






A Rational Protocol Treatment of 51% Attacks

Christian Badertscher¹ , Yun Lu²  , and Vassilis Zikas³

¹ IOHK, Zurich, Switzerland

`christian.badertscher@iohk.io`

² University of Edinburgh, Edinburgh, UK

`Y.Lu-59@sms.ed.ac.uk`

³ Purdue University, West Lafayette, USA

`vzikas@cs.purdue.edu`

Abstract. Game-theoretic analyses of cryptocurrencies and—more generally—blockchain-based decentralized ledgers offer insight on their economic robustness and behavior when even their underpinning cryptographic assumptions fail. In this work we utilize the recently proposed blockchain adaptation of the rational protocol design (RPD) framework [EUROCRYPT '18] to analyze 51% double-spending attacks against Nakamoto-style proof-of-work based cryptocurrencies. We first observe a property of the originally proposed utility class that yields an unnatural conclusion against such attacks, and show how to devise a utility that avoids this pitfall and makes predictions that match the observable behavior—i.e., that renders attacking a dominant strategy in settings where an attack was indeed observed in reality. We then propose a generic remedy to the underlying protocol parameters that provably deter adversaries controlling a majority of the system's resources from attacks on blockchain consistency, including the 51% double-spending attack. This can be used as guidance to patch systems that have suffered such attacks, e.g., Ethereum Classic and Bitcoin Cash, and serves as a demonstration of the power of game-theoretic analyses.

1 Introduction

The classical cryptographic analysis of blockchain ledgers establishes worst-case guarantees on their security either by proving central security properties [GKL15, PSs17], such as *consistency/common-prefix*—the stable parts of the chains held by honest parties are prefixes of one-another—*liveness*—new blocks with recent transactions keep being added—or by proving that the protocol realizes an ideal ledger functionality [BMTZ17]. Typically such analyses rely on an assumed limitation on the adversary's influence/presence in the system. In particular, the majority of an underlying resource—e.g., hashing power for proof-of-work (PoW)-based protocols such as Bitcoin [Nak08] and Ethereum [But13]

We refer to our full version [BLZ21] for the complete formal proofs and definitions.

© International Association for Cryptologic Research 2021

T. Malkin and C. Peikert (Eds.): CRYPTO 2021, LNCS 12827, pp. 3–32, 2021.

https://doi.org/10.1007/978-3-030-84252-9_1

(before version 2.0), or stake in Proof-of-Stake (PoS)-based protocols such as Algorand, Ouroboros, and Snow White [KRDO17, BGK+18, CM19, DPS19]—is owned/contributed by parties who honestly run the protocol.

Although such an analysis is instrumental for understanding the properties and limitations of the analyzed ledgers and gaining confidence in their security, it does not take into account a fundamental property of such systems, namely that the ledger’s state is often associated with some monetary value and therefore the protocol’s security might rely on how profitable an attack might be. Thus, in addition to the classical cryptographic analysis of such systems, it is useful to analyze their so-called *economic robustness*, namely their level of protection or susceptibility to attacks by an incentive-driven (also called rational) attacker. Such an analysis can fortify the security of these systems by providing a fallback rational assumption, e.g., assuming an incentives model of the attacker, security is maintained even when certain cryptographic assumptions fail, or indicate that the proven security is fragile by pointing out natural incentives that lead to violating the security assumptions. Additionally, it can offer a higher resolution picture of the systems guarantees—e.g., its tendency to decentralize [BKKS20]—and/or more realistic estimates of the parameters associated with its security properties—e.g., relation between the density of honest blocks (that is, the chain-quality parameter [GKL15]) and the properties of the communication network [ES14, NKMS16]. Perhaps, even more interesting, it can offer insight on the system’s behavior when the main (cryptographic) assumption fails, e.g., when the attacker controls a 51% fraction of the underlying resource of the blockchain protocol.

Motivated by the recent (repeated) 51% double-spending attacks that have drained millions of dollars from popular blockchain-based cryptocurrencies, we devise a game-theoretic analysis of such attacks for Nakamoto-style systems, e.g., Bitcoin, Bitcoin Cash/Gold, Ethereum (Classic), etc. We use the adaptation of the rational protocol design (RPD) framework by Garay *et al.* [GKM+13] to blockchains, which was recently proposed by Badertscher *et al.* [BGM+18], to analyze the utility of an attacker against these systems as a function of their basic parameters.

A central question to the relevance for practice of any game-theoretic analysis is to what extent the model and assumed utilities capture the incentives of real world attacks. Indeed, if the utilities are disconnected from reality, they can lead to counter-intuitive statements. We demonstrate an instance of such an artifact in [BGM+18] and propose a different class of utilities which is both natural and avoids this artifact. We validate our utility against a range of security parameters matching those of Ethereum Classic, a PoW-based system that fell victim to 51% double-spending attacks. We observe that when the payoff for double-spending is high, attacking is indeed a dominating strategy. That is, predictions of our utility choice match reality. We then use our framework to devise a generic tuning of one of the core parameters of such blockchains—namely, the number `cutOff` of most-recent blocks needed to be dropped to achieve the so-called common-prefix property with parameter `cutOff` (cf. [BMTZ17, BGM+18, GKL15])—to

deter any attacks on consistency by a rational attacker with our utility. Stated differently, we show how an incentive model can serve, possibly in addition to cryptographic assumptions, to find a robust protocol parameterization. This thereby demonstrates how our model and analysis can be used to improve the economic robustness of such blockchains, and offers a guide to how to “patch” such protocols to avoid future occurrences.

1.1 Related Literature

A number of works have focused on a rational analysis of decentralized ledgers and cryptocurrencies (e.g., [Ros11, CKWN16, ES14, Eya15, SBBR16, SSZ16, LTKS15, TJS16, NKMS16, PS17, GKW+16] to mention some). Typically, these works abstract away the computational aspects of cryptographic tools (signatures, hash-functions, etc.) and provide a game which captures certain aspects of the execution that are relevant for the rational analysis. In contrast, RPD uses a cryptographic simulation-based framework to incorporate these computational considerations into the analyzed game, ensuring that predictions about attacker behavior hold for the actual protocol and not only for an idealized version (unless the idealization is obtained via a cryptographic composition argument such as UC). Incorporating such computational considerations within a rational treatment is highly non-trivial (see [GKM+13, CCWr20] for a discussion). We discuss the RPD framework in more detail in the following section.

The term *51% (double-spending) attack* is defined in [Inv] as an attack where the adversary gains any majority (not necessarily just 51%) of mining power and reverses transactions in order to double-spend its coins, often by creating a deep fork in the chain. The site CoinDesk keeps track of news of 51% attacks [Coia], of which there are quite many: most recently, Verge suffered an attack with 200 days worth of transactions erased in Feb, 2021. Also recently, Ethereum Classic suffered three 51% attacks in the same month of August, 2020, prompting a solution called MESS to mitigate such attacks which still may not provide robust security [Coib]. Other recent victims of such attacks include well-known coins such as Bitcoin Gold (Jan 2020), and Bitcoin Cash (May 2019). A major avenue of 51% double-spending attacks is the use of rented hash power [For]. The site <https://www.crypto51.app/> gives rough estimates on the vulnerability of different coins, based on whether 51% of hashing power can be rented via a service called Nicehash. In some cases, e.g. Bitcoin Gold, it is estimated to only cost a few hundred dollars to have 51% of hashing power for 1 h.

Previous works have considered the ability of blockchain protocols to recover from 51% attacks. In [AKWW19], conditioned on honest majority being satisfied on expectation, Bitcoin was proven to be resilient against a (temporary) dishonest majority. In [BGK+20], no such condition is assumed and the authors give concrete recovery bounds as a function of the actual power of the adversary (captured as a budget to go over majority hashing power). We use the latter work for our analysis of the blockchain’s security against incentive-driven attackers.

The profitability of 51% double-spending attacks have also been analyzed in previous works. The work of [Bud18] explores these attacks through an economics

perspective, and leaving the cost of the attack as a parameter that is computed via simulations. The work of [JL20] computes probability of attack by modeling attacks as random walk of two independent Poisson counting processes (PCPs). In comparison, our rational analyses are done in the Rational Protocol Design (RPD) framework, where a fork is formally defined as a command in a UC ledger functionality. Another technique proposed is the Markov Decision Process (MDP) model, which is used by both [GKW+16] and [HSY+21]. In this model, the adversary takes a series of actions relevant to double-spending: adopting or overriding the honest party’s chain, waiting, or stopping. Solving the MDP allows these works to reason about the optimal double-spending adversary. While we do not analyze an optimal double-spending adversary, our model is more general. We do not restrict the actions of the adversary, which allows us to analyze conditions under which the protocol is secure against attacks on consistency by *any* incentive-driven adversary. Moreover, since standard MDP solvers cannot solve infinite state MDPs, the MDP is restricted to only consider situations where the chain length is less than some length c [GKW+16].

1.2 Our Results

We start by devising a utility in RPD which naturally captures the incentives of an attacker to provoke a double-spending attack. To this direction, we observe that the utility considered in [BGM+18] does not capture such an incentive. Intuitively, the reason is that the utility in [BGM+18] essentially only considers incentives related to the consensus layer of the protocol. This means that an attacker is rewarded when successfully mining a block, but is not rewarded depending on the block contents—i.e. what kinds of transactions are in the block. Their extension to a utility function to include transaction fees does not apply to double-spending attacks. In this case, the (only) reason to attack the blockchain stems from the existence of a super-polynomial transaction fee, and assuming a moderate range of fees, no incentive to attack is present. We discuss why super-polynomial quantities are generally problematic in Sect. 4. It follows from [BGM+18] that the attacker with these utility functions (and assuming moderate transaction fees) has no incentive to fork over mining honestly. Yet, looking at real-life double-spending attacks, this is clearly not the case. To capture double-spending, we introduce a special payoff that the attacker receives when successfully creating a deep-enough fork (i.e., orphans a sufficiently long valid chain). Intuitively, this payoff corresponds to the utility that the attacker receives when it double-spends by replacing the orphaned chain with his own.

Perhaps counter-intuitively, when analyzing Bitcoin¹ with this extended utility function, the attacker is still indifferent between forking and honest mining. We demonstrate this artifact and pinpoint the reason for it: Intuitively, the utility function from [BGM+18] (with or without the extra payoff for forking) rewards

¹ Our analysis uses Bitcoin as a representative example of Nakamoto-style blockchain ledgers, but similarly any blockchain protocol which realizes the ledger from [BMTZ17, BGK+18] could be analyzed.

the attacker by the same amount in all rounds in which it creates (mines) a block. This means that given any adversary that provokes a fork, there is always an honest-mining adversary who achieves more utility without forking by simply accumulating block rewards over a longer period of time. We distill the source of this issue in a property which we call *unbounded incentives*, and demonstrate that any utility which satisfies this property will make any deviation from passive mining a weakly *dominated* strategy.

We then devise a revision of this utility class which allows us to avoid the above counter-intuitive artifact. This utility, which satisfies a property we term *limited horizons*—a strong negation of unbounded incentives—has the property that the (actual) rewards of an adversary mining a block diminish with time. This is a natural way to avoid reasoning about extremely “long-lived” adversaries, i.e., that take decisions based on payoffs too far in the future, and captures features which are well-known in utility theory [Ber54]—intuitively, earning \$10 today is more attractive than \$1 million in 100 years, an example of the “St. Petersburg Paradox”. We next turn in analyzing the profitability of 51% double-spending attacks, by showing how our revised utility can actually capture them. We provide a range of payoffs for double-spending which would incentivize an attack. Then we visualize our result using concrete parameters estimated from those of Ethereum Classic, for which performing the attack is indeed a dominant strategy. This demonstrates that the above result can explain, in a game-theoretic framework, how recent victims of 51% attacks are vulnerable.

Finally, we discuss whether and how the blockchain protocol can be tuned so that such 51% double-spending attacks are deterred. In fact, we provide a much stronger tuning, which deters attacks on consistency by any incentive-driven adversary. The tuning depends on the costs (e.g. electricity or cost to rent hashing power), positive payoffs (e.g. block rewards and payoff for causing a fork, from double-spending or otherwise), and protocol parameters (e.g. the difficulty of creating a block). Intuitively, for any combination of these parameters, we show how the window size of the underlying blockchain protocol can be adjusted so that it is not rational for the attacker to perform this attack. At the core of this results is a lemma that relates the incentive model to an attack pattern, which coupled with the self-healing properties of Nakamoto-style PoW, leads to the desired estimate of a safe parameter. We view this as a demonstration that game theory can aid us in fortifying blockchains even when assumptions made by the cryptographic analyses fail.

2 Preliminaries

2.1 The Bitcoin Backbone Protocol

The abstraction of the Bitcoin protocol that is used in the cryptographic literature is known as the *Bitcoin backbone protocol* [GKL15, PSs17, BMTZ17] which we denote by Π^B . In this abstraction, Bitcoin is modeled as a round-based protocol, where a number of participants (the miners) are connected via a multicast

network with bounded delay Δ (unknown to the protocol). In every round, each party adopts the longest chain $\mathcal{C} = B_0 || \dots || B_k$ of block B_i (connected by hash-pointers) it has received so far, where B_0 is the unique genesis block of the system. Each party tries to extend this longest chain an by additional block, via running the PoW-lottery: an extension of chain \mathcal{C} by a new block B_{k+1} can only be valid, if its hash $H(B_{k+1})$ belongs to a dedicated small portion of the output domain of the function (typically, the hash must have a lot of leading zeros). In such analyses, the hash function is modeled using a random-oracle functionality \mathcal{F}_{RO} that returns uniform values upon each query. Therefore, when extending the chain, each party makes a certain number of *mining queries* per round (that is, RO-queries with candidate blocks B_{k+1} containing a random nonce to obtain the hash) and we call a mining query *successful*, if the output is below the threshold. In the setting with fixed PoW difficulty, we can assign a success probability p to each such mining query. Finally, if a miner is successful, it will send the new chain over the multicast network to all other miners.

Cryptographic Security. The main security guarantee² proven for the Bitcoin protocol is eventual *consistency*: every block that is deep enough can be considered immutable and only the most recent, `cutOff` number of blocks might be transient. This `cutOff`-consistency (where the cutoff parameter is often left implicit if clear from context) guarantee states that at any point in time, the prefix of \mathcal{C} consisting of $|\mathcal{C}| - \text{cutOff}$ blocks is common to all honest miners:

Definition 1 (Consistency). *Let $\mathcal{C}_1 \preceq \mathcal{C}_2$ denote the prefix-of relation, then the consistency guarantee (with parameter `cutOff`) states that at any two points in time $a \leq b$ in an execution, where party P at round a holds chain \mathcal{C}_1 and party P' at round b holds chain \mathcal{C}_2 , we have that $\mathcal{C}_1|_{\text{cutOff}} \preceq \mathcal{C}_2$, where the notation $\mathcal{C}|_k$ denotes the prefix of \mathcal{C} obtained by removing the most recent k blocks (and if k exceeds the length of \mathcal{C} , it is defined to correspond to the genesis block).*

In the cryptographic setting (without incentives), such a guarantee only holds if we restrict the adversary to have a minority of mining power. That is, given $n_a^{(r)}$ and $n_h^{(r)}$ denote the numbers of adversarial and honest mining queries in round r , respectively, then the protocol $\Pi^{\mathfrak{B}}$ is secure if in any round r the inequality $n_a^{(r)} < \theta_{\text{pow}} \cdot n_h^{(r)}$ holds, with $\theta_{\text{pow}} := (1 - p)^{(2\Delta+1)\text{T}_{\text{ub}}}$ being the well-established security threshold for Bitcoin (often stated in its linear approximation $1 - 2(\Delta + 1)p\text{T}_{\text{ub}}$) [GKL15, PSs17, BMTZ17], where the quantity T_{ub} denotes the upper bound on the number of mining queries per round. Throughout this work, we work in the so-called *flat model* of Bitcoin for notational simplicity [GKL15, BGM+18], where each miner gets one mining query per round (and the adversary’s power is the number of corrupted miners). We note that sometimes it is convenient to assume a lower bound T_{lb} on the number of mining queries (a.k.a. participation) per round, in particular when arguing about the

² While other security guarantees exist, such as *chain quality*, our focus in this paper is consistency.

guaranteed growth of the blockchain over time in combination with the security threshold. Finally, we point out that even if there are no adversarial players, an upper bound T_{ub} on the number of queries is necessary for security in the fixed difficulty setting, when aiming for a common prefix guarantee for some target parameter cutOff . As the failure probability of Bitcoin becomes negligible as a function of cutOff (more precisely, the relevant factor is of the order $2^{-\Omega(\text{cutOff})}$), we often treat it as a (of course polynomial-bounded) function $\text{cutOff}(\kappa)$ of a security parameter κ , and (in symbolic notation) $\text{cutOff} = \omega(\log(\kappa))$ is at least required to obtain a negligible probability of a failure.

Bitcoin Backbone and UC. The RPD framework is based on the UC framework. As such, the above Bitcoin backbone protocol $\Pi^{\mathbb{B}}$ is seen as a UC protocol as in [BMTZ17], where it is proven to UC-realize a strong transaction ledger functionality $\mathcal{G}_{\text{LEDGER}}$ under the honest majority assumption. We give here just the explanation of how the ideal consistency guarantee looks like: the functionality $\mathcal{G}_{\text{LEDGER}}$ ensures that at any point in time, there is only one unique ledger state (sequences of transactions packed in blocks), where the state is append-only (that is, whatever appears as a block in the state is immutable). Furthermore, different honest parties see different prefixes of this state, with the guarantee that these *views* are increasing and within a window of windowSize (a ledger parameter) blocks from the tip of the state. Note that the cut-off parameter of Bitcoin corresponds exactly to the size of that window in the realized ledger $\mathcal{G}_{\text{LEDGER}}$. More precisely, whenever Bitcoin satisfies Definition 1, then the above mentioned correspondence holds and the ledger state is a single chain of blocks [BMTZ17].

In UC, the protocol $\Pi^{\mathbb{B}}$ assumes a couple of hybrid functionalities. First, the round-based structure is achieved using UC-synchronous tools (assuming a clock functionality), a network, and a random oracle, where restrictions on the mining queries can be captured by functionality wrappers restricting the number of RO evaluations, e.g. [BMTZ17, GKO+20]. One extremely helpful aspect of UC in the context of RPD is the compatibility with the composition theorem [GKM+13]. In this work this is leveraged as follows. The Bitcoin backbone $\Pi^{\mathbb{B}}$ admits a modular structure that isolates the lottery aspect as a submodule of the system. Technically, the proofs in [BMTZ17, PSs17] show that whenever the PoW-lottery UC-realizes the *state exchange* functionality \mathcal{F}_{STX} (in [PSs17] the related concept is called $\mathcal{F}_{\text{tree}}$), the Nakamoto-style longest chain rule protocol (under the above honest-majority security threshold) realizes the ledger. This intermediate step is important due to two things: first, it models an idealized mining process where each mining query is an independent Bernoulli trial with success probability p (and hence abstracts away those real-life negligible probability events that would destroy independence), and second it abstracts away the low-level details of the chain structure (where e.g., “hash collisions” could cause disruptions). It is proven in [BMTZ17] that the proof-of-work layer of Bitcoin (in the random oracle model) UC-realizes \mathcal{F}_{STX} . Moreover, since it only abstracts the lottery part of the system, this realization does not depend on any security threshold. We can therefore leverage composition when analyzing the utilities of Bitcoin and work with the idealized lottery directly.

2.2 Rational Protocol Design

The Rational Protocol Design framework (RPD) allows us to analyze the security of the blockchain without assuming honest majority. Although consistency and other security properties are lost if an attacker can arbitrarily break honest majority, assuming attackers are *rational* offers an alternate method of limiting his actions. That is, although the attacker is free to act in any way (e.g. corrupt more than majority hashing power), he will only do so if it is profitable. Building on [BGM+18], our analysis is based on the Rational Protocol Design (RPD) framework introduced in [GKM+13]. RPD analyzes the security of protocols, such as Bitcoin, with respect to an incentive-driven adversary. In this model, a protocol designer D plays an *attack game* \mathcal{G} with an attacker A . First, the designer D comes up with a protocol Π . Then, the attacker A —who is informed about Π —comes up with an adversarial strategy \mathcal{A} to attack Π . The utility of the attacker (resp. designer) is then defined on the *strategy profile* (Π, \mathcal{A}) , and is denoted $u_A(\Pi, \mathcal{A})$ (resp. $u_D(\Pi, \mathcal{A})$). In this work, we focus on the attacker’s utility $u_A(\Pi, \mathcal{A})$.

The game \mathcal{G} is defined with respect to an attack model $\mathcal{M} = (\mathcal{F}, \langle \mathcal{F} \rangle, v_A, v_D)$. \mathcal{F} is the functionality which the designer would like to implement such as a ledger that provides certain ideal guarantees as described above. However, when certain assumptions, e.g. honest majority for Bitcoin, are not met (which as stated above we explicitly do not want to demand *a priori*), we cannot hope to get \mathcal{F} . Instead, the designer D ’s protocol Π (in our case, the Bitcoin protocol Π^B) only implements a weaker functionality. This weaker functionality that Bitcoin implements when lifting the honest majority assumption is proven to be $\mathcal{G}_{\text{WEAK-LEDGER}}^B$ in [BGM+18] and provided in our full version [BLZ21] for completeness. Intuitively, the weak ledger is derived from the stronger version [BMTZ17] by introducing a few weaknesses. For example, it allows the adversary to fork the ledger state and hence allows it to break consistency (this event corresponds to a deep reorganization of the blockchain in the real world). This is allowed by the FORK command in $\mathcal{G}_{\text{WEAK-LEDGER}}^B$. Given the views of the simulator and environment in an ideal world execution, the value functions v_A and v_D assign payoffs to the attacker and designer respectively, when certain events happen in the views, such as when the simulator forks the blockchain via $\mathcal{G}_{\text{WEAK-LEDGER}}^B$. Finally, utilities u_A and u_D are functions of payoffs (defined with v_A and v_D) of simulators that can simulate \mathcal{A} in Π in the environment \mathcal{Z} . Looking ahead, the goal of RPD is to find conditions under which a rational attacker would not invoke the weaknesses of $\mathcal{G}_{\text{WEAK-LEDGER}}^B$ (e.g., it is too costly to perform an attack). For example, if under a class of utilities, no rational attacker invokes the FORK command, then we essentially obtain a stronger ledger (i.e., the same except that this command is absent and hence the ledger state remains a unique chain) against attackers incentivized by this class of utilities.

2.3 Utility of the Attacker From [BGM+18]

We detail the attacker’s utility in [BGM+18], which in the RPD framework captures the expected payoff of a particular adversarial strategy \mathcal{A} in a given protocol Π (in our case $\Pi = \Pi^{\mathcal{B}}$). This payoff is calculated based on different *events* that occur in the real execution and the corresponding ideal experiment where a black-box simulator is attempting to simulate this adversarial strategy.

Specifically, the work of [BGM+18] considers the following events:

1. Event $W_{q,r}^{\mathcal{A}}$, for each pair $(q, r) \in \mathbb{N}^2$: The simulator simulates q mining queries by the adversary in round r of the simulated execution.
2. Event $I_{b,r}^{\mathcal{A}}$, for each pair $(b, r) \in \mathbb{N}^2$: The simulator inserts b blocks into the state of the ledger in round r , such that all these blocks were previously queries to the (simulated) random oracle by the adversary. Informally, this event occurs when an honest party views these blocks as “confirmed” (part of his own ledger state).

A different payoff is associated with each event. In order to make q mining queries and invoke event $W_{q,r}^{\mathcal{A}}$, the attacker must pay $q \cdot \text{mcost}$, where mcost is the cost of making a mining query (e.g. electricity cost per hash query). When b blocks made by the adversary are inserted into the ledger and event $I_{b,r}^{\mathcal{A}}$ occurs, the attacker receives payoff $b \cdot \text{breward} \cdot \text{CR}$. Here breward is the reward for making a block in the currency of the blockchain (e.g. Bitcoins), and CR is an exchange rate to the same currency used for mcost (e.g. USD).

Then, [BGM+18] defines the following attacker’s utility for a strategy profile (Π, \mathcal{A}) . Let $\mathcal{C}_{\mathcal{A}}$ denote the set of simulators that can emulate an adversary \mathcal{A} in the ideal world with access to the weaker ledger functionality $\mathcal{G}_{\text{WEAK-LEDGER}}^{\mathcal{B}}$, and \mathcal{Z} denote an environment. The *real payoff* of an adversary \mathcal{A} attacking the protocol is defined as the minimum payoff over all simulators in $\mathcal{C}_{\mathcal{A}}$. If $\mathcal{C}_{\mathcal{A}} = \emptyset$ (there are no simulators that can simulate \mathcal{A}) then $u_{\mathcal{A}}(\Pi, \mathcal{A}) = \infty$ by definition. Then, the utility $u_{\mathcal{A}}(\Pi, \mathcal{A})$ is the real payoff, maximized over all possible environments \mathcal{Z} (we assume for simplicity that environments are closed and run in polynomial time in the security parameter [Can01]).

$$u_{\mathcal{A}}(\Pi, \mathcal{A}) := \sup_{\mathcal{Z} \in \text{ITM}} \left\{ \inf_{\mathcal{S}^{\mathcal{A}} \in \mathcal{C}_{\mathcal{A}}} \left\{ \sum_{(b,r) \in \mathbb{N}^2} (b \cdot \text{breward} \cdot \text{CR} \cdot \Pr[I_{b,r}^{\mathcal{A}}]) \right. \right. \quad (1) \\ \left. \left. - \sum_{(q,r) \in \mathbb{N}^2} q \cdot \text{mcost} \cdot \Pr[W_{q,r}^{\mathcal{A}}] \right\} \right\}.$$

The work of [GKM+13] introduces the following notion of security against incentive-driven adversaries: No matter the utility achieved by an adversary \mathcal{A} running the protocol Π in the real world, there exists an adversary \mathcal{A}' running the dummy protocol with access to the ideal functionality \mathcal{F} that achieves the same or better utility. In other words, even the best adversary attacking Π , cannot achieve better utility than one who does not invoke any of the “bad

events” in $\langle \mathcal{F} \rangle$. Note that here \mathcal{F} can be any strengthening of its weaker version. For example, the weak ledger without the option to break consistency would be a strengthening of $\mathcal{G}_{\text{WEAK-LEDGER}}^{\mathbb{B}}$ in which case attack-payoff security implies that there is no incentive (even for a majority-controlling adversary) to create a fork (that is, a deep reorganization) even though he technically could be able to.

Strictly speaking, the utilities are also functions in the security parameter κ (the environment obtains the parameter as input in UC) but we omit it for notational simplicity. We note that as functions in the security parameter κ , the asymptotic behavior of the involved functions is the relevant aspect.

Definition 2 (Attack payoff security [GKM+13]). *Let $\mathcal{M} = (\mathcal{F}, \langle \mathcal{F} \rangle, v_{\mathcal{A}}, v_{\mathcal{D}})$ be an attack model inducing utility $u_{\mathcal{A}}$, and let $\Phi^{\mathcal{F}}$ be the dummy \mathcal{F} -hybrid protocol. A protocol Π is attack-payoff secure for \mathcal{M} if for all \mathcal{A} , there is an \mathcal{A}' such that $u_{\mathcal{A}}(\Pi, \mathcal{A}) \leq u_{\mathcal{A}}(\Phi^{\mathcal{F}}, \mathcal{A}') + \text{negl}(\kappa)$*

This notion of attack-payoff security does not necessarily mean an incentive-driven adversary will honestly follow the protocol—there is no restriction on the honesty of the actions of \mathcal{A}' in the above definition. To capture this stronger requirement in the context of Bitcoin, we also consider a stronger notion introduced by [BGM+18]: the attacker is incentivized to always choose a *front-running, passive-mining* adversary over any (potentially malicious) strategy. Informally, this passive adversary behaves exactly like an honest party (mining with all his hashing power and releasing a block he has found immediately), except the adversary’s messages are always delivered before the honest parties’ (front-running). Front-running gives the adversary an advantage since if an adversary’s block is concurrently competing with an honest party’s block to be appended to the longest chain, the adversary always wins.

Definition 3 (Front-running, passive-mining adversary [BGM+18]). *The front-running adversarial strategy $\mathcal{A} \in \mathbb{A}_{\text{fr}}$ is specified as follows: Upon activation in round $r > 0$, \mathcal{A} activates in a round-robin fashion all its (passively) corrupted parties, say p_1, \dots, p_t . When corrupt party p_i generates some new message to be sent through the network, \mathcal{A} immediately delivers it to all its recipients. In addition, upon any activation, any message submitted to the network $\mathcal{F}_{\text{N-MC}}$ by an honest party is maximally delayed.*

$\Pi^{\mathbb{B}}$ was proved to be strongly attack-payoff in [BGM+18] for the utility in Eq. 1. Informally, a protocol is strongly attack-payoff secure if there is always a passive adversarial strategy that is at least as good as any malicious strategy. In this work, we are also interested in the case where security does not hold: we say an adversary \mathcal{A} *breaks* strong attack-payoff security if $u_{\mathcal{A}}(\Pi, \mathcal{A})$ exceeds $u_{\mathcal{A}}(\Pi, \mathcal{A}')$ for any $\mathcal{A}' \in \mathbb{A}_{\text{fr}}$, by a non-negligible amount.

Definition 4 (Strongly attack-payoff secure [BGM+18]). *A protocol Π is strongly attack-payoff secure for attack model \mathcal{M} if there is a $\mathcal{A}' \in \mathbb{A}_{\text{fr}}$ such that for all \mathcal{A} , $u_{\mathcal{A}}(\Pi, \mathcal{A}) \leq u_{\mathcal{A}}(\Pi, \mathcal{A}') + \text{negl}(\kappa)$*

In our work, we will follow the approach from [BGM+18] that simplifies the proofs when analyzing the utilities from mining in the protocol $\Pi^\mathbb{B}$ by utilizing the composition theorem of RPD. As explained above, instead of analyzing the probabilities of payoff-inducing events for $\Pi^\mathbb{B}$ which uses the random oracle as the lottery, one can analyze probabilities for the *modular* ledger protocol w.r.t. an idealized lottery that makes use of the state exchange functionality \mathcal{F}_{STX} (for completeness, defined in full version [BLZ21]). In more detail: when a party (or the adversary in the name of a corrupted party) wishes to extend a chain, they would invoke \mathcal{F}_{STX} with a SUBMIT-NEW command, which performs a coin toss and informs him whether he is successful. If the party is successful, the functionality includes this new chain into a tree data structure and allows the party to multicast this new chain with a SEND command; this multicasting is done automatically for honest parties. Due to the correspondence of RO queries in the Bitcoin protocol and the SUBMIT-NEW-commands in the modularized Bitcoin protocol [BMTZ17], the events defined for $u_{\mathbb{A}}^\mathbb{B}(\Pi, \mathcal{A})$ (for the full Bitcoin protocol) above remain valid and meaningful also in this hybrid world, because the black-box simulator for the overall Bitcoin protocol simulates one RO-query (as a reaction to an input by a corrupted party) whenever the (black-box) simulator for the modular ledger protocol simulates one SUBMIT-NEW-command, as a reaction to the corresponding input by the same party [BGM+18].

3 Artifacts of Unbounded Incentives

In this section, we discuss an artifact of the utility function Eq. 1, which we will eliminate in the next section. Concretely, we prove that this RPD utility is inappropriate to capture the most realistic situation of attackers that attack the system, e.g., attempt a fork to profit from double-spending. To do so, we prove Lemma 1 and 2, which roughly show this surprising fact: if running the protocol (semi-)honestly is profitable in expectation, then there is no incentive for an adversary to fork. The intuitive reason for this is clear: Any fixed payoff for forking incurred by the adversary can be offset by an adversary who runs slightly longer (and still polynomially long) but does not fork. This, however, is an artifact of the asymptotic definition and does not reflect real-world incentive-driven attack scenarios, where mining is anticipated to be profitable—otherwise no one would mine—but attackers still perform forking attacks (in particular, in order to double-spend coins). We distill a property of the utility from [BGM+18] that is the reason this artifact, which we call *unbounded incentives*, and prove that any utility satisfying this property will suffer from the same artifact. Looking ahead to the following section, we will propose a natural adaptation of this utility function that does not suffer from the above artifact (and where in particular the duration of an attack actually starts to matter).

3.1 Demonstrating the Artifact

Let us first consider the straightforward adaptation of the utility from Eq. 1 to model the payoff (e.g. double-spending) an adversary gains by forking the ledger.

Define the event K as: There is a round r where the simulator uses the FORK command of the weak ledger functionality $\mathcal{G}_{\text{WEAK-LEDGER}}^{\mathbb{B}}$ (see [BLZ21] for formal definition) that allows the simulator to invoke a fork. Let $\mathbf{fpayoff}$ be the payoff for invoking the fork. Then, the utility $u_{\mathbf{f}}$ becomes:

$$u_{\mathbf{f}}(\Pi, \mathcal{A}) := \sup_{Z \in \text{ITM}} \left\{ \inf_{S^{\mathcal{A}} \in \mathcal{C}_{\mathcal{A}}} \left\{ \sum_{(b,r) \in \mathbb{N}^2} b \cdot \mathbf{breward} \cdot \mathbf{CR} \cdot \Pr[I_{b,r}^{\mathcal{A}}] \right. \right. \\ \left. \left. - \sum_{(q,r) \in \mathbb{N}^2} q \cdot \mathbf{mcost} \cdot \Pr[W_{q,r}^{\mathcal{A}}] + \mathbf{fpayoff} \cdot \Pr[K] \right\} \right\}. \quad (2)$$

Below, we show that for the utility function $u_{\mathbf{f}}$ above, the Bitcoin protocol $\Pi^{\mathbb{B}}$ is strongly attack-payoff secure as long as mining is profitable. Our proof takes advantage of the artifact of unbounded incentives: informally, first we show that the payoff of any polynomial-run-time adversary \mathcal{A} is bounded by a polynomial $p(\kappa)$ of the security parameter; then, we show that there is a passive, front-running adversary whose run-time is also polynomial (albeit bigger than that of \mathcal{A}), and who achieves at least $p(\kappa)$ utility.³

Lemma 1 (Attack payoff security with forking). *Let $T_{\text{ub}} > 0$ be the upper bound on total number of mining queries per round, $p \in (0, 1)$ be the probability of success of each mining query, and $\mathbf{cutOff} = \omega(\log(\kappa))$ be the consistency parameter. Let \mathcal{M} be a model whose induced utility $u_{\mathbf{f}}$ has parameters $\mathbf{fpayoff}, \mathbf{breward}, \mathbf{CR}, \mathbf{mcost} \geq 0$. The Bitcoin protocol $\Pi^{\mathbb{B}}$ is strongly attack-payoff secure in \mathcal{M} if $p \cdot \mathbf{breward} \cdot \mathbf{CR} - \mathbf{mcost} > 0$.*

3.2 A First Attempt to Eliminate the Artifact

Although we proved that Bitcoin is strongly attack payoff secure even with a payoff for forking, this is actually not a good sign, as this result does not reflect reality. In reality, attackers do fork blockchains to gain profit via e.g. double-spending transactions. Thus, the fact that we can prove Lemma 1 means that there must be a problem with our assumptions.

Why were we able to prove Lemma 1? It turns out the utility function we used has the weakness that it considers an attacker who does not care about the ephemeral payoff for forking—he can simply obtain more utility via block rewards if he just put in a bit more hashing power for mining. Thus, somewhat counter-intuitively, to model incentives for forking attacks, we must consider utilities that limit the amount of mining an attacker can do.

A first natural instinct may be to incorporate in the utility the (often substantial) initial investment (e.g. cost of buying mining rigs) an attacker must

³ We note that for the simple utility function presented in [BGM+18] other proof techniques could conclude attack-payoff security without the runtime-extension argument. The main point here is to demonstrate the importance of considering the attack duration in the utility function.

make before being able to participate in the blockchain protocol. This turns out to be not only a natural extension, but also a very simple one. Concretely, we capture this investment as *cost of party corruption*: in order to use party for mining, the adversary needs to corrupt him, which corresponds to acquiring its mining equipment. Formally, for each $g \in \mathbb{N}$ define C_g^A as follows: The maximum number of corrupted parties at any round is g . Let $\text{ccost}(g)$ be the cost of event C_g^A , i.e. corrupting g parties. Then we define the utility function:

$$u_{f,c}(\Pi, \mathcal{A}) := \sup_{\mathcal{Z} \in \text{ITM}} \left\{ \inf_{\mathcal{S}^A \in \mathcal{C}_A} \left\{ \sum_{(b,r) \in \mathbb{N}^2} b \cdot \text{breward} \cdot \text{CR} \cdot \Pr[I_{b,r}^A] \right. \right. \\ - \sum_{(q,r) \in \mathbb{N}^2} q \cdot \text{mcost} \cdot \Pr[W_{q,r}^A] \\ + \text{fpayoff} \cdot \Pr[K] \\ \left. \left. - \sum_{g \in \mathbb{N}} \text{ccost}(g) \cdot \Pr[C_g^A] \right\} \right\}. \quad (3)$$

Interestingly, as we see below, this natural extension is still insufficient to align the model with the reality that forking attacks occur. Indeed, even with this additional cost, we can still prove a result similar Lemma 1. Concretely, the following lemma shows that for $u_{f,c}$ above, we can prove the statement as the one in Lemma 1 about Π^B being attack-payoff secure by again exploiting the artifact of unbounded incentives.

Lemma 2 (Attack payoff security with forking, with cost of corruption). *Let $T_{\text{ub}} > 0$ be the upper bound on total number of mining queries per round, $p \in (0,1)$ be the probability of success of each mining query, and $\text{cutOff} = \omega(\log(\kappa))$ be the consistency parameter. Let \mathcal{M} be the model whose induced utility $u_{f,c}$ has parameters $\text{fpayoff}, \text{breward}, \text{CR}, \text{mcost} \geq 0$, $\text{ccost}(\cdot) : \mathbb{N} \rightarrow \mathbb{R}^+$. The Bitcoin protocol is strongly attack-payoff secure in \mathcal{M} if $p \cdot \text{breward} \cdot \text{CR} - \text{mcost} > 0$.*

3.3 The Source of the Artifact: Unbounded Incentives

Distilling the issue in above lemmas, we observe that that as long as the adversary keeps accumulating rewards as rounds are added to the protocol—i.e., mining remains profitable—he does not care about the payoff for forking: there always exists a polynomial-time, passively mining strategy that simply gains the same amount of utility by mining a bit more. However, not only do real-life attackers in fact profit from forks, even the assumption on the profitability of mining forever is unrealistic: any attacker is at least limited in time by e.g. the anticipated age of the universe, and cannot, in practice, keep accumulating utility in perpetuity.

Thus, to make accurate prediction about the attackability of a blockchain protocol the utility function must exclude the eternal profitability of passive mining. We generalize this intuition, by defining the notion of *unbounded incentives*:

a utility function has *unbounded incentives* if there is an adversarial strategy $\mathcal{A} \in \mathbb{A}_{\text{fr}}$ such that for any polynomial $h(\kappa)$, \mathcal{A} can gain better payoff than $h(\kappa)$. (Conversely, we will say that a utility has bounded incentives if there is no such passive adversary.).

It is straightforward to verify that the utilities we have seen so far have unbounded incentives, which explains the effect of the artifact exploited in the above lemmas. In fact, in the following there is a simple argument for a generic statement about the strong attack-payoff security of utility functions that have unbounded incentives.

Lemma 3. *Let \mathcal{M} be a model inducing a utility function $u_{\mathcal{A}}$. Assume for any adversary \mathcal{A} , in any real execution of the protocol his payoff is polynomially-bounded.⁴ If $u_{\mathcal{A}}$ has unbounded incentives for a protocol Π , then Π is strongly attack-payoff secure for \mathcal{M} .*

4 An RPD Analysis of Forks

In this section, we will tune our utility function to avoid the issue of *unbounded incentives* isolated in the previous section. A straw man approach would be to make **fpayoff** a super-polynomial function of the security parameter. But this would imply a very unnatural assumption, which, intuitively, corresponds to ensuring that the polynomially-bounded adversaries are *always* incentivized to fork. This would have the opposite effect and introduce a different artifact: it would make attack-payoff security impossible, and making a 51% attack always a dominant strategy no matter the systems parameters, contradicting the observable fact that many blockchains have not fallen to 51% attacks.

Instead, we make **breward** a function of time, which captures e.g., inflation, or simply that the adversary only plans to stay in the system for a limited amount of time. We refer to this adaptation of $u_{\text{f},c}$ as u_{buy} :

$$u_{\text{buy}}(\Pi, \mathcal{A}) := \sup_{Z \in \text{ITM}} \left\{ \inf_{S^{\mathcal{A}} \in C_{\mathcal{A}}} \left\{ \sum_{(b,r) \in \mathbb{N}^2} b \cdot \text{breward}(r) \cdot \text{CR} \cdot \Pr[I_{b,r}^{\mathcal{A}}] \right. \right. \\ - \sum_{(q,r) \in \mathbb{N}^2} q \cdot \text{mcost} \cdot \Pr[W_{q,r}^{\mathcal{A}}] \\ + \text{fpayoff} \cdot \Pr[K] \\ \left. \left. - \sum_{g \in \mathbb{N}} \text{ccost}(g) \cdot \Pr[C_g^{\mathcal{A}}] \right\} \right\}. \quad (4)$$

We also define a version of this utility u_{rent} (formally defined in full version [BLZ21]), which models the attacker renting hashing queries by replacing

⁴ This is true for the utility function $u_{\mathcal{A}}^{\text{B}}$ in Eq. 1 (as well as the utility functions we will consider)—no adversary can get payoff that is superpolynomial in the run time of the execution.

`mcost` with parameter `rcost` (rent cost) and setting `ccost`(\cdot) = 0. Renting especially has been observed in real attacks, such as the August 2020 attacks on Ethereum Classic [For].

Note that while `breward` is a function of time, we let the cost of a mining query, that is `mcost/rcost`, remain constant. We do so to model the attacker’s anticipated monetary budget to launch and maintain an attack, such as the costs for renting a certain amount of hashing power (which are generally paid up-front), or cost of electricity (which realistically appears to be relatively stable). Further, the parameter `fpayoff` should be seen as an abstract excess payoff for the attacker arising from forking that is able to capture various use-cases. In the prototypical (double-spend) example where the attacker sells some coins for fiat currency and later tries to regain the coins with a successful attack, it corresponds to this extra fiat inflow gained prior to attacking the blockchain. We note that the utility functions could be tweaked to allow for all parameters to be time-dependent without changing the results qualitatively as long as the relations among the parameters required by the definitions and theorems (which are time-dependent in our treatment already) still hold.

To capture realistic utilities, we restrict to instances of our utility function which satisfy what we call *limited horizons* (Definition 5). Roughly, limited horizons constrains utilities by requiring that passive mining eventually becomes unprofitable. Recall that in light of the St. Petersburg Paradox discussed in the introduction, rational parties become increasingly reluctant to invest some monetary budget for potential rewards gained only later in a randomized process (e.g. due to uncertainty about the future or other specific utility-relevant considerations like relative inflation between several quantities). We cast this general idea as a rather simple condition based on our utility function.

After defining limited horizons, in Sect. 4.1, we will first address a technical challenge imposed when payoff-parameters in the utility functions are non-constant. Then, in Sect. 4.2 we show that limited horizons implies bounded incentives (i.e., the opposite of unbounded incentives) through Lemma 5. More precisely, limited horizon is a strong negation⁵ of unbounded incentives. Looking ahead, we will prove that when utilities have limited horizons, there is always a large enough payoff for forking such that (strong) attack-payoff security is broken. Informally, a utility function u_{buy} (resp. u_{rent}) has limited horizons if there is a time limit after which passive mining becomes unprofitable.

Definition 5 (Limited Horizons). *We say u_{buy} in Eq. 4 (resp. u_{rent} , formally defined in our full version), parameterized by $\text{breward}(\cdot) : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$, $\text{mcost}, \text{fpayoff} \geq 0$, and non-decreasing function $\text{ccost}(\cdot) : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ (resp. $\text{breward}(\cdot) : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$, $\text{rcost}, \text{fpayoff} \geq 0$) satisfies limited horizons (resp. limited horizons with renting) if $\text{breward}(\cdot)$ is a non-increasing function such that $\exists x \in \mathbb{N} : p \cdot \text{CR} \cdot \text{breward}(x) < \text{mcost}$.*

⁵ Note that the *strong negation* of an assertion A is one which implies $\neg A$, but is not necessarily implied by $\neg A$.

Remark. Technically, u_{rent} is a special case of the case of u_{buy} (since the utilities are the same if we set $\text{mcost} = \text{rcost}$ and set $\text{ccost}(\cdot) = 0$); however semantically they are different: rcost represents the cost of renting a hashing query, which usually is much higher than mcost which represents the cost (e.g. electricity) of an adversary mining with his own equipment. Nevertheless, to reduce redundancies in the technical sections, we will analyze the utility u_{buy} in Eq. 4 (with a general $\text{ccost}(\cdot)$, including when $\text{ccost}(\cdot) = 0$), and state the results for the renting case as corollaries.

4.1 Addressing Technical Issue of Non-constant Payoff for Block Rewards

In this section, we address a technical issue with considering a non-constant **breward**—recall that in limited horizons, **breward** is a non-increasing function of time/round number. By our definition (which follows that of [BGM+18]), the event $I_{b,r}^A$ happens when b blocks are placed into the ledger of some honest party. This is intuitive—the block reward should be given only when the block is “confirmed” to be in the ledger. However, there is a delay between when a block is broadcasted, and when it makes it into the common prefix of an honest chain. This delay is a random variable which depends on the amount of (honest and corrupt) hashing power in the protocol, the network delay, and the adversary’s strategy. Fortunately, we can lower and upper bound such a delay (which we denote by t_{lb}, t_{ub} respectively), as we show in the following lemma. This will in turn allow us to avoid the complication of analyzing when blocks enter the ledger state and instead analyze when locks broadcasted by the adversary to honest parties (whose events are easier to analyze). Note that we choose to analyze time-of-block-broadcast, instead of time-of-block-creation, since the adversary may choose to withhold successfully-mined blocks instead of broadcasting them immediately, making time-of-broadcast more suitable for incorporating such adversarial strategies.

We first define a useful quantity $t_\delta^\Delta(q)$. As we will see, this quantity, which is derived from the *chain growth* property of Nakamoto-style blockchains, is the maximum time for honest chains to grow by `cutOff` blocks, given that in each round there are at least q honest mining queries.

Definition 6 (Maximum time to grow `cutOff` blocks). For network delay Δ , and $p, \delta \in (0, 1)$, we denote $t_\delta^\Delta(q) := \frac{\text{cutOff}}{(1-\delta)^\gamma}$, where $\gamma := \frac{h}{1+h\Delta}$ and $h := 1 - (1-p)^q$.

Let $t_{lb} := 0$ and $t_{ub} := t_\delta^\Delta(T_{ub})$. Let $B_{b,r}^A$ denote the event: At round r , the adversary broadcasts b blocks made by parties that are corrupted at the time of the blocks’ creation, and which are part of the longest chain at round r . Let u_{buy}^h be u_{buy} except $\sum_{(b,r) \in \mathbb{N}^2} b \cdot \text{breward}(r) \cdot \text{CR} \cdot \Pr[I_{b,r}^A]$ (which considers time of block confirmation) is replaced with $\sum_{(b,r) \in \mathbb{N}^2} b \cdot \text{breward}(r + t_{lb}) \cdot \text{CR} \cdot \Pr[B_{b,r}^A] = \sum_{(b,r) \in \mathbb{N}^2} b \cdot \text{breward}(r) \cdot \text{CR} \cdot \Pr[B_{b,r}^A]$ (which considers time of block broadcast).

Similarly, let u_{buy}^l replace the same term in u_{buy} with $\sum_{(b,r) \in \mathbb{N}^2} b \cdot \text{breward}(r + t_{ub}) \cdot \text{CR} \cdot \Pr[B_{b,r}^A]$. (See full version [BLZ21] for formal definitions.)

The following lemma tells us that instead of analyzing the utility function defined on when a block is confirmed in the ledger we can instead approximate by only analyzing when a block is broadcasted. This will be helpful in our proof of Lemma 5 on the utility of the optimal front-running, passive adversary.

Lemma 4 (Translating time-of-block-confirmation to time-of-block-broadcast: u_{buy}^h and u_{buy}^l). *For any utility function satisfying limited horizons (in fact, we only require that $\text{breward}(\cdot)$ is a non-increasing function), satisfies the following: For all adversaries \mathcal{A} , and front-running, passive \mathcal{A}' ,*

$$\begin{aligned} u_{\text{buy}}(\Pi^B, \mathcal{A}) &\leq u_{\text{buy}}^h(\Pi^B, \mathcal{A}) + \text{negl}(\kappa) \quad \text{and} \\ u_{\text{buy}}(\Pi^B, \mathcal{A}') + \text{negl}(\kappa) &\geq u_{\text{buy}}^l(\Pi^B, \mathcal{A}'). \end{aligned}$$

Proof. The first inequality is obvious: By limited horizons, giving block rewards using time-of-block-broadcast (i.e., u_{buy}^h) gives the attacker a higher payoff.

The second inequality: Let the environment be one which maintains T_{ub} parties in each round after r . The bound follows then from the chain-growth lower bound which states the minimum chain length increase during a time period, depending on the honest parties' hashing power and the network delay (cf. [BMTZ17, PSs17]). This concludes the proof. \square

4.2 Optimal Utility of Front-Running, Passive Adversaries

We show in this section if a utility satisfies limited horizons, then it also satisfies bounded incentives. We do so by proving the following optimal utility of a passive, front-running adversary. We define u_{honest}^h and u_{honest}^l which, as we will see in Lemma 5 below, are the upper and lower bounds on the optimal utility obtained by a front running, passive adversary in Π^B .

Definition 7 (Bounds u_{honest}^h and u_{honest}^l for optimal front-running, passive adversary). *We define the quantity*

$$\begin{aligned} u_{\text{honest}}^h(\text{breward}, \text{CR}, \text{mcost}, \text{ccost}) \\ := g \cdot p \cdot \text{CR} \cdot \sum_{x=1}^t [\text{breward}(x + t_{lb}) - \text{mcost}] - \text{ccost}(g) \end{aligned}$$

with

$$\begin{aligned} t &:= \arg \max_{x \in \mathbb{N}} (p \cdot \text{CR} \cdot \text{breward}(x + t_{lb}) \geq \text{mcost}), \\ g &:= \arg \max_{g \in [0, T_{ub}]} (mg - \text{ccost}(g)), \\ \text{for } m &:= \sum_{x=1}^t (p \cdot \text{CR} \cdot \text{breward}(x + t_{lb}) - \text{mcost}), \end{aligned}$$

and the quantity

$$u_{\text{honest}}^1(\text{breward}, \text{CR}, \text{mcost}, \text{ccost}) \\ := g \cdot p \cdot \text{CR} \cdot \sum_{x=1}^t [\text{breward}(x + t_{ub}) - \text{mcost}] - \text{ccost}(g)$$

with

$$t := \arg \max_{x \in \mathbb{N}} (p \cdot \text{CR} \cdot \text{breward}(x + t_{ub}) \geq \text{mcost}), \\ g := \arg \max_{g \in [0, T_{ub}]} (mg - \text{ccost}(g)), \\ \text{for } m := \sum_{x=1}^t (p \cdot \text{CR} \cdot \text{breward}(x + t_{ub}) - \text{mcost}).$$

We simplify the above upper and lower bounds on the optimal front-running, passive adversaries as u_{honest}^h and u_{honest}^l , when the parameters to the utility function are clear from context. As discussed before, although we prove the optimal passive adversary for u_{buy} , the renting case for utility u_{rent} is a direct corollary by setting $\text{ccost}(\cdot) = 0$ and $\text{mcost} = \text{rcost}$.

Intuitively, the following lemma is established by proving that (1) due to limited horizons, there is a fixed time t after which an optimal passive adversary will not mine, and (2) it is optimal for a passive adversary to corrupt parties statically. Then, we can re-write the utility of a front-running, passive adversary as a function of his running time t , and the number of parties he corrupts g . Optimizing for t and g gives us the optimal utility of this passive adversary.

Lemma 5 (Optimal utility of a front-running passive adversary, for incentives with limited horizons). *Let $T_{ub} > 0$ be the upper bound on total number of mining queries per round, $p \in (0, 1)$ be the probability of success of each mining query, and $\text{cutOff} = \omega(\log(\kappa))$ be the consistency parameter. Given parameters such that u_{buy} satisfies limited horizons and protocol $\Pi^{\mathcal{B}}$, for \mathcal{A} the optimal adversary in \mathbb{A}_{fr} , $u_{\text{buy}}(\Pi^{\mathcal{B}}, \mathcal{A}) \leq u_{\text{honest}}^h + \text{negl}(\kappa)$ and $u_{\text{buy}}(\Pi^{\mathcal{B}}, \mathcal{A}) + \text{negl}(\kappa) \geq u_{\text{honest}}^l$*

This lemma directly implies that any utility with limited horizons also has bounded incentives.

5 Analyzing 51% Attacks

We can now utilize our above framework to analyze one of the most common types of forking attacks, known as 51% *double-spending* attack [Inv]. We analyze a range of parameters for utility functions with limited horizons, for which a 51% double-spending adversary breaks the strong attack-payoff security of protocol $\Pi^{\mathcal{B}}$ (formalized by Theorem 1). In more detail, first we will show a general lemma

relating the number of honest/adversarial hashes per round, to the time it takes to fork with a 51% double-spending attack (Lemma 6). Then, in Theorem 1 we will show that if the payoff for a successful attack (**fpayoff**) satisfies certain conditions, then an adversary performing a 51% double-spending attack achieves better utility than any passive-mining strategy. This **fpayoff** is quantified as a function of the parameters of the protocol and the utility function.

We call the following strategy a *51% double-spending attack*: The adversary obtains any majority fraction (“51%” is just a colloquial name) of the hashing power, and uses it to secretly mine an extension of the currently longest chain (i.e., keeping successful blocks private to himself), and which he will release after some time. We say that a 51% double-spending attack is *successful* if, when released, the adversary’s secret chain is at least as long as the honest chain, and causes the ledger state of some honest party to fork (which in reality corresponds to a roll-back of more than **cutOff** blocks, in order to adopt the released attack chain). If this happens, some transactions on the reversed blockchain ledger state may become orphaned (no longer part of the ledger state), thus allowing the attacker to double-spend his coins.

5.1 Time to Fork

We start by showing a general lemma that relates the amount of honest and adversarial hashing power in a system, to the time to cause a fork via a 51% double-spending attack. That is, how long it takes for an adversary with majority hashing power to secretly create a chain that, when released, would cause an honest party to roll back, or discard, more than **cutOff** blocks of his own chain in order to adopt the new one.

Definition 8. *We say that an adversary \mathcal{A} causes a fork in a protocol Π if, except with negligible probability in κ , all simulators $\mathcal{S}^{\mathcal{A}} \in \mathcal{C}_{\mathcal{A}}$ (i.e. those which in fact simulate \mathcal{A} according to UC emulation) use the **FORK** command⁶.*

The **FORK** command, which allows forking the confirmed ledger state (and hence corresponds to rolling back more than **cutOff** blocks in the real world), is necessary and sufficient to simulate an adversary who succeeds in a 51% double-spending attack. We compute the (upper bound) time for a 51% double-spending adversary to fork, which is obtained by the time for honest parties to grow their chain by **cutOff** blocks (for which we can use guaranteed chain-growth of Nakamoto-style blockchains. Since the adversary has more hashing power (and thus more random oracle queries that can be issued sequentially) than the honest party, and since we assume $\text{cutOff} = \omega(\log(\kappa))$ and that the adversary does not interfere with the honest parties’ mining, this implies that the adversary’s secretly-mined chain will be longer than the honest parties’ chain, and be the only source for a large rollback, with overwhelming probability in κ .

⁶ If $\mathcal{C}_{\mathcal{A}} = \emptyset$, then in any case by definition the utility of \mathcal{A} is infinite.

Lemma 6 (Time to Fork with 51% Attack). *Let $\text{cutOff} = \omega(\log(\kappa))$, $[r, r+t]$ be any time interval (starting from some round $r \geq 0$) of $t \geq 1$ rounds, $\Delta \geq 1$ be the network delay, $p \in (0, 1)$ the success probability of one mining query.*

Then for all $\delta, \delta' \in (0, 1)$, $\alpha \geq \frac{1+\delta}{1-\delta}$, and $q \geq 1$ such that $t \geq t_{\delta'}^{\Delta}(q)$ (Definition 6) the following holds. Suppose in time interval $[r, r+t]$, (1) the honest parties make at least q mining queries per round, and (2) in total they make at most q_t queries. Then, the adversary \mathcal{A} who performs a 51% double-spending attack for at least αq_t queries during the time interval and then releases his secretly-mined chain, causes a fork in the protocol Π^B .⁷

A visualization. In Fig. 1, the (upper-bound) time to fork with exactly 51% corruption, is graphed against the total number of rigs in the system. The graph uses the formula from Lemma 6. We use current parameters for Ethereum Classic as the source of the concrete parameters for this figure, and refer the reader to our full version for more details.

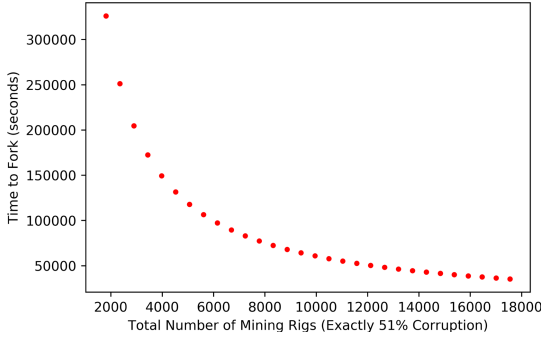


Fig. 1. Time to create a fork via 51% attack, versus the total number of mining rigs. Here the adversary corrupts exactly 51%.

5.2 Payoff of 51% Double-Spending Attacks

In this section, we prove Theorem 1 and its corollary Theorem 2, which quantify the size of the payoff for double-spending, under which a 51% double-spending attack can break strong attack-payoff security. That is, the attacker achieves better utility than *any* passive mining strategy. While one may think that it is always profitable to attack if there is no assumption on the honest majority of hashing power, there are a few things that may deter an attacker. For example, the costs of buying or renting mining equipment for the attack may become too high compared to the diminished block rewards as time goes on. Our statement below quantifies an amount of payoff for forking (e.g. how much an attacker

⁷ More concretely, he succeeds except with probability at most $\exp\left(-\frac{\delta^2 \alpha \mu}{2+\delta}\right) + \exp\left(-\frac{\delta^2 \mu}{2}\right) + \exp\left(-\frac{\delta'^2 t \gamma}{2+\delta}\right)$.

can double-spend) to incentivize a 51% double-spending attack. Intuitively, the result below says that as long as the payoff for forking (f payoff) is larger than the loss of utility from withholding blocks and corrupting a large number of parties to perform the attack, then there is a 51% attack strategy that is more profitable than *any* front-running, passive adversary.

Theorem 1 (51% Double-Spending Attacks that Break (Strong) Attack-Payoff Security (u_{buy})). *Let $T_{\text{ub}} > 2$ be the upper bound on total number of mining queries per round, $p \in (0, 1)$ be the probability of success of each mining query, and $\text{cutOff} = \omega(\log(\kappa))$ be the consistency parameter. Then, the protocol $\Pi^{\mathcal{B}}$ is not attack-payoff secure/strongly attack-payoff secure in any attack model \mathcal{M} whose induced utility function u_{buy} satisfies limited horizons, if for some $\delta \in (0, 1)$, $\alpha > 1$ and $g = \frac{T_{\text{ub}}}{1+\alpha}$ the following holds:*

$$\text{f payoff} > u_{\text{honest}}^{\text{h}} - \alpha \cdot g \cdot t_{\delta}^{\Delta}(g) \left(p \cdot \text{CR} \cdot \text{breward}(t_{\delta}^{\Delta}(g) + t_{\text{ub}}) - \text{mcost} \right) - \text{ccost}(\alpha g).$$

We state the case where the adversary mines with rented equipment (and uses utility function u_{rent}), as a direct corollary to Theorem 1.

Theorem 2 (51% Double-Spending Attacks that Break (Strong) Attack-Payoff Security (u_{rent})). *Let $T_{\text{ub}} > 2$ be the upper bound on total number of mining queries per round, $p \in (0, 1)$ be the probability of success of each mining query, and $\text{cutOff} = \omega(\log(\kappa))$ be the consistency parameter. Then, the protocol $\Pi^{\mathcal{B}}$ is not attack-payoff secure/strongly attack-payoff secure in any attack model \mathcal{M} whose induced utility function u_{rent} satisfies limited horizons, if for any $\delta \in (0, 1)$, $\alpha > 1$ and $g = \frac{T_{\text{ub}}}{1+\alpha}$ the following holds:*

$$\text{f payoff} > u_{\text{honest}}^{\text{h}} - \alpha \cdot g \cdot t_{\delta}^{\Delta}(g) \left(p \cdot \text{CR} \cdot \text{breward}(t_{\delta}^{\Delta}(g) + t_{\text{ub}}) - \text{mcost} \right).$$

5.3 Visualizations with Concrete Values

We will visualize Theorems 1 and 2 through Figs. 2 and 3. We consider two utility functions, one where the adversary buys mining equipment, and one where the adversary rents. We then graph the utilities of passive/non-passive adversaries, against the maximum fraction of corrupted parties. The concrete parameters are based on current (as of writing, Feb. 2021) parameters for Ethereum Classic. The outline is given in our full version.

In Fig. 2, we consider the incentives of a 51% attacker who rents his hashing power, using the price for renting of 1.96 BTC/TH/day (Bitcoin per terahash per day), at \$50,000/BTC. In this case, it is in fact not profitable to mine passively (and thus the optimal passive strategy is to not mine at all). However, when the adversary corrupts more than majority of hashing power, it may become profitable to mine in order to create a fork. It is less profitable for the adversary to corrupt a larger fraction of the parties, as cost of renting becomes too high. We remark that even when it is not profitable to mine (passively) using rented rigs, this does not exclude incentivizing honest parties from mining with e.g., bought equipment.

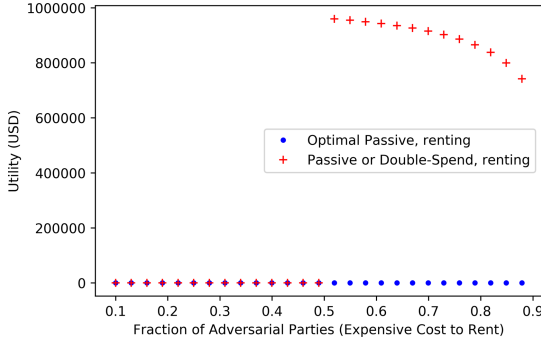


Fig. 2. Utility of the passive/51% double-spending attacker who rents hashing power, versus the fraction of adversarial parties. Here we consider an expensive cost to rent hashing power (1.96 BTC/TH/day, at \$50,000/BTC).

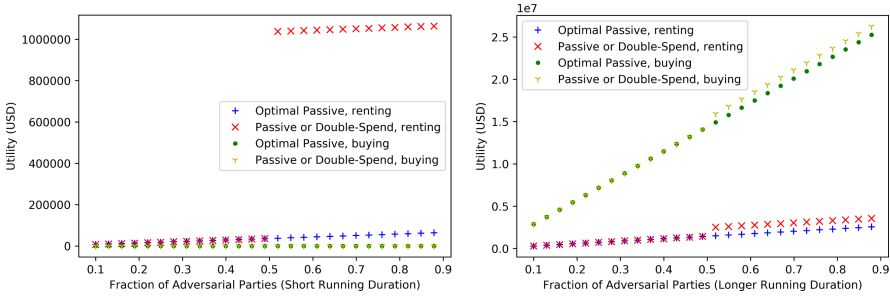


Fig. 3. Utility of the passive/51% double-spending attacker versus the fraction of adversarial parties. We consider an attacker who runs for a short duration (1 week) and a long duration (40 weeks).

In the next two examples in Fig. 3, we compare the utility of the attacker who mines with purchased rigs, and one who mines with rented rigs. For the attack who buys hashing power, each rig costs \$3000, and mining (electricity) costs \$0.000047/s. For the attacker who rents, for more interesting comparisons we consider a cheaper cost to rent hashing power (1.96 BTC/TH/day, at a cheaper \$22,000/BTC). We consider two scenarios: the attacker either (1) only plans to mine for a short duration of one week, or (2) plans to mine for a longer duration of 40 weeks (time is expressed in seconds in the code). For the purposes of the graphs, we account for the possible variance of Bitcoin-USD exchange rates by using an average exchange rate over the relevant period of time. In either case, to more closely model reality, we restrict the duration of the attack, where the adversary may obtain a majority of hashing power, to 3 days (which, in the code, simply means we do not show attacks that last longer than 3 days).

We see a big difference between the two scenarios. In the short duration case, it is much more profitable to mine or attack with rented rigs. In fact, it is not even profitable to fork using purchased rigs, as the cost of purchase is higher

than the payoff for double-spending. The long duration case is the opposite. Although it may be profitable to mine in both cases, it is vastly more profitable to mine and attack with purchased rigs than rented rigs. This agrees with our intuition and reality: the high initial investment of buying mining equipment is offset in the long run by the lower cost of mining. Moreover, an attacker who is only interested in mining in order to perform a 51% attack for a short time is incentivized to use hash renting services.

6 Mitigating 51% Attacks

In previous sections, we studied utility functions with limited horizons, in which an attacker is incentivized to perform a 51% double-spending attack and break (strong) attack-payoff security. In this section, we turn to analyzing how to *defend* against 51% attacks. Specifically, given an attacker’s utility function with limited horizons, and a cut-off parameter `cutOff` that achieves security in the honest majority setting, we show a way to amplify `cutOff` to obtain security against a rational (and possibly dishonest majority) attacker.

To show attack payoff security, one must show that for *any* adversarial strategy attacking the protocol, there is another adversary who attacks the dummy protocol with access to the ideal ledger functionality $\mathcal{G}_{\text{LEDGER}}$ ⁸, which achieves the same or better utility.

Even more difficult, we place very few restrictions on the adversary: he may corrupt any fraction of parties (e.g. more than majority) and perform any currently known (e.g. block withholding) or unknown strategy. The only restriction we place on the attacker is that he is incentive-driven. Fortunately, a rational attacker is limited by his utility function. As we show, given a utility function satisfying limited horizon, we are able to bound the amount of mining an incentive-driven adversary will do, even in presence of a payoff for forking. Then, by choosing a large enough consistency parameter `cutOff`, we ensure that attackers are disincentivized from creating a fork.

More specifically: We first present in Sect. 6.1 a result that shows that if an adversary’s hasing resources are limited by a budget B , then there is a bound on the interval of rounds where the blockchain is at risk of a consistency failure (Lemma 7). For this, we apply a result from [BGK+20] that, roughly, shows how fast a blockchain’s consistency can recover after an attack by an adversary with a given budget (the self-healing property of Bitcoin). Based on this fundamental property, we present in Sect. 6.2, the main result of the section: Given a utility function with limited horizons, we show a condition on the parameter `cutOff`, depending only on the utility function and protocol parameters, such that Π^B is attack-payoff secure. To do so, we show that an adversary who spends too much budget will begin to lose utility (Lemma 8), and then combine this result with that of Sect. 6.1.

⁸ Recall this is the ledger functionality that has consistency and liveness (following from the liveness results in [BGK+20]), but since we will amplify the cut-off parameter `cutOff`, we achieve less good parameters.

6.1 Budget to Vulnerability Period

Assume an instance of the Bitcoin backbone protocol with cut-off parameter ℓ . We distinguish here between ℓ and `cutOff` for clarity, since we will eventually amplify ℓ to obtain our final cut-off parameter `cutOff`.

Under the honest majority condition, we know that a consistency failure (expressed as the probability that blocks which are ℓ deep in an honest parties adopted chain can be reverted) appears with probability negligible in ℓ (and consequently also in any security parameter κ as long as $\ell = \omega(\log(\kappa))$). We now recall (and state a simple corollary from) the result from [BGK+20], which defines a relationship between an adversary's violation of honest-majority (measured as a so-called budget B by which it can violate the honest-majority condition) and the time until Bitcoin (and more generally, Nakamoto-style PoW chains) self-heals after the adversary returns to below 50% hashing power. That is, until Bitcoin can again guarantee consistency for the part of the chain that is at least ℓ blocks deep in a longest chain held by an honest party. The self-healing time depends on the budget and the parameter ℓ . Recall that θ_{pow} is the usual security threshold for Bitcoin as explained in Sect. 2.1.

Definition 9 ($(\theta_{pow}, \epsilon, T_{lb}, T_{ub}, B)$ -adversary [BGK+20]). *Let $\theta_{pow}, \epsilon \in (0, 1)$, $T_{lb}, T_{ub}, B \in \mathbb{N}$. A $(\theta_{pow}, \epsilon, T_{lb}, T_{ub}, B)$ -adversary is an adversary⁹ satisfying the following: At every round i , let n_a^i and n_h^i be the mining queries made by corrupt and honest parties in this round. Then, (1) For all i , $T_{lb} \leq n_a^i + n_h^i \leq T_{ub}$, and (2) For all i , $n_a^i \leq (1 - \epsilon) \cdot \theta_{pow} \cdot n_h^i + B_i$, where $B_i \geq 0$ and $\sum_i B_i = B$.*

We say the adversary attacks between rounds $a < b$ if $B_i = 0$ for any $i < a$ or $i > b$ (i.e. he spends all his budget between rounds a and b).

We say an adversary spends budget B over t rounds, if the adversary has budget B , and only spends it in rounds $r_1 < r_2 < \dots < r_t$, such that $\sum_i B_{r_i} = B$.

The behavior of a blockchain protocol under an attack by an $(\theta_{pow}, \epsilon, T_{lb}, T_{ub}, B)$ -adversary is described by a *vulnerability period*. The vulnerability period is an upper bound on number of rounds before and after an adversary performs the attack, such that protocol is still at risk of a (non-negligible) consistency failure.

Definition 10 (Consistency self-healing property and vulnerability period [BGK+20]). *A protocol is self-healing with vulnerability period (τ_l, τ_h) with respect to consistency, and against a $(\theta_{pow}, \epsilon, T_{lb}, T_{ub}, B)$ -adversary who attacks between rounds (a, b) , if the consistency failure event $\text{ConsFail}_\ell(r)$ occurs except with at most negligible probability unless $r \in [\rho_\alpha - \tau_l, \rho_\beta + \tau_h]$. $\text{ConsFail}_\ell(r)$ is defined as the event that ℓ -consistency is violated in an execution for rounds (r, r') , w.r.t. some round $r' > r$, and any two pairs of honest parties.*

In other words, outside of these “dangerous” rounds $[a - \tau_l, b + \tau_h]$, chains adopted by honest parties are guaranteed to diverge by at the most recent ℓ blocks. Below, [BGK+20] gives a characterization of the vulnerability period in terms of the budget B .

⁹ Here the environment is also included in this statement.

Theorem 3 ([BGK+20]). *A Nakamoto-style PoW blockchain with an upper bound T_{ub} of hashing queries per round, maximum network delay Δ , success probability p , and cut-off parameter ℓ satisfies the consistency self-healing property with vulnerability period $(\tau_l, \tau_h) = (O(B), O(B) + O(\ell))$ against any $(\theta_{\text{pow}}, \epsilon, T_{\text{lb}}, T_{\text{ub}}, B)$ -adversary, for any $\epsilon, T_{\text{lb}} > 0$.*

The vulnerability period only bounds the number of “bad” rounds before the attack, and after *all* the budget is spent. For our treatment, we consider a more applicable version of the vulnerability period. In Lemma 7, we show the maximum number of consecutive rounds where ConsFail may occur, by applying the above theorem in a piece-wise fashion. For example, if the adversary spends his budget over a long period of time (e.g., spend a bit of the budget, wait for 2 years, then spend more of his budget), the theorem is not directly suitable for our needs, but it is possible to isolate those “spending” rounds and applying the theorem to each such region. Then, since the total hashing power in the system is bounded, we can use this maximum consecutive “bad” rounds to bound the maximum number of blocks that can be rolled back at any given round.

Lemma 7 (Max consecutive consistency failure rounds and associated number of blocks and rollback). *In the same setting as above in Theorem 3, except with negligible probability the following holds: for any adversary with budget B , spent over t rounds (that is, for t different rounds i it holds that $B_i > 0$), there is a maximum number $R(B, t, \ell) = O(B) + O(\ell t)$ of consecutive rounds r_j where $\text{ConsFail}_\ell(r_j)$ occurs, during which at most $W(B, t, \ell) = 2T_{\text{ub}}p \cdot R(B, t, \ell)$ blocks are created.*

Looking ahead, this means that at any point in time, prefixes of honest parties’ chains must agree (except with negligible probability) when dropping the most recent $W(B, t, \ell) + \ell$ blocks. Here, we omit the dependency on p because we treat it as a constant parameter of the protocol.

6.2 Attack-Payoff Security

In this section we will show the following: For any utility function with limited horizons, we give a characterization of how to adjust the consistency parameter (depending on the protocol parameters and those of the utility function) such that Π^{B} is attack payoff secure. To do so, we will first upper bound the utility of adversaries who spends a total budget B over some time t , given a utility function u_{buy} with limited horizons (Lemma 8, and Corollary 1 for utility function u_{rent}). In Theorem 4, we then combine this lemma with the result of the previous subsection, and present our characterization of parameters for which Π^{B} is attack-payoff secure—i.e. for which forks are disincentivized.

Below, we quantify an upper bound $u_{\text{buy}}^{\text{ub}}(B, t)$ on the utility of *any* adversary spending budget of at least B over exactly t rounds, assuming the utility function satisfies limited horizons. Why are we interested in this quantity? Recall $W(B, t)$ —which informally represents an interval of blocks where consistency

might fail—increases with B and t . Looking ahead, we will find a large enough $W(B, t)$ that disincentivizes attacks (i.e., $u_{\text{buy}}^{\text{ub}}(B, t) < 0$). To show that the upper-bound $u_{\text{buy}}^{\text{ub}}(B, t)$ is useful, later we will show that it is possible use it to derive a maximum B, t , which we denote by \bar{B}, \bar{t} .

Lemma 8 (Upper bound utility of adversary spending budget at least B , over time t).

Suppose $u_{\text{buy}}(\Pi^{\mathcal{B}}, \mathcal{A})$ satisfies limited horizons. Then an adversary \mathcal{A} with budget at least $B > 0$, and who spends it over exactly $t \geq \frac{B}{T_{\text{ub}} - \bar{n}_a}$ rounds, achieves utility at most $u_{\text{buy}}(\Pi^{\mathcal{B}}, \mathcal{A}) \leq u_{\text{buy}}^{\text{ub}}(B, t)$ where

$$\begin{aligned} u_{\text{buy}}^{\text{ub}}(B, t) := & \sum_{x=1}^{t_h} T_{\text{ub}} \cdot (p \cdot \text{CR} \cdot \text{breward}(x) - \text{mcost}) \\ & + \sum_{x=t_h+1}^t (\bar{n}_a + 1) \cdot (p \cdot \text{CR} \cdot \text{breward}(x) - \text{mcost}) \\ & - \text{ccost} \left(\bar{n}_a + \frac{B}{t} \right) + \text{fpayoff} \end{aligned}$$

and where $t_h := \arg \max_{x \in \mathbb{N}} (p \cdot \text{CR} \cdot \text{breward}(x) \geq \text{mcost})$, $\bar{n}_a := \frac{(1-\epsilon) \cdot \theta_{\text{pow}} \cdot T_{\text{lb}}}{1 + (1-\epsilon) \cdot \theta_{\text{pow}}}$.

If $t < \frac{B}{T_{\text{ub}} - \bar{n}_a}$ (in this case it is not possible to spend budget B over t rounds) or $B \leq 0$, then $u_{\text{buy}}^{\text{ub}}(B, t)$ is undefined.

As a corollary, by setting $\text{ccost}(\cdot) = 0$ and $\text{mcost} = \text{rcost}$, we obtain an upper bound on the utility of any adversary who spends at least budget B , assuming the utility function satisfies limited horizons with renting.

Corollary 1. Suppose $u_{\text{rent}}(\Pi^{\mathcal{B}}, \mathcal{A})$ satisfies limited horizons with renting. Then an adversary \mathcal{A} who spends budget of at least $B > 0$ over exactly $t \geq \frac{B}{T_{\text{ub}} - \bar{n}_a}$ rounds, achieves utility at most $u_{\text{rent}}(\Pi^{\mathcal{B}}, \mathcal{A}) \leq u_{\text{rent}}^{\text{ub}}(B, t)$ where

$$\begin{aligned} u_{\text{rent}}^{\text{ub}}(B, t) := & \sum_{x=1}^{t_h} T_{\text{ub}} \cdot (p \cdot \text{CR} \cdot \text{breward}(x) - \text{rcost}) \\ & + \sum_{x=t_h+1}^t (\bar{n}_a + 1) \cdot (p \cdot \text{CR} \cdot \text{breward}(x) - \text{rcost}) + \text{fpayoff} \end{aligned}$$

and where $t_h := \arg \max_{x \in \mathbb{N}} (p \cdot \text{CR} \cdot \text{breward}(x) \geq \text{rcost})$, $\bar{n}_a := \frac{(1-\epsilon) \cdot \theta_{\text{pow}} \cdot T_{\text{lb}}}{1 + (1-\epsilon) \cdot \theta_{\text{pow}}}$.

A natural question is whether the upper bound $u_{\text{buy}}^{\text{ub}}(B, t)$ and $u_{\text{rent}}^{\text{ub}}(B, t)$ will be useful for bounding B, t . We remark that it is not even trivially clear whether they bounded, as the budget does not limit how many rounds (when honest majority is satisfied) the adversary can mine. Below, we show that there indeed

exist a maximum B, t for which $u_{\text{buy}}^{\text{ub}}(B, t) \geq 0$ (resp. $u_{\text{rent}}^{\text{ub}}(B, t) \geq 0$, which we denote by \bar{B}, \bar{t}). In our full version, we give examples of \bar{B}, \bar{t} using numbers from Ethereum Classic, Feb. 2021 (same as our other graphs).

Lemma 9. $\bar{B} := \arg \max_{B>0} (u_{\text{buy}}^{\text{ub}}(B, \cdot) \geq 0)$ and $\bar{t} := \arg \max_{t>0} (u_{\text{buy}}^{\text{ub}}(\cdot, t) \geq 0)$ exist, or $\forall t > 0, u_{\text{buy}}^{\text{ub}}(\cdot, t) < 0$. The same is true when replacing $u_{\text{buy}}^{\text{ub}}$ with $u_{\text{rent}}^{\text{ub}}$ in the statement.

Finally, we can make a general statement about the attack payoff security of protocol $\Pi^{\mathcal{B}}$ for any utility function satisfying limited horizons. Informally: our utility function limits how much budget B (spent over how many rounds t) any incentive-driven attacker could reasonably have. Then, if the window size is large enough to accommodate the largest of such budgets, the protocol is attack-payoff secure. In the following, $u_{\text{buy}}^{\text{ub}}(B, t)$, $u_{\text{honest}}^{\text{l}}$, and $W(B, t)$ are from Lemma 8, Lemma 5, and Lemma 7 respectively. The equivalent statement for utility function u_{rent} can be obtained by using the corollaries of the results above for the case of renting.

Theorem 4. Let $T_{\text{ub}}, T_{\text{lb}} > 0$ be the upper and lower bounds on total number of mining queries per round and $p \in (0, 1)$ be the probability of success of each mining query and let $\ell = \omega(\log(\kappa))$. Then, $\Pi^{\mathcal{B}}$ with consistency parameter `cutOff` is attack-payoff secure in any model \mathcal{M} , whose induced utility u_{buy} satisfies limited horizons, whenever the condition holds that

$$\text{cutOff} > \ell + \max_{(B,t): u_{\text{buy}}^{\text{ub}}(B,t) > u_{\text{honest}}^{\text{l}}} W(B, t, \ell).$$

The same statement is true replacing u_{buy} with u_{rent} and $u_{\text{buy}}^{\text{ub}}$ with $u_{\text{rent}}^{\text{ub}}$.

We remark the protocol is not *strong* attack-payoff secure: Recall a front-running adversary always maximally delays honest parties' messages. Intuitively, this reduces the mining power in the system, which delays the time between a block is broadcasted, and when it becomes part of the ledger. In effect, this reduces the payoff for block rewards for utilities with limited horizons.

Acknowledgment. Yun Lu and Vassilis Zikas acknowledge support from Sunday Group Inc.

References

- [AKWW19] Avarikioti, G., Käppeli, L., Wang, Y., Wattenhofer, R.: Bitcoin security under temporary dishonest majority. In: Goldberg, I., Moore, T. (eds.) FC 2019. LNCS, vol. 11598, pp. 466–483. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32101-7_28
- [Ber54] Bernoulli, D.: Exposition of a new theory on the measurement of risk. *Econometrica* **22**(1), 23–36 (1954)

- [BGK+18] Badertscher, C., Gazi, P., Kiayias, A., Russell, A., Zikas, V.: Ouroboros genesis: composable proof-of-stake blockchains with dynamic availability. In: Lie, D., Mannan, M., Backes, M., Wang, X. (eds.) ACM CCS 2018, pp. 913–930. ACM Press, October 2018
- [BGK+20] Badertscher, C., Gazi, P., Kiayias, A., Russell, A., Zikas, V.: Consensus redux: distributed ledgers in the face of adversarial supremacy. Cryptology ePrint Archive, Report 2020/1021 (2020). <https://eprint.iacr.org/2020/1021>
- [BGM+18] Badertscher, C., Garay, J., Maurer, U., Tschudi, D., Zikas, V.: But why does it work? A rational protocol design treatment of Bitcoin. In: Nielsen, J.B., Rijmen, V. (eds.) EUROCRYPT 2018. LNCS, vol. 10821, pp. 34–65. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78375-8_2
- [BKKS20] Brünjes, L., Kiayias, A., Koutsoupias, E., Stouka, A.-P.: Reward sharing schemes for stake pools. In: IEEE European Symposium on Security and Privacy, EuroS&P 2020, Genoa, Italy, 7–11 September 2020, pp. 256–275. IEEE (2020)
- [BLZ21] Badertscher, C., Lu, Y., Zikas, V.: A rational protocol treatment of 51% attacks. Cryptology ePrint Archive, Report 2021/897 (2021). <https://eprint.iacr.org/2021/897>
- [BMTZ17] Badertscher, C., Maurer, U., Tschudi, D., Zikas, V.: Bitcoin as a transaction ledger: a composable treatment. In: Katz, J., Shacham, H. (eds.) CRYPTO 2017. LNCS, vol. 10401, pp. 324–356. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63688-7_11
- [Bud18] Budish, E.: The economic limits of Bitcoin and the Blockchain. Technical report, National Bureau of Economic Research (2018)
- [But13] Buterin, V.: A next-generation smart contract and decentralized application platform (2013). <https://github.com/ethereum/wiki/wiki/White-Paper>
- [Can01] Canetti, R.: Universally composable security: a new paradigm for cryptographic protocols. In: 42nd FOCS, pp. 136–145. IEEE Computer Society Press, October 2001
- [CCWr20] Chung, K.-M., Hubert Chan, T.-H., Wen, T., Shi, E. (random author ordering): Game-theoretically fair leader election in $o(\log \log n)$ rounds under majority coalitions. Cryptology ePrint Archive, Report 2020/1591 (2020). <https://eprint.iacr.org/2020/1591>
- [CKWN16] Carlsten, M., Kalodner, H.A., Matthew Weinberg, S., Narayanan, A.: On the instability of Bitcoin without the block reward. In: Weippl, E.R., Katzenbeisser, S., Kruegel, C., Myers, A.C., Halevi, S. (eds.) ACM CCS 2016, pp. 154–167. ACM Press, October 2016
- [CM19] Chen, J., Micali, S.: Algorand: a secure and efficient distributed ledger. Theor. Comput. Sci. **777**, 155–183 (2019)
- [Coia] 51% attacks archives. <https://www.coindesk.com/tag/51-attack>. publisher=CoinDesk
- [Coib] Ethereum classic’s mess solution won’t provide ‘robust’ security against 51% attacks: Report. <https://www.coindesk.com/ethereum-classic-mess-security-51-attacks-report>. publisher=CoinDesk
- [DPS19] Daian, P., Pass, R., Shi, E.: Snow White: robustly reconfigurable consensus and applications to provably secure proof of stake. In: Goldberg, I., Moore, T. (eds.) FC 2019. LNCS, vol. 11598, pp. 23–41. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32101-7_2

- [ES14] Eyal, I., Sirer, E.G.: Majority is not enough: Bitcoin mining is vulnerable. In: Christin, N., Safavi-Naini, R. (eds.) FC 2014. LNCS, vol. 8437, pp. 436–454. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-45472-5_28
- [Eya15] Eyal, I.: The miner’s dilemma. In: 2015 IEEE Symposium on Security and Privacy, pp. 89–103. IEEE Computer Society Press, May 2015
- [For] Rented hash power for 51% attacks is a ‘huge vulnerability’ for proof-of-work blockchains, says etc labs ceo. <https://forkast.news/hash-power-51-attack-rent-huge-vulnerability-proof-of-work-blockchain/>. publisher=Forkast
- [GKL15] Garay, J., Kiayias, A., Leonardos, N.: The Bitcoin backbone protocol: analysis and applications. In: Oswald, E., Fischlin, M. (eds.) EUROCRYPT 2015. LNCS, vol. 9057, pp. 281–310. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-46803-6_10
- [GKM+13] Garay, J.A., Katz, J., Maurer, U., Tackmann, B., Zikas, V.: Rational protocol design: cryptography against incentive-driven adversaries. In: 54th FOCS, pp. 648–657. IEEE Computer Society Press, October 2013
- [GKO+20] Garay, J., Kiayias, A., Ostrovsky, R.M., Panagiotakos, G., Zikas, V.: Resource-restricted cryptography: revisiting MPC bounds in the proof-of-work era. In: Canteaut, A., Ishai, Y. (eds.) EUROCRYPT 2020. LNCS, vol. 12106, pp. 129–158. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45724-2_5
- [GKW+16] Gervais, A., Karame, G.O., Wüst, K., Glykantzis, V., Ritzdorf, H., Capkun, S.: On the security and performance of proof of work blockchains. In: Weippl, E.R., Katzenbeisser, S., Kruegel, C., Myers, A.C., Halevi, S. (eds.) ACM CCS 2016, pp. 3–16. ACM Press, October 2016
- [HSY+21] Han, R., Sui, Z., Yu, J., Liu, J., Chen, S.: Fact and fiction: challenging the honest majority assumption of permissionless blockchains. In: Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security, ASIA CCS 2021, pp. 817–831. Association for Computing Machinery, New York (2021)
- [Inv] 51% attack. <https://www.investopedia.com/terms/1/51-attack.asp>. publisher=Investopedia
- [JL20] Jang, J., Lee, H.-N.: Profitable double-spending attacks. Appl. Sci. **10**(23), 8477 (2020)
- [KRDO17] Kiayias, A., Russell, A., David, B., Oliynykov, R.: Ouroboros: a provably secure proof-of-stake blockchain protocol. In: Katz, J., Shacham, H. (eds.) CRYPTO 2017. LNCS, vol. 10401, pp. 357–388. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63688-7_12
- [LTKS15] Luu, L., Teutsch, J., Kulkarni, R., Saxena, P.: Demystifying incentives in the consensus computer. In: Ray, I., Li, N., Kruegel, C. (eds.) ACM CCS 2015, pp. 706–719. ACM Press, October 2015
- [Nak08] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system (2008). <http://bitcoin.org/bitcoin.pdf>
- [NKMS16] Nayak, K., Kumar, S., Miller, A., Shi, E.: Stubborn mining: generalizing selfish mining and combining with an eclipse attack. In: S&P (2016)
- [PS17] Pass, R., Shi, E.: FruitChains: a fair blockchain. In: Michael Schiller, E., Schwarzmann, A.A. (eds.) 36th ACM PODC, pp. 315–324. ACM, July 2017

- [PSs17] Pass, R., Seeman, L., Shelat, A.: Analysis of the blockchain protocol in asynchronous networks. In: Coron, J.-S., Nielsen, J.B. (eds.) EUROCRYPT 2017. LNCS, vol. 10211, pp. 643–673. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56614-6_22
- [Ros11] Rosenfeld, M.: Analysis of bitcoin pooled mining reward systems. CoRR (2011)
- [SBBR16] Schrijvers, O., Bonneau, J., Boneh, D., Roughgarden, T.: Incentive compatibility of Bitcoin mining pool reward functions. In: Grossklags, J., Preneel, B. (eds.) FC 2016. LNCS, vol. 9603, pp. 477–498. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-662-54970-4_28
- [SSZ16] Sapirshtein, A., Sompolinsky, Y., Zohar, A.: Optimal selfish mining strategies in Bitcoin. In: Grossklags, J., Preneel, B. (eds.) FC 2016. LNCS, vol. 9603, pp. 515–532. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-662-54970-4_30
- [TJS16] Teutsch, J., Jain, S., Saxena, P.: When cryptocurrencies mine their own business. In: Grossklags, J., Preneel, B. (eds.) FC 2016. LNCS, vol. 9603, pp. 499–514. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-662-54970-4_29