CERT.PL >_

Network traffic periodicity analysis of dark IP address space

Authors: Wojciech Sypniewski, Piotr Białczak

About CERT Polska

CERT Polska is a part of NASK (Research and Academic Computer Network) - a research institute that also operates the .pl TLD registry and offers a range of advanced telecom services. CERT Polska is the first CSIRT established in Poland. Since its inception in 1996, thanks to a dynamic activity in the CERT/CSIRT community, the team has become a recognized and experienced player in the field of computer security. The core activities provided by CERT Polska include incident handling and cooperation with similar units around the world, both in operational fields and research and development. CERT Polska is a member of various international forums and working groups, including FIRST (since 1998), TF-CSIRT (since 2000) and APWG (since 2010). In 2005, CERT Polska started a forum for the Polish abuse incident handling teams - Abuse Forum.

Contents

1. Introduction

Computer networks are subjects of many different kinds of threats, which, if not properly handled, result in lowering the availability and reliability of the services run within the network. In extreme cases, servers can get compromised and stop responding. Malicious actors are coming with more and more sophisticated methods of entering protected resources, therefore the defending side must also define better ways of detecting such threats.

Many of such attacks can be observed in the dark address space of the Internet – the unused IPv4 addresses. As no network services are present on these addresses, packets should not be sent there. However in practice various traffic types can be observed in those dark spaces, beginning with the misconfiguration of some client software (e.g. typos in IP address of hardware configuration), through automated port scanners looking for vulnerabilities, to echoes of Denial of Service (DoS) attacks. Such activity usually is performed without direct human interaction, therefore the frequency of sent or received packets is often regular, in contrary to the users' actions, which are hardly periodic.

Based on this assumption, the main goal of this report is to demonstrate that we can find periodicities in network traffic by splitting it into tuples containing IP address, port and protocol combinations and analyzing their frequency characteristics using Discrete Fourier Transform. We borrow such approach from the field of Digital Signal Processing. Periodicities found in the network traffic can be a good starting point for further analysis of anomalies and support real-time monitoring of the Internet.

Analysis of network traffic directed to the unused address space has been performed by CAIDA in project called Network Telescope.[1] The dark space is monitored mainly in order to recognize echoes of Denial of Service attacks. Other researchers looked at different aspect of network traffic analysis – its periodicity.[2,3] The results confirmed that Discrete Fourier Transform can be used as a method of traffic analysis and provides accurate data for further examination. However, within some of the publications rather high frequencies were analyzed (hundreds of Hz, that is for packets sent at short intervals under 1s) and on very short time spans (couple of seconds), as aim was to find Local Area Network short term periodicities, like ARP requests.[4]

After identifying the drawbacks of these approaches, we have decided to examine captured packets over long time spans (minutes, hours) to find network activity series, with periods defined in seconds or minutes rather than milliseconds.

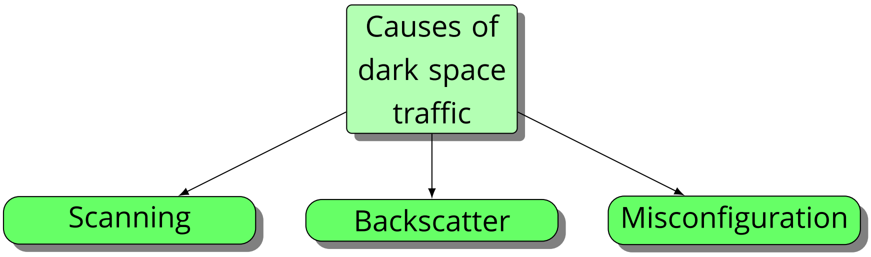
[1] https://www.caida.org/projects/network_telescope/

[2] *Using Signal Processing to Analyze Wireless Data Traffic*, C. Partridge et al., http://dl.acm.org/citation.cfm?id=570689

[3] *Filtering of Shrew DDoS Attacks in Frequency Domain*, Y. Chen et al., http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1550964

[4] *Identifying short-term periodicities in Internet traffic*, I. Grondman, B.Sc. thesis, 2006

Internet dark address space traffic is usually classified into three types according to its root cause: scanning, backscatter and misconfiguration.[5]



The scans seen by network telescopes are performed mostly by automatic systems, which scan big parts of the Internet in search for services or vulnerable hosts. Scans can also be conducted by hosts infected with Internet worms, which search for other vulnerable machines to spread the malware.
The backscatter traffic consists of reflections of DoS attacks. In this kind of attack, the malicious actor (usually group of machines) is sending big number of packets to the same machine and the same port in order to disrupt the service. In many cases the source IP addresses are spoofed by specifying different origin and sometimes addresses from dark space are selected. In such situation network telescopes can give good perspective and input data for analysis of this type of attacks.
The last type of dark address space traffic is caused by misconfiguration of software or hardware. Due to errors in provided IP addresses of services many clients send packets to non-existent servers, thus creating Internet background noise. Such behavior can be seen in NTP date server synchronization queries or in IPSec IKE dead peer search.

[5] *Internet background radiation revisited*, E. Wustrow et al., http://dl.acm.org/citation.cfm?id=1879149

The analysis was performed on data collected on an edge router with unallocated IPv4 addresses. As those addresses are not being used by any party, in principle there should be no traffic coming to this destination. However in reality the incoming traffic is relatively big. Our analysis used three datasets varying by the capture time length and the date they were captured on. A summary of these datasets is presented in Table 1.

Dataset date	**2011**	**May 2016**	**May 2016**
Number of packets	200 000	14 235 000	273 937 315
Time length (minutes)	8	84	1440
Unique source IP addresses	10 000	1 000 00	14 300 000
Unique destination IP addresses	133 000	257 000	257 000

Table 1: Summary of datasets used in analysis

In Table 2. we presented more detailed information about the longest dataset. Nearly half of the packets contained UDP datagrams. Also about third part of the traffic involved port 1194, which is assigned by IANA for OpenVPN. Also about 15% of all periodicities (which we define as regular activities observed for particular protocol, port and IP addresses detected by our system) occur in 3 second intervals, with precision of 1 second. Please take not, that presented measures are independent and their percentages do not sum up to 100%.

Measure	**Value**	**Percentage of all periodicities[%]**
Periodicity (1 s precision)	3s	15
Port	1194	31
Protocol	UDP	52

Table 2: Selected statistics of the longest analyzed dataset

4. Method of analysis

The flow of performed analysis is presented in fig. 1

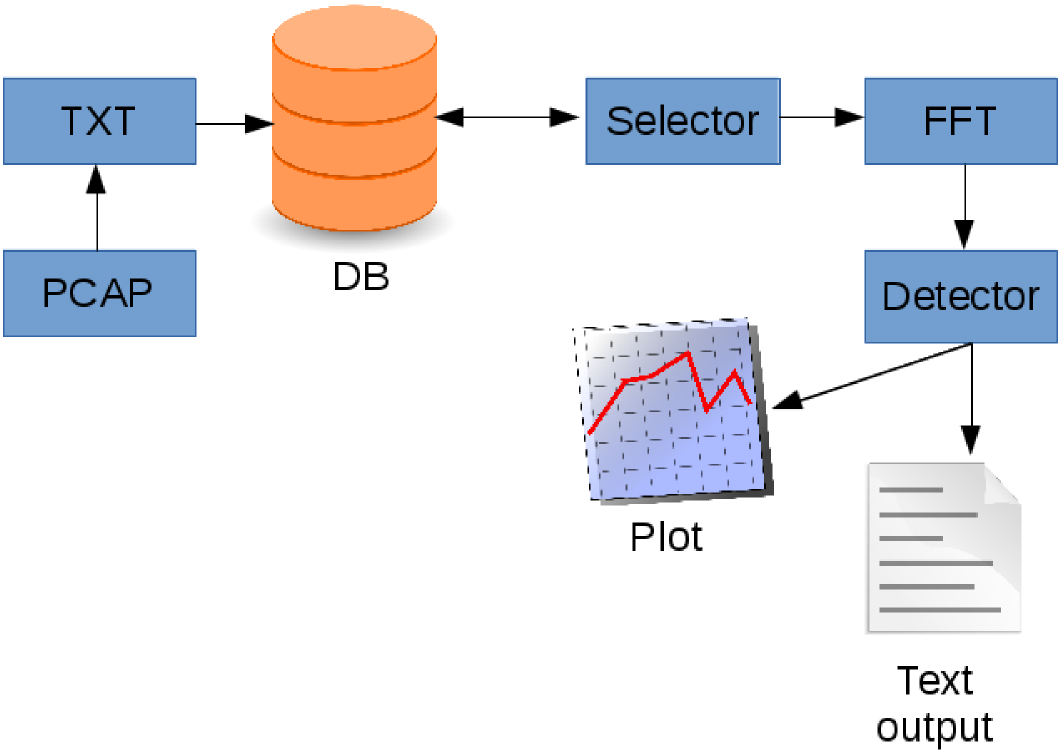


Figure 1: Analysis flow scheme

Below steps present the analysis process:

1. Transform the data from PCAP files into text files (keeping only IP addresses, ports, protocol and timestamp, not taking headers or payload into analysis)
2. Filter data for analysis (ICMP , UDP and TCP-SYN)
3. Insert traffic flow data into SQL database
4. Select first 10000 most active IP addresses (ones with most packets sent) with port and protocol. As we keep source and destination information, we can try up to four different combinations:
   - 4.1 Source IP, destination port, protocol
   - 4.2 Source IP, source port, protocol
   - 4.3 Destination IP, destination port, protocol
   - 4.4 Destination IP, source port, protocol
5. Analyze previously selected combinations by using DFT algorithm
   - 5.1 Sample the flows (count number of packet occurrences in 1 second time frame)

5.2 Use threshold to find peaks in the Discrete Fourier Transform plot and decide if the result is adequate

   i. Find frequency with the highest value (represented by the highest peak in the plot)
   ii. Mark its value as a reference point
   iii. Reduce value of reference point by 10% and check whether less than 5 other peaks exceeds this value. If not repeat this point.
   iv. If less than 5 such points exist, return them and consider plot as containing periodicities

5.3 Save potentially interesting frequency plots and text files containing information about peak frequencies in those outputs

4.1. Discrete Fourier Transform

If we put it very simply, Discrete Fourier Transform represents a signal in time domain into the frequencies which it is composed of. The result of this transform can be presented as a graph with peaks representing particular frequencies, on which signal was stronger. Otherwise, the output plot will be rather flat and with some distortions, but without any distinctive peaks on it. This means that there is no significant periodicity detected within such data. Based on the result of the DFT function, the analysis is performed in order to decide which of the output presents periodicity and which does not. The algorithm checks how many peaks can be found above defined threshold and one step below it. If no points are found step below the threshold, and only small number above, it is being marked as potentially periodic and the output plot together with its peaks frequencies are stored for further examination. In our analysis we considered number of consequent packets as the input data (packet timestamps were sampled in 1 second intervals). An example of a DFT plot from analyzed dataset is presented in fig. 2.

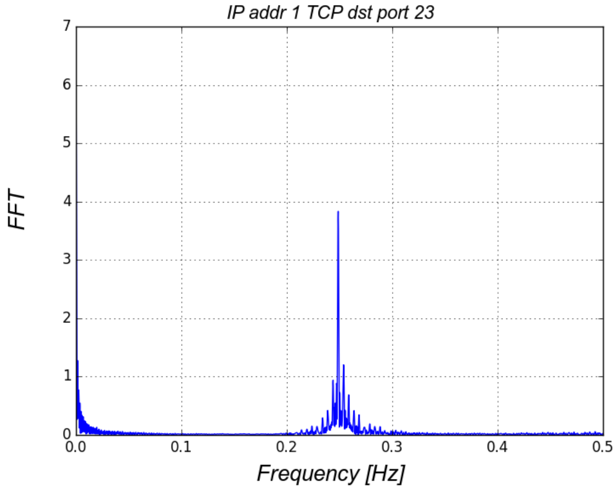


Figure 2: An example of a DFT plot. Most common characteristics found for port 23 in medium length dataset

5.1. Short time span dataset, year 2011

First dataset analyzed by us was eight minute long and it contained 200 thousand packets of network traffic from 2011 (as presented in Table 1). We used it mainly for code optimization and testing techniques needed for efficient parsing and automating process of periodicity finding. The time span of captured network traffic was relatively short, therefore we could only identify high frequencies, with the lowest ones approximately between 0,025 Hz and 0,015 Hz (40 and 65 seconds respectively). Example plots are presented in figures 3 – 5.

Most common protocol presenting periodic character was ICMP, what was most likely caused by misconfiguration of service and after-effects of DoS attacks. An example of such behavior is presented in fig. 4.

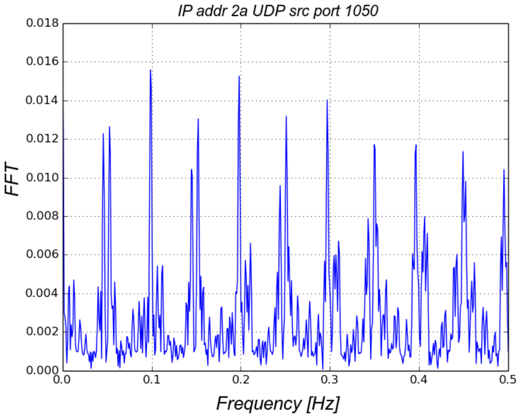


Figure 3: Short time dataset, 20 seconds interval (0,05Hz)

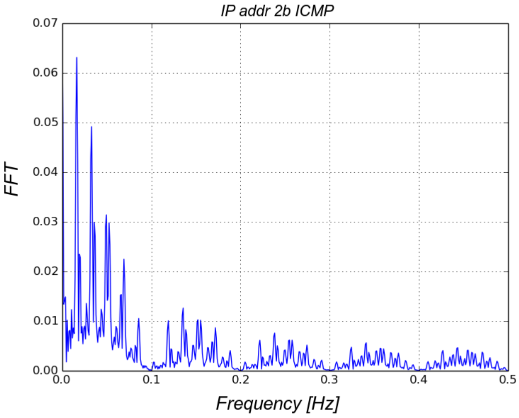


Figure 4: Short time dataset, 40 seconds interval (0,025Hz)

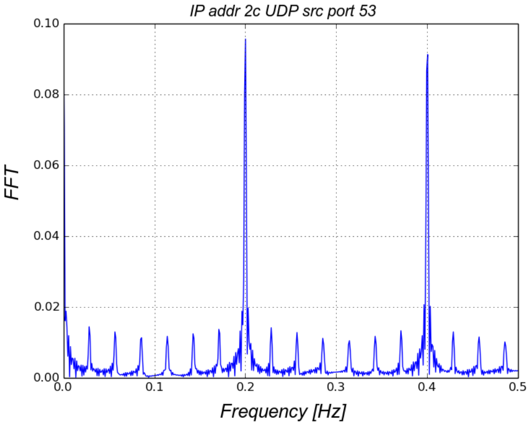


Figure 5: Short time dataset, reflection of a DoS DNS Attack

One of the most interesting anomalies we have detected was traffic coming from Google DNS servers (8.8.8.8 and 8.8.8.4). Closer analysis with Wireshark indicated that the servers tried to query Chinese NTP time server on cn.pool.ntp.org, but the domain name resolved to IP address of the dark address space observed by us.

5.2. Medium time span dataset, May 2016

Second data set contained more than 14 million packets captured within 84 minute period. It was analyzed primarily to check whether our method is fast enough to process longer and bigger network traces. Also we used it as a cross-check of anomalies found in the first and third dataset. Example plots are presented in figures 6 – 8.
The most significant difference in frequency characteristics comparing to the smaller dataset is the clearer shape of data plots, thus providing an easier way to identify periodicities, and a more precise way of specifying corresponding frequency. It can be seen by comparing respective plots in section 5.1 and here.

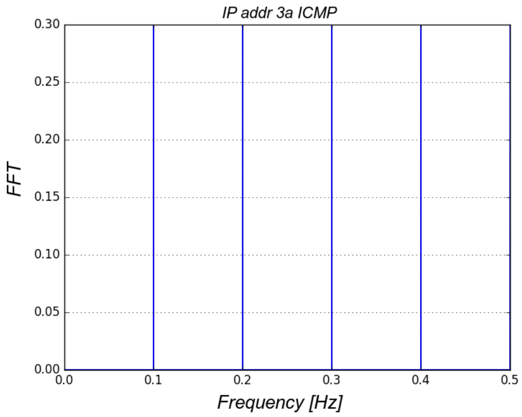


Figure 6: Medium time dataset, 10 seconds interval ICMP ping requests

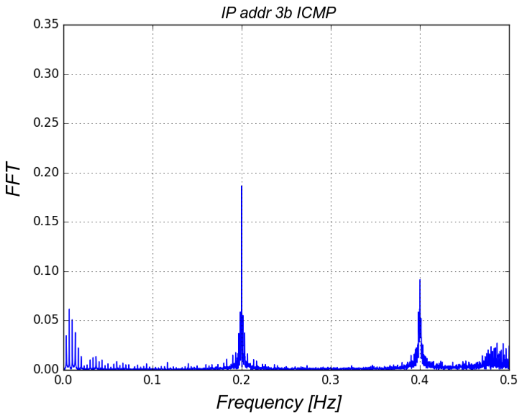


Figure 7: Medium time dataset, 20 seconds interval ICMP ping requests

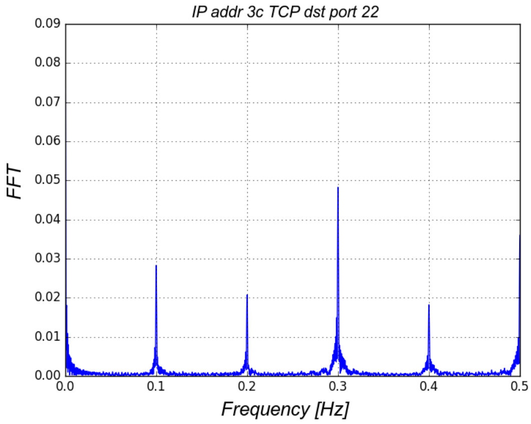


Figure 8: Medium time dataset, 10 seconds interval TCP port 22 activity

Additionally port 23, which is used default by Telnet service, is the most common port proven to have periodic network characteristic. DFT plot for an example IP address for this port is presented in fig. 2.

5.3. Long time span dataset, May 2016

The third dataset was obtained by monitoring backscatter traffic for 24 hours in May 2016. It was the longest set and as such gave the best perspective on probable periodicities. One of the reasons is strictly connected with the Fast Fourier Transform, which takes array of sampled data as an input parameter. In practice proportionally bigger datasets are having higher number of occurrences in a single sample and as a result more data is fed into Discrete Fourier Transform, which produce more precise graph output.

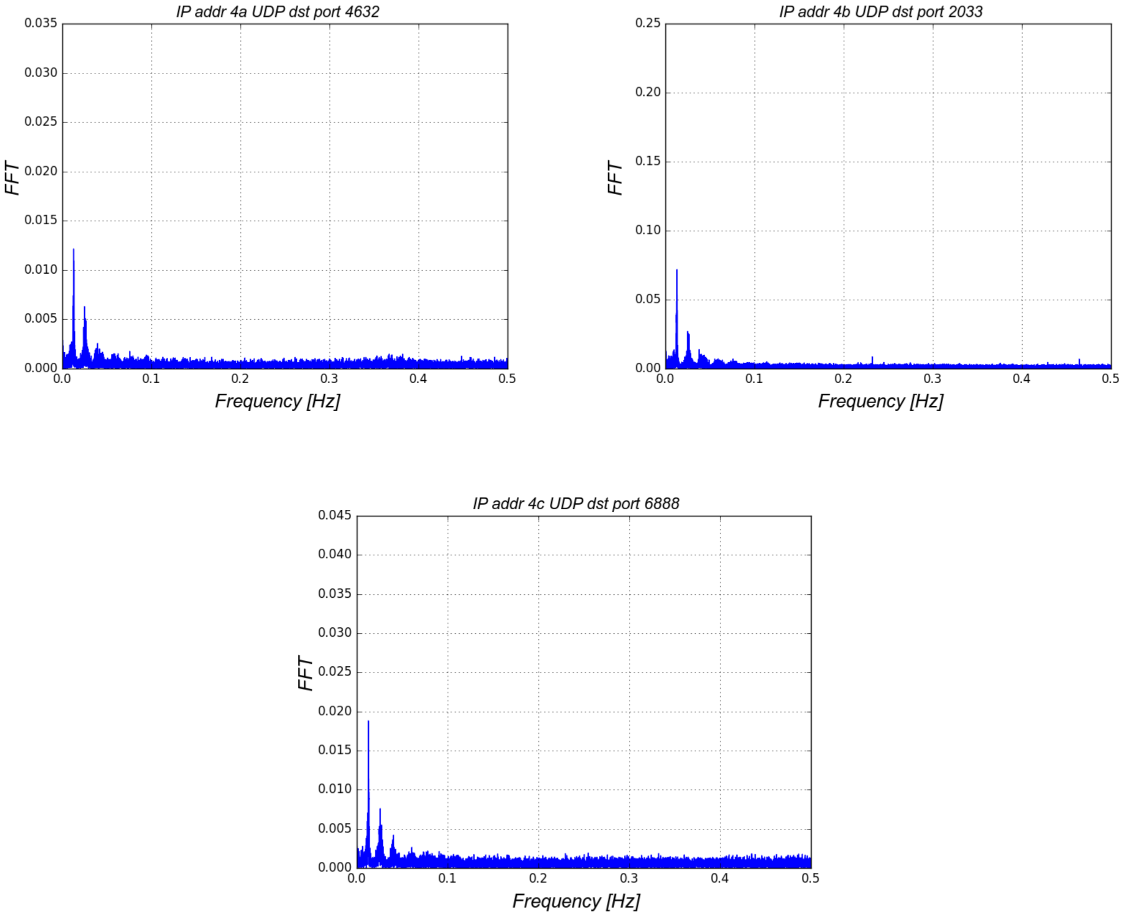


Figure 9: The longest time dataset. Similar characteristics among different ports and protocols, but within same source IP address.

As we presented before, we perform analysis by taking into account only three values from the packets, namely IP address, port and protocol. For the longest dataset, the combination of source IP address and destination port was used, as such tuple proven to be the most representative and results are seen to be the most accurate ones.
The dataset produced big amount of data to be analyzed, thus to optimize the process we chose group of ten thousand most common combinations of IP, port and protocol,

which have been selected for deeper study. 273 of those combinations were found to be periodic, and among them three common patterns were observed.

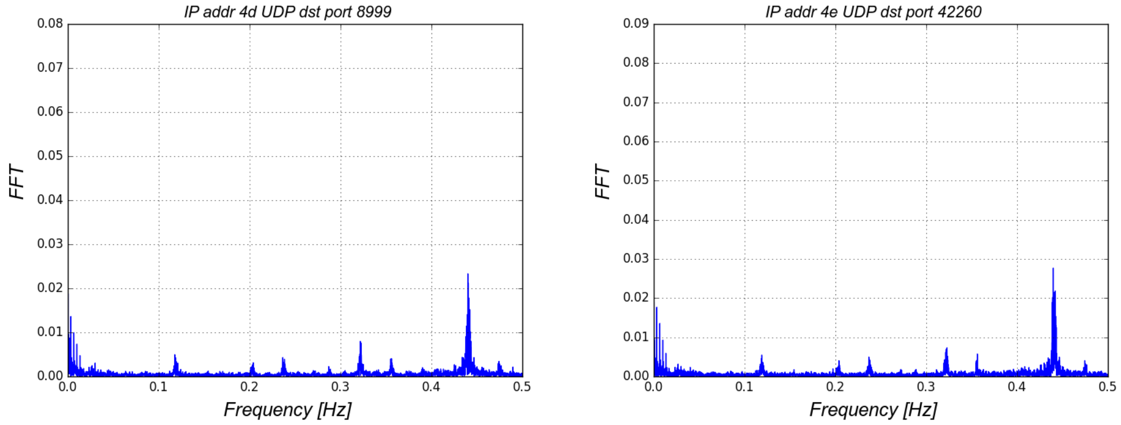


Figure 10: The longest time dataset. Similar characteristics among different ports and protocols, but within same source IP address.

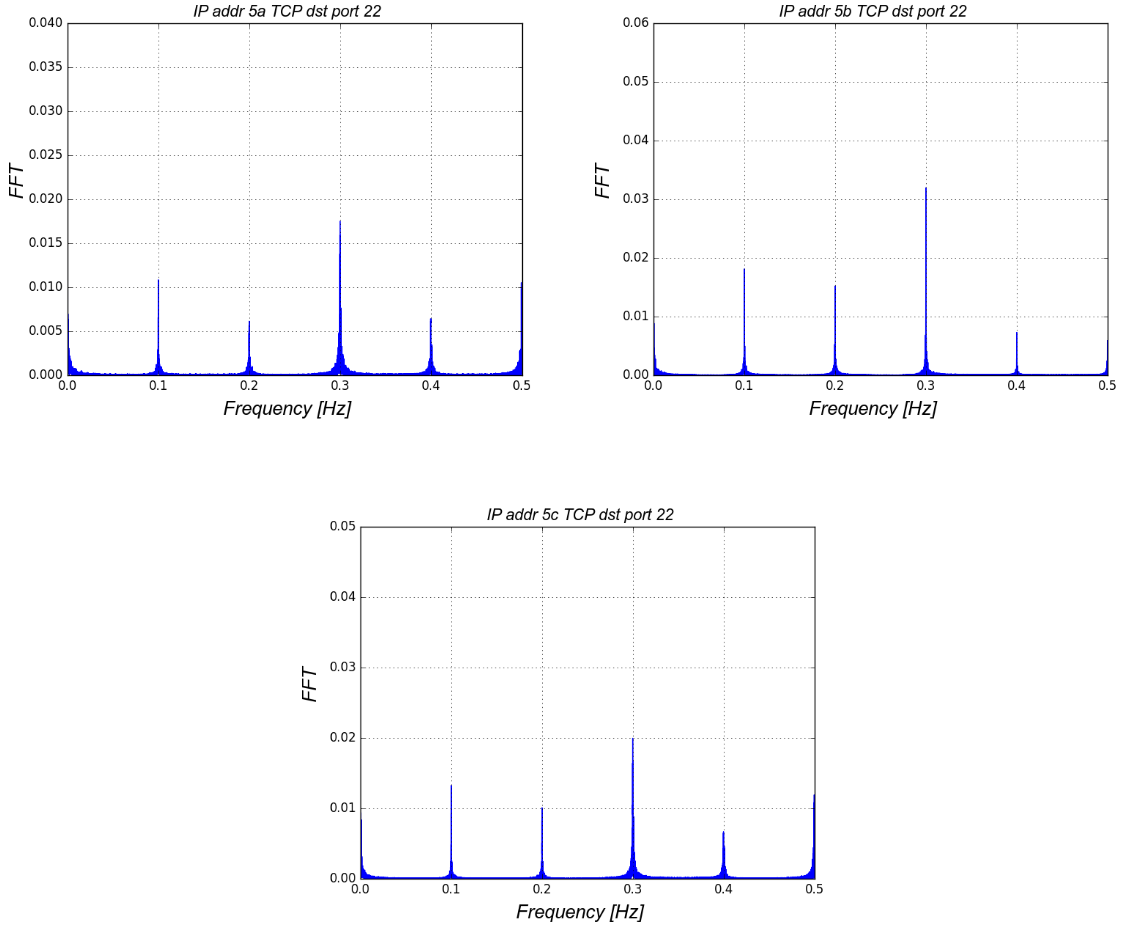


Figure 11: The longest time dataset. Similar characteristics among different IP addresses for TCP port 22.

In the first pattern couple of IP addresses sent data into different destination ports with almost identical frequency. Their Discrete Fourier Transform plots can be seen in fig. 9 and fig. 10. Most probable reason of such behavior is port scanning performed from the machines operating under those addresses. Nevertheless characteristics between different IP addresses varies significantly on frequency level which can be understood as different techniques used in the scanning itself. One IP address sent packets with low periodical frequency (~0.02 Hz, 1 packet every 50 s), whereas the other is found to be within high periodical frequencies (~0.45 Hz, about 1 packet every 2,5 s).

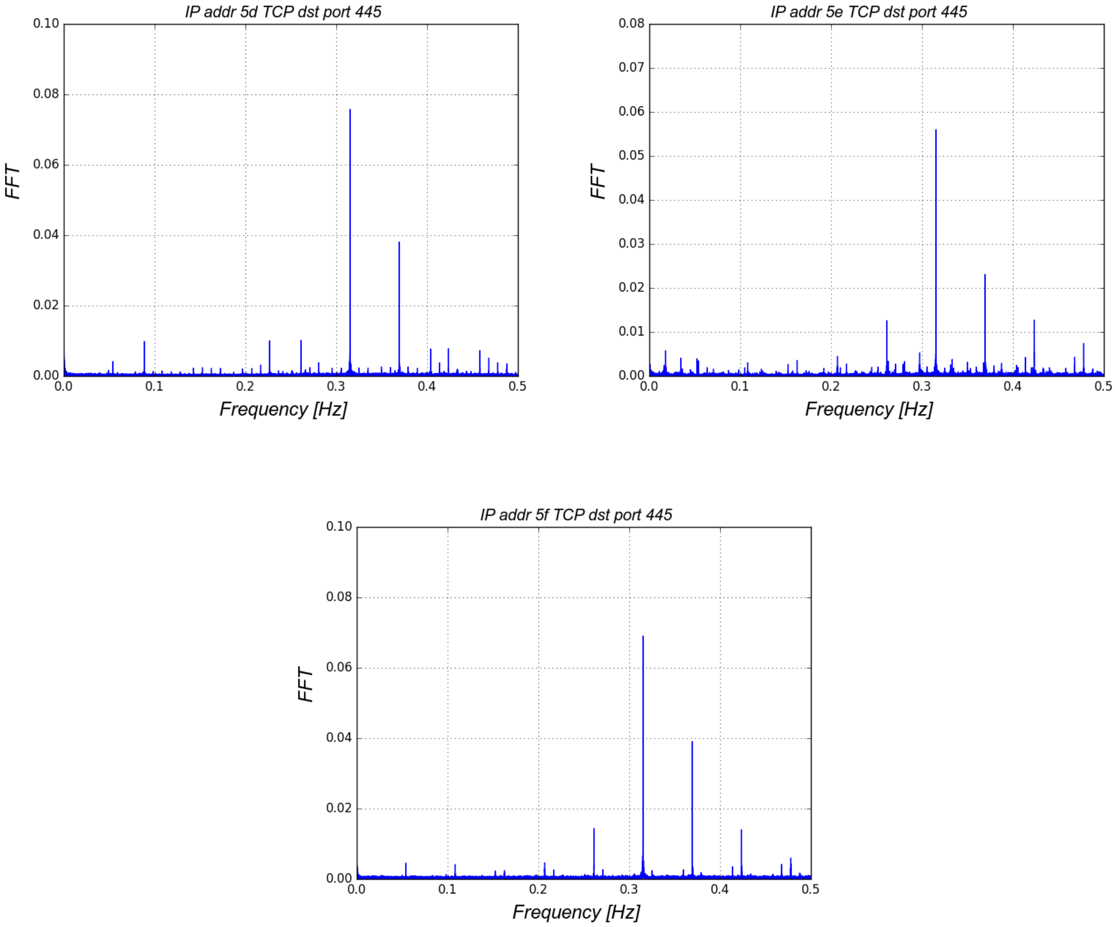


Figure 12: The longest time dataset. Similar characteristics among different IP addresses for TCP port 445.

In the second pattern we observed similar characteristics of network traffic for the same destination port, but originating from different addresses. Although the previous pattern was quite common and it was rather trivial to find its cause, this case does not have simple explanation. However, we can assume that traffic coming to a specific port is bound to a specific service, what is generally true for common legit traffic, eg. TCP SYN segment sent to port 22 usually determines connection attempt to SSH service and in that case characteristic of such connection is predefined and well-known (as in fig. 11). Surprisingly, such regularity can be also seen in traffic trying to reach the services which are not responding in any way and it proved to be repeatable among different IP addresses from different Autonomous Systems (fig. 12, 13). It could either mean all of the machines are part of one big network or it would mean that there is a common connection negotiation method for specific port/service among different clients.

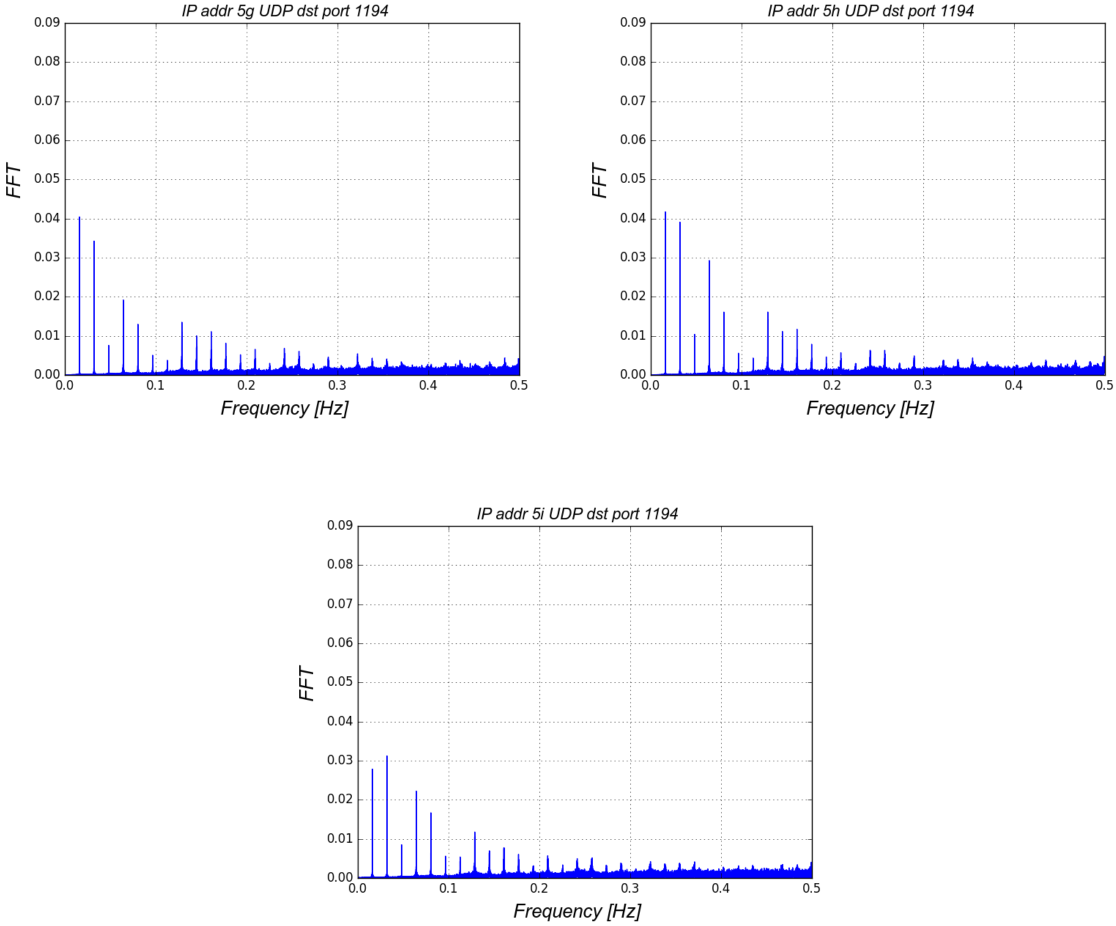


Figure 13: The longest time dataset. Similar characteristics among different IP addresses for UDP port 1194.

Similar situation occurred when comparing same destination protocol. Specifically different source IP addresses possessed comparable DFT plots in ICMP protocol analysis. Identification of patterns was relatively hard in that case, comparing to other protocols, as total of 42 IP addresses were found to share similar frequency spectrum, and within this group, three different patterns were found. The reason of such low similarity recognition can lie in protocol mechanism and behavior – in observed traffic ICMP requests were mainly ping requests (ICMP Echo request), which are designed to check if there is a machine on the other side responding with pong (ICMP Echo Reply). Additionally ICMP protocol is not bound to any specific service, like it happened in previous pattern describing same port similarities. As a result, it is relatively harder to find similarities and obtain context of data exchange than in other analyzed protocols. Example of one of the recognized frequency patterns can be seen in fig. 14

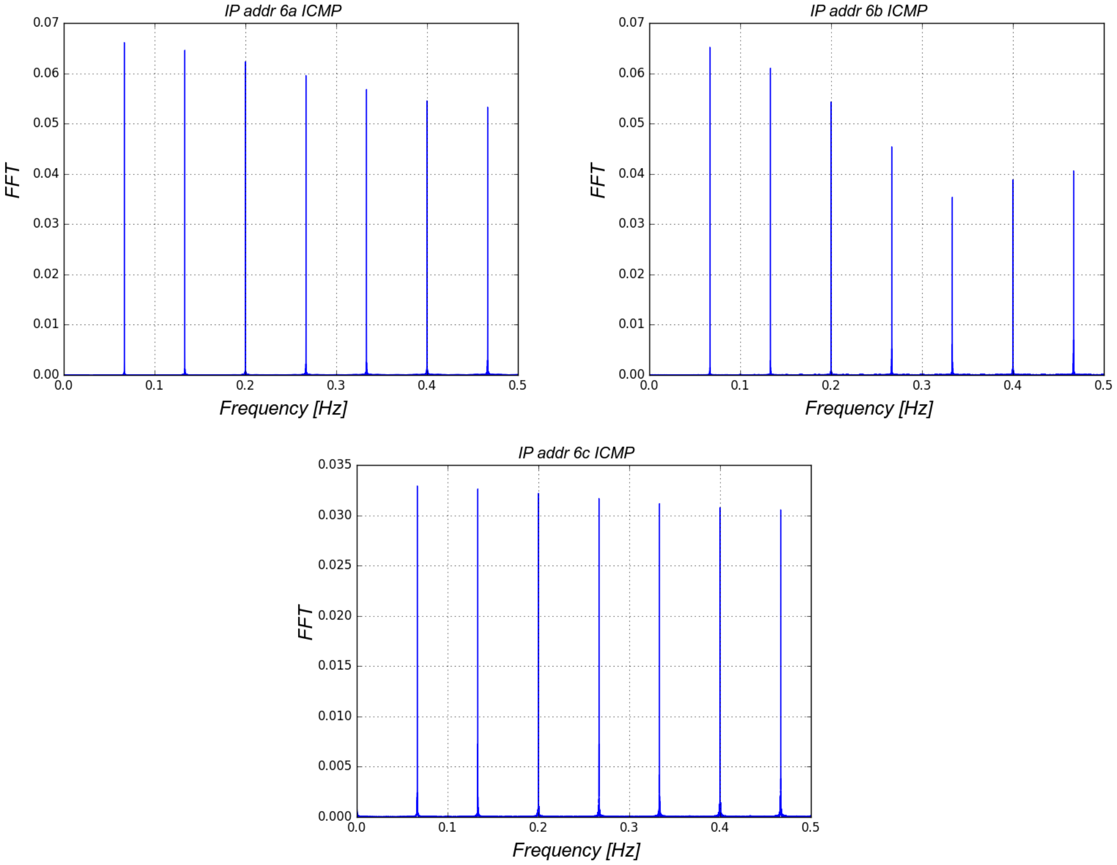


Figure 14: The longest time dataset. Similar characteristics among different source IP addresses for ICMP protocol.

6. Summary

After analysis we found that majority (199 out of 261, 77%) of identified periodicities possessed the same frequency characteristics when grouped by either same port, protocol or IP address.

Similar frequency characteristics by source IP is usually correlated with port scanning, where the same IP sends requests with the same frequency to multiple destination ports and possibly machines too. Another group of regularities is detected on the same destination port. Such case can be treated as a potential factor in defining anomalies in the gathered traffic in general, i.e. if frequency characteristic for the destination port is different than observed one during long term, then the alarm can be triggered and further analysis can be made in order to decide on the reason of such behavior.

As we have presented, the Discrete Fourier Transform can be successfully used to find periodicities in the dark address space network traffic. Such analysis can be a starting point for further investigation, for example after filtering out periodicities of chosen frequency. As above analysis method could be applied also to normal traffic, not only dark space, we hope that it could help you with developing your own analysis or systems.