Yara:

Down the Rabbit Hole Without Slowing Down

Dominika Regéciová | dominika.regeciova@avast.com | @regeciovad

Botconf 2022

TLP WHITE

A few notes about me

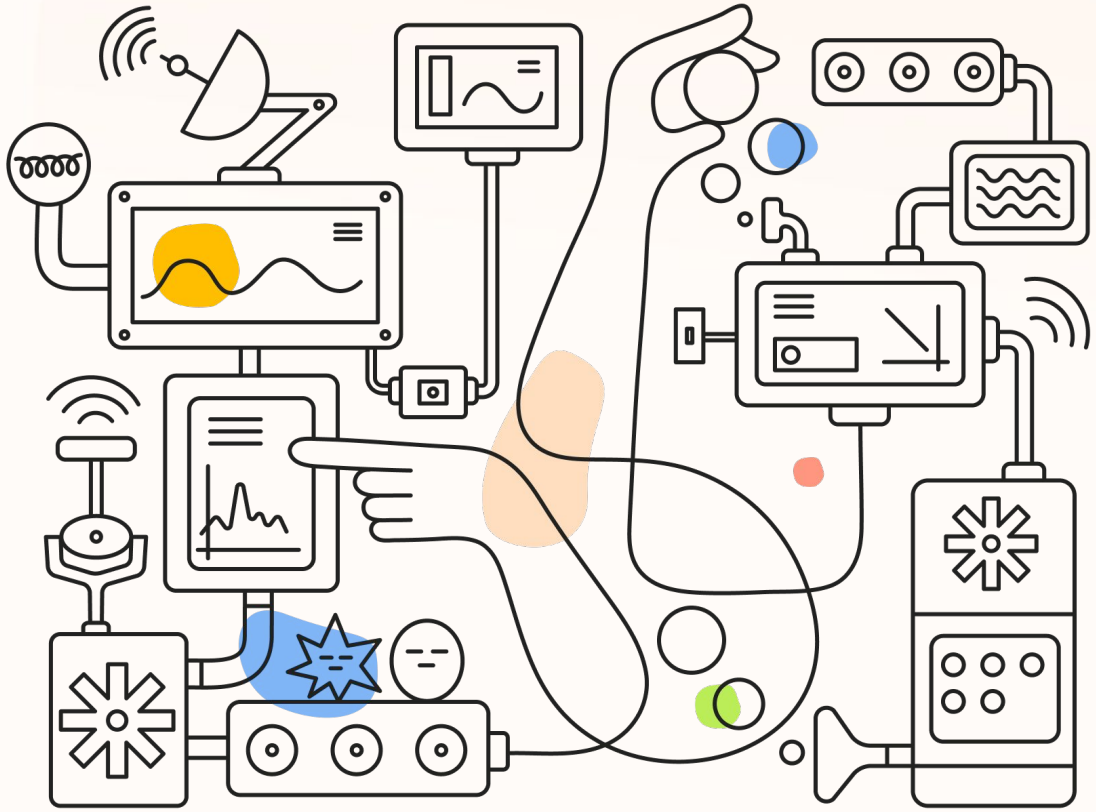
- Researcher at Avast
- Ph.D. student at FIT BUT in Brno
- Projects with ESA and Czech Police
- My research:
  - Formal models and languages in security
  - Pattern matching
  - Blockchain technology

What to expect from this talk

- What is Yara
- Yara Performance
- Changes in Yara

Avast

What is Yara?



Yara rules

```
import "math"

rule Botconf_malware
{
        meta:
                author = “John, Terry, and Caitlin”
                description = "detection based on this great conference"
        strings:
                $str = "cmd.exe" ascii wide nocase
                $re = /\w.*\d/
        condition:
                $str and $re and
                math.entropy(0, filesize) > 7.0 and
                uint16(0) == 0xFFFF
}
```

Yara rules

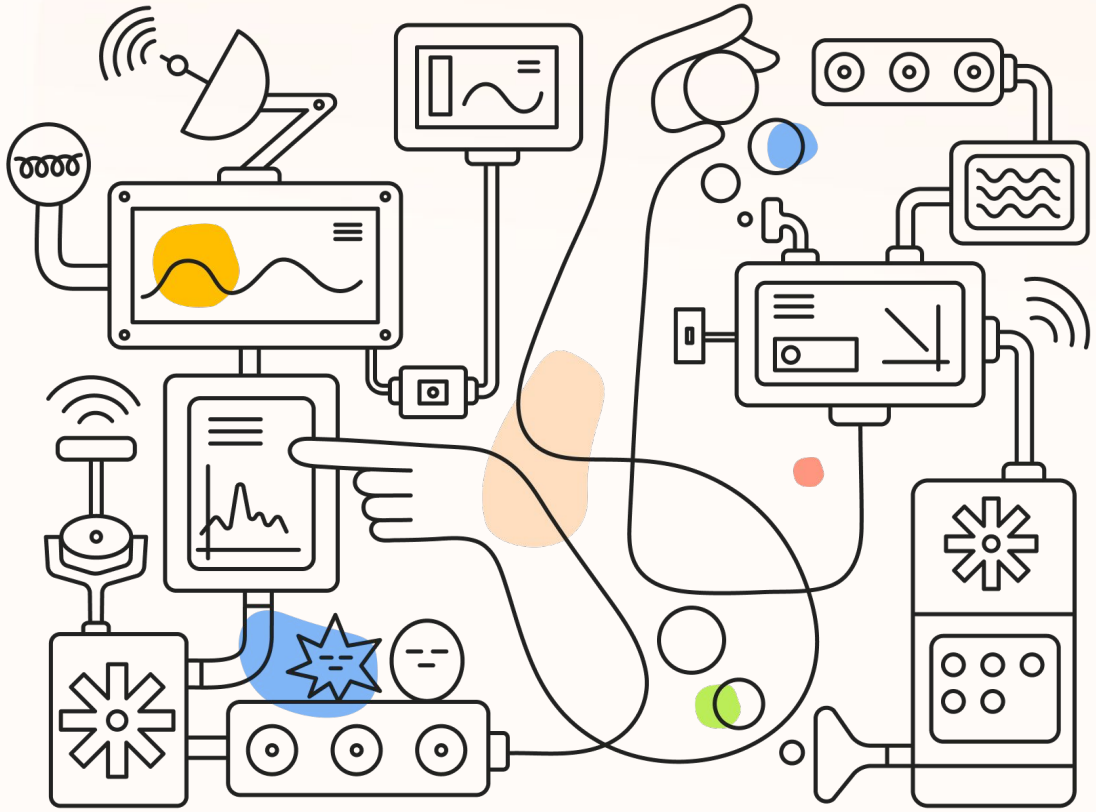
We want to scan the directory *secret_dir* recursively with our rule, which has 5.9 GB of data with 39,852 files:

./yara botconf_malware.yar -r secret_dir
warning: rule "Botconf_malware": too many matches for $re, results for this rule may be incorrect
warning: rule "Botconf_malware" in botconf_malware.yar(10): $re contains ., .+ or .{x,} consider using .{,N}, .{1,N} or {x,N} with a reasonable value for N*
warning: rule "Botconf_malware" in botconf_malware.yar(10): string "$re" may slow down scanning

Also, that scanning took almost 45 minutes 🤯. What can we do with it?

Avast

Yara Performance



Atoms selection from strings

- In static parts, we firstly match the substrings known as atoms
- The match is later confirmed from a list of potential matches in files
- The selection of atoms influences the speed of matching
- Atoms have 0 to 4 characters (0-length atom will match everything)

*/**abc**.{1,20}def/*
*/(one|two)**thre**e/*
*{ 00 00 00 00 [1-4] **01 02 03 04** }*
*/a**(c|d)**/* — Two atoms c and d.
/\w.\d/ => "" (0-length atom)* — This is bad for speed

Strings

- Use only modificators you really need

  $s1 = "cmd.exe" ascii only
  $s2 = "cmd.exe" ascii ascii only, same as $s1
  $s3 = "cmd.exe" wide 'UTF-16' only
  $s4 = "cmd.exe" ascii wide both ascii and 'UTF-16'

- Case-insensitive modificators

  $str = “cmd.exe” nocase will search all combinations such as Cmd., cMd.,...
  $re = /[Cc]md\.exe/ give you better results

Strings

- Be specific as possible
  $re = /\w.\d/* This is not good for matching (x0, a_1, abc3, whatever123,...)
  $re = /\w.{7, 8}\d/

- *Text string prefix also improves speed*
  $re1 = /.{0,2}Tom/ $re1 will find Tom, xTom, xxTom in "xxTom"
  $re2 = /Tom.{0,2}/ $re2 will find Tomxx in "Tomxx"
  $re = /C:\\.{7, 8}\d/

Too Many Matches

- Till Yara 4.1.0 - too many matches generated an error and the results could be invalid
- From version 4.1.0 a warning is raised, the scanning is finished, but the results still can be compromised (we still want to avoid it when possible)
- There is no one simple solution for this problem
- Possible causes and possible fixes:
  - The quantifiers *.** and *.+*, *.*?*
  - The quantifiers without upper bound such as *x{14,}*
  - Too large range (e. g. *x{1,300000}*)
  - Big jumps in the hexadecimal strings: *{00 01 02 [1 - 100] 04}*
  - Wild-cards characters - can they be specified more precisely, or could be string split into two, omitting the wild-cards character?
  - Alternations: can it be split into two or more strings?
  - Try to add specification for words matching (*fullword*, *\b*, ...)

Conditions

- Evaluation of static parts of rules are evaluated first
- Condition such as *filesize < 100 and $expensive_regex* will not help
- Short-circuit evaluation:

  // EXPENSIVE and CHEAP
  math.entropy(0, filesize) > 7.0 and uint16(0) == 0xFFFF

  // CHEAP and EXPENSIVE
  uint16(0) == 0xFFFF and math.entropy(0, filesize) > 7.0

- Integer loop optimization (both loops will stop iterating after the 1st time)

  for all i in (0..100): (false)
  for any i in (0..100): (true)

Yara rules

```
import "math"

rule Botconf_malware
{
        meta:
                author = “John, Terry, and Caitlin”
                description = "detection based on this great conference"
        strings:
                $re1 = /[Cc]md\.exe/
                $re2 = /C:\\\.{7,8}\d/
        condition:
                $re1 and $re2 and
                uint16(0) == 0xFFFF and
                math.entropy(0, filesize) > 7.0
}
```

Yara rules

```
import "math"

rule Botconf_malware
{
        meta:
                author = “John, Terry, and Caitlin”
                description = "detection based on this great conference"
        strings:
                $re1 = /[Cc]md\.exe/
                $re2 = /C:\\\.{7,8}\d/
        condition:
                $re1 and $re2 and
                uint16(0) == 0xFFFF and
                math.entropy(0, filesize) > 7.0
}
```

The scanning takes only 3 seconds! 😎

Additional tips and new features

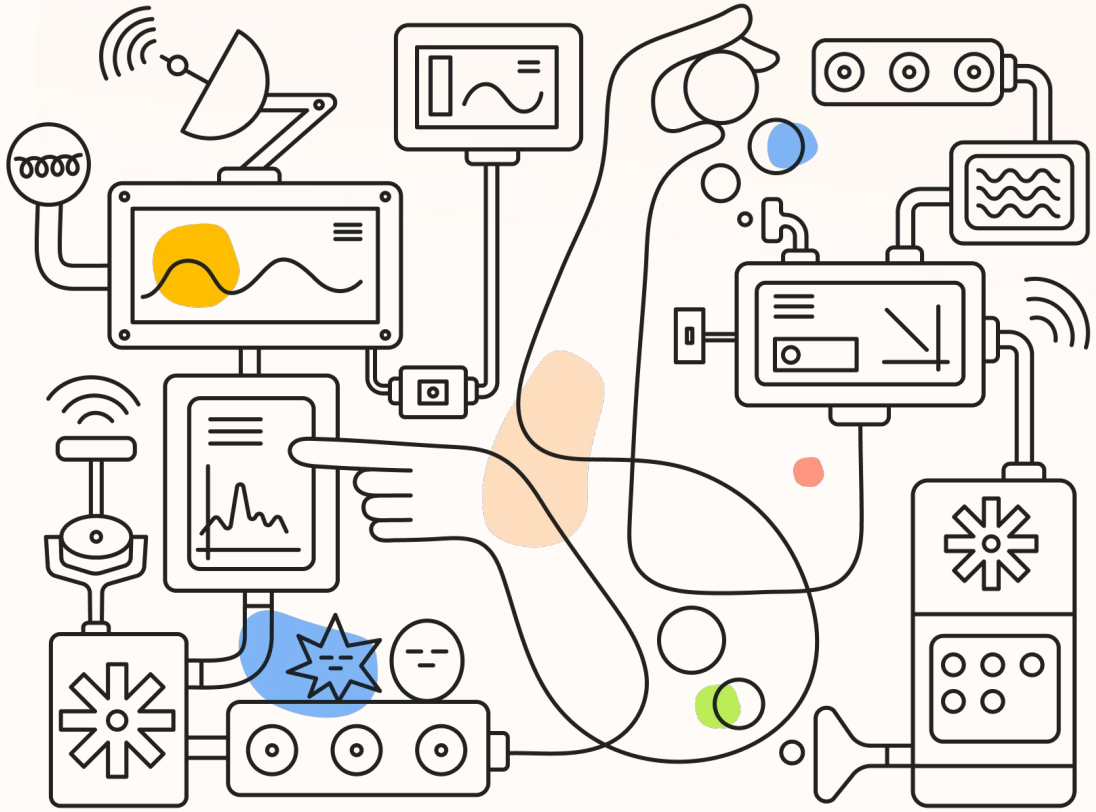
- *--no-follow-links* command-line option
- *--skip-larger* option for skipping files larger than a certain size while scanning directories
- New operator % for string sets. Example: *20% of them*
- New syntactic sugar allows writing *0 of ($a)* as *none of ($a*)*

More resources

- VirusTotal GitHub page
- Yara Documentation
- YARA Performance Guidelines
- yara_friends on Keybase

Avast

Our Changes in Yara



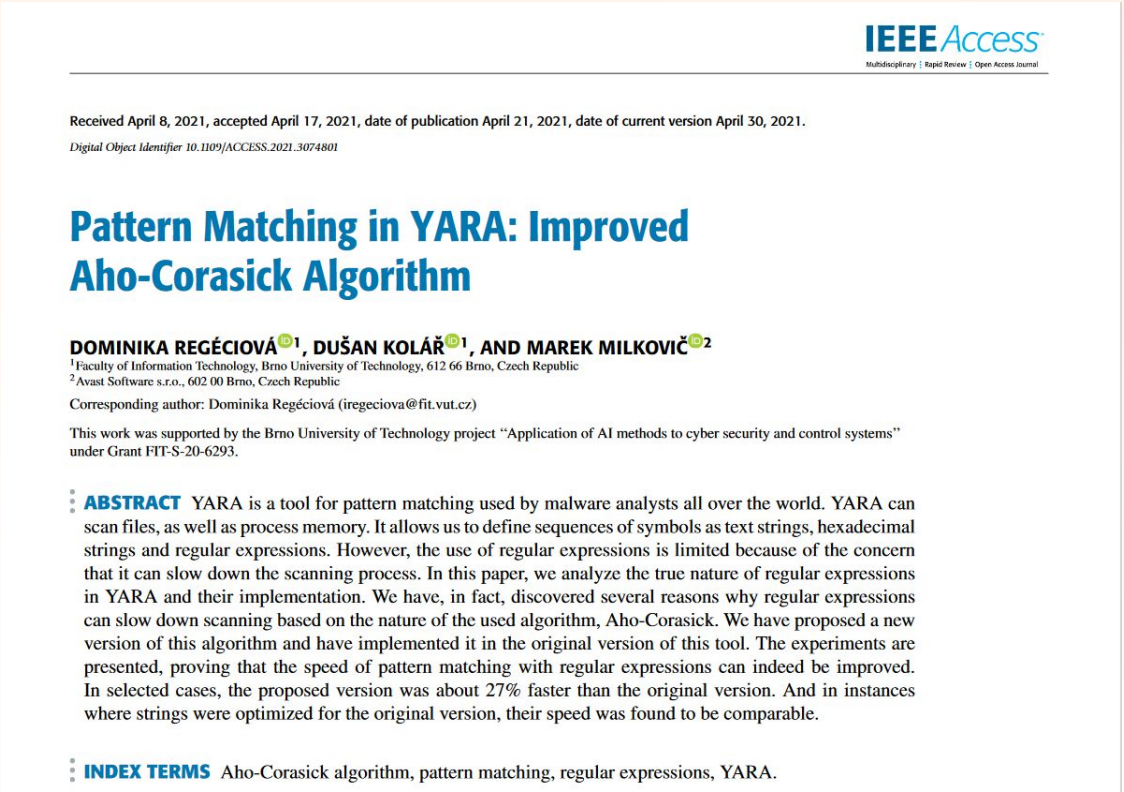
Motivation

This rule detects Bitcoin addresses in P2PKH and P2SH types

```
rule contains_btc_address
{
    strings:
        $btc_address = /[13][a-km-zA-HJ-NP-Z1-9]{25,34}/ fullword ascii wide
    condition:
        $btc_address
}
```

```
btc_address.yar(4): warning: $btc_address in rule btc_address is slowing down scanning
```

Results



IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 8, 2021, accepted April 17, 2021, date of publication April 21, 2021, date of current version April 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074801

Pattern Matching in YARA: Improved Aho-Corasick Algorithm

DOMINIKA REGÉCIOVÁ[1], DUŠAN KOLÁŘ[1], AND MAREK MILKOVIČ[2]

[1]Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic
[2]Avast Software s.r.o., 602 00 Brno, Czech Republic

Corresponding author: Dominika Regéciová (iregeciova@fit.vut.cz)

This work was supported by the Brno University of Technology project "Application of AI methods to cyber security and control systems" under Grant FIT-S-20-6293.

ABSTRACT YARA is a tool for pattern matching used by malware analysts all over the world. YARA can scan files, as well as process memory. It allows us to define sequences of symbols as text strings, hexadecimal strings and regular expressions. However, the use of regular expressions is limited because of the concern that it can slow down the scanning process. In this paper, we analyze the true nature of regular expressions in YARA and their implementation. We have, in fact, discovered several reasons why regular expressions can slow down scanning based on the nature of the used algorithm, Aho-Corasick. We have proposed a new version of this algorithm and have implemented it in the original version of this tool. The experiments are presented, proving that the speed of pattern matching with regular expressions can indeed be improved. In selected cases, the proposed version was about 27% faster than the original version. And in instances where strings were optimized for the original version, their speed was found to be comparable.

INDEX TERMS Aho-Corasick algorithm, pattern matching, regular expressions, YARA.

- Improved matching for strings:
  - The scanning with BTC addresses is ten times faster and without any warning
  - The scanning with the nocase option is about 27% faster

There is more...

Dominika Regéciová @regeciovad · 14. 11. 2021

I ran into a problem in #Yara when using the cuckoo module. No matter what cuckoo report I used, it gave me an ERROR_COULD_NOT_MAP_FILE error. I found the source of the problem and created a PR to fix the issue:

VirusTotal/yara

#1590 **Fix loading module**

0 comments 1 review 1 file +2 −0

regeciovad • November 13, 2021 1 commit

Avast

There is more...

Avast Threat Labs @AvastThreatLabs · 4. 11. 2021

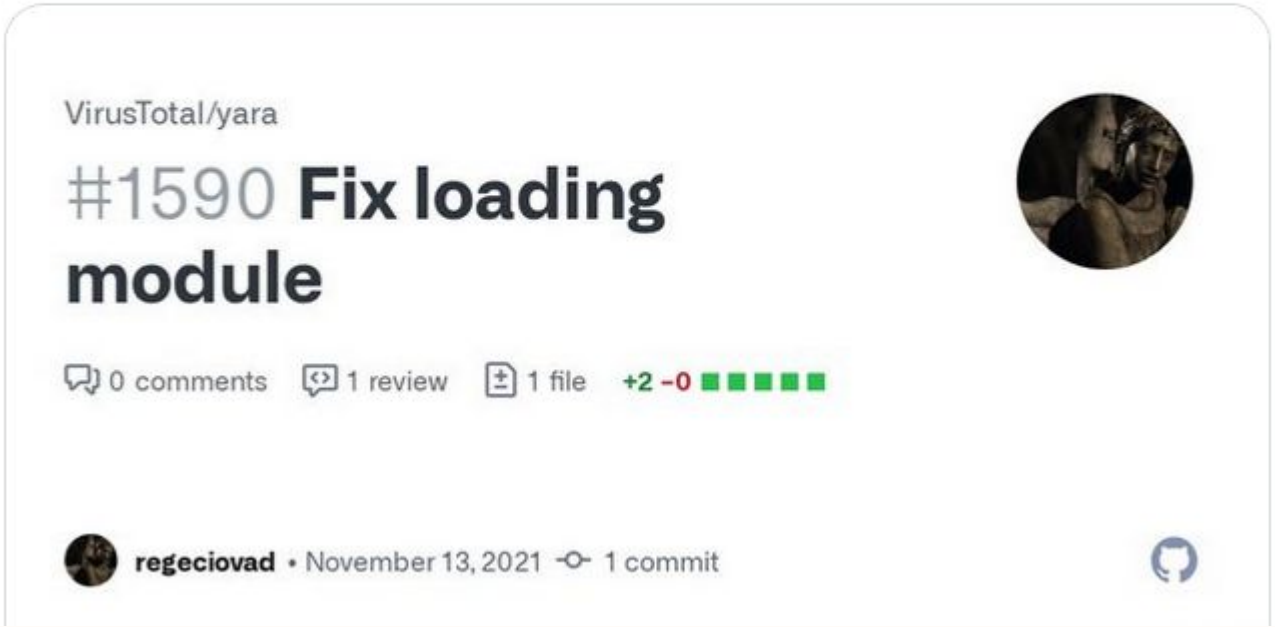
And last but not least in this batch of improvements, @regeciovad improved heuristic for atoms with repeating bytes resulting in faster matching with sequences containing repeated bytes (github.com/VirusTotal/yar...). As @plusvic said, awesome job Dominika!

> **Victor M. Alvarez** @plusvic · 14. 9. 2021
> @regeciovad has done an awesome job! twitter.com/cyb3rops/statu...

1 7

Dominika Regéciová @regeciovac

I ran into a problem in #Yara when what cuckoo report I used, it gave me an ERROR_COULD_NOT_MAP_FILE error. I found the source of the problem and created a PR to fix the issue:



Avast

There is more...

Avast Threat Labs @AvastThreatLabs · 4. 11. 2021

And last but not least in this batch of improvements, @regeciovad improved heuristic for atoms with repeating bytes resulting in faster matching with sequences containing repeated bytes (github.com/VirusTotal/yar...). As @plusvic said, awesome job Dominika!

Victor M. Alvarez @plusvic · 14. 9. 2021

Retweetnuli jste

Avast Threat Labs @AvastThreatLabs · 4. 11. 2021

#YARA is a tool (but also a language and even more) helping malware researchers to identify and classify malware samples (virustotal.github.io/yara/). We benefit from YARA at Avast, but we also give back to the community. Here you can find some of our recent contributions (🧵👇)

1 58 129

Zobrazit toto vlákno

Dominika Regé

I ran into a prob
what cuckoo re
error. I found th

VirusTotal/yar

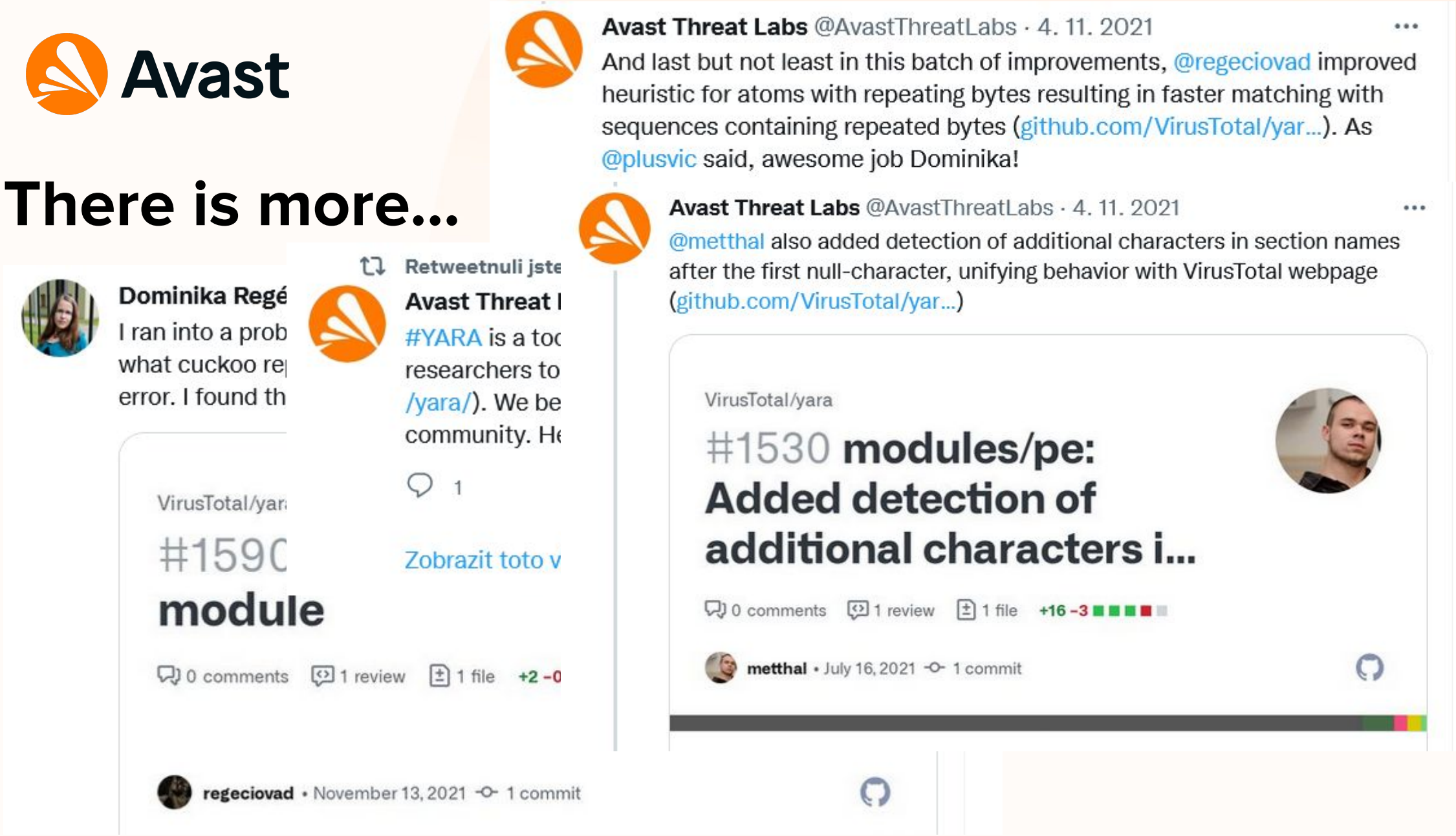
#159C

module

0 comments 1 review 1 file +2 -0

regeciovad • November 13, 2021 1 commit

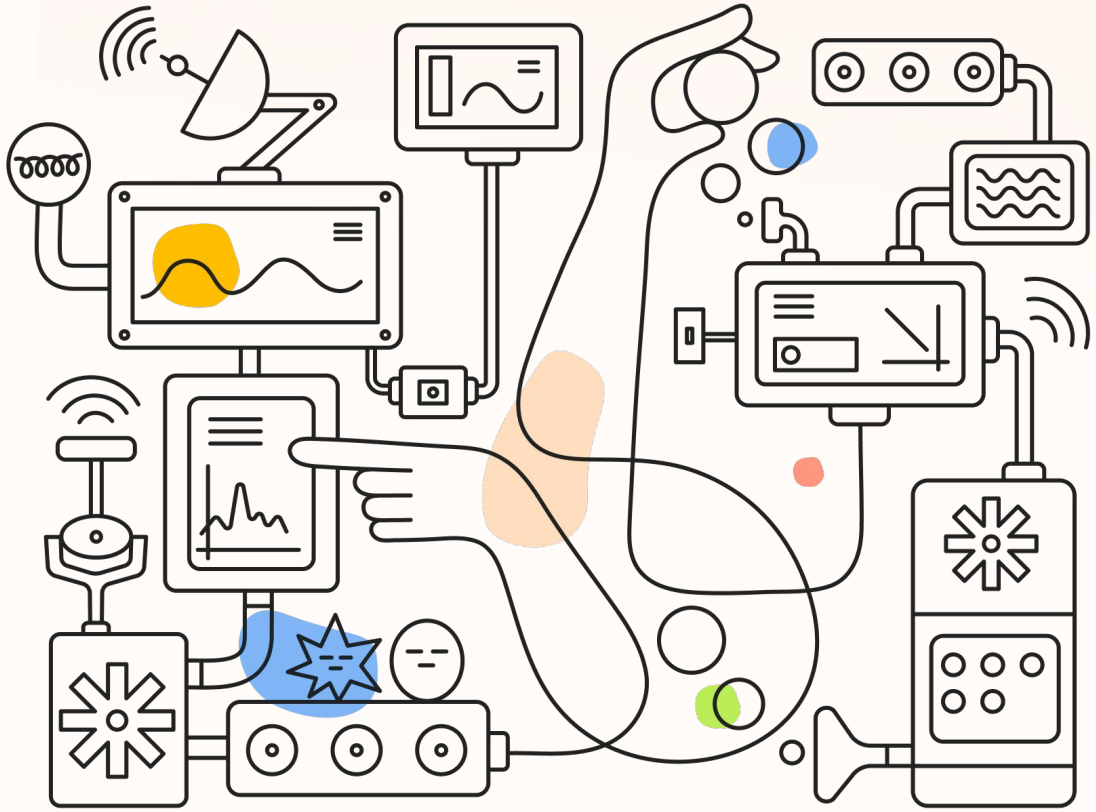
There is more...



More resources

- Paper Pattern Matching in Yara: Improved Aho-Corasick Algorithm
- Changes in Yara: PR (will be updated soon, I promise 😇)

Conclusion



Conclusion

- Yara is an amazing tool not only for malware analysis
- There is still space for improvements
- Spoilers for the next changes:
  - Behavioral analysis
  - Automated generation of Yara rules
  - Cuckoo module
- For more, follow me on Twitter and LinkedIn: regeciovad

Thank you!