0.25\}$. Tiered knowledge: an ordering with $\{2, 3, 4\}$ tiers. 40 instances per cell where the generator admits them (2,730 flip, 1,374 tiered)	A causal order per component from six language models at temperature zero under ten conditions, two with scrambled names (Table 14), read as an orientation or a decline for each of 369 undirected pairs; Qwen2.5 7B is the primary model, with a random and an empty state as controls
analyst's state	$G_0 = \mathrm{Meek}(\widehat{\mathcal{C}}, \mathcal{K})$; for a query (X, Y) the adjustment set $Z = \mathbf{O}(G_0)$ is committed at G_0 and held fixed under corruption	
radius	r_{val}, the fewest elementary revisions of the closure orientations from G_0 to a consistent state in which Z is invalid, computed from $\widehat{\mathcal{C}}$, $\mathcal{K}$, X, Y and Z alone, without sampling; its exact value comes with a certificate naming an orientation set whose retraction breaks Z	
corruption	$\mathcal{K}$ corrupted at a grid of intensities with 1,000 draws per grid point, by flips (claims reversed independently and uniformly at random) or by tiered errors (variables misplaced within the ordering); each draw is re-closed under Meek's rules, which flag contradictory draws and set them aside	
outcome	Whether Z remains a valid adjustment set, scored in each corrupted and re-closed state, the data-generating graph setting up only the CPDAG and the corruption; the surviving fraction, averaged over a ten-point intensity grid, is the survival AUC of the instance	
predictors	r_{val} against the analyst's counts $\lvert\mathcal{K}\rvert$ and $k_{G_0} = \lvert\mathcal{K}_{G_0}\rvert$, and against two oracle quantities no analyst can compute: the SHD from G_0 to the data-generating graph and, on published networks, the claim accuracy k_{acc}; each is evaluated at the uncorrupted state	
statistic	Kendall's τ_b between predictor and survival AUC; head-to-head comparisons by a paired bootstrap of the difference in τ_b on the same units	
scale	4,104 instances and 32.8M corruption draws, of which 15.1M tiered; bootstrap clustered by instance	528 queries per condition on 27 networks; 2,422 scored units (original cohort; 2,516 with the 94 valid empty sets) over the eight real-name conditions on 24 networks; bootstrap clustered by network

H ADDITIONAL EXPERIMENTAL DETAIL

Synthetic design. The synthetic track crosses component size $\{6, 8, 10, 12\}$ with the minimum, median and maximum achievable separation between treatment and adjustment set, and, for flipped knowledge, coverage $\{0.5, 1.0\}$ and base wrongness $\{0, 0.10, 0.25\}$; tiered knowledge uses $\{2, 3, 4\}$

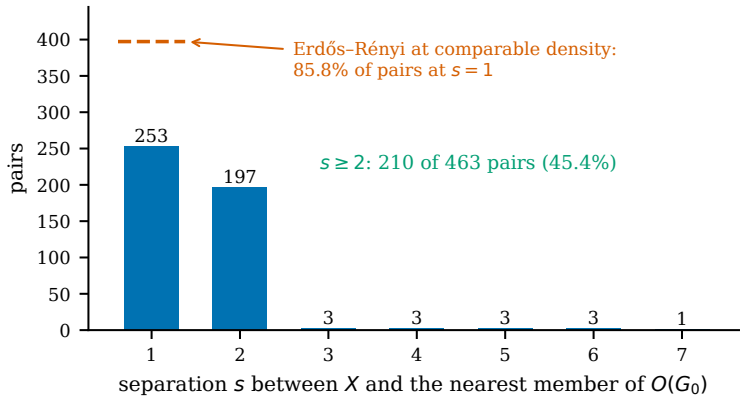

Figure 8: Separation s between the treatment X and the nearest member of the adjustment set, inside the undirected component of X, over the 463 treatment–outcome pairs of the published networks where s is defined. Separation of at least 2 occurs in 45.4% of these pairs, against 14.2% in Erdős–Rényi ensembles at comparable density (dashed marker).

tiers. There are 40 instances per cell where the generator admits them (2,730 flip and 1,374 tiered), and 1,000 draws per grid point: 17.68M flip and 15.11M tiered draws. Meek closure flags 84.4% of flip draws and 59.4% of tiered draws as contradictory, so an analyst sees these errors at once, as contradictions to resolve (our elicitation pipeline resolves them by dropping the least confident orientations). Independent reversals of a closed claim set collide with consistency more often than block-correlated relocations do. The radius is aimed at the errors closure cannot see: those that keep the knowledge consistent and still break Z.

Paired comparisons on synthetic structure. Table 10 reports the paired bootstrap difference in τ_b between r_{val} and five baselines. The count k_{G_0} coincides with $|\mathcal{K}|$ at full coverage, and where the two differ, r_{val} leads k_{G_0} by $+0.19$ to $+0.73$. The single-claim fragility φ_1 (Appendix H.2) rises with fragility, the opposite sense of a radius, so it enters as $-\varphi_1$; on that corrected sign r_{val} leads in six of nine strata, ties in two, and trails at half coverage with 25% base wrongness (-0.37), the stratum where r_{val} itself does not rank survival. Against the claim radius r_{claim} (Appendix H.2), on the rows where r_{claim} is not censored, r_{val} coincides with it on flips except at half coverage with 25% base wrongness, where the two tie, and leads on tiered knowledge at three and four tiers; r_{claim} is constant at two tiers, so the comparison is undefined there. Separation between treatment and adjustment set is the generator's own design variable, so it is a strong baseline here by construction: published networks reach large separations far more often than random graphs do (Figure 8), which is why the generator targets separation directly, and on this generator separation ranks survival about as well as the radius, ahead of it at three and four tiers (daggers in the table) and behind it elsewhere. The radius differs from separation on 56% of the 4,104 instances. Holding the generator's separation cell exactly fixed – so only instances where r_{val} itself still varies contribute – the radius still ranks survival, at $\tau_b = 0.49$ $[0.40, 0.56]$ on tiered knowledge (320 instances) and 0.38 $[0.31, 0.43]$ under flips (1,132 instances), so its ranking is not solely an artefact of the design; within these same cells the counts, SHD and $-\varphi_1$ are not distinguishable from zero, except $-\varphi_1$ on flips ($\tau_b = 0.18$).

Which predictor wins where they disagree. Instance-level paired τ_b on the 2,287 instances where $r_{\text{val}} \neq$ separation gives no single winner: separation leads in some design cells, r_{val} in others. The pair-level test is cleaner and avoids that composition effect: over every pair of instances whose r_{val}-order and separation-order strictly disagree, survival sides with r_{val} on 85% of such pairs at half coverage with no base wrongness and 82% at two tiers, but with separation on 64% at half coverage with 25% base wrongness, 62% at full coverage with 25% base wrongness, 69% at three tiers and 88% at four tiers; other cells are near even (`results/axis_robustness_p6_paired_v2/discordant/`). Conditioning the other way – holding r_{val} fixed within a design cell – separation orders survival only weakly, $\tau_b = 0.12$ $[0.05, 0.23]$ under flips and 0.04 $[-0.12, 0.20]$ on tiered knowledge (the latter over just 141 instances where separation still varies at fixed r_{val}), against 0.38 and 0.49 the other way round (above): the

radius carries information about survival that separation does not, and the reverse holds far less, or not at all in the tiered arm.

Robustness and accuracy come apart. SHD measures how wrong the knowledge is; r_{val} measures how much can go wrong before Z breaks. With full coverage and zero base wrongness G_0 is the data-generating DAG, so SHD is zero on all 480 flip instances, yet r_{val} still ranks survival, ahead of the claim count by $+0.564$ $[0.461, 0.668]$ (App. Table 10). SHD enters with its raw sign, higher meaning less accurate knowledge. Survival measures persistence of the adjustment guarantee, not correctness, so accuracy need not predict it in either direction, and its τ_b is positive in some strata and negative in others. The comparison with r_{val} does not hinge on this choice: $\tau_b(r_{\mathrm{val}}) - |\tau_b(\mathrm{SHD})|$ is positive in every stratum where SHD varies except flips at half coverage with 25% base wrongness, the stratum where r_{val} itself does not rank survival, and on the published networks r_{val}'s lead only grows if SHD is negated.

Table 10: Synthetic structure: paired bootstrap difference in τ_b, r_{val} minus each baseline, instance-clustered, 10,000 resamples, $N = 1{,}000$ draws per grid point. cov is knowledge coverage and bw base wrongness; flip and tiered corruption use different intensity axes and are not compared row to row. $-\varphi_1$ is the fragility φ_1 with its sign corrected (Appendix H.2); r_{claim} is compared on the rows where its search is not censored; "id." marks an exact identity between r_{val} and the baseline on the matched rows; "const." marks a baseline constant on the stratum, leaving the difference undefined. Bold: interval excludes zero in favour of r_{val}; †: interval excludes zero against r_{val}. SHD enters with its raw sign; ‡: the one lead that depends on it, since against $-$SHD the point difference is about -0.20 (see text). In the flip stratum with cov 1.0 and bw 0.00, G_0 is the data-generating graph, so SHD is constant. At two tiers exactly one claim breaks Z on every instance, so $\varphi_1 = 1/|\mathcal{K}|$ and $-\varphi_1$ orders instances as $|\mathcal{K}|$ does; the two differences there are therefore identical.

stratum	n	$\tau_b(r_{\mathrm{val}})$	Δ vs. sep.	Δ vs. $-\varphi_1$	Δ vs. r_{claim}	Δ vs. $\lvert\mathcal{K}\rvert$	Δ vs. SHD
flip, cov 0.5, bw 0.00	480	0.614	**0.153 [0.118, 0.190]**	**0.235 [0.197, 0.275]**	0 (id.)	**0.355 [0.245, 0.465]**	**0.384 [0.302, 0.466]**
flip, cov 0.5, bw 0.10	480	0.328	**0.083 [0.040, 0.128]**	−0.017 [−0.077, 0.040]	0 (id.)	**0.385 [0.278, 0.489]**	**0.338 [0.244, 0.434]**
flip, cov 0.5, bw 0.25	476	−0.027	−0.032 [−0.097, 0.032]	†−0.369 [−0.444, −0.296]	−0.030 [−0.072, 0.011]	**0.239 [0.145, 0.336]**	0.142‡ [0.060, 0.225]
flip, cov 1.0, bw 0.00	480	0.469	0 (id.)	**0.107 [0.078, 0.137]**	0 (id.)	**0.564 [0.461, 0.668]**	const.
flip, cov 1.0, bw 0.10	480	0.418	−0.000 [−0.019, 0.018]	**0.114 [0.055, 0.174]**	0 (id.)	**0.527 [0.430, 0.622]**	**0.594 [0.496, 0.690]**
flip, cov 1.0, bw 0.25	334	0.280	−0.033 [−0.085, 0.018]	−0.020 [−0.096, 0.060]	0 (id.)	**0.392 [0.269, 0.512]**	**0.381 [0.255, 0.508]**
tiered, 2 tiers	414	0.819	**0.085 [0.056, 0.115]**	**1.387 [1.340, 1.431]**	const.	**1.387 [1.340, 1.431]**	**0.543 [0.483, 0.604]**
tiered, 3 tiers	480	0.807	†−0.013 [−0.025, −0.002]	**0.274 [0.239, 0.309]**	**0.148 [0.124, 0.173]**	**1.263 [1.196, 1.327]**	**0.727 [0.664, 0.790]**
tiered, 4 tiers	480	0.748	†−0.031 [−0.044, −0.019]	**0.132 [0.104, 0.161]**	**0.034 [0.012, 0.058]**	**1.178 [1.109, 1.245]**	**0.613 [0.542, 0.684]**

Real structure: selection and balance. A unit is scorable when its optimal adjustment set is defined, its consistent extensions can be enumerated and its knowledge is nonempty. Units are otherwise excluded because the optimal set is undefined, because there are too many consistent extensions to enumerate (a frozen eligibility rule of the experimental pipeline, which read the optimal set off by enumerating extensions; the radius and the adjustment criterion need no enumeration), or because $|\mathcal{K}| = 0$. On the eight real-name conditions, 1,144 units were excluded as having an undefined optimal set: 898 have no optimal set, 152 have an empty optimal set that is not valid in G_0, and 94 have a valid empty optimal set. These 94 were excluded by an error in the scoring pipeline, which treated an empty set as undefined; the exclusion was not planned, and it was identified after the main analyses. Table 1, Table 11, Table 13 and the remaining analyses of this section use the original cohort. Scoring those 94 with the same protocol (all have $r_{\mathrm{val}} = 1$) and adding them leaves every conclusion unchanged: pooled $\tau_b(r_{\mathrm{val}}) = 0.32$ $[0.13, 0.49]$ (2,516 units), paired leads over k_{G_0} and $|\mathcal{K}|$ of $+0.14$ $[-0.06, 0.32]$ and $+0.16$ $[-0.07, 0.35]$, over SHD $+0.30$ $[0.06, 0.49]$, and within-state $\tau_b = 0.43$ $[0.07, 0.69]$ (93 states; `results/axis_robustness_llm_v2/table1_corrected.json`). Because entry into the scorable sample correlates with $|\mathcal{K}|$, which could by itself favour size-based predictors, Table 11 restricts to units scorable in at least a given number of the eight conditions. The interval of r_{val} excludes zero at every level, and among the predictors of Table 11 r_{val} is the only one for which it does so at $8/8$ and at $\geq 7/8$; separation, defined on 96 of the 248 balanced units (3 networks), is $+0.39$ $[0.22, 0.51]$ there (Table 1). Pooled over the scorable units (2,422 units, 24 networks), r_{val} also outranks SHD to the data-generating graph, by $+0.265$ $[0.038, 0.463]$ in a paired network-cluster bootstrap with 10,000 resamples. Paired against the counts, the pooled lead is not significant ($+0.11$ $[-0.09, 0.30]$ over k_{G_0}, $+0.13$ $[-0.08, 0.32]$ over $|\mathcal{K}|$), and it is still not significant on the balanced $8/8$ panel ($+0.12$ $[-0.22, 0.53]$, $+0.17$ $[-0.16, 0.59]$); a single predictor's interval excluding zero while another's does not is not itself a test that the two differ. Separation is defined on 1,389 of the 2,422 pooled units; there r_{val} leads it by $+0.04$ $[-0.18, 0.28]$, not significant. Inside a

single knowledge state, where $|\mathcal{K}|$ and k_{G_0} are constant, the comparison with knowledge size is the within-state result of Table 1, detailed in Appendix H.3; on the subset of those units where separation is also defined (1,213 units, 79 states), r_{val} leads separation by $+0.19$ $[0.07, 0.90]$ (15 network clusters, hence the wide upper bound), and the two agree on 56% of the units.

Table 11: Balanced subsets: τ_b [95% CI] for each predictor, restricting to units scorable in at least the stated fraction of the eight real-name conditions. Bold marks CIs that exclude zero.

level	r_{val}	k_{G_0}	$\lvert\mathcal{K}\rvert$	SHD
8/8 (31 queries, 248 units, 7 networks)	**+0.499 [0.240, 0.624]**	+0.380 [−0.074, 0.562]	+0.333 [−0.125, 0.484]	+0.285 [−0.167, 0.456]
≥7/8 (829 units, 16 networks)	**+0.251 [0.047, 0.479]**	+0.292 [−0.027, 0.590]	+0.262 [−0.030, 0.570]	+0.074 [−0.215, 0.281]
≥6/8 (205 queries, 1,375 units, 20 networks)	**+0.230 [0.048, 0.470]**	**+0.304 [0.085, 0.514]**	**+0.255 [0.048, 0.464]**	+0.053 [−0.144, 0.242]
pooled (2,422 units, 24 networks)	**+0.319 [0.122, 0.485]**	**+0.214 [0.079, 0.339]**	**+0.185 [0.071, 0.327]**	+0.054 [−0.077, 0.200]

Query-specificity. $|\mathcal{K}|$ and k_{G_0} are properties of the whole knowledge state: they assign the same score to every query answered against it. r_{val} is a property of the query (X, Y, Z). Two knowledge states with the same $|\mathcal{K}|$ can leave one adjustment set one retraction from failure and another deep inside a Meek-closed cascade, and a cardinality count cannot tell them apart.

Units of the corruption axis. The radius counts covering steps while the corruption process counts claims, so the two axes are not the same. On a small-component subsample where exact hop distances could be computed by enumeration, one corrupted claim corresponds to about 2.9 covering steps (s.d. 1.01), converging to twice the claim depth at depth five and beyond. The sign of every reported association agrees on both axes.

Contradictory draws. Meek closure flags a contradictory draw at once, and such draws are set aside from AUC (Section 6.1), since r_{val} is defined on consistent states. AUC is the arithmetic mean of survival over a fixed ten-point grid of fractions $0.1, \ldots, 1.0$ of the instance's own $|\mathcal{K}|$, not a numerical integral. On the raw grid, a point is undefined when every draw at its target depth is contradictory. For the mean, each point takes survival at the nearest depth on the whole curve with a defined value, ties broken toward the smaller depth, so the filled grid is complete whenever any depth is defined. Undefined raw grid points concentrate at high intensity: every tiered instance and 69% of flip instances have at least one. Recomputing AUC on each stratum's common support – the depths defined for every instance in it, 2,000 resamples – leaves every conclusion unchanged, including separation's lead on three and four tiers (-0.02 $[-0.03, -0.01]$, -0.04 $[-0.06, -0.03]$).

Unchanged corruptions. A draw can be consistent yet leave the closure unchanged. Let p_0 be the probability a consistent draw changes no closure orientation: under flips $p_0 = 0$ at every intensity, since an independent reversal always touches the closure; under tiered corruption p_0 falls from 0.86 to 0.11 to 0.02 over the first three intensities and is negligible beyond; over all intensities it is 0.42, 0.53 and 0.61 at two, three and four tiers. Rebuilding each instance's AUC from the draws that do change the closure state separates the two sources of survival. Under flips nothing changes ($p_0 = 0$), and $\tau_b(r_{\mathrm{val}}, \mathrm{AUC})$ is 0.39 pooled either way. Under tiered corruption the pooled value falls from 0.78 to 0.54 $[0.51, 0.57]$, and the fall comes from one stratum: at three and four tiers the changed-draw ranking is 0.76 $[0.73, 0.78]$ and 0.75 $[0.73, 0.77]$, against 0.81 and 0.75 on all draws, but at two tiers it is -0.03 $[-0.08, 0.03]$, against 0.82. At two tiers, therefore, the radius ranks how often a tier relocation leaves the analyst's closure untouched, not how the committed set fares once the state has changed; at three and four tiers the ranking survives the conditioning (instance bootstrap, 2,000 resamples; `results/axis_robustness_p6_paired_v2/changed_state_strata.json`).

H.1 Queries drawn without targeting separation

Generator. The synthetic tracks above draw separation from {minimum, median, maximum achievable} before the query is fixed, so every radius level is populated by construction. This sweep instead grows the component exactly as before but without a preallocated spine to a chosen treatment: every other component vertex is tested, after the graph exists, against the same admissibility gate the targeted generator applies (a valid adjustment set exists in the truth, the empty set is not trivially valid, some atomic knowledge-perturbation changes validity), the admissible set is collected in

full, and the treatment is drawn uniformly at random from it – the gate decides which vertices may be asked, never which one is. Separation is then measured, never targeted: it is the CPDAG distance from X to the nearest member of the analyst's committed set inside X's undirected component, kept (with status recorded, never dropped or imputed) when that set has no in-component member. The sweep crosses component size $\{6, 8, 10, 12\}$ with $\{2, 3, 4\}$ tiers, 40 instances per cell (480 total) and 1,000 corruption draws per grid point, otherwise identical to the tiered track of Section 6.2.

Table 12: Queries drawn without targeting separation: Kendall τ_b of r_{val} with survival AUC (instance bootstrap, 95% CI) and paired bootstrap difference r_{val} minus each baseline, 10,000 resamples, $n = 160$ per tier, 480 pooled. Bold: interval excludes zero in favour of r_{val}; †: interval excludes zero against r_{val}.

stratum	n	$\tau_b(r_{\mathrm{val}})$	Δ vs. sep.	Δ vs. k_{G_0}	Δ vs. r_{claim}	Δ vs. $\lvert\mathcal{K}\rvert$	Δ vs. SHD
2 tiers	160	0.447 [0.364, 0.517]	0.064 [−0.037, 0.165]	**0.444 [0.309, 0.575]**	**0.306 [0.201, 0.415]**	**0.541 [0.375, 0.701]**	**0.346 [0.214, 0.473]**
3 tiers	160	0.351 [0.224, 0.466]	†−0.128 [−0.239, −0.021]	**0.187 [0.021, 0.342]**	**0.102 [0.052, 0.159]**	**0.243 [0.064, 0.409]**	0.162 [−0.008, 0.325]
4 tiers	160	0.196 [0.074, 0.315]	†−0.226 [−0.325, −0.130]	−0.053 [−0.212, 0.111]	**0.057 [0.014, 0.106]**	−0.008 [−0.180, 0.169]	−0.087 [−0.235, 0.063]
pooled	480	0.292 [0.220, 0.360]	−0.053 [−0.125, 0.017]	0.074 [−0.012, 0.157]	**0.098 [0.064, 0.135]**	**0.118 [0.028, 0.207]**	**0.307 [0.199, 0.412]**

Marginally (table columns not shown), separation itself ranks survival at $\tau_b = 0.38$, 0.48 and 0.42 per tier and 0.34 $[0.28, 0.40]$ pooled, against r_{val}'s 0.45, 0.35, 0.20 and 0.29 $[0.22, 0.36]$: it leads r_{val} at three and four tiers and pooled, and the paired difference is significant only at three and four tiers (Table 12). Separation is measured on every instance here (no undefined rows), and differs from r_{val} on 77%, 63% and 46% of instances per tier (62% pooled). The radius is concentrated near its minimum: $r_{\mathrm{val}} = 1$ on 66% of instances (316/480), with mean 1.50 (s.d. 0.80) and a maximum of 5.

Interpretation. Under a query rule that does not target separation, separation is at least as good a marginal predictor of average-case survival as the radius, and better at three and four tiers; pooled, r_{val} keeps a significant lead over SHD and $|\mathcal{K}|$, ties k_{G_0}, and at four tiers is no longer distinguishable from either count or from SHD. This does not undercut the radius as a diagnostic: it is exact, comes with a certificate naming orientations whose retraction breaks Z, and is defined on queries – such as most of those in Section 6.3 – where separation is not. It does mean that, for a cheap average-case ranking on synthetic structure with no such certificate required, separation is a serious competitor, and the main text reports it as one rather than as a secondary check.

H.2 A SEPARATE COMPANION MEASURED IN ELICITED-CLAIM UNITS

Closure orientations and elicited claims. Everything reported so far concerns the breakdown radius r_{val}, which counts elementary revisions of $\mathcal{K}_{G_0}$; every result stated in closure orientations is a statement about it (Remark 1). This subsection defines a companion quantity in the unit of elicited claims, and statements about r_{val} do not transfer to it. The two units can come apart. Because R1–R4 propagate automatically, a single elicited claim can be responsible for many elements of $\mathcal{K}_{G_0}$, so a count in closure orientations and a count in claims need not agree even in order of magnitude. The claim radius r_{claim} is the minimum number of elicited claims whose retraction breaks the adjustment set.

Claims that propagate. In the real corpus a median of 5 elicited claims cascades to 12 closure orientations. In the most extreme case, the `paths` network, a single claim orients a 15-node chain and gives $r_{\mathrm{val}} = 14$. That value correctly states that 14 closure orientations must be revised before Z breaks. Read as a budget of 13 safe claims, however, it would be a unit error, since retracting that one claim breaks the adjustment set and the budget in claim units is zero. The single-claim fragility φ_1, the fraction of single claims whose withdrawal invalidates the set, reports such cases directly.

The elicitation of Section 6.3 provides claim sets of varying size and a matched random control: each item asks for a causal order over one chain component (Appendix I), and random states orient the same edges. The survival rankings of Section 6.3 and Appendix H.3 are stated in r_{val}; the results in this subsection are the ones stated in r_{claim}. On the primary model's knowledge, 141 of the 157 exact certificates name a claim whose withdrawal alone breaks Z (0.90 $[0.81, 0.97]$, 24 networks), a candidate to confirm first; other single claims may do the same. Deeper dependence also occurs: 8 of the 229 committed sets have $r_{\mathrm{claim}} \geq 3$ (0.035 $[0.005, 0.073]$).

The two units are positively associated. On the primary model's knowledge states, the claim radius and the breakdown radius are positively associated: Kendall $\tau_b = 0.66$ $[0.43, 0.84]$ over the 157 exact certificates (24 networks), and the ratio $r_{\text{val}}/r_{\text{claim}}$ equals 1 on at least half of them. A breakdown radius of two or more coincides with a single-claim failure on 7.8% $[2.5\%, 14.6\%]$ of the 204 queries where both are defined, and a breakdown radius of three or more on 1.0% $[0\%, 3.5\%]$. On elicited knowledge the cascade that separates the two units is therefore rare, and the closure count can be read alongside the claim count, though the two can order queries differently.

H.3 Elicited knowledge: further results

The results below support Sections 6.3 and 7. Every interval is a 95% percentile interval from a network-clustered bootstrap with 10,000 resamples, computed separately for each knowledge condition. The conditions and models are listed in Table 14.

The within-state ranking. The within-state result of Table 1 computes Kendall τ_b between r_{val} and AUC inside each knowledge state, one network under one condition. The states are combined by pooling pair counts, not by averaging statewise values: with C and D the concordant and discordant pairs and T_x, T_y the pairs tied only in r_{val} or only in AUC, each counted over pairs of units in the same state and summed over states, the statistic is $(C - D)/\sqrt{(C + D + T_x)(C + D + T_y)}$. Pairs tied in both are dropped, so each state is weighted by its number of pairs, and no statewise τ_b is formed or needs to be defined. Only states in which r_{val} varies enter; they are selected once on the full sample and not reselected in bootstrap replicates, which draw networks with replacement and include all of a drawn network's states (a network drawn twice contributes twice). The permutation test shuffles r_{val} within states, and none of 10,000 permutations reached the observed value. Of the 161 knowledge states on the eight real-name conditions, 92 show variation in r_{val}. They hold 1,973 units on 18 networks, and neither $|\mathcal{K}|$ nor k_{G_0} varies inside any of them; on these states the radius ranks survival at $\tau_b = 0.42$ $[0.05, 0.68]$. Leaving out one network at a time, the within-state τ_b ranges from 0.22 $[0.02, 0.67]$ (without `insurance`) to 0.64 $[0.43, 0.69]$ (without `barley`), 17 networks each, and no removal yields an interval that includes zero. Sixteen of the 18 networks give a positive value, and among the 92 states, 71 have more concordant than discordant pairs, 15 fewer and 6 as many. Table 13 gives the value for each condition; all eight real-name point estimates are positive. The two scrambled-name rows are of similar size (0.38 $[0.29, 0.77]$, 9 networks; 0.38 $[-0.13, 0.61]$, 14 networks), so we find no evidence that the within-state association rests on recall of the published networks.

Table 13: Within-state Kendall τ_b between r_{val} and AUC by elicitation condition, 95% network-clustered bootstrap intervals. The first eight rows are the real-name conditions pooled in Table 1; the last two use scrambled names. All eight real-name point estimates are positive.

condition	units	τ_b [95% CI]
Qwen2.5 7B (primary)	167	0.407 [0.275, 0.832]
Qwen2.5 7B, second presentation order	225	0.645 [−0.328, 0.879]
Llama 3.1 8B	231	0.740 [0.150, 0.795]
Gemma-3 27B	198	0.154 [−0.476, 0.633]
Qwen2.5 32B	187	0.221 [0.132, 0.876]
Llama 3.3 70B	333	0.541 [0.145, 0.725]
Qwen2.5 72B	339	0.365 [−0.021, 0.634]
Qwen2.5 72B, second presentation order	293	0.075 [−0.111, 0.678]
Qwen2.5 7B, scrambled names (9 networks)	171	0.376 [0.285, 0.771]
Qwen2.5 72B, scrambled names (14 networks)	227	0.383 [−0.128, 0.611]

Certificates across conditions. The share of exact claim certificates that name a single orientation is 0.90 $[0.81, 0.97]$ for the primary model, 0.94 $[0.85, 0.99]$ for knowledge copying the data-generating orientations, 0.90 $[0.84, 0.95]$ for random orientations and 0.92 $[0.86, 0.97]$ with scrambled names. Certificates with $r_{\text{claim}} \geq 3$ number 8 of 229 for the primary model, and 1, 12 and 3 for the three comparison conditions in the order above.

I THE RELEASED KNOWLEDGE STATES

This appendix is the datasheet of the knowledge states used in Section 6.3. A knowledge state here is a set of orientations for the edges a CPDAG leaves undirected, elicited from a language model, made consistent under Meek's rules, and paired with its consequence for the committed adjustment set. The data-generating graph of each network is used only to build its CPDAG and to score the states; no quantity the diagnostic computes uses it.

Networks. The questionnaire covers 33 published networks, Bayesian networks and DAGs drawn from applied studies, all taken from a single public archive of published networks. For each network we take the CPDAG of its data-generating DAG and ask about every chain component with at most 20 nodes. The components of `pathfinder` all exceed that cap, so answers exist on 32 networks. The analyses read subsets of these: the per-query records exclude the five networks above 500 nodes and cover 27, and the survival ranking has scorable units on 24. The within-state and balanced analyses of Appendix H.3 and Table 11 use those of the 24 networks whose states meet their conditions.

Items. The questionnaire holds 540 items. There are 480 open items, one per chain component (120 components, 369 undirected pairs) under each of two namings, the published variable codes and scrambled labels, and each of two presentation orders of the variables. The other 60 are control items, one per naming on each of 30 networks, which ask about 206 pairs the CPDAG already orients. One naming under one presentation order comprises 150 items: the 120 open items and the 30 control items.

Prompt. An item names the network's domain in one phrase, lists the variables of the component by published code or by scrambled label, states that some pairs are directly related but that the data cannot tell which of each pair causes the other, and lists those pairs. It then asks for a causal order of the variables, earliest cause first, omitting any variable the model cannot place, and for each pair the direction that order implies or `DECLINE`, with a confidence between 0 and 1. The order is the primary instrument. A stated direction that contradicts the model's own order is counted and the order is kept, a declined pair is an answer, and a pair the answer never mentions is recorded as not reached.

Models and decoding. Table 14 lists the ten conditions of Section 6.3. Each item was put to the model once per condition as a single message and decoded greedily, at temperature zero.

Table 14: Elicitation conditions. The first row is the primary model. The first eight conditions use the published variable codes and are pooled in Section 6.3; the last two use scrambled labels.

model	names	presentation order
Qwen2.5-7B-Instruct	published	first
Qwen2.5-7B-Instruct	published	second
Llama 3.1 8B	published	first
Gemma-3-27B-it	published	first
Qwen2.5-32B-Instruct	published	first
Llama-3.3-70B-Instruct	published	first
Qwen2.5-72B-Instruct	published	first
Qwen2.5-72B-Instruct	published	second
Qwen2.5-7B-Instruct	scrambled	first
Qwen2.5-72B-Instruct	scrambled	first

Repair. For each network the answered orientations are closed under Meek's rules against the CPDAG. When no consistent closure exists, the state is repaired as an analyst would repair it, by dropping orientations in increasing order of stated confidence until the closure succeeds, and the number dropped is recorded. All radii are computed on the repaired state. For the primary model the repair is needed on 6 of 32 networks and drops 43 of 214 orientations. It drops 57 of 196 for the same model under the second presentation order, 41 of 264 for Llama 3.1 8B and 12 of 163 under

scrambled names. The primary model keeps 171 of the 369 open pairs and declines 137; 53 of the 171 disagree with the data-generating graph, a network mean of 0.325 $[0.210, 0.449]$.

Matched states and per-query records. Three further states are built on the primary model's edges and scored on the same queries: a random state, which orients the edges uniformly among the choices that admit a consistent closure, in three replicates; the empty state, which asserts nothing; and, as a reference only, the state that copies the data-generating orientations. The release records each query under each condition on the 27 networks: 528 queries per condition and $1{,}444$ over the random replicates. Of the 528 queries, the primary model's knowledge commits a set on 229; on 155 it orients the component so that the outcome cannot descend from the treatment, and 144 are not amenable to adjustment. For each query, the release records what the analyst can compute without the data-generating graph: the analyst's graph, the committed set and whether one exists, the claim certificate searched to depth three with r_{claim} and the named orientations, the radius and the fraction φ_1. It records separately, for scoring only, what uses the data-generating graph: the agreement of each asserted orientation, whether the committed set holds there, and the effect and realised estimation error under the linear-Gaussian model.

Scrambled names and memorisation. Table 15 compares each model on published and scrambled names over the networks both versions cover. No difference that would indicate recall of the published networks has an interval excluding zero: the committed set does not fail more often under scrambled names, and the share of disagreeing orientations does not rise. Where both versions give an exact certificate, their claim radii agree as often as those of two random replicates. Asked directly for the source publication of each network from its variable list, Qwen2.5 7B and Llama 3.1 8B named it on none of 29 networks. The scrambled-name control therefore detects no memorisation at 7B or at 72B. The one difference whose interval excludes zero runs against recall: at 72B, published names give a larger share of certificates at $r_{\text{claim}} = 1$, so the familiar names yield more fragile sets. The converse check is weak as well: against $|\mathcal{K}|$-matched random orientations of the same pairs (20 seeds), the primary model's claim accuracy is 63.8% $[49.0, 78.1]$ against 56.1% $[51.5, 60.5]$, a paired per-network difference of $+0.04$ $[-0.10, 0.17]$, and $+0.09$ $[-0.02, 0.18]$ pooled over the eight real-name conditions. The elicited states are therefore used as realistic, varied knowledge inputs for the diagnostic, not as evidence that the models supply accurate domain knowledge; the diagnostic itself never reads the data-generating graph.

Table 15: Published minus scrambled names, primary model family, over the networks both versions cover (nets); the first four rows are differences. 95% network-clustered bootstrap intervals, $10{,}000$ resamples. The share of disagreeing orientations is given as a network mean and per query. Overlap is the Jaccard index of the queries that each version certifies, or certifies exactly; the last rows give the same quantities between two random replicates. The direct question about the source was put to the 7B and 8B models; Llama 3.1 8B also named the source on 0 of 29 networks.

quantity	7B	nets	72B	nets
set fails	$+0.009$ $[-0.149, 0.169]$	25	$+0.057$ $[-0.090, 0.210]$	26
disagreeing share, network mean	-0.029 $[-0.150, 0.089]$	30	-0.088 $[-0.241, 0.065]$	27
disagreeing share, per query	-0.037 $[-0.184, 0.109]$	23	$+0.004$ $[-0.136, 0.147]$	23
share with $r_{\text{claim}} = 1$	$+0.052$ $[-0.079, 0.183]$	25	$+0.154$ [0.065, 0.247]	26
exact r_{claim} agrees	0.841 [0.729, 0.926]	17	0.885 [0.731, 0.975]	19
(queries)	$53/63$		$77/87$	
agreement minus random replicates	$+0.079$ $[-0.045, 0.220]$	13	$+0.115$ $[-0.035, 0.259]$	14
overlap of certified queries	0.444 [0.352, 0.535]		0.529 [0.429, 0.635]	
overlap of exactly certified queries	0.283 [0.175, 0.386]		0.363 [0.223, 0.508]	
source publication named	$0/29$		–	
two random replicates: exact r_{claim} agrees $82/95 = 0.863$ [0.739, 0.957] (19 networks); overlap 0.638 [0.537, 0.739] of certified and 0.503 [0.371, 0.635] of exactly certified queries				

Stability of the answers. Re-querying the primary model on its 150 released items, with the same weights at temperature zero, returns the same answer for every asked pair on 108 items, 72.0% $[64.1, 78.3]$ over 32 networks, and on 76.9% $[71.5, 82.5]$ of the 575 pairs; the second presentation order and the scrambled names give item agreement of 73.3% $[64.3, 80.6]$ and 71.3% $[64.1, 77.5]$. Decoding was greedy at temperature zero ("Models and decoding" above); we do not record the

serving engine, numerical precision, batch composition or random seeds used at inference time. Greedy decoding at temperature zero is not a guarantee of bit-for-bit reproducibility under most serving stacks, where floating-point non-associativity interacting with non-deterministic GPU kernels and batch composition can change which token is selected; because the serving engine, numerical precision, batch composition and random seeds were not recorded, the cause of the disagreement above cannot be identified from the information available, and we cannot rule out genuine answer instability. This re-query does not by itself measure the stability of the model's elicited knowledge. The released states fix the knowledge, and every result on elicited knowledge in this paper is computed from them: analyses conditional on the released knowledge states are fully reproducible, whereas their acquisition (the elicitation itself) is not guaranteed to be.

Scope with respect to prior work. Elicited orientations on published networks already exist. CausalBN-Bench scores language-model edge directions against fourteen bnlearn graphs and releases its prompts and labels (Zhou et al., 2026), and Kitson & Constantinou (2026) score orientation confidences under disguised variable names against the reference graphs of 26 networks; we use scrambled names as the established control they are. The states released here go further: they restrict elicitation to the orientations the data cannot fix and pair each state with its consequence for a committed adjustment set, and the matched random and empty states on the same edges give each result a control.

Licences and contents of the release. The release, in the supplementary material, contains the questionnaire, the prompts, the answers of the ten conditions, the repaired states, and the per-query records, which include the matched random (three replicates) and empty states. Network structures are not redistributed; each network is identified by its published name and source. Model weights are available from their providers under their own licences. All knowledge in the release was elicited from the open-weight models of Table 14.