Information leakage via Certificate Transparency

BORIS GERRETZEN, University of Twente, The Netherlands

As the need for encryption on the internet grew, the secure sockets layer (SSL) and later TLS were introduced. These protocols rely on a cryptographic system called public-key cryptography. Where every user has a public and private key they use for encryption. To make sure that the key actually belongs to you, certificates were introduced, they are proof that you are the owner of a public key.
X.509 is the current standard format for public key certificates. Most of the certificate authorities that issue these certificates also publish them to certificate transparency logs. A number of certificates from various certificate transparency logs have been saved by the University of Twente, the domain names in these certificates have been analyzed to see what sensitive services can be identified by analyzing labels in the domain names.
I found that services that are indicated by the domain name label have their default ports exposed to the internet with rates varying from 1% to 90% depending on the service. In addition to this, I found that the amount of hosts with open ports that allow unauthenticated access ranges from 3% to 97%, depending on the service.

CCS Concepts: • **Networks** → *Web protocol security*; *Network privacy and anonymity*; Public Internet.

Additional Key Words and Phrases: certificate transparency, information leakage, DNS

1 INTRODUCTION

When the need for encrypted traffic on the internet grew, the secure sockets layer (SSL) was introduced.[1] SSL, and later transport layer security (TLS), provide secure communication over an insecure channel using cryptography. To do this, both SSL and TLS require certificates.[2] These certificates are used to prove that a public key belongs to a specific identity, for example a host name. The certificates are signed cryptographically by a certificate authority (CA), a trusted authority who verifies the ownership of the host name or identity by the person or institution requesting the certificate.
This introduces a major weakness into the internet infrastructure, the certificate authorities have to be trusted that they actually check the identity of the requester and do not give out certificates to malicious actors that try to get certificates for domains they do not own.

In 2011 it came to light that the Dutch certificate authority Diginotar had been compromised and was issuing fraudulent certificates for domains such as `google.com` and many more. [3] Because of this incident and others like it, a solution to this problem was needed. In 2012, a group of researchers submitted the first draft of what would become 'RFC 6962 - Certificate Transparency' (CT) to the IETF. [4]

The general idea of CT is to create a publicly accessible list that contains all of the certificates issued by a CA. The idea is not that the list itself prevents CA's from issuing fraudulent certificates, but because it is made public it can be audited by people and organizations to detect fraudulent certificates. [5]

A problem with certificate transparency as it is right now is that if some confidential or personal information is contained in the certificate, it is published in a CT log that everyone can see. This is especially a problem since certificate transparency is not known by every system administrator or other person who requests a certificate. Because domains often reference the service running on them, and 95% of websites use software that is out of date [6], this provides an easy way for a malicious actor to identify vulnerable services.g

In this paper, I will investigate what kinds of applications can be identified by looking at the labels in the domain names in certificates from scraped CT logs. For example `citrix.example.com` is a domain name that will most likely host a Citrix server.
Furthermore, I will check the IP addresses of the domains that include these labels in internet wide scans to confirm that they are actually running that service, and to see if any additional services can be discovered.

2 PROBLEM

In addition to securing their main domain, institutions usually also want a certificate to be valid for other domains, e.g. internal services. Or a company might want to give out subdomains for every employee, e.g. `boris.example.com`, the certificate would also need to be valid for these domains.
The problem with this is that some administrators who manage these certificates are not aware that all of this information gets placed in a publicly available list. The result of this can be that information that should be kept private is now available for everyone to see.
This paper will analyze what kinds of personal information can be found in the certificate transparency logs, and which applications can be identified by looking in the X.509 Common Name (CN) field and the Subject Alternative Name (SAN) field.

3 RESEARCH QUESTIONS

The problem statement will lead to the following research question:

How many and which sensitive applications and personal information can be found by analyzing domains in the SAN and CN fields of certificates gathered from CT logs?

This question can be answered with the following sub questions:

(1) What kinds of personal information can be gathered through the analysis of the domain names in X.509 certificates?
(2) Which domain name labels that indicate sensitive applications can be identified?

TScIT 37, July 8, 2022, Enschede, The Netherlands

(3) Which services can be confirmed to be running by combining the domain names from certificate transparency logs with internet wide scan data?

4 METHOD

Firstly, I conducted a literature review to find out which services are often exposed to the internet that should not be from a security standpoint. In addition to this, I attempted to find labels that are connected to these services. Unfortunately, I was unable to find a comprehensive study about common vulnerable services.

Therefore I queried the certificate transparency database for the top 10.000 labels. Duplicates and domains with less than 3 labels have been removed from this query. Duplicates happen because certificates expire after a certain amount of time. When the owner requests a new certificate it shows up in the certificate transparency logs again.
Domains with less than 3 labels have been removed because these domains are either the website for the service itself, e.g. `mysql.com`. The domains can also be domains that are not registered by a registrar and are routed internally in a network, e.g. `mysql.local`.
The top 15 labels of these filtered domains are listed in tables 1 and 2 for the CN and SAN field respectively.

For this research I used Google's Argon2021 certificate transparency log that contains about 1.15 billion certificates (see fig. 1). This CT log has been scraped by the University of Twente and stored on servers so it is available for usage by researchers.

Table 1. Aggregated CN labels

Label	#domains
*	44241982
www	37341025
mail	2712758
webmail	1591554
blog	1299181
webdisk	1110347
test	1093402
cpanel	1086551
dev	1072034
cpcalendars	1025990
cpcontacts	1006274
autodiscover	958598
shop	902037
api	894646

Table 2. Aggregated SAN labels

Label	#domains
*	77500809
www	75160497
mail	7699857
webmail	5693028
cpanel	4613758
webdisk	4490848
cpcalendars	4344437
cpcontacts	4328000
autodiscover	3535173
smtp	795169
pop	728355
ftp	706022
bucket	605163
m	587835

I then manually went through the list, noting down labels that indicate a service that could be a risk if exposed to the internet. These labels, with a short description and the number of domains can be found in table 3.

Thereafter, I queried a list of hosts that have the corresponding labels for these services in the domains listed in the CN and SAN fields. Like the aggregated results, duplicates and domains with less than 3 labels have been removed.

Table 3. Selected services and description

Service	Description	#domains
RDP	RDP is a remote desktop protocol developed by Microsoft	4343
SMB	SMB is a windows file sharing protocol	762
VNC	VNC is a graphical desktop sharing system	1271
Mongo	MongoDB is a document oriented database system	1996
Mongo-Express	Mongo-Express is a web interface to manage MongoDB databases	460
Elasticsearch	Elasticsearch is a search and analytics engine	7686

Following this, the IP-addresses of these domains were needed to perform a scan for open ports later, this was done with a custom Python script that resolves a list of domain names. Some of the resolves did not result in an IP-address. These results have been filtered out of the dataset.

To confirm if the service is actually running on a resolved domain name and to check if it publicly exposed, several ports were scanned according to the target service. These ports were chosen according to the ports these services are running on by default.
The scanning was done with a custom Python script that asynchronously scans a list of IP addresses and ports. The ports that have been scanned for each label can be found in table 4.

Table 4. Expected services and their default ports

Service	Expected ports
Mongo-Express	80, 27017, 8081, 8080
RDP	3389
SMB	139, 445
VNC	5800, 5900
MongoDB	27017
ElasticSearch	9200, 9300

In addition to only checking if the host has the default port for the target service opened, for some services additional things are checked.

(1) For all services that are web interfaces, I tried to connect to the host on the selected port via an http connection. If the host returns status code 200, the host-port combination is marked as accessible.
(2) For all host-port combinations where the port is 27017, a MongoDB connection is attempted with anonymous user credentials. If the login attempt is successful, a command is issued to get a list of the databases. If the login attempt and the list of databases are successful, the host-port combination is marked as accessible.

(3) Because of the small amount of domain names containing the label smb, I checked these by hand. I tried to connect using the guest user and attempted to get a list of shared folders. If this was successful, I marked the host-port combination as accessible.
(4) Because of the small amount of domain names containing the label vnc, I checked these by hand as well. I tried to connect to the host on the open port as an anonymous user without password. If this was successful, I marked the host-port combination as accessible.

5 SELECTED SERVICES

5.1 Mongo-Express

Mongo-Express is a web-based MongoDB admin interface. Database administrators can use it to view, edit, and remove databases in their connected MongoDB instance. If an attacker would be able to access this interface the damage could be catastrophic. Data from the databases could be stolen, sold, taken ransom, or just deleted. This is why it is important to secure the Mongo-Express dashboard in an adequate way.

5.1.1 Best practices. **Enable authentication / Change defaults**
Some administrators have disabled authentication for their Mongo-Express dashboards. This should be avoided at all costs, even if it is only facing the local network.
By default Mongo-Express uses the username:password combination admin:pass. [7] This should be changed to a more secure combination.

Restrict access
Access to the Mongo-Express dashboard could be configured to only be reachable from the internal network. It is then possible to connect to it from outside the network using a virtual private network (VPN). This way, if you are not connected to the VPN or not already inside the network, it is not possible to connect to the server.
This also protects the service in case a vulnerability is discovered. For example, a vulnerability with which an attacker could bypass authentication. In that case, the attacker still has no way to open a connection to the server, thus making it very hard for them to exploit this vulnerability.
If Mongo-Express has to be exposed to the internet and it is not possible to use a VPN, a firewall could be configured to only allow connections from a specific IP-address range.
This solution of isolating a service to the local network and only making it accessible to users with the right credentials or IP-address can be applied to all the following services listed.

5.2 Remote Desktop Protocol (RDP)

Remote desktop protocol, abbreviated to RDP, is a technical standard developed by Microsoft to access a desktop computer remotely. [8] It was introduced in 1998 as a part of Windows NT Server 4.0. [9] Instead of sending a video stream of the desktop to the client, it only sends the data required to render the screen on the client. With this technique the bandwidth required for the connection is substantially reduced.

5.2.1 Best practices. Because the RDP server is essentially a gateway for clients into the internal network of a company or organisation, it is important to enforce strict security measures on the incoming connections.[10] Because the software allows users to log into their accounts remotely, RDP is susceptible to social engineering.
If a user's credentials are compromised and the RDP server does not employ additional security features, an attacker can remotely log into the internal network of an organisation.

In addition to this, if an exploit is found in the RDP server software, attackers could exploit this to gain access to the server. This is what happened with a vulnerability called BlueKeep,[11] first reported in May 2019, and it's derivates collectively known as DejaBlue,[12][13] which were reported in August 2019.

These exploits allow an attacker to execute arbitrary code on the system running the RDP server, without needing any form of authentication whatsoever.
These risks can be mitigated by employing additional security measures, besides simple username / password authentication.

Multi-factor authentication (MFA)
By requiring multi-factor authentication for users attempting to log in, social engineering attacks become increasingly complex for attackers to execute. Multi-factor authentication requires the user to not only supply a password, but also another piece of information only the user can have/know.

5.3 Server message block (SMB)

The Server message block protocol (SMB) is a network file sharing protocol. It was originally developed at IBM but its first major release was in Microsoft Windows. It is used in Windows to create and connect to network drives and shared folders. [14]

5.3.1 Best practices. SMB has been the attack vector of choice for a number of high profile hacks in the past 10 years. For example, the WannaCry ransomware outbreak in 2017 used an exploit in Windows SMB server called EternalBlue to spread. [15] The exploit was likely discovered by the United State's National Security Agency and got stolen and released by a group of hackers called The Shadow Brokers. [16]
Like BlueKeep, mentioned in section 5.2, EternalBlue does not require any form of authentication to succeed. By sending a specially crafted packet to the server it is possible to execute arbitrary code remotely.[15] Furthermore, because the Windows SMB server is run as a kernel level service, the attackers get access to kernel level control over the machine.

Disabling SMB version 1 Development of the earliest SMB protocol started in early 1983 and Microsoft's implementation started in 1990. This makes the protocol more than 30 years old at this point. Several newer versions of the protocol have been released that add encryption, better authentication, and better performance.

Microsoft has deprecated SMBv1.0 in 2013 and it is no longer installed by default on new Windows installations.[17]

5.4 Virtual network computing (VNC)

VNC is a system that allows a users to control a desktop remotely with the remote framebuffer protocol (RFB). This is done by sending a video stream of the screen to the users, and relaying the mouse and keyboard inputs from the user back to the remote.[18]
A VNC system using the plain RFB protocol without any extensions is not very secure. The passwords and encryption keys can be sniffed from the network, but this requires an attacker to be able to intercept a successful connection. Furthermore, some versions of VNC only support passwords with a maximum length of 8 characters.

5.4.1 Best practices. **Avoid plain VNC**

A VNC system that only uses the RFB protocol will result in an unencrypted video stream over the internet, this makes it possible for an attacker to intercept it and observe the stream. This can be avoided by sending all VNC traffic through an secure shell (SSH) tunnel. In addition to this, extended versions of VNC exist that allow for passwords longer than 8 characters and integrations with authentication providers like Microsoft Active Directory.

5.5 MongoDB

MongoDB is a document oriented database program. Because MongoDB does not require authentication at all by default, administrators should set this up themselves. If the person responsible for the database neglects to setup authentication on the server, everyone is able to connect to the server.
This is potentially a large security risk and that is why I selected it as a potentially vulnerable service.

5.5.1 Best practices. Like the other services, isolating the service from the internet and making it only accessible to the local network is a good option to increase security. In addition to this, setting up authentication is necessary, even if the server is only exposed to the internal network.

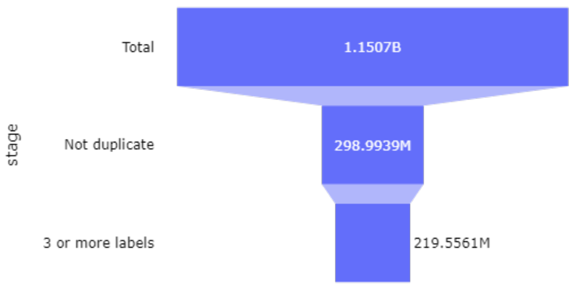
6 RESULTS

6.1 Collected labels

From the 1.15 billion records in the University of Twente certificate transparency database, after removing duplicates there are about 299 million unique certificates left. Furthermore, only domains that are composed of 3 or more labels are considered, e.g. `service.example.com` is counted, but `example.com` is not. This leaves approximately 220 million unique certificates. This can be seen in figure 1.

The first labels of these domains were then aggregated. The top 15 labels of these aggregated results can be found in table 1 and 2 for the CN and SAN respectively.

Fig. 1. Remaining number of certificates after filtering steps



6.2 Results per service

In the following section, the format for the figures for all services is identical.
The stage `domains` means the number of domains with that label, after filtering and duplication removal.
The stage `resolved` means the number of domains that got successfully resolved to an IP address.
The stage `port open` indicates how many host-port combinations are open. Keep in mind that because some labels are scanned on multiple ports, and that a host is marked as open if at least one host-port combination is marked as open. For example `mongo-express` hosts are scanned on 4 different ports, if one of them is marked as open, the host is marked as open. The stage `can connect` is only available for some services. It means that the additional check like opening a connection to a MongoDB instance succeeds. If this bar is not there, there are either no additional checks implemented for that label or none of them succeeded.

6.2.1 Elasticsearch. To find running Elasticsearch services, two labels were used. `elastic` and `elasticsearch`. The results can be seen in figure 2 and 3 for the CN and SAN fields respectively. I was unable to connect to any of the domains in the SAN field on port 9200, which is the default for Elasticsearch.

Notable findings
By manually inspecting the list of host-port combinations that returned a `200 OK` HTTP status code when approached on port 9200, I found some interesting results. I will not specifically name the domains because they have not yet fixed the problem. More on these considerations can be found in Firstly, I found the Elasticsearch instance of a multiple web shops. The largest of which has over 18 thousand product records.
Furthermore, I found a database that contains a lot of 'devices' and events from these devices, the total number of devices is over 50 thousand and the number of events is over 200 thousand. I do not know what the purpose of this database is, there are no other services running on the host.
In addition to this, I found 12 instances containing only a message from someone to the owner of the database. All of them are similar in that the intruder removed all of the indices but made a backup

they are willing to restore for some amount of bitcoin.
I also found a personal portfolio and the server of a web development company that are also exposed without any form of authentication. And finally, I the Elasticsearch instance of a company that specializes in chat bots for websites to help users. This instance contained a lot of chat messages between bots and users.

6.2.2 Mongo. The results for the CN and SAN field for the label `mongo` can be found in figures 4 and 5 respectively.
As is visible in these graphs, almost all hosts that have the port for MongoDB (27017) open, also do not have authentication enabled for their databases. This results in 9% of all hosts with the label mongo that do not have authentication enabled on their MongoDB instance.

6.2.3 Mongo-Express. The results for the CN and SAN field for the label `mongo-express` can be found in figures 6 and 7 respectively.
The number of host-port combinations in the can connect bar of these graphs is a little misleading. When manually inspecting the hosts, I noticed that there often was no instance of mongo-express running. Instead there was some other service that uses the same port(s). Because a host-port combination is marked as connectable if the HTTP request returns `200 OK`, other HTTP services running on the same ports influence these results.

Notable findings
Firstly, as with Elasticsearch, several of the databases are empty except for a ransom note.
Secondly, I found a database that belongs to an IT contractor that is building some sort of system for a large European city. It contains images from traffic cameras, air sensor data, shared cars, and more. I assume this Mongo Express instance is only used for development of the platform, because there also is a label dev in the domain name. Even though it is only used for development, developers should still ensure that the security is up to standards. Certainly if they are using real data.

6.2.4 RDP. The results for the CN and SAN field for the label `rdp` can be found in figures 8 and 9 respectively.

6.2.5 SMB. The results for the CN and SAN field for the label `smb` can be found in figures 10 and 11 respectively.
I did not make an additional check that confirms if anonymous connections are possible. This was not needed due to the low amount of results with an open port. I checked these results manually. Please note that I did not look at the actual files in the accessible shares, I only attempted to login to the server and retrieve a list of shares.

Notable findings
The findings for this label are not very interesting, it is mostly composed of home users who want to share some files with their network.
The most interesting share has two shares, one for .torrent input files and one for the downloaded output. This information was in the description for the shares.

6.2.6 VNC. The results for the CN and SAN field for the label `vnc` can be found in figures 12 and 13 respectively.
Because the amount of hosts with the VNC ports open is fairly low, I checked the results by hand. In the end I only found a single host that allowed for anonymous access. After checking the host on Shodan, I found out that the host hosts not only a VNC server, but also a Plex server and a Minecraft server.

6.3 Comparison

As can be seen in the figures in Appendix B and tables 5 and 6, there are differences in the security of domains in the CN and SAN fields. Domains in the CN field have ports open more often than domains in the SAN field. And domains in the CN field are also accessible more often than domains in the SAN field.
In addition to this, there are big differences between the services themselves. For MongoDB, almost all hosts that have the default port opened also do not have authentication enabled, (96.68%, 88.24% for CN, SAN resp.).

7 CONCLUSION

To answer my research question, the subquestions must be answered first. Because I was unable to use internet-wide scan data I had to scan hosts myself, more on this in section 8. Because of this there was not enough time left to answer sub question 1.

2. Which domain name labels that indicate sensitive applications can be identified
This question has been answered earlier, a list of these labels can be found in table 3.

3. Which services can be confirmed to be running by combining the domain names from certificate transparency logs with internet wide scan data?
Due to budgetary and time constraints, I was unable to utilize internet wide scan data. Instead I opted to scan the domains I needed myself.
Combining the labels with port scans and connection attempts greatly improves the accuracy of the results. As can be seen in Appendix B and tables 5 and 6, a large amount of the hosts do not have the default port opened, and even less are accessible.

This leads us to the main research question:
Which sensitive applications and personal information can be found by analyzing domains in the SAN and CN fields of certificates gathered from CT logs by the University of Twente?
As show in section 6, quite a lot of data is being leaked through certificate transparency logs. Not directly though the domain names but through the services that are exposed on them.

8 DISCUSSION

Originally I had planned to use an internet scan database like Censys[19] or Shodan instead of scanning all of the hosts myself.

Sadly Censys only allows for 25 thousand queries per month or access to the Google BigQuery dataset where you have to pay Google if you want to run a query.

I think an internet scan database would greatly improve the quality of the results because it contains port scan data from many ports, not just the default ports for the targeted service. This would mean that services that use a different port than the default one also get found.
In addition to this, internet scan databases store the response for all open ports. This can be used to determine if the service is actually running on that port or if it is something else. This would alleviate the issue I had with the label `mongo-express`.

Furthermore, this research can be adapted and improved to work on a realtime certificate stream like CertStream. This way less processing power is required because not all certificates have to be searched in order to find the target labels, it is simply a matter of time. Moreover, a system like this could alert operators early about possible misconfigurations instead of retroactively as is the case in this research.

As of the time of writing this, I have not contacted any of the owners of the vulnerable domains I found. I have started with collecting contact information and will keep working on this after publication of this research. My policy for informing operators and measures for not exposing found data any further can be found in appendix A

REFERENCES

[1] A. Freier, P. Karlton, Netscape Communications, P. Kocher, and Independent Consultant, “RFC 6101 - The Secure Sockets Layer (SSL) Protocol Version 3.0,” 2011. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc6101

[2] E. Rescrola and Mozilla, “RFC 8446 - The Transport Layer Security (TLS) Protocol Version 1.3,” 2018. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc8446

[3] Dutch Ministry of Justice and Security, “Frauduleus uitgegeven beveiligingscertificaat ontdekt,” 8 2011. [Online]. Available: https://web.archive.org/web/20120209233348/http://www.govcert.nl/dienstverlening/Kennis+en+publicaties/factsheets/factsheet-frauduleus-uitgegeven-beveiligingscertificaat-ontdekt.html

[4] B. Laurie, A. Langley, and E. Kasper, “draft-laurie-pki-sunlight-00,” 9 2012. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-laurie-pki-sunlight-00

[5] B. Laurie, A. Langley, E. Kasper, and Google, “RFC 6962,” 6 2013. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc6962

[6] N. Demir, T. Urban, K. Wittek, and N. Pohlmann, “Our (in)Secure Web: Understanding Update Behavior of Websites and Its Impact on Security,” 2021.

[7] “mongo-express/mongo-express: Web-based MongoDB admin interface, written with Node.js and express.” [Online]. Available: https://github.com/mongo-express/mongo-express

[8] Microsoft, “Understanding Remote Desktop Protocol (RDP) - Windows Server | Microsoft Docs.” [Online]. Available: https://docs.microsoft.com/en-us/troubleshoot/windows-server/remote/understanding-remote-desktop-protocol

[9] ——, “Microsoft Releases Windows NT Server 4.0 Terminal Server Edition - Stories,” 6 1998. [Online]. Available: https://news.microsoft.com/1998/06/16/microsoft-releases-windows-nt-server-4-0-terminal-server-edition/

[10] J. Ringold, “Security guidance for remote desktop adoption - Microsoft Security Blog.” [Online]. Available: https://www.microsoft.com/security/blog/2020/04/16/security-guidance-remote-desktop-adoption/

[11] Microsoft, “CVE-2019-0708 - Security Update Guide - Microsoft - Remote Desktop Services Remote Code Execution Vulnerability,” 5 2019. [Online]. Available: https://msrc.microsoft.com/update-guide/en-US/vulnerability/CVE-2019-0708

[12] ——, “CVE-2019-1182 - Security Update Guide - Microsoft - Remote Desktop Services Remote Code Execution Vulnerability,” 8 2019. [Online]. Available: https://msrc.microsoft.com/update-guide/en-US/vulnerability/CVE-2019-1182

[13] ——, “CVE-2019-1181 - Security Update Guide - Microsoft - Remote Desktop Services Remote Code Execution Vulnerability,” 8 2019. [Online]. Available: https://msrc.microsoft.com/update-guide/vulnerability/CVE-2019-1181

[14] ——, “Microsoft SMB Protocol and CIFS Protocol Overview - Win32 apps | Microsoft Docs,” 7 2021. [Online]. Available: https://docs.microsoft.com/en-us/windows/win32/fileio/microsoft-smb-protocol-and-cifs-protocol-overview

[15] D. Y. Kao and S. C. Hsiao, “The dynamic analysis of WannaCry ransomware,” *International Conference on Advanced Communication Technology, ICACT*, vol. 2018-February, pp. 159–166, 3 2018.

[16] S. B. Wicker, “The Ethics of Zero-Day Exploits—: The NSA Meets the Trolley Car,” *Commun. ACM*, vol. 64, no. 1, pp. 97–103, 12 2020. [Online]. Available: https://doi.org/10.1145/3393670

[17] Microsoft, “SMBv1 is not installed by default in Windows 10 version 1709, Windows Server version 1709 and later versions | Microsoft Docs,” 2 2021. [Online]. Available: https://docs.microsoft.com/en-us/windows-server/storage/file-server/troubleshoot/smbv1-not-installed-by-default-in-windows

[18] T. Richardson, Q. Stafford-Fraser, K. R. Wood, and A. Hopper, “Virtual network computing,” *IEEE Internet Computing*, vol. 2, no. 1, pp. 33–38, 1998.

[19] Z. Durumeric, D. Adrian, A. M. Bailey, Michael, and J. A. Halderman, “A Search Engine Backed by Internet-Wide Scanning,” in

Table 5. Shares of resolved domains with open ports and possibility to connect unauthenticated for the domains in the CN field

Service	Port open [% of resolved]	Access [% of open]
Elasticsearch	10.78%	15.54%
Mongo-Express	90.49%	23.72%
MongoDB	10.7%	98.68%
RDP	7.04%	0.0%
SMB	4.98%	0.0%
VNC	4.62%	0.0%

Table 6. Shares of resolved domains with open ports and possibility to connect unauthenticated for the domains in the SAN field

Service	Port open [% of resolved]	Access [% of open]
Elasticsearch	13.63%	0.0%
Mongo-Express	80.49%	3.03%
MongoDB	6.51%	88.24%
RDP	7.36%	0.0%
SMB	1.33%	0.0%
VNC	30.1%	0.0%

A ETHICAL CONSIDERATIONS AND POLICY FOR INFORMING VULNERABLE HOSTS

Because accessing remote systems that may contain sensitive data is required for this research., some ethical considerations have to be taken into account.

(1) I do not publish any of the domains with open ports or anonymously accessible services.
(2) The information listed in the notable findings sections for different services describes the data found but is not directly traceable to a specific domain. For example, the 'large European city'.
(3) Any information found during the manual checking of hosts is not stored, noted, or published.

These considerations make it possible for me to publish this research without compromising the safety of this data any further.
Because I have access to a list of vulnerable domains, I try to inform the operators of these domains of their security issues. However, because of the GDPR directive, contact information of the owners of domains can no longer be retrieved with a whois request. This significantly hampers my ability to contact thousands of vulnerable domains.

For the domains that I manually inspected and found to be containing accessible data, I tried to find a way to contact the owner. This is done by looking for a webpage containing contact information. For example for `mongo.example.com` I visit `example.com` and try to find contact information there. If there is no http server on the base domain, I use Shodan to find http servers on different ports of the host.

For the domains that I did not manually inspect but did find to be vulnerable I will execute whois requests to find out the registrar information of the domain. The domains will be aggregated according to their registrar. This way I can send one email per domain name registrar asking them to inform the owners of the domains of their security issues.

B GRAPHS

Fig. 2. Number of exposed services for labels elastic and elasticsearch in the CN field

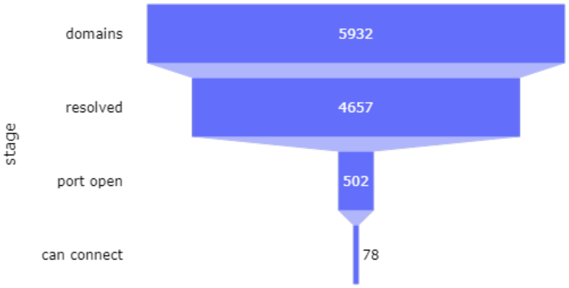


Fig. 3. Number of exposed services for labels elastic and elasticsearch in the SAN field

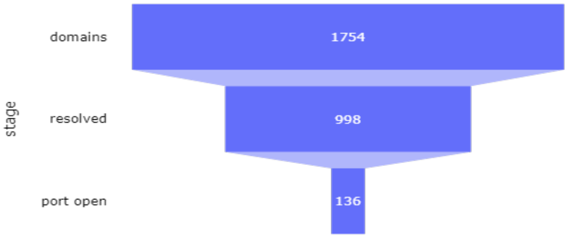


Fig. 4. Number of exposed services for label mongo in the CN field

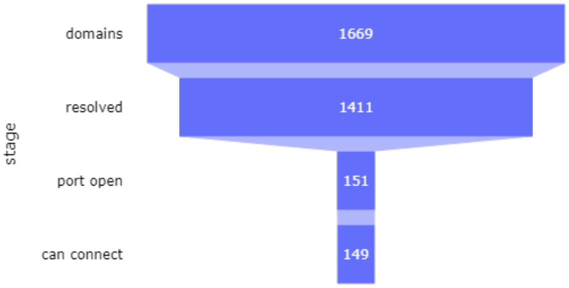


Fig. 5. Number of exposed services for label mongo in the SAN field

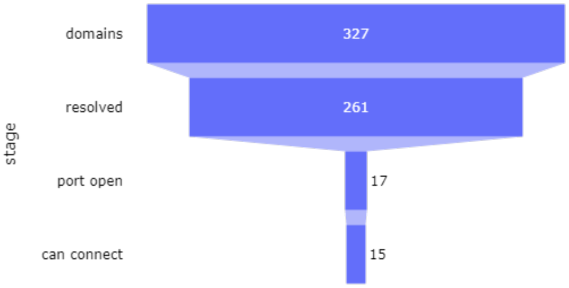


Fig. 6. Number of exposed services for label mongo-express in the CN field

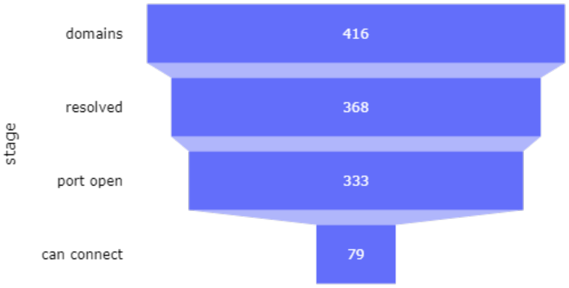


Fig. 7. Number of exposed services for label mongo-express in the SAN field

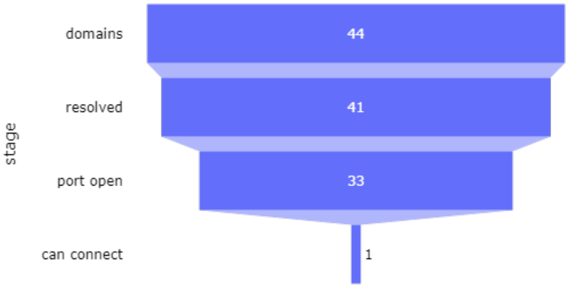


Fig. 8. Number of exposed services for label rdp in the CN field

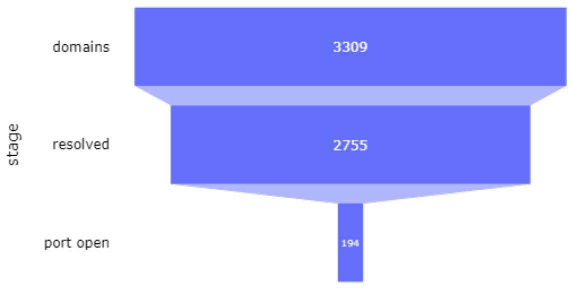


Fig. 9. Number of exposed services for label rdp in the SAN field

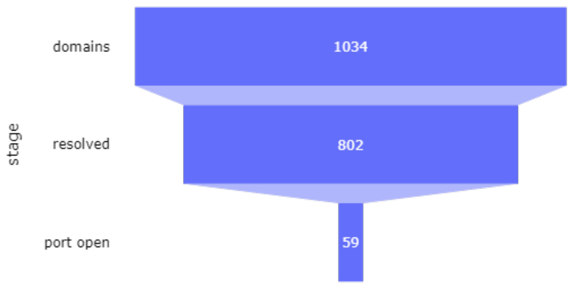


Fig. 10. Number of exposed services for label smb in the CN field

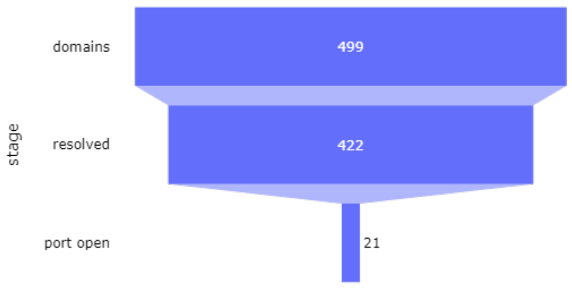


Fig. 11. Number of exposed services for label smb in the SAN field

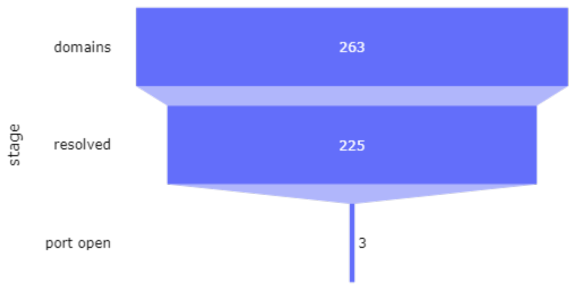


Fig. 12. Number of exposed services for label vnc in the CN field

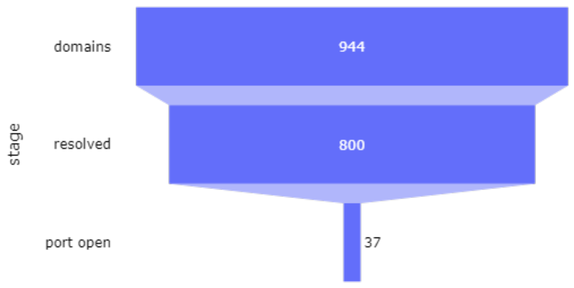
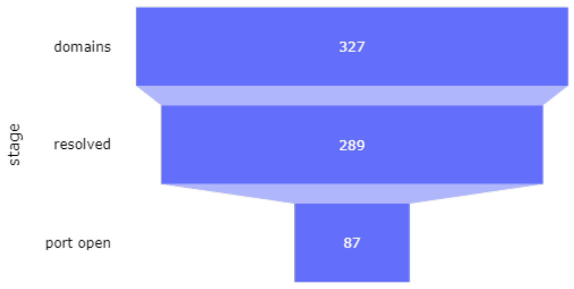


Fig. 13. Number of exposed services for label vnc in the SAN field


stage
domains
resolved
port open