Decentralization Cheapens Corruptive Majority Attacks

Stephen H. Newman
Princeton University, NJ, USA

Abstract

Corruptive majority attacks, in which mining power is distributed among miners and an attacker attempts to bribe a majority of miners into participation in a majority attack, pose a threat to blockchains. Budish bounded the cost of bribing miners to participate in an attack by their expected loss as a result of attack success. We show that this bound is loose. In particular, an attack may be structured so that under equilibrium play by most miners, a miner's choice to participate only slightly affects the attack success chance. Combined with the fact that most of the cost of attack success is externalized by any given small miner, this implies that if most mining power is controlled by small miners, bribing miners to participate in such an attack is much cheaper than the Budish bound. We provide a scheme for a cheap corruptive majority attack and discuss practical concerns and consequences.

2012 ACM Subject Classification Theory of computation → Algorithmic mechanism design; Applied computing → Digital cash

Keywords and phrases Blockchain, Majority Attack, Corruptive Majority Attack

Digital Object Identifier 10.4230/LIPIcs.AFT.2023.13

Acknowledgements Thanks to Matt Weinberg for substantial discussion, feedback, and advice.

1 Introduction

Blockchain- and consensus-based ledger protocols are generally susceptible to *majority attacks*, in which an attacker gains control of a majority of mining power and uses it to create a new canonical transaction history which diverges substantially from the original heaviest chain [11]. This attack may be highly profitable: on currency-only blockchains, this enables doublespend attacks and may cause chaos and/or devaluation, all of which may be used for economic gain. On blockchains that implement higher-level protocols and applications, such as Ethereum, attackers may also retroactively alter the state of smart contracts or other time-varying constructs, with similar consequences.

Historically, concerns about majority attacks have been dismissed as irrelevant to the current state of major cryptocurrencies. Budish argues that this is not necessarily the case in the long run (or even currently): the cost of an attack on the blockchain is proportional to the mining payout rate, so large transaction flow relative to this cost makes majority attacks profitable [3]. We show that under simple assumptions about the distribution of miner powers, the situation is far *worse* than Budish's upper bound suggests. Budish bounds the necessary payout to miners as the cost to them of attack success. While the cost of attack success is indeed well-estimated by Budish, individual miners' actions are typically not substantially causally correlated with attack success, and so it suffices to pay miners their cost of attack success *times the marginal increase in attack success probability which resulted from their actions*. As a result, in the by-design scenario where miners are small and therefore no individual or small group can exercise substantial control over a blockchain (though this is not always the stable state of affairs [1]), corruptive majority attacks are both cheap and hard to prevent, implying that cryptocurrencies are less game-theoretically stable than previously believed.

After a brief exposition of our model, we develop a simple framework for miner incentive analysis that illuminates the action/result-correlation aspect of their incentive problem. In particular, we show that for small miners, costs of participation in an attack are mostly externalized and sometimes very small. We then develop a more rigorous model of corruptive attacks, in which miners can change their behavior from timestep to timestep depending on state of the attack, and show that appropriately structured attacks succeed and are cheap with high probability. We give a practical example of such an attack on Bitcoin, including calculation of expected cost and discussion of profitability. We discuss novel and previously proposed economic-incentive-based prevention and mitigation strategies. We also give a brief overview of some of the noneconomic incentives that affect the feasibility/likelihood of such an attack.

Related Work

The consideration of majority attacks against digital currencies dates back to at least Nakamoto's work [11]. Bonneau noted the danger of bribery-based attacks and discussed attack methodologies and potential countermeasures in [2]. Since then, a variety of game-theoretic attack techniques (e.g. [8]) and practical methodologies (e.g. [9]) have been proposed. Judmeyer et al. provide a broad overview of bribery attack modeling in [7], including majority attacks. Several majority attacks have also been conducted against a variety of cryptocurrencies, with varying results, as summarized (as of 2019) by [12].

The central – and often ignored – obstacle in conducting a majority attack is the long-term cost to miners resulting from currency devaluation stemming from the attack, as first noted in [11]. Budish analyzed this concept in detail, providing a formal model of the cost to a majority miner and approximating it for Bitcoin in [3]. Moroz et al. responded with a model of an attack-counterattack game with a liquid mining power marketplace, arguing that in the Budish setting, the threat of counterattack served to deter majority attacks designed to allow doublespending [10].

There is also growing interest in the effects of incentive manipulation on a wider class of social-choice mechanisms without money. For instance, bribery [6] and coalition effects [14] have been studied in the context of voting, and there is continuing investigation into similar effects in matching markets and other contexts.

2 Modeling Mining and the Cost of Corruption

2.1 Mining Model

We assume a model of mining and miner incentives similar to that of Budish [3]. We assume a fixed set M of miners, with a function $\phi : M \to \mathbb{R}_{\geq 0}$ mapping miners to their powers (hashrates in the case of hash-based PoW coins, for instance), and denote the total power $\Phi \stackrel{\text{def}}{=} \sum_{m \in M} \phi(m)$. We assume that each block mined is mined by a random miner, chosen at time of mining, with independent probability $\frac{\phi(m)}{\Phi}$ of selecting miner m to mine any given block. A miner who mines a block receives reward R for doing so. We assume that miners may choose to dedicate mining power to an attacker whose goal is to execute a majority attack. For simplicity (and optimistically for stability against such attacks), we assume that miners who attack on a mining turn receive no block rewards that turn.

2.2 Miner Valuation Model

We assume that miners extract value exclusively from present and future mining activity. In particular, we will consider two primary costs to miners of selling mining power: the direct expected lost revenue from lost time mining, and the expected lost future revenue from the increased chance of a majority attack as a result of providing hashpower. We will assume (pessimistically for the attacker) that a successful majority attack will cost a miner exactly their time-discounted expected future revenue[1]. This will be indicated by vR for some value v to be bounded later.

As in Budish's work, the primary cost of attacks will lie in compensating miners for the change in the likelihood of attack success due to their participation – the cost of corruption. We wish to bound this.

2.3 Budish's Cost of Corruption

Budish argued (and we agree) that the cost to miners caused by an attack is bounded by a sum of two costs. The more obvious component is their expected *immediate loss*: their lost mining revenue as a result of devoting mining time to the attack. This may be trivially upper-bounded as

$$T\frac{\phi(m)}{\Phi}R$$

for an attack stretching over T timesteps, where R is the block reward, and will be negligible compared to the second component of the loss in most cases.

The more subtle loss which a miner incurs is their expected *devaluation loss*: their time-discounted future loss of revenue as a result of currency collapse. Following Budish's analysis, we may bound this by

$$v\frac{\phi(m)}{\Phi}R$$

for some constant multiplier v.[2] For instance, assuming constant distribution of mining power for a cryptocurrency which pays k fixed block rewards R per year, $v = \frac{1}{1-\sqrt[k]{0.95}} \approxeq 20k$ would reflect the sum of expected income for all time with 5%/year discounting applied. In most cases, however, v is substantially lower. For instance, assuming that R halves every few years (or that the total mining power of other miners is large and doubles every few years) we may bound v as the number of blocks in just a few years – and assumptions about ongoing costs of mining and revenue from selling mining equipment in a crash further reduce this estimate.

Combining the above two losses, we may weakly incentivize a miner $m \in M$ to participate in an attack of duration T by paying them $(T+v)\frac{\phi(m)}{\Phi}R$. Summing across miners, we see that it costs

$$(T+v)R$$

to pay all miners to participate.

[1] This corresponds to the assumption that a successful attack will prevent any future mining, while a failed attack does not change the value of future mining, causing the largest possible losses to miners.

[2] We assume that a coin will have no change in value under attack failure and will undergo total and permanent devaluation under attack success. This represents an optimistic assumption from the point of view of stability, as it drives the expected cost to miners of an attack (and therefore the cost of said attack) as high as possible.

As Budish points out, this may already be dangerous. TR is low compared to potential profit, and v may be as well – for blockchains where Φ is expected to rise rapidly (PoW coins, for instance), gross mining income for given hardware is expected to drop proportionately, implying that almost all income is obtained in the next few years of mining. For instance, assuming $v \cong 2 \cdot 10^5$, the number of Bitcoin blocks in a little under four years, we get an attack price above $1.2 \cdot 10^6$ BTC, slightly over 5% of BTC in existence. Even in cases where mining power does not substantially increase (some PoS protocols, for instance) and currency valuation is expected to remain approximately constant, v may be bounded by the number of blocks in twenty years, as $v\frac{\phi(m)}{\Phi}R$ is then enough to purchase a stock market portfolio which, if it yields 5%/year over inflation, will be more profitable than mining at relative power $\frac{\phi(m)}{\Phi}$ assuming constant currency valuation.

2.4 The Participation-Success Matrix

We first consider a binary model for action during an attack: during an attack, a player p may either *participate* or *refuse* to participate, and the attack will succeed or fail with probabilities determined by the participation/refusal of the various miners. Player p's reward will depend on both participation/refusal and success/failure.

We note something simple but interesting: when a player's reward depends mostly on success or failure, and success/failure is only slightly affected by participation/refusal, the marginal cost of participation/refusal is much lower than the difference in reward between the success and failure cases. Consider, for instance, a small miner who participates in a majority attack which is expected to permanently crash Bitcoin if it succeeds. Assume that their expected time-discounted future valuation was U, and that their expected mining reward over the course of the attack period is $u << U$. Then, excluding order-u increases in payout due to difficulty reduction, their mining payoff matrix is

	Attack succeeds	Attack fails
Participate	0	U
Refuse	u	$U + u$

For instance, considering the cases where the miner is sure that the attack will (succeed/-fail) regardless of their action, their expected loss as a result of participating in the attack is bounded by $(0/u)$. Relaxing this slightly, if they are sure that their participation affects attack success chance by at most ϵ, their expected losses are at most $u + \epsilon U$.

2.5 Bounding Expected Devaluation Loss

The difference between our bound on cost of corruption and Budish's is simple: while Budish proposed paying all miners their expected losses from an attack, we propose paying them only the *marginal increase* in their expected losses as a result of their participation in the attack. We consider an explicitly multiparty attack: the attacker attempts to corrupt many smaller miners into attack participation, and therefore must pay them only the damage they expect to cost themselves by joining. Under this view, Budish's cost of corruption is not tight – it pays miners enough to convince them to join even in a case where act/no-act and success/failure are perfectly correlated. This is generally far from the truth: the smaller a miner is, the less impact their choice has on their perceived likelihood of attack success (and therefore their expected loss).

Let $f(\frac{\phi(m)}{\Phi}) : [0,1] \to [0,1]$ be a bound on the probability, as estimated by a miner m, that miner m's participation in the attack will cause it to succeed (i.e. that it will not succeed if they do not participate, and will succeed if they do). Then we may tighten the bound on their devaluation loss to $f(\frac{\phi(m)}{\Phi})\frac{\phi(m)}{\Phi}Rv$ and the bound on their total loss to

$$\left(T + f\left(\frac{\phi(m)}{\Phi}\right)v\right)\frac{\phi(m)}{\Phi}R$$

3 A Thresholding Corruptive Attack

We now analyze the attack implied by the above payout rule. Assume for the moment that we can verify full participation in an attack. Pick a start time for the attack, and pay any miner m who participates in the attack

$$(T + f_{\max}v)\frac{\phi(m)}{\Phi}R$$

for $f_{\max}$ chosen such that $\sum_{m:f\left(\frac{\phi(m)}{\Phi}\right)\leq f_{\max}} \phi(m)$ is substantially greater than $\frac{\Phi}{2}$. If each miner acts rationally, any miner m with $f(\frac{\phi(m)}{\Phi}) \leq f_{\max}$ will participate, and our attack will succeed. Summing across all miners, total cost of the attack is bounded by

$$(T + f_{\max}v)\,R$$

which for small values of $f_{\max}$ is much smaller than the attack cost under Budish's analysis.

3.1 Estimating v

v depends on changing economic conditions, the specifics of the protocol under attack, and other real-world factors [3]. In Bitcoin and other PoW coins, for instance, v equal to the number of blocks in a few years (i.e. $\frac{\phi(m)}{\Phi}vR$ equal to undiscounted gross earnings over the next few years) may be a reasonable estimate, given reward halving, electricity costs, continual improvements in mining hardware, and regulatory concerns. Notably, payouts from attack participation provide substantial security as compared to mining, for the same real-world reasons as cited above. For more in-depth discussion of v, see [3].

3.2 Estimating f

We have two means of bounding f. First, under the assumption that $\sum_{m\in M:\frac{\phi(m)}{\Phi}\leq f_{\max}} \frac{\phi(m)}{\Phi}$ is substantially greater than $\frac{1}{2}$, $f\left(\frac{\phi(m)}{\Phi}\right)$ is likely bounded on the close order of $\frac{\phi(m)}{\Phi}$, as the expected distribution of participants should not concentrate except possibly at nearly all players participating or nearly no players participating (in which case $f(m)$ is very low). As such, $f_{\max}$ is only required to be big enough to make the above fraction substantially greater than $\frac{1}{2}$, and will therefore be small for a well-decentralized blockchain. The cheaper attack price that results from smaller miners under this bound corresponds to the fact that these smaller miners externalize more of the damage caused by participating – a miner's expected loss from participation is quadratic in their power, but their expected damage is linear.

Second, and more concerningly, if we can convince all miners that the attack is destined to succeed, f becomes 0, as no individual miner's choice will affect the attack. Convincing miners that the attack will almost surely succeed likewise forces f very low.

Both of these dynamics will prove relevant in the next model.

4 Refining the Model: A Block-By-Block Attack

The previous model relied on the assumption that miner behavior over the course of an attack will not change, and assumed a somewhat arbitrary bound on f. We now consider a multiparty refinement of the classical model for majority attacks. An attack is modeled as a T-step game with chance. The state of the game at the conclusion of timestep $t \in \{1, \ldots, T\}$ is given by an integer l_t indicating the length discrepancy between the attacking and defending chains (positive if the defending chain is longer), with l_0 the distance from the attack fork point to the tip of the heaviest chain at the start of the attack.

The attacker wins the game at the first timestep t where $l_t = 0$; if no such timestep has occurred by $t = T$, the attacker loses. As in [7] (in the case of zero native attacker power), miners are partitioned into two groups. We assume that a substantial set of the miners are honest (will always defend), either because we have not provided them sufficient economic incentive to attack, as will be the case for very large miners, or because of non-economic factors. We will model the remaining non-committed miners as rational agents that choose to mine for either the attacker or the defender based on expected time-discounted future rewards. At each step t, all miners decide to either attack or defend, and a random miner is chosen with probability proportional to their mining power. If that miner is defending, they mine a block on the canonical chain, setting $l_t = l_{t-1} + 1$. If attacking, they mine a block on the attacking chain, setting $l_t = l_{t-1} - 1$, and they receive a payout $c_{t-1,l}$ from the attacker for their choice to attack. Equivalently, we may model miners as making the choice to attack/defend after they are selected as the current round's block miner. At the end of the game, if the attacker lost, each miner m receives payout $v\frac{\phi(m)}{\Phi}R$, their expected future mining returns.

Our payout rule will not incentivize participation by miners of size $> \gamma\Phi$, so we assume for simplicity that all such miners are part of the honest pool – though if they do, it will aid the attack. We assume that the honest miners have combined mining power $\leq g_{\text{Def}}\Phi$. We will again exploit the fact that the actions of small miners (of power $\leq \gamma\Phi$, for our purposes) are only slightly correlated with attack success chance to construct a cheap attack.

4.1 A Payout Rule With Unique Subgame Perfect Nash Equilibrium

As before, our payouts will come in two varieties. Every time a miner mines a block on the attacking chain (as opposed to the defending chain), they lose their block reward R that would have been paid out on the defending chain. We therefore provide them a payout of R *via the defending chain* every time they mine an attacking block. This means that, no matter the eventual state of the attack, their block-reward payouts are identical across all of their strategy options, and their net rewards therefore follow those of the attacking game described above up to a constant. To incentivize miners to participate in the attack, it therefore suffices to provide a payout rule which enforces a unique Nash equilibrium of participation in said game. From an intuitive point of view, we hope to create a payment rule which causes the attack to have a very high chance of succeeding regardless of the behavior of any individual miner. This, in turn, should allow us to pay each individual miner very little, as their marginal loss as a result of participation will be low.

We first define a pseudo-value function

$$w^{\max}_{t,l} = \begin{cases} l = 0: & 0 \\ t = T, l > 0: & v\gamma R \\ t < T, l > 0: & g_{\text{Def}} w^{\max}_{t+1,l+1} + (1 - g_{\text{Def}}) w^{\max}_{t+1,l-1} + \gamma(w^{\max}_{t+1,l+1} - w^{\max}_{t+1,l-1}) \end{cases}$$

This is motivated by the following theorem:

▶ **Theorem 1.** *Fix a payout scheme where the miner who mines block t from the initial state $(t-1,l)$ receives (if they mined the block on the attacking chain) $c_{t-1,l} = (w^{\max}_{t,l+1} - w^{\max}_{t,l-1})$. This scheme admits a unique subgame perfect Nash equilibrium in the mining game described earlier, under which all non-committed miners participate in the attack on every block.*

Proof. We define the value function for miner m as

$$w_{t,l}(m) = \begin{cases} l = 0: & 0 \\ t = T, l > 0: & v\frac{\phi(m)}{\Phi}R \\ t < T, l > 0: & g_{\text{Def}} w_{t+1,l+1} + (1 - g_{\text{Def}}) w_{t+1,l-1} + \frac{\phi(m)}{\Phi} c_{t,l} \end{cases}$$

We induct backwards from $t = T$ on the joint hypotheses that:

1. $w_{t,l}(m)$ is equal to the expectation of the sum of attack payouts and time-discounted post-attack mining rewards for a miner m in the non-committed group playing the subgame perfect Nash equilibrium strategy.
2. Any non-committed miner selected at time t is incentivized to participate in the attack at step t given the above payout.

The first claim is trivial at $t = T$, as the attack's success/failure is determined at the end of step T, and our choice for $w_{T,l}$ reflects the expected future rewards under those circumstances. The second claim at $t = T$ follows fairly simply: if their decision would determine attack success, they are paid $v\gamma R > v\frac{\phi(m)}{\Phi}R$ (their expected loss from attack success), and if not, they are paid nothing (for uniqueness, assume a very small payment in this case).

Given that the hypotheses hold for $t > t_0$, we may show that hypothesis 1 holds for $t = t_0$: observe that as every non-committed miner will attack on the next step, the state (t_0, l) has chance g_{Def} of evolving to $(t_0 + 1, l + 1)$ and chance $(1 - g_{\text{Def}})$ of evolving to $(t_0 + 1, l - 1)$. Moreover, the chance that miner m will be selected to mine the next block (and therefore reap an additional reward of $c_{t_0-1,l} = (w^{\max}_{t_0,l+1} - w^{\max}_{t_0,l-1})$) is $\frac{\phi(m)}{\Phi}$.

It remains to show hypothesis 2 for $t = t_0$ given hypothesis 1 for $t \geq t_0$ and hypothesis 2 for $t > t_0$. We may observe that by hypothesis 1, the loss incurred for participating on step t_0 is $(w_{t_0,l+1}(m) - w_{t_0,l-1}(m))$. It therefore suffices to demonstrate that this is less than $w^{\max}_{t_0,l+1} - w^{\max}_{t_0,l-1}$. $w_{t,l}(m)$ is in fact positive-linear in $\frac{\phi(m)}{\Phi}$, and so as it is increasing in l, $(w_{t_0,l+1}(m) - w_{t_0,l-1}(m))$ is positive-linear in the same. At $\frac{\phi(m)}{\Phi} = \gamma$, we have $w_{t,l}(m) = w^{\max}_{t,l}$ and so $(w_{t_0,l+1}(m) - w_{t_0,l-1}(m)) = (w^{\max}_{t_0,l+1} - w^{\max}_{t_0,l-1})$; therefore, for $\frac{\phi(m)}{\Phi} \leq \gamma$, we have $(w_{t_0,l+1}(m) - w_{t_0,l-1}(m)) \leq (w^{\max}_{t_0,l+1} - w^{\max}_{t_0,l-1})$ as desired. ◀

The attacker must be able to credibly commit to their payout scheme to make the reward scheme accurate (and therefore for the equilibrium to hold). This is reasonable – presuming that their payout for attack failure is at least $v\gamma R$ plus the sum of their payouts (which we will see is small), they are incentivized to keep paying under the Nash equilibrium regardless of the game state (as their reward is inverse to that of a miner with power γv). This may also be generalized to arbitrary behavior, assuming that the attacker and the participant miners can agree on an expectation of future miner behavior.

4.2 Success Likelihood, Expected Attack Length, and Expected Attack Cost

We assume $g_{\text{Def}} + \gamma < \frac{1}{2}$. Attack success probability is at least the probability that, had the attack been allowed to continue for T steps whether or not l_t became 0, l_T would have been ≤ 0. We know that l_t decreases with probability $1 - g_{\text{Def}}$ and increases with probability g_{Def}, so after T steps, it has expectation $l_0 - T(1 - 2g_{\text{Def}})$ (assuming counterfactually that we continued the attack after $l_t = 0$). Then for $T > \frac{l_0}{1-2g_{\text{Def}}}$, Hoeffding's inequality gives that the probability that $l_T > 0$ is

$$\leq \exp\left[-\frac{\frac{1}{2}\left(l_0 - T(1-2g_{\text{Def}})\right)^2}{T}\right]$$

Expected time to attack conclusion is bounded by the expected stopping time of the biased random walk described above, equal to $\frac{l_0}{1-2g_{\text{Def}}}$.

We now prove a generic bound on $c_{t-1,l}$. First, let $\left(X^{t,l}\right)_{t,l}$ be a (t,l)-indexed collection of random walks on the integers s.t. $X^{t,l}$ starts at time t at position l, goes to time T, and increases/decreases at each step with probabilities $(g_{\text{Def}} + \gamma)$, $(1 - g_{\text{Def}} - \gamma)$ respectively. For $i \geq t$, let $X_i^{t,l}$ be the position of $X^{t,l}$ at time i. The choice of the step probabilities gives us that, by its definition, $w_{t,l}^{\max} = v\gamma R \Pr\left[\min_{i\in\{t,\ldots,T\}} X_i^{t,l} \leq 0\right]$. Now by coupling the walks $X^{t,l+1}$ and $X^{t,l-1}$ so that one increases on step i iff the other does, we have

$$\begin{aligned} w_{t,l+1}^{\max} - w_{t,l-1}^{\max} &= v\gamma R \Pr\left[\min_{i\in\{t,\ldots,T\}} X_i^{t,l+1} \in \{1,2\}\right] \\ &= v\gamma R \sum_{\tau=t}^{T} \Pr\left[X_\tau^{t,l+1} \in \{1,2\}\right] \Pr\left[\tau = \arg\min_{i\in\{t,T\}} X_i^{t,l+1} | X_\tau^{t,l+1} \in \{1,2\}\right] \end{aligned}$$

Each of the terms in the sum may be bounded. Proofs of these results are contained in the appendix.

▶ **Lemma 2.**

$$\Pr\left[X_\tau^{t,l+1} \in \{1,2\}\right] \leq \frac{1}{\sqrt{\tau - t}} \frac{(1 - g_{\text{Def}} - \gamma)^{5/2}}{\sqrt{2\pi}\,(g_{\text{Def}} + \gamma)^{7/2}}$$

▶ **Lemma 3.**

$$\Pr\left[\tau = \arg\min_{i\in\{t,T\}} X_i^{t,l+1} | X_\tau^{t,l+1} \in \{1,2\}\right] \leq e^{-\frac{1}{2}(T-\tau)(1-2g_{\text{Def}}-2\gamma)^2}$$

We also require a technical lemma:

▶ **Lemma 4.** *Let $0 < a \leq 1$. Then*

$$\sum_{i=t}^{T} \frac{e^{-a(T-i)}}{\max(\sqrt{i-t}, 1)} \leq \min\left(\frac{2}{\sqrt{T + 1 - t - 2\frac{\ln(1+\sqrt{T+1-t})}{a}}}, 1 + \frac{1}{1 - e^{-a}}\right)$$

Combined, these yield a straightforward bound on worst-case total attack cost:

▶ **Theorem 5.** *The worst-case cost of the $w^{\max}$-attack is bounded by*

$$TR + \sum_{t=1}^{T} c_{t-1,l_t} \leq TR + v\gamma R\left[\frac{4\ln(1+\sqrt{T})}{(1 - 2g_{\text{Def}} - 2\gamma)} + \sqrt{\frac{2}{\pi}} \frac{(1 - g_{\text{Def}} - \gamma)^{5/2}}{(g_{\text{Def}} + \gamma)^{7/2}} 2\sqrt{T}\right]$$

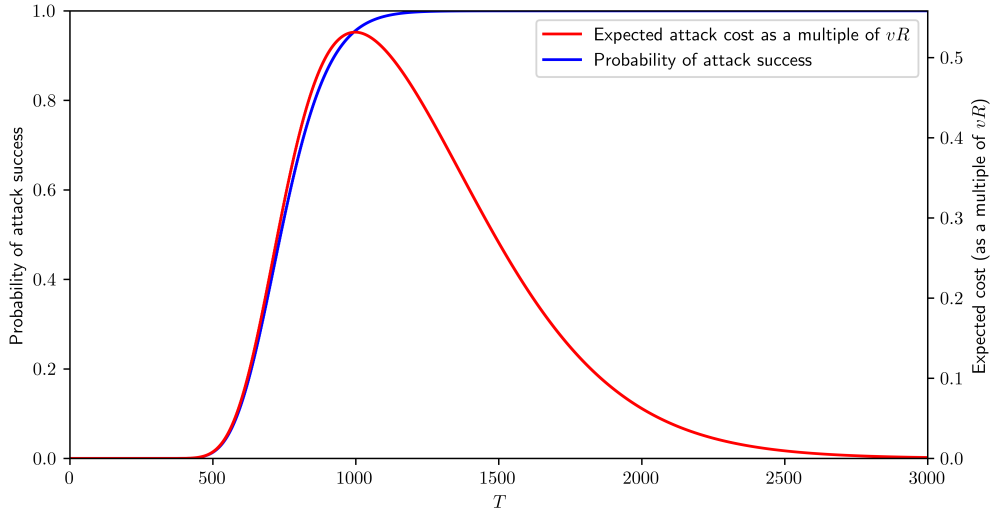

Figure 1 Probability of attack success and expected attack cost (excluding per-block payouts of R) for an attack with given T and $l_0 = 150, \gamma = 0.05, g_{\text{Def}} = 0.4$.

We also have a much lower bound on expected attack cost.

▶ **Theorem 6.** *The expected cost of the $w^{\max}$-attack is bounded by*

$$R\left(\frac{l_0}{2} + \frac{1}{2}\frac{l_0}{1-2g_{\text{Def}}}\right) + v\gamma RTe^{-\frac{(1-2g_{\text{Def}}-2\gamma)^2 T/2-(1-2g_{\text{Def}})l_0}{2}}$$

Critically, for sufficiently large T, we observe exponential decay in the expected attack cost beyond the per-block payout.

These bounds are far from tight. A graph of expected attack cost (excluding the per-block payouts of R) and attack success likelihood in terms of T for an attack against $l_0 = 150$ is included in Figure 1. The attack cost peaks when success is uncertain, as this is where participation/refusal has the strongest effect, and falls off rapidly as attack success becomes certain.

Two remaining points about the asymptotics are of interest. First, as l_0 increases, both attack success chance and attack cost converge more quickly in terms of $\frac{T}{l_0}$, as the expected result concentrates better. In particular, the premium over the naive cost of mining energy required to perform a range-l_0 attack in which $1 - g_{\text{Def}}$ of mining power is participating is logarithmic in l_0, not linear. Second, due to the exponential term, γ does not need to be low to make the attack cost very low. For instance, $\gamma = .2, g_{\text{Def}} = .25$ yields quite feasible attacks for $\frac{T}{l_0}$ substantially greater than $\frac{1-2g_{\text{Def}}}{(1-2g_{\text{Def}}-2\gamma)^2}$. However, only a mild amount of successful collusion (two miners of size .15, for instance) is needed to prevent this attack.

5 Relative Immunity of Proof-of-Stake: Ethereum

Ethereum contains a significant defense against attacks of this type: slashing [5]. As Ethereum severely penalizes stakers who validate on two incompatible chain branches, the internalized penalty for attack participation as compared to attack non-participation may be as high as the Budish payout if the attack is not expected to reduce the price of Ethereum. If attackers will lose their stake, the expected future payoff matrix is

	Attack succeeds	Attack fails
Participate	0	0
Refuse	0	$\frac{\phi(m)}{\Phi} vR$

This implies that short-range attacks (those in which attackers must possess vulnerable stake on the true chain) will cost close to the Budish estimate. Ethereum also has weak subjectivity: the property that any agent entering the network with access to a sufficiently recent honest network state can independently determine the present honest state [4]. Assuming access to such states for entrants, this renders long-range attack impossible, so decentralization does not meaningfully reduce Ethereum attack cost.

6 Practical Economic Considerations

Attackers face two primary logistical difficulties: estimating the total power of participants and coordinating the attack. For PoW blockchains, both of these are ameliorated by setting up a specialized mining pool (with similar structure to that of [2], albeit different intent). Such a pool could function normally until attack time, at which point it could begin sending work units for the attack chain, instead of the consensus chain. This approach has important secondary advantages: mining pools are commonly understood and easy to work with, and by offering low pool fees (or even paying pool participants slightly more than they mine, as suggested by [2] for a different purpose), switchover in advance of the attack could be incentivized while negligibly increasing attack cost. Stratum v2 supports header-only mining, allowing miners to attempt to mine an externally specified block, which should allow easy coordination of attack mining. The remaining coordination difficulty is in payout – corruption payouts cannot flow through the original chain due to likelihood of devaluation. Assuming that another cryptocurrency (for instance, a PoS-based one) is expected to be unaffected by the attack, it could be used as a payment medium; otherwise, a classical medium would be needed.

Regarding pool mining, we also note that the attack proposed in Section 4 can be enhanced to provide lower payout variability with small increase in cost: in addition to the payouts given to participants who mine blocks, duplicate the defending-chain payout and split it proportionate to mining power across the pool. This guarantees participants the same consistency of payout as they would have received had they been participating in a large mining pool, while keeping cost similar, and may incentivize risk-averse/small miners to participate.

7 Attacker Incentives and Dangers

We first estimate the cost of a medium-range majority attack on Bitcoin. Distribution of control of mining capacity is a closely guarded secret, but mining power appears to be significantly spatially and internationally distributed [13]. Assume that miners representing 60% of mining power will participate if we set $\gamma = 1/20$ (i.e. $g_{\text{Def}} = 0.4$). Consider an attack with a backwards range of $l_0 = 150$ blocks (about one day, for Bitcoin). Solving the recursion explicitly, expected attack cost for $T = 2500$ is $< (0.01v + 450)R$, and probability of success $\geq 1 - 10^{-12}$. Assuming $v = 158000$ (the number of Bitcoin blocks in a little over 3 years), expected attack cost is $\leq 1942R$, or a little under 12.2k BTC/310M USD at time of writing. Maximum potential attack cost is substantially higher – $2.38vR$, or about 2.35M BTC/60B USD – but given that attack cost concentrates with exponential falloff, this may be insurable. γ and g_{Def} may also be substantially overestimated here, leading to inflated cost bounds

– if we take $\gamma = 0.03, g_{\text{Def}} = 0.2$, maximum payout is less than 820k BTC/21B USD and expected payout is just over 1293 BTC/33M USD for a $T = 400$ attack (with attack success chance $\geq 1 - 10^{-7}$).

This attack is cheap enough to be profitable, and therefore dangerous. Open Bitcoin options volume has consistently held above $5 billion USD for four years, and has been substantially higher during peaks. Daily trading volume of BTC has stayed well above 200000 BTC for a similar period, and most BTC options are relatively short-term. Attackers able to successfully trade and liquidate options at these scales could make substantial profit.

One example of an attack strategy is as follows. First, buy B "covering" units of BTC, and short-sell a separate B units of BTC. Sell the covering units, and attack to revert that transfer. Provided that the attack succeeds, the recouped covering bitcoin may be used to cover the short position, and as its initial buy cost and sale profit are approximately equal, the net profit of the attack is equal to the gross income from the short sale. In the case of attack failure, the buy and sell costs of the covering units cancel, and cost is equal to the cost of covering the short positions minus the profit from selling them (i.e., it is as if the attackers had simply shorted BTC), less transaction costs from the purchase and sale. Even in this case, the attack may substantially devalue BTC, creating some profit from the short position.

If blockchain technologies become more mainstream, non-currency-oriented attacks will become increasingly appealing. When substantial value may be placed on non-currency attributes of the chain (such as smart contract or DAO state, ownership of a NFT, etc.) in unpredictable ways, many non-obvious attack incentives will exist. This also makes attack attribution more difficult.

In addition to potentially being profitable, a majority attack may significantly if not completely devalue the attacked currency, and possibly others as well. The economic impact of this could be substantial. At time of writing, the global market cap of cryptocurrencies stands at about $1 trillion USD, and various state or non-state actors may stand to gain substantially from a crash precipitated by an attack.

8 Economic Prevention and Mitigation

On a protocol level, this attack is equivalent to a 51% attack, and the same impossibility results apply to protocol-level defense. On a non-protocol level, [2] discusses several prevention measures. We discuss the most relevant of these, and some others, here.

8.1 Coalitions

In response to an incipient attack, a coalition of miners may agree on a mutual behavior contract designed to disincentivize attack participation. In particular, miners could commit to non-subgame-perfect equilibrium strategies, such as defending until a certain threshold of miners have been shown to participate in the attack, and then attacking thereafter, which stymie the given payout rule. However, credible commitment is difficult for miners, and such Nash equilibria are not robust to the attacker designing new payout schemes in response. We conjecture that given any such commitment scheme, the attacker can design a bribery system with expected payout and failure probability bounded by those of the subgame perfect Nash equilibrium and corresponding payout rule. For instance, in the above example, the attacker could increase payouts to miners until the threshold has been passed, and then remove them once it has.

Coalitions with leaders, such as some mining pools, present a separate danger to blockchains: if coalition mining power is controlled by a player whose expected future payout from attack failure is lower than vR, the controller may be incentivized to make the coalition participate even if it falls above the power threshold. For instance, if the controller of some mining pool receives 1% of the profits generated, they will be incentivized to cause the pool to participate if its power is $\leq 100\gamma\Phi$.

8.2 Social Consensus as Deterrence/Mitigation

If a sufficiently broad set of participants wishes, they may fork or otherwise alter the blockchain after an attack in an attempt to reverse the effects of the attack. This is unlikely to prevent profit-taking, as it must be done rapidly enough to fix all manipulated transactions before they can be used for profit. For instance, in the attack proposed in the previous section, the sale of B units of BTC must be fixed as canonical before it can be sold again in the attack chain. Moreover, this has obvious disadvantages for regular use, and does not prevent repeated sabotage attacks [3], which would effectively make the blockchain unusable (which can itself be profitable for an attacker).

8.3 Counterattacks and the Model of Moroz et al.

Moroz et al. [10] argue that in many cases, the threat of counterattacks renders double-spend attacks unprofitable. They analyze a model in which a single large transaction is in contention between the sender (the attacker) and the receiver (the defender), with the former wishing to establish a heaviest chain not including the transaction and the latter wishing the opposite. Each party is allowed to purchase mining power on an open market in order to attempt majority attacks to shift consensus, and take turns doing so. Moroz et al. find that under certain conditions, the only equilibrium strategy is to not attack in the first place.

However, their model (and therefore results) are inapplicable in our case, as both parties may have external rewards which do *not* depend on the final status of the transaction. The success of one or more attacks will almost certainly substantially reduce the value of the attacked cryptocurrency, independent of the final state of the blockchain. This itself may be a major source of utility for the attacker (see Section 7) and disutility for the defender (the defender will recoup coins of lesser value if they are successful, and if they hold additional units of the currency, those two will be devalued by attacks).

When this assumption is changed, the no-attack equilibrium established by Moroz et al. does not hold, and attack with no response is equilibrium in many practical cases. For instance, if a single successful attack will massively devalue a currency, but the attacker can make profit from this event, the attack is incentivized *even if it will be counterattacked*, and a counterattack is generally not incentivized.

8.4 Countercorruption

As noted in [2], miners above the threshold, targets of a large double-spend, and other entities with stake in the success of blockchains may be incentivized to attempt to bribe miners to *not* participate. While this may be theoretically sufficient in some cases, it is undesirable for a variety of reasons discussed in [2].

8.5 Extra Confirmations

Historically, requiring more confirmations has been seen as a natural way to secure transactions. However, the attack cost for even long-range attacks is relatively low, and the splitting tactic noted by [2] likely allows evasion unless many transactions require extremely long confirmation periods.

9 Non-Economic Prevention and Mitigation

We see that there exist potential large-scale corruptive majority attacks which are incentive-compatible for all participants and highly profitable for the attacker, and that economic forces are insufficient to disincentivize attack initiation or participation. However, this does not preclude non-economic forces preventing these attacks in practice. We briefly discuss three potential avenues of prevention.

9.1 Social Consensus

If a power-weighted majority of miners refuse to participate, the attack will be stymied [2]. However, for this to work, we require a broad coalition of miners to act against their own self-interest for philosophical or other non-economic reasons. Miners have regularly behaved selfishly in contravention of protocol [15], so confidence that they will behave selflessly in this case seems misplaced. For instance, miners did not leave F2Pool in large numbers despite its timestamp manipulation, implying that miners are willing to participate in activity with negative network consequences for mildly increased personal profit.

9.2 Force

Actors subject to the jurisdiction of a force (legal or otherwise) able to both detect attack organizers/participants and impose sufficient punishments against them are unlikely to perform/abet attacks. Most immediately, various state actors are likely to take a dim view of such an attack. In the United States, for instance, any attempt to organize such an attack would likely constitute securities market manipulation, and while the legal system is woefully unequipped to deal with the consequences of a double-spend attack, organizing (and perhaps merely participating in) one would likely incur substantial civil and criminal liability. This may be enough to dissuade most actors from performing a majority attack, but it will not be enough to disincentivize those who operate outside the bounds of the law and/or with other motivations. Even individuals subject to regulatory jurisdiction may be insufficiently deterred if the attack is conducted through privacy-preserving tools, as attribution in such a case might be difficult.

9.3 Non-Profitability

One of the incentives to carry out such an attack is profitability. Substantially reducing the viability of profiting off of the collapse of a cryptocurrency could reduce or eliminate the profit motive for majority attacks if its price is expected to collapse by the time the heaviest chain is revised.

10 Conclusion: What *Do* We Trust?

We clarify the analysis of miner incentives surrounding majority attacks, showing that the cost to bribe miners into attacking may be far less than previously believed. The cost is low enough that a corruptive majority attack could be profitable if combined with an appropriate strategy combining shorting and doublespending. Moreover, the value of the attack scales with the value of the blockchain and its associated assets/activity, and the cost of the attack decreases as the miner pool becomes more diverse and incentive-driven, implying that the danger will likely continue to increase as cryptocurrencies grow. Mitigating these attacks through mechanism design is nearly impossible, as doing so would require a method to either force small miners to internalize large social costs or prevent the attacker from profit-taking.

It is unclear whether these attacks can be prevented by non-economic factors. Further work is needed to determine the likelihood of attack in real-world contexts, but such analysis will necessarily be somewhat speculative. Even if these forces can prevent attacks in practice, the advantages of using distributed systems over more classical ledgers become substantially less clear when trust in their stability rests on widespread altruism, institutional force, or other non-economically-motivated behavior. Further analysis of the non-economic forces that may prevent majority attacks, and their implications for the usefulness and viability of blockchain technologies, is warranted.

References

1 Nick Arnosti and S Matthew Weinberg. Bitcoin: A natural oligopoly. *Management Science*, 68(7):4755–4771, 2022.

2 Joseph Bonneau. Why buy when you can rent? bribery attacks on bitcoin-style consensus. In *Financial Cryptography and Data Security: FC 2016 International Workshops, BITCOIN, VOTING, and WAHC, Christ Church, Barbados, February 26, 2016, Revised Selected Papers 20*, pages 19–26. Springer, 2016.

3 Eric Budish. The economic limits of bitcoin and the blockchain. Technical report, National Bureau of Economic Research, 2018.

4 Vitalik Buterin. Proof of stake: How i learned to love weak subjectivity, November 2014. URL: `https://blog.ethereum.org/2014/11/25/proof-stake-learned-love-weak-subjectivity`.

5 Vitalik Buterin and Virgil Griffith. Casper the friendly finality gadget. *arXiv preprint arXiv:1710.09437*, 2017.

6 Piotr Faliszewski, Jörg Rothe, and Hervé Moulin. Control and bribery in voting, 2016.

7 Aljosha Judmayer, Nicholas Stifter, Alexei Zamyatin, Itay Tsabary, Ittay Eyal, Peter Gaži, Sarah Meiklejohn, and Edgar Weippl. Sok: Algorithmic incentive manipulation attacks on permissionless pow cryptocurrencies. In *Financial Cryptography and Data Security. FC 2021 International Workshops: CoDecFin, DeFi, VOTING, and WTSC, Virtual Event, March 5, 2021, Revised Selected Papers 25*, pages 507–532. Springer, 2021.

8 Kevin Liao and Jonathan Katz. Incentivizing blockchain forks via whale transactions. In *Financial Cryptography and Data Security: FC 2017 International Workshops, WAHC, BITCOIN, VOTING, WTSC, and TA, Sliema, Malta, April 7, 2017, Revised Selected Papers 21*, pages 264–279. Springer, 2017.

9 Patrick McCorry, Alexander Hicks, and Sarah Meiklejohn. Smart contracts for bribing miners. In *Financial Cryptography and Data Security: FC 2018 International Workshops, BITCOIN, VOTING, and WTSC, Nieuwpoort, Curaçao, March 2, 2018, Revised Selected Papers 22*, pages 3–18. Springer, 2019.

10 Daniel J Moroz, Daniel J Aronoff, Neha Narula, and David C Parkes. Double-spend counterattacks: Threat of retaliation in proof-of-work systems. *arXiv preprint arXiv:2002.10736*, 2020.

11 Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260, 2008.

12 Savva Shanaev, Arina Shuraeva, Mikhail Vasenin, and Maksim Kuznetsov. Cryptocurrency value and 51% attacks: evidence from event studies. *The Journal of Alternative Investments*, 22(3):65–77, 2019.

13 Wei Sun, Haitao Jin, Fengjun Jin, Lingming Kong, Yihao Peng, and Zhengjun Dai. Spatial analysis of global bitcoin mining. *Scientific Reports*, 12(1):1–12, 2022.

14 Lirong Xia. The impact of a coalition: Assessing the likelihood of voter influence in large elections. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 1156–1156, 2023.

15 Aviv Yaish, Gilad Stern, and Aviv Zohar. Uncle maker:(time) stamping out the competition in ethereum. *Cryptology ePrint Archive*, 2022.

A Notation Reference

For convenience, we define the notations that appear in this paper here:

- M: the set of miners.
- Φ: the total mining power of all miners.
- $\phi : M \to [0, \Phi]$: the function mapping miners to their mining powers.
- T: the duration, in blocks, of an attack.
- R: the (expected) value of mining a block under the stable condition of the blockchain.
- v: a multiplier such that any miner m who expects per-block reward $\frac{\phi(m)}{\Phi}R$ has time-discounted total future profits $\leq v\frac{\phi(m)}{\Phi}R$. That this exists (and may be reasonably bounded) follows from time-discounting, reduction in rewards over time, and hardware improvement/death.
- $f\left(\frac{\phi(m)}{\Phi}\right)$: an upper bound on miner m's estimate of the probability that their participation/non-participation in the attack will determine whether it succeeds.
- $f_{\max}$: a threshold on f. We design an attack which will incentivize participation by miners with $f\left(\frac{\phi(m)}{\Phi}\right) \leq f_{\max}$.
- γ: a threshold on $\frac{\phi(m)}{\Phi}$ with similar purpose to the above.
- g_{Def}: the fraction of total power held by honest miners.

B Proofs of Results in Section 4

Lemma 2

Proof.

$$
\begin{aligned}
&\Pr\left[X_\tau^{t,l+1} \in \{1,2\}\right] \\
&= \binom{\tau - t}{\frac{\tau-t+l}{2}} (1 - g_{\text{Def}} - \gamma)^{\frac{\tau-t+l}{2}} (g_{\text{Def}} + \gamma)^{\frac{\tau-t-l}{2}} \\
&\quad + \binom{\tau - t}{\frac{\tau-t+l-1}{2}} (1 - g_{\text{Def}} - \gamma)^{\frac{\tau-t+l-1}{2}} (g_{\text{Def}} + \gamma)^{\frac{\tau-t-l+1}{2}}
\end{aligned}
$$

where the binomial coefficient $\binom{n}{k}$ is taken to be 0 if k is fractional. One term will always be 0, and so letting $m = \tau - t$, we may bound this by

$$\sup_{l \in \{0,1,\ldots,m\}} \binom{m}{\frac{m+l}{2}} (1 - g_{\text{Def}} - \gamma)^{\frac{m+l}{2}} (g_{\text{Def}} + \gamma)^{\frac{m-l}{2}}$$

Note that this is effectively only over l of the same parity as $\tau - t$.

Let $K_l = \binom{m}{\frac{m+l}{2}} (1 - g_{\text{Def}} - \gamma)^{\frac{m+l}{2}} (g_{\text{Def}} + \gamma)^{\frac{m-l}{2}}$. This admits a smooth extension to $l \in [0, \tau - t]$ via the gamma function. We may bound this extension (via Stirling approximation) by

$$K_l \leq \sup_{l \in [0,\tau-t]} \sqrt{\frac{m}{2\pi \frac{m+l}{2} \frac{m-l}{2}}} \frac{(m)^m}{\left(\frac{m+l}{2}\right)^{\frac{m+l}{2}} \left(\frac{m-l}{2}\right)^{\frac{m-l}{2}}} (1 - g_{\text{Def}} - \gamma)^{\frac{m+l}{2}} (g_{\text{Def}} + \gamma)^{\frac{m-l}{2}} \quad \leq$$

We observe that for integer l of parity $\tau - t$,

$$\begin{aligned} \frac{K_{l+2}}{K_l} &= \frac{\frac{m!}{(\frac{m-(l+2)}{2})!(\frac{m+(l+2)}{2})!}}{\frac{m!}{(\frac{m-l}{2})!(\frac{m+l}{2})!}} \frac{1 - g_{\text{Def}} - \gamma}{g_{\text{Def}} + \gamma} \\ &= \frac{m-l}{m+l+2} \frac{1 - g_{\text{Def}} - \gamma}{g_{\text{Def}} + \gamma} \end{aligned}$$

In particular, K_l is maximized across integers at the lowest l of parity $\tau - t$ s.t. the above ratio is ≤ 1 (as this ratio is decreasing in l). Solving $\frac{m-l}{m+l+2} \frac{1-g_{\text{Def}}-\gamma}{g_{\text{Def}}+\gamma} = l$ gives $l = m(1-2(g_{\text{Def}}+\gamma)) - 2(g_{\text{Def}}+\gamma)$. In particular, the integer $l^{\max}$ maximizing K_l is somewhere between $m(1 - 2(g_{\text{Def}} + \gamma)) - 3$ and $m(1 - 2(g_{\text{Def}} + \gamma))$. Plugging $l^* = m(1 - 2(g_{\text{Def}} + \gamma))$ into our bound on K_l gives

$$\frac{1}{2\sqrt{m}} \sqrt{\frac{1}{2(g_{\text{Def}} + \gamma)(1 - 2(g_{\text{Def}} + \gamma))}}$$

Moreover, we may note that

$$\begin{aligned} \frac{K_{l^*-c}}{K_{l^*}} &= \frac{\Gamma\left(\frac{m+l^*}{2}\right) / \Gamma\left(\frac{m+l^*-c}{2}\right)}{\Gamma\left(\frac{m-l^*+c}{2}\right) / \Gamma\left(\frac{m-l^*}{2}\right)} (1 - g_{\text{Def}} - \gamma)^{c/2} (g_{\text{Def}} - \gamma)^{-c/2} \\ &\leq \left(\frac{(m + l^*)(1 - g_{\text{Def}} - \gamma)}{(m - l^*)(g_{\text{Def}} + \gamma)}\right)^{c/2} \\ &= \left(\frac{1 - g_{\text{Def}} - \gamma}{g_{\text{Def}} + \gamma}\right)^{c} \end{aligned}$$

Then letting $l^{\max} \in [l^* - 3, l^*]$ be the maximizer of K_l, we obtain that

$$\sup_l K_l = K_{l^{\max}} \leq \frac{1}{\sqrt{2\pi m}} \sqrt{\frac{1}{(g_{\text{Def}} + \gamma)(1 - g_{\text{Def}} - \gamma)}} \left(\frac{1 - g_{\text{Def}} - \gamma}{g_{\text{Def}} + \gamma}\right)^3 \quad \blacktriangleleft$$

Lemma 3

Proof. It suffices to observe that

$$\begin{aligned}\Pr\left[\tau = \arg\min_{i\in\{t,T\}} X_i^{t,l+1} | X_\tau^{t,l+1} \in \{1,2\}\right] &\leq \Pr\left[\tau = \arg\min_{i\in\{\tau,T\}} X_i^{t,l+1} | X_\tau^{t,l+1} \in \{1,2\}\right] \\ &= \Pr\left[\tau = \arg\min_{i\in\{\tau,T\}} X_i^{\tau, X_\tau^{t,l+1}}\right] \\ &\leq \Pr\left[X_T^{\tau, X_\tau^{t,l+1}} \geq X_\tau^{\tau, X_\tau^{t,l+1}}\right] \\ &\leq e^{-\frac{1}{2}(T-\tau)(1-2g_{\text{Def}}-2\gamma)^2}\end{aligned}$$

by Hoeffding's inequality applied to $X_T^{\tau, X_\tau^{t,l+1}} - X_\tau^{\tau, X_\tau^{t,l+1}}$, which is by definition a sum of $T-\tau$ random variables which are independently -1 with probability $1-g_{\text{Def}}-\gamma$ and 1 with probability $g_{\text{Def}}+\gamma$. ◀

Lemma 4

Proof. $T = t$ is trivial. Otherwise, fix $t < k \leq T$. Then

$$\begin{aligned}\sum_{i=t}^{T} \frac{e^{-a(T-i)}}{\max(\sqrt{i-t},1)} &\leq \sum_{i=t}^{k-1} \frac{e^{-a(T-i)}}{\max(\sqrt{i-t},1)} + \sum_{i=k}^{T} \frac{e^{-a(T-i)}}{\max(\sqrt{i-t},1)} \\ &\leq \sum_{i=t}^{k-1} \frac{e^{-a(T-(k-1))}}{\max(\sqrt{i-t},1)} + \sum_{i=k}^{T} \frac{e^{-a(T-i)}}{\sqrt{k-t}} \\ &\leq \left(1+\sqrt{k-1-t}\right) e^{-a(T-(k-1))} + \frac{1}{\sqrt{k-t}} \frac{1}{1-e^{-a}}\end{aligned}$$

Setting $k = \max\left(\lceil T - 2\frac{\ln(1+\sqrt{T+1-t})}{a} + 1\rceil, t+1\right)$ yields that the above is

$$\begin{aligned}&\leq \min\left(\frac{1}{1+\sqrt{T+1-t}} + \frac{1}{\sqrt{T+1-t-2\frac{\ln(1+\sqrt{T+1-t})}{a}}}, 1+\frac{1}{1-e^{-a}}\right) \\ &\leq \min\left(\frac{2}{\sqrt{T+1-t-2\frac{\ln(1+\sqrt{T+1-t})}{a}}}, 1+\frac{1}{1-e^{-a}}\right)\end{aligned}$$

◀

Theorem 5

Proof. We have

$$\begin{aligned}c_{t-1,l} &= w_{t,l+1}^{\max} - w_{t,l-1}^{\max} \\ &= v\gamma R \sum_{\tau=t}^{T} \Pr\left[X_\tau^{t,l+1} \in \{1,2\}\right] \Pr\left[\tau = \arg\min_{i\in\{t,T\}} X_i^{t,l+1} | X_\tau^{t,l+1} \in \{1,2\}\right] \\ &\leq v\gamma R \frac{(1-g_{\text{Def}}-\gamma)^{5/2}}{\sqrt{2\pi}\,(g_{\text{Def}}+\gamma)^{7/2}} \sum_{\tau=t}^{T} \frac{e^{-\frac{1}{2}(T-\tau)(1-2g_{\text{Def}}-2\gamma)^2}}{\max\left(\sqrt{\tau-t},1\right)}\end{aligned}$$

by Lemmas 2 and 3. Applying Lemma 4 with $a = \frac{(1-2g_{\text{Def}}-2\gamma)^2}{2}$, combined with the trivial bound $c_{t-1,l} \leq w_{t,l+1}^{\max} \leq v\gamma R$ yields

$$c_{t-1,l} \le v\gamma R \min\left(\frac{(1-g_{\text{Def}}-\gamma)^{5/2}}{\sqrt{2\pi}\,(g_{\text{Def}}+\gamma)^{7/2}} \frac{2}{\sqrt{T+1-t-4\frac{\ln(1+\sqrt{T+1-t})}{(1-2g_{\text{Def}}-2\gamma)^2}}}, 1\right)$$
$$\le v\gamma R \min\left(\frac{(1-g_{\text{Def}}-\gamma)^{5/2}}{\sqrt{2\pi}\,(g_{\text{Def}}+\gamma)^{7/2}} \frac{2}{\sqrt{T+1-t-4\frac{\ln(1+\sqrt{T})}{(1-2g_{\text{Def}}-2\gamma)^2}}}, 1\right)$$

Then the result follows directly from partitioning the indices at $T+1-4\frac{\ln(1+\sqrt{T})}{(1-2g_{\text{Def}}-2\gamma)^2}$. ◀

Theorem 6

Proof. Fix two walks X and Y on the integers starting at l_0. At each step, let X increment with probability g_{Def} and decrement with probability $1-g_{\text{Def}}$, and let Y increment with probability $g_{\text{Def}}+\gamma$ and decrement with probability $1-g_{\text{Def}}-\gamma$. X will correspond to game state (until hitting 0 – i.e. the distributions of l_t and X_t are identical over positive integers), and Y_t is a virtual walk corresponding to the $w^{\max}$ recurrence.

Fix $k > l_0$. By Hoeffding's inequality on X_t, the probability that $l_t \ge k$ for given t is $\le e^{-\frac{(k+(1-2g_{\text{Def}})t)^2}{2t}} \le e^{-(1-2g_{\text{Def}})(\frac{(1-2g_{\text{Def}})t}{2}+k)}$. Taking a union bound across all t, the probability that there exists a timestep t with $l_t \ge k$ is

$$\le \sum_{i=1}^{T} e^{-(1-2g_{\text{Def}})(\frac{(1-2g_{\text{Def}})t}{2}+k)}$$
$$< \sum_{t=1}^{\infty} e^{-(1-2g_{\text{Def}})(\frac{(1-2g_{\text{Def}})t}{2}+k)}$$
$$< \frac{e^{-(1-2g_{\text{Def}})k}}{1-e^{-\frac{(1-2g_{\text{Def}})^2}{2}}}$$

We may observe by its recursion that

$$w^{\max}_{t,l} = v\gamma R \Pr[\min_{i\in[T]} Y_i > 0 | Y_t \le l]$$
$$\le v\gamma R \Pr[Y_T > 0 | Y_t \le l]$$
$$\le \begin{cases} (1-2g_{\text{Def}}-\gamma)(T-t) < l: & v\gamma R \\ (1-2g_{\text{Def}}-\gamma)(T-t) \ge l: & v\gamma R e^{-\frac{(l-(1-2g_{\text{Def}}-\gamma)(T-t))^2}{2(T-t)}} \end{cases}$$
$$\le v\gamma R e^{l(1-2g_{\text{Def}}-2\gamma)-\frac{(1-2g_{\text{Def}}-2\gamma)^2(T-t)}{2}}$$

where the concentration follows from Hoeffding's inequality. We may also bound the probability that the attack has not succeeded by the start of step t, conditioned on there not existing a timestep t with $l_t \ge k$, as

$$\Pr[X_{t-1} > 0 | \sup_{i\in[T]} X_i < k] \le \Pr[X_t \ge 0 | \sup_{i\in[T]} X_i < k]$$
$$\le \Pr[X_t \ge 0]$$
$$\le e^{l_0(1-2g_{\text{Def}})-\frac{(1-2g_{\text{Def}})^2 t}{2}}$$

by the same.

Finally, we observe that

$$c_{t-1,X_{t-1}} \leq w^{\max}_{t,X_{t-1}+1} \leq e^{(X_{t-1}+1)(1-2g_{\text{Def}}-2\gamma)-\frac{(1-2g_{\text{Def}}-2\gamma)^2(T-t)}{2}}$$

In particular, conditioned on $\forall i \in [T] : X_i < k$, we have

$$c_{t-1,l} \Pr[X_t \geq 0 | \forall i \in [T] : X_i < k] \leq e^{l_0(1-2g_{\text{Def}})+k(1-2g_{\text{Def}}-2\gamma)-T(1-2g_{\text{Def}}-2\gamma)^2/2}$$

Then we bound expected attack cost C (excluding the per-attacking-block payout) as

$$\begin{aligned}
\mathbb{E}[C] &\leq \mathbb{E}[C] \Pr[\exists t \in [T] : X_t \geq k] + \mathbb{E}[C | \forall t \in [T] : X_t < k] \\
&\leq \mathbb{E}[C] \Pr[\exists t \in [T] : X_t \geq k] + \sum_{t=1}^{T} c_{t-1,X_{t-1}<k} \Pr[X_{t-1} > 0 | \sup_{i \in [T]} X_i < k] \\
&\leq v\gamma R \left[T \frac{e^{-(1-2g_{\text{Def}})k}}{1 - e^{-\frac{(1-2g_{\text{Def}})^2}{2}}} + T e^{l_0(1-2g_{\text{Def}})+k(1-2g_{\text{Def}}-2\gamma)-T(1-2g_{\text{Def}}-2\gamma)^2/2} \right]
\end{aligned}$$

Solving $-(1-2g_{\text{Def}})k = l_0(1-2g_{\text{Def}}) + k(1-2g_{\text{Def}}-2\gamma) - T(1-2g_{\text{Def}}-2\gamma)^2/2$ yields $k = \frac{(1-2g_{\text{Def}}-2\gamma)^2 T/2 - (1-2g_{\text{Def}})l_0}{2-4g_{\text{Def}}-2\gamma}$ and therefore total expected cost

$$\begin{aligned}
&\leq v\gamma RT e^{-(1-2g_{\text{Def}})\frac{(1-2g_{\text{Def}}-2\gamma)^2 T/2-(1-2g_{\text{Def}})l_0}{2-4g_{\text{Def}}-2\gamma}} \\
&\leq v\gamma RT e^{-\frac{(1-2g_{\text{Def}}-2\gamma)^2 T/2-(1-2g_{\text{Def}})l_0}{2}}
\end{aligned}$$

The expected number of attacking blocks mined is bounded by $\frac{l_0}{2} + \frac{1}{2}\frac{l_0}{1-2g_{\text{Def}}}$ (as the biased random walk is expected to last $\frac{l_0}{1-2g_{\text{Def}}}$ steps, and if the attack lasts k blocks, the attackers mine at most $\frac{l+k}{2}$ of them. Then total expected attack cost is bounded by

$$R\left(\frac{l_0}{2} + \frac{1}{2}\frac{l_0}{1-2g_{\text{Def}}}\right) + v\gamma RT e^{\frac{(1-2g_{\text{Def}}-2\gamma)^2 T/2-(1-2g_{\text{Def}})l_0}{2}}$$ ◀