Consensus–Bayesian Anomaly Detection in Agentic Access Graphs

Pratyush Uppuluri [1] **Shilpa Noushad** [1] **Sajan Kumar** [1] **Jayanth Poosarla** [1]

Abstract

Enterprise environments increasingly include both human users and autonomous agents accessing shared software, knowledge, and infrastructure assets such as directories, repositories, and internal services. A single compromised or manipulated agent, especially under rapidly growing agentic capabilities, can propagate abnormal access across structurally coupled resources faster than existing threat-detection systems can react, making graph-level behavioral monitoring increasingly important. We model each user or agent as a node in a heterogeneous belief network over a multi-level access graph and each directory as a topic, following the consensus framework of Ye et al. (Ye et al., 2021). Healthy access patterns correspond to opinion-dynamics regimes in which the theory guarantees per-SCC(Strongly Connected Component) consensus, whereas malicious modifications to an agent's logical dependencies break the structural-homogeneity assumption underlying those guarantees and manifest as transient rises in cross-agent opinion variance. We turn this observation into a sequential Bayesian detector, in which a binary anomaly state is updated online through an exponential likelihood on the variance jump, with both static-prior and online-prior variants. Experiments on a synthetic seven-agent/seven-directory testbed show that the detector responds to homogeneity-breaking perturbations rather than benign within-SCC(Strongly Connected Component) weight changes, that SCC-aware variance localizes the attacked block more clearly than global variance, and that online-prior recursion saturates faster than a static-prior baseline as perturbation strength grows. SCC-level decomposition yields direct structural explainability for the flagged anomaly, producing a verdict format well matched to machine-speed threat regimes and agentic security-operations workflows.

[1]Purdue University. Correspondence to: Pratyush Uppuluri <puppulur@purdue.edu>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

1. Introduction

Enterprise environments are increasingly shared by both human users and autonomous agents. With proliferation of agentic infrastructure swarmed by foundation-model assistants, retrieval copilots, and tool-using agents now read code repositories, query knowledge bases, edit documents, and call internal APIs alongside employees. In many settings, these agents accumulate access patterns much like human users, but at higher speed, with lower observability, and with a much broader attack surface, in novel ways, e.g. "Content Injection: Exploiting differences in how humans and AI parse raw web data" (Franklin et al., 2024). A single prompt-injection or tool-hijacking event can make an agent issue requests its principals never intended. Industry threat-intelligence reports increasingly frame this as a shift toward *machine-speed* threats, where AI-enabled adversaries can identify, chain, and weaponize weaknesses faster than security teams of human analysts teamed with machines and conventional defenses can respond (Google Cloud Mandiant Threat Intelligence, 2024).

This whole new paradigm changes how anomalies appear in practice. Agentic adversaries can chain several individually low-severity steps into a much higher-impact compromise, blurring the practical distinction between a major exploit and a sequence of seemingly benign local accesses (Google Cloud Mandiant Threat Intelligence, 2024). As a result, defenders are advised to baseline normal interaction patterns and detect deviations from that baseline rather than rely only on signatures of known attacks (Google Cloud Mandiant Threat Intelligence, 2024). At individual user-action policy enforcement is therefore necessary, but not sufficient. Structural anomalies often emerge across sequences of otherwise individually legitimate accesses and only become visible when viewed against the broader graph structure of the user, agent, and resource ecosystem (Vectra AI, 2023; Microsoft Security Team, 2024; Google Cloud Mandiant Threat Intelligence, 2024).

Traditional User-Entity Behavior Analytics (UEBA) systems address threat detection primarily for human users

through anomaly detection on access logs, often using static baselines, LSTM-based models, or graph clustering (Khaliq et al., 2020; Yue Song, 2023; Cui et al., 2022). While useful, these approaches typically suffer from three limitations. First, they often require frequent retraining as behavior drifts. Second, they tend to produce opaque anomaly scores rather than structural explanations. Third, they flatten the genuinely *multi-level* structure of enterprise access graphs, where groups of users co-access groups of related directories, and where these block-structured patterns are often exactly where anomalies become most informative.

In this paper, we propose a novel consensus-Bayesian framework grounded in *opinion dynamics over heterogeneous belief systems* (Ye et al., 2021). Each user, whether human or autonomous, is treated as a node in an influence network with row-stochastic interpersonal influence matrix W. Each directory is treated as a topic, and logical dependencies between topics are encoded through agent-specific logic matrices C_i. Under fixed and structurally homogeneous belief systems, opinions converge to per-SCC(Strongly Connected Component) consensus values whose existence and form are characterized by Theorems 2-5 of Ye et al. (Ye et al., 2021). In our setting, anomalous access events, such as a compromised agent suddenly influencing directories outside its prior structural community or a malicious structural modification to the dependency graph, correspond to changes in C_i that violate the structural-homogeneity assumption in (Ye et al., 2021) or heightened relative activity (`read`, `write`, `delete`, `access`, `login`, `join`, etc.). The same theory that guarantees consensus in the healthy regime therefore also tells us what kind of structural mismatch can destroy that behavior. In practice, the post-perturbation system may fail to converge within the same SCC or may converge to a substantially different steady state, and both cases manifest as a transient rise in cross-agent opinion variance over the affected topics.

We turn this observation into a anomaly detector by combining three components: (i) an SCC-decomposition pipeline that routes each topic block to the appropriate Ye et al. theorem and produces per-block convergence signatures, (ii) a scaled cross-agent variance signal as a one-dimensional summary of regime violation, and (iii) a sequential Bayesian update over a binary anomaly state, with an exponential likelihood on the variance jump. The resulting anomaly probability π_t accumulates evidence over time. Just as importantly, SCC-level decomposition tells us not only that something abnormal happened, but also which structural block lost consensus, giving insights into target component.

Contributions.

1. We provide a formal mapping from enterprise access graphs with mixed human and autonomous agents to the heterogeneous-belief-system model of (Ye et al., 2021), in which malicious access events correspond to structural-homogeneity violations.
2. We clearly state the theoretical scope of the method: we use (Ye et al., 2021) Theorems 2-5 to characterize the normal/healthy regime and the post-perturbation regime separately, and treat the transient on which the detector fires as an empirical effect informed by, but not directly proved by, the theory.
3. We develop a sequential Bayesian detector with an explicit generative view of the anomaly state and the variance-jump likelihood, with both static-prior and online-prior variants.
4. We validate the framework on a synthetic but theoretically grounded seven-agent / seven-directory testbed, comparing static and online Bayesian inference, raw-variance and Frobenius-norm baselines, and false-positive behavior under benign structural shifts.

2. Related Work

Opinion dynamics and heterogeneous belief systems. Classical opinion-dynamics models study consensus on a single topic (DeGroot, 1974; Friedkin & Johnsen, 1990; Hegselmann & Krause, 2002). More recent work extends this setting to multi-topic, logically coupled belief systems. In particular, Ye et al. (Ye et al., 2021) study heterogeneous belief systems in which agents interact over a shared influence graph while carrying agent-specific logic matrices, and derive convergence and disagreement results through SCC-based analysis of the topic graph. Our work is inspired by this line of theory in fact leverages their theoretical framework coupled with Bayesian principles in a different setting: we focus on dynamic access-induced dependency structure and use the resulting consensus/disagreement behavior as a signal for anomaly detection.

Graph-based anomaly detection. Graph anomaly detection includes spectral and density-based methods (Akoglu et al., 2015), deep auto-encoder models (Ding et al., 2019), and contrastive self-supervised approaches (Liu et al., 2021). These methods typically define anomaly as deviation from learned graph statistics or latent representations. In contrast, our framework uses the *transient disruption of consensus structure* as the primary signal and leverages SCC decomposition to provide block-level interpretability.

Bayesian and sequential change detection. Bayesian online change-point detection (BOCPD) (Adams & MacKay, 2007) and classical sequential methods such as CUSUM and Shiryaev-Roberts (Page, 1954; Shiryaev, 1963) address the problem of detecting regime shifts over time. Our Bayesian

update plays a similar evidence-accumulation role, but the underlying signal is different: rather than a generic univariate time series, it is driven by SCC-aware opinion-variance shifts induced by evolving access-graph dependencies.

UEBA, insider-threat detection, and agentic security. UEBA methods are widely used for insider-threat and behavioral-anomaly detection in enterprise systems, including approaches based on access statistics, recurrent models, and clustering (Microsoft Security Team, 2024; Khaliq et al., 2020; Cui et al., 2022). These methods are often centered on human-user behavior and usually do not provide structural interpretation. At the same time, recent work on agentic-system security has highlighted new threat surfaces created by LLM-driven and tool-using agents, including prompt injection, tool misuse, and delegated multi-step behavior (Greshake et al., 2023; Zhan et al., 2024). Our work sits at the intersection of these areas: it studies structurally interpretable anomaly detection in mixed human-agent enterprise access graphs. To our knowledge, consensus-based monitoring of such agentic access graphs has not been studied in this form.

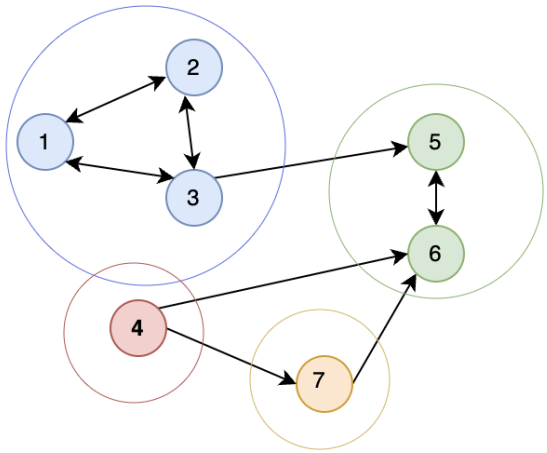

Figure 1. Directory/topic logical dependency graph induced by the agent-specific logic matrices. Nodes are directories/topics; arrows indicate logical dependencies between directory access propensities. Colored clusters mark strongly connected components; closed SCCs have no dependencies on outside topics, while open SCCs depend on at least one external topic.

3. Problem Formulation

3.1. Multi-level Agent-Directory Interaction Graph

Our enterprise model adopts a directory-conditioned, multi-level interpretation inspired by Ye et al. (Ye et al., 2021); the theorem-guided analysis is inherited from their framework, while the application-specific graph construction is adapted to enterprise access behavior.

Consider an enterprise with $n \geq 2$ directories $D_1, \ldots, D_n$ and $m \geq 1$ users (we use *user* hereafter for either humans or autonomous agents). Users interact with files through operations such as `read`, `write`, `delete`, `update`; these interactions induce two levels of structure, for scope of discussion we confine in this paper to these activities but can assume other forms such as logins etc.

Interpersonal influence graph $G[W]$. We define a row-stochastic interpersonal influence matrix $W \in \mathbb{R}^{m\times m}$ over the m users:

$$w_{ii} > 0, \qquad \sum_{j=1}^{m} w_{ij} = 1 \quad \forall i, \tag{1}$$

capturing how strongly each user's behavior is influenced by every other user.

Agent-specific directory-level logic graph $G[C_i]$. For each user u_i, the agent-specific logic matrix $C_i \in \mathbb{R}^{n\times n}$ encodes how directory D_p's expected access pattern depends on directory D_q's, as perceived by user u_i. Within each strongly connected component (SCC) of the directory-dependency graph for the nonnegative access-intensity construction, we use the row-normalized form:

$$\sum_{q=1}^{n} c_{pq,i} = 1, \quad c_{pq,i} \geq 0 \quad \forall p \in \mathrm{SCC}_j(i), \tag{2}$$

together with bidirectional symmetry inside an SCC, $c_{pq,i} = c_{qp,i}$ for $D_p, D_q \in \mathrm{SCC}_j$. From observed access intensities $A^i_{pq}(t)$ between directories D_p and D_q generated by user u_i we estimate

$$c_{pq,i}(t) = \frac{A^i_{pq}(t)}{\sum_{k\in\mathrm{SCC}_j(i)} A^i_{pk}(t)}. \tag{3}$$

SCC structure. Directories naturally cluster into interaction communities (e.g. a code-repository cluster shared by engineers, a payroll-data cluster shared by finance). We represent these as SCCs of $G[C_i]$, the per-user directory-dependency graph. Under Assumption 2 (Structural Homogeneity), the sparsity pattern of $G[C_i]$ is the same across users, so the SCC structure is well defined for the population. closed (no dependencies on topics outside the block) or open (depending on at least one external topic); the distinction will determine which Ye et al. theorem governs each block.

3.2. Opinion Dynamics on the Multi-Level Graph

Following (Ye et al., 2021), each user i maintains an opinion vector $x_i(t) \in \mathbb{R}^n$ over the n directories. We use the same

update law:

$$x_i(t{+}1) \;=\; \sum_{j=1}^{m} w_{ij}\, C_i\, x_j(t), \tag{4}$$

which reduces, for a single directory (topic) p with self-dependency $c_{pp,i}$, to

$$x_i^p(t{+}1) = c_{pp,i} \sum_{j=1}^{m} w_{ij} x_j^p(t) + \sum_{q \neq p} c_{pq,i}\, x_i^q(t). \tag{5}$$

Readers familiar with opinion dynamics may treat $x_i^p(t)$ as a scalar "access propensity" of user i for directory p at time t; the framework is agnostic to the concrete operationalization.

4. Theoretical Foundations and Their Scope

In this section, we make clear which parts of our detector are directly supported by Ye et al. (Ye et al., 2021) and where our empirical claims begin. We use the assumptions and theorem numbering from the cited preprint (arXiv:1812.05138); the published TAC version presents the same underlying results.

4.1. Assumptions Inherited from (Ye et al., 2021)

Assumption 1 (Logic-matrix structure)**.** *Each C_i has eigenvalue 1 semi-simple (or $|\lambda| < 1$); $c_{pp,i} > 0$ for all i, p; and $\sum_q |c_{pq,i}| = 1$ for all p, i.*

Assumption 2 (Structural homogeneity)**.** *For all users i, j, the $n \times n$ logic matrices C_i and C_j share the same zero/nonzero sparsity pattern over directories. Users may differ in weights and signs but not in which directory depends on which.*

Assumption 3 (Influence-graph topology)**.** *$W \in \mathbb{R}^{m\times m}$ (the user-user interpersonal influence matrix) is row-stochastic with $w_{ii} > 0$, and $G[W]$ is strongly connected.*

4.2. Convergence Theorems We Use

Theorem 1 (Global convergence; Theorem 2 of the preprint / Theorem 1 of the TAC paper)**.** *Under Assumptions 1–3, the process converges exponentially fast to some $y^* \in [-1, 1]^{nm}$ for any initial condition $y(0) \in [-1, 1]^{nm}$.*

Theorem 2 (Closed-SCC consensus; Theorem 3 of the preprint / Theorem 2 of the TAC paper)**.** *Under the conditions of Theorem 1, if each C_i is irreducible, then each topic k reaches a per-topic consensus $\lim_{t\to\infty} y^k(t) = \alpha_k \mathbf{1}_m$ for some $\alpha_k \in [-1, 1]$.*

Theorem 3 (Open-singleton consensus; Theorem 4 of the preprint / Theorem 3 of the TAC paper)**.** *For an open singleton SCC $J_j = \{p\}$ depending on external topics $\hat{J}_p$ that have already reached consensus values α_q, topic p reaches consensus $\alpha_p = \kappa$ if and only if there exists $\kappa \in [-1, 1]$ such that*

$$\kappa(I_m - \Gamma_{pp}) = \sum_{q\in\hat{J}_p} \alpha_q \Gamma_{pq}. \tag{6}$$

Equivalently, componentwise,

$$\kappa = \frac{\sum_{q\in\hat{J}_p} \alpha_q\, c_{pq,i}}{\sum_{q\in\hat{J}_p} |c_{pq,i}|} \quad \forall i. \tag{7}$$

Theorem 4 (Multi-topic open-SCC consensus; Theorem 5 of the preprint / Theorem 4 of the TAC paper)**.** *For a multi-topic open SCC J_j (a multi-directory SCC of $G[C_i]$) with external dependencies $\tilde{J}_j$ already at consensus, every directory (topic) $k \in J_j$ reaches a consensus $\alpha_k = \phi_k$ if and only if there exists a vector ϕ with entries in $[-1, 1]$ satisfying the joint agent-uniform condition of Equation (16) in the preprint (equivalently, Equation (15) in the TAC paper).*

We additionally use Corollary 2 of (Ye et al., 2021), which provides sufficient conditions for (6) to admit a solution and, in particular, shows that a single competing logical interdependency can be enough to break consensus.

Underlying opinion-dynamics update. Following Ye et al. (Ye et al., 2021), the core multidimensional update is

$$x_i(t+1) = \sum_{j=1}^{m} w_{ij}\, C_i\, x_j(t), \tag{8}$$

where $x_i(t) \in \mathbb{R}^n$ is the opinion vector of user i over the n directories, $W = [w_{ij}] \in \mathbb{R}^{m\times m}$ is the interpersonal influence matrix among the m users, and $C_i \in \mathbb{R}^{n\times n}$ is the agent-specific directory-dependency logic matrix of user i.

Reordered topic-wise form. Using the reordered topic-wise variables $y_k(t) = [x_1^k(t), \ldots, x_m^k(t)]^\top \in \mathbb{R}^m$ (the m users' opinions on directory k) and $\Gamma_{kj} = \mathrm{diag}(c_{kj,1}, \ldots, c_{kj,m}) \in \mathbb{R}^{m\times m}$, the same dynamics can be written as

$$y_k(t+1) = \sum_{j=1}^{n} \Gamma_{kj} W\, y_j(t), \tag{9}$$

or, in stacked block form,

$$y(t+1) = Ay(t), \tag{10}$$

with block entries $A_{kj} = \Gamma_{kj} W \in \mathbb{R}^{m\times m}$.

Open-topic steady-state relation. For an open directory (topic) p, once the predecessor directories on which it depends have already reached consensus, the corresponding steady-state relation is

$$y_p^* = (I_m - A_{pp})^{-1} \sum_{j\in\hat{J}_p} A_{pj}\, y_j^*, \tag{11}$$

which is then specialized by the singleton and multi-topic consensus conditions stated above.

Use in our simulation pipeline. The simulation pipeline is organized according to these theorem-guided SCC cases: closed irreducible blocks, open singleton blocks, and open multi-topic blocks.

4.3. Scope of Theoretical Guarantees

The above theorems characterize the *asymptotic* behavior of the process under *fixed* C_i. For our setting, two consequences are important to state clearly from the beginning, because they define exactly what the theory supports and where our empirical contribution starts:

1. **Healthy regime ⇒ consensus-forming behavior.** Under Assumptions 1–3, the healthy system evolves toward the consensus regimes characterized by Ye et al. In particular, closed irreducible blocks are consensus-forming, and open blocks are governed by the corresponding singleton or multi-topic consistency conditions. This is the theorem-guided steady-state behavior we use as the reference for an unattacked system.

2. **Anomaly regime lies outside Assumption 2.** A malicious modification that introduces a cross-SCC dependency in the logic matrices of one or more selected agents breaks structural homogeneity. Under Ye et al., this kind of mismatch can destroy the relevant consensus conditions—for example, a single competing logical interdependency can already break the singleton-topic condition—and can therefore move the affected block from a consensus-forming regime into a disagreement-prone one. Our detector does not require a separate convergence theorem for the perturbed case; it uses the failure of the healthy-regime consensus conditions as the structural mechanism behind anomaly formation.

The transient between regimes, during which our detector fires, is *not* characterized by Ye et al. That transient is an empirical claim: we observe and quantify it in Section 6, and report its sensitivity to perturbation strength, graph topology, and the Bayesian update parameters.

5. Detection Pipeline

5.1. SCC Decomposition and Theorem-Guided Evaluation

Following the decomposition framework of (Ye et al., 2021), we partition the topic index set into disjoint strongly connected components (SCCs), denoted by $\mathcal{J}_j$. After a suitable reordering of topics, the blocks are written as

$$\mathcal{J}_j \triangleq \left\{ \sum_{i=1}^{j} s_{i-1} + 1,\ \sum_{i=1}^{j} s_{i-1} + 2,\ \ldots,\ \sum_{i=1}^{j} s_{i-1} + s_j \right\},$$

where s_j is the number of topics in SCC $\mathcal{J}_j$ and $s_0 = 0$. This decomposition is central because the consensus conditions of Section 4 are naturally blockwise: closed irreducible SCCs, open singleton topics, and open multi-topic SCCs are governed by different theorem-guided cases.

Under Assumption 2, the logic matrices share a common zero/nonzero sparsity pattern, so a representative logical dependency graph is sufficient to identify the SCC blocks and their dependency structure. Each block is then classified as open or closed according to whether it depends on topics outside the block, and the resulting SCC graph induces the dependency order for evaluation.

The detection pipeline therefore proceeds block by block rather than by evolving the full system monolithically. Each SCC is processed only after the consensus information required from its predecessor blocks has been resolved, and the corresponding theorem-guided case is then applied. The resulting blockwise outputs provide both the steady-state reference used in the healthy regime and the localized trajectories used by the anomaly detector. The overall procedure is summarized in Algorithm 1; implementation details of SCC annotation, dependency bookkeeping, and theorem dispatch are deferred to Appendix 9.1.

Algorithm 1 Theorem-Guided SCC Evaluation

```
Input: representative logic matrix, agent logic matrices {C_i},
  influence matrix W, horizon T
Output: blockwise trajectories and final consensus values
extract SCC blocks and construct the block-dependency DAG
initialize pending blocks and external consensus store
while unprocessed blocks remain do
    select blocks whose predecessors have already been resolved
    for all ready block J_j do
        identify the corresponding theorem-guided case
        simulate the associated block dynamics for T steps
        record the block trajectory and final state
        pass resolved consensus values to downstream dependent blocks
    end for
end while
return all blockwise outputs
```

5.2. Variance Signal

For each topic p and step t, the cross-agent variance is $V_p(t) = \mathrm{Var}_i\big(x_i^p(t)\big)$. Theorems 2 and 4 imply $V_p(t) \to 0$ in closed SCCs (and bounded in open SCCs at consensus); structural-homogeneity violations inflate $V_p(t)$ during the

transient. We use the scaled mean

$$v(t) = \frac{1}{|J|} \sum_{p \in J} \mathrm{Var}_i\big(s\, x_i^p(t)\big), \quad \Delta v_t = \max\big(v(t) - v(t-1),\, 0\big) \tag{12}$$

as our regime-violation signal. The choice of variance over alternative signals (e.g. $\|C_i(t) - C_i(t-1)\|_F$ or spectral-gap proxies) is operational, not inherited from (Ye et al., 2021); we compare against $\|C\|_F$ as a baseline in Section 6.

5.3. Sequential Bayesian Update

We treat the presence of an anomaly as a latent binary state $H \in \{0, 1\}$, assumed approximately constant within a short detection window. At step t, we observe Δv_t. We posit the likelihood ratio

$$\frac{p(\Delta v_t \mid H{=}1)}{p(\Delta v_t \mid H{=}0)} \propto \frac{1 - e^{-\alpha \Delta v_t}}{e^{-\alpha \Delta v_t}}, \tag{13}$$

that is, the score $L_t = 1 - e^{-\alpha \Delta v_t}$ acts as the $H{=}1$ likelihood surrogate. Under conditional independence of $\{\Delta v_t\}$ given H, the sequential posterior is

$$\pi_t \;=\; \frac{L_t\, \pi_{t-1}^{\mathrm{eff}}}{L_t\, \pi_{t-1}^{\mathrm{eff}} + (1 - L_t)\,(1 - \pi_{t-1}^{\mathrm{eff}})}. \tag{14}$$

Two choices for the effective prior π_{t-1}^{eff} define two operating modes:

- **Static prior:** $\pi_{t-1}^{\mathrm{eff}} = \pi_0$ (fixed, e.g. $\pi_0 = 0.1$). Each step is treated as a one-shot test.
- **Online prior:** $\pi_{t-1}^{\mathrm{eff}} = \pi_{t-1}$ (the previous posterior). This can be viewed as a sequential evidence accumulator, in the spirit of classical change-detection procedures such as CUSUM (Page, 1954).

The underlying probabilistic model is intentionally simple: a constant-H two-state model with i.i.d. $\Delta v_t \mid H$. This is sufficient to make (14) a principled sequential posterior under the assumed model. Whether π_t is also a well-calibrated probability in practice, rather than simply a monotone anomaly score, remains an empirical question that we examine in Section 6.

Why this is well aligned with the theory. The results in Section 4 characterize the consensus-forming behavior of the healthy regime. In that regime, cross-agent variance shifts are expected to remain small, so the posterior tends to stay close to its prior. When Assumption 2 is broken by a malicious modification of C_i, the relevant consensus conditions can fail, and the affected SCC may move into a disagreement-prone regime. In that case, Δv_t increases and the posterior shifts toward $H{=}1$. The Bayesian layer provides noise robustness and sequential evidence accumulation, while the structural layer identifies *which SCC* is responsible for the shift.

Algorithm 2 Bayesian anomaly score (online prior)

Input: opinion matrices $x_{\text{prev}}, x_{\text{now}}$, prior π_{t-1}, scale s, exponent α
$v_{\text{cur}} \leftarrow \frac{1}{|J|} \sum_p \mathrm{Var}_i(s\, x_{\text{now}}^{(:,p)})$
$v_{\text{prev}} \leftarrow \frac{1}{|J|} \sum_p \mathrm{Var}_i(s\, x_{\text{prev}}^{(:,p)})$
$\Delta v \leftarrow \max(v_{\text{cur}} - v_{\text{prev}},\, 0)$
$L \leftarrow 1 - \exp(-\alpha \Delta v)$
$\pi_t \leftarrow \dfrac{L\, \pi_{t-1}}{L\, \pi_{t-1} + (1 - L)(1 - \pi_{t-1})}$
return π_t

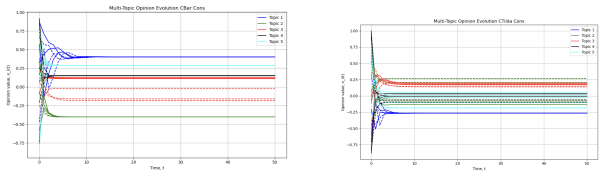

Figure 2. Replication of theorem-guided opinion evolution from Arxiv simulations (Ye et al., 2018). Left: nominal logic configuration. Right: perturbed logic configuration with disagreement induced by competing interdependencies.

6. Experiments

6.1. Implementation validation by replication

To validate correctness of the implementation, we first reproduce the multi-topic opinion-dynamics simulation described in (Ye et al., 2018), as illustrated in Figure 2. The purpose of this experiment is to verify that the SCC decomposition, theorem routing, and resulting opinion trajectories match the expected convergence and disagreement patterns implied by the original theory.

In particular, the replication confirms that:

- under the nominal logic configuration, theorem-routed SCC blocks exhibit the expected convergence behavior;
- when the perturbed logic matrix $\tilde{C}$ introduces competing interdependencies, the expected disagreement pattern appears;
- the resulting trajectory plots provide a sanity check that our implementation faithfully follows the theorem-guided simulation pipeline.

The implementation reproducing the Ye-style simulation together with the anomaly-detection code is available in the anonymized code repository at `https://anonymous.4open.science/r/icml-submission-code-808F/`.

6.2. Testbed

We use a synthetic but theoretically grounded enterprise with seven directories $D_1, \ldots, D_7$ and seven users $u_1, \ldots, u_7$.

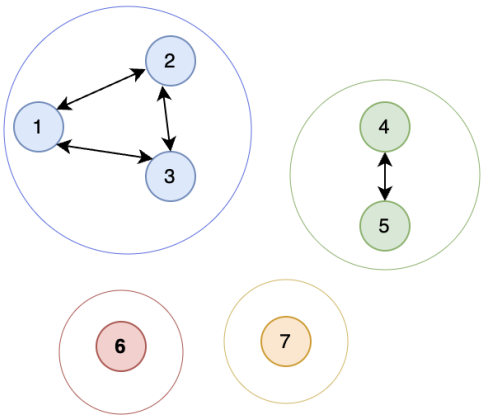

Figure 3. Baseline directory/topic SCC structure under the nominal logic matrix $\hat{C}$.

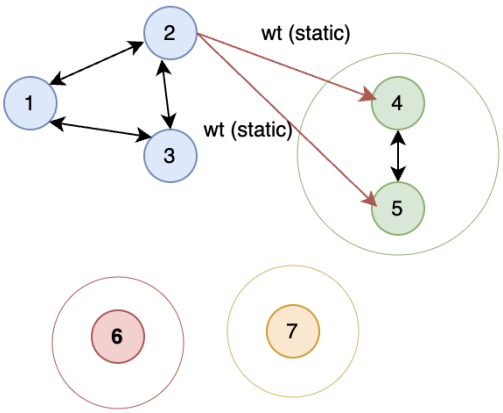

Figure 4. Perturbed logic structure used in the anomaly experiments. The perturbation is applied to the logic matrices of users u_4 and u_5; inside those matrices, directories D_4 and D_5 acquire a stronger dependency on directory D_2.

The matrix W encodes the user/agent influence structure used throughout the simulation:

$$W = \begin{bmatrix} 0.281 & 0.334 & 0.334 & 0 & 0 & 0.050 & 0 \\ 0.333 & 0.333 & 0.333 & 0 & 0 & 0 & 0 \\ 0.317 & 0.317 & 0.317 & 0 & 0 & 0.048 & 0 \\ 0.050 & 0 & 0 & 0.450 & 0.500 & 0 & 0 \\ 0.050 & 0 & 0 & 0.450 & 0.500 & 0 & 0 \\ 0.100 & 0.050 & 0 & 0.100 & 0.050 & 0.650 & 0.050 \\ 0 & 0 & 0 & 0 & 0 & 0.050 & 0.950 \end{bmatrix}.$$

The baseline logic matrix $\hat{C}$ is row-stochastic and induces four directory/topic SCC blocks:

$$\{D_1, D_2, D_3\}, \qquad \{D_4, D_5\}, \qquad \{D_6\}, \qquad \{D_7\}.$$

In the nominal regime, all users/agents use the same baseline directory-logic matrix $\hat{C}$. We simulate the baseline system for $T = 5000$ steps and use the resulting blockwise final states as the healthy reference.

Perturbation. To generate anomalous behavior, we modify the agent-specific logic matrices of users u_4 and u_5. In those perturbed matrices, we increase the topic-dependency entries $c_{4,2,i}$ and $c_{5,2,i}$, so that directories D_4 and D_5 depend more strongly on directory D_2, followed by row normalization. Operationally, the agent-specific logic-matrix list takes the form

$$[C_{\text{base}}]^3 + [C_{\text{pert}}(w_t)]^2 + [C_{\text{base}}]^2,$$

so only users u_4 and u_5 use the modified logic matrix. This breaks Assumption 2 across agents while keeping the anomalous topic dependency localized to the attacked directory block. More details are provided in Appendix 9.1..

6.3. Static vs. Online Bayesian Detection

We compare the attacked block under the baseline and perturbed systems and convert the resulting variance shift into a Bayesian anomaly score. The implementation uses the scaled-variance score from Section 5.3 with fixed parameters $s = 500$ and $\alpha = 0.01$.

For the *static-prior* detector, we sweep

$$w_t \in \{1, 2, 5, 10, 50, 100, 250, 500, 1000\},$$

using the fixed prior $\pi_0 = 0.1$ at every perturbation level. For the *online-prior* detector, we sweep

$$w_t \in \{2, 5, 10, 50, 100, 250, 500, 1000, 2000\},$$

and reuse the posterior from the previous perturbation level as the prior for the next one. The online setting therefore acts as a sequential evidence-accumulation view of the detector across an escalating perturbation schedule. Figure 5 shows the resulting curves.

Findings. First, the online-prior detector saturates at smaller perturbation strengths than the static-prior version, showing greater sensitivity to repeated positive evidence. Second, the posterior rises monotonically with perturbation strength in both modes, which is consistent with the increased variance shift induced by stronger cross-SCC topic perturbations. Third, the online curves are best interpreted as evidence accumulation across an ordered sweep, rather than as independent one-shot decisions at each w_t.

The additional ablation studies, reported in Appendix 9.2, evaluate detector robustness under asymmetric cross-SCC topic perturbations, benign within-SCC perturbations, and SCC-aware versus global variance scoring. An anonymized code repository containing the replication scripts, anomaly-detection simulations, and ablation studies is available at `https://anonymous.4open.science/r/icml-submission-code-808F/`.

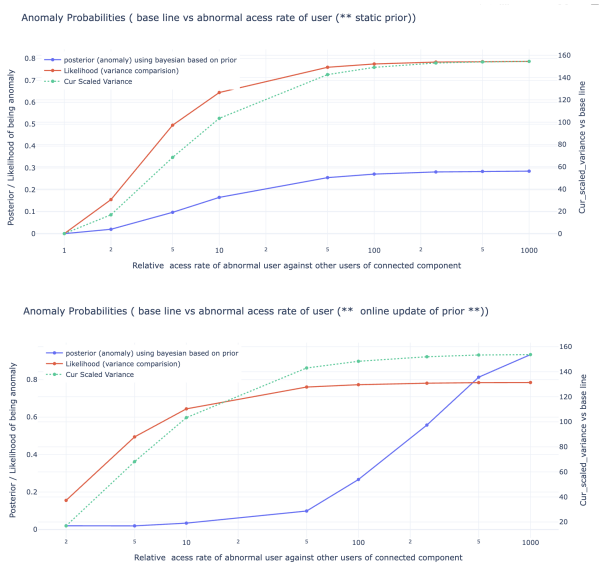

Figure 5. Bayesian anomaly score under static (top) and online (bottom) prior updates over a perturbation-strength sweep. The online detector accumulates evidence across rounds and saturates earlier than the static-prior baseline.

7. Discussion

Agentic-system implications. The framework is agnostic to whether the entities accessing directories are humans or autonomous agents. As enterprises deploy more agentic assistants and tool-using AI agents, the same multi-level access graph emerges, but with a different risk profile documented in recent threat-intelligence and agentic security work (Google Cloud Mandiant Threat Intelligence, 2024; Greshake et al., 2023; Zhan et al., 2024). This has three practical consequences. First, a compromised agent can issue cross-SCC access requests much faster than a human user, so detection latency matters; the online-prior detector's earlier flagging is therefore useful in machine-speed threat regimes. Second, attribution becomes harder under delegated agentic workflows; SCC-level decomposition gives a structural explanation, e.g., "SCC j lost consensus on topic p," without first requiring complete attribution to a specific human or agent. Third, indirect prompt-injection and tool-hijack attacks often aim to make an otherwise benign agent produce cross-trust-boundary behavior. Such behavior is naturally modeled as a structural-homogeneity violation in the agent-specific logic matrices C_i.

Fit with agentic security operations. Recent industry guidance positions agentic security operations centers, where autonomous triage and investigation agents consume alerts, gather evidence, and explain verdicts, as one operational response to machine-speed threats (Google Cloud Mandiant Threat Intelligence, 2024). Our pipeline produces an input that is well matched to this setting: a structurally localized verdict such as "π_t exceeded threshold because SCC j on topic p lost consensus after a cross-SCC topic dependency appeared in selected C_i." This decouples detection from response more cleanly than opaque UEBA scoring and gives an automated investigation loop a graph-level cause rather than only a scalar anomaly score.

Outside the theoretical framework. We have not characterized the transient dynamics under C-perturbations theoretically, shown calibration of π_t as a probability rather than a monotone anomaly score, or validated the method on real enterprise logs. These are natural next steps. The present work shows that Ye et al.'s structural-homogeneity assumption provides a useful mathematical handle for malicious access modifications, and that a sequential Bayesian detector can turn this handle into an online signal with structural explainability.

7.1. Limitations and Open Issues

Our evaluation uses a small synthetic testbed with $m = n = 7$ users and directories. Real deployments may involve thousands of entities and evolving access relationships. Although the method relies on lightweight Bayesian updates, SCC decomposition, and matrix computations, large-scale deployment would require incremental SCC maintenance, cached theorem-routing tables, and lazy re-evaluation of unchanged blocks. Future iterations of this research will focus on transitioning from this controlled testbed to real-world audit logs to evaluate how the SCC-decomposition handles the thousands of nodes and noisy relationships typical of global enterprise infrastructures.

The Bayesian likelihood in (13) is hand-crafted through the sensitivity parameter α. Future work should learn this parameter from historical audit data or replace the parametric likelihood with a more flexible model, especially to reduce false positives caused by benign role changes, team transfers, or temporary access expansions.

Finally, detection must be connected to operational responses such as alerting, access review, remediation, and audit explanations. Long-term consensus drift also requires secure recalibration, versioning, and privacy-aware updates of the logic matrices C_i.

8. Conclusion

We have shown that viewing enterprise access graphs through the lens of heterogeneous belief systems gives a principled mathematical handle on a class of malicious access events that violate the structural-homogeneity assumption underlying the consensus theorems of Ye et al. Such violations can appear as transient rises in cross-agent opinion variance, which we convert into an online anomaly score

through a sequential Bayesian update with a two-state latent indicator. The SCC-level decomposition makes the detector structurally interpretable, rather than reducing the event to an opaque anomaly score.

Empirically, the online-prior variant detects weaker perturbations earlier than the static-prior baseline. The ablations further show that SCC-aware variance localizes the affected block more clearly than global variance and that benign within-SCC reweighting does not trigger the detector in the same way as a homogeneity-breaking cross-SCC topic perturbation. This makes the framework well suited to agentic settings, where cross-SCC access perturbations may propagate at machine speed and early, structurally grounded detection becomes especially important.

More broadly, consensus-guided monitoring can serve as a latent, stealth detection layer within larger security systems. As highly capable, autonomous agents increasingly operate inside enterprise environments, threat actors will succeed only if their rogue agents can seamlessly mimic normal behavior across multi-level behavioral graphs. However, because our framework relies on hidden structural monitoring models, it detects the subtle topological shifts caused by agentic traversal. Given this global sensitivity, our proposed solution offers an important foundational component for next-generation enterprise defense pipelines.

References

Adams, R. P. and MacKay, D. J. C. Bayesian online changepoint detection. Technical Report arXiv:0710.3742, University of Cambridge, 2007. URL https://arxiv.org/abs/0710.3742.

Akoglu, L., Tong, H., and Koutra, D. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

Cui, J., Zhang, G., Chen, Z., et al. Multi-homed abnormal behavior detection algorithm based on fuzzy particle swarm cluster in user and entity behavior analytics. *Scientific Reports*, 12:22349, 2022. doi: 10.1038/s41598-022-26142-w.

DeGroot, M. H. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

Ding, K., Li, J., Bhanushali, R., and Liu, H. Deep anomaly detection on attributed networks. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pp. 594–602, 2019.

Franklin, M., Tomašev, N., Jacobs, J., Leibo, J. Z., and Osindero, S. AI agent traps. arXiv preprint, 2024. URL https://arxiv.org/abs/2409.11364.

Friedkin, N. E. and Johnsen, E. C. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3–4): 193–206, 1990.

Google Cloud Mandiant Threat Intelligence. Defending enterprise AI: Vulnerabilities and mitigations. https://cloud.google.com/blog/topics/threat-intelligence/defending-enterprise-ai-vulnerabilities, 2024. Industry threat-intelligence report on agentic and LLM-driven enterprise AI risks. Accessed: 2026-04-24.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023.

Hegselmann, R. and Krause, U. Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5 (3), 2002.

Khaliq, S., Tariq, Z. u. A., and Masood, A. Role of user and entity behavior analytics in detecting insider attacks. In *2020 IEEE International Conference on Cyber Warfare and Security (ICCWS)*, pp. 1–6, 2020.

Liu, Y., Li, Z., Pan, S., Gong, C., Zhou, C., and Karypis, G. Anomaly detection on attributed networks via contrastive self-supervised learning. In *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Microsoft Security Team. What is user and entity behavior analytics (ueba)? https://www.microsoft.com/en-us/security/business/security-101/what-is-user-entity-behavior-analytics-ueba, 2024. Accessed: 2025-03-15.

Page, E. S. Continuous inspection schemes. *Biometrika*, 41 (1/2):100–115, 1954.

Shiryaev, A. N. On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8 (1):22–46, 1963.

Vectra AI. 2023 state of threat detection report, 2023. URL https://www.vectra.ai/resources/2023-state-of-threat-detection. Accessed: March 14, 2025.

Ye, M., Liu, J., Wang, L., Anderson, B. D. O., and Cao, M. Consensus and disagreement of heterogeneous belief systems in influence networks. arXiv preprint, 2018. URL https://arxiv.org/abs/1812.05138.

Ye, M., Liu, J., Wang, L., Anderson, B. D. O., and Cao, M. Consensus and disagreement of heterogeneous belief systems in influence networks. *IEEE Transactions on Automatic Control*, 66(11):5266–5281, 2021. doi: 10.1109/TAC.2020.3046576.

Yue Song, J. Y. Insider threat detection based on user and entity behavior analytics. In *Proceedings of the 2023 ACM Conference on Security and Privacy*, pp. 123–134, 2023. URL https://dl.acm.org/doi/10.1007/978-3-031-75764-8_17. Online; accessed 2025-03-15.

Zhan, Q., Liang, Z., Ying, Z., and Kang, D. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics (ACL)*, 2024.

9. Appendix

9.1. Detailed SCC Annotation and Dependency-Aware Simulation

Algorithm 3 gives the implementation-level SCC processing routine used in our experiments. Given a representative logic matrix C, where $C[p, q] \neq 0$ means that topic p depends on topic q, the code first builds a theorem-routing table containing the SCC labels, open/closed status, local dependencies, external dependencies, dependency order, and theorem-guided dispatch label. Stable SCC labels $J_1, J_2, \ldots$ are assigned by sorting SCCs according to their minimum topic index.

The key convention is that an edge $J_a \to J_b$ in the SCC condensation graph means that block J_a depends on block J_b. Therefore, as shown in Algorithm 3, a block is simulated only after the blocks it depends on have been resolved and their external consensus values are available.

The theorem-routing labels in Algorithm 3 are used only for dispatching the appropriate simulation branch. The corresponding theoretical conditions are stated separately in Section 4.

Algorithm 3 Dependency-Aware Per-SCC Simulation

Input: theorem table `cdf`, logic matrices $\{C_i\}$, influence matrix W, horizon T
Output: blockwise trajectories and final block states
initialize pending blocks, completed blocks, external consensus store, and output tables
while pending blocks remain **do**
  identify all blocks whose dependency blocks are completed and whose external topic values are available
  if no block is ready **then**
    raise a dependency-resolution failure
  end if
  for all ready block J_j **do**
    read its topics, theorem-routing label, and required external dependencies
    if J_j is a closed multi-topic block **then**
      simulate the closed multi-topic branch
    else if J_j is a closed singleton block **then**
      simulate the singleton branch
    else if J_j is an open singleton block with resolved external dependencies **then**
      simulate the open-singleton branch using the stored external values
    else
      simulate the open multi-topic branch using the stored external values
    end if
    store the trajectory and final block state
    update the external consensus store with the resolved topic values from J_j
    mark J_j as completed and remove it from the pending set
  end for
end while
return stored trajectories and final values

9.2. Ablations

We run three ablations on the same seven-agent/seven-directory testbed. Unlike the main anomaly experiment, which increases the dependency from D_2 into both topics in the attacked block $\{D_4, D_5\}$, the ablation setting uses an asymmetric cross-SCC topic perturbation: in the selected perturbed agent logic matrices, the dependency from D_2 is increased into only one attacked topic, D_4. This creates within-SCC disagreement inside the attacked block and is used to stress-test whether SCC-aware variance localizes the affected component. We use a steady-state warm-up of $T_{\text{warm}} = 2000$, track the post-injection transient for $T_{\text{track}} = 20$ steps, and use the same scale factor $s = 500$.

E1: Evidence score versus baseline signals. For each cross-SCC topic weight

$$w_t \in \{1, 2, 5, 10, 50, 100, 250, 500, 1000\},$$

we record the maximum transient SCC-aware variance jump Δv_{SCC}, the Frobenius structural-change baseline $\|C_{\text{pert}} - C_{\text{base}}\|_F$, and the Bayesian evidence score

$$L(\Delta v) = 1 - \exp(-\alpha_{\text{E1}} \Delta v).$$

For visualization, α_{E1} is chosen so that the largest observed SCC-aware variance jump maps to $L(\Delta v) \approx 0.95$. This diagnostic scaling is used only to compare the shape of the bounded evidence score with the raw SCC-variance signal and the Frobenius baseline; the main detection experiment uses the fixed $\alpha = 0.01$. Figure 6 shows that the evidence score follows the SCC-aware variance evidence while remaining bounded, whereas $\|\Delta C\|_F$ measures only the magnitude of the logic-matrix perturbation.

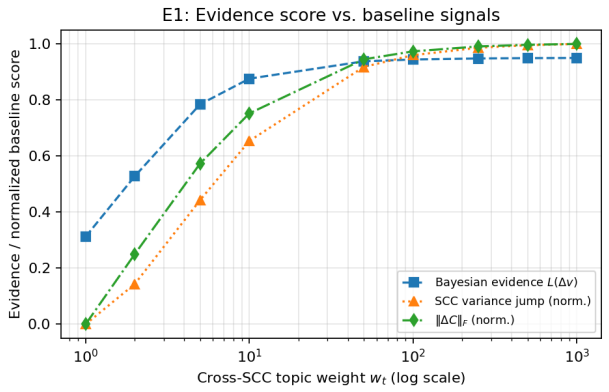

Figure 6. E1: bounded Bayesian evidence score versus SCC-aware variance jump and Frobenius structural-change baseline as the cross-SCC topic weight w_t increases. Raw baseline signals are min–max normalized.

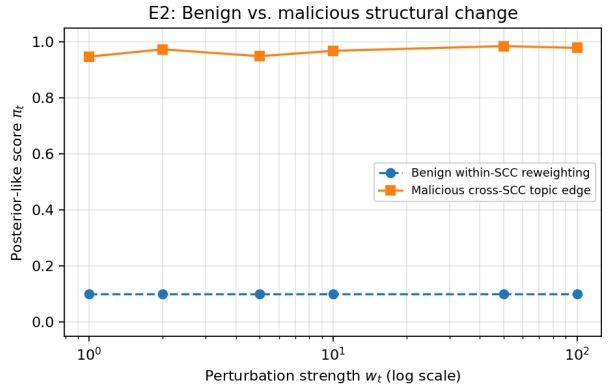

Figure 7. E2: posterior-like score under benign within-SCC reweighting versus a malicious cross-SCC topic edge. Only the homogeneity-breaking cross-SCC perturbation fires.

E2: Benign within-SCC change versus malicious cross-SCC change. We compare two perturbation types. The benign perturbation reweights an existing within-SCC topic edge and therefore preserves the shared sparsity pattern required by Assumption 2. The malicious perturbation introduces a cross-SCC topic dependency from D_2 into D_4 in the selected perturbed agent logic matrices, violating structural homogeneity. Figure 7 shows that the benign perturbation leaves the posterior-like score close to the prior, while the malicious perturbation drives the score sharply upward. This confirms that the detector responds to homogeneity-breaking cross-SCC topic changes rather than arbitrary within-SCC weight changes.

E3: SCC-aware versus global variance. At fixed $w_t = 10$, Figure 9 tracks the variance trajectory under three scopes: global variance, per-SCC variance for $\mathrm{SCC}_1 = \{D_1, D_2, D_3\}$, and per-SCC variance for the attacked block

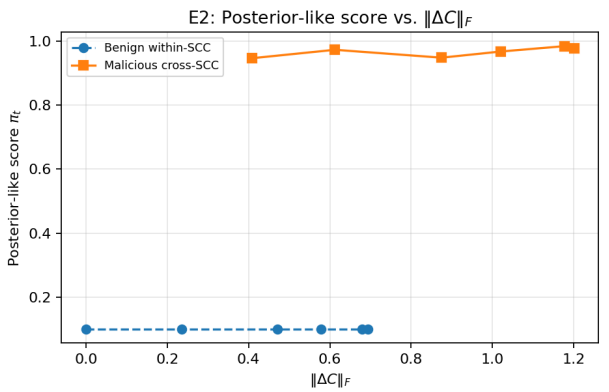

Figure 8. E2: posterior-like score plotted against $\|\Delta C\|_F$. The malicious cross-SCC perturbation produces a high score, whereas benign within-SCC reweighting remains near the prior even as the structural-change magnitude varies.

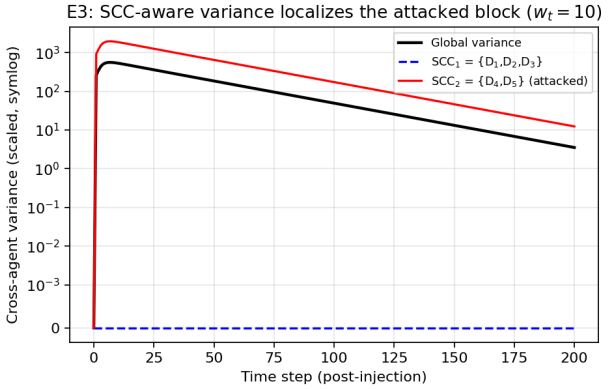

Figure 9. E3: post-injection variance trajectory at $w_t = 10$. SCC-aware variance localizes the attacked block $\{D_4, D_5\}$, whereas the global variance aggregates across affected and unaffected topics.

$\mathrm{SCC}_2 = \{D_4, D_5\}$. The SCC-aware view localizes the attack: variance in SCC_2 rises sharply after injection and then decays, while the unaffected block remains near zero. The global signal is less localized because it aggregates unaffected and affected components. This supports the use of SCC-aware variance in Algorithm 2 when SCC structure is available.

9.3. Asymmetric $m \neq n$ Implementation Sanity Check

To verify that the implementation does not rely on the special case $m = n = 7$, we additionally construct a six-agent/eight-directory variant inspired by the main anomaly-detection testbed. In this setting,

$$W \in \mathbb{R}^{6\times 6}, \qquad C_i \in \mathbb{R}^{8\times 8}, \qquad x_i(t) \in \mathbb{R}^8.$$

The baseline topic SCCs are

$$\{D_1, D_2, D_3\}, \qquad \{D_4, D_5\}, \qquad \{D_6, D_7\}, \qquad \{D_8\}.$$

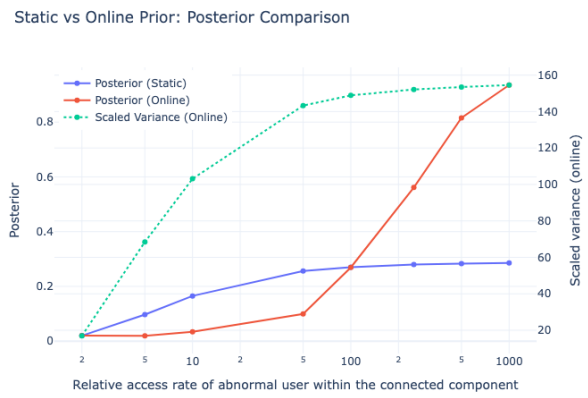

Figure 10. Comparative Dynamics of Static vs. Online Prior Updates. This figure illustrates the trade-off between the two Bayesian modes, showing that online-prior recursion enables the rapid evidence accumulation necessary for machine-speed threat detection

As in the main experiment, the anomaly perturbs only selected agent-specific logic matrices. Specifically, agents u_4 and u_5 receive the perturbed logic matrix, in which the dependency from D_2 into the attacked topic block $\{D_4, D_5\}$ is strengthened. This experiment is used only as an implementation sanity check for the corrected dimensional convention, not as a replacement for the main $m = n = 7$ benchmark.

Figure 11. Asymmetric $m = 6, n = 8$ sanity check under a static prior. The posterior-like anomaly score increases with the cross-SCC topic weight w_t, tracking the transient variance induced in the attacked topic block.

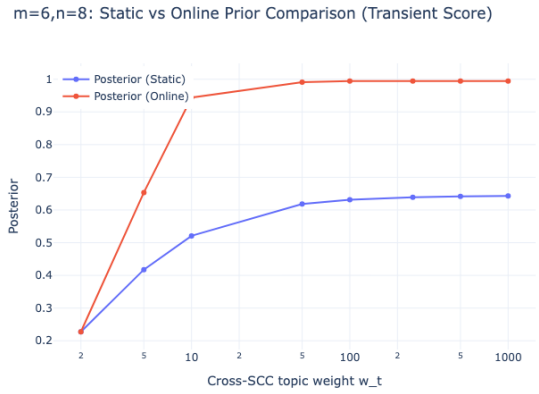

Figure 12. Static-prior versus online-prior anomaly scores for the asymmetric $m = 6, n = 8$ sanity check. The online-prior recursion accumulates evidence more rapidly, while the static-prior score increases more gradually with perturbation strength.