

Regular Expression Denial of Service

```
(\w+[-0-9A-Z_a-  
z]*(\|(\?|\w[\-\_0-9A-Z_a-  
z]*?)*)?([!+]|[:=][f\u00f1nos][%@]  
]?|:(?\d+|\+)[%@]?)?
```

How do we Regexp Match?

- Convert Pattern to NFA and simulate
 - Simulation linear in input length
- Convert Pattern to DFA and simulate
 - Simulation linear in input length
- Interpret the pattern nondeterministically
 - Quadratic or exponential time!
 - Allow recognition of irregular languages

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^*$ O

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^*$ O

aaaaaaaaaaaaaaaaaaaaaaaX

Time so far: n

Quadratic Time Example

$a^* [ab]^*$ O

aaaaaaaaaaaaaaaaaaaaaaX

Time so far: n

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^*$ O

aaaaaaaaaaaaaaaaaaaaaaX

Time so far: $n+1$

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^*$ O

aaaaaaaaaaaaaaaaaaaaaaX

Time so far: $n+1+2$

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^*$ O

aaaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^*$ O

aaaaaaaaaaaaaaaaaaaaaaaX

Time so far: $n+1+2+\dots+11$

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^* O$

aaaaaaaaaaaaaaaaaaaaaaX

Quadratic Time Example

$a^* [ab]^*$ O

aaaaaaaaaaaaaaaaaaaaaaaX

Time so far: $n+1+2+\dots+n = O(n^2)$

Worst cases

- Cubic time

- $a^* [ab]^* [ac]^* O$

- Exponential time

- $(a|aa)^* O$
 - $[a-z]^+ @ [a-z]^+ ([a-z\.]^+ \.)^+ [a-z]^+$

- Recognizes email addresses

- Taken from SpamAssassin

Worst cases

- Cubic time

- $a^* [ab]^* [ac]^* O$

- Exponential time

- $(a|aa)^* O$
 - $[a-z]^+ @ [a-z]^+ ([a-z\.]^+ \.)^+ + [a-z]^+$

- Recognizes email addresses

- Taken from SpamAssassin

Optimizations

May use other algorithms without this worst case

How?

What?

Proof?