ELEKTROTEHNIŠKI VESTNIK 89(4): 198-204, 2022
ORIGINAL SCIENTIFIC PAPER

Network telescope: insights from a decade of observations

Urban Sedlar

University of Ljubljana, Faculty of electrical engineering, Tržaška 25, 1000 Ljubljana, Slovenia
E-mail: urban.sedlar@fe.uni-lj.si

Abstract. Information and communication technologies have become the foundation of modern life due to their numerous advantages. However, their rapid introduction with an insufficient emphasis on security and protection unnecessarily exposes them to potential risks. In the paper, we focus on cybersecurity from the perspective of the growing amount of threats in modern networks. We describe the concept, operation, and purpose of a network telescope, i.e., a system that records unsolicited traffic to dark Internet Protocol addresses. We present the architecture of the system developed at the Faculty of Electrical Engineering, University of Ljubljana, and analyze the data collected in more than ten years of its operation. More than 2 billion events collected from 2011 onwards, exhibit an exponential trend of growth. Only five years ago, a vast majority of incoming packets was targeting ports of popular server management services (e.g., Telnet, SSH), while today the distribution is much more long-tailed and responds quickly to emerging vulnerabilities. We examine the nature of the collected data, describe the possible use cases, and present some data visualizations.

Keywords: cybersecurity, networks, internet protocol, network traffic, network telescope

Omrežni teleskop: ugotovitve iz desetletja opazovanj

Informacijsko-komunikacijske tehnologije so zaradi številnih prednosti postale temelj sodobnega življenja, vendar pa njihovo naglo uvajanje z nezadostnim poudarkom na zaščiti in varovanju po nepotrebnem izpostavlja sisteme potencialnim tveganjem. V tem članku se dotaknemo kibernetske varnosti z vidika naraščajoče količine groženj v internetnih omrežjih. Opišemo koncept, delovanje in namene uporabe omrežnega teleskopa, tj. sistema, ki beleži nepovabljen promet do temnih naslovov internetnega protokola. Predstavimo arhitekturo sistema, razvitega na Fakulteti za elektrotehniko, ter analiziramo podatke, zbrane v več kot 10 letih njegovega delovanja. Več kot 2 milijardi zbranih dogodkov od leta 2011 prikazuje eksponenten trend rasti. Še pred 5 leti je večina dohodnih paketov targetirala storitve za upravljanje strežnikov (npr. Telnet, SSH), danes pa je porazdelitev protokolov mnogo bolj uravnotežena in se hitro odziva na aktualne ranljivosti. V prispevku opišemo naravo zbranih podatkov in možne načine njihove uporabe ter podamo nekaj vizualizacij.

Ključne besede: kibernetska varnost, omrežja, internetni protokol, omrežni promet, omrežni teleskop

1 INTRODUCTION

In the past two decades, our lives have become increasingly dependent on modern technology. The digital transformation of our economy and society promises an increase in the productivity and efficiency of processes, facilitates communication and retrieval of information, and complements humans in their decision-making using machine learning. Furthermore, recent events, such as the COVID-19 pandemic, have shown that by using information and communication technologies (ICT), we can adapt a significant part of our work, education, and entertainment to online environments, and largely continue to function as a society.

Most of this has been made possible through rapid progress in two domains: exponential growth of computing capabilities (Moore's law) and the proliferation of the Internet, which enables universal connectivity of modern systems and devices. Internet technologies have effectively shrunk the world and obtaining information from the other side of the globe is now just a click away. But the trends and developments in this area show that this is only the beginning: virtual and augmented reality will bring an illusion of a physical presence in another location and enrich our environment with contextual information; sensor devices in our physical environment will serve to build better predictive and decision-making models; all of this will drive increased efficiency in modern industry and life in general. If today we can no longer imagine our work and life without computers and the Internet, this will be even truer in the future.

But even today, the dark side of digital transformation is becoming increasingly apparent through the lens of cyber security. By connecting our physical environments (smart locks, smart vehicles, industrial devices, smart lighting, etc.) to the Internet, and storing our most valuable data in cloud systems with external providers, we are preparing a fertile ground for attackers. The tools, techniques, and knowledge that are useful for compromising networked computers suddenly become useful for attacking critical infrastructure, disabling

Received 26 September 2022
Accepted 7 October 2022

vehicles on the road, and connecting to security cameras in the privacy of our homes or workplaces.

In such an environment, it becomes crucial to be aware of the scope and nature of cyber threats and to be able to obtain an appropriate *weather forecast* about the current conditions. Useful tools to achieve this goal are the so-called darknet sensors or network telescopes – systems that monitor the incoming traffic on unused IP addresses.

In the Laboratory for Telecommunications at the Faculty of Electrical Engineering, we have been monitoring since 2011 the activity on a /24 segment of public IP addresses (a total of 256 addresses) that have never been actively used, connected with any domain, or advertised as a part of any service. Over the 11 years of our data collection, we have detected an exponential growth of the traffic (Figure 2) and for the last 12 months, there have been on average between 40M and 61M events per month on these addresses (the period from June 2021 to June 2022). For a comparison with the beginnings of data collection: in June 2022 we captured a total of 54M events, while in the same period 11 years earlier (June 2011) we detected only 1.3M events.

In the paper, we describe in detail the architecture and operation of the system, as well as approaches to the analysis and visualization of the captured data. We also indicate how the data obtained in this way can be used in combination with other defense systems against possible cyberattacks.

2 Network telescope

A network telescope is a tool for monitoring poorly visible events on the Internet. The essence of its operation is monitoring the traffic on inactive (so-called dark) Internet Protocol (IP) addresses [1]. Inactivity in this context does not only mean that no one is *currently* responding on a specific address, but also that the amount of the expected legitimate traffic to these addresses is negligible. It is ideal, if the address has not been used for a very long time or has never been used at all, is not advertised anywhere, and has no records in the Domain Name System (DNS).

Since, by definition, all the Internet traffic that reaches such addresses is unsolicited, analyzing its characteristics can inform us about various illicit activities of Internet users. Analogous to the mirror size of an astronomical telescope, a larger number of inactive addresses increases the resolution of the network telescope and thus the ability to observe smaller events.

Because the telescope operates at the network layer (Internet Protocol), the traffic can contain any transport protocol that can be encapsulated in IP. Such protocols are identified by the protocol number field in the IP packet header. Both in regular internet usage as well as in our collected data, the most frequently used are: Transmission Control Protocol (TCP) with the protocol number 6; User Datagram Protocol (UDP) with the number 17; and Internet Control Message Protocol (ICMP) with the number 1; however, any of more than a hundred of other possibilities may also appear [2].

The traffic that reaches such dark IP addresses can be divided into two major types:

- Active traffic: the ICMP traffic is typically generated during ping scanning, while the TCP and UDP traffic is generated by port scanning. Searching for open ports is a common activity by which potential attackers find out on which ports the server is responding, and from the response, it is then possible to infer the active services, where a potential attack is possible [3] (either by guessing the password, performing *fuzzing* attacks with a random payload, etc.). Another possible source of the active traffic is the misconfiguration of systems that can be caused either by software, hardware, or human errors [4].
- Passive traffic: this is most often backscatter [4], i.e., the traffic resulting from the impersonation of the actual source IP address (IP spoofing), which is usually carried out during various targeted attacks. For example, if in a Distributed Denial of Service (DDoS) attack the attacker spoofs one of *our* IP addresses as the source address, then the target's response to the attack will return to our inactive IP address [5].

3 System description

Our laboratory network telescope uses a dark /24 subnet of public IP addresses (256 addresses). The data capture system runs on a virtual server (Ubuntu Linux) that has the entire subnet routed to it. Databases that serve the data feeds for all publicly facing visualizations are hosted on a separate server and thus don't interfere with the data capture in any way.

The entire setup contains the following building blocks (Figure 1):

- A router that routes the incoming traffic to the address of a capture server;
- *A LibPCAP* filter that removes any legitimate traffic (e.g., any traffic destined to the database server that also powers our visualizations[1]), and any other test traffic from other subnets of our laboratory;
- A geolocation component to determine an approximate location based on the IP address; for this, we rely on a local lookup database by MaxMind [6];
- A time-partitioned database (MySQL) in which we store basic data about all events: time stamp, source and destination IP address, protocol, destination port (in case of TCP, UDP and SCTP protocols), geo-located coordinates, ISO country code, etc.

[1] https://telescope.ltfe.org

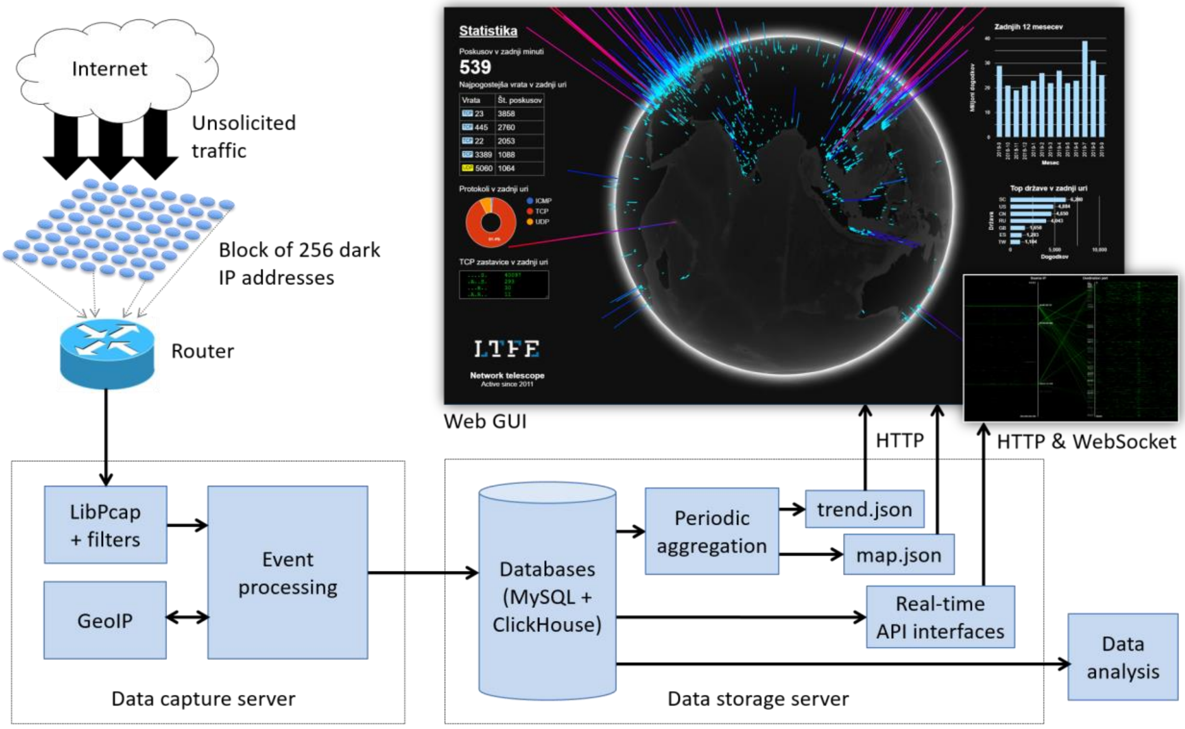

Figure 1. System architecture.

- For more performant complex queries, most of the dataset is also stored in a columnar database (ClickHouse, [8]), which is both more efficient with the storage (about 30% of the size) and significantly faster for analytic queries.
- Scripts for a periodic calculation of statistics for visualizations: a globe with the data for the last three hours, which is updated every five minutes, and a cumulative monthly chart for the last 12 months which is updated daily.

The system has had several major upgrades in the past, both in terms of functionality (logging of protocols, TCP flags, TCP windows, payload size, etc.) and capacity (optimization, scaling, and upgrades of the database due to the quickly growing amount of the data).

The system currently stores the following metadata of the received events:

- timestamp (in UTC to get a continuous time axis throughout the year),
- source IP address and source port (in the case of the TCP, UDP, or SCTP protocol),
- destination IP address and port,
- protocol number,
- geographic coordinates (approximate latitude and longitude) and ISO code of the country of origin,
- IP packet length,
- IP Time-to-Live (TTL),
- TCP flags, window size and segment length (in the case of the TCP protocol), and
- UDP packet length (in the case of the UDP protocol).

Analyses based on the extended set of metadata are thus not shown on the entire ten-year period, but only on more recent data, going back as much as five years.

4 Data Analysis

The database with more than a decade of data contains over 2.1 billion events (as of June 2022), exhibiting exponential growth throughout the years (see Figure 2).

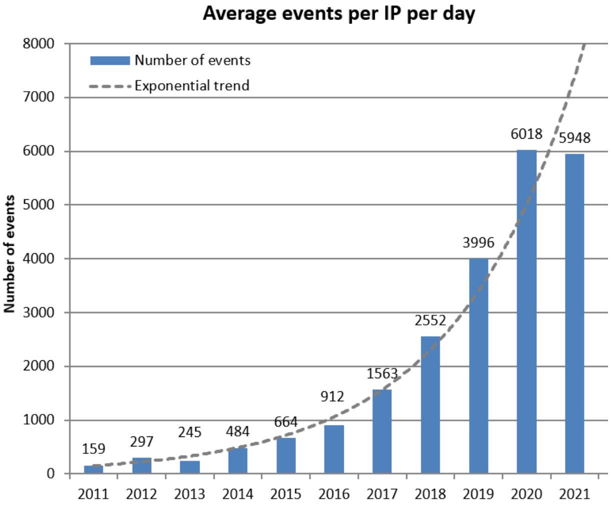

Figure 2. The number of the captured events by year normalized per IP address per day.

The largest share of the requests uses the TCP protocol; the cumulative data for 2021 show that there are 90.230% of TCP requests, followed by UDP (8.773%), ICMP (0.661%), GRE (0.334%), and 17 less common protocols. This statistic is surprising, as it shows that attackers and researchers in most cases do not even bother to test if the server responds to ICMP, but instead they directly scan the target port of the service they are targeting.

4.1 Determining the traffic source location

The location of traffic sources can be roughly determined in two ways: by using a database for the geolocation of IP addresses [6][7] and at the level of the owner of the Autonomous System (AS). One must take into account that the use of a Virtual Private Network (VPN) or the Tor anonymization network can hide the real IP of the attacker, which makes the address information even less reliable, but we estimate that due to the cheap provisioning of cloud servers, such approaches are less attractive due to poor cost-effectiveness. The source address can also be forged (spoofing) [9], which is often used mainly in Denial of Service (DoS) attacks and Distributed Denial of Service (DDoS) attacks [10], as there the attacker not only does not *need* an answer but expressly doesn't *want* it. The rough location of the origin is useful for interactive visualizations on our website and for traffic analyses according to the country of origin (Figure 3).

4.2 Source operating systems

The source operating system can be inferred based on a combination of the IP packet and TCP segment parameters. Typically, IP Time to live (TTL) and TCP window size are used for this [11][12]. Default initial parameters are usually defined in the network stack of the operating system and are consequently more difficult to change or even immutable (e.g., in closed-source operating systems). The distributions of both fields on a one-month data sample can be seen in Figure 5: IP TTL distribution on the upper axis of the graph, and TCP window size distribution on the left axis. The central part of the visualization shows the frequency of the combinations of both parameters, and some common combinations that are specific to individual operating systems. Example: the Windows operating systems use an initial IP TTL value of 128; this is reduced by 1 after each additional router hop. Based on the default TCP window size, we can distinguish Windows XP (window size 65535) and newer versions of Windows (window size 8192).

4.3 TCP traffic classification

As mentioned above, most of the traffic is represented by the TCP protocol. Due to the three-way handshake method, TCP is also the most suitable for separating individual sources of the traffic: port scanning, backscatter traffic, and misconfiguration traffic. For separation, we use the fields (flags) in the TCP segment header. In the literature, in the context of the network telescopes, the following classification of the traffic types is used according to the combination of the TCP flags [4]:

- SYN packets indicate active traffic and are considered port scanning
- Packets with flags RST, ACK, SYN+ACK, and RST+ACK are considered backscatter traffic
- The rest of the traffic is considered the result of misconfigurations.

Applying this classification to the data for the calendar year 2021 shows that the largest part of the received TCP segments has the SYN flag (97.55%) and is thus considered to be port scans; 2.43% of the segments represent backscatter traffic, and the remaining 0.02% is the misconfiguration traffic. Considering the current extremely varied approaches to port scanning, it is reasonable to assume that at least part of the remaining traffic is also related to the port scanning activities. We can gain two insights from this data: firstly, an increase in the port scanning activity and changes in the relative popularity of individual services (ports) over time, and secondly, variations in the amount and origin of the backscatter traffic. Figure 5 shows the relative amounts of events by the destination port for the last five years.

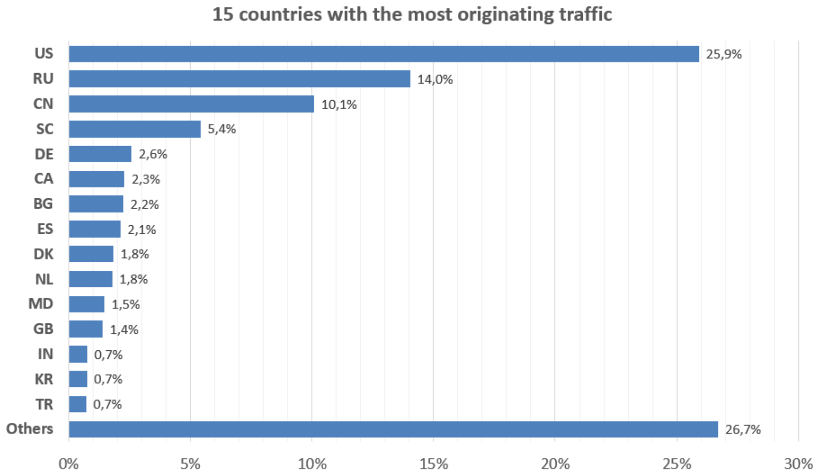

Figure 3. Countries with the largest share of the source traffic (based on cumulative data for 2021).

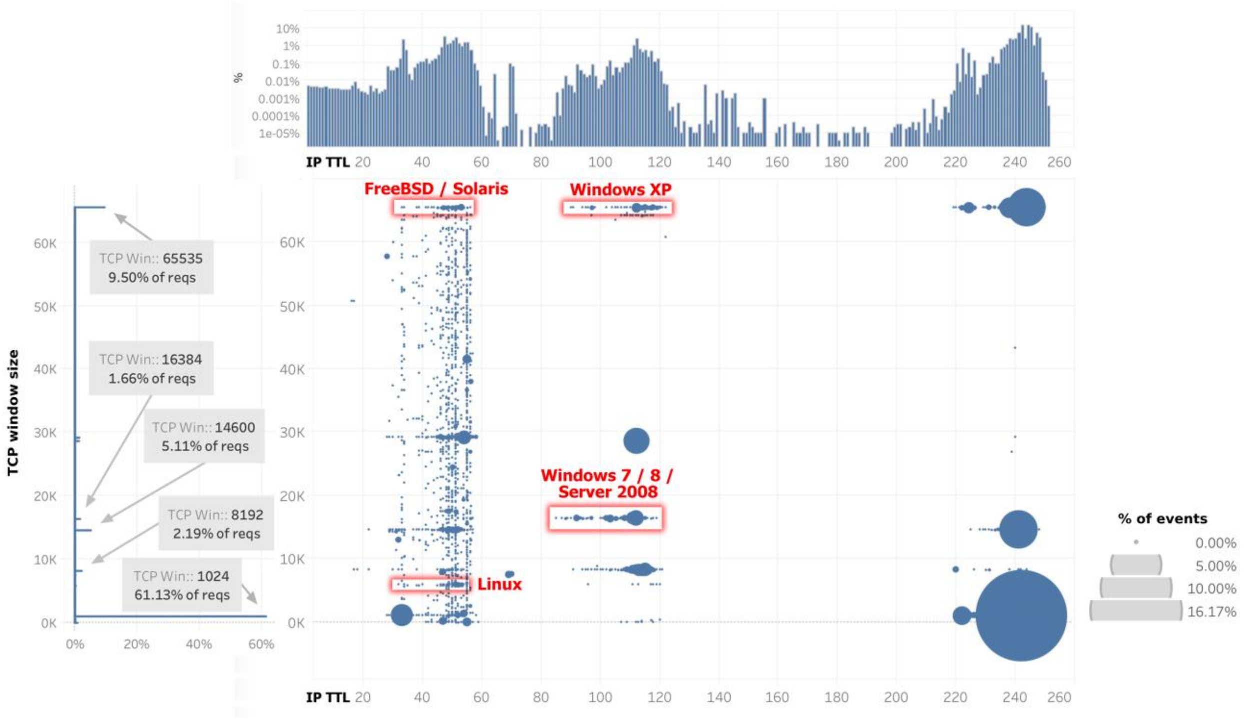

Figure 4. Distribution of the TCP window sizes and IP TTL with popular operating systems marked in red (sample for the month of July 2019).

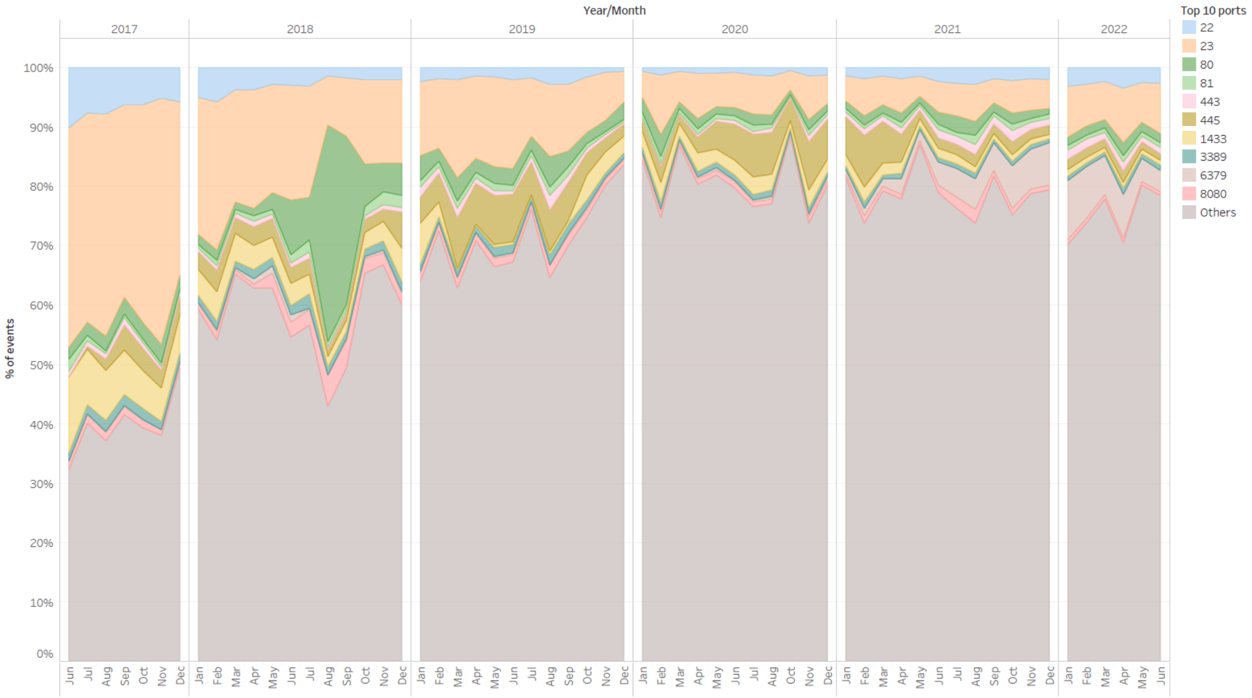

Figure 5. Distribution of the TCP ports over the last five years of data collection. The most common ten ports are shown in a color, and the rest of the traffic is shown in gray. A large drop in the proportion of the Telnet (port TCP/22) and SSH (port TCP/23) traffic can be seen from 2017 to 2020 (top two categories, respectively), as well as an increase in the long tail of services (bottom category – Others).

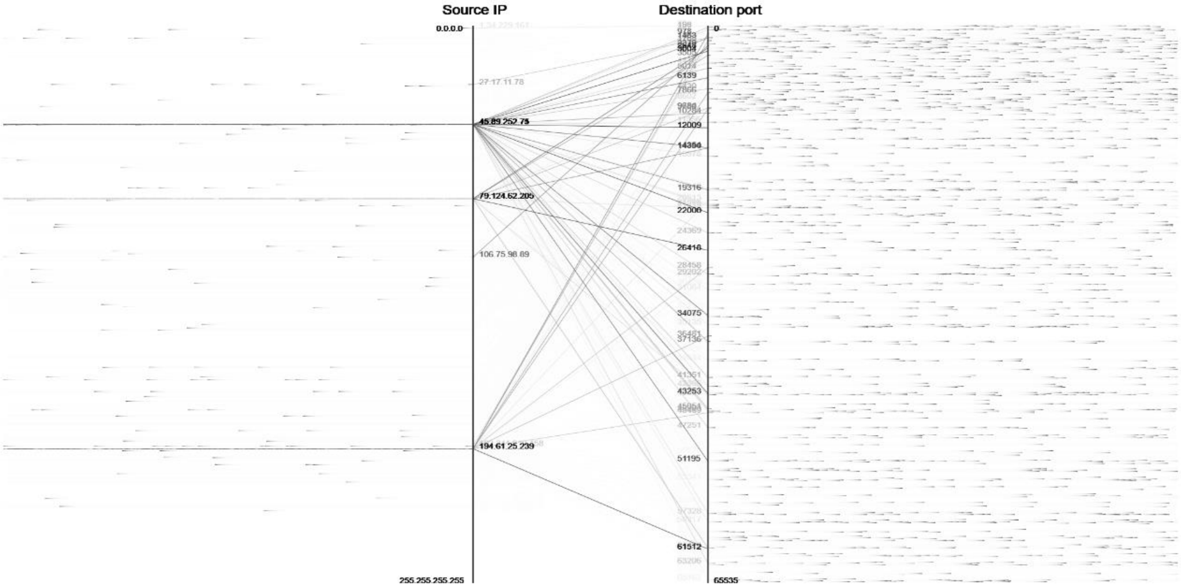

Figure 6. SecViz-inspired [14] traffic visualization with the entire IP address space on the left vertical axis, and ports from 0 to 65535 on the right axis. Fan-out shapes intuitively show a port scan from one or more IP addresses, the color of the lines encodes the used protocol.

4.4 *Visual analytics*

The discipline of visual analytics [11] is based on the premise that the human brain is an excellent pattern detector that can be effectively harnessed with proper data preparation and presentation. For this purpose, we prepared several visualizations; one example is shown in Figure 6. It is a variation of the SecViz visualization documented in the literature [14]. The visualization is also available on our website and allows an intuitive display of the mapping between the source IP address and destination ports. The fan-out patterns that appear represent a systematic port scanning attempt.

In addition to the 2D visual representation, a 3D display is also possible. It can become a particularly powerful technique using augmented and virtual reality (AR/VR) technologies. When channeling information to the human brain, sonification is also effective for easier recognition of one-dimensional time-dependent patterns.

5 DISCUSSION

The analysis of the data reveals some surprising facts. We expected a much higher proportion of the ICMP traffic (ping), but the amount is negligible despite the availability of several high-performance ping tools (e.g. zmap [15]) that allow a massively parallel ICMP scanning of the entire Internet address space in an extremely short time (i.e., approximately 45 minutes using a gigabit connection).

Another observation is that the number of requests has significantly increased and has approximately doubled every two years. For network security, this is an unfavorable trend. It means that a device with no firewall on a public IP address has an average time to the first contact of less than a minute (2021 data). Of course, there are variations in the used protocols. A device running Microsoft Windows with enabled Remote Desktop Services (open TCP port 3389) would see a connection attempt on that port in less than half an hour.

Most of the traffic is represented by attempts to establish a connection (SYN), where a significant relative (but not absolute) decline in the share of the SSH and Telnet connection attempts can be seen in the last five years at the expense of the increase of other types. It should be noted that the absolute amount of connection attempts on ports TCP/22 and TCP/23 has been fairly constant over the past five years. Information about such direct (active) traffic is useful in practice both for monitoring activities and trends in the field of Internet indexing services (researchers, search engines) and on the side of the attackers. This data can also be used for establishing the IP reputation, which can be used in the IDS/IPS systems and firewalls. Using large observation windows, it is also possible to detect a long-term activity that is hidden below the sensitivity threshold of most modern detection systems (e.g., port scans that are sufficiently throttled with the purpose of avoiding detection).

6 CONCLUSION

In the paper, we show the architecture and demonstrate the operation of a network telescope which has been active since 2011 at the Faculty of Electrical Engineering, University of Ljubljana.

Some initial findings are presented, such as the change in the volume and composition of the inbound traffic.

These findings have inspired many of our other efforts in the domain of threat intelligence, such as the development and deployment of various types of honeypots and attacker observation experiments.

Some of the important issues will be dealt with in our future work, such as an improved representation of the data for different types of analytics, and an analysis of the backscatter traffic to provide information about large cybersecurity events, such as the massive DDoS attacks.

ACKNOWLEDGMENTS

The research was in part supported by the Slovenian Research Agency and Government Information Security Office of Slovenia through the research project "Exposure of Modern Information and Communication Infrastructures to Cyber Attacks" (V2-2125) and the research programme "Decentralized solutions for the digitalization of industry and smart cities and communities" (P2-045).

REFERENCES

[1] Fachkha, Claude & Debbabi, Mourad. (2015). Darknet as a Source of Cyber Intelligence: Survey, Taxonomy and Characterization. IEEE Communications Surveys & Tutorials. 18. 1-1. 10.1109/COMST.2015.2497690.

[2] Internet Assigned Numbers Authority: Assigned Internet Protocol Numbers: https://www.iana.org/assignments/protocol-numbers/protocol-numbers.xhtml, accessed: 21.9.2018

[3] Sanghvi, H. P., and M. S. Dahiya. "Cyber reconnaissance: an alarm before cyber attack." International Journal of Computer Applications 63, no. 6 (2013).

[4] Balkanli, Eray, and A. Nur Zincir-Heywood. "On the analysis of backscatter traffic." 39th Annual IEEE Conference on Local Computer Networks Workshops. IEEE, 2014.

[5] David Moore, Geoffrey Voelker, and Stefan Savage. Inferring internet denial-of-service activity. In Proceedings of the 10th Usenix Security Symposium, 2001

[6] MaxMind GeoIP GeoIP® Databases & Services https://www.maxmind.com/en/geoip2-services-and-databases, accessed 21.9.2018

[7] Shavitt, Yuval, and Noa Zilberman. "A geolocation databases study." IEEE Journal on Selected Areas in Communications 29, no. 10 (2011): 2044-2056.

[8] Struckov, Alexey, Semen Yufa, Alexander A. Visheratin, and Denis Nasonov. "Evaluation of modern tools and techniques for storing time-series data." Procedia Computer Science 156 (2019): 19-28.

[9] Hastings, Nelson E., and Paul A. McLean. "TCP/IP spoofing fundamentals." In Conference Proceedings of the 1996 IEEE Fifteenth Annual International Phoenix Conference on Computers and Communications, pp. 218-224. IEEE, 1996.

[10] Ferguson, Paul, and Daniel Senie. Network ingress filtering: Defeating denial of service attacks which employ IP source address spoofing. No. rfc2267. 1998.

[11] Li, Ruoshi, Markus Sosnowski, and P. J. N. Sattler. "An Overview of OS Fingerprinting Tools on the Internet." Network 73 (2020).

[12] Zalewski, Michal. "p0f v3 (version 3.09 b)." (2014).

[13] Keim, Daniel, et al. "Visual analytics: Definition, process, and challenges." Information visualization. Springer, Berlin, Heidelberg, 2008. 154-175.

[14] S. Krasser, G. Conti, J. Grizzard, J. Gribschaw and H. Owen, "Real-time and forensic network data analysis using animated and coordinated visualization," Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop, 2005, pp. 42-49, doi: 10.1109/IAW.2005.1495932.

[15] The ZMap Project, https://zmap.io/, accessed 5.5.2022

Urban Sedlar received his PhD degree in 2010 from the Faculty of Electrical Engineering, University of Ljubljana, where he is currently employed as an assistant professor and senior researcher at the Laboratory for Telecommunications. His recent work focuses on the area of cybersecurity threat assessment using large-scale honeypots. He leads a national research project for cyber threat assessment in modern digital infrastructures. He has been involved in several EC and national research and development projects on the topics of cloud computing, the Internet of Things, and emergency response systems.