Fast Cryptographic Constants Identification in Retargetable Machine-code Decompilation

Peter Matula, DIFS FIT BUT
ID: 72

Excel@FIT

Motivation

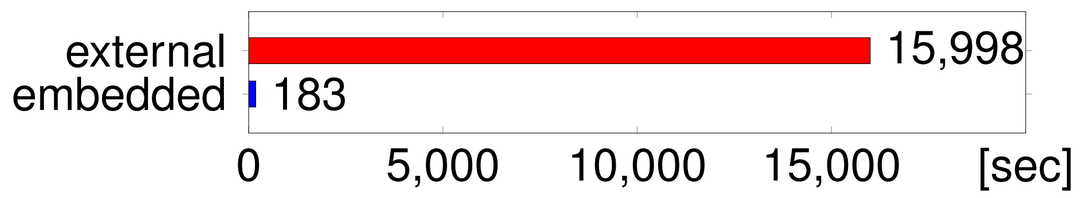
The task of a constants identification analysis is to search the binary for data sequences used by well-known algorithms. Such algorithms are often used by malware and their identification can ease up binary inspection. Analysis has the following goals:

- Identify data with known semantics.
- Assign known types to such data.
- Track usages of such data.

Signature Database

Embedded DB is loaded much faster.

Database of over 2000 signatures in YARA format is embedded into our sources to minimize the run-time parsing. Users can still provide additional signatures.

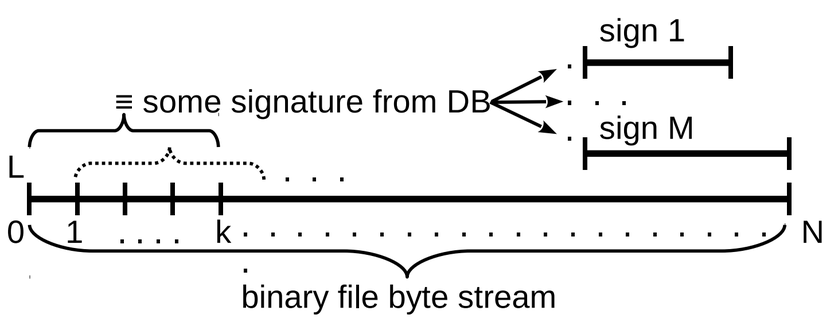


Fixed-length Signature Search

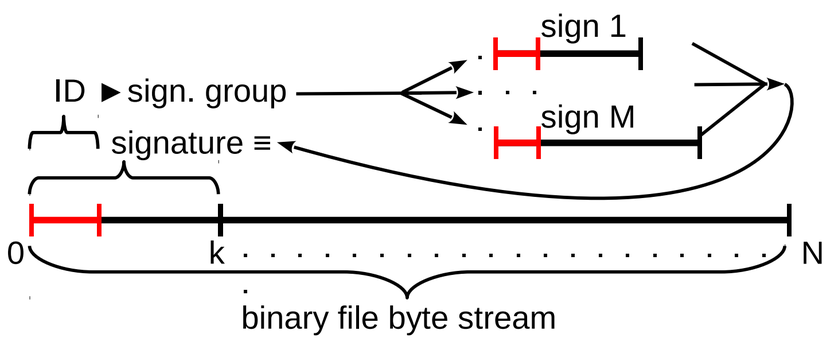
`Rar29,i16,dd3c3f1fbf59f348a164bc5a`

A fixed-length signature consists of name, data type, and a sequence of constant bytes for which the input binary is searched.

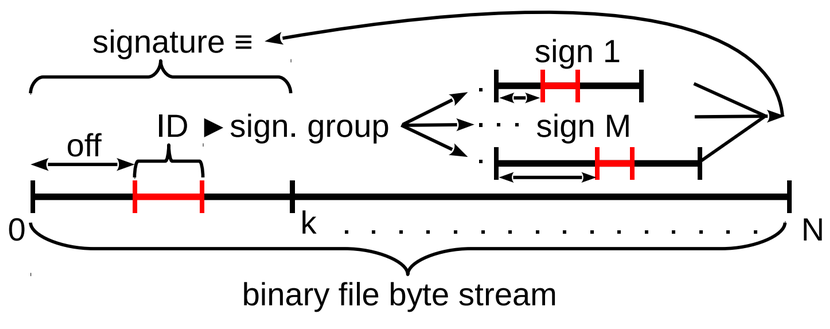
Naive Approach



Prefix-based Approach



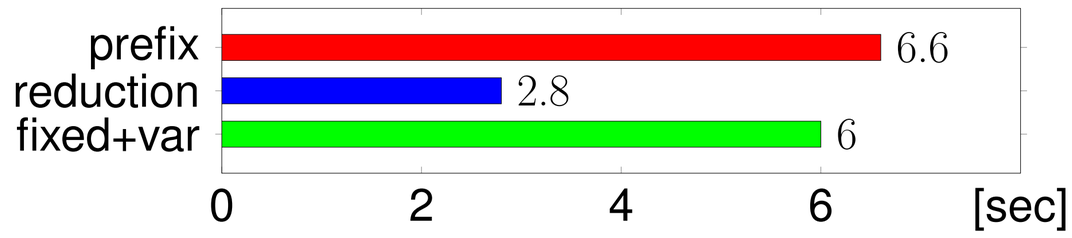
ID Collision and Prefix Hit Reduction



Variable-length Signature Search

`TEA,i32,c6ef3720[0-20]61c88647`

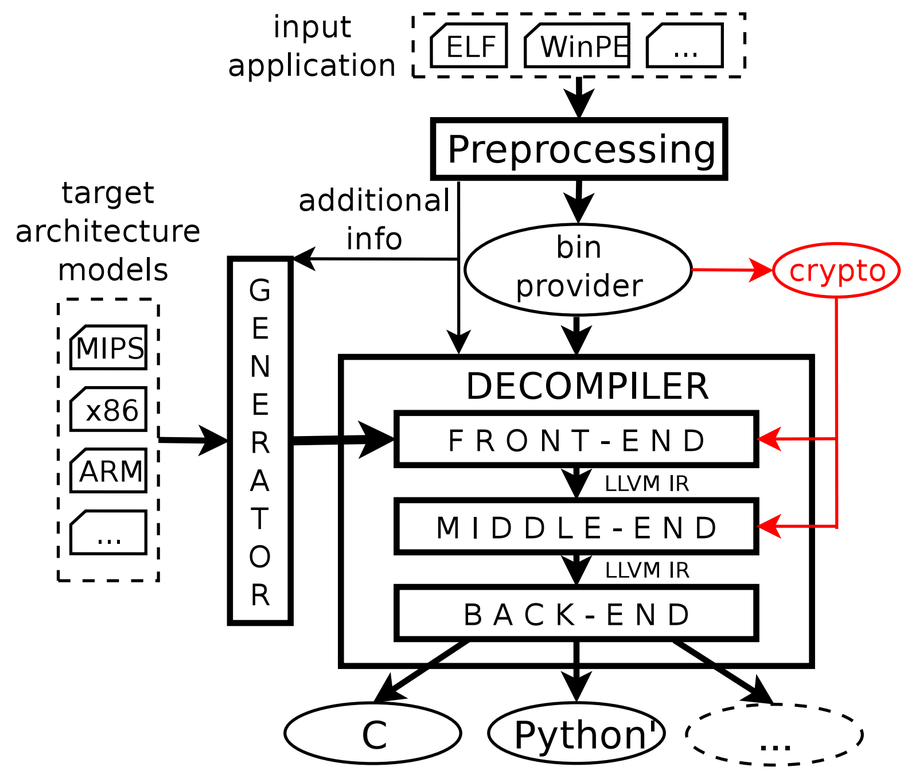
A variable-length signature consists of name, data type, and an alternating fixed and variable parts. `[0-20]` denotes arbitrary sequence from 0 to 20 bytes long.



AVG Retargetable Decompiler Integration

Results are utilized by AVG Decompiler.

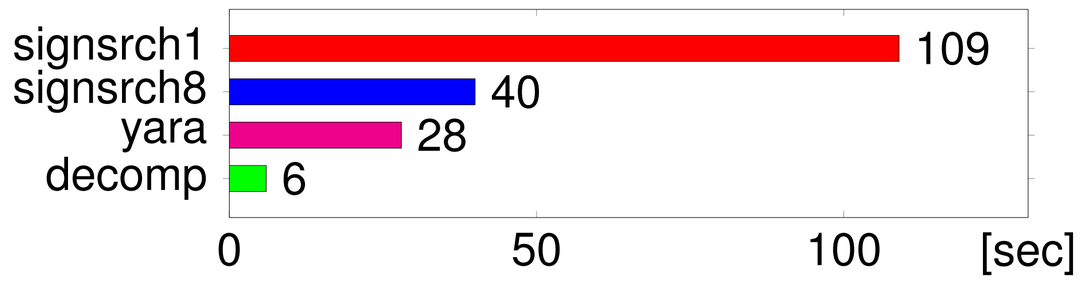
Front and middle end analyses can use gathered information to set more readable object names, tag functions, set precise data types, and create proper objects' initialization.



Experiments

Our analysis is faster than other detectors.

We compared our solution to other state-of-the-art detectors on a 130 MB real-world malware test suite.



Main Results

- Solution is extremely simple and fast.
- Results are utilized in AVG Decompiler.