École Polytechnique Fédérale de Lausanne

Recovering type information from compiled binaries to aid in instrumentation

by Louis Merlin

Master Thesis

Approved by the Examining Committee:

Prof. Dr. sc. ETH Mathias Payer
Thesis Advisor

Damian Pfammatter
External Expert

Antony Vennard
Thesis Supervisor

EPFL IC IINFCOM HEXHIVE
BC 160 (Bâtiment BC)
Station 14
CH-1015 Lausanne

July 19, 2021

Nobody tosses a dwarf !
— Gimli, son of Glóin

TODO Dedication

Acknowledgments

TODO Acknowledments

Lausanne, July 19, 2021 Louis Merlin

Abstract

The abstract serves as an executive summary of your project. Your abstract should cover at least the following topics, 1-2 sentences for each: what area you are in, the problem you focus on, why existing work is insufficient, what the high-level intuition of your work is, maybe a neat design or implementation decision, and key results of your evaluation.

Contents

Chapter 1

Introduction

Work on C++ began in 1979, as a "C with classes" [20]. Since then, the language has grown in popularity, and even surpassed C itself [18]. Examples of well-known C++ projects include the zoom [4] conferencing software, the gold linker [21] or [TODO: find another "well-known" example]. Nevertheless, reverse-engineering efforts have been focused towards C binaries, because analysis methods found this way will often work on C++ binaries too.

This has meant reverse engineering tools like ghidra [13], IDA Pro [16] or Binary Ninja [1] have treated non-C binaries as second-class citizen. These tools will often show C++ specific features as passing comments, failing to show the real implications of a try/catch block or a polymorphic class.

The blame can mostly be put on the complexity of C++ when compared to C. Whereas C translates quite naturally to assembly, abstractions specific to C++ require more work and complexity to be translated to asm. This also leads to important information being lost from C++ source code to binary, but also certain information remaining.

The recent RetroWrite [7] project by the HexHive lab is a static rewriting tool for x86_64 position-independent binaries. It enables the instrumentation of projects when we do not have access to the source code. This can include a legacy project, a closed-source product or even malware.

In this thesis we would like to present the **dis-cover** [11] static analysis tool, as well as improvements made to RetroWrite to support C++ binaries.

The dis-cover tool is able to extract information from a C++ binary, and re-inject it as debug information using the DWARF format into the binary. This enables other debugging tools to see and display this information.

[IF WE SUCCEEDED] In this thesis we would like to show how we brought C++ support to RetroWrite, and what research opportunities this will create.

[IF WE FAILED] In this thesis we will detail how we tried to bring C++ support to RetroWrite, and what remains to be done for the implementation to work.

Chapter 2

Background

2.1 C++ and polymorphism

The C++ programming language implements polymorphism. This feature enables complex code logic that can comply with external business logic for example. Polymorphic classes are defined by having at least one virtual method, which is inheritable and overridable. With this polymorphism comes type conversion, and more interestingly for us the **dynamic_cast** [5] expression. Dynamic casting is a safe kind of type conversion. It only will successfully cast if the value is of the referenced type or a base type of that type. In order to achieve that, the system must have some kind of information about the object's data type at run time.

This is where Run-Time Type Information (**RTTI**) comes into the picture. The system will use this RTTI to infer type inheritance for dynamic casting. We will now go into implementation details of RTTIs and VTables, which point to them.

To make RTTIs appear in your C++ binary, you will to define classes that inherit from each other, as well as at least one virtual method in one of these classes. Figure 2.1 shows an example of such classes.

```
class Animal {
public:
  virtual void speak() {}
};

class Dog {
public:
  virtual void speak() { cout << "Woof" << endl; }
}
```

Figure 2.1: Polymorphic classes in C++

For RTTIs to appear in your C++ binary, you will also need to instantiate these classes, and have some run time logic to make the binary non-deterministic [this not the right word I think, I mean to say that the binary depends on some input for its class-instantiating logic]. For example, a conditional dynamic cast between the two classes we defined in Figure 2.1. If you do not do these things, the class logic will be abstracted away by the compiler for optimization reasons.

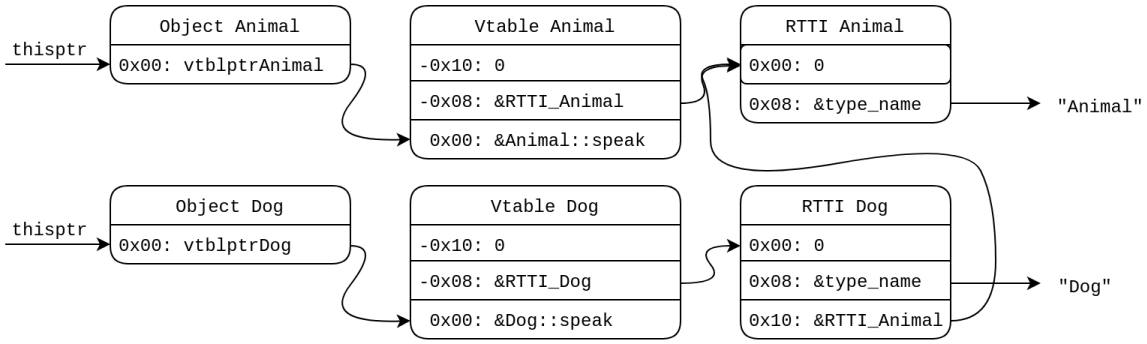

Figure 2.2: Overview of an example of VTables and RTTI in memory

The structure of VTables and RTTIs is detailed in Figure 2.2. All of this is defined in the Itanium C++ ABI [9]. An instance of a virtual class Animal will contain the **vtblptrAnimal**, a pointer to the virtual table (**VTable**) for the Animal class. This VTable will contain pointers to the virtual methods of the class, which are inherited and thus do not have to be duplicated for each class, they are simply pointed to. The first value preceding the VTable is a pointer to the RTTI of the class. The process finds out the class inheritance of a class instance by following this pointer.

The RTTI itself is composed of an offset (used for complex class inheritance structures) as well as a pointer to the type name (this type name is not removed by simple stripping of the binary, like with **objdcopy –strip-all**). This name is mangled using C++ mangling, and can trivially be demangled. This name is what we use to uniquely identify a class in a project (and make sure we have no duplicates). The next values of the RTTI are pointers to the RTTIs of the parent classes. See Figure 2.2 for an example, with the Dog RTTI containing a pointer to the Animal RTTI.

2.2 DWARF debugging standard

DWARF is the debugging standard used widely in conjunction with executable ELF files. It is often included in these ELF files in the **.debug_info** section (and other related sections). This debug information is made up of Debugging Information Entries (**DIEs**). These entries can contain information about variable names, method definitions, the compilation process, or more importantly for us, class names and class inheritance.

2.3 RetroWrite

Chapter 3

Design

Class recovery from compiled C++ binaries has been done multiple ways before: there are plugins for popular debugging tools, like ida_gcc_rtti [12] for IDA Pro, or Ghidra-Cpp-Class-Analyzer [19] for Ghidra; there are academic projects such as MARX [15] that rely on heuristics around VTables to find classes and inheritance information. The MARX project can recover classes when no RTTI is present (in the Chrome project for example, where the **fno-rrti** compilation flag is used), but not with 100% accuracy.

For this project we decided to focus on RTTIs which, as mentioned in the Background chapter, are available for a class hierarchy tree when there is at least one virtual method implemented by one of the classes. We decided to test if RTTIs occurred often in the wild : we proceed to download every Debian package that listed C++ as a dependency (get the list with **apt-cache rdepends libgcc1** on a Debian machine). Out of around 80'000 packages, 5827 of them list C++ as a dependency. Out of those, we extracted classes from 3194 (54%). We will share more details about this experiment in the Evaluation chapter.

We also considered extracting information about exceptions, which is another big feature that differentiates C++ from C. We decided not to focus on this feature, as it did not seem to bring as much to the project considering the evaluated development time. We did have to study and work with exceptions and exception handling frames for the augmentation of RetroWrite (see the end of the Implementation chapter).

Next comes the question : what should we do with all of this information ? What would be the most useful format to output the data in ? This is where DWARF [3] comes in the picture. DWARF is the debugging standard for programs. It is mostly used by developers trying to understand where their implementation fails, and by reverse engineers to get a better understanding of how a program was conceived (although DWARF information is usually stripped from proprietary software). This kind of debug data is mostly found in C and C++ projects. By having DWARF data as an output, the information would become readable by most modern reversing tools.

There is a current push for DWARF to become the lingua franca for reverse engineering tools, lead by researchers like Dr. Sergey Bratus [6] from DARPA. He has published a paper in 2011 where he was able to exploit certain features of DWARF to control program execution : Exploiting the hard-working DWARF [14].

With dis-cover, we are able to inject information in the debug and symbol sections of the binary, creating a new ELF file with all of this useful info included. We will go more into the implementation details in the next chapter.

Another focus of this project was the augmentation of RetroWrite for C++ capabilities. Today, RetroWrite supports the reversing of x86_64 position-independent binaries. There was also work to augment RetroWrite to support kernel code [17] and arm_64 [2]. We aimed to add C++ capabilities for x86_64 binaries only, as it seems like the biggest target.

Another potential use of the class information is to implement defenses around type confusion, working with the findings from the HexType [10] paper. We decided not to focus on this for our project, as it would have been out of the scope of a 6-months master thesis.

One of the defining properties of RetroWrite is that it does not use heuristics (and thus does not have probabilistic features and failures). Dis-cover has the same no-heuristics property, and thus could be used by the RetroWrite project without breaking the property.

Chapter 4

Implementation

We decided to write a python module for this project, as the python ecosystem has great reverse engineering packages, and because the other projects from the HexHive lab are usually python modules.

4.1 Finding RTTIs

The first thing dis-cover does is find ELF sections where the VTables and RTTIs could be hiding. This is usually **.rodata** (read-only data), but could sometimes also be related sections like **.data.rel.ro**, **.data.rel.ro.local**, or **.rdata**. As per the Itanium C++ ABI, the base VTable of a class will contain an "offset-to-top", which will be 0 in the primary base virtual table's case, followed by a pointer to an RTTI. We simply have to pattern-match for zeroes directly followed by a pointer to another part of the read-only data sections, and we have a potential RTTI pointer.

To check whether we have found an RTTI, we assert that the next value is a pointer to a string located in a read-only data section. If it is, we can extract it, demangle it, and we have a class and a name. The next values in the RTTI are pointers to the RTTIs the class inherits from. We can go through them and parse them if we have not already, to add this inheritance information to the original class.

This algorithm is $O(n)$, as adding a class only adds one more value to parse.

4.2 Creating DWARF data

Next, we want to add that information to the debug sections of a new ELF file.

In order to write DWARF data, the first step is defining the types we will be using, and their fields. This is done by writing bytes in the **.debug_abbrev** section. For example, we create an abbrev of type **class_type**, which has a **name**, and can have sub-field (children). Then, we create the abbrev of type **inheritance**, which has a **type** (a reference to the parent type).

We can then populate the **.debug_info** with classes and their inheritance data. DWARF data takes the form of a tree of values. We have to create a **compile_unit** value at the root, and then the branches will be **class_type**s. These **class_type**s will themselves have as children **inheritance** values if the class inherits from another class. Figure 4.1 shows a very simple example of this, with two class types and one inheriting from the other.

```
< 1><0x0000001a>    DW_TAG_class_type
                      DW_AT_containing_type        <0x0000001a>
                      DW_AT_calling_convention     DW_CC_pass_by_reference
                      DW_AT_name                   Shape
                      DW_AT_byte_size              0x00000008
< 1><0x00000026>    DW_TAG_class_type
                      DW_AT_containing_type        <0x00000026>
                      DW_AT_calling_convention     DW_CC_pass_by_reference
                      DW_AT_name                   Rectangle
                      DW_AT_byte_size              0x00000008
< 2><0x00000031>      DW_TAG_inheritance
                        DW_AT_type                   <0x0000001a>
```

Figure 4.1: Extract of a dwarfdump output showing simple inheritance

The strings themselves are stored in another section, **.debug_str**, and are referred to with their offset in that section.

4.3 Creating symbols

Symbols are used to have access to variable names, class names, function names or any other kind of text information when debugging a binary. We want to create two symbols for every class we have found during the analysis : one pointing to the VTable and one pointing to the RTTI, labeling them as such.

In order to create new symbol sections, we take the symbol table from the original binary (if there was one) and append the aforementioned symbols.

The two symbol sections are **.symtab**, which contains the information (offset, size, type, ...) for each symbol, and **.strtab**, which contains the strings related to these symbols.

4.4 Wrapping things together

4.4.1 Creating a fake ELF file

Once we have the three debug sections ready (**.debug_abbrev** with the debug types, **.debug_info** with the debug information, and **debug_str** with the strings) as well as the two symbol sections, we have to make them available to the user. The first step is building an empty ELF file with only these five sections in them.

We start by constructing a **program header table**. This table contains information about the offset and size of each segment of the binary (which segment is used for what, and their read/write permissions). The ELF file we're creating will not be run, but only used temporarily. Thus, we noticed that we did not have to create a valid program header table for the process to work. We simply copy this program header table almost as-is from the original binary.

Next, we use the individual sections we built earlier and construct the **section header table**. For every section present in the original binary, we create an entry in the section header table, reusing most values from the original section header table. The only values we modify is the offset. For every section that we have created, we add the appropriate row in the section header table. Every section name gets added to the **.shstrtab** section, as per the convention.

Finally, we construct the **elf header**, taking some of the values from the original binaries, and calculating some others from the size of the tables and sections we have built. We can now create a fake ELF file by appending the elf header, the program header table, the sections we built and the section header table.

4.4.2 Stripping the original ELF file

Next, we will create a stripped version of the original ELF file. We use the **objcopy –strip-all** command. This is to avoid section conflicts in the next step.

4.4.3 Combining the two ELF files

Now, we can use the ELF utility program **eu-unstrip** to combine the two ELF files we have created into one. The newly created combined ELF file will contain all of the code and data from the original file, as well as the debug and symbol sections we have created.

Chapter 5

Evaluation

5.1 Small case studies

In addition to the first version of dis-cover, we created three small programs highlighting different features of C++. One was using simple inheritance, one had a namespace (which we can and should recover as part of the analysis), and the last one was a use case of multiple inheritance (using the diamond problem).

We also created a script that would compile these 3 programs using different levels of compiler optimization. We could then use dis-cover to see if we could recover every class and the correct tree from the binaries. This served as a useful benchmark to check whether dis-cover was functioning correctly if we tried to apply changes to it.

These examples alone were not capable of letting us evaluate the full capabilities of dis-cover. We found a few big applications that could serve that purpose.

5.2 Big case studies

The first and smallest of these case studies was the **gold linker** [21]. We were able to find 571 classes in version 1.16 of the program. This provides a good benchmark, but the classes themselves do not make use of multiple inheritance (only a "simple" inheritance tree).

LibreOffice [8] on the other hand provided us with a great test case : the program is fragmented into many small libraries, containing some interesting uses of multiple inheritance. See Figure 5.1 for an example of multiple inheritance in the **libloglo.so** library from LibreOffice. This particular example was very important for verifying a big bug that was present in an early version of dis-cover. After fixing the bug, we were noticing around 10% more inheritance links in some projects

(but not more classes). Being able to go check with the open-source LibreOffice code that we had found the right inheritance links was extremely helpful.

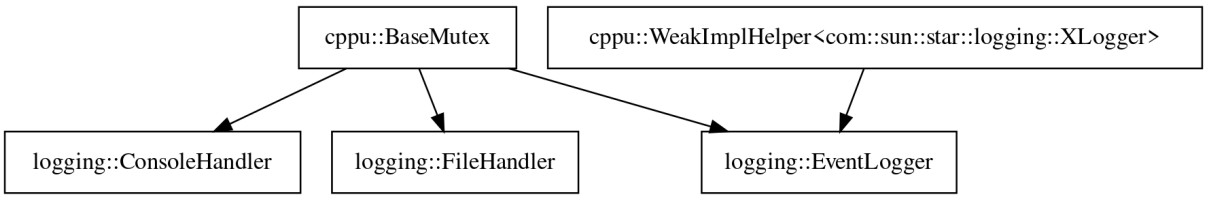

Figure 5.1: Partial class tree of libloglo.so from LibreOffice

Finally, we also were able to study the closed-source **zoom** [4] binary. This test case is very interesting in two ways. First, we are able to find **6039 classes** in the binary, with **5601 edges** in the class hierarchy graph. Second, as a large (80M) and complex ELF file, it served as a perfect benchmark for the performance of the algorithm. By adding a simple check earlier in the RTTI-spotting pipeline, we were able to speed up the analysis of the zoom binary 10 times. See Table 5.1 for the details of this benchmark.

ELF File and version	Old Algorithm	New Algorithm	Speedup Factor
zoom 5.5.7938.0228	52m42.123s	5m22.321s	10×
gold 1.16	0m5.123s	0m1.321s	5×
ceph-dencoder 15.2.13	1m30.123s	0m9.123s	10×

Table 5.1: Difference of computation time between algorithms on a machine running Ubuntu 20.04 with a 2.6 Ghz dedicated vCPU

As we have shown before, out of the 5827 Debian packages that list C++ as a dependency, we were able to extract classes from 3194 (54%). The total number of classes we found is 960'188, with 39% of them unique across all packages (unique name in unique tree).

Chapter 6

Related Work

The related work section covers closely related work. Here you can highlight the related work, how it solved the problem, and why it solved a different problem. Do not play down the importance of related work, all of these systems have been published and evaluated! Say what is different and how you overcome some of the weaknesses of related work by discussing the trade-offs. Stay positive!

This section is usually 3-5 pages.

Chapter 7

Conclusion

In the conclusion you repeat the main result and finalize the discussion of your project. Mention the core results and why as well as how your system advances the status quo.

Bibliography

[1] Vector 35. *Binary Ninja*. https://binary.ninja/. Accessed: 2021-06-06.

[2] Luca Di Bartolomeo. *ArmWrestling: efficient binaryrewriting for ARM*. http://hexhive.epfl.ch/theses/20-dibartolomeo-thesis.pdf. Accessed: 2021-06-15.

[3] DWARF Standards Committee. *The DWARF Debugging Standard*. http://www.dwarfstd.org/. Accessed: 2021-06-14.

[4] Zoom Video Communications. *Zoom Cloud Meetings*. https://zoom.us/. Accessed: 2021-06-09.

[5] cppreference.com. *dynamic_cast conversion*. https://en.cppreference.com/w/cpp/language/dynamic_cast. Accessed: 2021-06-10.

[6] DARPA. *Dr. Sergey Bratus*. https://www.darpa.mil/staff/dr-sergey-bratus. Accessed: 2021-06-15.

[7] Sushant Dinesh, Nathan Burow, Dongyan Xu, and Mathias Payer. "RetroWrite: Statically Instrumenting COTS Binaries for Fuzzing and Sanitization". In: *IEEE International Symposium on Security and Privacy*. 2020.

[8] The Document Foundation. *LibreOffice*. https://www.libreoffice.org/. Accessed: 2021-06-19.

[9] *Itanium C++ ABI*. https://itanium-cxx-abi.github.io/cxx-abi/abi.html. Accessed: 2021-06-15.

[10] Yuseok Jeon, Priyam Biswas, Scott Carr, Byoungyoung Lee, and Mathias Payer. "HexType: Efficient Detection of Type Confusion Errors for C++". In: *ACM Conference on Computer and Communication Security*. 2017.

[11] Louis Merlin. *dis-cover*. https://github.com/HexHive/dis-cover. Accessed: 2021-06-06.

[12] mwl4. *ida_gcc_rtti*. https://github.com/mwl4/ida_gcc_rtti. Accessed: 2021-06-12.

[13] NSA. *ghidra*. https://ghidra-sre.org/. Accessed: 2021-06-06.

[14] James Oakley and Sergey Bratus. "Exploiting the hard-working DWARF: Trojan and Exploit Techniques With No Native Executable Code". In: *Usenix Workshop on Offensive Technologies*. 2011.

[15] Andre Pawlowski, Moritz Contag, Victor van der Veen, Chris Ouwehand, Thorsten Holz, Herbert Bos, Elias Athanasopoulos, and Cristiano Giuffrida. “MARX: Uncovering Class Hierarchies in C++ Programs”. In: *Network and Distributed System Security Symposium*. 2017.

[16] Hex Rays. *IDA Pro*. `https://hex-rays.com/IDA-pro/`. Accessed: 2021-06-06.

[17] Matteo Rizzo. *Hardening and Testing Privileged Code through Binary Rewriting*. `http://hexhive.epfl.ch/theses/19-rizzo-thesis.pdf`. Accessed: 2021-06-15.

[18] stackoverflow. *Most Popular Technologies*. `https://insights.stackoverflow.com/survey/2020#most-popular-technologies`. Accessed: 2021-06-05.

[19] Andrew Strelsky. *Ghidra C++ Class and Run Time Type Information Analyzer*. `https://github.com/astrelsky/Ghidra-Cpp-Class-Analyzer`. Accessed: 2021-06-12.

[20] Bjarne Stroustrup. *When was C++ invented?* `https://www.stroustrup.com/bs_faq.html#invention`. Accessed: 2021-06-05.

[21] Ian Lance Taylor. *Gold Linker*. `https://en.wikipedia.org/wiki/Gold_(linker)`. Accessed: 2021-06-09.