Fileprints: Identifying File Types by n-gram Analysis

IEEE IAW June 2005

Wei-Jen Li
Ke Wang
Salvatore J. Stolfo
Benjamin Herzog
Columbia University

Agenda

- The Problem
  - Efficiently Identifying file types from content
  - Security related issues
- Related Work
- 1-gram models
  - Alternative modeling
- Experiments
- Performance Results

The Problem

- Can we *categorize* the type of an arbitrary file object effectively using (partial) binary content without extensive knowledge of syntax and deep parsing?
- Can we *verify* the type of file?
- Can we do so *efficiently* without specialized knowledge of every possible file type that exists now and will exist in the future?

Security Issues

- Change the file extension to hide true type
  - Malicious executable code can pretend to be a plain .txt
  - Document.txt.exe
  - By-pass security policy (users may not be allowed to email .doc files)
- Code obfuscation
  - A technique to protect source code
- Network environment
  - Files can be fragmented across packets

File Types

- How to identify file types?
  - Check the file extension
    - .doc, .pdf, .jpg etc
  - Read the file header
    - Unix *file* command (some file types have “magic numbers”)

Related Research

- N-gram analysis using machine learning [Matthew G. Schultz, Eleazar Eskin, and Salvatore J. Stolfo 2001, Jeremy Kolter and Marcus A. Maloof 2004]
  - Too expensive
- 1-gram analysis of network payload [Ke Wang, Salvatore J. Stolfo 2004]
  - Network packet content analysis
- Generate “fingerprints” of file types using byte-value distributions [Mason McDaniel and M. Hossain Heydari 2004]
  - Single model for each file type
  - Poor performance
  - Normalization issue

N-gram Analysis

111001110110011100110100110100101110001111111110

111001110110011100110100110100101110001111111110

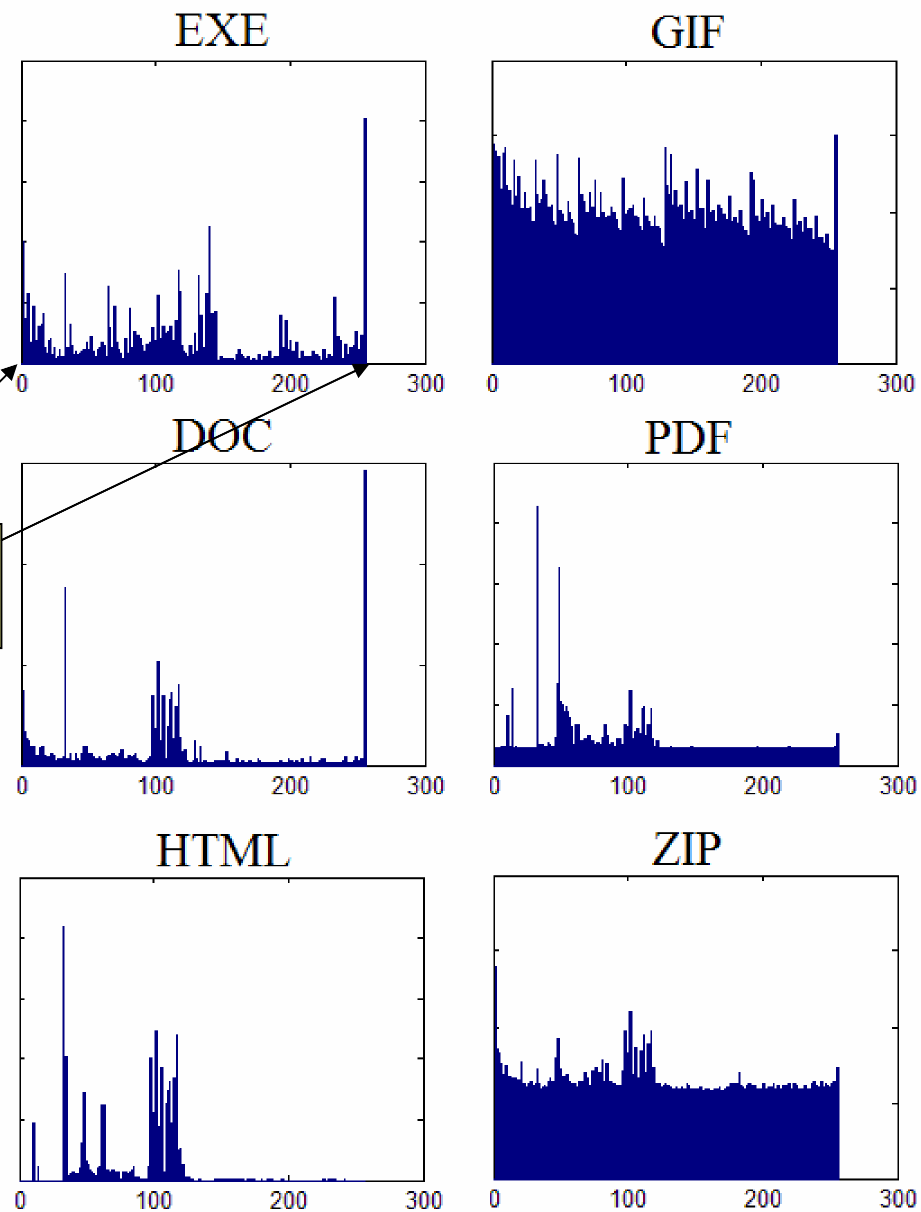
111001110110011100110100110100101110001111111110

- An n-gram is a subsequence of N consecutive tokens in a stream of tokens
- Compare the *distributions* of n-grams contained in a set of data to determine how consistent some new data may be with the set of data in question
- Each distribution is the average byte value frequency and their standard deviation

1-gram file binary content distribution

Byte Value 0
Avg Frequency

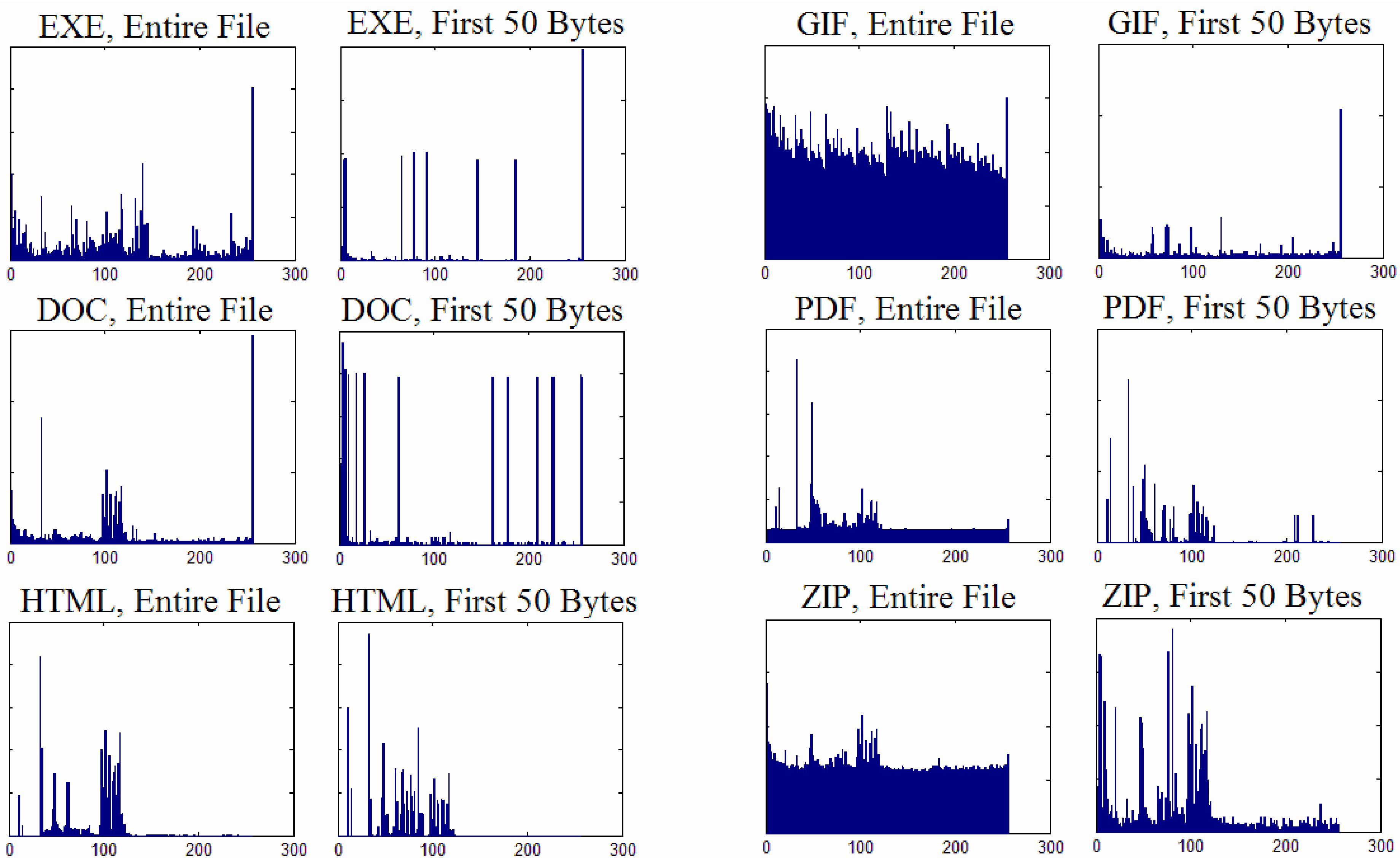
Byte Value 255
Avg Frequency



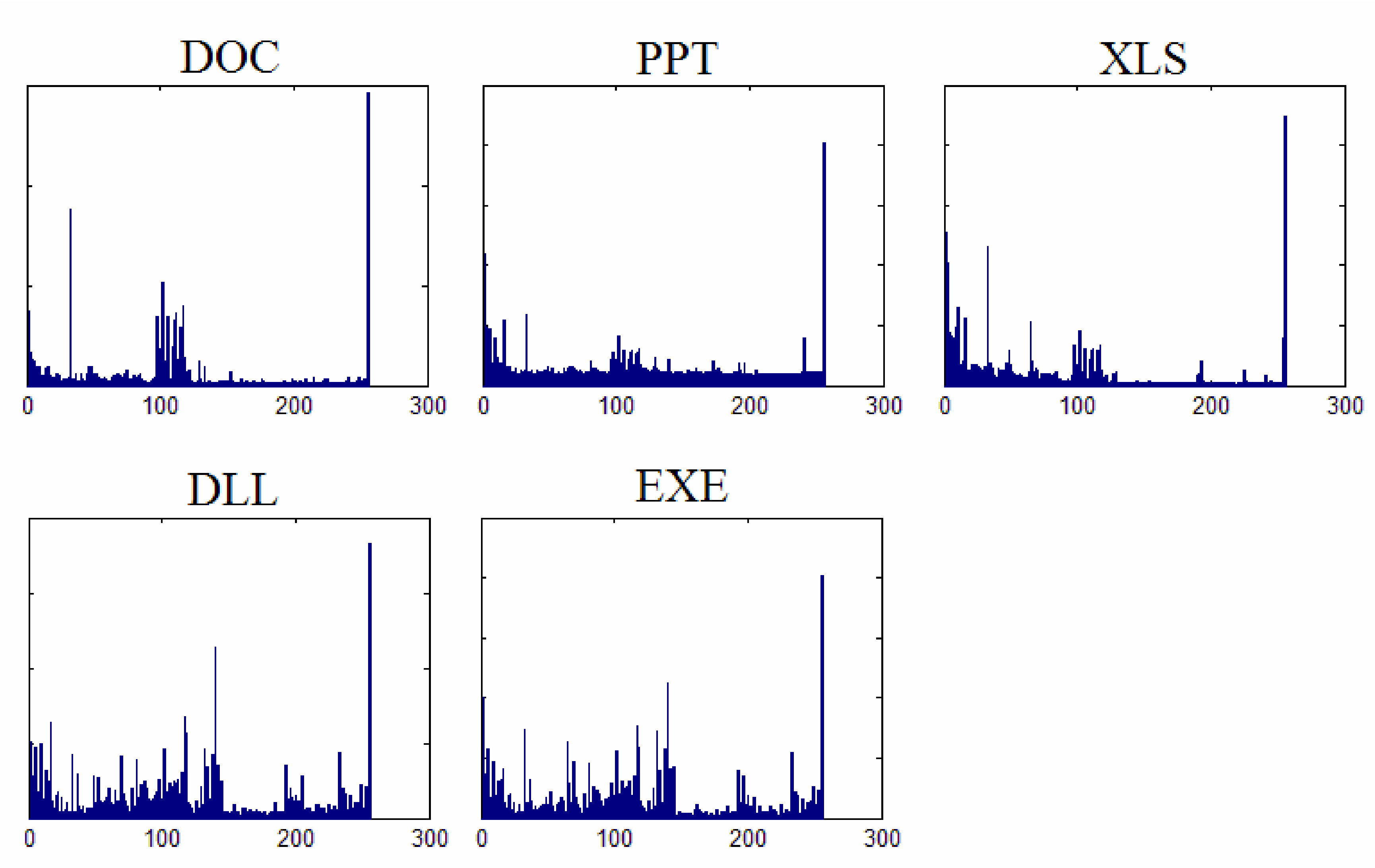
Observation:

Each File Type has a distinct Frequency Distribution

Modeling Methods, Truncation



Modeling Methods, Centroids



Comparing Distributions

- Simplified Mahalanobis Distance – comparing two distributions (mean and variance)

$$D(x,y) = \sum_{i=0}^{n-1} (|x_i - y_i| / (\sigma_i + \alpha))$$

- Compare unknown file distribution F to pre-trained model M_t , i.e. compute $D(F, M_t)$ for all models
- The smaller the distance, the more similar to the model
- Classify the file as the type that it has the smallest distance to, i.e.

Type(*F*) = t if M_t = argmin $\{D(F,M_i) \mid i=1,\ldots,n\}$

Experiment

- Dataset:
  - 800 files of 8 different file types
  - .EXE, .DLL, .GIF, .JPG, .PDF, .PPT, .DOC, .XLS
  - Files were randomly chosen from a Google search
  - 80% used for training models, and the remaining 20% for testing

Modeling methods

- One centroid file type model
  - Use one single model for each file type
  - For each file type *T*, build a model M_t
  - Build *n* models M_1, M_2….. M_n, from *n* different file types
  - Compute the distance of the testing file *F* to each model, and then *F* is classified to the model with the smallest distance

Modeling methods

- Multi-centroids file type model
  - For each file type *T*, build *k* models $M_{t1}, M_{t2} \ldots M_{tk}$ using *k-means* algorithm
  - Compute the distance of *F* to each model, and then *F* is classified to the model with the smallest distance
- Exemplar files used as centroids
  - Each trained file is an individual model
  - The # of models is the same to the number of training files
  - Compute the distance of *F* to each model, and then *F* is classified to the model with the smallest distance

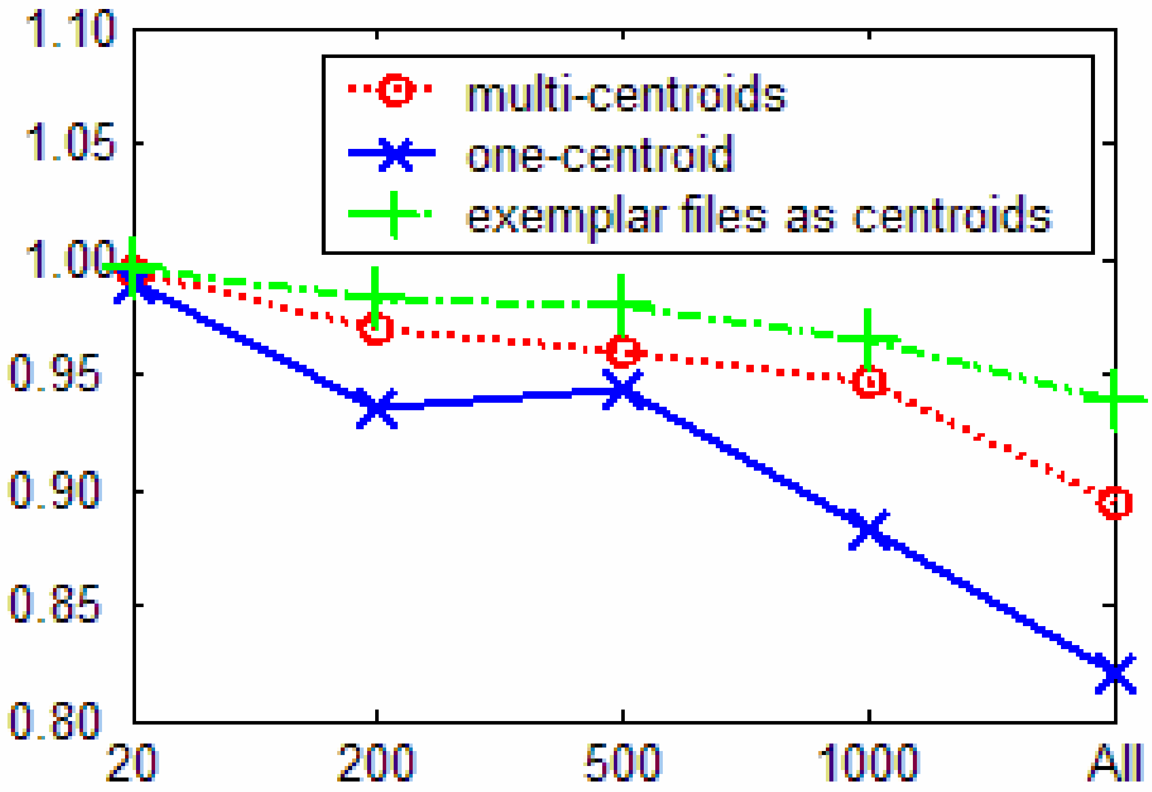
Performance

Truncation Size	EXE	GIF	JPG	PDF	DOC	AVG.
One-centroid file type classifying accuracy						
20	98.9%	100%	99%	100%	98.3%	98.9%
200	98.3%	91.1%	97%	82.8%	93.7%	93.6%
500	97%	97%	93.4%	80.4%	96.7%	94.3%
1000	97.3%	96.1%	93.5%	83.4%	82.6%	88.2%
All	88.3%	62.7%	84%	68.3%	88.3%	82%

Truncation Size	EXE	GIF	JPG	PDF	DOC	AVG.
Multi-centroids file type classifying accuracy						
20	99.9%	100%	98.9%	100%	98.8%	99.4%
200	97%	98.3%	96.6%	95%	97.2%	96.9%
500	97.2%	98.4%	94.8%	90%	96.9%	96%
1000	97%	95.1%	93.5%	90.7%	94.5%	94.6%
All	88.9%	76.8%	85.7%	92.3%	94.5%	89.5%

Truncation Size	EXE	GIF	JPG	PDF	DOC	AVG.
Classifying accuracy using exemplar files as centroids						
20	100%	100%	100%	100%	98.9%	99.6%
200	99.4%	91.6%	99.2%	100%	98.7%	98.2%
500	99%	93.6%	96.9%	99.9%	98.5%	98%
1000	98.9%	94.9%	96.1%	86.9%	98.6%	96.4%
All	94.1%	93.9%	77.1%	95.3%	98.9%	93.8%

Performance



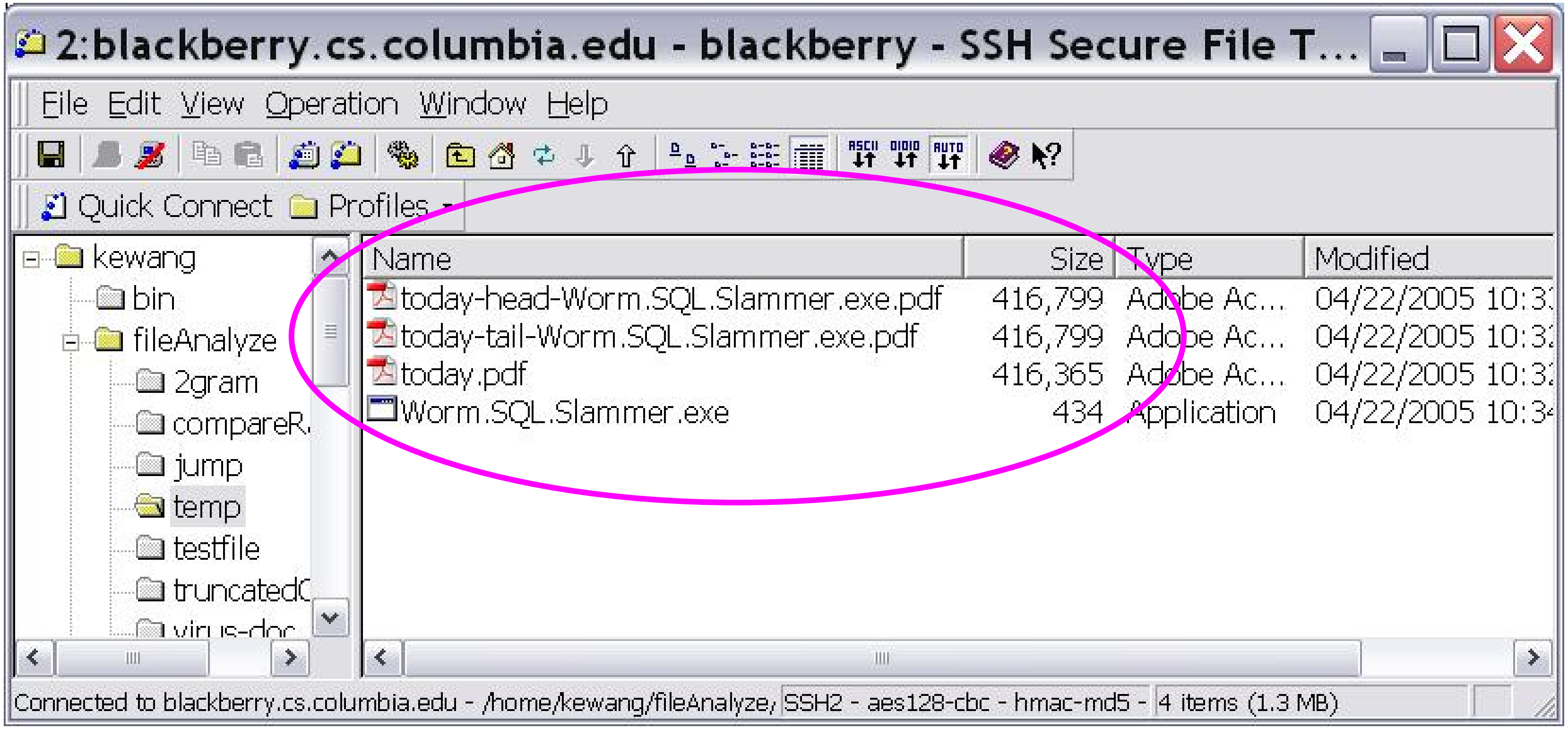
The classification accuracy -- comparison of three different methods.
X-axis: Size of truncation (in bytes). Y-axis: accuracy.

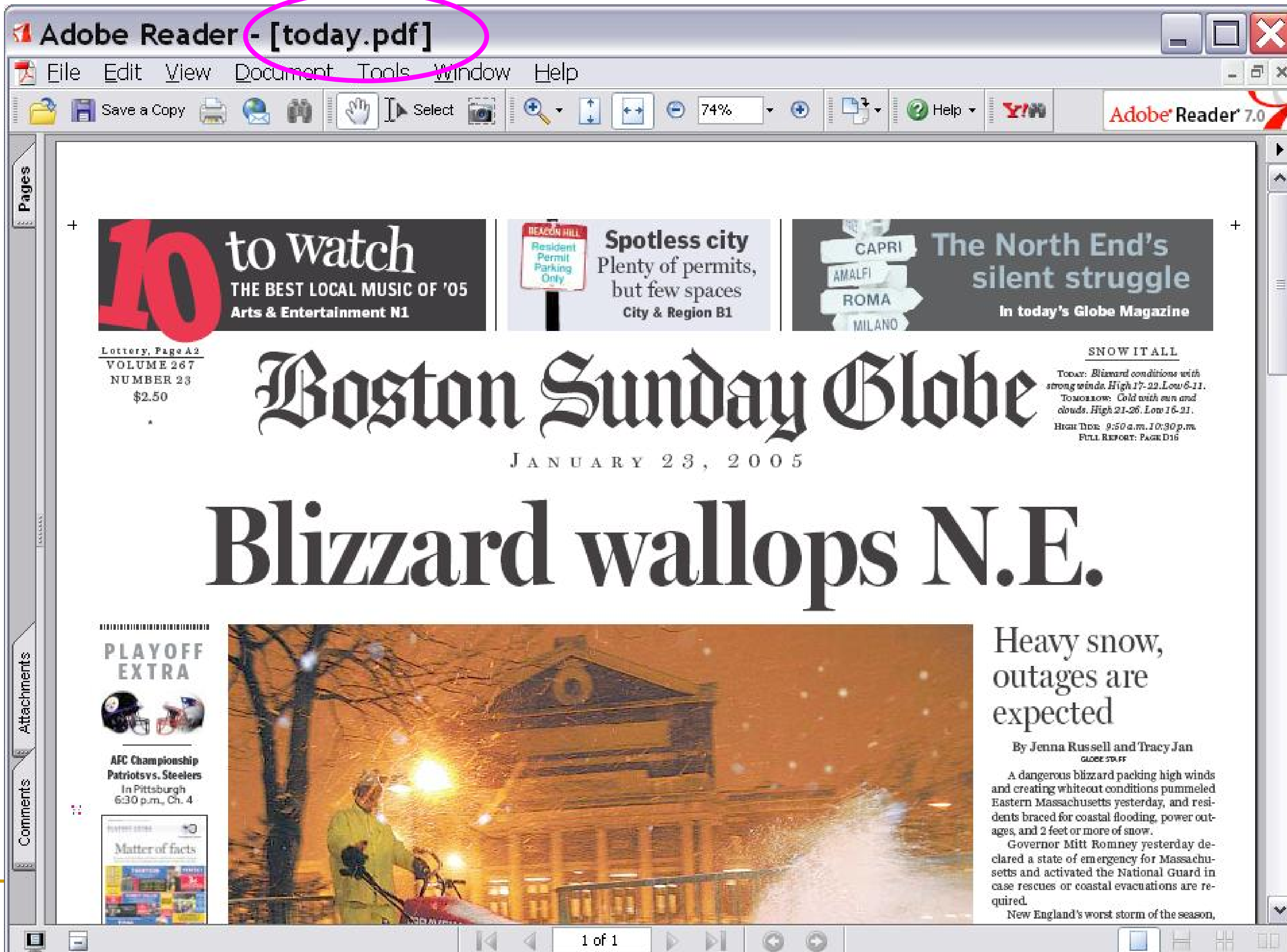
Findings

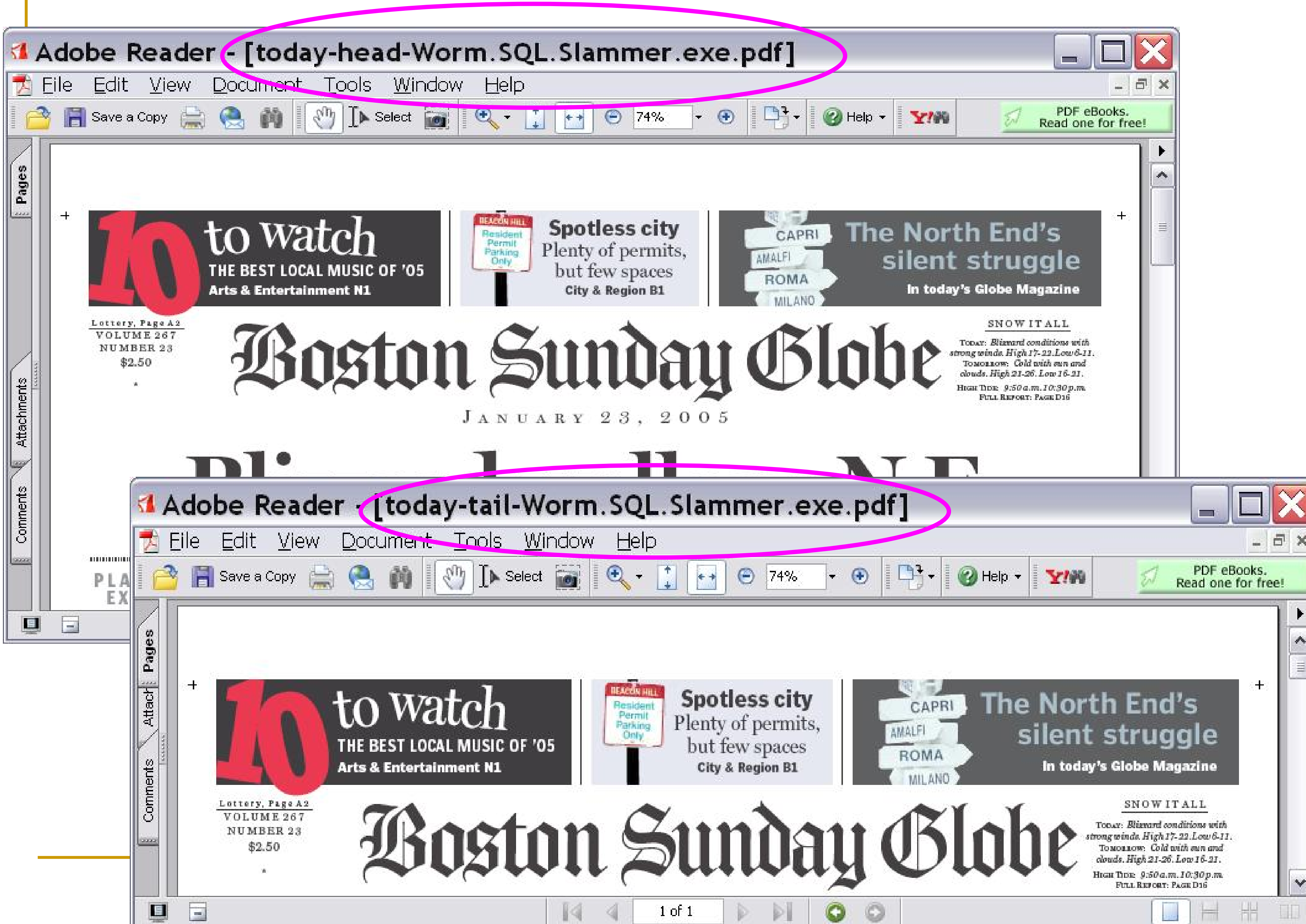
- 1-gram models are effective at identifying file types using purely binary content
- Prefix portion of files reveals their file type quite accurately
- It may be possible to detect ZERO-DAY (new) malcode within portions of files using this technique

Future and Ongoing Work

- Goal: Detecting malware embedded in normal files
  - Demonstrate it is possible to detect malware using fileprints, not signatures
- Preliminary experiment:
  - Inserted viral executables at the head of a random collection of pdf and doc files, tested whether pre-learned fileprints of known malware can be detected within the pdf file.
- Findings
  - Symantec AV missed several embedded viruses
  - A few pdf files infected by a virus can still be successfully opened by Acrobat.
  - For example, Slammer…


2:blackberry.cs.columbia.edu - blackberry - SSH Secure File T...
File Edit View Operation Window Help
Quick Connect
Profiles
kewang
bin
fileAnalyze
2gram
jump
temp
testfile
Name
Size
Type
Modified
today-head-Worm.SQL.Slammer.exe.pdf
416,799
Adobe Ac...
today-tail-Worm.SQL.Slammer.exe.pdf
416,799
Adobe Ac...
today.pdf
416,365
Adobe Ac...
Worm.SQL.Slammer.exe
434
Application
SSH2 - aes128-cbc - hmac-md5 -
4 items (1.3 MB)


Adobe Reader - [today.pdf]
File Edit View Document Tools Window Help
Save a Copy
Select
74%
Help
Adobe Reader 7.0
Pages
Attachments
Comments
10 to watch
THE BEST LOCAL MUSIC OF '05
Arts & Entertainment N1
Spotless city
Plenty of permits, but few spaces
City & Region B1
BEACON HILL
Resident Permit Parking Only
CAPRI
AMALFI
ROMA
MILANO
The North End's silent struggle
In today's Globe Magazine
Lottery, Page A2
VOLUME 267
NUMBER 23
$2.50
Boston Sunday Globe
JANUARY 23, 2005
SNOW IT ALL
TODAY: Blizzard conditions with strong winds. High 17-22. Low 6-11.
TOMORROW: Cold with sun and clouds. High 21-26. Low 16-21.
HIGH TIDE 9:50 a.m. 10:30 p.m.
FULL REPORT: PAGE D16
Blizzard wallops N.E.
PLAYOFF EXTRA
AFC Championship
Patriots vs. Steelers
In Pittsburgh
6:30 p.m., Ch. 4
Matter of facts
Heavy snow, outages are expected
By Jenna Russell and Tracy Jan
GLOBE STAFF
A dangerous blizzard packing high winds and creating whiteout conditions pummeled Eastern Massachusetts yesterday, and residents braced for coastal flooding, power outages, and 2 feet or more of snow.
Governor Mitt Romney yesterday declared a state of emergency for Massachusetts and activated the National Guard in case rescues or coastal evacuations are required.
New England's worst storm of the season,
1 of 1



Adobe Reader - [today-head-Worm.SQL.Slammer.exe.pdf]
File Edit View Document Tools Window Help
Save a Copy
Select
74%
Help
PDF eBooks. Read one for free!
10 to watch
THE BEST LOCAL MUSIC OF '05
Arts & Entertainment N1
Spotless city
Plenty of permits, but few spaces
City & Region B1
The North End's silent struggle
In today's Globe Magazine
Lottery, Page A2
VOLUME 267
NUMBER 23
$2.50
Boston Sunday Globe
JANUARY 23, 2005
SNOW IT ALL
Adobe Reader - [today-tail-Worm.SQL.Slammer.exe.pdf]
File Edit View Document Tools Window Help
Save a Copy
Select
74%
Help
PDF eBooks. Read one for free!
10 to watch
THE BEST LOCAL MUSIC OF '05
Arts & Entertainment N1
Spotless city
Plenty of permits, but few spaces
City & Region B1
The North End's silent struggle
In today's Globe Magazine
Lottery, Page A2
VOLUME 267
NUMBER 23
$2.50
Boston Sunday Globe
SNOW IT ALL
1 of 1

Conclusion

- Efficient 1-gram binary analysis to identify the type of a file from its binary byte-value distribution
  - Useful to detect security policy violations
  - Useful for identifying suspect files with possible malcious code
- The method introduces several nuances
  - the truncated modeling techniques and of multi-centroids for increased accuracy