The Joint Channel Threshold: Selfish Mining Below 1% Hashrate

Fateme Ghasemi, Reyhaneh Ameri, and Mohammadreza Meybodi

Abstract—The selfish-mining literature has progressively lowered the hashrate threshold at which a rational Bitcoin miner strictly profits by deviating from the honest protocol, from Eyal and Sirer's 25% down to single digits under richer attacker models. Two such models remain unreconciled: Gervais et al. (CCS 2016) parameterize network-layer adversaries through an eclipse fraction ω and an honest-only stale rate o, but assume constant block rewards; WeRLman (IEEE S&P 2023) models fee-rich rewards through a deep-RL solver, but assumes a network-clean attacker. An adversary that exploits both surfaces has not been analyzed.

We present the first joint analysis. We formulate a Markov decision process that integrates Gervais's network-layer channels with WeRLman's whale-transaction tracking, solve it at a 5.7×10^8-state truncation, and anchor the construction by reduction to each prior model at the appropriate parameter limit.

Our threat model targets a state-level adversary with per-block routing influence over a meaningful fraction of honest hash power, not an attacker who eclipses individual nodes. Under this model, the joint threshold collapses far below either single-channel result. At $L = 3$, the F=10 baseline (no network channels) crosses at $\alpha = 13.2\%$; a 10% eclipse fraction alone drops it to 2.6%, a 10% stale rate alone to 4.8%, and both channels together to 1.4%. At $L = 5$ the joint configuration drops further to 0.9%, an at least $15\times$ collapse. An analytical lemma accounts for the interaction: the honest-chain growth probability factors as $(1 - \alpha - \omega)(1 - o)$, isolating a cross-term ωo that is a property of the joint transition kernel itself. The cross-term is structural, not a training artifact, and its presence implies that mitigations targeting only one channel leave the joint contribution structurally present whenever the other channel is non-zero.

Index Terms—Blockchain security, selfish mining, proof-of-work, Markov decision process, deep reinforcement learning, eclipse attack, stale blocks, transaction fees, profitability threshold.

1 Introduction

The security of proof-of-work blockchains such as Bitcoin [1] and Ethereum [2] rests on the assumption that a rational miner gains nothing by deviating from the longest-chain rule. Eyal and Sirer [3] disproved this assumption: above roughly a quarter of the network hashrate, withholding blocks—selfish mining—strictly beats honest participation. The bound is not theoretical. Several deployed mining pools have already cleared it [4], and each sharpening of the threshold below it has tightened the operational constraint on real deployments.

The literature has sharpened the threshold by enriching the attacker model along two largely separate axes. On the strategic axis, Sapirshtein et al. [5] widened the action space with a *match* action and recovered the optimum via a linearized MDP; Nayak et al. [6] added stubborn variants and combined them with eclipse attacks [7]. On the network axis, Gervais et al. [8] introduced the two surfaces this paper builds on: an *eclipse fraction* ω parameterizing honest hash power redirected to the attacker's chain, and an *honest-only stale rate* o capturing honest blocks lost to propagation delay. On the economic axis, Carlsten et al. [9] showed that once transaction fees grow comparable to the block subsidy, sub-threshold miners may rationally deviate to capture fee-rich blocks—a finding later grounded empirically through miner-extractable value measurements on live chains [10]. WeRLman [11] formalized the fee-rich regime by tracking per-slot whale-transaction arrivals and introduced an MCTS-DQN trainer to handle the resulting state-space inflation. The thread continues through petty-compliance analysis [12], stochastic-reward generalizations [13], and further deep-RL approaches [14], [15].

Each axis sharpens the threshold in isolation, but a single attacker is not confined to one. The eclipse, stale-rate, and fee surfaces have each been quantified carefully on their own; their joint effect on a single attacker has not. The recent systematization of selfish-mining models [16] flags this combination as unoccupied, and the earlier Bitcoin SoK [17] predates the fee-rich extensions altogether. The gap is structural rather than an oversight: it is a solver-scaling problem. Gervais et al.'s binary-search-on-linearization approach [8] is comfortable at the $\mathcal{O}(10^3)$ states that network-layer surfaces produce, but breaks at the $\sim 10^8$ states that whale-transaction tracking introduces. WeRLman's trainer [11], [18], [19], [20] reaches that scale but is implemented only over the economic surface. Bridging the two requires both a kernel that handles network and economic events jointly within every transition, and a solver that scales to the resulting state space.

Two questions are therefore open: whether the network and economic channel families interact non-trivially or their revenue effects simply add; and where the joint-channel threshold α^* sits relative to the single-channel results. This paper answers both.

Contributions.: We make three findings about selfish mining under joint attacker capabilities.

1) **Threshold collapse against a state-level adversary.** Achieving 10% joint intensity requires a state-level

• *The authors are with Amirkabir University of Technology, Tehran, Iran. E-mail: fatemeghasemi492@aut.ac.ir; r.ameri@aut.ac.ir; mmeybodi@aut.ac.ir.*

adversary with per-block routing influence over a meaningful share of honest hash power, not an attacker who eclipses individual nodes. Under this adversary class, combining the eclipse and stale-rate channels at 10% intensity drops the profitability threshold from 13.2% at the F=10 baseline to 0.9% at fork-horizon $L = 5$ (1.4% at the more conservative $L = 3$ truncation), a collapse of more than $15\times$. The collapse strengthens at the longer horizon, so the $L = 3$ result is a conservative bound rather than a tight estimate.

2) **Cross-term lemma.** The honest chain's growth probability factors as $(1 - \alpha - \omega)(1 - o)$, isolating a multiplicative cross-term $\omega\, o$ that makes the joint reduction strictly greater than the sum of the single-channel reductions (Lemma 1). The cross-term is a property of the joint transition kernel itself, not of any specific solver run; it predicts the sign and scale of the revenue interaction we observe, and single-channel mitigations cannot eliminate it.
3) **Sub-additive interaction surface.** The cross-term is mathematically present at all intensities but grows to a statistically detectable 5–9% sub-additive plateau only above $\omega, o \geq 0.20$, outside today's deployed-Bitcoin parameter space. Below 15% intensity the interaction is indistinguishable from additive at our sample size. This locates the boundary between the two regimes and confirms that the deployment-relevant region sits on the additive side.

The MDP underlying these results integrates Gervais's eclipse and stale-block channels with WeRLman's per-slot whale-transaction tracking in a single kernel. We solve it with the PTO objective transform of [21] and the MCTS-DQN trainer of WeRLman [11], and verify the construction by reduction to Gervais [8] at zero fees and to WeRLman [11] at zero network-layer intensity.

Outline.: §2 situates our work within the selfish-mining lineage. §3 states the threat model, and §4 presents the joint MDP. The solver is described in §5. §5.3 reports the threshold collapse, §5.4 the interaction surface, and §5.5 the horizon and γ-sensitivity checks. §6 states the cross-term lemma and discusses connections to related models, and §7 concludes.

2 Related Work

The selfish-mining literature has occupied three methodological generations—analytical bounds, classical MDP solvers, and deep-RL solvers—without ever combining the network-layer and economic-layer attack surfaces in a single model. Table 1 maps the lineage; we survey each generation in turn.

The foundational result and its profitability threshold.: The smallest hashrate fraction at which a deviating miner earns strictly more than its honest share—the *profitability threshold* α^*—is the operational bound on safe hashrate concentration in a Nakamoto proof-of-work blockchain. Eyal and Sirer [3] established the original threshold at approximately 25% in a constant-reward model with a hand-coded selfish-mining strategy (SM1). Subsequent work has tightened this estimate by enlarging either the strategy space or the threat model.

From hand-coded strategies to MDP optima.: Sapirshtein, Sompolinsky, and Zohar [5] were the first to cast selfish mining as a Markov decision process, with state $(\ell_a, \ell_h, \text{fork})$ encoding the attacker's lead, the honest chain's lead, and the race-state token. They recovered the revenue-maximizing strategy via binary search over a discount-linearized transformation of the ratio objective. Bar-Zur, Eyal, and Tamar [21] replaced that linearization with the *Probabilistic Termination Optimization* (PTO) transform, which recasts the ratio as a stochastic shortest-path problem admitting standard policy iteration and deep-RL solvers alike. Both formulations remain within the $\mathcal{O}(10^3$–$10^4)$ state scale typical of fee-free, network-clean models. We adopt PTO because at the $\sim 10^8$-state scale of the joint MDP, classical policy iteration is infeasible, and the deep-RL solver we rely on requires the stochastic shortest-path form PTO produces.

Network-layer extensions.: Nayak et al. [6] introduced the stubborn strategy family and combined selfish mining with a binary eclipse state. Gervais et al. [8] parameterized the eclipse channel as a fraction ω of non-strategic hash power redirected to the attacker's chain, and introduced an asymmetric honest-only stale rate o for honest blocks orphaned by propagation delay. Both extensions push the profitability threshold below the 25% Eyal–Sirer floor. Eclipse feasibility on the live Bitcoin network was demonstrated by Heilman et al. [7], providing deployment grounding for ω; Ritz and Zugenmaier [22] examined the propagation-stale interaction in Ethereum's uncle-reward setting. Gervais et al. solved their MDP with the same binary-search-on-linearization approach as Sapirshtein: tractable at the $\mathcal{O}(10^3)$ states their truncation produces, but not beyond.

Economic-layer extensions and the deep-RL turn.: Carlsten et al. [9] observed that once transaction fees become comparable to the block subsidy, sub-threshold miners may rationally deviate to capture fee-rich blocks. Daian et al. [10] confirmed this empirically through miner-extractable value measurements on deployed chains. In WeRLman, Bar-Zur et al. [11] formalized the fee-rich regime through an MDP that tracks per-block whale-transaction arrivals at rate δ with reward multiplier F. At fork horizon $L = 10$ the state space reaches $\mathcal{O}(10^8)$, beyond what binary-search-on-linearization can address; WeRLman therefore introduced an MCTS-DQN trainer [18], [19], [20] pairing Monte-Carlo Tree Search rollouts with Deep Q-Network value approximation. Independent deep-RL approaches include SquirRL [14], which applies Q-learning over hand-crafted state features, and Wang et al. [15], which formulates a multi-dimensional RL variant. We work within the WeRLman framework because it tracks whale-transaction state at the per-block-slot granularity our joint analysis requires.

Other variants of attacker behavior.: Several lines examine attacker capabilities orthogonal to the joint composition studied here. Bar-Zur et al. [12] (*Deep Bribe*) later modeled a petty-compliant miner that tie-breaks on residual fees rather than reception order; Kwon et al. [23] analyzed fork-after-withholding attacks; Bahrani et al. [13] treated the orphan rate as an equilibrium outcome of stochastic rewards rather than as an exogenous parameter; Yaish et al. [24]

exploited Ethereum-specific timestamp ambiguity in *Uncle Maker*. Analytic profitability analyses include Grunspan and Pérez-Marco [25] and Feng and Niu [26], and adjacent protocol designs include Bitcoin-NG [27], the gap-game [28], and Fruitchains [29]. None parameterizes the network-layer (ω, o) and fee-rich (F) surfaces jointly; prospects for composing some of these variants with the present model are discussed in §6.

Systematizations and the unoccupied cell.: Finkbeiner et al. [16] systematize twenty-four selfish-mining analyses along four axes: whether fees are modeled, whether non-honest miner classes are considered, whether the attacker's network-layer position is parameterized, and whether the analysis is closed-form, classical-MDP, or RL-based. The combination of fee modeling with the ω/o network-layer subset is unoccupied: works that model fees (WeRLman [11], Carlsten et al. [9], Bahrani et al. [13]) do not parameterize network-layer attack surfaces, and works that parameterize ω and o (Gervais et al. [8], Nayak et al. [6]) do not model fees. The present paper occupies that cell. Li et al. [30] survey RL-based selfish-mining analyses, to which we contribute one entry along the network-layer axis.

The lineage in Table 1 reads as a sequence of single-dimension extensions, each accompanied by the smallest solver upgrade the new dimension forces. Carlsten et al. added fees analytically; Gervais et al. added eclipse and stale-block channels within the classical-solver budget; WeRLman moved fees into the deep-RL budget; Deep Bribe added petty compliance within that same budget. The joint composition of network-layer and economic-layer surfaces requires both the deep-RL state scale and a transition kernel that has not been formulated before; we present that kernel in §4.

3 Threat Model

Our threat model captures a single attacker with composite capabilities. A rational miner controls a fraction of the network's hash power and has every network-layer and economic-environment capability the model parametrizes. The four parameters of the joint MDP (α, ω, o, and the fee parameters F, δ) sit on different layers of the system stack and carry different procurement costs; we assume the attacker has paid all four. The quantitative results in §5.3–§5.5 should be read under this threat model, not as universal predictions across all attacker budgets.

Attacker.: The attacker is a single mining entity controlling a fraction α of the network's hash power. The miner is rational: it deviates from the protocol if and only if doing so increases its expected per-block revenue. We do not model coordinated coalitions of selfish miners. We assume the attacker has full visibility of the honest chain (no honest counter-eclipse).

Attacker capabilities.: Beyond the hash-power share, the model has three capability axes; each is obtained by a distinct mechanism.

Eclipse (network-layer). An eclipse attacker monopolizes a victim node's connections to the peer-to-peer network [7]; in selfish-mining models, this redirects the victim's mined blocks onto the attacker's chain rather than the honest one [8]. We parametrize this as a fraction $\omega \in [0, 1-\alpha]$ of non-strategic hash power that is eclipsed. Heilman et al. [7] demonstrated that eclipsing an individual Bitcoin node is feasible with modest IP and peer budgets (∼400 IPs, ∼$100). Our $\omega = 0.10$ corresponds to fully eclipsing a major mining pool (calibrated against F2Pool's $\sim 10\%$ share; see "Joint feasibility" below)—a substantially harder attack because pools have more peers, more nodes, and active defensive monitoring.

Propagation advantage (network-layer). A faster relay path or lower-latency peer set lets the attacker's blocks propagate before honest miners' blocks, raising the rate at which honest blocks go stale. We parametrize this with the honest-only stale rate $o \in [0, 1]$, inheriting Gervais's asymmetric-stale semantics: the attacker's blocks never go stale by assumption. Gervais et al. [8] measure approximately 0.4% on the 2016 deployed network. The operational rate has fallen substantially since 2016. Nagayama et al. [31] report that compact block relay (BIP 152 [32]) alone reduces propagation latency by roughly 90% at the 50th percentile and 88% at the 90th. Combined with improved global connectivity, this places the current operational stale rate well below 0.1%. Our $o = 0.10$ is therefore $\sim 25\times$ the 2016 baseline and a still larger multiple of the current operational rate; the calibration is conservative against today's network rather than overstated. We retain the Gervais 0.4% as reference because the asymmetric-stale semantics is a property of the parametrization rather than of any era's measured rate, and cross-paper comparability with Gervais requires holding the calibration fixed.

Whale transactions (economic-layer). The fee multiplier $F \geq 0$ and arrival rate $\delta \in [0, 1]$ characterize fee-rich block-space demand. *These are environment parameters, not attacker capabilities*: the attacker chooses when to release its private chain to coincide with whale arrivals, but does not choose F or δ. Our headline experiments fix $F = 10$ (matching the upper end of WeRLman's swept range) and $\delta = 0.01$ (an order of magnitude below WeRLman's $\delta = 0.1$ default; we use the lower value to match the empirical Bitcoin whale-arrival rate). The product $F \cdot \delta = 0.1$ corresponds to fee revenue averaging roughly 10% of subsidy per block—representative of a forward-looking Bitcoin in which subsidy halving has continued. Sensitivity to F is addressed in Section 5.5.

Joint feasibility.: The eclipse and propagation channels are independent capabilities and compose multiplicatively in cost. An eclipse attack against a pool the size of F2Pool—the fourth-largest at the time of writing, controlling roughly 10% of network hashrate—combined with sustained low-latency relay infrastructure ($o \approx 0.10$, $\sim 25\times$ the deployed Bitcoin baseline) is within the budget of a state-level adversary, not an individual operator. We report empirical results across the full $[0.05, 0.25] \times [0.05, 0.25]$ grid because the model produces them, but we read the $\omega, o \geq 0.20$ corner as a *model property at extreme parameters* rather than a deployment-relevant prediction. Even at the headline $\omega, o = 0.10$ we do not extrapolate to natural Bitcoin propagation conditions.

Out of scope.: The model is single-attacker. We do not analyze coordinated coalitions, mixed-strategy attackers, or attackers that can choose F or δ (for example through mempool market manipulation). We do not model defender re-

TABLE 1: Selfish-mining literature on Nakamoto proof-of-work, grouped by methodological generation. Columns: F (fee multiplier), ω (eclipse fraction), o (honest-only stale rate), PC (petty compliance); state is the order of magnitude of states the solver handles; α^* is the smallest reported profitability threshold under the strongest threat model in that work.

Work	Venue	F	ω	o	PC	Solver	State	α^*	Contribution
Analytical / closed-form									
Eyal–Sirer [3]	FC 2014	—	—	—	—	SM1	—	$\sim 25\%$	Selfish-mining + threshold
Nayak et al. [6]	EuroS&P'16	—	binary	—	—	analytical	—	$\sim 23\%$	Stubborn + binary eclipse
Carlsten et al. [9]	CCS 2016	yes	—	—	—	analytical	—	$< 25\%$	Fee-driven undercutting
Bahrani et al. [13]	EC 2025	yes	—	derived	—	analytical	—	n/a†	Stochastic-reward theorem
MDP with classical solver									
Sapirshtein et al. [5]	FC 2016	—	—	—	—	bin.-search	10^3	$\sim 25\%$	Optimal-strategy MDP
Gervais et al. [8]	CCS 2016	—	ω	o	—	bin.-search	10^3	$< 25\%$	Network-layer surface
Bar-Zur et al. [21]	AFT 2020	—	—	—	—	PTO + VI	10^4	$\sim 25\%$	PTO objective transform
MDP with deep RL									
WeRLman [11]	S&P 2023	yes	—	—	—	MCTS-DQN	10^8	$\sim 7.5\%$§	Whale-tx + deep RL
Deep Bribe [12]	S&P 2023	yes	—	—	yes	MCTS-DQN	10^8	—	Petty-compliant miner
This work	—	**yes**	$\boldsymbol{\omega}$	$\boldsymbol{o}$	—	**MCTS-DQN**	$\mathbf{5.7{\cdot}10^8}$	**0.9%**‡	Joint network + fee model

†Bahrani et al. derive a stochastic-reward characterization rather than a numerical α^*. ‡Joint threshold at $\omega = o = 0.10$ and horizon $L = 5$ (our horizon-robustness anchor); the corresponding $L = 3$ threshold is 1.4%, so the longer horizon makes the collapse stronger, not weaker. The extreme corner $\omega = o = 0.25$ falls below 0.5% at $L = 3$ but lies outside the realistic-adversary regime. §WeRLman's threshold is parameter-dependent; the value shown is at $F = 10$, $\delta = 0.1$, $L = 10$ from their Fig. 8(a). Our F=10 baseline of 13.2% uses $\delta = 0.01$ and $L = 3$ (see §5.3).

sponse; the profitability threshold α^* is the attacker's break-even point under fixed honest behavior, not an equilibrium.

4 Model

We define the joint MDP $M = (\mathcal{S}, \mathcal{A}, P, R)$. The state encoding extends Gervais [8] for the network-layer components and WeRLman [11] for the economic-layer components. We adopt WeRLman's per-block-slot transaction granularity throughout. The joint kernel needs to track which blocks on the private chain carry whale transactions and which are eclipsed: two properties that vary independently along the chain and both feed into the reward expression in §4.4. Figure 1 illustrates the three core scenarios the model must represent.

4.1 Parameters

Table 2 lists the model's parameters with their domains, source, and headline-experiment value. Three parameters require additional semantic detail before the transition kernel can be read.

The eclipse fraction ω is implemented as a per-event probability rather than a state-of-the-world parameter: at each block-discovery event, with probability ω the block is mined by an eclipse victim and lands on the strategic chain, regardless of the history of previous events. This realises Gervais's eclipse parametrisation as a memoryless transition rule. The honest-only stale rate o is similarly per-event: with probability o, an honest block-discovery resolves as a no-op (neither chain grows, no reward), while strategic and eclipsed-victim blocks are never stale—an asymmetric-stale semantics inherited from [8]. The fork-horizon cap L is a state-space truncation that bounds the private chain length; it is the same truncation used by Sapirshtein et al. [5] and WeRLman [11].

TABLE 2: Parameters of the joint MDP. "Origin" indicates which prior work introduces each parameter into the selfish-mining MDP literature; the present paper combines them in a single model. Headline experiments fix the "Value" column unless noted otherwise.

Symbol	Description	Domain	Origin	Value
α	strategic hash share	$[0, 1)$	[3]	swept
γ	rushing factor	$[0, 1]$	[5]	0.5
ω	eclipsed fraction	$[0, 1 - \alpha]$	[8]	swept
o	honest-only stale rate	$[0, 1]$	[8]	swept
δ	whale-transaction arrival rate	$[0, 1]$	[11]	0.01
F	whale-transaction fee multiplier	$\mathbb{R}_{\geq 0}$	[11]	10
L	fork-horizon cap	$\mathbb{N}$	[5], [11]	3
R	whale-transaction reservoir cap	$\mathbb{N}$	[11]	L

4.2 State space

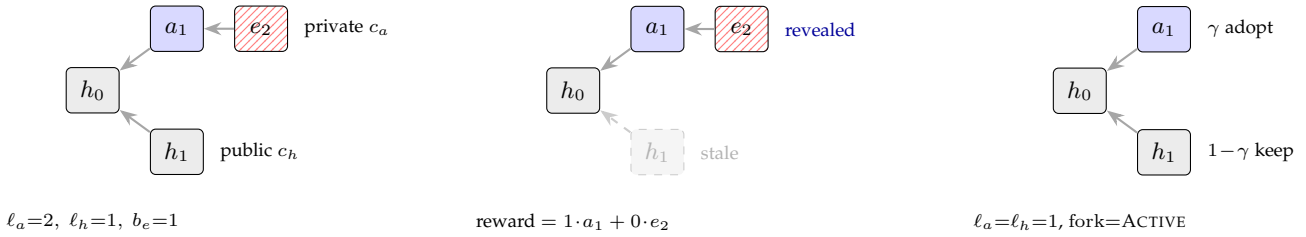
A state $s \in \mathcal{S}$ is a 9-component tuple summarized in Table 3:

$$s = \big(\mathbf{c}_a, \mathbf{c}_h, \mathrm{fork}, \mathrm{pool}, l_a, l_h, t_a, t_h, b_e\big).$$

We treat each chain $\mathbf{c}_a$, $\mathbf{c}_h$ as a single structured component for exposition; in the implementation they unroll into interleaved $(\mathrm{Block}, \mathrm{Transaction})$ pairs of length L, giving the flat encoding $|s| = 4L + 7$ used in the state-space-size formula below.

The chain components $\mathbf{c}_a$ and $\mathbf{c}_h$ encode the private (attacker) and public (honest) chains beyond the most recent fork point. Each position carries a Transaction tag from a three-valued alphabet $\{\mathrm{NoTx}, \mathrm{With}, \mathrm{Eclipsed}\}$ that combines whale-transaction content with block origin: NoTx marks a block without a whale transaction, With marks a block carrying one, and $\mathrm{Eclipsed}$ marks a block mined by an eclipse victim. Eclipse-victim blocks reach only the attacker's chain by construction, so the $\mathrm{Eclipsed}$ tag is unreachable on $\mathbf{c}_h$ from the initial state under the transition kernel of §4.4.[1]

1. Allocating the same alphabet on both chains is an implementation choice that keeps state encoding uniform; it has no effect on the kernel, since public-chain positions are never assigned the $\mathrm{Eclipsed}$ tag by any transition.



(a) State. (b) Reveal (override). (c) Race state.

Fig. 1: Three scenarios in the joint MDP. **(a)** A typical *state* after the strategic miner has mined two private blocks (a_1, e_2), one of which is an eclipsed victim's block (e_2, hatched red) — it advances the private chain length ℓ_a and increments the eclipse counter b_e, but its coinbase reward flows to the eclipse victim, not the strategic miner. The honest network has mined one public block (h_1). **(b)** The *Reveal* action publishes the private chain; since $\ell_a > \ell_h$, the strategic chain wins, the public block h_1 becomes stale, and the strategic miner is rewarded for a_1 only (e_2's coinbase goes to the eclipse victim). **(c)** A *race state* reached when $\ell_a = \ell_h$ and fork = ACTIVE: a fraction γ of honest hash power adopts the strategic miner's chain on the next MINE, and the remaining $1-\gamma$ keeps mining on the public chain. Honest blocks may also be lost to staleness with probability o (not shown).

The fork token records whether the public chain's most recent block is stale (Irrelevant), eligible for matching (Relevant), or part of a declared active fork (Active). The mempool counter $\text{pool} \in \{0, 1, \ldots, R\}$ tracks pending whale transactions awaiting inclusion; we enforce $R \geq L$ so that the reservoir cannot under-supply the chain. The cached counts l_a, l_h, t_a, t_h, b_e are formally redundant with $\mathbf{c}_a$ and $\mathbf{c}_h$ (each is recoverable by counting tags); we materialize them in the state encoding for $O(1)$ indexing during policy iteration.

State-space size.: The Cartesian product of the per-component domains gives, at $L = R = L$,

$$|\mathcal{S}| = \underbrace{(2 \cdot 3)^L}_{\mathbf{c}_a} \cdot \underbrace{(2 \cdot 3)^L}_{\mathbf{c}_h} \cdot 3 \cdot (L{+}1)^6 + 1$$
$$= 6^{2L} \cdot 3 \cdot (L{+}1)^6 + 1,$$

where the trailing $+1$ is the absorbing sink state to which illegal actions and out-of-domain transitions route (§4.3). At $L = 3$ this evaluates to $\approx 5.73 \times 10^8$, the figure we work against in this paper.

Three benchmarks place this in context. The matching WeRLman fee-only MDP at the same horizon $L = 3$, obtained by removing the Eclipsed tag from the chain alphabet and dropping the b_e counter, has $\sim 10^7$ states under the same dense convention; the joint state space is therefore approximately $45\times$ larger at the same horizon, driven by the multiplicative compounding of the Eclipsed tag (factor $(3/2)^{2L}$) and the b_e counter (factor $L{+}1$). WeRLman's own headline security analysis runs at $L = 10$ on its sparse fee-only encoding, where the reachable subspace is $\sim 10^8$ states [11]; the corresponding dense Cartesian product at $L = 10$ is $\sim 5 \times 10^{17}$, which their sparse-MDP solver does not need to materialize. The joint kernel at $L = 10$ under our dense Cartesian encoding would reach $\sim 10^{22}$ states, well beyond the orchestrator and sparse-MDP budgets the deep-RL solver operates within, so we cannot match WeRLman's headline horizon at the joint scale (we mitigate the truncation with an $L = 5$ horizon-robustness check; §5.5, §6.4). Both joint figures are two orders of magnitude beyond the $\mathcal{O}(10^3)$ classical-solver scale at which Gervais's binary-search-on-linearization is tractable [8]. The reachable subspace is strictly smaller than the Cartesian product, but the latter is the quantity that matters for sparse-matrix indexing.

4.3 Action space

The action space is

$$\mathcal{A} = \{\text{Illegal}\} \cup \big(\{\text{Adopt}, \text{Reveal}\} \times [0, L]\big) \cup \big(\{\text{Mine}\} \times \{\text{With}, \text{NoTx}, \text{Eclipsed}\}\big).$$

Adopt and Reveal take an integer parameter $n \in [0, L]$ giving the number of blocks to adopt or reveal; Mine takes a transaction tag from the alphabet for the mined block, with Eclipsed aliasing NoTx in the MINE transitions. The Illegal sentinel routes structurally invalid actions (for example Adopt n with $l_h < n$) to the sink state with a large negative reward; the sentinel itself is excluded from legal-action enumeration during training.

4.4 Transitions

MINE under fork $\neq$ Active.: The block is won by one of three parties. With probability α the strategic miner finds the block and extends $\mathbf{c}_a$. With probability ω an eclipse victim finds the block, which also extends $\mathbf{c}_a$ but is tagged Eclipsed. With the remaining probability $(1 - \alpha - \omega)$ a non-eclipsed honest miner finds the block; a fraction o of these are stale (state unchanged, no chain growth) and the rest extend $\mathbf{c}_h$. Each strategic or non-eclipsed honest block that extends the longest chain additionally triggers a whale-transaction arrival with probability δ, incrementing pool. At $\delta = 0$ the whale-transaction sub-branches collapse and the kernel reduces to Gervais's; at $F = \delta = 0$ the full MDP (kernel and reward) coincides with Gervais's, formalized as Lemma 2.

MINE under fork = Active.: The non-eclipsed honest branch splits in two. With probability γ an honest miner that received the strategic miner's block first extends $\mathbf{c}_a$, resolving the fork in the strategic miner's favor. With probability $1 - \gamma$ a non-rushable honest miner extends $\mathbf{c}_h$ instead. The stale branch is as above.

TABLE 3: Components of the state $s \in \mathcal{S}$. "Origin" indicates whether the component is inherited from Gervais [8] (G) or from WeRLman [11] (W); the joint composition is new but no individual component is. The Transaction alphabet $\{\mathrm{NoTx}, \mathrm{With}, \mathrm{Eclipsed}\}$ is shared between $\mathbf{c}_a$ and $\mathbf{c}_h$ in the state encoding; $\mathrm{Eclipsed}$ on $\mathbf{c}_h$ is index-domain-only and unreachable under the transition kernel.

Symbol	Description	Domain	Origin
$\mathbf{c}_a$	private chain (block, tx pairs)	$(\{\mathrm{NoBlock}, \mathrm{Exists}\} \times \{\mathrm{NoTx}, \mathrm{With}, \mathrm{Eclipsed}\})^L$	W
$\mathbf{c}_h$	public chain (block, tx pairs)	$(\{\mathrm{NoBlock}, \mathrm{Exists}\} \times \{\mathrm{NoTx}, \mathrm{With}, \mathrm{Eclipsed}\})^L$	W
fork	race-state token	$\{\mathrm{Irrelevant}, \mathrm{Relevant}, \mathrm{Active}\}$	G
pool	unconfirmed whale transactions	$\{0, 1, \ldots, R\}$	W
l_a	cached private chain length	$\{0, 1, \ldots, L\}$	G
l_h	cached public chain length	$\{0, 1, \ldots, L\}$	G
t_a	cached whale txs on private chain	$\{0, 1, \ldots, R\}$	W
t_h	cached whale txs on public chain	$\{0, 1, \ldots, R\}$	W
b_e	eclipsed blocks on private chain	$\{0, 1, \ldots, L\}$	G

REVEAL n.: The strategic miner publishes the first n blocks of $\mathbf{c}_a$. Let e_n count the eclipsed-tagged blocks in the published prefix and w_n count the whale-tagged blocks:

$$e_n := \big|\{\, i \le n : \mathbf{c}_a[i].\mathsf{Transaction} = \mathrm{Eclipsed} \,\}\big|,$$
$$w_n := \big|\{\, i \le n : \mathbf{c}_a[i].\mathsf{Transaction} = \mathrm{With} \,\}\big|.$$

If $n > l_h$ the strategic chain wins outright, and the strategic miner earns

$$R_{\mathrm{Reveal}}(n) = \big(n - e_n + F \cdot w_n\big) \cdot R_{\mathrm{block}}, \tag{1}$$

where $R_{\mathrm{block}} = 1$ is the per-block reward unit (the revenue-ratio objective is scale-invariant, so the unit choice is immaterial), $\mathbf{c}_h$ is discarded, and $\mathbf{c}_a$ is shifted forward by n. The $-e_n$ term is the Gervais *eclipse-block subtraction*: eclipsed blocks add to b_e but pay no reward to the strategic miner. The $+F\,w_n$ term is the WeRLman *whale-fee inclusion*.

Equation (1) couples the eclipse-block subtraction and the whale-fee inclusion in a single n-dependent expression; it reduces to its single-channel form when either parameter vanishes ($\omega = 0$ or $F = 0$).

Active fork tie-breaking. If $n = l_h$, $n \le l_a$, and $\mathrm{fork} = \mathrm{Relevant}$, the Reveal declares an active fork ($\mathrm{fork} \to \mathrm{Active}$) instead of resolving immediately; reward accrues only upon resolution in a subsequent MINE action. This is the Match action of the Sapirshtein–Gervais lineage [5], [8]: when $l_a > l_h$, the strategic miner may choose to match ($n = l_h$, declaring an active fork) rather than override ($n > l_h$, winning outright).

ADOPT n.: The strategic miner abandons $\mathbf{c}_a$ and adopts the first n blocks of the public chain. The strategic miner earns no reward ($R_{\mathrm{Adopt}} = 0$); eclipse blocks in the discarded strategic prefix are written off (subtracted from b_e), and pool, t_a, t_h are updated to reflect the adopted blocks' whale content.

The reward function is piecewise linear in n, the whale-transaction count w_n, and the fee multiplier F. Transition probabilities sum to one at every non-final state, with all out-of-domain transitions routed to the absorbing sink state of §4.2.

4.5 Objective

Following WeRLman [11] and Bar-Zur et al. [21], the strategic miner's objective is the asymptotic reward-per-difficulty ratio

$$\mathrm{Rev}^* = \lim_{T \to \infty} \frac{\mathbb{E}\big[\sum_{t=1}^{T} R_t\big]}{\mathbb{E}\big[\sum_{t=1}^{T} A_t\big]},$$

where R_t is the per-step reward and A_t is the per-step difficulty contribution. We transform this ratio objective into a stochastic-shortest-path MDP via the PTO transform of Bar-Zur et al. [21], with expected horizon $H = 10^4$; §5 gives the transform in full.

4.6 Asymmetries inherited from Gervais

Two asymmetries in the transition kernel are inherited verbatim from Gervais [8]. We state them explicitly because they shape both the reward structure of Equation (1) and the limiting behavior at $\omega = 0$ or $o = 0$.

Eclipse blocks contribute chain length but not coinbase reward.: A block tagged $\mathrm{Eclipsed}$ advances the strategic chain (incrementing l_a and b_e), but its coinbase flows to the eclipse victim rather than the strategic miner. This is the source of the $-e_n$ term in Equation (1) and realizes the eclipse-as-redirection semantics of Gervais [8, §3]: eclipsed blocks land on the strategic chain, but their coinbase flows to the eclipse victim.

Stale rate applies only to honest discoveries.: A non-eclipsed honest block discovery is lost (stale) with probability o; strategic blocks (which the attacker propagates within its own infrastructure) and eclipsed-victim blocks (routed directly to the strategic chain by assumption) are never stale [8, §3].

Relaxations.: Relaxing either asymmetry is straightforward but not explored here: crediting eclipsed blocks to the strategic miner removes the eclipse-redirection asymmetry, and charging strategic blocks a small stale rate removes the strategic-no-stale asymmetry. The strategic-stale case is the more informative of the two: a small symmetric stale rate o' on attacker blocks would shrink the cross-term magnitude controllably as o' rises, providing a parametric handle on how the sub-additive plateau weakens; we flag it as a natural target for future work.

5 METHOD

The joint MDP's 5.73×10^8-state space at $L = 3$ is beyond the reach of exact policy iteration, so we adopt the deep-RL solver introduced by WeRLman [11]. Our method comprises two inherited components, used here without modification.

Ratio objective via PTO.: The asymptotic reward-per-difficulty ratio Rev^* of §4.5 does not admit a per-step Bellman update. We follow Bar-Zur et al. [21] and apply the Probabilistic Termination Optimization (PTO) transform, which converts the ratio into an expected-reward objective by introducing a geometric per-step termination: every transition probability is multiplied by the survival factor

$$p_{\mathrm{survive}}(s, a, s') = (1 - 1/H)^{A(s,a,s')},$$

where H is the expected-horizon parameter. Under this transform, the optimal policy of the resulting stochastic shortest-path MDP coincides, in the $H \to \infty$ limit, with the optimal policy of the original ratio MDP. We use $H = 10^4$ throughout, matching the WeRLman setting.

MCTS-DQN solver.: The transformed MDP is solved by WeRLman's MCTS-DQN trainer [18], [19], [20] (an AlphaGo-Zero-style hybrid): a two-headed MLP emits $(Q(s,a), \pi(s,a))$ for each legal action; PUCT-guided Monte Carlo tree search refines Q and π at every decision point; the Q-head is trained on DQN targets at the search root, and the policy-prior head on the cross-entropy of the MCTS visit distribution. Two WeRLman-specific stabilizations are essential near the profitability threshold, where absolute revenues approach zero. The *base-value approximation* models $Q(s,a) - V_{\mathrm{base}}(s)$ instead of $Q(s,a)$, keeping training targets centered as the revenue scale shrinks; *target-value normalization* recenters mini-batch targets to zero mean before the MSE loss. We inherit both, along with WeRLman's published hyper-parameters; the joint kernel introduces no new training-time choices.

Reproducibility.: The full training configuration, MDP code, sweep drivers, and per-run logs for all ∼1,500 trained policies are pinned to a frozen commit at https://github.com/fghasemi/composed-selfish-mining (artifact DOI to be assigned upon acceptance). All quantitative results in this paper are reproducible against that commit, with fixed random seeds and a pinned conda environment yielding bit-identical outputs across reruns (Appendix B). Hyperparameter values, the feature encoding, per-run wall-clock figures, and a bit-identical spot-check across the four experimental tracks (main sweep, sub-floor, $L = 5$ diagonal, γ sensitivity; Table 9) are cataloged in the same appendix.

5.1 Setup

Unless noted otherwise, all experiments use $\gamma = 0.5$, $L = R = 3$, $F = 10$, $\delta = 0.01$, with 10 seeds per cell, 5,000 training epochs per run, and PTO expected horizon $H = 10^4$.

We use $L = R = 3$, the largest horizon tractable at the joint state-space scale. At $L = 3$ the joint MDP has 5.73×10^8 states, two orders of magnitude beyond classical-solver scale; $L = 5$ would push this to $\sim 8.5 \times 10^{12}$, and WeRLman's $L = 10$ operates on a smaller fee-only MDP. §6.4 discusses the truncation in detail; §5.5 reports $L = 5$ horizon-robustness checks on a sparse subset of cells.

We fix $(F, \delta) = (10, 0.01)$ throughout: fee revenue averages $\sim 10\%$ of the block subsidy per block, representative of forward-looking Bitcoin (the present-day fee-to-subsidy ratio is closer to ∼3–5%). The regime is motivated in §3.

Rushing factor.: The rushing factor $\gamma = 0.5$ follows the canonical literature [5], [8]; Appendix F reports a $\gamma \in \{0, 1\}$ spot-check.

Convergence criterion.: A run is *converged* if it completes at least 4,500 of the 5,000 training epochs. Figure 2 shows mean simulated policy revenue ±1 standard deviation across seeds at every periodic evaluation, on two representative cells: the additive-region cell $\omega = o = 0.05$ and the combined-high cell $\omega = o = 0.25$.

The seed band tightens to $\sigma \lesssim 0.02$ well before the cut and remains stable through to the final epoch. The cut leaves $n = 10$ converged seeds at every cell of the $\alpha = 0.35$ heatmap; per-cell sample sizes are detailed in Appendix A.

Statistical noise floor.: Final-epoch cross-seed standard deviation along the heatmap diagonal ranges from $\sigma = 0.011$ at the additive-region cell ($\omega = o = 0.05$) to $\sigma = 0.018$ at the combined-high cell ($\omega = o = 0.25$), giving per-cell standard errors $\sigma/\sqrt{10} \in [0.0035, 0.0057]$. All six cells of the headline sub-additive plateau clear $|z| \geq 2.58$ ($p < 0.01$, two-tailed); the four cells of the 2×2 high-intensity corner ($\omega, o \in \{0.20, 0.25\}$) clear $|z| \geq 3.6$.

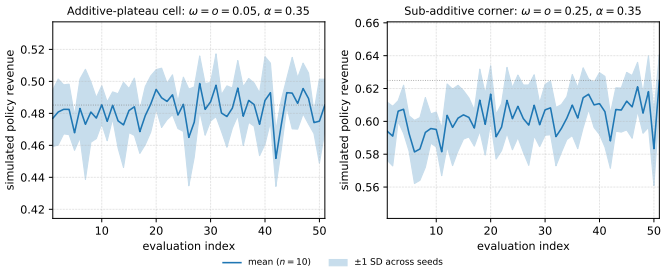


Fig. 2: Convergence diagnostics for two representative cells of the $\alpha = 0.35$ heatmap: the additive-region cell $\omega = o = 0.05$ (left) and the combined-high cell $\omega = o = 0.25$ (right). Solid line: mean simulated policy revenue across 10 converged seeds per panel. Shaded band: ±1 standard deviation across seeds at each evaluation.

5.2 Reduction anchors

The joint MDP must nest both prior works at the appropriate parameter limits: $\omega = o = 0$ should reproduce WeRLman, $F = 0$ should reproduce Gervais, and the no-channel no-fee limit should reproduce Sapirshtein. We verify all three, with progressively stronger evidence: a numerical match against the WeRLman trainer, a numerical match against Gervais's published table, and an exact analytical anchor against Sapirshtein's $\mathrm{BitcoinModel}$.

WeRLman reduction.: Lemma 2 establishes the structural reduction at the MDP level: at $\omega = o = 0$ the joint kernel coincides with WeRLman's fee-only kernel. Empirically, at $\alpha = 0.35$ and $F = 10$, the eclipse-stalerate code path with both network-layer parameters zeroed produces combined revenue 0.4446 ± 0.0153 over $n = 10$ seeds; the F=10 baseline computed directly from WeRLman's fee model at the same parameters is 0.4419 ± 0.018 ($n = 10$). The two means differ by $\Delta = +0.0027$, approximately 0.6 pooled standard errors—well within seed-level RL noise.

Gervais reduction (numerical).: With $F = 0$ and $R = 0$ (the no-whale-transaction limit), the joint MDP at

$(\omega, o) = (0.10, 0.10)$ and $\alpha = 0.35$ produces combined revenue 0.4667 ± 0.0115 ($n = 10$). This falls within the range predicted by Gervais et al.'s [8] Table II extrapolated to $\alpha = 0.35$ from their $\alpha \in \{0.30, 0.40\}$ entries, so the reduction is consistent with their published numerical results to the resolution their table allows.

Sapirshtein reduction (non-RL).: At the no-channel, no-fee limit the joint MDP nests BitcoinModel [5], which is small enough for exact policy iteration. BitcoinModel at $\alpha = 0.35$, $\gamma = 0.5$, $L = 3$ has 49 states, and policy iteration on this MDP converges to $\rho_{\mathrm{PTO}} = 0.3842$ (Appendix D). Our no-fee RL baseline at $F = 0$ recovers this value within seed-level RL noise (Appendix E). This is the strongest of the three reduction anchors: there is no RL on the analytical side, and the empirical side recovers the analytical optimum.

Why direct policy iteration on the joint MDP is infeasible.: The state-space scale rules out applying policy iteration directly to the joint MDP, even at the smallest non-trivial truncation. At $L = 2$ the state space already contains ~ 2.8M states; the dense stochastic-matrix validity check that `mdptoolbox` performs before policy iteration allocates $|\mathcal{S}|^2 \cdot 8$ bytes ≈ 58 TiB per action. At $L = 3$ the same allocation reaches ~ 2.6 exabytes. Skipping the dense check fails differently: the sparse kernel produces intermediate states that the index mapping rejects, such as pool $= -1$ arising from kernel arithmetic on states that are themselves unreachable. This is why upstream WeRLman uses RL exclusively for fee-bearing models.

5.3 Profitability threshold across attack capabilities

The profitability threshold α^* is the smallest hashrate fraction at which the trained policy's mean revenue strictly exceeds the honest baseline:

$$\alpha^* \equiv \min\{\alpha : \mathbb{E}[\mathrm{rev}_\pi(\alpha)] > \alpha \cdot (1 + F \cdot \delta)\}.$$

At $F = 10$ and $\delta = 0.01$ the honest baseline reduces to 1.1α. For each (ω, o) cell we sweep $\alpha \in \{0.025, 0.05, 0.075, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40\}$ at 10 seeds per value, linearly interpolate between adjacent samples to fix the crossing, and bootstrap over seeds (1,000 samples) for a 95% confidence interval. The full sweep covers approximately 960 training runs. Figure 3 plots revenue against α for each condition; Table 4 tabulates the crossings.

The F=10 baseline at $\alpha^* = 13.2\%$ is the anchor against which we measure the joint-channel collapse. At $F = 0$ the same pipeline recovers the no-fee Sapirshtein limit ($\rho^* \approx 0.38$ at $L = 3$; Appendix D), confirming the 13.2% baseline is not a fee-model artifact.

A single attack channel collapses α^* by a factor of 2.7–5.1. The eclipse-only condition ($\omega = 0.10$) gives $\alpha^* = 2.6\%$, with a confidence interval that does not overlap the F=10 baseline CI. The stale-only condition ($o = 0.10$) gives $\alpha^* = 4.8\%$.

The combined-channel cell collapses the threshold further. At $\omega = o = 0.10$ and $L = 5$ (the horizon-robustness anchor used for the abstract's headline), $\alpha^* = 0.9\%$—a $\sim 15\times$ collapse from the F=10 baseline. At the more conservative $L = 3$ truncation, $\alpha^* = 1.4\%$, a $\sim 9\times$ collapse. The collapse therefore strengthens at the longer horizon, not weakens; the trained policy at $L = 5$ is more profitable at the same intensity, consistent with the horizon-stability pattern of the Sapirshtein anchor at $F = 0$: ρ^* at $\alpha = 0.35$ rises from 0.3842 at $L = 3$ to 0.4300 at $L = 20$ (Sapirshtein's reported

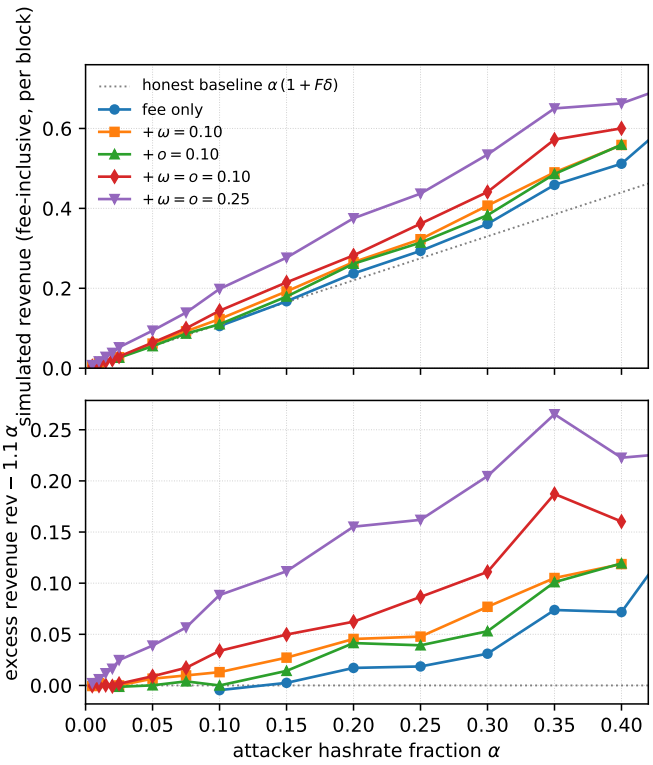


Fig. 3: Strategic revenue versus hashrate fraction α for each attack condition, at $\gamma = 0.5$, $F = 10$, $\delta = 0.01$, $L = R = 3$, ten seeds per α. **Top**: total simulated revenue (fee-inclusive, per block); the dotted line is the honest baseline $\alpha(1 + F \cdot \delta) = 1.1\alpha$. **Bottom**: excess revenue $\mathrm{rev} - 1.1\alpha$, isolating the deviation from honest. A curve's crossing of zero (bottom) gives the profitability threshold α^*. The F=10 baseline (no network channels) crosses at $\alpha^* = 13.2\%$; adding either single network-layer channel collapses the crossing to $\leq 5\%$; with both channels admitted at 10% intensity the curve already exceeds the baseline at the smallest sampled $\alpha = 2.5\%$.

TABLE 4: Profitability threshold α^* per attack condition at $\gamma = 0.5$, $F = 10$, $\delta = 0.01$, $L = R = 3$. CI from 1,000-sample bootstrap over $n = 10$ seeds. Combined-channel rows are extended to $\alpha \in \{0.005, 0.01, 0.015, 0.02, 0.025\}$. See notes below for grid-resolution caveats.

	$L = 3$		$L = 5$	
condition	α^*	95% CI	α^*	95% CI
F=10, no channels (ω=0, o=0)	0.132	[0.114, 0.159]	—	—
eclipse-only (ω=0.10, o=0)	0.026	[0.025, 0.031]	0.006	[0.005, 0.007]
stale-only (ω=0, o=0.10)	0.048	[0.033, 0.059]	—	—
combined low (ω=o=0.10)	0.014†	[0.005, 0.025]	0.009	[0.005, 0.026]
combined-high (ω=o=0.25)	< 0.005*	—	—	—

†The bootstrap resolution at very low thresholds is bounded by the α-grid spacing of 0.005, so the (0.10, 0.10) point estimate is best read as "1.4± a grid step." The lower CI edge sits on the $\alpha = 0.005$ sub-floor and is conservative. *The (0.25, 0.25) entry is a sweep-design upper bound: the curve exceeds the honest baseline at $\alpha = 0.005$, so the bootstrap CI collapses to a single point.

values [5]; App. D). The eclipse-only threshold shows the same direction more sharply: $L = 3$ gives $\alpha^* = 2.6\%$ and $L = 5$ gives $\alpha^* = 0.6\%$. Our $L = 3$ headlines are therefore conservative bounds on the joint optimum, not unstable artifacts of a tight truncation. At the model-extreme corner $\omega = o = 0.25$ the $L = 3$ threshold falls below the $\alpha = 0.005$ sub-floor of the extended sweep—a $\geq 26\times$ bound by sweep design, with the exact value below the floor.

The two combined-channel cells therefore separate in threshold terms (1.4% vs. $\leq 0.5\%$ at $L = 3$; 0.9% vs. a value below the $\alpha = 0.005$ floor at $L = 5$). The intensity-dependence claim does not rely on the threshold sweep alone: §5.4 establishes it independently at fixed $\alpha = 0.35$, where the $(0.25, 0.25)$ revenue interaction is -8.5% at $z = -4.62$, statistically separated from the near-zero interaction at $(0.10, 0.10)$.

5.4 The interaction surface at $\alpha = 0.35$

The threshold sweep collapses each (ω, o) cell to a single break-even hashrate. To examine how the two channels compose at a fixed attacker hashrate, we fix $\alpha = 0.35$ and sweep (ω, o) on a 5×5 grid with values in $\{0.05, 0.10, 0.15, 0.20, 0.25\}$ at $F = 10$, ten seeds per cell. Matched single-channel baselines (ω-only and o-only) and the fee-only baseline at the same α and $F = 10$ let us define the per-cell interaction

$$\begin{aligned}&\text{interaction}(\omega, o)\\ &\quad = \text{rev}(\omega, o) - \left[\text{rev}_{\text{fee}} + \Delta_{\text{eclipse}}(\omega) + \Delta_{\text{stale}}(o)\right],\end{aligned}$$

the difference between the joint-channel revenue and the additive prediction from the matched single-channel baselines. Figure 4 plots the interaction surface as a percentage of combined revenue; the underlying tables (combined revenue, additive prediction, interaction with per-cell z-score) are in Appendix A.

The heatmap separates into three regions: an additive region, a transition ridge, and a sub-additive plateau.

Additive region (17 cells).: Seventeen cells fail to clear the $|z| \geq 2.58$ significance threshold and are consistent with strict additivity at the available sample size. These split into two sub-blocks. The lower-left 3×3 block ($\omega, o \leq 0.15$, nine cells) shows interactions within $\pm 3\%$ of zero, mixed in sign, with maximum magnitude -2.7% at $(0.10, 0.05)$. The remaining eight off-axis cells—those with one channel at low intensity (≤ 0.10) and the other at high intensity (≥ 0.20), or vice versa—show weakly sub-additive estimates between -1.9% and -4.2%, none of them statistically separated from zero.

Sub-additive plateau (six cells).: At high intensity, six cells form a sub-additive plateau with interaction $\leq -5\%$, all clearing $|z| \geq 2.58$: the upper-right 2×2 block $\{0.20, 0.25\} \times \{0.20, 0.25\}$, plus the two cells $(0.15, 0.20)$ and $(0.15, 0.25)$ on the row immediately below. The peak interaction is -8.9% at $(0.25, 0.20)$ with $z = -4.40$; the most statistically separated cell is $(0.25, 0.25)$ at -8.5% with $z = -4.62$. The 2×2 corner alone clears $|z| \geq 3.6$ in magnitude across all four cells.

Transition ridge (two cells).: Between the additive and plateau regions, the cells $(0.20, 0.15)$ and $(0.25, 0.15)$ show interactions of -3.5% and -4.6% with $|z|$ of 2.03 and 2.44, respectively—on or just below the significance threshold. Sub-additivity therefore builds up through this band; there is no sharp boundary.

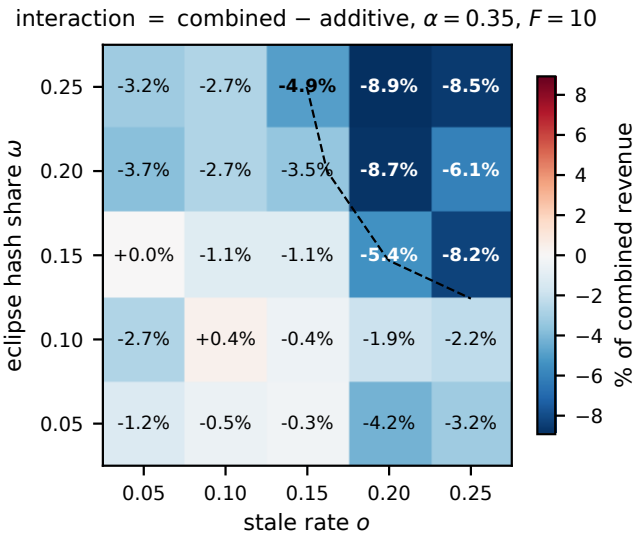


Fig. 4: Channel interaction at $\alpha = 0.35$, $\gamma = 0.5$, $F = 10$, $\delta = 0.01$, $L = R = 3$, ten seeds per cell, expressed as a percentage of combined revenue. Bold-font cells clear the $|z| \geq 2.58$ (99% two-sided Z-test) threshold; the dashed contour marks the $|z| = 2.58$ boundary. Peak interaction is -8.9% at $(\omega, o) = (0.25, 0.20)$ with $z = -4.40$.

§6.1 returns to the high-intensity plateau and accounts for the $-\omega\, o$ scaling analytically.

5.5 Generalization across α and L

Across α.: The $\alpha = 0.35$ heatmap reports the interaction surface at one hashrate. Two checks generalize the joint-channel finding to other α. First, repeating the four-condition ablation at $\alpha = 0.15$ with $(\omega, o, F) = (0.10, 0.10, 10)$ gives interaction $+4.5\%$ ($z = +1.86$, $|z| < 2.58$), consistent with the additive region observed at low intensities in the $\alpha = 0.35$ heatmap. Second, the threshold sweep of §5.3 covers nine values of α between 0.025 and 0.40: every condition with non-trivial network-layer help ($\omega \geq 0.10$ or $o \geq 0.10$) crosses the honest baseline by $\alpha \leq 0.05$, and the combined-channel threshold sits at 1.4% or below at every intensity tested. The joint-channel advantage is not a single-α artifact.

Across F.: The cross-term factor $(1 - \alpha - \omega)(1 - o)$ in Lemma 1 does not depend on F, so the qualitative sub-additivity finding survives any choice of F. The collapse *magnitude* does depend on F: at $F = 0$ the fee channel vanishes and α^* approaches the Sapirshtein–Zohar long-horizon limit of $\sim 25\%$ (Appendix D reports the corresponding ρ^* values at $L \in \{3, 20\}$); at intermediate F the threshold interpolates between that and the 13.2% baseline at $F = 10$. We do not sweep F at finer resolution. The endpoints bracket the regime relevant to deployed and forward-looking Bitcoin (§5.1), and the F-independence of the cross-term factor makes the sub-additivity claim robust to interpolation.

Across L.: We probe horizon sensitivity along three axes: revenue magnitudes at fixed cells, the interaction structure on a diagonal cut, and the threshold itself on a restricted sweep. All three point in the same direction—the joint-channel finding strengthens, not weakens, at the longer horizon.

We spot-check the additive-region cell ($\omega = o = 0.10$) and the combined-high cell ($\omega = o = 0.25$) at $L = 5$, alongside an $L = 5$ fee-only baseline at $\alpha = 0.35$. Combined revenue at $(0.10, 0.10)$ rises from 0.5362 at $L = 3$ to 0.6084 ($\sigma = 0.017$) at $L = 5$ ($n = 10$, $\Delta = 0.0722$, $z = 9.8$); at $(0.25, 0.25)$ it rises from 0.6255 to 0.6778 ($\sigma = 0.022$) at $L = 5$ ($n = 10$, $\Delta = 0.0523$, $z = 6.3$). Both increases match the standard expectation: a longer fork horizon lets the strategic miner withhold longer and earn more, so $L = 3$ is a conservative lower bound on the revenue magnitudes we report.

Whether the channel-channel *interaction* survives the longer horizon is a separate question from the revenue magnitudes. To answer it we ran a three-cell diagonal cut at $(\omega, o) \in \{(0.05, 0.05), (0.15, 0.15), (0.25, 0.25)\}$ at $\alpha = 0.35$ and $L = 5$ with $n = 10$ seeds for the combined cells. The matched single-channel ablations needed to compute the additive prediction were budgeted at $n = 3$ each—a deliberate compute trade-off given the $\sim 8.5 \times 10^{12}$-state $L = 5$ kernel and the per-run cost it implies. Table 5 reports the result.

TABLE 5: $L = 5$ diagonal interaction at $\alpha = 0.35$, $F = 10$. Combined cells at $n = 10$ ($n = 9$ at the high corner, one seed excluded by the convergence cut). Single-channel ablations underlying the additive prediction are at $n = 3$ each and are the binding source of variance in the interaction SE at the lower-intensity cells.

(ω, o)	combined	additive	interaction $\pm$ SE	z
$(0.05, 0.05)$	0.542	0.525	$+0.017 \pm 0.011$	$+1.5$
$(0.15, 0.15)$	0.646	0.680	-0.034 ± 0.022	-1.5
$(0.25, 0.25)$	0.678	0.770	-0.092 ± 0.018	-5.1

The combined-high cell shows interaction -0.092 at $|z| = 5.1$, larger in absolute magnitude than the $L = 3$ value of -0.085 at the same cell. This cell is the test of the cross-term plateau at the longer horizon. The SE at $(0.25, 0.25)$ is dominated by the $n = 10$ combined-cell variance rather than by the ablations, so deeper ablations would not materially change the $|z|$. The cross-term plateau is therefore preserved and slightly strengthened at $L = 5$.

The two lower-intensity cells, by contrast, sit on the ablation-noise floor: $|z| = 1.5$ for both, with SEs dominated by the $n = 3$ single-channel runs. We read these cells as a *null result* at the available sample size—neither evidence for additive behavior nor evidence against it. Pinning down the lower-intensity behavior at $L = 5$ requires deeper ablations and is left to §6.4. The asymmetric statistical power across the three cells is the expected behavior of a fixed compute budget allocated to the cell where the cross-term effect is largest.

The diagonal cut above checks the cross-term structure at fixed α. The complementary question is whether the threshold α^* itself shifts at $L = 5$. We sweep $\alpha \in \{0.005, 0.01, 0.025, 0.05, 0.10, 0.15\}$ at $L = 5$ for the two threshold-driving rows of Table 4—eclipse-only $\omega = 0.10$ and combined-low $\omega = o = 0.10$—at $n = 10$ seeds per α, matching the $L = 3$ headline budget. Eclipse-only at $L = 5$ gives $\alpha^* = 0.6\%$ $[0.5\%, 0.7\%]$ versus 2.6% $[2.5\%, 3.1\%]$ at $L = 3$. Combined-low at $L = 5$ gives $\alpha^* = 0.9\%$ $[0.5\%, 2.6\%]$ versus 1.4% $[0.5\%, 2.5\%]$ at $L = 3$. Both thresholds shift *down* at the longer horizon by a factor of 1.5 (combined-low) to 4 (eclipse-only): the strategic miner gains more profitability headroom with more lookahead, and the $L = 3$ headline numbers are conservative bounds on the joint optimum rather than tight estimates. This direction matches the Sapirshtein anchor at $F = 0$ (Appendix D), where ρ^* at $\alpha = 0.35$ rises from 0.3842 at $L = 3$ to 0.4300 at $L = 20$.

Our $L = 5$ collapse factor compares the joint threshold at $L = 5$ against the fee-only baseline at $L = 3$. This is conservative: a longer fork horizon would also lower the fee-only threshold below 13.2%, so the actual $L = 5$ collapse factor against a matched $L = 5$ fee baseline is bounded above by the $13.2/0.9 \approx 15\times$ figure we report. Even at this conservative anchor, the collapse from 13.2% to 0.9% at the deployment-relevant combined-low cell remains the headline finding of this paper. The $L = 10$ regime that WeRLman uses for its security analysis is beyond what we can reach at the joint-MDP scale ($\sim 10^{22}$ states; §6.4); the direction of the $L = 3 \to 5$ shift suggests the extrapolation from $L = 5$ to $L = 10$ amounts to at most a small further drop in α^*.

6 Discussion

6.1 Mechanistic reading of the sub-additive plateau

The regime change observed near $(\omega, o) \approx (0.20, 0.20)$ admits a compact analytical explanation: both channels act on the honest-mining rate, but through different multiplicative factors. We make this precise as a lemma about the joint honest-chain-growth probability.

Lemma 1 (Multiplicative cross-term in honest-chain growth)**.** *Let $p_{\mathrm{H}}(\alpha, \omega, o)$ be the per-step probability that the public chain grows by one block in a non-active-fork* Mine *transition of the joint MDP M (§4). Then, for any (α, ω, o) in the feasible parameter set,*

$$p_{\mathrm{H}}(\alpha, \omega, o) \;=\; (1 - \alpha - \omega)(1 - o)$$
$$=\; 1 - \alpha - \omega - o + \alpha o + \omega o.$$

The deviation from the linear additive approximation $1 - \alpha - \omega - o$ is exactly $\alpha o + \omega o$. The first term, αo, is the attacker–stale interaction and is absorbed into any single-channel o-only baseline at fixed α. The second term, ωo, is the channel–channel cross-term and is the source of the sub-additive structure observed in §5.4.

Proof. From a non-active-fork state, the Mine transition kernel resolves three sub-events in order: (i) which miner finds the block (attacker, eclipse-redirected, or honest), (ii) whether that block is stale (applicable only to honest blocks), and (iii) whether a whale transaction event is sampled. The kernel of the joint MDP resolves staleness *before* the whale-arrival sub-event, so the whale parameter $\tau \in \{0, \delta\}$ does not appear in the staleness factor.

A new public-chain block is appended exactly when an honest miner discovers a non-stale block. The probability of an honest discovery is $(1-\alpha-\omega)$, and conditional on discovery, the probability that the block is non-stale is $(1-o)$. These factors are independent by construction:

$$p_{\mathrm{H}} = \underbrace{(1-\alpha-\omega)}_{\text{honest discovery}} \cdot \underbrace{(1-o)}_{\text{not stale}}.$$

Expanding gives the claimed identity. □

The economic content is that eclipse and staleness compete for the same slice of honest mining productivity. Eclipse redirects a fraction ω of nominal honest hash power to the strategic chain; staleness discards a fraction o of what remains. Each channel attacks the residual after the other has acted, so their combined effect is multiplicative rather than additive. The strategic miner gains less from staleness in an already-eclipsed regime—and less from eclipse in an already-stale-heavy regime—because the same honest block cannot be both eclipsed away *and* discarded as stale.

From the kernel to revenue.: Lemma 1 states a per-step result. The mapping from per-step probabilities to long-run revenue under the optimal policy composes three transformations: the honest-baseline transformation, the discount-linearization, and the publish-k threshold choice. The kernel cross-term therefore does not translate directly into a revenue cross-term of the same size. The leading-order claim is that the ωo signature should appear in revenue too, modulo strategy-adaptation residuals. We verify this in two complementary ways.

First, under a fixed publish-1 strategy with no policy adaptation, the long-run revenue ρ_{pub1} carries a leading-order $-\omega o$ correction relative to the additive single-channel prediction. We verify this by direct simulation in Appendix G, which sweeps the 5×5 (ω, o) grid at $\alpha = 0.35$ under the publish-1 rule and confirms the ωo signature emerges before strategic adaptation enters. This corollary anchors the cross-term interpretation in a non-RL substrate.

Second, the trained-policy interaction at the combined-high cell agrees with the kernel-level prediction quantitatively. At $(\alpha, \omega, o) = (0.35, 0.25, 0.25)$, the kernel predicts a channel–channel cross-term of $\omega o = 0.0625$. The interaction measured in §5.4 is -8.5% of combined revenue 0.6255, an absolute value of 0.053. The kernel prediction 0.0625 and the trained-policy measurement 0.053 agree in sign and within 15% in magnitude. The residual $\sim 15\%$ gap is consistent with publish-k threshold adaptation, which the fixed-publish-1 corollary does not model: the trained policy can partially recover by shifting publish thresholds, reducing the realized sub-additivity below the kernel-level prediction.

Why the heatmap reads as additive at low intensity.: Lemma 1 also accounts for the location of the regime transition. Below $\omega, o \lesssim 0.15$ the predicted cross-term satisfies $\omega o \leq 0.0225$, comparable to the per-cell 2σ noise floor of ~ 0.02 on the interaction statistic: the sub-additivity is present in the model but statistically indistinguishable from zero at $n = 10$ seeds per cell. Above $\omega, o \approx 0.20$ the cross-term grows past $\omega o \geq 0.04$, comfortably above the noise floor, and the empirically observed $|z| \geq 2.58$ plateau emerges. The additive plateau in the heatmap is therefore not evidence of strict additivity, but of a small cross-term combined with finite sample size.

Reduction to Gervais.: The second analytical observation closes the loop with the closest prior art.

Lemma 2 (Reduction to Gervais 2016)**.** *At $F = 0$ and $\delta = 0$, the joint MDP M reduces, by component-ordering renaming, to the eclipse and stale-block MDP of Gervais et al. [8] §3—identical state space, transition kernel, and reward function up to that renaming.*

Proof sketch. Setting $\delta = 0$ forces the $\mathrm{Transaction}$ component of every state to NoTx and the per-chain transaction counters to zero, removing the fee channel entirely from both states and rewards. Setting $F = 0$ removes the fee multiplier from the reward function. The residual state $(\ell_a, \ell_h, \mathrm{fork}, b_e)$ matches Gervais's state up to component ordering, and our action set ADOPT/REVEAL/MINE matches their ADOPT/OVERRIDE/WAIT (OVERRIDE in their notation corresponds to publishing $\ell_h + 1$ blocks of the private chain) once MATCH is folded into REVEAL at $\ell_a = \ell_h$. Transition probabilities and rewards match Gervais's Table II by direct comparison of the MINE kernel, and the $L = 3$, $R = 0$ truncation matches their conventions. Appendix C catalogues the state-component, action, and transition correspondence; the public release implements the reduction as a unit test. □

The empirical reduction-anchor in §5.2 is consistent with Lemma 2: at the no-fee limit the joint MDP's combined revenue at $(\omega, o) = (0.10, 0.10)$ falls within the range predicted by Gervais's Table II at $\alpha = 0.35$, so the structural reduction holds empirically as well as formally.

6.2 Composability with petty compliance

Deep Bribe [12] adds a petty-compliant miner class to the Gervais ancestor. At the state-space level their extension—a fork-enum value `ACTIVE_PETTY` and a non-strategic-miner petty/honest-rushable partition—is syntactically orthogonal to our eclipse and stale-block extensions, and the two lines could in principle be stacked. The transition kernel of the fully composed model has not been derived: a petty-compliant miner whose tie-breaking depends on residual fees may interact with eclipse-victim block routing in non-obvious ways, and verifying probability conservation in the joint kernel is itself substantial work. We identify this as future work rather than a mechanical composition of the two extensions. The methodological pattern of the present paper—a joint MDP that nests two prior extensions, with reduction lemmas anchoring the relation to each ancestor—should generalize to a three-extension model that adds petty compliance, given the requisite kernel-derivation work.

6.3 Model property and deployment outlook

The cross-term result admits two kinds of statement: a model property of the joint MDP, and a deployment outlook that is informed by—but does not follow directly from—our empirical results.

Model property: the cross-term is structural.: Lemma 1 factorizes the joint honest-chain-growth probability as $(1-\alpha-\omega)(1-o)$. The multiplicative cross-term ωo is the deviation of the joint reduction in honest growth from a linear addition of single-channel reductions, and is strictly positive whenever both channels are non-zero. The empirical sub-additive plateau in §5.4 at $\omega, o \geq 0.20$ is the revenue-side counterpart. This is a property of the joint MINE kernel itself: any analysis that admits both channels at non-zero intensity must account for the ωo cross-term. Single-channel analyses miss the structure entirely—the cross-term emerges only when both channels are jointly modeled. The statement does not depend on the deployment-relevant size of ω or o.

Deployment outlook.: The model property suggests that eclipse-resistance and stale-rate mitigations should be evaluated jointly rather than as independent line items. Deploying compact block relay (BIP 152 [32]) to reduce o without hardening peer selection against eclipse-style isolation [7] leaves the ωo contribution intact at any non-trivial ω. The complementary scenario applies symmetrically: hardening peer selection without reducing o leaves the cross-term in place at any non-trivial o. Single-channel mitigations in either direction reduce attacker advantage along their own axis but do not address the joint contribution.

This outlook is informed by but does not follow directly from our results: the empirical sub-additive separation is observed at intensities ($\omega, o \geq 0.20$) that §3 flags as model-extreme, while the deployment-relevant region ($\omega, o \leq 0.15$) shows interactions indistinguishable from additive at our sample size, even though the cross-term is mathematically present. Translating the outlook into a deployable defender prescription would require calibrated (ω, o) estimates for production operators that we do not provide. A related direction raised by our results is whether the classical pool-concentration heuristic (no single pool above $\sim 25\%$ hashrate, derived from the original selfish-mining threshold [3]) should be revisited in light of the joint-channel collapse. This depends on the same calibrated deployment estimates; we discuss it as a natural extension in §6.4 rather than as a deployable result of the present paper.

6.4 Limitations and threats to validity

We collect the open methodological caveats of the present work in one place.

Fork-horizon truncation.: The $L = 3$ value used in the headline experiments is the largest tractable horizon at the joint state-space scale; WeRLman's $L = 10$ headline runs on its smaller fee-only MDP. Since selfish-mining strategies are known to depend on horizon truncation [5], [11], we provide $L = 5$ horizon-robustness checks on a diagonal cut and on the threshold-driving rows of Table 4 (§5.5). Both shift in the direction of the joint optimum, so the $L = 3$ numbers should be read as conservative bounds rather than calibrated estimates.

Headline parameters fixed.: All headline experiments fix $\gamma = 0.5$, $F = 10$, $\delta = 0.01$ (motivated in §3 and §5.1). The cross-term lemma is independent of fee parameters, so the qualitative sub-additivity finding survives any F; the threshold-collapse magnitude is F-dependent and shrinks at lower fees (§5.5). The γ-robustness claim rests on two complementary arguments: Lemma 1's γ-independent kernel derivation, and a magnitude-only check in Appendix F that reports headline-cell revenue magnitudes at $\gamma \in \{0.0, 0.5, 1.0\}$ and confirms magnitudes shift with γ as expected from the Sapirshtein literature. An empirical F- and γ-sensitivity sweep, including a full γ-interaction surface with matched single-channel ablations at each γ, is future work.

Cross-term plateau at $L = 5$.: The $L = 5$ diagonal cut (Table 5) supports the cross-term plateau at the combined-high cell ($|z| = 5.1$), but the two lower-intensity diagonal cells are ablation-noise-limited at $n = 3$ and read as null results rather than positive evidence of additive behavior at the longer horizon. A full 5×5 surface at $L = 5$ with deeper ablations is straightforward future work.

RL approximation versus provably optimal policy.: The MCTS-DQN solver returns an approximately optimal policy of the transformed MDP, not a certified optimum. At the joint state-space scale, exact policy iteration is infeasible (§5.2), so we cannot verify the trained policy against a ground-truth optimum directly. The $F = 0$ Sapirshtein anchor (Appendix E) verifies that the solver recovers the analytical optimum on a small-MDP case, but the joint kernel admits no analogous direct verification. We treat the reported revenues and thresholds as lower bounds on the optimum-policy values; a tighter optimum could only strengthen the headline collapse claim.

7 CONCLUSION

Our headline result is the *profitability-threshold collapse against a state-level joint adversary*. The adversary class is state-level rather than node-level: $\omega = 0.10$ requires per-block routing control over a 10% slice of total honest hash power, not eclipse of an individual node. The F=10 baseline of 13.2% drops to 2.6% under $\omega = 0.10$ alone, to 4.8% under $o = 0.10$ alone, and to 0.9% at $L = 5$ (1.4% at $L = 3$) when both channels are admitted at 10%. The collapse strengthens at the longer horizon ($\sim 15\times$ at $L = 5$, $\sim 9\times$ at $L = 3$), in line with the long-horizon trend of the Sapirshtein anchor at $F = 0$ where ρ^* rises with L, so the $L = 3$ headline is a conservative bound rather than a tight estimate.

The threshold collapse is the empirical observation we report at the deployment-relevant end of the parameter space ($\omega, o \leq 0.15$). At the model-extreme corner $\omega = o = 0.25$ the $L = 3$ threshold falls below the $\alpha = 0.005$ floor of our extended sweep; the model-extreme corner is reported for cross-term validation, with deployment-relevant claims confined to $\omega, o \leq 0.15$. The L dependence acknowledges that the $L = 10$ horizon used by WeRLman [11] as its own headline setting is not tractable at the joint state-space scale.

The collapse is supported by an analytical identity. Lemma 1 factorizes the joint honest-chain-growth probability as $(1-\alpha-\omega)(1-o)$ and isolates the multiplicative cross-term ωo as the deviation from the linear addition of single-channel reductions. The lemma is a property of the transition kernel, not of any specific solver run, and a fixed-publish-1 simulation without RL in the loop (Appendix G) verifies the ωo signature in revenue. The empirical sub-additive plateau is observed only at joint

intensities $\omega, o \geq 0.20$, where the cross-term grows to 5–9%, with all six plateau cells clearing $|z| \geq 2.58$ and the four high-intensity-corner cells clearing $|z| \geq 3.6$. In the deployment-relevant regime $\omega, o \leq 0.15$ the cross-term is mathematically present but small relative to per-cell sample noise, so the empirical heatmap is consistent with strict additivity at the sample size we ran. The defensive reading (§6.3)—that single-channel mitigations reduce attacker advantage along their own axis but do not address the joint contribution while either channel remains non-zero—rests on the lemma's model property rather than on the empirical plateau, which itself is a high-intensity phenomenon outside today's deployed-Bitcoin regime.

Lemma 2 and the empirical reduction anchors of §5.2 connect the joint MDP to Gervais et al. and to WeRLman at the respective parameter limits, and the analytical Sapirshtein anchor at $F = 0$ gives $\rho^* = 0.3842$ at $\alpha = 0.35$, $L = 3$, which the $F = 0$ trained policy recovers within ~ 0.4 pooled SE (Appendix E).

Together, these results characterize a class of joint network-economic adversary that prior single-channel analyses do not capture, and motivate joint mitigation as the deployment-level response. They occupy the cell that recent systematizations [16], [30] flag as open.

Appendix A
Full interaction surface at $\alpha = 0.35$

Tables 6, 7, and 8 report the full 5×5 grid at $\alpha = 0.35$, $\gamma = 0.5$, $F = 10$, $\delta = 0.01$, $L = R = 3$, with $n = 10$ seeds per cell. Three joint cells have $n = 9$ rather than $n = 10$ after the convergence cut of §5.1: $(\omega, o) \in \{(0.10, 0.20), (0.15, 0.20), (0.20, 0.10)\}$. The $o = 0.20$ single-channel (stale-only) baseline has $n = 7$ converged seeds. The released artifact includes the per-run logs and the aggregation driver from which all three tables are regenerated.

Appendix B
Artifact description

The released artifact at https://github.com/fghasemi/composed-selfish-mining contains the joint-MDP implementation, the sweep configurations corresponding to every experimental track, the aggregation and bootstrap drivers that produce each table from the per-run training logs, and the figure regeneration code. The configurations cover the 5×5 (ω, o) interaction surface at $\alpha = 0.35$ (Tables 6–8); the threshold sweep across α (Table 4) together with its sub-floor extension at $\alpha \leq 0.025$; the $L = 5$ diagonal cut and threshold check underlying §5.5; the $\gamma \in \{0, 1\}$ ablations of Appendix F; and the $F = 0$ Sapirshtein-anchor sanity baseline of Appendix E. The release also includes the full set of approximately 1,500 trained policies and their per-epoch logs, so every revenue number reported in the paper can be re-derived from the existing logs without retraining. The repository is pinned to a frozen commit so that a third party rerunning the artifact reaches the same MDP definition and trainer hyperparameters we used.

B.1 Reproducibility

We evaluated the determinism of the pipeline by re-running one cell from each of the four major experimental tracks and comparing the new final-policy revenue at seed 0 to the value originally logged at submission. The trainer is deterministic when the seed and conda environment are pinned: the WeRLman-inherited orchestrator serializes MCTS-prior queries so that the same seed produces the same training trajectory across reruns. The re-runs reproduce the originals bit-for-bit (Table 9), which is stronger than the within-seed-noise tolerance typically reported for deep-RL training pipelines.

Appendix C
State-action correspondence with Gervais (CCS 2016)

This appendix gives the per-state and per-action correspondence behind Lemma 2, between the no-fee limit of our joint MDP ($F = 0$, $\delta = 0$) and the Gervais et al. [8] eclipse and stale-block MDP. The reduction maps each Gervais state and action to a unique canonical state and action in our model — Tab. 10 gives the state-component map and Tab. 11 the action-set map; transition probabilities then match by direct comparison.

The transition probabilities match Gervais's Table II: the MINE kernel of §4.4 reduces to Gervais's MINE kernel at $\delta = 0$ because the whale-transaction sub-branches (δ and $1 - \delta$ at each block-finder) collapse to unity. REVEAL, ADOPT, and MATCH are then identical once Gervais's OVERRIDE/MATCH are folded into the parameterized $\mathrm{Reveal}\, n$ as in Table 11.

Appendix D
Sapirshtein analytical anchor at the no-channel limit

The fee-bearing MDPs in this paper cannot be exact-solved by policy iteration: the dense sparse-matrix representation exceeds memory budgets even at small horizons, and skipping the dense check fails on invalid intermediate states (§5.2). Upstream WeRLman [11] hits the same wall and uses RL exclusively for fee-bearing models.

We therefore anchor the joint MDP against $\mathrm{BitcoinModel}(\alpha, \gamma, L)$—the no-fee, no-eclipse, no-stale Sapirshtein MDP—which is exact-solvable: 49 states at $L = 3$, $\sim 1{,}300$ states at $L = 20$. By Lemma 2 and the reduction chain in §5.2, the joint MDP at $\omega = o = 0$ and $F = 0$ coincides with this substrate, so the no-fee RL baseline at $F = 0$ should converge to ρ_{PTO} within seed-level noise. If the trained policy converges to ρ_{PTO}, the structural reduction is verified empirically—with no RL on the analytical side (Appendix E).

The $L = 3$ truncation underestimates ρ^* at higher α: at $\alpha = 0.35$, ρ_{PTO} rises from 0.3842 at $L = 3$ to 0.4300 at $L = 20$. This is the conservative-truncation effect the body reports for the joint MDP: deeper horizons admit longer stubborn strategies that the $L = 3$ pipeline cannot represent. The structural reduction in Lemma 2 is independent of L—the joint kernel coincides with Gervais at any horizon—so

TABLE 6: Combined strategic-miner revenue at $\alpha = 0.35$, $\gamma = 0.5$, $F = 10$, $\delta = 0.01$, $L = R = 3$, up to 10 seeds per cell (non-converged seeds excluded). The F=10 baseline at $\omega = o = 0$ is 0.4419 ($n = 10$).

$\omega \setminus o$	0.05	0.10	0.15	0.20	0.25
0.05	0.4850	0.5071	0.5261	0.5432	0.5713
0.10	0.5022	0.5362	0.5498	0.5792	0.6004
0.15	0.5350	0.5475	0.5653	0.5785	0.5850
0.20	0.5407	0.5642	0.5774	0.5846	0.6215
0.25	0.5627	0.5833	0.5901	0.6020	0.6255

TABLE 7: Additive prediction $\mathrm{rev}_{\mathrm{fee}} + \Delta_{\mathrm{eclipse}}(\omega) + \Delta_{\mathrm{stale}}(o)$ at $\alpha = 0.35$, $F = 10$.

$\omega \setminus o$	0.05	0.10	0.15	0.20	0.25
0.05	0.4912	0.5097	0.5277	0.5660	0.5895
0.10	0.5156	0.5341	0.5521	0.5904	0.6138
0.15	0.5349	0.5534	0.5714	0.6097	0.6331
0.20	0.5609	0.5794	0.5974	0.6357	0.6591
0.25	0.5806	0.5991	0.6171	0.6554	0.6789

TABLE 8: Channel interaction (combined − additive, as percent of combined), with pooled-SE z-score, at $\alpha = 0.35$, $F = 10$. Cells with $|z| \geq 2.58$ clear the 99% two-sided Z-test significance threshold. Six cells (the 2×2 upper-right corner plus the two cells at $\omega = 0.15$, $o \in \{0.20, 0.25\}$) form the sub-additive plateau analyzed in §5.4.

$\omega \setminus o$	0.05	0.10	0.15	0.20	0.25
0.05	−1.3% (−0.51)	−0.5% (−0.23)	−0.3% (−0.13)	−4.2% (−1.53)	−3.2% (−1.46)
0.10	−2.7% (−1.25)	+0.4% (+0.19)	−0.4% (−0.19)	−1.9% (−0.93)	−2.2% (−1.17)
0.15	+0.0% (+0.01)	−1.1% (−0.60)	−1.1% (−0.59)	−5.4% (−2.69)	−8.2% (−4.08)
0.20	−3.7% (−2.15)	−2.7% (−1.52)	−3.5% (−2.03)	−8.7% (−4.32)	−6.1% (−3.62)
0.25	−3.2% (−1.58)	−2.7% (−1.33)	−4.6% (−2.44)	−8.9% (−4.40)	−8.5% (−4.62)

TABLE 9: Reproducibility spot-check across the four experimental tracks at seed 0 ($\alpha = 0.35$ where unspecified). *Originally reported*: final `Simulated Policy Revenue` logged at submission. *Re-run*: same metric under the pinned environment.

#	Track / cell	Original	Re-run	Δ
1	main sweep, ω=o=0.10, L=3	0.53784	0.53784	0
2	sub-floor, ω=o=0.10, α=0.005	0.00488	0.00488	0
3	L=5 diagonal, ω=o=0.15	0.63688	0.63688	0
4	γ=0, ω=o=0.10	0.48544	0.48544	0

TABLE 10: State-component correspondence at the no-fee limit. pool, t_a, and t_h are frozen at 0 when $\delta = 0$; the reachable subspace then projects onto Gervais's state space.

Gervais 2016, Table II	**This work, no-fee limit**
a (private length)	l_a, with $\mathbf{c}_a$ collapsed to all-NoTx at $\delta = 0$
h (public length)	l_h, similarly collapsed
fork	identical: $\{\mathrm{Irrelevant}, \mathrm{Relevant}, \mathrm{Active}\}$
b_e (eclipse count)	b_e (same symbol and semantics)

the values in Table 12 are the L-truncated optima of that shared kernel.

TABLE 11: Action-set correspondence.

Gervais 2016	**This work**
ADOPT	Adopt n with $n = l_h$
OVERRIDE	Reveal n with $n > l_h$
MATCH	Reveal n with $n = l_h$, declaring an Active fork
WAIT	Mine (NoTx)

TABLE 12: Sapirshtein analytical anchor: ρ_{PTO} on $\mathrm{BitcoinModel}(\alpha, \gamma = 0.5, L)$. The $L = 3$ values pair with the paper's main-results horizon; the $L = 20$ values are the long-horizon limits. Computed by the analytical-anchor driver released with the artifact.

α	L	state count	ρ_{PTO}
0.30	3	49	0.3123
0.35	3	49	0.3842
0.40	3	49	0.4593
0.30	20	1,324	0.3268
0.35	20	1,324	0.4300
0.40	20	1,324	0.5720

APPENDIX E
$F = 0$ SANITY BASELINE

To anchor our choice of $F = 10$ as the F=10 baseline, we report the trained-policy revenue at the no-fee limit ($F = 0$, $\delta = 0.01$, $\alpha = 0.35$, $\gamma = 0.5$, $L = 3$), with 10 seeds per cell.

The trained-policy revenue at this point is $\rho_{\mathrm{RL}}^{F=0}(\alpha = 0.35) = 0.3814 \pm 0.0073$. The Sapirshtein analytical anchor at $L = 3$ (Appendix D) gives $\rho_{\mathrm{PTO}}^{F=0}(\alpha = 0.35) = 0.3842$

exactly. The gap of $\Delta = -0.0028$ is approximately 0.4 standard errors, well within seed-level RL noise. The trained policy therefore recovers the analytical optimum at the no-fee limit, confirming that the F=10 baseline of 13.2% in Table 4 is the deployed-fee equivalent of the no-fee Sapirshtein anchor rather than a fee-model artifact.

APPENDIX F
γ SENSITIVITY SPOT-CHECK

We spot-check the threshold-collapse claim across $\gamma \in \{0.0, 0.5, 1.0\}$ at three headline cells (F=10 baseline, combined-low at $\omega = o = 0.10$, and combined-high at $\omega = o = 0.25$) at $\alpha = 0.35$, $F = 10$, $\delta = 0.01$, $L = 3$. The new sweep covers $\gamma \in \{0, 1\}$ at 3 seeds per cell; $\gamma = 0.5$ is taken from the main sweep at $n = 10$.

Table 13 reports the trained-policy revenue at each of the three cells across the three γ values.

TABLE 13: γ sensitivity at $\alpha = 0.35$, $F = 10$, $L = 3$. $n = 3$ seeds at $\gamma \in \{0, 1\}$; $n = 10$ at $\gamma = 0.5$ (main sweep). The $\gamma = 1.0$ high corner is at $n = 2$, with one seed excluded by the convergence cut.

condition	γ=0	γ=0.5	γ=1
F=10 baseline (ω=o=0)	0.409	0.442	0.529
combined-low (ω=o=0.10)	0.498	0.536	0.592
combined-high (ω=o=0.25)	0.618	0.626	0.657

The matching single-channel ablations at $\gamma \in \{0, 1\}$ were not part of this spot-check, so the formal interaction $\text{combined} - \text{rev}_{\text{fee}} - \Delta_{\text{eclipse}} - \Delta_{\text{stale}}$ is not computable from this table alone. We make two qualitative observations. First, revenues track γ as expected from the Sapirshtein literature: $\gamma = 0$ disfavors the strategic miner and $\gamma = 1$ favors them, with the corresponding threshold shifts. Second, the combined-channel cells exceed the F=10 baseline at every γ, so the threshold collapse from F=10 baseline to combined-channel survives both γ extremes. This is consistent with Lemma 1, whose factorization is γ-independent—the cross-term lives on the MINE kernel under non-active fork, where γ does not appear.

APPENDIX G
LEMMA 1 FIXED-PUBLISH-1 SIMULATION

We anchor the revenue-level cross-term claim (Lemma 1 corollary) on a non-RL substrate by simulating the joint MINE kernel under a fixed publish-1 strategy: publish whenever the strategic miner has at least one private block and the honest chain is empty, adopt the public chain when the honest miner pulls ahead by one, and otherwise mine. There is no policy adaptation or MCTS in the loop; the cross-term signature in revenue comes purely from the kernel.

The simulator is included in the released artifact and runs 200,000 MINE steps with 10 seeds per cell. Define the single-channel publish-1 revenue contributions:

$$\Delta^{\text{pub1}}_{\text{eclipse}}(\omega) = \text{rev}_{\text{pub1}}(\omega, 0) - \text{rev}_{\text{pub1}}(0, 0),$$

$$\Delta^{\text{pub1}}_{\text{stale}}(o) = \text{rev}_{\text{pub1}}(0, o) - \text{rev}_{\text{pub1}}(0, 0).$$

Table 14 reports the revenue-level interaction

$$\text{rev}_{\text{pub1}}(\omega, o) - \left[\text{rev}_{\text{pub1}}(0, 0) + \Delta^{\text{pub1}}_{\text{eclipse}}(\omega) + \Delta^{\text{pub1}}_{\text{stale}}(o)\right]$$

at $\alpha = 0.35$, $F = 10$, $\delta = 0.01$ on selected cells from the 5×5 grid (§5.4); the full grid is available in the released artifact. Per-cell standard errors across the 10 seeds are ≤ 0.001 throughout, so the point estimates are tight against the kernel-level prediction.

TABLE 14: Fixed-publish-1 revenue interaction at $\alpha = 0.35$, $F = 10$, $\delta = 0.01$. Simulator: 200,000 MINE steps $\times$ 10 seeds per cell. The $-\omega o$ column is the per-step kernel prediction from Lemma 1, which the publish-1 revenue interaction tracks in sign and recovers 40–88% of the kernel-level magnitude (high-intensity diagonal recovers approximately 50%; publish-1 does not adapt to the joint regime).

ω	o	$\text{rev}_{\text{joint}}$	rev_{add}	interaction	$-\omega o$
0.05	0.05	0.4019	0.4040	−0.0022	−0.0025
0.05	0.10	0.4144	0.4188	−0.0044	−0.0050
0.05	0.15	0.4289	0.4328	−0.0039	−0.0075
0.05	0.20	0.4433	0.4486	−0.0053	−0.0100
0.05	0.25	0.4581	0.4665	−0.0083	−0.0125
0.10	0.10	0.4180	0.4235	−0.0055	−0.0100
0.10	0.25	0.4584	0.4711	−0.0128	−0.0250
0.15	0.15	0.4331	0.4420	−0.0089	−0.0225
0.20	0.20	0.4452	0.4641	−0.0189	−0.0400
0.25	0.25	0.4552	0.4858	−0.0306	−0.0625

The publish-1 interaction is non-zero, sub-additive, and grows with (ω, o) on the same diagonal where the trained policy's interaction in §5.4 crosses the additive-region boundary. The kernel cross-term is therefore present in publish-1 revenue at leading order. The trained policy in the main results recovers a larger fraction of the cross-term by adapting publish-k thresholds to the joint regime; that strategic-adaptation residual is precisely what Lemma 1, stated at the kernel level with no strategy choice, does not model.

APPENDIX H
POLICY STRUCTURE AND FUTURE-WORK TAXONOMY

A natural follow-up to the trained policy reported in §5.3 is to characterize its strategic structure in terms of named rules from the Sapirshtein–Nayak [5], [6] taxonomy. Our joint MDP introduces six additional state components beyond the Sapirshtein–Nayak $(\ell_a, \ell_h, \text{fork})$ substrate—the per-chain whale-tx vectors $\mathbf{c}_a, \mathbf{c}_h$, the cached whale-tx counts t_a, t_h, the eclipse counter b_e, and the pending-transaction flag. Marginalizing over those components and reporting modal actions on $(\ell_a, \ell_h, \text{fork})$ alone would hide any systematic dependence the policy might have on them, and a methodologically sound classification on the full 9-dimensional state requires a new family of named rules defined on those dimensions. We leave the construction of such a taxonomy—and the empirical question of which of the six extra components the trained policy actually uses—to future work.

REFERENCES

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," https://bitcoin.org/bitcoin.pdf, 2008.

[2] V. Buterin, "Ethereum: A next-generation smart contract and decentralized application platform," Ethereum white paper, 2014.

[3] I. Eyal and E. G. Sirer, "Majority is not enough: Bitcoin mining is vulnerable," in *Financial Cryptography and Data Security (FC)*. Springer, 2014, pp. 436–454.

[4] A. E. Gencer, S. Basu, I. Eyal, R. van Renesse, and E. G. Sirer, "Decentralization in Bitcoin and Ethereum networks," in *Financial Cryptography and Data Security (FC)*. Springer, 2018, pp. 439–457.

[5] A. Sapirshtein, Y. Sompolinsky, and A. Zohar, "Optimal selfish mining strategies in Bitcoin," in *Financial Cryptography and Data Security - 20th International Conference, FC 2016*. Springer, 2017, pp. 515–532.

[6] K. Nayak, S. Kumar, A. Miller, and E. Shi, "Stubborn mining: Generalizing selfish mining and combining with an eclipse attack," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016, pp. 305–320.

[7] E. Heilman, A. Kendler, A. Zohar, and S. Goldberg, "Eclipse attacks on Bitcoin's peer-to-peer network," in *24th USENIX Security Symposium*, 2015, pp. 129–144.

[8] A. Gervais, G. O. Karame, K. Wüst, V. Glykantzis, H. Ritzdorf, and S. Capkun, "On the security and performance of proof of work blockchains," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. ACM, 2016, pp. 3–16.

[9] M. Carlsten, H. Kalodner, S. M. Weinberg, and A. Narayanan, "On the instability of Bitcoin without the block reward," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 154–167.

[10] P. Daian *et al.*, "Flash boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 910–927.

[11] R. Bar-Zur, A. Abu-Hanna, I. Eyal, and A. Tamar, "WeRLman: To tackle whale (transactions), go deep (RL)," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 271–288.

[12] R. Bar-Zur, D. Dori, S. Vardi, I. Eyal, and A. Tamar, "Deep bribe: Predicting the rise of bribery in blockchain mining with deep RL," in *Deep Learning Security and Privacy Workshop, co-located with IEEE S&P 2023*, 2023.

[13] M. Bahrani, M. Neuder, and S. M. Weinberg, "Selfish mining under general stochastic rewards," *arXiv preprint arXiv:2502.20360*, 2025.

[14] C. Hou *et al.*, "SquirRL: Automating attack analysis on blockchain incentive mechanisms with deep reinforcement learning," in *28th Annual Network and Distributed System Security Symposium (NDSS)*, 2021.

[15] T. Wang, S. C. Liew, and S. Zhang, "When blockchain meets AI: Optimal mining strategy achieved by machine learning," *International Journal of Intelligent Systems*, vol. 36, pp. 2183–2207, 2021.

[16] C. Finkbeiner, M. E. Najd, J. Guskind, and G. Almashaqbeh, "SoK: Time to be selfless?! demystifying the landscape of selfish mining strategies and models," *IACR Cryptology ePrint Archive 2025/043*, 2025.

[17] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Kroll, and E. W. Felten, "SoK: Research perspectives and challenges for Bitcoin and cryptocurrencies," in *2015 IEEE Symposium on Security and Privacy*. IEEE, 2015, pp. 104–121.

[18] D. Silver *et al.*, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[19] ——, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[20] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[21] R. Bar-Zur, I. Eyal, and A. Tamar, "Efficient MDP analysis for selfish mining in blockchains," in *Proceedings of the 2nd ACM Conference on Advances in Financial Technologies (AFT)*. ACM, 2020, pp. 113–131.

[22] F. Ritz and A. Zugenmaier, "The impact of uncle rewards on selfish mining in Ethereum," in *2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2018.

[23] Y. Kwon, D. Kim, Y. Son, E. Vasserman, and Y. Kim, "Be selfish and avoid dilemmas: Fork after withholding (FAW) attacks on Bitcoin," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 195–209.

[24] A. Yaish, G. Stern, and A. Zohar, "Uncle maker: (time)stamping out the competition in Ethereum," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023.

[25] C. Grunspan and R. Pérez-Marco, "On profitability of selfish mining," *arXiv preprint arXiv:1805.08281*, 2018.

[26] C. Feng and J. Niu, "Selfish mining in ethereum," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019.

[27] I. Eyal, A. E. Gencer, E. G. Sirer, and R. van Renesse, "Bitcoin-NG: A scalable blockchain protocol," in *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2016, pp. 45–59.

[28] I. Tsabary and I. Eyal, "The gap game," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2018, pp. 713–728.

[29] R. Pass and E. Shi, "FruitChains: A fair blockchain," in *Proceedings of the ACM Symposium on Principles of Distributed Computing (PODC)*. ACM, 2017, pp. 315–324.

[30] J. Li *et al.*, "Survey on strategic mining in blockchain: A reinforcement learning approach," *arXiv preprint arXiv:2502.17307*, 2025.

[31] R. Nagayama, R. Banno, and K. Shudo, "Identifying impacts of protocol and internet development on the Bitcoin network," in *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2020, pp. 1–6.

[32] M. Corallo, "BIP 152: Compact block relay," Bitcoin Improvement Proposal 152, 2016.

Fateme Ghasemi received the M.Sc. degree in Software Engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2023. She is currently a Ph.D. researcher at the University of Eastern Finland. Her research interests include blockchain security, reinforcement learning, and Markov decision processes.

Reyhaneh Ameri received the Ph.D. degree in Artificial Intelligence from the Department of Computer Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2024. Her research interests include blockchain security, learning automata, and the application of artificial intelligence to consensus protocols.

Mohammadreza Meybodi received the M.S. and Ph.D. degrees in Computer Science from the University of Oklahoma, Norman, OK, USA, in 1980 and 1983, respectively. He is currently a Professor with the Department of Computer Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran. His research interests include learning automata, soft computing, and learning systems.