UNIVERSITÉ Concordia UNIVERSITY

NCFTA National Cyber-Forensics and Training Alliance CANADA

FINGERPRINTING ANDROID PACKAGING

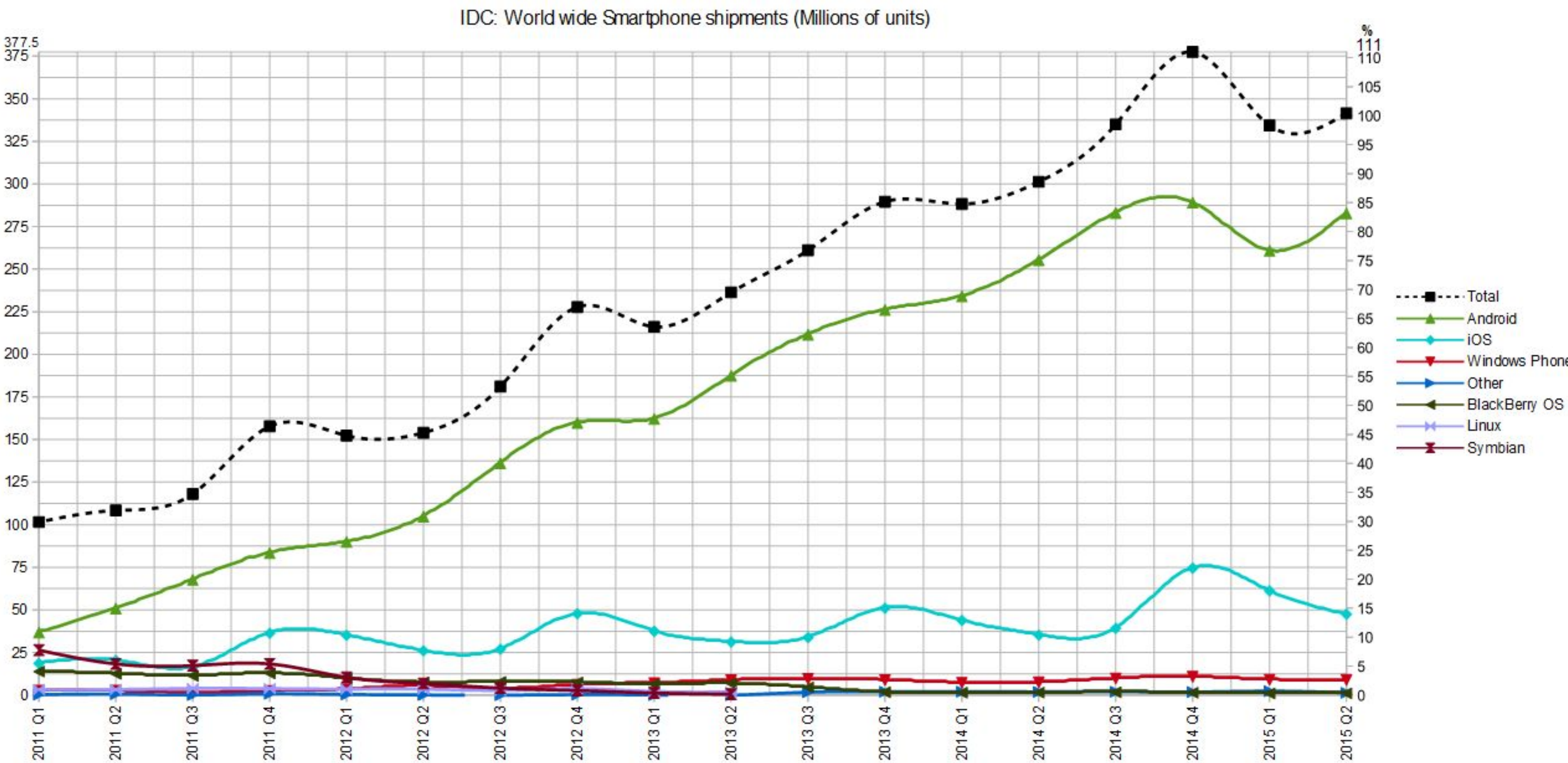
GENERATING DNA'S FOR MALWARE DETECTION

ElMouatez B. Karbab, Mourad Debbabi, Djedjiga Mouheb

August 8, 2016

DFRWS USA Conference 2016

Motivation



https://commons.wikimedia.org/wiki/File:MobileOS_market_share_till_2014_Q2.png

National Cyber-Forensics and Training Alliance CANADA

Motivation (cont'd)

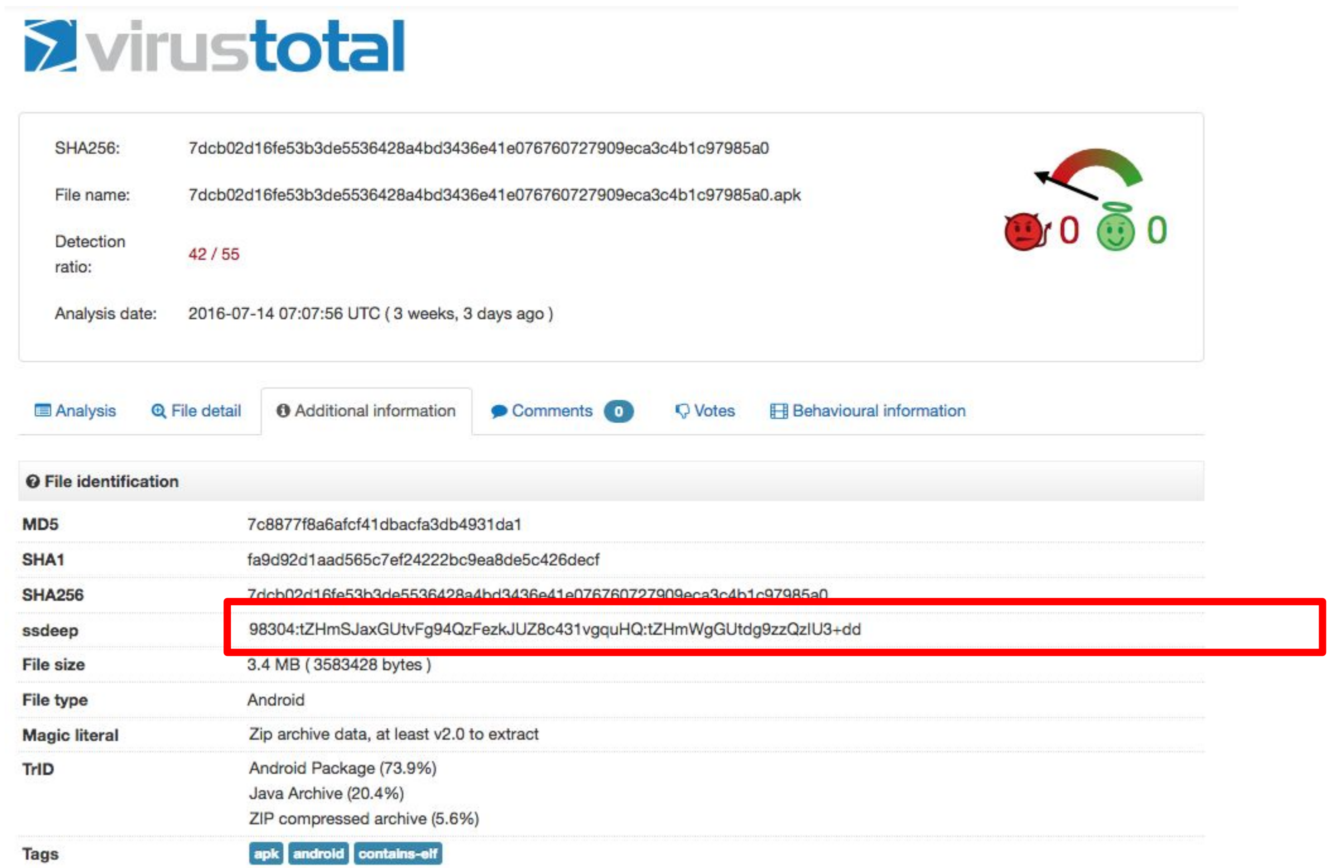
- Crypto Hash: MD5, SHA1…

2886dba4
c8c519f1
e6e44416
9580f18b

Malware Fingerprint Using Cryptographic Hashing

Motivation (cont'd)

- Fuzzy Hash: ssdeep, mv-Hash …



Malware Fingerprint Using Fyzzing Hashing

Current Fuzzy Hashing Problem

- Ignoring the underneath structure and semantics of the malicious package.
- Fuzzy hashing suffers from its single fingerprint bounded with a maximum size.
- Android Packaging Noise.

Objectives

- Develop a more accurate, yet broad, fuzzy fingerprinting technique for Android malicious apps packages.
- Design and implement a framework for Android malware detection and family attribution on top of the developed fuzzy fingerprint.

APK-DNA

- Fuzzy fingerprint covers the underneath Android app structure.

- An automatic framework for Android malware detection using APK-DNA
  - Peer-matching approach
  - Family-fingerprinting approach

Android Package (APK)

Android Package (APK)

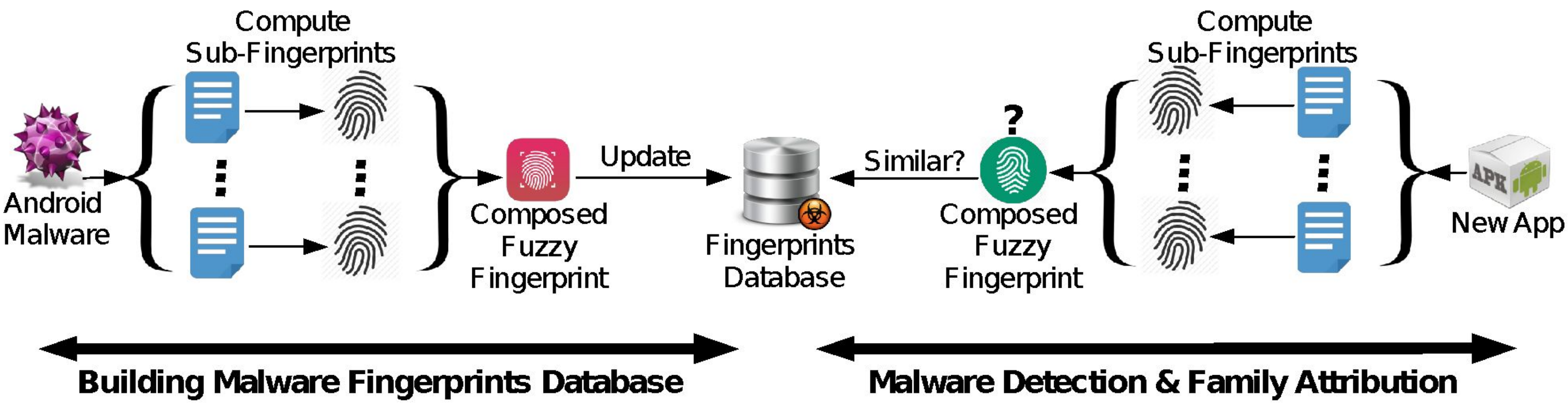
Unziping

Contents

Name	Size	Compressed
META-INF	3 Files	
ALIAS_NA.RSA	1.3 KiB	1.0 KiB
ALIAS_NA.SF	1.7 KiB	839 B
MANIFEST.MF	1.6 KiB	733 B
assets	1 File	
libparser.so	291.7 KiB	291.7 KiB
lib	1 Folder	
armeabi	3 Files	
res	13 Folders	
drawable-hdpi	1 File	
drawable-ldpi	1 File	
drawable-mdpi	1 File	
layout	1 File	
raw	1 File	
raw-de	1 File	
raw-es	1 File	
raw-fr	1 File	
raw-it	1 File	
raw-ja	1 File	
raw-zh-CN	1 File	
raw-zh-TW	1 File	
xml	1 File	
AndroidManifest.xml	4.1 KiB	1.2 KiB
classes.dex	146.5 KiB	67.8 KiB
resources.arsc	18.2 KiB	5.2 KiB

NCFTA
National Cyber-Forensics and Training Alliance CANADA

Approach

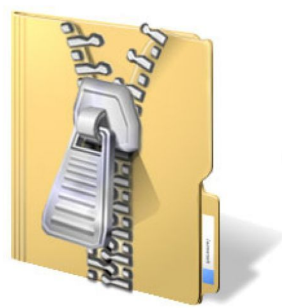


NCFTA
National Cyber-Forensics
and Training Alliance CANADA

Android Package (APK)

Android Package (APK)

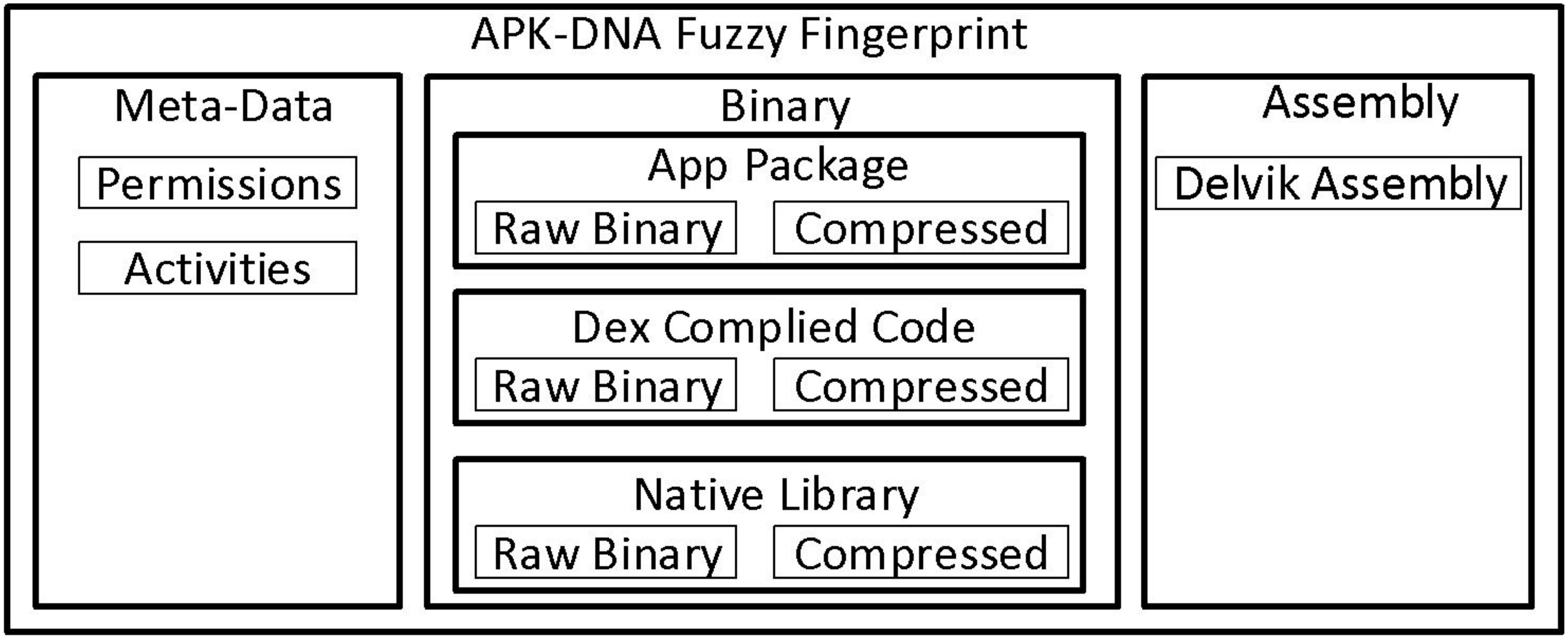
Unziping



Contents

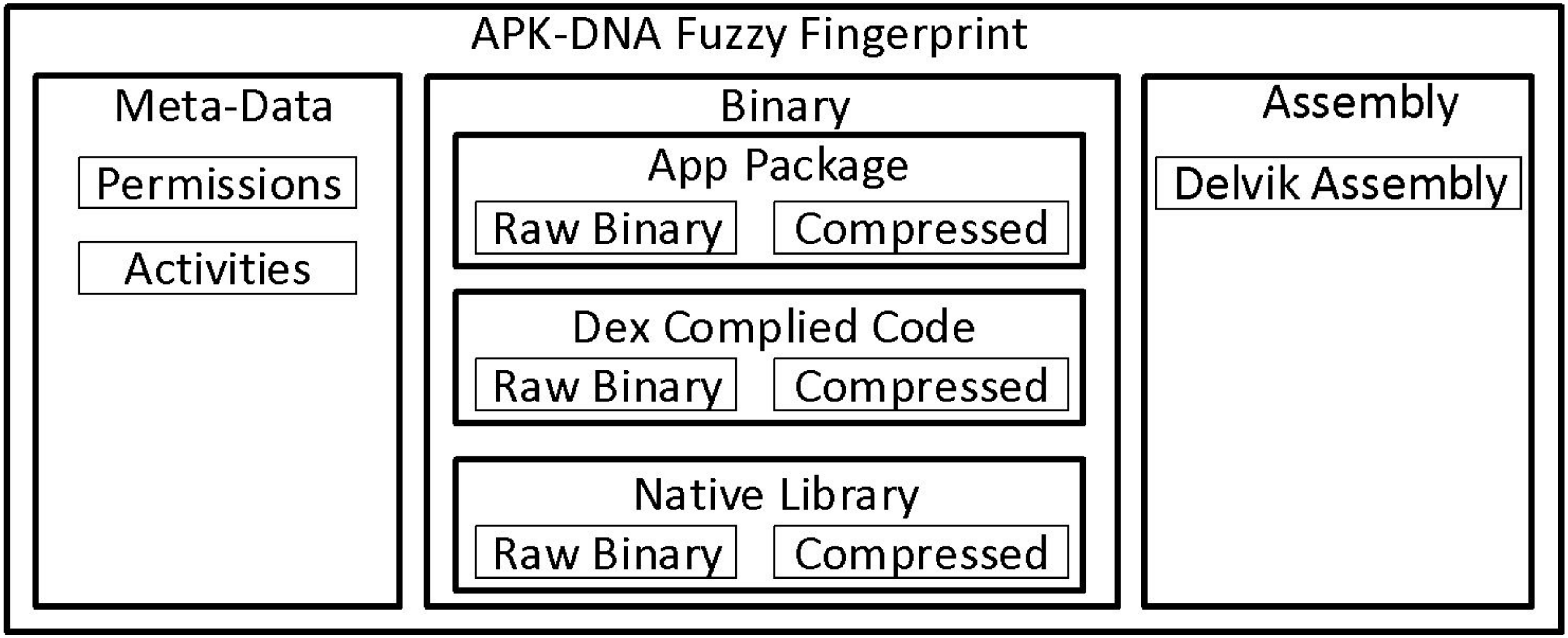
Name	Size	Compressed
META-INF	3 Files	
ALIAS_NA.RSA	1.3 KiB	1.0 KiB
ALIAS_NA.SF	1.7 KiB	839 B
MANIFEST.MF	1.6 KiB	733 B
assets	1 File	
libparser.so	291.7 KiB	291.7 KiB
lib	1 Folder	
armeabi	3 Files	
res	13 Folders	
drawable-hdpi	1 File	
drawable-ldpi	1 File	
drawable-mdpi	1 File	
layout	1 File	
raw	1 File	
raw-de	1 File	
raw-es	1 File	
raw-fr	1 File	
raw-it	1 File	
raw-ja	1 File	
raw-zh-CN	1 File	
raw-zh-TW	1 File	
xml	1 File	
AndroidManifest.xml	4.1 KiB	1.2 KiB
classes.dex	146.5 KiB	67.8 KiB
resources.arsc	18.2 KiB	5.2 KiB

APK-DNA Structure



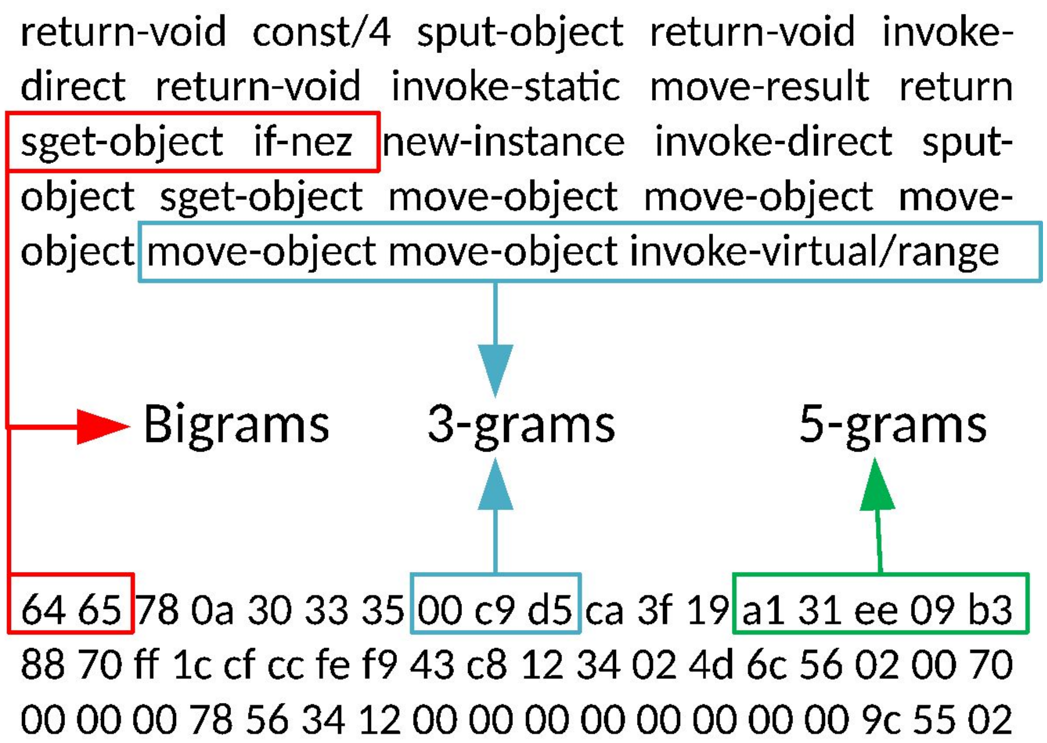
National Cyber-Forensics and Training Alliance CANADA

APK-DNA Structure



NCFTA
National Cyber-Forensics and Training Alliance CANADA

APK-DNA Structure (cont'd)



First Instructions and Bytes of AnserverBot Malware

NCFTA
National Cyber-Forensics and Training Alliance CANADA

Feature Hashing Technique

Algorithm 1: Feature Vector Computation

```
input  : N-grams: Set,
         L: Feature Vector Length
output: Binary Feature Vector
features_vector = new bitvector[L];
for Item in N-grams do
    H = hash(Item) ;
    feature_index = H mod L ;
    features_vector[feature_index] = 1 ;
end
```

Sub-Fingerprint Similarity

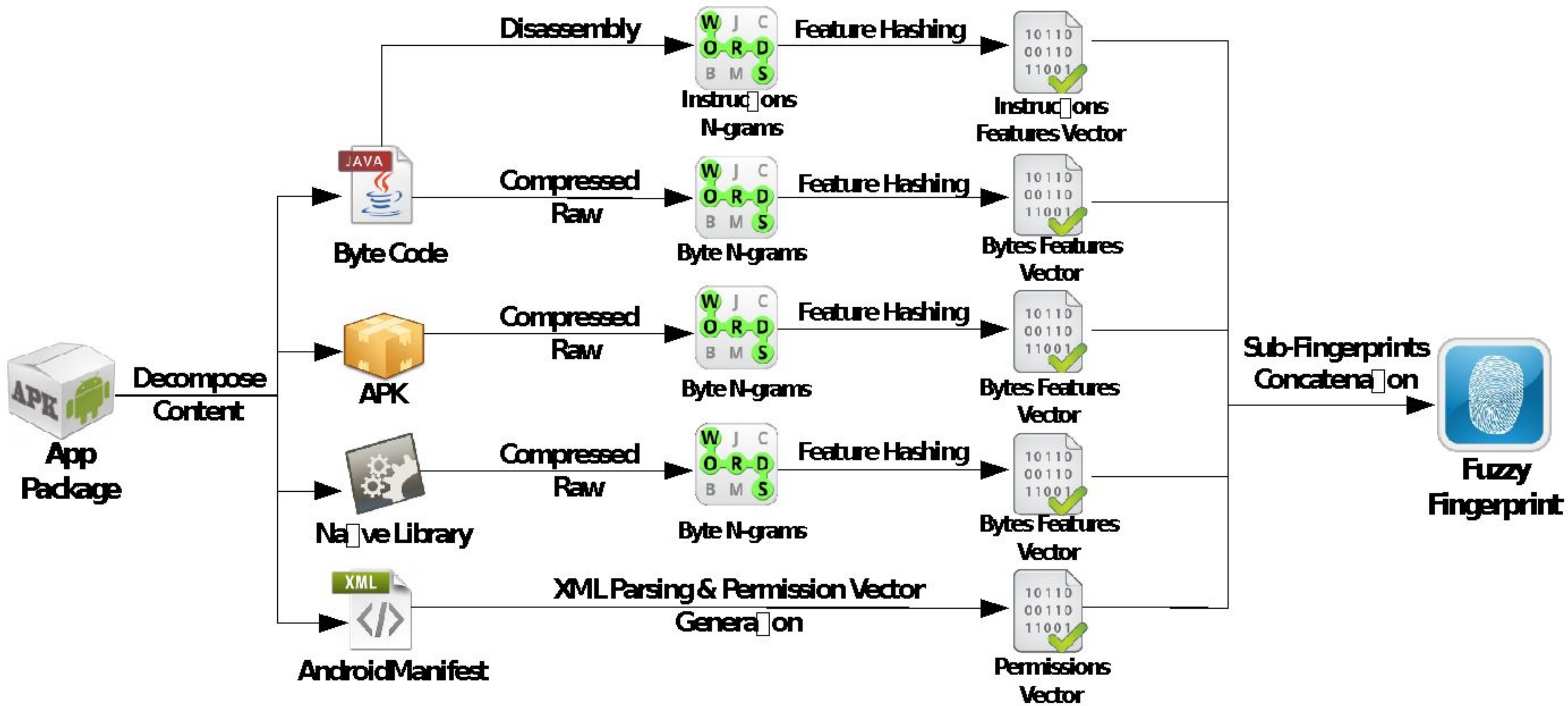
$$Jaccard(X,Y) = \frac{|X \bigcap Y|}{|X \bigcup Y|}$$
$$0 \leq Jaccard(X,Y) \leq 1$$

$$Jaccard\_bitwise(A,B) = \frac{Ones(A.B)}{Ones(A+B)}$$
$$0 \leq Jaccard\_bitwise(X,Y) \leq 1$$

A = 1011 0110 1001 0001
B = 0011 1010 1001 0000
A+B = 1011 1110 1001 0001
A.B = 0011 0010 1001 0000

Ones(A+B) = 9
Ones(A.B) = 5
Jaccard(A,B) =5 / 9

Methodology Overview



NCFTA
National Cyber-Forensics and Training Alliance CANADA

APK-DNA Fingerprint Similarity

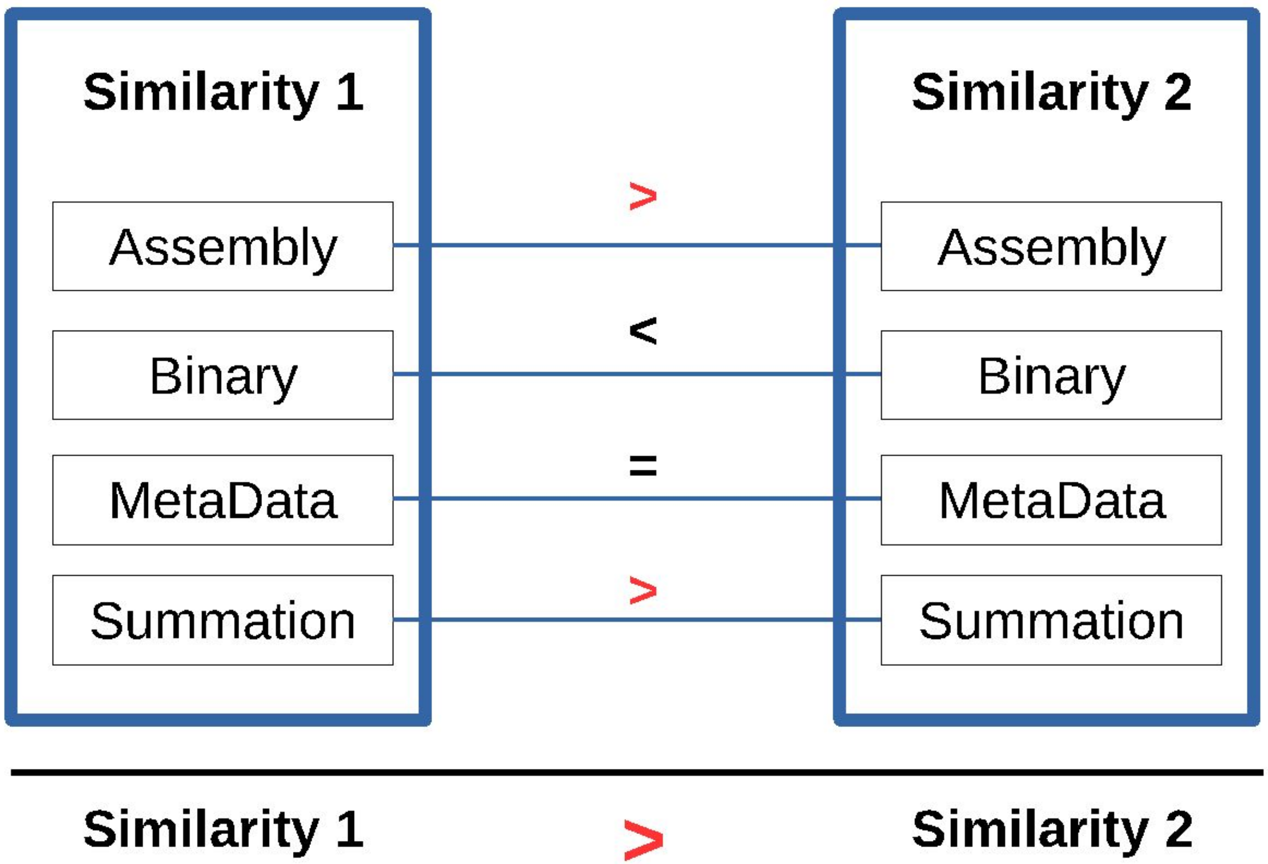
- Compare APK-DNA Similarities.

Algorithm 2: APK-DNA Similarity Computation

```
input  : APK-DNA A : list
         APK-DNA B : list
output: similarity-list: list
similarity-list = empty-list();
for content in content categories do
    similarity = Jaccard_bitwise(A[content],B[content]) ;
    similarity-list.add(similarity);
end
summation = sum(similarity-list);
similarity-list.add(summation);
```

NCFTA
National Cyber-Forensics and Training Alliance CANADA

APK-DNA Fingerprint Similarity



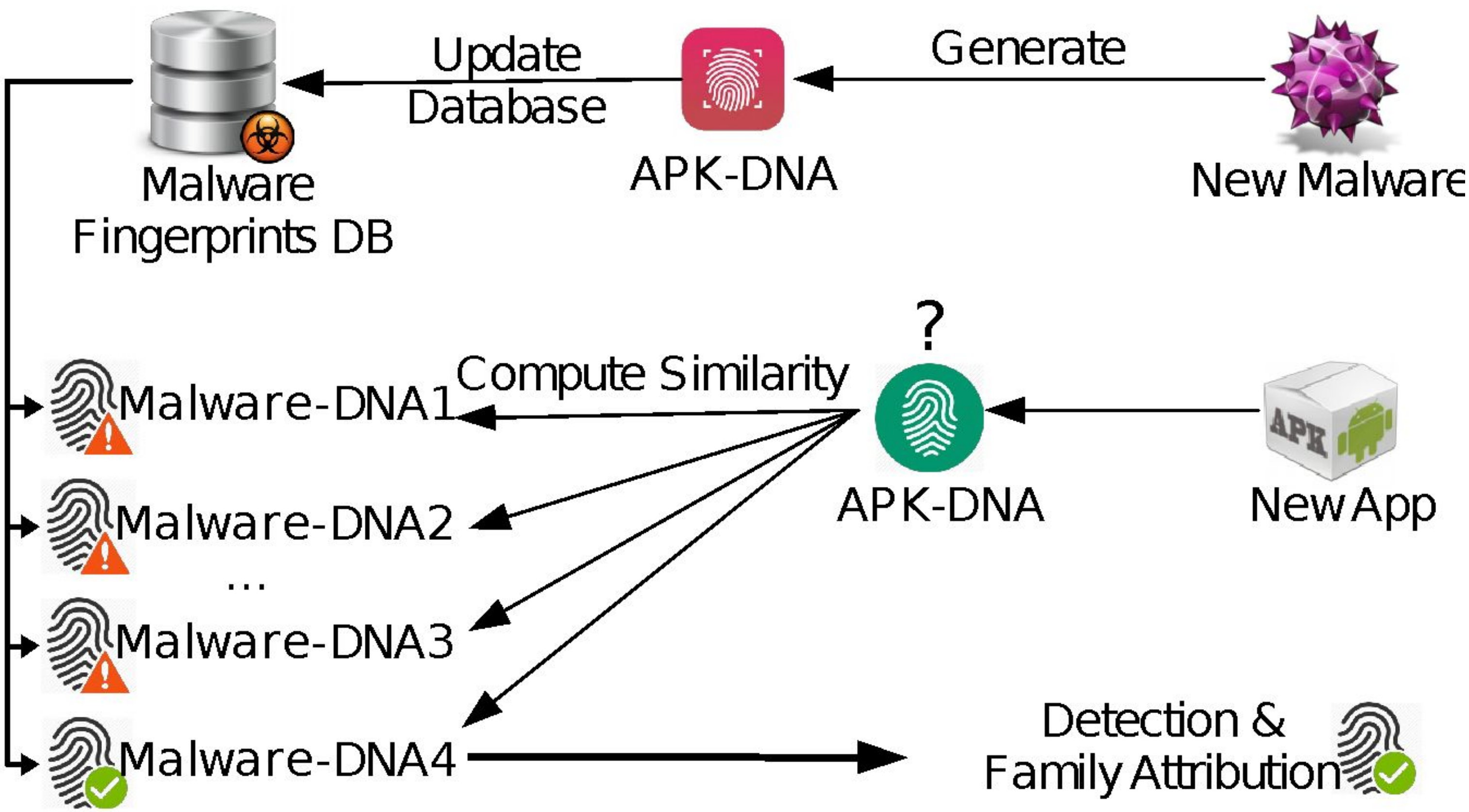
NCFTA
National Cyber-Forensics and Training Alliance CANADA

Framework Approaches

- Peer-Matching
- Family-Fingerprint

National Cyber-Forensics and Training Alliance CANADA

Peer-Matching approach



Family-Fingerprinting Approach

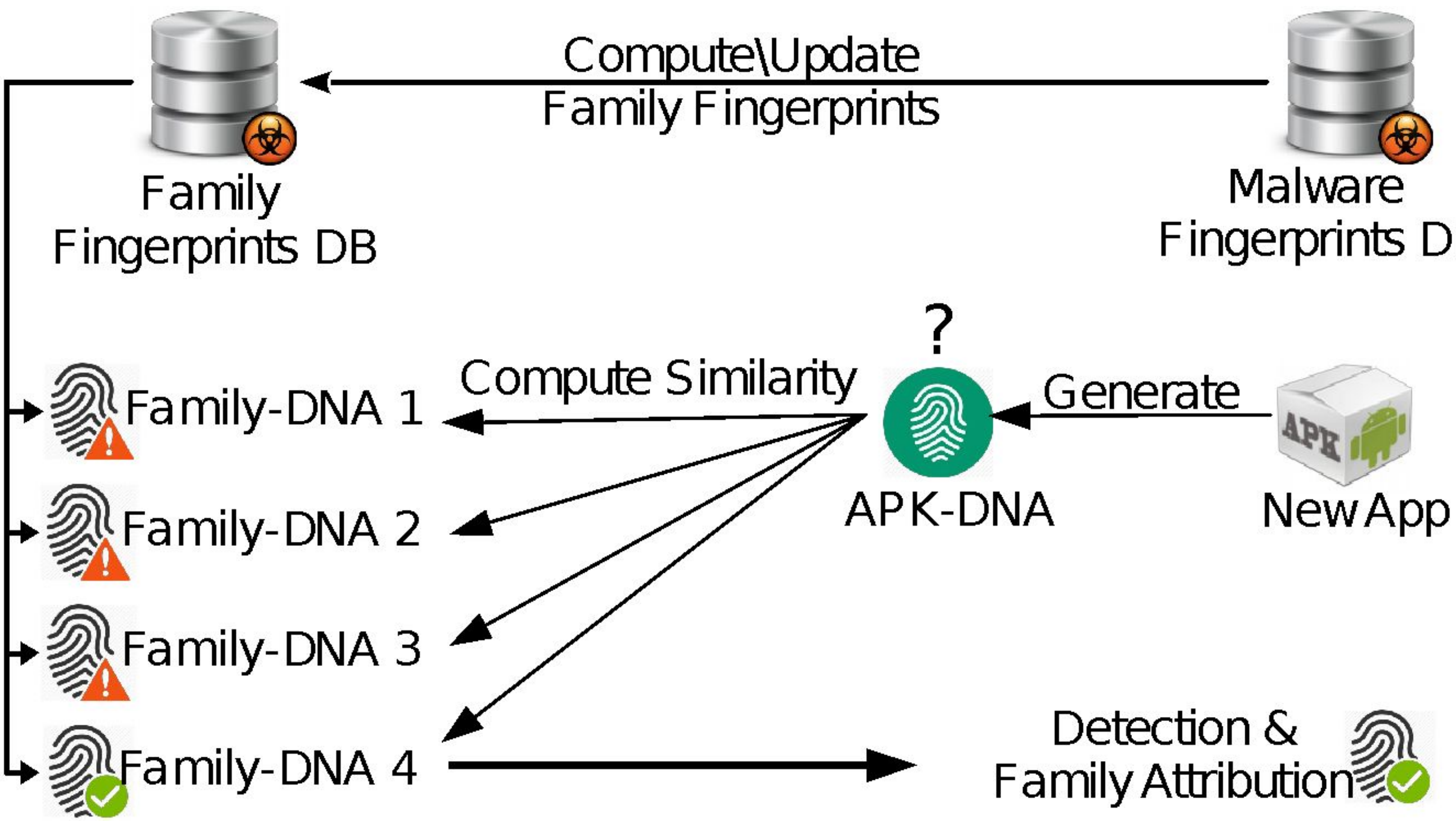
- Compute
- Overview

Family-Fingerprinting Compute

Algorithm 4: Family Fingerprint Computation

```
input  : Malware Family X Fingerprints: Set
output: Family X Fingerprint: FP_X
FP_X = new bitvector[Zeros];
for fprint in Fingerprints do
    FP_X{meta} = FP_X{meta} or fprint{meta};
    FP_X{bin} = FP_X{bin} or fprint{bin};
    FP_X{asm} = FP_X{asm} or fprint{asm};
end
```

Family-Fingerprinting Overview



NCFTA
National Cyber-Forensics and Training Alliance CANADA

Evaluation

- Android Malware Dataset.
- Evaluation Metrics.

NCFTA
National Cyber-Forensics and Training Alliance CANADA

Evaluation Dataset

- Android Malware Genome Project Dataset

#	Malware Family/Apps	Number of Samples
0	AnserverBot	187
1	KMin	52
2	DroidKungFu4	96
3	GoldDream	47
4	Geinimi	69
5	BaseBridge	122
6	DroidDreamLight	46
7	DroidKungFu3	309
8	Benign Apps	100

Evaluation Metrics

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

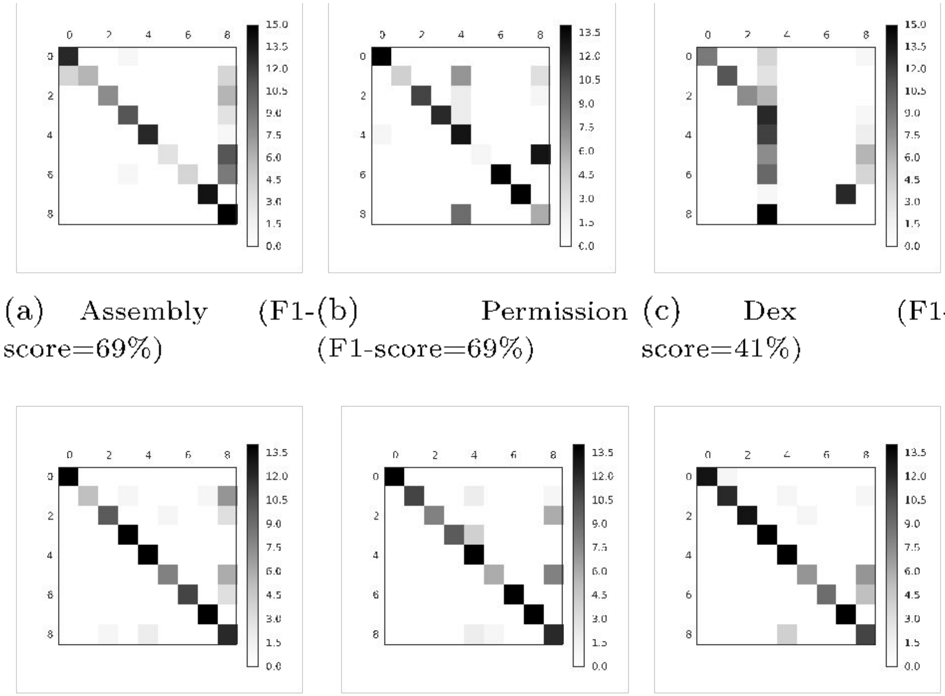
Results

- Family-Fingerprint Approach
- Peer-Matching Approach
- Voting Similarity vs Merging Similarity.

Results (Family-Fingerprint)

Fingerprint Setup	*F1-Score*	*Precision*	*Recall*
Assembly	69%	88%	68%
APK	33%	36%	32%
Permission	69%	84%	70%
Dex	41%	46%	43%
Assembly, Permission, Dex, APK	81%	88%	80%
Assembly, Permission	82%	88%	81%
Assembly, Permission, Dex	**85%**	**89%**	**84%**
Best Fingerprint Setup	**85%**	**89%**	**84%**

Table 2: Accuracy Results of the Family-Fingerprinting Approach



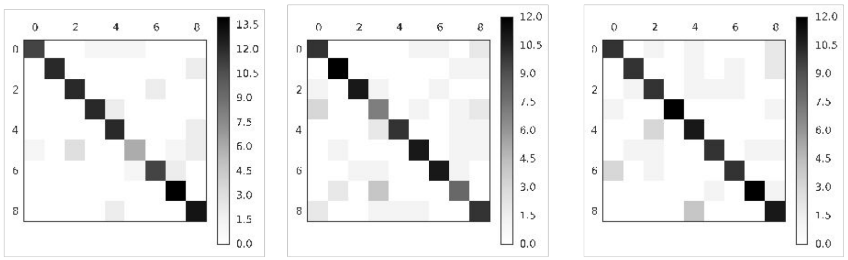
(a) Assembly (F1-score=69%)
(b) Permission (F1-score=69%)
(c) Dex (F1-score=41%)
(d) Assembly, Permission, Dex, APK (F1-score=81%)
(e) Assembly, Permission (F1-score=82%)
(f) Assembly, Permission, Dex (F1-score=85%)

Figure 8: Confusion Matrices of Family-Fingerprint Approach for each Fingerprint Setup

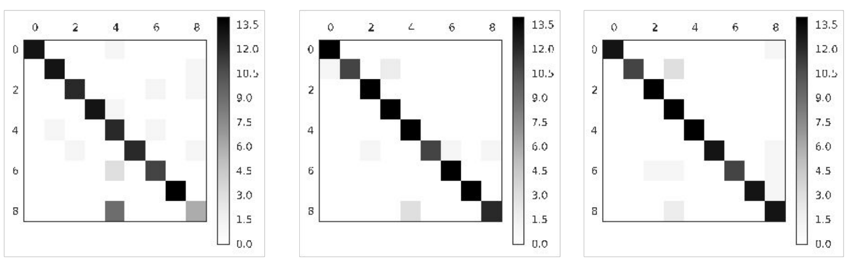
Results (Peer-Matching)

Fingerprint Setup	*F1-Score*	*Precision*	*Recall*
Assembly	91%	91%	90%
Apk	46%	48%	44%
Permission	81%	82%	80%
Dex	80%	90%	84%
Assembly, Permission, Dex, APK	84%	91%	81%
Assembly, Permission, Dex	93%	94%	93%
Assembly, Permission	**94%**	**95%**	**94%**
Best Fingerprint Setup	**94%**	**95%**	**94%**

Table 3: Accuracy Result of Peer-Matching Approach



(a) Assembly (F1-score=91%) (b) Permission (F1-score=81%) (c) Dex (F1-score=80%)



(d) Assembly, Permission, Dex, APK (F1-score=84%) (e) Assembly, Permission (F1-score=94%) (f) Assembly, Permission, Dex (F1-score=93%)

Figure 9: Confusion Matrices of Peer-Matching Approach for each Fingerprint Setup

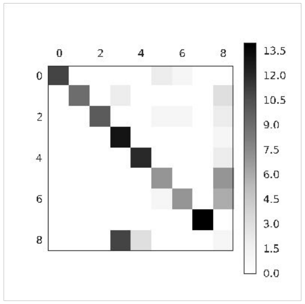
National Cyber-Forensics and Training Alliance CANADA

Results (Voting VS Merging)

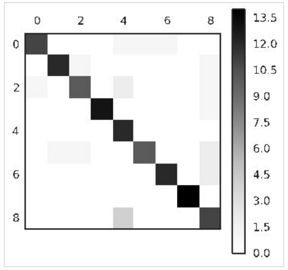
- Detection Performance using Peer-Voting Vs Merged Similarity.

Fingerprint Setup	*F1-Score*	*Precision*	*Recall*
Merged in Family-Approach	72%	84%	72%
Peer-Voting in Family-Approach	**85%**	**89%**	**84%**
Merged in Peer-Approach	86%	87%	86%
Peer-Voting in Peer-Approach	**94%**	**95%**	**94%**

Table 4: Accuracy Result of ROAR Using Merged Fingerprint



(a) Merged Fingerprint in Family-Fingerprint Approach (F1-score=72%)



(b) Merged Fingerprint in Peer-Matching Approach (F1-score=86%)

Figure 10: Confusion Matrices Of ROAR Approaches Using Merged Fingerprint

NCFTA
National Cyber-Forensics and Training Alliance CANADA

Summary

- Comprehensive fuzzy fingerprinting design for investigating Android malware variations, APK-DNA.
- Captures not only the binary of the APK file, but also both its structure and semantics.
- Framework for Android malware detection Using APK-DNA, peer-matching and family-fingerprint approaches.
- The evaluation demonstrated very promising results

Acknowledgement

- Research members of the NCFTA Canada lab headed by Prof. Mourad Debbabi
- The anonymous reviewers and Timothy Vidas for their insightful comments.
- The DFRWS USA 2016 organization comity.

NCFTA
National Cyber-Forensics and Training Alliance CANADA

Questions

- Thank you

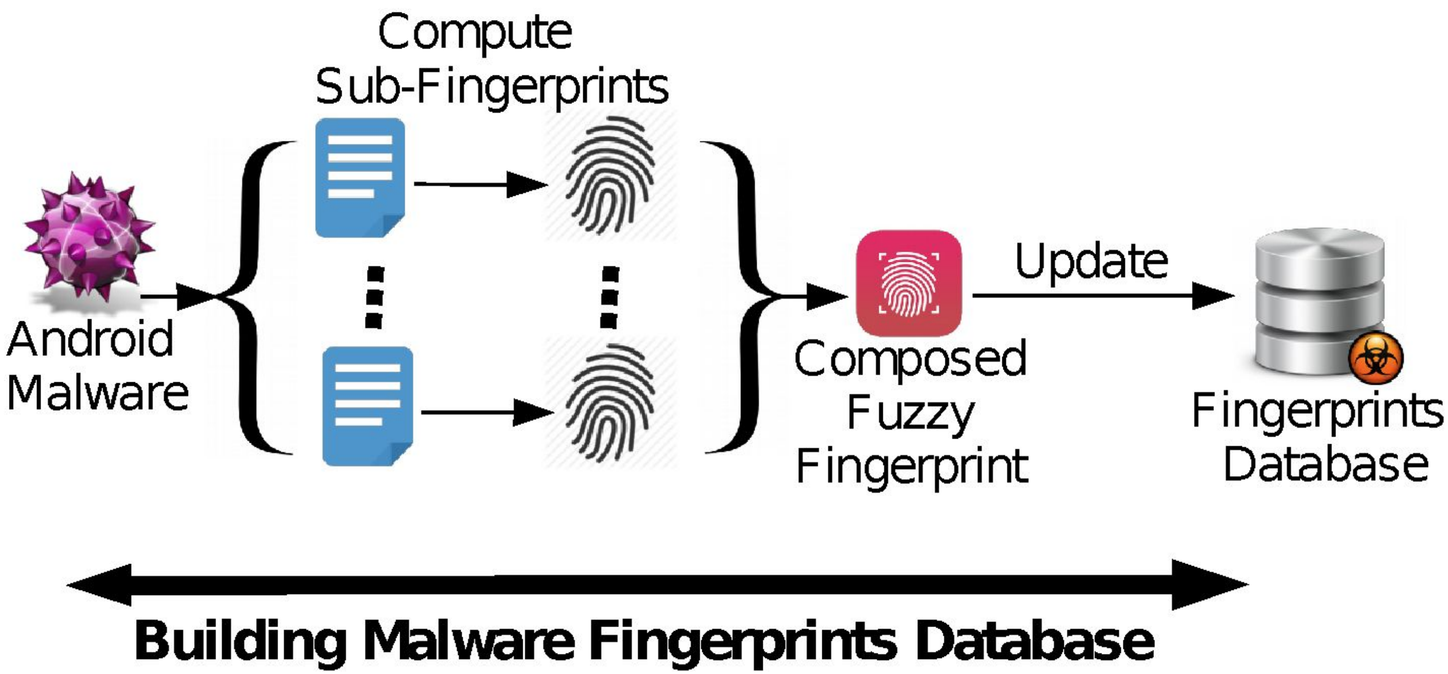
ElMouatez Billah Karbab
Ph.D. Student
Cyber Security R&D

e_karbab@encs.concordia.ca
mtz@ncfta.ca

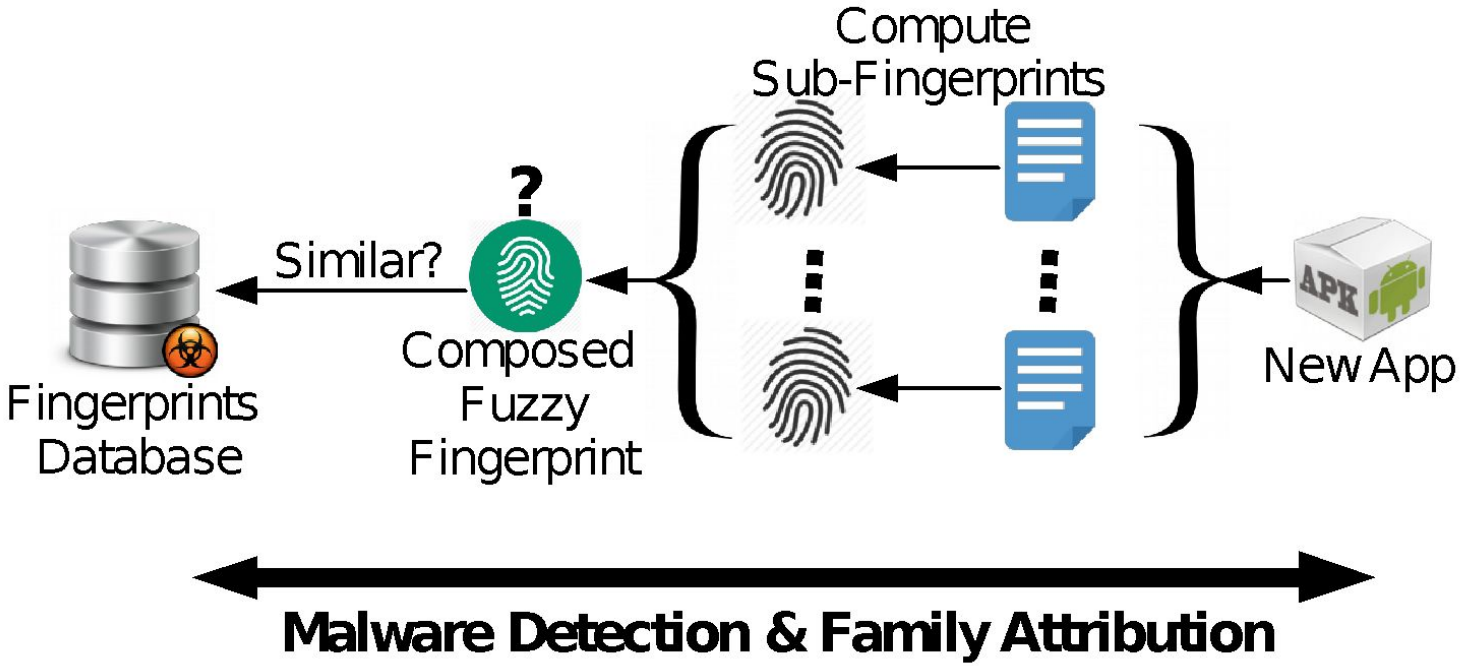
National Cyber-Forensics and Training Alliance CANADA

Appendix

Approach (cont'd)



Approach (cont'd)



NCFTA
National Cyber-Forensics
and Training Alliance CANADA

APK-DNA Fingerprint Similarity

Algorithm 3: Peer-Fingerprint Voting Mechanism

```
input  : similarity-list A-B : list
         similarity-list A-C : list
output: Decision
A-B-count = 0 ;
A-C-count = 0 ;
for content in content categories do
  | if A-B[content] >A-C[content] then
  |   | A-B-count += 1;
  |   else
  |   | A-C-count += 1;
  |   end
end
if A-B-count >A-C-count then
  | Decision = A-B;
else
  | Decision = A-C;
end
```

NCFTA
National Cyber-Forensics and Training Alliance CANADA