RELIABLE ROUTING WITH THOMPSON SAMPLING: DISCOVERY BARRIERS AND ROUTE CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

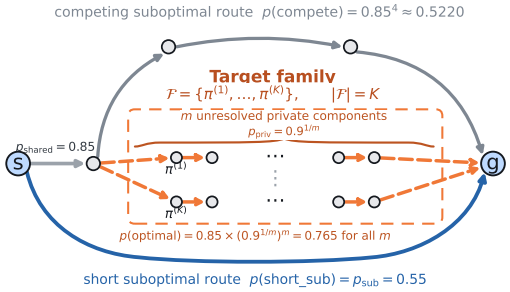
Learning the most reliable route under initially unknown link success probabilities is a fundamental problem in stochastic routing, where route reliability is multiplicative. Thompson sampling (TS), with component-wise Bayesian learning and information sharing across overlapping routes, is highly competitive on such problems. However, multiplying independent posterior samples can sharply suppress the probability that an underexplored route receives a competitive score. Consequently, TS may repeatedly select suboptimal competitors while a more reliable route remains unexplored. We formalize this phenomenon as a compositional discovery barrier, driven by unresolved family-private edges, whose uncertainty persists as competing routes are explored. We establish a general posterior-dependent upper bound on route-selection probability and corresponding lower bounds on first-selection round and cumulative regret. For standard TS with $\mathrm{Beta}(1,1)$ priors, we show that $\log \mathbb{E}[\tau_{\mathcal{F}}] \geq m \log m - \log K - O_p(m)$ on a controlled overlapping path-trap family with online competitor learning, where $\tau_{\mathcal{F}}$ is the first selection of one of K target routes, each containing m unresolved private edges, and p is the reliability of the competitor's exclusive edge. At a fixed positive optimality gap, this delay yields linear regret over horizons that scale factorially with m. Guided by this analysis, we propose Route-Calibrated Thompson Sampling (RC-TS), which compensates for multiplicative pessimism through a theoretically motivated calibration of edge-level prior optimism. The calibrated prior preserves competitive tail probabilities, yielding a first-discovery upper bound uniform in m on the same instance family for fixed p and calibration λ. Experiments under both semi-bandit and cascade feedback show that RC-TS substantially improves robustness in challenging cold-start regimes while preserving competitive performance on ordinary routing instances.

1 INTRODUCTION

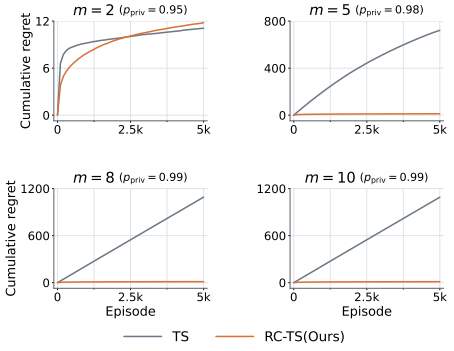
Learning the most reliable route when link success probabilities are initially unknown is a fundamental problem in stochastic routing, with applications in communication over unreliable links (Chen et al., 2011), transportation networks (Prakash, 2020), stochastic logistics networks (Lin & Yeh, 2010), and payment routing networks (Waugh & Holz, 2020). When an end-to-end attempt succeeds only if every selected link succeeds, independent link outcomes induce a multiplicative route reliability $\mu(\pi) = \prod_{e \in \pi} p_e$. We study two feedback models. Under selected-route semi-bandit feedback (Wang & Chen, 2018b; Hüyük & Tekin, 2019), the learner observes all selected link outcomes; under cascade feedback (Kveton et al., 2015a; Zhong et al., 2021a), execution stops at the first failed link and reveals the attempted prefix. In both settings, the learner must discover highly reliable routes online while limiting costly failures during exploration.

Thompson sampling (TS) (Thompson, 1933; Russo et al., 2018) combines simple Beta–Bernoulli updates, information sharing across overlapping routes, and efficient shortest-path optimization. Although combinatorial TS admits logarithmic-in-horizon regret under suitable assumptions (Wang & Chen, 2018b), its dependence on action size can be severe. Zhang & Combes (2021) establish super-exponential discovery delays in a multiplicative example with a fixed competing score, and Jin et al. (2026) address coordinated optimism through shared-seed sampling. We study how this exploration difficulty depends on route overlap and evolving competitor posteriors, and how prior calibration addresses it while retaining independent Beta sampling.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Routing instance. A one-link short route ($p = 0.55$) and a competing route ($0.85^4 \approx 0.522$) compete with K target routes. Each target shares one link ($p = 0.85$) with the competing route and contains m unresolved private edges (dashed), each with $p_{\text{priv}} = 0.9^{1/m}$. Target reliability therefore remains 0.765 for every m.



(b) Cumulative regret of $\text{Beta}(1, 1)$ TS and RC-TS under semi-bandit feedback with $K = 4$. As m increases, standard TS exhibits prolonged periods of approximately linear regret despite fixed target reliability and increasingly reliable private edges. Curves show means over 200 runs.

Figure 1: Compositional discovery barrier of Thompson sampling.

Figure 1(a) illustrates the mechanism. Two suboptimal routes compete with K target routes, each containing m initially unobserved edges accessible exclusively through the target family. Shared-edge samples cancel in pairwise route comparisons, leaving the respective exclusive-edge products. Exploring competitors updates shared edges, while family-private edges retain their priors until a target is selected. Under independent $\text{Beta}(1, 1)$ priors, their sampled product $Z_m = \prod_{i=1}^{m} U_i$, with $U_i \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$, satisfies $\Pr(Z_m \geq q) \leq [\log(1/q)]^m/m!$ for every fixed $q \in (0, 1)$. When competitor samples typically exceed a competitive threshold, this small tail probability suppresses the selections needed to resolve private uncertainty. We call this interaction the *compositional discovery barrier*. Figure 1(b) shows its cost: as m grows, target reliability remains fixed and individual private edges become more reliable, yet TS exhibits prolonged periods of approximately linear regret. The barrier already arises under semi-bandit feedback, where a target selection reveals all its edge outcomes; cascade feedback adds censoring during execution.

These observations raise two questions: *How does private-edge uncertainty produce discovery delay as competing posteriors evolve?* And *can prior calibration restore effective discovery while preserving the simplicity of TS?*

We answer the first question through selection probabilities and their online consequences. First, we derive a posterior-dependent route-selection upper bound that separates target-side and reference-side exclusive uncertainty. We then analyze a controlled overlapping path-trap family that allows a common suffix shared with the competitor and edge overlap among targets. The competitor's exclusive edge is observed in every pre-discovery round, enabling an explicit analysis of learning from cold start under both feedback models. Writing $\tau_{\mathcal{F}}$ for the first selection of a target, standard TS satisfies $\log \mathbb{E}[\tau_{\mathcal{F}}] \geq m \log m - \log K - O_p(m)$, where p is the reliability of the competitor's exclusive edge. For fixed p and K, discovery delay thus grows factorially with the number of private edges, even at a fixed positive optimality gap. This delay also yields linear regret over factorially long horizons.

To answer the second question, we propose *Route-Calibrated Thompson Sampling* (RC-TS). The algorithm retains independent Beta sampling and standard updates, using the prior $\text{Beta}(1, \lambda/H)$, where H upper-bounds route length and $\lambda > 0$ controls prior optimism. A uniform prior induces mean sampled score 2^{-L} on an unobserved route of length L; inverse-length calibration keeps this mean at least $e^{-\lambda}$. More strongly, we prove that the product of any $m \leq H$ edges retaining their calibrated priors exceeds each fixed threshold $q < 1$ with probability bounded below by a positive constant depending only on q and λ. Combining this competitive-tail guarantee with online competitor concentration gives $\mathbb{E}[\tau_{\mathcal{F}}] = O_{p,\lambda}(1)$ on the same overlapping path-trap family under both feedback models. For fixed p and λ, this upper bound is uniform in m, establishing a sharp

separation from standard TS during first discovery. Experiments on controlled and general routing benchmarks show that RC-TS substantially reduces cold-start exploration delays while retaining competitive performance under both feedback models. Our contributions are three-fold:

1. We characterize the *compositional discovery barrier* through unresolved family-private edges and derive a posterior-dependent selection bound accounting for route overlap.
2. We establish a first-discovery separation on controlled overlapping path-trap instances with online competitor learning: factorial-scale delay for standard TS and a length-uniform upper bound for RC-TS under both feedback models.
3. We develop inverse-length prior calibration with a uniform competitive-tail guarantee and evaluate the resulting RC-TS algorithm on controlled and general routing benchmarks.

2 Related Work

Cascading feedback and reliable routing. Cascading bandits model sequential observations in recommendation (Kveton et al., 2015a), and combinatorial cascading bandits extend this framework to conjunctive rewards, including reliable routing with first-failure feedback (Kveton et al., 2015b). Both fit the CMAB framework with probabilistically triggered arms (Chen et al., 2016). Triggering-probability-modulated smoothness yields regret bounds that avoid inverse minimum-triggering-probability factors (Wang & Chen, 2017), while variance-aware methods improve action-size dependence (Liu et al., 2024). These UCB-style methods explore through optimistic confidence indices, and we include them as baselines. Our analysis complements these guarantees by characterizing a selection bottleneck of independent posterior sampling under both semi-bandit and cascade feedback.

Exploration difficulty in combinatorial TS. Combinatorial TS admits logarithmic-in-horizon regret bounds under suitable assumptions for semi-bandit and probabilistically triggered feedback (Wang & Chen, 2018a; Hüyük & Tekin, 2019; 2020). Subsequent work improves statistical efficiency and addresses approximation oracles (Perrault et al., 2021; Perrault, 2023). These finite-time guarantees can contain terms growing exponentially with action size. Zhang & Combes (2021) establish severe exploration delays, including a multiplicative example with super-exponential first-discovery delay against a fixed competing score. We build on this mechanism by identifying persistent family-private uncertainty in overlapping routes and analyzing competitor learning from cold start. On a controlled overlapping path-trap family, we establish factorial-scale discovery delay for standard TS and a length-uniform discovery-time upper bound for calibrated TS.

Sampling distributions and coordinated exploration. TS-Cascade coordinates exploration in cascading recommendation through a shared Gaussian perturbation (Zhong et al., 2021b). For linear combinatorial semi-bandits, Zhang & Combes (2024) develop Gaussian sampling with polynomial dimension dependence. More recently, Jin et al. (2026) introduce shared-seed coupling and establish polynomial regret guarantees for general CMABs under monotonicity and bounded smoothness. These results demonstrate the benefits of designing sampling distributions and dependence across components. RC-TS preserves independent Beta sampling and conjugate updates through a calibrated prior. Its theoretical guarantee concerns first discovery on the specified instance family, while its empirical evaluation covers general routing graphs.

Prior sensitivity and optimism in TS. TS performance depends on prior choice (Liu & Li, 2016), and optimistic variants have been developed for single-arm bandits (May et al., 2012). We study how component priors determine competitive tail probabilities after multiplicative composition. For any fixed $b > 0$, products of independent $\mathrm{Beta}(1, b)$ samples exhibit super-exponentially decaying probabilities of exceeding a fixed positive threshold as route length grows. With $b = \lambda/H$, every initially unobserved route of length at most H exceeds each fixed threshold $q \in (0, 1)$ with probability bounded below by a positive constant depending only on q and λ. This tail guarantee provides a direct exploration justification for inverse-length prior calibration.

3 Problem Formulation

We consider a directed graph $G = (V, E)$ with distinct source $v_s \in V$ and goal $v_g \in V$. Each edge $e \in E$ has an unknown success probability $p_e \in (0, 1)$. A *route* is a simple path $\pi = (e_1, \ldots, e_k)$

from v_s to v_g, with $e_i = (v_{i-1}, v_i)$, $v_0 = v_s$, and $v_k = v_g$, and we write $|\pi| = k$. Let Π denote the nonempty set of all such routes, and let $H \geq \max_{\pi \in \Pi} |\pi|$ be an upper bound on route length; for simple paths, $H = |V| - 1$ is always valid. Similar formulations appear in (Talebi et al., 2017; Zhu & Modiano, 2018; Prakash, 2020).

Reliability and regret. In each round $t = 1, \ldots, T$, the learner selects a route $\pi_t \in \Pi$. Edge outcomes $X_{e,t} \sim \text{Bernoulli}(p_e)$ are independent across edges and rounds, and an attempt succeeds if and only if every edge on π_t succeeds. The reliability of route π is therefore $\mu(\pi) = \prod_{e \in \pi} p_e$. Let $\mu^\star = \max_{\pi \in \Pi} \mu(\pi)$ and $\Delta(\pi) = \mu^\star - \mu(\pi)$. We measure performance by the expected cumulative pseudo-regret $R(T) = \mathbb{E}\big[\sum_{t=1}^{T} \Delta(\pi_t)\big]$. The two feedback models below share this objective and differ in which edge outcomes are revealed.

Feedback models. Let $\pi_t = (e_1, \ldots, e_k)$ and let $\mathcal{O}_t$ denote the set of edges whose outcomes are observed in round t. Under *semi-bandit feedback* (Wang & Chen, 2018b; Hüyük & Tekin, 2019), $\mathcal{O}_t = \{e_1, \ldots, e_k\}$. Under *cascade feedback* (Kveton et al., 2015a; Zhong et al., 2021a), the route is traversed in order and execution stops at the first failure. Define $J_t = \min\big(\{i \in [k] : X_{e_i,t} = 0\} \cup \{k+1\}\big)$. Then $\mathcal{O}_t = \{e_i : 1 \leq i \leq \min(J_t, k)\}$. Conditional on the selected route and the preceding history, edge e_i is observed with probability $\prod_{j<i} p_{e_j}$. Under semi-bandit feedback, every selected edge is observed.

Component-wise Thompson sampling. Let $\mathcal{H}_{t-1}$ denote the history before round t, and let $S_{e,t}$ and $F_{e,t}$ count the observed successes and failures of edge e in $\mathcal{H}_{t-1}$. Component-wise TS (Thompson, 1933; Russo et al., 2018; Wang & Chen, 2018b) places an independent $\text{Beta}(\alpha_0, \beta_0)$ prior on each edge, with $\alpha_0, \beta_0 > 0$. Conditional on $\mathcal{H}_{t-1}$, it independently draws $\tilde{p}_{e,t} \sim \text{Beta}(\alpha_0 + S_{e,t},\ \beta_0 + F_{e,t})$ for every edge and selects

$$\pi_t \in \arg\max_{\pi \in \Pi} \tilde{\mu}_t(\pi), \qquad \tilde{\mu}_t(\pi) = \prod_{e \in \pi} \tilde{p}_{e,t}. \tag{1}$$

Each edge receives one sample per round, shared across all routes containing that edge. Standard TS uses the uniform prior $(\alpha_0, \beta_0) = (1, 1)$, whereas RC-TS (Section 5) uses $(1, \lambda/H)$ with $\lambda > 0$. Route selection requires one shortest-path computation with edge costs $-\log \tilde{p}_{e,t}$. Each round also includes edge sampling and posterior updates: for every $e \in \mathcal{O}_t$, the success and failure counts increase by $X_{e,t}$ and $1 - X_{e,t}$, respectively; all remaining counts retain their previous values. We call edge e *unresolved* at round t if $S_{e,t} + F_{e,t} = 0$, so its posterior equals its initial prior.

4 Compositional Discovery Barrier

TThis section characterizes the compositional discovery barrier through route-selection probabilities and derives its consequences for discovery time and regret under online posterior updates. Throughout, we consider a *target family* $\mathcal{F} \subsetneq \Pi$ of $K \geq 1$ routes. An edge is *private to* $\mathcal{F}$ if it belongs to a target route and every feasible route containing it lies in $\mathcal{F}$. Such an edge may be shared by several targets, and its outcome is observed only when a target is selected. Let $\tau_{\mathcal{F}} := \inf\{t \geq 1 : \pi_t \in \mathcal{F}\}$. Under cold-start initialization, every family-private edge retains its initial prior before discovery.

4.1 Failure of Component-Wise Thompson Sampling

Table 1 reports results on the *path-trap* instance of Figure 1(a), in which each of the K optimal routes contains m private edges each that are initially unresolved, while shorter suboptimal routes compete for selection. Results under cascade feedback appear in Appendix A. Under both feedback models, standard TS incurs regret $R(T)$ that grows with m and, in several instances, a low convergence rate, defined as the fraction of runs in which the greedy route induced by the final edge estimates is optimal (Section 6). The semi-bandit case is particularly informative. Because every edge on the selected route

Table 1: Thompson-sampling performance (Mean regret | Conv. rate) across K/m at $T = 10{,}000$. Highlighted indicate top performance.

K/m	TS	TS-Cascade	**RC-TS (Ours)**
1/5	1674.91 | 0.4	64.05 | 1.0	**11.65** | 1.0
2/6	1949.68 | 0.0	109.09 | 1.0	**12.28** | 1.0
4/6	1713.61 | 0.0	146.87 | 0.8	**11.03** | 1.0
4/8	2149.98 | 0.0	258.15 | 0.5	**10.45** | 1.0

is observed ($\mathcal{O}_t = \pi_t$), a single selection of a target route would reveal all of its private edges, yet discovery remains severely delayed. The bottleneck therefore lies in selection rather than observation: target routes rarely receive a competitive sampled score$\tilde{\mu}_t(\pi)$, and thus never acquire the observations needed for learning.

4.2 A General Route-Selection Upper Bound

We first bound the probability of selecting a given route at an arbitrary posterior state. Write $\alpha_{e,t} = \alpha_0 + S_{e,t}$ and $\beta_{e,t} = \beta_0 + F_{e,t}$ for the posterior parameters of edge e at round t. Fix a target route $\pi \in \Pi$ and a reference route $\rho \in \Pi \setminus \{\pi\}$, both chosen based on $\mathcal{H}_{t-1}$ before the round-t samples are drawn, and let $U = \pi \setminus \rho$ and $V = \rho \setminus \pi$. Because shared edges contribute the same sample to both scores, $\tilde{\mu}_t(\pi) \geq \tilde{\mu}_t(\rho)$ if and only if $\prod_{e\in U} \tilde{p}_{e,t} \geq \prod_{e\in V} \tilde{p}_{e,t}$.

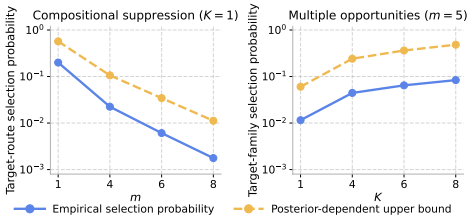


Figure 2: Empirical selection probabilities and the bounds equation 3–equation 4 at a fixed posterior state.

Proposition 1 (Posterior-dependent route-selection bound)**.** If $V \neq \emptyset$, then

$$\Pr(\pi_t = \pi \mid \mathcal{H}_{t-1}) \leq \min\left\{1,\ \inf_{0<\theta<\min_{e\in V}\alpha_{e,t}} \prod_{e\in U} \frac{\mathrm{B}(\alpha_{e,t}+\theta,\beta_{e,t})}{\mathrm{B}(\alpha_{e,t},\beta_{e,t})} \prod_{e\in V} \frac{\mathrm{B}(\alpha_{e,t}-\theta,\beta_{e,t})}{\mathrm{B}(\alpha_{e,t},\beta_{e,t})}\right\}. \tag{2}$$

Proof sketch. If TS selects π, then in particular π must outperform ρ. Let $X = \prod_{e\in U} \tilde{p}_{e,t}$ and $Y = \prod_{e\in V} \tilde{p}_{e,t}$. For any admissible $\theta > 0$, Markov's inequality gives

$$\Pr(\pi_t = \pi \mid \mathcal{H}_{t-1}) \leq \Pr(X \geq Y \mid \mathcal{H}_{t-1}) \leq \mathbb{E}\big[(X/Y)^\theta \mid \mathcal{H}_{t-1}\big].$$

Conditional independence of the edge samples factorizes the expectation, and the positive and negative moments of Beta random variables give equation 2. (See Appendix B) □

Proposition 1 separates target-side from reference-side uncertainty. Suppose U contains a set U_0 of m edges that are unresolved at round t. Each such edge retains its prior and contributes $\mathbb{E}[\tilde{p}^{\theta}_{e,t} \mid \mathcal{H}_{t-1}] = \mathrm{B}(\alpha_0+\theta,\beta_0)/\mathrm{B}(\alpha_0,\beta_0)$, which equals $(1+\theta)^{-1}$ under the uniform prior of standard TS. Since every other positive-moment factor is at most one, standard TS satisfies

$$\Pr(\pi_t = \pi \mid \mathcal{H}_{t-1}) \leq \min\left\{1,\ \inf_{0<\theta<\min_{e\in V}\alpha_{e,t}} (1+\theta)^{-m} \prod_{e\in V} \frac{\mathrm{B}(\alpha_{e,t}-\theta,\beta_{e,t})}{\mathrm{B}(\alpha_{e,t},\beta_{e,t})}\right\}. \tag{3}$$

The factor $(1+\theta)^{-m}$ captures the suppression induced by multiplying unresolved samples, while the remaining product accounts for the reference route's posterior uncertainty; the optimization over θ balances the two. For a fixed reference posterior, the bound is nonincreasing in m. This suppression depends on the prior as well as on composition. Under the prior $\mathrm{Beta}(1, b)$, each unresolved edge instead contributes $\Gamma(1+\theta)\Gamma(1+b)/\Gamma(1+\theta+b)$, which tends to one as $b \to 0$; Section 5 exploits this by choosing b inversely proportional to route length.

The bound extends directly to the target family $\mathcal{F}$. Let $\bar{s}_t(\pi)$ denote the right-hand side of equation 2 for $\pi \in \mathcal{F}$ with a chosen reference route, and let $\bar{s}_t = \max_{\pi\in\mathcal{F}} \bar{s}_t(\pi)$. Since the selection events are mutually exclusive,

$$\Pr(\pi_t \in \mathcal{F} \mid \mathcal{H}_{t-1}) \leq \min\Big\{1, \sum_{\pi\in\mathcal{F}} \bar{s}_t(\pi)\Big\} \leq \min\{1, K\bar{s}_t\}. \tag{4}$$

Unresolved edges thus control the bound for each target, while the family size K enters through aggregation across targets.

Figure 2 compares empirical one-round selection probabilities with these bounds at a fixed posterior state. Each target route has m unresolved edges with $\mathrm{Beta}(1, 1)$ posteriors, and the reference route

has one exclusive edge with a $\mathrm{Beta}(8,4)$ posterior. We estimate selection probabilities by repeatedly sampling from these fixed posteriors and selecting the route with the largest sampled score. For a single target ($K = 1$), the selection probability decreases sharply as m increases. For a family of K targets ($m = 5$), it increases with K, reflecting more opportunities to select a target. The bounds capture both trends and remain above the empirical estimates throughout the evaluated range.

4.3 Discovery-Time and Regret Lower Bounds

We now relate one-round selection probabilities to discovery under online posterior updates. Suppose each of the K targets contains at least m initially unobserved family-private edges. Under uniform priors, these edges remain unresolved until $\tau_{\mathcal{F}}$. For $q \in (0,1)$, define $M_t := \max_{\rho \in \Pi \setminus \mathcal{F}} \tilde{\mu}_t(\rho)$ and $\eta_t(q) := \Pr(M_t < q \mid \tau_{\mathcal{F}} \geq t)$. The uniform-product tail and a union bound yield

$$\Pr(\tau_{\mathcal{F}} = t \mid \tau_{\mathcal{F}} \geq t) \leq K \frac{[\log(1/q)]^m}{m!} + \eta_t(q). \tag{5}$$

A suitable uniform bound on $\eta_t(q)$ gives a conditional discovery-time lower bound (Appendix C). The following construction handles competitor learning explicitly from cold start.

Definition 1 (Overlapping path-trap)**.** Fix $0 < p < \nu < 1$, $m \geq 2$, and $K \geq 1$. The feasible routes are exactly $\rho = (e_0) \circ C$ and $\pi_k = P_k \circ C$, $k = 1, \ldots, K$. Here e_0 connects v_s to a junction u with reliability p; each P_k is an m-edge path from v_s to u with reliability ν; and C is a common suffix from u to v_g. The paths P_k may overlap with one another and are edge-disjoint from $\{e_0\}$ and C. Writing $w = \prod_{e \in C} p_e$ (with $w = 1$ for an empty suffix), the target family has reliability $\mu_{\mathcal{F}} = w\nu$ and gap $\Delta_{\mathcal{F}} = w(\nu - p)$.

The shared suffix cancels in sampled comparisons:

$$\tilde{\mu}_t(\pi_k) \geq \tilde{\mu}_t(\rho) \quad \Longleftrightarrow \quad \prod_{e \in P_k} \tilde{p}_{e,t} \geq \tilde{p}_{e_0,t}. \tag{6}$$

Before discovery, e_0 is observed every round under both feedback models, while target-prefix edges retain their priors.

Theorem 1 (Cold-start discovery delay and regret of TS)**.** On an overlapping path-trap, initialize every edge with an independent $\mathrm{Beta}(1,1)$ prior. There exists $m_0(p,K) = O(1+\log K+\log(1/p))$ such that, for $m \geq m_0(p,K)$, standard TS under either selected-route semi-bandit or cascade feedback satisfies

$$\Pr(\tau_{\mathcal{F}} > T) \geq \left(\frac{3}{4} - T \frac{K[\log(2/p)]^m}{m!} \right)_+ \tag{7}$$

for every integer $T \geq 0$, where $(x)_+ = \max\{x, 0\}$. Consequently,

$$\mathbb{E}[\tau_{\mathcal{F}}] \geq \frac{m!}{8K[\log(2/p)]^m}, \qquad R(T) \geq \frac{\Delta_{\mathcal{F}} T}{2} \quad \text{for} \quad T \leq \frac{m!}{4K[\log(2/p)]^m}. \tag{8}$$

Proof sketch. Couple TS to an auxiliary process that observes e_0 every round and resamples each target-prefix edge from its prior, sharing samples across overlapping targets. By equation 6, the processes agree until discovery. For $q = p/2$, early discovery and any subsequent competitor sample below q each have probability at most $1/8$, using a burn-in depending only on p. Each remaining round contributes at most $K[\log(2/p)]^m/m!$ by the uniform-product tail and a union bound, valid for correlated target scores. This gives equation 7; tail summation and the pre-discovery gap give equation 8. Appendix C supplies the details. □

For fixed p, K, w, ν, the theorem establishes factorial-scale discovery delay at a fixed positive gap, despite shared-edge learning. Section 5 gives a discovery-time upper bound uniform in m for calibrated TS on the same family.

5 Route-Calibrated Thompson Sampling

We introduce *Route-Calibrated Thompson Sampling* (RC-TS), which uses the prior $(\alpha_0, \beta_0) = (1, \lambda/H)$ in component-wise TS equation 1. Algorithm 1 summarizes the procedure. Each edge

Algorithm 1 Route-Calibrated Thompson Sampling (RC-TS)

Require: Route set Π, length bound $H \geq \max_{\pi\in\Pi} |\pi|$, calibration $\lambda > 0$, horizon T
1: $S_e \leftarrow 0$, $F_e \leftarrow 0$ for every $e \in E$
2: **for** $t = 1, \ldots, T$ **do**
3: Sample $\tilde{p}_{e,t} \sim \mathrm{Beta}(1 + S_e,\ \lambda/H + F_e)$ independently for every $e \in E$
4: Select $\pi_t \in \arg\max_{\pi\in\Pi} \prod_{e\in\pi} \tilde{p}_{e,t}$ {shortest path with costs $-\log \tilde{p}_{e,t}$}
5: Execute π_t and observe $X_{e,t}$ for $e \in \mathcal{O}_t$ {semi-bandit or cascade}
6: **for** $e \in \mathcal{O}_t$ **do**
7: $S_e \leftarrow S_e + X_{e,t}, \quad F_e \leftarrow F_e + 1 - X_{e,t}$
8: **end for**
9: **end for**

receives one sample per round, shared across all routes containing it. Route selection uses shortest-path computation with costs $-\log \tilde{p}_{e,t}$, together with edge sampling and observed-outcome updates.

Calibration principle. For an initially unobserved route of length L with independent $\mathrm{Beta}(1, b)$ edge priors, the mean sampled score is $(1 + b)^{-L}$. Thus the uniform prior induces mean score 2^{-L}. Maintaining a reference mean $\zeta \in (0, 1)$ gives

$$(1 + b)^{-L} = \zeta \quad \Longleftrightarrow \quad b = \zeta^{-1/L} - 1 = \frac{-\log \zeta}{L} + O(L^{-2}), \tag{9}$$

identifying inverse-length scaling. To use a common prior across overlapping routes, we set $b = \lambda/H$, where H is the route-length bound from Section 3. Every initially unobserved route of length $\ell \leq H$ then satisfies

$$\mathbb{E}[\tilde{\mu}_1(\pi)] = \left(1 + \frac{\lambda}{H}\right)^{-\ell} \geq e^{-\lambda\ell/H} \geq e^{-\lambda}. \tag{10}$$

Smaller λ increases prior optimism. The reference choice $\lambda = 1$ recovers $b = 1/L$ when $H = L$.

The following lemma strengthens mean calibration into a direct guarantee on competitive samples.

Lemma 1 (Uniform competitive-tail guarantee)**.** Let $U_1, \ldots, U_L$ be independent $\mathrm{Beta}(1, b)$ variables with $Lb \leq \lambda$. For every $q \in (0, 1)$,

$$\Pr\left(\prod_{i=1}^{L} U_i \geq q\right) \geq \eta_\lambda(q) := \frac{1}{2}\left[\min\left\{1, \frac{1-q}{2\lambda}\right\}\right]^{\lambda}. \tag{11}$$

Proof sketch. Write $W_i = 1 - U_i$ and $\varepsilon = \min\{1, (1 - q)/(2\lambda)\}$. The event $A = \{\max_i W_i \leq \varepsilon\}$ has probability $\varepsilon^{Lb} \geq \varepsilon^{\lambda}$. Moreover, $\mathbb{E}[\sum_i W_i \mid A] = Lb\varepsilon/(1 + b) \leq (1 - q)/2$. Markov's inequality gives conditional probability at least $1/2$ that $\sum_i W_i \leq 1 - q$. On this event, $\prod_i U_i \geq 1 - \sum_i W_i \geq q$. □

For $b = \lambda/H$ and $L \leq H$, the lower bound depends only on q and λ. It applies to any set of edges retaining their initial priors, including the private prefixes in Definition 1. For comparison, fixed $b > 0$ yields

$$\Pr\left(\prod_{i=1}^{L} U_i \geq q\right) \leq \frac{\Gamma(1 + b)^L [\log(1/q)]^{bL}}{\Gamma(1 + bL)},$$

whose logarithm is at most $-bL \log L + O_{b,q}(L)$. Appendix C.4 gives the derivation. We now combine the competitive-tail guarantee with online learning of the competing edge.

Theorem 2 (Length-uniform discovery time of RC-TS)**.** On an overlapping path-trap from Definition 1, initialize every edge with an independent $\mathrm{Beta}(1, \lambda/H)$ prior, where $H \geq \max_{\pi\in\Pi} |\pi|$ and $\lambda > 0$. Set

$$q_p := \frac{1 + p}{2}, \qquad c_p := \frac{(1 - p)^2}{8}, \qquad C_p := \frac{4e^{-c_p}}{(1 - e^{-c_p})^2}.$$

Under either selected-route semi-bandit or cascade feedback,

$$\mathbb{E}[\tau_{\mathcal{F}}] \leq \frac{2}{\eta_\lambda(q_p)} + C_p = O_{p,\lambda}(1). \tag{12}$$

The bound is uniform in m, K, and target-prefix overlap. Expected pseudo-regret accumulated strictly before discovery equals $\Delta_{\mathcal{F}}(\mathbb{E}[\tau_{\mathcal{F}}] - 1)$.

Proof sketch. Cancel the common suffix and couple RC-TS to an auxiliary process that observes e_0 every round and resamples each target-prefix edge from its prior. Samples are shared across targets containing the same edge. The processes agree until discovery.

Fix one target prefix. By Lemma 1, its product exceeds q_p with probability at least $\eta_\lambda(q_p)$, independently across auxiliary rounds. After $n \geq 1$ observations, the competitor sample Y_n satisfies $\Pr(Y_n > q_p) \leq 2e^{-c_p n}$, where probability is averaged over the auxiliary outcome history. For $T \geq 2$ and $j = \lfloor T/2 \rfloor$, a union bound gives

$$\Pr(\tau_{\mathcal{F}} > T) \leq (1 - \eta_\lambda(q_p))^j + \frac{2e^{-c_p j}}{1 - e^{-c_p}}.$$

Summing this tail gives equation 12. The argument uses a single target, accommodating arbitrary overlap among target prefixes. Appendix C.4 provides the full proof. □

With $\lambda = 1$, Theorems 1 and 2 yield, for $m \geq m_0(p, K)$,

$$\mathbb{E}_{\mathrm{TS}}[\tau_{\mathcal{F}}] \geq \frac{m!}{8K[\log(2/p)]^m}, \qquad \mathbb{E}_{\text{RC-TS}}[\tau_{\mathcal{F}}] \leq \frac{16}{1-p} + C_p. \tag{13}$$

For fixed p and K, this establishes a factorial-scale lower bound for standard TS and an upper bound uniform in m for RC-TS. The separation concerns first discovery on the specified overlapping-route family.

6 Experiments

We evaluate the discovery behavior predicted by Theorems 1 and 2, then compare performance on general routing benchmarks. Additional experiments on gap sensitivity, graph-size scaling, calibration, and further domains appear in Appendix A.

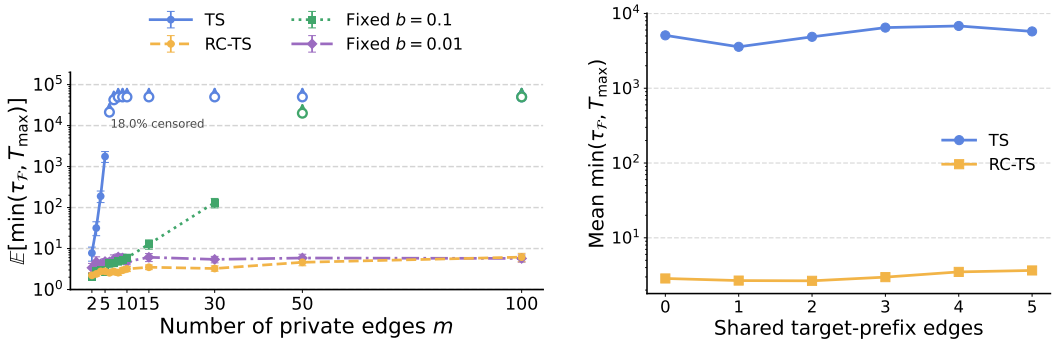
Setup. We study parallel path-traps, the empty-suffix, edge-disjoint special case of Definition 1, and target-prefix overlap in the overlapping-route construction. General benchmarks include Erdős–Rényi graphs (Erdös & Rényi, 1959), grids, layered DAGs, and reduced quantum networks (Wehner et al., 2018; Chakraborty et al., 2020). Methods share instances, reliabilities, and interaction budgets; RC-TS uses $H = |V| - 1$ and $\lambda = 1$. We compare with uniform-prior TS, shared-seed CTS^3 (Jin et al., 2026), TS-Cascade (Zhong et al., 2021b), CombCascade (Kveton et al., 2015b), BCUCB-T (Liu et al., 2024), and a random-routing baseline. Feedback-specific implementations and tuning are detailed in Appendix D.

Metrics. We report cumulative pseudo-regret and the fraction of runs whose final empirical greedy route is optimal, assigning estimate 0 to unobserved edges. Discovery plots report the sample mean of $\min(\tau_{\mathcal{F}}, T_{\max})$, the discovery time truncated at the run horizon. Censored runs contribute $T_{\max}$; the corresponding population quantity is a lower bound on $\mathbb{E}[\tau_{\mathcal{F}}]$.

Discovery and prior calibration. Figure 3a extends the private-edge sweep to $m = 100$. Standard TS rapidly reaches the truncation regime, while RC-TS remains within roughly 3–6 rounds throughout this range, consistent with the length-uniform upper bound. The fixed prior $b = 0.1$ initially performs well but incurs much longer delays as m grows. The more optimistic fixed prior $b = 0.01$ remains competitive over the tested range, reaching about 6 rounds at $m = 100$. The theory distinguishes these finite-range results from the asymptotic suppression for any fixed $b > 0$: inverse-length calibration preserves a uniform competitive-tail guarantee.

Discovery with overlapping targets. Figure 3b varies shared target-prefix edges from 0 to 5. TS has mean truncated discovery times of several thousand rounds, while RC-TS stays near 3 rounds across the sweep. This supports effective discovery with correlated target scores, in line with Theorem 2.

General routing performance. Table 2 reports results under both feedback models on Erdős–Rényi and grid graphs. CTS^3 attains the lowest mean regret in both semi-bandit settings, while RC-TS attains the lowest mean regret and full observed convergence in both cascade settings. On Grid



(a) Private-edge scaling.

(b) Target-prefix overlap.

Figure 3: Discovery under increasing private-route length and target overlap. Both panels show mean truncated first-discovery time on logarithmic vertical axes. (a) Parallel path-traps with $p = 0.55$, $\nu = 0.70$, and $K = 1$ under semi-bandit feedback over 500 runs, with $T_{\max} = 5 \times 10^4$. Open markers indicate settings containing censored runs. Fixed-prior baselines use $\mathrm{Beta}(1, b)$. (b) TS and RC-TS as the number of shared target-prefix edges varies. Construction and run settings are detailed in Appendix A.

Table 2: General routing performance at $T = 10{,}000$ over 50 seeds. Entries show mean $\pm$ standard deviation of cumulative pseudo-regret | convergence rate. Highlighting identifies the lowest mean regret in each setting.

Method	**Erdős–Rényi**		**Grid**	
	Semi-bandit	**Cascade**	**Semi-bandit**	**Cascade**
TS	64.2 ± 18.5 | 1.000	81.5 ± 21.0 | 1.000	22.4 ± 5.2 | 1.000	32.5 ± 9.6 | .800
RC-TS ($\lambda = 1.0$)	48.3 ± 11.2 | 1.000	**58.1 ± 14.3 | 1.000**	24.8 ± 10.8 | .800	**29.1 ± 1.8 | 1.000**
CTS^3	**39.7 ± 7.0 | 1.000**	76.2 ± 21.1 | 1.000	**6.4 ± 1.9 | 1.000**	57.3 ± 36.3 | .700
TS-Cascade	395.4 ± 25.8 | 1.000	412.6 ± 22.1 | 1.000	128.1 ± 4.8 | 1.000	139.4 ± 4.1 | 1.000
BCUCB-T	612.8 ± 41.3 | 1.000	625.4 ± 43.2 | 1.000	251.6 ± 6.5 | .800	263.7 ± 7.2 | .800
CombCascade	879.5 ± 33.1 | 1.000	891.0 ± 29.8 | 1.000	245.2 ± 8.1 | 1.000	236.9 ± 6.9 | 1.000
Random	1948.1 ± 55.2 | .000	1929.8 ± 64.1 | .000	612.4 ± 20.3 | .000	558.2 ± 15.2 | .000

under cascade feedback, RC-TS achieves 29.1 ± 1.8 regret and convergence rate 1.0, compared with 57.3 ± 36.3 and 0.7 for CTS^3. On Grid under semi-bandit feedback, CTS^3 leads with 6.4 ± 1.9 regret, compared with 24.8 ± 10.8 for RC-TS. These results illustrate feedback-dependent performance of prior calibration and coordinated sampling. Full benchmark results and all remaining ablations appear in Appendix A.

7 Discussion and Conclusion

We characterized the compositional discovery barrier: multiplicative posterior sampling suppresses the selections needed to resolve family-private uncertainty. On a controlled overlapping path-trap family, standard TS exhibits factorial-scale discovery delay and linear regret over factorially long horizons at a fixed positive gap. RC-TS calibrates independent Beta priors to a route-length bound, preserving competitive tail probabilities and yielding a first-discovery upper bound uniform in private-route length for fixed p and λ. Experiments support effective discovery and competitive routing performance: CTS^3 leads the reported semi-bandit benchmarks, while RC-TS leads their cascade counterparts.

Our guarantees cover shared suffixes and overlapping target prefixes with a single competing route. Multiple adaptive competitors and full cumulative-regret guarantees on general graphs remain important directions. Prior optimism has instance-dependent costs, and calibration depends on the ratio between route length and H; adapting it to observed structure may improve this trade-off. The competitive-tail analysis also motivates length-calibrated priors for other multiplicative decision problems, including multi-stage pipelines and conjunctive recommendation.

AI USE STATEMENT

In this work, we used generative AI tools for debugging code and for checking mathematical derivations for potential errors (e.g., typos, sign mistakes, inconsistent notation, or missing conditions). We have not used generative AI tools for generating the core research ideas, formulating the proposed method, or carrying out the theoretical derivations and proofs; all of these were conceived and developed entirely by the authors. Additionally, we used generative AI tools for improving the layout and visual presentation of figures. We have reviewed all AI-assisted work. Specifically, all code fixes suggested by AI tools were manually inspected, and the affected experiments were re-run and verified by the authors; every potential issue flagged by AI tools in the formulas was independently re-derived and checked by the authors before any change was made, and suggestions judged incorrect were discarded; and AI-assisted figure adjustments were limited to formatting and presentation, with all plotted data and results verified against our original experimental outputs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not raise any concerns with respect to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Detailed experimental settings, model architectures, training procedures, and computing resources, are provided in Appendix D. Complete proofs of all theoretical results, along with the necessary assumptions, are also given in Appendix.

REFERENCES

Kaushik Chakraborty, David Elkouss, Bruno Rijsman, and Stephanie Wehner. Entanglement distribution in a quantum network: A multicommodity flow-based approach. *IEEE Transactions on Quantum Engineering*, 1:1–21, 2020.

Phoebus Chen, Karl H. Johansson, Paul Balister, Bela Bollobas, and Shankar Sastry. Multi-path routing metrics for reliable wireless mesh routing topologies. *arXiv preprint arXiv:1103.5957*, 2011.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50): 1–33, 2016. URL https://arxiv.org/abs/1407.8339. arXiv:1407.8339.

P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

Alihan Hüyük and Cem Tekin. Analysis of thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 1322–1330, 2019.

Alihan Hüyük and Cem Tekin. Thompson sampling for combinatorial network optimization in unknown environments. *IEEE/ACM Transactions on Networking*, 28(6):2836–2849, 2020. doi: 10.1109/TNET.2020.3025904. arXiv:1907.04201.

Tianyuan Jin, Heyang Zhao, Vincent Y. F. Tan, and Quanquan Gu. Avoiding $\exp(k^*)$ scaling for thompson sampling in combinatorial semi-bandits: From multiple seeds to a single seed. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, pp. 3825–3855, 2026.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 767–776, 2015a.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Combinatorial cascading bandits. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1450–1458, 2015b. URL https://arxiv.org/abs/1507.04208. arXiv:1507.04208.

Yi-Kuei Lin and Cheng-Ta Yeh. Optimal carrier selection based on network reliability criterion for stochastic logistics networks. *International Journal of Production Economics*, 128(2):510–517, 2010. doi: 10.1016/j.ijpe.2010.07.001.

Che-Yu Liu and Lihong Li. On the prior sensitivity of thompson sampling, 2016. URL https://arxiv.org/abs/1506.03378.

Xutong Liu, Jinhang Zuo, Siwei Wang, Carlee Joe-Wong, John C. S. Lui, and Wei Chen. Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms, 2024. URL https://arxiv.org/abs/2208.14837.

Benedict C. May, Nathan Korda, Anthony Lee, and David S. Leslie. Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, 13:2069–2106, 2012. URL https://jmlr.org/papers/v13/may12a.html.

Pierre Perrault. When combinatorial thompson sampling meets approximation regret, 2023. URL https://arxiv.org/abs/2302.11182.

Pierre Perrault, Etienne Boursier, Vianney Perchet, and Michal Valko. Statistical efficiency of thompson sampling for combinatorial semi-bandits, 2021. URL https://arxiv.org/abs/2006.06613.

A. Arun Prakash. Algorithms for most reliable routes on stochastic and time-dependent networks. *Transportation Research Part B: Methodological*, 138:202–220, 2020. doi: 10.1016/j.trb.2020.05.013.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018. doi: 10.1561/2200000070.

Sadegh Talebi, Zhenhua Zou, Richard Combes, Alexandre Proutiere, and Mikael Johansson. Stochastic online shortest path routing: The value of feedback. *IEEE Transactions on Automatic Control*, 2017.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1703.01610. arXiv:1703.01610.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5114–5122. PMLR, 2018a. URL https://proceedings.mlr.press/v80/wang18a.html.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5114–5122, 2018b.

Finnegan Waugh and Ralph Holz. An empirical study of availability and reliability properties of the bitcoin lightning network. *arXiv preprint arXiv:2006.14358*, 2020.

Stephanie Wehner, David Elkouss, and Ronald Hanson. Quantum internet: A vision for the road ahead. *Science*, 362:eaam9288, 10 2018. doi: 10.1126/science.aam9288.

Raymond Zhang and Richard Combes. On the suboptimality of thompson sampling in high dimensions. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2102.05502. arXiv:2102.05502.

Raymond Zhang and Richard Combes. Thompson sampling for combinatorial bandits: Polynomial regret and mismatched sampling paradox, 2024. URL https://arxiv.org/abs/2410.05441.

Zixin Zhong, Wang Chi Cheung, and Vincent Y. F. Tan. Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research*, 22(218):1–66, 2021a.

Zixin Zhong, Wang Chi Cheung, and Vincent Y. F. Tan. Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research*, 22(218):1–66, 2021b. URL http://jmlr.org/papers/v22/20-447.html. arXiv:1810.01187.

Ruihao Zhu and Eytan Modiano. Learning to route efficiently with end-to-end feedback: The value of networked structure. *arXiv preprint arXiv:1810.10637*, 2018.

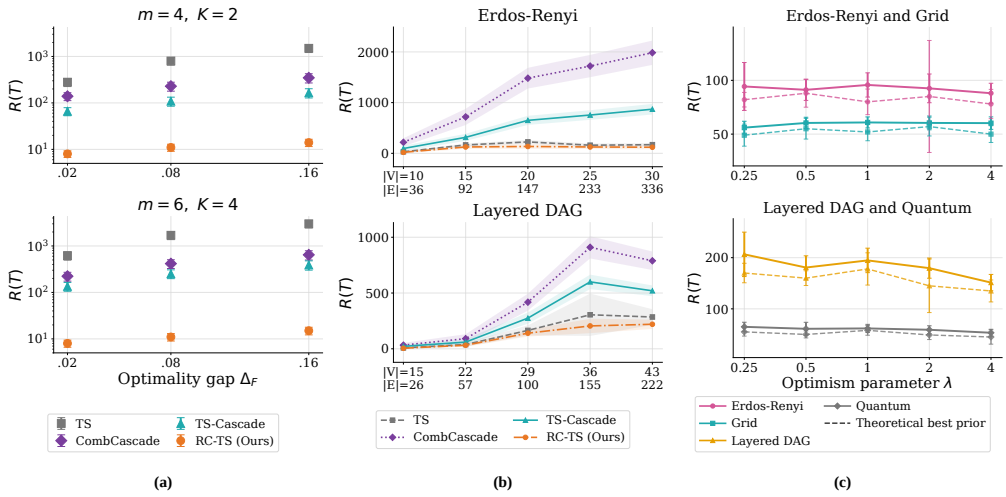


Figure 4: Final regret $R(T)$ under cascade feedback at $T = 10{,}000$ over 20 seeds (mean $\pm$ one standard deviation; lower is better). (a) Shared-edge path-traps as the gap $\Delta_{\mathcal{F}}$ varies (log scale). (b) Graph size; axis labels give $|V|$ and $|E|$. (c) Sensitivity to λ; solid curves use H and dashed curves use the optimal route length L.

A Additional Experimental Results

Table 3: Benign-domain benchmark results (Layered DAG & Quantum) under both feedbacks at $T = 10{,}000$ over 20 seeds. Each entry reports final cumulative regret $R(T)$ as mean $\pm$ standard deviation | convergence rate. Highlighted values indicate the lowest mean regret in each setting.

Method	Layered DAG		Quantum	
	Semi-bandit	Cascade	Semi-bandit	Cascade
TS	154.8 ± 29.7 | 1.000	182.3 ± 31.5 | 1.000	34.8 ± 5.8 | 1.000	42.6 ± 6.5 | 1.000
RC-TS ($\lambda = 1.0$)	118.2 ± 21.5 | 1.000	**131.7 ± 22.9 | 1.000**	39.1 ± 7.5 | 1.000	**41.6 ± 7.9 | 1.000**
CTS[3]	**78.7 ± 23.9 | 1.000**	165.7 ± 44.3 | .900	**9.8 ± 1.8 | 1.000**	59.2 ± 42.2 | 1.000
TS-Cascade	593.4 ± 30.4 | 1.000	627.1 ± 34.8 | 1.000	173.2 ± 6.8 | 1.000	189.5 ± 8.3 | 1.000
BCUCB-T	1121.8 ± 25.7 | 1.000	1158.4 ± 29.1 | 1.000	372.6 ± 12.9 | 1.000	394.1 ± 15.0 | 1.000
CombCascade	900.5 ± 31.1 | 1.000	914.2 ± 30.6 | 1.000	366.5 ± 18.0 | 1.000	375.8 ± 18.7 | 1.000
Random	3409.0 ± 33.9 | .000	3357.0 ± 29.8 | .000	425.7 ± 20.9 | .000	438.7 ± 19.2 | .000

Table 4: Thompson-sampling performance (Mean regret | Conv. rate) across K/m at $T = 10{,}000$ under **cascade feedback**. Highlighted cells indicate top performance.

K/m	TS	TS-Cascade	RC-TS (Ours)
1/5	1820.45 | 0.2	82.40 | 0.9	**12.10 | 1.0**
2/6	2105.30 | 0.0	145.60 | 0.8	**13.05 | 1.0**
4/6	1985.12 | 0.0	192.35 | 0.6	**11.82 | 1.0**
4/8	2350.80 | 0.0	310.20 | 0.4	**11.15 | 1.0**

Path-trap robustness. Figure 4(a) varies the gap $\Delta_{\mathcal{F}}$ on shared-edge path-traps with $(m, K) = (4, 2)$ and $(6, 4)$. The plotted regret of TS grows with the gap parameter and reaches about 10^3. With multiple competing routes, expected regret accumulated strictly before discovery by horizon T is $\sum_{\rho \in \Pi \setminus \mathcal{F}} \Delta(\rho) \mathbb{E}[N_\rho^{\mathrm{pre}}(T)]$, where $N_\rho^{\mathrm{pre}}(T) = \sum_{t=1}^{T} \mathbf{1}\{t < \tau_{\mathcal{F}}, \pi_t = \rho\}$. Each competing route thus contributes its own gap and selection count. TS-Cascade and CombCascade reduce it to the order of 10^2, while RC-TS stays near 10 throughout. This instance includes two competing routes and extends the empirical evaluation beyond the single-competitor family in our theorems.

Scaling with graph size. Figure 4(b) varies the graph size. RC-TS matches or improves upon TS at every size, with a clearer advantage on larger layered DAGs, while the regret of TS-Cascade and CombCascade grows steeply with size.

Sensitivity to calibration. Figure 4(c) varies λ under cascade feedback. Solid curves use the practical prior $\mathrm{Beta}(1, \lambda/H)$, and dashed curves use the reference prior $\mathrm{Beta}(1, \lambda/L)$, where L is the optimal route length; the latter recovers $b = 1/L$ at $\lambda = 1$. Regret is nearly flat in λ on Erdős–Rényi, Grid, and Quantum, while layered DAGs favor larger λ. The practical prior tracks the reference with slightly higher regret in these experiments. The reference uses the optimal route length as oracle information, illustrating the empirical effect of the calibration scale.

A.1 Discovery Experiment Protocols

Private-edge scaling. The parallel construction is the empty-suffix, edge-disjoint special case of Definition 1. For the experiment in Figure 3a, the competing edge has reliability $p = 0.55$ and the single target has reliability $\nu = 0.70$. Assigning each of its m edges probability $\nu^{1/m}$ keeps target reliability fixed. The experiment uses semi-bandit feedback, 500 runs, and $T_{\max} = 5 \times 10^4$, with the sweep extending to $m = 100$. The fixed-prior baselines use $\mathrm{Beta}(1, b)$ for $b \in \{0.1, 0.01\}$ and the same Beta–Bernoulli updates as TS. RC-TS uses $\lambda = 1$ and $H = |V| - 1$; for this single-target parallel graph, $H = m$.

Target overlap. Figure 3b reports discovery as the number of shared target-prefix edges varies from 0 to 5. The relevant structure is described in Definition 1: target prefixes may share edges, and all routes share a suffix. The number of family-private edges on each target is m, including edges shared with other targets. Since the practical calibration uses $H = |V| - 1$, reporting the vertex count at each overlap level specifies its effect on the prior.

Truncation and run counts. Discovery plots use the empirical mean of $\min(\tau_{\mathcal{F}}, T_{\max})$. Runs censored at the horizon contribute $T_{\max}$, and the population truncated expectation is bounded above by $\mathbb{E}[\tau_{\mathcal{F}}]$. Open markers in Figure 3a identify settings containing censored runs. Run counts are experiment-specific: Table 2 reports 50 seeds, while Table 3 and Figure 4 report 20 seeds.

B Proof of Proposition 1

Proof. Condition on $\mathcal{H}_{t-1}$, so the routes π, ρ and all posterior parameters are fixed. The edge samples are conditionally independent, with $\tilde{p}_{e,t} \sim \mathrm{Beta}(\alpha_{e,t}, \beta_{e,t})$.

Since TS maximizes the sampled route score, selecting π implies $\tilde{\mu}_t(\pi) \geq \tilde{\mu}_t(\rho)$. Shared edges contribute identical, strictly positive samples to both scores and cancel. Thus, writing $U = \pi \setminus \rho$, $V = \rho \setminus \pi$, $X = \prod_{e \in U} \tilde{p}_{e,t}$, and $Y = \prod_{e \in V} \tilde{p}_{e,t}$, we obtain

$$\Pr(\pi_t = \pi \mid \mathcal{H}_{t-1}) \leq \Pr(X \geq Y \mid \mathcal{H}_{t-1}).$$

Empty products are interpreted as one.

For any $0 < \theta < \min_{e \in V} \alpha_{e,t}$, Markov's inequality gives

$$\Pr(X \geq Y \mid \mathcal{H}_{t-1}) \leq \mathbb{E}\left[(X/Y)^{\theta} \mid \mathcal{H}_{t-1}\right].$$

Because U and V are disjoint, conditional independence factorizes this expectation as

$$\mathbb{E}\left[(X/Y)^{\theta} \mid \mathcal{H}_{t-1}\right] = \prod_{e \in U} \mathbb{E}[\tilde{p}_{e,t}^{\theta} \mid \mathcal{H}_{t-1}] \prod_{e \in V} \mathbb{E}[\tilde{p}_{e,t}^{-\theta} \mid \mathcal{H}_{t-1}].$$

For $W \sim \mathrm{Beta}(\alpha, \beta)$ and $s > -\alpha$, direct integration yields

$$\mathbb{E}[W^s] = \frac{1}{B(\alpha, \beta)} \int_0^1 w^{\alpha+s-1}(1-w)^{\beta-1}\, dw = \frac{B(\alpha+s, \beta)}{B(\alpha, \beta)}.$$

The admissible range of θ ensures that all negative moments above are finite. Substituting these moments gives

$$\Pr(\pi_t = \pi \mid \mathcal{H}_{t-1}) \leq \prod_{e \in U} \frac{B(\alpha_{e,t} + \theta, \beta_{e,t})}{B(\alpha_{e,t}, \beta_{e,t})} \prod_{e \in V} \frac{B(\alpha_{e,t} - \theta, \beta_{e,t})}{B(\alpha_{e,t}, \beta_{e,t})}.$$

Taking the infimum over admissible θ and combining with the probability bound of one proves the proposition. □

C Discovery-Time and Regret Analysis

C.1 General Conditional Discovery Bound

Write $h_t := \Pr(\tau_{\mathcal{F}} = t \mid \tau_{\mathcal{F}} \geq t)$ and $d_m(q) := [\log(1/q)]^m/m!$. We first establish Equation equation 5 and its conditional discovery-time consequence.

Uniform-product tail. Let $U_1, \ldots, U_m$ be independent uniform variables on $(0, 1)$. Their negative log-product satisfies

$$-\sum_{i=1}^{m} \log U_i \sim \mathrm{Gamma}(m, 1).$$

With $L_q := \log(1/q)$,

$$\Pr\left(\prod_{i=1}^{m} U_i \geq q\right) = \frac{1}{(m-1)!}\int_0^{L_q} x^{m-1}e^{-x}\,dx \leq \frac{L_q^m}{m!} = d_m(q).$$

Online selection probability. Before discovery, every target contains m edges whose posteriors remain independent $\mathrm{Beta}(1, 1)$ distributions. All other sampled edge probabilities are at most one, so a target's sampled route score is bounded above by the product of these m samples. Whenever $M_t \geq q$, selecting a target requires at least one target score to reach q. A union bound, first conditional on the observed history and then averaged over histories satisfying $\tau_{\mathcal{F}} \geq t$, gives $h_t \leq Kd_m(q) + \eta_t(q)$. This establishes Equation equation 5.

Corollary 1 (Conditional factorial-scale discovery delay)**.** Suppose that, for fixed $q \in (0, 1)$ and $c \geq 0$, $\eta_t(q) \leq cKd_m(q)$ for every t. If $a := (1 + c)Kd_m(q) < 1$, then

$$\Pr(\tau_{\mathcal{F}} > T) \geq (1 - a)^T, \qquad \mathbb{E}[\tau_{\mathcal{F}}] \geq \frac{1}{a} = \frac{m!}{(1 + c)K[\log(1/q)]^m}.$$

Proof. The assumed condition gives $h_t \leq a$. Iterating conditional survival probabilities yields $\Pr(\tau_{\mathcal{F}} > T) = \prod_{t=1}^{T}(1 - h_t) \geq (1 - a)^T$. The tail-sum identity then gives $\mathbb{E}[\tau_{\mathcal{F}}] \geq \sum_{T=0}^{\infty}(1 - a)^T = 1/a$. □

The corollary states a sufficient condition on survival-conditioned competitor scores. The next proof establishes the cold-start result directly through an auxiliary process and cumulative event bounds.

C.2 Shared-Edge Auxiliary Process

Consider Definition 1 and a common edge prior $\mathrm{Beta}(1, b)$, with $b > 0$. Write $E_{\mathcal{F}} := \bigcup_{k=1}^{K} P_k$, $\mu_{\mathcal{F}} = w\nu$, and $\Delta_{\mathcal{F}} = w(\nu - p)$. The common suffix contributes the same strictly positive sampled factor to every route. Cancelling it reduces selection to comparing the target-prefix products with the sample of e_0.

Let $B_1, B_2, \ldots$ be independent $\mathrm{Bernoulli}(p)$ outcomes and let $S_n = \sum_{j=1}^{n} B_j$, with $S_0 = 0$. In auxiliary round $n + 1$, draw

$$Y_n \mid S_n \sim \mathrm{Beta}(S_n + 1, n - S_n + b).$$

Independently draw $U_{e,n} \sim \mathrm{Beta}(1, b)$ for every $e \in E_{\mathcal{F}}$, using one sample per edge shared across all targets containing it. Define

$$Z_{k,n} := \prod_{e \in P_k} U_{e,n}, \qquad D_n := \{\max_{k \leq K} Z_{k,n} > Y_n\}.$$

All target-edge draws are independent across edges and auxiliary rounds, and independent of the competitor outcome and sampling randomness. Target products within a round may be dependent. The auxiliary process observes the competing edge after every round.

Couple the actual algorithm to these draws and outcomes until its first target selection. Before that selection it always chooses $\rho = (e_0) \circ C$, whose first edge e_0 is observed under both feedback models. Its posterior therefore matches the auxiliary posterior, while target-prefix edges retain their initial priors. Shared-suffix samples cancel regardless of their posterior updates. Hence, up to probability-zero ties,

$$\tau_{\mathcal{F}} = 1 + \inf\{n \geq 0 : D_n\}, \qquad \{\tau_{\mathcal{F}} > T\} = \bigcap_{n=0}^{T-1} D_n^c. \tag{14}$$

An infimum over an empty set is interpreted as infinity. All probabilities concerning Y_n in the following proofs are unconditional auxiliary-process probabilities, averaged over S_n, rather than survival-conditioned probabilities.

C.3 Proof of Theorem 1

For standard TS, take $b = 1$ and define

$$c_- := \frac{p^2}{16}, \qquad n_p := \left\lceil \frac{1}{c_-} \log \frac{16}{1 - e^{-c_-}} \right\rceil, \qquad m_0(p, K) := \left\lceil \log_2\!\big(4Kn_p(n_p+1)\big) \right\rceil. \tag{15}$$

Since $1 - e^{-c_-} \geq c_-/2$ for $0 < c_- \leq 1$, these choices give $n_p = O(p^{-2}[1 + \log(1/p)])$ and hence $m_0(p, K) = O(1 + \log K + \log(1/p))$.

Proof. We separately control early discovery, low competitor samples after burn-in, and large target samples.

Early discovery. Conditional on $S_n = s$, the density of Y_n is

$$f_{n,s}(y) = (n+1)\binom{n}{s} y^s (1-y)^{n-s} \leq n + 1.$$

Hence $\Pr(Y_n < z \mid S_n = s) \leq (n+1)z$. Each target score has mean $\mathbb{E}[Z_{k,n}] = 2^{-m}$, and is independent of Y_n and S_n. A union bound over targets, valid for overlapping prefixes, therefore gives $\Pr(D_n) \leq K(n+1)2^{-m}$. For $m \geq m_0(p, K)$, a union bound gives

$$\Pr\left(\bigcup_{n=0}^{n_p-1} D_n\right) \leq \frac{Kn_p(n_p+1)}{2^{m+1}} \leq \frac{1}{8}. \tag{16}$$

Competitor concentration. Set $q = p/2$. Hoeffding's inequality yields $\Pr(S_n < 3np/4) \leq e^{-np^2/8}$. The beta–binomial identity gives

$$\Pr(Y_n < q \mid S_n = s) = \Pr(\mathrm{Bin}(n+1, q) \geq s + 1)\,.$$

For $s \geq 3np/4$, the threshold exceeds the binomial mean by at least $np/4$. For $n \geq 1$, another application of Hoeffding's inequality therefore gives

$$\Pr(Y_n < q \mid S_n = s) \leq \exp\left(-\frac{n^2p^2}{8(n+1)}\right) \leq e^{-np^2/16}.$$

Averaging over S_n yields $\Pr(Y_n < q) \leq 2e^{-c_- n}$. By the definition of n_p,

$$\Pr\left(\bigcup_{n=n_p}^{\infty} \{Y_n < q\}\right) \leq \frac{2e^{-c_- n_p}}{1 - e^{-c_-}} \leq \frac{1}{8}. \tag{17}$$

Target product tail. For independent uniform variables, $-\log Z_{k,n}$ has a $\mathrm{Gamma}(m, 1)$ distribution. Writing $a_q = \log(1/q)$,

$$\Pr(Z_{k,n} \geq q) = \frac{1}{(m-1)!} \int_0^{a_q} x^{m-1} e^{-x}\, dx \leq \frac{a_q^m}{m!}.$$

Thus, with $a := K[\log(2/p)]^m/m!$, the probability that any target score exceeds q in a given round is at most a.

For $n \geq n_p$, the event D_n implies either $Y_n < q$ or $\max_k Z_{k,n} \geq q$. Combining this observation with equation 16–equation 17 and taking a union bound over the first T rounds gives

$$\Pr(\tau_{\mathcal{F}} > T) \geq \left(\frac{3}{4} - aT\right)_+, \qquad T \in \mathbb{Z}_{\geq 0}. \tag{18}$$

Discovery time and regret. Summing the survival probabilities and comparing the decreasing summand with its integral gives

$$\mathbb{E}[\tau_{\mathcal{F}}] \geq \sum_{T=0}^{\infty} \left(\frac{3}{4} - aT\right)_+ \geq \int_0^{\infty} \left(\frac{3}{4} - ax\right)_+ dx = \frac{9}{32a} \geq \frac{1}{8a}.$$

Let $R(T)$ denote expected cumulative pseudo-regret. On $\{\tau_{\mathcal{F}} > t\}$, the algorithm selects the competitor in round t and incurs gap $\Delta_{\mathcal{F}}$. Hence

$$R(T) \geq \Delta_{\mathcal{F}} \sum_{t=1}^{T} \Pr(\tau_{\mathcal{F}} > t).$$

When $T \leq 1/(4a)$, every term in this sum is at least $1/2$, yielding $R(T) \geq \Delta_{\mathcal{F}} T/2$. Substituting the definition of a proves the theorem. □

C.4 Competitive Tails and Proof of Theorem 2

C.4.1 Proof of Lemma 1

Let $U_1, \ldots, U_L$ be independent $\mathrm{Beta}(1, b)$ variables with $Lb \leq \lambda$, and fix $q \in (0, 1)$. Set $W_i = 1 - U_i$ and $\varepsilon = \min\{1, (1-q)/(2\lambda)\}$. The distribution function of W_i is x^b for $x \in [0, 1]$. Consequently, for $A = \bigcap_{i=1}^{L} \{W_i \leq \varepsilon\}$,

$$\Pr(A) = \varepsilon^{Lb} \geq \varepsilon^{\lambda}, \qquad \mathbb{E}[W_i \mid A] = \frac{b}{\varepsilon^b} \int_0^{\varepsilon} x^b \, dx = \frac{b\varepsilon}{b+1}.$$

Independence justifies conditioning each variable on its own truncation event. Thus

$$\mathbb{E}\left[\sum_{i=1}^{L} W_i \mid A\right] = \frac{Lb\varepsilon}{b+1} \leq \lambda\varepsilon \leq \frac{1-q}{2}.$$

Markov's inequality gives $\Pr(\sum_i W_i \leq 1 - q \mid A) \geq 1/2$. For $w_i \in [0, 1]$, induction yields $\prod_i (1 - w_i) \geq 1 - \sum_i w_i$. It follows that

$$\Pr\left(\prod_{i=1}^{L} U_i \geq q\right) \geq \frac{1}{2}\varepsilon^{\lambda} = \eta_{\lambda}(q). \tag{19}$$

For $b = \lambda/H$, the condition holds for every $L \leq H$.

C.4.2 Fixed-Prior Product Tail

For $U \sim \mathrm{Beta}(1, b)$, $V = -\log U$ has density

$$f_V(v) = be^{-v}(1 - e^{-v})^{b-1}, \qquad v > 0.$$

If $b \geq 1$, use $1 - e^{-v} \leq v$ to obtain $f_V(v) \leq bv^{b-1}$. If $0 < b < 1$, use $1 - e^{-v} \geq ve^{-v}$ to obtain $f_V(v) \leq bv^{b-1}e^{-bv} \leq bv^{b-1}$. Writing $\ell_q = \log(1/q)$, independence and the Dirichlet integral yield

$$\begin{aligned}
\Pr\left(\prod_{i=1}^{L} U_i \geq q\right) &= \Pr\left(\sum_{i=1}^{L} V_i \leq \ell_q\right) \\
&\leq b^L \int_{\substack{v_i \geq 0 \\ \sum_i v_i \leq \ell_q}} \prod_{i=1}^{L} v_i^{b-1} \, dv_1 \cdots dv_L \\
&= \frac{\Gamma(1+b)^L \ell_q^{bL}}{\Gamma(1+bL)}.
\end{aligned} \tag{20}$$

For fixed $b > 0$ and $q \in (0, 1)$, Stirling's formula gives

$$\log \Pr\left(\prod_{i=1}^{L} U_i \geq q\right) \leq -bL \log L + O_{b,q}(L).$$

This includes the uniform-product bound when $b = 1$.

C.4.3 Length-Uniform Discovery Bound

Use the auxiliary process of Appendix C.2 with $b = \lambda/H$ and $H \geq \max_{\pi \in \Pi} |\pi|$. Define

$$q := \frac{1+p}{2}, \qquad \eta := \eta_\lambda(q), \qquad c_+ := \frac{(1-p)^2}{8}, \qquad C_p := \frac{4e^{-c_+}}{(1-e^{-c_+})^2}. \tag{21}$$

Here c_+ is the constant denoted c_p in the theorem.

Proof. We combine target optimism with competitor concentration.

Target optimism. Fix any one target, say P_1. Its m distinct edges have independent prior samples within each auxiliary round. Since $mb \leq \lambda$,

$$\Pr(Z_{1,n} \geq q) \geq \eta. \tag{22}$$

The events $\{Z_{1,n} \geq q\}$ are independent across auxiliary rounds. This fixed-target argument accommodates shared edges among the remaining targets.

Competitor concentration. For every $b > 0$ and $n \geq 1$, we establish

$$\Pr(Y_n > q) \leq 2e^{-c_+ n}. \tag{23}$$

Let $r = (p + q)/2$. Since $S_n \sim \mathrm{Bin}(n, p)$, Hoeffding's inequality gives

$$\Pr(S_n > nr) \leq \exp[-2n(r-p)^2] = \exp[-n(q-p)^2/2].$$

For an integer $s \leq nr < n$, the parameter $n - s$ is positive. Increasing the second beta parameter shifts the distribution stochastically toward zero, so the beta–binomial identity gives

$$\Pr(Y_n > q \mid S_n = s) \leq \Pr(\mathrm{Beta}(s+1, n-s) > q) = \Pr(\mathrm{Bin}(n, q) \leq s)\,.$$

By Hoeffding's inequality, the last probability is at most $\exp[-2n(q-r)^2] = \exp[-n(q-p)^2/2]$. Averaging over S_n proves equation 23, because $(q-p)^2/2 = c_+$.

Survival probability. For $T \geq 2$, put $j = \lfloor T/2 \rfloor$. If $Z_{1,n} \geq q$ and $Y_n \leq q$ at any index $n \in \{j, \ldots, T-1\}$, then D_n occurs almost surely, so discovery has occurred by round T. Therefore equation 14 implies

$$\{\tau_{\mathcal{F}} > T\} \subseteq \bigcap_{n=j}^{T-1} \{Z_{1,n} < q\} \cup \bigcup_{n=j}^{T-1} \{Y_n > q\},$$

up to probability-zero ties. Independence of the fixed target's samples across rounds and a union bound for competitor events yield

$$\begin{aligned} \Pr(\tau_{\mathcal{F}} > T) &\leq (1-\eta)^{T-j} + \sum_{n=j}^{T-1} 2e^{-c_+ n} \\ &\leq (1-\eta)^j + \frac{2e^{-c_+ j}}{1-e^{-c_+}}. \end{aligned} \tag{24}$$

All bounds here use unconditional auxiliary-process probabilities.

Expected discovery time. Use the bound of one for $T = 0, 1$ and pair $T = 2j$ with $T = 2j + 1$:

$$\begin{aligned}\mathbb{E}[\tau_{\mathcal{F}}] &= \sum_{T=0}^{\infty} \Pr(\tau_{\mathcal{F}} > T) \\ &\le 2 + 2\sum_{j=1}^{\infty} \left[(1-\eta)^j + \frac{2e^{-c_+ j}}{1-e^{-c_+}}\right] \\ &= \frac{2}{\eta} + C_p = O_{p,\lambda}(1). \end{aligned} \tag{25}$$

This bound is uniform in m, K, target-prefix overlap, and valid choices of H for fixed p and λ.

Pre-discovery regret. Every round strictly before discovery selects ρ, whose gap is $\Delta_{\mathcal{F}} = w(\nu - p)$. Thus

$$\mathbb{E}\left[\sum_{t=1}^{\tau_{\mathcal{F}}-1} (\mu_{\mathcal{F}} - \mu(\pi_t))\right] = \Delta_{\mathcal{F}}(\mathbb{E}[\tau_{\mathcal{F}}] - 1).$$

This proves the theorem under both feedback models. □

Scope of the separation. For $\lambda = 1$, $\eta = (1-p)/8$, so the upper bound becomes $16/(1-p) + C_p$. Together with Theorem 1, this yields a factorial-scale TS lower bound and a length-uniform RC-TS upper bound for fixed p and K on the overlapping path-trap family. The TS regret bound describes a finite-horizon linear regime; the RC-TS regret identity concerns rounds strictly before the first target selection.

D Baseline Implementation Details

Shared-seed baseline. CTS^3 (Jin et al., 2026) coordinates component samples through a common uniform random seed and inverse distribution functions. For the feedback conventions used here, cascade observations supply outcomes from the attempted prefix, while semi-bandit observations supply outcomes from every selected edge.

Method	Parameter	Value
TS	Beta prior	$\mathrm{Beta}(1, 1)$
	Probability floor	10^{-9}
TS-Cascade	Sampling distribution	Gaussian
	Gaussian coupling	Shared Z_t per episode
	Probability range	$[0, 1]$
	Log floor	10^{-12}
CombCascade	Confidence coefficient	1.5
	Probability ceiling	1
BCUCB-T	Variance coefficient	6
	Additive coefficient	9
	Theory confidence scale c	1
	Tuned confidence scale c	0.25
	Unobserved-edge estimate	1
	Probability ceiling	1
	Tie-breaking scale	10^{-8}
	Greedy-policy floor	$p_{\min}$
CTS^3	Beta prior	$\mathrm{Beta}(1, 1)$
	Marginal sampling	Exact Beta posterior quantile
	Posterior coupling	Shared $U_t \sim \mathrm{Unif}(0, 1)$ per episode
	Probability floor	10^{-9}
Random	Maximum restart attempts	1000

Table 5: Hyperparameters and implementation details.