Detecting Obfuscated Function Clones in Binaries using Machine Learning

Michael Pucher, Christian Kudera, Georg Merzdovnik

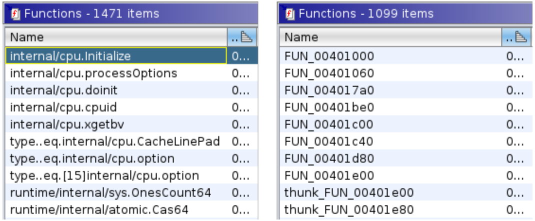
2022

University of Vienna / SBA Research

Security and Privacy / Applied Security

Function Clone Detection

- Binary function clones are common
  - Vendoring/static linking of OSS
  - Embedded firmware
- Malware also reuses code
  - Use clone detection for version tracking
- **But:** Malware uses **obfuscation**



Contributions

- End-to-end function clone detection framework: **OFCI**
    - Rely on open-source tools
    - Make use of recent NLP research
- Method for preserving call-graph information: **Call-ID**
- Analysis of performance **issues with obfuscated clone detection**
- Analysis of **virtualized function clone detection** performance

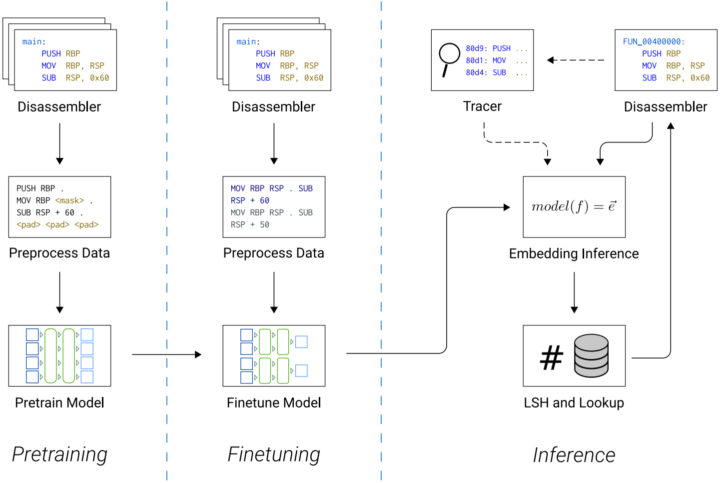
Research Questions

- **RQ1:** Can existing approaches be reduced in complexity?
  - Large models in recent architectures
  - **Goal:** Maintain same level of performance
- **RQ2:** Can additional features improve detection results?
  - In this case: Preserve call graph information
- **RQ3:** Combined with dynamic analysis for virtualization obfuscation?
  - Has not been analyzed before

OFCI

- Obfuscated Function Clone Identification (OFCI)
  - Full code largely not available for recent work
  - Provide framework for feature extraction from binaries
  - Implemented as Ghidra plugin
- Integrate ALBERT as language model
  - Recent approaches use BERT-based architectures for program analysis
  - OFCI uses ALBERT on disassembly text
  - Use PyTorch and Transformers for stock implementation
- Create instruction traces of virtualized code
  - Intel Pin

OFCI: Architecture Overview



OFCI: Feature Extraction

```
LEA     RSI, [RBX + RBP*0x8]
MOV     EDI, R13D
XOR     R8D, R8D
SUB     EDI, R12D
MOV     ECX, 0x40f9e0
MOV     EDX, 0x40d0bc
CALL    <EXTERNAL>::getopt_long
CMP     EAX,-0x1
JZ      LAB_004019a8
```

→

```
LEA RSI , [ RBX + RBP * 08 ] . MOV
EDI , R13D . XOR R8D , R8D . SUB
EDI , R12D . MOV ECX , e0 f9 40 .
MOV EDX , bc d0 40 . CALL 00 . CMP
EAX , - 01 . JZ a8 19 40 .
```

OFCI: ALBERT

- BERT-like architectures are used in recent work
  - ALBERT reduces number of trainable parameters
- Pre-Training (Self-Supervised)
  - Train the network on large corpus of disassembly text
  - Masked language modeling: Mask tokens and have network guess them
- Fine-Tuning (Supervised)
  - Sample function pairs from the dataset
  - Assign labels (similarity)
  - Training objective: cosine similarity of embeddings = label

OFCI: Tracing

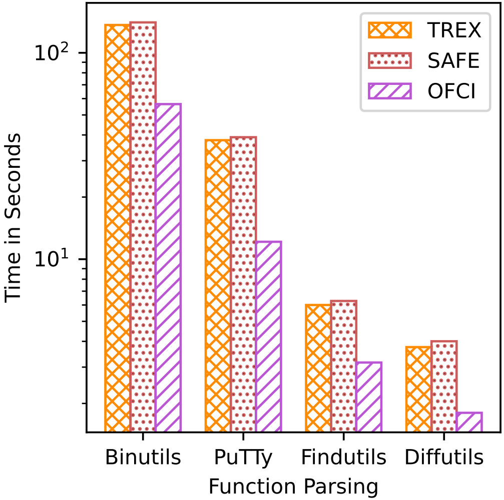
- Virtualization Obfuscation
  - Code is compiled into bytecode
  - Executed by small virtual machine
  - If functions differ in bytecode $\rightarrow$ clones not detectable statically
- **Idea:** Generate execution trace
  - Trace instructions
  - Implicitly capturing bytecode behavior
- Use Intel Pin and import in Ghidra

Evaluation

- Use dataset from recent work
  - Contains real-world O-LLVM obfuscated binaries
  - Create synthetic Tigress dataset for virtualization
- Train model
  - NVIDIA GTX Titan X for Pre-Training (1 week)
  - NVIDIA RTX 2070 Super for Fine-Tuning (24h)
- Compare against reported values
  - Similarity: ROC
  - Clone Search: Precision@1

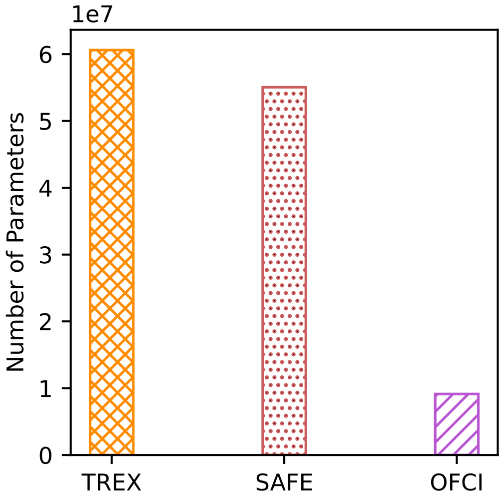
Evaluation: Feature Extraction

- Ghidra Analysis is slow
  - Decompiler needed for full disassembly
  - Most expensive analysis
- However, parsing is fast
  - **SAFE** parsing performance is underreported
  - **OFCI** is still faster

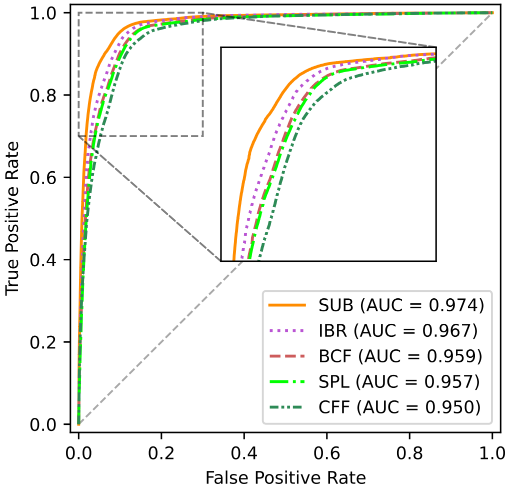


RQ1: Complexity Reduction

- Largest SotA model: **TREX**
  - 60M parameters, 700MB
- OFCI requires 17% of space at worst
  - When compared to similar models
  - 9M parameters, 35MB
- No speedup in evaluation use



RQ1: Obfuscations - ROC



	Obfuscated Pairs	
Project	TREX	OFCI
Diffutils	**0.990**	0.959
Findutils	**0.990**	0.967
GMP	**0.990**	0.869
ImageMagick	**0.989**	0.910
Libmicrohttpd	**0.991**	0.905
SQLite	**0.993**	0.956
Average	**0.990**	0.929

RQ1: Obfuscations - Precision@1

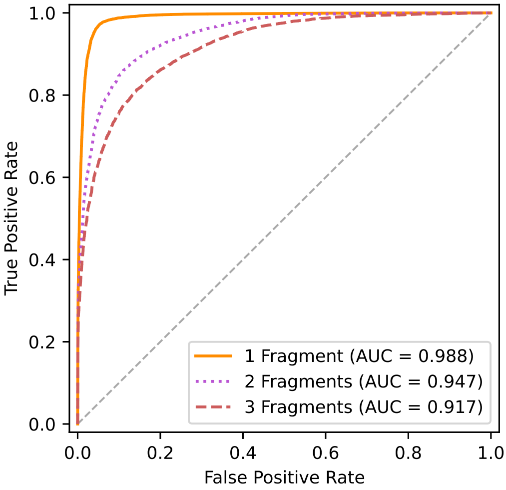
Obf.	Approach	GMP	LibTomCrypt	ImageMagick	OpenSSL	Average
bcf	Trex	**0.926**	**0.938**	**0.934**	**0.898**	**0.924**
	Asm2Vec	0.802	0.920	0.933	0.883	0.885
	Ofci	0.158	0.121	0.224	0.093	0.149
cff	Trex	**0.943**	**0.931**	**0.936**	**0.940**	**0.930**
	Asm2Vec	0.772	0.920	0.890	0.795	0.844
	Ofci	0.169	0.178	0.156	0.043	0.136
sub	Trex	**0.949**	**0.962**	**0.981**	**0.980**	**0.968**
	Asm2Vec	0.940	0.960	0.981	0.961	0.961
	Ofci	0.249	0.214	0.283	0.169	0.229

RQ1: Function Search Performance

- Reduction of model complexity
  - Other approaches use additional features (e.g. TREX and microtraces)
- Definition of similarity
  - Datasets rely on function names
  - Same name and different semantics?
- Dataset differences
  - Dataset is from TREX, but TREX doesn't use the whole dataset

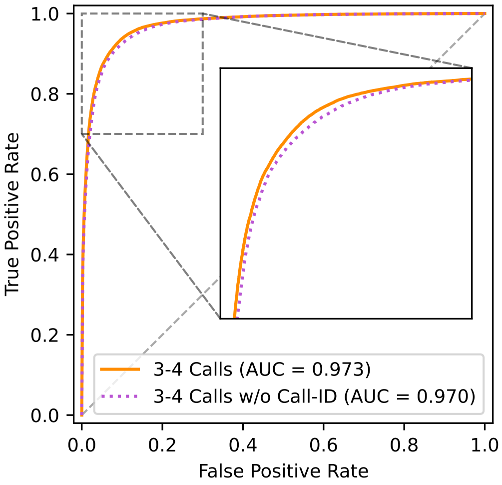
Function Fragments

- All approaches have input limits
  - Discard everything after limit
- OFCI combines multiple fragments
  - Steep drop if > 1 fragment
- Different approaches needed
  - Transformers too expensive

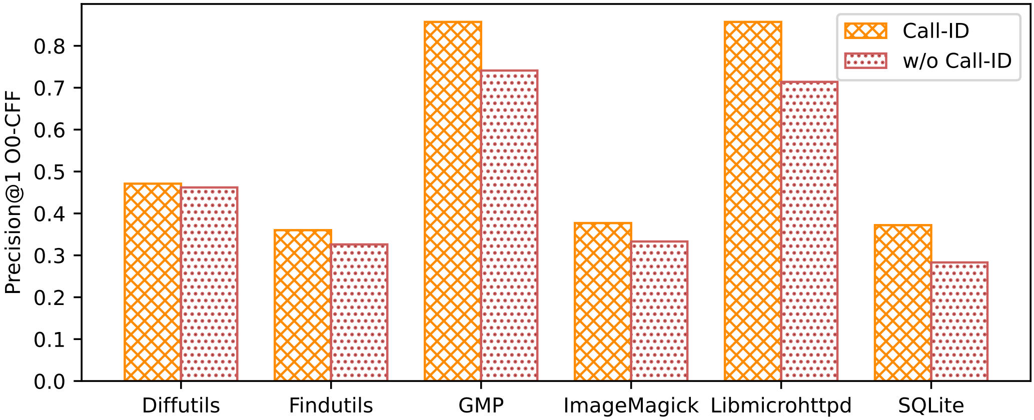


RQ2: Call-ID

- Does Call-ID have an effect?
  - Yes, but slight
  - Slightly favorable across all ROC-AUC measurements
- Effect is bigger with more calls
- Bigger effect needs model redesign

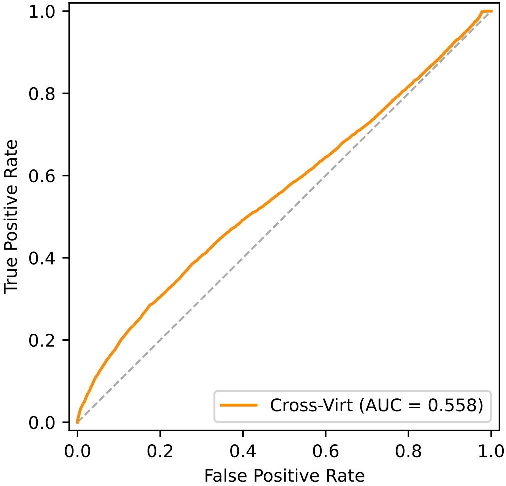


RQ2: Call-ID - Precision@1



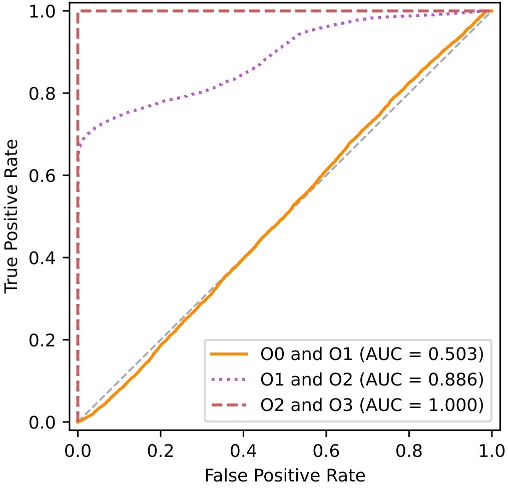
RQ3: Virtualization Obfuscated Code

- ROC-AUC close to random
  - Precision@1: 0 for *O0-Virt*
  - Precision slightly above random for *Cross-Virt*
- Random results here can affect results of other evaluations



RQ3: Virtualized Code - Issues

- Issues with dataset
  - Synthetic dataset
  - No real-world functions
- Synthetic functions are too ”simple”
  - No changes between O2 and O3
- Only small differences in traces
  - **Also:** Input size limitations!



Conclusion

- Introduced the OFCI framework
  - Efficient feature extraction
  - Open-source tools
  - Reduced model size
- Introduced Call-ID
  - Slight effect, worth pursuing in future work
- Analysis of performance issues with obfuscated clone detection
  - Issues with training data selection and similarity definition
- Analysis of virtualized function clone detection performance
  - Tracing approach not effective without modification

Questions?

Backup Slides

RQ3: Virtualized Code - Dataset Generation

- New approach for dataset generation is needed
  - Large-scale (real-world) datasets with Tigress currently not possible
  - Not possible to apply to popular open-source binaries
  - Needs re-design of Tigress
- Models working with less data instead?
  - There will always be cases like Tigress...

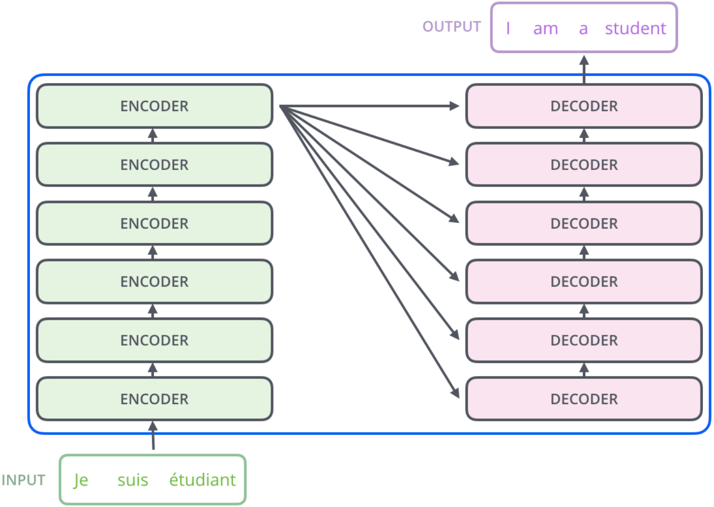
Ghidra Analysis Timings

Category	Analysis	Export
O0	88	8
O1	95	6
O2	90	6
O3	103	7
BCF	102	6
CFF	36	7
IBR	410	5
SPL	124	7
SUB	57	6
EA	42	11
VIRT	248	17
VIRT-EA	271	17

Hyperparameters

Parameter	Value
hidden_size	768
intermediate_size	3072
num_attention_heads	12
max_position_embeddings	514
num_hidden_layers	8
vocab_size	868
pretrain_batch_size	520
finetune_batch_size	522
peak_learning_rate	0.00005
transformers_version	4.11.3

Transformers



Transformers: Self-Attention