Scattered context grammars and generic reverse compilation

Lukáš Ďurfina

UIFS FIT VUT

BRNO UNIVERSITY OF TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

Contents

- Motivation and aims
- Ideas
- Scattered context grammar
- Application in reverse compilation

Motivation

Why?

- a lot of space for research
- usefull applications

Probably the most usefull application

- every year milions of new unique malware programs
- families of malware with similar code and behaviour
- polymorphic and metamorphic viruses
- antivirus software, IDS

Way of reverse compilation

Facts

- part of reverse engineering
- do not produce code in high level language
- find and mark patterns
- recognize specific behaviour

Aim

Give answer

YES or NO

And what was a question?

- same behaviour
- level of an accuracy

Ideas of solution

Base

- imagine executable as word of some language L
- we need mechanism for comparing these words

Ideas of solution

Transformation

- transformational grammar
- we are [not]? able to transform one word (executable) to another

Transformation

- transformational grammar
- we are [not]? able to transform one word (executable) to another
  - $W_A \stackrel{?}{\Rightarrow} W_B$

Transformation

- transformational grammar
- we are [not]? able to transform one word (executable) to another
  - $W_A \stackrel{?}{\Rightarrow} W_B$
  - $W_A \Rightarrow W_{C_1}, W_B \Rightarrow W_{C_2}, W_{C_1} \stackrel{?}{=} W_{C_2}$

Ideas of solution

Generation

- we are [not]? able to create generative grammar G
- G can generate “all” words (executables) with specific behaviour

Ideas of solution

Generation

- we are [not]? able to create generative grammar G
- G can generate “all” words (executables) with specific behaviour
- G can also parse them

Ideas of solution

Generation

- we are [not]? able to create generative grammar G
- G can generate "all" words (executables) with specific behaviour
- G can also parse them
  - LL-parsing

Conditions

- has strong generative power
- can describe the problem
- is not too much complex for implementation

Definition

$$G = (V, T, P, S)$$

V is the total alphabet
T is the set of terminals, $T \subset V$
S is the start symbol, $S \in V - T$
P is a finite set of productions of the form

$$(A_1, \ldots, A_n) \rightarrow (x_1, \ldots, x_n),$$

where $A_1, \ldots, A_n \in V - T, x_1, \ldots, x_n \in V^*$

Derivation step naturaly describes problem

For $(A_1, \ldots, A_n) \to (x_1, \ldots, x_n) \in P$ and

$$u = u_1 A_1 \ldots u_n A_n u_{n+1}$$
$$v = u_1 x_1 \ldots u_n x_n u_{n+1}$$

we write $u \Rightarrow v[(A_1, \ldots, A_n) \to (x_1, \ldots, x_n)]$

Reasons for SCG

Strong generative power

$\mathscr{L}(SC) = \mathscr{L}(RE)$

Reasons for SCG

Strong generative power

$\mathscr{L}(SC) = \mathscr{L}(RE)$

Comparison with PSCG

allow ϵ rules

Reasons for SCG

Example

```
a = countA();                    b = countB();
b = countB();                    a = countA();
return magic(a,b);               return magic(a,b);
```

Example

```
a = countA();                    b = countB();
b = countB();                    a = countA();
return magic(a,b);               return magic(a,b);
```

$(\langle b = countB() \rangle, \langle a = countA() \rangle) \rightarrow (\langle a = countA() \rangle, \langle b = countB() \rangle)$

Example

```
a = countA();                    a = countA();
b = countB();                    b = countB();
return magic(a,b);               return magic(a,b);
```

$(\langle b = countB() \rangle, \langle a = countA() \rangle) \rightarrow (\langle a = countA() \rangle, \langle b = countB() \rangle)$

Example

```
a = countA();
b = countB();
c = countC();
return magic(a,b);
```

Example

```
int a = countA();
int b = countB();
volatile int c = countC();
return magic(a,b);
```

Example

A: `int a = countA();`
B: `int b = countB();`
C: `volatile int c = countC();`
$\langle AB \rangle$: `return magic(a,b);`

$(A, B, C, \langle AB \rangle) \rightarrow (A, B, \epsilon, \langle AB \rangle)$

Example

A: `int a = countA();`
B: `int b = countB();`
C: `volatile int c = countC();`
$\langle AB \rangle$: `return magic(a,b);`

$(A, B, C, \langle AB \rangle) \rightarrow (A, B, \epsilon, \langle AB \rangle)$

Example

```
int a = countA();
int b = countB();
return magic(a,b);
```

Definition

$$G = (V, T, P, I)$$

V is the total vocabulary
T is the set of terminals (or the *output vocabulary*), $T \subset V$
P is a finite set of productions of scattered context productions
I is the *input vocabulary*, $I \subset V$

Derivation step

$T(G, K)$: transformation T that G defines from $K \subseteq I^*$
$T(G, K) = \{(x, y) : x \Rightarrow_G^* y, x \in K, y \in T^*\}$

Example

$$G = (V, T, P, I)$$

$V = \{A, B, C, a, b, c\}$
$T = \{a, b, c\}$
$I = \{A, B, C\}$
$P = \{(A, B, C) \rightarrow (a, bb, c)\}$

We can take input sentence $AABBCC$:

$$AABBCC \Rightarrow_G aABbbcC \Rightarrow_G aabbbbcc$$

Description model for executables

Action blocks

- granularity
- connections

ASM code transformation

Example

```
INC ax                    INC bx
INC bx                    INC ax
MUL bx                    MUL bx
ADD bx, ax                MOV ebx, eax
PUSH eax                  PUSH eax
```

Example

$\langle INCA \rangle$	$\langle INCB \rangle$
$\langle INCB \rangle$	$\langle INCA \rangle$
$\langle MULB \rangle$	$\langle MULB \rangle$
$\langle ADDBA \rangle$	$\langle MOVBA \rangle$
$\langle PUSHA \rangle$	$\langle PUSHA \rangle$

Example

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle ADDBA \rangle$
$\langle PUSHA \rangle$

$\langle INCB \rangle$
$\langle INCA \rangle$
$\langle MULB \rangle$
$\langle MOVBA \rangle$
$\langle PUSHA \rangle$

Rules:
$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$

Example

$\langle INCA \rangle$	$\langle INCB \rangle$
$\langle INCB \rangle$	$\langle INCA \rangle$
$\langle MULB \rangle$	$\langle MULB \rangle$
$\langle ADDBA \rangle$	$\langle MOVBA \rangle$
$\langle PUSHA \rangle$	$\langle PUSHA \rangle$

Rules:

$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$

Example

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle ADDBA \rangle$
$\langle PUSHA \rangle$

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle MOVBA \rangle$
$\langle PUSHA \rangle$

Rules:
$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$

Example

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle ADDBA \rangle$
$\langle PUSHA \rangle$

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle MOVBA \rangle$
$\langle PUSHA \rangle$

Rules:

$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$

$(\langle MULB \rangle, \langle MOVBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$

Example

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle ADDBA \rangle$
$\langle PUSHA \rangle$

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle MOVBA \rangle$
$\langle PUSHA \rangle$

Rules:

$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$

$(\langle MULB \rangle, \langle MOVBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$

Example

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle ADDBA \rangle$
$\langle PUSHA \rangle$

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle PUSHA \rangle$

Rules:

$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$

$(\langle MULB \rangle, \langle MOVBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$

Example

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle ADDBA \rangle$
$\langle PUSHA \rangle$

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle PUSHA \rangle$

Rules:

$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$
$(\langle MULB \rangle, \langle MOVBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$
$(\langle MULB \rangle, \langle ADDBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$

Example

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle ADDBA \rangle$
$\langle PUSHA \rangle$

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle PUSHA \rangle$

Rules:

$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$
$(\langle MULB \rangle, \langle MOVBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$
$(\langle MULB \rangle, \langle ADDBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$

Example

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle PUSHA \rangle$

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle PUSHA \rangle$

Rules:

$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$
$(\langle MULB \rangle, \langle MOVBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$
$(\langle MULB \rangle, \langle ADDBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$

Example

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle PUSHA \rangle$

$\langle INCA \rangle$
$\langle INCB \rangle$
$\langle MULB \rangle$
$\langle PUSHA \rangle$

Rules:

$(\langle INCB \rangle, \langle INCA \rangle) \rightarrow (\langle INCA \rangle, \langle INCB \rangle)$
$(\langle MULB \rangle, \langle MOVBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$
$(\langle MULB \rangle, \langle ADDBA \rangle, \langle PUSHA \rangle) \rightarrow (\langle MULB \rangle, \epsilon, \langle PUSHA \rangle)$

Example

$$G = (V, T, P, I)$$

$$V = \{\langle INCA\rangle, \langle INCB\rangle, \langle MULB\rangle, \langle ADDBA\rangle, \langle MOVBA\rangle, \langle PUSHA\rangle, INCA, INCB, MULB, ADDBA, MOVBA, PUSHA\}$$

$$T = \{INCA, INCB, MULB, ADDBA, MOVBA, PUSHA\}$$

$$I = \{\langle INCA\rangle, \langle INCB\rangle, \langle MULB\rangle, \langle ADDBA\rangle, \langle MOVBA\rangle, \langle PUSHA\rangle\}$$

$$P = \{(\langle INCB\rangle, \langle INCA\rangle) \rightarrow (\langle INCA\rangle, \langle INCB\rangle),$$
$$(\langle MULB\rangle, \langle MOVBA\rangle, \langle PUSHA\rangle) \rightarrow (\langle MULB\rangle, \epsilon, \langle PUSHA\rangle),$$
$$(\langle MULB\rangle, \langle ADDBA\rangle, \langle PUSHA\rangle) \rightarrow (\langle MULB\rangle, \epsilon, \langle PUSHA\rangle),$$
$$(\langle INCA\rangle) \rightarrow (INCA), \dots\}$$

Thank you for your attention

Literature

A. Meduna and J. Techet.
Scattered Context Grammars and their Applications.
WIT Press, 2010.

P. Szor.
Počítačové viry - analýza útoku a obrana.
Zoner Press, 2006.