Enter the Hydra: Towards Principled Bug Bounties and Exploit-Resistant Smart Contracts
Lorenz Breidenbach, Cornell Tech, IC3, ETH Zurich; Philip Daian, Cornell Tech, IC3; Florian Tramer, Stanford; Ari Juels, Cornell Tech, IC3, Jacobs Institute
https://www.usenix.org/conference/usenixsecurity18/presentation/breidenbach

This paper is included in the Proceedings of the 27th USENIX Security Symposium.
August 15–17, 2018 • Baltimore, MD, USA
978-1-939133-04-5

Open access to the Proceedings of the 27th USENIX Security Symposium is sponsored by USENIX.
Enter the Hydra: Towards Principled Bug Bounties and Exploit-Resistant Smart Contracts

Lorenz Breidenbach  
lorenzb@inf.ethz.ch  
Cornell Tech, IC3†  
ETH Zürich

Philip Daian  
phil@cs.cornell.edu  
Cornell Tech, IC3†  
Stanford

Florian Tramèr  
tramer@cs.stanford.edu  
Stanford

Ari Juels  
juels@cornell.edu  
Cornell Tech, IC3†  
Jacobs Institute

Abstract

Bug bounties are a popular tool to help prevent software exploits. Yet, they lack rigorous principles for setting bounty amounts and require high payments to attract economically rational hackers. Rather than claim bounties for serious bugs, hackers often sell or exploit them.

We present the Hydra Framework, the first general, principled approach to modeling and administering bug bounties that incentivize bug disclosure. Our key idea is an exploit gap, a program transformation that enables runtime detection, and rewarding, of critical bugs. Our framework transforms programs via N-of-N-version programming, a variant of classical N-version programming that runs multiple independent program instances.

We apply the Hydra Framework to smart contracts, small programs that execute on blockchains. We show how Hydra contracts greatly amplify the power of bounties to incentivize bug disclosure by economically rational adversaries, establishing the first framework for rigorous economic evaluation of smart contract security. We also model powerful adversaries capable of bug withholding, exploiting race conditions in blockchains to claim bounties before honest users can. We present Submarine Commitments, a countermeasure of independent interest that conceals transactions on blockchains.

We design a simple, automated version of the Hydra Framework for Ethereum (ethereum.org) and implement two Hydra contracts, an ERC20 standard token and a Monty-Hall game. We evaluate our implementation for completeness and soundness with the official Ethereum Virtual Machine test suite and live blockchain data.

1 Introduction

Despite theoretical and practical advances in code development, software vulnerabilities remain an ineradicable security problem. Vulnerability reward programs—aka. bug bounties—have become instrumental in organizations’ security assurance strategies. These programs offer rewards as incentives for hackers to disclose software bugs. Unfortunately, hackers often prefer to exploit critical vulnerabilities or sell them in gray markets.

The chief reason for this choice is that the bugs eligible for large bounties are generally weaponizable vulnerabilities. The financial value of critical bugs (0-days) in gray markets may exceed bounty amounts by a factor of as much as ten to one hundred [2]. For example, while Apple offers a maximum 200k USD bounty, a broker intermediary such as Zerodium purportedly offers 1.5 million USD for certain iPhone jailbreaks. In some cases hackers can monetize vulnerabilities themselves for large payouts [15, 11]. Modest bounties may thus fail to successfully incentivize disclosure by rational actors [43].

Pricing bounties appropriately can also be hard because of a lack of research giving principled guidance. Payments are often scheduled arbitrarily based on bug categories and may not reflect bugs’ market value or impact. For example, Apple offers up to 100k USD for generic bugs defined as “Extraction of confidential material protected by the Secure Enclave Processor” [43].

Finally, bounties present a problem of fair exchange. A bounty payer does not wish to pay before reviewing an exploit, while hackers are wary of revealing exploits and risking non-payment (e.g., [26, 4, 54]). This uncertainty creates a market inefficiency that limits incentives for rational hackers to uncover vulnerabilities.

We introduce the Hydra Framework, the first principled approach to bug bounty administration that addresses these challenges. Our framework deters economically rational actors, including black-hat hackers, from exploiting bugs or selling them in underground markets. We focus on smart contracts as a use case to demonstrate our framework’s power analytically and empirically.
The Hydra Framework. Our key idea is to build support for bug detection and bounties into software at development time using a concept that we call an exploit gap. This is a program transformation that makes critical bugs detectable at runtime, but hard to exploit.

We propose an exploit gap technique that we refer to as N-of-N-version programming (NNVP). A variant of classical N-version programming, NNVP leverages multiple versions of a program that are independently developed, or otherwise made heterogeneous. In the Hydra Framework, these program versions, or heads, are executed in parallel within a meta-program called a Hydra program.

In stark contrast to N-version programming’s goal of fault tolerance (i.e., where the program attempts to produce a correct output even in the face of partial failures), NNVP focuses on error detection and safe termination. If heads’ outputs are identical, a Hydra program runs normally. If the outputs diverge for some input, a dangerous state is indicated and the program aborts and pays out a bounty. The basic idea is depicted in Figure 1.

A bug is only exploitable if it affects all Hydra heads identically. If failures are somewhat uncorrelated across heads, a bug in one head is thus unlikely to affect the Hydra program as a whole. Moreover, an adversary that breaks one head and, instead of claiming a bounty, tries to generalize the exploit, risks preemption by honest bounty hunters. We show that even when an exploit’s market value exceeds the bounty by multiple orders of magnitude, economically rational hackers are incentivized to disclose bugs rather than attempt an exploit.

A Hydra Framework for smart contracts. We focus on smart contracts, programs that execute on blockchains such as Ethereum [14]. They are especially well suited as a use case given several distinctive properties:

- Heightened vulnerability: Smart contracts are often financial instruments. Bugs usually directly affect funds, changing value, creating a stable bounty marketplace.
- Unique economic properties: A smart contract’s cryptocurrency balance is often a direct measure of an exploit value. This facilitates principled bounty price setting in our framework. Moreover, blockchain protocols are often secured through both cryptography and economic guarantees. For the first time, we lift similar economic safety guarantees to the smart-contract level, creating programs with measurable economic security.
- Bounty automation: Application of our framework to and by smart contracts can award bounties automatically. The result is a fair exchange of bugs for bounty and guaranteed payment for the first valid submitted bug. Bounties are transparent to bounty hunters and can be adjusted dynamically to reflect contracts’ changing value, creating a stable bounty marketplace.
- Graceful termination: Smart contracts are not (yet) mission critical software and can often be aborted with minimal adverse effects, as required for NNVP. Remediation of the DAO and Parity multisig attacks involved refunding users, a mechanism considered in this paper.

We design a Hydra Framework for Ethereum and evaluate it on two applications, an ERC20 token [55] and a Monty Hall game [56]. In both cases, we produce three independent implementations of a common contract specification, using three different programming languages in the Ethereum ecosystem. The Hydra Framework automatically instruments these contract “heads” so that they interact with a common Hydra meta-contract. The meta-contract acts as a generic proxy that delegates incoming transactions to each head in turn and pays out a bounty in the event of a disagreement between the heads. Our Hydra ERC20 token is deployed on the Ethereum main network (with a 3000 USD bounty), the first principled, automated and trust-free bug bounty. Our framework is applicable to over 76% of Ethereum contracts in use. Our full framework code, tests, and experiments are available at thehydra.io.

Major challenges. Several papers [33, 21] criticize traditional N-version programming, observing that multiple versions of a program often exhibit correlated faults—an ostensible hitch in our framework.

We revisit these papers and show that NNVP achieves an appealing cost-benefit trade-off, by abandoning fault-tolerance in favor of error detection. Compared to the majority voting scheme used in N-version programming, partial independence is greatly amplified by NNVP, which requires agreement by all heads. Previous experimental results in fact show that NNVP can achieve a large exploit gap in Hydra programs. In particular, we review high-profile smart contract failures, showing that
NNVP would have addressed many of them.

A second challenge arises in automating bug bounties for smart contracts. Decentralized blockchain protocols allow adversaries to perform front-running—ordering their transactions ahead of those of honest users [51]. As a result, a naïvely implemented bounty contract is vulnerable to bug-withholding attacks: upon finding a bug in one head, a hacker can withhold it and try to compromise all heads to exploit the full contract. If an honest user discovers a bug, the hacker front-runs her and claims the bounty first. Thus, withholding carries no cost for the hacker, removing incentives for early disclosure.

We propose Submarine Commitments, a countermeasure of independent interest that temporarily conceals a bounty claim among ordinary transactions, preventing a hacker from observing and front-running a claim. We analyze its security, we present a strong, formal adversarial model that encompasses front-running and other attacks. We introduce a countermeasure of independent interest called Submarine Commitments and prove that it effectively prevents bug withholding.

Contributions. Our main contributions are:

• The Hydra Framework: We propose, analyze, and demonstrate the first general approach to principled bug bounties. We introduce the idea of an exploit gap and explore N-of-N-version programming (NNVP) as a specific instantiation. We demonstrate the power of NNVP Hydra programs in revisiting the N-version programming literature and provide the first quantifiable notion of economic security for smart contracts.

• Bug withholding and Submarine Commitments: We identify the subtle bug-withholding attack. To analyze its security, we present a strong, formal adversarial model that encompasses front-running and other attacks. We introduce a countermeasure of independent interest called Submarine Commitments and prove that it effectively prevents bug withholding. Front-running is a widespread, costly flaw in blockchain applications more general than bug withholding [8] [10] [12] [51], and Submarine Commitments provide a mitigation usable for exchanges, auctions, and other systems.

• Implementation: We implement a Hydra Framework for Ethereum and instantiate it for an ERC20 token and a Monty Hall game. We measure costs of running multi-headed contracts on-chain and showcase Hydra’s soundness and applicability, concluding that our framework can automatically transform the majority (76%) of contracts used in Ethereum while passing all official virtual machine tests. Our bounty-backed, three-headed Hydra ERC20 token is live on Ethereum.

2 Preliminaries and Notation

Programs. Let $f$ denote a stateful program. From a state $s$, running $f$ on input $x$ produces output $y$ and updates $s$. For an input sequence $X = [x_1, x_2, \ldots]$, we denote by $\text{run}(f, X) := [y_1, y_2, \ldots]$ a serial execution trace of $f$ starting at the initial state and outputting $y_i$ on input $x_i$.

Exploits. For a program $f$, let $I$ be an abstract ideal program that defines $f$’s intended behavior. I.e., for any input $X$, $\text{run}(I, X)$ is the correct output. The input space is assumed to be bounded and input sequences are finite.

We assume that a program may produce a fallback output if it detects that the execution is diverging from intended behavior (e.g., throwing an exception on a stack overflow). The ideal program $I$ never outputs $\bot$. If a program $f$ outputs $\bot$ on some input $x_i$, then all subsequent outputs in that execution trace will also be fallbacks. A program’s execution trace is a fallback trace if it agrees with the ideal program up to some input $x_i$ and then outputs $\bot$. Let $A \subseteq B$ denote that sequence $A$ is a strict prefix of sequence $B$. The set of fallback traces is then

$$\mathcal{Y}_\bot := \{Y \mid \exists i, y_i \subseteq \text{run}(I, X) \land \mathcal{N}_{j=i+1}(y_j = \bot)\},$$

We define an exploit against $f$ as any input sequence $X$ for which $f$’s output is neither that of the ideal program nor a fallback trace. If $E(f, I)$ denotes the exploit set of $f$ with respect to $I$, then $X \in E(f, I)$ if and only if $\text{run}(f, X) \notin \mathcal{Y}_\bot \cup \{\text{run}(I, X)\}$. Note that the notions of ideal program, fallback output, and exploit are oblivious to the representation of the program’s internal state.

Exploit gaps and bug bounties. A program transformation $T$ combines $N \geq 1$ programs into a program $f^* := T(f_1, f_2, \ldots, f_N)$. Our notion of exploit gap aims to capture the idea that $f^*$ has fewer exploits than the original $f_i$. However, directly relating the sizes $|E(f^*, I)|$ and $|E(f_i, I)|$ is problematic as these quantities are hard to measure. Instead, we define a probabilistic notion of exploit gap, for input sequences sampled from a distribution $D$ (e.g., the distribution of user inputs to a program).

Definition 1 (Exploit Gap). A program transformation $T(f_1, f_2, \ldots, f_N) := f^*$ introduces an affirmative exploit gap for a distribution $D$ over inputs sequences $X$ if

$$\text{gap} := \frac{\Pr_{X \in D} [X \in \bigcup_{i=1}^N E(f_i, I)]}{\Pr_{X \in D} [X \in E(f^*, I)]} > 1. \tag{1}$$

The exploit gap is empirically measurable and its magnitude reflects the likelihood that an input sequence that is an exploit for some $f_i$ does not affect $f^*$.

A transformed program $f^*$ that always returns $\bot$ induces a large exploit gap, yet has no utility. We therefore also require the following notion of availability.

Definition 2 (Availability Preservation). Let $F(f)$ be the set of inputs with fallbacks, i.e. $X \in F(f)$ iff $\text{run}(f, X) \in \mathcal{Y}_\bot$. Then a program transformation $T$ is availability-preserving iff $F(f^*) \subseteq \bigcup_{i=1}^N (E(f_i, I) \cup F(f_i))$. 


To be availability-preserving and yield an exploit gap, a program transformation may trade availability for correctness. That is, a transformed program may fallback on inputs that are exploits for some of the original programs.

Given a transformation $\mathcal{T}$ that induces an exploit gap, a natural bug bounty for a deployed program $f^*$ rewards bugs in the original programs $f_i$. Such a bug bounty scheme satisfies three important properties:

1. The bugs are efficiently verifiable, via differential testing: if $\text{run}(f_i, X) \neq \text{run}(f^*, X)$, then the input $X$ is an exploit against $f_i$ or $f^*$ or both.
2. A claimable bug need not be an exploit on $f^*$. If the exploit gap is large ($\text{gap} \gg 1$), then a discovered bug likely affects one of the programs $f_i$ but not $f^*$.
3. The bugs are valuable. If $\text{gap} > 1$, fixing bugs in the $f_i$ eventually reduces the probability of exploits in $f^*$.

**Achieving an exploit gap.** Generically, dynamic runtime checks (e.g., stack canaries, under- or overflow detection) can yield an availability-preserving exploit-gap: the checks result in a fallback output (e.g., a runtime exception), where the original program had an exploit.

A broadly applicable method for achieving an exploit-gap is via redundancy and fault-tolerance, e.g., Recovery Blocks [46] or N-version programming [17]. These transformations operate on $N > 1$ programs and aim at full availability (i.e., no fallback outputs), a natural requirement in mission-critical systems.

We focus on N-version (or multiversion) programming, which we build upon in Section 3. This software paradigm consists in three steps [17, 6]:

1. A specification is written for the program’s functionality, API, and error handling. It further defines how to combine outputs of different versions (see Step 3).
2. $N$ versions of the program specification are developed. Independence among versions is promoted via isolation (i.e., minimal interactions between developers) and diversity (i.e., different programming languages, or technical backgrounds of developers).
3. The $N$ versions are run in parallel and their outputs combined via some voting scheme. N-version programming traditionally uses majority voting between programs to induce an exploit gap [17, 6].

### 3 N-of-N-version Programming

N-version programming assumes that heterogeneous implementations have weakly correlated failures [17]. Many experiments have challenged this view [33, 21], questioning the cost-benefit trade-off of the paradigm. Our thesis is that smart-contract ecosystems present a number of key properties that render multiversion programming and derived bug-bounty schemes attractive.

The main differentiator between the traditional setting of N-version programming, and ours, is the role of availability. Prior works consider mission-critical systems and thus favor availability over safety in the face of partial failures. For instance, Eckhardt et al. [21] explicitly ignore the "error-detection capabilities" of N-version programming. This setup is not suitable for smart-contracts: As in centralized financial institutions (e.g., stock-markets [48]), the cost of a fault typically trumps that of a temporary loss of resource availability.

Ethereum’s community exemplified its preference for safety in this trade-off, when attackers found an exploit in the Parity Multisig Wallet [11] and stole user funds. A consortium of “white-hat hackers” used the same bug to move user’s funds to a safe account. Despite funds being unavailable for weeks, and reimbursement depending on the consortium’s goodwill, the action was acclaimed by the community and affected users. The simple escape hatch in this scenario (i.e., move funds to a safe account) was deemed a successful alternative to an actual exploit.

We propose trading availability for safety in N-version programming, by replacing the goal of fault-tolerance by one of error detection and safe termination. Suppose that programs $f_1, \ldots, f_N$ have no fallback outputs (i.e., $F(f_i) = \emptyset$). Then majority voting yields a program $f^*$ that also satisfies $F(f^*) = \emptyset$, but the exploit gap may be small. At the other end of the spectrum, we propose N-of-N-version programming (NNVP), wherein $f^*$ aborts unless all of the $N$ versions agree. NNVP is an availability-preserving transformation that induces a much larger exploit gap ($f^*$ only fails if all the $f_i$ fail simultaneously).

Table 1 lists prominent Ethereum smart contract failures. We discuss these in more detail in the extended version of this paper [13], and argue that a majority could have been abated with NNVP.

#### 3.1 Revisiting N-version Programming

We revisit experiments on the cost-effectiveness of N-version programming, in light of our NNVP alternative.

Knight and Leveson [34] first showed that the null-hypothesis of statistical independence between program failures should be rejected. Yet, such correlated failures only invalidate the N-version independence of partial failures. For instance, Eckhardt et al. [21] found that the correlation between individual versions’ faults could be too high to be considered cost-effective, with a majority vote between three programs reducing the probability of some fault classes by only a small factor (as we show in Appendix A, some of the workloads in [21] yield an exploit gap of $\text{gap} \approx 5$ using majority voting between three programs).

Fortunately, NNVP provides a better cost-benefit
Table 1: Selected smart contract failures and potential exploit gaps. The list is extended from [27]. For each incident, we report the value of affected funds (data from [1]), the cause of the exploited vulnerability, as well as the (hypothetical) potential for fault independence between multiple contract versions. Green lines indicate settings in which a Hydra contract would have likely induced a large exploit gap and prevented the exploit. Yellow and red lines indicate incidents that Hydra addresses only partially or not at all. Asterisks indicate ERC20 compatible contracts, like our bounty described in Section 6. More details are in the extended version of this paper [13].

<table>
<thead>
<tr>
<th>Contract name</th>
<th>Exploit value (USD)</th>
<th>Root cause</th>
<th>Independence source</th>
<th>Exploit gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parity Multisig 2 [50]</td>
<td>300M</td>
<td>Delegate call+exposed self-destruct</td>
<td>programmer/language?</td>
<td>✓/✗</td>
</tr>
<tr>
<td>The DAO* [18]</td>
<td>150M</td>
<td>Re-entrancy</td>
<td>language</td>
<td>✓</td>
</tr>
<tr>
<td>Proof of Weak Hands [7]</td>
<td>1M</td>
<td>Arithmetic overflow</td>
<td>programmer+language</td>
<td>✓</td>
</tr>
<tr>
<td>SmartBillions [49]</td>
<td>500K</td>
<td>Bug in caching mechanism</td>
<td>programmer</td>
<td>✓</td>
</tr>
<tr>
<td>HackerGold (HKG)* [40]</td>
<td>400K</td>
<td>Typo in code</td>
<td>programmer+language</td>
<td>✓</td>
</tr>
<tr>
<td>MakerDAO* [47]</td>
<td>85K</td>
<td>Re-entrancy</td>
<td>language</td>
<td>✓</td>
</tr>
<tr>
<td>Rubixi [16]</td>
<td>&lt;20K</td>
<td>Wrong constructor name</td>
<td>programmer+language</td>
<td>✓</td>
</tr>
<tr>
<td>Governmental [16]</td>
<td>10K</td>
<td>Exceeds gas limit</td>
<td>None?</td>
<td>✗</td>
</tr>
</tbody>
</table>

3.2 Smart Contracts are NNVP-Friendly

In addition to favoring safety over availability, other properties of smart contract ecosystems (and Ethereum in particular) render NNVP bug bounties attractive:

- **High risk for small applications.** Smart contracts store large financial values in small applications with an exceptionally high “price per line of code” (some token contracts hold over 1M USD per line [1]). Contract code is stored on a public blockchain and exploits often directly extract or destroy stored funds. Yet developing multiple versions is typically cheap in absolute terms.

- **Principled bounty pricing.** A contract’s balance is often a direct measure of an exploit’s market value. This facilitates our analysis of principled bounty pricing that incentivizes early disclosure of bugs (see Section 4).

- **Bounty automation.** Smart contracts enable automation of the full bounty program, from bug detection (with differential testing) to rollback to bounty payments. Bounties administered by smart contracts can satisfy *fair exchange* of bounties for bugs and *guaranteed payment* for disclosure of valid bugs [53]. Bounties are also *transparent* (i.e., the bounty is publicly visible on the blockchain) and may be *dynamically* adjusted to reflect a contract’s changing exploit value. The result is a stable, decentralized bounty market.

- **Programming language diversity.** Many exploits in Ethereum arose due to specific language idiosyncrasies. The multiple interoperable languages for Ethereum enable potentially diverse implementations.

3.3 The Hydra Contract

Hydra consists of two program transformations. The first, $T_{NNVP}$, uses the NNVP paradigm to yield an availability-preserving exploit gap. $T_{NNVP}$ combines $N$ smart contracts (or heads) $f_1, \ldots, f_N$ into a contract $f^*$, which delegates incoming calls to each head. If all outputs match, $f^*$ returns the output; otherwise, $f^*$ reverts all state changes and returns ⊥.

The idea is depicted in Figure 2. The heads are individually deployed and *instrumented* such that they only interact with the Hydra meta-contract (MC). The MC is the logical embodiment of the contract functionality (i.e., the MC holds all assets, and interfaces with external contracts and clients). To maintain consistency while interacting with external contracts, the MC checks that all heads agree on which external interaction to perform, executes the interaction *once*, and distributes the obtained response (if any). Our design and implementation of the $T_{NNVP}$ transformation for Ethereum smart contracts is described in Section 6.

The second transformation $T_{Bounty}$ is responsible for paying out a bounty and providing escape-hatch functionality. It transforms a program $f^*$ into a program $\hat{f}$ which forwards any input to $f^*$ and then returns $f^*$’s output, unless $f^*$ returns ⊥. In the latter case, $\hat{f}$ will pay out a bug bounty to its caller and enter an *escape hatch mode*.

**Escape hatches.** Ideally, bugs could be patched online. This is hard in Ethereum as smart contract code cannot be updated after deployment [41]. Best practices [23] suggest enhancing smart contracts with an *escape hatch*...
mode, which enables the contract’s funds to be retrieved, before its eventual termination and redeployment.

The design of the escape hatch mode depends on the application, but there are some universal design criteria:

- **Security**: The escape hatch’s correctness requires special care, as it will not be protected by NNVP.
- **Availability**: The escape hatch must be available for the contract’s entire lifetime, or assets could end up stuck.
- **Distributed trust**: All assets should be returned to their rightful owners, or distributed among multiple parties.

For instance, contract funds could be sent to an audited *multisig* contract (possibly implemented as a Hydra contract itself), to distribute trust among multiple parties.

### 4 Economic Analysis of Hydra Bounties

We formally analyze the exploit gap induced by the Hydra contract, and derive a bounty pricing model to incentivize bug disclosure. We assume that bounties are paid out immediately upon bug disclosure. In Section 5, we refine our analysis in the blockchain model, wherein an adversary may reorder messages sent to smart contracts.

#### 4.1 Bug Finding as a Stochastic Process

We consider a set of parties that try to find vulnerabilities in a Hydra contract $f^*$ composed of $N$ heads $f_1, \ldots, f_N$. For simplicity, we slightly overload notation and identify an exploit with the input that ultimately causes the contract’s outputs to depart from the ideal behavior $\mathbb{I}$ (although the internal state of $f^*$ may have been corrupted earlier). That is, $x$ is an exploit if $\text{run}(f^*, X \cup \{x\}) \neq \text{run}(\mathbb{I}, X \cup \{x\})$, where $X$ is the sequence of all inputs previously submitted to $f^*$ and $\cup$ denotes concatenation.

If an honest party finds an input $x$ that yields an exploit for at least one of the heads ($\exists i \in [1, N] : x \in E(f_i, \mathbb{I})$), then the party is awarded a bounty of value $\text{Sbounty}$ and the contract’s escape hatch is triggered. If a malicious party finds an exploit against the full Hydra contract ($x$ is an exploit for each head), the party can use this exploit to steal the entirety of the contract’s balance, **Sbalance**.

We model bug finding as a Poisson process with rate $\lambda_i$, which captures a party’s work rate towards finding bugs. We assume that parties sample inputs $x$ from a common distribution of potential exploits $D$. We then recover our exploit gap notion (Definition 1) by considering the difference in arrival times of two random events: (1) a party discovers a flaw in one of the heads; (2) a party finds a full exploit. The waiting times for both events are exponentially distributed with respective rates $\lambda_i$ and $\lambda_i \cdot \text{gap}^{-1}$.

Let us first consider the strong assumption of independent program failures. For a head $f_i$, let $p$ be the probability that an input $x \in D$ is an exploit for $f_i$. We get

$$\text{gap} = \frac{\Pr_{x \in D} \left[ x \in \bigcup_{i=1}^N E(f_i, \mathbb{I}) \right]}{\Pr_{x \in D} \left[ x \in E(f^*, \mathbb{I}) \right]} = \frac{1 - (1 - p)^N}{p^N},$$

which grows exponentially in $N$, for $p \in (0, 1)$.

The gap can be empirically estimated using Equation (1). For the test suites considered in the experiments of Eckhardt et al. [21], the average gap for three program variants is 4400 (see Appendix A for details).

#### 4.2 Analyzing Economic Incentives

We assume a set of honest parties with combined work rate $\lambda_H$. These bounty hunters only try to exchange bugs for bounties. Note that a bug in all heads (i.e., a full exploit) cannot be detected and rewarded by the meta-contract $f^*$. We thus let $\lambda_H$ be the rate at which honest parties find bugs that affect $1 \leq k < N$ heads.

To analyze economic incentives of bounties, we consider malicious parties which, if given an exploit, would deplete the contract’s balance. W.l.o.g., we model a single adversary $A$ with work rate $\lambda_A$. Indeed, for $m$ (non-colluding) adversaries with work rates $\lambda_1, \ldots, \lambda_m$, it suffices to analyze the party with rate $\lambda_A = \max_{1 \leq i \leq m} \lambda_i$. If the bounty incentivizes this party to act honestly, less efficient parties will have the same incentive.

Let $T_{\text{H}}$ be the waiting time until an honest party finds a bug. $T_{\text{H}}$ is exponentially distributed with rate $\lambda_H$. Let $T_M$ be the waiting time until $A$ finds an exploit against $f^*$, which is exponential with rate $\lambda_A \cdot \text{gap}^{-1}$. We analyze two cases: (1) $A$ finds an exploit against $f^*$, and (2) $A$ finds a bug for a strict subset of the heads.
In the first case, it is clear that $A$ has no incentive to disclose, unless the bounty exceeds the contract’s value. This is the situation of a “traditional” bounty scheme. However, the probability of this bad event occurring is

$$\Pr[T_M < T_H] = \frac{\lambda_H \cdot \text{gap}^{-1}}{\lambda_H + \lambda_M \cdot \text{gap}^{-1}} = \frac{\lambda_M}{\lambda_H \cdot \text{gap} + \lambda_M},$$

which naturally decays as the exploit gap increases.

In the second case, a bounty can incentivize early disclosure. Suppose $A$ found a bug in a head. If $A$ discloses it, her payout is $\text{payout}_H := $bounty. Instead, if she conceals the bug and continues searching for exploits, she risks a payout of 0 if another party claims the bounty first. Her expected payout, $\text{payout}_M$, is thus

$$\Pr[T_M < T_H] \cdot \text{balance} = \frac{\lambda_M}{\lambda_H \cdot \text{gap} + \lambda_M} \cdot \text{balance}.$$

Let $\alpha := \frac{\lambda_H}{\lambda_M}$. Then, honest behavior is incentivized if

$$\frac{\text{payout}_H}{\text{payout}_M} > 1 \iff \text{bounty} > \frac{1}{\alpha \cdot \text{gap} + 1} \cdot \text{balance}.$$ 

We may assume that $\lambda_M = \lambda_H$ (i.e., $A$’s work rate is equal to the combined work rate of honest parties). Then, for independent program failures (see Equation (3)) the bounty decays exponentially in the number of heads $N$.

Thus, given estimates of $\alpha$ and gap, we get a principled bounty pricing that incentivizes bug disclosure. For example, in the experiment of Eckhardt et al. [21], a three-headed Hydra could sustain a bounty 3 to 4 orders of magnitude below an exploit’s value.

This analysis also provides insight into why bounties are paid when bugs are not necessarily actively exploitable against the target system. If $\text{bounty}$ is too small, all economically rational players will attempt to privately weaponize any partial exploits they develop. Traditional bounties operate off similar intuition, with tiers of exploit values to boost participation (e.g., [24]).

5 The Bug-Withholding Problem

Our analysis in Section 4 assumed that a bounty is paid immediately when a bug is claimed. Hereafter, we refine our analysis by modeling bounty smart-contract execution with respect to a powerful adversary, that can cheat users by exploiting blockchain network protocols. We highlight the bug-withholding attack and propose and analyze a solution called Submarine Commitments.

Front-running. The issue is that transactions may not be ordered in blocks by network submission time. When a user sends a bounty-claim transaction $\tau$ to the network, an adversary may front-run the user, and insert its own bounty-claim $\tau'$ earlier in the chain [51]. It does this by ensuring faster network propagation of $\tau'$ or by causing a miner to order $\tau'$ before $\tau$, e.g., by paying a higher fee (more gas in Ethereum) or corrupting the miner.

Front-running opens up a bug bounty system to bug-withholding attacks. Suppose an adversary has found a bug in one or more heads in a Hydra contract, and aims to find a stronger exploit against all heads. If another party in the meantime claims the bounty, the adversary’s progress is wiped out: It loses all potential payoff on its already discovered bugs. By front-running, though, the adversary can ensure it claims the bounty first, thus nullifying any economic incentives for early disclosure.

We propose a formal model for blockchain security, expressed as an ideal functionality $F_{\text{withhold}}$. It captures front-running, but is far stronger than previous models (e.g., Hawk [35]). We present a basic bug-bounty contract BountyContract in $F_{\text{withhold}}$. Refining our analysis of Section 4, we show how bug withholding breaks incentives for bug disclosure in BountyContract. We show that commit-reveal schemes are an insufficient defense, and therefore introduce Submarine Commitments. We prove, in an $F_{\text{withhold}}$-hybrid world, that using Submarine Commitments for BountyContract drastically reduces the payoff of a bug-withholding adversary.

5.1 Adversarial Model

We model an adversary $A$ that can front-run a victim. In our model, $A$ can mount strong history-revision attacks, overwriting blocks at the head of the blockchain, and can delay any transaction by a bounded number of blocks.

This reflects an adversary’s ability to monitor transactions, mount network-level attacks, control client accounts, and even corrupt or bribe miners to alter legitimate blocks. Previous models, e.g., [35], considered weaker attacks in which $A$ can arbitrarily reorder transactions in a pending block. They are equivalent to weak history-revision attacks with only a single block.

In our model, $A$ itself constructs the blockchain. $A$ controls all but one honest player, denoted $P_0$. ($P_0$ models the collective behavior of all honest players.) $A$ can reorder $P_0$’s transactions by: (1) Rewinding the blockchain from its head, i.e., mounting a history-revision attack, for a sequence of up to $\rho$ blocks; and (2) Delaying the posting on the blockchain of a transaction by $P_0$ by up to $\delta$ blocks. We call such an adversary $A$ a $(\delta, \rho)$-adversary.

Our adversarial model takes the form of an ideal functionality $F_{\text{withhold}}$ characterizing an $(\delta, \rho)$-adversary $A$. We give details on $F_{\text{withhold}}$ in the extended version of this paper [13].

Notation. Let $\mathcal{B} = \{B_1, \ldots, B_{\text{Height}}\}$ be a blockchain, i.e., an ordered sequence of blocks. Here, $\text{Height}$ is the
5.2 The BountyContract Smart Contract

Within the $F_{withhold}$-hybrid model, we specify a contract
BountyContract to administer a single bug bounty, using
a simple commit-reveal scheme to prevent adversarial
copying and resubmission of bugs. BountyContract has parameters $\Delta > \delta + \rho$, $\Delta depot$ and $\Delta bounty$. It takes
as input a commitment to a bug in some block $B_i$ (via
transaction "commit"), which must be revealed before
block $B_{i+\Delta}$ (via transaction "reveal"). After a delay $\Delta$, the
player with the first validly revealed commitment may
claim the bounty (via transaction "claim"). A "commit"
incurs a cost of $\Delta depot$, to prevent $A$ from committing
in every block and revealing only if $P_0$ also reveals.

We assume a function isvalidbug that determines
whether a submitted bug is valid. In the $F_{withhold}$-hybrid
model, BountyContract is fed a height-$n$ blockchain $B$, which is replaced after being generated by $F_{withhold}$, i.e.,
transactions are executed as ordered by $F_{withhold}$ in $B$.

Bug withholding in BountyContract. The contract in
Figure 3 uses a cryptographic commit-reveal scheme, a
central folklore solution to certain front-running att-
acks [31]. This works if $A$ cannot post a valid com-
mmitment itself until it sees a victim’s reveal. For
instance, BountyContract prevents $A$ from trying to learn
and steal the committed bug from an honest player $P_0$.

Unfortunately, this approach does not protect against
front-running in the $F_{withhold}$-hybrid model if $A$ is with-
holding a bug it already knows. Here, $A$ waits until $P_0$
sends a "commit". $A$ then knows that $P_0$ is trying to
claim a bounty, and can front-run $P_0$’s commitment by
posting her own "commit" ahead in the blockchain.

This problem arises in many other scenarios, e.g., to-
ken sales or auctions, where a user must send funds to
place her bid, thus exposing the bid on the blockchain.

Impact of bug withholding. In our analysis of Hydra
bug bounties in Section 4.2, we assumed that $A$ risks for-
feiting a payout of $\Delta bounty$ if she conceals a bug. How-
ever, front-running has the potential of removing incen-
tives for early disclosure, as $A$ can ensure a payout of
$\Delta bounty$ by front-running the honest bounty hunter.

If $A$ conceals a bug, she finds a full exploit before the
bounty is claimed with probability $q := \Pr[T_M < T_H]$. Otherwise she front-runs and steals the bounty. Her ex-
pected payout is $q \cdot \Delta balance + (1 - q) \cdot \Delta bounty$.

If $A$ discloses the bug, her payout is $\Delta bounty$. To in-
centivize disclosure, we need $\Delta bounty > \Delta balance$, as in
a standard bounty with no exploit gap. We now show a so-
lution that thwarts bug-withholding attacks in Ethereum,
thus re-instantiating positive incentives to disclose bugs.

5.3 Submarine Commitments

We present a bug-withholding defense called a Subma-
rine Commitment. This is a powerful, general solution
to the problem of front-running that may be of indepen-
dent interest, as it can be applied to smart-contract-based
auctions, exchange transactions, and other settings.

As the name suggests, a Submarine Commitment is a
transaction whose existence is temporarily concealed,
but can later be surfaced to a target smart contract. It may
be viewed as a stronger form of a commit-reveal scheme.
Achieving Submarine Commitments is challenging in
systems like Ethereum, however, because message con-
tents and currency in all transactions are in the clear.

Briefly, in Ethereum, to commit in a Submarine Com-
mmitment scheme, $P$ posts a transaction $\tau$ that sends (non-
refundable) currency $\Delta val > \Delta depot$ to an address $\hat{addr}$. This address is itself a commitment of the form

$$\hat{addr} = H(\hat{addr}(Contract), H(\hat{addr}(P), key), data),$$

for $H$ a commitment scheme (e.g., hash function in the
ROM), $key$ a randomly selected witness (e.g., 256-bit string), and $data$ other ancillary information. $P$’s address
is included in the commitment to prevent replay by $A$.
To reveal, $P$ sends key to Contract. A Submarine Com-
mmitment scheme includes an operation DepositCol-
lection that permits Contract to recover $\Delta val$ using $\hat{addr}(P)$
and key. This scheme has these key properties:

1. Commit: As key is randomly selected, $\hat{addr}$ is indis-
tinguishable from random in the view of $A$.
\( \tau \) has no ascertainable connection to Contract, and looks to \( A \) like an ordinary send to a fresh address.

2. **Reveal**: After learning key, Contract can compute \( \text{addr} \) as above and verify that \( \$val \) was sent correctly. Via \( \text{DepositCollection} \), Contract recovers \( \$val \) thus avoiding unnecessary burning of funds.

Thus if \( A \) does not know \( P \)'s address (honest bounty hunters could use a mixer), and \( \$val \) is sampled from an appropriate distribution of values \( \$val \geq \$deposit \), \( A \) cannot distinguish transaction \( \tau \) from other sends to fresh addresses. As we show in Appendix B.2, such sends are common in Ethereum and, for a reasonable commit-reveal period (e.g., 25 minutes), form an anonymity set of hundreds of transactions with a diverse range of values among which \( \$val \) is statistically hidden. Notably, the anonymity set represents 2-3% of transaction traffic over the commit-reveal window. Two concrete Submarine Commitment constructions are in Appendix B.

### 5.4 Analysis of Submarine Commitments

We prove that Submarine Commitments strongly mitigate bug withholding in BountyContract. Our analysis uses a game-based proof in the \( \mathcal{F}_{\text{withhold}} \)-hybrid world. Details are in [13], although our model is understandable without detailed knowledge of \( \mathcal{F}_{\text{withhold}} \).

**Withholding game: \( \text{Exp}_{\mathcal{A}}^{bntyrace} \).** Figure 4 shows the simple game used in our security analysis, denoted by \( \text{Exp}_{\mathcal{A}}^{bntyrace} \). The game is played between an honest user \( P^* = P_b \), and a user \( P_t \) controlled by \( A \). W.l.o.g., \( P^* \) models a collection of honest players, while \( P_t \) models players controlled by \( A \). \( A \) interacts with \( P^* \) in the ideal functionality \( \mathcal{F}_{\text{withhold}} \). Let \( \Delta > \delta + \rho \), where \( \delta \) and \( \rho \) are the number of blocks by which \( A \) can delay or rewind in \( \mathcal{F}_{\text{withhold}} \). The experiment considers an interval of \( n \) blocks in a blockchain \( B \) of length \( n' = n + \Delta \).

In this game, a player can send only two messages: ("commit", \( \$deposit \)), and "reveal". To model Submarine Commitments, we assume that \( P^* \)'s commit message is opaque to \( A \), i.e., \( A \) cannot detect its presence in a block and it does not count toward the block’s size.

For clarity’s sake, we first analyze Submarine Commitments outside the Poisson framework of Section 4. Our results also hold in that setting, with a slightly tighter bound for Theorem 3 below (see [13] for a proof).

Instead, we consider a blockchain interval of \( n \) blocks, wherein \( P^* \) commits in a block chosen uniformly at random. That is, \( P^* \) posts ("commit", \( \$deposit \)), in the block at index \( \text{commblock}_{P^*} \leftarrow \{1, n \} \). \( P^* \) posts a "reveal" in block \( \text{revblock}_{P^*} = \text{commblock}_{P^*} + \rho \).

\( A \) wins the game if she posts a valid "commit" before \( P^* \) does, and also posts a corresponding "reveal" to claim the bounty. We let \( p_{\text{wins}} = \Pr[(\text{TRUE}, \cdot) \leftarrow \text{Exp}_{\mathcal{A}}^{bntyrace}] \), as a first goal, an economically rational adversary \( A \)'s aims to maximize its expected payoff, namely

\[
E[\text{payoff}] = p_{\text{wins}} \cdot \$\text{bounty} - E[\$\text{cost}].
\]

Of course, \( A \) can always post a "commit" in block \( 1 \), followed by a "reveal" within \( \Delta \) blocks, in which case it achieves \( p_{\text{wins}} = 1 \) with \( \text{payoff} = \$\text{bounty} - \$\text{deposit} \), which is optimal. But then it achieves no withholding.

**Results.** A compelling withholding strategy for \( A \) is to reveal a bug only by front-running \( P^* \), i.e., a pure front-running strategy. That is, if \( P^* \) sends a "reveal" in block \( B_j \), then \( A \) learns that \( P^* \) posted a "commit" in block \( B_{j-\rho} \). \( A \) can rewind and post its own "reveal" earlier than \( P^* \). But \( A \) can rewind most \( \rho \) blocks (i.e., block \( B_{j-\rho} \) cannot be erased), so \( A \) only succeeds if it has previously posted a "commit" in the interval \( [B_{j-\rho}, B_{j-\rho}] \).

We show that for natural parameters, \( A \) achieves no benefit, i.e., positive expected payoff, via pure front-running. Intuitively, this is because front-running is expensive: Since \( A \) observes a "commit" message from \( P^* \) too late to remove it by rewinding, \( A \) must post "commit" messages continuously to ensure that it can front-run \( P^* \). The proof of the following theorem is in the extended version of this paper [13].

**Theorem 3.** Let \( \Delta \geq 4 \) and \( \$\text{deposit} > \frac{10(\Delta+1)}{b_{\text{by}}} \$\text{bounty} \). Then a pure front-running adversary has \( E[\text{payoff}] < 0 \).

This result is fairly tight and enables practical parameterizations of BountyContract, as this example shows.

**Example 1.** Consider a bounty in Ethereum, with 15-second block intervals. Suppose that \( \$\text{bounty} = 100,000 \) USD, that the period over which \( A \) competes with honest bounty hunters is one week, and that a commitment must be revealed in \( \Delta = 100 \) blocks. Then given \( \$\text{deposit} \geq 278 \) USD, a pure front-running adversary cannot achieve a positive expected payoff (i.e., \( E[\text{payoff}] > 0 \)).
Of course, $A$ could use other strategies. In the extended version of this paper [13], we consider a generalized $\alpha$-revealing strategy that involves conditional pre-emptive bug disclosure. We show that this strategy does no better than pure front-running.

6 Design and Implementation

We implemented a decentralized automated bug bounty for Ethereum smart contracts. We describe the main technical deployment challenges, and explain our design.

The EVM. The Ethereum Virtual Machine (EVM) is a simple stack-based architecture [57]. Smart contracts can access three data structures: a stack, volatile memory, and permanent on-chain storage.

Execution of a contract begins with a transaction sent to the blockchain, specifying the called contract, the call arguments, and an amount of ether, Ethereum’s currency. The EVM executes the contract’s code in a sequential, deterministic, single-threaded fashion. Operations can read and write to stack, memory or storage, and spawn a new call frame (with a fresh memory region) by calling other contracts. Each instruction costs a fixed amount of gas, a special resource used to price transactions.

Contracts can exceptionally halt, revert all changes made in the current call frame (e.g., storage updates, transfers of ether), and report an exception to the callee.

6.1 An EVM Execution Environment

To achieve the full power of our Hydra bug bounty, $N$ smart contract versions are run on the blockchain. While we could also run a bounty program off-chain (for a single deployed contract), this would not provide an exploit gap, a key property in our analysis of attacker incentives.

The main challenge is the implementation of the “Execution Environment” [17, 6], the agent that coordinates the $N$ heads and combines their outputs. Its complexity should be minimal, as it is part of the Trusted Computing Base (TCB) of our application: a bug in the coordinating agent is likely an exploit against the Hydra contract.

A proxy meta-contract. As we showed in Figure 2, the logical embodiment of a Hydra contract is a proxy meta-contract (MC), which coordinates $N$ deployed contract versions (or heads). Clients and other contracts only interact with the MC. The heads only respond to calls from the MC, and do not hold any ether themselves.

The MC delegates all incoming calls to each head, and verifies that the obtained outputs match. If so, it returns that output. Otherwise, it throws an exception, to revert all changes made by the heads. The $T_{Bounty}$ transformation described in Section 3.3 is implemented as a simple wrapper around the MC, which catches the above exception, pays out a bounty, and enters an escape-hatch mode.

Maintaining consistent blockchain interactions. As the EVM execution is deterministic, the result of a contract call is fully determined by the call’s input, the contract code and the current blockchain state. If smart contracts were executed in isolation, the above proxy contract would thus be sufficient. However, most smart contracts also interact with the blockchain, e.g., by accessing information about the current transaction (such as the sender’s address) or by calling other contracts, and the MC must thus guarantee consistency among the heads.

We illustrate the issue in Figure 5 with a Solidity code snippet (top-left) and corresponding EVM opcodes (bottom-left). The function $f(x)$ makes a call to $g(x)$ in the calling contract ($\text{msg.sender}$) and reimburses any sent ether ($\text{msg.value}$). If used as a head in a Hydra contract, this code snippet presents multiple issues.

1. \text{CALLVALUE} and \text{CALLER} are modified when the MC delegates a call to the head. \text{CALLER} will now be the MC’s address, and \text{CALLVALUE} will be zero.
2. The heads cannot send ether as they do not hold any.
3. With $N$ heads, $g(x)$ is called $N$ times instead of once.

The heads might also obtain different return values.

To resolve these issues, the heads are \textit{instrumented} prior to deployment so that all interactions with the blockchain are mediated by the MC. While these modifications could be made in a high-level language (e.g., Solidity), we opt for a more generic, automated, and globally applicable solution that operates on the EVM opcodes of a compiled contract (the instrumentation is thus agnostic to the language used to develop the heads). Opcode instructions are essentially of two types:

- \textit{Environment Information}. We ensure that all heads share the view of a common Hydra contract. The \texttt{ADDRESS} opcode (which returns the current contract’s address) is modified to return the MC’s address. The heads reject all calls that do not emanate from the MC. The MC also forwards \texttt{CALLVALUE} and \texttt{CALLER} to the heads as extra call arguments, to make the proxy delegation transparent. These opcodes are overwritten accordingly in the heads to read from the call data.
- \textit{System Operations}. Opscodes that interact with other blockchain entities (e.g., calling a contract, reading account balances, or logging messages) are rewritten as \textit{callbacks} to the MC. The MC checks consistency among the heads’ callbacks and issues the required operations on their behalf. The instrumentation requires some extra volatile memory to store callback argu-
Tail-call optimization. A design pattern for smart-contracts (“Checks-Effects-Interactions” [23]) suggests that interactions with other blockchain entities should occur last in a call. For contracts that follow this paradigm, a tail-call optimization can be applied to callbacks.

Instead of calling into the MC, the heads simply append any required call or log operations to the calls’ return value. Operations that read blockchain state (e.g., balance checks) are not instrumented. The MC then collects the return values from all heads, verifies consistency, and executes all interactions before returning.

Figure 5: EVM instrumentation of Hydra heads (simplified example). (Top left) Solidity function that calls $\text{g}(x)$ in the calling contract (msg.sender) and sends back all ether (msg.value). (Bottom left) EVM bytecode for the call to $\text{g}(x)$. (Top right) Functionally equivalent Solidity code for the instrumented bytecode. (Bottom right) Instrumented bytecode: CALLVALUE and CALLER are read from function arguments. All call data is stored in memory and used as arguments for a callback to the MC. (Top right) Exception handling. Recall that the EVM halts when contracts perform illegal operations, e.g., explicitly throwing exceptions or running out of gas. Ideally, we would classify any divergence in the heads’ behavior as a bug and pay a bounty. However, it is easy to set gas amounts so that one head runs out of gas, yet others succeed. Explicit exceptions are thus instrumented to return a special value to the MC, so as to be distinguished from an out-of-gas exception. If all heads throw an explicit exception, the MC propagates the exception to the caller.

6.2 Limitations

Our Hydra head instrumenter, written in Haskell, applies simple opcode rewriting rules (see Figure 5), which are verified to preserve program invariants such as stack and memory layout. Our modifications impact the heads’ gas consumption, yet the overhead is minor (see Section 7). Rewriting opcodes also modifies the layout of the bytecode, so all JUMP instructions are updated accordingly.

The instrumentation applies to contracts written in any high-level language that compiles to the EVM, and requires no changes to the EVM. We have not yet implemented callbacks for the infrequent CREATE and SELFDESTRUCT opcodes. We do not yet support opcodes that modify a head’s code (e.g., DELEGATECALL). These are often used to load libraries into a contract. Using such opcodes would require the library code to be instrumented itself, which is possible in principle. We note however that code delegation is typically at odds with the multiversion programming philosophy: if all heads call the same library contract, a library bug could yield an exploit. We leave Hydra-based libraries to future work.
7 Evaluation

This paper’s goal is not to rigorously measure correlations between smart contract faults, but to propose a novel principled bug bounty framework built upon an assumed exploit gap. We leave a thorough analysis of smart contract failure patterns to future work. We evaluate our framework under standard software metrics: TCB size, soundness, applicability and performance. We conclude with a discussion of our development process.

Workloads. To test soundness, applicability and performance of Hydra contracts, we use three workloads: (1) The official suite of test contracts for the EVM\(^1\); (2) All contracts used in Ethereum between Dec. 7 2017 and Feb. 7 2018; and (3) two representative smart-contract applications developed by the authors. Implementations of Submarine Commitments in Ethereum, and a thorough analysis of the resulting anonymity sets for bounty claiming transactions are in Appendix B.

We developed a generic ERC20 contract [55] for token transfers, and a Monty Hall Lottery, wherein two participants play a multi-round betting game [56]. In both cases, three authors independently developed one head in each of Solidity, Serpent, and Vyper, the main programming languages in Ethereum. These languages have different design tradeoffs (in terms of ease-of-use, low-level features or security) and are by themselves a valuable source of diversity between our Hydra heads.

- **The Hydra ERC20 token:** The ERC20 token-transfer API has been thoroughly peer reviewed [55], and is supported by most of the highest-dollar contracts in Ethereum (as of February 2018, the combined market cap of the top ten Ethereum tokens is over 20 billion USD [1]). Notably, the exploit in the DAO [15] was partially present in the code managing tokens.

  Our three-headed Hydra token is deployed on the main Ethereum network and can be used as a drop-in replacement for any ERC20 token, e.g., in the DAO [15] and ether.camp [40] contracts. When a user submits a token order, the MC delegates to all heads and validates the order upon agreement. Our initial bounty is 3000 USD, which we will increase as the choice takes the form of a cryptographic commitment that is later opened to reveal the winner. If either party aborts, the other party can claim both the reward and bet after a fixed timeout. The specification leaves the internal representation of the game open to developers.

- **A Hydra Monty-Hall lottery:** In this game, one party, the house, first hides a reward behind one of \(n\) doors. The player bets on the winning door, and the house opens \(k\) other non-winning doors. The player may then change his guess. If he guessed correctly, the player wins the reward; otherwise the house collects the bet.

  A fourth author wrote a specification describing the contract’s API and behavior. The house’s initial door

<table>
<thead>
<tr>
<th>Opcode</th>
<th>Contracts</th>
<th>Transactions</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODECOPY</td>
<td>50,147 (14%)</td>
<td>5,646,607 (27%)</td>
<td>medium</td>
</tr>
<tr>
<td>CALLCODE</td>
<td>30,109 (8%)</td>
<td>1,213,064 (6%)</td>
<td>hard</td>
</tr>
<tr>
<td>SELFDESTRUCT</td>
<td>24,707 (7%)</td>
<td>739,249 (4%)</td>
<td>easy</td>
</tr>
<tr>
<td>DELEGECALL</td>
<td>19,749 (6%)</td>
<td>2,695,326 (13%)</td>
<td>hard</td>
</tr>
<tr>
<td>CREATE</td>
<td>11,559 (3%)</td>
<td>1,143,961 (5%)</td>
<td>easy</td>
</tr>
<tr>
<td>Other</td>
<td>6681 (2%)</td>
<td>195,569 (1%)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Frequency of main unsupported opcodes. For blocks 4690101 to 5049100 on the Ethereum network, we count how many transactions use an opcode that currently cannot be handled by our Hydra Framework. We further record the fraction of unique smart contract codes that contain those opcodes, and the difficulty in adding support for each opcode.

TCB size. Our design from Section 6.1 is generic, and covers both of our target applications (and the majority of our other workloads, see below). The instrumenter for Hydra heads is written in 1500 lines of Haskell, and applies simple code parsing and rewriting rules. The MC’s proxy functionality is implemented in EVM assembly. We also wrote an MC in Solidity (185 lines) that applies tail-call optimization to callbacks. As the Hydra Framework is application-agnostic, we believe this is a reasonable TCB. It should also be relatively easy to write a formal specification for the simple functionality of the MC and instrumenter, although we have not attempted this.

Completeness and correctness. To evaluate completeness of our Hydra instrumenter, we consider all Ethereum transactions for blocks 4690101 to 5049100 (Dec. 7 2017 to Feb. 7 2018). For each transaction, we test whether our instrumenter supports the evaluated code (see Section 6.2 for unsupported opcodes). We find that 61% out of 21M transactions, or 76% of 350K unique smart contracts, are compatible with Hydra. Table 2 breaks down the contracts that Hydra currently cannot handle. This analysis supports the fact that Hydra could be usable for the majority of Ethereum contracts, both by deployed code and transaction volume.

We verify soundness by running the official EVM test suite\(^1\) on Hydra contracts. That is, we replace every contract in the test suite by a Hydra contract, and ensure all observable side effects (e.g. logs, external calls, return values, computation outputs) are unchanged. This test suite is used to evaluate EVM implementations, including executable formal specifications of the virtual machine [27]. It is thus critical that the suite be comprehensive: any gap in coverage represents a potential con-
Gas costs. Running \( N \) copies of a smart contract incurs an overhead on gas consumption. Some Ethereum projects, notably the Vyper language, already trade gas efficiency for security. Moreover, a transaction’s gas cost can be offloaded onto the contract owner, thus dispensing users from Hydra’s gas overhead. In any event, for small yet common workloads, the main gas cost of a transaction is the fixed “base fee”. As the MC calls all the heads in a single transaction, this fee is amortized, leading to sub-linear scaling of the gas-cost for \( N \)-headed Hydras.

Figure 6 compares gas costs for Hydra contracts with 1-5 heads to a linear scaling of a single non-instrumented contract. We show results for the five non-static calls in the ERC20 API, and for a full Monty Hall game (five transactions), with and without tail-call optimization.

For the ERC20 contract, the main cost is the transaction’s base fee of 21,000 gas. A call to the MC incurs an overhead of about 8000 gas (independent of the number of heads) or about \( 0.08 \text{ USD}^2 \). Each function call ends in a LOG callback to the MC (to log an “Approval” or “Transfer” event, as mandated by the ERC20 specification). The withdraw function also sends ether to the calling party. Applying the tail-call optimization results in significant savings for these callback-heavy functions.

Completing a game of Monty Hall requires long-term storage of many game parameters which overshadows the base fee costs (each stored word costs 20,000 gas). As each head stores the data independently, the scaling is close to (but still below) linear in this case. The tail call optimization still results in savings at the end of the game, when the winnings are sent to the house or player.

Evaluation of gas costs (and anonymity set sizes) for Submarine Commitments are in Appendix B. These costs only affect the transaction that claims the bounty.

Observations on the development process. After writing three heads independently, we commonly tested our contracts for discrepancies and found multiple bugs in each head, none of which impacted all heads simultaneously. Examples include a misunderstanding of the ERC20 API, integer overflows, “off-by-one” errors in the Monty Hall game, and a vulnerability to an only recently discovered EVM anti-pattern that lets a contract silently increase another contract’s ether balance via the SUICIDE opcode. Notably, all these bugs could have been exploited against a single contract, yet none of them appear useful against all heads simultaneously.

In addition to the exploit gap induced by Hydra, the NNVP development process itself increased the quality of our contracts. For the Monty Hall, ensuring compatibility between heads required writing a detailed specification, which revealed many blind spots in our original design. Moreover, differential testing [42] (verifying agreement between heads on random inputs) was remarkably simpler for exercising multiple code paths for the Monty Hall game, compared to a standard test suite.

8 Related Work

Software assurance and fault-tolerance are well-studied topics with an extensive literature. N-version programming [17, 6, 22] in particular was introduced decades ago and challenged in influential studies [21, 33] (see Section 2). Nagy et al. use N-version programming to construct honey-pots for detecting web exploits [45].

Bitcoin and, more importantly, Ethereum [14] have popularized smart contracts [52] and script-enhanced cryptocurrency [30]. Research on smart contract security is burgeoning and includes: Analysis of common contract bugs [19, 38, 5], static analysis and enhancements for Solidity [38], formal verification tools [9, 28, 3], design of “escape hatches” [41], DoS defenses for miners [39], trusted data feeds [58], formal EVM semantics [27, 29], and automated exploitation tools [36]. While promising, none of these tools and techniques have yet seen mainstream adoption, nor do they relate directly to our explorations in this paper.

In a closely related work, Tramè et al. [53] consider using smart contracts for bug bounties (using SGX), but not the converse, i.e., bounties for smart contracts.

Bug withholding is related to selfish-mining [25], where a miner withholding blocks to later nullify other
miners’ work. As selfish mining operates at the block level and bug withholding at the application level, they differ in their mechanisms, analysis, and implications.

Submarine Commitments hide bounty claims among normal Ethereum transactions and relate to cover traffic techniques such as anonymity networks (e.g., Tor [20]), network-based covert channels [44], steganography and watermarking [32]. Submarine Commitments differ in that they assume ultimate opening of a hidden value.

Several works [35, 31, 53] model blockchain-level adversaries. They consider an adversary that can reorder transactions within a given block, however, and not the much stronger model of chain-rewriting we explore here.

9 Conclusion

We have presented the Hydra Framework, the first principled approach to administering bug bounties that incentivize honest disclosure. The framework relies on a novel notion of an exploit gap, a program transformation that enables bug detection at runtime. We have described one such strategy, N-of-N-version programming (NNVP), a variant of N-version programming that detects divergences between multiple program instances.

We have applied our framework to smart contracts, highly valuable and vulnerable programs that are particularly well suited for fair and automated bug bounties. We have formally shown that Hydra contracts incentivize bug disclosure, for bounties orders of magnitude below an exploit’s value. We have modeled strong bug-withholding attacks against on-chain bounties, and analyzed Submarine Commitments, a generic defense to front-running that hides transactions in ordinary traffic.

Finally, we have designed and evaluated a Hydra Framework for Ethereum, and rigorously tested its soundness and applicability to the majority of Ethereum contracts today. We used this framework to construct a Hydra ERC20 token and Monty Hall game. The former is live in production on Ethereum, and represents the first principled and trust-free bug bounty offering.

Acknowledgements

We thank Paul Grubbs and Rahul Chatterjee for comments and feedback. This research was supported by NSF CNS-1330599, CNS-1514163, CNS-1564102, and CNS-1704615, ARL W911NF-16-1-0145, and IC3 Industry Partners. Philip Daian is supported by the National Science Foundation Graduate Research Fellowship DGE-1650441. Lorenz Breidenbach was supported by the ETH Studio New York scholarship.

References

[34] KNIGHT, J. C., AND LEVESON, N. G. A reply to the criticisms of the Knight & Leveson experiment. ACM SEN 15, 1 (1990), 24–35.
A Analysis of NNVP in the NASA Experiment

We briefly justify the results we obtained when applying our NNVP paradigm for the experimental results in [21]. The experiment consisted of 20 different program versions evaluated on six work-loads (corresponding to different initial system states). For \( y \in [0, 20] \), Eckhardt et al. report \( g(y) \), the empirical proportion of inputs in each of their test suites that induce a failure in exactly \( y \) out of 20 programs. They do not distinguish whether the failures are identical or not. Compared to our setting of Section 2, Eckhardt et al. further consider a distribution over programs. That is, the \( N \) programs to be aggregated are chosen at random from the pool of 20 programs.

Following the notation and analysis for majority-voting in [21], we define the empirical probability \( \hat{p}_{\text{maj}} \) that a majority of the \( N \) programs (randomly chosen from the 20) fail simultaneously (see [21, Equation 6]):

\[
\hat{p}_{\text{maj}} = \sum_{y=0}^{20} \binom{N}{y}^{-1} \sum_{l=0}^{N-y} \binom{N-l}{y} \binom{20-y}{N-1} g(y). \tag{5}
\]

Similarly, we define the empirical probability \( \hat{p}_{\text{NNVP}} \) that all \( N \) chosen programs fail simultaneously on a given input:

\[
\hat{p}_{\text{all}} = \sum_{y=0}^{20} \binom{N}{y}^{-1} \binom{y}{N} g(y). \tag{6}
\]

Finally, to recover our definition of an exploit gap in Equation (1), we define the probability \( \hat{p}_{\text{one}} \) that at least one of the programs fails:

\[
\hat{p}_{\text{one}} = \sum_{y=0}^{20} \binom{N}{y}^{-1} \sum_{l=1}^{N-y} \binom{y}{l} \binom{20-y}{N-l} g(y). \tag{7}
\]

We can then define two different exploit gaps, one for traditional N-version programming with majority voting, and one for NNVP (where we abort unless all programs fail identically). We have

\[
\text{gap}_{\text{maj}} = \frac{\hat{p}_{\text{one}}}{\hat{p}_{\text{maj}}} \quad \text{and} \quad \text{gap}_{\text{NNVP}} = \frac{\hat{p}_{\text{one}}}{\hat{p}_{\text{all}}}, \tag{8}
\]

where the inequality for \( \text{gap}_{\text{NNVP}} \) is because NNVP only fails if all programs fail identically (the results in [21] only give us an upper bound for this probability).

Using the estimated values \( g(y) \) from [21], we obtain:

<table>
<thead>
<tr>
<th>( N )</th>
<th>Majority Voting ( \hat{p}_{\text{maj}} )</th>
<th>NNVP ( \hat{p}_{\text{all}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5 ≤ gap_{maj} ≤ 189</td>
<td>74 ≤ gap_{NNVP} ≤ 14,845</td>
</tr>
<tr>
<td>3</td>
<td>13 ≤ gap_{maj} ≤ 3399</td>
<td>5544 ≤ gap_{NNVP} ≤ 801,741</td>
</tr>
</tbody>
</table>

In all cases, the lowest exploit gap is obtained for the third work-load (denoted \( S_{1,0} \) in [21]), which has the lowest failure rate overall.

If we combine all work-loads into one, and assume that hackers sample uniformly from the test inputs used in the experiment, we obtain:

<table>
<thead>
<tr>
<th>( N )</th>
<th>Majority Voting ( \hat{p}_{\text{maj}} )</th>
<th>NNVP ( \hat{p}_{\text{all}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>N.A.</td>
<td>gap_{NNVP} ≥ 79</td>
</tr>
<tr>
<td>3</td>
<td>gap_{maj} = 7</td>
<td>gap_{NNVP} ≥ 4409</td>
</tr>
<tr>
<td>5</td>
<td>gap_{maj} = 709</td>
<td>gap_{NNVP} ≥ 282,605</td>
</tr>
</tbody>
</table>

Note that NNVP makes sense even in the case \( N = 2 \), and yields gaps that are multiple orders of magnitude greater than the ones obtained with majority voting.

B Submarine Commitment Constructions

In this section, we present two constructions for Submarine Commitments. The first, in Appendix B.1, is our preferred construction. It is simple and efficient, but only realizable with changes to Ethereum awaiting adoption of EIP-86. The second, in Appendix B.3 is more involved and expensive, but realizable today.

We note that players could in principle conceal true commitments by sending dummy (regular) commitments with random values \( \text{val} \geq \text{deposit} \)—so that they are indistinguishable from real commitments—but have a “dummy” flag that can be revealed to trigger a refund. This approach turns out to be complicated and unworkable, although. A community of users would not in general have an incentive to generate dummy traffic and incur transaction fees. A would-be claimant could generate dummy traffic to conceal her true commitment, but then the very inception of dummy traffic would signal a pending claim and incentivize \( A \) to release its withheld bug. These problems motivate the use of Submarine Commitments instead.

B.1 EIP-86-Based Construction

Our simple realization of Submarine Commitments in Ethereum leverages a new EVM opcode, CREATE2, introduced in EIP-86 (EIP stands for “Ethereum Improvement Proposal”) and scheduled to be included in the upcoming “Constantinople” hardfork. CREATE2 creates new smart contracts, much like an already existing CREATE opcode. Unlike CREATE, which does not include a user-supplied value, CREATE2 computes the address of
the created contract \(C\) as \(H(\text{addrCreator}, \text{salt}, \text{code}C)\), where \(\text{addrCreator}\) is the address of the contract’s creator, \(\text{salt}\) is a 256-bit salt value chosen by the creator, \(\text{code}C\) is the EVM byte code of \(C\)’s initcode, and \(H\) is Keccak-256.

To realize a Submarine Commitment, we can use \(\text{salt}\) to encode the inputs to the commit, key and \(\text{addr}(P)\). Let \(\text{Forwarder}\) be a contract that sends any money received at its address to \(\text{BountyContract}\). A Submarine Commitment involves these functions:

- **Commit**: \(P\) selects a witness key \(\tau \leftarrow \{0,1\}^\ell\) for suitable \(\ell\) (e.g., \(\ell = 256\)). \(P\) sends \(\text{deposit}\) to address
  \[\text{addr} = H(\text{addr(\text{BountyContract})}, H(\text{addr}(P), \text{key}), \text{code}),\]
  where \(\text{addr(\text{BountyContract})}\) is \(\text{BountyContract}\)’s address and \(\text{code}\) is \(\text{Forwarder}\)’s EVM initcode.

- **Reveal**: \(P\) sends key and \(\text{commitBlk}\) (the block number in which \(P\) committed) to \(\text{BountyContract}\). \(\text{BountyContract}\) verifies that the contract indeed occurred in block \(\text{commitBlk}\) (e.g. using Appendix B.2).

- **DepositCollection**: \(\text{BountyContract}\) creates an instance of \(\text{Forwarder}\) at address \(\text{addr}\) using \(\text{CREATE}\). A call to \(\text{Forwarder}\) sends \(\text{deposit}\) to \(\text{BountyContract}\).

### B.2 Merkle-Patricia Proof Verification

In order for Submarine Commitments to be secure against front-running attacks, we need to verify that the commit transaction indeed occurred in block \(\text{commitBlk}\). Otherwise, an adversary can wait until she observes the “reveal” transaction \(\tau\) and then front-run \(\tau\) by including a backdated “commit” and corresponding “reveal” in front of \(\tau\). We can prevent this attack by having Contract verify that “commit” was indeed sent in block \(\text{commitBlk}\) and that at least \(\rho\) blocks have elapsed since \(\text{commitBlk}\) upon receiving a “reveal”. (Recall that the adversary can roll back the blockchain by at most \(\rho\) blocks.)

Unfortunately, Ethereum provides no native capability for smart contracts to verify that a transaction occurred in a specific block. However, Ethereum’s block structure enables efficient verification of Merkle-Patricia proofs of (non-)inclusion of a given transaction in a block [37]: all transactions in a block are organized in a Merkle-Patricia Tree [57] mapping transaction indices to transaction data. The root hash of this tree is included in the block header and the block header is hashed into the *block hash*, which can be queried from inside a smart contract by means of the \(\text{BLOCKHASH}\) opcode.

We implemented this verification procedure in a smart contract that takes a block number, the transaction data, and a Merkle-Patricia proof of transaction inclusion as inputs, and outputs \textit{accept} or \textit{reject}. We benchmarked the gas cost of this contract by verifying the inclusion of 25 transactions from the Ethereum blockchain. The proof verification has a mean cost of 207,800 gas (approximately 2.08 USD\(^2\)). Note that this cost is only incurred when a bounty is being claimed, and has no impact on “normal” transactions.

**Proof of Cheat.** We can reduce the gas cost of our Submarine Commitment scheme by not performing a Merkle-Patricia proof verification on every “reveal”; instead of requiring parties to prove that their “commit” occurred in \(\text{commitBlk}\), we only require them to provide \(\text{commitBlk}\) and the transaction data, but \textit{no Merkle-Patricia proof}. A party \(P\) can then submit a *Proof of Cheat*, a Merkle-Patricia proof demonstrating that an adversary \(A\) backdated their transaction: \(A\) had to claim the existence of a non-existing transaction; therefore, there will either be a different transaction or no transaction at the purported transaction index in block \(\text{commitBlk}\). If the proof of cheat is accepted, \(A\)'s \(\text{deposit}\) is given to \(P\) and \(A\)'s “commit” and “reveal” are voided.

Competing parties can easily check each other’s commits for correctness off-chain and provide a Proof of Cheat if they witness a cheat. In this setting, \(P\) benefits from catching a malicious competitor \(A\) in two ways: \(A\)'s claim is voided (potentially netting \(P\) the \(\$\text{bounty}\) and \(A\)'s \(\text{deposit}\) is given to \(P\)).

### B.3 CREATE-Based Construction

In Appendix B.1, we gave a construction of Submarine Commitments that requires the yet-to-be-introduced \(\text{CREATE}\) opcode. Hereafter, we show a different construction relying on the \(\text{CREATE}\) opcode, available in Ethereum today. However, the \(\text{CREATE}\)-based construction is simpler and has 98.5% lower gas costs than the \(\text{CREATE}\)-based construction during deposit collection (75,000 gas vs 5,000,000 gas, or 0.75 USD vs 50.00 USD respectively\(^2\)).

When a contract \(C\) creates a new contract \(C_{\text{new}}\) using the \(\text{CREATE}\) opcode, \(C_{\text{new}}\’s\) address is computed as \(H(\text{addr}(C), \text{nonce}(C))\), where \(\text{nonce}(C)\) a monotonic counter of the number of contracts created by \(C\). (Ethereum’s state records this nonce for each contract.) By chaining a series of contract creations and encoding information in the associated nonce values, we can compute an address for Submarine Commitments. Let \(\text{Contract}\) be the contract that will receive \(\text{nonce}\) values, we can compute an address for Submarine Commitments. Let \(\text{Forwarder}\) be a simple contract that has two functions both of which abort if they aren’t being called by \(\text{Contract}\):

- **Clone** uses \(\text{CREATE}\) to spawn another \(\text{Forwarder}\) instance at address \(H(\text{addr}(\text{Forwarder}), \text{nonce}(\text{Forwarder}))\).
- **Forward** sends all funds held by the contract to \(\text{Contract}\).
We now describe the three functions that make up a Submarine Commitment:

- **Commit**: $P$ selects a witness key $\epsilon \leftarrow \{0, 1\}^k$ and computes $x := H'(\text{addr}(\text{Contract}), \text{key})$ for a suitable $\ell$ and hash function $H'$ with codomain $\{0, 1\}^\ell$. Let $A := \text{addr}(\text{Contract})$ and let $E : \{0, 1\}^\ell \rightarrow \{0, \ldots, b-1\}^k$ be the function that takes an integer (encoded as a binary string) and reencodes it as a string of length $k$ in base $b$. $P$ sends $\text{Sdeposit}$ to address $\text{addr} = H(H(\ldots H(A, E(x)_1 + 1), \ldots, E(x)_{k-1} + 1), E(x)_k + 1)$.

- **Reveal**: $P$ sends key and a Merkle-Patricia proof that she committed in the correct block (see Appendix B.2) to BountyContract.

- **DepositCollection**: BountyContract repeatedly calls the $\text{Clone}$ function of appropriate Forwarder instances until a Forwarder is created at $\text{addr}$. (See Figure 7 for details.) BountyContract then calls $\text{Forward}$ to make this instance send the the deposit to BountyContract.

Choosing $n$ and $b$. Since we aren’t concerned with collision attacks on $H'$, $n = 80$ provides sufficient security. For $n = 80$, in the ROM, a choice of $b = 4$ minimizes the expected number of contract creations $\log_b(2^n) (1 + \frac{1}{b-1})$. In practice, we instantiate $H'$ as a truncated version of Keccak-256 as this is the cheapest cryptographic hash function available in the EVM. In our prototype, a DepositCollection call costs 5,000,000 gas with these parameters.

### B.4 Analysis of Anonymity Set Size

Submarine Commitments rely on concealing “commit” transactions in an anonymity set of unrelated transactions: to prevent bug-withholding attacks, the “commit” of the Submarine Commitment scheme must remain concealed until the “reveal” is broadcast. Since a “commit” is indistinguishable from a benign transaction sending ether to a fresh address, a transaction to an address $A$ is a part of the anonymity set if:

- The (external) transaction is a regular send of a non-zero amount of ether with an empty data field.
- $A$ has never received or sent any transactions.
- $A$ has no associated code (i.e. $A$ is not a contract).
- $A$ is not involved in any other transactions (internal or external) during the commit window.

In the experiment $\text{Exp}_A$ analyzed in Section 5.4, a commitment is revealed after $\rho$ blocks, where it is assumed that the adversary can rewind up to $\rho$ blocks in the blockchain. Figure 8 shows the size of the anonymity set as a function of this commitment window $\rho$. Even for $\rho = 100$ (i.e. a 25 minute rewind window at 15 seconds/block), average blocks still contain two transactions in the anonymity set. Furthermore, 34 of the 48 blocks we studied (70%) contained at least one transaction that is part of the anonymity set. In a full commitment window of size $\rho = 100$, we get an anonymity set of approximately 200 transactions, over 2% of all transactions in period.

As Figure 9 shows, the transaction values in the anonymity set span a wide range. Commitments with an associated value between 0.0001 ether and 10 ether (approximately 10,000 USD\textsuperscript{2}) are easily concealed.