Evaluating Detection & Treatment Effectiveness of Commercial Anti-Malware Programs

Jose Andre Morales, Ravi Sandhu, Shouhuai Xu

Institute for Cyber Security

University of Texas at San Antonio

Introduction

- Evaluate CAmp's detection and treatment effectiveness against malicious objects
- Redefine true positives (TP) to include treatment effectiveness
- Evaluate 4 current CAmp's in three tests reflecting realistic scenarios
- Results suggest our approach is a more realistic evaluation of CAmp effectiveness than current trends.

Current Evaluation Trends

- In ranking CAmp's for users to purchase
  - Detection accuracy is king
  - Treatment not rigorously tested
- More realistic approach is to evaluate both detection and treatment
  - Treatment just as important as detection and must be equally measured
  - Detection alone does not give the full picture of a CAmp's effectiveness

Desired Characteristics

- Camp Should:
  - Automatically detect and treat malware
  - Correctly inform the user of system status
  - Not leave active threats on a system
  - Minimize treatment choices left up to the user
- From a user perspective, these are desirable characteristics making their life easier!

CAmp Components

- CAmp A(S) where S is any input accepted by A for detection and treatment of malicious objects.
- A( ) consists of two sub-components
  - Detection A_D( )
  - Treatment A_T( )
  - Assumption: a malicious objects is always detected first then treated

Detection $A_D( )$

- Classic measure of detection accuracy
- $A_D(S) \rightarrow (TP,TN,FP,FN)$
  - true positives (TP)
  - true negatives (TN)
  - false positives (FP)
  - false negatives (FN)
  - S can be a single object (file, process) or many objects (directory, or system in infected state)

Treatment $A_T( )$

- Measures treatment effectiveness
- Input comes from output of $A_D(S)$
- $A_T(TP,FP) \rightarrow (TP_A, TP_O, TP_N, FP)$
  - TP_A :TP with automatic treatment
  - TP_O :TP with treatment chosen via user option
  - TP_N :TP which did not receive treatment
- Redefining TP incorporates treatment outcomes in a standardized form

Evaluating A(S)

- A(S) → (FN,FP,TN,TP_A,TP_O,TP_N)
- Evaluates both detection and treatment effectiveness of a CAmp
- Only interested in malicious objects
  - We set FP=0, TN=0 thus S = TP + FN
  - Tests designed to guarantee as much as possible
- Effective detection → high TP & low FN
- Effective detection + treatment → high TP_A & low FN

Evaluation Tests

- 4 CAmps (trial versions): Kaspersky, ESET, BitDefender, ZoneAlarm.
- Set of 974 malware samples
- CWSandbox 27 October 2009 upload
- 3 tests emulating realistic scenarios a user may face when dealing with malware
- VMWare running Windows XP-SP2
- Snapshot scanned and assured malware free prior to testing

Calculating TP_A, TP_O,TP_N

- CAmp log file labels used to calculate results
- All labels in TP_A verified to do what label suggests
- All labels in TP_N verified as leaving malware active
- Misleading labels leave active malware on system
- TP_O calculated by counting number of malware samples a user is asked to choose treatmen

CAmp Name	Log File Labels	
	TP_N	TP_A
Kaspersky Internet Security 2010	Detected	Disinfected, Quarantined, Deleted
ESET Smart Security V4.2	Detected	cleaned by deleting - quarantined cleaned by deleting, cleaned
ZoneAlarm Extreme Security V9.1	File Repaired Failed	File Repaired Failed (Quarantined) Infected, File Repaired
BitDefender Total Security 2010	fail disinfect	deleted, disinfected, quarantined

Test 1

- A static file scan on a folder of 974 known malware samples, FN = 974 – TP
- TP rates higher than TP_A meaning detection + treatment less effective then detection alone

CAmp Name	TP	FN	TP_A	TP_O	TP_N	TP Rate	TP_A Rate	FN Rate
Kasperskey	921	53	912	9	0	94.6%	93.6%	5.4%
ESET	936	38	814	118	4	96.1%	83.5%	3.9%
ZoneAlarm	920	54	896	24	0	94.5%	91.9%	5.5%
BitDefender	951	23	653	288	10	97.6%	67.0%	2.4%

Malware used in Tests 2 & 3

- Used 3 sets of 4 malware samples, each set executes together harmoniously
- Active at time of testing

1st malware set	
Trojan.Pasta.anq - 2574eda157245099b0ab2dbc1be2d980	Backdoor.Poison.xtr - 37e8695cc7be98a6ae637c240f09d6c0
Trojan.Buzus.lba - 58242cb6fbf79d7e7ea616e17acf7e11	Virus.Xorala - 7018d9ed260232cd4983ed4f4b59a9c6
2nd malware set	
Net-Worm.Allaple.e - 75da1173c9325b926a58d83ac4949915	Packed.Krap.n - 75e3fc3b0cf524293c8bef77e2f2dc43
Worm.Win32.Fujack.dg - 75fbfe7a92bf11b59e3e6616b4cfc8db	Trojan-Spy.Pophot.hbn - 760dfe7eb26db590eeb0f54b7340e2f9
3rd malware set	
Trojan.Banker.afhd - 760f3a3f7520378278b84a25dd79dcd7	Backdoor.Sinowal.eed - 760f71e67ef40f75c8de084559f4a807
Packed.Tdss.i - 7628ab6d1aa42671f29efde798ac1225	Trojan-Spy.Zbot.ack - 76348e3e8016ce0663635ad6f7b8cf0e

Test 2

- Install a CAmp in a clean state, infect the system with malware for 3 minutes and perform detection and treatment, FN=TP-4
- Almost every case malware detected when attempting to execute, $TP=TP_A$
- One case TP=12, a detected malware seems to have executed before treatment, newly infected objects not detected

Test 2

1st malware set									
CAmp Name	TP	FN	TP_A	TP_O	TP_N	S	TP Rate	TP_A Rate	FN Rate
Kasperskey	4	0	4	0	0	4	100%	100%	0%
ESET	4	0	4	0	0	4	100%	100%	0%
ZoneAlarm	2	2	2	0	0	4	50.0%	50.0%	50.0%
BitDefender	4	0	4	0	0	4	100%	100%	0%
2nd malware set									
Kasperskey	4	0	4	0	0	4	100%	100%	0%
ESET	12	0	10	2	0	12	100%	83.3%	0%
ZoneAlarm	4	0	3	1	0	4	100%	75.0%	0%
BitDefender	4	0	4	0	0	4	100%	100%	0%
3rd malware set									
Kasperskey	4	0	4	0	0	4	100%	100%	0%
ESET	3	1	3	0	0	4	75.0%	75.0%	25.0%
ZoneAlarm	3	1	2	1	0	4	75.0%	50.0%	25.0%
BitDefender	4	0	4	0	0	4	100%	100%	0%

Test 3

- Execute malware for 3 minutes, then install a CAmp and perform detection and treatment in the infected state
- Most difficult for CAmps to handle, broad range of TP, TP_A and FN rates
- FN calculated using Anubis and CWSandbox
  - Compared log files to Analysis reports
  - .EXE files in report and not in log file marked FN

Test 3

1st malware set									
CAmp Name	TP	FN	TP_A	TP_O	TP_N	S	TP Rate	TP_A Rate	FN Rate
Kasperskey	199	22	198	1	0	221	90.0%	89.5%	10.0%
ESET	692	7	682	8	2	699	99.0%	97.5%	1.0%
ZoneAlarm	51	31	26	22	3	82	62.2%	31.7%	37.8%
BitDefender	241	23	193	42	6	264	91.3%	73.1%	8.7%
2nd malware set									
Kasperskey	1617	10	1614	3	0	1627	99.4%	99.2%	0.6%
ESET	19	15	19	0	0	34	55.9%	55.8%	44.1%
ZoneAlarm	16	13	3	11	2	29	55.2%	10.3%	44.8%
BitDefender	915	10	900	15	0	925	98.9%	97.2%	1.1%
3rd malware set									
Kasperskey	33	13	32	1	0	46	71.8%	69.5%	28.2%
ESET	6	12	2	3	1	18	33.3%	11.1%	66.7%
ZoneAlarm	1	14	0	1	0	15	6.7%	0%	93.3%
BitDefender	85	14	42	41	2	99	85.9%	42.4%	14.1%

Discussion

- Many cases TP_A lower than TP, implying detection + treatment not as effective as detection alone
- Infected state (Test 3) most difficult case
- FN, TP_O in all 3 tests, TP_N in only 2 tests
- Many malware left active on system, either not detected or detected & not treated

	Test 1		Test 2		Test 3		Overall	
CAmp Name	TP	TP_A	TP	TP_A	TP	TP_A	TP	TP_A
Kasperskey	94.6%	93.6%	100%	100%	87.0%	86.0%	93.8%	93.3%
ESET	96.1%	83.5%	91.6%	86.1%	75.5%	74.2%	87.7%	81.2%
ZoneAlarm	94.5%	91.9%	75.0%	58.3%	41.3%	14.0%	70.2%	54.7%
BitDefender	97.6%	67.0%	100%	100%	92.0%	70.8%	96.5%	79.2%

One More Thing…

- CAmps G-Data, AVG results not included
  - AVG did not install, improperly ran, BSOD
  - G-Data only produced FN & TP_O & no TP_A
    - Very high detection rate
    - Automatic treatment disabled in trial version?

New Results

- 5000 samples
  - CWSandbox: Drew samples from 5 random dates
    - 2009: Nov 4, Dec 8; 2010: Jan 28, Jun 29, Aug 25

	Test 1		Test 2		Test 3		Overall	
CAmp Name	TP	TPA	TP	TPA	TP	TPA	TP	TPA
Kaspersky	95.7	95.3	100	100	85	84.2	93.6	93.2
ESET	94.8	81.7	93.4	84.1	78.4	76.2	88.9	80.7
ZoneAlarm	93.6	90.2	76.3	55.8	38.8	11	69.6	52.3
BitDefender	96.8	64.8	97	97	93.4	68.3	95.7	76.7

Conclusions - 1

- New approach to evaluate detection & treatment effectiveness of a CAmp with standardized output
- Redefined TP to include treatment results
- Tests show detection & treatment less effective than detection alone
- Misleading labels, malware left active
- Users unaware of system's real security status

Conclusions - 2

- CAmps need to improve detection & treatment
- Should minimize TP_O & TP_N
- Maximize TP_A
- CAmps need to be tested rigorously and incorporate treatment resulting in a more realistic evaluation than current trends.

Self-Defense Mechanisms

- Camp processes can be disabled and terminated with simple commands
- Poor self defense
- Leaves system vulnerable
- Not able to perform static or behavior based malware scans
- Gives malware the upper hand.

THANK YOU!

QUESTIONS?