An Empirical Study of Blockchain System Vulnerabilities: Modules, Types, and Patterns

Xiao Yi
The Chinese University of Hong Kong
Hong Kong SAR, China
yx019@ie.cuhk.edu.hk

Daoyuan Wu*
The Chinese University of Hong Kong
Hong Kong SAR, China
dywu@ie.cuhk.edu.hk

Lingxiao Jiang
Singapore Management University
Singapore, Singapore
lxjiang@smu.edu.sg

Yuzhou Fang
The Chinese University of Hong Kong
Hong Kong SAR, China
yzfang@ie.cuhk.edu.hk

Kehuan Zhang
The Chinese University of Hong Kong
Hong Kong SAR, China
khzhang@ie.cuhk.edu.hk

Wei Zhang
Nanjing University of Posts and
Telecommunications
Nanjing, China
zhangw@njupt.edu.cn

ABSTRACT

Blockchain, as a distributed ledger technology, becomes increasingly popular, especially for enabling valuable cryptocurrencies and smart contracts. However, the blockchain software systems inevitably have many bugs. Although bugs in smart contracts have been extensively investigated, security bugs of the underlying blockchain systems are much less explored. In this paper, we conduct an empirical study on blockchain's system vulnerabilities from four representative blockchains, Bitcoin, Ethereum, Monero, and Stellar. Specifically, we first design a systematic filtering process to effectively identify 1,037 vulnerabilities and their 2,317 patches from 34,245 issues/PRs (pull requests) and 85,164 commits on GitHub. We thus build the first blockchain vulnerability dataset, which is available at https://github.com/VPRLab/BlkVulnDataset. We then perform unique analyses of this dataset at three levels, including (i) file-level vulnerable module categorization by identifying and correlating module paths across projects, (ii) text-level vulnerability type clustering by natural language processing and similarity-based sentence clustering, and (iii) code-level vulnerability pattern analysis by generating and clustering code change signatures that capture both syntactic and semantic information of patch code fragments.

Our analyses reveal three key findings: (i) some blockchain modules are more susceptible than the others; notably, each of the modules related to consensus, wallet, and networking has over 200 issues; (ii) about 70% of blockchain vulnerabilities are of traditional types, but we also identify four new types specific to blockchains; and (iii) we obtain 21 blockchain-specific vulnerability patterns that capture unique blockchain attributes and statuses, and demonstrate that they can be used to detect similar vulnerabilities in other popular blockchains, such as Dogecoin, Bitcoin SV, and Zcash.

*Corresponding author.

CCS CONCEPTS

• **Security and privacy → Software and application security**.

KEYWORDS

Blockchain Security, System Vulnerability, Data Mining

1 INTRODUCTION

While blockchain was first invented as a transaction ledger of the Bitcoin cryptocurrency [68], it is now serving as a fundamental component of many cryptocurrencies, the total market capitalization of which is close to two trillion USD in February 2022 [40]. Smart contract platforms (e.g., Ethereum [35] and Hyperledger Fabric [30]) and decentralized computing platforms (e.g., Interplanetary File System [34] and Blockstack [29]) further evolved the blockchain technology into various decentralized applications, such as DeFi (Decentralized Finance) [80], smart contract oracles [83, 84], decentralized identities [65], decentralized IoT management [73], and decentralized app markets [37]. To protect the decentralization of these systems and secure those finance-critical cryptocurrencies, security is a top priority of many blockchains.

Prior research on blockchain security focused on smart contract vulnerability detection and network analysis. Many static program analysis tools, e.g., Oyente [64], Zeus [51], Securify [76], Gigahorse [47], and ETHBMC [42], have been proposed to detect vulnerable smart contracts via symbolic execution and model checking. Dynamic tools [38, 49, 70, 75] and learning-based tools [45, 60, 63] were also invented. Besides smart contract analysis, some works analyzed network traffic hijacking [31] and mining [44] attacks and performed transaction attack analysis [39, 53, 85, 87]. In contrast, blockchains' system-level security issues are much less explored in academic research. To the best of our knowledge, there was only one study [77] in this direction. It specifically analyzed 946 blockchain bugs, with only 18 security bugs covered and four analyzed.

Xiao Yi, Daoyuan Wu, Lingxiao Jiang, Yuzhou Fang, Kehuan Zhang, and Wei Zhang

In this paper, we aim to *systematically understand blockchain system vulnerabilities* by conducting an empirical vulnerability study of the representative blockchains in four directions, including the classic Bitcoin [68], the smart contract platform Ethereum [35], the anonymous coin Monero [69], and the payment network Stellar [61]. They are not only popular in the cryptocurrency market but also backed up with solid technical papers.

As depicted in Figure 1, the first step and challenge of our study is to *effectively* collect vulnerable issues and their patches of those four blockchains. This is difficult because there is very little CVE information associated with blockchain projects (unlike other vulnerability mining studies [56, 81, 86]), and the large number (over 34K) of raw blockchain bugs in our crawled database makes manual vulnerability filtering[1] ineffective. To address this, we propose a vulnerability filtering framework based on the intuition that vulnerabilities have unique characteristics from various aspects, and we can gradually identify candidate vulnerabilities by analyzing attributes of the code commits, files, labels, and keywords. Eventually, we obtain 1,037 vulnerabilities and their 2,317 patches as our blockchain vulnerability dataset.

Based on this unique dataset, we study three key yet unexplored aspects of blockchain vulnerabilities, including susceptible blockchain modules, common blockchain vulnerability types, and blockchain-specific patch code patterns. To this end, we perform the file-, text-, and code-level vulnerability analysis as follows.

Firstly, we conduct the module analysis by inspecting patched files. However, inspecting each individual file is time-consuming because there are 2,362 unique patch file paths. Therefore, we propose to identify the *module path*, i.e., the folder name that could summarize the module of enclosed files (e.g., the "rpc/" folder indicates the RPC module). We further correlate module paths across different blockchains by identifying a reference blockchain architecture and mapping different module paths into this architecture. This module categorization allows us to obtain a layered map of blockchain vulnerabilities in different modules and pinpoint susceptible blockchain modules. We find that some modules are more susceptible than the others, such as the highly susceptible ones related to consensus, wallet, and networking, each with over 200 vulnerabilities.

Secondly, we perform the type analysis by analyzing vulnerability text, more specifically, vulnerability titles. This is because a vulnerability type is typically captured by the title of an issue/PR (pull request), e.g., Bitcoin PR #17640 "wallet: Fix uninitialized read in bumpfee(...)," where "uninitialized read" is the type. To eliminate noisy words and generate good-quality clusters about types, we leverage the part-of-speech analysis of NLP (natural language processing) to first extract *type keywords* before we conduct actual clustering. By extracting type keywords in various situations and identifying a suitable clustering algorithm (and its setting), we successfully map 75.8% of the vulnerabilities into the clusters of different types and analyze the top 20 types that affect at least ten vulnerabilities each. Among these types, we identify four new vulnerability types that are directly related to blockchain transaction, block, peer/node, and wallet key/password. We also show that traditional vulnerability types still hold 62%~78% of all the blockchain

vulnerabilities. Furthermore, we analyze the type differences across different blockchain projects.

Thirdly, we conduct the pattern analysis by analyzing vulnerability patch code. In particular, we focus on blockchain-specific vulnerability types since the code patterns of traditional vulnerability types are well-known. To facilitate similar patch code into the same cluster, we design and generate the *code change signatures* that concisely capture both syntactic and semantic information of patch code fragments. By clustering 3,251 code fragments into 174 clusters of code change signatures, we identify 21 blockchain-specific vulnerability patterns that check unique blockchain attributes (e.g., the sender address, transaction order, block header, and gas limit) and validate various blockchain statuses during node synchronization, peer validation, wallet, and database operations. We further leverage these patterns to discover 20 similar vulnerabilities in other popular blockchains, notably, Dogecoin, Bitcoin SV, and Zcash, which have a collective market capitalization of over 25 billion USD as of January 2022. Most of our vulnerability reports have been confirmed and are under patching, with only two being rejected. This demonstrates the real-world impact of our vulnerability patterns. A thorough detection of blockchain system vulnerabilities based on the patterns extracted in this paper will be our future work.

To sum up, the main contributions of this paper are as follows:

- We design a systematic filtering process to curate a unique vulnerability dataset and will release it to the research community. The link of the dataset is already available at https://github.com/VPRLab/BlkVulnDataset.
- We develop a set of new methods to analyze blockchain vulnerabilities, build a knowledge base on previously unknown patterns of the vulnerabilities and their fixes.
- We reveal three key findings about blockchain system vulnerabilities in terms of their susceptible modules, various vulnerability types, and specific vulnerability patterns. Moreover, we demonstrate the usage of these vulnerability patterns by detecting 20 similar vulnerabilities in other popular blockchains.

The rest of this paper is organized as follows. We first provide the background of studied blockchains and their bug-fixing process in §2 and describe our systematic data collection in §3. We then present our multi-level vulnerability analysis in §4, §5, and §6, respectively. §7 summarizes the related works. Finally, §8 concludes this study.

2 BACKGROUND

2.1 Four Representative Blockchains Studied

In this paper, we study the representative blockchains that are (i) popular in the cryptocurrency market, (ii) in different directions of blockchain usages, and (iii) backed up with solid technical papers. Under these three conditions, we select the classic Bitcoin [68], the smart contract platform Ethereum [35], the anonymous coin Monero [69], and the payment network Stellar [61]. Next, we present their basic information and the development status on GitHub.

Bitcoin introduces the concept of *blockchain* [68] and uses it as a distributed ledger to record transactions for public verification. As of January 2022, the Bitcoin cryptocurrency (or BTC) has the top one market capitalization of more than 832 billion USD. The Bitcoin software was released in 2009, and it is actively maintained by over

---

[1]That said, we need to recognize or differentiate *real* vulnerabilities from *regular* bugs.
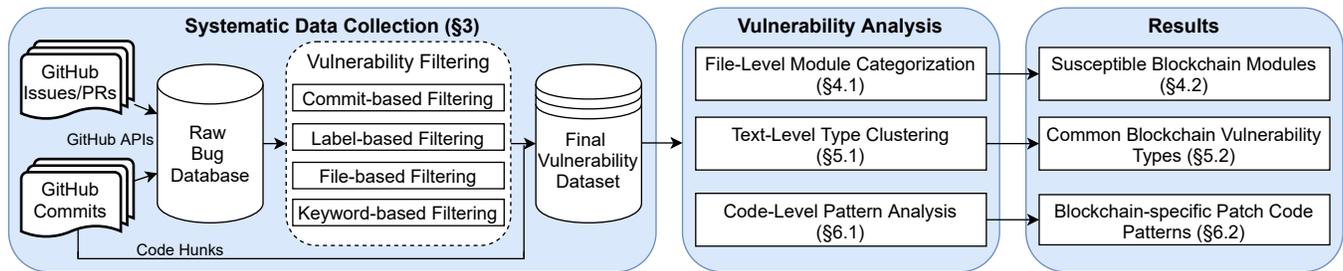
Figure 1: The overall workflow of our blockchain vulnerability study.

850 contributors on GitHub in a repository called **bitcoin/bitcoin**. The primary programming language of Bitcoin is C++.

Ethereum is the first blockchain system with the capability of constructing Turing-complete *smart contracts* [35], which contain a set of pre-defined rules and regulations for self-execution. To maintain the operation of Ethereum, it creates a native cryptocurrency called Ether (or ETH), which is the second largest cryptocurrency with a market capitalization of more than 410 billion USD as of January 2022. The Ethereum software was released on GitHub in 2015, and its Go implementation is maintained by over 700 contributors in a repository called **ethereum/go-ethereum**.

Monero aims to mitigate the privacy leakage in blockchain systems, since each blockchain transaction is transparent and could leak some sensitive information. To do so, Monero uses an obfuscated ledger [69] to prevent the transaction details (e.g., transaction source, amount, and destination) from being revealed to outside observers. As of January 2022, the Monero coin (XMR) is ranked 47th with a market capitalization of over 3.8 billion USD. The Monero software was released on GitHub in 2014, and it is maintained by over 250 contributors in a repository called **monero-project/monero**. The primary language of Monero is C++.

Stellar is a blockchain-based *payment network* [61] that can perform cross-border money transfer in seconds. It uses a novel consensus protocol called *Stellar Consensus Protocol* (SCP) [61] for fast and secure transactions among untrusted participants. The native cryptocurrency of Stellar is called XLM, which is ranked 30th with a market capitalization of around 6.7 billion USD as of January 2022. The Stellar software was released on GitHub in 2015, and it is currently maintained by more than 80 contributors in a repository called **stellar/stellar-core**. Similar to Bitcoin and Monero, the primary language of Stellar is also C++.

2.2 Bug-fixing Process in Blockchain Projects

It is also necessary to understand the typical bug-fixing process of blockchain projects hosted as open-source projects on GitHub in order to collect and analyze their vulnerabilities and patches. A **commit** is a set of changes submitted by developers into a project repository; a commit can change anything, ranging from changing source code to modifying document files or merging multiple previous commits. A change consisting of a consecutive sequence of added/deleted lines is also known as a **hunk**. A **patch** is a collection of changes or commits that can be applied to a set of files via a patching tool. An **issue** is often a report on a project's GitHub page; it may describe a potential bug or sometimes an enhancement or a

question, and may come with fixes and solutions. A **pull request (PR)** is the proposed commit for a project from a separate clone of the project; it can be pulled from the project clone and accepted into the original project based on the review of managing developers. For simplicity, we do not explicitly distinguish an issue and a PR in this paper since the latter often contains a bug description too. Indeed, GitHub itself mixes up the usage of issue/PR numbers.

3 SYSTEMATIC DATA COLLECTION

As shown in Figure 1, the first and a critical step of our study is to collect a good-quality blockchain vulnerability dataset across multiple blockchain systems that satisfies two conditions: (i) cover as *many* vulnerabilities as possible in the studied blockchains (i.e., minimizing false negatives); and (ii) introduce as *few* non-vulnerability bugs as possible in the dataset (i.e., minimizing false positives).

Some other vulnerability studies [56, 81, 86] leverage the CVE (Common Vulnerabilities and Exposures) or Bulletin (i.e., bug bounty) information to collect vulnerability data. However, we found that there is very little CVE/Bulletin information about most blockchains because blockchain vulnerabilities are critical and often patched directly via the reports from bug bounty programs without releasing a CVE. For example, Ethereum (go-ethereum) had only four CVEs released before our data collection while Bitcoin had 33 CVEs.

We take a different way—directly analyze the blockchain projects' issues and commits in their GitHub repositories and extract the vulnerable ones from them. We first crawl all blockchain bugs and organize them into a raw bug database (in §3.1). The major challenge is how to recognize or differentiate *real* vulnerabilities[2] from a large number of *regular* bugs. To address the challenge, we propose a novel vulnerability filtering framework (in §3.2) that systematically and effectively filters out regular bugs and extracts blockchain vulnerabilities. We eventually obtain the first dataset of blockchain system vulnerabilities (in §3.3), comprising more than 1K vulnerabilities identified from over 34K issues. It could not be done via manual analysis or via prior training-based patch identification [74, 88] since (i) there is no ground-truth training set for blockchain vulnerabilities and (ii) the learning-based nature of those techniques tends to identify only the similar bugs or vulnerabilities.

---

[2]In this paper, we adopt a broad definition of vulnerabilities that considers *the bugs with security impact* as vulnerabilities.

3.1 Crawling and Organizing Blockchain Bugs

As illustrated in Figure 1, our blockchain bug database is constructed from two data sources, the *issues* and *commits*, by leveraging GitHub APIs[3]. For the issues, we collect all the information of each closed issue/PR, including the issue title, issue body, comments, events, and bug category labels. We consider only *closed* issues/PRs because *open* issues are not confirmed bugs yet and certainly have no patches. Note that even for closed issues, they may not be the real bugs and could have no patches (i.e., they were simply closed by developers). For the commits, we first crawl *all* the commits of a repository and then determine which commits are bug-related. For each commit, we collect its title, commit message, affected files, and id/URL; for some commits filtered according to §3.2, their actual code change hunks are also collected and processed according to §3.3 and used for code-level pattern analysis in §6. We have collected a total of 34,245 *closed* issues/PRs and 85,164 commits as the raw dataset at the end of February 2020. The detailed breakdown of these issues/PRs and commits across four blockchain projects is available in Table 2.

With the raw data collected, a non-trivial task is to organize and correlate the issues with their corresponding commits. Specifically, we need to determine all the relevant commits for a given issue/PR — if an issue/PR has no patch commits, it is not a real bug and will be filtered out. By summarizing the issue/PR and commit's GitHub structures, we observe three kinds of information we can leverage for such correlation. First, we leverage the issue page's event information (e.g., XXX mentioned this issue and YYY added a commit) and retrieve the commit URLs from those events. For example, in https://github.com/bitcoin/bitcoin/issues/595, we obtain the commit URL via the event of "laanwj added a commit that referenced this issue." Second, for a PR like https://github.com/bitcoin/bitcoin/pull/9366, we can directly retrieve its commit lists at its "Commits" tab page. Although these two kinds of information is useful for most issues/PRs, some commits may not appear in the events of issues or commit lists of PRs. To overcome this, our script analyzes all the 85,164 commits' titles and messages and identifies issue/PR numbers from them. With these strategies, we successfully build the relationship between the issues and commits and finish constructing the raw bug database shown in Figure 1.

3.2 A Vulnerability Filtering Framework

To evolve the raw bug database into the final vulnerability dataset, we design a systematic vulnerability filtering framework expressed as a seven-step process (i.e., S0~S4b in Table 1) to effectively differentiate vulnerabilities from regular bugs with minimal manual work. The intuition is that vulnerabilities have unique characteristics at various aspects, and we can gradually identify candidate vulnerabilities by analyzing attributes of the code commits, files, labels, and keywords. As shown in Table 1, we perform the filtering at the following four aspects:

Commit-based filtering. Firstly, in the step S0, we leverage the most straightforward characteristic that *a closed vulnerability must associate with code commits*. In other words, an issue/PR without any commit could be excluded directly. Since we have already built

---

[3]https://docs.github.com/en/rest/reference/commits and https://docs.github.com/en/rest/reference/issues

Table 1: Intermediate results of the filtering in each step.

Action	Commit	File		Label		Keyword	
	S0	S1	S2	S3a	S3b	S4a	S4b
Include/ Exclude	–10,101	–3,798	–1,522	56	–4,400	1,227	–6,330
Remain	24,144	20,346	18,824	18,768	14,368	13,141	6,811

the relationship between issues/PRs and commits in §3.1, we easily exclude 10,101 issues/PRs out of the entire 34,245 issues/PRs.

File-based filtering. Secondly, we leverage two characteristics of patch files to filter out the bugs that are certainly not vulnerabilities. The basic idea of these two characteristics is that *the patch of a vulnerable issue/PR must make some real code changes*, including changing files with actual source code and not containing only test code. Specifically, in the step S1, we determine the file types with actual source code (by their file suffixes) for four blockchains. An issue/PR whose commits do not modify any file in these types should be excluded. For example, there are 152 different file types for Bitcoin's commits, but only these seven file types, ['.cpp', '.h', '.py', '.sh', '.cc', '.c', '.java'], contain actual source code whereas other file types like '.yml' and '.mk' are unlikely related to vulnerabilities. This step filters out 3,798 more issues/PRs, then the remaining 20,346 are further filtered by the step S2. Specifically, S2 excludes the test-only commits and their associated issues/PRs. With the file-based filtering, we exclude 22% (5,322/24,144) of the issues/PRs.

Label-based filtering. Thirdly, we leverage the characteristic of the labels of issues/PRs: *certain words in the labels could indicate whether an issue/PR is related to a vulnerability or not*. For example, the 'Privacy' label marks privacy-related bugs in the Bitcoin project and the 'obsolete:vuln' label indicates the early-stage vulnerabilities of Ethereum. To avoid false positives, we are conservative in specifying vulnerability labels — we assign only three labels (i.e., the 'Privacy', 'obsolete:vuln', and special label 'SEC-XXX' that appeared in the beginning of issue/PR titles) and mark their corresponding 56 issues/PRs explicitly as vulnerabilities in the step S3a. In contrast, there are much more labels clearly indicating non-vulnerability issues/PRs. Specifically, out of the entire 87 labels from four blockchain projects, we manually determine that 48 of them are *not* related to vulnerabilities, such as 'Refactoring', 'Docs', and 'type:feature'. With these labels, we filter out their associated 4,400 issues/PRs in the step S3b. After this step, we have narrowed the filtering scope from 34,245 to 14,368 issues/PRs, a reduction of 58%.

Keyword-based filtering. Lastly, we directly check issues/PRs' text based on the characteristic that *some keywords could indicate an issue/PR vulnerable whereas others could imply an issue/PR not related to vulnerabilities*. To this end, we first perform a word count analysis on the words in issue/PR titles and bodies, sort these words by their appearance frequency, and exclude the words that appear only once. We then group the words by their semantic similarity using the spaCy [25] NLP library. Since similar words are grouped together, we manually go through all the clusters to obtain a set of vulnerability-related words (Step S4a) or non-vulnerability words (Step S4b). Specifically, we obtain 62 clusters of vulnerability-related words and 79 clusters of non-vulnerability words, which allows us to automatically identify 1,227 vulnerable issues/PRs and exclude 6,330 irrelevant issues/PRs in the step S4a and S4b, respectively.

Table 2: Metadata of the raw and vulnerability datasets.

Repository	Raw Bug Database		Vulnerability Dataset	
	Closed Issues/PRs	Commits	Vulnerable Issues/PRs	Patch Commits
Bitcoin	16,731	41,706	442	942
Ethereum	9,321	23,764	365	826
Monero	5,918	12,656	178	286
Stellar	2,275	7,038	52	263
Total	**34,245**	**85,164**	**1,037**	**2,317**

Eventually, our filtering framework extracted 1,283 (=1,227+56) suspicious issues/PRs (in the step S3a and S4a) from the entire 34,245 issues/PRs. We have manually examined all these candidates and confirmed that 1,059 of them were actually vulnerability-related. This suggests that our filtering achieves a precision of 82.5% in identifying true vulnerabilities. It is also worth noting that our filtering framework may potentially have a high recall in identifying all patched vulnerabilities in the projects although there is no ground-truth for exact measurement, since it handles at least 80.1% (27,434/34,245) of all the issues/PRs; although the remaining 6,811 after step S4b are discarded, we believe that they have a low chance of being vulnerabilities due to no relevant keywords.

3.3 The Vulnerability Dataset and Its Metadata

We then retrieve the actual code hunks for the identified 1,059 issues/PRs from their corresponding 2,933 commits. This allows us to further exclude 22 issues/PRs because they associate with "invalid" code commits through the code hunk analysis. Specifically, we identified 586 duplicate code commits whose code hunks were the same (e.g., https://github.com/bitcoin/bitcoin/commit/d4781ac6 and https://github.com/bitcoin/bitcoin/commit/8a445c56), for which we kept just one code commit for each duplicate pair. We also found 30 empty code commits where we were not able to obtain their code hunks due to disappeared (e.g., https://github.com/bitcoin/bitcoin/commit/7e193ff6) or large diffs (e.g., https://github.com/ethereum/go-ethereum/commit/34dde3e2). As a result, our final vulnerability dataset consists of 1,037 vulnerability-related issues/PRs and their 2,317 commits, as shown in Table 2. It is worth noting that while items in our dataset are all security patches, some of them are not conventionally technical vulnerabilities but more like security enhancements, such as upgrading weak crypto algorithms to strong ones. In this paper, we do not distinguish them.

In Table 2, we also list the metadata of each blockchain project. We can see that Bitcoin and Ethereum contribute 77.8% of the vulnerabilities in our dataset, whereas the percentages of Monero and Stellar vulnerabilities are relatively low. This is mainly because Bitcoin and Ethereum have much more code commits than the other two blockchains, holding a similar percentage (76.9%) of the entire 85,164 commits. Additionally, we notice that Stellar has around the same number of patches as Monero, whereas the number of issues/PRs is three times lower (56 v.s. 178). The main reason is that Stellar developers tended to use one PR to cover multiple-bug fixes at the early stage of Stellar development.

Based on this unique dataset, we perform a comprehensive vulnerability analysis at three different levels in §4, §5, and §6.

4 FILE-LEVEL MODULE CATEGORIZATION

At the first-level of our study, we perform the module analysis of patched files. We first propose a lightweight method for categorizing vulnerable modules in §4.1, and then present the categorization result and its implication in §4.2.

4.1 Identifying and Correlating Module Paths for Vulnerable Module Categorization

We found that 1,037 vulnerable issues/PRs (or more precisely, 2,317 patch commits) totally generated 2,362 unique file paths (544 in Bitcoin, 1,376 in Ethereum, 251 in Monero, and 191 in Stellar), which makes inspecting each individual file time-consuming. Therefore, we propose to identify the *module path*, i.e., the folder name that could summarize the module of enclosed files (e.g., the "rpc/" folder indicates the RPC module). For some paths of generic names (e.g., the "src/" folder), we consider its sub-folders as module paths. Since Ethereum's folder structure is more complicated than the other three projects, we also consider three additional folders (the "core/", "swarm/", and "eth/" folders) as generic, and consider their sub-folders as module paths. Eventually, we obtain a total of 146 module paths (28 in Bitcoin, 71 in Ethereum, 26 in Monero, and 21 in Stellar) from 2,317 patch commits in the four studied blockchains.

Further, since different blockchains have different path names for the same module (e.g., the Consensus module of Bitcoin/Ethereum is in "consensus/" while that of Stellar is in "src/scp/"), we need to correlate those module paths *across projects*. Our solution is to identify a reference blockchain architecture and map different module paths into this architecture. Since many blockchains are based on Bitcoin, we use Bitcoin Core's architecture [22] as our reference. For easier understanding, we separate the entire architecture into four layers [18], as shown in Figure 2, and unify the traditional Miner, Mempool, and Validation Engine components into the Consensus module. We then manually map those 146 module paths into our blockchain architecture one by one.

It is worth noting that a vulnerable issue/PR may affect multiple modules, so the sum of the numbers of vulnerabilities of all the modules is larger than 1,037. Also, some patch commits change only the files directly under the generic "src/" folder and do not have module paths. We inspect all such patch files (107 in Bitcoin, 31 in Ethereum, 6 in Monero, and 4 in Stellar) and map their corresponding vulnerabilities into the modules in Figure 2 based on the patch file names.

4.2 Susceptible Blockchain Modules

Figure 2 shows the result of our module categorization in a layered map of blockchain modules and the numbers of vulnerabilities in those modules. We can see that modules in the Policy, Peer, Network layers each introduce around one-fourth of the vulnerabilities, while the UI modules and other uncategorized modules contribute the remaining 30%. Among all modules, we find that some modules are more susceptible than the others. Notably, the modules related to Consensus, Wallet, and NetConn contain over 200 issues each. Other modules about RPC, GUI/CMD, and Storage are also susceptible, affecting around 100 issues each. We observe that:

UI Layer		8.47%		21.74%
GUI/CMD: 141			Utils: 129	
Network Layer		24.92%		
RPC: 160	NetConn: 206	P2P: 49	External: 82	
Peer Layer		22.22%		
Storage: 93	Wallet: 214	Node: 63	EVM: 28	
Policy Layer		22.64%		
Crypto: 67	Consensus: 265	Policy: 45	Others: 123	
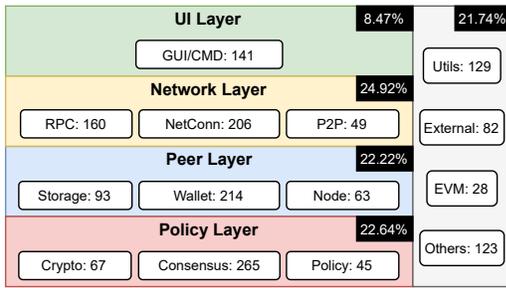
Figure 2: A layered map of blockchain vulnerabilities in different modules.

- The Consensus module covers the consensus (e.g., the Proof-of-Work mechanism [68]), miner, block/transaction related components. Unfortunately, it was affected by 265 vulnerabilities, with the major module path from the "consensus/" folder. Other module paths include "miner/", "ethchain/", "src/cryptonote_core/", "src/scp/", and "src/ledger/".
- In the Peer layer, the Wallet module handles transactions for each peer and the Storage module manages the storage of those transactions. As shown in Figure 2, the Wallet module was affected by 214 vulnerabilities, which are mainly from the "src/wallet/" and "accounts/" module paths. In contrast, the Storage was affected by 93 vulnerabilities, all of which are from database-related module paths, such as "src/blockchain_db/", "src/leveldb/", and "ethdb/".
- The NetConn and RPC modules collectively incurred the most blockchain vulnerabilities in our dataset. As a distributed system by nature, blockchain systems heavily rely on network synchronization and RPC (Remote Procedure Call). Since it deals with complex network communication of different peers, multiple security issues could occur, such as data race, deadlock, resource leak, and denial-of-service.
- Surprisingly, the GUI/CMD module is also a major source, with 141 vulnerabilities from the module paths like "src/qt/", "ethereal/ui/", "src/daemon/", and "cmd/". The underlying faults vary, but segfault and deadlock are typical bugs.

5 TEXT-LEVEL TYPE CLUSTERING

At the second-level of our study, we conduct the type analysis by analyzing vulnerability text. In this section, we first present a NLP-based approach for clustering vulnerability types in §5.1, and then summarize the clustering results and showcase common blockchain vulnerability types in §5.2, including the ones not known before.

5.1 NLP-based Analysis of Vulnerability Titles for Type Clustering

We find that a vulnerability type is typically captured by the title of an issue/PR page, e.g., Bitcoin PR #17640 "wallet: Fix uninitialized read in bumpfee(...)," where "uninitialized read" is the type. However, simply clustering issue/PR titles does not generate good-quality clusters about vulnerability types because each title could have some noises. For instance, in the earlier example, "wallet" and

"bumpfee" would affect the clustering quality. To address this problem, we propose a novel NLP-based method to first extract *type keywords* before we conduct actual clustering. This method is based on a grammatical pattern of vulnerability titles we observed, that a type is often a noun phrase located in between a verb (e.g., "fix") and a preposition (e.g., "in"). Figure 3 shows an intuitive illustration. Overall, our approach consists of two major steps: NLP-based keyword extraction and clustering the obtained type keywords. Before these two steps, we also need to perform some pre-processing.

Pre-processing. To this end, we remove useless words and formalize remaining words in the vulnerability titles. Specifically, the useless words include (i) the module/version information (e.g., the word before ":", such as the "wallet" above, or the word inside "[]", such as "[rpc]" or "[RELEASE]"), (ii) the special word (e.g., "SEC-*" for Ethereum and one-character word like "a"; note that numbers and symbols like "–" or "(...)" could be automatically handled by tokenizing), and (iii) noun-like adjective words (e.g., "possibility of" and "use of"). After cleaning useless words, we further formalize the remaining words by setting them to the lower case and tokenizing them via the NLP nltk [24] library's RegexpTokenizer. During this process, we also unify a few words (e.g., replacing all "tx"/"txs"/"txns" using "transaction"). In Table 3, we list several example titles our script automatically cleaned.

NLP-based keyword extraction. According to the grammatical pattern shown in Figure 3, our objective is to find the *target* verb and preposition that could determine the range of type words. However, one vulnerability title may contain multiple verbs or prepositions. Moreover, some verbs mainly act as nouns in our context, such as "check" and "leak". Based on these two reasons, we do not directly use the nltk [24] library's pos_tag() for a real-time part-of-speech analysis. Instead, we perform a *pre*-analysis of words' parts of speech in our cleaned vulnerability titles and build a vocabulary of verbs and prepositions and count their frequencies in our dataset. Eventually, we obtain a list of 33 verbs and 21 prepositions and rank them by frequencies. Table 4 shows the top 10 frequently used verbs and prepositions in our dataset.

Based on our vocabulary of verbs and prepositions and their frequencies, we are able to automatically locate the target verb and preposition for a *cleaned* vulnerability title in various situations using the following rules:

- If only one verb and one preposition exist and the preposition appears after the verb (with one or more words in between), such a verb and preposition, e.g., the word fix and in of the example E1 in Table 3, are the target words.
- If there is no verb but the preposition exists (e.g., the example E2) or there is no preposition but the verb exists (e.g., the example E3), the preposition or the verb will be determined as the target.
- If multiple verbs appear in a title, the one with the *highest* frequency will be regarded as the target verb. For example, in Figure 3 (or the example E4), the word fix has higher frequency than the word read in our vocabulary, fix is used as the target verb.
- If multiple prepositions appear in a title, the *first* one appearing after the target verb (with one or more words in between) is determined as the target preposition. For instance, in the example E5 in Table 3, both words on and in are prepositions, but since the word on appears before in, on is then determined as the target preposition.

Table 3: Examples of the cleaned issue/PR titles and their corresponding type keywords extracted.

ID	Raw Title	Cleaned Title	Type Keywords
E1	accounts: fix two races in the account manager	['fix', 'two', 'races', 'in', 'the', 'account', 'manager']	['two', 'races']
E2	blockchain_db: sanity check on tx/hash vector sizes	['sanity', 'check', 'on', 'transaction', 'hash', 'vector', 'sizes']	['sanity', 'check']
E3	[net] Avoid possibility of NULL pointer dereference	['avoid', 'null', 'pointer', 'dereference']	['null', 'pointer', 'dereference']
E4	wallet: Fix uninitialized read in bumpfee(…)	['fix', 'uninitialized', 'read', 'in', 'bumpfee']	['uninitialized', 'read']
E5	Prevent DOS attacks on in-flight data structures	['prevent', 'dos', 'attacks', 'on', 'in', 'flight', 'data', 'structures']	['dos', 'attacks']
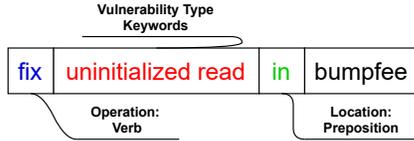
Figure 3: An example issue/PR title to illustrate the grammatical pattern of vulnerability titles we observed.

Table 4: The top 10 frequently used verbs and prepositions.

Verb	add	remove	fix	make	fixed
	set	avoid	improve	handling	added
Preposition	in	for	on	of	with
	from	by	before	if	after

- If none of above applies for a vulnerability title, we conclude that it has no target word.

After recognizing the target verb and preposition for each vulnerability title, the keywords in between the two target words are extracted as the type for the vulnerability. However, as we list above, some cleaned titles may end up with only one target word or even no any target word. We handle those special titles as follows:
- If only the target verb exists, all words after the target verb will be regarded as the type keywords.
- If only the target preposition exists, all words before the target preposition will be treated as the type keywords.
- If no target word exists, the entire cleaned title becomes the type keywords.

Clustering type keywords. With the extracted type keywords, we aim to cluster them based on their semantic meaning rather than their appearance as a string of letters. Thus, after embedding all the keywords into the vector space using word2vec [66], we choose the Word Mover's Distance (WMD) [55] as the similarity metric. Another reason for applying WMD is that it performs well on short sentences like our type keywords. Then, we calculate their pairwise similarity with WMD and generate a large similarity matrix.

The last step is to cluster the type keywords based on the similarity matrix. To reach an optimal clustering result, we tested four clustering algorithms: K-means [32], Gaussian Mixture [27], Agglomerative Clustering [26], and Affinity Propagation (AP) [43]. The first three algorithms require a pre-defined number of clusters as the key parameter, while AP needs a damping factor. For the first three algorithms, we tried a wide range of cluster numbers from 25 to 225 with an interval of 2. For AP, we tried the damping factor from 0.5 to 1 with an interval of 0.01. We kept other parameters

unchanged as default. After clustering with the given parameters, we computed the Silhouette Coefficient score [71] to determine the performance of the corresponding combination. As a result, Agglomerative clustering with 125 clusters was the best setting for our similarity matrix, which reached a coefficient score of 0.66.

5.2 Common Blockchain Vulnerability Types

According to Table 5, we obtain not only the traditional vulnerabilities, such as race condition and sanity check, but also blockchain-specific vulnerabilities. Among the top 20 vulnerability types, we find that seven of them are related to blockchains' characteristics. In particular, the 130 (22.1%) vulnerabilities from four types (T4, T7, T9, and T12) are blockchain-specific, which are related to blockchains' transaction, block, peer/node, and wallet key/password. Additionally, we have three more vulnerability types, T2, T14, and T20, that have some portions of their vulnerabilities related to blockchains' features. The rest of 366 (62.4%) vulnerabilities are solely the traditional vulnerabilities, not specific to blockchains.

Next, we explain three categories of these blockchain types: *specific*, *partially specific*, and *traditional*. For the patterns of blockchain-specific vulnerabilities, we will present them in §6.2.

Blockchain-specific vulnerability types. Since transactions, blocks, gas fees are the unique characteristics of blockchain systems, the type T4 and T7 record a large number of such new vulnerabilities. Examples are Bitcoin PR #8312 "Fix mempool DoS vulnerability from malleated transactions" and Ethereum PR #1354 "gpo non-existent block checks". Moreover, as a peer-to-peer software by nature, blockchains could suffer from peer/node vulnerabilities. By inspecting 28 such vulnerabilities in the type T9, we find that they are mainly related to the unique P2P features in blockchains, such as header sync and block validation. Examples include Bitcoin PR #10345 "timeout for headers sync" and Ethereum issue #604 "SEC-41 Peer TD in NewBlockMsg not verified". Lastly, blockchain systems often provide wallets to end users, which cause the new vulnerabilities related to wallet keys and passwords in the type T12. For example, Bitcoin PR #10308 describes the vulnerability patch of "[wallet] securely erase potentially sensitive keys/values".

Partially blockchain-specific vulnerability types. We also observe three vulnerability types partially specific to blockchains, i.e., T2, T14, and T20. Specifically, 64 vulnerabilities in the type T2 performed various checks, e.g., error and length checks, and some of them checked blockchain-related properties. For example, Bitcoin issue #1167 "check for duplicate transactions earlier" for DoS prevention, and Ethereum PR #20546 "check propagated block malformation on receipt". In contrast, the type T14 and T20 fixed more traditional vulnerabilities related to RPC calls and database corruption (due to exceptional closing), with a few vulnerabilities directly related to blockchains. Examples of blockchain-related

Xiao Yi, Daoyuan Wu, Lingxiao Jiang, Yuzhou Fang, Kehuan Zhang, and Wei Zhang

Table 5: The top 20 blockchain vulnerability types that affect at least ten vulnerabilities in our dataset.

ID	Type	# Vulnerability Issues/PRs					Specific?*
		All	B°	E°	M°	S°	
T1	Race Condition	77	14	48	10	5	–
T2	Check/Validation	64	36	14	10	4	◑
T3	Resource Leak	47	24	12	9	2	–
T4	Transaction Related	43	24	9	6	4	✔
T5	Deadlock	36	16	13	6	1	–
T6	Go Panic	36	0	36	0	0	–
T7	Block Related	34	9	21	4	0	✔
T8	Denial-of-Service	31	17	11	3	0	–
T9	Peer/Node Related	28	12	11	3	2	✔
T10	Sanity Check	28	11	3	13	1	–
T11	Overflow	27	11	8	6	2	–
T12	Wallet Key/Password	25	12	6	7	0	✔
T13	Uninitialized Read	19	14	0	5	0	–
T14	RPC Related	16	9	5	2	0	◑
T15	Out-of-Bound	14	9	4	1	0	–
T16	Off-by-One	14	5	2	7	0	–
T17	Segfault	13	13	0	0	0	–
T18	Memory Pool	12	10	1	1	0	–
T19	Nil Pointer Deref	12	6	5	1	0	–
T20	Database Corruption	11	4	3	4	0	◑
Sum	–	**587**	**256**	**212**	**98**	**21**	–
(%)		56.6	24.7	20.4	9.5	2.0	

*: ✔ means *most* in this type are blockchain-specific and ◑ means *some* are specific.
°: B, E, M, and S represent Bitcoin, Ethereum, Monero, and Stellar, respectively.

vulnerabilities are Ethereum PR #19401 "implement cli-configurable global gas cap for RPC calls" and Monero issue#706 "DB corruption" due to unfinished blockchain tasks.

Traditional vulnerability types in blockchains. Besides blockchain-specific vulnerabilities, Table 5 also shows that 366 vulnerabilities are solely from the 13 traditional vulnerability types. The top types, such as race condition, deadlock, and denial-of-service, are more frequent probably because it is difficult for blockchain systems to avoid them due to the sync among distributed nodes.

Further analysis. According to the detailed distribution of vulnerability types across different blockchain projects in Table 5, we make three observations. First, Ethereum has more than half of the T1 (Race) vulnerabilities, much higher than the other three. After investigating all the race-related vulnerability issues/PRs, we identify that the Swarm [28] subsystem is the major cause. Specifically, Swarm is only available in Ethereum and used for distributed storage and content distribution. Second, we notice that T6 (Go Panic) appears only in Ethereum because only Ethereum is implemented in Go. Moreover, since Go is a memory-safe language, Ethereum has fewer memory-related (T13, T18) vulnerability issues/PRs than Bitcoin. Third, we find that Monero has the most number of T10 (Sanity Check) and T16 (Off-by-One) vulnerabilities, while Stellar has the least number of vulnerability types since it is relatively new.

6 CODE-LEVEL PATTERN ANALYSIS

At the third-level of our study, we perform the pattern analysis by analyzing vulnerability patch code. In particular, we focus on blockchain-specific vulnerability types (i.e., the seven types mentioned in §5.2) since the code patterns of traditional vulnerability types like race condition, deadlock, overflow, and uninitialized read

are well-known (e.g., [36, 62, 79, 82]). In this section, we first propose our approach to summarizing patch code patterns in §6.1, and then present blockchain-specific code patterns in §6.2.

6.1 Generating and Clustering Code Change Signatures for Vulnerability Patterns

To obtain vulnerability code-level patterns, our objective is to put similar patch code *changes* into the same cluster so that analysts can summarize patterns from each cluster. To this end, we need an effective representation of code changes so that it keeps important semantic information yet ignores unimportant or noisy information. We call this representation the *code change signature*. Table 6 illustrates the evolution process from raw code hunks to their code fragments (i.e., contiguous lines of code) and the corresponding code change signatures using three examples. Taking the code in Table 6b and 6c as an example, both patches check whether the sender of a transaction is valid. However, if the variable name senderAddr is different, the similarity between their raw code fragment change (i.e., the syntactic changes indicated by F2 and F3) would be low. To capture the *essential* changes in patch code, we do not use the syntactic changes but their code change signatures like S2 and S3, the details of which will be illustrated during their generation.

Next, we introduce our approach to generating code change signatures and clustering them. Before these two major steps, we first clean up code hunks and turn them into fragments, and then align up the changed lines of code in each fragment.

Cleaning and splitting each code hunk into fragments. The raw code hunks we retrieved contain not only *meaningful* diff code but also test code, neighboring context (e.g., in-line and block comments, unchanged code lines, #include and import statements), and modification of none-code files (e.g., mark-down, JSON, and text files). Therefore, we first initiate a cleaning process [81] to keep only the actual diff code hunks and separate them into individual fragments by continuous '+' and '-' lines. Taking the code hunk in Table 6a as an example, it is separated into four code fragments, F1-1 (line 2-5), F1-2 (line 8-9), F1-3 (line 15-16), and F1-4 (line 19-21) after removing the neighboring context lines (i.e., line 1, 6-7, 10-14, 17-18, 22-23, and the comments in line 3 and 4).

Aligning up changed lines of code in each fragment. Before we generate each code fragment's change signature based on deleted and added lines in it, we need to first pair up the changed lines of code since only some code fragments have one-to-one line change (i.e., at most one '-' line and one '+' line). For example, in Table 6, only the fragments F1-2 and F1-3 have one-to-one line change. For a multiple-line change in other fragments, we measure the edit distance similarity between each '-' line and all '+' lines and pair the one with the highest similarity. For instance, line 3 in Table 6c is paired with line 8 since it has the highest similarity with line 8 as compared with all the other lines. However, some lines could be simply deleted or added, causing their similarity with all other lines to be low. We handle this by *not* pairing the lines with the highest similarity of less than 0.5. As a result, line 3 in Table 6a will not be paired with line 5 due to the low similarity.

Generating the signatures of code changes. After determining the paired lines of code, we extract their syntactic changes [81] to generate the signatures with the following alterations:

Table 6: The evolution from raw code hunks to their code fragments and code change signatures.

(a) Example 1: Monero commit 1d5e8f46.

src/crypto/tree-hash.c - void tree_hash(		
const char (*hashes)[HASH_SIZE], size_t count, char *root_hash) {		
Code Hunk (line 1–23)		
1		`size_t cnt = tree_hash_cnt( count );`
2	-	`char ints[cnt][HASH_SIZE];`
3	-	`memset(ints, 0 , sizeof(ints)); // zero out as extra...`
4	+	`char *ints = calloc(cnt, HASH_SIZE); // zero out as extra...`
5	+	`assert(ints);`
6		`memcpy(ints, hashes, (2 * cnt - count) * HASH_SIZE);`
7		`for ({OMIT}) {`
8	-	`cn_fast_hash(hashes[i], 64, ints[j]);`
9	+	`cn_fast_hash(hashes[i], 64, ints + j * HASH_SIZE);`
10-14		`...`
15	-	`cn_fast_hash(ints[i], 64, ints[j]);`
16	+	`cn_fast_hash(ints + i * HASH_SIZE, 64, ints + j * HASH_SIZE);`
17-18		`}}`
19	-	`cn_fast_hash(ints[0], 64, root_hash);`
20	+	`cn_fast_hash(ints, 64, root_hash);`
21	+	`free(ints);`
22		`}`
Code Fragments		
F1-1	-	`char ints[cnt][HASH_SIZE];`
	-	`memset(ints, 0 , sizeof(ints));`
	+	`char *ints = calloc(cnt, HASH_SIZE);`
	+	`assert(ints);`
F1-2	-	`cn_fast_hash(hashes[i], 64, ints[j]);`
	+	`cn_fast_hash(hashes[i], 64, ints + j * HASH_SIZE);`
F1-3	-	`cn_fast_hash(ints[i], 64, ints[j]);`
	+	`cn_fast_hash(ints + i * HASH_SIZE, 64, ints + j * HASH_SIZE);`
F1-4	-	`cn_fast_hash(ints[0], 64, root_hash);`
	+	`cn_fast_hash(ints, 64, root_hash);`
	+	`free(ints);`
Code Change Signatures		
S1-1	`VAR[][] ==> calloc() memset() assert()`	
S1-2	`cn_fast_hash()`	
S1-3	`cn_fast_hash()`	
S1-4	`cn_fast_hash() free()`	

(b) Example 2: Ethereum commit b765e2d1.

core/transaction_pool.go - func (pool *TxPool)		
ValidateTransaction(tx *types.Transaction) error {		
Code Hunk (line 1–9)		
1-2		`return fmt.Errorf("tx.v != (28 || 27) = % v", v)}`
3	+	`senderAddr := tx.From()`
4	+	`if senderAddr == nil || len(senderAddr) != 20 {`
5	+	`return fmt.Errorf("invalid sender")`
6	+	`}`
7		`/* XXX this kind of validation needs to happen elsewhere...`
Code Fragment		
F2	+	`senderAddr := tx.From()`
	+	`if senderAddr == nil || len(senderAddr) != 20`
	+	`return fmt.Errorf("invalid sender")`
Code Change Signature		
S2	`From() if NIL || LEN return ERR`	

(c) Example 3: Ethereum commit 7c24cd79.

chain/transaction_pool.go - func (pool *TxPool)		
ValidateTransaction(tx *types.Transaction) error {		
Code Hunk (line 1–10)		
1		`//sender := pool.{OMIT}.proState.GetAccount(tx.Sender())`
2	-	`sender := pool.{OMIT}.CurrentState().GetAccount(tx.Sender())`
3	+	`senderAddr := tx.Sender()`
4	+	`if senderAddr == nil {`
5	+	`return fmt.Errorf("Invalid sender")`
6	+	`}`
7	+	`sender := pool.{OMIT}.CurrentState().GetAccount(senderAddr)`
8		`totAmount := new(big.Int).Set(tx.Value)`
Code Fragment		
F3	-	`sender := pool.{OMIT}.CurrentState().GetAccount(tx.Sender())`
	+	`senderAddr := tx.Sender()`
	+	`if senderAddr == nil`
	+	`return fmt.Errorf("Invalid sender")`
	+	`sender := pool.{OMIT}.CurrentState().GetAccount(senderAddr)`
Code Change Signature		
S3	`GetAccount() Sender() if NIL return ERR`	

- *(Recognizing and marking the type of statements.)* We first determine the control-flow statements by six reserved keywords, `if`, `for`, `while`, `return`, `throw`, and `defer`. If a control-flow statement is identified, we keep not only their type keyword but also their logical operators, e.g., "`||`" in line 4 in Table 6b. If a statement does not contain any control-flow keyword, we regard it as a function call if it includes a function or an assignment statement if it does not. For example, neither line 2 and line 4 in Table 6a have a control-flow keyword, but line 4 contains a function `calloc()`, so we regard line 4 as a function call and line 2 as an assignment.
- *(Preserving the name only for a function call.)* For function calls, we found that the function name itself is often enough to capture the statement nature despite parameter changes. Therefore, in code change signatures, we eliminate the function's parameters and caller variables. For example, we eliminate the three parameters of `cn_fast_hash()` in Table 6a and keep its function name only. As a result, it is easy for the generated three signatures (S1-2/3/4) to be in the same cluster. The symbols for calling a function vary, including `.`, `->`, and `::`. Additionally, if a function is called by another in a statement, we consider the last one as the actual function call of this statement, e.g., `GetAccount()` in Table 6c.

- *(Abstracting variable names and variable values.)* We also abstract variable names and variable values for a more concise signature. Specifically, we substitute variable names with the keyword VAR. If a variable is an array, we further add one or more `[][]`, such as `VAR[][]` for line 2 in Table 6a. For variable values, we define six keywords for the substitution of different types of values: NIL for `nil`, `null`, and `none`; BOL for `true` and `false`; NUM for numbers; TXT for strings; LEN/SIZE for size-related functions (e.g., `len()`, `length()`, `size()`, and `sizeof()`); and ERR for error functions.

Clustering code change signatures. As mentioned earlier, we cluster code change signatures from the vulnerabilities of blockchain-specific types. Since the RPC-related and database corruption types (i.e., the type T14 and T20 in Table 5) have only one or two blockchain-specific vulnerabilities, there is no need to cluster their signatures. Eventually, our target is 3,251 code fragments from 194 vulnerabilities of the type T2, T4, T7, T9, and T12 (see §5.2). The clustering process is similar to that in §5.1. One difference is that the WMD similarity is no longer applicable because code fragment signatures cannot be mapped to the token-based vector space. Therefore, we choose the Normalized Levenshtein distance [57] as the metric for calculating the similarity between code fragment signatures.

Table 7: 21 blockchain-specific patch code patterns obtained from the clustering result of 3,251 code fragments.

Type	ID	Description	Pattern (in the revised code change signature with some generalizations)	Example*
Transaction Related	P1	Check the transaction sender address	From|Sender|address() if ==NIL||LEN()!=NUM | IsValid() return ERR()	E: #272 [4]
	P2	Check the size of transactions in a pool	GetSerializeSize()|SIZE() if > MAX_STANDARD_TX_SIZE return BOL	B: #2273 [2]
	P3	Shuffle the transaction order; otherwise, fingerprinting	clear() selected_coins() shuffle() push_back()	B: #12699 [14]
	P4	Prevent the duplicated transaction	BOOST_FOREACH() insert() if SIZE() != SIZE() return DoS()	B: #1167 [1]
	P5	Prevent the malformed transaction	if !IsStandardTx() return DoS()	B: #8312 [8]
	P6	Prevent the double-spent transaction (relay)	if RelayableRespend() VAR = BOL	B: #4514 [3]
Block Related	P7	Validate the new header not from an invalid block	if IsValid() while!= insert() return DoS()	B: #11531 [11]
	P8	Check the gas limit in a block header	CalcGasLimit() if Cmp() != NUM return ERR()	E: #389 [5]
	P9	Check the block timestamp	time() int64() if <	M: #5902 [20]
	P10	Validate some block fields (number and hash) not null	GetBlockByNumber()|Hash() if != NIL	E: #1354 [6]
	P11	Do not connect a corrupted block	if CorruptionPossible() return AbortNode()|BOL	B: #12561 [13]
	P12	Prevent a malformed block to be propagated or forked	if CalcUncleHash()!=UncleHash() | if DeriveSha()!=TxHash() break	E: #20546 [23]
Peer/Node Related	P13	Disconnect after the timeout of header synchronization	if GetBlockTime() <= GetAdjustedTime()-NUM return BOL	B: #10345 [10]
	P14	Disconnect outbound peers on the invalid chain	if GetHash()!= fDisconnect = BOL	B: #11568 [12]
	P15	Drop the remote peer on an invalid or unverified TD	if Cmp(BlockTd) != 0	E: #604 [7]
Wallet Key/Password	P16	Immediately wipe the memory for critical secret keys	rct2sk()|MLSAG_Gen() memwipe()|memory_cleanse() return	M: #4268 [16]
	P17	Try to keep the wallet address in testnet or memory	generate() if != MAINNET || create_address_file ERR()	M: #3315 [15]
	P18	Do not skip asking for password when watch-only	if ask_password()	M: #4791 [17]
Other Check	P19	Check the validity of Quorum set	if !isQuorumSetSane() ERR() if throw invalid_argument()	S: #2233 [21]
RPC Related	P20	Enforce a gas cap of caller to protect against DoS	if := RPCGasCap() if != NIL if Cmp() > NUM Warn()	E: #19401 [19]
DB Corruption	P21	Avoid corruption due to unfinished blockchain tasks	CRITICAL_REGION_LOCAL()	M: #706 [9]

*: This column lists one example issue/PR for each code pattern, where B, E, M, and S represent Bitcoin, Ethereum, Monero, and Stellar, respectively.

To find a suitable clustering algorithm here, we also tested the four algorithms in §5.1, i.e., K-means, Gaussian Mixture, Agglomerative Clustering, and Affinity Propagation (AP). For the first three algorithms that require a pre-estimation of the number of clusters, we compute the Silhouette Coefficient score in a wide range of cluster numbers, but the result is not satisfactory. Therefore, we choose AP as our code clustering algorithm since it does not require pre-setting the number of clusters and performs well with a gradual tuning of the damping factor to 0.78. Under this setting, we eventually obtain a total of 174 clusters for further pattern analysis.

6.2 Blockchain-specific Patch Code Patterns

After clustering code change signatures, we inspect all the clusters and generalize the code patterns from them. Table 7 lists the 21 evidently blockchain-specific vulnerability patterns. They are organized in seven categories by their types (see §5.2), and most check-related patterns have been categorized into the detailed types.

Transaction-related patterns. We have identified six patterns (P1–P6) related to blockchain transactions. They check the sender (P1), size (P2), and order (P3) of a transaction, and prevent duplicated (P4), malformed (P5), double-spent (P6) transactions. Specifically, **P1** checks a sender address function to guarantee non-null values with valid lengths. **P2** checks the maximum size of transactions allowed in a pool. Besides the sender and length, the order of transactions could incur privacy risks like fingerprinting if not randomized. To address this and as the case of Bitcoin #12699, **P3** is to clear the original order, shuffle it, and push_back the new order. Both **P4** and **P5** check the blockchain structure to prevent duplicated or malformed transactions; otherwise, DoS could happen. **P6** prevents double-spent transaction relays via Relayable-Respend(), which was checked in Bitcoin #4515 and #4450.

Block-related patterns. We identify another six patterns (P7–P12) related to blockchain blocks. As a basic blockchain unit, a block stores multiple transactions and will be appended to the chain according to the consensus protocols. However, vulnerabilities could happen if the header (P7), gas limit (P8), and timestamp (P9) of a new block is invalid, or if some block fields are not null (P10), or if a corrupted (P11) or malformed (P12) block is identified. Specifically, **P7** validates that a newly appended block header is not from an invalid block. **P8** checks the gas limit in a block header, where gas is the fee for running smart contracts in Ethereum [35]. **P9** checks whether the timestamp in a block is less than the current time(), such as Monero #5902 and Ethereum #1355. **P10** validates the block fields like number and hash, and guarantees they are not null. Both **P11** and **P12** check the structure of a block to prevent a corrupted or malformed block being connected or forked.

Peer/node-related patterns. We also identify three patterns (P13–P15) related to peer/node synchronization and validation. Specifically, **P13** checks the time of block header synchronization, and if it is timed out, the node would disconnect. Bitcoin #10345 shows such an example in Bitcoin #10345, where the timeout is checked via GetBlockTime(). A similar case is Bitcoin #5463 for the block download timeout. Additionally, **P14** and **P15** perform the validation of remote peers and drop them if they fail. For example, Bitcoin #11568 and #11446 in P14 validate the hash of outbound peers via GetHash(). P15, on the other hand, checks the Ethereum-specific TD (Total Difficulty) field of a peer and guarantees the advertised TD actually deliverable, as in Ethereum #604 and #1451.

Wallet-related patterns. We further identify three patterns related to the blockchain wallet. First, since secret keys of a blockchain wallet are critical, **P16** immediately wipes the memory via memwipe() (Monero #4268) or memory_cleanse() (Bitcoin #10308) after generating some secrets. Second, the addresses in a wallet are also sensitive and should be kept in testnet or memory. For example, Monero #3315 in **P17** adds a create_address_file option for the address generating function generate() to create an address file only in the testnet environment. Similarly, Bitcoin #787 keeps the address table in memory and only writes to file when necessary. Third, a blockchain wallet requires users to always input passwords for critical operations. For example, Monero #4791 in **P18** performs such password checks via ask_password().

Other blockchain-specific patterns. From P19 to P21, we summarize the last three kinds of blockchain-specific patterns. Specifically, **P19** checks the validity of a Stellar-specific concept called *Quorum*, which represents a set of nodes that are sufficient to reach an agreement in the Stellar network [61]. For example, Stellar #2233 and #2209 check the sanity of Quorum via `isQuorumSetSane()`. **P20** is a RPC-related pattern, which restricts the gas cap of RPC calls. If the requested gas exceeds the cap limit via `RPCGasCap()`, the caller should be warned (see Ethereum #19401). The last pattern, **P21**, asks a blockchain client to gracefully shutdown itself when there are unfinished block synchronization and processing. This can be done by setting a global blockchain lock via `CRITICAL_REGION_LOCAL()`, as shown in Monero #706.

6.3 Applying the Obtained Patterns for Vulnerability Detection

While the focus of this paper is not vulnerability detection, we demonstrate the impact of obtained patterns by applying them to detect the same kinds of vulnerabilities in other blockchain projects. Specifically, in the blockchain world, it is normal for new blockchains to fork or partially reuse the code of classic blockchains, such as Bitcoin and Ethereum. These "forked" blockchains thus could encounter similar vulnerabilities that appeared in Bitcoin and Ethereum. Here we demonstrate a simple *direct* search of the vulnerable clones and leave a *variant* search to the future work.

Among the top 100 cryptocurrencies[4], we identified that 11 blockchains were forked from Bitcoin or had similar codebase as Bitcoin, including the rank #6 Dogecoin, #11 Bitcoin Cash, #12 Litecoin. For each blockchain b and a given vulnerability pattern p, we leverage the vulnerable file f, vulnerable function method m, and surrounding code c to first locate a clone of the original (unpatched) code (of p). We then determine whether the cloned code is vulnerable or not by checking if a patch (of p) has been applied. We use the pattern P3, P7, P11, P13, and P14 (see §6.2) that caused Bitcoin vulnerabilities in recent years to detect their cloned ones in the 11 projects. The results are worrisome: six projects are affected by at least one vulnerability, and two projects, Dogecoin and Bitcoin SV, even suffer from all five kinds of vulnerabilities. Among a total of the 20 vulnerabilities discovered, only ten use the same file and function name as the original vulnerability. This suggests the importance of generalized code patterns over the exact signatures relying on file names or function names.

We have reported all the 20 vulnerabilities to their corresponding vendors (via Email, GitHub, Discord, and Dash's Bounty) and offered them fix suggestions in late June and early July 2021. A summary of our vulnerability reporting is available at https://tinyurl.com/fse-227. Dogecoin promptly confirmed all of our five reports and planned to fix them in their next minor release version. Zcash, Horizen, and Ravencoin also confirmed our reports and are coordinating with their developers for fixes. Dash checked our report on P3 and believed that they had applied a different patch by sorting inputs and outputs based on BIP69. Bitcoin SV replied to two of our reports on P7 and P14; however, they seem not keen to fix them. We are still waiting for Bitcoin SV's replies on the other three reports.

7 RELATED WORK

Blockchain vulnerability research. Existing blockchain security studies mainly focus on smart contract vulnerability detection and transaction- or network-level analysis of the blockchains. For *smart contract vulnerability detection*, both static and dynamic program analysis tools have been proposed. For instance, Oyente [64], Zeus [51], Securify [76], Gigahorse [47], and ETHBMC [42] detected vulnerable smart contracts via symbolic execution, while Contract-Fuzzer [49] and ConFuzzius [75] used fuzzing inputs to detect smart contract vulnerabilities, and Sereum [70] and SODA [38] monitored run-time contract execution to detect on-chain attacks in modified EVMs. For *transaction-level analysis*, Karame et al. [53] analyzed the double-spending resilience of Bitcoin fast payments. TxSpector [85] studied Ethereum transactions by replaying historical transactions and recording EVM bytecode-level traces. DeFiPoser [87] proposed methods for discovering profit-generating transactions in DeFi protocols just in time. For *network-level analysis*, Apostolaki et al. [31] analyzed routing attacks by hijacking BGP prefixes and showed that such attacks could delay the propagation of blocks without being detected. Gao et al. [44] showed that by power adjusting and bribery racing, attackers could increase their mining rewards.

Mining-based vulnerability detection. Code clone detection is a long-standing research topic in the software engineering area [41, 46, 59, 67, 72, 78]. Existing approaches are mainly based on detecting duplicated token subsequences or identifying exact or similar subtrees in abstract syntax tree (AST) representations. For *token-based* approaches, CCFinder [52], CP-Miner [58], and ReDeBug [48] are the representative work. Recently, a token-based approach, VUDDY [54], generated code fingerprints via abstraction and normalization to speed up code clone detection. For *tree-based* approaches, e.g., DECKARD [50] and CloneDR [33], they considered code's structural information by generating ASTs and embedding them into a vector space for similarity comparison.

8 CONCLUSION

In this paper, we conducted the first empirical study of blockchain system vulnerabilities and their security patches using four representative blockchains. To enable this study, we proposed a vulnerability filtering framework to effectively identify 1,037 vulnerabilities and their 2,317 patches from 34,245 issues/PRs and 85,164 commits on GitHub. Based on this unique dataset, we performed three levels of analyses. Our analysis revealed three key findings of blockchain system vulnerabilities, including (i) the modules related to consensus, wallet, and networking are highly susceptible, each with over 200 issues; (ii) around 70% of blockchain vulnerabilities are in traditional types, but we also identify four new types specific to blockchains; and (iii) we are able to obtain 21 blockchain-specific vulnerability patterns that check unique blockchain attributes and validate various blockchain statuses, and demonstrate that they can be applied to detect similar vulnerabilities in other top blockchains. In the future, we will perform a thorough detection of blockchain system vulnerabilities based on the patterns extracted in this paper.

ACKNOWLEDGEMENTS

---

[4]Based on the market cap at https://coinmarketcap.com/ on 15 June 2021.

REFERENCES

[1] 2012. Bitcoin Pull Request #1167. https://github.com/bitcoin/bitcoin/pull/1167.
[2] 2013. Bitcoin Pull Request #2273. https://github.com/bitcoin/bitcoin/pull/2273.
[3] 2014. Bitcoin Pull Request #4514. https://github.com/bitcoin/bitcoin/pull/4514.
[4] 2015. Ethereum Pull Request #272. https://github.com/ethereum/go-ethereum/pull/272.
[5] 2015. Ethereum Pull Request #389. https://github.com/ethereum/go-ethereum/pull/389.
[6] 2015. Ethereum Pull Request #389. https://github.com/ethereum/go-ethereum/pull/1354.
[7] 2015. Ethereum Pull Request #604. https://github.com/ethereum/go-ethereum/pull/604.
[8] 2016. Bitcoin Pull Request #8312. https://github.com/bitcoin/bitcoin/pull/8312.
[9] 2016. Monero Pull Request #706. https://github.com/monero-project/monero/pull/706.
[10] 2017. Bitcoin Pull Request #10345. https://github.com/bitcoin/bitcoin/pull/10345.
[11] 2017. Bitcoin Pull Request #11531. https://github.com/bitcoin/bitcoin/pull/11531.
[12] 2017. Bitcoin Pull Request #11568. https://github.com/bitcoin/bitcoin/pull/11568.
[13] 2018. Bitcoin Pull Request #12561. https://github.com/bitcoin/bitcoin/pull/12561.
[14] 2018. Bitcoin Pull Request #12699. https://github.com/bitcoin/bitcoin/pull/12699.
[15] 2018. Monero Pull Request #3315. https://github.com/monero-project/monero/pull/3315.
[16] 2018. Monero Pull Request #4268. https://github.com/monero-project/monero/pull/4268.
[17] 2018. Monero Pull Request #4791. https://github.com/monero-project/monero/pull/4791.
[18] 2019. Bitcoin Core 0.11 (ch 1): Overview. https://en.bitcoin.it/wiki/Bitcoin_Core_0.11_(ch_1):_Overview.
[19] 2019. Ethereum Pull Request #19401. https://github.com/ethereum/go-ethereum/pull/19401.
[20] 2019. Monero Pull Request #5902. https://github.com/monero-project/monero/pull/5902.
[21] 2019. Stellar Pull Request #2233. https://github.com/stellar/stellar-core/pull/2233.
[22] 2020. Bitcoin Core: The Reference Implementation. https://cypherpunks-core.github.io/bitcoinbook/ch03.html.
[23] 2020. Ethereum Pull Request #20546. https://github.com/ethereum/go-ethereum/pull/20546.
[24] 2020. NLTK: Natural Language Toolkit. https://www.nltk.org/.
[25] 2020. spaCy: Industrial-Strength Natural Language Processing. https://spacy.io/.
[26] 2021. Agglomerative Clustering. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html.
[27] 2021. Gaussian Mixture Models. https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html.
[28] 2021. Swarm. https://github.com/ethersphere/swarm.
[29] Muneeb Ali, Jude Nelson, Ryan Shea, and Michael J Freedman. 2016. Blockstack: A global naming and storage system secured by blockchains. In *USENIX ATC*.
[30] Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, and Yacov Manevich. 2018. Hyperledger Fabric: A distributed operating system for permissioned blockchains. In *Proc. ACM EuroSys*.
[31] Maria Apostolaki, Aviv Zohar, and Laurent Vanbever. 2017. Hijacking Bitcoin: Routing Attacks on Cryptocurrencies. In *Proc. IEEE Symposium on Security and Privacy*.
[32] David Arthur and Sergei Vassilvitskii. 2007. K-Means++: The Advantages of Careful Seeding. In *Proc. ACM SODA*.
[33] Ira D. Baxter, Andrew Yahin, Leonardo Moura, Marcelo Sant'Anna, and Lorraine Bier. 1998. Clone Detection Using Abstract Syntax Trees. In *Proc. ACM ICSE*.
[34] Juan Benet. 2014. IPFS-content addressed, versioned, p2p file system. *CoRR arXiv* abs/1407.3561 (2014).
[35] Vitalik Buterin. 2014. A next-generation smart contract and decentralized application platform. *white paper* (2014).
[36] Yan Cai and Wing-Kwong Chan. 2014. Magiclock: Scalable detection of potential deadlocks in large-scale multithreaded programs. *IEEE Transactions on Software Engineering* (2014).
[37] Mengjie Chen, Daoyuan Wu, Xiao Yi, and Jianliang Xu. 2021. AGChain: A Blockchain-based Gateway for Permanent, Distributed, and Secure App Delegation from Existing Mobile App Markets. *CoRR arXiv* abs/2101.06454 (2021).
[38] Ting Chen, Rong Cao, Ting Li, Xiapu Luo, Guofei Gu, Yufei Zhang, Zhou Liao, Hang Zhu, Gang Chen, Zheyuan He, Yuxing Tang, Xiaodong Lin, and Xiaosong Zhang. 2017. SODA: A Generic Online Detection Framework for Smart Contracts. In *Proc. ISOC NDSS*.
[39] Ting Chen, Yuxiao Zhu, Zihao Li, Jiachi Chen, Xiaoqi Li, Xiapu Luo, Xiaodong Lin, and Xiaosong Zhang. 2018. Understanding Ethereum via Graph Analysis. In *Proc. IEEE INFOCOM*.
[40] Consultancy-me.com. 2022. Determining the real market capitalization of crypto assets. https://www.consultancy-me.com/news/4692/determining-the-real-market-capitalization-of-crypto-assets.
[41] Chunrong Fang, Zixi Liu, Yangyang Shi, Jeff Huang, and Qingkai Shi. 2020. Functional code clone detection with syntax and semantics fusion learning. In *Proc. ACM ISSTA*.
[42] Joel Frank, Cornelius Aschermann, and Thorsten Holz. 2020. ETHBMC: A Bounded Model Checker for Smart Contracts. In *Proc. USENIX Security*.
[43] Brendan J. Frey and Delbert Dueck. 2007. Clustering by Passing Messages Between Data Points. *Science* (2007).
[44] Shang Gao, Zecheng Li, Zhe Peng, and Bin Xiao. 2019. Power Adjusting and Bribery Racing: Novel Mining Attacks in the Bitcoin System. In *Proc. ACM CCS*.
[45] Zhipeng Gao, Lingxiao Jiang, Xin Xia, David Lo, and John Grundy. 2020. Checking Smart Contracts with Structural Code Embedding. *IEEE Transactions on Software Engineering* (2020).
[46] Sina Gholamian. 2021. Leveraging Code Clones and Natural Language Processing for Log Statement Prediction. In *Proc. ACM ASE*.
[47] Neville Grech, Lexi Brent, Bernhard Scholz, and Yannis Smaragdakis. 2019. Gigahorse: Thorough, Declarative Decompilation of Smart Contracts. In *Proc. ACM ICSE*.
[48] Jiyong Jang, Abeer Agrawal, and David Brumley. 2012. ReDeBug: Finding Unpatched Code Clones in Entire OS Distributions. In *Proc. IEEE Symposium on Security and Privacy*.
[49] Bo Jiang, Ye Liu, and W. K. Chan. 2018. ContractFuzzer: Fuzzing Smart Contracts for Vulnerability Detection. In *Proc. ACM ASE*.
[50] Lingxiao Jiang, Ghassan Misherghi, Zhendong Su, and Stephane Glondu. 2007. DECKARD: Scalable and Accurate Tree-Based Detection of Code Clones. In *Proc. ACM ICSE*.
[51] Sukrit Kalra, Seep Goel, Mohan Dhawan, and Subodh Sharma. 2018. ZEUS: Analyzing Safety of Smart Contracts. In *Proc. ISOC NDSS*.
[52] Toshihiro Kamiya, Shinji Kusumoto, and Katsuro Inoue. 2002. CCFinder: A Multilinguistic Token-based Code Clone Detection System for Large Scale Source Code. *IEEE Transactions on Software Engineering* (2002).
[53] Ghassan O. Karame, Elli Androulaki, and Srdjan Capkun. 2012. Double-Spending Fast Payments in Bitcoin. In *Proc. ACM CCS*.
[54] Seulbae Kim, Seunghoon Woo, Heejo Lee, and Hakjoo Oh. 2017. VUDDY: A Scalable Approach for Vulnerable Code Clone Discovery. In *Proc. IEEE Symposium on Security and Privacy*.
[55] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings to Document Distances. In *Proc. IMLS ICML*.
[56] Frank Li and Vern Paxson. 2017. A Large-Scale Empirical Study of Security Patches. In *Proc. ACM CCS*.
[57] Yujian Li and Bi Liu. 2007. A normalized Levenshtein Distance metric. *IEEE Transactions On Pattern Analysis and Machine Intelligence* (2007).
[58] Zhenmin Li, Shan Lu, Suvda Myagmar, and Yuanyuan Zhou. 2004. CP-Miner: A Tool for Finding Copy-paste and Related Bugs in Operating System Code. In *Proc. USENIX OSDI*.
[59] Chenyao Liu, Zeqi Lin, Jian-Guang Lou, Lijie Wen, and Dongmei Zhang. 2021. Can Neural Clone Detection Generalize to Unseen Functionalities*f*. In *Proc. ACM ASE*.
[60] Zhenguang Liu, Peng Qian, Xiang Wang, Lei Zhu, Qinming He, and Shouling Ji. 2021. Smart Contract Vulnerability Detection: From Pure Neural Network to Interpretable Graph Feature and Expert Pattern Fusion. In *Proc. IJCAI*.
[61] Marta Lokhava, Giuliano Losa, David Mazières, Graydon Hoare, Nicolas Barry, Eli Gafni, Jonathan Jove, Rafał Malinowsky, and Jed McCaleb. 2019. Fast and Secure Global Payments with Stellar. In *Proc. ACM SOSP*.
[62] Kangjie Lu, Marie-Therese Walter, David Pfaff, Stefan Nümberger, Wenke Lee, and Michael Backes. 2017. Unleashing Use-Before-Initialization Vulnerabilities in the Linux Kernel Using Targeted Stack Spraying. In *Proc. ISOC NDSS*.
[63] Oliver Lutz, Huili Chen, Hossein Fereidooni, Christoph Sendner, Alexandra Dmitrienko, Ahmad-Reza Sadeghi, and Farinaz Koushanfar. 2021. ESCORT: Ethereum Smart COntRacTs Vulnerability Detection using Deep Neural Network and Transfer Learning. *CoRR* abs/2103.12607 (2021).
[64] Loi Luu, Duc-Hiep Chu, Hrishi Olickel, Prateek Saxena, and Aquinas Hobor. 2016. Making Smart Contracts Smarter. In *Proc. ACM CCS*.
[65] Deepak Maram, Harjasleen Malvai, Fan Zhang, Nerla Jean-Louis, Alexander Frolov, Tyler Kell, Tyrone Lobban, Christine Moy, Ari Juels, and Andrew Miller. 2021. CanDID: Can-Do Decentralized Identity with Legacy Compatibility, Sybil-Resistance, and Accountability. In *Proc. IEEE Symposium on Security and Privacy*.
[66] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proc. NIPS*.
[67] Kawser Wazed Nafi, Tonny Shekha Kar, Banani Roy, Chanchal K Roy, and Kevin A Schneider. 2019. Clcdsa: cross language code clone detection using syntactical features and api documentation. In *Proc. ACM ASE*.
[68] Satoshi Nakamoto. 2008. Bitcoin: A peer-to-peer electronic cash system. *white paper* (2008).
[69] Shen Noether. 2015. Ring SIgnature Confidential Transactions for Monero. *IACR Cryptol. ePrint Arch.* (2015).
[70] Michael Rodler, Wenting Li, Ghassan O. Karame, and Lucas Davi. 2019. Sereum: Protecting Existing Smart Contracts Against Re-Entrancy Attacks. In *Proc. ISOC NDSS*.

[71] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* (1987).

[72] Hitesh Sajnani, Vaibhav Saini, Jeffrey Svajlenko, Chanchal K. Roy, and Cristina V. Lopes. 2016. SourcererCC: Scaling Code Clone Detection to Big-Code. In *Proc. ACM ICSE.*

[73] Bo Tang, Hongjuan Kang, Jingwen Fan, Qi Li, and Ravi S. Sandhu. 2019. IoT Passport: A Blockchain-Based Trust Framework for Collaborative Internet-of-Things. In *Proc. ACM SACMAT.*

[74] Yuan Tian, Julia L. Lawall, and David Lo. 2012. Identifying Linux Bug Fixing Patches. In *Proc. ACM ICSE.*

[75] Christof Ferreira Torres, Antonio Ken Iannillo, Arthur Gervais, and Radu State. 2021. ConFuzzius: A Data Dependency-Aware Hybrid Fuzzer for Smart Contracts. In *Proc. IEEE European Symposium on Security and Privacy.*

[76] Petar Tsankov, Andrei Dan, Dana Drachsler-Cohen, Arthur Gervais, Florian Bünzli, and Martin Vechev. 2018. Securify: Practical Security Analysis of Smart Contracts. In *Proc. ACM CCS.*

[77] Zhiyuan Wan, David Lo, Xin Xia, and Liang Cai. 2017. Bug Characteristics in Blockchain Systems: A Large-Scale Empirical Study. In *Proc. ACM MSR.*

[78] Huaijin Wang, Pingchuan Ma, Yuanyuan Yuan, Zhibo Liu, Shuai Wang, Qiyi Tang, Sen Nie, and Shi Wu. 2022. Enhancing DNN-Based Binary Code Function Search With Low-Cost Equivalence Checking. *IEEE Transactions on Software Engineering* (2022).

[79] Zhilong Wang, Xuhua Ding, Chengbin Pang, Jian Guo, Jun Zhu, and Bing Mao. 2018. To detect stack buffer overflow with polymorphic canaries. In *Proc. DSN.*

[80] Sam M Werner, Daniel Perez, Lewis Gudgeon, Ariah Klages-Mundt, Dominik Harz, and William J Knottenbelt. 2021. SoK: Decentralized Finance (DeFi). *CoRR arXiv* abs/2101.08778 (2021).

[81] Daoyuan Wu, Debin Gao, Eric K. T. Cheng, Yichen Cao, Jintao Jiang, and Robert H. Deng. 2019. Towards Understanding Android System Vulnerabilities: Techniques and Insights. In *Proc. ACM AsiaCCS.*

[82] Meng Xu. 2020. *Finding Race Conditions in Kernels: The Symbolic Way and the Fuzzy Way.* Ph. D. Dissertation. Georgia Institute of Technology.

[83] Fan Zhang, Ethan Cecchetti, Kyle Croman, Ari Juels, and Elaine Shi. 2016. Town Crier: An authenticated data feed for smart contracts. In *ACM CCS.*

[84] Fan Zhang, Deepak Maram, Harjasleen Malvai, Steven Goldfeder, and Ari Juels. 2020. DECO: Liberating web data using decentralized oracles for TLS. In *Proc. ACM CCS.*

[85] Mengya Zhang, Xiaokuan Zhang, Yinqian Zhang, and Zhiqiang Lin. 2020. TXSPECTOR: Uncovering Attacks in Ethereum from Transactions. In *USENIX Security.*

[86] Mingyi Zhao, Jens Grossklags, and Peng Liu. 2015. An Empirical Study of Web Vulnerability Discovery Ecosystems. In *Proc. ACM CCS.*

[87] Liyi Zhou, Kaihua Qin, Antoine Cully, Benjamin Livshits, and Arthur Gervais. 2021. On the Just-In-Time Discovery of Profit-Generating Transactions in DeFi Protocols. In *Proc. IEEE Symposium on Security and Privacy.*

[88] Yaqin Zhou and Asankhaya Sharma. 2017. Automated Identification of Security Issues from Commit Messages and Bug Reports. In *Proc. ACM FSE.*