Malware Analysis Using Visualized Image Matrices

Tzu-Ming Huang

CISC850
Cyber Analytics

Overview

- malware visual analysis method
  - convert binary files into images
    - Reduce computation – major block
  - similarity calculation method between these images

Method Overview

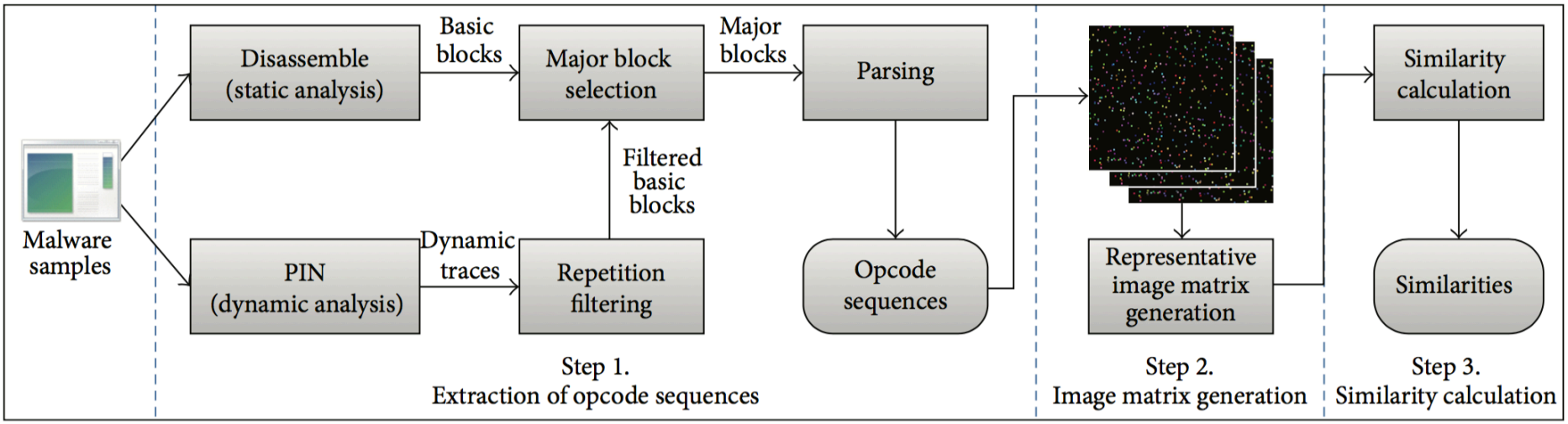


FIGURE 1: Overview of the proposed method.

Extract opcode sequences from binary

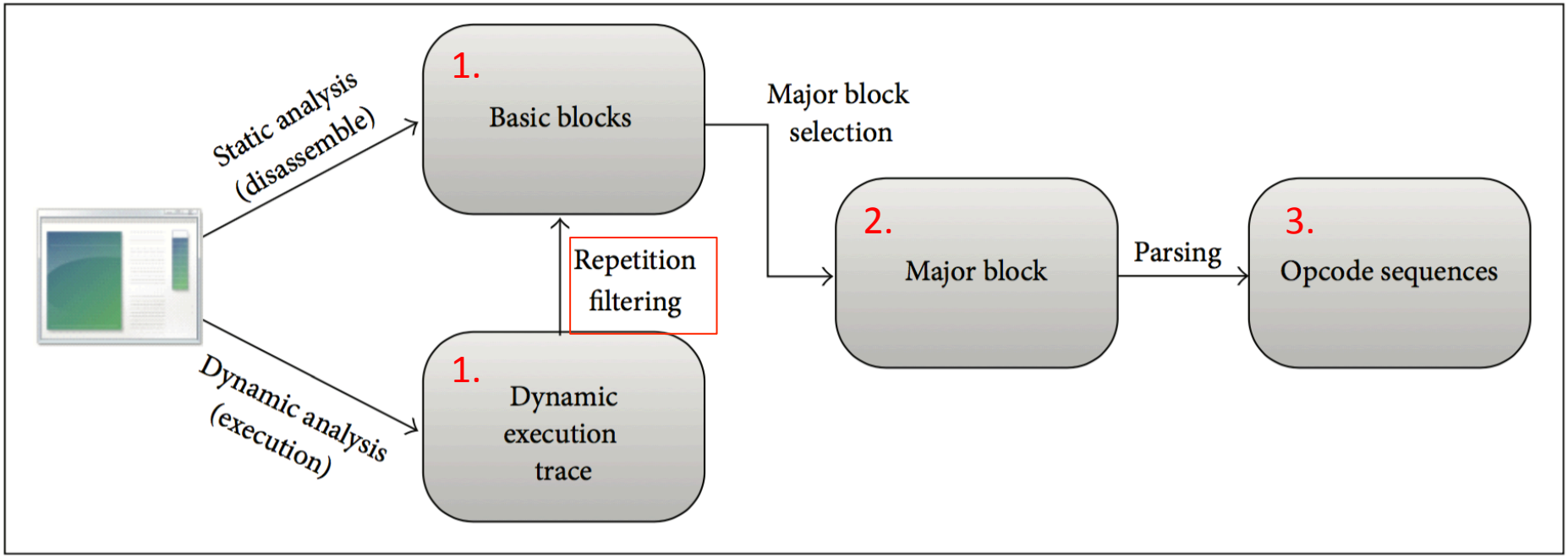


FIGURE 2: Opcode sequence extraction procedure.

Repetition Filtering

```
0042A68B  Main    JBE   SHORT Exploit_.0042A69C
0042A68D  Main    MOV   AL, BYTE PTR DS: [EDX]
0042A68F  Main    INC   EDX
0042A690  Main    MOV   BYTE PTR DS: [EDI], AL
0042A692  Main    INC   EDI
0042A693  Main    DEC   ECX
0042A694  Main    JNZ   SHORT Exploit_.0042A68D
0042A68D  Main    MOV   AL, BYTE PTR DS: [EDX]
0042A68F  Main    INC   EDX
0042A690  Main    MOV   BYTE PTR DS:[EDI],AL
0042A692  Main    INC   EDI
0042A693  Main    DEC   ECX
0042A694  Main    JNZ   SHORT Exploit_.0042A68D
0042A68D  Main    MOV   AL, BYTE PTR DS: [EDX]
0042A68F  Main    INC   EDX
0042A690  Main    MOV   BYTE PTR DS: [EDI], AL
0042A692  Main    INC   EDI
0042A693  Main    DEC   ECX
0042A694  Main    JNZ   SHORT Exploit_.0042A68D
```

Extract opcode sequences from binary

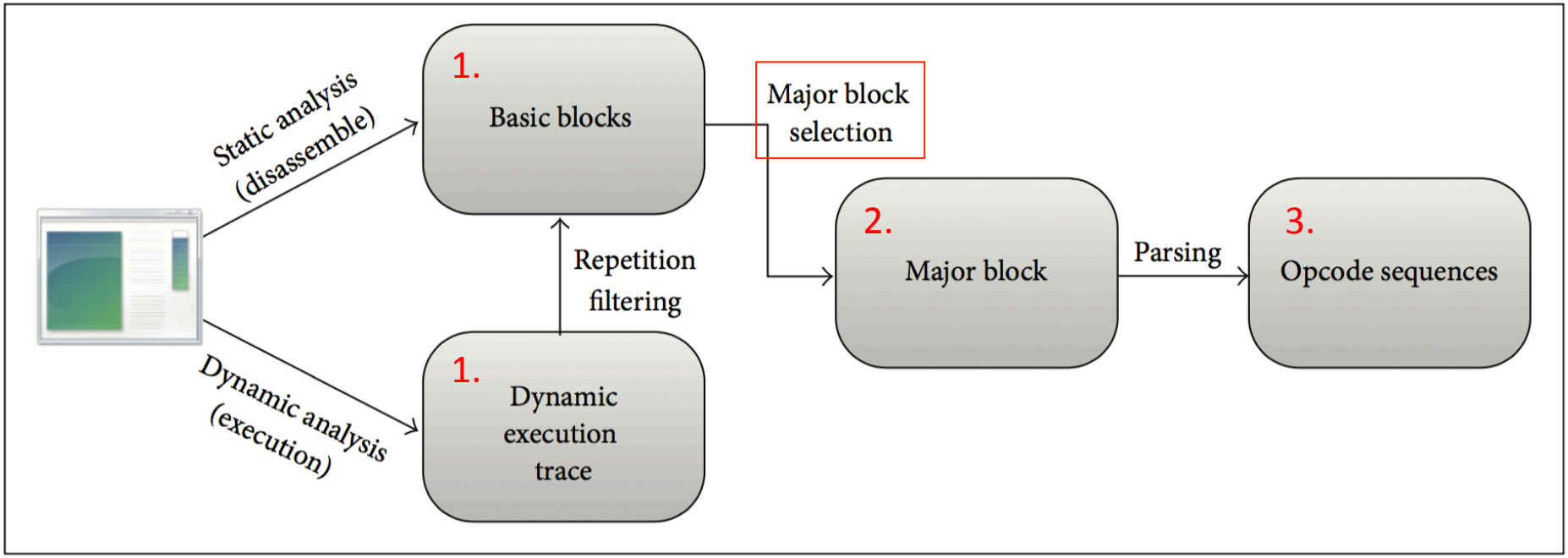
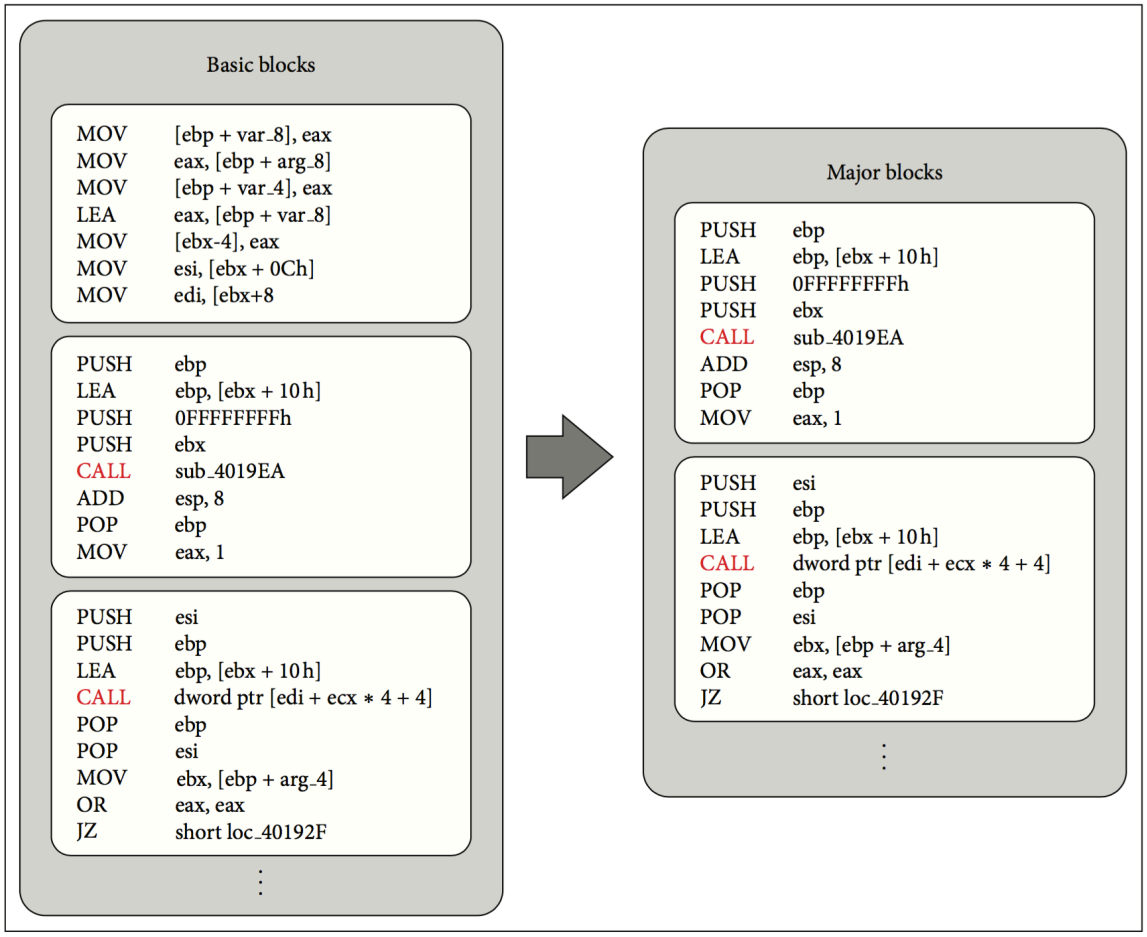


FIGURE 2: Opcode sequence extraction procedure.

Major Block Selection

- Not all of the basic blocks (file header, meaning less blocks)
- Target suspicious behavior
- Blocks include “CALL” instruction

Major Block Selection



Extract opcode sequences from binary

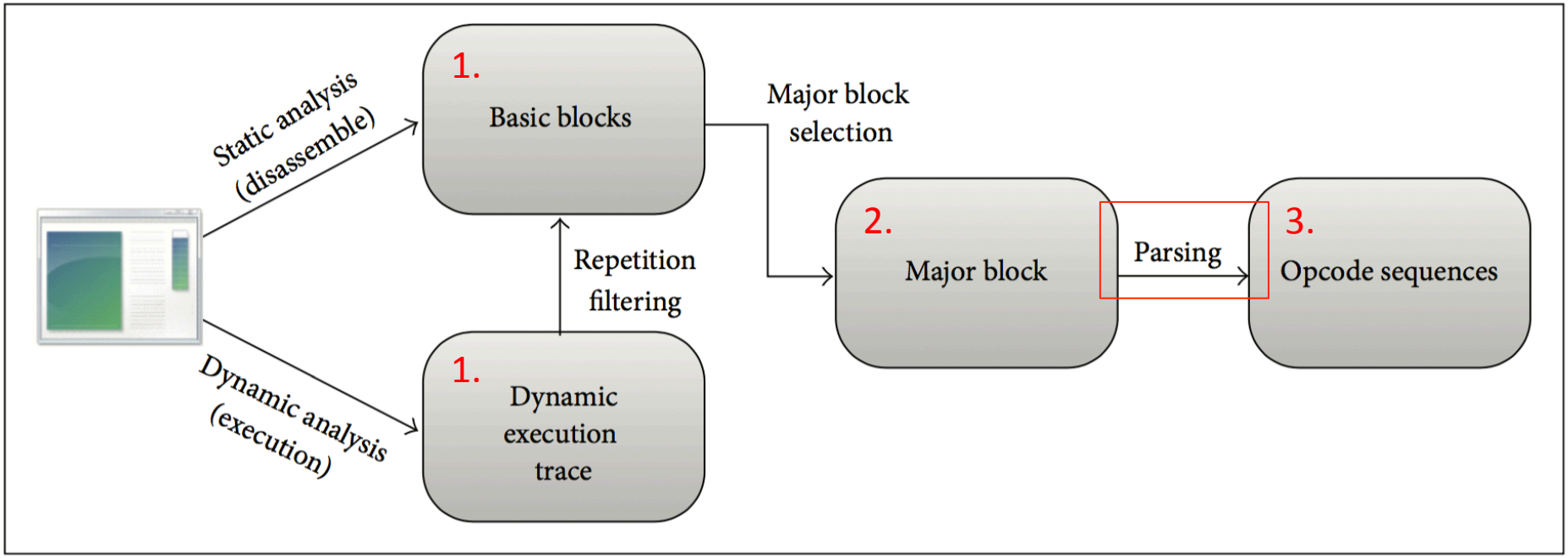
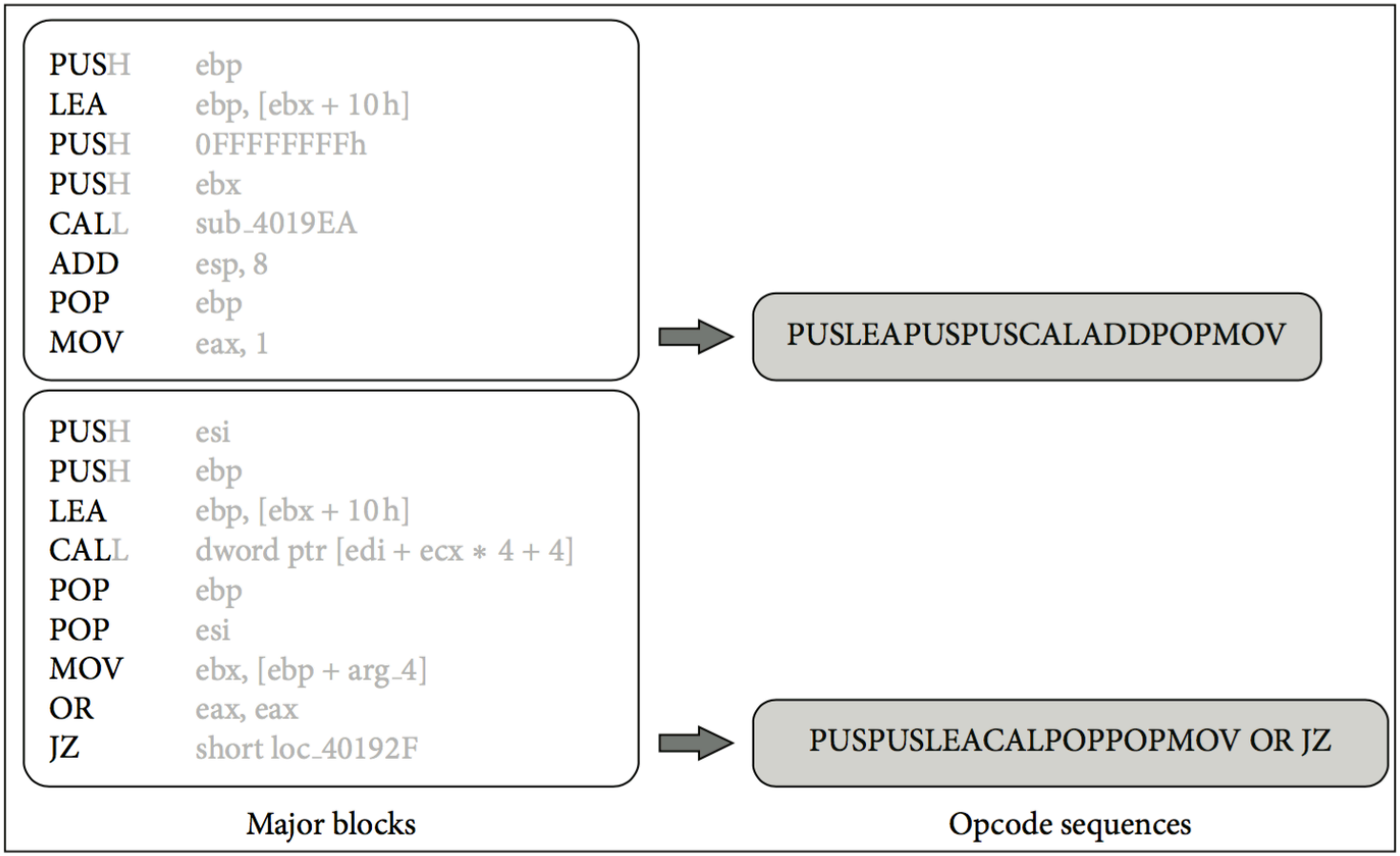


FIGURE 2: Opcode sequence extraction procedure.

Parsing Opcode Sequence

- First three characters of opcode
  - 41.4% of opcodes have3 characters
  - Meaning is maintained
  - Eg. PUSH -> PUS; CALL -> CAL; OR?
- These three-character opcodes are concatenated together

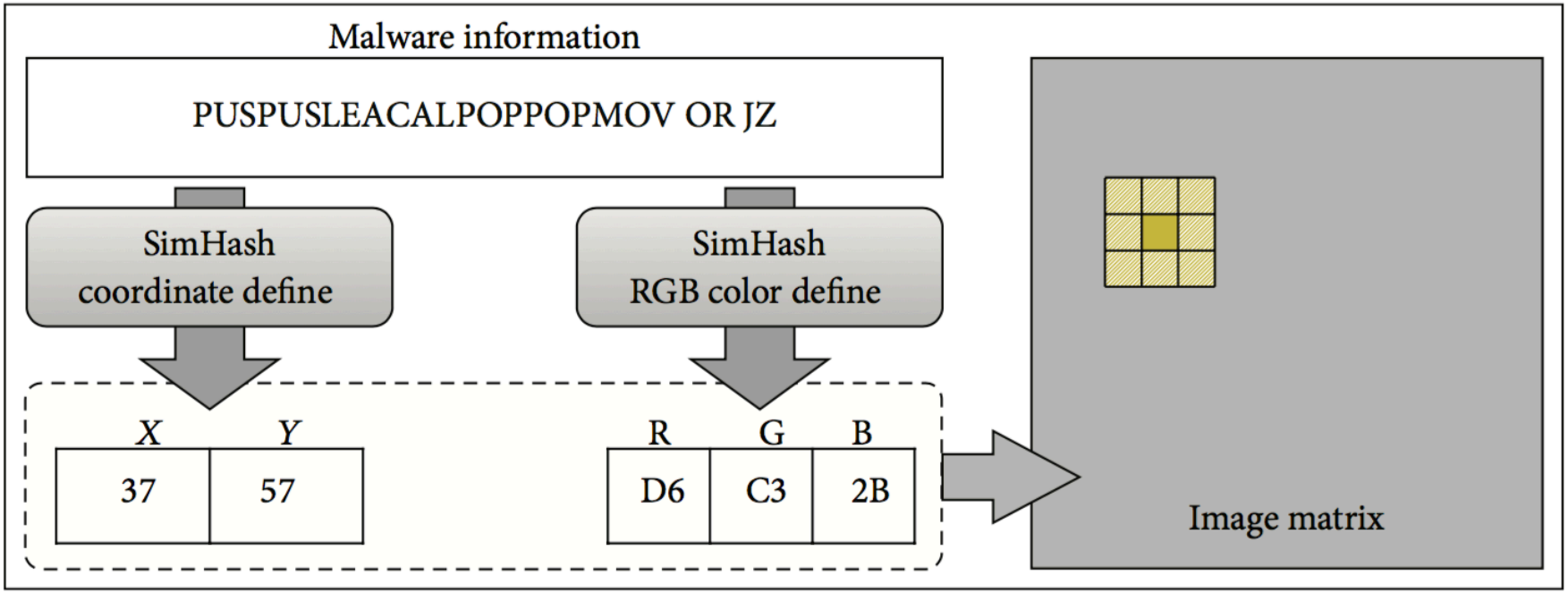
Parsing Opcode Sequence



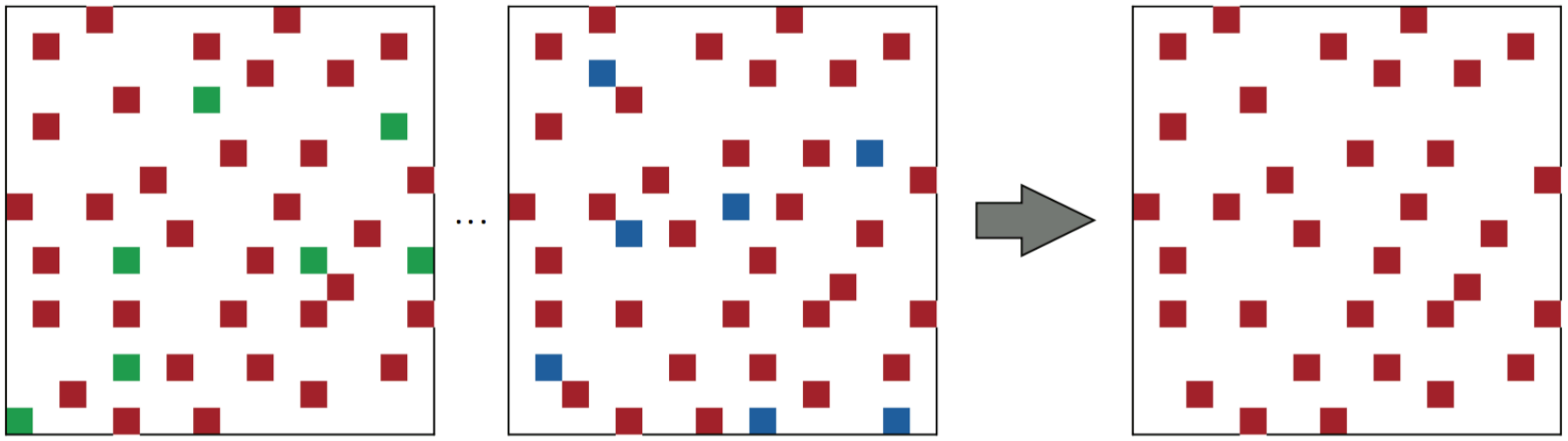
Generate Image Matrix

- Use hash function (*SimHash*) to decide X-Y coordinate and RGB colors of the pixels
- Length and width of matrix are 2^n (8)
- If hash in same X-Y coordinate, simply sum the RGB colors value

Generate Image Matrix



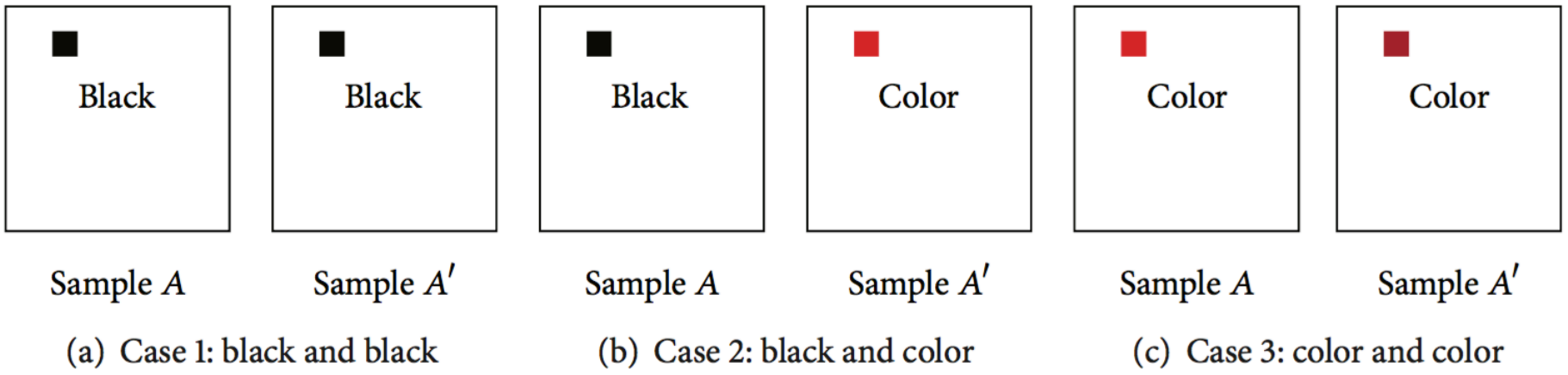
Choose Representative Image Matrix



Similarity Calculation Using Image Matrix

- Faster performance than opcode string comparison
- Finding pairs in string: $O(n^2)$
- Simhash and calculate similarity in image: $O(n)$

Similarity Calculation Using Image Matrix



(a) Case 1: black and black (b) Case 2: black and color (c) Case 3: color and color

Similarity Calculation Using Image Matrix

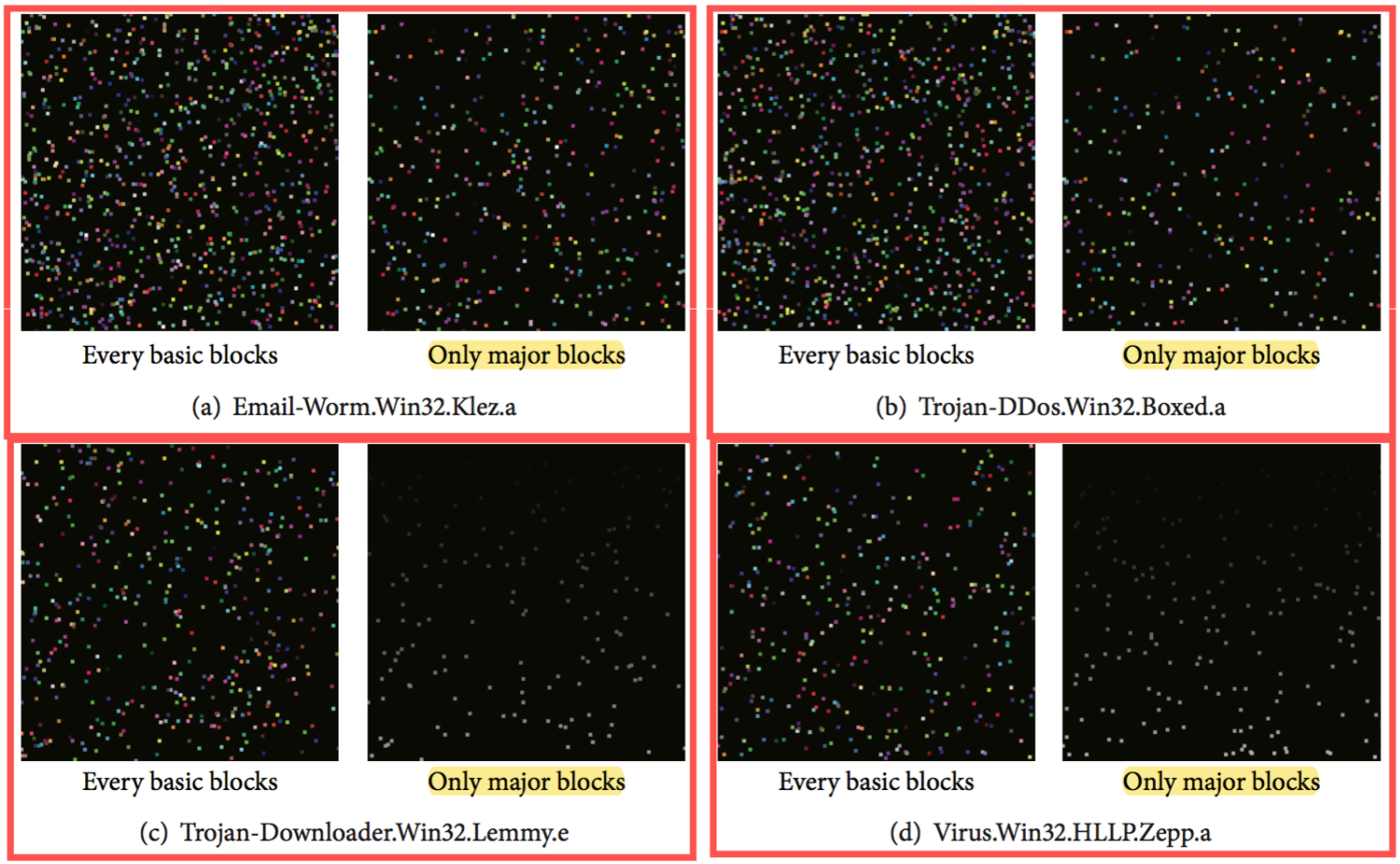
- vector angular-based distance measurement algorithm
  - Pixels are viewed as 3D vector

$$\delta\left(x_i, x_j\right) = \left[1 - \frac{2}{\pi}\cos^{-1}\left(\frac{x_i \cdot x_j}{\left|x_i\right|\left|x_j\right|}\right)\right]\left[1 - \frac{\left|x_i - x_j\right|}{\sqrt{3 \cdot 255^2}}\right]$$

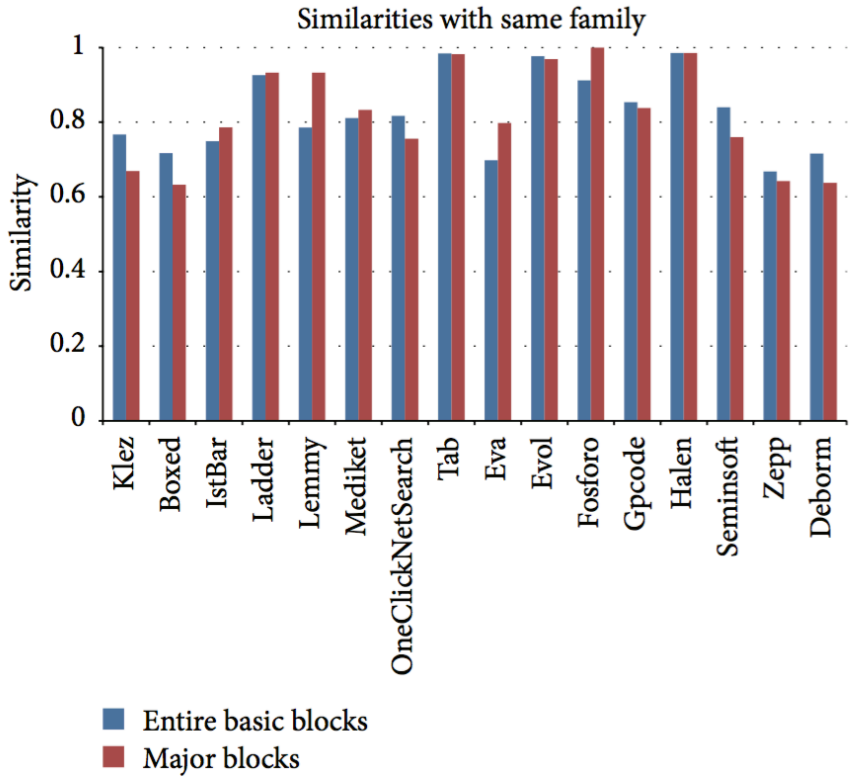
Similarity Calculation Using Image Matrix

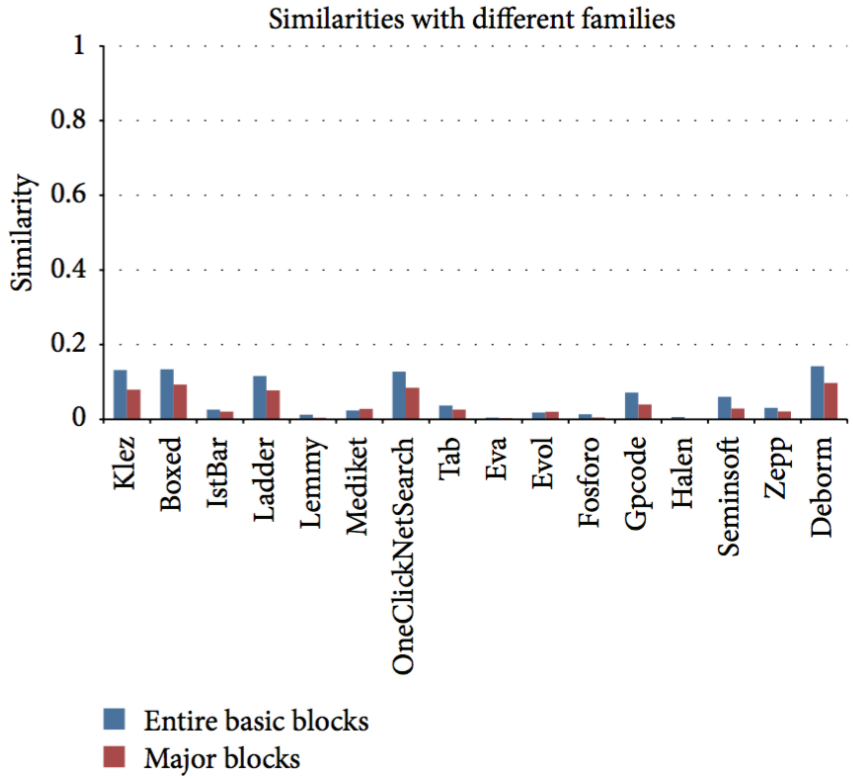
$$\text{Sim}(A, B) = \frac{\text{sum of pixel similarity values in case 3}}{\text{\# of pixels in case 2 and case 3}}.$$

Experiment: Major Blocks Selection?

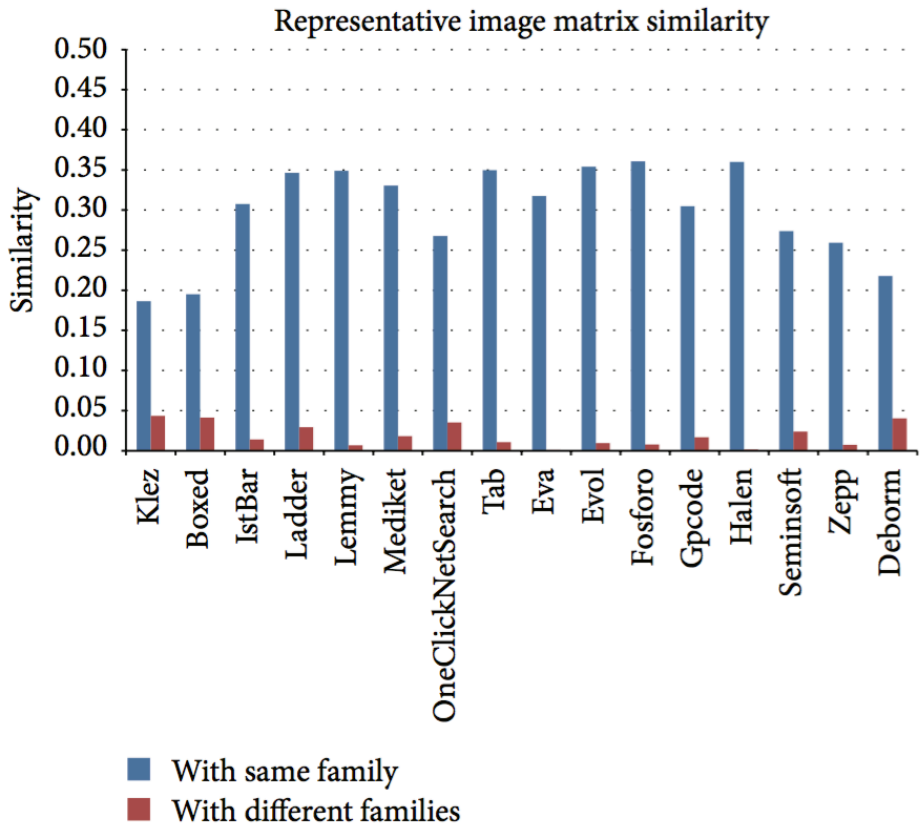


Experiment: Major Blocks Selection?





Experiment: Feasibility



Experiment: Feasibility

- Similarity of sample malwares from same family: 0.19 ~ 0.36
- Similarity of sample malwares from different family: < 0.05
- Classification accuracy = 0.9896