Experimental study of Binary Diffing Resilience on Obfuscated Programs

DIMVA 2025

Roxane Cohen <rcohen@quarkslab.com>
Quarkslab & LAMSADE, CNRS
Robin David <rdavid@quarkslab.com>
Quarkslab
Riccardo Mori <rmori@quarkslab.com>
Quarkslab
Florian Yger <florian.yger@insa-rouen.fr>
LITIS - INSA Rouen
Fabrice Rossi <fabrice.rossi@dauphine.psl.eu>
CEREMADE - PSL Dauphine-University

Binary diffing

Definition

Goal is **comparing** two *(or more)* binaries to analyze their differences. It usually done using functions with a `1-to-1` mapping computation.
(which can be problematic when functions are merged or split)

Use-cases:

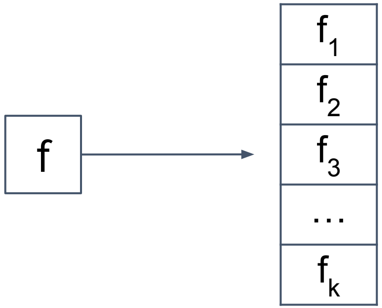
- malware diffing *(analysing updates, or common components between two variants)*
- patch analysis / 1-day analysis *(understanding if patch is correct, or what is 1-day about)*
- statically linked libraries identification *(static binary against some libs)*
- symbol porting *(e.g: IDA annotations to a new version of a binary)*

⇒ **Problematic**: Need to diff obfuscated binaries

Diffing ain’t Similarity

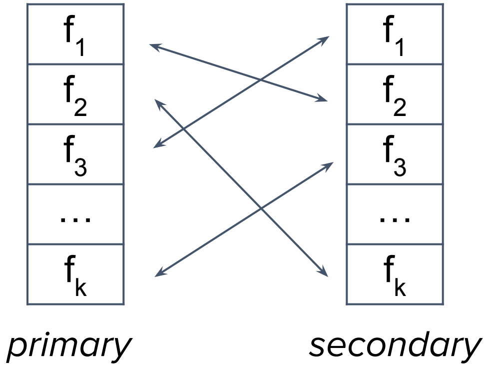
Similarity

Which function is the **most similar** to *f* among a pool of size *k* ?



Matching

What is the **best mapping** between functions of primary and secondary ?



Diffing = Similarity + Matching

(from similarity scores, create an assignment…)

Diffing **normal** binaries

Diffing

- Multiple granularity (function, basic-block, instruction) *[1, 2, 3]*
- Binary features : Call-Graph, Control-Flow Graph, assembly code *[1, 2, 3, 4]*

Similarity *(only)*

- Usually at the function level
- Well adapted in a cross-compiler, cross-architecture and cross-optimization setting *[5, 6, 7, 8]*
- Binary features : function only

[1] Dullien and al. **Structural comparison of executable objects**, 2004

[2] Dullien and al. **Graph-based comparison of executable objects**, 2005

[3] ***https://github.com/joxeankoret/diaphora***

[4] Mengin and al. **Binary Diffing as a Network Alignment Problem via Belief Propagation**, 2021.

[5] Wang and al. **jTrans: Jump-Aware Transformer for Binary Code Similarity.** 2022

[6] Li and al.**Graph Matching Networks for Learning the Similarity of Graph Structured Objects.** 2019

[7] Marcelli and al. **How Machine Learning Is Solving the Binary Function Similarity Problem.** 2022

[8] He and al. **Code is not Natural Language: Unlock the Power of Semantics-Oriented Graph Representation for Binary Code Similarity Detection.** 2024

Diffing **obfuscated** binaries

State-of-the-Art

Diffing

- Semantic features (symbolic) adapted for matching different granularities (basic-block or path) *[1, 2]*
- Obfuscation techniques that adversarially disturbs differs *[3]*

Similarity *(only)*

- Small experiments on OLLVM-only obfuscated binaries *[4, 5]*
- Limited set on obfuscations / obfuscation types

[1] Luo and al. **Semantics-based obfuscation-resilient binary code similarity comparison with applications to software plagiarism detection.** 2014

[2] Gao and al. **Binhunt: Automatically finding semantic differences in binary programs**. 2008

[3] Zhang and al. **Khaos: The Impact of Inter-procedural Code Obfuscation on Binary Diffing Techniques**, 2023

[4] Kim and al. **Revisiting Binary Code Similarity Analysis using Interpretable Feature Engineering and Lessons**, 2022

[5] Ding and al. **Asm2vec: Boosting static representation robustness for binary clone search against code obfuscation and compiler optimization**, 2019

Why diffing obfuscated binaries ?

Using multiple binary variants to infer knowledge between binaries

- An attacker obtains a **“plain”** binary and an **“obfuscated”** newer variant
- An attacker gets its hands on **two obfuscated variants** *(of the same program)*

Core concept:

- Idea: Multiple binary variants can help to **draw correlations** between program content
- Advantage: Comparing binaries **without** having to deobfuscate them.
- Why: weaken the obfuscation security*

ApkDiff: Matching Android App Versions Based on Class Structure, De Ghein and al., 2022

**cannot compute the same property before and after the obfuscation is applied*

Problematics & Contributions

Current limitations

- Standard differs are not suited for obfuscated binaries
- No satisfactory dataset (not enough data, code snippet, only OLLVM…)
- Limited work on diffing in an obfuscated setting

Contributions

- Creating a realistic and large obfuscated dataset
- Comparing differs ability to recover correspondence between obfuscated binaries in two settings : **plain-vs-obfuscated** and **obfuscated-vs-obfuscated**
- Evaluating an obfuscation / obfuscator robustness according to its ability to prevent computing the correspondence between obfuscated binaries

Our Dataset: ObfuBench

	Passes	Pass type	zlib	lz4	minilua	sqlite	freetype
Tigress	Copy	Inter	✔	✔	✔	✔	✔
	Split	Inter	✔	✔	✔	✔	✔
	Merge	Inter	✔	✔	✗	✗	~
	CFF	Intra	✔	✔	✔	✔	✔
	Virtualize	Intra	✔	✔	~	~	✗
	Opaque	Intra	✔	✔	✔	✗	~
	EncodeArithmetic (Enc.A)	Data	✔	✔	✔	✔	✔
	EncodeLiterals (Enc.L)	Data	✔	✔	✔	✔	✔
	Mix	Intra & Data	✔	✔	✔	~	~
	Mix + Split	All	✔	✔	✔	~	~
OLLVM-14	CFF	Intra	✔	✔	✔	✔	✔
	Opaque	Intra	✔	✔	✔	✔	✔
	EncodeArithmetic (Enc.A)	Data	✔	✔	✔	✔	✔
	Mix	Intra & Data	✔	✔	✔	✔	✔

Projects strongly limited by Tigress ability to obfuscate whole projects *(its file merging is limited)*

⇒ Dataset available at: *https://github.com/quarkslab/diffing_obfuscation_dataset*

Evaluating diffing and similarity tools

Differs

Use standard binary differs:

- BinDiff *[1, 2]*
- Diaphora *[3]*
- QBinDiff *[4]*

Similarity tools

Use state-of-the-art similarity approaches

- *Asm2vec [5]*
- *JTrans [6]*
- *GMN [7]*

⇒ Combined with Hungarian algorithm (**optimal** but **n^3**)

[1] Dullien and al. **Structural comparison of executable objects**, 2004

[2] Dullien and al. **Graph-based comparison of executable objects**, 2005

[3] ***https://github.com/joxeankoret/diaphora***

[4] Mengin and al. **Binary Diffing as a Network Alignment Problem via Belief Propagation**, 2021.

[5] Ding and al. **Asm2Vec: Boosting Static Representation Robustness for Binary Clone Search against Code Obfuscation and Compiler Optimization.** 2019

[6] Wang and al. **jTrans: Jump-Aware Transformer for Binary Code Similarity.** 2022

[7] Li and al.**Graph Matching Networks for Learning the Similarity of Graph Structured Objects.** 2019

Diffing Evaluation

Comparing the *Ground-Truth* functions pairs and the differ's functions pairs ?

True Positives
good match correctly identified

False Positives
wrong match identified

True Negative
Not a match considered as-is

False Negative
Good match **not** identified

Precision =

Recall =

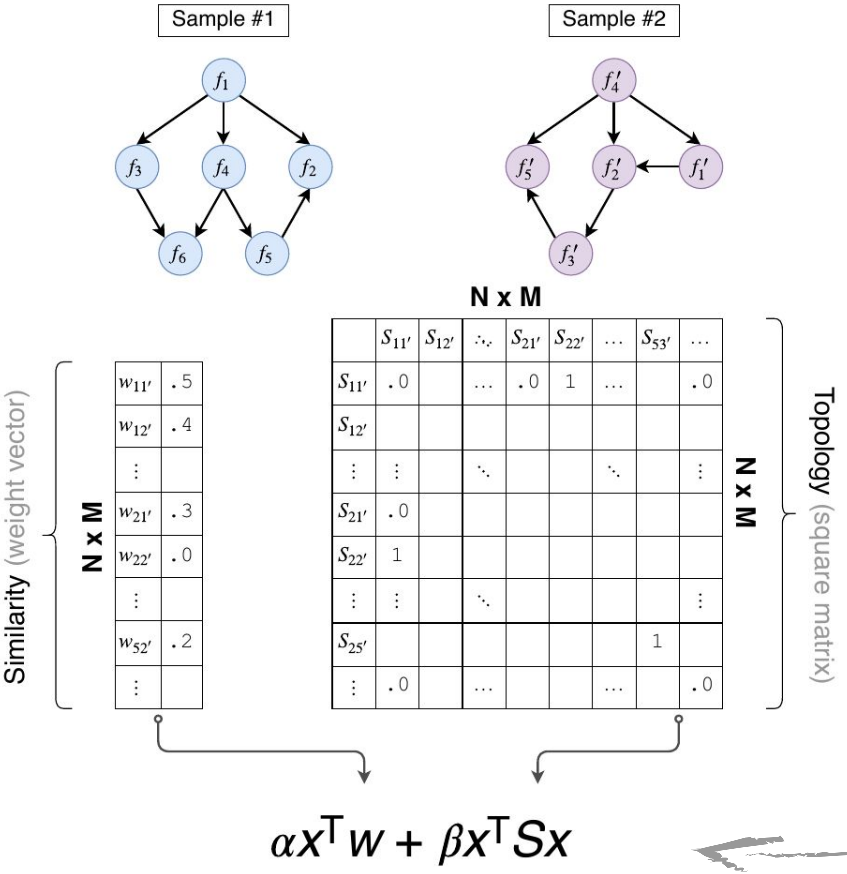
⇒

F1-score = $2 \times \frac{P \times R}{P + R}$

QBinDiff: A Modular differ

Available features

- CG features *(caller, callees...)*
- CFG features *(#bb, #edges, #loops, cyclomatic complexity...)*
- Assembly features *(grouped mnemonics, Pcode mnemonics...)*



Goal

Solve an instance of the ***Network Alignment Problem***

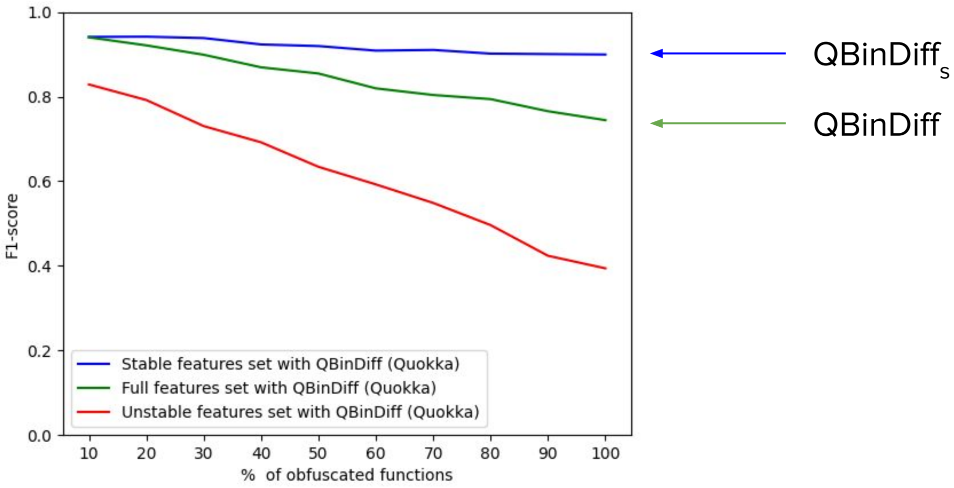
Arbitrate between **function similarity** and **call-graph topology** to be more resilient if one of them is altered *(+ still use imported functions as anchors)*

Resilient features against an obfuscation

	BBlockNb	SCComponents	BytesHash	Cyclomatic Complexity	MDIndex	JumpNb	SmallPrimeNumbers	MaxParentNb	MaxChildNb	MaxInsNb	MeanInsNb	InsNb	GraphMeanDegree	GraphDensity	GraphNbComponents	Graph Diameter	GraphTransitivity	GraphCommunities	Address	DatName	FuncName	ChildNb	ParentNb	RelativeNb	LibName	ImpName	Constant	StrRefs	MnemonicSimple	MnemonicTyped	GroupsCategory	ReadWriteAccess	
Merge Split	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓
Copy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓
Intra	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Data	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗

Stable or unstable QBinDiff features for different obfuscation passes

Feature impact on diffing



QBinDiff feature impact : stable, full and unstable features

(Control-Flow Graph Flattening f1-score evolution)

Characterize the obfuscation ⇒ adapt the features for better diffing results

Diffing: plain-vs-obfuscated *(OLLVM)*

- f1-score comparison
- ObfuBench dataset *(stripped binaries)*
- the higher, the better
- Columns:
  - General: all functions together
  - Obfuscated: solely obfuscated functions

OLLVM-14 `plain-obfuscated`		General f1-score				Obfuscated f1-score			
		Mix	*CFF*	*Opaque*	*Enc.A*	*Mix*	*CFF*	*Opaque*	*Enc.A*
10%	Bindiff	0.81	0.78	0.78	0.79	**0.72**	0.71	0.69	0.75
	Diaphora3	0.79	0.78	0.78	0.80	0.45	0.59	0.61	0.78
	GMN	0.66	0.64	0.65	0.69	0.24	0.27	0.36	0.52
	Asm2vec	0.56	0.53	0.56	0.59	0.32	0.40	0.44	0.61
	JTrans	**0.85**	0.85	0.85	**0.86**	0.67	**0.81**	0.79	**0.82**
	QBinDiff	0.84	0.85	0.85	**0.86**	0.58	0.78	0.76	0.81
	QBinDiff_s	-	**0.86**	**0.86**	0.81	-	0.80	**0.81**	0.77
50%	Bindiff	0.72	0.76	0.74	0.79	**0.62**	0.67	0.65	0.73
	Diaphora3	0.64	0.70	0.71	0.79	0.40	0.57	0.59	0.78
	GMN	0.44	0.47	0.50	0.62	0.20	0.25	0.31	0.49
	Asm2vec	0.32	0.37	0.44	0.59	0.23	0.32	0.40	0.61
	JTrans	0.70	0.82	0.80	**0.86**	0.57	0.77	0.73	**0.81**
	QBinDiff	**0.74**	0.84	0.82	**0.86**	0.60	0.76	0.73	0.80
	QBinDiff_s	-	**0.85**	**0.85**	0.80	-	**0.79**	**0.79**	0.73
100%	Bindiff	0.53	0.65	0.65	0.78	0.47	0.56	0.57	0.72
	Diaphora3	0.40	0.50	0.61	0.79	0.35	0.49	0.56	0.77
	GMN	0.23	0.26	0.33	0.53	0.18	0.22	0.28	0.48
	Asm2vec	0.15	0.17	0.32	0.59	0.17	0.21	0.35	0.60
	JTrans	0.53	0.71	0.73	**0.86**	0.50	0.70	0.70	**0.81**
	QBinDiff	**0.65**	0.74	0.80	0.85	**0.59**	0.69	0.71	0.80
	QBinDiff_s	-	**0.77**	**0.84**	0.79	-	**0.73**	**0.78**	0.73

Diffing: plain-vs-obfuscated *(OLLVM)*

OLLVM scores are high, no matter the differ, the type or level of obfuscation

OLLVM-14 `plain-obfuscated`		General f1-score				Obfuscated f1-score			
		Mix	*CFF*	*Opaque*	*Enc.A*	*Mix*	*CFF*	*Opaque*	*Enc.A*
10%	Bindiff	0.81	0.78	0.78	0.79	**0.72**	0.71	0.69	0.75
	Diaphora3	0.79	0.78	0.78	0.80	0.45	0.59	0.61	0.78
	GMN	0.66	0.64	0.65	0.69	0.24	0.27	0.36	0.52
	Asm2vec	0.56	0.53	0.56	0.59	0.32	0.40	0.44	0.61
	JTrans	**0.85**	0.85	0.85	**0.86**	0.67	**0.81**	0.79	**0.82**
	QBinDiff	0.84	0.85	0.85	**0.86**	0.58	0.78	0.76	0.81
	$QBinDiff_s$	-	**0.86**	**0.86**	0.81	-	0.80	**0.81**	0.77
50%	Bindiff	0.72	0.76	0.74	0.79	**0.62**	0.67	0.65	0.73
	Diaphora3	0.64	0.70	0.71	0.79	0.40	0.57	0.59	0.78
	GMN	0.44	0.47	0.50	0.62	0.20	0.25	0.31	0.49
	Asm2vec	0.32	0.37	0.44	0.59	0.23	0.32	0.40	0.61
	JTrans	0.70	0.82	0.80	**0.86**	0.57	0.77	0.73	**0.81**
	QBinDiff	**0.74**	0.84	0.82	**0.86**	0.60	0.76	0.73	0.80
	$QBinDiff_s$	-	**0.85**	**0.85**	0.80	-	**0.79**	**0.79**	0.73
100%	Bindiff	0.53	0.65	0.65	0.78	0.47	0.56	0.57	0.72
	Diaphora3	0.40	0.50	0.61	0.79	0.35	0.49	0.56	0.77
	GMN	0.23	0.26	0.33	0.53	0.18	0.22	0.28	0.48
	Asm2vec	0.15	0.17	0.32	0.59	0.17	0.21	0.35	0.60
	JTrans	0.53	0.71	0.73	**0.86**	0.50	0.70	0.70	**0.81**
	QBinDiff	**0.65**	0.74	0.80	0.85	**0.59**	0.69	0.71	0.80
	$QBinDiff_s$	-	**0.77**	**0.84**	0.79	-	**0.73**	**0.78**	0.73

Diffing: plain-vs-obfuscated *(OLLVM)*

OLLVM-14 plain-obfuscated		General f1-score				Obfuscated f1-score			
		Mix	*CFF*	*Opaque*	*Enc.A*	*Mix*	*CFF*	*Opaque*	*Enc.A*
10%	Bindiff	0.81	0.78	0.78	0.79	**0.72**	0.71	0.69	0.75
	Diaphora3	0.79	0.78	0.78	0.80	0.45	0.59	0.61	0.78
	GMN	0.66	0.64	0.65	0.69	0.24	0.27	0.36	0.52
	Asm2vec	0.56	0.53	0.56	0.59	0.32	0.40	0.44	0.61
	JTrans	**0.85**	0.85	0.85	**0.86**	0.67	**0.81**	0.79	**0.82**
	QBinDiff	0.84	0.85	0.85	**0.86**	0.58	0.78	0.76	0.81
	$QBinDiff_s$	-	**0.86**	**0.86**	0.81	-	0.80	**0.81**	0.77
50%	Bindiff	0.72	0.76	0.74	0.79	**0.62**	0.67	0.65	0.73
	Diaphora3	0.64	0.70	0.71	0.79	0.40	0.57	0.59	0.78
	GMN	0.44	0.47	0.50	0.62	0.20	0.25	0.31	0.49
	Asm2vec	0.32	0.37	0.44	0.59	0.23	0.32	0.40	0.61
	JTrans	0.70	0.82	0.80	**0.86**	0.57	0.77	0.73	**0.81**
	QBinDiff	**0.74**	0.84	0.82	**0.86**	0.60	0.76	0.73	0.80
	$QBinDiff_s$	-	**0.85**	**0.85**	0.80	-	**0.79**	**0.79**	0.73
100%	Bindiff	0.53	0.65	0.65	0.78	0.47	0.56	0.57	0.72
	Diaphora3	0.40	0.50	0.61	0.79	0.35	0.49	0.56	0.77
	GMN	0.23	0.26	0.33	0.53	0.18	0.22	0.28	0.48
	Asm2vec	0.15	0.17	0.32	0.59	0.17	0.21	0.35	0.60
	JTrans	0.53	0.71	0.73	**0.86**	0.50	0.70	0.70	**0.81**
	QBinDiff	**0.65**	0.74	0.80	0.85	**0.59**	0.69	0.71	0.80
	$QBinDiff_s$	-	**0.77**	**0.84**	0.79	-	**0.73**	**0.78**	0.73

The obfuscation level deteriorates only slightly the scores

Diffing: plain-vs-obfuscated *(OLLVM)*

BinDiff, JTrans and QBinDiff are the best “adversaries”

OLLVM-14 plain-obfuscated		General f1-score				Obfuscated f1-score			
		Mix	*CFF*	*Opaque*	*Enc.A*	*Mix*	*CFF*	*Opaque*	*Enc.A*
10%	Bindiff	0.81	0.78	0.78	0.79	**0.72**	0.71	0.69	0.75
	Diaphora3	0.79	0.78	0.78	0.80	0.45	0.59	0.61	0.78
	GMN	0.66	0.64	0.65	0.69	0.24	0.27	0.36	0.52
	Asm2vec	0.56	0.53	0.56	0.59	0.32	0.40	0.44	0.61
	JTrans	**0.85**	0.85	0.85	**0.86**	0.67	**0.81**	0.79	**0.82**
	QBinDiff	0.84	0.85	0.85	**0.86**	0.58	0.78	0.76	0.81
	$QBinDiff_s$	-	**0.86**	**0.86**	0.81	-	0.80	**0.81**	0.77
50%	Bindiff	0.72	0.76	0.74	0.79	**0.62**	0.67	0.65	0.73
	Diaphora3	0.64	0.70	0.71	0.79	0.40	0.57	0.59	0.78
	GMN	0.44	0.47	0.50	0.62	0.20	0.25	0.31	0.49
	Asm2vec	0.32	0.37	0.44	0.59	0.23	0.32	0.40	0.61
	JTrans	0.70	0.82	0.80	**0.86**	0.57	0.77	0.73	**0.81**
	QBinDiff	**0.74**	0.84	0.82	**0.86**	0.60	0.76	0.73	0.80
	$QBinDiff_s$	-	**0.85**	**0.85**	0.80	-	**0.79**	**0.79**	0.73
100%	Bindiff	0.53	0.65	0.65	0.78	0.47	0.56	0.57	0.72
	Diaphora3	0.40	0.50	0.61	0.79	0.35	0.49	0.56	0.77
	GMN	0.23	0.26	0.33	0.53	0.18	0.22	0.28	0.48
	Asm2vec	0.15	0.17	0.32	0.59	0.17	0.21	0.35	0.60
	JTrans	0.53	0.71	0.73	**0.86**	0.50	0.70	0.70	**0.81**
	QBinDiff	**0.65**	0.74	0.80	0.85	**0.59**	0.69	0.71	0.80
	$QBinDiff_s$	-	**0.77**	**0.84**	0.79	-	**0.73**	**0.78**	0.73

Diffing: plain-vs-obfuscated *(OLLVM)*

QBinDiff_s > QBinDiff

OLLVM-14 `plain-obfuscated`		General f1-score				Obfuscated f1-score			
		Mix	*CFF*	*Opaque*	*Enc.A*	*Mix*	*CFF*	*Opaque*	*Enc.A*
10%	Bindiff	0.81	0.78	0.78	0.79	**0.72**	0.71	0.69	0.75
	Diaphora3	0.79	0.78	0.78	0.80	0.45	0.59	0.61	0.78
	GMN	0.66	0.64	0.65	0.69	0.24	0.27	0.36	0.52
	Asm2vec	0.56	0.53	0.56	0.59	0.32	0.40	0.44	0.61
	JTrans	**0.85**	0.85	0.85	**0.86**	0.67	**0.81**	0.79	**0.82**
	QBinDiff	0.84	0.85	0.85	**0.86**	0.58	0.78	0.76	0.81
	QBinDiff_s	-	**0.86**	**0.86**	0.81	-	0.80	**0.81**	0.77
50%	Bindiff	0.72	0.76	0.74	0.79	**0.62**	0.67	0.65	0.73
	Diaphora3	0.64	0.70	0.71	0.79	0.40	0.57	0.59	0.78
	GMN	0.44	0.47	0.50	0.62	0.20	0.25	0.31	0.49
	Asm2vec	0.32	0.37	0.44	0.59	0.23	0.32	0.40	0.61
	JTrans	0.70	0.82	0.80	**0.86**	0.57	0.77	0.73	**0.81**
	QBinDiff	**0.74**	0.84	0.82	**0.86**	0.60	0.76	0.73	0.80
	QBinDiff_s	-	**0.85**	**0.85**	0.80	-	**0.79**	**0.79**	0.73
100%	Bindiff	0.53	0.65	0.65	0.78	0.47	0.56	0.57	0.72
	Diaphora3	0.40	0.50	0.61	0.79	0.35	0.49	0.56	0.77
	GMN	0.23	0.26	0.33	0.53	0.18	0.22	0.28	0.48
	Asm2vec	0.15	0.17	0.32	0.59	0.17	0.21	0.35	0.60
	JTrans	0.53	0.71	0.73	**0.86**	0.50	0.70	0.70	**0.81**
	QBinDiff	**0.65**	0.74	0.80	0.85	**0.59**	0.69	0.71	0.80
	QBinDiff_s	-	**0.77**	**0.84**	0.79	-	**0.73**	**0.78**	0.73

Diffing: plain-vs-obfuscated *(OLLVM)*

OLLVM-14 plain-obfuscated		General f1-score				Obfuscated f1-score			
		Mix	CFF	Opaque	Enc.A	Mix	CFF	Opaque	Enc.A
10%	Bindiff	0.81	0.78	0.78	0.79	**0.72**	0.71	0.69	0.75
	Diaphora3	0.79	0.78	0.78	0.80	0.45	0.59	0.61	0.78
	GMN	0.66	0.64	0.65	0.69	0.24	0.27	0.36	0.52
	Asm2vec	0.56	0.53	0.56	0.59	0.32	0.40	0.44	0.61
	JTrans	**0.85**	0.85	0.85	**0.86**	0.67	**0.81**	0.79	**0.82**
	QBinDiff	0.84	0.85	0.85	**0.86**	0.58	0.78	0.76	0.81
	QBinDiff$_s$	-	**0.86**	**0.86**	0.81	-	0.80	**0.81**	0.77
50%	Bindiff	0.72	0.76	0.74	0.79	**0.62**	0.67	0.65	0.73
	Diaphora3	0.64	0.70	0.71	0.79	0.40	0.57	0.59	0.78
	GMN	0.44	0.47	0.50	0.62	0.20	0.25	0.31	0.49
	Asm2vec	0.32	0.37	0.44	0.59	0.23	0.32	0.40	0.61
	JTrans	0.70	0.82	0.80	**0.86**	0.57	0.77	0.73	**0.81**
	QBinDiff	**0.74**	0.84	0.82	**0.86**	0.60	0.76	0.73	0.80
	QBinDiff$_s$	-	**0.85**	**0.85**	0.80	-	**0.79**	**0.79**	0.73
100%	Bindiff	0.53	0.65	0.65	0.78	0.47	0.56	0.57	0.72
	Diaphora3	0.40	0.50	0.61	0.79	0.35	0.49	0.56	0.77
	GMN	0.23	0.26	0.33	0.53	0.18	0.22	0.28	0.48
	Asm2vec	0.15	0.17	0.32	0.59	0.17	0.21	0.35	0.60
	JTrans	0.53	0.71	0.73	**0.86**	0.50	0.70	0.70	**0.81**
	QBinDiff	**0.65**	0.74	0.80	0.85	**0.59**	0.69	0.71	0.80
	QBinDiff$_s$	-	**0.77**	**0.84**	0.79	-	**0.73**	**0.78**	0.73

Asm2vec and GMN binary similarity tools *(+ matching)* show disappointing performances

Diffing: plain-vs-obfuscated *(OLLVM)*

Slight difference between general f1-score & obfuscated f1-score, depending on the tool used and the obfuscation

OLLVM-14	plain-obfuscated	General f1-score				Obfuscated f1-score			
		Mix	CFF	Opaque	Enc.A	Mix	CFF	Opaque	Enc.A
10%	Bindiff	0.81	0.78	0.78	0.79	**0.72**	0.71	0.69	0.75
	Diaphora3	0.79	0.78	0.78	0.80	0.45	0.59	0.61	0.78
	GMN	0.66	0.64	0.65	0.69	0.24	0.27	0.36	0.52
	Asm2vec	0.56	0.53	0.56	0.59	0.32	0.40	0.44	0.61
	JTrans	**0.85**	0.85	0.85	**0.86**	0.67	**0.81**	0.79	**0.82**
	QBinDiff	0.84	0.85	0.85	**0.86**	0.58	0.78	0.76	0.81
	$QBinDiff_s$	-	**0.86**	**0.86**	0.81	-	0.80	**0.81**	0.77
50%	Bindiff	0.72	0.76	0.74	0.79	**0.62**	0.67	0.65	0.73
	Diaphora3	0.64	0.70	0.71	0.79	0.40	0.57	0.59	0.78
	GMN	0.44	0.47	0.50	0.62	0.20	0.25	0.31	0.49
	Asm2vec	0.32	0.37	0.44	0.59	0.23	0.32	0.40	0.61
	JTrans	0.70	0.82	0.80	**0.86**	0.57	0.77	0.73	**0.81**
	QBinDiff	**0.74**	0.84	0.82	**0.86**	0.60	0.76	0.73	0.80
	$QBinDiff_s$	-	**0.85**	**0.85**	0.80	-	**0.79**	**0.79**	0.73
100%	Bindiff	0.53	0.65	0.65	0.78	0.47	0.56	0.57	0.72
	Diaphora3	0.40	0.50	0.61	0.79	0.35	0.49	0.56	0.77
	GMN	0.23	0.26	0.33	0.53	0.18	0.22	0.28	0.48
	Asm2vec	0.15	0.17	0.32	0.59	0.17	0.21	0.35	0.60
	JTrans	0.53	0.71	0.73	**0.86**	0.50	0.70	0.70	**0.81**
	QBinDiff	**0.65**	0.74	0.80	0.85	**0.59**	0.69	0.71	0.80
	$QBinDiff_s$	-	**0.77**	**0.84**	0.79	-	**0.73**	**0.78**	0.73

Diffing: plain-vs-obfuscated *(Tigress)*

Tigress `plain-obfuscated`		General f1-score										Obfuscated f1-score									
		Mix	Mix + Split	Copy	Merge	Split	CFF	Virtualize	Opaque	Enc.A	Enc.L	Mix	Mix + Split	Copy	Merge	Split	CFF	Virtualize	Opaque	Enc.A	Enc.L
10%	Bindiff	0.80	0.72	0.80	0.76	0.74	0.82	0.81	0.79	0.84	0.84	**0.69**	0.02	0.21	0.56	0.09	0.75	0.72	0.69	0.86	0.81
	Diaphora3	0.75	0.67	0.76	0.73	0.70	0.76	0.74	0.75	0.78	0.79	0.34	0.02	0.46	0.34	0.08	0.52	0.04	0.66	0.77	0.78
	GMN	0.51	0.46	0.57	0.52	0.47	0.53	0.48	0.53	0.54	0.59	0.04	0.01	0.34	0.15	0.02	0.08	0.01	0.25	0.16	0.49
	Asm2vec	0.49	0.42	0.51	0.55	0.46	0.49	0.45	0.51	0.56	0.57	0.15	0.01	0.28	0.35	0.03	0.20	0.03	0.45	0.52	0.58
	JTrans	0.80	0.74	0.82	0.78	0.76	0.84	0.80	0.81	0.85	0.85	0.55	0.02	**0.54**	0.35	0.04	0.72	0.36	**0.74**	0.87	0.88
	QBinDiff	**0.84**	**0.77**	**0.86**	0.82	0.81	0.87	0.83	**0.84**	**0.89**	**0.89**	0.55	**0.05**	0.25	0.59	0.19	0.73	0.35	0.69	**0.92**	**0.91**
	$QBinDiff_s$	-	-	0.85	**0.84**	**0.82**	**0.89**	**0.89**	0.79	0.83	0.84	-	-	0.25	**0.60**	**0.22**	**0.85**	**0.79**	0.65	0.90	0.86
50%	Bindiff	0.63	0.38	0.65	0.63	0.45	0.73	0.66	0.65	0.81	0.83	**0.52**	0.02	0.19	0.43	0.07	0.67	0.60	0.57	0.83	0.79
	Diaphora3	0.57	0.33	0.69	0.59	0.43	0.63	0.46	0.69	0.73	0.78	0.28	0.01	0.48	0.29	0.08	0.46	0.01	0.64	0.74	0.82
	GMN	0.30	0.19	0.49	0.35	0.26	0.32	0.27	0.38	0.38	0.58	0.02	0.01	0.36	0.13	0.02	0.06	0.00	0.20	0.13	0.50
	Asm2vec	0.27	0.16	0.41	0.40	0.26	0.30	0.18	0.40	0.49	0.59	0.10	0.00	0.29	0.28	0.04	0.14	0.01	0.39	0.50	0.64
	JTrans	0.64	0.41	0.71	0.56	0.47	0.76	0.52	**0.74**	0.63	0.85	0.46	0.01	**0.54**	0.16	0.03	0.69	0.23	**0.71**	0.67	**0.90**
	QBinDiff	**0.68**	**0.45**	**0.75**	0.70	0.56	0.79	0.68	**0.74**	**0.87**	**0.89**	0.49	**0.03**	0.26	0.51	0.17	0.72	0.53	0.66	**0.90**	**0.90**
	$QBinDiff_s$	-	-	0.74	**0.73**	**0.58**	**0.87**	**0.81**	0.68	0.82	0.83	-	-	0.26	**0.58**	**0.20**	**0.84**	**0.77**	0.59	0.88	0.87
100%	Bindiff	0.33	0.10	0.48	0.44	0.22	0.60	0.51	0.47	0.77	0.80	0.23	0.01	0.20	0.28	0.06	0.56	0.48	0.41	0.80	0.68
	Diaphora3	0.27	0.09	0.64	0.38	0.25	0.46	0.10	0.61	0.66	0.76	0.19	0.01	0.50	0.28	0.07	0.43	0.01	0.61	0.71	0.78
	GMN	0.10	0.05	0.42	0.23	0.13	0.12	0.11	0.24	0.25	0.57	0.01	0.01	0.35	0.12	0.02	0.05	0.00	0.19	0.12	0.49
	Asm2vec	0.08	0.04	0.29	0.29	0.13	0.11	0.02	0.32	0.43	0.56	0.08	0.00	0.24	0.32	0.03	0.11	0.00	0.37	0.48	0.65
	JTrans	**0.46**	**0.21**	0.62	0.32	0.28	0.68	0.20	**0.66**	0.60	0.83	**0.43**	0.01	**0.54**	0.14	0.03	0.68	0.16	**0.68**	0.66	**0.89**
	QBinDiff	0.40	0.19	**0.65**	0.57	0.36	0.71	0.49	0.63	**0.85**	**0.87**	0.33	**0.02**	0.26	0.48	0.15	0.70	0.46	0.61	**0.89**	0.87
	$QBinDiff_s$	-	-	0.64	**0.61**	**0.39**	**0.84**	**0.72**	0.56	0.81	0.82	-	-	0.27	**0.55**	**0.18**	**0.83**	**0.72**	0.53	0.86	0.84

Diffing: plain-vs-obfuscated *(Tigress)*

Tigress `plain-obfuscated`		General f1-score										Obfuscated f1-score									
		Mix	Mix + Split	Copy	Merge	Split	CFF	Virtualize	Opaque	Enc.A	Enc.L	Mix	Mix + Split	Copy	Merge	Split	CFF	Virtualize	Opaque	Enc.A	Enc.L
10%	Bindiff	0.80	0.72	0.80	0.76	0.74	0.82	0.81	0.79	0.84	0.84	**0.69**	0.02	0.21	0.56	0.09	0.75	0.72	0.69	0.86	0.81
	Diaphora3	0.75	0.67	0.76	0.73	0.70	0.76	0.74	0.75	0.78	0.79	0.34	0.02	0.46	0.34	0.08	0.52	0.04	0.66	0.77	0.78
	GMN	0.51	0.46	0.57	0.52	0.47	0.53	0.48	0.53	0.54	0.59	0.04	0.01	0.34	0.15	0.02	0.08	0.01	0.25	0.16	0.49
	Asm2vec	0.49	0.43	0.51	0.55	0.46	0.49	0.45	0.51	0.56	0.57	0.15	0.01	0.28	0.35	0.03	0.20	0.03	0.45	0.52	0.58
	JTrans	0.80	0.74	0.82	0.78	0.76	0.84	0.80	0.81	0.85	0.85	0.55	0.02	**0.54**	0.35	0.04	0.72	0.36	**0.74**	0.87	0.88
	QBinDiff	**0.84**	**0.77**	**0.86**	0.82	0.81	0.87	0.83	**0.84**	**0.89**	**0.89**	0.55	**0.05**	0.25	0.59	0.19	0.73	0.35	0.69	**0.92**	**0.91**
	$QBinDiff_s$	-	-	0.85	**0.84**	**0.82**	**0.89**	**0.89**	0.79	0.83	0.84	-	-	0.25	**0.60**	**0.22**	**0.85**	**0.79**	0.65	0.90	0.86
50%	Bindiff	0.63	0.38	0.65	0.63	0.45	0.73	0.66	0.65	0.81	0.83	**0.52**	0.02	0.19	0.43	0.07	0.67	0.60	0.57	0.83	0.79
	Diaphora3	0.57	0.33	0.69	0.59	0.43	0.63	0.46	0.69	0.73	0.78	0.28	0.01	0.48	0.29	0.08	0.46	0.01	0.64	0.74	0.82
	GMN	0.30	0.19	0.49	0.35	0.26	0.32	0.27	0.38	0.38	0.58	0.02	0.01	0.36	0.13	0.02	0.06	0.00	0.20	0.13	0.50
	Asm2vec	0.27	0.16	0.41	0.40	0.26	0.30	0.18	0.40	0.49	0.59	0.10	0.00	0.29	0.28	0.04	0.14	0.01	0.39	0.50	0.64
	JTrans	0.64	0.41	0.71	0.56	0.47	0.76	0.52	**0.74**	0.63	0.85	0.46	0.01	**0.54**	0.16	0.03	0.69	0.23	**0.71**	0.67	**0.90**
	QBinDiff	**0.68**	**0.45**	**0.75**	0.70	0.56	0.79	0.68	**0.74**	**0.87**	**0.89**	0.49	**0.03**	0.26	0.51	0.17	0.72	0.53	0.66	**0.90**	**0.90**
	$QBinDiff_s$	-	-	0.74	**0.73**	**0.58**	**0.87**	**0.81**	0.68	0.82	0.83	-	-	0.26	**0.58**	**0.20**	**0.84**	**0.77**	0.59	0.88	0.87
100%	Bindiff	0.33	0.10	0.48	0.44	0.22	0.60	0.51	0.47	0.77	0.80	0.23	0.01	0.20	0.28	0.06	0.56	0.48	0.41	0.80	0.68
	Diaphora3	0.27	0.09	0.64	0.38	0.25	0.46	0.10	0.61	0.66	0.76	0.19	0.01	0.50	0.28	0.07	0.43	0.01	0.61	0.71	0.78
	GMN	0.10	0.05	0.42	0.23	0.13	0.12	0.11	0.24	0.25	0.57	0.01	0.01	0.35	0.12	0.02	0.05	0.00	0.19	0.12	0.49
	Asm2vec	0.08	0.04	0.29	0.29	0.13	0.11	0.02	0.32	0.43	0.56	0.08	0.00	0.24	0.32	0.03	0.11	0.00	0.37	0.48	0.65
	JTrans	**0.46**	**0.21**	0.62	0.32	0.28	0.68	0.20	**0.66**	0.60	0.83	**0.43**	0.01	**0.54**	0.14	0.03	0.68	0.16	**0.68**	0.66	**0.89**
	QBinDiff	0.40	0.19	**0.65**	0.57	0.36	0.71	0.49	0.63	**0.85**	**0.87**	0.33	**0.02**	0.26	0.48	0.15	0.70	0.46	0.61	**0.89**	0.87
	$QBinDiff_s$	-	-	0.64	**0.61**	**0.39**	**0.84**	**0.72**	0.56	0.81	0.82	-	-	0.27	**0.55**	**0.18**	**0.83**	**0.72**	0.53	0.86	0.84

Tigress associated f1-score are significantly lower than OLLVM, especially for inter-procedural obfuscation

Results **BinKit** dataset

Binkit	Plain-obfuscated					Obfuscated-obfuscated				
	bool	cpio	cflow	ccd2cue	a2ps	bool	cpio	cflow	ccd2cue	a2ps
BinDiff	0.9	0.63	0.78	0.94	0.7	0.8	0.42	0.61	0.84	0.44
Diaphora3	0.66	0.6	0.71	0.71	0.63	0.57	0.45	0.4	0.57	0.39
GMN	0.41	0.39	0.30	0.53	0.22	0.40	0.39	0.31	0.53	0.23
Asm2vec	0.37	0.29	0.22	0.55	0.15	0.34	0.25	0.19	0.38	0.13
JTrans	0.86	0.80	0.84	0.90	0.69	0.70	0.55	0.55	0.66	0.42
QBinDiff	0.96	0.92	0.91	0.98	0.82	0.9	0.82	0.82	**0.91**	0.7
$QBinDiff_s$	**0.97**	**0.94**	**0.93**	**0.99**	**0.87**	**0.92**	**0.86**	**0.86**	**0.91**	**0.80**

Same trend than the previous ObfuBench experiment, even more pronounced

Real-World example : XTunnel

XTunnel

- Malware designed by APT-28
- Obfuscated with Opaque Predicates [1]
- **Handmade ground-truth** *(costly)*

	General f1-score	Obfuscated f1-score
BinDiff	0.966	0.303
QBinDiff_s	0.97	0.915

Around 400 obfuscated functions for ~ 3500 functions

(f1-score two samples in a ***plain-obfuscated*** *setting)*

[1] Bardin and al. **Backward-bounded dse: Targeting infeasibility questions on obfuscated codes**. 2017

Conclusion

- Using multiple program variants helps to weaken the obfuscation
- Differs and especially Qbindiff work well on obfuscated programs *(even for 100% of obfuscation)*
- Intra-procedural obfuscation and data obfuscation are sensitive to this attack, contrary to inter-procedural obfuscation that impedes differs and similarity tools abilities
- Valid for a large scale obfuscated dataset *(contribution)* and BinKit dataset
- Valid on real-world malware samples

Thank you

Contact information:

Email: contact@quarkslab.com

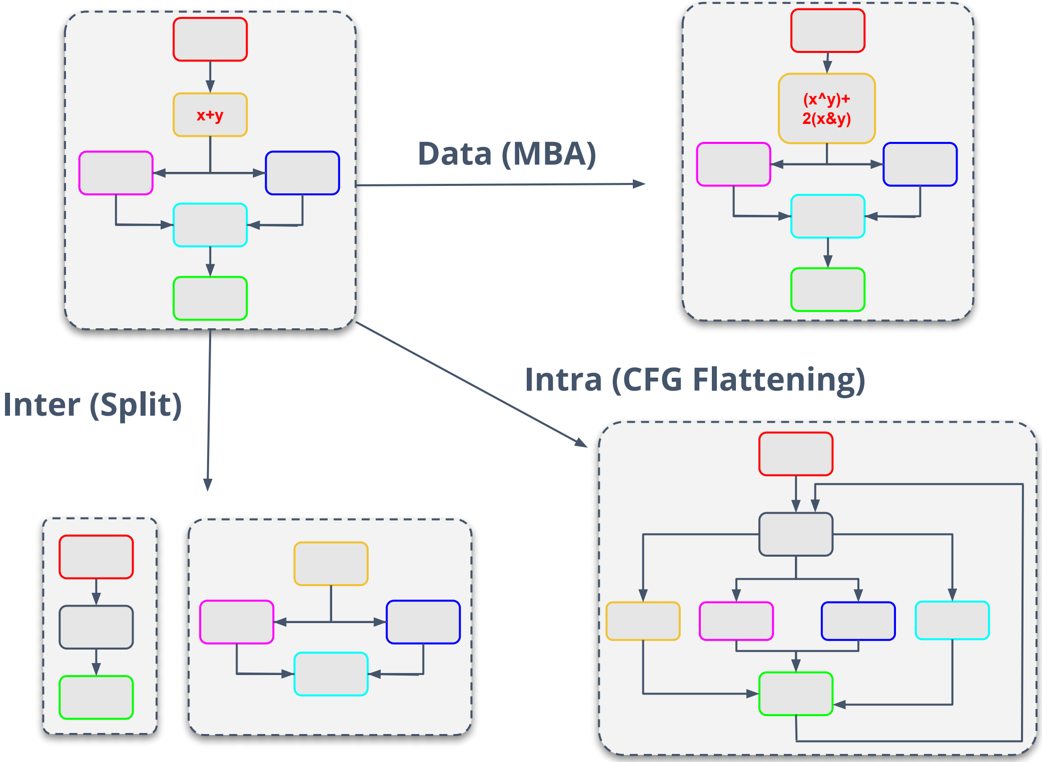
Phone: +33 1 58 30 81 51

Website: quarkslab.com

@quarkslab

Quarkslab

Obfuscation



Definition

All the techniques used to alter the syntactic properties of a program without modifying its semantics *(preserving soundness)*

Obfuscation types (static)

- Inter-procedural *(between functions)*
- Intra-procedural *(inside functions)*
- Data *(operations, constants, strings, etc.)*

Diffing solutions

	Binary diffing				Binary similarity + Matching			
	Diaphora	Bindiff	QBinDiff	DeepBinDiff	Asm2vec	JTrans	GMN	SAFE
Exporter	SQLite	Binexport	BinExport Quokka	Assembly text	Assembly text	Assembly text	ACFG	Assembly text
Technique	Ranked heuristics	Call-Graph Propagation	Belief Propagation	Enhanced word2vec	word2vec	transformer	GNN	word2vec & self-attentive network
Modularity	++	+	+++	+	+	+	+	+
Settings	Function-level & One-to-many	Function-level	Function-level	Basic-Block level	Function-level	Function-level	Function-level	Function-level