Control Flow Graphs as Malware Signatures

Guillaume Bonfante, *Matthieu Kaczmarek* and Jean-Yves Marion

CARTE – LORIA – INPL

TCV'07 – Nancy

Semantics in malware detection

- Semantic aspects in detection:
  - *Specification and prover*
    Webster, Malcaolm (JCV06).
  - *Abstract interpretation*
    Dalla Preda, Christodorescu, Jha, Debray (POPL07).
  - *Instruction normalization*
    Lakhotia, Mohammed (WCRE04).
  - *Instruction normalization and GFCs*
    Bruschi, Martignoni, Monga (TR06).
  - *Data flow*
    Venable, Chouchane, Karim, and Lakhotia (DIVMA05).
- Detection based on GFCs: a first step.

1 Design of the detector

2 Relevance and practicability

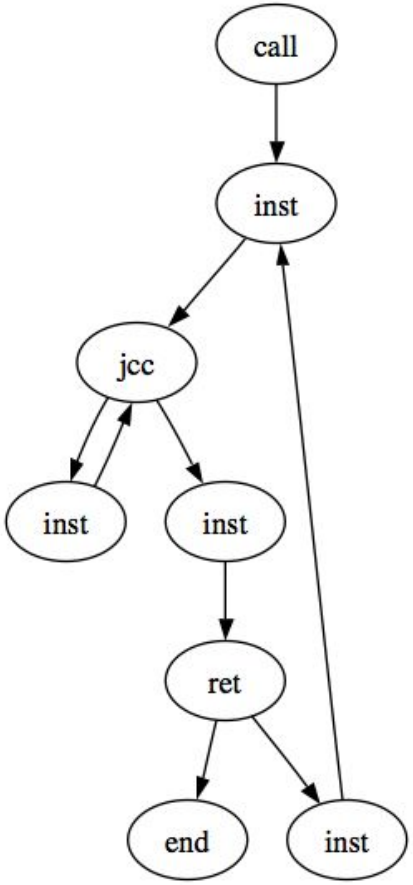
3 Soundness w.r.t mutations

CFGs from x86 machine code

- Four kinds of flow instructions.
  - `jmp`: unconditional jump, one successor.
  - `jcc`: conditinal jump, two succesors.
  - `call`: function call, one successor.
  - `ret`: function return, unbounded successors.
- Kind `inst`: contiguous blocks of non-flow instructions.

An example of CFG extraction

```
     call @1
     inc  dx
@1   mov  ax, 1
     mov  cx, 8
@2   mul  cx
     dec  cx
     cmp  cx, 0
     jne  @2
     ret
```



Detection strategy

- Focus on malware detection.
- Extract CFGs from known malware to build a database.
- A program is detected if its CFG is in the database.

Experimental results

Size of CFGs	$0-100$	$101-500$	$501-3000$	>3000	Overall
Sane	76	126	325	223	750
Malware	1024	590	358	106	2278
False-pos	23	7	1	0	31
Ratios	30%	5.6%	0.3%	0.0%	4.1%

Table: Results of the experiments

Statistical methods	Kephart, Arnold VBIC95	$0,5\%$ - 34%
Neural networks	Tesauro, al. IEEE96	1%
Data Mining methods	Schultz, al. IEEE01	2.2% - 47.5%
Heuristics in industry	Gryaznov VBIC99	0.2%

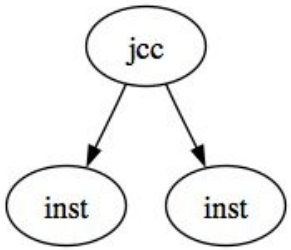
Practicability (MBP 2,16GHz)

- Building the database: 17 min (144 ko/s).
- Scanning: 10 min (340 ko/s).

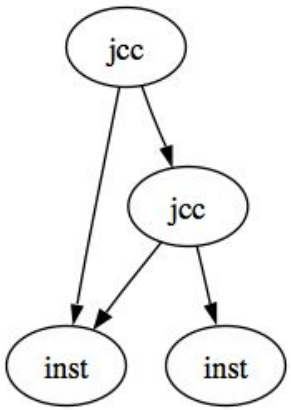
jcc reduction

- Reduce triangles into caps.

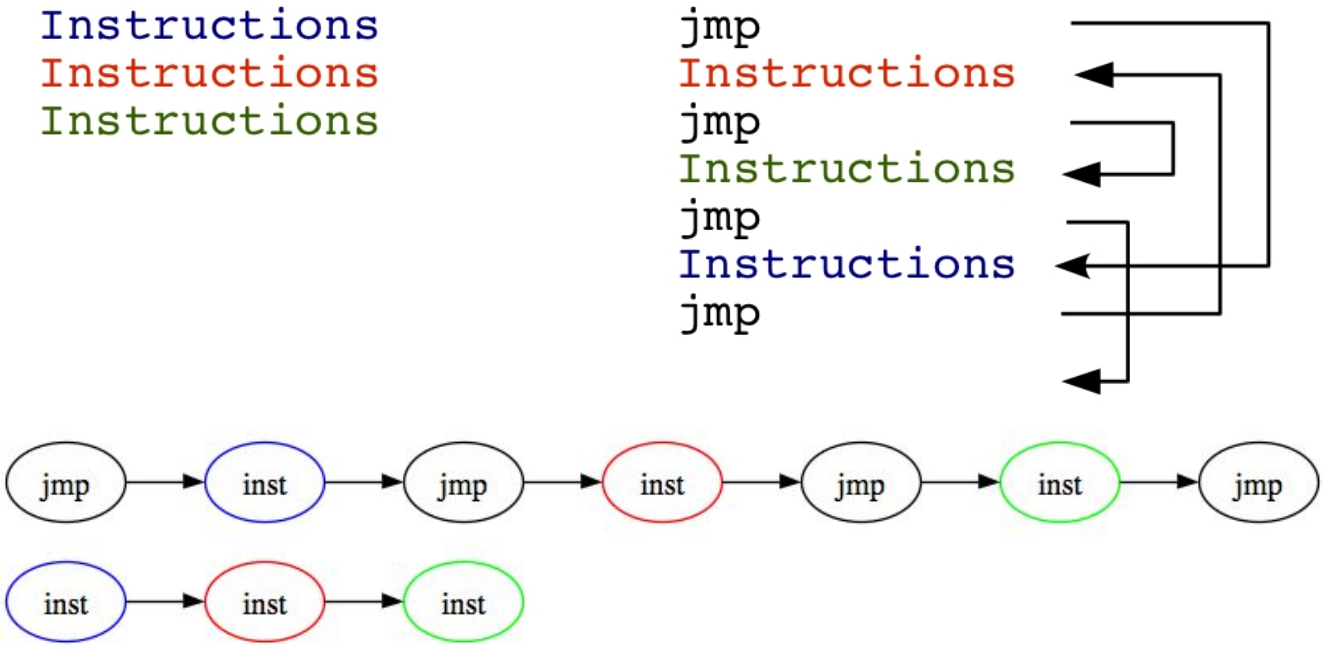
```
jna @@@ /* CF == 1 or ZF == 1*/
```




```
jb @@@ /* CF == 1 */
jz @@@ /* ZF == 1*/
```

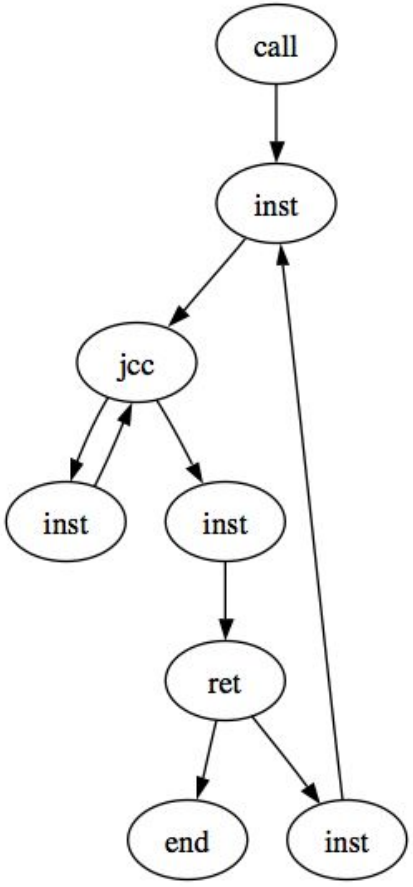


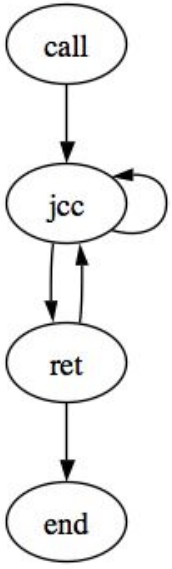
Code reordering



Code reordering

- Lines do not provide flow information.





Experimental results

Size of CFGs	$0-100$	$101-500$	$501-3000$	>3000	Overall
Sane	91	145	347	167	750
Malware	1297	528	356	67	2278
False-pos	23	7	1	0	31
Ratios	25.3%	7%	0.3%	0.0%	4.1%

Table: Results of the experiments with reduction

Further research

- Enhance graph extraction and graph matching.
- More reductions.
- Viral infection: sub-graph isomorphism
  Bruschi, Martignoni, Monga (TR06).
- Associate with sub-string matching