When entropy meets Shannon

splunk.com/en_us/blog/tips-and-tricks/when-entropy-meets-shannon.html

April 21, 2016

By Splunk April 21, 2016

This is the third post on URL analysis, please have a look at the two other posts for more context about what can be done with Splunk to analyze URLs:

- Splunking 1 million URLs
- Hunting that evil typosquatter

You will find in this article information on how one can detect DNS tunnels. While you can find lots of very useful apps on Splunkbase to help you analyze DNS data, it is always good for curious individuals to discover some techniques being used underneath.

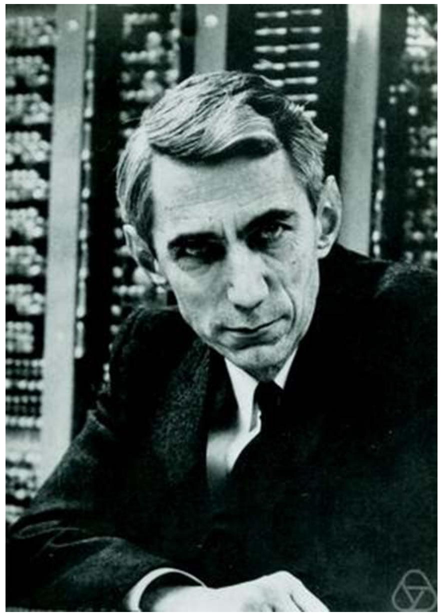
A lot of captive portals are bypassed everyday by anyone able to run a DNS request, if someone can run on their machine the following command:

```
$ host splunk.com
splunk.com has address 54.69.58.243
...
```

Without being authenticated on the captive portal, then they can use any service on the internet using a DNS tunnel. There are a lot of tools out there to create those tunnels. And for a great paper on the topic, I encourage you to read the Detecting DNS Tunneling from

SANS Institute.

Claude Shannon to the rescue!



Claude Shannon
By Jacobs, Konrad - https://opc.mfo.de/detail?photo_id=3807, CC BY-SA 2.0 de, Link

Long time ago, the venerable Claude E. Shannon wrote the paper "A Mathematical Theory of Communication", which I strongly encourage to read for its clarity and amazing source of information.

He invented a great algorithm known as the Shannon Entropy which is useful to discover the statistical structure of a word or message.

If you consider a word, being a discrete source of the finite number of characters type which can be considered, for each possible character there will be a set of probabilities which would produce various outputs. There will be an entropy for each character. This entropy on the chosen word is defined as the average of the output weighted on the probability of occurrence of the characters.

The previous paragraph can easily be translated into the following Python code (taken from the excellent URL Toolbox on Splunkbase:

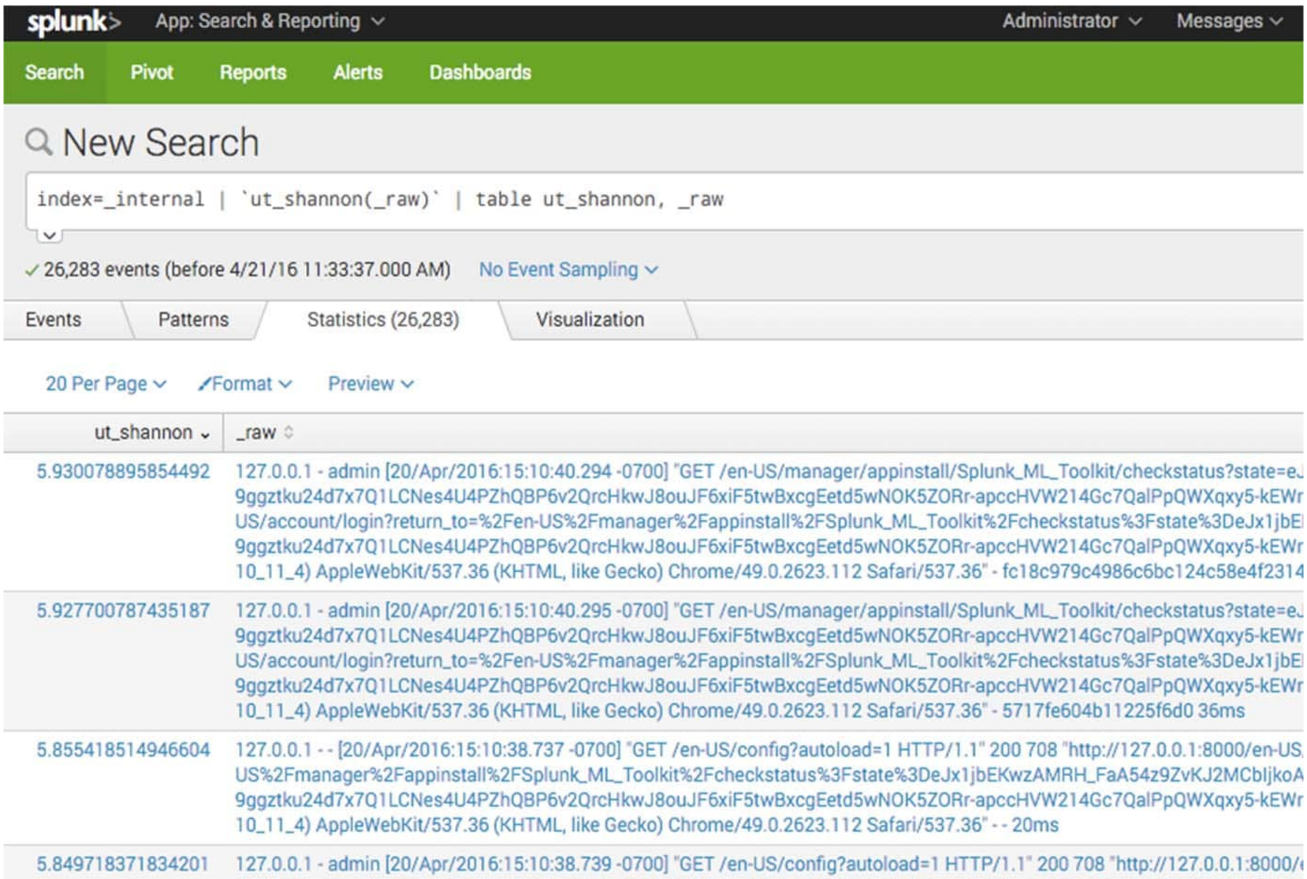
```
def shannon(word):
    entropy = 0.0
    length = len(word)
```

```
occ = {}
for c in word :
    if not c in occ:
        occ[ c ] = 0
    occ += 1

for (k,v) in occ.iteritems():
    p = float( v ) / float(length)
    entropy -= p * math.log(p, 2) # Log base 2

return entropy
```

Which can be run directly from any word you can have in Splunk:



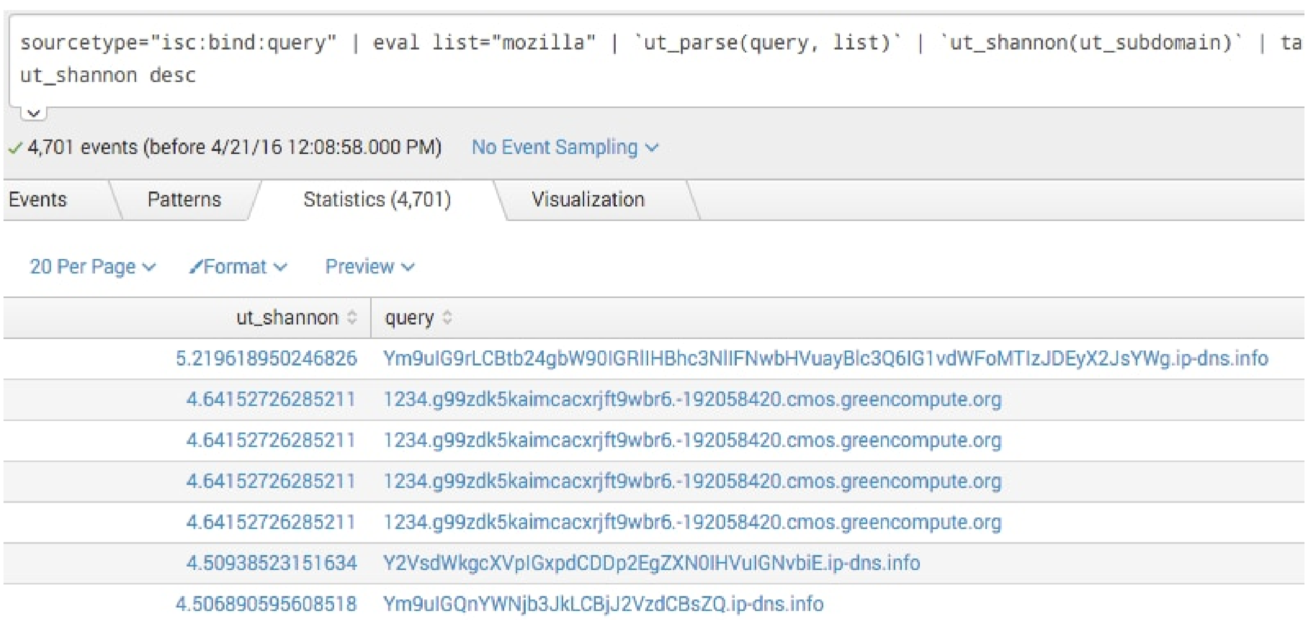
As you can see, the score is pretty high, which makes sense since there is a high variety of frequency over those data. If we click on the ut_shannon field to sort in reverse order, this is what you could get:

ut_shannon ^	_raw
2.0	open
2.0	open
2.0 2.0	Done
2.0 2.0	Done
3.5465935642949384	VERSION=6.4.0
3.5511911744656968	splunkd is not running.
3.551191174465697	splunkd is not running.
3.6835423624332306	BUILD=f2c836328108

As one can see, words of low characters distribution get a low score.

Catching DNS tunnels from subdomains in URLs

```
If we run the following query, interesting results are shown:
sourcetype="isc:bind:query" | eval list="mozilla" | `ut_parse(query, list)` |
`ut_shannon(ut_subdomain)` | table ut_shannon, query | sort ut_shannon desc
```



ut_shannon	query
5.219618950246826	Ym9uIG9rLCBtb24gbW90IGRlIHBhc3NlIFNwbHVuayBlc3Q6IG1vdWFoMTIzJDEyX2JsYWg.ip-dns.info
4.64152726285211	1234.g99zdk5kaimcacxrjft9wbr6.-192058420.cmos.greencompute.org
4.64152726285211	1234.g99zdk5kaimcacxrjft9wbr6.-192058420.cmos.greencompute.org
4.64152726285211	1234.g99zdk5kaimcacxrjft9wbr6.-192058420.cmos.greencompute.org
4.64152726285211	1234.g99zdk5kaimcacxrjft9wbr6.-192058420.cmos.greencompute.org
4.50938523151634	Y2VsdWkgcXVpIGxpdCDDp2EgZXN0IHVuIGNvbiE.ip-dns.info
4.506890595608518	Ym9uIGQnYWNjb3JkLCBjJ2VzdCBsZQ.ip-dns.info

As you can see in the results here, the high score come from tunnels made to the domain ip-dns.info as well as something which is unknown but could also be a tunnel: traffic towards greencompute.org

I hope this post helps you to see tools and methodologies one can use to find out unusual activity strictly based on the DNS traffic. More to come…

Thanks!
Sebastien Tricaud

Posted by

Splunk
