Declarative Demand-Driven Reverse Engineering

Yihao Sun*, Jeffrey Ching†, and Kristopher Micinski‡
Department of Electrical Engineering and Computer Science, Syracuse University
Email: *ysun67@syr.edu, †eching01@syr.edu, ‡kkmicins@syr.edu

Abstract—Binary reverse engineering is a challenging task because it often necessitates reasoning using both domain-specific knowledge (e.g., understanding entrypoint idioms common to an ABI) and logical inference (e.g., reconstructing interprocedural control flow). To help perform these tasks, reverse engineers often use toolkits (such as IDA Pro or Ghidra) that allow them to interactively explicate properties of binaries. We argue that deductive databases serve as a natural abstraction for interfacing between visualization-based binary analysis tools and high-performance logical inference engines that compute facts about binaries. In this paper, we present a vision for the future in which reverse engineers use a visualization-based tool to understand binaries while simultaneously querying a logical-inference engine to perform arbitrarily-complex deductive inference tasks. We call our vision declarative demand-driven reverse engineering (D³RE for short), and sketch a formal semantics whose goal is to mediate interaction between a logical-inference engine (such as Soufflé) and a reverse engineering tool. We describe a prototype tool, d3re, which are using to explore the D³RE vision. While still a prototype, we have used d3re to reimplement several common querying tasks on binaries. Our evaluation demonstrates that d3re enables both better performance and more succinct implementation of these common RE tasks.

I. INTRODUCTION

Binary reverse engineering (henceforth RE) is the process by which we start with some input binary (sequence of bytes) and employ various reasoning principles to explicate its behavior when executed as code. While RE tasks are often partially automated (e.g., via decompilation), full automation is often impossible: the extreme semantic expressivity afforded to binaries (including encrypted code, stripped symbol tables, etc.) often necessitates open-ended exploration and case-specific reasoning. Recent literature suggests that many practitioners follow an iterative approach involving several rounds of hypothesis formation and validation/ falsification, often assissated via a combination of static and dynamic analysis [1]–[3].

To rapidly interact with a binary, RE practitioners often use reverse engineering tools such as Ghidra [4], IDA Pro [5], or Radare2 [6]. The goal of these tools is to allow an RE to quickly explore the binary and visualize it (typically interactively, via a GUI or CLI) in a variety of ways. For example, a reverse engineer looking for a time bomb may first search for calls to the system’s time function, and then walk backwards to understand whether each call is associated with legitimate or malicious behavior. In doing so, the RE may need to reason about, e.g., indirect control flow, or even identify the time function (in a stripped binary). Because REs are expert users, and often skilled programmers, RE tools provide programmatic interfaces that enable REs to systematize reasoning tasks via extensions. A broad range of popular extensions exist for several tools which perform such tasks as loading the results of static analyses [7], [8], interacting with debuggers [9], and identifying common cryptographic-relevant code [10].

In this paper, we argue that deductive databases (e.g., Datalog) serve as a natural abstraction boundary between RE tools and logical inference tasks over binaries. We envision a future in which a reverse engineer interactively explores a binary using an RE tool while simultaneously querying arbitrarily-complex logical properties written in a terse declarative style. We call this Declarative Demand-Driven Reverse Engineering (henceforth D³RE). In D³RE, an RE interacts with a deductive database by giving inputs (e.g., the currently highlighted address) to a rule-based deductive inference system written in a declarative language such as Datalog. Rules inductively compute relations over facts about the binary. As an example, consider a relation direct call \( \in \text{Addr} \times \text{Addr} \) which relates call site addresses (offsets within the binary) to procedure invocation target addresses. In our vision, D³RE allows REs to interactively compute with and visualize the results of queries over these declarative rules.

We see D³RE as a natural extension of several observations about the state of the art. First, many existing RE tools assemble databases to index various properties (e.g., addresses, symbols, etc.) of binaries for quick exploration. Deductive databases further allow REs to write arbitrary logical queries which are computed maximally efficiently via, e.g., compilation to relational algebra kernels as done in Soufflé. Deductive databases have also enabled several recent advances in binary analysis demonstrating both efficiency and robustness over conventional techniques. For example, the Datalog-based disassembler ddiasm achieves both faster and more-precise disassembly than other state-of-the-art disassemblers, and OOAnalyzer uses Prolog to enable declarative recovery of classes from compiled C++ code.

In this short paper we describe our progress in implementing a prototype tool, d3re, which we are building to realize the D³RE vision. d3re allows REs to interactively define and calculate queries of arbitrary complexity over large production binaries and then visualize their results using Ghidra. To implement d3re we have designed an interface, which we call the mediator, that sits between a traditional Datalog solver and an RE tool. We briefly formalize this interaction between the RE tool and logic solver in Section III, and go on to describe

1When unambiguous, we will use the term RE both to mean the process of reverse engineering and a reverse engineering practitioner.
our prototype Ghidra extension that enables visualizing the results of binary analyses in our tool. Using this formalism, we describe how d3re readily enables a broad range of binary analyses and sketch a vision for how we believe D³RE will prove to be a natural ergonomic for reverse engineering.

We have measured the robustness of d3re in several ways. We wanted to know whether d3re could truly live up to our vision of being a natural replacement for the kinds of scripts REs already use in their day-to-day work. To evaluate this, we reimplemented a set of currently-existing Ghidra scripts. We happily observed that d3re was not only an ergonomic advantage (allowing us to write succinct but obviously-correct queries) but also a performance advantage. For example, many Ghidra scripts play tricks to avoid unnecessarily complexity that would arise in a straightforward implementation, e.g., iterating over a set of functions in a loop to check a property resulting in super-linear complexity. In d3re, the Datalog solver was naturally able to compile and organize work in an optimal way. We discuss this and other results in Section IV. We conclude with a brief overview of related work and our outlook on future directions in Section V.

Specifically, we claim the following three contributions:

• A formalization of our metadatabase as a database of databases used to optimize subsequent invocations of the Datalog solver (Section III).

• A prototype tool, d3re, consisting of a server which wraps ddisasm with logic to enable chaining multiple subsequent calls via the metadatabase. Also included in d3re is an extension to the Ghidra RE toolkit to enable visualizing results computed using d3re.

• An evaluation of d3re on a set of benchmarks demonstrating positive initial results indicating that d3re could replace present-day binary analysis infrastructure (e.g., Ghidra scripts) and directly enable more efficient and succinct implementation.

II. OVERVIEW

In this section, we demonstrate the vision and application of D³RE by illustrating how a reverse engineer might explicate a vulnerability due to an uninitialized global variable. We consider a particular binary, CROMU00038, from DARPA's Cyber Grand Challenge which contains a function pointer which is uninitialized when an invalid flag is set in the metadata portion of an input file [11], [12]. We demonstrate how our prototype tool, d3re, can be used to build a declarative query to find uninitialized function entry points and visualize them within Ghidra. We do not claim that d3re can immediately or automatically discover vulnerabilities—in this section we try to focus on how its declarative reasoning instead enables rapidly exploring a binary to uncover some property.

The vulnerable segment of code is a use-before-definition bug shown in Figure 1. The swap_word function is initialized inside of the main function based on a value parsed in a TIFF header—if the flag does not match 0x4949 or 0x4d4d the function is left uninitialized and the call on line 17 crashes.

```c
// swap_short and swap_word only initialized
// within if
if (tiff_hdr->Byte_Order == 0x4949) {
    printf("Intel formatted integers\n");
    swap_word = intel_swap_word;
} else if (tiff_hdr->Byte_Order == 0x4d4d) {
    printf("Motorola formatted integers\n");
    swap_word = motorola_swap_word;
} #ifdef PATCHED
else {
    printf("Invalid header values\n");
    _terminate(-1);
} #endif
// might cause an uninitialized variable bug here
offset = swap_word(tiff_hdr->Offset_to_IFD);
```

Fig. 1: Uninitialized variable vulnerability in CROMU0038 source code

Fig. 2: d3re REPL session used in this overview.

*Loading the binary:* To begin an analysis of a binary, an RE will load the binary into a reverse engineering tool. In our current implementation of d3re, a user opens two processes simultaneously: a GUI-based instance of Ghidra, and a terminal running d3re’s REPL. The user can explore the binary using all of the normal features of Ghidra and use all of its conventional analyses (e.g., to recover entrypoints). However, d3re’s REPL communicates with Ghidra so that when d3re’s analysis finishes Ghidra’s views update as appropriate.

*Initial processing:* It is conventional that reverse engineering tools will apply a set of analyses to a binary to disassemble it and index various items such as entrypoints and callsites. In d3re, the user builds queries in Datalog starting from a large initial set of Datalog rules that build on top of ddisasm, a Datalog-based disassembly engine [13]. Analogously to the indexing and analysis operations provided by Ghidra (and other RE tools), d3re invokes ddisasm once to build an initial database.

Building on top of ddisasm was initially a strategic choice—ddisasm already includes facilities to parse object files and transform them into input databases in the style required by Soufflé. Initially, we extended ddisasm’s set of rules with additional user-specific queries—a slow process, as ddisasm can take several minutes to run on large binaries.
The RE tool (such as currentAddress, the currently-selected address) as input relations, perform logical inference, and leave their output in relations such as comment(addr,"vuln").

Like ddisasm, d3re uses the Soufflé Datalog engine to perform logical inference over binaries. Users of d3re can incrementally build up more rules in the interactive REPL (shown in Figure 2). Currently, our REPL allows loading rules by loading new files—we plan on adding direct support for new rules, along with error-reporting feedback soon.

Knowing there was an uninitialized global function pointer being used, a user of d3re might first define a set of relations to build up def-use-chains of global variables. Datalog code to implement these queries is illustrated in Figure 3. The last two rules build up a relation use_before_def_global(EA_used,GA,Name), which infers that at address EA_def, the global variable (at address GA) is defined and used at address EA_used at operand index Index. While this is a relatively coarse query, we envision the user could run the query on the binary to visualize a large answer set. In our setting, this can be done using the highlight or comment commands, which display the data marked to be highlighted by the most recent result computation.

Based on our definitions in Figure 3, we can define a relation for variables which are possibly used before they are defined. We demonstrate this in Figure 4. In the first rule, we say that if there is some usage of a global variable at some address, but in that address, we can’t find any definition related to it, then we will consider variable there as an uninitialized variable; The second clause says that for some usage of a global variable even if it has some definition associated with it, if that definition is nullptr, we will still consider that there is a use-before-def vulnerability here.

**Refining the query:** In d3re, users can easily access the result of the rule and all facts generated by ddisasm through the GUI by writing into output tables using d3re rules. Unfortunately, our above query produces over 50 possible results—checking each occurrence would still be a timely endeavor. Next, we narrow down the query space to the range of just the main function. We use an auxiliary predicate code_in_range, which we seed with constants for the beginning and end of the main function we gain from inspecting the binary in Ghidra.

```prolog
use_before_def_global(EA_used,GA,Name) :-
  code(EA), instruction_get_dest_op(EA,Index,_,),
  pc_relative_operand(EA,Index,dest),
  defined_symbol(dest,_,"OBJECT","GLOBAL",_,_).

use_before_def_global(EA_used,GA,Name) :-
  code_in_range(19490,21704),
  !def_used_global( EA_used,GA,EA_used,_),
  defined_symbol(GA,_,"OBJECT","GLOBAL",_,Name).
```

### Fig. 4: uninitialized variables

2 A relevant analogy might be that ddisasm is the standard library of d3re.
After new rules are applied, the output of the program becomes empty: however, this does not specify the program is free from the vulnerability. First, because of our constraint, only the main function is searched, bugs may still hide in other functions. Secondly, if all usage of a variable is before it’s definition, null pointer error can still appear: programmers may initialize a variable to NULL and use several non-total branches to initialize the pointer, leaving the pointer uninitialized at the join point when no switch fires. We modify our rules to account for this:

```prolog
use_before_def_global(EA_used, GA, Name) :-
  code_in_range(from, to), EA_used >= from,
  EA_used < to,
  used_global(EA_used, GA, Index),
  def_used_global(EA_used, GA, _,_),
  !def_used_global(_, GA, EA_used, _),
  defined_symbol(GA, _, "OBJECT", "GLOBAL", _, Name).
```

This change results in 19 addresses to search, and combining these results with use-def information in the previous step and intra-procedural control-flow graph in Ghidra, we can fairly easily infer that the global variable swap_word is initialized to 0 at address 0x4c2a, that both conditional jumps 0x4f80 and 0x4fb8 fail, and observe a subsequent usage of swap_word at 0x5017 which will trigger a crash.

We conclude this section by remarking upon the nature of our analyses. Our analyses would be considered naïve by the standards of industrial static analyses. Indeed, our reasoning is not even sound—we can restrict ourselves to looking at results for only one function or ignore complex behavior. Still, we believe that this iterative ad-hoc reasoning is a technique many reverse engineers already employ—the vision of D³RE is to harmoniously leverage state-of-the-art deductive reasoning engines while performing human-guided RE tasks.
We define an EDB metadatabase as a graph of EDBs with labeled edges, \((\Delta, \mathcal{L})\), where \(\Delta\) is a set of EDBs, each EDB enumerating tuples for a given set of relations, and \(\mathcal{L}\) is a relation in \(\Delta \times Rules \times \Delta\). When we process a program, \(P\), using an input EDB, we traverse the graph \((\Delta, \mathcal{L})\) to find the most optimal, compatible EDB to start execution of \(P\). Aided by Datalog’s monotonicity, we define an EDB as compatible if it was produced by a subset of rules (or facts) from the input program / EDB. We conclude our formalism sketch by remarking that \((\Delta, \mathcal{L})\), given our usage, also forms a lattice.

B. Implementation of d3re

D3re is implemented in two parts: a REPL that communicates with Ghidra’s GUI and a background service to manage the metadatabase and run the Datalog engine. The REPL currently communicates with Ghidra via a third-party extension named ghidra_bridge [15], which we plan to replace imminently with an extension using protocol buffers.

To execute a Datalog program \(P\), d3re analyzes the file using the logic sketched in the above section to determine an optimal compatible EDB to use. In the common case, a user will gradually accumulate a stream of programs \(P, P', P''\) consisting of a mix of rules and assumptions. In the future, we envision that certain assumptions (e.g., about calling conventions) may be implemented as GUI extensions rather than, e.g., manually-enumerated facts. After each run, the metadatabase will index the output facts and associate them with the program \(P\), establishing an edge in the aforementioned graph. In our experiments, we refer to this as “caching.”

IV. Evaluation

We evaluated d3re both qualitatively, by implementing several queries, and quantitatively by measuring its performance in benchmarks. While d3re is still a work in progress, we had several hypotheses we aimed to test as we designed and conducted these experiments. First, we wanted to understand whether d3re provided the necessary building blocks to enable replacing currently-existing Ghidra scripts. Second, we wanted to understand whether d3re could offer performance competitive with the kinds of Ghidra scripts that reverse engineers typically use. Last, we wanted to understand the performance of Ghidra for performing several repeated queries that might mirror a realistic end-to-end workload using d3re.

Ghidra Script Replication Study: we wanted to determine whether d3re could realistically be used to accomplish the kinds of tasks that reverse engineers face on a day-to-day basis. This is an admittedly challenging question, which we plan to eventually evaluate in several ways including user studies. However, as initial work in this direction we arbitrarily selected four Ghidra scripts listed in the awesome-ghidra GitHub repository [16]. The scripts we chose are listed in Table I, along with their corresponding lines of code in Python / Datalog. While Ghidra scripts may consist of a mix of Python and Java, our experience is that most scripts use a small subset of the Python API. The first three are relatively small and find instructions that match a specific template, e.g., non-xor finds xor instructions that aren’t zeroing registers, and overflow heuristically searches for potential overflows in
calls to common functions such as `strcpy`. Our largest was `findcrypto`, which looks for common cryptographic constants.

**Qualitative Results of our Replication Study:** Our experience using `d3re` to replace Ghidra scripts must be understood in the context that we are expert users and the developers of `d3re`. However, we are pleasantly surprised that `d3re` enabled us to succinctly write equivalent implementations of each Ghidra script: we rewrote each script in substantially less Datalog code. This is because the declarative nature of Datalog eliminates the need for much of the conventional ceremony around, e.g., looping over instructions and checking against a type that we found in our evaluation scripts. Key to D³RE’s success, we believe, is its ability to directly use relations from `ddisasm`: we found that much of the necessary work of, e.g., filtering instructions by their type or operand was very useful at achieving succinct Datalog in practice. We are in the initial planning stages of developing a reverse engineering tutorial (or mini-course) around `d3re`, and are hoping to use this to recruit developers to get more realistic assessment of `d3re`’s usability by professional REs.

**Quantitative Results of our Replication Study:** We hoped that `d3re`, being based on a high-performance Datalog solver, would offer performance competitive with Ghidra’s current scripts. Each of our evaluation scripts processed the entire binary and would highlight or label certain instructions. To test the Ghidra scripts, we used Python’s standard time function before and after the script’s work finished. We evaluated the corresponding Datalog program by using Soufflé’s internal performance timers. We then benchmark Ghidra vs. `d3re` on a corpus of six binaries (all sized less than 10MB), five from `ddisasm`’s test suite and Soufflé, shown at the top of Table II. We used the latest versions of each pre-built in the latest Arch Linux, but we used a pre-built version of Soufflé. For each script, we waited for all of Ghidra’s typical analyses to finish, and similarly we ran `ddisasm` to build up the initial input database for `d3re`.

The body of Table II compares the runtime of each Ghidra script versus its corresponding implementation in `d3re`. The single occurrence of – indicates that Ghidra did not finish within an hour. Broadly, we found that `d3re` outperformed Ghidra for each of the scripts in our replication study. As we had hoped, `d3re`’s design allowed us to leverage useful relations from `ddisasm`. We found that many scripts do things like naively loops over sets of functions or symbols to locate some property. By contrast, the declarative style of `d3re` allowed us to write these not only more succinctly (e.g., Datalog naturally aggregates results) but also more efficiently—Soufflé optimally compiles input programs to efficient relational algebra kernels that loop only when necessary. We did observe various ways in which `d3re`’s limitations could cause performance issues. For example, the `findcrypto` script scans the binary for 256-segments of code. `d3re` is built on Soufflé, which supports 64-bit primitive ints, but not 256-byte sequences. Thus, we had to build up sequences via a set of Datalog rules, causing inefficient memory representation due to the necessary duplication due to representing subsequences as Datalog facts.

**Evaluating End-to-End Behavior in Subsequent Invocations:** To understand the effect of caching via repeated calls to `d3re`, we ran four subsequent analysis queries in a row by using both our caching-based approach and without caching (wherein we started only with the results of `ddisasm`). Our results are shown in Table III: the time of the cached run (C) is shown above the time for the correspond sequential run (S). As each query builds on the previous, we expect caching to reduce the amount of work and commensurately reduce the runtime. stack_var finds stack-allocated variables, while heap_var calculates stack variables holding pointers to heap values based on stack_var, static_var and unl_static attempt to find uninitialized global variables. Overall, we found rule caching was especially important on larger binaries versus sequential runs, justifying our choice to structure the metadatabase as a graph.

### V. Related Work and Conclusion

We conclude with a brief discussion of proximately-related work that lies at the synthesis of reverse engineering and static/dynamic analysis, and contextualize this work in terms of our aspirations for the future of D³RE. There has been extensive work using logic programming, and in particular Datalog, for static analysis of higher-level languages such as Java [17]–[19]. The success of the Soufflé Datalog engine has inspired recent adoption of logic programming within the binary analysis community. For example, Datalog Disassembly uses Soufflé to achieve both faster and more-precise disassembly than the state-of-the-art disassembler Ramblir [13]. Similarly, OOAalyzer uses XSB-Prolog, a version of Prolog implemented as a library [20]. We are currently reimplementing OOAalyzer in `d3re` targeting C++ Linux binaries. We feel particularly excited about this direction because we believe Soufflé will be immediately more scalable than XSB-Prolog.

While there are a broad range of plugins for Ghidra and IDA Pro to load the results of static analyses, we believe `d3re` is the first to focus on the combination of open-ended deductive logical inference and rapid interactivity (enabled by our metadatabase). We believe the most closely related work is Ponce [21], which enables GUI-based symbolic execution. We plan to integrate symbolic execution into `d3re` as a long-term goal, inspired by the recent work of Formulog [22].

Our goal in this work was to introduce a new vision for reverse engineering, D³RE, wherein expert users rapidly query high-performance logical inference engines to help them accomplish their day-to-day work in RE, vulnerability construction, and penetration testing. Visualization-based tools such as Ghidra are of immense value in understanding a binary, but have fundamentally different design considerations than high-performance logical inference engines (such as Soufflé). Recent work in compiling Datalog to parallel relational algebra (e.g., Gilray et al. [23]) has enabled a new frontier in scale of Datalog-based analyses. We hope that developments such as these will someday enable realizing fully the vision of D³RE to help reverse engineers perform powerful static binary analyses at unprecedented scale.
REFERENCES


