Source Code Obfuscation: A Technique for Checkmating Software Reverse Engineering

Edward E. Ogheneovo and Chidi K. Oputeh,

1,2Department of Computer Science, University of Port Harcourt, Port Harcourt, Nigeria

ABSTRACT: With the widespread use of personal computers and advent of the Internet technology, computer security has suffered a lot of setbacks. Computer users especially hackers these days are finding ways to get access to certain information on the Internet which ordinarily they suppose to have use passwords to get access to. The result is that these scrupulous users bypass username and passwords to access information that are not well protected. One such problem is software cracking. Attackers and hackers bypass protection copyright, download software and use them without permission. Software cracking is the process of bypassing the registration and payments options on a software product to remove copyright protection safeguards or to turn a demo version of software into a fully function version without paying for it. In this paper, we used source code obfuscation technique to checkmate software reverse engineering which is one of the techniques hackers often used to crack software. In this technique, the original code is mangled such that the resultant code, the obfuscated code, is difficult to analyze by the hackers, although; the code retains its originality. Our result shows that the obfuscated code cannot be easily decoded. Hence, it prevents hackers from cracking the code.

KEYWORDS: Software obfuscation, reverse engineering, software cracking, obfuscated code

I. INTRODUCTION

Security is a very crucial aspect of computer science and it plays a major role in the advancement of Information Technology especially with the growing use of computer software for e-commerce, e-banking and for governance in all areas of life where national security is a very critical issue [10]. It is therefore relevant to study this area with increased efforts so that software produced and made available for the general public can be more secure and reduce the possibilities of being attacked by crackers or hackers [7][8]. This will give more confidence to the software developers and manufacturers to continue to create and develop software knowing that their software are saved and cannot be hacked or cracked. However, if hackers decide to hack the software, they can be sure that their software can stand the test of time. Code obfuscation [3][11] focuses on the protection of a software program and the content that the program protects. There have been billions of dollars spent each year by the industries especially for software piracy and digital media piracy. The achievement of the content/software security in a huge segment is based on the ability of protecting software code against tampering and identifying the attackers who issue the pirate copies [6][9]. Linn and Debray [12] concentrates on the attacker identification and forensic examination. The author discussed about a proactive detection approach for defeating an on-going attack before the cooperation has occurred. The author also describes another detection approach for post compromise attacker identification. Especially, the author takes into account the real world scenarios where the application programs connect with their vendors so often, and a discovery of attack can bar a hacker from further business. Code obfuscation focuses to protect code against both static and dynamic study and there exists another approach to protect against code analysis, namely self-modifying code [4]. This approach provides the opportunity to create code at runtime, rather than changing it statically. Practically, self-modifying code is highly restricted to the monarchy of viruses and malware. Yet, some publications regard self-modifying code as an approach to protect against static and dynamic analysis. Madou et al. [13][14] for instance regard dynamic code generation. The author proposed an approach where functions are generated earlier to their first call at runtime. Moreover, clustering is presented in such a way that a general template can be utilized to generate each function in a cluster, carrying out a least amount of alterations. The decryption at runtime technique is equal with code generation, apart from the fact that the decryption key can depend on other code. Moreover, it lessens re-encrypting the viability of code during execution [15][16]. However, the technique does not clearly protect a function template after the function executed.

II. REVERSE ENGINEERING PROCESS

The reverse engineering process [2] begins by extracting detailed design information, and from that extracting a high-level design abstraction. Detailed (low-level) design information is extracted from the source code and existing design documents. This information includes structure charts and data descriptions to describe processing details. The high-level design representation is extracted from the recovered detailed design and
expressed using data-flow and control-flow diagrams. The recovered design is the same as the extracted design. The procedure steps are discussed below. Figure 1 summarizes the procedure.

- **Collect information**: Collect all possible information about the program. Sources of information include source code, design documents and documentation for system calls and external routines. Personnel experienced with the software should also be identified.

- **Examine information**: Review the collected information. This step allows the person(s) doing the recovery to become familiar with the system and its components. A plan for dissecting the program and recording the recovered information can be formulated during this stage.

- **Extract the structure**: Identify the structure of the program and use this to create a set of structure charts. Each node in the structure chart corresponds to a routine called in the program. Thus the chart records the calling hierarchy of the program. For each edge in the chart, the data passed to a node and returned by that node must be recorded.

- **Record functionality**: For each node in the structure chart, record the processing done in the program routine corresponding to that node. A Program Design Language (PDL) can be used to express the functionality of program routines. For system and library routines the functionality can be described in English or in a more formal notation.

- **Record data-flow**: The recovered program structure and PDL can be analyzed to identify data transformations in the software. These transformation steps show the data processing done in the program. This information is used to develop a set of hierarchical data flow diagrams that model the software.

- **Record control-flow**: Identify the high-level control structure of the program and record it using control-flow diagrams. This refers to high-level control that affects the overall operation of the software, not to low-level processing control.

- **Review recovered design**: Review the recovered design for consistency with available information and correctness. Identify any missing items of information and attempt to locate them. Review the design to verify that it correctly represents the program.

- **Generate documentation**: The final step is to generate design documentation. Information explaining the purpose of the program, program-overview, history, etc, will need to be recorded. This information will most probably not be contained in the source code and must be recovered from other sources [2].

---

**Fig. 1**: Reverse engineering procedure  
**Source**: Chikofsky and Cross [2]
implies that an attacker has full control over the host system and can therefore perform static and/or dynamic analysis techniques at will.

**Static Analysis:** This technique is applied on non-executing code and comprises static disassembling and subsequent static examination steps. Disassembling code is usually done using either linear sweep or recursive traversal. Linear sweep simply scans over the code, disassembling instructions, assuming that every instruction is followed by another instruction. GNU’s (UNIX-like Operating system) gdb (GNU debugger) [16] uses this technique. Recursive traversal takes control flow into account. However, as some branches are input-dependent, usually not all target addresses can be statically derived and disassembled. Linear sweep is easily fooled by inserted data bytes [12], while recursive traversal often gives incomplete results. As a reaction, Schwarz et al. [16] propose to use a combined approach. Additionally, Krügel et al. [11] treat every memory address as a potential start of an instruction. Consequently, they filter out overlapping assembly results based on a control flow graph based approach. Additionally to disassembly, a decompilation step could map low-level code to more high-level constructs [3]. These high-level structures might facilitate human inspection more than abstract assembly listings. For certain languages such as Java or .NET, it is easy to decompile bytecode into source code [15]. Even if static analysis cannot derive all control flow, it is considered more complete than dynamic analysis as it examines all possible paths while dynamic analysis only considers the executed path. Figure 2 gives a schematic overview of the disassembly and decompilation phase for reverse engineering code. Recovered high-level code aims to approximate the original source code.

**Dynamic Analysis:** Dynamic techniques are performed on executing code. This involves tracing of executed instructions, register contents, data values, etc. While this type of attack is more powerful than a static attack, it may be more time consuming or more complex. First, it requires a platform similar to that of the target code. Secondly, a program might be equipped with anti-debugging techniques hindering dynamic analysis. Debugging and emulation belong to the dynamic analysis. Commercial debuggers include SoftICE [1] and IDA Pro [5].

### III. METHODOLOGY

Code obfuscation attempts to transform a program into an equivalent one that is more difficult to manipulate and reverse engineer. Code obfuscation attempts to make the task of reverse engineering a program daunting and time consuming. This is done by transforming the original program into an equivalent program, which is much harder to understand, using static analysis. Considering our obfuscation mechanism, we define obfuscation as follows:
Let $T(P)$ be program, transformation of program $P$. $T$ is an obfuscating transformation, if $T(P)$ has the same observable behavior as $P$. In addition $T$ must follow conditions:

- if program $P$ fails to terminate or terminates with an error condition, then $T(P)$ may or may not terminate
- otherwise $P$ terminates and $T(P)$ must terminate and produce the same output as $P$

![Obfuscation Mechanism Module](image)

In this mechanism, the encrypted serial number generation code segment and the other code segments are, both, subjected to obfuscation. This is done by writing an obfuscation routine within the source code that will be responsible for applying the transformation on the source code itself. Since the entire source code will be obfuscated, it is very difficult for any cracker or hacker to crack or have access to the internals of the program. Any cracker will lose the interest to crack any program that our obfuscation mechanism is applied on. This is because the sight of the program code after been obfuscated is mangled that it does not follow the normal chronological order of writing a program neither does it look like any other program seen around.

The technique involves lexical, control and data transformations. Lexical transformations alter the actual source code, such as C code. This transforms the original source code into a lexically equivalent form by mangling names and scrambling identifiers. Such transformations make it a daunting task to reverse engineer a program. Control transformations alter the control flow of the program by changing branch targets to an ambiguous state. The code for the program is shuffled such that the original branch targets are no longer correct. During this shuffling, the new targets are calculated, and code is inserted in place of the old branch instruction to acquire its new target address. Data transformations rearrange data structures such that they are not contiguous. Data can be transformed all the way down to the bit level.

Our obfuscation mechanism utilizes an algorithm for the obfuscation of the serial number generation code segment and the other code segment. The pseudo code is written below:

**Pseudo code 1: Obfuscation Mechanism**

1. Set string literals to array of numbers
2. Extract integer address of the array
3. Add index variable to the arguments list of a recursive function
4. Extract 3 lower case l's out of the array and insert a Boolean statement
   - IF the index variable is at the proper position of the array THEN
   - print T
   - END IF
Add a second index to keep track of l’s that is printed
Subtract the number from the index variable to access the real array element
Shift around the location of the conditional operators
WHILE switching some of the integers for character literals
RENAME a few of the variables
END WHILE
RENAME functions to something similar to variable names
GET RID of all unnecessary argument type specifiers
MANIPULATE the formatting to obscure the flow of conditional operators and mask distribution between function name and main

Our obfuscation mechanism utilizes an algorithm for the obfuscation of the serial number generation code segment and the other code segment. This is shown in algorithm 1.

Algorithm 1: Obfuscation Mechanism

1. Init sl = strings
2. Init ia = address of sl
3. Init al = argument list of recursive function
4. FOR x = 1 to length of al
5. al = al + sl[x]
6. END FOR
7. Insert 3 lower case l’s into a Boolean statement
8. FOR x = 1 to length of al
9. IF x is at the proper position of sl, then
10. print ‘l’
11. END IF
12. Init al[x+1] = printed “l”
13. SUBTRACT n from x to get element of sl
14. SWAP conditional operators
15. WHILE (Swapping between integers and characters)
16. RENAME variables
17. END WHILE
18. END FOR
19. RENAME functions to look like variable names
20. ELIMINATE argument type specifiers

IV. RESULTS AND DISCUSSIONS

We cracked software (crackme.exe) that has a serial key authentication attribute. We want to attempt to use the same cracking rules to crack the hybridized self-modifying mechanism. In doing this, we will have some program generated outputs to show the results as we undergo the cracking processes. The hybridized self-modifying code is compiled and the executable named ObfusSec.exe is generated at compilation time. We start by running the source code to check the serial number parameters. Figure 4 shows the assembly language representation of the source code used for implementing this work. The figure shows the offset location of the jmp call. When the crackme.exe is run, it usually prompt the user to enter a serial number.

![Fig. 4: Hiew displaying the Offset location of the jmp call](image-url)
In Figure 5, the C++ source code of our model is opened using the Code::Blocks 13.12 IDE. We run the code and observed the behavior it displays. The resulting code is not readable and shows that the code has been obfuscated and converted to the form that is difficult to understand by someone intending to crack the code.

![Fig. 5: The source code of ObfusSec.exe](image)

In Figure 6, the C++ source code of our model is built and compiled. Then, a screen display showing an authentication scheme pops up. You are required to enter your key. If a wrong key is entered, it will display “invalid” showing that the key is not the correct one. This way the software cannot be cracked. However, if a key different from the correct key is entered and the word “invalid” is not displayed, it means that the code has been cracked.

![Fig. 6: Screen display to enter key](image)

![Fig. 7: Screen display showing the text string "Invalid Key, please try again"](image)
In Figure 7, we opened our executable file, created by running our C++ source code, in the Hacker's Disassembler. We then searched for the text string "Invalid Key, please try again". We discovered that the text string cannot be found in the executable file. The reason for this is because the serial number authentication scheme has been encrypted and obfuscated. So the text string can never be seen by the cracking tool. This means that the executable file cannot be cracked since every cracking process requires that you should search for the displayed text strings after entering the wrong serials before cracking can take place. Without detecting the text string generated by entering the wrong key, we cannot know the offset address of the jump call required for modification of the executable to take place. This means our hybridized self - modifying mechanism cannot be cracked. Thus the wrong key is typed and the text strings "Invalid key, you may need to try again" is shown on the screen display. This text string is very important to us because we need it to commence our cracking processes. We have to search for the text string "Invalid key, you may need to try again " using the search button at the top of the dialogue box in the Hacker's Disassembler. If we can locate the text string in the hacker's disassembler, then we can proceed with the cracking processes.

V. CONCLUSION

Security is a very crucial aspect of computer science and it plays a major role in the advancement of Information Technology especially with the growing use of computer software for e – commerce, e – banking and for governance in all areas of life where national security is a very critical issue. It is therefore relevant to study this area with increased efforts so that software produced and made - available for the general public can be more secured and reduce the possibilities of being attacked by crackers or hackers. In this paper, we used source code obfuscation technique to checkmate software reverse engineering which is one of the techniques hackers often used to crack software. Using this technique, the original code is mangled such that the resultant code, the obfuscated code, is difficult to analyze by the hackers, although; the code retains its originality. We discovered that the text string cannot be found in the executable file. The reason for this is because the serial number authentication scheme has been encrypted and obfuscated. So the text string can never be seen by the cracking tool. This means that the executable file cannot be cracked since every cracking process requires that you should search for the displayed text strings after entering the wrong serials before cracking can take place. Without detecting the text string generated by entering the wrong key, we cannot know the offset address of the jump call required for modification of the executable to take place. This means our hybridized self - modifying mechanism cannot be cracked.

REFERENCES

Source Code Obfuscation: A Technique For...

Appendix: The Obfuscated version of the C++ code

```c++
// ObfuscSec.cpp
#include <stdlib.h>
#include <stdio.h>
#include <iostream>
#include<iostream>
#include<fstream>
#include <iomanip>
#include <sstream>
#include <windows.h>
#define O int
#define __OxCXs return
#define O_Ox__ }
#define __O_ {
#define Ax0xBEGIN void
#define _O_ Ax0xBEGIN
#define _C_ char
#define OxcC_ _C_
#define OxK__0x654hu string
#define _010011101 bool
#define oxFEABD if
#define o0Oo while
#define o0 cout
typedef unsigned __int64 uint64_t;
using namespace std;
OxK__0x654hu Ox11(ifstream&);
OxK__0x654hu _O__(OxK__0x654hu toEncrypt);
class HashInfo
__O_ OxC_ _hashval[100];OxcC_ configHash[100];O run_state;OxcC_ appkey[20];
public:Ax0xBEGIN setHashVal(OxK__0x654hu);OxcC_ * getHashVal();Ax0xBEGIN setState(O);
O getState();Ax0xBEGIN setConfigHash(OxK__0x654hu);OxcC_ * getConfigHash();
Ax0xBEGIN setAppKey(OxK__0x654hu);OxcC_ * getAppKey();O_Ox__;
Ax0xBEGIN HashInfo :: setHashVal(OxK__0x654hu hashv)__O_ strncpy(hashval, hashv.c_str(), sizeof(hashval));hashval[sizeof(hashval) - 1] = 0;}
Ax0xBEGIN HashInfo :: setState(O)
__O_ run_state = state;}
O getHashInfo :: getState()__O_ __OxCXs run_state;O_Ox__;
OxcC_ * HashInfo :: getHashVal()__O_ __OxCXs hashval;O_Ox__
Ax0xBEGIN HashInfo :: setConfigHash(OxK__0x654hu hashv)
__O_ strncpy(configHash, hashv.c_str(), sizeof(configHash));configHash[sizeof(configHash) - 1] = 0;O_Ox__
OxcC_ * HashInfo :: AppConfigHash(OxK__0x654hu configKey)
__O_ strncpy(configKey, configKey, sizeof(configKey));configKey[sizeof(configKey) - 1] = 0;O_Ox__
Ax0xBEGIN HashInfo :: setAppKey(OxK__0x654hu appKey)
```
Source Code Obfuscation: A Technique For...

```
O MAX(O x, O y)__O_  oxFEABD  ((x) > (y))__OxCXs x;else
__OxCXs y;O_Ox__
//---Read more---

C_ ox11L[10] = __O_ \n__O_Ox_ 0'o70'BE-0xBD1',\n__O-Ox_ _obfuscated2'123',\n__O-Ox_ 0x1234';A1h2c#d45'000000006'0xuiif7';_8',_XXxXXX9'\nO_Ox_--__0xEEABD();A0xBEGIN___sxBEGIN(O,O_OxC_--_0);\n__O_01(O_Ox_;O_01(O_OxK_0x654hu);\nO main(O argc, const OxC_ *argv[])__O_ HashInfo sobj;\n_010011101 rit_new_hash = false;ifstream in("config.dat");\noxFEABD (in.is_open())__O_ o0Oo(!in.eof())__O_ in.read((OxC_*) &sobj, sizeof(sobj));\nxOFEABD (in.eof())__O_ OxK_0x654hu mhash; oxFEABD (sobj.getState()==0)__O_ O__
OxK_0x654hu chash; stringstream cst;\no0 << sobj.getHashVal(); OxK_0x654hu __O_ oxFEABD (sobj.getAppKey())
OxK_0x654hu new_computed_hash = mhash; sobj.setHashVal(new_computed_hash);
rit_new_hash = true;O_Ox__ else __O_
OxK_0x654hu f2.open(argv[0], ios::in | ios::binary | ios::ate);
oxFEABD (!f2.is_open())__O_ cerr << " Error opening main executable file 2 n" << endl;
OxK_0x654hu close();\nOxK_0x654hu new_computed_hash = mhash; sobj.setHashVal(new_computed_hash);
OxK_0x654hu ifstream f2; f2.open(argv[0], ios::in | ios::binary | ios::ate);
OxK_0x654hu ifstream f2; f2.open(argv[0], ios::in | ios::binary | ios::ate);
```

```
O player = 1,choice; C_mark;do __O_ _xBEGIN(0xE0F > 0 ? 0 : 0xFF , 0xAB);\nplayer=(player%2)?1:2; o0 << "Player " << player << "", enter a number: " ;cin >> choice;
else oxFEABD (choice == 5 && ox11L[5] == '5')ox11L[5] = mark;else oxFEABD (choice == 6 && ox11L[6] == '6')ox11L[6] = mark;else oxFEABD (choice == 7 && ox11L[7] == '7')ox11L[7] = mark; else oxFEABD (choice == 8 && ox11L[8] == '8')ox11L[8] = mark; else oxFEABD (choice == 9 && ox11L[9] == '9')ox11L[9] = mark; else __O_ o0<<"Invalid move ";player--;cin.ignore();cin.get();\nOx_Ox_--__0xEEABD();player++;Ox_Ox_ o0Oo(i==1); __$_BEGIN(0xE0F>2}1,2,2,'B');\nxOFEABD (i=1?true:false) o0<<"<<=\"aPlayer " <<=\" win ";\nelse o0<<=>=\"aGame draw\":cin.ignore();cin.get();system("pause");__OxCXs 0;
```
Source Code Obfuscation: A Technique For...