2nd International Workshop on Theory of Computer Viruses

Nancy France – May 10-11 2007

Characterizing Virus Replication

FLORIDA INTERNATIONAL UNIVERSITY
Miami's public research university

Jose Andre Morales
Peter J. Clarke
Yi Deng

Introduction - 1

- Viruses spreading faster everyday – (flash, warhol worms)
- Caused $67 Billion Dollars in Corporate Damage (2006 FBI survey)
- Virus Authors Well Funded by organized crime and terrorist groups
- The purpose of the Virus today is to spread other malware as payload.

Introduction - 2

- Signature Detection still centerpiece of today’s antivirus systems
- Average 6 hours to update with new signatures

Poor Protection Against Unknown and Fast Spreading Viruses – False negatives

- Behavior Based Detection Better Defense Against New Unknown Viruses

Introduction - 3

- Drawbacks to Behavior Based Detection
  - Can only detect specific class or groups of viruses or under specific conditions
  - High false positive production
  - The conditions used for this type of detection not consistently present in all viruses
    - Accessing privilieged areas
    - Port Opening
    - Registry modification

Introduction - 4

- This paper characterizes Virus Replication
  - Fundamental characteristic of a virus
  - Guaranteed to be in any malware classified as a virus
  - Consistently present in all viruses
  - Limited ways to attempt replication
  - A good vector to use in the detection of known and unknown viruses.

Presentation Outline

- Concept of a Virus Replication Sequence introduced.
  - Illustrated with an FSA
- Two Detection Models presented
  - Operation Sequence Matching
  - Frequency Measures
- Preliminary Results

Background

- Seminal Work of Cohen, Adleman and Von Neumann describe virus replication
- Cohen: Virus replication on a Turing Machine transferring symbols from one part of a tape to the other.
- Adleman: described infection as the replication aspect of a virus with recursive functions
- Von Neumann: created self reproducing automata showing replication can be defined with computational models

Characterizing Replication - 1

- Definition of a Virus
  - Strict: a program that infects other programs by modifying them to include a possibly evolved version of itself (Cohen)
  - Less Strict: a program that recursively and explicitly copies a possibly evolved version of itself (Szors)
- Both express replication as the qualifying fundamental characteristic of a virus.

Characterizing Replication - 2

- Under these definitions, a malware program is classified as a virus:
  - if and only if it has the ability to replicate.
  - It can be inferred that replication is the only characteristic of a virus consistently present in all viruses.

Characterizing Replication - 3

- Cohen’s Turing Machine shows: read, write, search as essential to replication
- To infect, a virus must gain access to the target. Once a target is infected it may need to be closed to be used by the system
- Therefore Open and Close are also needed for replication.

Characterizing Replication - 3

- Virus replication consists of an ordered sequence of execution of some combination of the following general operations:

  open, read, write, search and close.

- Operations transition the virus to a new state called a ***replication state q in Q***
- The operations that cause the transitions are members ***p*** of the ***replication set P***

Characterizing Replication - 4

- Characterize Virus Replication with an FSA
- FSA E is a 5-tuple (Σ, Q, s, f, Delta) where:
  - Σ is the alphabet of E. Elements of Σ are specific operations p belonging to the replication set P.
  - Q is the finite set of replication states {o, r, w, s, c}
  - s in Q is the start state of E
  - f in Q is the final state of E
  - Delta: Q x P → Q
- Replication states: o=opened, r=read, w=written, s=searched, c=closed
- The output of FSA E is called a replication sequence.

Characterizing Replication - 5

$$E_1 = \text{start} \xrightarrow{O_1} \mathit{opened} \xrightarrow{R_2} \mathit{read} \xrightarrow{S_3} \mathit{read} \xrightarrow{W_4} \mathit{written} \xrightarrow{W_5} \mathit{written} \xrightarrow{R_6} \mathit{read} \xrightarrow{F_7} \mathit{searched} \xrightarrow{C_8} \mathit{closed}$$

- The Figure above is a sample replication sequenc of FSA E

- O_1 R_2 S_3 W_4 W_5 R_6 F_7 C_8 are operations p belonging to the replication set P

- Each operation p is followed by the replication state q in Q that p transitions the virus into

- A replication sequence captures the complete replication process of a virus.

Operation Sequence Matching - 1

- This detection model is done in 4 steps:
  1. Build a training set of random virus samples
  2. Record the complete replication sequence of each virus
  3. Extract replication subsequences
  4. Match replication subsequences in a process to detect virus replication behavior
- Steps 1-3 training session, Step 4 detection session

Operation Sequence Matching - 2

- An Example:
- Complete Sequence

$$E_1 = \mathrm{start} \xrightarrow{O_1} opened \xrightarrow{R_2} read \xrightarrow{S_3} read \xrightarrow{W_4} written \xrightarrow{W_5} written \xrightarrow{R_6} read \xrightarrow{F_7} searched \xrightarrow{C_8} closed$$

- Valid Subsequenes

$$E_{21} = \xrightarrow{R_2} read \xrightarrow{S_3} read \xrightarrow{W_4} written$$

$$E_{31} = \xrightarrow{S_3} read \xrightarrow{W_4} written \xrightarrow{W_5} written \xrightarrow{R_6} read$$

- A process's complete replication sequence containing E_{21} or E_{31} as a subsequence will be flagged as viral.

Replication State Frequency - 1

- This model is done in 3 steps:
  1. Build a training set of random virus samples
  2. Calculate occurrence percentage for each replication state for the entire training set
  3. Match occurrence percentage in a process to detect virus replication behavior
- Steps 1-2 training session, Step 3 detection session
- This model counts the number of replication states.
- Assumption: Viruses attempt to replicate many times leading to high use of operations p in P resulting in high frequency of replication states that should be more than benign processes.

Replication State Frequency - 2

- An Example:

Training Set

$$E_1 = \xrightarrow{O_1} \mathit{opened} \xrightarrow{R_2} \mathit{read} \xrightarrow{R_3} \mathit{read} \xrightarrow{W_4} \mathit{written} \xrightarrow{W_5} \mathit{written} \xrightarrow{W_6} \mathit{written} \xrightarrow{F_7} \mathit{searched}$$

$$E_2 = \xrightarrow{S_1} \mathit{read} \xrightarrow{W_2} \mathit{written} \xrightarrow{W_3} \mathit{written} \xrightarrow{R_4} \mathit{read} \xrightarrow{L_5} \mathit{closed} \xrightarrow{F_2} \mathit{searched}$$

$$E_3 = \xrightarrow{F_1} \mathit{searched} \xrightarrow{L_2} \mathit{closed} \xrightarrow{O_3} \mathit{opened} \xrightarrow{R_4} \mathit{read} \xrightarrow{C_5} \mathit{written} \xrightarrow{S_6} \mathit{read} \xrightarrow{C_7} \mathit{written}$$

$$E_4 = \xrightarrow{W_1} \mathit{written} \xrightarrow{R_2} \mathit{read} \xrightarrow{F_3} \mathit{searched} \xrightarrow{L_4} \mathit{closed} \xrightarrow{O_5} \mathit{opened} \xrightarrow{R_6} \mathit{read} \xrightarrow{T_7} \mathit{written} \xrightarrow{W_8} \mathit{written} \xrightarrow{L_9} \mathit{closed}$$

Detection Set

$$E_1 = \xrightarrow{O_1} \mathit{opened} \xrightarrow{T_2} \mathit{written} \xrightarrow{W_3} \mathit{written} \xrightarrow{L_4} \mathit{closed}$$

$$E_2 = \xrightarrow{F_1} \mathit{searched} \xrightarrow{O_2} \mathit{opened} \xrightarrow{G_3} \mathit{searched} \xrightarrow{C_4} \mathit{written} \xrightarrow{S_5} \mathit{read} \xrightarrow{W_6} \mathit{written} \xrightarrow{W_7} \mathit{written} \xrightarrow{L_8} \mathit{closed}$$

$$E_3 = \xrightarrow{O_1} \mathit{opened} \xrightarrow{R_2} \mathit{read} \xrightarrow{W_3} \mathit{written} \xrightarrow{L_4} \mathit{closed}$$

Replication State Frequency - 3

- Frequency of a state calculated by dividing number of times a state occurs by the total number of all occurred states
- If a process's occurrence percentage equal or surpass the results of the training session, its flagged as viral
- Example Detection done for opened and written.

Results:

- Training Session: opened=10%, read=27%, written=34%, searched=14%, closed=14%
- Detection Session:
  - E1: opened=25%, read=0%, written=50%, searched=0%, closed=25% VIRAL
  - E2: opened=12%, read=12%, written=38%, searched=25%, closed=12% VIRAL
  - E3: opened=25%, read=25%, written=25%, searched=0%, closed=25% NOT VIRAL, written 25% !>= 34%

Testing and Preliminary Results - 1

- Testing of training and detection sessions in progress
- Sample set of 112 viruses in 4 groups: email worms, P2P worms, Network worms and Win32 viruses.
- 4 test sets created of size 28, 56, 84, 112, each set with equal number of the 4 groups.

Testing and Preliminary Results - 2

Preliminary Results for set of 28

- Training session of operation sequence matching gave 154,659 matched subsequences with 77,677 being unique.
- Some subsequences appeared in upto 13 out of 28 viruses
- Only 8 subsequences needed to detect all 28 viruses.

Contributions

- Characterizing virus replication using an FSA lays foundation for new theoretical results
- Showing that a virus can be detected based on replication allows for development of new detection models
- Using replication to detect unknown viruses gives immediate protection against unknown viruses and allows time for antivirus companies to release defense updates

Conclusion & Future Work -1

- Virus replication can be characterized and used to detect known viruses
- Preliminary results of operation sequence matching suggest replication can also detect unknown viruses.
- The results also suggest consistency in replication allowing a small set of sequences to detect a large set of viruses.

Conclusion & Future Work - 2

- Future work includes:
  - Complete testing and result analysis
  - Perform false negative and false positive testing
  - Create new detection models using data mining and machine learning techniques
  - Strengthen the characterization to encompass more properties specific to virus replication

This was supported in part by the National Science Foundation under Grant No. HRD-0317692.

Questions? – ¿Preguntas?
質問 - **вопросы** - sawaal
domande – soru - **ερωτήσεις** - 問題