

The Masked SYNger: Investigating a Traffic Phenomenon

blog.rapid7.com/2020/05/28/the-masked-synger-investigating-a-traffic-phenomenon/

matthew berninger

May 28, 2020

Last updated at Wed, 16 Dec 2020 17:23:45 GMT

At the beginning of 2020, Rapid7 and other researchers began noticing increased scanning activity against a variety of TCP ports. Through our daily monitoring of connections to our Heisenberg honeynet, as well as discussions with other community members such as Andrew Morris of GreyNoise, we felt confident that we were seeing something new—certainly not part of our “normal” traffic to the honeypots. The first public mention of this activity was actually on Jan. 3, when [@Andrew__Morris](#) tweeted:

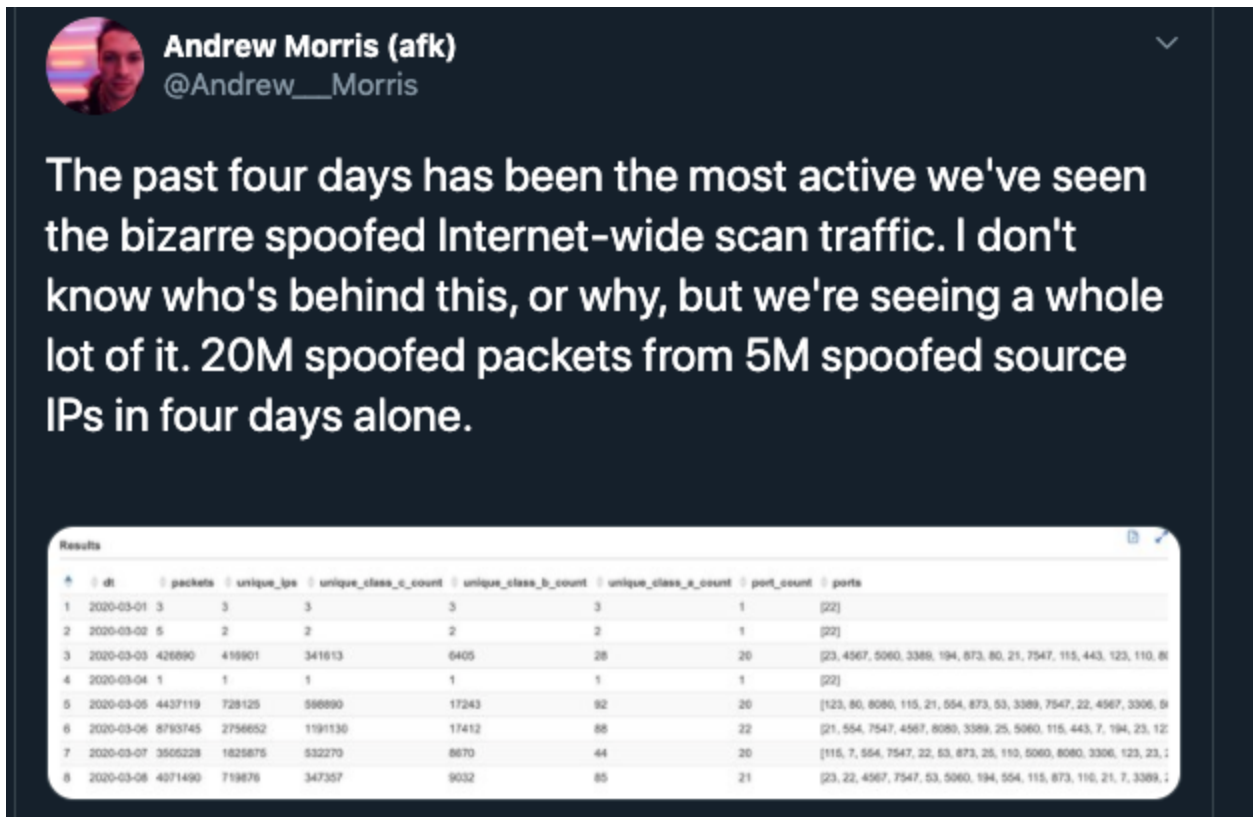
Andrew Morris (afk)
@Andrew__Morris

Wew okay, so there's some REALLY bizarre internet-wide scan traffic happening right now in [@GreyNoiseIO](#) . it started two/three hours ago, quieted down for a few minutes, then just started up again. I'm going to do my best to stream of conscious what's going on while we dig in:

Anomalies		74123	▲ 682%
Increases over the monthly average			
01/03/2020		36772	▲ 1175%
Port / Protocol	Unique IPs		
23 / TCP	79820 ▲ 489%	34871	▲ 3006%
25 / TCP	62976 ▲ 2882%	70007	▲ 924%
47721 / TCP	20 ▲ 432%		
21 / TCP	67387 ▲ 2814%	36313	▲ 1211%
53 / TCP	20960 ▲ 2724%	37787	▲ 2937%
5060 / TCP	38672 ▲ 2853%		
22 / TCP	76101 ▲ 683%	68391	▲ 1834%
36081 / TCP	26 ▲ 463%		
7 / TCP	38208 ▲ 1373%	38303	▲ 2347%

1:10 PM · Jan 3, 2020 · [Twitter Web App](#)

Over the following weeks, we noticed strange behavior on some ports like TCP 123, but had not noticed a continued phenomenon until March. On March 5, Greynoise again confirmed that they were also seeing the same sustained behavior we had seen in January and February.



The tweet text reads: "The past four days has been the most active we've seen the bizarre spoofed Internet-wide scan traffic. I don't know who's behind this, or why, but we're seeing a whole lot of it. 20M spoofed packets from 5M spoofed source IPs in four days alone."

The table below shows the results of network scans over an 8-day period in March 2020. The columns represent date, packets, unique IP addresses, and various scan counts for different ports.

#	dt	packets	unique_ip	unique_class_c_count	unique_class_b_count	unique_class_a_count	port_count	ports
1	2020-03-01	3	3	3	3	1	1	{22}
2	2020-03-02	5	2	2	2	1	1	{22}
3	2020-03-03	426890	419901	341813	6405	28	20	{23, 4567, 5060, 3389, 194, 873, 80, 21, 7547, 115, 443, 123, 110, 8}
4	2020-03-04	1	1	1	1	1	1	{22}
5	2020-03-05	4437119	728125	590890	17243	92	20	{123, 80, 8080, 115, 21, 554, 873, 53, 3389, 7547, 22, 4567, 3306, 9}
6	2020-03-06	8793745	2756652	1191130	17412	88	22	{21, 554, 7547, 4567, 8080, 3389, 25, 5060, 115, 443, 7, 194, 23, 12}
7	2020-03-07	3505228	1825875	532270	8670	44	20	{115, 7, 554, 7547, 22, 53, 873, 25, 110, 5060, 8080, 3306, 123, 23, 1}
8	2020-03-08	4071490	719876	347357	9032	85	21	{23, 22, 4567, 7547, 53, 5060, 194, 554, 115, 873, 110, 21, 7, 3389, 1}

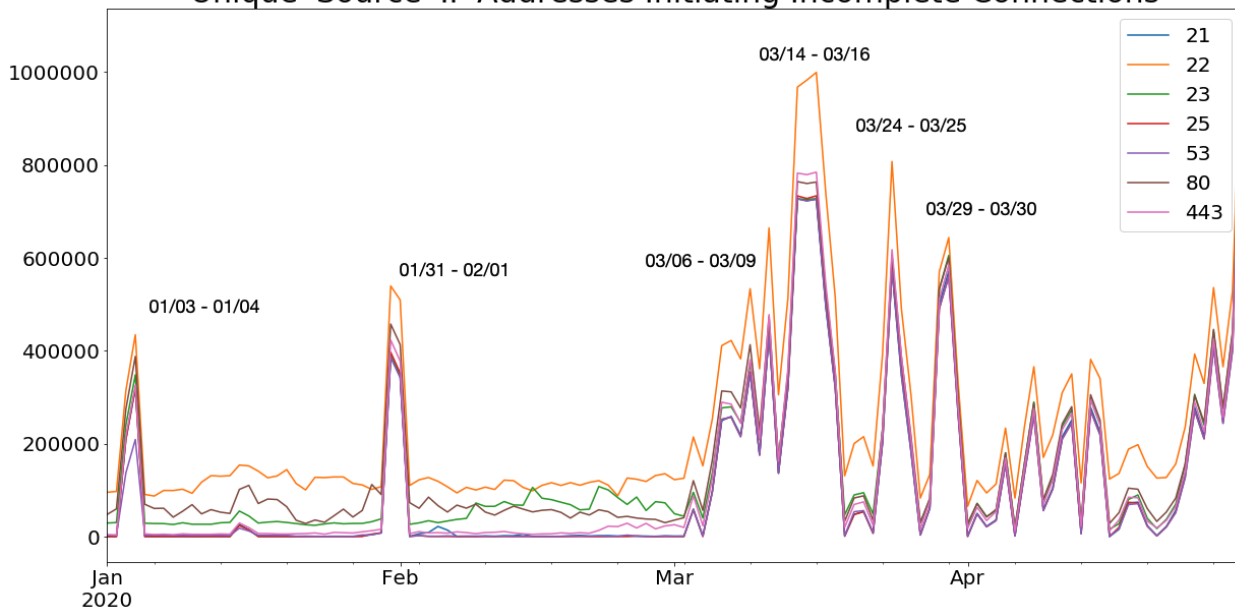
Once March hit, we were seeing consistently high numbers of unique IP addresses connecting to our honeypots. Primarily, we observed incredibly high volumes of SYN scanning activity against these TCP ports:

- 21 (FTP)
- 22 (SSH)
- 23 (Telnet)
- 25 (SMTP)
- 53 (DNS, more commonly seen on UDP [RFC5966](#))
- 80 (HTTP)
- 110 (POP3)
- 123 (NTP, more commonly seen on UDP [RFC5905](#))
- 443 (HTTPS)

The graph below shows the pattern of activity, namely the sharp spikes and sustained increase in the number of distinct IP addresses seen scanning these ports every day.

Unique 'Source' IP Addresses Initiating Incomplete Connections

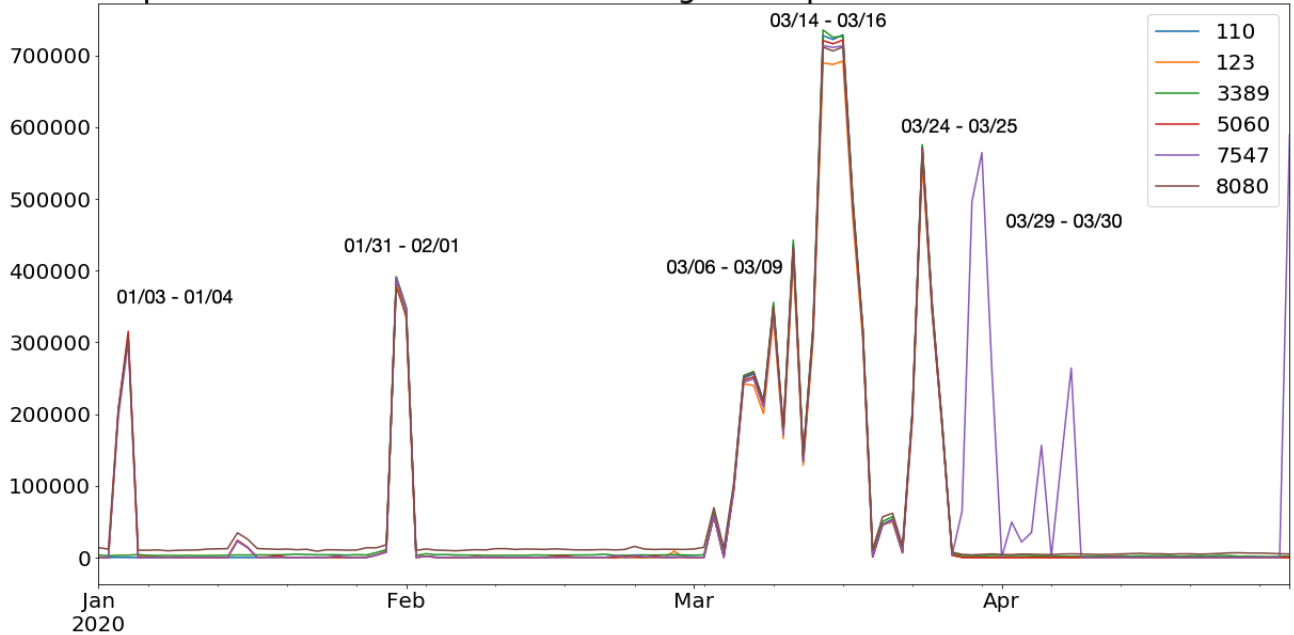
04/30-05/02



We have also seen the same activity against these ports, which later dropped off in April.

- 110 (POP3)
- 123 (NTP)
- 3389 (RDP)
- 5060 (SIP)
- 7547 (TR-069)
- 8080 (HTTP-alt)

Unique 'Source' IP Addresses Initiating Incomplete Connections - Other Ports



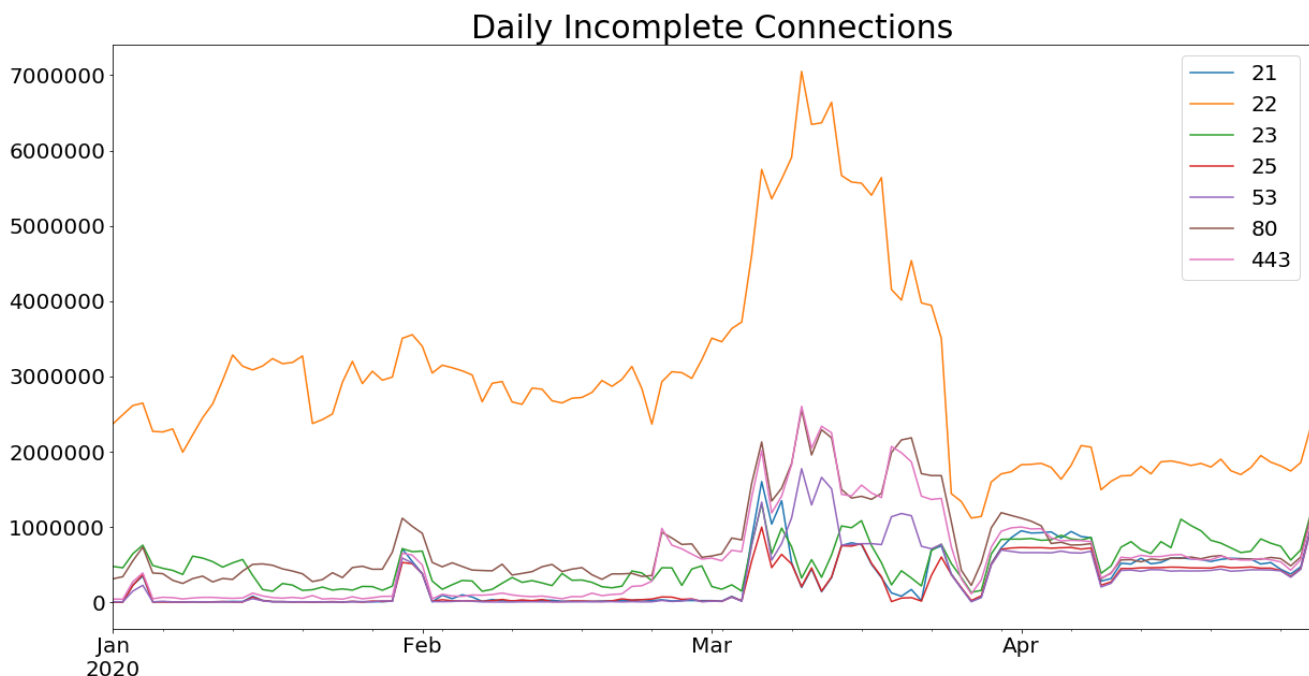
We suspect that other ports have been included on certain days, but they have not been as consistently uniform in their source IP volume. Some of these are ports you might expect, such as 3389 or 1433, but also included less common ports such as 5060, 111, and 17.

It is important to note the sheer numbers here. While honeypot traffic analysis is no stranger to large numbers, the daily scale of unique IPs for this specific traffic is in the hundreds of thousands. The IPs seen are mostly different every one or two days, so it is not even the same million or so source IPs involved the whole time. All told, we estimate that the activity we have seen between January and May 2020 involves more than **100 million** unique IP addresses (insert Dr. Evil meme). For comparison, the largest botnets in history are estimated to be in the range of 1 million to 10 million hosts.

This is one of the reasons we believe this activity to be spoofed—not in fact actually coming from these supposed sources. If this activity truly represented nodes under control by a single entity, the sheer number of IP addresses involved would be the largest botnet ever created by several orders of magnitude.

Also of note is that while the total number of connections is similarly staggering and generally follows the same pattern, it does not follow the source IP pattern *exactly*. Nor is it uniform between ports the same way distinct IPs were. This likely means two things:

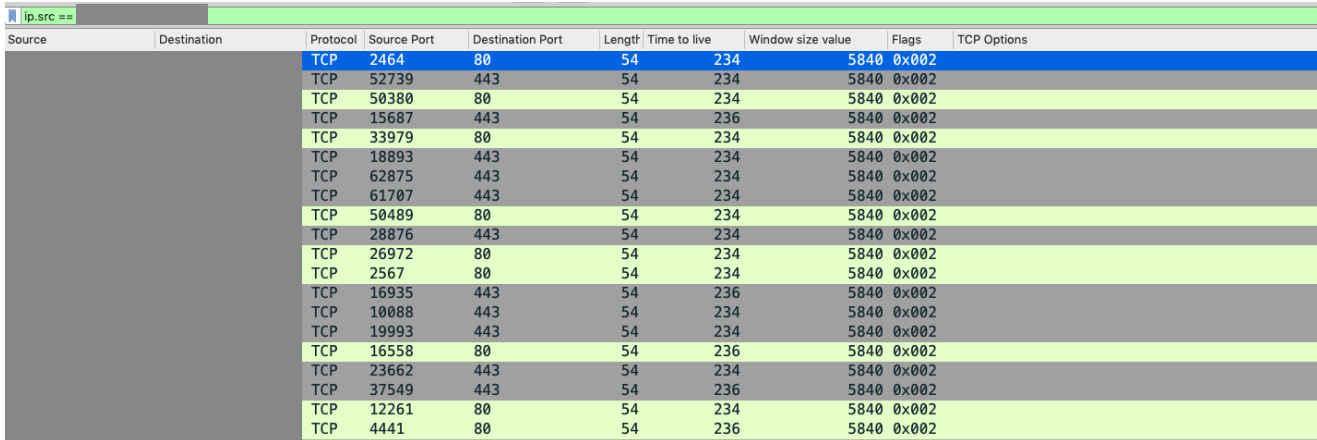
1. Some spoofed IPs “send” more traffic than others (much more on this later). If each spoofed IP sent the same number of packets, connections would simply be a multiple of the number of distinct IPs.
2. Packet numbers are not evenly distributed among ports, even among spoofed IPs. Otherwise, the lines below would be more parallel to one another.



So, in summary, we have a historic number of spoofed “sender” IP addresses sending a historic amount of traffic on common ports. Also of note is that these are all TCP ports, including ports 53 and 123 where UDP would be a more common protocol. The next step was to look at the traffic and see if these connections had anything in common.

What makes this traffic unique?

Furthering our theory of spoofing, in every case we have examined, across multiple ports, these scans never establish a full TCP handshake, typically sending only SYN packets. An example is below, filtered on just the spoofed “source” IP, which we have redacted.



Source	Destination	Protocol	Source Port	Destination Port	Length	Time to live	Window size value	Flags	TCP Options
		TCP	2464	80	54	234	5840	0x002	
		TCP	52739	443	54	234	5840	0x002	
		TCP	50380	80	54	234	5840	0x002	
		TCP	15687	443	54	236	5840	0x002	
		TCP	33979	80	54	234	5840	0x002	
		TCP	18893	443	54	234	5840	0x002	
		TCP	62875	443	54	234	5840	0x002	
		TCP	61707	443	54	234	5840	0x002	
		TCP	50489	80	54	234	5840	0x002	
		TCP	28876	443	54	234	5840	0x002	
		TCP	26972	80	54	234	5840	0x002	
		TCP	2567	80	54	234	5840	0x002	
		TCP	16935	443	54	236	5840	0x002	
		TCP	10088	443	54	234	5840	0x002	
		TCP	19993	443	54	234	5840	0x002	
		TCP	16558	80	54	236	5840	0x002	
		TCP	23662	443	54	234	5840	0x002	
		TCP	37549	443	54	236	5840	0x002	
		TCP	12261	80	54	234	5840	0x002	
		TCP	4441	80	54	236	5840	0x002	

In our dataset, we generally find these connections have the following characteristics:

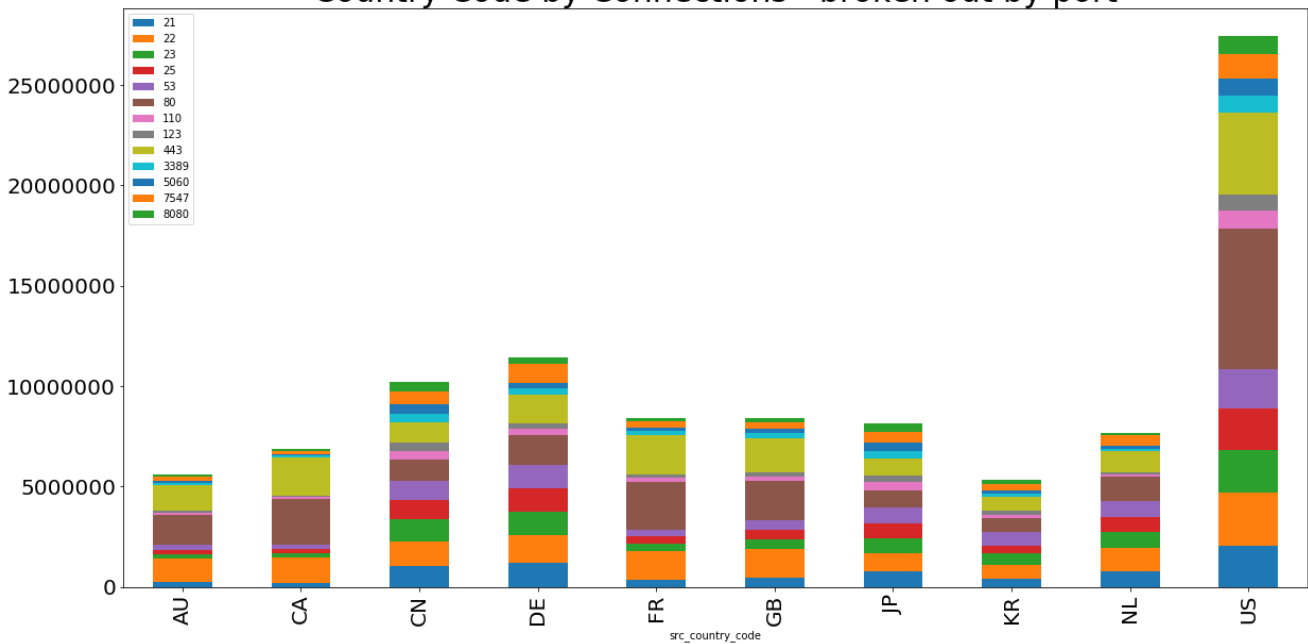
- Frame Length is always 54.
- TTL values are variable and sometimes do not make much sense. They’re either very low, or very high, and do not seem to fit typical TTL values for normal operating systems. Packets with the exact same source and same destinations, sent milliseconds apart, will very often have varying TTL values.
- Notably, the “TCP Options” field is empty. This is uncharacteristic of normal TCP traffic, and is a strong indicator of synthetic traffic.
- Window size is generally 5840 or 29200.

There are other strange idiosyncrasies with the traffic. For example, SYN packets will often reuse source port numbers, sometimes using non-ethereal source ports like 21 or 22. There are some rare instances where the packets contain SYN and PSH flags, or the RST flag. However, a handshake is never completed. It is, of course, possible that some of these idiosyncrasies are from unrelated activity, but it is difficult to tell, given the scale of the activity, exactly which IPs are involved and which are not. After all, this very confusion may in fact be the purpose of the activity.

Where is this pretending to come from?

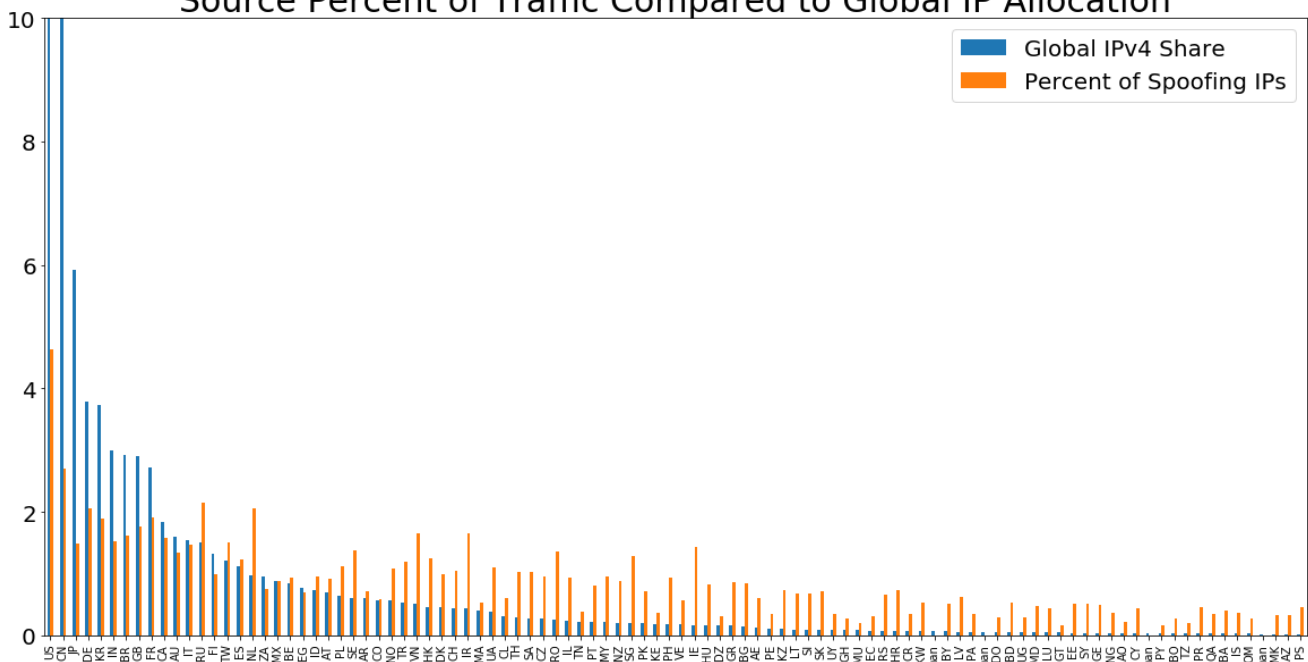
Grouping by Country Code (while also realizing the sources were certainly spoofed), we get these results for the top 10 “source” countries:

Country Code by Connections - broken out by port



These results are not very surprising, since many of these nations contain large hosting infrastructure and would be expected in most lists of top internet traffic. Perhaps the actors here are simply sampling spoofed IPs from the public internet. If they were, then the distribution of source IPs in this traffic would roughly mirror global IP allocation. However, the distribution of the top 100 source country codes seen in spoofed traffic does not match global IP allocation.

Source Percent of Traffic Compared to Global IP Allocation



Whereas the Global IPv4 allocation is highly concentrated on the left—the U.S. is allocated roughly 34% (bar was cut off for scale)—the spoofing activity is disproportionately high among “smaller” players. This indicates that the purported “sources” of the spoofed traffic likely do not follow a random selection from IPv4 address allocation—rather, there is something more manual behind which source IPs are used.

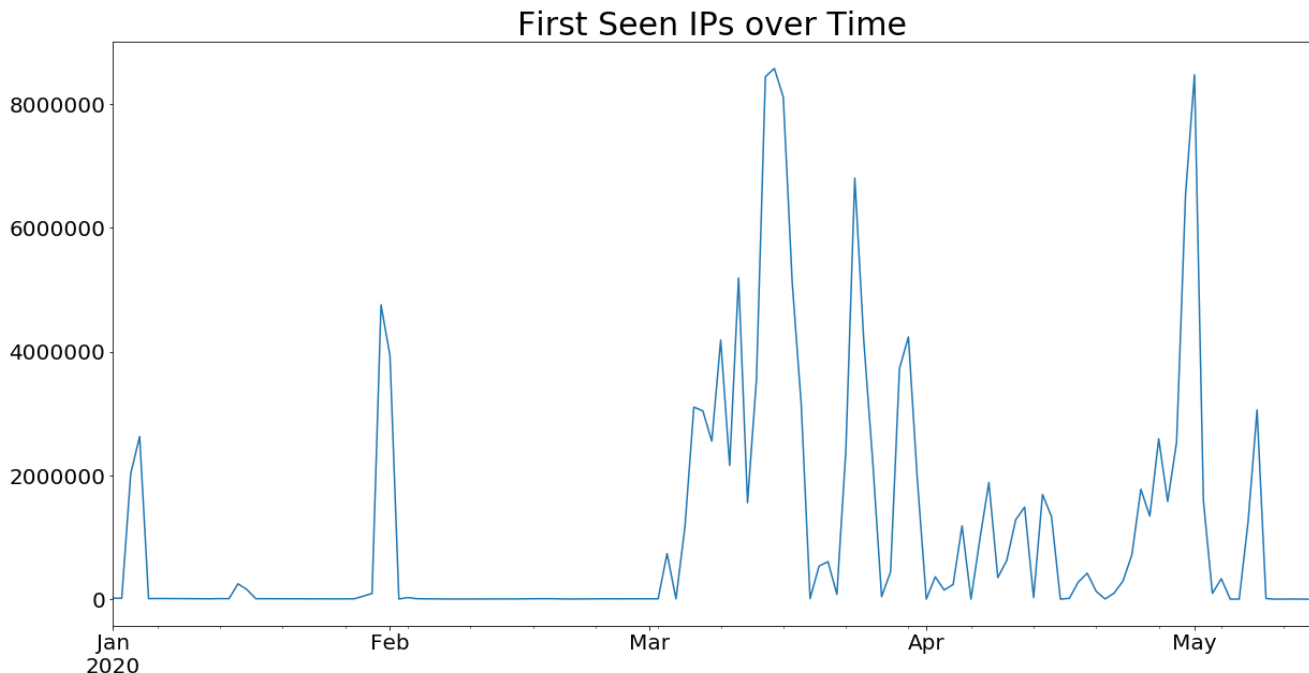
“Sender” IP analysis

As mentioned earlier, the number of packets sent varies wildly by source IP. Many source IPs send one and only one packet, while others send hundreds of thousands. Below, we will examine the breakdown of source IPs and various ways we are identifying anomalies.

We pulled every IP address that had appeared hitting our honeypots on these ports, then built features for each IP in the activity. These features included:

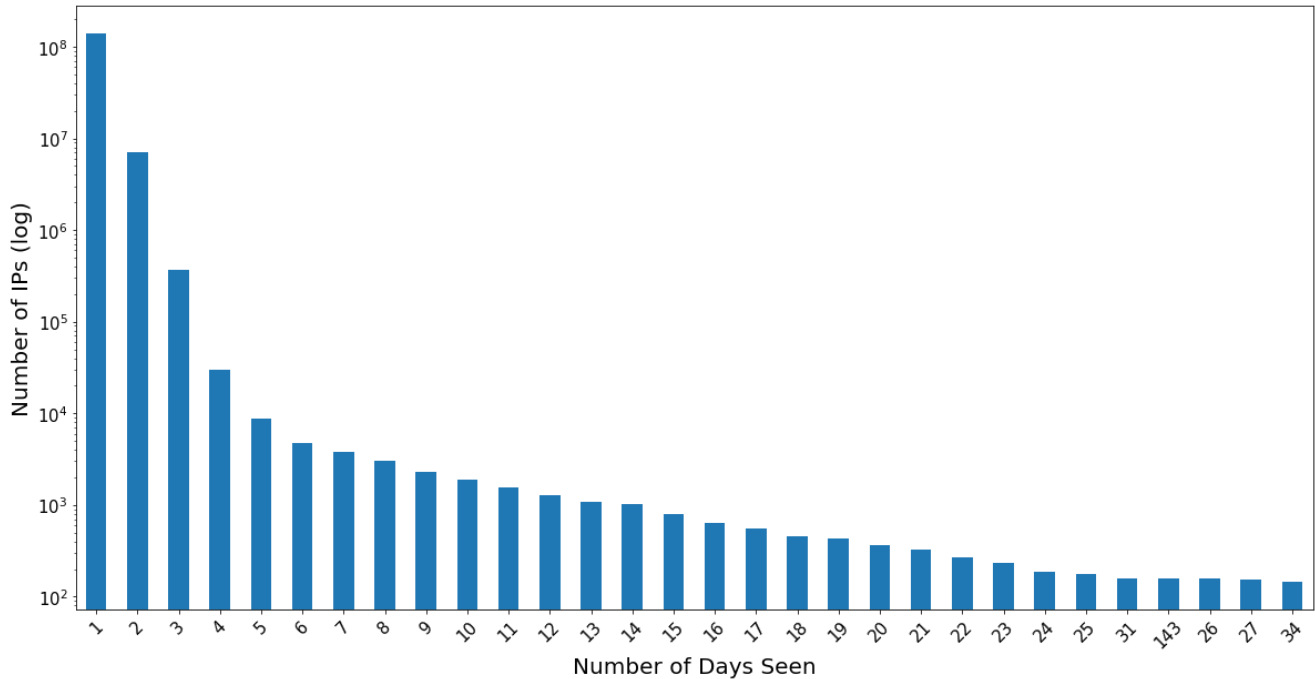
- **First Seen:** First seen from January 2020 until now
- **Last Seen:** Within the same period
- **Duration:** Delta in days between First and Last Seen
- **Number of Days:** The number of distinct days during that time, when the IP appeared as a source of spoofed SYN scans to our honeynet.
- **Connections:** Total number of connections from the IP to our honeynet over 2020 so far.

The first thing we wanted to look for was an idea of turnover.

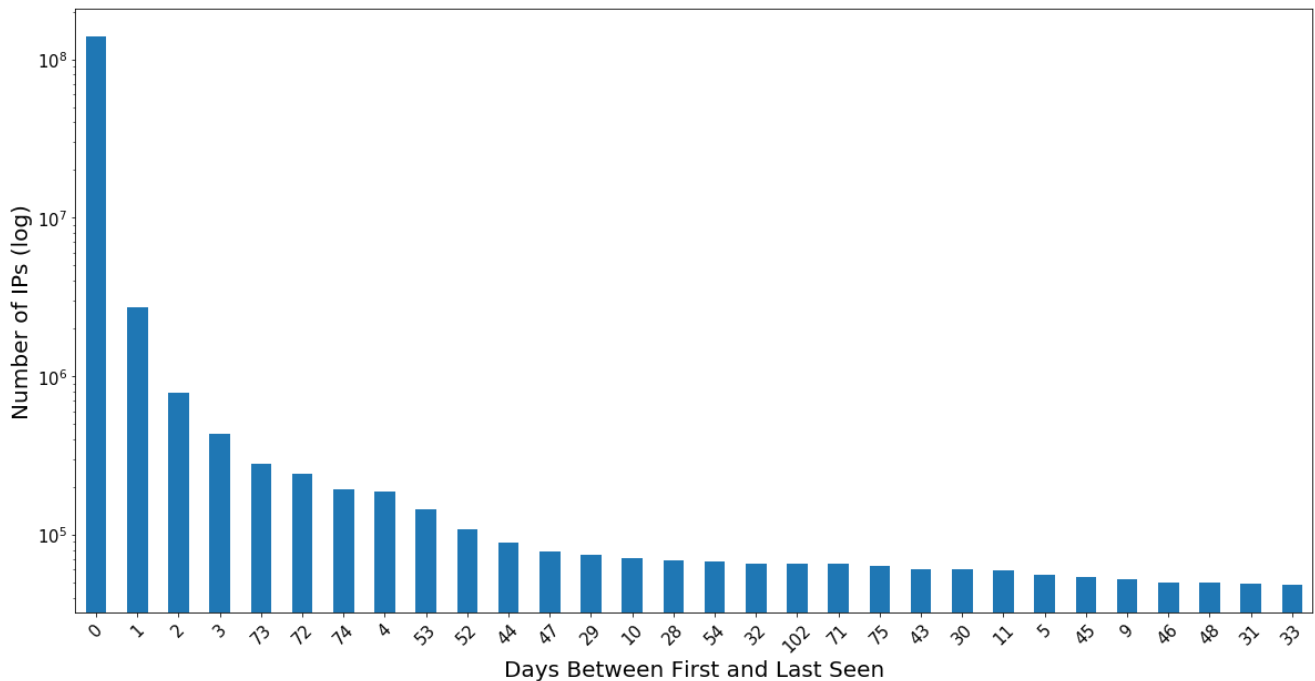


The above graph shows the timing of when IPs first appear in the traffic. As you can see, much of the traffic on any given day is performed by IPs participating for the first time.

Additionally, we can look at the total number of days that an IP is used in this activity.



Note that the y axis is on a log scale, so the **vast** majority of spoofed source IPs appear only one or two days total. Let's also look at the Duration—the total days elapsed between First Seen and Last Seen:



As you can see, not surprisingly, the vast majority of IPs are only seen for one day, with a net duration of zero.

However, it is interesting that there are still a fair number of IPs in the long tails here—those that appear on multiple days, or whose appearances are many days apart, or both.

You may have also noticed an interesting anomaly in the decreasing durations:

duration	count
0	139185530
1	2710925
2	785571
3	434538
73	280230
72	240738
74	192048
4	187374
53	143929
52	108490

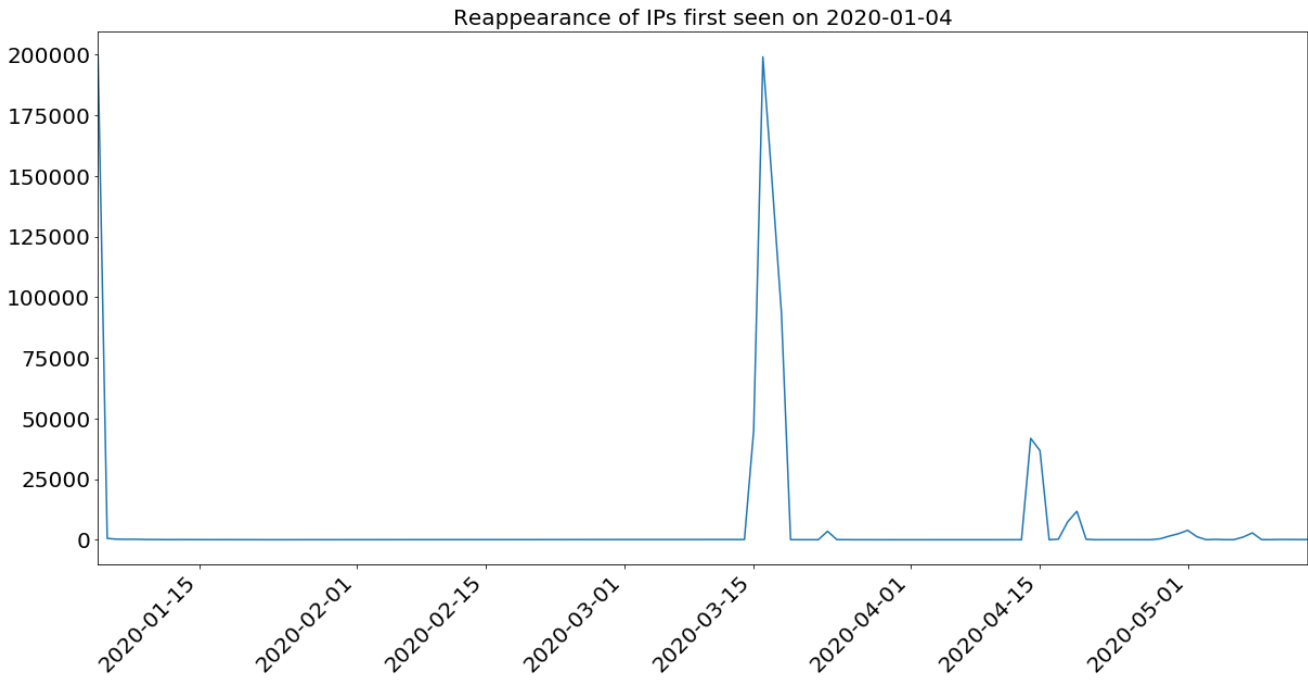
Where we would expect the “day interval” to simply count up from 0 (only appeared one day) to 1,2,3, etc., we see this cluster of IPs for which the Duration is between 72–74 days. Further down, we see a grouping of 52 and 53 day intervals. When do these IPs first appear and last appear?

Once we filtered on that data, we got consistent appearance of a couple very specific dates:

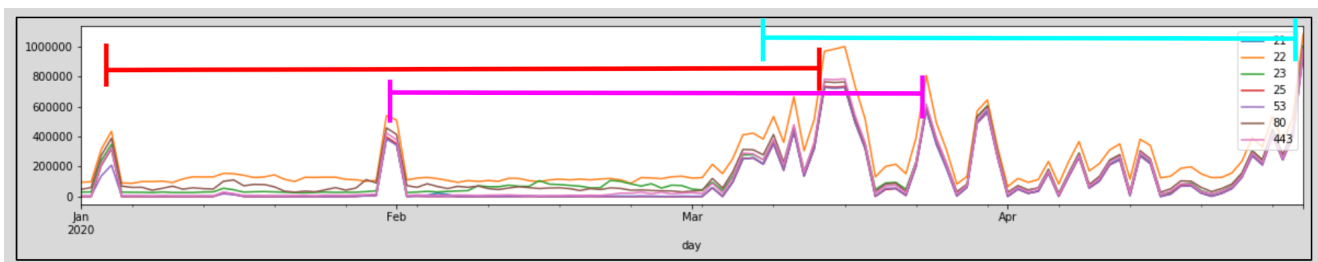
```
df3['day_list'].value_counts()[ :20]
```

```
[20200104, 20200316]      194591
[20200104, 20200317]      131349
[20200103, 20200316]      105561
[20200104, 20200318]       78510
[20200103, 20200317]       70952
[20200103, 20200315]       24043
[20200103, 20200104, 20200316]  23839
[20200103, 20200104, 20200317]  15950
[20200104, 20200316, 20200317]  12651
[20200131, 20200412]       10741
```

This indicates a very specific interval of appearance for the IPs seen on Jan. 3 and Jan. 4. Even though there is another spike of activity at the end of January, we do not see those IPs then. In fact we don’t see them again until until March 15.



Repeating this analysis a couple of times gives us a few intervals of reappearance:



Jan. 3, 4 → March 15,16,17

Jan 3,4 → April 14, 15

Jan. 31, Feb. 1 → March 24 and 25

March 15,16,17 → April 30 to May 2nd

This further indicates manual operation of some extent—this activity is neither purely random nor evenly distributed. While we cannot prove intent or motivation for these intervals, they are not the same length, nor do they involve the same number of addresses each time. There is some reuse for each large spike, with decreasing overlaps over time. This likely indicates some degree of human tinkering on the back end.

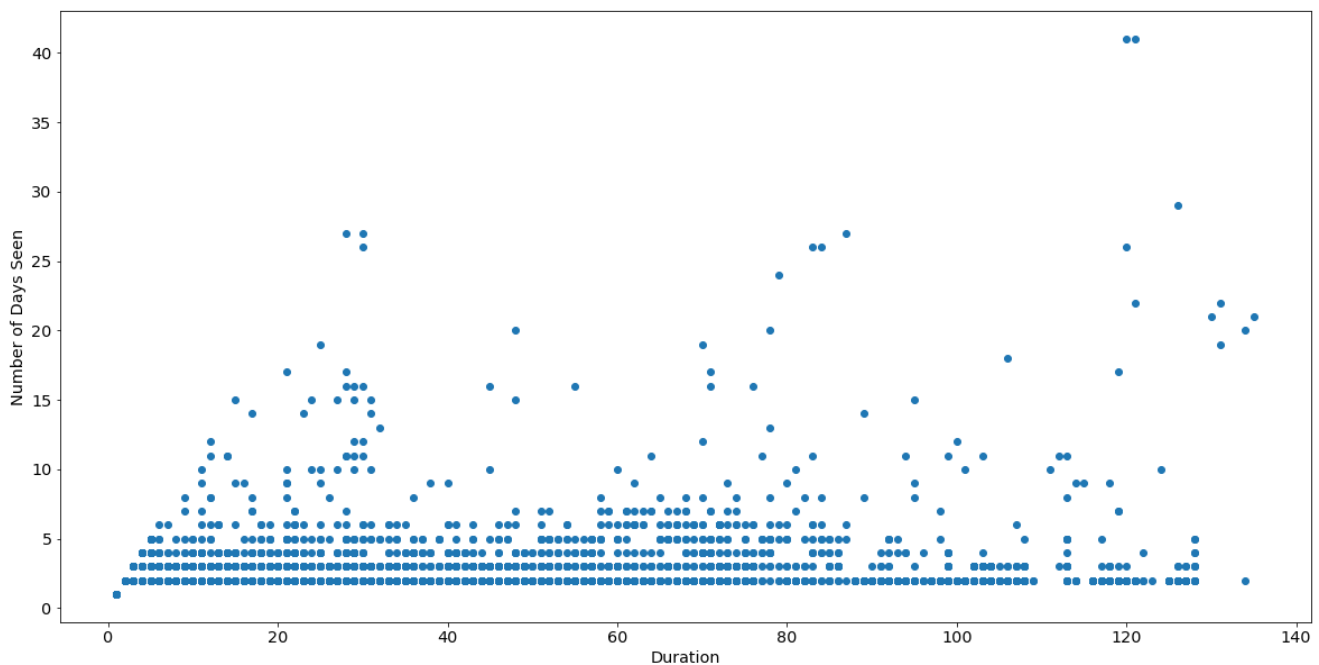
The signal and the noise

Finally, while **most** of the activity involves IP addresses used for one or two days, there are a small minority of IPs issuing huge numbers of requests, and appear on many days—but, importantly, not **all** days. It is possible that these IPs serve a specific purpose or that the traffic they send is in some ways different in nature than the rest of the traffic.

For this analysis, we focused only on traffic sent to TCP port 25 (typically used for SMTP), which resulted in the 'S1' conn_state, and took place between (and including) the months of January and May 2020. Focusing on a single port removed noise and reduced the dataset to something that could more easily be examined and iterated on. The same analysis should be done for other common ports seen in this traffic.

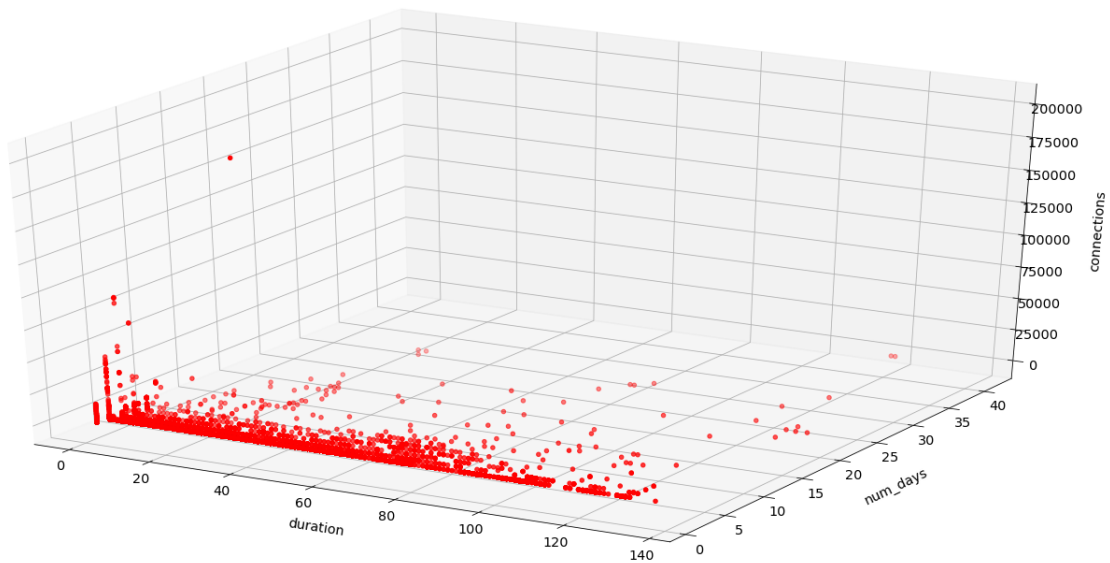
For every IP address in the data, we built a profile, including first seen, last seen, number of days total, the 'duration' between the first and last date, and other stats like 'day density' (the fraction of the duration taken by days seen). For example, an IP seen on only two days, but those days are a month apart, would have a much lower 'density' than an IP seen on two consecutive days.

Here is the breakdown by duration and number of days:



As you can see, the vast majority of IPs exist near the bottom, each a small number (1–2) of days total. However, some of the IPs with 2–5 days have long durations—that is, the few days they appear are very spread out. Lastly, there are very few IPs with a large number of days total, over a long period of time. Let's factor in total number of connections as well (apologies for my 3D graphing skills, I'm no [@hrbrmstr](#)):

Here, we can definitely spot one outlying spoofed source IP, the red dot near the top.



Using this approach, we are building shorter lists of interesting IP addresses to perform deeper analysis to find differences in the content. While we have been able to spot these anomalies, thus far, traffic from these “top talkers” has not substantially differed from the other packet profiles noted above. It is of course possible that the anomalies we see are strange artifacts of routing, misconfiguration, or some other factor unrelated to intent.

So what? Speculation and wildly unfounded theories

Frustratingly, we cannot point to an existing threat that we can say with any likelihood is responsible for this activity. Additionally, given our available evidence, this does not appear to currently pose a threat to organizations. Most organizations with basic network security will simply block this type of traffic. Furthermore, the traffic is not concentrated enough on any single destination to be considered a large-scale DoS attack. So, why is it happening? Here, we can only offer speculative theories. Caution: these theories can provide new investigative directions, but in the end, only consistent evidence matters.

Cover for collection

Given the likely spoofed nature of the traffic, the originating party would only receive response data if they were in a position to collect the responses. Therefore, an actor would need to have some other listening capability to collect this spoofed traffic for scanning information. It is possible that some of the traffic is not spoofed and could hide in the noise, such that the actor would receive the responses from traffic sent by their actual infrastructure. So, either the actor has the ability to intercept responses, or this traffic could be providing a cover for actual scanning activity. However, this theory has holes. While May 2020 does seem an auspicious time to gain information about rapidly changing internet exposure, *many* other data sources already exist that could provide this information. Furthermore, why draw attention to SYN scanning by doing so much of it?

Poisoning threat intel

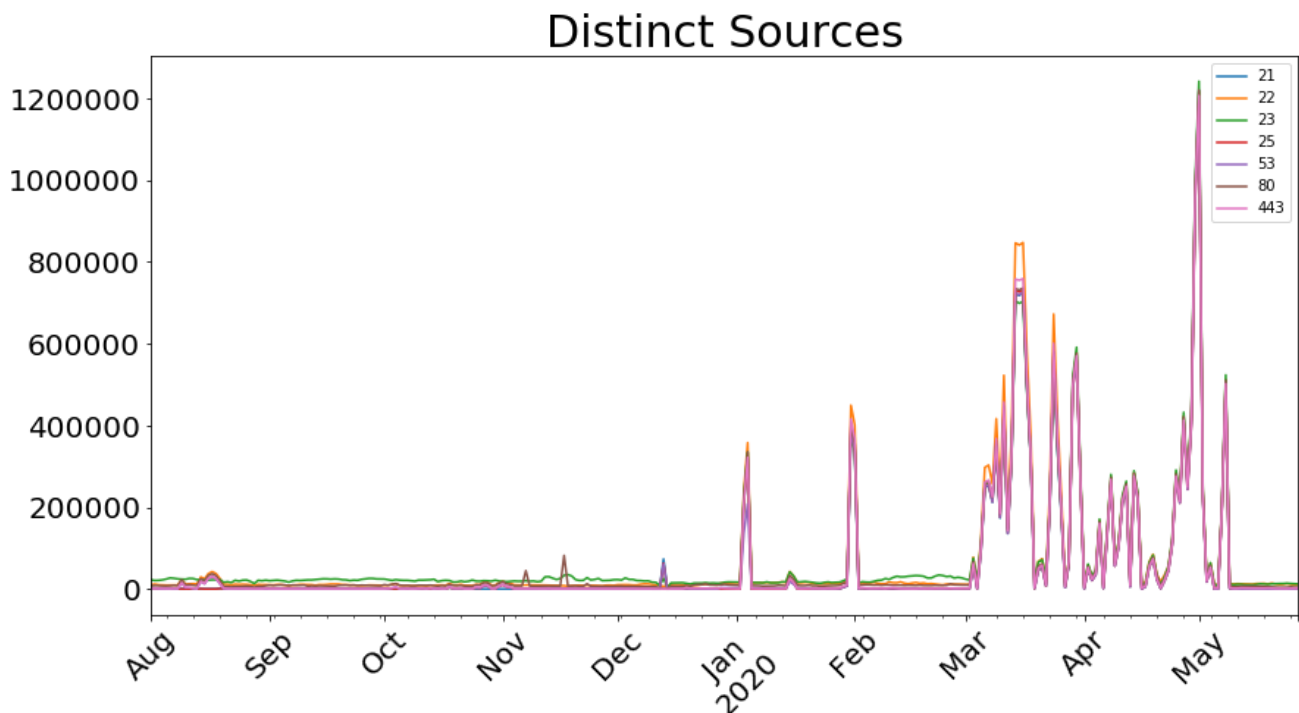
It is possible that this could be an effort to “poison” automated threat intel feeds by suddenly inundating them with millions of IP addresses purportedly performing scanning. In investigating some of these spoofed sources, we have seen them appear on recently updated blacklists and threat intel feeds. This is possible, but the scale of the activity is large enough to be unnecessary for this goal.

Testing

Much of what we have outlined above looks like potential testing. An initial spike occurs in January, followed by a lull in activity and then another “test” at the end of January. Later, sustained traffic occurs, with small tweaks appearing here and there. An actor with interest in deploying TCP spoofing for more destructive purposes could simply be testing their capability. The broad spread of distinct spoofed IPs does not create denial-of-service (DoS) conditions now, but that could change. It is possible that this is someone testing or demonstrating capability.

Recent activity and next steps

On April 30, May 1, and May 2 2020, we saw the highest levels of connections from this activity, followed by a sharp drop-off. Activity spiked again around May 8, but it has disappeared since then. Now that we have some rough signatures for what the traffic looks like, we are able to detect these spikes as they occur. Additionally, while it is impossible to confirm, we have seen some traffic which looks similar to this dating as far back as 2017, though not nearly at the levels seen this year.



Finally, a caveat: Our visibility through Heisenberg is by definition limited to our honeypots. While other organizations have corroborated these trends, we do not have a large amount of details on traffic being sent elsewhere.

The above theories are illustrative of what *might* be the intention here, but we do not currently have solid evidence to support a most probable theory, or eliminate others. Additionally, at the time of writing, we have no evidence to confirm that this presents an exigent threat. The simple fact is we do not know *why* this is happening, only that we are seeing it, it is new and strange, and other organizations have corroborated these trends. Given the unprecedented scope and volume, we felt it was worth publishing our research to begin a discussion among researchers, and hopefully understand this better.

Many thanks to Andrew Morris of GreyNoise and others in the community for providing information and feedback on this research.



Never miss a blog

Get the latest stories, expertise, and news about security today.