SELECTIVE TAINT: Efficient Data Flow Tracking With Static Binary Rewriting

Sanchuan Chen  Zhiqiang Lin  Yinqian Zhang
The Ohio State University
{chen.4825, lin.3021, zhang.834}@osu.edu

Abstract
Taint analysis has been widely used in many security applications such as exploit detection, information flow tracking, malware analysis, and protocol reverse engineering. State-of-the-art taint analysis tools are usually built atop dynamic binary instrumentation, which instruments at every possible instruction, and rely on runtime information to decide whether a particular instruction involves taint or not, thereby usually having high performance overhead. This paper presents SELECTIVE TAINT, an efficient selective taint analysis framework for binary executables. The key idea is to selectively instrument the instructions involving taint analysis using static binary rewriting instead of dynamic binary instrumentation. At a high level, SELECTIVE TAINT statically scans taint sources of interest in the binary code, leverages value set analysis to conservatively determine whether an instruction operand needs to be tainted or not, and then selectively taints the instructions of interest. We have implemented SELECTIVE TAINT and evaluated it with a set of binary programs including 16 coreutils (focusing on file I/O) and five network daemon programs (focusing on network I/O) such as nginx web server. Our evaluation results show that the binaries statically instrumented by SELECTIVE TAINT has superior performance compared to the state-of-the-art dynamic taint analysis frameworks (e.g., 1.7x faster than that of libdft).

1 Introduction
One of the mostly used techniques in software security is dynamic taint analysis [28], also called dynamic data flow tracking (DDFT), which tracks the data flow of interest during program execution and has been widely used in many security applications, such as exploit detection [14, 28–30], information flow tracking [34, 41], malware analysis [18, 39], and protocol reverse engineering [10, 19]. However, the implementation of taint analysis often has high performance overhead. For instance, a state-of-the-art dynamic taint analysis framework libdft [17] imposes about 4x slowdown for gzip when compressing a file.

There has been a body of research that seeks to improve the performance of taint analysis. For instance, Jee et al. [16] applied compiler-like optimizations to eliminate redundant logic in taint analysis code. SHADOW REPLICA [15] improved the performance by decoupling taint logic from program logic, minimizing the information needed to communicate, and optimizing the shared data structures between them. TAINTPIPE [25] explored a parallel and pipeline scheme. STRAIGHT TAINT [24] combined an online execution state tracing and offline symbolic taint analysis for further performance improvement.

Interestingly, these general DDFT frameworks and their optimizations are built atop dynamic binary instrumentation (DBI), particularly Intel’s PIN [22], to instrument the taint analysis logic at runtime. We believe a fundamental reason of using DBI for these frameworks is to basically avoid the code discovery challenge from static binary analysis. Note that PIN is a DBI tool, and it dynamically disassembles, compiles, and reassembles the executed code at runtime without any code discovery issues. The core module of PIN is a virtual machine (VM) that consists of a just-in-time (JIT) compiler, an emulator, and a dispatcher. PIN also has a rich set of APIs used for Pintool’s implementations. However, the VM and APIs both add additional performance overhead to a taint analysis tool.

Unlike DBI, static binary instrumentation (SBI) inserts the analysis code directly into the native binary and thus avoids the unnecessary DBI overhead incurred by JIT and emulation. Meanwhile, SBI would have fewer context switches, since the rewritten binary has a better code locality. While it is challenging to perform static binary analysis, recently there are substantial advancements in static binary rewriting and reassembling (e.g., UROBOROS [37], RAMBLR [36], MultiVERSE [6], and recently Datalog Disassembly [13]). Therefore, it is worthwhile revisiting the taint analysis and study the feasibility of using static binary rewriting for more efficient taint analysis.

In addition to the use of DBI, existing taint analysis frameworks also instrument the binary code at every possible instruction that can contribute the information flow, and rely on the execution context to determine whether there is a need to taint the corresponding operand. However, if a static analysis could figure out precisely the instructions that will never get involved in taint analysis (e.g., via some conservative static analysis), it would have not instrumented them. Therefore, enabling taint analysis to selectively instrument the binary code statically is viable and highly desired.
In this paper, we propose \textsc{SelectiveTaint}, an efficient selective taint analysis framework for binary code with static binary rewriting. There are two salient features in \textsc{SelectiveTaint}. First, it directly removes the overhead from dynamic binary translation, and is built atop SBI instead of DBI. Second, it scans taint sources of interest in the binary, statically determines whether an instruction operand will be involved in taint analysis by leveraging the value analysis (VSA) \cite{3,4}, and then selectively taints the instructions of interest. There are well-known challenges that \textsc{SelectiveTaint} must address, such as how to deal with point-to (i.e., alias) analysis inside binary code. \textsc{SelectiveTaint} solves this problem by conservatively identifying the memory addresses that will never be involved in taint, and then taint the rest.

We have implemented \textsc{SelectiveTaint} atop SBI and evaluated it with a variety of applications consisting of 16 coreutils, and five network daemon programs such as Nginx web server. We use these programs as the benchmarks because they represent both file I/O and network I/O, the two most common input channels used by real world programs. The evaluation results show that \textsc{SelectiveTaint} is 1.7x faster than that of \textsc{libdft}, a state-of-the-art dynamic taint analysis framework. We formally prove that \textsc{SelectiveTaint} is soundy (mostly sound) \cite{21}, and also confirm the soundness of \textsc{SelectiveTaint} by using it to detect real-world exploits against the memory corruptions vulnerabilities in a variety of software including image decoder, audio normalization, and assembler.

In short, we make the following contributions:

\begin{itemize}
  \item We propose \textsc{SelectiveTaint}, the first static-binary-rewriting based selective taint analysis framework, to substantially improve the performance overhead incurred by earlier DBI-based approaches.
  \item We also present a conservative tainted instruction identification approach, which statically identifies the instructions that will never involve tainted memory or registers by using VSA and then conservatively taints the rest instructions.
  \item We have implemented \textsc{SelectiveTaint}, and tested with 16 coreutils and five network daemons. The evaluation results show that \textsc{SelectiveTaint} has superior performance compared to the state-of-the-art taint analysis tools such as \textsc{libdft}.
\end{itemize}

2 Background

2.1 Taint Analysis

Taint analysis is a widely used program analysis technique that tracks the flow of data of interest as they propagate during the program execution \cite{28}. It is also referred as dynamic data flow tracking (DDFT) or dynamic taint analysis (DTA), which is usually implemented using virtualization or DBI and can be performed per-process \cite{17} or system-wide \cite{39}.

\begin{figure}

\begin{lstlisting}
void process(int client_sock, char *buffer, int size)
{
    char ch;
    int read_size = recv(client_sock, buffer, 2048, 0);
    if(read_size > 0)
    {
        ch = buffer[0];
        if(ch >= 'a' && ch <= 'z')
          buffer[0] = ch -32;
        write(client_sock, buffer, read_size);
        memset(buffer, 0, 1024);
    }
}

int server(int client_sock)
{
    int i = 0;
    char buffer[1024] = {0};
    for(i = 0; i < 3; i++)
    {
        process(client_sock, buffer, 1024);
    }
    return 0;
}
\end{lstlisting}

Figure 1: A simplified running example

\end{figure}

Taint analysis needs taint tags, which are markings associated to registers and memory to indicate their taint status. Taint tags can have different (1) granularities to mark the taintedness for a bit, a byte, a word, or a block of data, and (2) sizes to indicate the taintedness to be a bit—tainted or not, or an integer—which input byte tainted the data. A finer granularity enhances taint analysis precision but adds performance costs, e.g., the storage cost for tag-related data structure, whereas a coarser granularity offers less precision but better performance. When a tag size is a single bit, it can be used to represent whether a corresponding register or memory location is tainted or not; when it is an integer, it can represent which part of the input (e.g., a particular byte offset) has been propagated to the tainted registers or memory locations.

A taint analysis typically consists of three components: taint sources, taint propagation, and taint sinks. In the following, we use a simplified networking program illustrated in Figure 1, as a running example, to demonstrate how a typical taint analysis works.

\begin{itemize}
  \item **Taint sources.** Taint sources are program points or memory locations where data of interest is introduced. Typically, taint analysis is interested in user input coming from locally or remotely. For example, in Figure 1, if we are interested in the remote input, we will taint the data stored in \texttt{buffer} right after entering the system when calling \texttt{libc function recv} at line 4.
  \item **Taint propagation.** Taint tags are propagated during the program execution according to the taint propagation rules, which are specified with respect to the semantics of each instruction, e.g., the specific operands in the instruction, and also the side-effect of the instruction. For instance, for instruction \texttt{ADD src, dst}, a taint propagation rule could specify that the new tag of \texttt{dst}
\end{itemize}
Abstract addresses and value sets. In Figure 1, at line 7 ch is assigned the tainted data of buffer[0] and at line 9 buffer[0] is calculated based on tainted ch, which has a data dependency, whereas at lines 8-9 whether buffer[0] is assigned or not depends on the outcome of the predicate in the if statement, which involves a tainted ch with a control dependence between buffer[0] and ch. Note that most of the DDFT works (e.g., [15–17, 24, 25]) only consider taint propagation based on data dependencies.

- **Taint sinks.** Taint sinks are specific program instructions where taint analysis checks the existence of taint tags of interest for various security applications such as detecting control flow hijacks or information flow leakage. Common taint sinks are control flow transfer instructions for detecting control flow hijack attacks, or output system calls (e.g., write, send) for detecting information leakage (e.g., a tainted secret leaked out). In Figure 1, line 10 could be a taint sink for information leakage detection, since it is the libc function write that writes the content starting at buffer to client_socket.

### 2.2 Value Set Analysis

Value set analysis (VSA) [3, 4] is a static program analysis technique. It over-approximates the set of possible values that each data object of the program could hold at each program point, and it uses a value set to represent the set of memory addresses and numeric value quantities.

**Memory regions and abstract locations.** VSA uses an abstract memory model that separates the address space into multiple disjoint areas that are referred to as memory regions. Memory regions in VSA consist of: a global region for memory locations storing uninitialized and initialized global variables, a stack region per function for memory locations of activation record of a procedure, and a heap region per heap allocation for memory locations allocated by a particular malloc-type of function call site. An abstract location, i.e., an a-loc, is a variable-like entity which spans from one statically known location to next statically known location (not including it).

**Abstract addresses and value sets.** An abstract address in VSA is represented by a pair (memory-region, offset). A set of abstract addresses can be represented using:

\[
\{i | \text{rgn}_i \rightarrow \{o_1^i, o_2^i, \ldots, o_n^i\}\}
\]

More specifically, when there are at most one stack memory region and one heap memory region, the value set can be specified as 3-tuple [4]:

\[
(\text{global} \rightarrow O^g, \text{stack} \rightarrow O^s, \text{heap} \rightarrow O^h)
\]

abbreviated as \((O^g, O^s, O^h)\). A set of memory offsets in each memory region is represented by a strided-interval (SI): \(s[1,u]\), where \(s\) is the stride, \(l\) and \(u\) are lower bound and upper bound. For instance, \((\{1, 3, 5\}, 1, 1)\) could be represented using SI as \((2[1,5], 1, 1)\).

The analysis is performed on a control-flow graph (CFG) in which each node represents an instruction (not a basic block as VSA is calculated for each instruction) and each edge represents a control flow transfer. A transfer function that characterizes the instruction semantics is associated with each edge. Note that since the address values and numeric values are interleaved in the binary, VSA tracks address values and numeric values at the same time.

### 2.3 Binary Instrumentation

Binary instrumentation is the process of instrumenting binary with additional analysis code added and meanwhile maintaining the original functionality. It is a widely used technique for many important security applications such as malware analysis and binary code hardening. Binary instrumentation could be either static or dynamic.

**Static binary rewriting.** Static binary instrumentation (SBI), also known as static binary rewriting, modifies the binary file directly. Static binary rewriting can be performed in three ways, as summarized in RAMBLR [36]: (1) trampoline-based, (2) lifting and recompiling, (3) symbolization [40] and reassembling [37]. Specifically, in trampoline-based approaches, hooks which detour the control flow to trampolines are added to the binary. In contrast, for lifting and recompiling, the binary code will be first lifted into an intermediate representation (IR), then inserted with the code of interest in the IR, and finally compiled back. The first two approaches have been known in the community for years. Recently, symbolization and reassembling approach was proposed, in which a rewriter needs to identify the locations pointed by memory references first, and then symbolize those references. The process of converting numeric references back to symbols is called symbolization. After symbolization, the rewriter could correctly relocate binary in reassembling. The first two approaches impose significant overhead and the last approach may mix code with data and may not correctly separate them.

**Dynamic binary instrumentation.** Dynamic binary instrumentation (DBI) recovers the code while program is executing, which can correctly separate program code from data. However, compared with static approaches, DBI has high performance overhead. There are generally two ways to implement DBI: using a trampoline, or using just-in-time (JIT) compiling. The trampoline approach replaces the instruction with a trampoline at run-time which jumps to the instrumented analysis code, and the JIT compiling approach dynamically compiles the binary on the fly.

### 3 Challenges and Insights

#### 3.1 Challenges

To clearly illustrate the challenges of selective taint analysis, we still use the example code shown in Figure 1. This
program receives three messages from a client (line 19-22), capitalizes the first character in each message if needed (line 8-9), and sends the messages back to the client (line 10). It has a buffer overflow vulnerability at line 4, when receiving the input with size larger than 1,024 bytes. The taint source of our interest is the network input stored in array buffer, which is tainted right after the execution of libc function recv. The taint sink of our interest is the control flow transfer instruction ret of function server at line 23, assume our objective is to detect the control flow hijacks.

Performing selective binary code taint analysis using static binary rewriting is by no means trivial. Unlike DBI-based approaches where taint analysis logic is instrumented at runtime, a SBI-based approach has to analyze and rewrite the binary statically. In addition to the challenges from static binary disassembling and rewriting (they are orthogonal to the problem we aim to solve in this paper), SELECTIVETAINT has to address at least the following unique challenge—how to determine whether a disassembled instruction needs to be instrumented by taint analysis. If so, we have to also rewrite it accordingly based on the taint semantics (e.g., whether this instruction introduces a taint sources, contributes to taint propagation, or it is a taint sink).

Essentially, the problem becomes how to determine the taintedness of an instruction according to its operands including both memory addresses and registers without executing the binary. Determining the taintedness of registers is easier compared to memory addresses, since registers can be directly identified based on names whereas a memory address cannot be easily resolved. Therefore, determining the taintedness for memory addresses is much harder in SBI. More specifically, different from DTA in which a memory address has a single runtime address at each program point, static binary taint analysis can only conservatively infer the possible values for a symbolic memory address at each program point. Except for global memory addresses, symbolic addresses of stack and heap are only in relative addresses when performing the static analysis. In addition, there are also unknown inputs (from a command line, local files and keystrokes, or remote network packet) that also make the problem hard.

### 3.2 Insights

It is obvious that in order to address the aforementioned challenges, it requires the inference of possible values of both registers and memory cells at each program point. Fortunately, a key enabling technique in this direction is the VSA [3, 4], which seeks to compute the possible values at each symbolic memory address and register. Therefore, with VSA, we could determine whether a particular memory address or register involves taint or not, e.g., whether it is an alias to the address of our interest, or it will hold the propagations of the tainted data.

To see exactly how VSA helps our analysis, we show the value set analysis results of our running example along with its assembly code in Table 1. At the prologue of function server, the initial esp has a value set of (1, 0x0, 1), since the stack pointer address for a function is unknown statically. After executing push %ebp at 0x8048687, esp has a value set of (1, -0x4, 1). The analysis continues, and computes the rest of the VSA for each register and memory operand. With the statically computed VSA, we can easily see that ebx at 0x80486a9 and eax at 0x80486c6 have the same value sets (1, -0x410, 1), and thus these two registers are actually aliased. In fact, both of them refer to the address of the local variable buffer defined in function server.

To statically analyze which instructions need to be tainted, a straw-man approach is to statically maintain tainted value sets (i.e., value sets of registers and symbolic memory that need to be tainted) at each program point. In particular, this approach checks whether the value set of any of the operand of an instruction is a subset of the tainted value set, if so, this instruction is added into the tainted instruction set; meanwhile the register or symbolic memory of the corresponding operand is also added to the tainted value sets if the taint will be propagated to this operand, and the corresponding taint rule is used to taint this instruction.

However, when analyzing real-world binaries, VSA may lose its precision due to various factors such as imprecise control flow graph (CFG), sophisticated static point-to analysis (which is an undecidable problem [31]), and unknown inputs. Consequently, as illustrated in Figure 2, we may not be able to get the ideal tainted instruction set I for the instructions that need to be tainted, and instead the VSA identified the must-tainted instruction set I¹ (i.e., instructions must be tainted) can have false negatives because of the imprecision mentioned, but it will not have false positives for the must-analysis (the worst case is it can be empty, if the must-analysis cannot decide anything). On the other hand, by using VSA, we can also identify must-not-tainted instruction set I² that will never be involved in taint analysis. Therefore, in order not to have any false negatives (no missing of attacks) when using taint analysis, we eventually decide to taint the instructions that are not in I². The worst case of our algorithm is that the identified I² is Ø, which means we taint all instructions similarly to other DBI-based taint analysis. Our key objective is to confidently enlarge I² as much as possible (note that we will not have false positives when being conservative).

As in our running example, in Table 1, the instructions in light gray are identified as in must-not-tainted instruction set I², the instructions in dark gray are identified as in must-tainted instruction set I¹, and all instructions not in light gray are our conservatively tainted instructions. For each instruction, a must-not-tainted value set V¹ is maintained and if the value set of any of its operand is a subset of V¹, this instruction is added to I¹. For instance, for instructions at 0x804861a and 0x804861d before taint introduction at 0x8048620, must-not-tainted value set V¹ equals value set S, which contains all possible values at this execution point (recall VSA is a
Table 1: The assembly code snippets of our running example. Instructions in light gray are identified by our analysis as in must-not-tainted instruction set $\mathcal{I}_u$, and Instructions in dark gray are in must-tainted instruction set $\mathcal{I}_t$.  

Scope and Assumptions. In this work, we focus on x86 binaries with ELF format running atop Linux platform. We assume the binary code is not obfuscated, and we are able to get their correct disassembly. For proof-of-concept, we demonstrate the use of taint analysis to track the untrusted user input through static binary rewriting, and detect the memory exploits by just using a single bit (tainted or not) in our taint record. Also, our static binary rewriting is based on DYINST [7]. While it is not perfect, it has been widely used...
4 Detailed Design

In this section, we present the detailed design of SELECTIVE_TAINT. As illustrated in Figure 3, there are four key components inside:

- **CFG Reconstruction** (§4.1). When given an application binary, we will first disassemble and build its CFG starting from the main function. If there is a library call, we will resolve the calling target and use the function summaries to decide whether further instrumentation of the library is needed. If an indirect jmp/call is encountered, we will perform backward slicing [36] and use the VSA and type information to resolve the target.

- **Value Set Analysis** (§4.2). VSA [3] has become a standard technique in static binary analysis for determining the possible values of a register or a symbolic memory address. We use the VSA to help identify the instruction operands that are never involved in the taint analysis.

- **Taint Instruction Identification** (§4.3). Selective tainting essentially aims to identify the instructions that are involved in the taint analysis. With the identification of $\mathcal{I}_u$ by VSA, we then start from the instructions that introduce the taint sources, and systematically identify the rest of instructions that are not in $\mathcal{I}_u$.

- **Binary Rewriting** (§4.4). Having identified the instructions that need to be tainted, we then use the static binary rewriting to insert the taint analysis logic including tracking of the taint sources and taint propagations as well as the taint checks at the taint sinks.

4.1 CFG Reconstruction

The first step of SELECTIVE_TAINT is to disassemble and rebuild the CFG, when given an application binary. This is quite a standard process and the only additional challenge is to identify the control flow targets of the indirect calls and jumps, as they are important to compute the VSA. To get the CFG, we first reconstruct the possible control flow targets using the RAMBLR [36] approach, and in case of undecided target (e.g., jmp/call eax), we use the following approaches:

**Handling Indirect Call.** We adopt and implement two forward-edge CFI identification approaches, namely TYPEARMOR [35] and τCFI [26], to recover the type information (i.e., parameter count and parameter type) about actual and formal parameters at the callsites and callee functions. By connecting the matching callsites and callees regarding these type information, we build a CFG which is an over-approximation of actual CFG. The type information is generated by running liveness analysis at indirect call sites and use-def analysis at callees.

**Handling Indirect Jump.** We first use VSA to resolve the indirect jump target and connect the jump target if it is solved. Otherwise, we determine whether the function that contains the indirect jump uses any external data references (e.g., global variable addresses): if not, we connect all of the possible basic block starting address in this function as the potential jump target (we still consider it local); otherwise, we connect the jump target with all function entry addresses. The rationale is we notice the inter-procedural jumps we encountered are from compiler optimizations, and basically compiler optimizes the call instruction with an indirect jump. We therefore connect the indirect jump in this way to get an over-approximation of the CFG.

4.2 Value Set Analysis

**Our VSA Algorithm.** A key technique inside SELECTIVE_TAINT is the VSA [3], which is a context-sensitive and flow-sensitive whole program analysis. As described in Algorithm 1, our whole_program_vsa first initializes the ValueSet for each instruction in the program with an initial esp, initial empty heap, and initial memory cell values resolved from original binary. Then, function VSA is called to analyze each function func, which is of work list style with multiple iterations on each individual instruction until no changes are discovered (i.e., reached a fixed point). The context and value sets are adjusted depending on the type of instruction opcode, e.g., for call/ret instruction, inter-procedural analysis is performed and the environment is adjusted accordingly including changing the current stack region and matching formal and actual parameters.

**Practical Challenges.** While the idea of calculating VSA is simple, it has a number of practical challenges when used for data flow tracking, such as context-sensitive, flow-sensitivity, and alias analysis. In the following, we describe these challenges and also how we have addressed them below:
Algorithm 1: Whole Program Value Set Analysis

Function whole_program_VSA(CFG, ValueSet):
1: ValueSet, context ← init()
2: VSA(entryFunc, context)

Function VSA(CFG, ValueSet, func, context):
3: worklist ← [entryInst]
4: while worklist ≠ ∅ do
5: i ← pop(worklist)
6: if callInst(i) then
7: newContext ← adjustContext(context, callee(i))
8: VSA(CFG, ValueSet, callee(i), newContext)
9: if retInst(i) then
10: adjustContext(context, caller(i))
11: if condInst(i) then
12: newValueSet ← ValueSet_{cond_i}
13: else
14: newValueSet ← ValueSet_{cond_i}
15: exit
16: if uninitialized(op_i) then
17: if newValueSet_{cond_i} ≠ ValueSet_{cond_i} then
18: newValueSet_{cond_i} ← newValueSet_{cond_i} ∪ newValueSet_{cond_i}
19: push(worklist, succs(i))
20: push(worklist, pred(i))

(I) Handling context-sensitivity. It will be overly complicated if a function is called multiple times when performing the inter-procedural analysis in a CFG. We therefore augment our VSA with a cloning-based context sensitivity analysis [38]. Basically, we have a separate analysis for each function clone per calling context. More specifically, we generate a function clone for every acyclic path through a program call graph and, for cyclic paths, we merge all functions in a strongly connected component to have a single function context for them as in [38].

(II) Handling flow-sensitivity. Since VSA is flow-sensitive and per-instruction, it is an engineering challenge to inspect each instruction statically. We therefore borrow the idea of how symbolic execution interprets each instruction and updates the corresponding symbolic states. Essentially, when perform our flow sensitivity analysis, we need to interpret each instruction, and updates the VSA based on its semantics. Since symbolic execution is well studied (with many open source tools), we do not describe how we implement our interpreter and instead we abstract it as a simple EXE operation (line 18) in algorithm 1, which is responsible to capture the value set changes for each analyzed instruction (working as a transfer function in static analysis). In particular, all incoming value sets are merged on a per register and memory cell basis as input value sets and are fed into the static reasoning engine EXE to update the value sets of each registers and memory cells for this analyzed instruction i according to its semantics, and this updated value set forms the output of our static analysis for this particular instruction. A work list keeps looping and works on each instruction as such, until a fixed point is reached.

(III) Handling a-locs with unknown values or addresses. Performing VSA on binary suffers from the lack of dynamic information (e.g., calling context, and concrete memory addresses). One major issue when applying VSA on real-world binary is uninitialized variables and their aliases. Among these uninitialized variables, some are used in address calculation, which leads to a-locs with unknown addresses. To conservatively taint instructions, we need to infer the value of these unknown addresses; otherwise the reads and writes to them would indicate the reads and writes to the whole address space.

In case VSA encounters an a-loc with uninitialized values or addresses due to system inputs for instance, the special handling is shown in line 16-17 in algorithm 1. In particular, our analysis will assume the uninitialized a-loc to have any value, i.e., with the value set (T, T, T). In practice, we have identified the following three cases in which VSA cannot determine the corresponding addresses:

(i) Unknown values from command line input (CLI), e.g., argv[]. The argv elements are pointers which is uninitialized at analysis-time. As shown in Figure 4a, instruction at 0x804b362 reads argv[0] which is unknown at analysis-time.

(ii) Unknown addresses or values passed from missing callers. Even we use approaches such as TypeArmor to recover CFG, there are still some callee functions without callers and the calling context is missing for these callee. As shown in Figure 4b, the function version_etc_arn has no identified callers, and thus, the value of parameter at instruction 0x804b7a7 is uninitialized.

(iii) Unknown addresses or values due to library functions and system calls. For instance, fopen64 function returns a pointer which is a pointer to FILE struct that is uninitialized at analysis-time as illustrated in Figure 4c.

4.3 Taint Instruction Identification

After our whole program VSA analysis, we next need to identify the instructions that need to be instrumented for the taint analysis with the computed VSA. To this end, we have to decide whether a memory address involves taint or not, which essentially leads to problem of point-to (i.e., alias) analysis. However, due to the imprecision of the static point-to analysis, we may not be able to resolve all memory addresses with VSA [3, 4], and instead we focus on identifying the addresses that will never be involved in taint analysis for each specific instruction (since VSA is flow sensitive). Initially, all instructions will be marked tainted (i.e., they will all be instrumented for taint analysis). As described in §3.2, our key objective is to minimize this set, by identifying and enlarging the must-not tainted set. In the following, we describe how we achieve this.
bzip2
0804b296 <main>:
  804b296: push %ebp
  804b297: mov %esp,%ebp
...
  804b2a2: mov 0xc(%ebp),%esi
...
  804b362: mov (%esi),%edx
(a) Entry-function uninitialized variable

comm
0804b7a0 <version_etc_arn>:
  804b7a0: push %ebp
  804b7a1: push %edi
  804b7a2: push %esi
  804b7a3: push %ebx
  804b7a4: sub $0x5c,%esp
  804b7a7: mov 0x74(%esp),%eax
  804b7ab: mov 0x70(%esp),%esi
  804b7af: mov 0x78(%esp),%edx
  804b7b3: mov 0x7c(%esp),%ecx
  804b7b7: test %eax,%eax
  804b7b9: mov 0x80(%esp),%ebx
  804b7c0: mov 0x84(%esp),%edi
(b) Incomplete CFG caused uninitialized variable

cut
08049dd0 <cut_file>:
  8049dd0: push %ebp
  8049dd1: push %edi
  8049dd2: push %esi
  8049dd3: push %ebx
  8049dd4: sub $0x4c,%esp
...
  8049df3: call 8048e70 <fopen64@plt>
  8049df8: test %eax,%eax
  8049dfa: mov %eax,%ebp
(c) fopen64 uninitialized variable

Figure 4: Example code of uninitialized variable in whole program VSA

4.3.1 Must-not Tainted Analysis

In order to statically identify instructions never involved in taint analysis, we should know the must-not-tainted value set, which is an opposite, more conservative counter-part of the intuitive tainted value set, at each program point. This is also a data flow analysis problem, and we have to inspect each instruction to decide whether its operand will never be involved in taint or not.

Identification Policy. Must-not-tainted set is based on the following policy: (1) instructions unreachable from taint sources are removed from the must-not-tainted set (which is one of the differences compared to DBI-based taint implementations), e.g., in Figure 5a, the instruction at 80491b7, which is at the beginning of the program, is removed from must-not-tainted set as 804b4e1 is the first instruction that introduces the taint; (2) instructions with operands of potentially tainted or unknown value sets are removed from must-not-tainted set such as the instruction at 8055c41 that may contain tainted data from __IO_getc function return value in Figure 5b; (3) instructions whose operands hold literal values are added to must-not-tainted set since none of the operands will be tainted, e.g., instruction inc %ebp whose operand contains a literal value in register ebp as shown in Figure 5c is added to must-not-tainted set; (4) instructions whose opcode indicates they will not be involved in taint propagation are added to must-not-tainted set, e.g., control-flow instructions (e.g., jmp in Figure 5d) and compare and test instructions (e.g., cmp, and test). The must-not tainted value set will propagate along with data flow, and it is a must analysis.

80491b7: mov %eax,0x8052160
...
804b4e1: call 8048d70 <read@plt>
(a) Unreachable instructions

8055c3c: call 8048f30 <_IO_getc@plt>
8055c41: mov %eax,0edx
(b) Potentially tainted instructions

8096a07: inc %ebp
(e) Untainted operand instructions

8062456: jmp 806238b <mbslen+0x8b>
(d) None taint-propagation instructions

Figure 5: Example code of the corresponding identification policy

Resolving operand’s addresses. To conservatively track the must-not tainted value sets, we have to look into different types of memory access of an instruction operand: (1) for constant memory address, e.g., [0x8000200], we can easily infer that it is a global variable rather than a local variable or a heap variable and the must-not tainted value sets of that address can be updated based on the constant memory address, e.g., if this constant memory address may be tainted, the constant memory address would be removed from our must-not-tainted set; (2) for a memory access based on ESP register, which we call stack pointer addressing, e.g., [esp + 0x4], we identify it as a stack variable, the stack region and offset can be obtained through our whole program analysis caller/callee stack information; (3) for a memory access without ESP register, e.g., [eax], this is tricky since we may not know whether it is a stack, global or heap variable; we thus use the VSA result to decide the value set of the memory access: if the VSA cannot decide whether the memory access address is tainted or not, we conservatively remove it from the must-not tainted set.

Resolving operand’s values. Once the algorithm meets an instruction operand that is uninitialized (it can lead to an
alias that cannot be resolved), as mentioned in §4.2, we conservatively taint the associated variables, depending on the specific cases:

(i) **Unknown value from CLI** (e.g., Figure 4a). Based on where the input value is going to be stored, we assign a corresponding uninitialized value for these variables. For instance, we will assign an uninitialized value for a stack variable which belongs to the stack of the caller of main and is prior to the stack of main function, and proceed the must-not analysis as usual.

(ii) **Unknown value passed from missing callers** (Figure 4b). A caller function passes function parameters to a callee function, causing aliasing between actual parameters and formal parameters. When CFG reconstruction cannot determine the caller for a callee, it results in unknown value from the missing callers. We conservatively remove all of the memory access instructions in the function and all of the data uses of these variables outside the function from the must-not tainted set. To optimize our analysis, we do not taint all global variables, and instead we taint the data based on their types (the type inference is described below) in global sections such as .data and .bss.

(iii) **Unknown value due to library function calls and system calls** (Figure 4c). We taint these unknown variables according to the semantics of library functions and system calls. For instance, the pointer returned by fopen is put in the must-not tainted set at the program point right after the library call and the pointer returned is assigned a value set in a special heap region.

**Variable type inference.** To taint instructions more precisely, we perform a simple variable type inference to determine whether a variable is a pointer or not. We care them because we want to identify the potential pointers that can hold the tainted buffer. The analysis is based on whether a variable is dereferenced or whether it is a pointer type parameter or return value of known library functions as type sinks [20]. For instance, movzbl (%ebx),%eax indicates the variable stored at ebx is a pointer, and also variable stored at edi in the following snippet is a pointer as it is passed to the first parameter of strchr library function. With variable type inference, we could only taint pointer variable of interest when an unknown pointer is dereferenced instead of tainting all variables.

```
movl 0x0a,0x4(%esp)
movl %edi,(%esp)
call <strchr@plt>
```

**Our Algorithm.** Specifically, the must-not tainted analysis algorithm as shown in algorithm 2 first scans the whole binary for possible taint sources, e.g., read system call and recv system call (line 2). Each identified taint source serves as a starting point of our analysis. The initial tainted buffer has two major characteristics: start address and length. As our evaluation shows, we are able to identify the value set of the starting address and length of the buffer that introduces the taint. Otherwise, if either upper bound or lower bound of buffer’s starting address or buffer length cannot be determined, our analysis triggers a warning and terminates, since it may indicate program vulnerability (line 4-5). The analysis is of a work-list style and iterates over each instruction until the UntaintedSet and TaintedInst remain unchanged (reached a fixed point). For each instruction i in the program (line 7-10), we first compare the incoming value sets of instructions and formal parameters. When CFG reconstruction cannot determine the callers for a callee, it proceeds the must-not analysis as usual.

```
Algorithm 2: Must-not Tainted Analysis
1 Function MustNotTainted(UntaintedSet, TaintedInst, ValueSet);
2 input : set of must-not tainted data object UntaintedSet, set of tainted instructions TaintedInst, value set ValueSet
3 output : set of tainted data object UntaintedSet, set of tainted instructions TaintedInst
4 Source ← TaintSourceScan(Bin);
5 Init(buffer_start_addr, buffer_length, ValueSet, Source);
6 if unbounded([UntaintedSet BUFFER] ∨ unbounded([ValueSet BUFFER] [buffer_length])); then
7 exit();
8 while changed do
9   foreach instruction i do
10      if ValueSet ENTRY [opaddr] ⊆ UntaintedSet then
11         TaintedInst ← TaintedInst ∪ {i};
12         Transfer(UntaintedSet, ValueSet);
13      end if
14   end for each instruction
15   TaintSourceScan(Bin);
16   if ValueSet ENTRY [opaddr] ⊆ UntaintedSet then
17      UntaintedSet ← UntaintedSet − ValueSet ENTRY [opaddr] ;
18      if ValueSet ENTRY [opaddr] ⊆ UntaintedSet ∧
19         evalToConcrete(ValueSet ENTRY [opaddr]) then
20         UntaintedSet ← UntaintedSet −
21         ValueSet ENTRY [opaddr] ;
22      end if
23   end if
24 end while changed do
25 end Function
```

we taint the destination operand and remove it from UntaintedSet as shown in line 16. If the source operand
is deemed untainted and we know the exact concrete address of the destination operand, we enlarge our UntaintedSet by adding destination operand value set to UntaintedSet as illustrated in line 17-18. Otherwise, we conservatively taint all of the possible memory address involved in instruction $i$.

**Example.** We use the instruction 0x8048634: `mov %al, −0x44(%ebp)` listed in Table 1 to demonstrate how to use our propagation rules to update the must-not tainted set. Specifically, since the source operand of instruction 0x8048634 is not in UntaintedSet, this instruction is added to TaintedInst for further taint propagation logic instrumentation and since the taint propagation rule for this instructions is in the form of:

$$\text{tag}(op_{dest}) \leftarrow \text{tag}(op_{dest}) | \text{tag}(op_{src})$$

According to algorithm 2, UntaintedSet is updated:

- UntaintedSet = UntaintedSet $-$ UntaintedSet
- UntaintedSet = UntaintedSet $-$ UntaintedSet

### 4.3.2 Soundness Analysis of SelectiveTaint

We define **false negatives** to be the instructions that SelectiveTaint considers to not be tainted but the ground truth indicates it should, and **false positives** to be the instructions that SelectiveTaint considers to be tainted but the ground truth indicates it should not. Therefore, we can afford to have false positives, since the overly tainted instructions will only impact the performance overhead (we could have avoided tainting them). In contrast, we should avoid having false negatives; otherwise, we could miss the attacks if we do not taint these instructions.

In the following, we provide a formal analysis of our must-not tainted analysis to prove that it is soundy [21] (mostly sound), i.e., it hardly introduces false negatives, which means all instructions in the must-not-tainted instruction set $\mathcal{I}_u$ generated by our must-not tainted analysis are indeed not-tainted. It is soundy, as there could be the imprecise CFG and VSA in practice.

**Figure 6** shows the formal representation of must-not tainted analysis. Basically, for removing or adding an instruction $i$ in $\mathcal{I}_u$, one of the four primary inference rules, i.e., rule Unreachable, UnknownOperand, UntaintedOperand, or NonPropagateOpcode has to be applied:

- In **Unreachable** rule, if there is no path from taint sources to instruction $i$ and no path from instruction $i$ to taint sources, then $i$ is removed from the must-not-tainted instruction set $\mathcal{I}_u$;
- In **UnknownOperand** rule, if there exists an operand with unknown value set, then $i$ is removed from the must-not-tainted instruction set $\mathcal{I}_u$;
- In **UntaintedOperand** rule, if for each operand $o$ of instruction $i$, its value set is a subset of must-not-tainted value set $\mathcal{V}_u$, then $i$ is added to the must-not-tainted instruction set $\mathcal{I}_u$;
- In **NonPropagateOpcode** rule, if for each operand $o$ of instruction $i$, there is no side effect on operand $o$ after executing $i$, then $i$ is added to the must-not-tainted instruction set $\mathcal{I}_u$.

The rest rules of **Figure 6** are auxiliary inference rules. Rule **Reachable** indicates, if instruction $i_2$ is a successor of instruction $i_1$, then $i_2$ is reachable from $i_1$. Similarly, **TransReachable** rule indicates, if instruction $i_3$ is a successor of instruction $i_2$ and instruction $i_2$ is a successor of instruction $i_1$, $i_3$ is reachable from $i_1$. Note the succ predicate in TransReachable rule includes the control flow transfer for fall-throughs, unconditional jumps, conditional jumps, calls, returns, and other control flow transfer which is represented as an edge in the whole program control flow graph. **LiteralOperand** rule indicates, if the operand value set has a type of literal, it is added to must-not-tainted value set $\mathcal{V}_u$ and **LabelOperand** rule indicates, if the operand’s value set has a type of label, it is added to must-not-tainted value set $\mathcal{V}_u$. **TaintSource** and **TaintPropagate** rules infer that the tainted value set is removed from must-not-tainted value set $\mathcal{V}_u$ and when taint propagates from source operand to destination operand of an instruction, the destination operand is also tainted. **PCRegChangeOpcode** and **StatusRegChangeOpcode** rules indicate that, after executing instruction $i$, the value sets of instruction operands are not changed and only the value sets of program counter or status registers are changed, then instruction $i$ is added to must-not-tainted instruction set $\mathcal{I}_u$.

**Theorem 1.** Must-not-tainted analysis is sound, except for the precision loss due to imprecise CFG and VSA results (thereby making it soundy or mostly sound).

**Proof.** We prove this theorem with induction.

1. In the first iteration of the analysis, the must-not-tainted set $\mathcal{I}_u$ is $\varnothing$. Must-not-tainted analysis is sound since every instruction is taint.

2. We next prove that if in the $k^{th}$ iteration, the must-not-tainted analysis is sound, w.r.t. precise CFG and VSA results, it also holds in $(k + 1)^{th}$ iteration.

Suppose the must-not-tainted set $\mathcal{I}_u$ in the $k^{th}$ and $(k + 1)^{th}$ iteration are $\mathcal{I}_u^k$ and $\mathcal{I}_u^{k+1}$, respectively. Given any instruction $i$, whether instruction $i$ is added to or removed from the must-not-tainted instruction set $\mathcal{I}_u^{k}$ has four cases:

1.1) **Instruction $i$ cannot be reached from taint sources and taint sources cannot be reached from instruction $i$**. The Unreachable rule will remove $i$ from $\mathcal{I}_u^k$. In this case, no paths lies between instruction $i$ and taint sources, which indicates incomplete CFG and thus instruction $i$ can be potentially tainted, and therefore safely removing $i$ from $\mathcal{I}_u^k$ will result in a sound $\mathcal{I}_u^{k+1}$.

**CFG imprecision.** As reconstructing CFG is a hard problem in practice, case 2.1 is sound except for the imprecise CFG.
Primary Inference Rules

Instructions:

\[ \exists l \in \text{source}, i \sim i, i \sim i \quad \text{UNREACHABLE} \]

\[ \forall o \in \text{op}(i), V[o] \subseteq V_a \quad \text{UNTANTEDOperand} \]

\[ \exists o \in \text{op}(i), V[o] = (\perp, \perp, \perp) \quad \text{UNKNOWNOperand} \]

\[ \forall o \in \text{op}(i), V[o] \not\equiv V[o] \quad \text{NONPROPAgateOpcode} \]

Auxiliary Inference Rules

Control-flows:

\[ \text{REACHABLE} \quad \text{TRANSREACHABLE} \]

Operands:

\[ \text{LITERALOperand} \quad \text{LABELOperand} \]

\[ \text{TaintSource} \quad \text{TaintPropagate} \]

Opcodes:

\[ \text{PCRegChangeOpcode} \quad \text{StatusRegChangeOpcode} \]

Figure 6: The formal inference rules of our must-not tainted analysis. \( V_a \): Must-not-tainted value set, \( \mathcal{I}_u \): Must-not-tainted instructions, \( \equiv \): Equal values after executing \( i \), \( V \): VSA result map.

As shown in §4.1, SELECTivetaint matches callers and callees based on forward-edge CFI identification approaches and matches jumps and jump targets with basic block starting addresses within the same function or function entry addresses. These methods may introduce false negatives and produce imprecise CFG for real-world binaries.

(2.2) \textbf{One or more operands of instruction i have an unknown value set.} The UNKNOWNOperand rule will remove \( i \) from \( \mathcal{I}_u^{k+1} \). In this case, instruction \( i \) can propagate taints, and therefore safely removing \( i \) from \( \mathcal{I}_u^k \) will result in a sound \( \mathcal{I}_u^{k+1} \).

\textbf{VSA imprecision.} Though VSA may introduce imprecision, this rule conservatively removes all instructions with unknown value sets from \( \mathcal{I}_u^k \).

(2.3) \textbf{All operands of instruction i are subsets of must-not-tainted value set \( V_a^k \).} \( V_a^k \) is updated based on rules in the Operands rule group in Figure 6. Per rule LITERALOperand and LABELOperand, if an operand is of type literal or label, its value cannot propagate taint and it is added to \( V_a^k \). Per rule TaintSource and TaintPropagate, at taint source and taint propagation instructions, \( V_a^k \) gets updated by removing the tainted value set. If all operands of instruction \( i \) are subsets of must-not-tainted value set \( V_a^k \), it means all values in the operands are must-not-tainted, and it is sound after adding it to \( \mathcal{I}_u^{k+1} \) from \( \mathcal{I}_u^k \).

\textbf{VSA imprecision}. As VSA is an undecidable problem, it may introduce imprecision when VSA fails to identify whether an operand is actually a subset of \( V_a^k \) and when \( V_a^k \) is updated. Thus, in practice, this rule leads to a soundy analysis, i.e., mostly sound except for the imprecision caused by imprecise VSA.

(2.4) \textbf{Instruction opcode has no impact on taint propagation.} In this case, instruction \( i \) should be added to \( \mathcal{I}_u \), as the instruction does not involve in any explicit handling of tainted data. Particularly, an instruction may only have side effects on program counter or status register but not its operands, and in these cases no taint is involved and instruction \( i \) should be added to \( \mathcal{I}_u \), as in rules PCRegChangeOpcode and StatusRegChangeOpcode, which results in a sound \( \mathcal{I}_u^{k+1} \).

Therefore, must-not-tainted analysis of SELECTivetaint is soundy, due to the imprecision introduced by current limitations of undecidable CFG reconstruction and VSA results in binary (otherwise, it is sound).

4.4 Binary Rewriting

Based on the identified tainted instructions, we instrument taint propagation logic via binary rewriting for each instruction just as how a conventional DFT performs. The only
We have implemented an open source version of SELECTIVE-TAINT atop angr [33] and Dyninst [7]. Specifically, (1) we used angr to build a CFG for the binary, then implemented our own forward-edge control flow target identification based on TypeArmor [35] and tCFI [26], i.e., using our own VSA to determine unsolved call sites targets, connecting unsolved call sites to functions with the same parameter count and parameter type, and connecting unsolved indirect jumps with basic block starting address or all function entry addresses; (2) we implemented our own flow-sensitive and context-sensitive whole program VSA, which is used to determine the value set held at each program point; (3) based on the generated CFG and VSA results, we implemented taint instruction identification using the rules described in Figure 6 to identify the untainted instructions; and (4) after that we use Dyninst, which is widely used in recent studies [27, 35], to statically rewrite the binary. The total implementation of SELECTIVE-TAINT consists of 7,000 python code, and 22,000 C/C++ code. SELECTIVE-TAINT is tested in Ubuntu 14.04 32 bit OS to be compatible with the legacy 32 bit version libdft.

**Algorithm 3: SELECTIVE-TAINT Algorithm**

```plaintext
Algorithm 3: SELECTIVE-TAINT Algorithm
1 Function SelectiveTaint(Bin):
   input : original binary Bin
   output : instrumented binary NewBin
2   Init (UntaintedSet, TaintedInsn, ValueSet)
3   while changed do
4      CFG ← CfgReconstruction(Bin, CFG, ValueSet)
5      ValueSet ← whole_program_VSA(CFG, ValueSet)
6      UntaintedSet, TaintedInsn ← MustNotTainted(UntaintedSet, TaintedInsn, ValueSet)
7      NewBin ← Rewriting (Bin, TaintedInsn)
```

With the support from our CFG reconstruction, value set analysis, and taint instruction identification, we then sequentially combine these three analyses in a loop body and the set of ValueSet, UntaintedSet and TaintedInsn get gradually changed until a fixed point is reached. In particular, as illustrated in algorithm 3, at line 2, we first initialize the three sets of ValueSet, UntaintedSet and TaintedInsn with the taint source information. We will reach a fixed point after iterations of sequentially applying CFG reconstruction, value set analysis and taint instruction selection algorithm (line 3-6). When all the taint sources are processed, at line 7, our binary rewriter rewrites the original binary with taint propagation logic on selected instructions.

Note that for performance reasons, we use a function summary approach to process standard libraries such as libc, which is inspired by how RAMBLR [36] handled libraries. That is, we will not statically rewrite the instructions in the library, and instead we rewrite the callers to perform direct taint tracking, e.g., introducing taint, and propagating taint according to the corresponding parameters. For instance, when we notice memcpy call, we will directly taint the destination memory based on the data in the source memory.

**5 Implementation**

In intra-procedural analysis, the value set for each a-loc is calculated in a worklist algorithm until a fixed point is reached. In inter-procedural analysis, a function summary is generated based on intra-procedural analysis results to summarize the value set changes of each function. The static analysis finishes when all value sets in the whole program remain unchanged.

**Taint Instruction Selection.** We examine and maintain a must-not tainted a-loc set for each program point based on the value sets generated by VSA and the type of each instruction generated by CAPSTONE [2] disassembler. When must-not tainted a-loc sets reach a fixed-point, each instruction is examined as tainted or untainted based rules in Figure 6, i.e., we conservatively assume an a-loc is tainted whenever we cannot determine its taintedness. Unlike libdft which is implemented using Pin [22], we do not go into the dynamic library functions, and instead, we use a function summary for each library functions to track taint propagation.

**Binary Rewriting.** Our rewriter is implemented with a bit tag size and a byte tag granularity using Dyninst [7] binary instrumentation and analysis framework. Dyninst is an anywhere, anytime binary instrumentation framework which could be used in both static binary rewriting at compile-time or dynamic instrumentation at run-time. We favor Dyninst as it is a the-state-of-the-art tool in binary rewriting which is used in a variety of tools and its robust API implementation.

**CFG Reconstruction.** To implement the analyzer, first, we recover the control flow graph (CFG) of the binary using angr. Basically, we use angr to find every basic block address, its containing instructions and the predecessors and successors of each basic block. Afterwards, the remaining unsolved indirect control flow transfer targets are further resolved using our method described in §4.1.

**Value Set Analysis.** We first initialize the corresponding ValueSet with the data extracted from original binary, e.g., section and segment information, initial data values in .rodata and .data sections. With respect to the variables in different memory region: for (1) stack variable, we track the value set of stack pointer SP in different calling context, examine and identify whether a variable is a stack variable and in which function the variable is defined, i.e., in whose stack frame the variable resides; for (2) global variable, we track the value set of variables and check if it could be evaluated to an address in code segment or data segment as a global variable; for (3) heap variable, we track the call instructions for malloc-family library calls to determine whether it is a heap variable.
### Table 2: Statistics of the instrumented instructions by SELECTIVEAINT and libdft

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Input Functions</th>
<th># Func.</th>
<th># Inst.</th>
<th>SELECTIVEAINT</th>
<th>libdft</th>
<th>Executed Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Utilities</td>
<td></td>
<td></td>
<td>Instrument. Inst. (%)</td>
<td>Tainted Inst. (%)</td>
<td></td>
</tr>
<tr>
<td>tar</td>
<td>read, fscnf</td>
<td>967</td>
<td>65,795</td>
<td>45,630 (69.35%)</td>
<td>4,083 (6.21%)</td>
<td>688</td>
</tr>
<tr>
<td>gzip</td>
<td>read, _IO_getc</td>
<td>220</td>
<td>15,173</td>
<td>10,076 (66.41%)</td>
<td>2,067 (13.62%)</td>
<td>40</td>
</tr>
<tr>
<td>bzip2</td>
<td>fread, fgetc</td>
<td>127</td>
<td>16,195</td>
<td>11,160 (68.91%)</td>
<td>3,524 (2.18%)</td>
<td>51</td>
</tr>
<tr>
<td>scp</td>
<td>read</td>
<td>145</td>
<td>6,390</td>
<td>4,238 (66.32%)</td>
<td>1,875 (29.34%)</td>
<td>10</td>
</tr>
<tr>
<td>cat</td>
<td>read, fscnf</td>
<td>174</td>
<td>8,003</td>
<td>5,366 (67.05%)</td>
<td>548 (6.85%)</td>
<td>19</td>
</tr>
<tr>
<td>comm</td>
<td>fscnf</td>
<td>150</td>
<td>4,659</td>
<td>3,254 (69.84%)</td>
<td>918 (19.70%)</td>
<td>9</td>
</tr>
<tr>
<td>cut</td>
<td>fgetc, fscnf, __fread_chk</td>
<td>181</td>
<td>6,587</td>
<td>4,343 (65.93%)</td>
<td>742 (11.26%)</td>
<td>13</td>
</tr>
<tr>
<td>grep</td>
<td>read, fscnf, fread_unlocked</td>
<td>410</td>
<td>28,858</td>
<td>19,500 (67.57%)</td>
<td>3,693 (12.80%)</td>
<td>129</td>
</tr>
<tr>
<td>head</td>
<td>read, fscnf</td>
<td>149</td>
<td>6,533</td>
<td>4,517 (69.14%)</td>
<td>497 (7.61%)</td>
<td>13</td>
</tr>
<tr>
<td>nl</td>
<td>fscnf</td>
<td>252</td>
<td>20,677</td>
<td>14,082 (68.10%)</td>
<td>684 (3.31%)</td>
<td>77</td>
</tr>
<tr>
<td>od</td>
<td>fgetc, fscnf, fread_unlocked, __fread_unlocked_chk</td>
<td>188</td>
<td>10,696</td>
<td>7,143 (66.78%)</td>
<td>1,640 (15.33%)</td>
<td>24</td>
</tr>
<tr>
<td>ptx</td>
<td>fread, fscnf</td>
<td>297</td>
<td>25,906</td>
<td>17,503 (67.56%)</td>
<td>5,478 (21.15%)</td>
<td>106</td>
</tr>
<tr>
<td>shred</td>
<td>fscnf, __readchk, fread_unlocked</td>
<td>198</td>
<td>9,195</td>
<td>6,404 (66.65%)</td>
<td>1,673 (18.19%)</td>
<td>21</td>
</tr>
<tr>
<td>tail</td>
<td>read, fscnf</td>
<td>216</td>
<td>10,966</td>
<td>7,251 (66.12%)</td>
<td>825 (7.52%)</td>
<td>26</td>
</tr>
<tr>
<td>truncate</td>
<td>fscnf</td>
<td>168</td>
<td>8,952</td>
<td>6,006 (67.09%)</td>
<td>491 (5.48%)</td>
<td>22</td>
</tr>
<tr>
<td>uniq</td>
<td>fscnf</td>
<td>165</td>
<td>5,539</td>
<td>3,822 (69.00%)</td>
<td>815 (14.71%)</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network daemons</th>
<th></th>
<th></th>
<th></th>
<th>SELECTIVEAINT</th>
<th>libdft</th>
<th>Executed Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Utilities</td>
<td></td>
<td></td>
<td>Instrument. Inst. (%)</td>
<td>Tainted Inst. (%)</td>
<td></td>
</tr>
<tr>
<td>exim</td>
<td>fgetc, fread, fscnf, __IO_getc, recv</td>
<td>876</td>
<td>140,847</td>
<td>93,058 (66.07%)</td>
<td>8,160 (5.79%)</td>
<td>2,843</td>
</tr>
<tr>
<td>memcached</td>
<td>read, fget, recvfrom</td>
<td>286</td>
<td>19,319</td>
<td>13,676 (70.79%)</td>
<td>852 (4.41%)</td>
<td>62</td>
</tr>
<tr>
<td>proftpd</td>
<td>read, fget, __read_chk</td>
<td>1,037</td>
<td>153,306</td>
<td>106,821 (69.68%)</td>
<td>19,181 (12.51%)</td>
<td>3,048</td>
</tr>
<tr>
<td>lighttpd</td>
<td>read, fread</td>
<td>466</td>
<td>31,130</td>
<td>21,713 (69.75%)</td>
<td>5,437 (17.47%)</td>
<td>156</td>
</tr>
<tr>
<td>nginx server</td>
<td>read, pread64, readv, recv</td>
<td>1,277</td>
<td>133,666</td>
<td>92,041 (68.86%)</td>
<td>12,905 (9.65%)</td>
<td>2,249</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other applications</th>
<th></th>
<th></th>
<th></th>
<th>SELECTIVEAINT</th>
<th>libdft</th>
<th>Executed Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoX 14.4.2</td>
<td>read, fread, fscnf, __IO_getc, recv</td>
<td>1,159</td>
<td>112,762</td>
<td>67,808 (60.13%)</td>
<td>9,583 (8.50%)</td>
<td>1,791</td>
</tr>
<tr>
<td>TinTin++ 2.01.6</td>
<td>read, fread, fget, gnutls_record_recv, fget</td>
<td>831</td>
<td>93,618</td>
<td>74,389 (79.46%)</td>
<td>9,839 (10.51%)</td>
<td>975</td>
</tr>
<tr>
<td>dcrw 9.28</td>
<td>fread, fscnf, __read_chk, __IO_getc, fget, jpeg_read_header</td>
<td>292</td>
<td>70,358</td>
<td>42,840 (60.89%)</td>
<td>497 (13.57%)</td>
<td>510</td>
</tr>
<tr>
<td>ngiflib 0.4</td>
<td>fread, __IO_getc</td>
<td>43</td>
<td>2,135</td>
<td>1,477 (69.18%)</td>
<td>649 (30.40%)</td>
<td>3</td>
</tr>
<tr>
<td>Gravity 0.3.5</td>
<td>read, getline</td>
<td>1,124</td>
<td>86,783</td>
<td>65,636 (75.63%)</td>
<td>16,959 (19.54%)</td>
<td>773</td>
</tr>
<tr>
<td>MP3Gain 1.5.2</td>
<td>fread, _IO_getc</td>
<td>144</td>
<td>17,573</td>
<td>9,900 (56.34%)</td>
<td>5,934 (33.77%)</td>
<td>53</td>
</tr>
<tr>
<td>NASM 2.14.02</td>
<td>fread, fget, jpeg_read_header</td>
<td>826</td>
<td>87,456</td>
<td>66,601 (76.15%)</td>
<td>7,276 (8.32%)</td>
<td>838</td>
</tr>
<tr>
<td>Jhead 3.00</td>
<td>fread, fgetc</td>
<td>118</td>
<td>9,815</td>
<td>6,805 (69.33%)</td>
<td>621 (6.33%)</td>
<td>19</td>
</tr>
</tbody>
</table>

### 6 Evaluation

In this section, we present the evaluation results. To see the improvements over dynamic taint analysis, we compare SELECTIVEAINT with libdft [17]. The version of Intel Pin used to build libdft was 2.14 (build 71313), and we slightly modified the nullpin and libdft tool and adopted them in our experiment settings. Also, to see the advancements of the selective taint analysis, we also implemented a static taint analysis tool and adopted them in our experiment settings. We use four commonly used Unix utilities, i.e., the GNU versions of tar (version 1.27.1), gzip (version 1.3.13), bzip2 (version 1.0.3), scp from OpenSSH (version 3.8), 12 file content processing utilities cat, comm, cut, grep, head, nl, od, ptx, shred, tail, truncate, uniq from coreutils (version 8.21) and grep (version 2.16), and we also use email server exim (version 4.80), general-purpose distributed memory caching system Memcached (version 1.4.20), FTP server ProFTPD (version 1.3.5), web server lighttpd (version 1.4.35) and nginx (version 1.4.0), and eight recent programs (each of which contains a memory corruption vulnerability) to evaluate SELECTIVEAINT. We first evaluate its effectiveness by looking into the details of how SELECTIVEAINT performs in §6.1, and then report the performance overhead of the rewritten binaries in §6.2. Finally, we demonstrate its security applications with real world binaries in §6.3.

#### 6.1 Effectiveness

We report the effectiveness of how SELECTIVEAINT performs with the common Unix utilities tar, gzip, bzip2, scp, cat, comm, cut, grep, head, nl, od, ptx, shred, tail, truncate, uniq, network daemons exim, memcached, proftpd, lighttpd, nginx, and eight other applications in Table 2. The first column shows the 29 C/C++ programs in the benchmark we used in our evaluation, followed by the 2nd column of the input functions detected by SELECTIVEAINT. Note that the input function is the function that introduces the taint sources. Next, we report the total number of functions contained in the benchmark program in the 3rd column, which provides an estimation of the complexity of the program. Then, we show the total number of instructions identified in the binary in the 4th column. Our STATICTAINTALL statically rewrites all of these instructions, similarly to how dynamic taint analysis instruments them. This will provide an upper bound of how SELECTIVEAINT would perform in the worst case (by statically taint them all). Next, we show the total number of instructions that need to be statically
instrumented by SelectiveTaint in the 5th column followed by the total number of executed unique instructions that really involved in taint analysis in the 6th column, and this number is obtained by running the corresponding benchmark by using the default configured input with libdft, which will provide a lower bound of the number of unique tainted instructions. For fair comparison, we did not count the instructions in the library from the libdft trace since SelectiveTaint will not instrument them. Finally, we report how long SelectiveTaint performs to process each of the benchmarks in the last column.

We can observe from Table 2 that our static analysis works well in these benchmarks, and we have reduced the possible tainted instructions to about 56.34% - 79.46% compared to StaticTaintAll. While ideally we would like to instrument only the instruction involved in the taint analysis (which is a subset of the instructions identified by SelectiveTaint), as detected by the libdft which shows about 1.37% - 33.77% of these instructions are essentially needed in the taint analysis at run-time with an average of 6.85% instructions, we will not be able to achieve this by purely static analysis.

False Positives and False Negatives. We have defined false positives and false negatives in Section 4.3.2. By examining the instructions tainted by SelectiveTaint and libdft, we observe SelectiveTaint reports no false negative but false positives. False positives indicate SelectiveTaint is conservative and has over-tainted instructions, and such false positives are acceptable (it will not miss any attacks). Meanwhile, no false negative indicates our approach is a sound over-approximation of the tainted instructions. This is attributed to the conservative rules in Figure 6; for instance, we remove the value set from untainted value set whenever we cannot determine the taintedness of that value set. Note the 6th column in Table 2 is generated by running the tested benchmarks, which may not explore all instructions that should be tainted and thus the ground truth is unavailable and we only observe false positives without quantifying them.

Internal Statistics. We also measured the statistics of SelectiveTaint in Table 3 to understand its inner-workings. Columns 2-3 are CFG construction details, i.e., the number of initial CFG edges and the number of final CFG edges after our CFG construction. We can observe our CFG construction can add hundreds of edges to the CFG using the techniques described in §4.1. Columns 4-8 are value set analysis statistics, which are the number of a-locs in the analysis, the unknown a-locs due to command line parameters, argument aliasing when missing callers, and library function calls. We can observe our approach identifies a large number of unknown a-locs in each category. Columns 9-12 are numbers in taint instruction identification, such as the number of initially untainted value sets in the first iteration, the number of final untainted value sets, the intra-procedural iteration times, and the inter-procedural iteration times. We can observe that the number of untainted value sets get smaller through analysis iterations, which means our analysis does propagate untaintedness and remove potentially tainted value sets from the must-not-tainted set $U_t$. The intra-procedural analysis generally has hundreds of iterations while the inter-procedural analysis has significantly fewer iterations, from which we can observe the intra-procedural analysis reaches the fixed point with more iterations than inter-procedural analysis.

Performance. With respect to the performance (e.g., the analysis time) of SelectiveTaint itself, we notice it sometimes consumes tens of minutes to finish analyzing a program. This is understandable, since SelectiveTaint will inspect each instruction and calculate VSA for each of them. Meanwhile, the analysis has to be run twice: first calculating the VSA, and then determining the taintedness. We notice it took more than 50 minutes to process the proftpd FTP server, whereas for small binaries, e.g., scp, it could take just a few seconds.

6.2 Efficiency

Next, we measure the performance overhead of the rewritten binaries. To compare with libdft, we run the binaries with the default configured input, with nullpin (a simple implementation to evaluate Intel Pin platform overhead), and libdft with a bit level taint. We run the corresponding benchmark with and without rewriting to understand the additional overhead. All of the experimental results were obtained with 10 runs and then normalized by dividing each average result against native unmodified executables.

Unix Utilities. Figure 7a shows the normalized runtime overhead of nullpin, libdft, StaticTaintAll, SelectiveTaint, when running with 16 Unix Utilities, compared with the native execution. We can notice that libdft imposes a slowdown ranging from 1.39 (tar) to 5.86x (scp), whereas StaticTaintAll and SelectiveTaint impose 1.10x (tar) to 3.85x (bzip2) and 1.02x (tar) to 3.41x (grep), respectively. StaticTaintAll outperforms libdft in all benchmarks with 1.26x - 1.87x faster with an average of 1.53x and similarly SelectiveTaint performs even 1.36x - 2.41x faster than that of libdft with an average of 1.77x.

Network Daemons. One ideal use case for SelectiveTaint would be for the protection of network daemons. We thus use exim, memcached, proftpd, lighttpd, nginx as benchmarks to thoroughly evaluate its overhead. In particular, we tested nullpin, libdft, StaticTaintAll, and SelectiveTaint on these five daemons, in which exim is tested by sending email messages, memcached by getting values with keys from database, and ftp and http daemons via requesting files from the daemons. All four tools including libdft performs no more than 4x slowdown. The biggest slowdown for libdft is 3.76x (exim). StaticTaintAll imposes 1.25x - 2.23x slowdown and outperforms libdft by
1.10x-1.77x. **SELECTIVEAINT** performs even better with 1.12x-1.91x which outperforms libdf by 1.12x-2.08x.

### 6.3 Security Case Studies

**Protecting nginx web server.** To show that our tools could be used to detect real-world attacks, we first implemented a buffer overflow attack detector and used it to protect the nginx web server. To test its effectiveness, we generated an exploit based on the buffer overflow vulnerability CVE-2013-2028. By leveraging this vulnerability, an attacker could send a malformed request that triggers an integer signedness error which further causes a stack-based buffer overflow. This bug can be used in a denial-of-service attack or cause arbitrary code execution. Without any surprise, our **SELECTIVEAINT** detects the exploit at the `ret` instruction because the return value stored on the stack is tainted.

**Protecting other binaries against recent memory exploits.** We further tested eight recent real world software vulnerabilities from Common Vulnerabilities and Exposures (CVE), which are listed in Table 4. The collected vulnerabilities covered a broad range of software vulnerabilities, including buffer overflow vulnerability, double free vulnerability, and integer underflow vulnerability, which manifested in varied programs such as sound processing utilities SoX, programming language interpreter Gravity, and audio normalization software MP3Gain.

We implemented the corresponding exploits to compromise these vulnerabilities and validate whether **SELECTIVEAINT** is able to detect the attacks. For instance, to exploit CVE-2017-1000437 vulnerability in Gravity 0.3.5, we developed a malformed `gravity` normalization software source code file, which overflowed the program stack to rewrite the return address with payloads in source code file. To exploit CVE-2019-7629 vulnerability in TinTin++, we set up a simple game server with exploits that keep sending crafted message which overflowed the multiplayer online game client `TinTin++`. Then, the tested binaries were instrumented with **SELECTIVEAINT**. In all cases, **SELECTIVEAINT** successfully detects the exploits which shows **SELECTIVEAINT** can facilitate real world vulnerability detection in various software.

### 7 Limitations and Future Work

**Augmenting static analysis with dynamic information.** As static analysis lacks dynamic information, **SELECTIVEAINT** has unknown values from multiple sources as shown in §4.2 and also VSA is an over-approximation of the possible values...
Figure 7: The performance overhead of the tested benchmarks

![Graph showing performance overhead of tested benchmarks](image)

Table 4: The tested vulnerable software and their vulnerabilities

<table>
<thead>
<tr>
<th>Program</th>
<th>Category</th>
<th>Vulnerability</th>
<th>CVE ID</th>
<th>STATICTAINTALL</th>
<th>SELECTIVE TAINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoX 14.4.2</td>
<td>Sound Processing Utilities</td>
<td>Buffer Overflow</td>
<td>CVE-2019-8356</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TianTan++ 2.01.6</td>
<td>Multiplayer Online Game Client</td>
<td>Buffer Overflow</td>
<td>CVE-2019-7629</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>drcaw 9.2x</td>
<td>Raw Image Decoder</td>
<td>Buffer Overflow</td>
<td>CVE-2018-19655</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ngifbf 0.4</td>
<td>GIF Format Decoding Library</td>
<td>Buffer Overflow</td>
<td>CVE-2018-11575</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Gravity 0.3.5</td>
<td>Programming Language Interpreter</td>
<td>Buffer Overflow</td>
<td>CVE-2017-1000437</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MP3Gain 1.5.2</td>
<td>Audio Normalization Software</td>
<td>Buffer Overflow</td>
<td>CVE-2017-14411</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NASM 2.14.02</td>
<td>Assembler and Disassembler</td>
<td>Double Free</td>
<td>CVE-2019-8343</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Jhead 3.00</td>
<td>Exif Jpeg Header Manipulation Tool</td>
<td>Integer Underflow</td>
<td>CVE-2018-6612</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Nginx 1.4.0</td>
<td>Web Server</td>
<td>Buffer Overflow</td>
<td>CVE-2013-2028</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

8 Related Work

Dynamic Data Flow Tracking (DDFT). Over the past decades, many DDFT systems were built. DYTAN [11] is one of the first such a tool that allows customized taint analysis, and it can also track implicit information flows due to control dependences. However, its performance overhead can be as high as 50x when performing dynamic taint analysis with both control- and data-flow based propagation and 30x for data-flow based propagation alone. Saxena et al. [32] propose
a static technique that recovers higher level structures from x86 binaries and apply it to the context of taint tracking. Unlike our approach, their stack analysis trades off analysis accuracy over performance, e.g., stack analysis ignores global and heap memory while VSA tracks global and heap memory. Also, Saxena et al.’s analysis mainly enables optimizations such as tag-sharing, but our analysis is the first static taint analysis to selectively instrument instructions. MINEMU [8] aims to design efficient emulator, using new memory layout and SSE registers to improve the taint analysis. Being a highly-optimized DDFT framework, libdft [17] shows a faster performance than previous efforts; for instance, libdft imposes about 4x slowdown for gzip when compressing or decompressing files.

Recent efforts [5, 15, 16, 24, 25] further improve the performance overhead of DDFT. For instance, Jee et al. propose Taint Flow Algebra (TFA) [16] that separates the program logic from data flow tracking, transforms the data flow tracking logic into an intermediate representation, and then performs classic compiler optimizations. SHADOWREPLICA [15] runs DFT in parallel in a shadow thread for each application thread and uses an off-line application analysis phase which leverages both static and dynamic analysis approaches to minimize the information needed to communicate between both threads. TAINTPIPE [25] uses pipelined symbolic taint analysis that both parallelizes and pipelines taint analysis. STRAIGHTTAINT [24] logs control flow profiling and execution state when taint seeds were first introduced. Most recently, IODINE [5] uses an optimistic hybrid analysis which restricts predicated static analysis to elide a runtime monitor only when it is proven to be a safe elision. Different from our approach, Iodine is built atop LLVM IR, which requires source code of the target application, whereas SELECTIVE TAINT is a binary only approach.

**Binary Rewriting.** There is also a large body of work on static binary rewriting. Most recent efforts include UROBOROS [37], RAMBLR [36], MULTIVERSE [6], probabilistic disassembly [23], and DDISASM [13]. UROBOROS [37] is a tool which repeatedly dissembles the executable such that the generated code could be reassembled back to working binaries. RAMBLR [36] further analyzes the assumptions of UROBOROS and finds multiple complex corner cases that must be considered in symbolization. RAMBLR applies advanced static analyses, e.g., VSA, and achieves great performance for a static rewriter. MULTIVERSE [6] is the first static binary rewriting tool that systematically rewrites x86 COTS binaries without heuristics. Probabilistic disassembly [23] uses probabilities to model the uncertainty caused by the information loss during compilation and assembling. Features such as data flow and control flow features are leveraged to compute a probability for each address in the code space to indicate the likelihood to be an instruction. DDISASM [13] combines static analysis and heuristics in Datalog and shows that Datalog’s inference process suits for the disassembling. In our implementation, while we could have used the most recent work such as MULTIVERSE, we use Dyninst [7] instead due to its rich APIs.

**Alias Analysis on Binary.** Prior efforts on binary alias analysis either introduce an IR and use Datalog to reason about points-to relations [9], or introduce sets for values held at each program point (e.g., abstract address sets [12], or symbolic value sets [1]). The alias relation of two variables is determined by whether the abstraction sets of these two variables intersect, e.g., intersection of abstract address sets [12], symbolic value sets [1], and points-to predicates results [9], respectively. Also, a number of earlier efforts (e.g., [1, 12]) do not further resolve indirect jumps in CFG and reconstruct more CFG edges and yet this limits the analysis precision. They also assume no system calls. However, system calls may introduce uninitialized value sets into the system, and the work by Debray et al. [12] uses less general sets of values, which is residue-based (module k), whereas we use all possible values.

9 Conclusion

We have presented an efficient static analysis based data flow tracking framework SELECTIVE TAINT. Unlike previous taint analysis that uses dynamic binary instrumentation, SELECTIVE TAINT is built atop static binary rewriting. The key insight is to use VSA to identify the instructions that never involve taint analysis, and then rewrite the rest to implement the taint analysis. We have tested SELECTIVE TAINT with 29 binary programs including 16 Unix utilities, five network daemons, and eight vulnerable applications and observed a superior performance, which is 1.7x faster than that of the state of the art dynamic taint analysis tools.

Acknowledgment

We are grateful to our shepherd Vasileios P. Kemerlis as well as the anonymous reviewers including those from the artifact evaluation committee for their very constructive feedback. We also would like to thank Haohuang Wen for his assistance during the evaluation. This research was supported in part by DARPA award N6600120C4020, NSF awards 1750809 and 1834215, and ONR award N00014-17-1-2995.

Availability

The source code of SELECTIVE TAINT and also the benchmark programs used during the evaluation have been made public available at https://github.com/OSUSecLab/SelectiveTaint.