# Breaking Blockchain Rationality with Out-of-Band Collusion

Haoqian Zhang[1], Mahsa Bastankhah[1], Louis-Henri Merino[1], Vero Estrada-Galiñanes[1], and Bryan Ford[1]

École polytechnique fédérale de Lausanne(EPFL)
haoqian.zhang,mahsa.bastankhah,louis-henri.merino,vero.estrada,bryan.ford@epfl.ch

**Abstract.** Blockchain systems often rely on rationality assumptions for their security, expecting that nodes are motivated to maximize their profits. These systems thus design their protocols to incentivize nodes to execute the honest protocol but fail to consider out-of-band collusion. Existing works analyzing rationality assumptions are limited in their scope, either by focusing on a specific protocol or relying on non-existing financial instruments. We propose a general rational attack on rationality by leveraging an external channel that incentivizes nodes to collude against the honest protocol. Our approach involves an attacker creating an out-of-band bribery smart contract to motivate nodes to double-spend their transactions in exchange for shares in the attacker's profits. We provide a game theory model to prove that any rational node is incentivized to follow the malicious protocol. We discuss our approach to attacking the Bitcoin and Ethereum blockchains, demonstrating that irrational behavior can be rational in real-world blockchain systems when analyzing rationality in a larger ecosystem. We conclude that rational assumptions only appear to make the system more secure and offer a false sense of security under the flawed analysis.

## 1 Introduction

Blockchain systems often rely on rationality assumptions to ensure their security by providing financial incentives for adhering to the honest protocol. For example, in Proof-of-Work, miners are incentivized to work on the longest chain as it increases their expected chances of having their blocks accepted in the blockchain. Similarly, in Proof-of-Stake, such as the one recently adopted by Ethereum [4], validators are disincentivized from malicious behavior, such as signing two blocks with the same height, due to the loss of part of their deposits. These incentive mechanisms seem to secure these systems as any entity deviating from the honest protocol would have a lower or negative expected return.

However, as many previous works demonstrated [6, 12, 1, 8, 5, 10], those mechanisms might not be incentive-compatible, *i.e.*, there exists a more profitable alternative strategy that deviates from the honest protocol. For instance, selfish mining is a strategy to increase miners' expected return by deviating from the longest-chain rule expected by the Bitcoin mining protocol [6]. Whale attacks

incentivize miners to fork the chain to include an off-the-blockchain transaction with a substantial transaction fee [12].

Whereas those previous works focus on specific protocols within individual blockchain systems, we question the incentive mechanism at a meta-level: Are those blockchain systems rely on rationality assumptions secure in general? We try to answer this research question by considering attacks beyond their ecosystem taking into account the broader influences of the outside world on the system. What is considered irrational behavior within their ecosystem might be rational when analyzing rationality in the context of a larger ecosystem.

We demonstrate that rationality assumptions can be defeated by attacks driven by rationality. Specifically, an attacker creates an out-of-band bribery smart contract that incentivizes nodes to double-spend the attacker's transactions. In return, the attacker can then share the profits from the double-spending with colluded consensus nodes, offering a financial incentive for them to commit the attack in the first place.

A closely related work by Ford and Böhme [7] also offer a general rational attack on rationality. However, their attack method relies on financial instruments that are either non-existent or not well-established in the cryptocurrency markets. We, on the other hand, eliminate the need for non-existent financial instruments and significantly relaxes the requirements to launch the attack.

To prove that out-of-band collusion breaks blockchain systems' rationality assumptions, we propose a game theory model and use it to analyze a blockchain system before and after launching our attack. We find that in the absence of the attack, following the honest protocol is a strict Nash equilibrium that discourages nodes from deviating; however, in the presence of our attack, the honest protocol becomes a weakly dominated strategy. In particular, we identify a finite sequence of deviations from the honest protocol where each deviating node obtains at least the same reward as before the deviation. This sequence ultimately leads to a state where all the nodes follow our attack. Furthermore, we prove that following our attack is a strict Nash equilibrium, thus disincentivizing further deviation.

We provide an outline of the steps required to break the longest-chain rule in Bitcoin and the deposit-slashing protocol in Ethereum. Our work implies that rationality assumptions only appear to make the system more secure and provide a false sense of security.

## 2 Assumptions Underlying the Attack

This section introduces the following assumptions for our attack model:

**Assumption 1:** We consider the target system $S$ to be an open financial payment network operating on blockchain rails, where any client can initiate a transaction. $S$ is maintained by a set of rational nodes $\mathcal{N} = \{1, 2, \ldots, n\}$ who seek to maximize their profits. We assume that each node, $i \in \mathcal{N}$, has the power of $v_i$, *i.e.*, the voting power to decide the next block in the blockchain system. For

example, the voting power in a Proof-of-Work blockchain is the nodes' computational power and the voting power in a Proof-of-Stake blockchain is nodes' stake amount, whereas the voting power in a practical Byzantine Fault Tolerance(PBFT) blockchain is the existence of an approved node. We normalize the power distribution such that the sum of all the nodes' power is equal to 1: $\sum_{i=1}^{n} v_i = 1$. For simplicity, we assume that the number of nodes and their power distribution remains constant; however, our model also applies to the dynamic number of nodes with smooth power changes.

**Assumption 2:** We assume the existence of an open system $S'$ that supports smart contracts and has access to a perfect oracle mechanism $\mathcal{O}$ that can access real-time state information on $S$ without manipulation. To avoid $S'$ and $\mathcal{O}$ being attacked by the same rational attack, we assume that $S'$ and $\mathcal{O}$ do not rely on any rationality assumption, and their security assumptions hold. For example, $S'$ could be a PBFT-styled blockchain, where at most $f$ of $3f + 1$ nodes can fail or misbehave, and $\mathcal{O}$ can solely rely on trusted hardware [3] to provide truthful information from $S$.

**Assumption 3:** The system $S$ leverages, in some fashion, rationality assumptions to incentivize nodes to follow the $S$-defined honest protocol $\mathcal{P}_h$. Mathematically, we assume there is a well-known power threshold $t$ such that, within a time period, if $\mathcal{N}_h \subset \mathcal{N}$ with $\sum_{i \in \mathcal{N}_h} v_i > t$ follows the honest protocol $\mathcal{P}_h$, for $i \in \mathcal{N}_h$ expects to receive a reward of $\mathcal{R}_{h,i} > 0$, and for $i \notin \mathcal{N}_h$ expects to obtain a reward $\mathcal{R}_{d,i}$. We assume that $\forall i \in \mathcal{N}, \mathcal{R}_{d,i} < \mathcal{R}_{h,i}$. $\mathcal{R}_{d,i}$ can be negative, *i.e.*, a node receives punishment for deviating from $\mathcal{P}_h$.

**Assumption 4:** We assume the existence of a malicious protocol $\mathcal{P}_m$ that differs from the expected behavior such that, within the same time period, if $\mathcal{N}_m \subset \mathcal{N}$ with $\sum_{i \in \mathcal{N}_m} v_i > t$ follows the malicious protocol $\mathcal{P}_m$, for $i \in \mathcal{N}_m$ can expect to receive a reward of $\mathcal{R}_{m,i}$, and for $i \notin \mathcal{N}_m$ can expect to obtain a reward of $\mathcal{R}_{d',i}$. We assume that $\forall i \in \mathcal{N}, \mathcal{R}_{d',i} < \mathcal{R}_{m,i}$ and $\mathcal{R}_{m,i} > \mathcal{R}_{h,i}$ as the malicious protocol is only worthwhile for attackers if it provides them with greater rewards. In Section 5.1, we show that there always exists a malicious protocol capable of double-spending attacks to satisfy this assumption in real-world blockchain systems.

**Assumption 5:** We assume that the underlying consensus requires $t \geq \frac{1}{2}$ to avoid nodes split into two independent functional subsets. We also assume that no single node can abuse the system, meaning that $\forall i \in \mathcal{N}, v_i < t$. For simplicity, we assume that if neither $\mathcal{P}_h$ nor $\mathcal{P}_m$ has enough nodes to execute, $S$ loses liveness, and nobody gets any reward.

**Algorithm 1:** Bribery smart contract to incentivize collusion

**Init** *Upon creating the bribery smart contract*:
  Set $T_e$ as the expiration time
  Set $\mathcal{P}_m$ as the malicious protocol
  Deposit $\mathcal{D}_m$ by the magnate
  $\mathcal{N}_m \leftarrow \varnothing$
  $order \leftarrow \mathcal{P}_h$

**Commit** *Upon receiving node i's commitment request*:
  $\mathcal{N}_m \leftarrow \mathcal{N}_m \cup i$
  Deposit $\mathcal{D}_i$ by $i$

**Attack** *Upon* $\sum_{i \in \mathcal{N}_m} v_i > t$:
  $order \leftarrow \mathcal{P}_m$

**Distribute** *Upon receiving the request from* $i \in \mathcal{N}_m$ *for the first time*:
  **if** *Attack is successful and i has executed* $\mathcal{P}_m$ **then**
    Distribute $v_i\mathcal{D}_m + \mathcal{D}_i$ to $i$
  **end**
  **if** *Attack is not successful and* $T_{now} > T_e$ **then**
    Distribute $\mathcal{D}_i$ to $i$
  **end**

## 3 Rational Attack on Rationality

This section presents our attack on rationality at a high-level. We begin by demonstrating that no rational node would execute $\mathcal{P}_m$ without collusion. We then introduce an attacker who creates a *Bribery Smart Contract* on $S'$ that incentivizes the nodes on $S$ to launch the attack.

**Without Collusion:** In the absence of collusion between nodes, each node is incentivized to follow the honest protocol $\mathcal{P}_h$; no single rational node will deviate from $\mathcal{P}_h$ as the expected reward is lower than that of following $\mathcal{P}_h$ ($\mathcal{R}_{d,i} < \mathcal{R}_{m,i}$ in Assumption 3). Therefore, when there is no collusion, $S$ is secure under the rational assumption (we present a game theory analysis in Section 4.1). However, one cannot optimistically assume that such collusion will not exist.

**Magnate-Coordinated Collusion:** When an $S'$ exists, an attacker (referred to as a *magnate*) can use it to coordinate collusion between nodes (Assumption 2). To defeat $S$, the magnate can create a bribery smart contract to attract nodes (referred to as *minions* and denoted by $\mathcal{N}_m$).

We use the double spending attack induced by the magnate as an example to illustrate a possible malicious protocol $\mathcal{P}_m$. The magnate needs to use a bribery smart contract to specify the transaction to be reverted, and order minions to work on a fork that allows the magnate to double-spend the transaction. To

ensure the attack's success, the magnate must guarantee that each node can expect a higher reward, *i.e.*, $\mathcal{R}_m > \mathcal{R}_h$. In the case of this double-spending attack, each node can still expect to receive the rewards that a node executing $\mathcal{P}_h$ would typically get, such as block rewards and transaction fees. However, nodes can now expect to receive a share of the profits obtained by the magnate through double-spending by having the nodes execute $\mathcal{P}_m$. Therefore, the magnate has successfully produced a reward $\mathcal{R}_m$ strictly greater than $\mathcal{R}_h$. Note that the double spending attack is just one example of a malicious protocol. As long as the malicious protocol $\mathcal{P}_m$ produces a higher reward, *i.e.*, $\mathcal{R}_m > \mathcal{R}_h$, it works in our model to defeat rationality.

We outline the design of the bribery smart contract (Algorithm 1) on $S'$ that would enable the magnate to execute the attack successfully. All parties must be held accountable if any party defects to ensure a successful attack in practice. During the creation of the smart contract, the magnate thus deposits $\mathcal{D}_m$ to be shared among the nodes if the attack is successful. In addition, when joining the bribery smart contract, each minion is required to deposit $\mathcal{D}_i$ to be slashed in case of a defect. When the minions' total voting power exceeds $t$, the bribery smart contract orders them to execute $\mathcal{P}_m$. The smart contract then can monitor the attack through the oracle $\mathcal{O}$ (Assumption 2) and upon success, returns the deposits with a share of $\mathcal{D}_m$ to each minion. If the magnate fails to attract enough minions to commit the attack, the deposits are still returned to each minion after an expiration time, making the commitment of the attack by a node risk-free. The magnate can also require a large $\mathcal{D}_i$ as each colluded node expects to get back $\mathcal{D}_i$ eventually (we discuss how to choose $\mathcal{D}_i$ in Section 4.2). However, if a minion does not follow the order from the bribery smart contract, their deposit is burned, thus incentivizing each minion to follow the order.

Given the bribery smart contract, a rational node is incentivized to commit and execute $\mathcal{P}_m$, as, intuitively, every node can benefit. If a node does not participate in the attack, it can, at most, obtain $\mathcal{R}_h$. However, if a node joins the attack, it will receive at least $\mathcal{R}_h$ with the opportunity of increasing its reward to $\mathcal{R}_m$. We offer a game theory analysis on node collusion in Section 4.2. We emphasize that, in this attack, the magnate does not even need to control any part of $S$ or $S'$, making such an attack doable with a low barrier to launch.

## 4 Game Theoretic Analysis

In this section, we formalize the behavior of $S$ nodes and examine the possibility of deviation first without any collusion and then with collusion through the bribery smart contract on $S'$.

In the absence of collusion, following the honest protocol $\mathcal{P}_h$ is a strict Nash equilibrium, meaning that no player will deviate as deviation leads to a lower payoff. However, in the presence of the bribery smart contract, following the protocol $\mathcal{P}_h$ is a weakly dominated strategy and thus is no longer a strict Nash equilibrium. In particular, we identify a sequence of deviations from $\mathcal{P}_h$ where each deviant node obtains at least the same payoff as before. We show that this

sequence of deviations ends with following the bribery smart contract orders. Furthermore, we prove that following the bribery smart contract orders is a strict Nash equilibrium, yielding the maximum payoff of the game. As a result, no rational player would deviate from it.

Additionally, we provide a bound on the amount of money that minions should deposit in the bribery smart contract to ensure that they do not deviate from the bribery smart contract's commands.

### 4.1 Game 0: Without Collusion

We model the behavior of the nodes in the absence of any external factors as a strategic-form game $\Gamma_0 = (\mathcal{N}, \{\mathtt{S}_h, \mathtt{S}_m\}^n, \text{Utility}_i^0(.)_{i\in\mathcal{N}})$. $\mathcal{N} = \{1, 2, \ldots, n\}$ is the set of nodes (players) of the game. Each node $i$ has power $v_i$ such that $\sum_{i\in\mathcal{N}} v_i = 1$. Each player can choose the honest strategy $\mathtt{S}_h$ (corresponding with the protocol $\mathcal{P}_h$) or the malicious strategy $\mathtt{S}_m$ (corresponding with the protocol $\mathcal{P}_m$). We denote the chosen strategy of node $i$ by $s_i$.

We define $V_h$ as the total power of the nodes that choose strategy $\mathtt{S}_h$ and $V_m$ as the total power of the nodes which follow $\mathtt{S}_m$, *i.e.*,

$$V_h \coloneqq \sum_{i\in\mathcal{N}} v_i 1_{\{s_i=\mathtt{S}_h\}}$$

$$V_m \coloneqq \sum_{i\in\mathcal{N}} v_i 1_{\{s_i=\mathtt{S}_m\}} = 1 - V_h.$$

Finally, we define the utility function of node $i$, $\text{Utility}_i^0(.)$, which is a function of $i$'s and other players' strategies as follows:

$$\text{Utility}_i^0(s_1, \ldots, s_n) = \begin{cases} \mathcal{R}_{h,i} & \text{If } s_i = \mathtt{S}_h \quad \& \quad V_h > t \\ \mathcal{R}_{d',i} & \text{If } s_i = \mathtt{S}_h \quad \& \quad V_m > t \\ \mathcal{R}_{d,i} & \text{If } s_i = \mathtt{S}_m \quad \& \quad V_h > t \\ \mathcal{R}_{m,i} & \text{If } s_i = \mathtt{S}_m \quad \& \quad V_m > t \\ 0 & \text{If } V_h, V_m \leq t \end{cases}$$

$$\text{with } \mathcal{R}_{h,i} > \mathcal{R}_{d,i}, \mathcal{R}_{m,i} > \mathcal{R}_{h,i} > 0, \mathcal{R}_{m,i} > \mathcal{R}_{d',i}.$$

Suppose $V_h > t$, *i.e.*, majority power is dedicated to the strategy $\mathtt{S}_h$, player $i$ obtains reward $\mathcal{R}_{h,i}$ by following $\mathtt{S}_h$ and obtains $\mathcal{R}_{d,i}$ otherwise. Similarly, when the majority adopts $\mathtt{S}_m$, player $i$ obtains reward $\mathcal{R}_{m,i}$ by following $\mathtt{S}_m$ and gets $\mathcal{R}_{d',i}$ otherwise. We assume that $\mathcal{R}_{m,i} > \mathcal{R}_{h,i}$ (Assumption 4). If both $V_h$ and $V_m$ are smaller than $t$, all the nodes receive a payoff of 0 (Assumption 5).

**Theorem 1.** *In the strategic-form game $\Gamma_0$ if $\forall i \in \mathcal{N}$, $\mathcal{R}_{d,i} < \mathcal{R}_{h,i}$ and $\max_{i\in\mathcal{N}} v_i \leq t$, the strategy $\mathcal{S}_h$ is a strict Nash Equilibrium.*

*Proof.* We should prove that when all nodes play strategy, $\mathtt{S}_h$, and an arbitrary node $i$ deviates to $\mathtt{S}_m$, $i$ obtains less payoff. We use overline to denote a variable if $i$ deviates.

When everybody plays $\mathtt{S}_h$, $V_h = 1$, and if $i$ deviates then $\overline{V_h} = 1 - v_i$. One of the following two cases will occur:

- If $v_i < 1 - t$, $\overline{V_h} > t$; therefore, even if $i$ deviates, $\mathcal{P}_h$ executes, and $i$ gets $\mathcal{R}_{d,i}$ which is strictly less than $\mathcal{R}_{h,i}$.
- If $v_i \geq 1 - t$, $\overline{V_h} \leq t$ and $\mathcal{P}_h$ does not execute with enough power in $S$ if $i$ deviates. As we assumed that $v_i \leq t$ and $i$ is the only player that plays $\mathcal{P}_m$, we will have $\overline{V_m} = v_i < t$; therefore, $\mathcal{P}_m$ executes with enough nodes neither and every node, including $i$, receives utility 0. As $\mathcal{R}_{h,i} > 0$, $i$ gets less payoff if deviates.

Theorem 1 implies that in the absence of any external factors, given an initial honest behavior in $S$, deviating from $\mathcal{P}_h$ has strictly less utility. Therefore, nodes do not deviate from the honest protocol.

### 4.2 Game 1: Magnate-Coordinated Collusion

We define Game $\Gamma_1 = (\mathcal{N}, \{\mathtt{S}_h, \mathtt{S}'_m\}^n, \text{Utility}_i^1(.)_{i\in\mathcal{N}})$ to describe $S$ in the presence of an external factor: the bribery smart contract (Algorithm 1). Each node has two strategies $\mathtt{S}_h$, $\mathtt{S}'_m$. $\mathtt{S}_h$ is the honest strategy as described before. $\mathtt{S}'_m$ denotes the strategy of committing to the bribery smart contract and following its commands. We can interpret $\mathtt{S}'_m$ as a colluding version of $\mathtt{S}_m$ which nodes only run $\mathcal{P}_m$ if they are sure that enough voting power is dedicated to $\mathcal{P}_m$.

Similarly, we denote the overall power of players who choose $\mathtt{S}_h$ by $V_h$; furthermore, we denote the overall power of minions (players who choose strategy $\mathtt{S}'_m$) by $V'_m$ with relation $V_h + V'_m = 1$. Note that $V'_m$ does not necessarily represent the real power dedicated to $\mathcal{P}_m$ because if $V'_m \leq t$ then the bribery smart contract orders minions to follow $\mathcal{P}_h$ and no one follows $\mathcal{P}_m$; only when $V'_m > t$, the bribery smart contract orders minions to follow the protocol $\mathcal{P}_m$.

To incentivize minions to follow the bribery smart contract's orders unconditionally, the bribery smart contract requires the minions to deposit some money at the time of commitment. Magnate should choose a large enough deposit such that it rules out any order violation. In Theorem 2, we find a deposit function that satisfies this necessity.

**Theorem 2.** *If the bribery smart contract sets the deposit for all the minions as described in the equation 1, under no circumstances any minion has the incentive to deviate from the bribery smart contract commands.*

$$D > \max_{i\in\mathcal{N}}(\mathcal{R}_{m,i} + \max\{|\mathcal{R}_{d,i}|, |\mathcal{R}_{d',i}|\}) \tag{1}$$

*Proof.* Consider node $i$ that has committed to the bribery smart contract and has deposited value $\mathcal{D}_i$. $i$ receives a payoff $x$ if it follows the bribery smart contract commands and gets a payoff $y - \mathcal{D}_i$ if it deviates from the commands where $x, y$ are valid utility values, i.e., $x, y \in \{\mathcal{R}_{m,i}, \mathcal{R}_{h,i}, \mathcal{R}_{d,i}, \mathcal{R}_{d',i}\}$ and their value depend on the strategy of other players. Our objective is to select $\mathcal{D}_i$ in such a way that deviates from the commands of the bribery smart contract are

always more detrimental than any other strategy, regardless of what strategies other players are pursuing. Hence, the following should hold for any valid $x, y$:

$$y - \mathcal{D}_i < x \rightarrow \mathcal{D}_i > y - x$$

We know that as $\mathcal{R}_{h,i}, \mathcal{R}_{m,i} > 0$, $(\max\{\mathcal{R}_{h,i}, \mathcal{R}_{m,i}\} + \max\{|\mathcal{R}_{d,i}|, |\mathcal{R}_{d',i}|\}) = \mathcal{R}_{m,i} + \max\{|\mathcal{R}_{d,i}|, |\mathcal{R}_{d',i}|\}$ is an upper bound on $y - x$; therefore, it suffices to choose $D > \max_{i \in \mathcal{N}}(\mathcal{R}_{m,i} + \max\{|\mathcal{R}_{d,i}|, |\mathcal{R}_{d',i}|\})$

The implication of Theorem 2 is that if a rational node commits to the bribery smart contract, it always follows the bribery smart contract commands. Therefore there are only two possible strategies for the nodes, either playing the honest strategy or committing all of their power to the bribery smart contract and following its orders. If we use a deposit function that does not satisfy equation 1, in some cases, some minions might benefit by deviating from the bribery smart contract orders and dedicating less power to the specified protocol by the bribery smart contract even if they have committed to the bribery smart contract. Thus Theorem 2 is essential for defining $\Gamma_1$. Now we can define the utility function of the game $\Gamma_1$ as follows:

$$\text{Utility}_i^1(s_1, \ldots, s_n) = \begin{cases} \mathcal{R}_{h,i} & \text{If } s_i = \mathtt{S}_h \quad \& \quad V_h > t \\ \mathcal{R}_{d',i} & \text{If } s_i = \mathtt{S}_h \quad \& \quad V'_m > t \\ \mathcal{R}_{h,i} & \text{If } s_i = \mathtt{S}'_m \quad \& \quad V_h > t \\ \mathcal{R}_{m,i} & \text{If } s_i = \mathtt{S}'_m \quad \& \quad V'_m > t \\ \mathcal{R}_{h,i} & \text{If } V_h, V'_m \leq t \end{cases}$$

$$\text{with } \mathcal{R}_{m,i} > \mathcal{R}_{h,i} > 0, \mathcal{R}_{m,i} > \mathcal{R}_{d',i}.$$

The key difference between game $\Gamma_1$ and $\Gamma_0$ is that the minions are now colluding and as a result, they will not execute protocol $\mathcal{P}_m$ when $V_h > t$ to avoid the penalty $\mathcal{R}_{d,i}$.

**Theorem 3.** *In the strategic-form game $\Gamma_1$, the strategy $\mathtt{S}_h$ is not a strict Nash equilibrium, and even further, if any subset of nodes deviates from $\mathtt{S}_h$ to $\mathtt{S}'_m$, the deviating nodes always get at least the same payoff as if they were playing strategy $\mathtt{S}_h$.*

*Proof.* Without the deviation $V_h = 1$, $V'_m = 0$ and every node $i$ obtains reward $\mathcal{R}_{h,i}$. We denote the set of nodes that deviate from $\mathtt{S}_h$ to $\mathtt{S}'_m$ as $\mathcal{N}_m$, while the rest of the nodes $\mathcal{N} - \mathcal{N}_m$ play strategy $\mathtt{S}_h$. We use the overlined variable to show the value of that variable if deviation takes place.

- If the overall power of $\mathcal{N}_m$ is equal or less than $t$, i.e., $\overline{V'_m} \leq t$, the bribery smart contract will order running protocol $\mathcal{P}_h$; therefore, the members of $\mathcal{N}_m$ will run $\mathcal{P}_h$. As other nodes also run $\mathcal{P}_h$, all the nodes no matter if they are a member of $\mathcal{N}_m$ or not will get the same reward as before, i.e., $\mathcal{R}_{h,i}$.

– If the overall power of $\mathcal{N}_m$ is greater than $t$, i.e., $\overline{V'_m} > t$, the bribery smart contract will order running protocol $\mathcal{P}_m$; therefore, the members of $\mathcal{N}_m$ will run $\mathcal{P}_m$ and will obtain reward $\mathcal{R}_{m,i}$, and the rest of the nodes will get the utility $\mathcal{R}_{d',i}$. As $\mathcal{R}_{d',i} < \mathcal{R}_{m,i}$, the nodes that deviate will get a better payoff, and the nodes that do not deviate are better off by deviating.

**Theorem 4.** *In the strategic-form game $\Gamma_1$, if $\mathcal{R}_{d',i} < \mathcal{R}_{m,i}$ and $\mathcal{R}_{h,i} < \mathcal{R}_{m,i}$, the strategy $\mathtt{S}'_m$ is a strict Nash Equilibrium.*

*Proof.* When all the nodes play $\mathtt{S}'_m$ we have $V'_m = 1$, and every node $i$ obtains reward $\mathcal{R}_{m,i}$. If player $i$ deviates to $\mathtt{S}_h$, one of the following two cases will occur:

– If $v_i < 1 - t$, $\overline{V'_m} = 1 - v_i > t$; thus, the bribery smart contract orders to run $\mathcal{P}_m$ and $i$ will receive $\mathcal{R}_{d',i} < \mathcal{R}_{m,i}$.
– If $v_i \geq 1 - t$, $\overline{V'_m} = 1 - v_i \leq t$; thus, the bribery smart contract orders to follow $\mathcal{P}_h$ and every node, as well as $i$, gets the honest reward $\mathcal{R}_{h,i} < \mathcal{R}_{m,i}$.

**Implication:** In a functional system where nodes execute the honest protocol without any collusion, no node has the incentive to deviate. However, with collusion, strategy $\mathtt{S}_h$ becomes a weakly dominated Nash equilibrium. Specifically, any colluding subset of nodes would receive at least the same payoff as before. Hence, it is rational for them to deviate in order to seek a higher payoff. Once the subset with power larger than $t$ deviates, the nodes strictly benefit from deviation (as $\mathcal{R}_{m,i} > \mathcal{R}_{h,i}$); thus, we expect $S$ to transition to a state where everybody plays $\mathtt{S}'_m$. From this point, as $\mathtt{S}'_m$ is a strict Nash equilibrium, no party will deviate from it. In summary, we have identified a sequence of deviations where each node receives at least the same payoff as before, and eventually, the system settles into a strict Nash equilibrium and remains there.

Coming back to the example of a double-spending attack organized by a magnate, Theorem 3 states that starting from a healthy system $S$, if any subset of nodes commit their power to the bribery smart contract and run the double-spending attack if the bribery smart contract orders so, the minions will never get a less payoff than playing the honest strategy. Moreover, Theorem 4 suggests that starting from a situation where all the nodes commit to the bribery smart contract and execute the double spending attack, if a node deviates and plays the honest strategy, the deviant node gets strictly less payoff after deviation.

## 5 Sketch to Break Real-World Blockchain Systems

We illustrate a malicious protocol that generally exists in real-world blockchain systems, and then we discuss how we can use it to attack Bitcoin and Ethereum.

### 5.1 Double-Spending as Malicious Protocol

We present there always exists a malicious protocol $\mathcal{P}_m$ enabling double-spend attacks in $S$, illustrated in Figure 1. A colluded node executes the $\mathcal{P}_m$ when

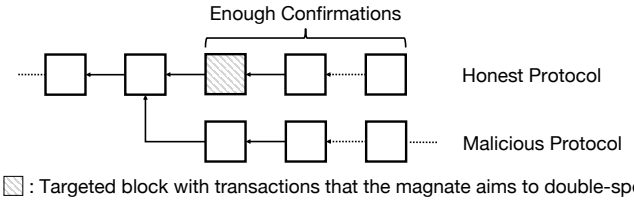


Fig. 1: In a real-world blockchain system, given an honest protocol $\mathcal{P}_h$, the magnate can always construct a malicious protocol $\mathcal{P}_m$ with a higher total reward by double-spending transactions through reverting a confirmed block.

the block that contains the target transactions receives enough block confirmations. The protocol aims to revert the block by working a fork, which allows the magnate to double-spend the transactions confirmed previously. When the fork becomes the valid chain, $\mathcal{P}_m$ finishes.

### 5.2 Breaking the Longest-Chain Rule in Bitcoin

Bitcoin's protocol incentivizes the nodes to adopt the longest-chain rule when mining a new block. This behavior assumption applies to the rationality principle: As long as more than 50% of the nodes follow the longest-chain rule, any rule-deviating node would reduce its expected chance to mine new accepted blocks and thus its expected reward. Therefore, the longest-chain rule is consistent with our Assumption 3.

We now sketch the attacking method based on double-spending. Once a magnate selects a transaction to double spend, they create a bribery smart contract with the malicious protocol $\mathcal{P}_m$ in an attempt to reverse the transaction by creating a fork. The magnate is required to put up a deposit $\mathcal{D}_m$ proportional to the expected reward for double spending this transaction. Similar to an auction contract, the magnate also specifies a time $T_e$ when the contract expires.

Once the bribery smart contract is published, any rational node is incentivized to join the bribery smart contract and, when enough nodes have joined, follow $\mathcal{P}_m$ due to the expected reward increase over following $\mathcal{P}_h$. The bribery smart contract requires nodes to deposit $\mathcal{D}_i$ in case they defect. $\mathcal{D}_i$ needs to be more than the block rewards and transactions fees that can be reverted by the fork. If the bribery smart contract successfully attracts more than 50% of the nodes, then the nodes launches the attack. While launching the attack, each node submits proofs to the bribery smart contract that it is following $\mathcal{P}_m$. Since Bitcoin uses Proof-of-Work as the underlying consensus algorithm, proofs can be hash results that satisfy a difficulty requirement, similar to how miners prove their work to a mining pool [11].

### 5.3 Breaking the Deposit-Slashing Protocol in Ethereum

In the recent upgrade of the Merge [4], Ethereum changed its consensus algorithm to Proof-of-Stake. To incentivize honest nodes and punish malicious ones, Ethereum adopts a deposit-slashing protocol, where each node must deposit some cryptocurrency. A node can withdraw its deposit entirety when exiting the consensus group if no other node can prove that it violated the protocol. Ethereum utilizes the deposit-slashing protocol to punish the double-sign behavior, *i.e.*, a node signs two blocks with the same height, thus mitigating the double-spending issues.

The magnate can adopt a similar strategy to break the deposit-slashing protocol. The magnate still tries to double spend transactions to create additional rewards for the colluded nodes. The colluded nodes need to work on the fork indicated by the magnate after the targeted transaction is confirmed. By doing so, each colluded node needs to sign two blocks with the same height, a behavior violating the deposit-slashing protocol. Thus, the colluded node is subject to be slashed if anyone submits the proof to the blockchain. However, as long as all the colluded nodes do not allow the proof to be included on the blockchain in the first place, the slashing will never happen.

To prove that a node has executed the $\mathcal{P}_m$, the bribery smart contract has to verify that it has voted to the fork indicated by the magnate and has not voted for any block with proof potentially slashing other colluded nodes before exiting the consensus group. The second condition effetely delays the verification time; However, as long as the magnate attracts enough nodes, the magnate is in total control of the blockchain before the colluded nodes exit the consensus group.

## 6 Discussion

Our work reveals the weakness of blockchain systems that depend on rationality for security. Despite this weakness, to the best of our knowledge, no major cryptocurrency has suffered from rational attacks [16, 2], even with the usual concentration of voting power in the hands of a few [13].

The absence of such an attack may result from other factors. First, it may be because the attack is hard to communicate and coordinate, *i.e.*, every node must be aware of such a bribery smart contract, rendering such attacks hard to be realized in real-world blockchain systems. Second, cryptocurrency stakeholders may be unwilling to conduct such an attack due to the potential loss of faith in the cryptocurrency market, leading to significant price drops; thus, it is irrational to launch such an attack if we consider the monetary value of the cryptocurrency [2]. Finally, some actors may choose not to participate in such an attack out of altruism, even though the strategy does not maximize their profits.

Nevertheless, our theoretical conclusion is that rationality is insufficient for security; thus, its use results in a false sense of security, and such an attack could happen at any moment. Our work implies that to build a secure blockchain system, we have to rely on non-rational assumptions, such as threshold assumptions

(*i.e.*, a certain percentage of the nodes are truly honest, even though this would lead to profit loss) and police enforcement (*e.g.*, nodes would face legal prosecution if not following the honest protocol).

## 7 Related Work

The earliest work attacking blockchain rationality is selfish mining, demonstrating that the Bitcoin mining protocol is not incentive-compatible [6]. They prove that, in the current Bitcoin architecture, even if the adversary controls less than 50% of the hashing power, it can launch the attack successfully and earn more benefits than honest behavior.

Following the selfish mining attacks, several works attack blockchain incentive mechanisms, such as whale attacks [12], block withholding [5], stubborn mining [15], transaction withholding [1], empty block mining [8], and fork after withholding [10]. However, these previous works only discuss the attacks in a specific protocol.

Ford et al. first outline a general method to attack rationality, arguing that rationality is self-defeating when analyzing rationality in the context of a large ecosystem [7]. Although the attack generally applies to any blockchain system, it builds upon some non-existing financial instruments, indicating the attack is not practical any time soon. To our knowledge, our work is the first practical and general attack on rationality assumptions for various blockchain systems.

Finally, utilizing smart control to incentivize malicious behaviors is a well-known strategy in the blockchain space. McCorry et al. present various smart contracts that enable bribing of miners to achieve a strategy that benefits the briber [14]. Juels et al. propose criminal smart contracts that encourage the leakage of confidential information [9].

## 8 Conclusion

This paper proposes an attacking method that breaks the rationality assumptions in various blockchain systems. The attack utilizes an out-of-band smart contract to establish the collusion between nodes coordinated by a magnate. Unlike previous works which attack rationality for a specific protocol or rely on non-existent financial instruments, our method is more general and practical. Our result indicates that the rationality assumptions do not increase the system's security and might provide a false sense of security under the flawed analysis.

## Acknowledgments

This research was supported in part by U.S. Office of Naval Research grant N00014-19-1-2361, the AXA Research Fund, the PAIDIT project funded by ICRC, the IC3-Ethereum Fund, Algorand Centres of Excellence programme

managed by Algorand Foundation, and armasuisse Science and Technology. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding sources.

## References

1. Babaioff, M., Dobzinski, S., Oren, S., Zohar, A.: On Bitcoin and red balloons. In: ACM Conference on Electronic Commerce. pp. 56–73. ACM (2012)
2. Badertscher, C., Garay, J., Maurer, U., Tschudi, D., Zikas, V.: But why does it work? a rational protocol design treatment of bitcoin. In: Advances in Cryptology–EUROCRYPT 2018: 37th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Tel Aviv, Israel, April 29-May 3, 2018 Proceedings, Part II 37. pp. 34–65. Springer (2018)
3. Costan, V., Devadas, S.: Intel sgx explained. Cryptology ePrint Archive (2016)
4. The merge (2022), https://ethereum.org/en/upgrades/merge/, accessed: 2022-10-03
5. Eyal, I.: The miner's dilemma. In: IEEE Symposium on Security and Privacy (Oakland). pp. 89–103. IEEE (2015)
6. Eyal, I., Sirer, E.G.: Majority is not enough: Bitcoin mining is vulnerable. Communications of the ACM **61**(7), 95–102 (2018)
7. Ford, B., Böhme, R.: Rationality is self-defeating in permissionless systems. arXiv preprint arXiv:1910.08820 (2019)
8. Houy, N.: The Bitcoin mining game. Available at SSRN: https://ssrn.com/abstract=2407834 (March 2014)
9. Juels, A., Kosba, A., Shi, E.: The ring of gyges: Using smart contracts for crime. aries **40**, 54 (2015)
10. Kwon, Y., Kim, D., Son, Y., Vasserman, E., Kim, Y.: Be selfish and avoid dilemmas: Fork after withholding (FAW) attacks on Bitcoin. In: ACM Conference on Computer and Communications Security (CCS). pp. 195–209. ACM (2017)
11. Lewenberg, Y., Bachrach, Y., Sompolinsky, Y., Zohar, A., Rosenschein, J.S.: Bitcoin mining pools: A cooperative game theoretic analysis. In: Proceedings of the 2015 international conference on autonomous agents and multiagent systems. pp. 919–927 (2015)
12. Liao, K., Katz, J.: Incentivizing blockchain forks via whale transactions. In: International conference on financial cryptography and data security. pp. 264–279. Springer (2017)
13. Mariem, S.B., Casas, P., Romiti, M., Donnet, B., Stütz, R., Haslhofer, B.: All that glitters is not bitcoin–unveiling the centralized nature of the btc (ip) network. In: NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium. pp. 1–9. IEEE (2020)
14. McCorry, P., Hicks, A., Meiklejohn, S.: Smart contracts for bribing miners. In: International Conference on Financial Cryptography and Data Security. pp. 3–18. Springer (2018)
15. Nayak, K., Kumar, S., Miller, A., Shi, E.: Stubborn mining: Generalizing selfish mining and combining with an eclipse attack. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P). pp. 305–320. IEEE (2016)
16. Wang, Z., Lv, Q., Lu, Z., Wang, Y., Yue, S.: Forkdec: accurate detection for selfish mining attacks. Security and Communication Networks **2021**, 1–8 (2021)