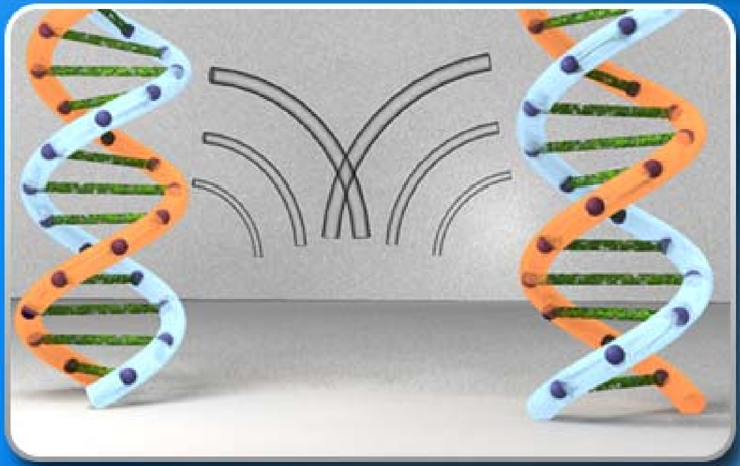
Next Generation Intelligent nexGIN RC Networks Research Center

Embedded Malware Detection using Markov n-grams



M. Zubair Shafiq[1], Syed Ali Khayam[2], Muddassar Farooq[1]

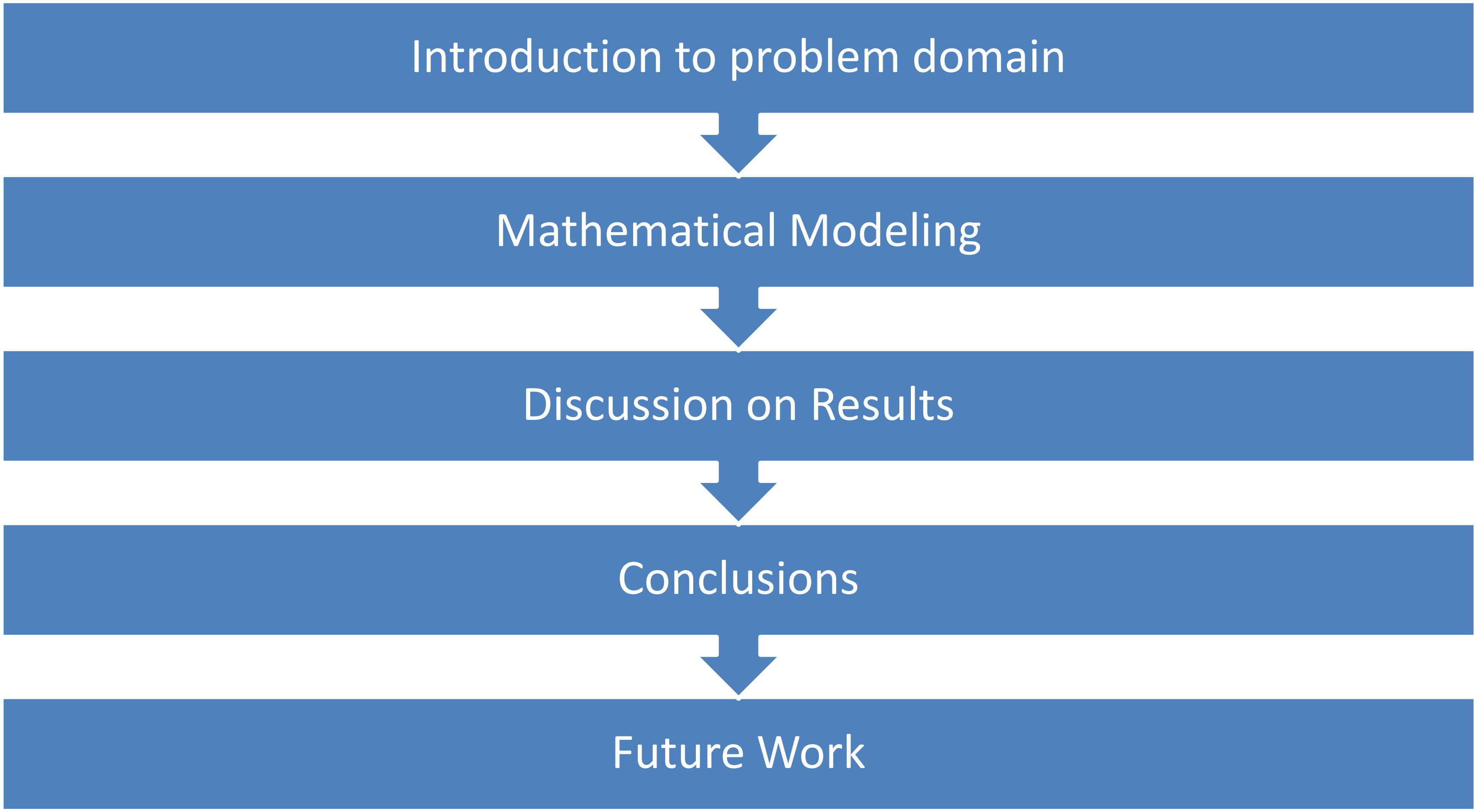
[1] Next Generation Intelligent Networks Research Center
National University of Computer & Emerging Sciences
Islamabad, Pakistan
http://www.nexginrc.org

[2] School of Electrical Engineering & Computer Sciences
National University of Sciences & Technology
Rawalpindi, Pakistan
http://wisnet.niit.edu.pk

Next Generation Intelligent nexGIN RC Networks Research Center

NUST

Agenda



Introduction to problem domain

Mathematical Modeling

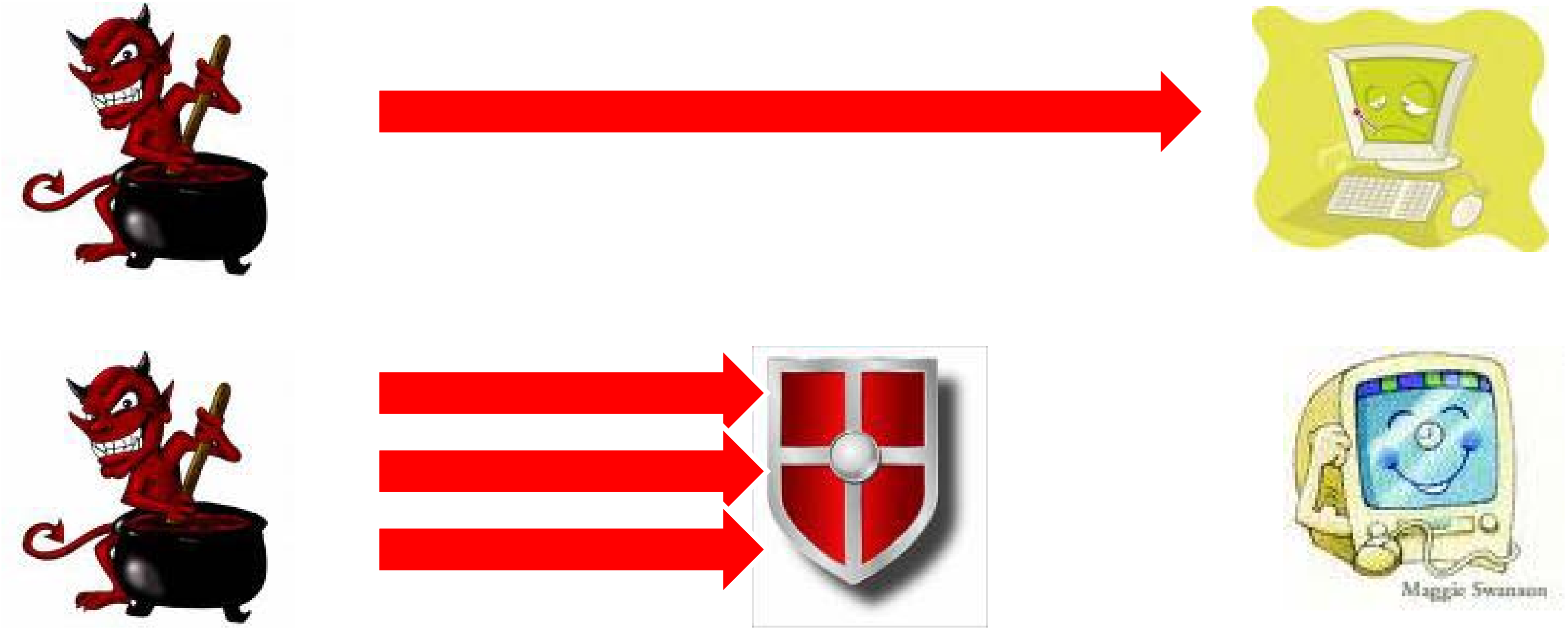
Discussion on Results

Conclusions

Future Work

Introduction to Problem Domain

Introduction



Problem with State-of-the-art antivirus

Signature matching

- Identify sequence of instructions unique to a virus => virus signature
- Match program to a database of signatures

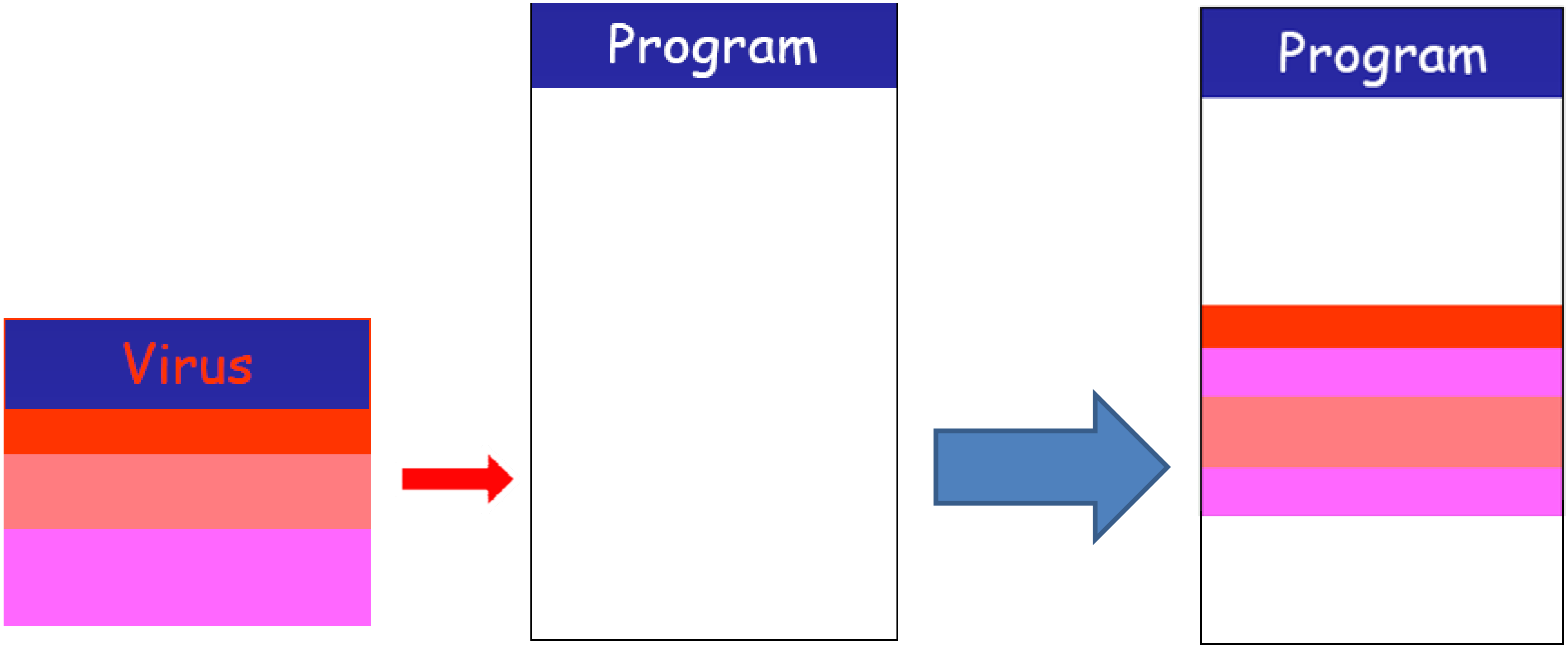
Problems

- Inability to detect zero-day attacks
- Scan only starting portions of files due to high overhead and false alarms; vulnerable!
- Size of signature database cannot scale in future

Next Generation Intelligent nexGIN RC Networks Research Center

NUST

Embedded Malware



benign program

infected program

Mathematical Modeling

Our Approach

Anomaly detection

- Differentiate anomalous behavior from the normal workflow
- Primarily utilize *statistical modeling*

Statistical model of benign DOC, EXE, JPG, MP3, PDF, ZIP files using Markov n-grams

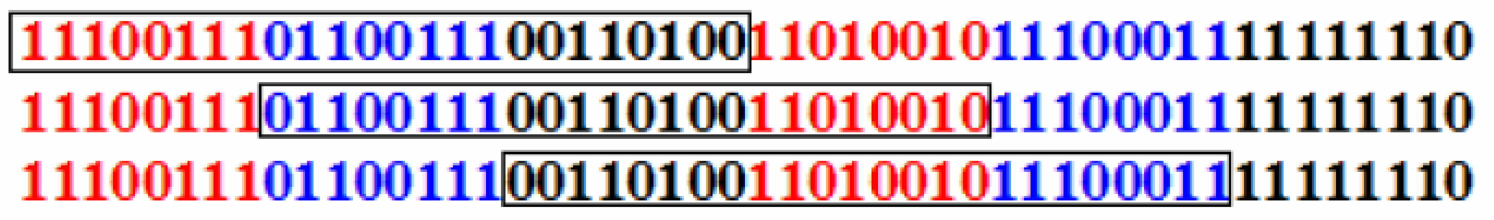
Feature extraction using well known information-theoretic measure, entropy rate

Threshold based detection using Gaussianity of sum of sampled entropy rate distribution

n-gram analysis

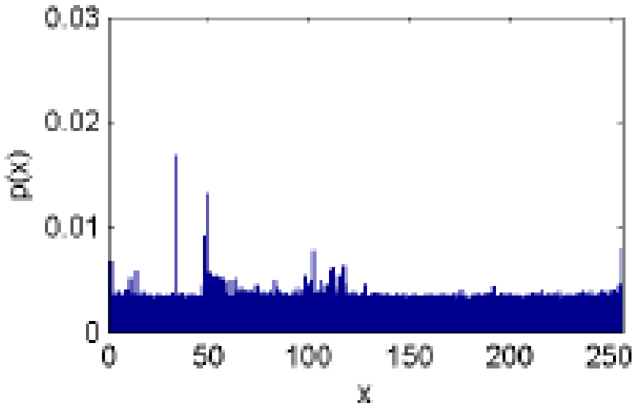
n-gram Definition [en-wikipedia]

- An *n*-gram is a sequence of *n* symbols in a given sequence

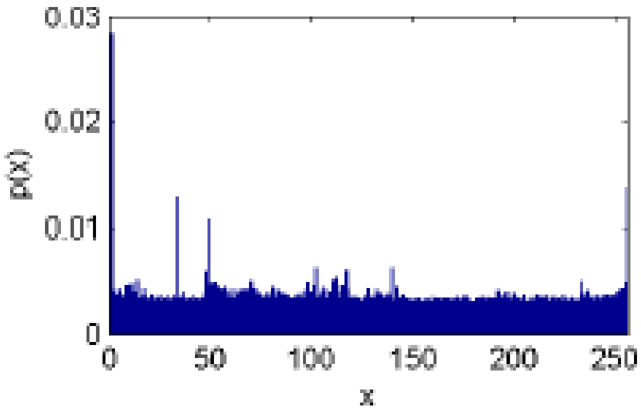


3-gram

Whole file n-gram analysis



1-gram benign PDF



1-gram infected PDF

Whole file, 1-gram analysis DOES NOT show significant perturbations ☹

Block wise 1-gram analysis

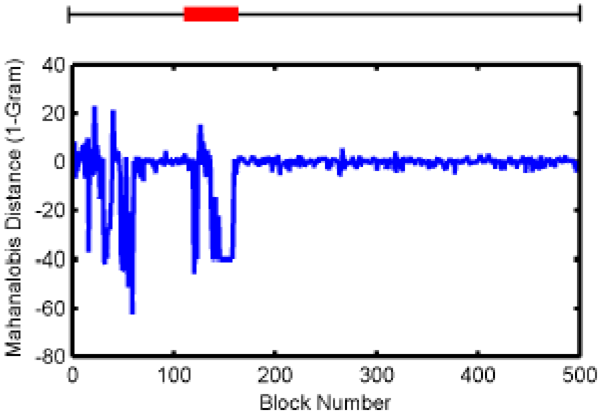
Calculate Mahanalobis distance between benign model and given file in a block wise manner (blocksize = 1000bytes)

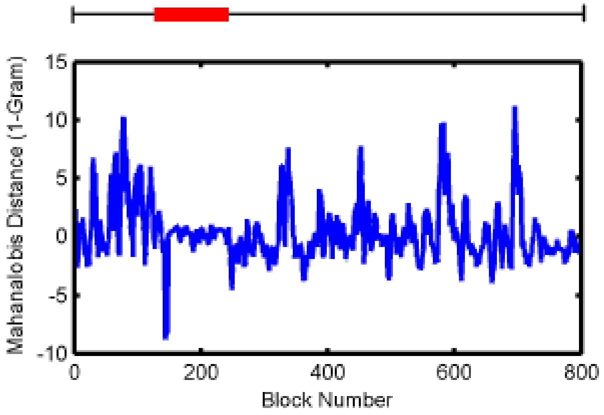
Benign model distribution is:

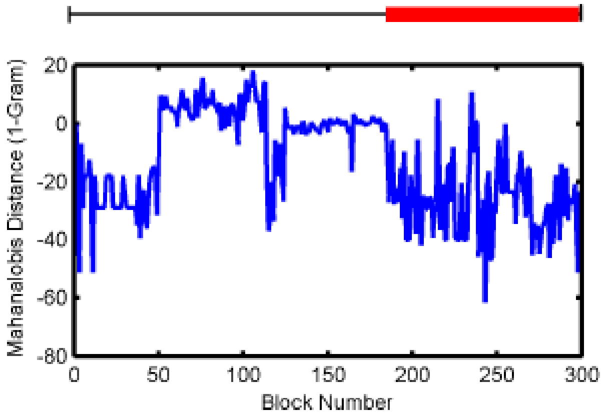
- average byte value frequency
- their standard deviation

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

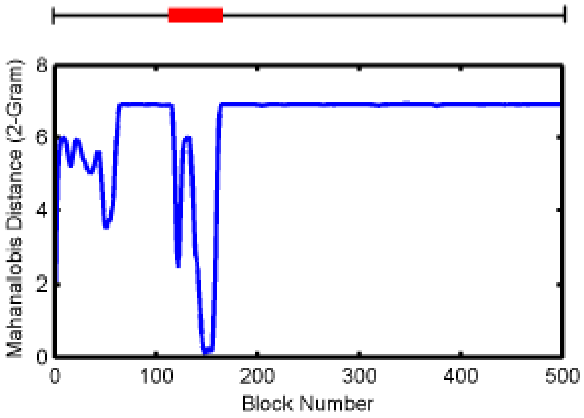
Block wise 1-gram analysis

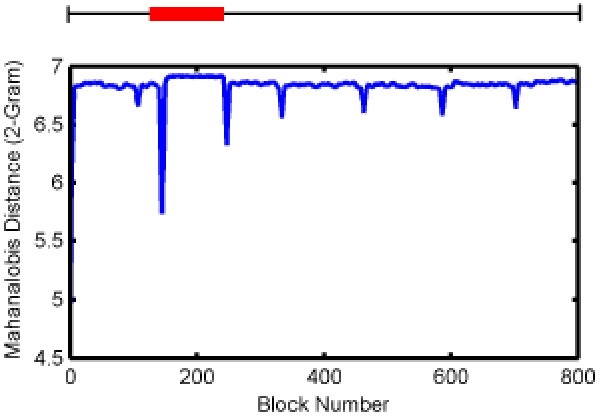


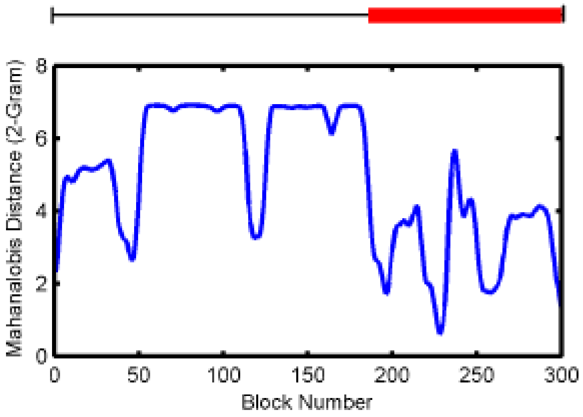




Block wise 2-gram analysis







Analysis of results

2-gram is better than 1-gram; at the cost of exponentially higher computational complexity!

Cannot increase *n* due to a fixed block size of 1000 bytes!

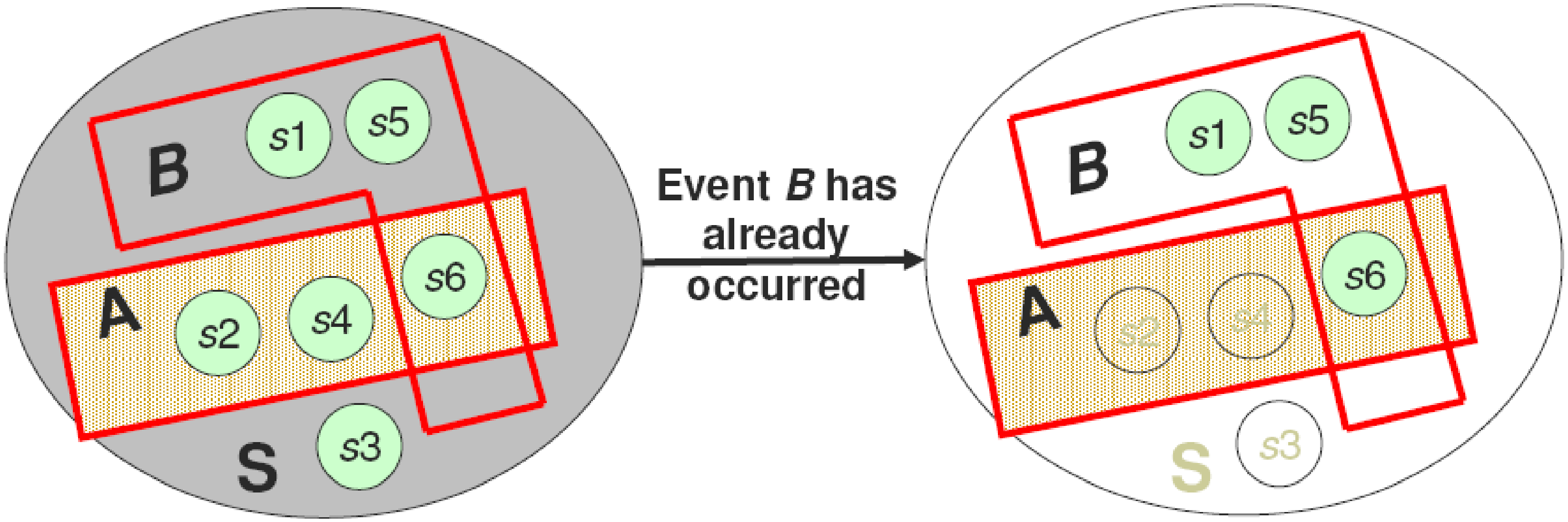
2-gram distribution is a joint distribution

Some redundant information in joint distribution that can be removed using *conditional distribution*. HOW?

Advantage of Conditional Distribution

$$Pr\{A|B\} = \frac{Pr\{A \bigcap B\}}{Pr\{B\}}$$

Reduction in sample space



From Conditional Distribution to Markov Chain

A conditional distribution can be directly mapped to a Markov chain

Symbols of distribution mapped to states in Markov Chain

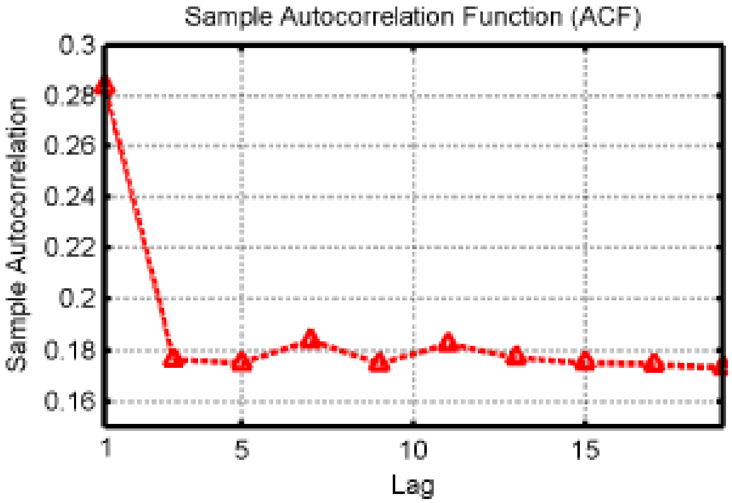
How to select order of Markov chain?

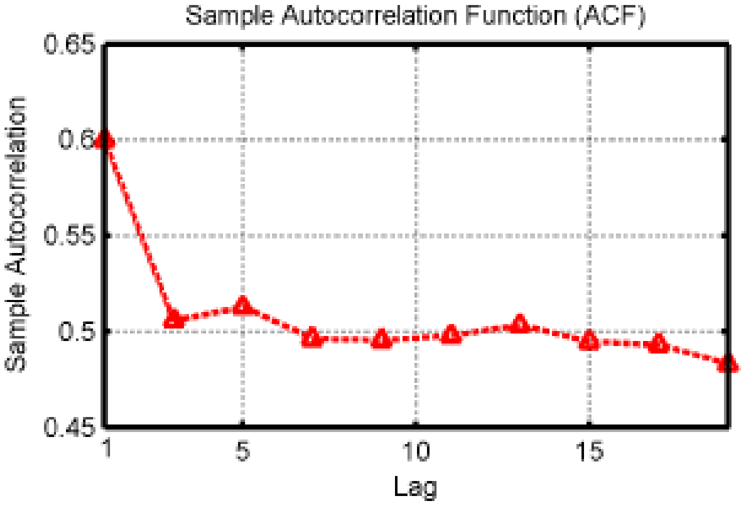
Byte level autocorrelation!

Byte-level autocorrelation results

Random information beyond lag=1

1st order Discrete Time Markov Chain

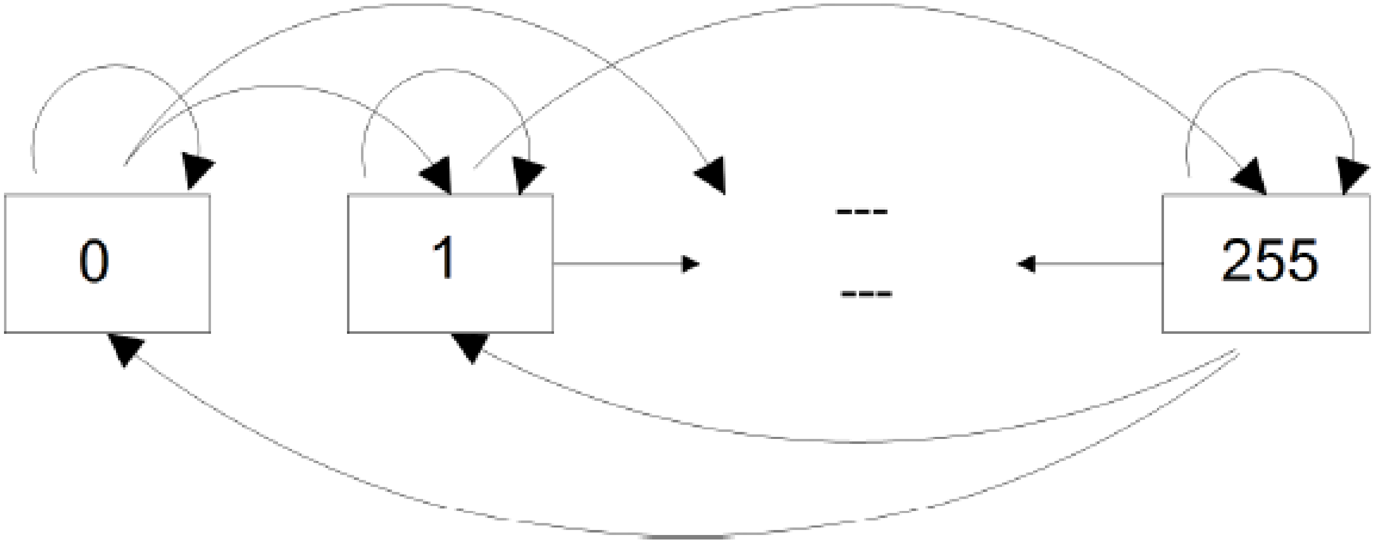




Markov n-gram Model

1st order, 256 state Markov Chain

Can be constructed using conditional 2-gram distribution



Quantification feature

average information in the Markov Chain

Entropy rate gives

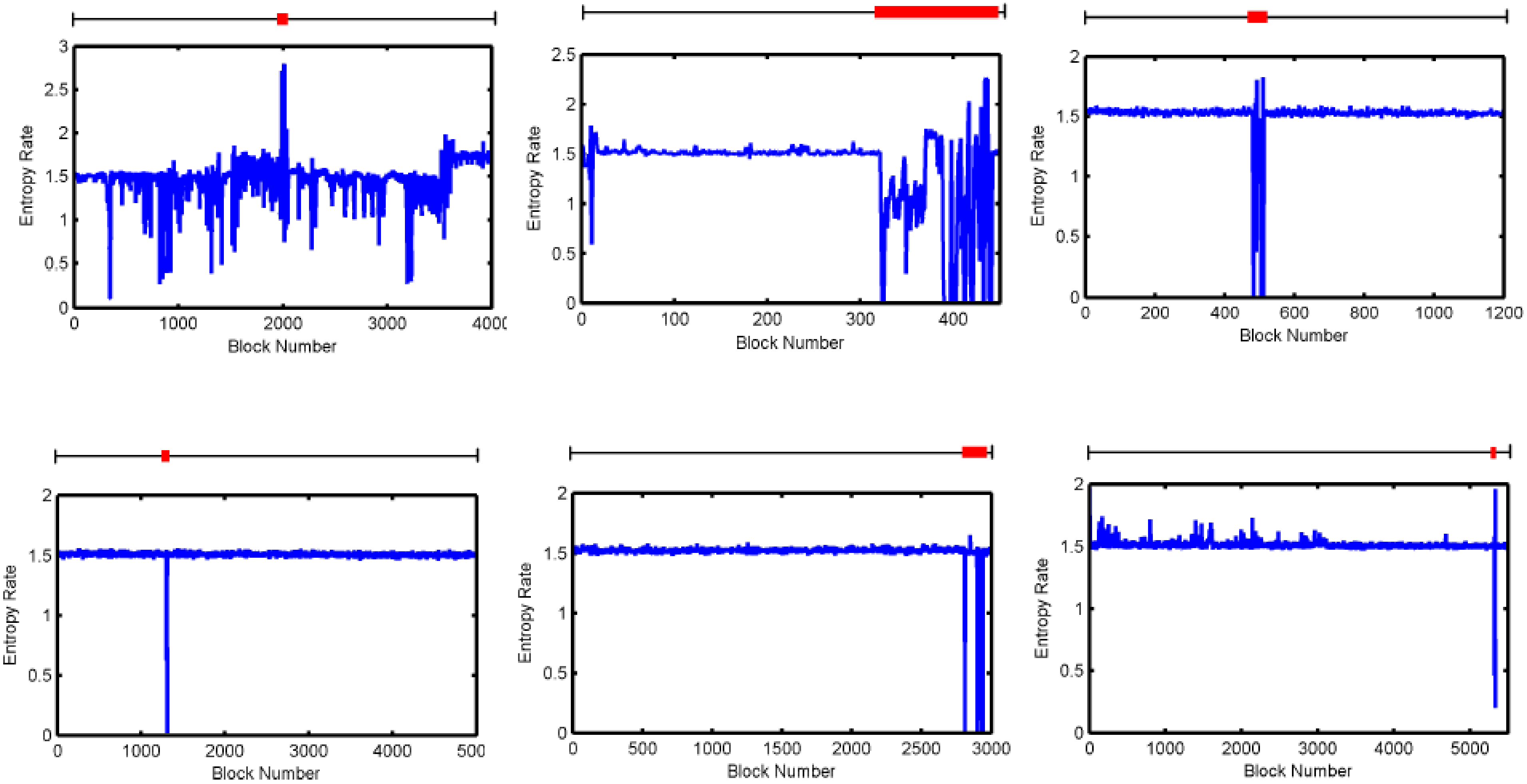
- Time density of average information in the Markov chain

$$R = \lim_{N \to \infty} \frac{H(X_1, X_2, ..., X_n)}{N}$$

For a 256 state Markov chain

$$R = \sum_{i=0}^{255} \pi_i H(X_i)$$

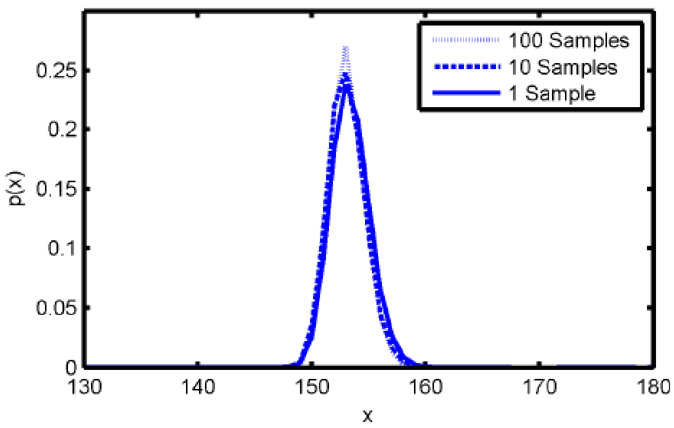
Entropy rate Results

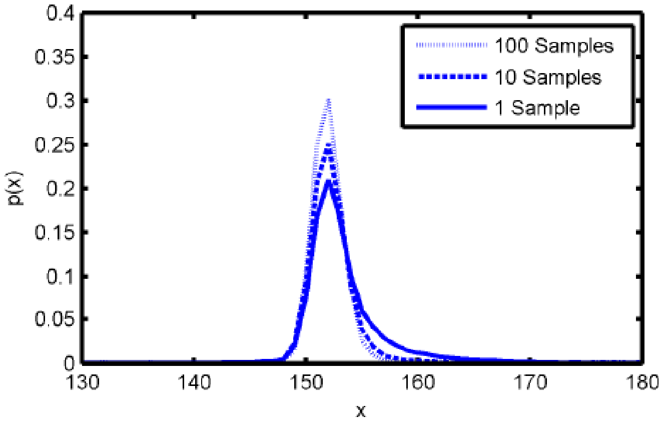


Threshold selection

Sum of sampled entropy rate distributions approach Gaussianity*

Threshold selected at μ ± 5σ (99.99%)





* Direct consequence of central limit theorem

Results

	Mahanalobis n-gram Detector (%)	Markov n-gram Detector (%)	Percentage Improvement (%)
		MP3	
TP rate	63.8	95.0	**31.2**
FP rate	32.3	0.2	**32.1**
		JPG	
TP rate	76.3	95.4	**19.1**
FP rate	35.7	2.7	**33.0**
		PDF	
TP rate	75.4	84.5	**9.1**
FP rate	46.8	31.8	**15.0**
		ZIP	
TP rate	60.0	90.4	**30.4**
FP rate	29.9	8.3	**21.6**
		EXE	
TP rate	54.1	84.9	**30.8**
FP rate	47.3	16.7	**10.6**
		DOC	
TP rate	65.6	66.3	**0.7**
FP rate	48.8	29.2	**19.6**

Conclusion & Future Work

Conclusion

Advantages

- Ability do identify location of malware
- Significantly improved *TP rate* and *FP rate* as compared to Mahanalobis detector

Limitations

- Still slightly high false positive rate for DOC and PDF files
- Embedded Malware -> dormant malware
- Mimicry Attacks

Future Work

How to reduce false positive rate?

Complement with signature based detector

- Inability to detect zero-day attacks ☹

Multiple features and correlation

- Initial results have shown promise ☺
- Subject of forthcoming journal publication…

Patent pending on current work

Next Generation Intelligent nexGin RC Networks Research Center

NUST

Thank you!

Contact authors for queries/suggesstions:

zubair.shafiq@nexginrc.org
khayam@niit.edu.pk
muddassar.farooq@nexginrc.org