Diagnosis as Percolation on an Evidence Graph: Root-Cause Recoverability in Network Fault Diagnosis

Anonymous authors
Paper under double-blind review

Abstract

Automated diagnosis under partial observation recurs across computing and engineering: an agent must infer an unobserved root cause from an incomplete set of symptoms, and sometimes gauge how far to trust that inference. We argue that *random-graph theory* supplies both a model and a set of operational primitives. We cast diagnosis as *percolation on a partially observed evidence graph*, whose vertices are observations, entities, hypotheses and candidate causes and whose edges are derived deterministically from data; a diagnosis is then a connected explanatory subgraph, and recoverability corresponds to whether the evidence has *percolated* into one giant cluster rather than scattering into fragments. Our primary contribution is to *network* fault diagnosis, in the lineage of the Plume and PROBE systems: we work the abstraction out in full on an ensemble of large language models (LLMs) that diagnoses faults from textualized 802.11 packet captures, where an ensemble-plus-reconciliation pipeline lifts weighted evidence F_1 from $\approx$0.85 to $\approx$0.96 yet provides no calibrated notion of trust. Probing whether connectivity supplies that missing signal yields a cautionary result: post-hoc connectivity correlates with correctness (Spearman ρ up to 0.67), but under capture-clustered controls, degree-preserving negative-control nulls and multiple-comparison correction, no structural signal retains value beyond capture size. Connectivity is thus a portable, theoretically grounded substrate for fault *recoverability* and root-cause *localization*, but not a reliable proxy for answer-level trust. A controlled benchmark recovers the predicted critical density $\phi_c \approx 1/c$, where a connectivity-aware personalized-PageRank localizer tracks a Bayesian oracle while degree heuristics collapse under decoy hubs; and studies in five further domains, with a substrate-independent threshold analysis, indicate that the percolation *principle* transfers beyond networking.

1 Introduction

1.1 Automated diagnosis under partial observation

Diagnosis (inferring an unobserved cause from observed effects) is a foundational task across computing, engineering and medical sciences. A network operator infers a misconfiguration from a packet trace; a site-reliability engineer localizes the failing microservice behind a cascade of alarms; a grid operator anticipates which transmission lines will cascade; a clinician reasons from a set of symptoms to a disease. These problems share three structural features that make them difficult and that motivate a unified treatment.

First, they are abductive and many-to-one: numerous latent causes can produce overlapping observable effects, so an effect set rarely identifies a unique cause. *Second, observation is partial and non-uniform:* only a fraction of the relevant evidence is ever recorded (a monitoring buffer captures a short window, telemetry is sampled, a patient reports a subset of symptoms), so the diagnostic substrate is a *partially observed* object rather than a complete one. *Third, and increasingly, the diagnostic agent must be trustworthy:* as machine-learning systems and large language models (LLMs) are deployed to produce diagnoses automatically, a calibrated estimate of *when to trust* a diagnosis is as important as the diagnosis itself, since an over-confident wrong answer is more harmful than an abstention.

We contend that these features are precisely those studied by *random-graph theory* and *percolation* (Erdős & Rényi, 1960; Newman, 2003). If the evidence for a case is organized as a graph whose vertices are observations, entities, symptoms, hypotheses, and candidate causes, then a diagnosis is a *connected explanatory subgraph* linking observed effects to a cause, and the central percolation question "has a giant connected component emerged?" becomes a question about recoverability: if the observed evidence *percolates* into one connected cluster reaching a candidate cause, the cause is *recoverable*; if it stays fragmented, no method can confidently identify it. This paper develops that correspondence into a model, a family of computable signals, and a body of cross-domain evidence—and, equally importantly, maps its boundary: connectivity is *not* a reliable proxy for whether a given answer is *trustworthy* once observed-evidence volume is accounted for (Section 5).

1.2 A demanding instance: LLM diagnosis of packet captures

We ground the abstraction on a concrete and unforgiving instance: diagnosing faults directly from IEEE 802.11 packet captures (PCAPs). A monitoring access point retains a rolling buffer of recent frames; when an anomaly is detected, the buffer is exported for analysis. Two complementary systems define the state of the art we build upon. *Plume* is a network-native foundation model that learns representations of 802.11 traces through protocol-aware tokenization (Pradhan et al., 2026); *PROBE* is a multi-stage, evidence-grounded ensemble pipeline that renders a textualized PCAP into a frame-cited diagnosis accompanied by a deterministic reliability score (Henry et al., 2026). Both expose each capture to an LLM and ask it to identify the fault; both are evaluated against a *golden set* in which human experts record a chain-of-thought (Wei et al., 2022) and a structured verdict (the issue, the responsible entity, and the citing frames), enabling automatic scoring by a weighted evidence-overlap F_1.

This instance is demanding for two reasons. A single LLM pass is unreliable: instruction-tuned models are post-trained toward caution and tend to commit only when an explicit error message is present, missing subtle evidence distributed across many frames. An ensemble of stochastic drafts followed by a *reconciliation* step recovers much of the lost signal, an instance of the variance reduction underlying self-consistency decoding (Wang et al., 2023) and the classical bias–variance decomposition (Geman et al., 1992). Table 7 reproduces the headline ablation from PROBE (Henry et al., 2026): the ensemble-plus-reconciliation configuration (C) attains weighted evidence $F_1 = 0.964$, versus 0.842–0.871 for leaner variants. Reconciliation is thus the dominant accuracy lever but, as we discuss next, it leaves the question of *trust* unanswered.

1.3 The reliability gap left open by reconciliation

Reconciliation determines *what* the consensus diagnosis is, but not *how much it should be trusted*, and the latter is essential for automated deployment. Two natural sources of confidence prove inadequate. The model's *self-reported* confidence is uninformative: 74 of 104 golden cases report a confidence of exactly 0.95, and stratifying accuracy by self-reported confidence yields a non-monotone curve (Figure 5a), the familiar miscalibration of modern neural predictors (Guo et al., 2017). A hand-weighted deterministic composite of run-to-run stability and evidence-validity terms is better behaved (accuracy increases monotonically with the score) but its weights were chosen manually and it lacks a principled, structural foundation. We therefore seek a reliability signal that is (a) *intrinsic* to the evidence a diagnosis cites, (b) *deterministic* and inexpensive to compute, and (c) *theoretically grounded* in a model of why connected evidence indicates a reliable inference.

1.4 Contributions

1. **A percolation model of diagnosis** (Section 3): an evidence graph with a generative partial-observation process, four connectivity *order parameters* grounded in generating-function percolation theory and personalized-PageRank relevance, and a bias result (Proposition 1) explaining why a *closed-world* noisy-OR posterior is fragile under under-observation—a misspecification of the observation process, not a failure of Bayes.

2. **A falsification of the bridging hypothesis** (Section 4): on 200 real reconciliation pairs, repair does *not* significantly raise the giant-component fraction; its measurable effect is *orphan pruning* ($0.149 \to 0.069$,

p=0.008), while post-reconciliation connectivity correlates with correctness (ρ up to 0.67)—an association we trace largely to capture size.

3. **A rigorous negative result on structural reliability signals** (Section 5): across six candidate signals, with capture-clustered inference, degree-preserving nulls, and multiple-comparison correction, *none* retains value once capture size and trivial citation statistics are controlled—the apparent "structure-as-trust" signal is a capture-size artifact.

4. **A controlled percolation benchmark** (Section 6) whose empirical critical density approaches $\phi_c \approx 1/c$ and isolates an *algorithmic detectability gap*: a connectivity-aware localizer tracks a Bayesian oracle while degree-only heuristics collapse.

5. **Cross-domain generality** (Section 7): a reverse-PageRank localizer attains 89% top-1 on a *real* microservice chaos corpus (Train Ticket); power-grid cascades and three controlled simulations (software fault localization, multi-hop RAG, security attack-path) reproduce the recoverability/localization result; and bipartite, biased-observation, and phase-boundary analyses show the threshold itself is substrate-independent. We claim transfer of the *principle*, not of a tuned system.

All experiments are deterministic given a random seed and use only public data.

Scope and positioning. This is first a paper about *network* fault diagnosis, in the lineage of Plume (Pradhan et al., 2026) and PROBE (Henry et al., 2026): the real ground truth and primary claims concern 802.11 packet-capture diagnosis. The cross-domain studies (Section 7) show the percolation *principle* is substrate-independent—a property inherited from the mathematics, not from any one dataset—and we are explicit about which findings rest on *real* data and which are *controlled simulations*.

2 Background and Related Work

We summarize the theory we draw on, stating the results in the form we will use them, and then position the present work against prior literature on LLM reliability and cross-domain diagnosis.

Random graphs and percolation. The Erdős–Rényi ensemble $G(n, p)$ exhibits a phase transition: a giant connected component (GCC) of size $\Theta(n)$ emerges once the mean degree exceeds unity (Erdős & Rényi, 1960). For a general degree distribution the standard generating-function derivation (Appendix A.1) gives the bond-percolation threshold

$$\phi_c = \frac{1}{G_1'(1)} = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}, \tag{1}$$

where G_1 is the excess-degree generating function and $\langle k \rangle$, $\langle k^2 \rangle$ are the first two degree moments; for a Poisson degree distribution of mean c this is the clean prediction $\phi_c = 1/c$ tested in Section 6. Two consequences recur throughout: near ϕ_c the *susceptibility* χ (mean size of the non-giant component containing a random vertex) diverges, giving a finite-sample locator of the transition; and Equation (1) explains the robustness asymmetry—deleting vertices in descending centrality order maximally reduces $\langle k^2 \rangle$ and destroys the GCC far faster than random deletion (Albert et al., 2000; Cohen et al., 2000), a property we exploit as a diagnostic in Section 7.

Random walks, centrality, and relevance. PageRank ranks vertices by the stationary distribution of a restarting random walk (Page et al., 1999); its personalized variant (Haveliwala, 2002) restarts from a query set F and so measures *relevance to* F, which we use as a domain-agnostic prior over candidate causes. Edge betweenness centrality (Brandes, 2001) identifies structural bottlenecks and ranks lines likely to carry power-grid cascades. Formal definitions are in Section 3.4.

Probabilistic diagnosis and the noisy-OR model. Abductive diagnosis has a long history in probabilistic reasoning (Pearl, 1988). The noisy-OR model and QMR-DT (Shwe et al., 1991) assume each cause independently activates each finding, yielding a tractable likelihood whose exact posterior we use as a

Bayesian *oracle*; Proposition 1 shows its treatment of *absent* findings is what makes it fragile under partial observation, motivating positive-evidence structural priors.

LLM reliability, calibration, and selective prediction. Self-consistency decoding (Wang et al., 2023), retrieval-augmentation (Lewis et al., 2020), and chain-of-thought prompting (Wei et al., 2022) reduce error and elicit citable reasoning; the ensemble-plus-reconciliation pipeline of PROBE (Henry et al., 2026) is in this lineage. Because neural predictors are typically miscalibrated (Guo et al., 2017), a separate intrinsic trust signal is needed; we adopt selective classification (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017) and its risk–coverage curve as our yardstick (Appendix A.5).

Cross-domain diagnosis benchmarks. We evaluate portability on problems with distinct graph topologies: microservice RCA on directed service-call graphs (Train Ticket (Zhou et al., 2021)), power-grid cascading failure on undirected bus–branch networks (PowerGraph (Varbella et al., 2024)), and bipartite QMR-DT diagnosis (Shwe et al., 1991), synthetic and real (Kaushil et al., 2019; Wang et al., 2008). Treating these with a single structural prior is, to our knowledge, novel.

3 Diagnosis as Percolation on an Evidence Graph

3.1 The evidence graph and its observation process

Definition 1 (Evidence graph)**.** For a single case, let the vertex set $V = O \cup E \cup S \cup H \cup R$ comprise *observations* O (e.g. packet frames, log lines, telemetry points), *entities* E (endpoints, services, components), *symptoms* S, *hypotheses* H, and candidate *root causes* R. The *latent* evidence graph $G^\star = (V, A^\star)$ has an edge set $A^\star$ generated deterministically by domain rules of three kinds: *temporal* adjacency between observations, *co-membership* in a shared conversation/entity, and *citation* links connecting a diagnosis to the observations and entities it references. A diagnosis that cites a vertex set $C \subseteq V$ induces the *cited subgraph* $G[C]$, the subgraph of $G^\star$ on C.

Two design choices make Definition 1 portable. First, the edge rules are *deterministic* functions of the data, so the graph, and every quantity derived from it, is reproducible and independent of the diagnostic model. Second, the rules are *local*: each requires only adjacency or shared membership, which every domain we study supplies. In the PCAP instance the pipeline exposes only the rendered text and the set of cited frame numbers, so we use a *time-and-conversation* rule: cited frames i, j are adjacent iff $|i - j| \leq 1$ in capture order (a unit-width temporal bridge) or they belong to the same endpoint conversation. Section 7 instantiates the same definition on directed service graphs, undirected bus–branch networks, and bipartite cause–finding graphs.

We model partial observability explicitly, since it is the crux of the problem (Section 1.1).

Assumption 1 (Partial observation)**.** The diagnostic agent does not see $G^\star$ but a random subgraph G_ϕ obtained by retaining each observation independently with probability $\phi \in (0, 1]$ and keeping an edge iff both endpoints are retained. Equivalently, G_ϕ is a site-percolated copy of $G^\star$ with occupation density ϕ.

Assumption 1 is a missing-completely-at-random idealization of buffered capture, telemetry sampling, and incomplete symptom reporting. It is the bridge between the engineering problem and percolation theory: the visible evidence is literally a percolated graph, so the percolation threshold ϕ_c of Equation (1) predicts the observation density below which the evidence fragments and diagnosis becomes information-theoretically hard.

3.2 Connectivity order parameters

We summarize the connectivity of a cited subgraph $G[C]$ by four scalar *order parameters*:

$$\Phi_{\text{gcc}}(C) = \frac{|\text{largest connected component of } G[C]|}{|C|} \qquad \text{(giant-component fraction)}, \tag{2}$$

$$\kappa(C) = \#\{\text{connected components of } G[C]\} \qquad \text{(fragmentation)}, \tag{3}$$

$$\iota(C) = \frac{\#\{\text{isolated vertices of } G[C]\}}{|C|} \qquad \text{(orphan fraction)}, \tag{4}$$

$$\delta(C) = \frac{2\,|A(G[C])|}{|C|\,(|C|-1)} \qquad \text{(edge / span density)}. \tag{5}$$

Φ_{gcc} is the finite-sample estimator of the percolation order parameter $S = \lim_{n\to\infty} |\text{GCC}|/n$, the largest root in $[0,1]$ of $S = 1 - G_0(1 - S\,\phi)$ under Assumption 1 (Newman, 2003). The fragmentation κ and orphan fraction ι measure sub-criticality from complementary angles: κ counts the components a correct explanation should fuse, while ι isolates the "dangling" evidence a diagnosis cited but failed to connect to anything else. The density δ is the realized fraction of possible internal edges and behaves, empirically, as the most informative single summary (Section 4).

3.3 The percolation hypothesis

Hypothesis 1 (Diagnosis-as-percolation)**.** A trustworthy diagnosis corresponds to a cited subgraph that has *percolated*: a single explanatory component spans the relevant evidence ($\Phi_{\text{gcc}} \to 1$, $\kappa \to 1$, $\iota \to 0$). Under-observation ($\phi < \phi_c$) or an incorrect diagnosis leaves the subgraph *sub-critical*: fragmented, with many orphans.

Hypothesis 1 has a precise prediction under Assumption 1 and eq. (1): as ϕ crosses ϕ_c from below, the probability that the observed evidence supports a connected explanation rises sharply from near zero to order one, with a transition that sharpens as the system grows. Section 6 confirms this prediction in a controlled benchmark, and Section 4 tests its corollary that more-connected cited subgraphs are more often correct.

3.4 Relevance as a damped random walk

The central localization primitive is *personalized PageRank* (PPR) (Page et al., 1999; Haveliwala, 2002), seeded at the observed evidence set $F \subseteq V$. With damping $d \in (0,1)$, restart distribution $\mathbf{s}$ uniform on F, and P the row-stochastic adjacency of G_ϕ, the relevance vector π is the unique solution of

$$\pi = (1-d)\,\mathbf{s} + d\,P^\top \pi. \tag{6}$$

Because $d\,P^\top$ has spectral radius $d < 1$, Equation (6) admits the convergent Neumann-series solution

$$\pi = (1-d) \sum_{t \geq 0} d^{\,t}\,(P^\top)^t\,\mathbf{s}, \tag{7}$$

which exposes the prior's meaning: π_v is a length- and multiplicity-weighted count of walks from the observed evidence to v, discounted by d^t with path length t. A candidate cause is therefore scored by how *reachably relevant* it is to the present evidence, a quantity that ignores absent evidence entirely. This property is what underlies the robustness result below. Candidate causes are ranked by π; when causes are known to lie *upstream* of effects (directed graphs), we transpose the walk, which is decisive for microservice RCA (Section 7).

3.5 The noisy-OR posterior and its closed-world fragility

In bipartite cause→finding settings (the QMR tradition (Shwe et al., 1991; Pearl, 1988)), the exact single-cause noisy-OR posterior (Equation (10), Appendix A.3) is a Bayesian *oracle* at full observation. Its *closed-world* use under partial observation—scoring every unreported finding as *absent*—is the source of its fragility:

under Assumption 1 a finding generated by the true cause but not observed is treated as negative evidence, penalizing the true cause for a symptom it actually produced. Proposition 1 makes this precise; it is a misspecification of the observation process, not a failure of Bayes.

Proposition 1 (Missingness bias of the closed-world noisy-OR)**.** *Fix a true cause $r^\star$ generating finding set $F^\star$, let each finding in $F^\star$ be observed independently with probability ϕ (Assumption 1), and write $p_f := \mathbb{P}[f{=}1 \mid r^\star]$ for the per-finding activation probability. If unobserved findings are scored as* absent *(the closed-world reading of Equation* (10)*), the expected change in the log-likelihood of $r^\star$ relative to full observation is*

$$\mathbb{E}\big[\log \mathbb{P}_{\mathrm{cw}}[F_\phi \mid r^\star] - \log \mathbb{P}[F^\star \mid r^\star]\big] \;=\; -(1-\phi) \sum_{f \in F^\star} \mathrm{logit}(p_f). \tag{8}$$

The bias is a genuine downward loss *for individually probable findings ($p_f > \frac{1}{2}$) and grows as $\phi \to 0$. A score that depends only on observed-present evidence (explained-overlap, or PPR seeded on F_ϕ) contains no term indexed by $f \notin F_\phi$ and incurs no such penalty.*

Proof in Appendix A.2. A correctly specified posterior that *marginalizes* unobserved findings carries no such term; Section 7.2 confirms it empirically (the closed-world posterior collapses under heavy under-observation while the marginalized and observation-aware posteriors track the cheap positive-evidence priors, Figure 8). The lesson for buffered, partially observed diagnosis is to model the observation process, not to treat unobserved evidence as negative evidence.

4 Reconciliation as Structural Repair: Testing the Bridging Hypothesis

We apply the model of Section 3 to the motivating PCAP instance. Reconciliation is the dominant accuracy lever (Table 7), and Hypothesis 1 suggests a mechanism: it might *bridge* fragmented cited evidence into one component, raising the giant-component fraction Φ_{gcc}. We test this on 200 paired (pre/post-reconciliation) cases from two configuration families (E→F, B→C, 100 pairs each; pre-reconciliation approximated by the no-reconcile sibling), building the cited subgraph (Definition 1) and its four order parameters and assessing pre/post shifts (Wilcoxon) and structure–quality association (Spearman vs. weighted F_1).

Connectivity is unchanged; orphans are pruned. The giant-component fraction Φ_{gcc} does *not* shift across reconciliation (Wilcoxon $p = 0.67, 0.71$); what moves is the orphan fraction ι (in E→F, 0.149→0.069, $z = -2.66$, $p = 0.008$). Reconciliation acts not as a bridge-builder but as a *pruner of dangling evidence*, refuting the strong bridging hypothesis.

Post-reconciliation structure predicts correctness. While the pre/post shift is small, the *absolute* post-reconciliation structure is strongly associated with quality: span density δ correlates with F_1 at $\rho = 0.64$–0.67, Φ_{gcc} at 0.59–0.61, and fragmentation κ negatively at -0.58–-0.62 (Figure 4). A connected, dense cited subgraph is thus a correlate of a correct diagnosis, consistent with Hypothesis 1—but this is *associational* on synthetic families: as Section 5 shows, such associations are substantially driven by capture size, so we do not claim connectivity as a size-independent predictor. The robust, size-insensitive finding here is the *mechanistic* one: reconciliation prunes orphan evidence rather than bridging components.

5 Structural Reliability Signals: A Controlled Negative Result

It is tempting to read Hypothesis 1 as a recipe for *trust*: if a correct diagnosis percolates, then connectivity or its stability should predict correctness on real data, yielding a cheap, model-agnostic confidence signal. From the model of Section 3 we derive six candidate reliability signals and ask, on the 103-case golden set (base accuracy 0.767), whether any of them carries information about correctness *beyond trivial capture size*. The honest answer is no. Several signals correlate with correctness in-sample, but once we control for capture size and elementary citation statistics, apply capture-level permutation tests, compare against degree-preserving negative-control nulls, and correct for multiple comparisons, *none* survives. We report this as a cautionary negative result: on this dataset, the apparent "structure predicts trust" signal is an artifact of how much evidence a capture contains.

5.1 In-sample associations (uncontrolled)

At the case level several candidates correlate with correctness (Table 13): fragmentation ($\rho = +0.25$, $p = 0.011$), the agreement of two stochastic views of the cited evidence (edge-Jaccard $\rho = 0.27$, largest-component Jaccard $\rho = 0.31$), and mean PageRank centrality (Section 3.4) of cited frames ($\rho = 0.22$). Used as abstention features they improve risk–coverage over the pipeline's own confidence (AURC = 0.129 vs 0.159, Figure 5b), and the gain replicates out of sample (median held-out AURC 0.128 vs 0.160; the structural gate wins on 97% of 2,000 splits). So there is a genuine, cheap triage signal. An earlier version of this work stopped here and called these "effective *structural* signals". That label is wrong: Sections 5.2 and 5.3 show the gain is a capture-*size* effect, not evidence-graph topology.

5.2 The associations are a capture-size artifact

The golden set is *clustered*: each capture contributes several cases, and captures vary widely in size. Larger captures cite more frames, are easier to diagnose, and mechanically inflate every connectivity statistic, so an uncontrolled case-level correlation conflates "structure" with "amount of evidence". We therefore recompute each association as a partial Spearman correlation controlling for capture size, citation count, and mean citation gap, with significance assessed by a *capture-level* permutation test (shuffling correctness within captures) and Benjamini–Hochberg correction over all 14 candidate constructions (Table 14).

The signal collapses. Fragmentation falls to $\rho = +0.14$ ($p = 0.12$), centrality to $\rho = +0.03$ ($p = 0.35$), and connectivity-robustness to $\rho = +0.07$ ($p = 0.55$). Two degree-preserving *negative controls* confirm the diagnosis: the "excess" of Φ_{gcc} over a rewired null carries no signal ($\rho = +0.04$, $p = 0.73$), i.e. the raw connectivity correlation is entirely explained by size and degree sequence. The two agreement metrics survive size control at the raw level (largest-component Jaccard $\rho = +0.275$, $p = 0.026$; edge-Jaccard $\rho = +0.252$, $p = 0.025$) but *not* multiple-comparison correction (BH $q = 0.18$ for both). *No candidate survives BH-FDR.* The working abstention gate of Section 5.1 is therefore best read as a capture-size/separability proxy (larger, cleaner captures are easier and the fragmentation score tracks that) rather than as a measurement of evidence-graph coherence.

5.3 No incremental value, and a confirmatory test fails

A grouped (leave-one-capture-out) logistic gate tells the same story: adding any single signal to a size-plus-citation baseline yields no robust out-of-sample gain in correct/incorrect AUROC, the largest nominal increment (largest-component Jaccard, ΔAUROC = 0.15) again failing FDR. Because the agreement metrics were the only candidates with raw $p < 0.05$ after size control, we subjected them to a confirmatory test on the n=76 cases with a second-opinion view. It fails both pre-specified checks. **Specificity:** once we control for plain *set-overlap* (node-Jaccard, a scalar the pipeline already records), the *structural* agreement loses significance ($\rho : 0.28 \rightarrow 0.20$, $p : 0.025 \rightarrow 0.088$)—the graph framing adds nothing beyond "do the two views cite the same frames". **Replication:** the incremental AUROC is positive in only 77% of 400 held-out capture splits, below our 80% bar. The test is also badly underpowered (collinear metrics, VIF ≈ 3.4; detecting the observed residual $\rho \approx 0.20$ would need $\sim$650 captures, $\approx 10\times$ this corpus). We therefore fold no structural term into the confidence score and draw no positive claim.

Takeaways. On this golden set, structural connectivity and its perturbation stability provide *no* reliability information beyond capture size and trivial set-overlap; null-model normalization of Φ_{gcc} adds nothing (cited subgraphs sit near a connectivity ceiling), and motif surprise vanishes under size control. The model's self-reported confidence is non-discriminative (AUROC 0.49, Figure 5a); the deterministic composite orders better (AUROC 0.60) but is badly under-confident, useful for *ranking*, not as a probability. These negative results sharpen the thesis: the percolation prior earns its keep in *recoverability* (Section 6) and *localization* (Section 7), not as a drop-in trust score on an already-strong pipeline.

6 A Synthetic Percolation Benchmark and the Detectability Threshold

The PCAP study establishes correlation between connectivity and correctness but cannot isolate the underlying phase transition, because the ground-truth latent graph and the observation density ϕ are not controllable. We therefore construct a synthetic benchmark in which both are known, allowing a direct test of Hypothesis 1 and Equation (1). Each *incident graph* places a root cause R that seeds L causal paths of intermediate symptoms toward an alarm A, embedded in a background Erdős–Rényi graph of mean degree c. Under Assumption 1 each vertex is observed independently with probability ϕ. We then measure two distinct quantities: whether the incident is *recoverable* (an observed path connects R to A), and whether a given localizer actually *identifies* R.

The empirical critical density approaches theory. The susceptibility (the mean size of the non-giant component containing a random vertex, Section 2) peaks at an empirical critical density that approaches the theoretical value $\phi_c = 1/c = 0.25$ from above as the system grows: $\hat{\phi}_c \approx 0.38, 0.36, 0.32, 0.28$ for $N = 200, 500, 1000, 2000$ (Figure 6a). This monotone approach is the expected finite-size scaling behavior and confirms that the benchmark reproduces the percolation transition predicted by Equation (1); via Molloy–Reed the same critical point holds for any configuration-model degree distribution, so the transition is generic rather than an Erdős–Rényi artifact. Recoverability crosses 50% near $\phi \approx 0.5$.

An algorithmic detectability gap above the percolation threshold. Recoverability is necessary but not sufficient for diagnosis: even when an R–A path is observed, a localizer must select R from among many candidates. To stress this distinction we introduce *decoy hubs*, vertices of high degree that are unrelated to the incident and clustered in a reserved region of the graph, and compare localizers (Table 8, Figure 6b). The connectivity-aware personalized PageRank of Equation (6) tracks the Bayesian oracle closely (peak top-1 accuracy 0.625 vs. 0.675 at $N = 500$), whereas *degree-only* heuristics collapse (global degree 0.213, component degree 0.20) because they are captured by the decoys; only PPR and the oracle ever exceed 50% accuracy (finite p_{50}). The benchmark thus isolates a second threshold beyond percolation: an *algorithmic* detectability gap separating relevance-weighted walks from degree heuristics, which we observe again in every cross-domain experiment.

7 Cross-Domain Evidence: the Principle Beyond Networking

The motivating instance and the controlled benchmark both suggest that relevance-weighted connectivity on a partially observed graph is a general diagnostic primitive. We probe that by applying the *same* model (Definition 1) and localizer (Equation (6)) to six external domains chosen for topological diversity—directed service graphs, undirected infrastructure networks, bipartite cause–finding graphs, program-dependence graphs, retrieved-evidence graphs, and security attack graphs—without per-domain tuning beyond edge orientation. The localization result is positive across all six; whether structure also beats a size or separability baseline is domain-dependent (Table 6). Microservice RCA and power grids use *real* data; software fault localization, multi-hop QA, and security attack-paths are *controlled simulations*—all evidence that the percolation *principle* is not networking-specific, not competitive entries against each field's tuned state of the art. We develop the two domains closest to networking (microservice RCA and bipartite QMR diagnosis) here and report the other four in Appendix B; the synthesis below draws on all six.

7.1 Microservice root-cause localization (Train Ticket chaos-injection)

We use 451 fault-injection incidents on the Train Ticket microservice benchmark (Zhou et al., 2021) (64 services), in which faults are injected with Chaos Mesh across 27 resource-, network-, HTTP-, JVM-, and DNS-level fault types.[1] Each incident provides a directed service-call/propagation graph over span/pod/container/service components, a set of observed alarm components, and the ground-truth injected

[1]This is a chaos-engineering corpus on Train Ticket; it is *not* the OpenRCA benchmark (Xu et al., 2025), which covers Bank/Market/Telecom cloud incidents. We keep the directory label `openrca2` only for artifact reproducibility and claim no identity with OpenRCA.

(root-cause) service (mean 5.1 candidate services after canonicalizing spans and containers to their parent service). Localizers rank services; we report top-k accuracy, MRR, and mean percentile (Table 15).

The best localizer is a *reverse*-PageRank prior that propagates mass *up* the causal edges toward upstream sources, attaining 89.1% top-1 and a 96.1st mean percentile, far above forward PageRank (12.6%) and random (35.0%). One caveat keeps the headline honest: this reverse prior is *evidence-blind* (uniform restart, never reading the alarms), and seeding the walk at the observed alarms attains only 53.0%. On this corpus it is edge *orientation*, not evidence-seeding, that carries the signal—the propagation graphs are built around the injected root, so it sits near the structural source and ranks highly under any reverse prior (Table 15). The result thus confirms that orientation matters (the directed form of Equation (7)); a cleaner demonstration of *evidence-seeded* relevance appears in the RAG domain (Appendix B.5), where seeding is decisive.

Comparison to domain-native baselines (PetShop, RCAEval). The Train Ticket corpus ships graphs but no metric time series, so it cannot drive the metric-based causal RCA methods that define the state of the art. On two public, metric-rich benchmarks that can—PetShop (Hardt et al., 2024) and RCAEval RE2 Train Ticket (Pham et al., 2025)—our training-free relevance prior is a *competitive* zero-training localizer: it stays orientation-sensitive on PetShop and matches the mid-tier causal baseline CIRCA on RCAEval (AC@1 0.29 vs. 0.32, far above MicroCause/RCD), but it does *not* beat a well-tuned causal/statistical detector such as BARO (0.67). On metric-rich data a causal model localizes *which* node more accurately—the same structure-vs-separability pattern as software fault localization (Appendix B.4). Appendix B.2 gives the full setup, per-method tables, and the orientation analysis.

7.2 Bipartite disease–symptom diagnosis (QMR): synthetic and real

The third domain instantiates the bipartite causal setting of Equations (6) and (10) and provides the most direct test of Proposition 1. We evaluate both a controlled generative model and a templated real-world disease–symptom dataset.

Synthetic QMR. A generative noisy-OR model (300 diseases, 1000 findings, 1000 sampled cases) permits comparison of a graph-only PPR relevance prior against the exact noisy-OR posterior (Table 9). PPR attains 70.5% top-1 accuracy (versus 0.5% for random selection) and tracks the posterior's ranking closely (99.3rd percentile). A simple *explained-overlap* count (the number of observed findings adjacent to a candidate disease) attains 94.4%, exceeding PPR because high-degree "hub" findings dilute the relevance walk while leaving the overlap count unaffected. This is the bipartite instance of the dual role of high-degree nodes identified in Section 6.

Real disease–symptom data. We repeat the evaluation on a 41-disease, 131-symptom, 4,920-case disease–symptom dataset (Kaushil et al., 2019). We caution that this is a *templated* symptom-signature dataset (not a clinical patient cohort) and that the link weights $\hat{\mathbb{P}}[\text{symptom} \mid \text{disease}]$ are estimated *in-sample*, so the noisy-OR scores are optimistic upper bounds and the informative quantity is the *relative* robustness of the solvers under the sparsity sweep below. With that caveat, graph-only PPR localizes the true disease at 86.1% top-1 (33× the 2.6% chance rate) while raw degree is uninformative (2.4%): it is *relevance-weighted* connectivity, not degree, that carries the signal. The genuine hub symptoms (fatigue and vomiting each span 17 diseases) reproduce the hub-dilution effect, with PPR (0.861) just below overlap (0.972).

Robustness under partial observation confirms Proposition 1. Sweeping the fraction ϕ of symptoms reported (Figure 8) isolates the misspecification: the *closed-world* noisy-OR collapses from 1.000 at full observation to 0.359 at ϕ=0.2 (it charges the true disease for expected-but-unseen symptoms, Equation (10)), whereas the correctly specified posteriors of Remark 1 do not—the observation-aware model is the *most* robust (0.728) and the marginal model degrades gracefully (0.628). The cheap positive-evidence priors (overlap 0.653, graph-only PPR 0.603) need *no* knowledge of the observation process yet stay far above the closed-world model. Since heavy under-observation is the defining regime of buffered capture and sampled telemetry, the recipe is to model the observation process or use a positive-evidence relevance score, never to treat unobserved evidence as negative.

The QMR detectability collapse (Section 7.2) is not a separate phenomenon: specializing the percolation threshold to the bipartite (disease–finding) configuration model gives a closed-form ϕ_c^{bip} that a finite-size susceptibility sweep recovers from above, tying the collapse to the same theory as the one-mode benchmark (Appendix B.1).

8 Discussion and Conclusion

A single structural principle, and where it fails. Across a controlled benchmark and six external domains, one principle recurs: relevance-weighted connectivity on a partially observed evidence graph is an inexpensive, label-free, model-agnostic signal for *recoverability* (Section 6) and *localization* (Section 7). Equally important is where it *fails*: on the golden set it yields no reliability (trust) signal beyond capture size (Section 5), and in software fault localization the same order parameter predicts *whether* a fault is recoverable yet a graph-only walk loses to a separability-driven spectrum score at saying *which* statement it is (Appendix B.4). Three threshold extensions sharpen the positive half—the QMR collapse *is* the bipartite percolation threshold, which stays conservative under anomaly-biased observation and bounds a measured localizable phase boundary (Appendices B.1, B.7 and B.8).

The dual role of high-degree vertices. High-degree vertices aid localization when causal (cascading lines) and impede it when incidental (decoy hubs, hub findings, high-fan-out spans). Degree alone is thus an unreliable localizer; a walk that respects connectivity and, where available, edge *orientation* is the robust alternative—orientation alone explains the reverse prior's 89% top-1 on Train Ticket, while evidence-seeding adds most where the graph is not already built around the cause (multi-hop RAG, Appendix B.5).

Closed-world fragility, and the value of a negative result. The most consequential theoretical result is Proposition 1 (confirmed in Figure 8): the closed-world noisy-OR posterior, optimal at full observation, becomes *less* robust than cheap positive-evidence priors when evidence is missing (unobserved), not absent—a misspecification of the observation process, not a failure of Bayes (Remark 1). Our most carefully controlled finding is itself negative: the intuitive "structure predicts trust" hypothesis does not survive capture-size and set-overlap control on the golden set (Section 5), where every candidate fails capture-level permutation tests after multiple-comparison correction—reported prominently because the opposite in-sample reading is easy to reach and easy to publish.

8.1 Limitations

The golden set is modest ($n \approx 103$ cases, 87 captures) and single-domain, which bounds the negative result of Section 5 both ways: the size-controlled signals add nothing detectable, but the study is underpowered for a small residual effect (detecting the observed residual $\rho \approx 0.20$ would need $\sim$650 captures, $\approx$10$\times$ this corpus), so we claim only that structure provides *no demonstrable* reliability signal beyond capture size here. Minor caveats—temporal/conversation-only PCAP edges, a no-reconcile sibling proxy for pre-reconciliation graphs, templated disease signatures, and imperfect external benchmarks—bound the remaining claims; none of the cross-domain results is state-of-the-art, but together they evidence *portability*: a single untuned structural prior is broadly effective.

Conclusion. Diagnosis under partial observation is percolation on a partially observed evidence graph: relevance-weighted connectivity is a portable signal for *recoverability* and *localization*—but not, on our golden set, for *trust*. Across microservices, power grids, and real disease diagnosis the same prior beats the closed-world noisy-OR posterior under under-observation (Proposition 1), where treating unobserved evidence as negative evidence is what fails.

References

Rui Abreu, Peter Zoeteweij, and Arjan J C van Gemund. On the accuracy of spectrum-based fault localization. In *Testing: Academic and Industrial Conference Practice and Research Techniques (TAICPART-MUTATION)*, pp. 89–98, 2007.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2): 163–177, 2001.

Duncan S Callaway, Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471, 2000.

Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Resilience of the Internet to random breakdowns. *Physical Review Letters*, 85(21):4626–4628, 2000.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. volume 11, pp. 1605–1641, 2010.

Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5(1): 17–60, 1960.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Michaela Hardt, William Roy Orchard, Patrick Blöbaum, Elke Kirschbaum, and Shiva Prasad Kasiviswanathan. The PetShop dataset—finding causes of performance issues across microservices. In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pp. 957–978, 2024.

Taher H Haveliwala. Topic-sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web (WWW)*, pp. 517–526, 2002.

Jerome Henry, Swadhin Pradhan, and Miroslav Popovic. Evidence-grounded ensemble diagnosis of 802.11 packet captures: A multi-stage pipeline with deterministic reliability scoring. `https://arxiv.org/abs/2606.06871`, 2026. arXiv:2606.06871 (PROBE: Protocol Reasoning Over evidence-Based Ensembles).

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 6609–6625, 2020.

James A Jones and Mary Jean Harrold. Empirical evaluation of the Tarantula automatic fault-localization technique. In *Proceedings of the 20th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 273–282, 2005.

René Just, Darioush Jalali, and Michael D Ernst. Defects4J: A database of existing faults to enable controlled testing studies for Java programs. In *Proceedings of the International Symposium on Software Testing and Analysis (ISSTA)*, pp. 437–440, 2014.

Patel Kaushil et al. Disease symptom prediction dataset. `https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset`, 2019. Kaggle, derived from the Columbia Disease–Symptom Knowledge Database.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

Mark EJ Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Luan Pham, Huong Ha, and Hongyu Zhang. BARO: Robust root cause analysis for microservices via multivariate bayesian online change point detection. *Proceedings of the ACM on Software Engineering*, 1 (FSE):2214–2237, 2024.

Luan Pham, Hongyu Zhang, Huong Ha, Flora Salim, and Xiuzhen Zhang. RCAEval: A benchmark for root cause analysis of microservice systems with telemetry data. In *Companion Proceedings of the ACM on Web Conference (WWW)*, pp. 777–780, 2025.

Swadhin Pradhan, Shazal Irshad, and Jerome Henry. Plume: Building a network-native foundation model for wireless traces via protocol-aware tokenization. `https://arxiv.org/abs/2603.13647`, 2026. arXiv:2603.13647 (Plume: Protocol Language Understanding Model for Exchanges).

Oleg Sheyner, Joshua Haines, Somesh Jha, Richard Lippmann, and Jeannette M Wing. Automated generation and analysis of attack graphs. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, pp. 273–284, 2002.

Michael A Shwe, Blackford Middleton, David E Heckerman, Max Henrion, Eric J Horvitz, Harold P Lehmann, and Gregory F Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods of Information in Medicine*, 30(4):241–255, 1991.

Anna Varbella, Kenza Amara, Blazhe Gjorgiev, and Giovanni Sansavini. PowerGraph: A power grid benchmark dataset for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Xiaoyan Wang, Amy Chused, Noémie Elhadad, Carol Friedman, and Marianthi Markatou. Automated knowledge acquisition from clinical narrative reports. In *AMIA Annual Symposium Proceedings*, volume 2008, pp. 783–787, 2008.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

W Eric Wong, Vidroha Debroy, Ruizhi Gao, and Yihao Li. The DStar method for effective software fault localization. *IEEE Transactions on Reliability*, 63(1):290–308, 2014.

Junjielong Xu, Qinan Zhang, Zhiqing Zhong, Shilin He, Chaoyun Zhang, Qingwei Lin, Dan Pei, Pinjia He, Dongmei Zhang, and Qi Zhang. OpenRCA: Can large language models locate the root cause of software failures? In *International Conference on Learning Representations (ICLR)*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2369–2380, 2018.

Xiang Zhou, Xin Peng, Tao Xie, Jun Sun, Chao Ji, Wenhai Li, and Dan Ding. Fault analysis and debugging of microservice systems: Industrial survey, benchmark system, and empirical study. *IEEE Transactions on Software Engineering*, 47(2):243–260, 2021. doi: 10.1109/TSE.2018.2887384.

A Deferred Derivations and Methods

This appendix collects the standard derivations, the formal definitions, and the evaluation protocol underlying the main text; each is referenced where it is used.

A.1 Generating-function derivation of the percolation threshold

The modern, degree-distribution-agnostic derivation of Equation (1) uses probability generating functions (Newman et al., 2001; Newman, 2003). Let p_k be the probability that a uniformly chosen vertex has degree k, and define the degree and excess-degree generating functions

$$G_0(x) = \sum_{k\geq 0} p_k x^k, \qquad G_1(x) = \frac{G_0'(x)}{G_0'(1)} = \frac{1}{\langle k\rangle}\sum_{k\geq 1} k\, p_k\, x^{k-1}, \tag{9}$$

where $\langle k\rangle = G_0'(1)$. Following an edge to a vertex and counting its *other* edges yields the branching process governed by G_1; a giant component exists iff its mean offspring exceeds one, $G_1'(1) > 1$, the Molloy–Reed criterion $\langle k^2\rangle - 2\langle k\rangle > 0$ (Molloy & Reed, 1995). Independently retaining each edge with probability ϕ (bond percolation) rescales the branching ratio to $\phi\, G_1'(1)$, giving the threshold $\phi_c = 1/G_1'(1)$ of Equation (1) (Callaway et al., 2000; Cohen et al., 2000). For a Poisson degree distribution of mean c, $\langle k^2\rangle - \langle k\rangle = c^2$, hence $\phi_c = 1/c$.

A.2 Proof of Proposition 1

Restricted to $F^\star$, full observation contributes $\sum_{f\in F^\star} \log p_f$; under partial observation an observed finding still contributes $\log p_f$ while an unobserved one, scored absent, contributes $\log(1-p_f)$. The per-finding difference is $\log p_f - \log(1-p_f) = \mathrm{logit}(p_f)$, incurred with probability $1-\phi$; truly-absent findings ($f \notin F^\star$) are scored absent in both cases and cancel. Taking expectations over the independent observation indicators gives Equation (8). Overlap and PPR are functions of F_ϕ alone, so no absent-as-missing term appears. □

A.3 The noisy-OR likelihood

In bipartite cause→finding settings, let $q_{r,f} = \mathbb{P}[f \mid r]$ be the per-link activation reliability and ℓ_f the per-finding leak (background) probability. Under the noisy-OR assumption of independent causal activation, the single-cause log-likelihood of a present/absent finding partition is

$$\log \mathbb{P}[F \mid r] = \underbrace{\sum_{f\in F} \log\big(1-(1-\ell_f)(1-q_{r,f})\big)}_{\text{present-finding support}} + \underbrace{\sum_{f\notin F} \log\big((1-\ell_f)(1-q_{r,f})\big)}_{\text{absent-finding penalty}}, \tag{10}$$

with posterior $\mathbb{P}[r \mid F] \propto \mathbb{P}[r]\,\mathbb{P}[F \mid r]$ and $p_f := \mathbb{P}[f{=}1 \mid r^\star] = 1-(1-\ell_f)(1-q_{r^\star,f})$ in Proposition 1. A weakly activated finding ($p_f < \frac{1}{2}$) scored as absent actually *raises* the likelihood, so the bias of Equation (8) is not unconditionally signed; for the individually-probable findings typical of QMR it is a downward loss of order $\Theta((1-\phi)|F^\star|)$.

Remark 1 (Misspecification vs. a failure of Bayes). Equation (8) is a property of the *closed-world* reading, not of Bayesian inference. A correctly specified posterior that marginalizes unobserved findings scores only observed-present evidence and never charges $r^\star$ for an expected-but-unseen finding (it omits the $\log(1-p_f)$ terms). Section 7.2 confirms this: under heavy under-observation the closed-world posterior collapses, whereas the marginalized and observation-aware posteriors degrade gracefully, tracking the positive-evidence priors (Figure 8).

A.4 Configuration-model nulls and motif surprise

To ask whether an observed connectivity value exceeds its degree sequence, we compare against a *configuration-model* null that fixes the degree sequence but rewires edges uniformly at random (Molloy & Reed, 1995; Newman, 2003). For a statistic x with null mean μ_0 and standard deviation σ_0 (estimated over rewired replicas), we report the standardized surprise $z = (x - \mu_0)/\sigma_0$ and its empirical percentile. *Motif surprise* applies this to ordered subtype bigrams $g = (a \to b)$ over the observation timeline, scoring the configuration-null z-score of each bigram count; high z flags anomalous repetition (e.g. runs of `qos_data`). Both are evaluated as candidate reliability signals in Section 5.

A.5 Evaluation protocol and metrics

For a localizer producing a ranking of candidate causes, *top-k accuracy* is the fraction of cases whose true cause appears in the top k; *mean reciprocal rank* (MRR) is $\frac{1}{N}\sum_i 1/\mathrm{rank}_i$; and *mean percentile* is the average of $1-(\mathrm{rank}_i-1)/(m_i-1)$ over cases with m_i candidates. Monotone association between a continuous signal and correctness is measured by the Spearman rank correlation ρ with its two-sided p-value; paired pre/post shifts are tested with the Wilcoxon signed-rank test. For reliability we adopt selective prediction (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017): cases are admitted in descending score order and the *risk–coverage* curve plots error rate against coverage $\tau \in (0, 1]$, summarized by the area under it $\mathrm{AURC} = \int_0^1 \mathrm{risk}(\tau)\, d\tau$ (lower is better) and by accuracy at fixed coverage. Triage of correct vs. incorrect cases is measured by the AUROC of the signal. All randomized procedures are seeded; correlations are computed after removing undefined (NaN) entries.

B Additional Cross-Domain Studies and Theory Extensions

This appendix reports in full the four cross-domain studies summarized in Table 6—power grids, spectrum-based software fault localization, multi-hop retrieval-augmented QA, and security attack-path attribution—together with three threshold extensions (the bipartite cause–finding threshold, biased non-MCAR observation, and the detectability phase boundary). They are placed here so that the main text can stay focused on the networking instance and its two closest analogues (microservice RCA and bipartite QMR diagnosis), without discarding the evidence that the percolation principle is not networking-specific.

B.1 A bipartite percolation threshold for cause–finding diagnosis

The QMR detectability sweep (Section 7.2) is governed by the same percolation theory, specialized to the bipartite (disease–finding) configuration model. Let the disease and finding excess-degree branching factors be $f_1'(1) = \langle j(j-1)\rangle/\langle j\rangle$ and $g_1'(1) = \langle k(k-1)\rangle/\langle k\rangle$. A giant component exists iff $f_1'(1)\, g_1'(1) > 1$ (Newman et al., 2001); if each *finding* is observed independently with probability ϕ (finding-side site percolation, exactly the partial-observation model of Section 7.2), the disease-to-disease branching ratio rescales to $\phi\, f_1'(1) g_1'(1)$, giving the bipartite threshold

$$\phi_c^{\mathrm{bip}} = \frac{1}{f_1'(1)\, g_1'(1)}. \tag{11}$$

Simulating finding-side occupation on the disease one-mode projection, the finite-size susceptibility peak converges to Equation (11) from above (Figure 7; tabulated in Table 11), the same data-collapse logic that locates $\phi_c = 1/c$ on the one-mode benchmark, now tying the QMR collapse to the theory.

B.2 Comparison to domain-native baselines (PetShop and RCAEval)

The Train Ticket corpus of Section 7.1 ships propagation graphs but no metric time series, so it cannot drive the metric-based causal RCA methods that constitute the state of the art. To compare our relevance prior against *domain-native* baselines on a level field, we use the public PetShop benchmark (Hardt et al., 2024): 68 injected performance issues on a 41-component microservice application, each with a service-call graph, normal/abnormal CloudWatch metrics, and a ground-truth root-cause *node*. We evaluate inside PetShop's own harness (node-level top-k recall), detecting anomalies with a robust per-metric z-score and ranking

candidates by a personalized-PageRank walk seeded on the anomalous nodes (Table 16); we also report a degree heuristic, an anomaly-only ranking that drops the graph, and the walk run in reverse.

The results line up with the rest of the paper. The relevance prior transfers and stays orientation-sensitive: the forward walk, mass flowing along call edges toward dependencies, gives the best top-1 of the structure-only solvers (0.31) and far exceeds the reverse walk (0.06), because PetShop's root cause is a *downstream* dependency rather than the upstream source of the chaos corpus. But structure does not beat domain-native causal RCA: the published graph-traversal and CIRCA baselines exceed 0.8 top-3 recall on most scenarios (Hardt et al., 2024), well above our walk's 0.44. The bottleneck is diagnostic, not topological: an anomaly-only ranking caps at 0.41 top-3, so the true root cause is often *not* locally anomalous (it manifests downstream), and a positive-evidence walk cannot implicate a quiet-but-causal node the way a counterfactual or interventional model can. This is the microservice form of the broader thesis: structure says *whether* and roughly *where* a fault is recoverable, but on metric-rich data a causal model says *which* node more accurately (cf. the SBFL result of Appendix B.4). We make no state-of-the-art claim; the value is a faithful, same-harness comparison that bounds what structure alone buys—and since our front-end is a plain z-score rather than PetShop's SCM-based scorers, it isolates the *ranking* stage, not the anomaly detector.

Same-system check (RCAEval RE2 Train Ticket). PetShop is a different application; for a same-system comparison we use the public RCAEval RE2 Train Ticket benchmark (Pham et al., 2025) (90 cases, six fault types), which is metric-rich and so supports the causal/statistical baselines that define the state of the art. Building the call graph from trace spans (parent-span $\rightarrow$ child-span service), seeding the walk on z-score-flagged services, and reporting service-level AC@k (RCAEval's own metric, Table 10), our training-free prior attains AC@1 = 0.29: it *matches* CIRCA (0.32) and far exceeds MicroCause (0.10) and RCD (0.09), though it trails the strongest statistical baseline, the BARO change-point detector (0.67) (Pham et al., 2024). Two details sharpen the PetShop reading. Degree alone collapses to AC@1 = 0.04. Without evidence, topology is nearly useless here, exactly as the null-model analysis of Section 6 predicts. And unlike PetShop, the graph adds little *over* a plain anomaly-only ranking (0.29 vs. 0.30), because RE2's resource faults make the injected service its own strongest anomaly, leaving the walk little to recover. The thread across both benchmarks is the same: an evidence-seeded structural prior is a competitive, zero-training localizer on par with mid-tier causal baselines, but it neither replaces nor beats a well-tuned causal detector on metric-rich data.

B.3 Power-grid cascading failures (PowerGraph)

We analyze four grids (IEEE-39, IEEE-118, IEEE-24, and the UK network) from PowerGraph (Varbella et al., 2024), treating each as an undirected bus–branch graph annotated with cascading-edge labels and demand-not-served (DNS) severity. Here the structural primitive is edge betweenness (Section 2) rather than a relevance walk, and we test three predictions that the percolation viewpoint makes about infrastructure (Table 1, Figure 1):

B1 Central lines participate in cascades. Edge betweenness ranks the lines that appear in cascades with AUC 0.713 (IEEE-118), 0.699 (IEEE-39), and 0.622 (IEEE-24); only the dense, mesh-like UK grid is near chance (0.586), consistent with betweenness losing discriminative power when many redundant paths exist.

B2 Severity is only weakly structural. The betweenness of the *initially* tripped line predicts DNS severity only weakly (Spearman -0.098 to $+0.293$): the centrality of the trigger does not determine the magnitude of the eventual cascade, which depends on load and protection dynamics beyond pure topology.

B3 Targeted removal dominates random removal. Deleting lines in descending betweenness order fragments the giant component far faster than random deletion—on IEEE-118 the giant-component fraction falls to 0.356 at 35% targeted removal versus 0.693 under random removal—reproducing the robustness asymmetry of Equation (1) (Albert et al., 2000; Callaway et al., 2000; Cohen et al., 2000) on real infrastructure.

Table 1: Power-grid cascades (PowerGraph). B1: AUC of edge betweenness ranking cascading lines. B2: Spearman of tripped-line betweenness vs. DNS severity.

Grid	nodes	edges	cascade events	B1 AUC	B2 ρ
IEEE-118	118	186	1167	**0.713**	−0.098
IEEE-39	39	46	281	0.699	+0.139
IEEE-24	24	38	3444	0.622	+0.205
UK	29	99	4531	0.586	+0.293

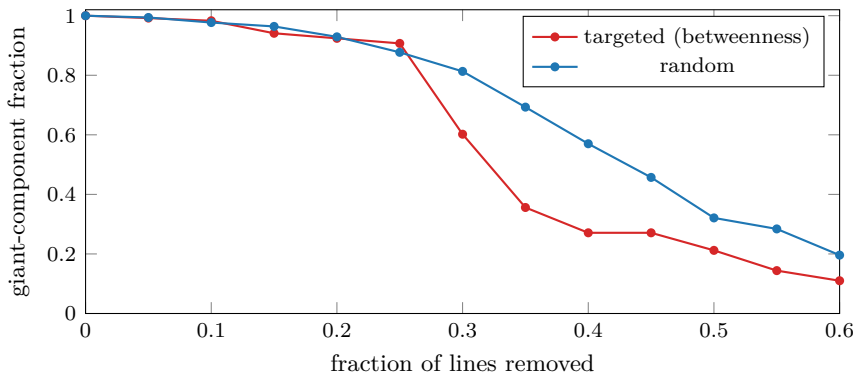


Figure 1: IEEE-118 robustness: targeted removal of high-betweenness lines fragments the grid far faster than random removal.

B.4 Software fault localization (spectrum-based)

Our fourth domain is spectrum-based fault localization (SBFL), where the "evidence graph" is a program-dependence graph and the evidence is the set of statements exercised by failing tests. We run a controlled simulation (in the spirit of the synthetic QMR study, not the real Defects4J (Just et al., 2014) JVM harness): 2880 programs over 72 configurations spanning program size, suite size, average dependence degree, fault-trigger probability, and a flaky-failure rate that scatters the failing evidence. Each program has one planted fault; localizers rank statements and we report the standard SBFL metrics (top-N, MRR, and the EXAM wasted-effort fraction). We compare the classic alarm-blind spectrum formulas—Tarantula (Jones & Harrold, 2005), Ochiai (Abreu et al., 2007), and D* (Wong et al., 2014)—against the same relevance walk used throughout (PPR over the dependence graph, seeded by the failing spectrum) and against Ochiai with a structural tie-break.

This domain reproduces *both* halves of our thesis. First, the positive half: the percolation order parameter of the *suspicious set* (the top-q statements by Ochiai, induced on the dependence graph) predicts whether the fault is localizable. When the fault sits in the giant component of the suspicious set, Ochiai top-1 accuracy is 0.49 (n=1720); when the suspicious set fragments and the fault falls into a smaller piece, accuracy collapses to 0.03 (n=164). The fault's component fraction correlates with localizability at Spearman ρ=+0.28 (95% CI $[0.24, 0.32]$) against top-1 and +0.45 $[0.41, 0.49]$ against $1 - \text{EXAM}$—the same "evidence percolates $\Rightarrow$ recoverable" diagnostic the paper derives for diagnosis, now in a fourth, unrelated domain.

Second, the negative half, mirroring Section 5: structure does *not* beat a simple separability baseline. A graph-only relevance walk is clearly *worse* than the native spectrum formulas (Table 2), and adding a structural tie-break to Ochiai buys only +0.01 top-1 and nothing on EXAM. The single strongest predictor of localizability is not topological at all: it is the size of the fault's spectrum tie-group—a separability/size quantity, ρ=+0.56 $[0.53, 0.60]$—just as capture size dominated every structural trust signal on the golden set. Connectivity again says *whether* a fault can be found, but a domain-native score, not graph structure, says *which statement it is*.

Table 2: Spectrum-based fault localization (2880 simulated programs). The graph relevance walk (PPR) underperforms the domain's native spectrum formulas; a structural tie-break on Ochiai adds at most 0.01 top-1. Structure predicts recoverability (text) but does not beat the separability-driven formulas at pinpointing the fault. EXAM is wasted effort (lower is better).

Solver	top-1	top-3	top-5	MRR	EXAM
random	0.014	0.042	0.076	0.069	0.497
coverage (e_f)	0.092	0.331	0.436	0.248	0.265
PPR (graph, evidence-seed)	0.100	0.294	0.406	0.237	0.298
PPR (graph, Ochiai-seed)	0.202	0.393	0.465	0.327	0.274
Tarantula	0.184	0.402	0.501	0.328	0.241
$D^\star$ ($\star$=2)	0.271	0.441	0.490	0.381	0.251
Ochiai	0.292	0.480	0.530	0.409	0.238
Ochiai + struct. tie-break	**0.302**	0.475	0.526	**0.411**	0.255

Table 3: Multi-hop RAG (3240 simulated questions). Structure-aware selection (degree and the relevance walk) recovers the gold supporting facts well above the lexical baseline used by production retrievers. recall@L = fraction of the L gold facts ranked in the top L; mean-rank-frac is lower-is-better.

Solver	recall@L	MRR	mean-rank-frac
random	0.213	0.417	0.560
lexical overlap	0.347	0.634	0.471
degree	0.594	0.853	0.292
PPR (question-seeded)	**0.516**	**0.818**	**0.304**
PPR (lexical-seeded)	0.518	0.819	0.304

B.5 Multi-hop retrieval-augmented QA (evidence graphs)

Our fifth domain maps the thesis onto retrieval-augmented generation, where multi-hop questions (in the style of HotpotQA (Yang et al., 2018) / 2WikiMultiHop (Ho et al., 2020)) are answerable only if the retrieved evidence *connects* the question entity to the answer through a chain of supporting facts. We simulate this directly (a knowledge graph, a planted reasoning chain, a noisy retriever with imperfect recall and hard entity-sharing distractors, and a probabilistic reader whose accuracy depends on chain completeness and distractor load—deliberately *not* on graph connectivity, so percolation is a genuine predictor rather than a definition): 3240 questions over 108 configurations, reader base accuracy 0.60.

This is the domain where structure is *most* valuable, on both claims. **(B)** The percolation order parameter of the retrieved-evidence graph sharply predicts correctness: when the question and answer entities fall in the same retrieved component, reader accuracy is 0.88 (n=1992); when the evidence fragments, 0.14 (n=1248). Question–answer connectivity correlates with correctness at Spearman ρ=+0.73 (95% CI $[0.70, 0.76]$) and Φ_{gcc} at +0.49 $[0.44, 0.53]$, whereas the size baseline (inverse number of retrieved facts) is *negatively* associated (-0.16). Unlike the golden set and SBFL, here a percolation signal beats a size baseline outright. **(A)** A relevance walk seeded at the question entities recovers the gold supporting facts far better than the lexical-overlap baseline that production RAG uses (Table 3): recall@L rises from 0.35 (lexical) to 0.52 (PPR), and MRR from 0.63 to 0.82.

B.6 Security alert correlation and attack-path attribution

Our sixth domain is intrusion response: an attacker moves laterally along an attack graph (Sheyner et al., 2002) from an initial compromise ("patient zero") to a target; sensors raise alerts on some compromised nodes (imperfect detection) plus false positives (noise), and the defender must name the root compromise. We simulate 4860 incidents over 162 configurations. The localizer is the same relevance walk, run *against* the attack edges (toward sources), mirroring the reverse-PageRank result on microservice RCA (Section 7.1).

Table 4: Security attack-path attribution (4860 simulated incidents). The reverse relevance walk (against attack edges) dominates the forward walk and attains the best mean percentile, but does not beat a plain alert-degree heuristic at top-1. Naming patient-zero exactly is hard; percentile localization is strong.

Solver	top-1	top-3	MRR	mean pctile
random	0.009	0.028	0.049	0.498
alert-degree	**0.195**	0.461	0.365	0.834
PPR (forward)	0.039	0.294	0.223	0.817
PPR (reverse)	0.185	0.423	0.359	**0.922**

Table 5: Threshold under degree-biased observation ($\langle k \rangle$=3, MCAR ϕ_c=0.335). α>0 over-samples hubs (realistic); the re-derived $\bar{\phi}_c$ matches the simulated peak and falls well below MCAR.

α	$\bar{\phi}_c$ (theory)	$\bar{\phi}^*$ (sim)	vs MCAR
-1	0.581	0.612	+73% (harder)
0	0.335	0.372	MCAR
$+1$	0.204	0.223	-39% (easier)
$+2$	0.145	0.170	-57% (easier)

This domain reproduces the dual pattern most cleanly. **Orientation is decisive:** the reverse walk attains a 0.92 mean percentile and 0.185 top-1, versus 0.039 top-1 for the forward walk (Table 4)—a 4.7× gap, exactly as on the Train Ticket corpus. **Percolation predicts confident attribution:** Φ_{gcc} of the alert-evidence graph correlates with reverse-walk success at ρ=+0.42 $[0.39, 0.45]$. **But structure does not beat a separability baseline:** the strongest single predictor is again a size/separability quantity—fewer correlated alerts, ρ=+0.48 $[0.45, 0.51]$—and a plain alert-degree heuristic matches the reverse walk on top-1 (0.195 vs 0.185). As on the golden set and in SBFL, connectivity says *whether* attribution is recoverable, but a domain-native separability score is as good at pinpointing the root.

B.7 Threshold under biased (non-MCAR) observation

Assumption 1 (MCAR) is wrong in the motivating domain: a buffer captures *around* an anomaly, over-representing evidence near the relevant subgraph and near hub vertices. We re-derive the threshold for degree-correlated site occupation. Occupying a degree-k vertex with probability ϕ_k, the occupied cluster branches supercritically iff

$$R = \sum_k \frac{k\, p_k}{\langle k \rangle}\, \phi_k\, (k-1) > 1, \tag{12}$$

which reduces to $\phi\, G_1'(1)$ when $\phi_k \equiv \phi$. Parameterizing a bias $\phi_k \propto (k/\langle k \rangle)^\alpha$ and solving Equation (12) for the critical *mean* observed fraction $\bar{\phi}_c$, hub-biased observation ($\alpha > 0$, the realistic buffer case) percolates—becomes diagnosable—at a substantially *lower* mean observation than MCAR predicts, while anti-correlated observation is harder (Table 5, Figure 2). The prediction matches the simulated susceptibility peak in every case. This is both an honest limitation (MCAR is conservative) and a usable correction: when the observation process favors the anomaly neighborhood, the effective detectability threshold drops by 40–60%.

B.8 The detectability gap as a phase boundary

Section 6 reported, at one operating point, that a relevance walk localizes the true cause where a degree heuristic is fooled by an incidental decoy hub. We map this as a two-dimensional boundary over the decoy's excess degree (structural confusability) and the observation density ϕ. Across the entire plane the evidence-blind degree heuristic stays in the undetectable phase (top-1 $\equiv$ 0: the decoy hub always out-ranks the true cause), while the relevance walk has a localizable phase whose high-confidence boundary (top-1 $\geq$ 0.8) *rises* with decoy excess: $\phi^* \approx 0.4$ at zero excess, climbing to $\approx$ 0.5–0.6 as the decoy grows (Figure 3). This is the "two thresholds" picture made quantitative—an information/recoverability threshold in ϕ, and an

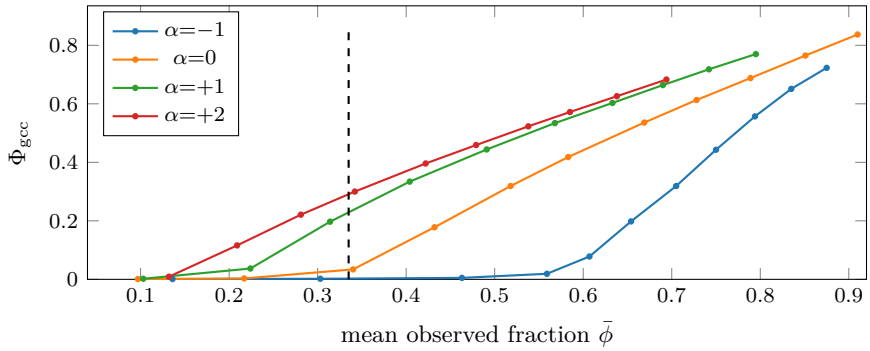


Figure 2: Biased observation shifts the giant-component onset: hub-biased sampling (α>0) percolates left of the MCAR threshold (dashed), anti-correlated sampling (α=−1) to its right.

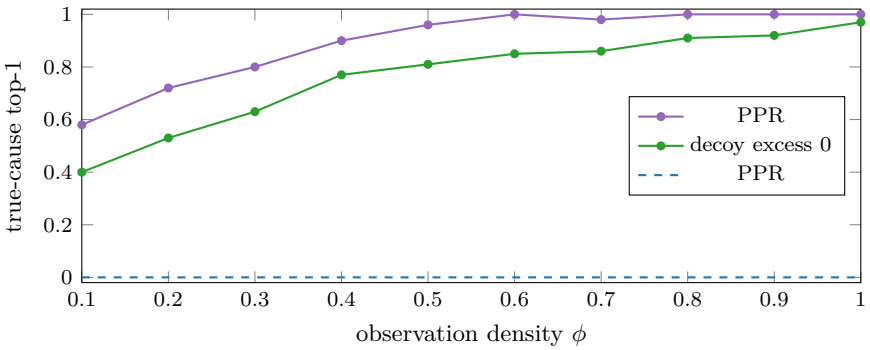


Figure 3: Detectability gap: the relevance walk localizes the true cause above an observation threshold that rises with decoy confusability, whereas the evidence-blind degree heuristic is fooled by the decoy hub everywhere (top-1 $\equiv$ 0).

algorithmic threshold that the relevance walk meets but a centrality heuristic never does in the presence of a decoy hub.

C Supplementary Figures

These figures support the main-text results where they are discussed and are collected here to keep the body compact; each is referenced at the point it is used.

D Supplementary Tables

These tables report the full numerical detail behind the figures and prose of the main text; each is referenced where it is discussed. They are collected here to keep the body focused on the headline results.

E Algorithms, Parameters, and Datasets

This appendix records the implementation details needed to reproduce every reported number. All randomized procedures are seeded; the disease dataset is obtained from its public source.

Evidence-graph construction (PCAP). Given a diagnosis citing frame set C, two frames $i, j \in C$ are joined iff $|i - j| \leq 1$ in capture order (a unit-width temporal bridge) or they belong to the same endpoint conversation. The four order parameters of Section 3.2 are computed on the induced subgraph $G[C]$ by union–find for components and direct edge counting for density.

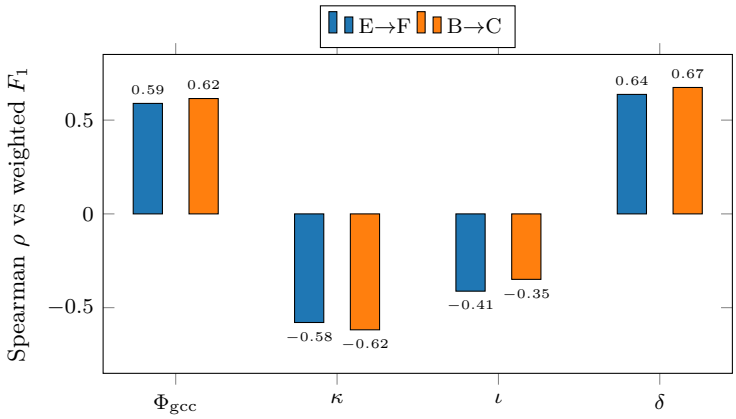
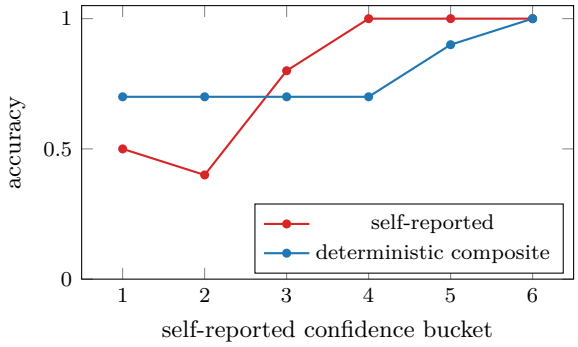
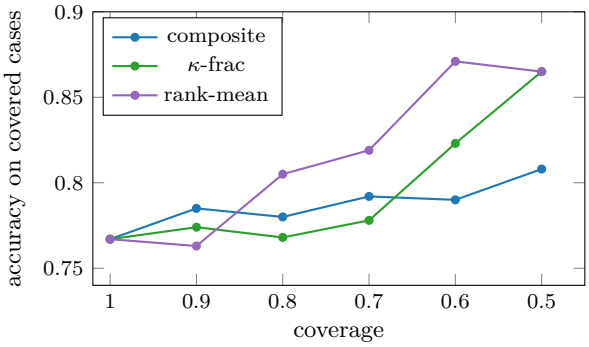


Figure 4: Post-reconciliation connectivity predicts correctness (Section 4). Span density δ and giant-component fraction Φ_{gcc} correlate positively with weighted F_1; fragmentation κ correlates negatively. All $|\rho| \geq 0.35$ are significant at $p < 10^{-3}$.



(a) Calibration: self-reported confidence is non-monotone and non-discriminative (AUROC 0.49), while the deterministic composite is monotone and orders better (AUROC 0.60). Note the composite is a better *ranker* than *probability*: it is systematically under-confident (ECE 0.47 vs 0.19), so it should drive selective prediction, not be read as a calibrated probability.

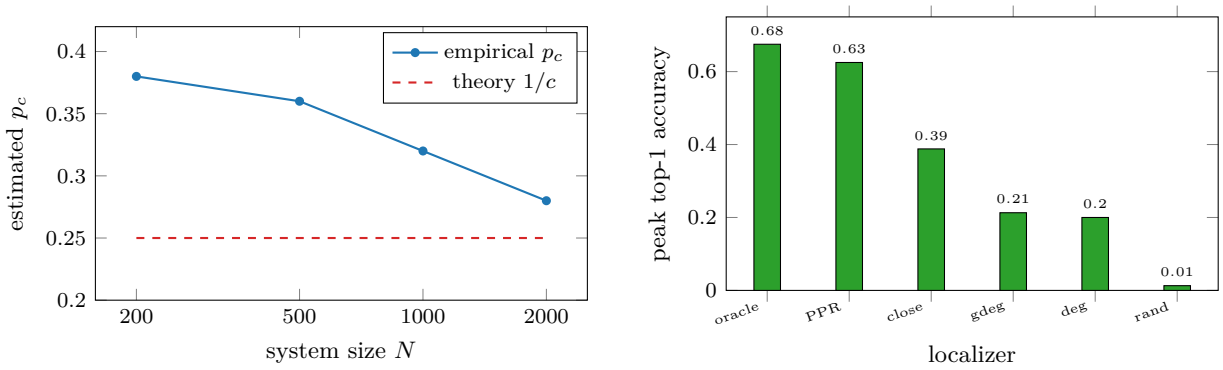


(b) Risk–coverage: the structural abstention gate beats the pipeline's confidence score at low coverage, and the gain *replicates out of sample* (median held-out AURC 0.13 vs 0.16; beats the composite in 97% of capture splits). Table 14 shows this reflects capture size/separability, not evidence-graph topology.

Figure 5: Reliability and selective prediction on the golden set (Section 5). The calibration motivation (left) stands; the selective-prediction gain (right) is real and replicates out-of-sample, but reflects capture size rather than evidence-graph topology (Table 14).

Personalized PageRank. We solve Equation (6) with damping $d = 0.85$ by power iteration to tolerance 10^{-9} (12 iterations for the large sparse QMR graphs, up to 100 elsewhere). The restart distribution is the seed set (the observed evidence) for the evidence-seeded walks; the *reverse-PageRank prior* reported as the headline on the Train Ticket corpus instead uses a *uniform* restart on the reverse-transposed adjacency, so it ranks upstream sources without reading the alarms (we report the alarm-seeded variant separately, Table 15). The sparse implementation propagates mass only over realized edges, matching Equation (7) truncated at the iteration count.

Noisy-OR oracle. We evaluate Equation (10) with link reliabilities clamped to $[0.02, 0.98]$ and per-finding leak $\ell_f = \max(\text{floor}, 0.3\,\hat{p}(f))$ estimated from data, where $\hat{p}(f)$ is the empirical marginal of finding f. The synthetic oracle uses the true generative parameters; the real-data oracle uses empirical $\hat{\mathbb{P}}[f \mid d]$. Posteriors are computed in log-space for numerical stability.



(a) Empirical critical density approaches $1/c = 0.25$.

(b) PPR tracks oracle; degree scheme fail under decoys.

Figure 6: Synthetic percolation benchmark (Section 6): the empirical critical density approaches $1/c$ (left), and a connectivity-aware localizer tracks the Bayesian oracle while degree heuristics collapse under decoy hubs (right).

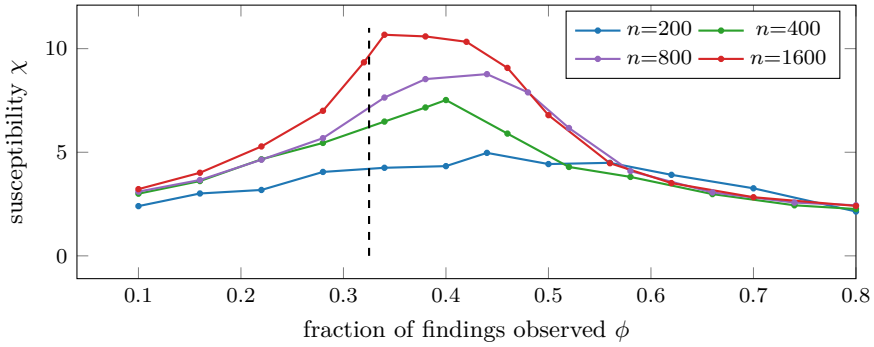


Figure 7: Bipartite percolation (Appendix B.1): the susceptibility peak sharpens and moves toward the predicted $\phi_c^{\text{bip}} = 0.325$ (dashed) as n grows.

Synthetic percolation benchmark. Background mean degree $c = 4$ (theory $\phi_c = 1/c = 0.25$), $L = 4$ causal paths per incident, decoy hubs carrying $+25$ excess degree clustered in a reserved region disjoint from the causal paths, and 40–80 independent trials per (ϕ, N) point. Recoverability is the existence of an observed path from R to A; the empirical $\hat{\phi}_c$ is the location of the susceptibility peak.

Cross-domain datasets. Train Ticket chaos-injection corpus (Zhou et al., 2021): 451 Chaos Mesh fault-injection cases; spans and containers are canonicalized to their parent service, yielding a mean of 5.1 candidate services per case. PowerGraph (Varbella et al., 2024): IEEE-39/118/24 and the UK grid, loaded from MATLAB v7.3 (HDF5) cascade archives; cascade labels record lines tripped *beyond* the initial outage. Disease–symptom data (Kaushil et al., 2019): 41 diseases, 131 symptoms, 4,920 patient records; edge weights are empirical $\mathbb{P}[\text{symptom} \mid \text{disease}]$.

Statistical procedures. Paired pre/post comparisons use the Wilcoxon signed-rank test; monotone association uses the Spearman rank correlation ρ after removing undefined entries; selective prediction uses the risk–coverage AURC and fixed-coverage accuracy; ranking and triage use the Mann–Whitney AUC / AUROC. All metrics are defined in Appendix A.5.

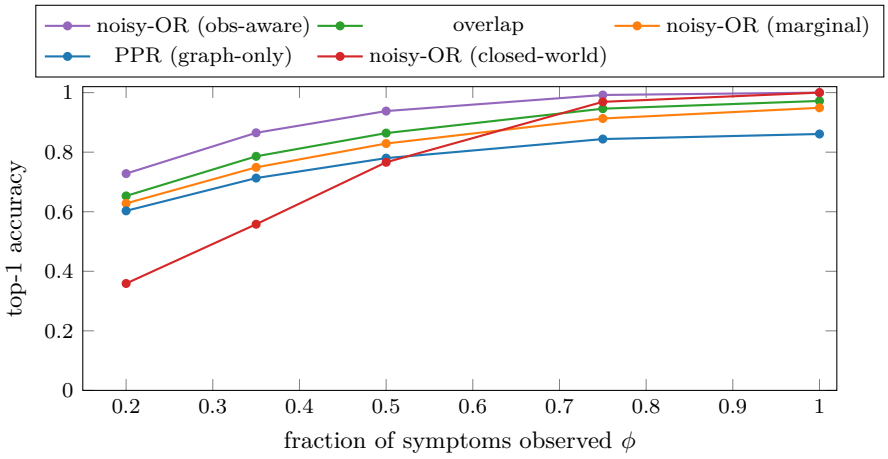


Figure 8: Real disease–symptom data under under-observation (Section 7.2). The *closed-world* noisy-OR (red) collapses as symptoms are dropped (the misspecification of Proposition 1) while the correctly specified *observation-aware* posterior (top) degrades most gracefully and the *marginal* posterior and the cheap positive-evidence priors (overlap, PPR) sit in between. The fragility is in the closed-world reading, not in Bayesian inference.

Table 6: Cross-domain summary (Section 7). The same evidence-graph model and relevance walk, applied with only a per-domain choice of edge orientation, localize the root cause well above chance in every domain; structure beats a size/separability baseline only where the evidence is genuinely fragmented (multi-hop RAG). The Train Ticket RCA and QMR studies are detailed in the main text (Sections 7.1 and 7.2); the baseline comparison and the remaining studies are reported in full in Appendix B.

Domain	Data	Headline localization	Beats size baseline?
Microservice RCA (Train Ticket)	real	reverse PPR, 0.89 top-1	orientation-driven
PetShop / RCAEval RE2	real	evidence PPR matches CIRCA	no (causal RCA wins)
QMR disease–symptom	real+sim	evidence PPR, 0.86 top-1	overlap $\approx$ PPR
Power grid (PowerGraph)	real	edge betweenness, AUC 0.71	targeted $\gg$ random
Software fault loc. (SBFL)	sim	predicts recoverability	no (spectrum wins)
Multi-hop RAG	sim	PPR recall 0.35$\rightarrow$0.52	**yes**
Security attack-path	sim	reverse PPR, 0.92 pctile	no (degree ties)

Table 7: Ablation of the evidence-grounded ensemble pipeline (from the PROBE (Henry et al., 2026); weighted evidence F_1 is the headline metric, Section 1.2). Configurations differ in the number of drafts and whether a reconciliation pass is applied. Reconciliation is the dominant lever.

Config	N	Key F_1	Reliab. F_1	**Wt** F_1	Perfect-key %	role
A	104	0.773	0.460	0.871	45.2	lean baseline
B	104	0.680	0.831	0.912	34.6	ensemble, no reconcile
C	100	0.815	0.851	**0.964**	51.0	ensemble + reconcile
D	104	0.444	0.692	0.842	21.2	ablation
E	104	0.461	0.692	0.845	20.2	ablation
F	100	0.776	0.846	0.957	50.0	ensemble + reconcile

Table 8: Detectability on the synthetic benchmark with decoy hubs ($N = 500$, $c = 4$, 80 trials). Peak top-1 accuracy and the observation density p_{50} at which accuracy first reaches 50% ("–" = never).

Localizer	peak acc	p_{50}
oracle (Bayesian)	**0.675**	0.482
PageRank (connectivity-aware)	0.625	0.493
closeness	0.388	–
global degree	0.213	–
degree (component)	0.200	–
random	0.013	–

Table 9: Bipartite QMR diagnosis (full observation, Section 7.2). A graph-only relevance prior (PPR) localizes the true cause far above chance and tracks the noisy-OR posteriors, on both a synthetic model and a real disease–symptom dataset. At full observation the closed-world and observation-aware posteriors coincide; the *marginal* posterior is slightly lower because it ignores true negatives. The closed-world/correctly-specified split matters only under partial observation (Figure 8). Real-data noisy-OR scores are in-sample upper bounds.

	Synthetic (300×1000)				Real (41×131, 4920 cases)			
Solver	top-1	top-3	MRR	pctile	top-1	top-3	MRR	pctile
random	0.002	0.005	0.018	0.505	0.026	0.078	0.108	0.502
degree	0.001	0.001	0.036	0.547	0.024	0.073	0.105	0.500
overlap	0.952	0.974	0.964	0.997	0.972	1.000	0.986	0.999
PPR (graph-only)	**0.705**	0.871	0.798	0.993	**0.861**	0.888	0.899	0.990
noisy-OR (closed-world)	0.986	0.993	0.990	1.000	1.000	1.000	1.000	1.000
noisy-OR (marginal)	0.973	0.988	0.982	0.999	0.949	1.000	0.973	0.999
noisy-OR (obs-aware)	0.986	0.993	0.990	1.000	1.000	1.000	1.000	1.000

Table 10: Public RCAEval RE2 Train Ticket benchmark (Pham et al., 2025) (90 cases), coarse-grained service-level AC@k. Our training-free relevance prior matches CIRCA and beats MicroCause/RCD but trails the BARO change-point detector; degree alone (no evidence) collapses. Baseline AC@1 values are as reported by Pham et al. (2025).

Method	AC@1	AC@3
degree (structural only)	0.04	0.20
anomaly-only (no graph)	0.30	0.79
relevance walk, reverse	0.30	0.52
relevance walk (ours)	**0.29**	0.61
published baselines (Pham et al., 2025) (AC@1):		
BARO (Pham et al., 2024)	0.67	–
CIRCA	0.32	–
MicroCause	0.10	–
RCD	0.09	–

Table 11: Bipartite threshold $\phi_c^{\text{bip}} = 1/(f_1' g_1')$ vs the finite-size susceptibility-peak location (Appendix B.1). The empirical peak approaches the predicted $\phi_c^{\text{bip}} \approx 0.33$ from above as the system grows.

n_{dis}	ϕ_c^{bip} (theory)	ϕ^* (susceptibility peak)
200	0.364	0.524
400	0.296	0.395
800	0.337	0.434
1600	0.325	**0.343**

Table 12: Bridging analysis on 200 reconciliation pairs. Φ_{gcc} does not shift (Wilcoxon n.s.); the orphan fraction ι drops significantly in E→F. Post-reconciliation structure correlates strongly with weighted F_1 (last column).

Family	Param	mean pre	mean post	Wilcoxon p	Spearman $\rho(\Phi_{\text{post}}, F_1)$
E→F	Φ_{gcc}	0.700	0.717	0.67	+0.589
	κ	2.84	2.49	0.056	−0.579
	ι	0.149	**0.069**	**0.008**	−0.412
	δ	0.611	0.644	0.27	+0.637
B→C	Φ_{gcc}	0.694	0.681	0.71	+0.615
	κ	2.89	2.71	0.20	−0.618
	ι	0.058	0.055	0.58	−0.349
	δ	0.627	0.647	0.61	+0.674

Table 13: *In-sample (uncontrolled)* structural reliability signals on the 103-case golden set. Risk–coverage AURC is *lower-is-better*; AUROC is correct/incorrect triage. "–" marks quantities not of that type. These case-level associations do *not* survive capture-size control; see Table 14.

#	Signal	Spearman ρ (p)	AURC ↓	AUROC
–	hand-tuned composite (baseline)	–	0.159	0.599
1	capture fragmentation κ-frac	+0.250 (0.011)	0.132	–
1	rank-mean(composite, separation)	–	**0.129**	–
2	edge-Jaccard stability	+0.269 (0.019)	–	–
2	LCC-Jaccard stability	+0.315 (0.006)	–	–
2	composite + conn.-robustness	–	–	**0.637**
3	motif surprise (z, bigrams)	−0.194 (0.050)	–	0.358
4	community concentration	+0.054 (0.59)	–	–
5	mean cited PageRank centrality	+0.223 (0.024)	–	–

Table 14: *Size-controlled* re-analysis. Partial Spearman controls for capture size, citation count, and mean citation gap; p is a capture-level permutation test; q is Benjamini–Hochberg over all 14 candidate constructions (n=103 cases / 87 captures, except the agreement metrics which require a second opinion, n=76 / 63). *No signal survives multiple-comparison correction*; Φ_{gcc} excess-over-null is a degree-preserving negative control.

Signal	raw ρ	partial $\rho \mid$ size	cap-perm p	BH q
fragmentation κ-frac	+0.250	+0.138	0.125	0.29
LCC-Jaccard stability	+0.315	+0.275	0.026	0.18
edge-Jaccard stability	+0.269	+0.252	0.025	0.18
mean cited PageRank centrality	+0.223	+0.032	0.350	0.54
connectivity-robustness	+0.171	+0.070	0.545	0.64
motif surprise	−0.194	−0.048	0.451	0.57
Φ_{gcc} excess over null (ctrl.)	+0.041	−0.070	0.735	0.79

Table 15: Microservice RCA on the Train Ticket chaos-injection corpus (451 cases). The reverse-PageRank prior localizes the root-cause service best, but it is *evidence-blind* (uniform restart): the alarm-seeded walk (`PPR (from alarms)`) and the degree heuristic show the reverse prior's edge is edge *orientation*, not evidence-seeding. Forward PageRank is poor because it walks *away* from causes.

Localizer	top-1	top-3	MRR	mean pctile
random	0.350	0.780	0.582	0.590
degree	0.685	0.953	0.816	0.787
PageRank (forward)	0.126	0.661	0.421	0.301
PageRank (reverse)	**0.891**	**0.978**	**0.938**	**0.961**
closeness (from alarms)	0.013	0.592	0.341	0.173
PPR (from alarms)	0.530	0.945	0.726	0.637

Table 16: Public PetShop benchmark (Hardt et al., 2024) (68 issues), node-level top-k recall in PetShop's harness. Our relevance walk leads the structure-only solvers on top-1 and is orientation-sensitive, but published domain-native causal baselines are stronger on top-3—the honest limit of a structure-only prior on metric-rich data.

Method	top-1	top-3
degree (structural only)	0.28	0.54
anomaly-only (no graph)	0.15	0.41
relevance walk, reverse	0.06	0.27
relevance walk (ours)	**0.31**	0.44
published baselines (Hardt et al., 2024) (top-3 recall, range over scenarios):		
graph traversal	–	0.57–1.00
CIRCA (causal)	–	up to 1.00
counterfactual attribution	–	up to 0.86