HUNTING FOR METAMORPHIC ENGINES

Mark Stamp
&
Wing Wong

August 5, 2006

Outline

I. Metamorphic software
II. Virus construction kits
III. How “effective” are metamorphic engines?
- Method used to compare two pieces of code
- Similarity within virus families
- Similarity between virus families

IV. Can metamorphic viruses be detected?
- Commercial virus scanners
- Hidden Markov models (HMMs)
- Similarity index

V. Conclusion

PART I

Metamorphic Software

What is Metamorphic Software?

- Software is metamorphic provided
  - All copies do the same thing
  - Internal structure of copies differs
- Today most software is cloned
- Why metamorphic?
  - Virus/worm avoids signature detection
  - Increase “genetic diversity” of software

Genetic Diversity of Software?

- Suppose a program has a buffer overflow
- If we clone the program
  - One attack works against *every* copy
  - Break once, break everywhere (BOBE)
- If instead, we create metamorphic copies
  - Each copy still has a buffer overflow
  - Same attack does not work against every metamorphic copy
  - Break once break everywhere (BOBE) resistance
  - Sorta like genetic diversity in biology

Evolution of Virus

- Viruses first appeared in the 1980s
  - Fred Cohen
- Viruses must avoid signature detection
  - Virus can alter its “appearance”
- Techniques employed
  - encryption
  - polymorphic
  - metamorphic

Evolution of Virus - *Encryption*

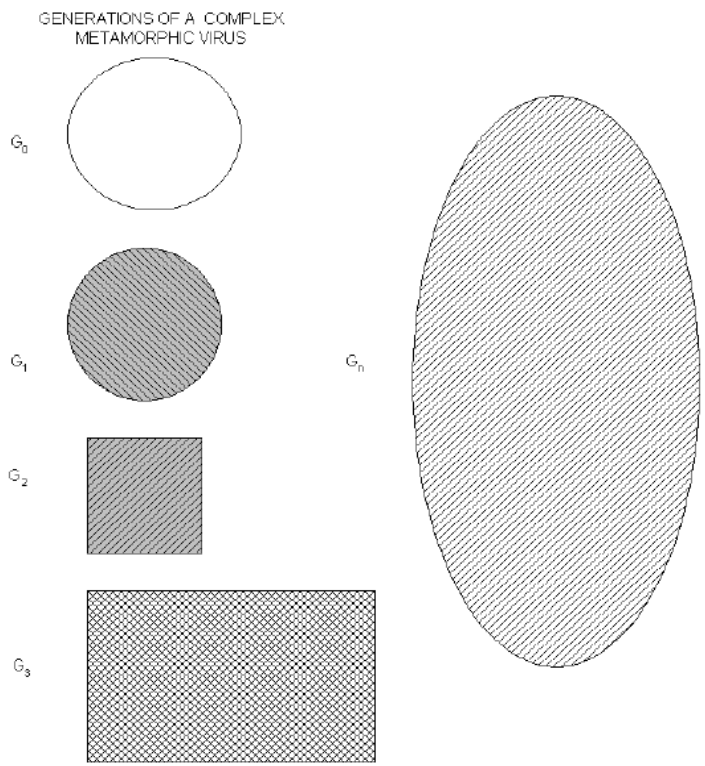
- Virus consists of
  - decrypting module (decryptor)
  - encrypted virus body
- Different encryption key
  - different virus body signature
- Weakness
  - decryptor can be detected

Evolution of Virus – *Polymorphic Viruses*

- Try to hide signature of decryptor
- Can use *code emulator* to decrypt putative virus dynamically
- Decrypted virus body is constant
  - Signature detection is possible

Evolution of Virus – *Metamorphic Viruses*

- Change virus body
- Mutation techniques:
  - permutation of subroutines
  - insertion of garbage/jump instructions
  - substitution of instructions



PART II

Virus Construction Kits

Virus Construction Kits – PS-MPC

- According to Peter Szor:

  "... PS-MPC [*Phalcon/Skism Mass-Produced Code generator*] uses a generator that effectively works as a code-morphing engine..... the viruses that PS-MPC generates are not [only] polymorphic, but their decryption routines and structures change in variants..."

Virus Construction Kits – G2

- From the documentation of G2 (*Second Generation virus generator*):

  "... different viruses may be generated from identical configuration files..."

Virus Construction Kits - NGVCK

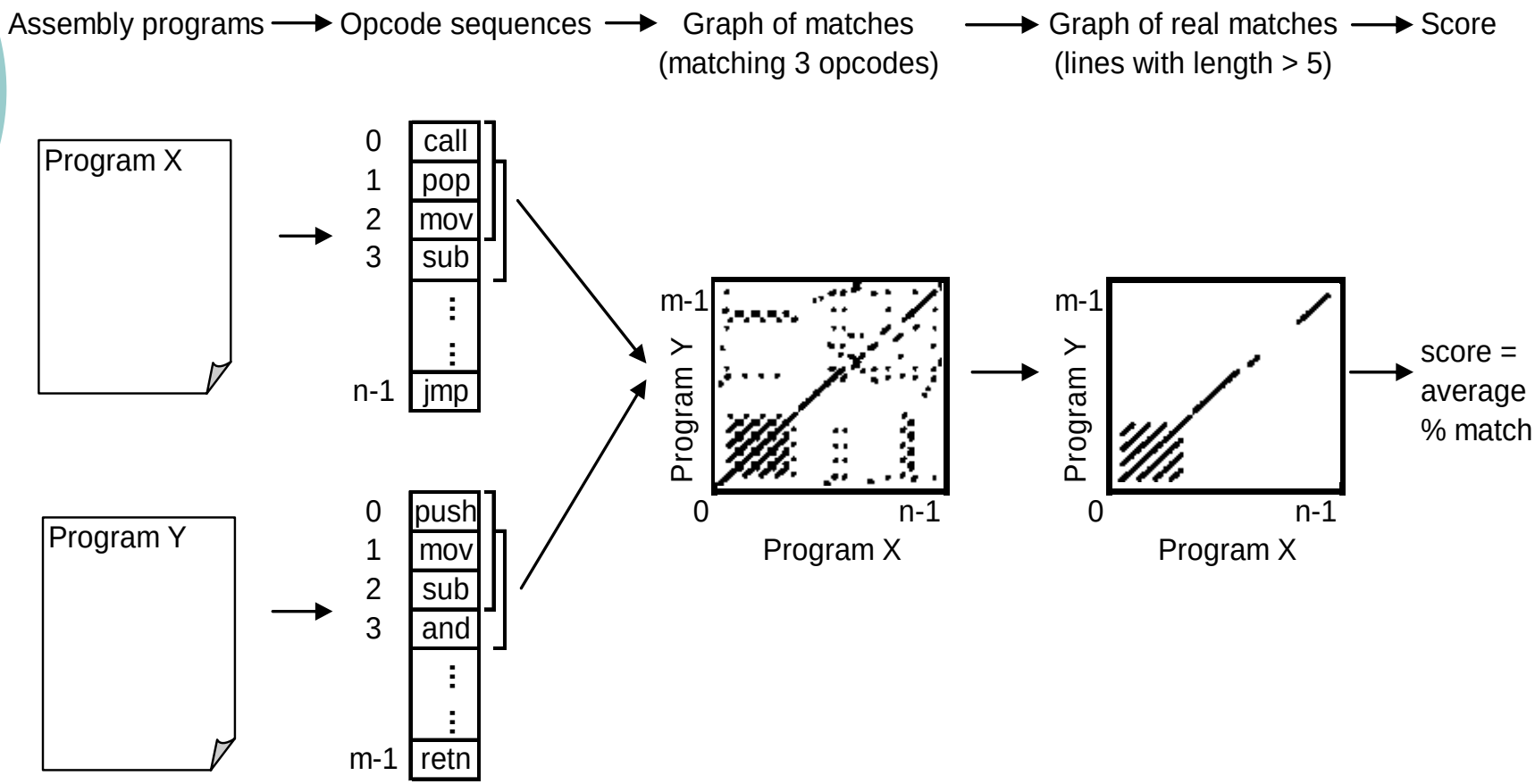
- From the documentation of NGVCK (*Next Generation Virus Creation Kit*):

  "... all created viruses are completely different in structure and opcode..... impossible to catch all variants with one or more scanstrings..... nearly 100% variability of the entire code"

PART III

How Effective Are Metamorphic Engines?

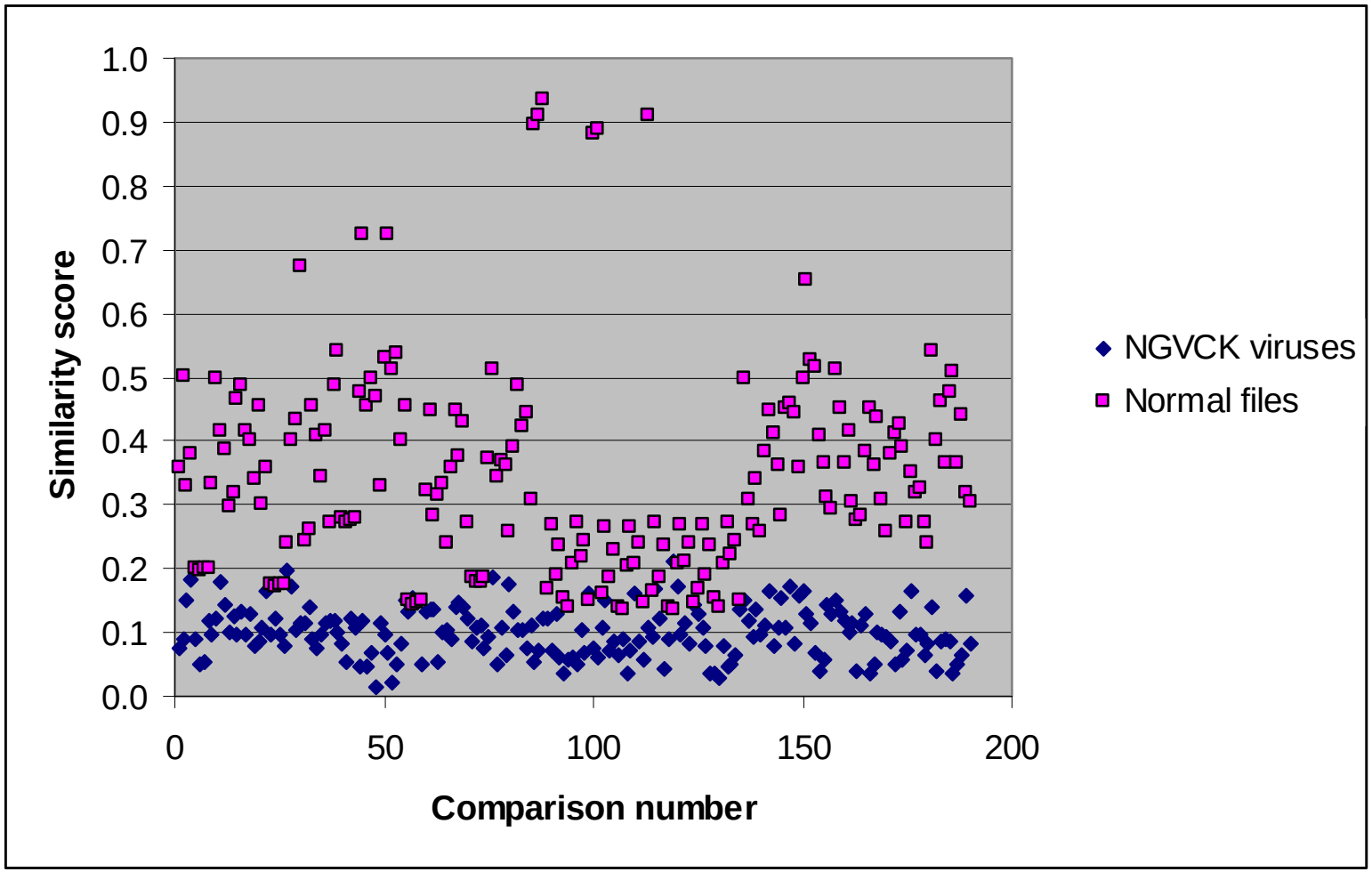
Method to Compare Two Pieces of Code



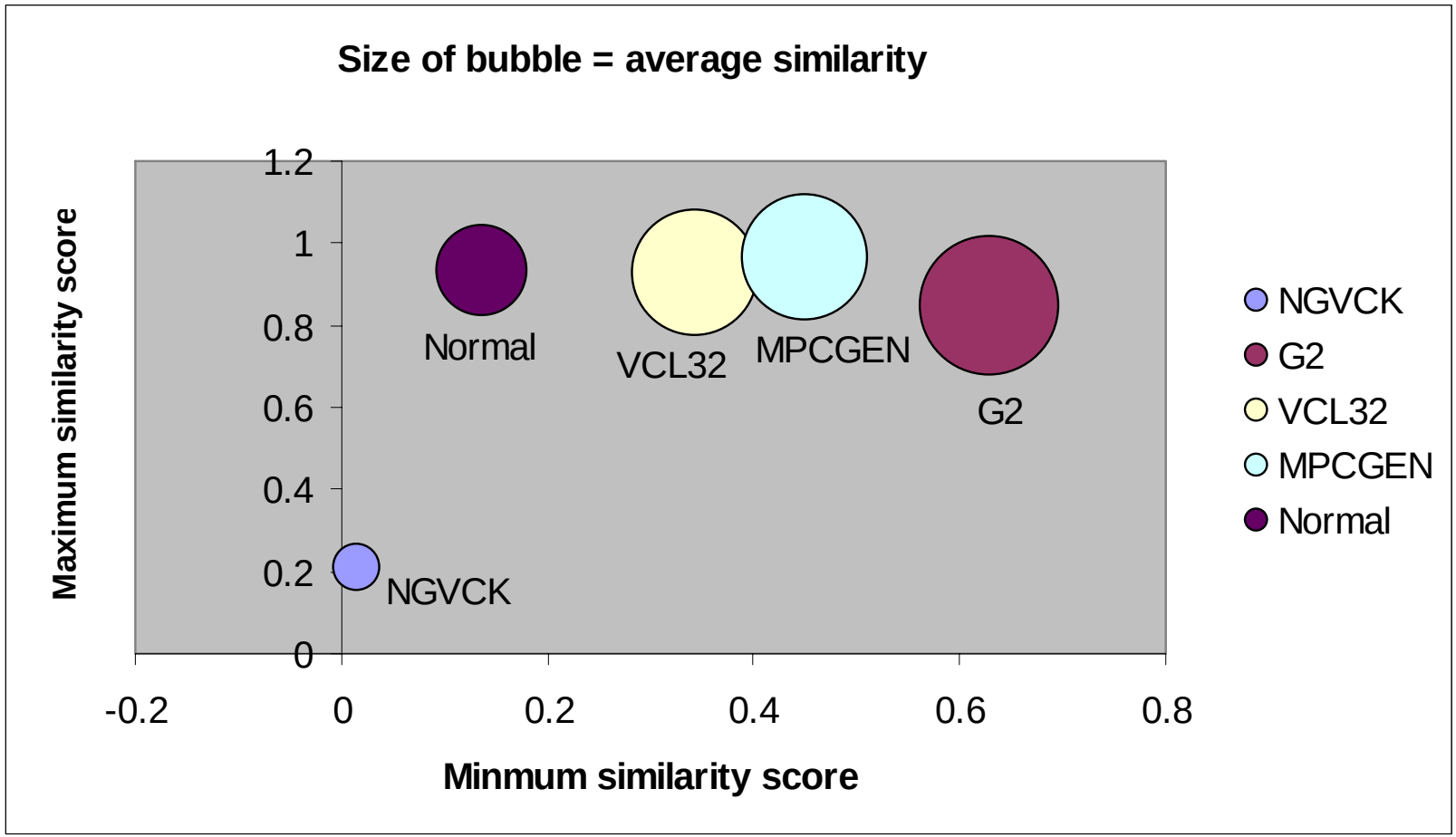
Similarity within Virus Families – Test Data

- Four generators, 45 viruses
  - 20 viruses by NGVCK
  - 10 viruses by G2
  - 10 viruses by VCL32
  - 5 viruses by MPCGEN
- 20 normal utility programs from the Cygwin DLL

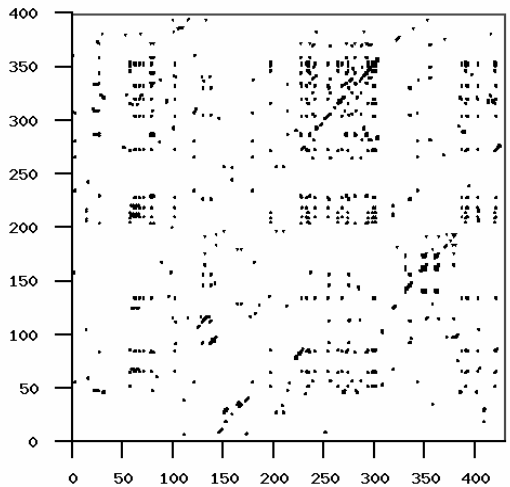
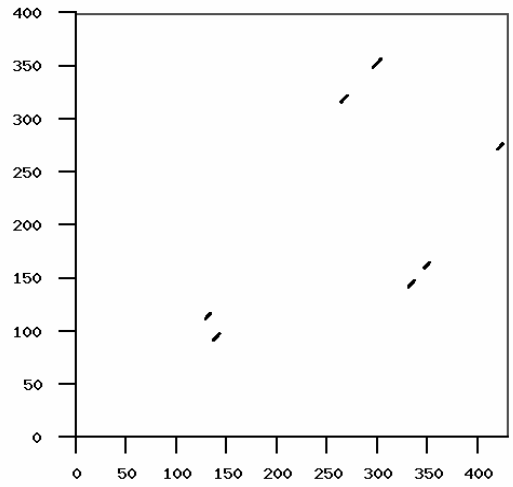
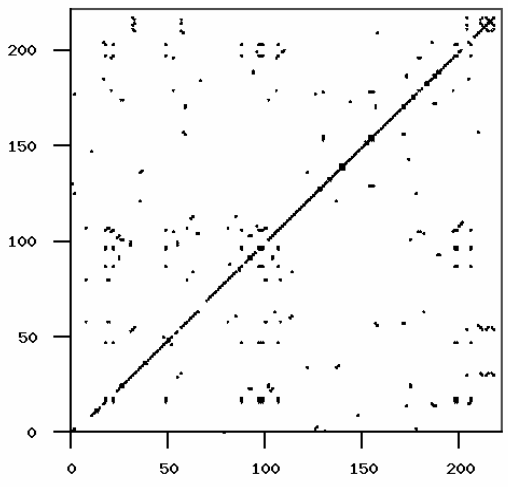
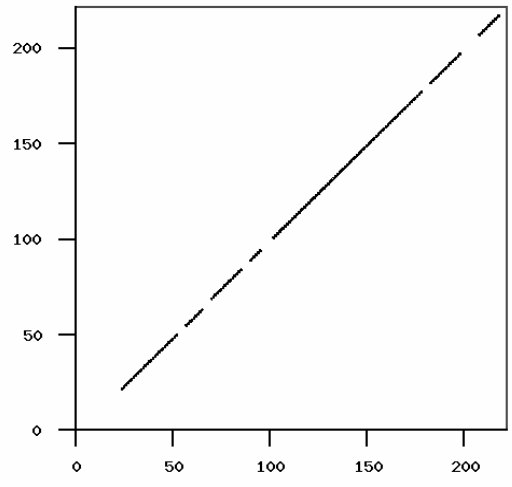
Similarity within Virus Families – Test Result



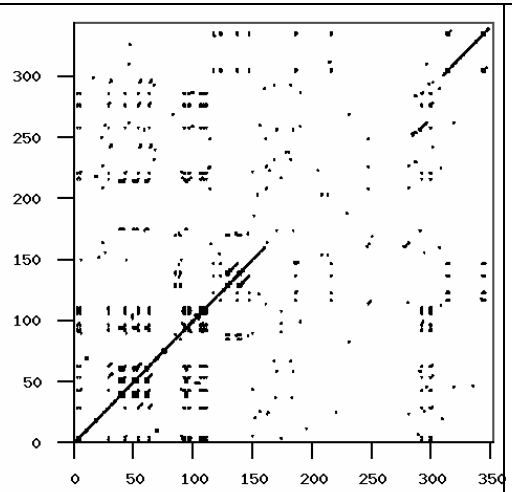
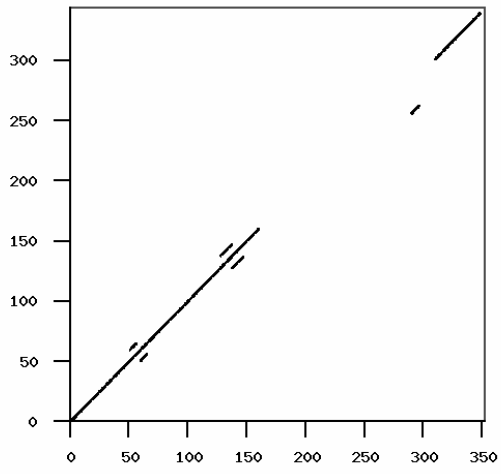
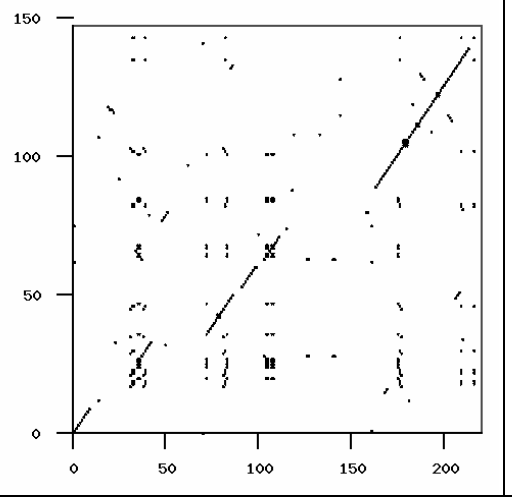
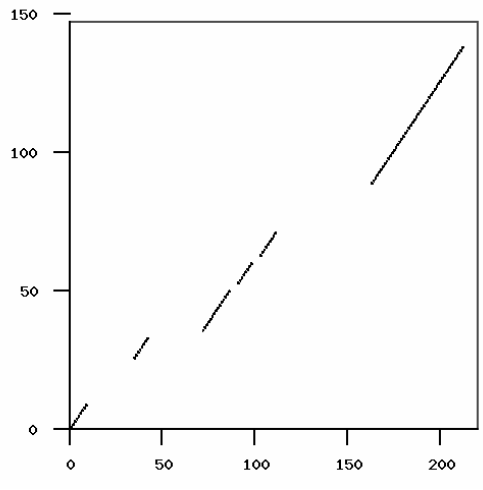
Similarity within Virus Families – Test Result



Similarity within Virus Families – Test Result

IDA_ NGVCK0- IDA_ NGVCK8 (11.9%)		
IDA_G4- IDA_G7 (75.2%)		

Similarity within Virus Families – Test Result

IDA_VCL 0- IDA_VCL 9 (60.2%)		
IDA_MPC 1- IDA_MPC 3 (58.0%)		

Similarity among Virus Families

- NGVCK versus other viruses
  - **0%** similar to G2 and MPCGEN viruses
  - **0 – 5.5%** similar to VCL32 viruses (43 out of 100 comparisons have score > 0)
  - **0 – 1.2%** similar to normal files (only 8 out of 400 comparisons have score > 0)

Similarity among Virus Families

- NGVCK
  - Highest degree of metamorphism of kits tested
  - Virtually no similarity to other viruses or normal programs

PART IV

Can Metamorphic Viruses Be Detected?

Detection with Commercial Virus Scanners

- Tested three virus scanners
  - eTrust version 7.0.405
  - avast! antivirus version 4.7
  - AVG Anti-Virus version 7.1
- Each scanned 37 files
  - 10 NGVCK viruses
  - 10 G2 viruses
  - 10 VCL32 viruses
  - 7 MPCGEN viruses

Detection with Commercial Virus Scanners

- Results
  - eTrust and avast! detected **17** (G2 and MPCGEN)
  - AVG detected **27** viruses (G2, MPCGEN and VCL32)
  - **none** of NGVCK viruses detected

Detection with Hidden Markov Models (HMMs)

- Use *hidden Markov models* (HMMs) to represent *statistical properties* of a set of metamorphic virus variants
  - Train the model on family of metamorphic viruses
  - Use trained model to determine whether a given program is *similar* to the viruses the HMM represents

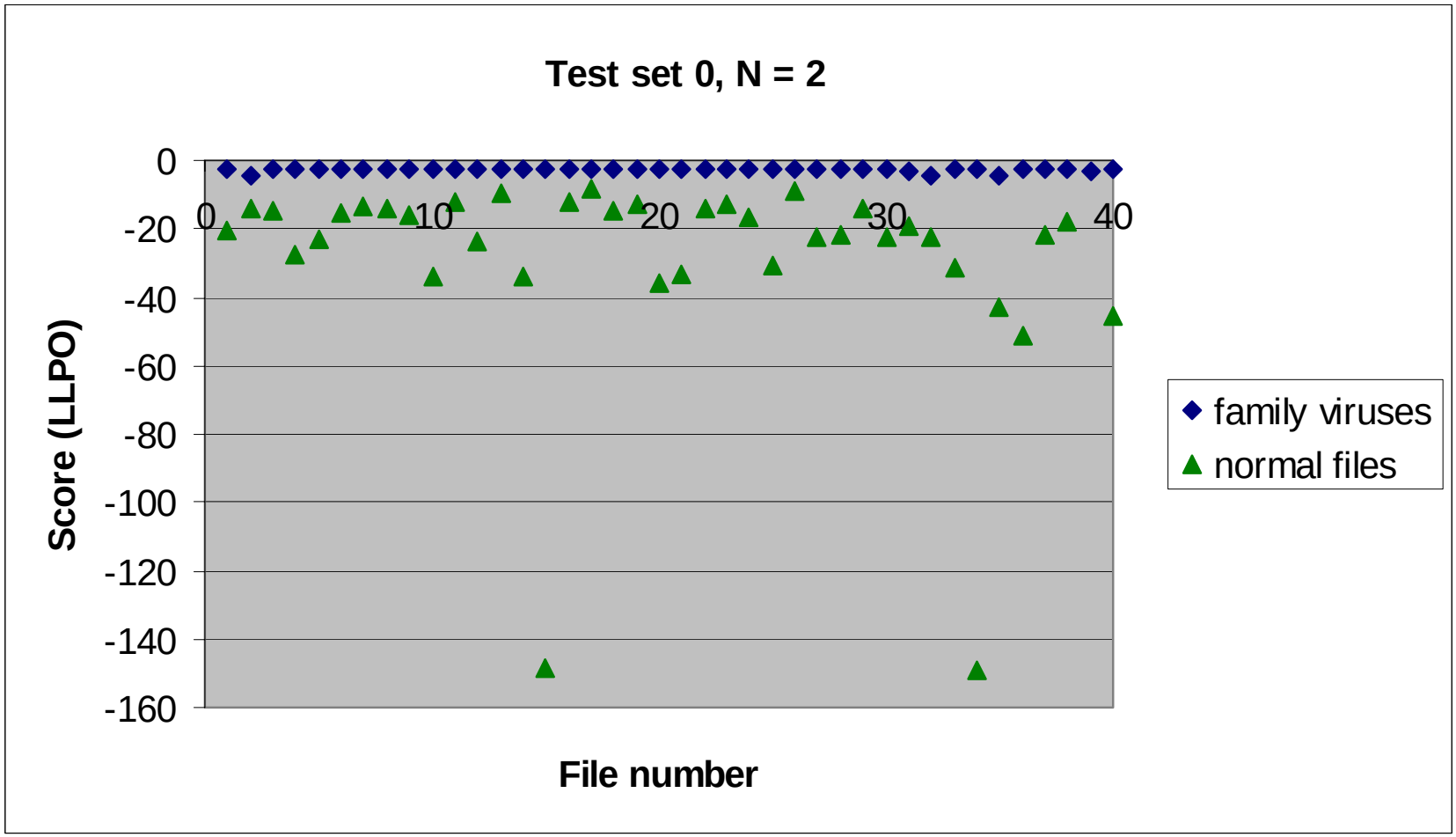
Detection with HMMs – Theory

- A trained HMM
  - maximizes the probabilities of observing the training sequence
  - assigns high probabilities to sequences similar to the training sequence
  - represents the “average” behavior if trained on multiple sequences
  - represents an entire virus family, as opposed to individual viruses

Detection with HMMs – Data Used

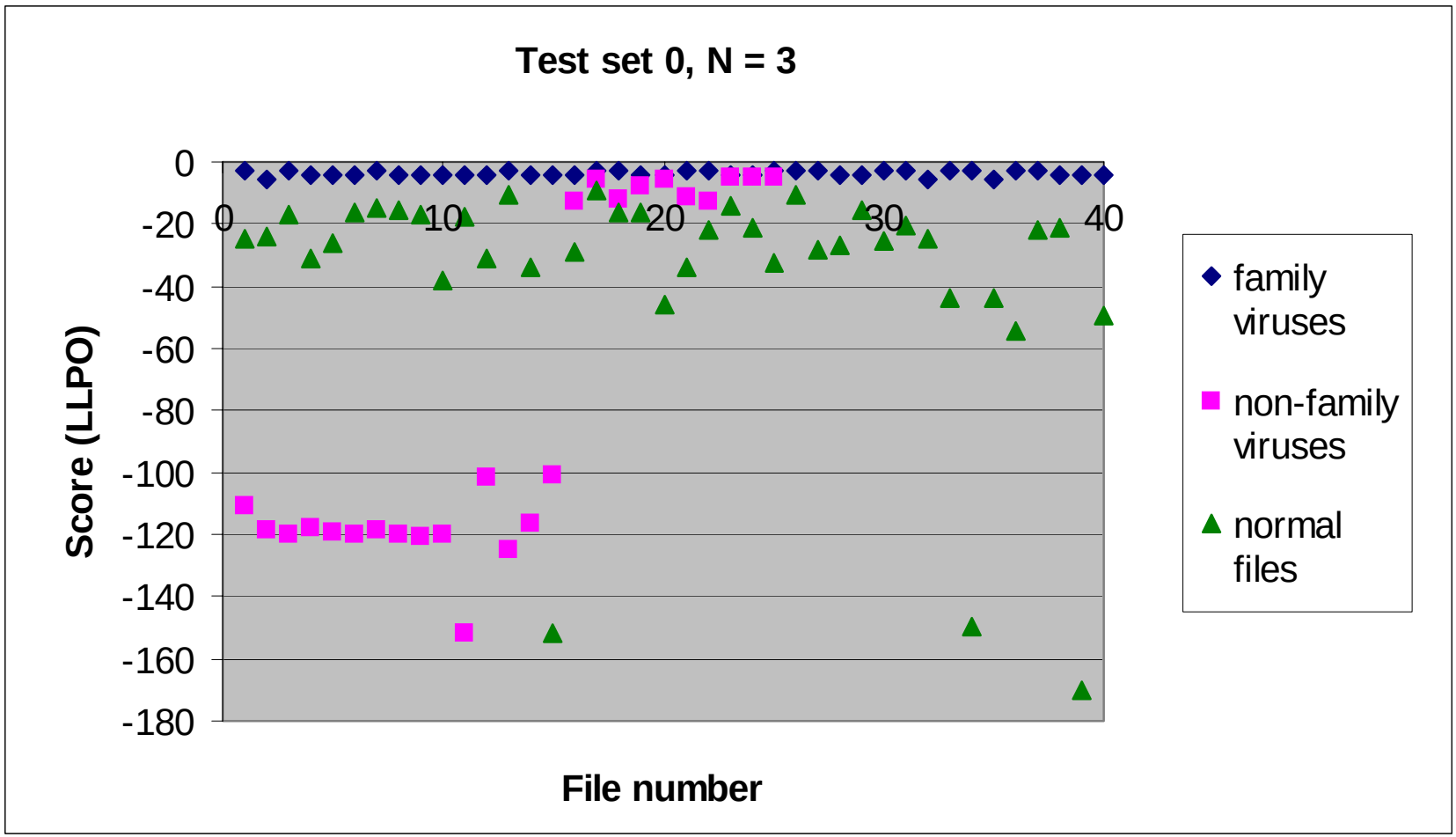
- *Data set*
  - 200 NGVCK viruses
- *Comparison set*
  - 40 normal exes from the Cygwin DLL
  - 25 other "non-family" viruses (G2, MPCGEN and VCL32)
- Many HMM models generated and tested

Detection with HMMs – Experimental Result



Detection with HMMs – Experimental Result

- Detect some other viruses “for free”



Detection with HMMs – Experimental Result

- Summary
  - All normal programs distinguished
  - VCL32 viruses had scores close to NGVCK family viruses
  - With proper threshold, 17 HMM models had 100% detection rate and 10 models had 0% false positive rate
  - No significant difference in performance between HMMs with 3 or more hidden states

Detection with HMMs – The Trained Models

- Converged probabilities in HMM matrices may give insight into the *features* of the viruses it represents
- We observed
  - opcodes grouped into states
  - most opcodes in one states only
- What does this mean?
  - We are not sure…

Detection with Similarity Index

- Straightforward *similarity index* approach
  - To determine whether a program belongs to the NGVCK virus family, compare it to any randomly chosen NGVCK virus
  - Similarity to non-NGVCK code is small
  - Can use this fact to detect metamorphic NGVCK variants

Detection with Similarity Index

- Experiment
  - compare 105 programs to selected NGVCK virus
- Results
  - 100% detection, 0% false positive
- Same results using other NGVCK virus

PART V

Conclusion

Conclusion

- Metamorphic generators vary greatly
  - NGVCK has highest metamorphism (**10%** similarity on average)
  - Other generators far less effective (**60%** similarity on average)
  - Normal files **35%** similar on average
- However
  - NGVCK viruses are “too different” from other viruses and normal programs

Conclusion

- NGVCK viruses not detected by commercial scanners we tested
- Hidden Markov model (HMM) detects NGVCK (and other) viruses with high accuracy
- NGVCK viruses also detectable by similarity index

Conclusion

- All viruses tested were detectable because
  - High similarity within family and/or
  - Too different from normal programs
- Effective use of metamorphism requires both
  - A high degree of metamorphism and
  - Some similarity to other programs

References

- P. Szor, *The Art of Computer Virus Research and Defense*, Addison-Wesley, 2005
- M. Stamp, *Information Security: Principles and Practice*, Wiley Interscience, 2005