

# RoBin: Facilitating the Reproduction of Configuration-Related Vulnerability

Ligeng Chen<sup>†</sup>, Jian Guo<sup>†</sup>, Zhongling He<sup>†</sup>, Dongliang Mu<sup>§\*</sup> and Bing Mao<sup>†</sup>

<sup>†</sup>National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

{chenlg, zhe}@smail.nju.edu.cn, jaguo2014@outlook.com, maobing@nju.edu.cn

<sup>§</sup>School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan, China  
dzm91@hust.edu.cn

**Abstract**—Vulnerability reproduction paves a way in debugging software failures, which need intensive manual efforts. However, some key factors (*e.g.*, software configuration, trigger method) are often missing, so we can not directly reproduce the failure without extra attempts. Even worse, highly customized configuration options of programs create a barrier for reproducing the vulnerabilities that only appear under some specific combinations of configurations.

In this paper, we address the problem mentioned above — reproducing the configuration-related vulnerability. We try to solve it by proposing a binary similarity-based method to infer the specific building configurations via the binary from crash report. The main challenges are as follows: precise compilation option inference, program configuration inference, and source-code-to-binary matching. To achieve the goal, we implement RoBin, a binary similarity-based building configuration inference tool. To demonstrate the effectiveness, we test RoBin on 21 vulnerable cases upon 4 well-known open-source programs. It shows a strong ability in pinpointing the building configurations causing the vulnerability. The result can help developers reproduce and diagnose the vulnerability, and finally, patch the programs.

**Index Terms**—Binary Code Similarity, Building Configurations, Reproduction, Vulnerability.

## I. INTRODUCTION

Despite the best efforts of software developers, software systems inevitably contain defects that may be leveraged as vulnerabilities. In the real world, these vulnerabilities could lead to many notorious cyberattacks, such as Stuxnet, WannaCry Ransomware [1], HeartBleed [2]. Since modern software systems are becoming more complex and release cycles are getting shorter, testing teams are unable to identify all the possible vulnerabilities before a software release. Thus software systems are typically released with many underlying vulnerabilities and end-users may experience software failures. Considering the performance overhead, when software crashes, only a crash report with the necessary information will be sent to the developers. The developers first need to reproduce the failure so can they diagnose and patch the vulnerability.

**Existing Works.** For software failures, the reproduction of underlying vulnerabilities is a key stage in the vulnerability diagnosis. Without reproduction, software developers are unable to diagnose the root causes and eventually verify whether vulnerabilities are fixed or not. There are several existing

works that attempt to reproduce the underlying vulnerabilities and enables software developers to check the input and program state that lead to software failures. Mu et al. [3] show that vulnerability reproduction is difficult in the real world and there are many missing key factors in the vulnerability reports, such as building configurations, Proof-of-Concept (PoC), trigger method, etc. REPT [4] leverages online trace recording from Intel Processor Tracing (Intel PT) and offline binary analysis to facilitate reverse debugging of software failures. Hercules [5] takes advantage of symbolic execution to generate one input which reproduces the give software crash. BugRedux [6] conducted one empirical analysis, to understand the trade-offs between different execution data recording and the effectiveness of approaches.

**Limitations.** However, some vulnerabilities only occur when the specific code segment controlled by specific program configurations is compiled into the binary, which we call *configuration-related vulnerabilities*. Only can the developers reproduce the failure when they get the binary containing the vulnerable code segment. Nevertheless, though they can reproduce the failure, software developers and security analysts still need to instrument source code with sanitizers (*e.g.*, Address Sanitizer [7]) to locate and diagnose the vulnerability. To be concrete, if we want to patch the program, it is vital to hold the source code and non-stripped vulnerable binary at the same time. So it is quite significant to figure out the program configurations of the configuration-related vulnerabilities.

For the developers, the crash report is the most accessible resource, which is collected by several tools (*e.g.*, Windows Error Reporting [8], breakpad [9]), and records the memory dump with other signals at the crashing time. It is reasonable to figure out the building configurations from code segments extracted from the report. To the best of our knowledge, none of the previous works attempts to infer the building configurations from the crash report to facilitate the reproduction of configuration-related vulnerabilities.

**Our Approach.** To this end, we present a system called RoBin, a practical solution to figure out the configurations in the building process from the code segment of the crash report. The method is beneficial to reproduce the vulnerabilities in the highly configurable programs, even for the further diagnosis and patch process. In this work, we mainly focus on inferring the provenance of *building configurations* (that is how the

\* Corresponding author.

binary is compiled with specific *compilation options* and *program configurations*) from the crash report. To consolidate our work, we conduct one empirical analysis of how the different building configurations affect the similarity of generated binaries.

To achieve this goal, we present a comparison-based strategy to extract the configurations from the code segment of the crash report (called *crash report binary* later). Firstly, by recursively generating similar binaries with the crash report binary, we efficiently and precisely search the most possible provenance of *compilation options*. To this step, we hold the crash report binary and the generated binary with the highest similarity on hand. Then, we capture the difference between crash report binary and generated binary, and map it by the self-defined features with the source code to infer the provenance of *program configuration(s)*. Finally, if we successfully get the correct building configurations, we can reproduce the failure, and the program can be carefully diagnosed and patched without any missing information.

To demonstrate the utility of RoBin, it is evaluated on a set of 21 real-world configuration-related vulnerabilities via 4 well-known programs. Our experiment shows that RoBin could effectively figure out the building configurations needed to generate the vulnerable binary and reproduce the failure, with only 2 cases failed with limited information.

**Contributions.** The contributions are as follows:

- We conduct an investigation on how the different building configurations influence the generated binaries on binary similarity. Based on the discovery, we propose a comparison-based strategy to infer the provenance of building configurations.
- We implement a prototype called RoBin on Linux to facilitate software developers and security analysts in reproducing the configuration-related vulnerabilities.
- We demonstrate the effectiveness of the system on 21 real-world vulnerabilities. 19 vulnerabilities are successfully reproduced, and the output results are available for further applications, such as root cause diagnosis, vulnerability patching.

## II. BACKGROUND AND MOTIVATIONS

### A. Variability of Binary Building

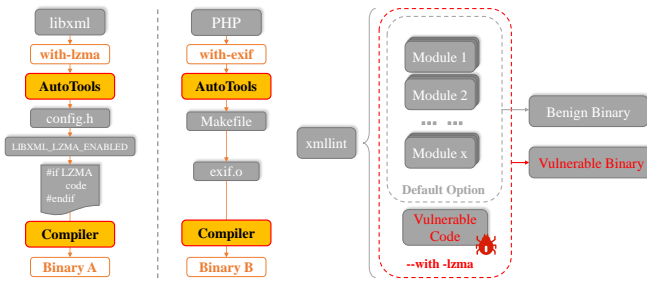


Fig. 1. Configs and programs. Fig. 2. Example of CVE-2018-14567

**Configuration-related Vulnerability.** Configuration-related vulnerabilities only exist under a specific combination of program configurations in a software system.

GNU AutoTools [10], like the GNU compilation system, is a toolset that helps developers adapt source code to various Unix-like systems. Based on the GNU AutoTools compilation system, developers can use Makefile at the file level or at a more fine-grained level to control the content during the compilation process. As shown in the Figure 1, libxml uses macros (e.g., *LIBXML\_LZMA\_ENABLED*) to control whether the code is compiled into an executable file. When running the config script, it will read the value of the program configurations, to control whether the source code is compiled into the executable file during conditional compilation. As for PHP, the program configurations are realized through file-level (e.g., *exif.o*) control configuration. It modifies the Makefile to control the compilation process. This mechanism is necessary to support a wide variety of hardware and specific program settings, but it introduces problems in reproducing vulnerabilities in those highly configurable systems.

**Motivation Example.** Take the following situation as an example. If a specific program configuration controls the vulnerable code, which is missed in the building process, the generated binary is non-vulnerable and the reproduction fails without observing any abnormal behaviors. The vulnerability can not be reproduced without any extra information. Take the case shown in Figure 2 as an example. We get a benign binary by compile the *xmllint* with default option, while we get a malicious one with the extra Macro option *--with-lzma*.

So it is obvious to come out a naive solution — enabling all of the configurations trying to include the vulnerable code. However, it is not practical in the real world after our thorough investigation. On the one hand, some configurations are mutually exclusive. It is impossible to compile all the source code into binary at one time. On the other hand, other configurations in the building process might influence the execution of the vulnerable program, and the generated binary with all possible program configurations will fail to reproduce the target vulnerability. Therefore, we need to treat or include each program configuration conservatively in case they could lead to the failure of reproduction.

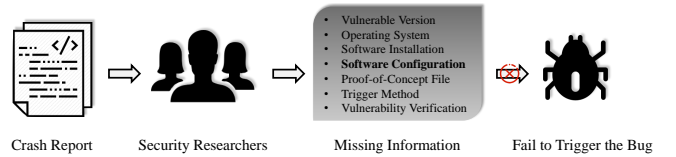


Fig. 3. Software Building Process.

### B. Vulnerability Reproduction

Nowadays software vendors are increasingly relying on the power of software users to identify security vulnerabilities. Once software developers receive those failure reports from software users, they will reproduce and identify the underlying vulnerabilities promptly. However, previous works have shown failure reports usually lack much key information on vulnerability reproduction [3], [11]. To understand the reproducibility of software vulnerabilities, one study [3] conducted

an empirical analysis of real-world security vulnerabilities, intending to quantify their reproducibility. The study shows that vulnerability reproduction is difficult and needs intensive manual efforts. And software installation is usually a missing key factor that makes vulnerability reproduction difficult, which is shown in Figure 3. The software installation builds the vulnerable binary and installs it to one specific path. And the key point in this step is how to build the vulnerable binary from source code. That is, in our scenario, it is vital to get the precise program configuration(s).

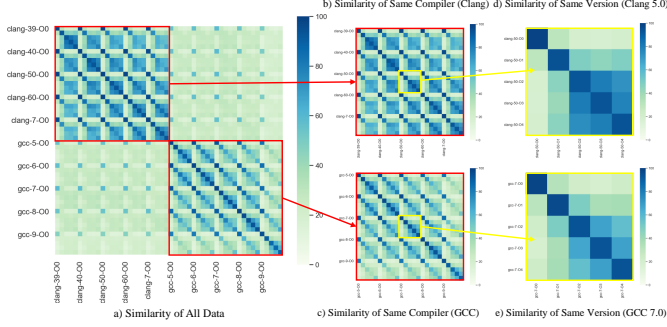


Fig. 4. Cross-Compiler, Cross-Optimization Level, Cross- Compiler Version Binary Similarity Comparison by BinDiff

### III. ANALYSIS ON BINARY COMPARISON

Like the aforementioned, the specific value of program configurations is hard to be extracted from crash report binary with scattered information. So it need to be split the goal into two stages: **compilation option inference stage** and **program configuration inference stage**. The reason for firstly inferring the compilation options is because the features of them are distributed globally across the binary, and the information of program configurations need to be matched more precisely. Once the compilation options are confirmed, RoBin can concentrate on finding the configurations without being distracted by other information.

In this section, an empirical study is conducted on the binary similarity measured by BinDiff, to investigate how the binaries behave differently that are compiled from the same source code but with different compilation options. The investigation result can guide us to figure out the suspected *compilation options* by leveraging the feedback of BinDiff.

To scope the field, RoBin is experimented on 2 compilers (GCC and Clang), 5 optimization levels (O0, O1, O2, O3, Os separately), and 5 versions for each compiler (5.0, 6.0, 7.0, 8.0, 9.0 for GCC, 3.9, 4.0, 5.0, 6.0, 7.0 for Clang). Under this setting, a piece of source code has  $50 (2 \times 5 \times 5)$  possible combinations of compilation options. Figure 4 presents the cross-comparison between 50 sets of binaries compiled with different compilation options, measured by BinDiff. Figure 4 a) shows the full scene of the result. Due to the limited space, part of the description of the grids is omitted. From the top row to the bottom, the grid separately denotes Clang-39-O0, Clang-39-O1, Clang-39-O2, Clang-39-O3, Clang-39-Os, Clang-40-O0, ..., and so on. The color of the grid represents the grade

of similarity between the two binaries. The darker, the more similar. For instance, the grid in row 1 column 2 denotes the similarity between the binaries compiled with Clang-39-O0 and Clang-39-O1, which are not that similar. As for the sub-figure b) and c), it is easier to capture the trend of similarity of the same compiler. And the sub-figure d) and e) that shows the trend in the same version but with different optimization levels.

According to the figures, the following findings inspire the further procedure of inferring compilation options:

- **Findings for Sub-Figure a)** ❶ Binaries compiled with -O0 behave quite differently from -O1, -O2, -O3, and -Os, no matter which compiler. ❷ In the vertical comparison, binaries compiled by the same compiler has a higher similarity under the same optimization level.
- **Findings for Sub-Figure b) and c)** ❸ Neighboring compiler versions (*i.e.*, GCC-5 and GCC-7 are the neighboring compiler versions of GCC-6) process more similar than others, which will slightly influence how the optimization level affect the binary. ❹ Binaries compiled with the same optimization level behave more similar from the same compiler, no matter which version.
- **Findings for Sub-Figure d) and e)** ❺ Neighboring optimization levels (*i.e.*, -O1 and -O3 are the neighboring optimization levels of -O2) make the binary more similar, but only ranging from -O0 to -O3. -Os behave more similar with -O2 rather than -O3.

### IV. PROBLEM STATEMENT

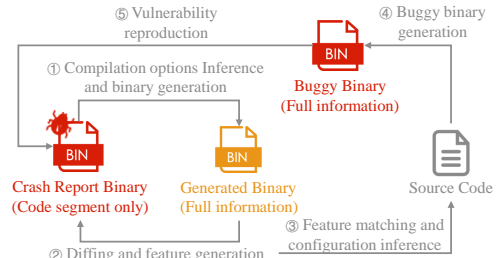


Fig. 5. Brief Workflow according to RoBin

#### A. Problem Definition

- **Input:** crash report binary  $crB_t$ , open source program  $S_x$
- **Intermediate Result:** generated binary  $gB_s$
- **Output:** buggy binary  $bB_n$

According to the brief workflow shown in Figure 5, RoBin leverages generated binary  $gB_s$  as an intermediate result to bridge the crash report binary  $crB_t$  and open source program  $S_x$ , and finally generates the buggy binary  $bB_n$ .

❶ Given a piece of crash report binary  $crB_t$ , we firstly infer the provenance of compiler  $C_i$ , optimization level  $OL_j$  and compiler version  $V_k$  by the recursive generation step as  $R(crB_t) = \{C_i, OL_j, V_k\}$ , where in this work  $C = \{GCC, Clang\}$ ,  $OL = \{O0, O1, O2, O3, Os\}$ ,  $V = \{GCC 5.0, GCC 6.0, GCC7.0, GCC 8.0, GCC 9.0, Clang 3.9, Clang 4.0, Clang 5.0, Clang6.0, Clang 7.0\}$ .

With the information of compilation options above, we can generate a binary  $gB_s = G(S_x, C_i, OL_j, V_k)$  from the open source program  $S_x$ . The similarity of two binaries  $Sim(crB_t, gB_s)$  exceeds the preset threshold, which indicates that they are exactly the same except for the program configuration(s). ② Finally, with the difference  $Diff(crB_t, gB_s)$  between the binaries and the features  $F(S_x)$  captured from the source code, ③ we are able to find the relationships and infer the setting of configuration(s)  $Cinfig_y = \{config_1 = True, config_2 = Flase, \dots\}$ .

Till this step, we have all of the building configurations and the source code on hold, ④ so that we can generate the buggy binary  $bB_n = G(S_x, C_i, OL_j, V_k, Config_y)$ . ⑤ And it is natural to reproduce the vulnerability with the buggy binary  $bB_n$ , which can also assist most of the downstream applications in the security field.

## V. SYSTEM DESIGN

Figure 6 delineates the architecture of RoBin. The whole system contains three major components: **Compilation Options Inference**, **Configuration Inference** and **Vulnerability Reproduction**.

To summarize, our work focus on facilitating the reproduction of configuration-related vulnerabilities by hierarchically inferring the compilation options and program configurations.

### A. Compilation Options Inference

In section III, we summarize some findings from the analysis results. Based on them, we search the compilation options by the following steps.

①**Whether -O0 or NOT?** Firstly, we compile the source code by default version of GCC (or Clang) separately with -O0 and -O2. By comparing with the crash report binary, the similarity produced by BinDiff will tell whether the provenance of the optimization level is -O0.

②**Which Compiler?** Then, we compile the code separately by GCC and Clang with the same optimization level produced by the former step. The one that has a higher similarity with the crash report binary can tell the provenance of the compiler.

③**Optimization Level Confirmation.** If the provenance of the optimization level is -O0, we do not have to go through this step. Conversely, we generate the binary by the same compiler and its version with -O1, -O2, O3 to confirm the right optimization level by the highest similarity.

④**Compiler Version Confirmation.** Last but not least, we confirm the version of the compiler. By comparing the similarity of neighboring versions, we can finally get all of the compilation options.

Leading by the steps above, we are feasible to figure out the specific compilation options from the large search space. Due to the limitation that the mechanism of BinDiff is rule-based, we can not guarantee that the generated binary has the full similarity with the crash report binary. But the binary with a relatively high similarity is prepared for the next stage.

We take an example to show the procedure. As shown in Figure 7, we present the process of how to figure out

the provenance of compilation options of *xmllint* compiled with compilation options *GCC-7-O2*. All the percentages in the figure are the similarity comparison result between the generated binary and the crash report binary produced by BinDiff.

In the first step, we confirm ①whether it is compiled with -O0. By compiling the source code with the same default version of Clang (randomly choose the compiler) but different optimization levels, we can make sure the optimization level is NOT -O0. Then, we compare the generated binary compiled with ②different default compiler. The result indicated the provenance of the compiler is GCC. Next, by generating the binary with neighboring optimization levels, we can confirm the ③exact optimization level -O2. Finally, by listing all of the candidate compiler versions, we can figure out the ④specific compiler version. It takes 8 times (we do not repeatedly compare binaries with the same compilation options) of comparison to figure out the options. And the process of compilation options inference meets an end till now.

### B. Binary Diffing

This stage mainly contains three parts. According to Figure 6, we need go through **Binary Diffing**, **Source Code Extraction** and **Bin-Source Mapping** to figure out the specific program configuration(s).

To map the differences from binary to source code, we firstly leverage IDA Pro [12] and Bindiff [13] to identify identical and similar functions between the generated binary and the crash report binary. So the rest of the parts are the candidates controlled by the program configurations.

Due to the behavior of program configurations is unpredictable, some are 'enable' and some are 'disable', so the size of the relationship between the generated binary and the crash report binary is uncertain, which is presented in Table ???. To confirm our observation, we investigate our data set. The result shows three kinds of situation as follows:

- Some functions only appear in the generated binary;
- Some functions only appear in the crash report binary;
- Functions appear in both of them.

Because the size of the crash report is limited to compress the content, the crash report binary always does not contain symbolic information. Therefore, we need to position the differences as much as possible in the generated binary. According to the situations above, we give different solutions.

For the first situation, we realize those functions are not compiled into the vulnerable binary or disabled by the configurations. In the subsequent Bin-Source mapping step, the constraints of these types of functions can be obtained, and only the logical relationship 'and' is needed as the connection. To obtain the final constraints, we only need to reverse them. However, conflicts may arise during the procedure. Our strategy is to discard conflict functions. On the one hand, discarding them has little impact on the result. On the other hand, conflicts are very rare and almost all of them appear in function re-definitions. If we do not find any configurations, it may be a false alarm of BinDiff.

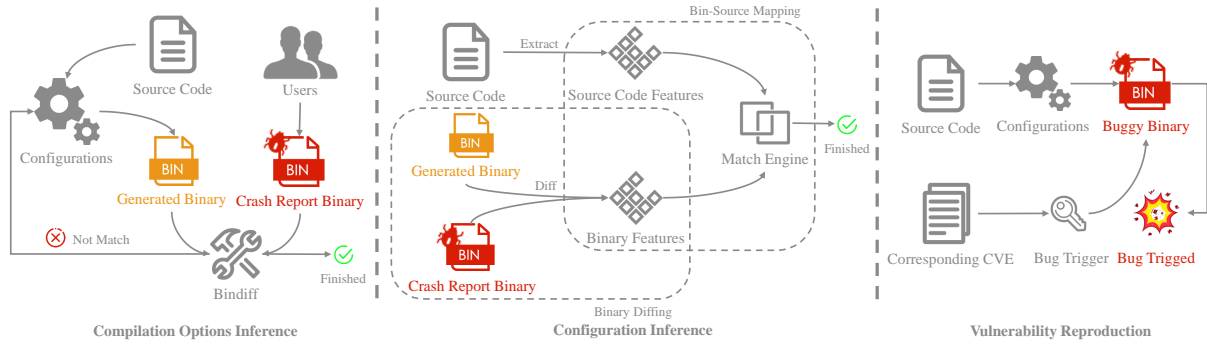


Fig. 6. Workflow of building configuration inference

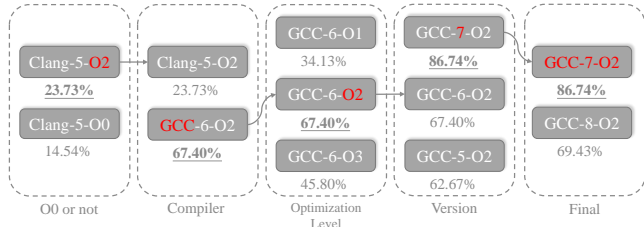


Fig. 7. An Show Case of Compilation Options Inference

In the second situation, it is exactly opposite to the first one. The process is similar to the first situation. The only difference is that we do not need to reverse the constraint. But we need to extract features separately from the crash report binary and source code due to the missing symbolic information of the crash report binary. After feature matching, a set of constraints is obtained, and the provenance of program configuration can be extracted by the solver.

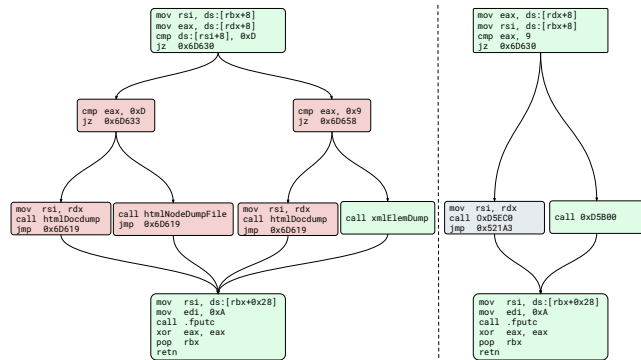


Fig. 8. Binary Diffing Example of Generated Binary and Crash Report Binary

For the last situation, we continue to detect and highlight binary differences between two variants of the same function. We use some fine-grained methods to map those binary differences to the corresponding source code. If the functions are exactly the same, we can easily extract and map the constraints. If the functions are partially matched the rest partially matched functions, it is due to the different value of Macros inside the function, which leads to the difference of internal structures.

To better analyze the result, we take a piece of binary code from *XMLlint* and visualize part of the structure of the generated binary and crash report binary with the help of IDA

Pro and BinDiff.

As shown in Figure 8, the green parts are the base blocks that exist in both of the binaries, the red parts are the base blocks that only exist in the generated binary, and the gray part is the base block that only exists in the crash report binary.

Till now, we have the rough constraints extracted by the diffing result from generated binary and crash report binary. We still need to map them to the source code, so that we can figure out the specific program configurations (*i.e.*, configuration files, macros).

### C. Feature Extraction

The exact correspondence between binary and source code is an open and difficult problem in the security field. So we come up with the second-best solution. Based on the extraction result of the binary diffing, we can get the corresponding source code range. However, it is impossible to determine which macro-related code fragment in the source code appears in the crash report binary. In addition, when there are more than one macros in the source code range, it is challenging to directly infer the specific configurations from the generated binary. Therefore, we separately extract the source code fragment features related to the macro and search for the source code feature in the binary to determine which code fragment is compiled into the binary. To tackle the problem, we need to find out some *features* to assist our work. The ideal feature needs to be **unique** and **stable**. The feature should not appear multiple times within the specified range, and they should be stable enough which cannot be affected by the different compilation options from source code to binary as well.

### D. Feature Generator

With the aforementioned features, we add some structural features *if*, *condition*, *then* branch, and *else* branch. When combined with structural features, we can make the semantic features more unique to improve efficiency.

**Source Code Feature Generator.** TypeChef has been designed to check for incompatible types and developer errors in untested configuration combinations. So we leverage TypeChef to generate a macro AST, called VAST, and extract source code features from it. Each extracted feature has the constraints of the features. **Binary Feature Generator.** We leverage IDA Pro to extract the features from binaries. If more accurate



information is needed, we will use some procedural analysis techniques, such as *binary function parameters*, *data streams* to assist us in more precise binary matching.

### E. Feature Matching Engine

In the feature matching stage, we start with a coarse-grained matching strategy for efficiency. When it fails, we complement it with a fine-grained matching strategy.

**Coarse-grained Matching.** In the coarse-grained matching, we mainly consider the simple features, *string*, *constant* and *function call*. For instance, we directly search a string in the source code to see whether the corresponding string information exists in the binary function. If it exists, the match succeeds. If it fails, we select other features to match again and come out with the confidence of the feature identification. Assuming that none of the features are matched successfully, it can be considered that the code does not exist in the binary. Furthermore, we can use this strategy to match the function level. However, this matching strategy is relatively simple and rough, and some of the segments can not be successfully matched.

**Fine-grained Matching.** In case all the rules of coarse-grained matching fail, we will turn to use a fine-grained matching strategy. Since some features can not be a matching anchor alone, so we combine them with extra structural features (e.g., branch structure, loop structure) to become a more unique identifier. Such as, *if* statements in source code correspond to the instruction *cmp* in binary code. As shown in Figure 9, if the function *foo* is considered alone, once the function call *foo* can be found more than once in the binary, so it is not certain whether the macro-related code is compiled into the binary. But when the combined features are introduced, considering the features of the *if* structure, the constraint can be extracted in the comparison condition, and the function call *foo* can be extracted in the statement block. So when we leverage the context information (i.e., structural feature), the combined information can be used as the smallest matching unit to accurately match the functions affected by the macro.

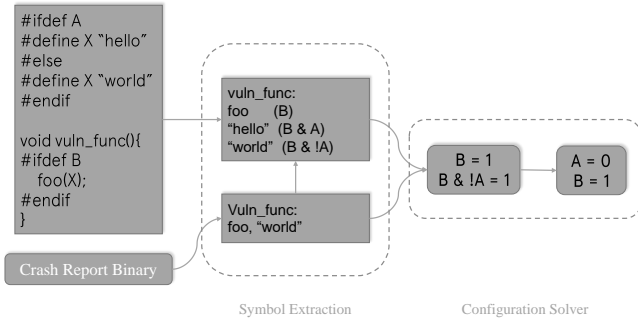


Fig. 9. Example for the Solver and Matching Engine

### F. Solver

By feature extraction and feature matching, we can confirm whether a code segment controlled by program configuration(s) exists in the target binary (generated binary or crash report binary). Furthermore, we need to figure out which

program configuration controls the code segment. To infer the precise configuration(s) of the program, we speculate the specific values of the program configurations according to the result of feature matching and the existing condition of the source code from VAST. Take the case in Figure 9 as an example, feature matching can find the string *world* in the binary matches the string in the source code. And there is a function call *foo* in the binary, with VAST we can get the condition of *foo*'s existence is *B*, and the existence condition of the string *world* is  $\neg A$ . Combining the analysis and matching results of the source code, a satisfactory value of macros *A* and *B* can be obtained by the constraint solver.

### G. Vulnerability Reproduction

At this stage, we overcome the challenges mentioned in the previous sections and infer the compilation options and program configurations from the crash report binary. With the source code and the inferred building configurations, we are not only able to reproduce the failure, but also can precisely locate the vulnerability. Furthermore, we can leverage the result to diagnose and patch the vulnerable part.

## VI. EVALUATION

### A. Procedures and Settings

Due to the inconsistency of CVE and the bug report [3], [14], only part of the vulnerabilities is crash reports available. So we simulate the scene to evaluate our work. It will be our future work that facilitating the reproduction of vulnerabilities by leveraging limited and incomplete information.

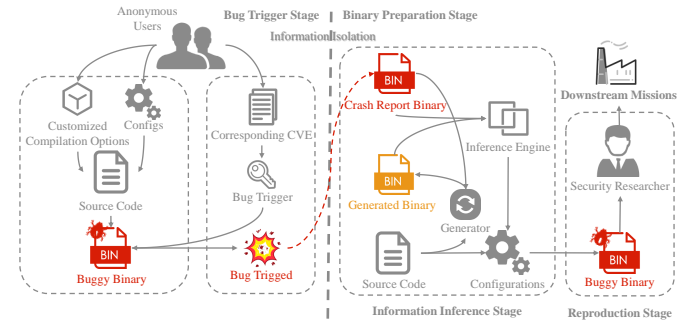


Fig. 10. Procedure of Evaluation

**Evaluation Procedures.** As shown in Figure 10, we present the procedure of setting a simulated scene for our experiment. Firstly, we divide the experiment candidates into two groups: **Anonymous Users** (*users* for short) and **Security Researchers** (*researchers* for short). Two groups of people are completely information isolated, which is consistent with what happens in the real world.

Anonymous users correspond to a variety of users in the real world. They randomly use the customized compilation options and set the program configurations according to the CVE report to compile the source code into a buggy binary. With the bug trigger, they are able to trigger the specific vulnerability. Then, they sent the bug report containing the vulnerable code segment to the researchers.

	CVE ID	Compilation Options Inference					Configuration Inference		Reproduction	
		Version	$t_{Build}$ (s)	compilation options	$T_{Infer}$ (#)	$\beta$	Config.	$t_{Extract}$ (s)	Default	Robin
libxml	CVE-2015-8035	2.9.1	22.67	GCC-7-O0	5	98.22%	with-lzma	28.96	X	✓
	CVE-2017-18258	2.9.6	18.38	GCC-5-O3	7	86.53%	with-lzma	33.32	X	✓
	CVE-2018-9251	2.9.8	19.78	GCC-8-O1	7	95.33%	with-lzma	182.17	X	✓
	CVE-2018-14567	2.9.8	19.24	Clang-6-O1	7	94.60%	with-lzma	181.37	X	✓
OpenSSL	CVE-2014-3513	1.0.1i	65.28	GCC-5-O3	7	73.05%	with-srtp	102.19	✓	✓
	CVE-2014-3568	1.0.1i	66.54	GCC-7-O0	5	98.94%	no-ssl3	43.68	X	X
	CVE-2014-3569	1.0.1i	67.45	Clang-6-O1	7	97.61%	no-ssl3	44.52	X	X
	CVE-2016-6304	1.1.0	81.16	GCC-8-O2	7	95.61%	with-ocsp	193.47	✓	✓
PHP	CVE-2007-1001	5.2.0	109.09	Clang-6-O1	7	87.89%	with-gd	184.33	X	✓
	CVE-2016-6294	7.0.0	308.3	GCC-5-O2	8	87.50%	enable-intl	1712.87	X	✓
	CVE-2016-6297	7.0.0	306.87	Clang-5-O3	7	92.80%	enable-zip	1708.22	X	✓
	CVE-2019-9025	7.3.0	378.53	GCC-6-O0	5	99.13%	enable-mbstring	2549.78	X	✓
	CVE-2019-9638	7.0.0	305.64	GCC-8-O3	6	85.44%	enable-exif	1729	X	✓
	CVE-2019-9641	7.0.0	307.35	Clang-39-O0	6	99.26%	enable-exif	1701.33	X	✓
	CVE-2019-9640	7.0.0	307.21	GCC-7-O2	8	97.01%	enable-exif	1717.82	X	✓
	CVE-2019-9639	7.0.0	309.87	GCC-7-O1	8	98.82%	enable-exif	1712.34	X	✓
proftpd	CVE-2009-3639	1.3.1	9.47	GCC-8-O1	7	99.05%	mod_tls	17.98	X	✓
	CVE-2010-4652	1.3.1	8.79	Clang-6-O1	7	98.50%	mod_sql	18.23	X	✓
	CVE-2013-4359	1.3.3d	9.41	GCC-6-O2	6	98.70%	mod_sftp	31.93	X	✓
	CVE-2015-3306	1.3.5b	10.09	Clang-5-O3	7	98.69%	mod_copy	39.31	X	✓
	CVE-2016-3125	1.3.3d	9.45	GCC-5-O0	5	97.84%	mod_tls	32.46	X	✓

TABLE I  
EVALUATION RESULT OF VULNERABILITIES IN EACH STAGE

When the researchers receive the report, they are able to process the inference of compilation options and program configurations. Finally, with all the information and materials on hold, they can reproduce the failure. What’s more, the result also can benefit the downstream applications of vulnerability reproduction, such as, root cause diagnosis, etc.

### B. Experiment Result

**Compilation Options Inference.** Due to the procedure of our method, we need to recursively generate the binary. To minimize the time of building each program, we optimized the search strategy according to the analysis result of binary similarity.

In this work, we mainly focus on the vulnerabilities compiled from GCC and Clang. Firstly, *users* will customize the compilation options. With the vulnerability-related program configurations, they can produce buggy binaries. After confirming the vulnerabilities can be triggered, they pack the code segment into crash reports and send them to the *researchers*.

At the beginning, *researchers* randomly set the default options as *gcc-version6.0-00* or *clang-version5.0-00*. They will leverage the script to automatically figure out the compilation options guided by the optimized search strategy aforementioned. The result is shown in Table I. Column 4 denotes the building time  $t_{Build}$  in seconds, column 5 denotes the options customized by the *users*, column 6 denotes the number of times to generate the binary and compare it with the crash report binary, and column 7 denotes the confidence  $\beta$  of the last binary comparison produced by BinDiff. On average, we need to generate binaries about 6.6 times. In particular, with about 6 to 7 times of attempts, we can figure out the compiler, compiler version, and optimization level precisely from the binary with high confidence.

Overall, the time-consuming  $t_{CO}$  of this procedure is calculated as  $t_{CO} = t_{Build} \times T_{Infer}$ , which is in a wide range, from 47.23 seconds to 2478.76 seconds. The consumption of time

is related to the characteristics (e.g., size, complexity, etc.) of the program. And in this stage, all the customized compilation options of vulnerabilities are successfully inferred.

**Retrieval Program Configuration.** After inferring the compilation options for each program, we retrieve the program configurations in a more fine-grained way mentioned in the previous section. According to Table I, column 8 denotes the vulnerability-related program configurations, and column 9 denotes the consumption of time to extract the features and map them. We spend most of our time parsing AST (Abstract Syntax Tree) of the programs. It is under our expectation that the rest of the processes take less than 20 seconds. The result also inspires us to improve the speed of parsing AST in future work.

**Configuration-Related Vulnerability Reproduction.** With the inferred compilation options and program configurations on hold, we are able to reproduce the failures. As shown in the Table I, column 10 indicates that most of the vulnerabilities can not be reproduced with default options, and column 11 represents that only 2 of them can not be reproduced with the assistance of RoBin due to the limited information provided. Intuitively, RoBin can effectively reproduce the configuration-related vulnerabilities. Interestingly, 2 vulnerabilities from OpenSSL can be reproduced with default building configurations. But RoBin still can contribute to pinpointing the vulnerable range in the default program configurations rather than other options. This ability is not available with other tools.

## VII. RELATED WORKS

**Vulnerability Reproduction.** The reproduction of vulnerabilities is quite significant to maintain the security of open-source programs. Some works [3], [11] contribute to investigating the situation. One work [3] conducted an empirical analysis of real-world security vulnerabilities with the goal of quantifying their reproducibility. The presented result indicates that the vital information for reproduction is always scattered

or missing due to the large gap in expert knowledge between users and developers.

**Compilation Tool-chain Inference.** To bridge the gap between reproduction and developers, various works propose a different kind of method to identify the provenance of compilation tool-chain from the binary. Luca [15] tries to infer the provenance of the compiler by utilizing the graph neural networks to extract binary features. Bincomp [16] proposes a hierarchical model to infer the compiler, version, and optimization levels from binary. What's more, Himalia [17] and its future work [18] demonstrate the problem of distinguishing the provenance of optimization levels and make the method quite precise and efficient.

**Vulnerable Code Analysis.** Abal [19] presents a qualitative analysis of the Linux kernel for the researcher. TypeChef [20], the tool also does variability-aware static analysis of software systems to detect compile and link-time errors introduced by the C preprocessor, while Vampyr [21] proposes a static analysis of kernel drivers. And KConfigReader [22] analyzes the Linux kernel and transforms it into a newly presented formula. OSSPATCHER [23] presents a method to build the relations between OSS variants and their compiled counterparts in binaries, aiming to figure out the potential vulnerabilities. During the procedure, OSSPATCHER also tries to figure out the corresponding program configuration by the source-to-binary mapping technique.

These works mainly focus on directly analyzing source code and binary related to the vulnerabilities. In contrast, RoBin gets the most of the information that causes the program to crash, trying to facilitate the reproduction of configuration-related vulnerabilities. We pay more attention to rebuild the building configurations of the vulnerabilities and try to reproduce how the users go from building open source software to cause the crash. The output materials and information are quite beneficial for the downstream applications, *e.g.*, root cause diagnosis, precise source code patch, etc.

## VIII. CONCLUSION

In this paper, we present RoBin, a comparison-based approach to figure out the building configurations to facilitate the reproduction of configuration-related vulnerabilities. RoBin can infer the compilation options assisted by BinDiff. Furthermore, we overcome the challenge of inferring the program configurations by proposing a source-to-binary mapping method. Finally, the on-hand materials help to reproduce the failures and benefit the downstream applications.

## ACKNOWLEDGMENT

This work was supported in part by grants from the Chinese National Natural Science Foundation (NSFC 61272078 and NSFC 62032010) and the Fundamental Research Funds for the Central Universities (HUST: 2020kfyXJJS129).

## REFERENCES

- [1] "WannaCry Ransomware Attack." [https://en.wikipedia.org/wiki/WannaCry\\_ransomware\\_attack](https://en.wikipedia.org/wiki/WannaCry_ransomware_attack), 2017.
- [2] Synopsys, "The heartbleed bug," 2014.
- [3] D. Mu, A. Cuevas, L. Yang, H. Hu, X. Xing, B. Mao, and G. Wang, "Understanding the reproducibility of crowd-reported security vulnerabilities," in *27th USENIX Security Symposium (USENIX Security)*, Usenix Security '18, 2018.
- [4] W. Cui, X. Ge, B. Kasikci, B. Niu, U. Sharma, R. Wang, and I. Yun, "REPT: Reverse debugging of failures in deployed software," in *Proceedings of 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, OSDI '18, 2018.
- [5] V.-T. Pham, W. B. Ng, K. Rubinov, and A. Roychoudhury, "Hercules: Reproducing crashes in real-world application binaries," in *Proceedings of the 37th International Conference on Software Engineering (ICSE)*, ICSE '15, 2015.
- [6] W. Jin and A. Orso, "Bugredux: Reproducing field failures for in-house debugging," in *Proceedings of the 34th International Conference on Software Engineering (ICSE)*, ICSE '12, 2012.
- [7] K. Serebryany, D. Bruening, A. Potapenko, and D. Vyukov, "Address-sanitizer: A fast address sanity checker," in *Proceedings of the 2012 USENIX Conference on Annual Technical Conference*, USENIX ATC '12, 2012.
- [8] K. Glerum, K. Kinshumann, S. Greenberg, G. Aul, V. Orgovan, G. Nichols, D. Grant, G. Loihle, and G. Hunt, "Debugging in the (very) large: Ten years of implementation and experience," in *Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles*, SOSP '09, 2009.
- [9] Google, "Breakpad." [https://github.com/google/breakpad/blob/master/docs/getting\\_started\\_with\\_breakpad.md](https://github.com/google/breakpad/blob/master/docs/getting_started_with_breakpad.md), 2007.
- [10] Google, "Gnu autotools." [https://en.wikipedia.org/wiki/GNU\\_Autotools](https://en.wikipedia.org/wiki/GNU_Autotools), 1992.
- [11] E. I. Laukkanen and M. V. Mantyla, "Survey reproduction of defect reporting in industrial software development," in *2011 International Symposium on Empirical Software Engineering and Measurement*, pp. 197–206, IEEE, 2011.
- [12] Hex-Rays, "Disassembler, ida pro," 2010.
- [13] Google, "Bindiff." <https://www.zynamics.com/bindiff.html>, 2016.
- [14] Y. Dong, W. Guo, Y. Chen, X. Xing, Y. Zhang, and G. Wang, "Towards the detection of inconsistencies in public security vulnerability reports," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 869–885, 2019.
- [15] L. Massarelli, G. A. Di Luna, F. Petroni, L. Querzoni, and R. Baldoni, "Investigating graph embedding neural networks with unsupervised features extraction for binary analysis," in *Proceedings of the 2nd Workshop on Binary Analysis Research (BAR)*, 2019.
- [16] A. Rahimian, P. Shirani, S. Alrbaee, L. Wang, and M. Debbabi, "Bincomp: A stratified approach to compiler provenance attribution," *Digital Investigation*, vol. 14, pp. S146–S155, 2015.
- [17] Y. Chen, Z. Shi, H. Li, W. Zhao, Y. Liu, and Y. Qiao, "Himalia: Recovering compiler optimization levels from binaries by deep learning," in *Proceedings of SAI Intelligent Systems Conference*, pp. 35–47, Springer, 2018.
- [18] S. Yang, Z. Shi, G. Zhang, M. Li, Y. Ma, and L. Sun, "Understand code style: Efficient cnn-based compiler optimization recognition system," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2019.
- [19] I. Abal, C. Brabrand, and A. Wasowski, "42 variability bugs in the linux kernel: a qualitative analysis," in *Proceedings of the 29th ACM/IEEE international conference on Automated software engineering (ASE)*, ASE '14, 2014.
- [20] A. Kenner, C. Kästner, S. Haase, and T. Leich, "Typechef: Toward type checking #ifdef variability in c," in *Proceedings of the 2nd International Workshop on Feature-Oriented Software Development (FOSD)*, FOSD '10, 2010.
- [21] A. Ziegler, V. Rothberg, and D. Lohmann, "Analyzing the impact of feature changes in linux," in *Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems*, pp. 25–32, 2016.
- [22] C. Kästner, "Differential testing for variational analyses: Experience from developing kconfigreader," *arXiv preprint arXiv:1706.09357*, 2017.
- [23] R. Duan, A. Bijlani, Y. Ji, O. Alrawi, Y. Xiong, M. Ike, B. Saltafornaggio, and W. Lee, "Automating patching of vulnerable open-source software versions in application binaries.," in *NDSS*, 2019.