The Pure Price of Anarchy of Pool Block Withholding Attacks in Bitcoin Mining

Colleen Alkalay-Houlihan, Nisarg Shah
Department of Computer Science
University of Toronto
{colleen,nisarg}@cs.toronto.edu

Abstract

Bitcoin, a cryptocurrency built on the blockchain data structure, has generated significant academic and commercial interest. Contrary to prior expectations, recent research has shown that participants of the protocol (the so-called "miners") are not always incentivized to follow the protocol. We study the game induced by one such attack – the pool block withholding attack – in which mining pools (groups of miners) attack other mining pools. We focus on the case of two pools attacking each other, with potentially other mining power in the system.

We show that this game always admits a pure Nash equilibrium, and its pure price of anarchy, which intuitively measures how much computational power can be wasted due to attacks in an equilibrium, is at most 3. We conjecture, and prove in special cases, that it is in fact at most 2. Our simulations provide compelling evidence for this conjecture, and show that players can quickly converge to the equilibrium by following best response strategies.

1 Introduction

There has been significant interest in Bitcoin ever since it was first proposed by the pseudonymous Satoshi Nakamoto in 2008 (Nakamoto 2008). In recent years, popular fascination with Bitcoin has increased in tandem with its price.

Bitcoin was originally proposed as a decentralized cryptocurrency that at last presented a compelling solution to the double-spending problem. Bitcoin is built on a data structure called the *blockchain*, which is a global distributed ledger that stores relevant information on all historical transactions in the system. Blocks that are sufficiently far back from the leading edge of the blockchain are considered immutable.

Anyone can join or leave the Bitcoin system at any time, so Bitcoin is secured against attacks on the ledger through a *proof of work* system. Significant computational resources must be expended in order for a new block to be generated, through randomly guessing at nonces that, combined with other fixed and time-stamped input, result in a hashed output that satisfies certain criteria. Since there are electricity and hardware costs associated with generating these proofs of work (solutions), participants who generate a successful solution are rewarded with a predetermined number of newly generated Bitcoins (fixed reward), as well as fees for the transactions included in the block. In our work, we only consider the fixed Bitcoin rewards or, equivalently, assume that transaction fees per block are fixed. The process of generating new blocks is called *mining* and participants in this process are called *miners*.

The difficulty of mining is dynamically adjusted so that a new block is mined every ten minutes in expectation. Due to the vast amounts of computational resources currently being expended on Bitcoin mining, an individual miner has a very small probability of being the first one to generate a block, and thus may need to wait many months or years before receiving a positive reward. Although the expected reward from mining is still positive, the variance from mining individually is unacceptably high. As a result, miners form *mining pools*. Each member of a mining pool looks for proofs of work on behalf of the pool, and the Bitcoins awarded for a block mined by any member of the mining pool are shared amongst all members of the pool.

In a decentralized system such as Bitcoin, it is not possible for pool managers to know the true computational power of a miner, so miners send pool managers partial solutions, called *shares*, to allow the pool managers to estimate their computational power. The generally accepted principle is that miners' reward should be somewhat in proportion to their true computational power expended, though there are many different mining pool reward structures; see Rosenfeld (2011) for a thorough review. It is not possible for an invidividual miner in the pool to steal a solution and claim it for himself, since the input for any solution includes the destination wallet for the Bitcoin rewards for that block. Changing the destination for the rewards would cause the hash of the original input to no longer be a valid solution.

Some mining pools are *open*, i.e., they allow any miner to join or leave the pool at will. Eyal (2015) demonstrated that open mining pools may have an incentive to perform *pool block withholding attacks* on each other. In this attack, a pool uses some of its miners to infiltrate the other pools. These miners perform work in the other pools, and release shares in order to obtain a portion of the rewards generated by these other pools, but never release a full solution. While a pool sacrifices some of its computational power in the attack, Eyal showed that the rewards obtained from the other pools can potentially compensate for the loss and increase the overall reward of the pool. The analysis is performed under the assumption that the total Bitcoin reward is fixed, which is justified given that the difficulty of mining adjusts over time based on how frequently blocks are mined.

Eyal argued that, in certain cases, the game induced – *the miner's dilemma* – shares similarities to the eponymous prisoner's dilemma game because in an equilibrium, the pools attack each other and all earn fewer rewards than they would have if they had all mined honestly. In the classic prisoner's dilemma, the price of anarchy – the social loss due to players not cooperating in an equilibrium – can be unbounded. Is this a trait also shared by the miner's dilemma?

1.1 Our Results

We study the miner's dilemma game induced between the mining pools due to the possibility of pool block withholding attacks; we focus on the case of two open (attackable) pools, with potentially additional mining power in the sytem. We show that every game admits a pure Nash equilibrium, and study its *pure price of anarchy*, which we define as the total computational power of the two open pools divided by the least total computational power they use for honest mining in any pure Nash equilibrium.

Note that the traditional definition of the pure price of anarchy would measure loss of player rewards — rather than computational power — in an equilibrium. We believe that measuring loss of computational resources is more important, and show that our definition is in fact an upper bound on the traditional pure price of anarchy. That is, all our upper bounds apply to the traditional definition as well.

We show that the pure price of anarchy is at most 3, and that this bound is not tight. We conjecture that every game in fact admits a unique pure Nash equilibrium, the pure price of anarchy is at most 2 (even with more than two open pools), and the bound of 2 is realized only when there are precisely two open pools with equal computational power and no other solo miners or inaccessible pools. We prove the conjecture in several special cases, and provide compelling empirical evidence in the general case. Our experiments also show that players converge to the (conjectured unique) pure Nash equilibrium quickly by following iterative best responses.

1.2 Related Work

The initial widespread belief underlying Bitcoin was that the protocol only requires a majority of miners to be "honest" – that is, follow the protocol faithfully – in order for the system to be secure against attacks such as transaction manipulations and double-spending. Moreover, it was generally assumed that miners' best strategy to maximize their expected reward (in Bitcoins) was to follow the protocol and behave honestly in all regards.

However, Raulo (2011) and then Rosenfeld (2011) showed that this was not the case. For example, when mining in pools, miners can increase their expected reward under several commonly used pool reward schemes by switching their mining power between different pools. Lewenberg et al. (2015) looked at this from a cooperative game theoretic perspective, demonstrating that for some network parameters, some participants are always incentivized to switch between mining pools.

Furthermore, Eyal and Sirer (2014) showed that honest miners must command at least two-thirds of the Bitcoin mining power to ensure that the system is secure against spending attacks, such as double-spending. See the survey by Zohar (2017) for a partial overview of the results.

Carlsten et al. (2016) studied a game in which the fixed Bitcoin rewards for mining a block are negligible and the total mining reward is dominated by the transaction fees, a scenario which will occur in the future, as fixed Bitcoin rewards per block mined decay over time. They showed that in this scenario, miners can, in certain cases, obtain a greater reward by intentionally forking a block with high transaction fees to dishonestly obtain the rewards therein. They also showed that selfish mining can be profitable for a miner with an arbitrarily low share of the total computational power and who is arbitrarily poorly connected in the network.

Rosenfeld (2011) discussed the so-called *block withholding attack*, in which a miner either delays releasing a full proof of work ("delayed release") or only ever releases shares and never full proofs of work to the pool ("full withholding"). Schrijvers et al. (2017) studied incentive compatibility of different pooled mining reward functions against the delayed release version of the block withholding attack. They proved or disproved incentive compatibility of common pool reward functions, and designed a new incentive compatible reward function, aptly named "the IC reward function."

Eyal (2015) and Luu et al. (2015) both examined full block withholding attacks, though their reward functions differ. Luu et al. assume that a pool divides its own (honest) earnings among all miners in the pool, but divides the earnings from infiltrating (and performing block withholding attacks on) other pools only among its "loyal" miners, implying that pools can distinguish which miners within their pools are honest and which are performing block withholding attacks, but only when dividing rewards from attacking other pools, and not when dividing rewards from honest mining in their own pool. This appears to be less plausible than the reward structure used in the work of Eyal, which assumes that a pool divides all of its rewards – from both honest mining and infiltrating other pools – among both its loyal miners and the miners from other pools performing an attack on the pool.

Our work uses the reward structure proposed by Eyal, in which a pool divides all earnings among all miners mining in the pool in proportion to the number of shares they submit. Eyal proved that with any number of pools, all pools mining honestly and not attacking each other is never a pure Nash equilibrium. Eyal also stated, without proof, that there is always a pure Nash equilibrium in the case of two pools.[1] We prove the claim in our work.

[1] In fact, Eyal makes a specific claim and argues that its validity can be checked "using symbolic computation tools." We show that his claim is incorrect, and prove a corrected version of the claim.

2 Model

Following the work of Eyal (2015), we consider a (closed) Bitcoin system with two open mining pools (players). There may be additional mining power outside of these two pools, due to either solo miners or inaccessible mining pools. We assume that the number of open pools, the mining power of each pool, and the total mining power in the system (including any solo miners and inaccessible pools) are fixed.

We also assume that miners are only ever loyal to one pool, and their loyalty does not change. When mining in their own pool, miners honestly report shares as well as full proofs of work. However, each mining pool might use some (or none) of its miners to infiltrate the other pool. These miners mine in the other pool, report shares, but never report full solutions, i.e., perform a full block withholding attack.

Let m denote the total mining power in the system. For $i \in \{1, 2\}$, let m_i denote the mining power of pool (player) i; thus, $m \geq m_1 + m_2$. The values of m, m_1, and m_2 define a game. In this game, the strategy of pool i is to choose $x_i \in [0, m_i]$, which is the amount of its mining power used to attack the other pool. In this work, we are focused only on *pure strategies*, i.e., deterministic values of x_i.

2.1 Reward Functions

Given a *strategy profile* $\mathbf{x} = (x_1, x_2)$, the rewards of pools 1 and 2 are computed as follows. Note that the effective mining power contributed by pool i is $m_i - x_i$. Thus, the total effective mining power in the system is $m - x_1 - x_2$. Since the difficulty of the proofs of work is frequently adjusted, the direct revenue to pool i from the Bitcoin system, denoted $R_i(x_1, x_2)$, is proportional to the fraction of the effective mining power contributed by the pool.

$$R_1(x_1, x_2) = \frac{m_1 - x_1}{m - x_1 - x_2}, \;\; R_2(x_1, x_2) = \frac{m_2 - x_2}{m - x_1 - x_2}.$$

Without loss of generality, we assume that $m_1 + m_2 - x_1 - x_2 > 0$. This is because if $m_1 + m_2 - x_1 - x_2 = 0$ (i.e., $x_1 = m_1$ and $x_2 = m_2$), then neither pool receives any reward, and either pool can obtain a strictly higher reward by simply retaining a non-zero amount of its computational power to mine honestly. Specifically, $(x_1 = m_1, x_2 = m_2)$ cannot be a pure Nash equilibrium (formally defined below).

The net revenue of each pool comes from two sources: 1) direct revenue from the Bitcoin system (for mining honestly in its own pool) and 2) revenue from infiltrating the other pool. The net revenue is divided among all miners mining in the pool in proportion to their computational power, which includes both the honest miners from this pool and the dishonest miners sent by the other pool (the pool is unable to distinguish between these honest and dishonest miners). Thus, the revenue per unit mining power paid by pools 1 and 2, respectively, are

$$r_1(x_1, x_2) = \frac{R_1(x_1, x_2) + x_1 r_2(x_1, x_2)}{m_1 + x_2},$$
$$r_2(x_1, x_2) = \frac{R_2(x_1, x_2) + x_2 r_1(x_1, x_2)}{m_2 + x_1}.$$

Each pool i wants to maximize its reward function r_i. Note that the total reward to pool i is $m_i r_i(x_1, x_2)$, but maximizing this is equivalent to maximizing $r_i(x_1, x_2)$ because m_i is fixed.

Substituting in the equations for $R_1(x_1, x_2)$ and $R_2(x_1, x_2)$, solving the above system of equations, and performing some simple algebra, we rewrite the reward functions directly in terms of players' strategies:

$$\begin{aligned} r_1(x_1, x_2) &= \frac{m_1 m_2 + m_1 x_1 - x_1^2 - x_1 x_2}{(m - x_1 - x_2)(m_1 m_2 + m_1 x_1 + m_2 x_2)}, \\ r_2(x_1, x_2) &= \frac{m_1 m_2 + m_2 x_2 - x_2^2 - x_1 x_2}{(m - x_1 - x_2)(m_1 m_2 + m_1 x_1 + m_2 x_2)}. \end{aligned} \tag{1}$$

We say that a strategy profile $\mathbf{x}^* = (x_1^*, x_2^*)$ is a pure Nash equilibrium if each player's reward function is maximized, given the strategy of the other player, i.e., if

$$x_1^* = \arg\max_{x_1 \in [0, m_1]} r_1(x_1, x_2^*), \;\; x_2^* = \arg\max_{x_2 \in [0, m_2]} r_2(x_1^*, x_2).$$

Since we assumed $m - x_1 - x_2 \geq m_1 + m_2 - x_1 - x_2 > 0$, both reward functions are continuous and differentiable on the entire domain. Their partial derivatives are provided in the full version of the paper. [2]

Further, $x_1^* \in [0, m_1]$ maximizes $r_1(\cdot, x_2)$ only if one of the following conditions is met; a symmetric statement holds for $x_2^* \in [0, m_2]$ maximizing $r_2(x_1, \cdot)$.

a. The partial derivative vanishes, i.e., $\partial_{x_1} r_1(x_1^*, x_2) = 0$.

b. $x_1^* = 0$ and $\partial_{x_1} r_1(0, x_2) \leq 0$.

c. $x_1^* = m_1$ and $\partial_{x_1} r_1(m_1, x_2) \geq 0$.

2.2 (Pure) Price of Anarchy

In a non-cooperative game, the *(pure) price of anarchy* is a measure of the ratio between the optimal social welfare and the worst social welfare in any (pure) Nash equilibrium (Koutsoupias and Papadimitriou 1999; Papadimitriou 2001). In our case, it is socially desirable that more mining power be used to mine honestly because the more power expended on honest mining, the more secure against attacks the system becomes (Zohar 2017). Hence, we define the pure price of anarchy as follows.

$$\text{PPoA} = \frac{m_1 + m_2}{\min_{(x_1^*, x_2^*) \in N} m_1 + m_2 - x_1^* - x_2^*}.$$

Here, N is the set of all pure Nash equilibria. In words, we essentially define the social welfare in strategy profile (x_1, x_2) to be the total mining power used by the two players to mine honestly (i.e., $m_1 + m_2 - x_1 - x_2$).

Note that the traditional definition of social welfare is the sum of rewards to the two players. In a strategy profile (x_1, x_2), this would be

$$\begin{aligned} m_1 r_1(x_1, x_2) + m_2 r_2(x_1, x_2) &= R_1(x_1, x_2) + R_2(x_1, x_2) \\ &= \frac{m_1 + m_2 - x_1 - x_2}{m - x_1 - x_2}. \end{aligned}$$

[2] https://www.cs.toronto.edu/~nisarg/papers/bitcoin_poa.pdf

We can now easily see that the traditional definition of pure price of anarchy is upper bounded by our definition:

$$\frac{(m_1 + m_2)/m}{\min_{(x_1^*, x_2^*) \in N} (m_1 + m_2 - x_1^* - x_2^*)/(m - x_1^* - x_2^*)} \leq \frac{m_1 + m_2}{\min_{(x_1^*, x_2^*) \in N} m_1 + m_2 - x_1^* - x_2^*}.$$

Thus, our upper bounds on the pure price of anarchy apply to the traditional definition as well.

Similarly, another interesting definition of the social welfare in strategy profile (x_1, x_2) is $m - x_1 - x_2$, i.e., the *total* power in the system used to mine honestly (as opposed to the power used only by the two players to mine honestly). It is easy to see that the pure price of anarchy according to this definition is also upper bounded by our version of the pure price of anarchy.

3 Extreme Equilibria

We begin by examining extreme strategies ($x_1 = 0$, $x_1 = m_1$, $x_2 = 0$, or $x_2 = m_2$) that players might choose in a pure Nash equilibrium, and if they do, how much computational power might be wasted in such an equilibrium. Our first result shows that if at least one player is honest in a pure Nash equilibrium, then at most half of the total power is wasted in the equilibrium.

Lemma 1. *If (x_1^*, x_2^*) is a pure Nash equilibrium where $x_1^* = 0$ or $x_2^* = 0$, then $x_1^* + x_2^* < \frac{m_1+m_2}{2}$. Moreover, if $x_1^* = 0$, then we have*

$$x_2^* = \begin{cases} \frac{m_1\left(m_2 + \sqrt{m^2 - (m+m_1)m_2} - m\right)}{m - m_1 - m_2} & \text{if } m > m_1 + m_2, \\ \frac{m_2}{2} & \text{if } m = m_1 + m_2. \end{cases}$$

A symmetric statement holds if $x_2^* = 0$.

Proof. Suppose (x_1^*, x_2^*) is a pure Nash equilibrium. Assume $x_1^* = 0$ (the case of $x_2^* = 0$ is symmetric). Then, $x_2^* \in \arg\max_{x_2 \in [0, m_2]} r_2(0, x_2)$. First, note that

$$\partial_{x_2} r_2(0, x_2) = \frac{m_1^2 m_2 + 2m_1(m_2 - m)x_2 + (m_1 + m_2 - m)x_2^2}{m_2(m - x_2)^2(m_1 + x_2)^2}$$

Next, it is easy to check that $\partial_{x_2} r_2(0, 0) > 0$ (respectively, $\partial_{x_2} r_2(0, m_2) < 0$); this requires substituting $x_2 = 0$ (respectively, $x_2 = m_2$) and checking that the numerator of the partial derivative is strictly positive (respectively, negative), using the facts that $m_1, m_2 > 0$ and $m \geq m_1 + m_2$. Hence, $x_2^* = 0$ or $x_2^* = m_2$ cannot be maximizers.

The only possibility, therefore, is a solution to $\partial_{x_2} r_2(0, x_2) = 0$. By solving a quadratic equation, it can be checked that the unique solution to this equation is

$$x_2^* = \begin{cases} \frac{m_1\left(m_2 + \sqrt{m^2 - (m+m_1)m_2} - m\right)}{m - m_1 - m_2} & \text{if } m > m_1 + m_2, \\ \frac{m_2}{2} & \text{if } m = m_1 + m_2. \end{cases}$$

In the latter case, we trivially have $x_1^* + x_2^* = 0 + \frac{m_2}{2} < \frac{m_1+m_2}{2}$. In the former case, showing that $x_1^* + x_2^* = x_2^* < \frac{m_1+m_2}{2}$ requires simple algebra. A detailed proof is provided in Lemma 5 in the full version of the paper. □

Lemma 2. *(x_1^*, x_2^*) cannot be a pure Nash equilibrium if $x_1^* = m_1$ or $x_2^* = m_2$.*

Proof. We show that (m_1, x_2^*) cannot be a pure Nash equilibrium; the case of (x_1^*, m_2) is symmetric.

For (m_1, x_2^*) to be a pure Nash equilibrium, we need $r_1(m_1, x_2^*) \geq r_1(0, x_2^*)$; otherwise, player 1 would have an incentive to deviate from the equilibrium. Note that

$$r_1(m_1, x_2^*) = \frac{m_1(m_2 - x_2^*)}{(m - m_1 - x_2^*)(m_1 m_2 + m_1^2 + m_2 x_2^*)},$$
$$r_1(0, x_2^*) = \frac{m_1 m_2}{(m - x_2^*) \cdot m_2(m_1 + x_2^*)}.$$

It is easy to check that $r_1(m_1, x_2^*) \geq r_1(0, x_2^*)$ is equivalent to

$$\begin{aligned} &(m_2 - x_2^*)(m - x_2^*)(m_1 + x_2^*) \\ &\qquad \geq (m - m_1 - x_2^*)(m_1 m_2 + m_1^2 + m_2 x_2^*) \\ &\Leftrightarrow (x_2^*)^3 \geq (x_2^*)^2(m - m_1) + x_2^* m_1(m - m_1 - m_2) \\ &\qquad + m_1^2(m - m_1 - m_2). \end{aligned}$$

Because $m - m_1 \geq m_2 \geq x_2$ (thus, $(x_2^*)^2(m - m_1) \geq (x_2^*)^3$) and $m_1 > 0$, the inequality above can hold only if $m = m_1 + m_2$ and $x_2^* \in \{0, m_2\}$. For $x_2^* = 0$, we showed in the proof of Lemma 1 (and it is easy to check) that $x_1^* = m_1$ would not be the best response for player 1. Also, we noted earlier that (m_1, m_2) cannot be a pure Nash equilibrium because both players receive zero reward, and either player could do better by using some of its power to mine honestly. This concludes the proof. □

4 Symmetric Case

We are now ready to analyze the pure price of anarchy. We begin by examining the special case where both pools have equal computational power (i.e., $m_1 = m_2$). In this case, we can analytically express the unique pure Nash equilibrium (which is symmetric), and establish an upper bound of 2 on the pure price of anarchy.

Lemma 3. *Let $m_1 = m_2 = \frac{m}{k}$, where $k \geq 2$. Then, the unique pure Nash equilibrium (x_1^*, x_2^*) is given by*

$$x_1^* = x_2^* = \frac{m}{4k}\left(2k - 1 - \sqrt{4k^2 - 4k - 7}\right).$$

The pure price of anarchy is at most 2, and equal to 2 if and only if $m_1 = m_2 = \frac{m}{2}$.

Proof. Suppose (x_1^*, x_2^*) is a pure Nash equilibrium. We first show that we cannot have $x_1^* = 0$ or $x_2^* = 0$. Suppose $x_1^* = 0$. Then, Lemma 1 provides the only possible value for x_2^*. It is now easy to check that $\partial_{x_1} r_1(0, x_2^*) > 0$, when $m_1 = m_2 = m/k$. This implies that given strategy x_2^* of pool 2, $x_1^* = 0$ cannot be the best response for pool 1. A symmetric argument shows that we also cannot have $x_2^* = 0$. Additionally, Lemma 2 shows that we cannot have $x_1^* = m_1$ or $x_2^* = m_2$ in the pure Nash equilibrium.

Hence, (x_1^*, x_2^*) is a pure Nash equilibrium only if $\partial_{x_1} r_1(x_1^*, x_2^*) = \partial_{x_2} r_2(x_1^*, x_2^*) = 0$. Substituting $m_1 = m_2 = m/k$, we obtain that both derivatives vanish only if

$$\left(\frac{m}{k}\right)^3 - (m - x_2^*)(x_1^* + x_2^*)^2 \\ + 2\left(\frac{m}{k}\right) x_1^*(x_1^* + x_2^* - m) + \left(\frac{m}{k}\right)^2 (2x_1^* + x_2^*) = 0 \tag{2}$$

and

$$\left(\frac{m}{k}\right)^3 - (m - x_1^*)(x_1^* + x_2^*)^2 \\ + 2\left(\frac{m}{k}\right) x_2^*(x_1^* + x_2^* - m) + \left(\frac{m}{k}\right)^2 (x_1^* + 2x_2^*) = 0. \tag{3}$$

Subtracting Equation (3) from Equation (2), we obtain

$$(x_1^* - x_2^*) \left[2\left(\frac{m}{k}\right)(x_1^* + x_2^* - m) + \left(\frac{m}{k}\right)^2 - (x_1^* + x_2^*)^2\right] = 0$$

$$\Leftrightarrow \begin{cases} x_1^* = x_2^*, \text{ or} \\ 2\left(\frac{m}{k}\right)(x_1^* + x_2^* - m) + \left(\frac{m}{k}\right)^2 - (x_1^* + x_2^*)^2 = 0. \end{cases}$$

The second equality is a quadratic equation in $x_1^* + x_2^*$, which has real roots only if

$$\frac{4m^2}{k^2} \geq 4 \cdot \left(\frac{2m^2}{k} - \frac{m^2}{k^2}\right) \Leftrightarrow \frac{1}{k^2} \geq \frac{1}{k},$$

which does not hold because $k \geq 2$. Hence, we must have $x_1^* = x_2^*$. Setting $x_2^* = x_1^*$ in Equation (2), we obtain

$$4\,(x_1^*)^3 + (x_1^*)^2 \left(\frac{4m}{k} - 4m\right) + x_1^* \left(\frac{3m^2}{k^2} - \frac{2m^2}{k}\right) + \frac{m^3}{k^3} = 0. \tag{4}$$

We see Equation (4) has three solutions to x_1^*:

$$x_1^* = \frac{-m}{2k} \quad \text{or} \quad x_1^* = \frac{m}{4k}\left(2k - 1 \pm \sqrt{4k^2 - 4k - 7}\right).$$

Note that $x_1^* = -m/(2k)$ is clearly invalid because we know that $x_1^* > 0$. Also, note that

$$x_1^* = \frac{m}{4k} \cdot \left(2k - 1 + \sqrt{4k^2 - 4k - 7}\right) \geq \frac{m}{4k} \cdot 4 = \frac{m}{k},$$

where the first transition holds because $2k - 1 + \sqrt{4k^2 - 4k - 7}$ is an increasing function of k for $k \geq 2$ and achieves its lowest value of 4 at $k = 2$. This implies that $x_1^* = x_2^* = m/k$, which we know cannot be a pure Nash equilibrium (Lemma 2). Hence, the only possible equilibrium is given by

$$x_1^* = x_2^* = \frac{m}{4k}\left(2k - 1 - \sqrt{4k^2 - 4k - 7}\right).$$

It is easy to check that this is indeed a pure Nash equilibrium of the game. Now, the pure price of anarchy is given by

$$\frac{m_1 + m_2}{m_1 + m_2 - x_1^* - x_2^*} = \frac{\frac{2m}{k}}{\frac{2m}{k} - \frac{m}{2k}\left(2k - 1 - \sqrt{4k^2 - 4k - 7}\right)} \\ = \frac{4}{5 - 2k + \sqrt{4k(k-1) - 7}}.$$

This is a strictly decreasing function of k for $k \geq 2$. Hence, it achieves its highest value of 2 if and only if $k = 2$. □

5 No Other Miners

Next, we examine a different special case in which the two pools can have arbitrary computational power relative to each other, but there is no computational power in the system besides these two pools. That is, we assume that $m = m_1 + m_2$. We note that this case is mainly of theoretical interest because we can establish the uniqueness of pure Nash equilibrium and analytically express the equilibrium. In practice, with $m = m_1 + m_2$ one of the pools holds at least 50% of the computational power in the system, a scenario under which important security guarantees of the Bitcoin system do not hold.

Lemma 4. *When $m = m_1 + m_2$, the unique pure Nash equilibrium is given by*

$$(x_1^*, x_2^*) = \begin{cases} \left(0, \frac{m_2}{2}\right) & \textit{if } m_1 \leq \frac{m_2}{4}, \\ \left(\frac{m_1}{2}, 0\right) & \textit{if } m_2 \leq \frac{m_1}{4}, \\ \left(\frac{\sqrt{m_1 m_2}\,(2\sqrt{m_1} - \sqrt{m_2})}{\sqrt{m_1} + \sqrt{m_2}}, \right. & \textit{otherwise.} \\ \left. \qquad \frac{\sqrt{m_1 m_2}\,(2\sqrt{m_2} - \sqrt{m_1})}{\sqrt{m_1} + \sqrt{m_2}}\right) & \end{cases}$$

The pure price of anarchy is at most 2, and equal to 2 if and only if $m_1 = m_2 = \frac{m}{2}$.

The proof is presented in the full version of the paper, due to space constraints.

Eyal (2015) claimed (see Section VI of his paper) that there is always a unique solution (x_1^*, x_2^*) to the system of equations $\partial_{x_1} r_1(x_1^*, x_2^*) = 0$ and $\partial_{x_2} r_2(x_1^*, x_2^*) = 0$, and that this solution is the unique pure Nash equilibrium of the game. In our proof of Lemma 4, we show that this claim is incorrect. Specifically, when $m = m_1 + m_2$, this system has a feasible solution if and only if $m_1 \geq m_2/4$ and $m_2 \geq m_1/4$. Otherwise, we obtain an equilibrium at an endpoint ($x_1^* = 0$ or $x_2^* = 0$) with a non-vanishing partial derivative.

6 General Case

We are now ready to present our results in the general case that applies to all miner's dilemma games. First, we show that a pure Nash equilibrium always exists, implying that the pure price of anarchy is well-defined. The proof is in the full version of the paper. For $m = m_1 + m_2$, we note that Lemma 4 already proves the required result. For $m > m_1 + m_2$, we show that the sufficient conditions for the existence of a pure Nash equilibrium given by Glicksberg (1952) are satisfied.

Theorem 1. *Every two-player game of miner's dilemma admits a pure Nash equilibrium.*

We are now ready to present the main result of the paper.

Theorem 2. *The pure price of anarchy of every two-player miner's dilemma game is strictly less than 3.*

Proof. Define $t = m - m_1 - m_2$. Lemmas 3 and 4 give us the desired result when $m_1 = m_2$ or when $t = 0$. Hence, we assume $m_1 \neq m_2$ and $t > 0$. In fact, without loss of generality, we assume $m_1 < m_2$. We want to show that for any pure Nash equilibrium (x_1^*, x_2^*), $x_1^* + x_2^* < (2/3) \cdot (m_1 + m_2)$.

We have already established this for extreme equilibria when $x_1^* \in \{0, m_1\}$ or $x_2^* \in \{0, m_2\}$ (Lemmas 1 and 2). Hence, we restrict our attention to equilibria that satisfy $\partial_{x_i} r_i(x_1^*, x_2^*) = 0$ for $i \in \{1, 2\}$. In this case, we have

$$\begin{aligned}&\partial_{x_1} r_1(x_1^*, x_2^*) + \partial_{x_2} r_2(x_1^*, x_2^*) = 0 \quad \Leftrightarrow \\ &m_1(2m_2 - x_1^* - 2x_2^*) - (2t - x_1^* - x_2^*)(x_1^* + x_2^*) \\ &\qquad\qquad\qquad - m_2(2x_1^* + x_2^*) = 0. \qquad (5)\end{aligned}$$

We can now express x_2^* in terms of x_1^* and t:

$$\begin{aligned}x_2^* &= m_1 + t + \frac{m_2}{2} - x_1^* \\ &\pm \frac{1}{2}\sqrt{4m_1^2 + (m_2 + 2t)^2 + 4m_2x_1^* - 4m_1(m_2 - 2t + x_1^*)}.\end{aligned}$$

It is easy to check that the negative root is the only feasible solution. Recall that our goal is to show an upper bound on $x_1^* + x_2^*$. Because we have x_2^* in terms of x_1^* and t, we can express $x_1^* + x_2^*$ in terms of x_1^* and t. Define

$$\begin{aligned}f(x_1, t) &= m_1 + t + \frac{m_2}{2} \\ &\pm \frac{1}{2}\sqrt{4m_1^2 + (m_2 + 2t)^2 + 4m_2x_1 - 4m_1(m_2 - 2t + x_1)}.\end{aligned}$$

Our goal is to show that $f(x_1^*, t) < (2/3)(m_1 + m_2)$. Let us maximize $f(x_1, t)$. First, we see that

$$\partial_t f(x_1, t) < 0 \Leftrightarrow -4m_2(2m_1 - x_1) - 4m_1x_1 < 0,$$

which is true for all feasible x_1 and t. Hence, we obtain that for all x_1 and $t > 0$, $f(x_1, t) < f(x_1, 0)$. It thus remains to show that $f(x_1^*, 0) \leq (2/3)(m_1 + m_2)$. Note that

$$\begin{aligned}f(x_1, 0) &= m_1 + \frac{m_2}{2} \\ &- \frac{1}{2}\sqrt{4m_1^2 - 4m_1m_2 + m_2^2 - 4m_1x_1 + 4m_2x_1}.\end{aligned}$$

Next, observe that

$$\partial_{x_1} f(x_1, 0) = \frac{m_1 - m_2}{\sqrt{(m_2 - 2m_1)^2 + 4(m_2 - m_1)x_1}} < 0,$$

where the last inequality follows because we assumed $m_1 < m_2$. Hence, we obtain that

$$\begin{aligned}f(x_1^*, t) < f(x_1^*, 0) &\leq f(0, 0) \\ &= m_1 + \frac{m_2}{2} - \frac{1}{2} \cdot |m_2 - 2m_1| \\ &= \min(2m_1, m_2) \\ &\leq \frac{1}{3}(2m_1) + \frac{2}{3}(m_2) \\ &= \frac{2}{3}(m_1 + m_2), \qquad (6)\end{aligned}$$

as desired. □

While we only prove existence of a pure Nash equilibrium and establish a weak upper bound of 3 on the pure price of anarchy, we conjecture that the following stronger result should hold.

Conjecture 1. *In every two-player miner's dilemma game, there exists a unique pure Nash equilibrium, the pure price of anarchy is at most* 2*, and it is equal to* 2 *if and only if* $m_1 = m_2 = m/2$.

We note that in the extreme cases where $m_1 \ll m_2 \approx m$ or $m_2 \ll m_1 \approx m$, the pure price of anarchy also reaches arbitrarily close to 2. But because we require $m_1, m_2 > 0$, it never achieves the bound of 2.

We have proved this conjecture in two special cases: the symmetric case, i.e., $m_1 = m_2$ (Section 4), and the case of no other miners, i.e., $m = m_1 + m_2$ (Section 5). One promising direction for establishing the uniqueness of pure Nash equilibrium in the general case is to leverage the result by Rosen (1965), and show that the sufficient conditions they provide are satisfied in our game. For establishing a tighter upper bound of 2, we note that it may be possible to tighten the analysis in the proof of Theorem 2. Specifically, the upper bound we establish on $f(x_1^*, 0)$ in Equation (6) is in fact an upper bound on $f(x_1, 0)$ for every x_1. Noting that x_1^* is in fact the strategy of pool 1 in a pure Nash equilibrium, we may be able to leverage additional structure of x_1^*, and prove a tighter upper bound on $f(x_1^*, 0)$.

7 Experiments

In this section, we experimentally analyze the miner's dilemma game with two as well as three players, and provide compelling empirical evidence towards Conjecture 1. Additionally, we show that players following best response dynamics quickly converge to the pure Nash equilibrium.

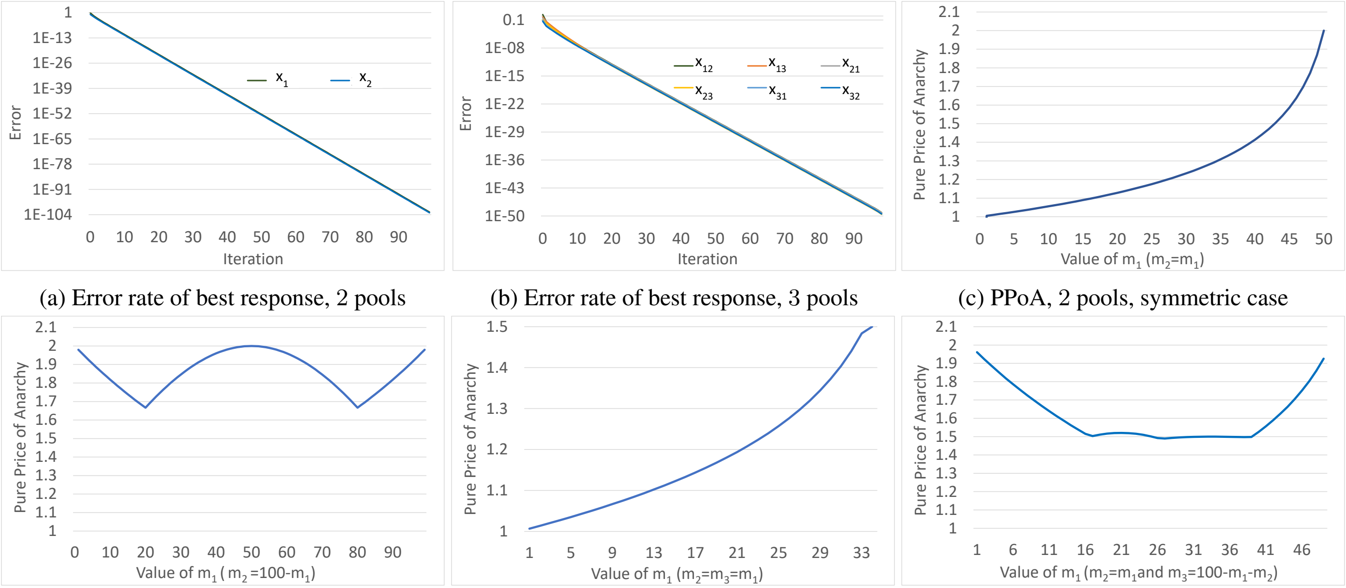
Let $p \in \{2, 3\}$ denote the number of pools. For pool i, let $x_{i,j}$ denote the amount of computational power pool i uses to attack pool j. We refer to the $x_{i,j}$'s as the attack rates. Thus, $\sum_j x_{i,j} \leq m_i$.

We begin by analyzing the best response dynamics. We set $m = 100$, and sample 100 integral values of $\{m_i : i \in [p]\}$ (i.e., 100 different games) uniformly at random subject to the constraint that $\sum_{i=1}^{p} m_i \leq m$. In each game, we simulate iterative best responses starting from honest mining. That is, we begin with $x_{i,j}^0 = 0$ for all i, j. In iteration k, the players optimize their attack rates one by one. That is, $\left\{x_{i,j}^k : j \neq i\right\}$ are optimized given $\left\{x_{i',j}^k : i' < i, j \neq i'\right\} \cup \left\{x_{i',j}^{k-1} : i' > i, j \neq i'\right\}$.

After each iteration, we measure how far the current strategy profile is from the pure Nash equilibrium $\left\{x_{i,j}^* : i \neq j\right\}$; specifically, we define the error at the end of iteration k to be $\sum_{i \neq j} |x_{i,j}^k - x_{i,j}^*|$. For the two-pools case, we were able to numerically find the (unique) pure Nash equilibrium using Mathematica. For the three-pools case, we were not able to do so, and used the strategy profile at the end of iteration 100 as a proxy for the pure Nash equilibrium. We remark that after a few iterations, the change in the attack rates drops significantly. All computations were done in Mathematica using a precision of 250 digits in base 10.

Figures 1a and 1b show log-plots of the error with the number of iterations in the two-pools and three-pools cases, respectively. As we can see, the error decreases exponentially with iterations. In particular, the error drops to less than 10^{-10} in less than 10 iterations for the two-pools case and in less than 20 iterations for the three-pools case.

Next, we focus on the pure price of anarchy in the case of two and three pools. Again, in the case of two pools, we



(a) Error rate of best response, 2 pools

(b) Error rate of best response, 3 pools

(c) PPoA, 2 pools, symmetric case

(d) PPoA, 2 pools, no other miners

(e) PPoA, 3 pools, symmetric case

(f) PPoA, 3 pools, no other miners

Figure 1: Error rates of best response dynamics and pure price of anarchy in miner's dilemma games.

solve the pure Nash equilibrium numerically using Mathematica, while in the case of three pools, we run the best response dynamics until the change is at most 10^{-10} (i.e., until iteration k such that $\sum_{i \neq j} |x^k_{i,j} - x^{k-1}_{i,j}| \leq 10^{-10}$), and use the final strategy profile as proxy for the pure Nash equilibrium. All computations are done with 250 digits (in base 10) of precision in Mathematica. In all our experiments, we set $m = 100$.

Figures 1c and 1d show the PPoA with two pools in the symmetric case ($m_1 = m_2 \in [1, 50]$) and in the case of no other miners ($m_1 \in [1, 99], m_2 = m - m_1$), respectively. Note that in these cases, Lemmas 3 and 4 already provide an analytic expression for the PPoA, which is precisely the quantity plotted in these figures.

Figure 1e shows the PPoA with three pools in the symmetric case where $m_1 = m_2 = m_3 \in [1, 33]$. While the trend is similar to the case of two pools, the PPoA is never higher than 1.5. We explain this phenomenon analytically in the full version of the paper.

Figure 1f shows the PPoA with three pools in the no other miners case where $m = m_1 + m_2 + m_3$, and $m_1 = m_2 \in [1, 49]$. We fixed $m_1 = m_2$ to keep a single free parameter. Here, while PPoA reaches close to 2 in the two edge cases similarly to the two pools case, it only reaches 1.5 in the symmetric case where $m_1 = m_2 = m_3 = m/3$. In fact, it seems that there is an (almost) flat region in the graph, which would be very interesting to explain analytically.

8 Discussion

Our work leaves a number of immediate open questions. Settling our Conjecture 1 (establishing uniqueness of pure Nash equilibrium and proving an upper bound of 2 on the pure price of anarchy) is the first important step. The next step would be to generalize our analysis to the case of more than two pools.

More broadly, game-theoretic analysis of Bitcoin system is still in its infancy, and there are a number of interesting research agendas including analysis of other types of reward functions (in which a pool does not simply divide the reward among miners in proportion to the number of shares submitted) (Schrijvers et al. 2017; Rosenfeld 2011), other types of attacks (e.g., closed pools can also attack open pools), and even other proof systems (e.g., proof of stake).

The ultimate goal in this direction would be to take a mechanism design viewpoint, and design a cryptocurrency system which is either immune to game-theoretic attacks, or which guarantees that desirable equilibria are achieved when the players strategize.

References

Carlsten, M.; Kalodner, H.; Weinberg, S. M.; and Narayanan, A. 2016. On the Instability of Bitcoin Without the Block Reward. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 154–167. ACM.

Eyal, I., and Sirer, E. G. 2014. Majority is not enough: Bitcoin mining is vulnerable. In *Eighteenth International Conference on Financial Cryptography and Data Security*, 436–454.

Eyal, I. 2015. The Miner's Dilemma. In *Proceedings of the 2015 IEEE Symposium on Security and Privacy*, 89–103.

Glicksberg, I. L. 1952. A further generalization of the Kakutani fixed theorem, with application to Nash equilibrium points. *Proceedings of the American Mathematical Society* 3(1):170–174.

Koutsoupias, E., and Papadimitriou, C. 1999. Worst-case equilibria. In *Annual Symposium on Theoretical Aspects of Computer Science*, 404–413. Springer.

Lewenberg, Y.; Bachrach, Y.; Sompolinsky, Y.; Zohar, A.; and Rosenschein, J. S. 2015. Bitcoin mining pools: A cooperative game theoretic analysis. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 919–927. International Foundation for Autonomous Agents and Multiagent Systems.

Luu, L.; Saha, R.; Parameshwaran, I.; Saxena, P.; and Hobor, A. 2015. On Power Splitting Games in Distributed Computation: The Case of Bitcoin Pooled Mining. In *Computer Security Foundations Symposium (CSF), IEEE.*

Nakamoto, S. 2008. Bitcoin: A peer-to-peer electronic cash system. *www.bitcoin.org*.

Papadimitriou, C. 2001. Algorithms, games, and the internet. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 749–753. ACM.

Raulo. 2011. Optimal pool abuse strategy. *http://bitcoin.atspace.com/poolcheating.pdf*.

Rosen, J. B. 1965. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society* 520–534.

Rosenfeld, M. 2011. Analysis of Bitcoin Pooled Mining Reward Systems. *arXiv preprint, arXiv 1112.4980.*

Schrijvers, O.; Bonneau, J.; Boneh, D.; and Roughgarden, T. 2017. Incentive compatibility of bitcoin mining pool reward functions. In *International Conference on Financial Cryptography and Data Security*, 477–498.

Zohar, A. 2017. Securing and scaling cryptocurrencies. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, volume 17, 5161–5165.