J Comput Virol (2006) 2:135–147
DOI 10.1007/s11416-006-0024-y

ORIGINAL PAPER

Testing and evaluating virus detectors for handheld devices

Jose Andre Morales · Peter J. Clarke · Yi Deng · B. M. Golam Kibria

Received: 15 March 2006 / Accepted: 20 June 2006 / Published online: 2 September 2006

Abstract The widespread use of personal digital assistants and smartphones gives securing these devices a high priority. Yet little attention has been placed on protecting handheld devices against viruses. Currently available antivirus software for handhelds is few in number. At this stage, the opportunity exists for the evaluation and improvement of current solutions. By pinpointing weaknesses in the current antivirus software, improvements can be made to properly protect these devices from a future tidal wave of viruses. This research evaluates four currently available antivirus solutions for handheld devices. A formal model of virus transformation that provides transformation traceability is presented. Two sets of ten tests each were administered; nine tests from each set involved the modification of source code of two known viruses for handheld devices. The testing techniques used are well established in PC testing; thus the focus of this research is solely on handheld devices. Statistical analysis of the test results show high false negative production rates for the antivirus software and an overall false negative production rate of 47.5% with a 95% confidence interval between 36.6% and 58.4%. This high rate shows that current solutions poorly identify modified versions of a virus. The virus is left undetected and capable of spreading, infecting and causing damage.

J. A. Morales (✉) · P. J. Clarke · Y. Deng
School of Computing and Information Sciences,
Florida International University, University Park,
Miami, FL 33199, USA
e-mail: jmora009@cis.fiu.edu

P. J. Clarke
e-mail: clarkep@cis.fiu.edu

Y. Deng
e-mail: deng@cis.fiu.edu

B. M. G. Kibria
Department of Statistics, Florida International University,
University Park, Miami, FL 33199, USA
e-mail: kibriag@fiu.edu

1 Introduction

On June 14, 2004, the first computer virus infecting handheld devices was identified [9]. The first virus to infect handhelds[1] running Windows Mobile operating system was released July 17, 2004 [7]. This was the beginning of a new era for the virus and antivirus community. At that time there were little if any antivirus solutions available. An overwhelming majority of users were vulnerable to any possible viral attack. In a reactionary effort, security companies released antivirus solutions for the infected devices that only protected against these specific viruses. Still today many handhelds do not have some form of antivirus software installed.

This research evaluates current antivirus solutions for handhelds with the objective of identifying problems in their detection mechanisms. To achieve this objective we introduce a formal model to represent virus transformations and use the model in the generation of test cases. The work of Cohen [15] served as the inspiration for the creation of this model. This model provides detailed traceability of the transformations produced by a virus. The transformed viruses can be precisely ordered by creation time and transformation type. The approach taken was to create test cases that are modifications of two

[1] Smartphones and personal digital assistants will be collectively referred to as handheld devices or handhelds throughout this paper.

already identified viruses and load them into the handheld running the antivirus software. That is, we wanted to test the detection accuracy of the antivirus software against virus modifications. Specifically, the tests were designed with the goal of producing false negatives, which occur when an infected object is not detected as infected, by the virus detectors. Testing virus detectors for production of false negatives has been extensively performed on PCs [12,24,27] and is well documented. Therefore this research focuses only on testing handheld devices. A high false negative production rate would reveal virus detection weaknesses in the software. The test environment consisted of a Pocket PC running the Microsoft Windows Mobile operating system and the antivirus software. The tested antivirus software is specifically designed for this platform and currently available to the public. At the time this research was conducted there were only two known viruses for the Pocket PC running Windows Mobile. In this research both of these viruses were used for testing. This research extends the research presented by Morales et. al. in [3].

To our knowledge, this research is the first to evaluate current antivirus solutions for the Windows Mobile operating system and for handheld devices in general. The flaws and problems discovered by this research can help lay the foundation for future study and work in virus detection for handheld devices. The results of this work can be made public via vulnerability databases, such as the National Vulnerability Database [25]. This research also provides insight on the application of testing methodologies to a new platform in the emerging area of handheld devices. Currently there is no standard set of test cases for virus detectors on this platform. Testing related organizations like Eicar.com and av-test.org also have not yet addressed this issue. The test cases created here can be applied to the development of a standardized set of test cases for this platform and these devices.

In the next section we overview the terminology used in the paper. Section 3 describes related work on testing virus detectors. Section 4 describes a formal model for virus transformation and the test categories used to generate the test cases. Section 5 describes the tests we performed and Sect. 6 our results. Some conclusion are made in Sect. 7.

2 Computer viruses and handheld devices

A computer virus is defined as a program that can infect other programs by modifying them to include a possibly evolved version of itself [15]. Computer viruses have become very sophisticated in detection avoidance, fast spreading and causing damage. A highly populated taxonomy of viruses exists with each classification having its own challenges for successful detection and removal [12,27]. Today viruses are regarded as a real global threat and viewed as a weapon usable by those bent on creating large scale interruption of everyday life [10,34].

The problem of viral detection was studied by Cohen which showed that detecting a virus is not decidable [15]. Many detection algorithms have been presented [26], each with its advantages and disadvantages. Virus detection can be classified as one of two forms: signature based and behavior based [12,27]. Signature based detectors work by searching through objects for a specific sequence of bytes that uniquely identify a specific version of a virus. Behavior based detectors identify an object as being viral or not by scrutinizing the execution behavior of a program [16]. Behavior based detection is viewed by many including the authors of this research as key to the future of virus detection [1,5,14] because of its ability to detect unknown viruses.

The evolution of virus detectors has moved parallel with the release of viruses in a reactionary manner [21]. As new viruses with new techniques were identified, antivirus researchers rushed to include these new tactics in their software [4,12,27]. This evolution has produced a learning curve, with virus authors and antivirus researchers as both teacher and student. Antivirus companies need to develop security solutions for handheld devices that defend against the types of viruses seen in the past without having to go through the same learning curve for a second time.

A handheld device can be described as a pocket sized device with computing capabilities. Two types of handheld devices are relevant to this paper: the personal digital assistant, also called pda, and the smartphone. A pda is used as a personal organizer that includes a contact list, calendar of events, voice recorder, notes, and more. A smartphone can be viewed as a cellular phone integrated with a pda. Both of these types of handhelds share some basic limitations such as: limited screen size, variable battery life, small storage space, operating system installed with limited resources and reduced processing capabilities [17,18]. These limitations may not allow for antivirus software to be as powerful as those found in desktop PCs. Signature databases and detection functionalities are limited in size and scope. This can possibly result in more viruses being able to easily spread and avoid detection in an environment with weak security. Some handheld device security issues have been previously addressed in [11,28–31].

3 Related work

An earlier version of this research was presented by Morales et. al. in [3]. We have extended that research by adding a new set of test cases for a second virus for handhelds known as Brador [8]. In addition, formal statistical analysis of the results was conducted providing rigorous results and conclusions of the administered tests (see Sect. 6).

In this paper we use a black-box approach to test the antivirus solutions for handheld devices. Black-box testing is an approach that generates test data solely from the application's specification [22]. Since the software under test is proprietary, we employ the end-user view of the software as our specification. This specification is the detection of objects infected with a virus. There are several techniques used to generate test cases based on the specification of a software system [20]. Two of these techniques are input space partitioning, and random testing [20]. Partition testing uses domain analysis to partition the input-output behavior space into sub domains such that any two test points chosen in a sub domain generates the same output value [6]. Random testing involves the selection of test points in the input space based on some probability distribution over the input space [22]. To generate the input data for our test cases we used a combination of input space partitioning and random selection of test points. Due to the limited access to the full specification of the antivirus software, we informally apply partition testing and random testing. We intuitively apply these techniques using the results of previous studies in testing antivirus software.

The research presented here is motivated by the work done by Christodorescu and Jha [24]. Their research proposed methods of testing malware detectors based on program obfuscation [12,27]. They used previously identified viruses to test the resilience of commercially available antivirus software for PCs. Christodorescu and Jha address two questions in their work: (1) the resistance of malware detectors to obfuscations of known malware; (2) can a virus author identify the algorithm used in a malware detector based on obfuscations of the malware. The approach they used to answer these questions involved: the generation of test cases using program obfuscation, the development of a signature extraction algorithm, and the application of their methodology to three commercial virus scanners. Filiol presented more rigorous work on signature extraction algorithms and prevention in [13]. The results of Christodorescu and Jha's work indicated that the commercial virus scanners available for PCs are not resilient to common obfuscation transformations. We use a similar approach to test the virus detection ability for handheld devices. Unlike the work by Christodorescu and Jha [24], we are limited by the number of viruses available for handheld devices. This limitation is based on the fact that virus authors have just only started to write viruses targeting handheld devices. Our experiments use similar transformations on the source code of the malware to generate test cases.

Marx [2] presents a comprehensive set of guidelines for testing anti-malware software in the "real world". Marx claims that many of the approaches used to test anti-malware software in research do not translate into appropriate testing strategies for small business and home office use. He further states that the focus of testing for the real world should be to create tests that are as exact as possible. That is, tests that focuses on on-demand, on-access, disinfection and false positive testing of the anti-malware software products. Although his article is targeted for data security managers and professional testers, he outlines procedures that should be taken when performing anti-virus software testing in any environment. The work done by Marx [2] was used as a reference guideline for this research. Other relevant research on the subject of testing virus detectors can be found in [32,33].

4 Testing and evaluation

In this section we present a formal model for the transformation of viruses and show how this model is used to generate the test cases for our study. A characterization of the transformation stage that a virus enters when executing is given. Detection of a transformed virus, viral infection and the production of false negatives and false positves are also characterized. Descriptions of each of the five test categories are also given.

4.1 Formal model of virus transformation

As stated in Sect. 2, a computer virus is defined as a program that can infect other programs by modifying them to include a possibly evolved version of itself [15]. A virus $v \in V$ where V is the set of all possible viruses, enters during its execution a transformation stage R where one or more possibly evolved versions of v written v', are produced and written to some location, expressed in Eq. (1). Successful transformation occurs when v' has preserved the original intended execution behavior $XB()$ of v, expressed in Eq. (2).

$$R_i(p_j, v, s) \equiv p_{ij}(v, s) = v' \tag{1}$$

In Eq. (1), R_i is the currently running transformation instance. A specific type of transformation is denoted as

$p \in P$ where $P = \{T, H, B, L, C\}$, for example B means substitution (see Sect. 4.2 for descriptions of these values). The number of transformations that have occurred are in i, the current value of i is the ith transformation to have taken place. The variable j holds the value representing the number of times, jth occurrence, a specific transformation type p has occurred, $p = H$ and $j = 3$ means that the transformation type H has been used in 3 transformations up to this point. The virus to be transformed is denoteted by v. The element s provides p the details for a specific transformation. For example if $p = B$ then s may contain the line numbers to substitute and the new lines to use for substitution (see Sect. 4.2 for details of s for each transformation type). The element s can be characterized as a data structure composed of two types of attributes: (1). *an association attribute* and (2). *transformation detail attributes*. The association attribute is a single attribute that defines the transformation instance and test category p_{ij} that s is associated with. The transformation detail attributes consist of one or more attributes that specify the details needed for the specified transformation to execute successfully as described in the above example. The transformed version of v is denoted by v'. When R_i occurs, the operation is always independent from every other occurrence of R. The virus v used as input by R is always the same; it is the virus currently executing that invokes R. The output of R, written v', is always a possible evolution of v. The number of v's that is produced is equal to the value of i. In each occurrence of R, the only input that may change is the information held in the attributes in s. Thus the output v' of two occurrences of R may be the same if the attributes of s was unchanged in both operations and the same transformation type p was used. The attribute of s defining the transformation instance and test category p_{ij} will be different for each occurrence of R.

$$\begin{aligned} &\text{If } (XB(v') = XB(v)) \text{ Then } R_i(p_j, v, s) = \text{Success} \\ &\quad \text{Else } R_i(p_j, v, s) = \text{Failure} \end{aligned} \tag{2}$$

The transformed virus v' can equivalently be written as v_{ijk} where k is the symbol for the transformation type used in a specific transformation R_i. The variable k is added to differentiate the value of j for each transformation type p. This is necessary to illustrate that there are multiple instances of j, one for each transformation type p that is used. Each j has its own value representing the j number of times p has been used. Therefore, if $j = 2$ and $k = C$, we know that this is the second time compression is used. Consider the following example: assume virus v has finished one execution of itself. During this execution five transformations occurred. The transformation types used were: 1 substitution of source code, 2 compressions, 1 insertion of trash source code and 1 label renaming. Applying the notation in Eq. (1) to this example, we produce the results shown in Eq. (3).

$$\begin{aligned} &R_1(B_1, v, s) \equiv B_{11}(v, s) = v'_{11B} \\ &R_2(C_1, v, s) \equiv C_{21}(v, s) = v'_{21C} \\ &R_3(C_2, v, s) \equiv C_{32}(v, s) = v'_{32C} \\ &R_4(H_1, v, s) \equiv H_{41}(v, s) = v'_{41H} \\ &R_5(L_1, v, s) \equiv L_{51}(v, s) = v'_{51L} \end{aligned} \tag{3}$$

We can see from Eq. (3) that placing the outputs v' in order of creation is simple. The notation facilitates identifying each virus v' by order of creation and input transformation type. Note that virus v_{21C} and v_{32C} may have been transformed the same or differently from one another. This is, as previously noted, dependent on the information held in the attributes of s.

A virus detector written D, is a software program meant to detect and remove viruses before infecting a computer system [15]. When detection is complete only one of two outcomes can result. The detection was successful or there was a failure. A successful detection implies the correct identification of a virus infected object O_v. This implies that infection, expressed I, of the object O with a virus v is true. That is, the sequence of bits representing v is contained within the sequence of bits representing O. Thus v becomes a subsequence of O. The object could be a file, an address in memory, or some other information stored in a computer system. All objects O are assumed non-viral before detection starts. Notation for virus infection and detection is presented in Eq. (4).

$$\begin{aligned} &I(O) = \text{True iff } v \text{ is a subsequence of } O \\ &\text{If } D(O) = \text{Success then } I(O) \end{aligned} \tag{4}$$

A failed detection produces one of two outcomes: a false positive, FP, or a false negative, FN. A false positive occurs when a non viral object is detected as being viral. A false negative occurs when a virus infected object is not detected as being viral. A small amount of false positives is tolerable, but false negatives must be avoided always. This is formalized in Eq. (5).

$$\begin{aligned} &\text{If } D(O) = FP \text{ then } D(O) \wedge \neg I(O) \\ &\text{If } D(O_v) = FN \text{ then } \neg D(O) \wedge I(O) \end{aligned} \tag{5}$$

Note in Eq. (5) the assumption is made in the case of $D(O_v) = FN$ that the object is already infected with a virus thus justifying the use of the symbol O_v.

4.2 Test categories

The test cases generated, using a non-strict approach to input space partitioning and random testing, can be classified in five categories. These are transposition of source code, insertion of trash source code, substitution of source code, label renaming and compression of the virus executable. These categories were chosen due to the facilitation each one gives virus detectors to produce a false negative [24]. These categories are also characteristic of polymorphism [4,12,27] and metamorphism [27], powerful techniques used by virus authors. Test case implementations of each category are presented in Sect. 5.

4.2.1 Transposition of source code

Transposition is the rearrangement of statements in the source code. This makes the virus look differently by reorganizing its physical appearance. It still preserves the original intended execution behavior. Transposition can be done randomly or in specific areas. The whole body of the source code or only pieces of it can be transposed as long as the original intended execution behavior is preserved. Applying Eq. (1) to this category produces the Eq. (6)

$$R_i(T_j, v, s) \equiv T_{ij}(v, s) = v'_{ijT} \tag{6}$$

where $p = T$ indicates transposition and the transformation detail attributes of s specify the line numbers of the source code to transpose. Transposition can result in changing the area of source code that is used as the signature by virus detectors. This is a result of a change in the byte sequence of the executable version of the virus. The transposition can also result in an increase in the byte size of the virus executable. This is due to the addition of commands that preserve the original intended execution behavior. These changes make transposition of source code a possible cause of a virus detector producing a false negative.

4.2.2 Insertion of trash source code

This category inserts new code into the original source code. This new code consists of instructions that do nothing to change, alter or affect the intended behavior of the original source code. It does, in some cases, change the byte size of the executable version of the virus. By changing the byte size of the executable, some virus detectors may produce a false negative more easily. This occurs in the case where the detector uses the length of the entire virus as part of the detection process. Thus a change in this length could result in the detector misreading the virus. What the newly inserted code does is inconsequential as long as it does not change the intended behavior of the original source code. Using Eq. (1), trash source code insertion is expressed in Eq. (7)

$$R_i(H_j, v, s) \equiv H_{ij}(v, s) = v'_{ijH} \tag{7}$$

where $p = H$ denotes trash insertion. The transformation detail attributes of s store the trash code to be inserted and specify source code locations of where to insert them.

4.2.3 Substitution of source code

The removal of lines of source code is replaced with different lines of code. The lines of code used for replacement are not copied from other areas of the code body. The replacement lines can be the same size as the original. They can also be deliberately shortened or lengthened. This is done to manipulate the overall byte size of the virus executable. The lines that are to be replaced cannot be in an area that can disrupt the original intended execution behavior. This implies that this process cannot be random. Careful selection of lines to replace can assure preservation of execution behavior. Applying Eq. (1) produces Eq. (8).

$$R_j(B_j, v, s) \equiv B_{ij}(v, s) = v'_{ijB} \tag{8}$$

$p = B$ specifies substitution and the transformation detail attributes of s specify which lines to replace and also has stored the lines to replace them with. A virus detector can produce a false negative under this category for one of two possible reasons. First, the substituted lines can change the source code used as a signature by the detector for a given virus. Second, as discussed before, if the byte size is not preserved it could cause the detector to identify it as benign. This occurs in cases where the length of the virus is used in detection.

4.2.4 Label renaming

This category involves the substitution of label names in the source code for new names. A label is synonymous with a procedure or function name in a high level language. The label is a pointer to an address space where the instructions to be executed are located. A label therefore points to a set of instructions that are always executed when the label is referenced. The new labels can be kept the same byte size as the original one and also can be purposely changed to a different size. In addition, the corresponding calls to these labels must be updated to ensure original intended execution behavior. The label names chosen for substitution should be those that reference blocks of instructions essential to the virus

execution such as: finding a file to infect, opening a file for infection and infecting the file. A virus detector can produce a false negative in this category only when a signature includes a label or a call to a label that has been modified. If no labels are included in the virus signature and the length of the entire virus is not used for detection, the possibility of a false negative production is greatly reduced. This category is expressed in Eq. (9) by applying Eq. (1).

$$R_i(L_j, v, s) \equiv L_{ij}(v, s) = v'_{ijL} \tag{9}$$

where $p = L$ signifies label renaming and transformation detail attributes of s store a list of the label names to replace and the new names to replace them with.

4.2.5 Compression of a virus executable

This category is the compression of the original virus executable. Compression is done by a commercial product or private software belonging to the virus author. The original intended execution behavior is fully preserved. When a virus transforms it can evolve into a new version of itself that is self compressed. This new version makes no modifications to alter the execution as it is originally intended. Virus detectors can produce a false negative under this category by failing to match the virus signature. The compression may create a new byte sequence in achieving an overall byte size reduction. This in turn may cause the source code used for the virus signature to be completely modified and thus detection is almost impossible. Virus compression is expressed in Eq. (10) using Eq. (1).

$$R_i(C_j, v, s) \equiv C_{ij}(v, s) = v'_{ijC} \tag{10}$$

$p = C$ represents compression and transformation detail attributes of s store the file name for the compressed version.

5 Test implementation

As of the writing of this paper there were only two known viruses for the Windows Mobile platform: WinCE. Duts.A and Backdoor.Brador.A [7,8]. Test were conducted on both of these two viruses. The source code for both of these is readily available to the public [7,8]. The Duts virus consists of 531 lines of source code. This virus was created as a proof of concept code by virus author Ratter formerly of the virus writers group 29A. The Brador virus consists of 602 lines of source code. This virus was detected in the wild and classified as a Trojan horse. Both of these expose some of the vulnerabilities already present in the Windows Mobile operating system. They are written in the ARM processor assembly language.

5.1 Testing environment

Four commercially available antivirus products for handheld devices were tested: *Norton, Avast!, Kaspersky, and Airscanner.com*. The handheld device used for testing was a Toshiba 2032SP Pocket PC running Windows Mobile 2002 (version 3.0.11171, build 11178) with full phone functionality provided by Sprint PCS. The central processing unit is the ARM processor SA1110. The Operating System of the PC used was Windows XP service pack 2. Before administering the test cases a control test was given. The original two viruses were tested for detection to assure each antivirus product properly identified it. *Each of the two sets of ten test cases were allowed to fully execute to assure that infection of the system was occurring. Thus showing the original intended execution behavior of the virus had been preserved after modifications were made.* On administering the control test of the Brador virus, Kaspersky failed to detect it. This was the only failed control test that was encountered. In spite of this Kaspersky was kept in the final results and used in the statistical analysis (see Sect. 6).

5.2 Description of test cases

The test cases were introduced to the handheld device via the synchronization functionality from a PC. The version used here was Microsoft ActiveSync version 3.7.1 build 4034. The antivirus software performed a complete virus scan with every test. Before testing commenced the antivirus software was checked for updates from the software company's website including the latest virus signature database. Relevant segments of code for several test cases are also provided.

5.2.1 Transposition of source code

Test case 1.1 We chose a set of blocks of source code and inserted labels to each of the blocks. The area of the source code chosen for this is the area where the actual file infection takes place, thus assuring probable execution of the transposed source code. Then with the use of branch statements each labeled block branched to the next block in the set thus preserving the original execution order. As a final step, all the blocks were rearranged and taken out of its original physical order. Example 1 is an implementation of this starting at line 308 of the Duts virus source code.

1. Example of test case 1.1

Original source Code	Modified source code
`ldr r8, [r0, #0xc]`	`section19`
`add r3, r3, r8`	`ldr r8, [r0, #0xc]`
`str r3, [r4, #0x28]`	`add r3, r3, r8`
	`str r3, [r4, #0x28]`
`sub r6, r6, r3`	`bl section20`
`sub r6, r6, #8`	
	`section21`
`mov r10, r0`	`mov r10, r0`
`ldr r0, [r10, #0x10]`	`ldr r0, [r10, #0x10]`
`add r0, r0, r7`	`add r0, r0, r7`
`ldr r1, [r4, #0x3c]`	`ldr r1, [r4, #0x3c]`
`bl _align_`	`bl _align_`
	`bl section22`
	`section20`
	`sub r6, r6, r3`
	`sub r6, r6, #8`
	`bl section21`

Test case 1.2 This involved manipulation of values held in various registers at a given moment during the execution. In assembly language, registers are used extensively to hold values and addresses. The manipulation of these values was done via addition and/or subtraction of a value in a particular register. Moving the value to other registers was also used. The result was an extended piece of source code that took a value, modified it via 2 to 5 instructions and finished by placing back the original value in the original register. This transformation preserved the execution order of the virus and the intended values held in the registers at a given instant in execution. Example 2 is an implementation starting at line 71 of the Brador virus source code.

2. Example of test case 1.2

Original source code	Modified source code
`add R1, R1, R6`	`add R1, R1, R6`
`add R6, R6, #0x1000`	`add R1, R1, #4`
	`add R1, R1, #6`
	`sub R1, R1, #10`
	`sub R1, R1, #10`
	`add R6, R6, #8`
	`add R6, R6, #4`
	`sub R6, R6, #12`
	`add R6, R6, #0x1000`

5.2.2 Insertion of trash source code

Test case 2.1 This involved a copy of an original single line of code. The line was pasted back into the source code immediately following the original one. This did not change the behavior because the line of source code chosen consists of the instruction `DCB` which defines a byte with a string value. This insertion only increased the byte size of the file by the size of the line of code.

Test case 2.2 In this test, the same instruction as in test case 2.1 was inserted right after five lines of source code. The five lines were not in successive order and deliberately chosen to cover the whole body of the source code. Each chosen line represented an essential part of the execution sequence such as: finding a file to infect and reading the stack pointer. The insertion did not affect the intended execution of the code and increased the files byte size by length of the insert line multiplied by five. Example 3 shows the implementation of this test case in the source code of the Brador virus.

3. Example of test case 2.2

```
DCB "Just Looking"
```

Inserted after each of the following lines

```
Line 57  ldr R4, =FILE_ATTRIBUTE_NORMAL
Line 119 ldr R0, =hostname
Line 198 ldr R1, =FIONBIO
Line 387 ldr R0, victims_socket
Line 436 str R3, [SP, #8]
```

5.2.3 Substitution of source code

Test case 3.1 Here we do a replacement of line 514 of the Duts virus source code as shown in example 4. The substitution preserved the length of the original line while making a modification to a subsection of it. This was done to make a modification that did not affect the byte size of the virus. This substitution did not affect the intended execution of the virus. Finally, it is worth noting that the format of the two lines is indeed identical with respect to spaces and character alignments.

4. Example of Test case 3.1

```
DCB "This is proof of concept code.
 Also, i wanted to make avers happy."
```

Replaced with

```
DCB "This is foorp fo tpecnoc code.
 Also, i wanted to make avers happy."
```

Test case 3.2 This test is similar to test case 3.1. We replaced the same line 514 of the virus source code with an almost identical one. This new line also had a modification to a subsection of it. The modification was not the same as that of the first test. This modification made the length of the line smaller than the original and thus also decreased the overall byte size. Also the character and space alignment was not preserved. Example 5 is

the performed line substitution in the Duts virus source code:

5. Example of test case 3.2

```
DCB "This is proof of concept code.
 Also, i wanted to make avers happy."
```

Replaced with

```
DCB "This is poc code.
 Also, i wanted to make avers happy."
```

Test case 3.3 Here we again substituted line 514 of the virus source code with a new one. The new line of code was maximally modified while still preserving the ability to assemble the source code. The line used for replacement was the same length as the original line but space and character alignment were purposely not preserved. Example 6 is the actual substitution in the Duts virus source code.

6. Example of test case 3.3

```
DCB "This is proof of concept code.
 Also, i wanted to make avers happy."
```

Replaced with

```
DCB "dkfjvd dkfje dkfdsfg kd934,
 d kdick 3949rie jdkckdke 345r dlie4 vhg."
```

5.2.4 Label renaming

The labels that were used for substitution were purposely kept the same byte size and also made different sizes in the tests. Also the corresponding calls or branches to these labels were also modified to ensure original execution behavior. The label names chosen for substitution referenced blocks of instructions essential to the virus execution such as: finding a file to infect, opening a file for infection and infecting the file.

Test case 4.1 This test was a simple reversal of four label names found throughout the source code. The byte size was preserved. Also character alignment was preserved. Two of the labels, appearing in lines 79 and 397 of the Duts virus source code were renamed as shown in example 7.

7. Example of test case 4.1

Line number	Original source code	Modified source code
79	`find_next_file`	`next_file_find`
397	`open_file`	`file_open`

Test case 4.2 In this test, the label names were purposely made longer thus increasing the byte size. In this test the character and space alignment were not preserved. Two of these labels, located at lines 79 and 482 of the Duts virus source code were renamed as shown in example 8.

8. Example of test case 4.2

Line number	Original source Code	Modified source code
79	`find_next_file`	`next_file_to_find_for_use`
482	`ask_user`	`user_ask_question_to_continue`

5.2.5 Compression of a virus executable

Test case 5.1 Compression of the virus executable was done by compressing the executable version of the original virus using commercially available software. The software PocketRAR [36] was chosen for this task. This choice was made based on the experience of using the software and there is a version available for Windows Mobile. The compressed file was placed in the handheld device and opened to view its contents. Then the virus scan was performed. This was done to find out if the antivirus software would not only detect the virus in compressed form but also delete it or at a minimum keep it from executing.

6 Test results

Table 1 shows results of applying the tests described in Sect. 5.2 with the Duts virus. Table 2 shows results of applying the tests with the Brador virus. Column 1 is the test categories. Column 2 is the individual tests in the order described in Sect. 5. Columns 3 through 6 contain the individual test results for the antivirus software used in the test executions. The last row shows the false negative production rate of each of the software tested. A value of 0 represents detection failure, thus the virus was not detected and deleted and was still capable of execution. A value of 1 represents detection success and deletion of the infected file. A value of 2 denotes successful detection but not deletion, this value was added for the special case of compression. Note that a value of 0 is a false negative.

6.1 Inference analysis

Table 1 shows that Norton had the highest false negative production rate while the Avast! had the lowest for the Duts virus. Table 2 shows Kaspersky had the

Table 1 Virus scanner test results and false negative production by software for Duts virus

	Norton	Avast!	Kaspersky	Airscanner.com
Original virus	1	1	1	1
Transposition				
Test 1.1	0	1	0	0
Test 1.2	0	1	1	0
Trash insertion				
Test 2.1	0	1	1	1
Test 2.2	0	0	0	0
Substitution				
Test 3.1	1	1	1	1
Test 3.2	0	1	1	1
Test 3.3	1	1	0	0
Label renaming				
Test 4.1	1	1	1	1
Test 4.2	1	1	1	1
Compression				
Test 5.1	2	2	2	2
False negative %	60	20	40	50

highest false negative production rate while Airscanner.com had the lowest for the Brador virus. If Kaspersky were to be disqualified due to failing the control test than Norton would have the highest false negative rate. We have opted to keep Kaspersky in the analysis to see the test results. Not including scanning the original virus, a total number of 80 tests were performed. Of these, 42 tests were successful detections, leaving 38 as failures. This is an overall 47.5% false negative production rate which is very high and unacceptable.

In the test for compression of source code, a special note should be taken regarding the behavior of the virus. The compression software apparently creates a temporary copy of the contents of a compressed file when the files are viewed. The virus scanner detects and deletes this temporary copy, however, the original virus file can still be executed from within the compressed file view. Thus the compression software does not allow the antivirus to delete the contents of a compressed file. We count this as a failure because the virus is still in the handheld device, even though it was detected, and can still be executed.

Table 3 shows cumulative false negative production rates for all the administered tests on both viruses. Columns 1 and 2 are similar to Table 1, Columns 3 and 4 shows successful and failed detections, and Columns 5 and 6 show false negative production rates by individual test and test category. Compression had the highest false negative production rate followed by transposition of source code and insertion of trash source code. In the individual test results, test 2.2 caused almost all the antivirus software to produced false negatives. Yet test 2.1 caused a comparably smaller false negative production amount. This shows the insertion of trash source code within actual lines of instruction code is enough to cause the detector to incorrectly identify the file as viral. The transposition test category, the first test produced the most false negatives. The insertion of branch statements in the source code results in a different physical appearance while maintaining the same execution behavior proved to be very effective in avoiding detection.

In the substitution of source code category the false negative produced in test 3.2 hints that a slight decrease in the byte size of the virus executable may cause the virus to go undetected. In test 3.3 , we purposely made space and character alignments different than the orig-

Table 2 Virus scanner test results and false negative production by software for Brador virus

	Norton	Avast!	Kaspersky	Airscanner.com
Original virus	1	1	0	1
Transposition				
Test 1.1	0	0	0	1
Test 1.2	0	0	0	1
Trash insertion				
Test 2.1	0	1	0	1
Test 2.2	0	0	0	1
Substitution				
Test 3.1	1	1	0	1
Test 3.2	0	1	0	1
Test 3.3	1	1	0	1
Label renaming				
Test 4.1	1	1	0	1
Test 4.2	1	1	0	1
Compression				
Test 5.1	2	2	0	2
False negative %	60	40	100	10

Table 3 Cumulative false negative production by individual test and category for both viruses

	Successful detection	Failed detection	Per test False negative %	Test category False negative %
Transposition				
Test 1.1	2	6	75%	68.75
Test 1.2	3	5	62.50	
Trash Insertion				
Test 2.1	5	3	37.50	62.50
Test 2.2	1	7	87.50	
Substitution				
Test 3.1	7	1	12.50	29.17
Test 3.2	5	3	37.50	
Test 3.3	5	3	37.50	
Label Renaming				
Test 4.1	7	1	12.50	12.50
Test 4.2	7	1	12.50	
Compression				
Test 5.1	0	8	100	100

inal line of source code while keeping the byte size the same which caused the same amount of false negative production to occur as in test 3.2. In the label renaming category preserving and purposely changing the byte size of the labels did not affect the virus detectors. This implies that changing the byte size may have the affect of avoiding detection if the byte size reduction is done in certain areas of the source code. Also one can infer that labels may not be used by the virus signatures. When a byte size reduction causes a false negative production, the modified area might be of critical importance to the detector deciding if the code is viral or not.

6.2 Statistical analysis

This section presents a formal statistical analysis of the results presented in Tables 1,2 and 3. Computer software STATA 9.0 [35] was used to analyze the data. For testing about the proportions we refer Zar [19] and Scheaffer and McClave [23] among others. The analysis performed were: (1) false negative production for the test results of each virus and overall; (2) false negative pairwise test between the tested software for each virus and overall; (3) false negative pairwise test between test categories. Confidence intervals were established for all the performed statistical analysis.

6.2.1 False negative production analysis

1. Analysis of the complete data set for Duts and Brador (80 observations) The 95% confidence interval for the proportion of overall false negative production is between 0.366 to 0.584. We have tested the following hypotheses. From both the confidence interval and tests in Table A, we may conclude that the overall proportion of false negative production is about 50%.

Table A Analysis for overall data set (80)

H_0	$<$	$\neq$	$>$
0.30	0.9997	0.0006	0.0003
0.40	0.9145	0.1709	0.0855
0.50	0.3274	0.6547	0.6726
0.60	0.0112	0.0225	0.9888
0.70	0.0000	0.0000	1.0000

2. Analysis for the Duts data set (40) The 95% confidence interval for the proportion of false negative production is between 0.272 to 0.578. This interval and the tests in Table B support that the proportion of false negative production is about 50%.

Table B Analysis for the Duts data set (40)

H_0	$<$	$\neq$	$>$
0.30	0.9578	0.0845	0.0422
0.40	0.6266	0.7469	0.3734
0.50	0.1714	0.3428	0.8086
0.60	0.0119	0.0239	0.9881
0.70	0.0001	0.0001	0.9999

3. Analysis for the Brador data set (40) The 95% confidence interval for the proportion of false negative production is between 0.370 to 0.680. This interval and the tests in Table C support that the proportion of false negative production is about 50%.

Table C Analysis for the Brador data set (40)

H_0	$<$	$\neq$	$>$
0.30	0.9990	0.0019	0.0010
0.40	0.9467	0.1066	0.0533
0.50	0.6241	0.7518	0.3759
0.60	0.1665	0.3329	0.8335
0.70	0.0079	0.0157	0.9921

Comparing the results in Tables A though C, it is clear that the proportion of false negative production with the Brador virus is higher than the proportion of false negative production with the Duts virus. This is because Kaspersky had the highest proportion of false negative production in comparison to the other software for the Brador virus.

6.2.2 False negative pairwise Analysis by software

1. Pairwise comparison between softwares for both Duts and Brador (n=20) The estimated proportion of false negative production along with their corresponding confidence intervals are given in Table D. From Table D it can be observed that both Avast! and Airscanner.com have the same proportion of false negative production and this amount is lower than both Norton and Kaspersky. Kaspersky performed the worst.

Table D Proportion of false negative production for both viruses (20)

Variable	Proportion	Confidence Interval
Norton	0.60	(0.385, 0.814)
Avast!	0.30	(0.099, 0.501)
Kaspersky	0.70	(0.499, 0.901)
Airscanner.com	0.30	(0.099, 0.500)

From Table E we observe that statistically there is no difference between the proportion of false negative

Table E False negative pairwise test, confidence intervals by software (both viruses) (n=20)

	P-value			Confidence Interval
$H_0 : P_1 - P_2 = 0$	$P_1 - P_2 < 0$	$P_1 - P_2 \neq 0$	$P_1 - P_2 > 0$	
Norton and Avast!	0.9717	0.0565	0.0283	(0.0060, 0.5940)
Norton and Kaspersky	0.2537	0.5073	0.7463	(−0.3940, 0.1940)
Norton and Airscanner.com	0.9717	0.0565	0.0283	(0.0060, 0.5940)
Avast! and Kaspersky	0.0057	0.0114	0.9943	(−0.6840, −0.1160)
Avast! and Airscanner.com	0.5000	1.0000	0.5000	(−0.2840, 0.2840)
Kaspersky and Airscannner.com	0.9943	0.0114	0.0057	(0.1150, 0.6840)

Table F Proportion of false negative production for Duts virus (10)

Variable	Proportion	Confidence Interval
Norton	0.60	(0.262, 0.878)
Avast!	0.20	(0.052, 0.556)
Kaspersky	0.40	(0.122, 0.738)
Airscanner.com	0.50	(0.187, 0.813)

Table H Proportion of false negative production for Brador virus (10)

Variable	Proportion	Confidence Interval
Norton	0.60	(0.262, 0.878)
Avast!	0.40	(0.122, 0.738)
Kaspersky	1.00	(0.692, 1.000)
Airscanner.com	0.10	(0.003, 0.445)

production by Norton and Kaspersky and Avast! and Airscanner.com. It also confirms that both Avast! and Airscanner.com have the lowest proportion of false negative production in comparison to the others.

2. Analysis of Duts data (10 observations) The estimated proportion of false negative production along with their corresponding confidence interval are given in Table F. From Table G we may conclude that statistically there is no differences among the software except between Norton and Avast!. It also indicates that Avast! produced the lowest proportion of false negative production in comparison to the others.

3. Analysis of Brador data (10 observations) The estimated proportion of false negative production along with their corresponding confidence interval are given in the Table H. From table I we may conclude that statistically there is no any difference between the proportion of false negative production by Norton and Avast! and Kaspersky and Airscanner.com. It is noted that Airscanner.com has the lowest proportion of false negative production followed by Avast! and Norton. However, Avast! gives the lowest proportion of false negative production for Duts data.

6.2.3 Analysis by individual categories

From Tables J and K we can conclude with 95% confident that the Label Renaming produced the lowest false negative production rate followed by substitution, Trash Insertion, Transposition and Compression. Thus both label and substitution might be considered for successfully detecting the virus.

Overall the data showed high false negative production rates for the tested antivirus software. The overall false negative production rate is about 47.5% with 95% confidence limits between 36.6% and 58.4%. That means, we are 95% confident that there is at least 37% and at most 58% of false negatives that can be produced. We observed that both Avast! and Airscanner.com provided the lowest proportion of false negative production among the tested software. However, their proportion of false negative production can go up as high as 50%. For the Duts virus, Avast! produced the lowest rate of false negative production and Airscanner.com produced the lowest rate for Brador. Among the categories, both Substitution and Label Renaming provided the lowest

Table G False negative pairwise test, confidence intervals by software comparison (Duts)

	P-value			Confidence Interval
$H_0 : P_1 - P_2 = 0$	$P_1 - P_2 < 0$	$P_1 - P_2 \neq 0$	$P_1 - P_2 > 0$	
Norton and Avast!	0.9661	0.0679	0.0339	(0.0080, 0.7920)
Norton and Kaspersky	0.8145	0.3711	0.1855	(−0.2294, 0.6294)
Norton and Airscanner.com	0.6735	0.6531	0.3265	(−0.3339, 0.5339)
Avast! and Kaspersky	0.1646	0.3291	0.8354	(−0.5920, 0.1920)
Avast! and Airscanner.com	0.0798	0.1596	0.9202	(−0.6969, 0.0969)
Kaspersky and Airscanner.com	0.3265	0.6531	0.6735	(−0.5339, 0.3339)

Table I False negative pairwise test, confidence intervals by software comparison (Brador)

	P-value			Confidence Interval
$H_0 : P_1 - P_2 = 0$	$P_1 - P_2 < 0$	$P_1 - P_2 \neq 0$	$P_1 - P_2 > 0$	
Norton and Avast!	0.8145	0.3711	0.1855	(−0.2294, 0.6294)
Norton and Kaspersky	0.0127	0.0253	0.9873	(−0.7035, −0.0964)
Norton and Airscanner.com	0.9905	0.0191	0.0095	(0.1440, 0.8560)
Avast! and Kaspersky	0.0017	0.0034	0.9983	(−0.9036, −0.2964)
Avast! and Airscanner.com	0.9393	0.1213	0.0607	(−0.0560, 0.6560)
Kaspersky and Airscanner.com	1.0000	0.0001	0.0000	(0.7140, 1.0859)

Table J Proportion of false negative production for individual test categories

Variable	Proportion	Confidence Interval
Transposition	0.6875	(0.4134, 0.8898)
Trash Insertion	0.6250	(0.3543, 0.8480)
Substitution	0.2917	(0.1262, 0.5109)
Label Renaming	0.125	(0.0155, 0.3835)
Compression	1.000	(0.6306, 1.0000)

Table K False negative pair wise test and corresponding confidence interval for category

	P-value			Confidence Interval
$H_0 : P_1 - P_2 = 0$	$P_1 - P_2 < 0$	$P_1 - P_2 \neq 0$	$P_1 - P_2 > 0$	
Transposition and trash insertion	0.6451	0.7097	0.3549	(−0.2659, 0.3909)
Transposition and substitution	0.9932	0.0137	0.0068	(0.1049, 0.6868)
Transposition and label	0.9994	0.0012	0.0006	(0.2834, 0.8415)
Transposition and compression	0.0378	0.0756	0.9622	(-0.5396, -0.0854)
Trash Insertion and substitution	0.9817	0.0367	0.0183	(0.0344, 0.6322)
Trash Insertion and label	0.9983	0.0035	0.0017	(0.2127, 0.7873)
Trash Insertion and compression	0.0228	0.0455	0.9772	(-0.6122, -0.1377)
Substitution and label	0.8910	0.2162	0.1081	(-0.0770, 0.4102)
Substitution and compression	0.0003	0.0005	0.9997	(-0.8902, -0.5265)
Label and compression	0.0000	0.0000	1.0000	(-1.0370, -0.7130)

proportion of false negative production. During the test case creation, we were not aware if the signature used by a detector was modified. Many of the successful detections could have occurred because the transformation did not affect the virus signature. Overall, with a 47.5% false negative production rate, there is clearly room for improvement.

7 Conclusion

We have presented a technique of testing handhelds based on a formal model of virus transformation. The results show multiple flaws in current virus detectors for handheld devices. The statistical analysis of the test results produced high false negative rates for each antivirus product and an extremely high overall false negative production rate of 47.5% with with 95% confidence limits between 36.6% and 58.4%. These results suggest that current virus detectors are purely simple signature based detection. The formal model shows how detailed traceability of the virus transformations can be done. Future work includes the detailed study of false negative productions in any of the given tests. Byte size changes, substitution and transposition of source code and compression require further study to improve virus detection under these conditions. Currently we have a great archive of knowledge of viruses for PCs. This information can be used to produce sophisticated virus scanners for handheld devices given their limitations. Ideally, this will occur expeditiously and preemptively to help avoid infections of future viruses for handheld devices.

Acknowledgments This was supported in part by the National Science Foundation under Grant No. HRD-0317692. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied by the above agencies. The authors thank Gonzalo Argote-Garcia and Konstantin Beznosov for their contributions to this research.

References

1. Conry-Murray, A.: Behavior blocking stops unknown malicious code. Netw. Mag. (2002) http://www.networkmagazine.com
2. Marx, A.: A guideline to anti-malware-software testing. In: European Institute for Computer Anti-Virus Research (EICAR) 2000 Best Paper Proceedings, 2000. pp. 218–253.
3. Morales, J.A., Clarke, P.J., Deng, Y.: Testing and evaluation of virus detectors for handheld devices. In: The Proceedings of NIST Workshop on Software Security Assurance Tools, Techniques, and Metrics (SSATTM), pp. 67–74 (2004)
4. Nachenberg, C.: Computer virus-antivirus coevolution. Commun. ACM, **40**(1):46–51 (1997). http://doi.acm.org/10.1145/242857.242869
5. Nachenberg, C.: Behavior blocking: the next step in anti-virus protection. Security Focus, March (2002) http://www.securityfocus.com/infocus/1557
6. Ntafos, S.C.: On random and p1artition testing. In: ISSTA '98: Proceedings of the 1998 ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 42–48 ACM Press, New York (1998).http://doi.acm.org/10.1145/271771.271785
7. Peikari, C., Fogie S., Ratter/29A.: Details emerge on the first windows mobile virus. informit.com (2004) http://www.informit.com/articles/article.asp?p=337069.
8. Peikari, C., Fogie, S., Ratter/29A, Read, J.: Reverse engineering the first pocket pc trojan. Sams Publishing (2004) http://www.samspublishing.com/articles/article.asp?p=340544.
9. Symantec antivirus research center: http://securityresponse.symantec.com/avcenter/
10. Denning, D.: Cyberterrorism testimony before the special oversight panel of terrorism committee on armed services, house of representatives, May (2000) http://www.cs.georgetown.edu/ denning/infosec/cyberterror.html
11. Mackey D., Gossels J., Johnson, B.C.: Securing your handheld devices. The ISSA Journal, April (2004). http://www.systemexperts.com/tutors/ISSAHandheldArticle.pdf
12. Filiol, E. Computer Viruses: from Theory to Applications. IRIS International series, Springer, Berlin Heidelberg Newyork Verlag, (2005). ISBN 2-287-23939-1
13. Filiol, E.: Malware pattern scanning schemes secure against black box analysis. J. Comput. Virol., EICAR 2006 Special Issue, (2), 1 (2006)
14. Messmer, E.: Behavior blocking repels new viruses. Network World Fusion, January (2002) http://www.nwfusion.com/news/2002/0128antivirus.html
15. Cohen, F.: A Short Course on Computer Viruses. Wiley Professional Computing (1994). ISBN 0-471-00769-2
16. Schneider, F.: Enforceable security policies. ACM Trans. Inf. Syst. Security, **3**(1):30–50, 2000. http://doi.acm.org/10.1145/353323.353382
17. Vahid, F., Givargis, T.: Embedded System Design a Unified Hardware/Software Introduction. Wiley (2002) ISBN 0-471-38678-2
18. Francia, G.: Embedded system programming. J Comput Sci Colleges **17**(2), (2001)
19. Zar, J.H.: Biostatistical Analysis. Prentice-Hall, New Jersey (1999). Second edition, ISBN 0-130-81542-X
20. Zhu, H., Hall, P., May, J.: Software unit test coverage and adequacy. ACM Comput. Surve. **29**(4), 366–427 (1997)
21. Ibm research. virus timeline. http://www.research.ibm.com/antivirus/timeline.htm
22. Myers, G.J. The Art of Software Testing. Wiley (2004). Second edition, ISBN 0-471-46912-2
23. Sheaffer, R.L., McClave J.T.: Probability and Statistics for Engineers. International Thomson Publishing and Wadsworth Publishing Company (1996) Fourth edition, ISBN 0-534-20964-5
24. Christodorescu, M., Jha, S.: Testing malware detectors. ISSTA '04: Proceedings of the 2004 ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 34–44 (2004) http://doi.acm.org/10.1145/1007512.1007518
25. National vulnerability database. http://nvd.nist.gov/
26. Singh, P., Lakhotia, A.: Analysis and detection of computer viruses and worms: an annotated bibliography. In: ACM SIGPLAN Notices, Vol.37, pp. 29–35 (2002) http://doi.acm.org/10.1145/568600.568608
27. Szor, P.: The Art of Computer Virus Research and Defense. Symantec Press and Addison-Wesley (2005). ISBN 9-780321-304544
28. Symantec Security White Paper: Wireless handheld and smartphone security. Technical report, Symantec Corporation (2003). http://www.symantec.com
29. Ford, R.: The wrong stuff? IEEE Security Privacy (2004)
30. Fogie, S.: Pocket pc abuse: to protect and destroy. In: Black Hat USA (2004) http://www.airscanner.com/pubs/BlackHat2004.pdf
31. Foley, S.: Dumigan, R.: Are handheld viruses a significant threat? Commun ACM **44**(1):105–107 (2001) http://doi.acm.org/10.1145/357489.357516
32. Gordon, S., Howard, F.: Antivirus software testing for the new millennium. In: Proceedings of National Information Systems Security Conference (NISSC), (2000). http://csrc.nist.gov/nissc/2000/proceedings/papers/038.pdf
33. Gordon, S., Ford, R.: Real world anti-virus product reviews and evaluations - the current state of affairs. In: Proceedings of the 1996 National Information Systems Security Conference (1996)
34. Gordon, S., Ford, R.: Computer crime revisited: the evolution of definition and classification. In: European Institute for Computer Anti-Virus Research (EICAR) (2006)
35. Stata release 9.0. Stata Corporation (1999); College Station, Texas
36. Winrar. http://www.win-rar.com/