UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

NUEVAS ESTRATEGIAS DE ANÁLISIS DE DATOS DE ESCANEOS DE LA RED CHILENA PARA EL MONITOREO PERIÓDICO DE SU SEGURIDAD

TESIS PARA OPTAR AL GRADO DE MAGISTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

EDUARDO ANDRÉS RIVEROS ROCA

PROFESOR GUÍA:
ALEJANDRO HEVIA ANGULO

MIEMBROS DE LA COMISIÓN:
JAVIER BUSTOS JIMÉNEZ
ÉRIC TANTER
PATRICIO GALDAMES SEPÚLVEDA

SANTIAGO DE CHILE
2020

Resumen

Las amenazas informáticas dirigidas a computadores conectados a Internet generan problemas de gran impacto en nuestra sociedad, debido a la importancia que esta red tiene en nuestras vidas, la cual se observa en el carácter cada vez más sensible de los datos personales almacenados y las operaciones realizadas en estos dispositivos. Por lo tanto, es importante que los administradores de sistemas se mantengan informados de la existencia y alcance de vulnerabilidades que puedan ser aprovechadas por algún agente malicioso, tomando las medidas necesarias para parcharlas y resguardar la información que manejan.

Una forma de aportar a la prevención de estos ataques es a través del monitoreo activo (escaneos de puertos/protocolos) y pasivo (*honeypots y darknets*) de Internet, y en particular, el escaneo en búsqueda de dispositivos vulnerables a amenazas conocidas. La tecnología existente en la actualidad permite realizar escaneos sobre subredes de miles de millones de IPs versión 4 en unas pocas horas, por lo que monitorear el espacio de dispositivos entero asignado a un país o región en específico es hoy factible. Sin embargo, la sola recolección y revisión de estos datos permite encontrar un conjunto reducido de estos problemas, por lo que se requiere de estrategias más sofisticadas que las actualmente usadas para identificar problemas claves en seguridad y asegurar una buena salud en el ecosistema de dispositivos conectados a las redes estudiadas.

Este trabajo propone dos grupos de estrategias novedosas para el uso de datos de monitoreo de red, que pueden ayudar a prevenir y detectar problemas de confiabilidad, resiliencia y seguridad de subredes de Internet. El primero consiste en el análisis de concentración de servicios sobre los dominios del *Top Level Domain* de un país, cuyos servicios asociados están enfocados al uso de sus habitantes, por lo que la existencia de métricas recurrentes de confiabilidad y resiliencia permite elaborar políticas que prevengan deficiencias en seguridad y disponibilidad a futuro. El segundo considera el uso de datos de reportes de vulnerabilidades de IPs de múltiples fuentes, con el objetivo de clasificar las máquinas vulnerables según cantidad y tipo de reportes recibidos. El presente trabajo analiza el impacto y la eficiencia de ambas estrategias en el contexto de la red chilena. Los procedimientos detallados, sin embargo, pueden ser replicables en otros conjuntos de subredes de gran tamaño sin problemas. Además, para la ejecución eficiente y periódica de las estrategias propuestas, se aporta con el diseño y la implementación de un sistema de código abierto que facilita la recopilación, el procesamiento y la transformación de los datos de escaneos, mejorando la calidad y rapidez de entrega de los resultados de monitoreo. Si bien el sistema está diseñado a partir de las necesidades del Laboratorio de Seguridad Computacional de la U. de Chile (CLCERT), su uso puede ser extendido a cualquier grupo interesado en la realización de esta práctica.

A mis mascotas, Lolita y Manchita

Agradecimientos

A mis padres, hermanos y abuelos, que me apoyaron y motivaron durante todo mi paso por el colegio y la Universidad.

A Kyra, por apoyarme desde siempre y darse la lata de ayudarme corrigiendo *typos*.

A mis amigos de primer año *(Queue)*, por hacer soportable el Plan Común y la transición a la rutina universitaria.

A la gente de *La Radio Integral* por su motivación, energía y entusiasmo en revivir y mantener un proyecto tan especial como lo es una radio estudiantil.

A todos los participantes de *Comunidad Felicidad* por las ganas de hacer cosas distintas en la facultad y por querer una pelota gigante.

A la gente del *Diario Integral* por apoyar durante harto tiempo con una disciplina tan distinta a lo que acostumbrábamos a hacer en la U, pero a la vez tan necesaria, como lo es el periodismo estudiantil.

A los equipos del *Área de Infotecnologías* y del *Centro de Tecnologías UCampus*, por todo lo que aprendí con ellos durante mis prácticas y pasantías.

A la gente de la *salita del DCC* por la compañía y amistad, y en especial a mi equipo del *Centro de Alumnos del 2017*, por su dedicación a llevar el trabajo de representar y organizar a nuestros compañeros de carrera, algunos incluso hasta el día de hoy.

A la gente del *CLCERT* por lo aprendido en temas de seguridad informática y un poco de criptografía. En especial al Profesor Alejandro Hevia por la disposición, la paciencia y el entusiasmo en cada idea surgida durante la tesis y otros proyectos.

Al sitio web *Flaticon* por la mayoría de los íconos usados en las figuras de este trabajo.

A los estudiantes de *HackCC* que me ayudaron a notar lo importante que sigue siendo para los estudiantes del departamento la seguridad informática.

A la gente de *NIC Chile Research Labs* por el grato ambiente, los juegos de mesa y las oportunidades de aprendizaje, conocimiento y crecimiento académico.

Y a mis compañeros/as de magister. Avanzar en la tesis en grupo es lo mejor.

Tabla de Contenido

Introducción **1**
1. La Internet . . . 1
   1.1. Capa de Red y Protocolo de Internet (IP) . . . 2
   1.2. Capa de Transporte y protocolos TCP y UDP . . . 5
   1.3. Capa de Aplicación . . . 8
2. Estrategias para el uso de datos de escaneo de la red chilena . . . 13
   2.1. Estructura del trabajo . . . 13
   2.2. Hipótesis . . . 15
   2.3. Objetivo General . . . 15
   2.4. Objetivos Específicos . . . 15
   2.5. Contribuciones . . . 15

1. Antecedentes **17**
1.1. Medición de Internet para Monitoreo de Seguridad computacional . . . 17
   1.1.1. Mediciones Activas y Pasivas . . . 18
   1.1.2. Técnicas Generales de Medición en Seguridad . . . 19
   1.1.3. Técnicas Específicas de Medición en Seguridad . . . 21
   1.1.4. Uso de Mediciones en conjunto con otros datos de Seguridad . . . 24
   1.1.5. Dificultades Generales de Mediciones y Escaneos de Internet . . . 24
   1.1.6. Ámbito de la Investigación Realizada . . . 25
1.2. Organizaciones y Software Orientados a Monitoreo de Internet . . . 25
   1.2.1. Organizaciones . . . 25
   1.2.2. Software . . . 26
   1.2.3. Ámbito de Diseño e Implementación de Software realizado . . . 28
1.3. La "Red Chilena" en cada contexto . . . 28
   1.3.1. Según Sistemas Autónomos . . . 28
   1.3.2. Según relevancia para usuarios chilenos . . . 30
   1.3.3. Según servicios asociados a dominios .CL . . . 30
   1.3.4. Según información de proveedores externos . . . 31
   1.3.5. Definiciones a usar . . . 32

2. Revisión Preliminar de Datos de Escaneo manejados **33**
2.1. El CLCERT . . . 33
   2.1.1. Datos Manejados . . . 33
2.2. Análisis Preliminar de Datos Manejados . . . 35
   2.2.1. Datos del CLCERT . . . 35

2.2.2. Datos de Censys . . . 43
2.2.3. Escaneo de Protocolos, Software y Versiones . . . 46
2.2.4. Datos de Malware Fuentes Reservadas . . . 48
2.3. Dificultades . . . 56
2.3.1. Existencia de lagunas de datos en algunos intervalos de tiempo . . . 56
2.3.2. Existencia de datos con alta varianza en cortos intervalos de tiempo . 57
2.3.3. Diferencias de resultados entre datos de distintas fuentes . . . 57
2.4. Conclusiones . . . 61

3. Análisis de Concentración de Servicios Dependientes del ccTLD chileno 62
3.1. Antecedentes del estudio . . . 62
3.1.1. NIC Chile . . . 63
3.1.2. Servicios a estudiar . . . 64
3.1.3. Datos a utilizar y recopilar . . . 64
3.2. Herramientas usadas y desarrolladas . . . 66
3.2.1. Escaneo usando ZMap y Mercury . . . 66
3.2.2. Importación a OSR . . . 66
3.2.3. Procesamiento de datos . . . 67
3.3. Análisis de datos recopilados . . . 67
3.3.1. Análisis sobre todos los dominios chilenos . . . 68
3.3.2. Análisis de FQDNs Gubernamentales . . . 77
3.4. Consideraciones . . . 78
3.4.1. Recomendaciones . . . 78
3.4.2. Limitaciones . . . 79
3.4.3. Trabajo Futuro . . . 80
3.4.4. Conclusiones . . . 81

4. Estrategias de análisis de datos de Escaneo usando múltiples fuentes 83
4.1. Antecedentes . . . 83
4.1.1. Motivación . . . 83
4.1.2. Red Chilena en este contexto . . . 84
4.1.3. Estrategias Propuestas . . . 84
4.2. Comparación histórica de datos de escaneo de protocolos . . . 86
4.2.1. Datos históricos del CLCERT . . . 86
4.2.2. Datos Históricos de Censys . . . 92
4.2.3. Conclusiones de Estrategia . . . 92
4.3. Comparación de datos de escaneo de protocolos de múltiples proveedores . . . 93
4.3.1. Comparación entre ambas fuentes . . . 93
4.3.2. Uso de Tercera Fuente . . . 94
4.3.3. Conclusiones de Estrategia . . . 97
4.4. Uso de datos de abandono en estimación de máquinas vulnerables . . . 98
4.4.1. Definición de Abandono . . . 98
4.4.2. Abandono por Certificados . . . 99
4.4.3. Abandono por Software Obsoleto . . . 102
4.4.4. Conclusiones de Estrategia . . . 112
4.5. Conclusiones . . . 112
4.5.1. Limitaciones del trabajo . . . 112

4.5.2. Trabajo Futuro 116
4.5.3. Conclusiones 117

5. OSR: Un Observatorio de Seguridad para la Red Chilena 119
5.1. El Sistema Actual 119
5.1.1. Infraestructura y Procesos 120
5.1.2. Problemas 121
5.2. Diseño del nuevo sistema 122
5.2.1. Requisitos 123
5.2.2. Decisiones de Diseño 124
5.2.3. Limitaciones 126
5.3. Implementación del Nuevo Sistema 127
5.4. Funcionamiento del nuevo sistema 127
5.4.1. Máquina Principal 127
5.4.2. Archivos de Configuración 127
5.4.3. Rendimiento 129
5.5. Trabajo Futuro 131
5.5.1. Nuevos procesos 131
5.5.2. Nuevas entradas y salidas 131
5.5.3. Scheduler interno de tareas 132
5.5.4. Sitio Web con Datos Agregados 132
5.6. Conclusiones 132

Conclusión 133

Bibliografía 141

Anexos 141

Anexo A. Análisis de ccTLD sobre dominios de gobierno 141
A.1. Distribución de dominios gubernamentales 141
A.2. Concentración de dominios gubernamentales 142
A.3. Uso en dominios gubernamentales de Proveedores Conocidos 147
A.4. Infraestructura compartidas en dominios gubernamentales 147
A.5. Consideraciones sobre dominios gubernamentales 148

Anexo B. Detalles de Implementación del Sistema OSR 149
B.1. Herramientas a utilizar 149
B.2. Módulos implementados 149
B.2.1. Comandos 150
B.2.2. Conexiones Remotas 151
B.2.3. Consultas SQL 151
B.2.4. Notificaciones 152
B.2.5. Modelo de datos 152
B.2.6. Tareas 154
B.2.7. Procesos 155
B.2.8. Entradas 157
B.2.9. Salidas 158

Anexo C. Validación histórica de datos de escaneo de protocolos de Censys 159
C.1. Datos históricos de Censys 159
C.1.1. Análisis global de IPs encontradas 159
C.1.2. Continuidad de las IPs por protocolo 160

Índice de Tablas

1.1. Historial IPs según AS 29
1.2. Historial Red Chilena según GeoLite2 32

2.2. Resumen de Escaneo de protocolos del CLCERT 37
2.1. Resumen de escaneos de puertos del CLCERT 38
2.3. Resumen Escaneos de Protocolos Censys 47
2.4. Resumen Resultados Fuente #1 52
2.5. Resumen Resultados Fuente #2 52
2.6. Coincidencias Escaneo Protocolos Censys y CLCERT 59
2.7. Escaneos promedio por semana por puerto Censys 60

3.1. Dominios válidos y accesibles por tipo 67
3.2. Valores MX usados por Google 69
3.3. Ranking 5 Sistemas Autónomos con más dominios chilenos 74
3.4. Ranking 5 países con más dominios chilenos 76
3.5. Número de dominios chilenos asociados con proveedor conocido 76
3.6. Número de dominios chilenos que comparten IP entre RRs distintos 77

4.1. Comparación Resultados Tercera Fuente 97
4.2. Cantidad de máquinas con y sin certificado vencido por tipo 99
4.3. Versiones de Software revisadas 106
4.4. Cantidad de IPs asociadas a cada servicio revisado en estrategia de abandono 107

5.1. Uso de Espacio en Disco 129
5.2. Duración de Procesos en OSR 130

A.1. Ranking 5 Sistemas Autónomos con más FQDNs gubernamentales 145
A.2. Ranking 5 países con más FQDNs gubernamentales 146
A.3. Número de FQDNs gubernamentales asociados con proveedor conocido . . . 147
A.4. Número de FQDNs gubernamentales que comparten IPs entre RRs distintos 147

Índice de Ilustraciones

1. Modelo OSI vs. Modelo Internet . . . 2
2. Estructura Paquete IP . . . 3
3. Ejemplo de Delegación de Subredes . . . 4
4. Estructura de un paquete UDP . . . 5
5. Estructura de un Paquete TCP . . . 7
6. Inicio Conexión TCP Cliente-Servidor . . . 7
7. Ejemplo de delegación de FQDNs en DNS . . . 9
8. Consulta DNS Iterativa . . . 10

1.1. Escaneo Activo . . . 18
1.2. Escaneo Pasivo . . . 19
1.3. Problemas de Escaneos . . . 25
1.4. Red Chilena según Sistemas Autónomos . . . 28
1.5. Red Chilena según Rankings de Uso . . . 29
1.6. Red Chilena según TLD chileno . . . 30
1.7. Red Chilena según Servicio Externo . . . 31

2.1. Resumen Resultados Escaneos CLCERT de Puertos . . . 39
2.2. Resumen Resultados Escaneos CLCERT de Protocolos . . . 39
2.3. Comparación Escaneos CLCERT Puertos y Protocolos . . . 42
2.4. Histograma Datos de Certificados CLCERT . . . 45
2.5. Resumen Resultados Escaneos de Protocolos Censys . . . 48
2.6. Comparación Escaneos Censys y CLCERT . . . 50
2.7. Histograma Malware por Fuente #1 . . . 54
2.8. Histograma Malware por Fuente #2 . . . 55

3.1. Mapeo IPs a RRs . . . 63
3.2. Resumen Servicios a Analizar, sus RR y sus puertos . . . 65
3.3. Distribución según RRs, IPs, ASNs y Países en RRs de dominios chilenos . . . 71
3.4. Concentración de Servicios en RRs de dominios chilenos . . . 73

4.1. Ranking visibilidad de IPs CLCERT . . . 88
4.2. Comparación Histórica Escaneo Protocolos CLCERT . . . 91
4.3. Comparación Intersección Censys y CLCERT . . . 96
4.4. Histograma Máquinas con Certificado Vencido por Tipo . . . 100
4.5. IPs de Certificados vencidos versus datos de Malware bots . . . 103
4.6. IPs de Certificados vencidos versus datos de Malware bruteforce . . . 104
4.7. IPs de Certificados vencidos versus datos de Malware darknet . . . 105

4.8. Comparación diferentes niveles de obsolescencia 108
4.9. Comparación Máquinas con Software Obsoleto 110
4.10. IPs de Software Obsoleto versus datos de Malware bots 113
4.11. IPs de Software Obsoleto versus datos de Malware bruteforce 114
4.12. IPs de Software Obsoleto versus datos de Malware darknet 115

5.1. Máquinas del CLCERT 120
5.2. Procesos del CLCERT 121
5.3. Problemas al sincronizar archivos en CLCERT 122
5.4. Formatos de Reportes de Escaneos 122

A.1. Distribución según RRs, IPs, ASNs y países en RRs de FQDNs gubernamentales 143
A.2. Concentración de Servicios en RRs de FQDNs gubernamentales 144

B.1. Funcionamiento General del OSR 150
B.2. Modelo de Datos del OSR 156

C.1. Ranking visibilidad de IPs Censys 161
C.2. Comparación Histórica Escaneo Protocolos Censys 164

Introducción

En este trabajo se estudian estrategias de uso de datos de monitoreo activo y pasivo para la evaluación de la seguridad computacional de la red chilena. Estas estrategias consideran el desarrollo de una plataforma de recopilación de datos de escaneo con el objetivo de facilitar su ejecución periódica. A lo largo de este documento, se detalla el proceso de investigación y desarrollo realizado y se exponen y justifican las estrategias elaboradas.

Para apoyar la comprensión de este trabajo, esta sección entrega una descripción breve de la infraestructura de la Internet, enfocándose en las capas de red, transporte y aplicación, el espacio en el que la investigación se desenvuelve. También se detallan los objetivos, alcances y aportes del trabajo de tesis realizado.

Parte de este trabajo puede verse como una continuación y extensión de la investigación de Eduardo Acha del 2017 [1], por lo que algunas de las definiciones de esta introducción son similares a las presentadas en el documento citado. Las descripciones en esta sección, sin embargo, se focalizarán en los conceptos necesarios de manejar para comprender de mejor forma los capítulos posteriores.

1. La Internet

La *Internet* es una red interconectada global de computadores sobre la cual corren distintos servicios de comunicación. Su definición completa se puede encontrar en el RFC 1122 [55], y es posible hacer paralelos de infraestructura con la propuesta por el Modelo OSI, el cual establece un conjunto de capas, cada una con funciones específicas , las cuales permiten el funcionamiento de una red de comunicaciones en un sistema de información.

La equivalencia entre capas de los modelos mencionados se puede observar en la figura 1. Las capas OSI de Aplicación, Presentación y Sesión equivalen aproximadamente en el modelo de Internet en la capa de aplicación, mientras que las capas de transporte, red y enlace son iguales en ambos modelos. La capa física no se menciona en el modelo de la Internet especificado en el RFC 1122.

En esta sección se explicarán brevemente las capas de la Internet más importantes para comprender este trabajo. Estas son la capa de red, la capa de transporte y la capa de aplicación.

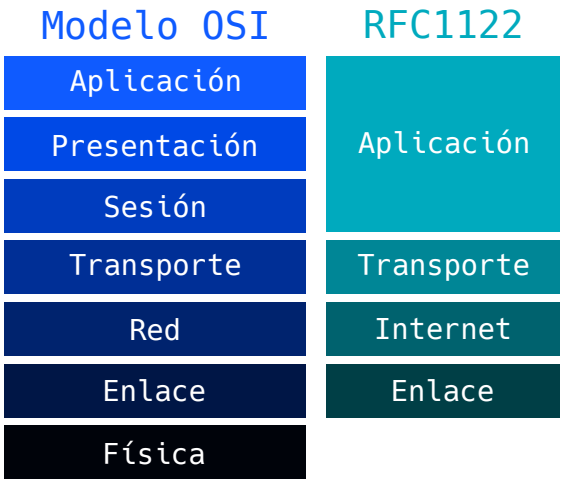


Figura 1: Comparación de las capas del modelo OSI con las del modelo de la Internet, especificadas en el RFC 1122.

1.1. Capa de Red y Protocolo de Internet (IP)

La capa de red está encargada del enrutamiento de los paquetes de datos entre distintas redes. En la Internet, el protocolo encargado de esta capa se denomina IP o Internet Protocol, el cual, dependiendo de la versión del protocolo, requiere que cada dispositivo conectado directamente a la red posea un número que lo identifica de 32 (IPv4) o 128 bits (IPv6).

En esta breve descripción, se hablará solamente de las direcciones IPv4 y del protocolo ICMP, el cual es usado para reportar errores de ruteo de paquetes en esta capa. Se omitirán descripciones acerca de cómo se enrutan los paquetes en la Internet, las cuales pueden ser consultadas en los RFCs respectivos [64].

Paquete IPv4

La estructura de un paquete IPv4 se encuentra definida en el RFC 791 [70]. Se puede ver en la figura 2, y contiene los siguientes campos:

- **Versión:** Número de 4 bits que indica la versión del paquete IP. En el caso de IPv4, este número es 4.
- **IHL:** Número de 4 bits que indica la cantidad de bloques de 32 bits que componen la cabecera. Este número debiese ser al menos 5 (en el caso en que no hayan opciones). Usándolo como desfase, permite saltar directamente a los datos del paquete IP.
- **Differentiated Services Code Point (DSCP):** Definido en el RFC 2474 [60] como reemplazo del campo *Type of Service*.
- **Explicit Congestion Notification (ECN):** Definido en el RFC 3168 [63], se traduce como *Notificación Explícita de Congestión* y permite notificar congestión de red al receptor sin necesidad de perder paquetes.
- **Largo Total:** Tamaño en bytes de los datos adjuntos en el paquete.
- **Identificación:** Valor usado para identificar fragmentos del mismo tipo, de forma de

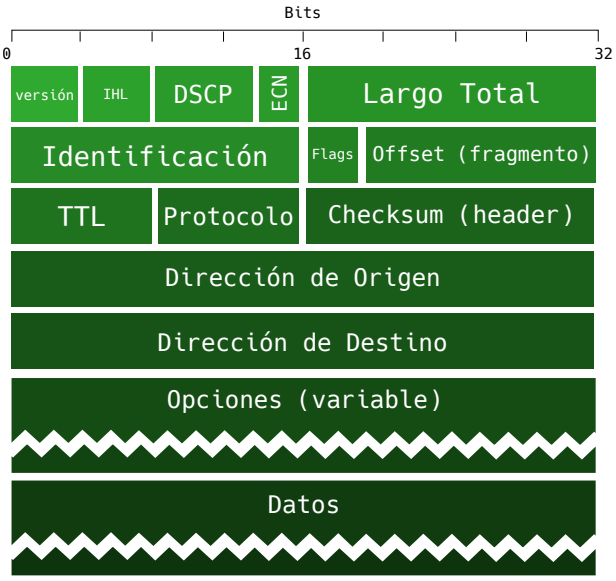


Figura 2: Estructura de un paquete IP.

permitir su reconstitución.

- **Flags:** Secuencia de tres bits utilizados como marcadores o *flags* en operaciones de fragmentación.
- **Offset de Fragmento:** Indica la posición del fragmento al momento de reconstituirse.
- **Time to Live (TTL):** Valor que es decrementado en 1 cada vez que un paquete es delegado a un dispositivo enrutador o *router*. Cuando el valor de *TTL* de un paquete es 0, se descarta.
- **Protocolo:** Protocolo usado en la capa de transporte.
- ***Checksum* del Header:** Permite realizar una verificación básica de errores del header del paquete, los cuales se pudieron haber producido por corrupción del mismo al viajar hasta su destino.
- **Direcciones de origen y destino:** Usadas para determinar el remitente y el receptor del paquete, respectivamente.
- **Opciones (opcional):** Esta sección puede o no aparecer en un paquete, y representa algunas opciones configurables que pueden especificar en parte o completamente la ruta de un paquete.
- **Datos:** Los datos enviados en el paquete IP.

Del conjunto anterior, los campos más importantes para este trabajo son tanto la dirección de origen como la de destino. Además debido al enfoque de este trabajo, se usará desde este momento el término "IP" para referirse exclusivamente a las direcciones "IPv4", a menos que se indique lo contrario.

Protocolo ICMP

El RFC 792 [71] define un tipo de mensaje especial en el protocolo IP, denominado *Internet Control Message Protocol* o ICMP. Este protocolo permite enviar mensajes de consulta o error

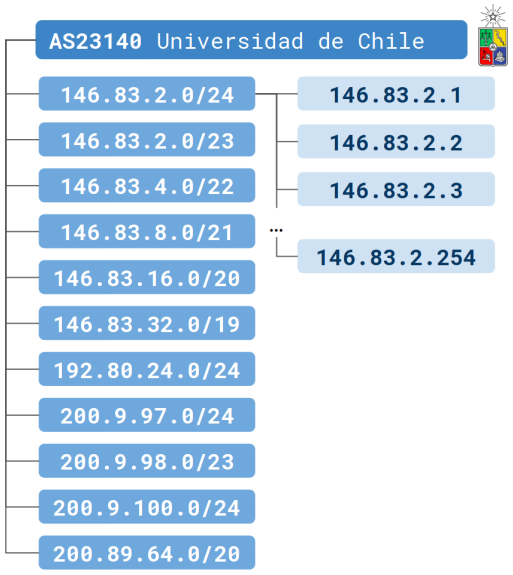


Figura 3: Ejemplo de delegación de subredes. En la imagen, el sistema autónomo 23140 tiene a su cargo las subredes mencionadas. Cada subred contempla un conjunto de IPs consecutivas, determinadas según la máscara de red usada.

de vuelta al remitente del paquete IP, en caso de problemas de enrutamiento. A continuación, algunos ejemplos de errores ICMP detallados en el RFC respectivo:

- **Destination Unreachable:** Error recibido cuando no es posible entregar el paquete al destinatario, ya sea porque no se encuentra una ruta hacia la IP de destino, porque no se maneja el protocolo o porque el paquete lleva activado *don't fragment* (no fragmentar), pero debe ser fragmentado para ser enviado a destino.
- **Time Exceeded:** Error recibido cuando un enrutador detecta que el campo TTL del paquete recibido es cero o cuando el paquete se fragmenta y su receptor no alcanza a recibir todos los fragmentos en un tiempo determinado.
- **Source Quench:** Error recibido cuando un enrutador no tiene la memoria suficiente para colocar el paquete en cola. También lo puede enviar el dispositivo de destino si los paquetes son enviados tan rápido que no alcanza a procesarlos.
- **Parameter Problem:** Error recibido cuando los parámetros de la cabecera del paquete IP no son consistentes.
- **Redirect:** Error recibido de parte de algún enrutador que solicita al remitente cambiar la ruta por la cual envía el paquete.

La diferenciación de estos errores será usada en el trabajo actual para determinar la existencia o inexistencia de máquinas o servicios asociados a una IP en particular.

Rangos de direcciones IP

Los rangos de direcciones de IP (o subredes) existentes son asignados por la IANA *(Internet Assigned Numbers Authority)* a distintos RIRs *(Regional Internet Registries)*. Cada uno de los 5 RIRs existentes en la actualidad asigna conjuntos de rangos más pequeños a ISPs *(Internet Service Providers)* y organizaciones que los requieren para usar en sus dispositivos.

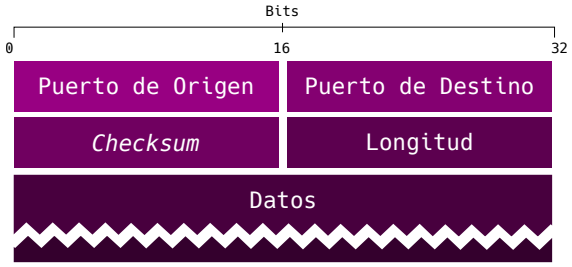


Figura 4: Estructura de un paquete UDP.

Las organizaciones que reciben estos rangos de IPs son denominadas Sistemas Autónomos o *AS* por su nombre en inglés *(Autonomous Systems)* y son designadas por un número único *(AS Number o ASN)*. Las asignaciones de subredes se suelen realizar en tamaños dependientes de las necesidades de los sistemas autónomos. Un ejemplo de asignaciones se puede observar en la figura 3.

1.2. Capa de Transporte y protocolos TCP y UDP

La capa de transporte entrega servicios comunicación de extremo a extremo a las aplicaciones, posibilitando el envío de información entre máquinas en la red de forma directa. En esta capa se encuentran dos protocolos de uso masivo en la Internet: UDP (*User Datagram Protocol*) y TCP (*Transmission Control Protocol*). Si bien existen otros protocolos, esta sección se enfocará en explicar solamente los ya mencionados, debido a su rol en el desarrollo de este trabajo.

Tanto en los protocolos TCP como UDP, la transmisión de información está asociada a *puertos*, los cuales son identificados por un número de 16 bits, y permiten diferenciar las comunicaciones de distintas aplicaciones que pudiesen estar corriendo en una máquina. Algunos puertos se encuentran estandarizados por RFCs o por uso generalizado, tema que se ahondará en la próxima sección.

El protocolo UDP se encuentra definido en el RFC 768 [69], y permite enviar *datagramas* entre distintas máquinas conectadas por IP sin necesidad de establecer una conexión, lo que lo hace liviano, pero al mismo tiempo poco confiable. La figura 4 muestra la estructura de un paquete UDP, la cual está compuesta por los siguientes campos:

- **Puerto de origen:** Número de 16 bits que permite diferenciar distintas conexiones UDP provenientes de una misma IP.
- **Puerto de destino:** Número de 16 bits que permite diferenciar distintas conexiones UDP realizadas a una misma IP.
- **Checksum:** Campo que permite realizar una verificación básica de errores del datagrama UDP, los cuales se pudieron haber producido por corrupción del paquete enviado al viajar hasta su destino.
- **Largo:** Tamaño en bytes de los datos adjuntos en el datagrama.
- **Datos:** Los datos enviados en el datagrama.

El protocolo TCP se encuentra definido en el RFC 793 [72], y se diferencia de UDP en que este protocolo se encarga de establecer una conexión confiable entre las máquinas que desean comunicarse, a costas de mayor complejidad y tamaños de segmentos más grandes. La figura 5 muestra la estructura de un paquete TCP, la cual está compuesta por los siguientes campos:

- **Puerto de origen:** Número de 16 bits que, en conjunto con otros valores, permite diferenciar distintas conexiones TCP provenientes de una misma IP.
- **Puerto de destino:** Número de 16 bits que, en conjunto con otros valores, permite diferenciar distintas conexiones TCP realizadas a una misma IP.
- **Número de secuencia:** Número incremental que identifica el orden del paquete en una conexión, de forma de que el receptor lo incorpore en la posición correcta.
- **Número de Confirmación:** Número que permite al receptor de un paquete dar aviso del próximo paquete esperado de parte del otro participante.
- **Offset de Datos:** Indica el tamaño de la cabecera del paquete en *words* (conjuntos de 4 bytes), incluyendo la sección de opciones.
- **Flags:** marcadores de 1 bit de longitud que indican características especiales del segmento.
- **Tamaño de Ventana:** Indica al receptor la cantidad de datos extra que puede recibir su emisor en el futuro, otorgando control de flujo al protocolo.
- **Checksum:** Similar al caso UDP, entrega una verificación simple del estado de validez del paquete recibido, capturando algunos casos de corrupción del segmento en caso de pérdida de información durante el viaje.
- **Puntero Urgente:** En el caso que la flag *URG* se encuentre activada, el segmento incluye información urgente desde su inicio hasta el byte indicado por este puntero.
- **Opciones (variable):** Configuraciones extra y opcionales.
- **Datos:** Los datos llevados por el segmento.

Cabe recordar que la información de tamaño de paquete TCP es calculable con la información del paquete IP que lo envuelve.

Tanto en el caso de TCP como de UDP, los conceptos más importantes para el desarrollo de este trabajo son los puertos de entrada y de salida.

Dado que el protocolo TCP es un protocolo que mantiene el estado de la conexión en ambos extremos, es necesario iniciar esta conexión antes de transmitir información con un servidor. Quien inicia la conexión es considerado como el cliente, y lo hace siguiendo los pasos mostrados en la figura 6:

- El cliente envía paquete con *flag* SYN al servidor y un número de secuencia aleatorio.
- El servidor contesta al cliente con un paquete con la *flag* SYN activada, un número de secuencia aleatorio y un ACK asociado al número de secuencia del paquete recibido.
- El cliente contesta el paquete del servidor con un ACK asociado al número de secuencia del paquete recibido, iniciándose de esta forma la conexión.

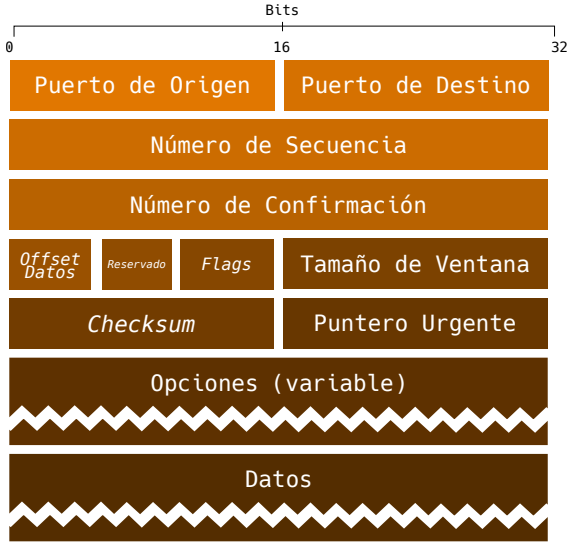


Figura 5: Estructura de un paquete TCP.

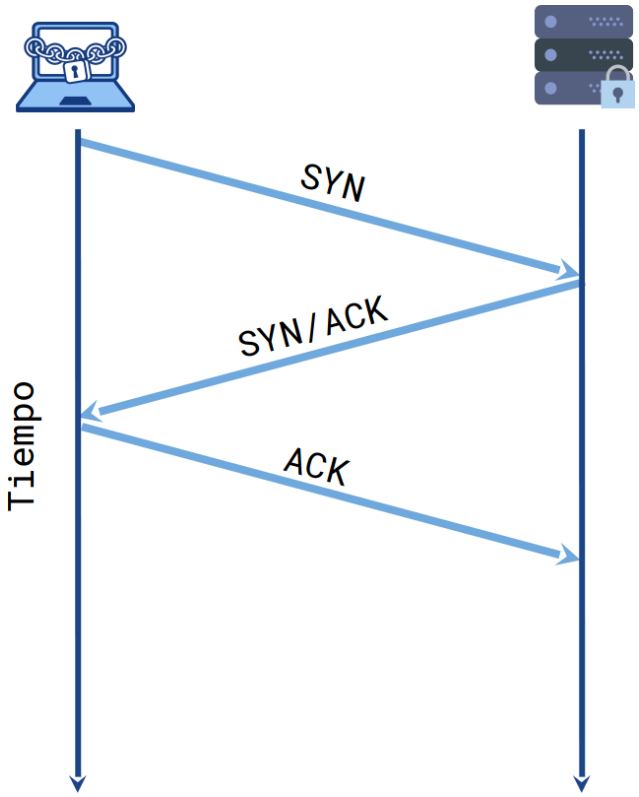


Figura 6: Pasos para realizar una conexión TCP entre un cliente y un servidor. El ícono de servidor

1.3. Capa de Aplicación

El RFC 1122 [55] define la capa de aplicación como la capa superior de la suite IP, y es en la cual corren los protocolos específicos de cada aplicación. Es importante recordar que, al compararse con el modelo OSI, esta capa agrupa aproximadamente las 3 capas superiores de ese modelo (sesión, presentación y aplicación).

Como se mencionó en la sección anterior, tanto en los protocolos TCP como UDP sobre IP se suele estandarizar el uso de distintos puertos para el inicio de comunicación de aplicaciones estándar. Esta estandarización es importante para facilitar el descubrimiento de servicios específicos en las máquinas conectadas a la Internet.

A continuación se explicará en detalle el sistema de nombres de dominio, el cual posee un rol importante en parte del desarrollo de este trabajo.

Servicio DNS

Según lo especificado en el RFC1035 [54], el sistema de nombres de dominio (o *DNS* por sus siglas en inglés) provee un mecanismo distribuido para dar nombre a recursos de la Internet, de forma que estos nombres puedan ser usados en distintas redes, familias de protocolos y organizaciones, entre otras instancias. En la práctica, el sistema de nombres de dominio permite traducir un FQDN *(Fully Qualified Domain Name)* en otro valor, el cual puede ser una dirección IP, otro FQDN o una cadena de texto.

La organización jerárquica de estos dominios parte en un conjunto de 13 servidores raíz administrados por la IANA y proveídos por varias organizaciones en el mundo. Para mejorar su disponibilidad y resiliencia, estos servidores se encuentran replicados en muchas máquinas, todas compartiendo la misma IP, en un esquema denominado *anycast*.

En segunda posición de jerarquía están los TLDs (Top Level Domains), que representan una unidad organizativa de dominios y están expresados en la subcadena de texto más a la derecha de un dominio, separada del dominio por un punto.

Actualmente, existen tanto TLDs asociados a organizaciones político-administrativas autónomas que estén definidas por el estándar ISO 3166-1 [83] como otros de uso general. Cada administrador de TLD (Nombrado a veces como NIC o Network Information Center) delega los dominios bajo él según políticas propias, y generalmente se requiere de un pago periódico para poseer el derecho de administración del dominio y todos los subdominios asociados de él.

Algunos TLDs (como el chileno) permiten el registro de SLDs *Second Level Domains* directamente, mientras que otros (como el del Reino Unido) definen SLDs de uso público en los cuales se pueden inscribir dominios (como *co.uk*). Un ejemplo de delegación se puede observar en la figura 7, en la cual se observa que el FQDN `dcc.uchile.cl` posee 4 capas de delegación distintas (IANA, NIC Chile, Universidad de Chile y Departamento de Ciencias de la Computación).

El administrador de la zona de dominio debe asignarle al dominio un servidor autoritativo, el cual estará encargado de ser la fuente oficial de información relacionada con la zona. En el

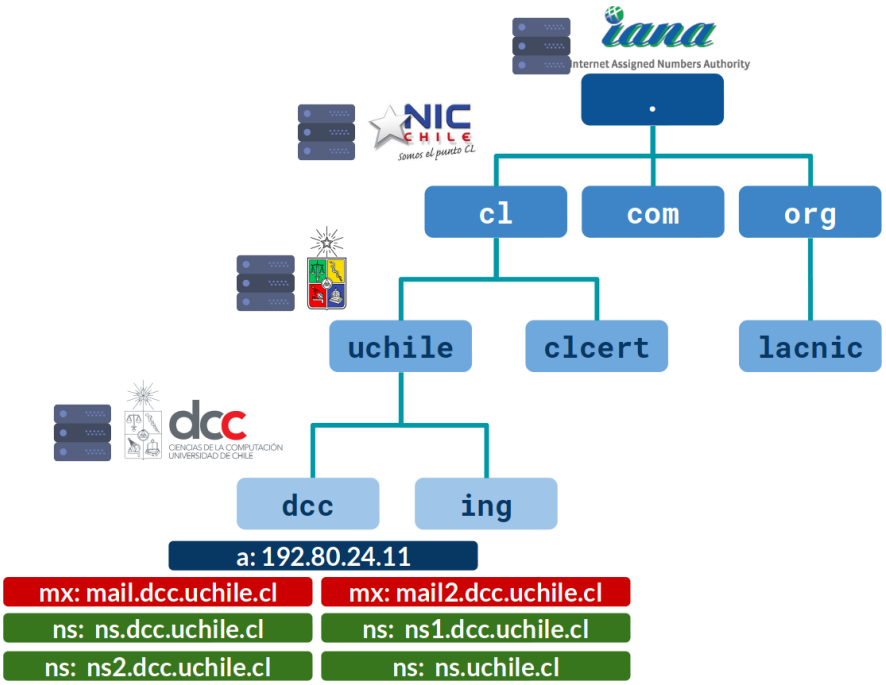


Figura 7: Ejemplo de delegación de FQDNs en DNS. La zona raíz, denotada por un punto (.), administrada por IANA, delega la administración de los TLD a organizaciones específicas. En el caso del dominio `.cl`*, éste es delegado a NIC.cl, quien a su vez delega* `uchile.cl` *a la Universidad de Chile, la cual delega* `dcc.uchile.cl` *al Departamento de Cs. de la Computación de la Universidad.*

servidor autoritativo de cada zona se almacenan *Resource Records* (o RRs), los que corresponden básicamente a un par llave-valor (contando en algunos casos con algunas propiedades extra) y definen asociaciones entre FQDNs y recursos de la Internet, tales como IPs, otros FQDNs, cadenas de texto arbitrarias, etc.

Existen muchos tipos de RRs, pero los más importantes para este trabajo son los siguientes:

- **A (y AAAA)**: Permite asociar un FQDN con una dirección IPv4 (o IPv6 en el caso de AAAA).
- **MX**: Permite asociar un FQDN con otro FQDN, de forma de delegar el manejo del servicio de entrega de correos electrónicos (mencionado brevemente más adelante) a los servidores apuntados por un registro A (o AAAA) de aquel valor. Este registro, además del FQDN, contiene un campo de *prioridad*, el cual se usa en los casos de múltiples valores de este RR, para determinar con cuál probar primero al enviar un correo electrónico.
- **NS**: define el servidor autoritativo para el FQDN al cual está asociado este registro.
- **TXT**: Asocia una cadena de texto al FQDN. Este RR se suele usar para verificaciones de dominios y para guardar información útil para otros servicios asociados al dominio.
- **CNAME**: Define un nombre canónico para un FQDN, heredando directamente todos los valores de RR del FQDN asociado.

Todos los registros tienen además un campo denominado *Time to Live* (o TTL), el cual le indica al cliente el tiempo en segundos que puede mantener en caché un resultado de consulta

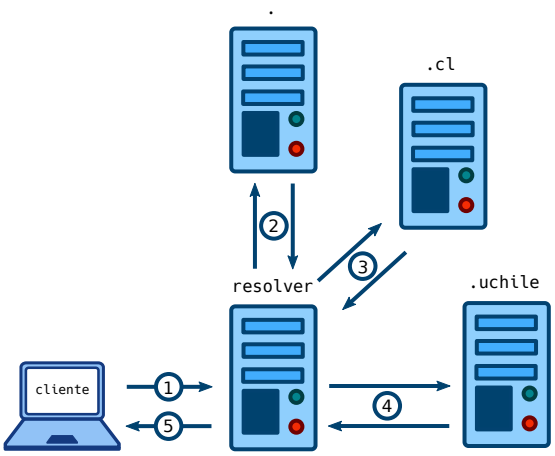


Figura 8: Infografía de proceso de consulta DNS de tipo iterativa.

DNS antes de tener que preguntarlo nuevamente. Esto disminuye la carga de los servidores autoritativos, pero dificulta la propagación inmediata en casos de actualización de los valores de los RRs.

Aparte de los servidores autoritativos, se puede contar con servidores secundarios que pueden replicar la información oficial, mejorando la confiabilidad del servicio al reducir la cantidad de puntos de fallo únicos de la zona. Estos servidores suelen ser asociados como registros extra de tipo NS.

El uso de este servicio se realiza a través de la consulta directa al puerto 53 (tanto en protocolos TCP como UDP) a máquinas conocidas como *Resolvers DNS*. Estas máquinas se conectan con los servidores DNS autoritativos y realizan las consultas de sus usuarios en su nombre. Los resolvers suelen ser configurados por defecto por el proveedor de servicio de Internet del usuario, pero en algunos casos, el usuario puede usar un resolver público, provisto por una empresa externa. Por ejemplo, la empresa de servicios de Internet *Cloudflare* dispone de una IPs de fácil memorización que funciona como resolver DNS pública: *1.1.1.1*.

El proceso de funcionamiento de un resolver puede ser recursivo o iterativo. Además, los resolver suelen implementar optimizaciones que reducen su carga, como guardar una caché de respuestas de consultas DNS existentes y no existentes.

La figura 8 muestra los pasos generalizados seguidos al realizar resolving de tipo iterativo, y los pasos se describen a continuación.

1. El cliente realiza una consulta sobre un dominio y un tipo a su resolver.
2. El resolver revisa si la consulta está en su caché, y en caso de no estar, pregunta al servidor de raíz sobre el servidor encargado del TLD. En caso de existir, el servidor entrega la dirección IP del servidor encargado del TLD.
3. Se repite el paso anterior, pero ahora sobre el TLD. El TLD entrega el servidor encargado del dominio.

4. Con la información del paso anterior, el resolver pregunta al servidor del dominio sobre la información solicitada.
5. El resolver reenvía la respuesta del servidor encargado del dominio al cliente.

Servicio de Correo Electrónico

El servicio de correo electrónico fue inventado en la década de 1970 por Raymond Tomlinson, y permite el envío de mensajes digitales entre computadores conectados a una misma red.

Desde un punto de vista técnico, el servicio de correo electrónico se puede dividir en varias partes: Una correspondiente a la transmisión del mensaje mismo (Protocolo SMTP definido en RFC 821 [73] y actualizado en los RFC 2821 [62] y 5321 [65]), otra relacionada con su ruteo (RFC 974 [74], actualizado por el RFC 2821), y otra correspondiente a su revisión remota (Protocolos POP3 (definido en el RFC 1225 [56]) e IMAPv4 (definido en el RFC 1730 [57])).

La investigación presentada en este documento requiere fundamentalmente de la comprensión del proceso de ruteo de correos electrónicos, por lo que esta sección presentará brevemente ese mecanismo.

El RFC 2821 [62] menciona que en el protocolo SMTP el ruteo es determinado a través de la búsqueda de registros MX en el dominio al cual el correo electrónico es enviado, los cuales son resueltos para determinar un conjunto de direcciones IP candidatas. En caso de no existir registros de tipo MX, pero sí al menos un registro de tipo A para el dominio, se usan esas IP.

En caso de existir más de una IP candidata para entrega del correo electrónico, se intentan éstas en el orden de prioridad establecido en el campo respectivo del registro MX.

World Wide Web y HTTP

World Wide Web (*WWW*) es un sistema de información que funciona en la Internet y que permite la obtención de documentos a partir de identificadores de recursos uniformes *URLs*. La comunicación de estos documentos entre servidor y cliente se realiza utilizando el protocolo de transferencia de hipertexto (HTTP), el cual fue desarrollado por Tim Berners-Lee a inicios de la década de 1990. Los usuarios de Internet pueden acceder al contenido entregado por *servidores HTTP* utilizando software especializado denominado Navegador.

Tanto HTTP como HTTPS (la versión encriptada del protocolo) forman parte del conjunto de protocolos más usados de la Internet en estos tiempos. La definición de una de sus versiones más populares (HTTP/1.1) se puede encontrar en el RFC 2616 [61]. Además, se está masificando el uso de una nueva versión del protocolo HTTP, denominada HTTP/2 [67], y en estos momentos se encuentra en desarrollo la versión 3 de este protocolo [48], la cual está basada en el protocolo de capa de transporte QUIC [20].

El protocolo HTTP es un protocolo diseñado para no llevar estado de las conexiones. Sin embargo, se han desarrollado estrategias para poder realizar consultas con estado definido en casos en que se requiere (como cuando no se desea pedir en cada solicitud credenciales

de inicio de sesión en un sitio que requiere autenticación), consistentes en *cookies* y *sesiones*. Además, a pesar de haber sido pensado inicialmente para entregar sitios estáticos, tecnologías relativamente modernas como *Javascript* [47] y *AJAX* [46] permiten el desarrollo de sitios web dinámicos y responsivos, de comportamiento similar al de una aplicación de computador nativa.

Con respecto a su funcionamiento, cabe mencionar que tanto las consultas como las respuestas HTTP/1.1 se componen de una cabecera y un cuerpo. En la cabecera se encuentra información como método usado (los más comunes son GET y POST, distinguiéndose fundamentalmente porque el segundo puede enviar datos en el cuerpo o *body* en la consulta y el primero no), una URL (*Uniform Resource Locator*, compuesta por el FQDN o IP y una ruta similar a la de las carpetas de un sistema operativo basado en Unix) y una lista de cabeceras o *headers*, las cuales incluyen datos como nombre y versión del software usado para servir el protocolo y "marcadores"s que facilitan el intercambio de información entre cliente y servidor (conocidos generalmente como *cookies*).

En el caso de las respuestas HTTP, además se incluye un código de respuesta, el cual puede representar que la consulta fue exitosa (códigos 200), que el recurso se encuentra en una nueva dirección (códigos 300) o que ocurrió algún error en el lado del cliente (códigos 400) o servidor (códigos 500).

Generalmente, el acceso a las páginas web se realiza a través del ingreso de una URL a la barra de direcciones de un navegador. Si la URL contiene un FQDN al inicio, es necesario realizar una consulta DNS sobre ese dominio para determinar la IP a la cual el navegador debe conectarse. En el caso de HTTP no encriptado, el navegador consulta al puerto TCP 80 del servidor por defecto, mientras que en el caso de HTTPS esta consulta se realiza por defecto al puerto 443. La máquina corriendo un servidor HTTP puede hospedar más de un sitio web, distinguiendo qué sitio entregar según el FQDN usado para realizar la consulta, el cual queda registrado como una cabecera de consulta hacia el servidor. Esta posibilidad de hospedar múltiples sitios en una misma máquina o IP se conoce como *servidores virtuales*.

Otros protocolos importantes

Además de los protocolos anteriores, varios dispositivos conectados a la Internet suelen servir los servicios de FTP (puerto 21), SSH y SFTP (puerto 22), Bases de datos (3306 en MySQL, 5432 en Postgres, etc), entre otros. El presente documento asumirá que el lector entiende en forma básica para qué se usa cada uno de los protocolos mencionados.

Software usados para proveer servicios

Dado que la mayoría de los protocolos mencionados en este documento están especificados por documentos que explican su funcionamiento, no es raro contar con más de un *software* capaz de actuar como servidor o cliente del protocolo.

La información del nombre y versión de este software en algunos casos es enviada en alguno de los pasos de comunicación del protocolo. Cuando ésta se recibe como primer mensaje, suele decirse que se encuentra dentro del *banner* del protocolo.

Más adelante se verán estrategias y técnicas donde se aprovecha la existencia del *banner* para desarrollar escaneos más avanzados.

2. Estrategias para el uso de datos de escaneo de la red chilena

La importancia de la Internet en nuestra vida cotidiana la vuelve un blanco importante de ataques informáticos, debido al cada vez mayor valor económico que tienen los procesos que ocurren en ella. Lo anterior motiva a personas, empresas y gobiernos a cuidar preventivamente este espacio, con el objetivo de limitar y prevenir algunas de las amenazas existentes en la actualidad.

Si bien existen varias formas de aportar con este objetivo, el trabajo de Eduardo Acha [1] sigue la línea de la realización de escaneos de puertos y protocolos sobre un subconjunto de la Internet denominado como chileno, de manera de tener una idea de la cantidad y variedad de dispositivos que entregan servicios a la población del país. Sin embargo, no se ha profundizado en el uso de estos datos más allá de enumeración de dispositivos, versiones y configuraciones.

Este trabajo busca extender la investigación de Acha, planteando dos nuevas categorías de estrategias de monitoreo a través del uso fundamental de los datos de escaneo recopilados por él. Estas estrategias buscan colaborar con la determinación dependiendo del contexto de subconjuntos críticos de la red chilena, sobre los cuales puede ser necesario focalizar el monitoreo para evitar problemas a gran escala, ya sea por número de usuarios afectados como por el valor de los recursos potencialmente comprometidos.

Además, estas métricas desarrolladas requieren contar con infraestructura técnica que permita automatizar, procesar y agregar los datos recopilados, por lo que este trabajo también considera el diseño e implementación de un sistema que sostenga estos análisis en el tiempo.

Finalmente, se probarán estas métricas desarrolladas usando datos de escaneos propios y externos sobre dispositivos conectados a la "Internet Chilena", esto es, sobre el conjunto de direcciones IP asociadas a Chile[1].

2.1. Estructura del trabajo

Este trabajo se compone de cuatro partes interdependientes. La primera parte consiste en la presentación de antecedentes sobre la medición de Internet en el contexto de seguridad computacional y categorías de clasificación de subconjuntos de Internet como chilena, mientras que la segunda parte revisa preliminarmente los datos históricos recopilados por herramientas de escaneo del CLCERT de la Universidad de Chile durante los últimos 3 años. Las dos últimas partes contienen los aportes principales de este trabajo de investigación, los cuales consisten en la elaboración de estrategias que ayuden a detectar vulnerabilidades y amenazas a partir del uso de datos de escaneos de puertos y protocolos activos y pasivos y el diseño y la creación de una herramienta para importar, procesar y transformar estos datos. Transversalmente, el trabajo es validado mediante el uso de las herramientas desarrolladas

[1]Como veremos en el Capítulo 1, este concepto, aunque intuitivo, no es fácil de precisar.

sobre un conjunto de datos de escaneos realizados en la Red Chilena y en el contexto del trabajo investigativo realizado por el CLCERT.

Presentación de Antecedentes

El primer capítulo de este trabajo expone conceptos que es necesario manejar para la comprensión de sus aportes, entre el cual se encuentra una explicación detallada de algunos tipos de medición y escaneo de Internet relacionados con seguridad computacional. Para cada uno de estos tipos, se presentan los trabajos previos de los cuales se obtuvieron las bases de las propuestas realizadas y se detallan algunos proyectos, software y empresas importantes en el rubro estudiado. Como aporte importante de este trabajo, se encuentra la proposición de distintos métodos para determinar el subconjunto de Internet denominado "red chilena", analizando sus ventajas y desventajas.

Análisis preliminar de datos de Escaneo

El segundo capítulo del trabajo presenta y analiza de forma preliminar los datos recopilados por las herramientas de escaneo del Laboratorio de Investigación en Seguridad Computacional y Criptografía Aplicada (CLCERT). Este capítulo sirve como una actualización del trabajo realizado por Acha en su tesis de magíster, extendiendo el análisis de los datos desde inicios del año 2017 hasta fines del año 2019. Asímismo, se realiza un análisis similar sobre los datos de escaneos de protocolos de un proveedor externo (Censys, presentado en el capítulo 1), con el objetivo de motivar una comparación en mayor profundidad de ambos resultados, ponderando sus diferencias en función del software usado para su recopilación y la frecuencia y ubicación de sus escaneos.

Elaboración de Estrategias de Análisis y Validación de Vulnerabilidades

Con respecto a la parte investigativa del trabajo, los datos manejados por el CLCERT serán utilizados para ejecutar las estrategias de detección de vulnerabilidades mencionadas a continuación:

- **Análisis de concentración de servicios asociados a dominios .CL:** Desarrollada en el capítulo 3 de este informe, esta estrategia busca determinar el estado de concentración de algunos servicios asociados a un ccTLD específico, a través de la realización de escaneos DNS y el procesamiento de estos datos, de modo de dar recomendaciones a los administradores de los dominios para evitar problemas de disponibilidad y resiliencia en caso de falla de alguno de los proveedores, y tener en cuenta qué proveedores de servicios son más importantes dentro del ccTLD, según la cantidad y tipo de dominios que dependen de su funcionamiento.
- **Análisis de datos de Escaneo usando múltiples fuentes**: Desarrolladas en el capítulo 4, estas estrategias buscan entregar distintos niveles de clasificación a las máquinas encontradas en escaneos y reportes a partir de información histórica, de fuentes de datos de otros tipos o de metadatos especiales, además de entender mejor las diferencias entre distintas fuentes que realizan los mismos tipos de escaneos, pero con distinto software y ubicación geográfica.

Un problema recurrente con los datos de escaneo manejados por el CLCERT es que estos son descargados y procesados de forma manual. Actualmente, tanto datos externos como de origen propio se recopilan corriendo scripts registrados en el *crontab* de las máquinas manejadas por el CLCERT. Los datos luego se guardan como archivos *CSV* o *JSON*, y en caso de requerir procesamiento, se crean scripts de *Shell* o *Python* específicos para cada uso.

Para solucionar los problemas de escalabilidad y mantenibilidad de la situación actual, se diseñó e implementó una alternativa de código abierto que importa, procesa y transforma los datos de las distintas fuentes que maneja el CLCERT. La exploración en detalle de los problemas de la situación actual que dan razón a la necesidad de elaboración de este sistema y el proceso de implementación de la herramienta son explicados en el Capítulo 5 del presente trabajo.

2.2. Hipótesis

Es posible proponer, diseñar e implementar estrategias de monitoreo de la red chilena, a través del uso y análisis de datos obtenidos por herramientas de escaneo de puertos y protocolos, evaluando su rendimiento de forma experimental sobre los dispositivos integrantes de la *Red Chilena*.

2.3. Objetivo General

Proponer, diseñar, implementar y comprobar estrategias de monitoreo, de la red mediante el uso y análisis de datos de escaneos activos y pasivos.

2.4. Objetivos Específicos

- Elaborar definiciones consistentes del concepto de "red chilena" justificando su importancia en el contexto de cada escaneo propuesto.
- Proponer y evaluar dos categorías de estrategias de análisis y monitoreo, utilizando datos de escaneos activos y pasivos.
- Ejecutar las estrategias propuestas mediante el uso de datos de escaneos de la red chilena tanto locales como externos.
- Desarrollar una nueva herramienta de código abierto que permita automatizar la recolección, procesamiento y transformación de información relacionada con escaneos activos y pasivos.
- Desarrollar la infraestructura necesaria para la implementación futura de un portal de acceso público con datos agregados de escaneos activos y pasivos manejados por el CLCERT.

2.5. Contribuciones

- Aportar al mundo académico con estrategias de análisis de vulnerabilidades de dispositivos y sistemas informáticos conectados a Internet.

- Contribuir al mundo del código abierto con el desarrollo de una nueva herramienta de recopilación y agregación de datos de escaneos activos y pasivos.
- Aportar con la recopilación de datos para la mejora de políticas de ciberseguridad en Chile.

Capítulo 1

Antecedentes

Esta sección menciona los antecedentes necesarios para comprender las motivaciones y el contexto del trabajo realizado. En primer lugar, se presenta un resumen del estado del arte con respecto a la realización de mediciones de Internet para el monitoreo de la seguridad computacional. Luego, se listan algunas organizaciones y herramientas trabajando con escaneos y monitoreo de Internet desde la perspectiva de investigación y seguridad computacional. Finalmente, se establecen definiciones para el concepto de "Red Chilena", algunas de las cuales son usadas en los análisis posteriores.

1.1. Medición de Internet para Monitoreo de Seguridad computacional

El carácter abierto y distribuido de la Internet permite el acceso potencial a servicios y sistemas de forma remota desde cualquier parte del mundo. Si bien es verdad que este acceso puede ser limitado según criterios como ubicación geográfica, dirección IP u horario, en diversas situaciones estas limitaciones no son aplicadas, tanto por corresponder a servicios de acceso público (como sitios web o servidores de correo electrónico) como por problemas de configuración de parte de quienes los administran.

Por otro lado, es importante notar que en la red IPv4 existen a lo más 2^{32} dispositivos conectados a Internet que pueden prestar cada servicio en un puerto determinado. Este límite está determinado por la cantidad de direcciones IPv4 diferentes que pueden existir, sin considerar rangos reservados.

Durante mucho tiempo, la realización de escaneos completos a la Internet no era posible en tiempos razonables usando *hardware* de acceso general. Sin embargo, la tecnología de hoy sí permite consultar iterativamente ciertos servicios en tiempos dentro del rango de minutos a días, según el tipo de consulta deseada. Esto da pie a la posibilidad que entidades, sin importar su motivación, realicen de forma automatizada estas consultas sobre gran parte o incluso toda la Internet IPv4. En contraste, el acercamiento anterior no funciona en redes IPv6, debido a que el espacio completo de IPs versión 6 disponible es del orden de 2^{128}, lo que es 2^{96} veces más grande que el espacio IPv4. En este caso, se han pensado en otro tipo

Figura 1.1: El Escaneo Activo envía mensajes a todos los dispositivos de una subred determinada. Se registran las respuestas de estos dispositivos con el objetivo de determinar cuántas máquinas poseen un puerto abierto o reciben consultas sobre un protocolo específico.

de estrategias, como la planteada en [42], pero que no se considerarán en este trabajo debido a la poca penetración de IPv6 en Chile en estos momentos.

Esta sección se concentrará en explicar el estado del arte con respecto a investigación, herramientas y técnicas de medición de la Internet, enfatizando en aquellas que tienen aplicaciones para el monitoreo de seguridad de las redes y sus equipos. En primer lugar, se explicará la clasificación de mediciones dependiendo de si ellas comienzan las comunicaciones o si las reciben. Posteriormente, se mencionarán algunas categorías generales de mediciones que aplican a más de un tipo de servicio o protocolo. Más adelante, se detallarán estrategias de medición y escaneo diseñadas para servicios específicos que suelen ser estudiados en términos de seguridad, disponibilidad o resiliencia, además de estrategias que intentan combinar los escaneos con fuentes de información de seguridad obtenidas por otros medios. Finalmente, se enumerarán algunas organizaciones, grupos de investigación y empresas que utilizan estas mediciones para monitorear el estado de seguridad de la Internet, se mencionarán algunas de las limitaciones conocidas de las estrategias de monitoreo y se contextualizará el trabajo realizado dentro de las categorías mencionadas.

1.1.1. Mediciones Activas y Pasivas

Dependiendo de si el escáner o el dispositivo escaneado inicia las comunicaciones en las mediciones de internet, estas se pueden clasificar en dos tipos: activas y pasivas.

Las *mediciones activas* consideran el envío de "mensajes" a los dispositivos que se desea estudiar. Lo anterior significa que a mayor cantidad de dispositivos revisados, es necesario producir una mayor cantidad de tráfico. Al mismo tiempo, cada máquina medida debe consumir recursos propios para contestar (o ignorar) las consultas realizadas, lo cual puede variar dependiendo del protocolo. La figura 1.1 ilustra este caso.

En contraposición, las *mediciones pasivas* consideran la recepción de "mensajes" de parte de dispositivos conectados a la red estudiada. En este caso, no se eligen los dispositivos que son considerados en el escaneo. Al contrario del caso anterior, en este caso no se debiese generar tráfico extra en la red observada, ni tampoco debiese aumentar directamente el consumo de recursos de las máquinas escaneadas. La figura 1.2 ilustra este caso de medición.

En ambos casos, nada limita que las comunicaciones entre el escáner y los dispositivos escaneados se mantengan. De ser así, ambos tipos de escaneo se comportan de forma similar, produciendo tráfico y consumiendo recursos dependiendo del protocolo escaneado.

Figura 1.2: En el escaneo pasivo, una máquina recibe todos los paquetes enviados por investigadores o agentes maliciosos hacia una subred en específico (flechas rojas), pudiendo o no responderlos según el tipo de configuración (flechas naranjas). Estos paquetes se almacenan para su posterior análisis.

1.1.2. Técnicas Generales de Medición en Seguridad

A continuación, se explicarán algunas categorías generales de técnicas de medición de internet orientadas a seguridad computacional. Se consideran como generales debido a que son aplicables a un gran número de servicios. Sin embargo, su falta de especialización produce que, por defecto, no entreguen una gran cantidad de información en comparación a técnicas más especializadas.

Escaneo de Puertos

El escaneo de puertos en la revisión del estado de un puerto de un protocolo de capa de transporte (como por ejemplo, TCP o UDP), de forma de determinar si éste se encuentra en uno de los siguientes estados:

- **Puerto abierto,** si la máquina que escanea logra determinar que es posible establecer comunicación satisfactoria con el puerto del dispositivo escaneado.
- **Puerto cerrado,** en el caso contrario.
- **Puerto filtrado,** si la máquina que escanea determina que existe un *firewall* entre el dispositivo escaneado y ella.

Las condiciones anteriores se pueden determinar a partir del uso de algunos de los escaneos siguientes, los cuales son detallados por la guía oficial de *Nmap* (herramienta de escaneo de redes internas) [38]:

- **Escaneo TCP SYN:** Es el escaneo más común para el protocolo TCP debido a que no requiere usar una gran cantidad de recursos computacionales en ningún lado de la conexión. Sin embargo, requiere contar con *permisos de administrador* en la máquina que realiza el escaneo. Consiste en enviar paquetes SYN al puerto de los dispositivos escaneados, los cuales de seguir correctamente el protocolo TCP, responderán con un

paquete SYN/ACK en caso de tener habilitado el puerto escaneado. Inmediatamente después de recibir respuesta del dispositivo escaneado, la máquina que escanea marca el puerto como abierto y envía un paquete RST, el cual fuerza el cierre de conexión. Si se recibe de respuesta un paquete RST, se asume que el puerto está cerrado, y si no se recibe respuesta o se recibe un error ICMP *Destination Unreachable*, se considera el puerto como filtrado.

- **Escaneo TCP ACK:** Permite diferenciar si un puerto TCP está filtrado o no, enviando un paquete con la *flag* ACK a él. Si la respuesta es un paquete con la *flag* RST, el puerto puede estar abierto o cerrado. Si no hay respuesta, el puerto está filtrado.
- **Escaneo UDP:** Permite determinar el estado de un puerto UDP, enviando un paquete vacío o con un payload definido al puerto del dispositivo escaneado. En caso de recibir un error ICMP de tipo *Destination Unreachable* y código *Port Unreachable*, el puerto se considera cerrado. En caso de recibir un error ICMP de tipo *Destination Unreachable* de código distinto a *Port Unreachable*, el puerto se considera filtrado. En caso de recibir respuesta, el puerto se considera abierto. En caso de no recibir respuesta alguna, el puerto puede estar abierto o filtrado.

Una limitación importante al uso de *NMap* es que la herramienta se encuentra diseñada para su uso en escaneos de redes locales, por lo que su uso no es apropiado para realizar escaneos de grandes porciones de IPs debido al tiempo que demora la realización de los escaneos. En este contexto, se han desarrollado herramientas optimizadas para la ejecución de escaneos en solo minutos, contando con equipamiento y recursos de costo razonable para un grupo de investigación pequeño. Estas herramientas serán presentadas en la sección *Software y Organizaciones de Monitoreo de Internet* de este capítulo.

Darknets y Honeypots

Con respecto a las *Darknets*, este tipo de escaneo pasivo no entrega respuesta alguna luego de recibir un mensaje, simulando ser una máquina con todos los puertos filtrados. Sin embargo, sí lleva registro de todos los paquetes recibidos. El trabajo mostrado en [8] explica la construcción de una *Darknet* o *Blackhole* de Internet, asociado a una subred de 2^{24} IPs contiguas, de forma de aumentar las posibilidades de detectar un escaneo completo a la Internet de parte de un tercero. Los recursos computacionales asociados a la máquina que opera la Darknet requieren que ésta sea capaz de procesar la cantidad de paquetes que está recibiendo. La mayor utilidad de este tipo de escaneo pasivo es que entrega información en tiempo real de puertos y protocolos más escaneados en cierto momento, entregando información parcial que pudiese permitir la deducción de amenazas conocidas pero no divulgadas por otros grupos de investigadores o de atacantes.

Por otro lado, una *Honeypot* es una máquina (física o virtual) configurada para emular un sistema o dispositivo vulnerable. Cuando se usa una gran cantidad de ellas, la infraestructura se denomina *Honeynet*. Estos dispositivos se conectan a Internet escuchando todo tipo de mensaje enviado a ellas asociadas a bloques de subredes como en el caso de las *darknet*. En caso de entender algún paquete recibido, ellas contestan simulando el comportamiento de este dispositivo vulnerable, por lo que requieren más recursos computacionales debido a que necesitan armar y enviar respuestas y potencialmente mantener estados de conexión. El trabajo mostrado en [80] explica en mejor detalle el funcionamiento y la puesta en marcha

de una *Honeynet*

Este último tipo de escaneo permite entender de mucho mejor forma los métodos que utilizan los atacantes para abusar y vulnerar estos dispositivos, pero su alcance queda limitado por la variedad de los servicios implementados por los servidores. Por lo tanto, si se quiere emular una mayor cantidad de dispositivos y sistemas, se requiere de una mayor cantidad de desarrollo, mantención y poder computacional, lo que puede ser restrictivo para algunas organizaciones al compararse con una *darknet*.

1.1.3. Técnicas Específicas de Medición en Seguridad

Estas técnicas y estrategias consideran un dominio en específico de aplicación en Internet, y adaptan sus consultas y respuestas al protocolo estudiado. Por lo tanto, su alcance no es tan grande como el de las estrategias generales revisadas anteriormente, pero es posible obtener mucha más información de ellas. En esta sección se revisan servicios específicos relacionados con las técnicas propuestas más adelante.

Escaneo de Protocolos y Banners

El escaneo de puertos no es suficiente para detectar potenciales vulnerabilidades en dispositivos conectados a la Internet, pero ayuda a disminuir el universo de máquinas a escanear según si mantiene abierto algún puerto asociado al protocolo revisado. En el caso de protocolos de capa de aplicación complejos, suele ser necesario emular uno o más pasos de la interacción cliente-servidor para determinar la existencia o inexistencia de una vulnerabilidad, además de mantener almacenados los estados de estas conexiones.

A continuación, se mencionan algunos ejemplos de técnicas relacionadas con escaneos de protocolos que ayudan a encontrar vulnerabilidades. Algunas de estas fueron presentadas en el trabajo de tesis de Eduardo Acha [1] ya mencionado.

- **Escaneo de versiones y aplicaciones a partir de Banners:** Algunos protocolos, como SMTP y HTTP, entregan una respuesta con información específica de la aplicación usada en el dispositivo escaneado y su versión. Un escaneo de banners permite a la entidad que realiza el proceso de escaneo recopilar estadísticas de versiones y aplicaciones usadas para atender cierto protocolo, lo cual es útil para determinar el impacto potencial de una vulnerabilidad en casos en que esta aplique solo en ciertas versiones y aplicaciones.
- **Escaneo de vulnerabilidades con Payloads especiales:** Algunas vulnerabilidades pueden ser escanadas enviando un *payload* especial a cada máquina con el puerto del protocolo abierto, de tal forma que es posible determinar si el dispositivo es o no vulnerable a partir de la respuesta recibida, sin comprometer información sensible.

Algunas herramientas usadas para realizar estos escaneos se presentan en la sección *Software y Organizaciones de Monitoreo de Internet* de este capítulo.

Mediciones Sobre Servicios DNS

Como se vio en la introducción, el servicio DNS cumple un rol fundamental en la infraestructura de los servicios de Internet más usados por las personas. Lo anterior motiva la realización de mediciones orientadas a entender los alcances de posibles problemas, enfocadas fundamentalmente en disponibilidad y resiliencia del servicio.

La investigación acerca de medición de servicios DNS a nivel de la Internet completa posee un gran número de trabajos relacionados, tanto a nivel nacional como internacional. Como uno de los trabajos iniciales, es posible mencionar el realizado por Danzig y otros el año 1992 [23], el cual explora el rendimiento del servicio DNS en intervalos de 24 horas, entregando evidencia experimental de la penalización en rendimiento entregada por algunos algoritmos populares en ese tiempo que buscaban mejorar la resiliencia del servicio.

Un proyecto activo hasta el día de hoy, desarrollado el año 2016 por van Rijswijk-Deij y otros, corresponde al observatorio OpenINTEL [50, 75], el cual escanea y almacena los resultados obtenidos sobre la inspección diaria 50 % de los dominios existentes, entre los cuales se incluyen los *TLD* `.com`, `.org` y `.net`, además de algunos TLD de países. Este sistema realiza consultas de más de una decena de tipos de RRs, al contrario de muchos de los estudios DNS mostrados en esta sección. El trabajo realizado en la Sección 3 tiene similitudes con la metodología de este proyecto, tanto en metodología como en motivaciones. Sin embargo, en el caso del trabajo actual y al contrario del ya citado, las mediciones no tienen por objetivo principal el entender la penetración de servicios *cloud* de correo electrónico.

En el caso internacional también es posible mencionar el estudio de robustez de DNS sobre *Second Level Domains* asociados a los TLDs *.org*, *.net* y *.com* [5]. Este trabajo también sigue una metodología similar al presentado en el Capítulo 3, pero con un mayor énfasis en datos históricos y sin considerar clasificaciones temáticas de los dominios revisados, enfocándose solamente en el top 1 million del ranking Alexa [33].

El trabajo más reciente encontrado en el ámbito de medición DNS corresponde a [28], el cual presenta el desarrollo y la operación de un "Observatorio de DNS", observando billones de consultas DNS de forma pasiva, desde un gran número de sondas distribuidas en la red. Este observatorio busca entender mejor tanto el tráfico de consultas del protocolo como la concentración de importancia entre servidores DNS, proponiendo además ideas para mejorar el rendimiento frente al mal funcionamiento de algunos algoritmos de eficiencia. La mayor diferencia de este trabajo con el desarrollado en el Capítulo 3 es que las mediciones en él son pasivas, lo que permite entregar resultados agregados en tiempo real.

Con respecto a investigación relacionada con el servicio DNS en Chile y Latinoamérica, se encuentran algunos estudios realizados sobre el espacio de dominios chileno en los últimos 15 años. El primero fue publicado por José Urzúa el año 2005 [88] y revisaba el cumplimiento de recomendaciones RFC en la configuración de los servidores DNS enlazados a los dominios `.cl`, incluyendo los porcentajes totales de cumplimiento por cada recomendación.

El segundo estudio latinoamericano revisado fue publicado por el Observatorio LACNIC, iniciativa llevada por *NIC Chile Research Labs* [3] que se enfocó en revisar tanto cantidad de registros NS como el estado general de la implementación de DNSSEC en un subconjunto de

los dominios administrados por LACNIC.

Mediciones Sobre Certificados SSL/TLS

La importancia que hoy en día tiene el uso de certificados SSL/TLS para el manejo de datos encriptados en distintos protocolos de Internet motiva el estudio de la penetración de éstos en distintos protocolos y los parámetros de seguridad con los que suelen ser configurados. A continuación, se nombran algunas de las iniciativas que se han levantado en el último tiempo relativas a este tipo de mediciones.

Durante el año 2010, la *Electronic Frontier Foundation* presenta en DEFCON 18 un *Observatorio SSL*, el cual se enfoca en la recopilación centralizada de un gran número de certificados SSL y TLS con el objetivo de motivar posteriores investigaciones. Sin embargo, a la fecha el proyecto fue descontinuado, y recomienda actualmente el uso de las fuentes de datos Censys [10] y crt.sh [37].

Con respecto a trabajo más reciente relacionado a mediciones en la Internet completa, es posible mencionar el trabajo de Durumeric y otros [27], y el de VanderSloot y otros [89], ambos realizados en el contexto de investigación de Censys. El primer trabajo revisa certificados HTTPS a través de escaneos con la suite *ZMap* el año 2013. El segundo trabajo entrega una vista del ecosistema de certificados TLS según los datos agregados del año 2016 de Transparencia de Certificados y de la organización Censys, trabajando y comparando 99 % de los certificados observados a través de estas fuentes.

En el caso de estudios enfocados a la red chilena, el trabajo de tesis de Eduardo Acha [1] considera el escaneo de certificados SSL/TLS sobre la red chilena durante un año, entregando resultados específicos relativos a los servicios ejecutados en nuestro país. Esta información es actualizada hasta el mes de noviembre de 2019 en el Capítulo 2 de este trabajo.

Mediciones de Vulnerabilidades en Internet

El estudio del impacto de distintos tipos de vulnerabilidades a lo largo del tiempo es de interés especial debido a que permite comprender de mejor manera el comportamiento y el impacto de malware que las abusa, además de la evolución de cómo se despliegan los parches para prevenir efectos no deseados.

Como ejemplo de detección vulnerabilidades relacionadas con mediciones a partir de escaneos de Internet, es posible mencionar un estudio por Durumeric y otros con el objetivo de entender mejor el impacto en Internet de la vulnerabilidad *Heartbleed*, relacionada con la librería criptográfica OpenSSL [24]. Esta vulnerabilidad es detectable a partir del uso de escaneos de puertos y protocolos especializados, y la herramienta *Mercury* cuenta con un módulo que permite detectar estos problemas.

Un caso de análisis reciente corresponde a la vulnerabilidad *BlueKeep* [18], descubierta en Mayo de 2019. Esta vulnerabilidad afectó en su momento a al menos un millón de dispositivos corriendo Windows con una implementación del protocolo RDP vulnerable, permitiendo ejecución de código de forma remota. De las máquinas afectadas, al menos 2.933 correspondían a máquinas en la red chilena, según estudio del CLCERT [11]. Desde el mes de septiembre,

este problema empezó a ser explotado en la Internet por distintas variantes de malware, lo que enfatiza la necesidad de detección de máquinas vulnerables de forma temprana para la prevención de problemas de seguridad mayores.

1.1.4. Uso de Mediciones en conjunto con otros datos de Seguridad

Existe trabajo relacionado al uso de datos de mediciones de Internet en conjunto con otra información que permita evaluar el estado de seguridad de las máquinas escaneadas. Debido a su enfoque en seguridad computacional, se consideró que su revisión puede ser provechosa en el contexto de este trabajo.

En la publicación de Jamie O'Hare y otros el año 2018 [49], los autores proponen la cruza de datos de escaneo de la plataforma Censys con información existente en la Base de Datos Nacional de Vulnerabilidades de Estados Unidos (NVD), manejada por el *National Institute of Standards and Technology* (NIST) del mismo país. De esta forma, se puede automatizar la detección y reporte de vulnerabilidades ya conocidas, detectando servicios con versiones vulnerables en los escaneos realizados por Censys. Específicamente, la publicación realizó análisis de *banners* de servidores web, de modo de determinar automáticamente cuántos de éstos se encontraban vulnerables a problemas de seguridad según sus versiones. Cabe destacar que esta propuesta considera solamente el uso de datos de fuentes externas, sin discutir o proponer la posibilidad de la realización de una revisión de primera fuente sobre los datos de escaneo utilizados.

Dentro de la misma línea, este trabajo cita otras investigaciones que utilizan información de plataformas como Censys y Shodan para la detección de vulnerabilidades en varios subconjuntos de redes, tales como Internet de las Cosas *(IoT, por sus siglas en inglés)* [4].

Un estudio similar al anterior, pero en el contexto de análisis de dispositivos de tipo SCADA *(Supervisory Control and Data Acquisition)*, fue realizado en por Sagar Samtani y otros [76]. Esta investigación utiliza como fuentes de datos de escaneo la plataforma Shodan, y al igual que el trabajo mencionado anteriormente, cruza estos datos con información de la NVD para determinar gravedad de vulnerabilidades según el puntaje asignado en esta plataforma.

1.1.5. Dificultades Generales de Mediciones y Escaneos de Internet

Los escaneos de Internet activos y pasivos tienen ciertas limitaciones que hay que tener en cuenta al momento de interpretar sus resultados, como las explicadas en la figura 1.3. Por ejemplo, un escaneo activo clasifica como vulnerables dispositivos que en realidad son *honeypots*, y en caso que la *honeypot* esté bien implementada, no debiese poder distinguir entre ella y el dispositivo real. Además, ambos escaneos requieren que los servicios no bloqueen a las IPs de las máquinas escaneadoras, pero debido a que los escaneos replican técnicas también utilizadas por agentes maliciosos, al transcurrir mayor tiempo realizando escaneos periódicos, es más probable que estos sean detectados y las direcciones asociadas a estas máquinas sean bloqueadas.

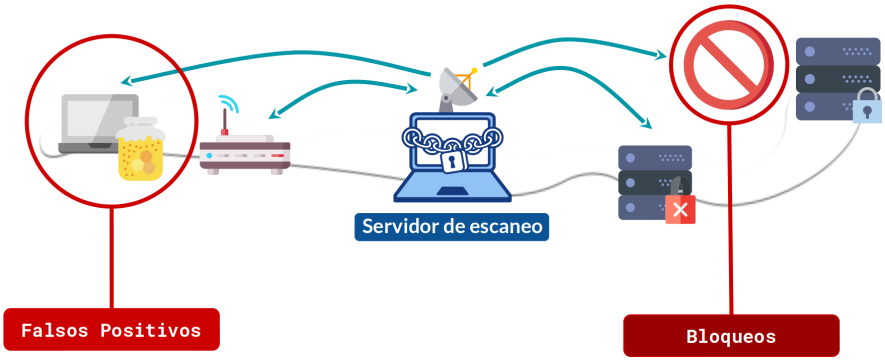


Figura 1.3: Algunos problemas relacionados con escaneos, como por ejemplo, la dificultad de llevar registro de una subred en caso que la IP encargada de hacer los escaneos sea bloqueada, y la dificultad para distinguir dispositivos vulnerables reales y honeypots de otros investigadores.

Con respecto a escaneos de tipo pasivo, también existe el riesgo de que la sola publicación de resultados de escaneo con objetivos preventivos permita a los atacantes conocer las ubicaciones de los dispositivos de monitoreo, tal como se menciona en el trabajo de Shinoda y otros [78]. Sin embargo, el trabajo anteriormente citado entrega algunas recomendaciones para evitar este tipo de detecciones.

1.1.6. Ámbito de la Investigación Realizada

Las estrategias propuestas en este documento se incluyen en algunas de las categorías mencionadas hasta ahora en esta sección. En específico, las presentadas en el capítulo 3 están bastante relacionadas a la categoría **Medición DNS** a partir de escaneos de tipo **activo**. El trabajo del capítulo 4 utiliza principalmente escaneos de puertos y protocolos tipo **activo**, además **del uso conjunto de sus mediciones con otros datos de seguridad**.

1.2. Organizaciones y Software Orientados a Monitoreo de Internet

En esta sección se presentan tanto algunas organizaciones cuyo rol involucra el monitoreo de Internet, como también el software de acceso libre que se puede utilizar para realizar distintos tipos de escaneos orientados a seguridad computacional. Finalmente, se justifica el trabajo de diseño e implementación realizado a partir de las alternativas existentes en la actualidad.

1.2.1. Organizaciones

Existen varias agrupaciones enfocadas en seguridad que realizan monitoreo de Internet a través de métodos de medición pasivos y activos de forma periódica en la actualidad. Estas organizaciones venden o entregan voluntariamente estos datos a otras entidades interesadas o encargadas de la seguridad de la red en cada ubicación geográfica. A continuación, se nombran algunas de ellas:

- **Censys:** *(Libre, Limitada y Pagada)* Censys es una empresa fundada por los investiga-

dores de seguridad que desarrollaron la suite ZMap [25, 2]. Usando los datos obtenidos con estas herramientas, desarrollaron un portal de búsqueda de resultados de escaneos en el espacio completo IPv4 [26]. Este portal provee de acceso limitado a consultas a cualquier persona de forma gratuita, pero requiere del pago de una suscripción para un acceso total a la información. También existen planes especiales gratuitos para organizaciones de investigación que no persigan fines comerciales.

- **Shodan:** *(Libre, Limitada y Pagada)* Autodenominado como "el buscador de dispositivos IoT", Shodan recopila en su sitio web escaneos sobre más de 300 puertos en toda la Internet, utilizando infrastructura ubicada en 7 países y con la capacidad de recopilar hasta 500 banners por segundo [79]. La plataforma ofrece un uso limitado gratuito de la API y planes empresariales pagados que permiten monitorear hasta 300 mil IPs simultáneamente. Cabe destacar también que la exportación de los datos manejados por ellos requiere pagar por "créditos" de un solo uso aparte de la membresía inicial.
- **Team Cymru:** *(Libre y Pagada)* Corresponde a una agrupación formada en 1998 que busca entender el cómo y el por qué de la actividad maliciosa en Internet. [22] Actualmente, recopilan información importante para su uso en análisis de la Internet, manteniendo bases de datos libres como *IP to ASN*, y otras pagadas como seguimiento de malware.
- **ShadowServer:** *(Solo para organizaciones asociadas)* Corresponde a una organización sin fines de lucro cuyo objetivo es *trabajar de forma altruista detrás de escena para hacer de la Internet un espacio seguro para todos* [29]. Entregan reportes y datos de escaneos a CERTs *Computer Emergency Response Team* y CSIRTs (*Computer Security Incident Response Teams*) Nacionales y administradores de bloques de redes de forma gratuita.
- **ZoomEye:** (*Libre y pagada*) Servicio que presta una funcionalidad muy similar a la de Shodan, permitiendo encontrar servicios no publicados corriendo en toda la Internet [90]. Para la detección de estos servicios usan dos desarrollos propietarios, denominados Xmap y WMap.
- **Aposemat:** (*Libre*) StratosphereIPS[36] es un grupo de Investigación en Seguridad de República Checa que lidera proyectos de protección frente a amenazas informáticas. Uno de ellos es *Aposemat*, el cual consiste en la liberación de datos de botnets capturadas a partir de una *honeynet* en su control.
- **RIPE Atlas:** (*Libre*) Corresponde a un a plataforma integral de medición de internet manejada por *RIPE NCC* [43], compuesta por más de 10 mil dispositivos dispositivos ubicados en distintas partes de la red europea. La red Atlas mide una gran cantidad de tráfico en la Internet en su territorio, como por ejemplo, consultas DNS, *pings*, *traceroutes*, entre otros.
- **Rapid7 Project Sonar:** (*Libre para investigadores*) Conjunto de datos liberado de forma recurrente por la empresa *Rapid7*[52], que incluye escaneos de tipo DNS, HTTP, certificados SSL, datos de honeypots y escaneos de puertos TCP y UDP a nivel mundial.

1.2.2. Software

Para la correcta realización de monitoreo y escaneos replicables, es necesario contar con herramientas de código abierto, dado que soluciones comerciales limitan tanto en la opacidad de los procesos ejecutados como en los recursos económicos necesarios para poder realizar

estas estrategias de monitoreo. En esta sección, se listan algunos programas orientados a realizar tareas de escaneo de Internet que son de código abierto.

Software para Escaneos de Puertos

A continuación se nombran algunas herramientas de código abierto especializadas en escaneos de puertos a gran escala.

- **ZMap:** Su página oficial define al *Proyecto ZMap* como una *colección de herramientas de código abierto que permite a investigadores realizar estudios a gran escala sobre [máquinas] anfitrionas y servicios que componen la internet pública* [86]. Específicamente, la herramienta que realiza escaneos de puertos se llama *ZMap* y permite escanear el espacio IPv4 completo en menos de una hora [25, 2]. La herramienta implementa escaneos TCP SYN, ICMP y UDP para aplicaciones específicas.
- **Masscan:** Herramienta con objetivos similares a los de ZMap, su página en Github declara poder escanear la internet completa en menos de 6 minutos, enviando 10 millones de paquetes por segundo desde una sola máquina [31]. Para lograr su objetivo, utiliza una implementación TCP/IP propia. Su método de escaneo para TCP es usando paquetes SYN, como se explicó anteriormente.

Software para Escaneos de Protocolos y Servicios

Para realizar estos escaneos, existen herramientas de código abierto especializadas, entre las cuales se pueden nombrar las siguientes.

- **Suite ZMap:** Parte de la suite ZMap mencionada anteriormente, La herramienta *Zgrab* permite la recopilación de información de protocolos comunes, tales como HTTP, SMTP, POP3, SSH, Telnet, SFTP, entre otros. La herramienta *ZDNS* permite realizar y capturar consultas DNS. La herramienta *ZCertificate* permite capturar certificados X.509. [86]
- **Mercury:** Herramienta desarrollada en el CLCERT como parte de la tesis de Magister de Eduardo Acha [1]. Permite la recopilación de información de protocolos comunes, como HTTP, SMTP, SSH, DNS, Ethereum, certificados, entre otros.

Manejo de Datos Recopilados en Mediciones

Aparte del escaneo y detección de amenazas y vulnerabilidades a partir de datos de monitoreo de Internet, es necesario mantener esta información en algún sistema que permita su agregación y procesamiento, en especial si esta información proviene de más de una fuente.

En el transcurso de esta investigación, no se encontró ninguna plataforma con las mismas intenciones que las declaradas por el desarrollo del Capítulo 5. Muchas organizaciones, como *Censys*, desarrollan una suite de herramientas, las cuales tratan de solucionar de forma independiente estos problemas [86].

En general, la falta de interés en el desarrollo de este software también se observa en que la mayoría de los datos de escaneo son entregados en su forma pura de parte de las

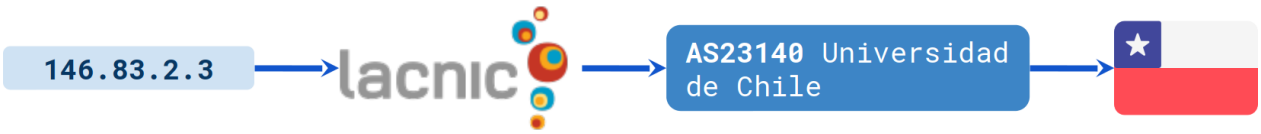


Figura 1.4: Una forma de determinar el país de una IP es considerar el país del sistema autónomo asociado a ella, dato que debiese estar registrado en el RIR correspondiente. En este ejemplo, debido a que la IP consultada es parte del Sistema Autónomo de la Universidad de Chile, ésta se considera como una IP chilena.

organizaciones que los generan, dejando a los grupos de investigación la tarea de desarrollar herramientas que les permitan importarlos a sus propios sistemas.

1.2.3. Ámbito de Diseño e Implementación de Software realizado

Debido a la falta de herramientas especializadas de código abierto para el manejo de datos de escaneo de Internet, el Capítulo 5 de este trabajo entrega el proceso de diseño y desarrollo de una herramienta que permite el levantamiento expedito de un sistema de administración de datos de monitoreo activo de la Internet, compatible con escaneos **activos** y **pasivos**. Esta herramienta permite su **uso conjunto con mediciones de otros datos de seguridad**, facilitando su agregación para la entrega de reportes de forma periódica y automática.

1.3. La "Red Chilena" en cada contexto

Como se mencionó anteriormente, el CLCERT tiene como objetivo monitorear y analizar los problemas de seguridad computacionales de Chile. Por lo tanto, es necesario almacenar y procesar la mayor cantidad de datos dentro del conjunto de "la red chilena".

Sin embargo, las características de la Internet vistas en la introducción hacen que resulte bastante complejo definir fácilmente el concepto de "red chilena", dado que los componentes físicos y lógicos que permiten su funcionamiento pueden ubicarse en cualquier parte del mundo, y los servicios ofrecidos sobre ella pueden estar dirigidos a, o ser utilizados mayoritariamente por distintos grupos geográficos.

Una definición adecuada y clara de qué es "red chilena" para cada contexto es fundamental si se desea entender la relevancia de los resultados obtenidos tanto sobre los datos de escaneo como sobre los datos derivados de estos. A continuación, se mostrarán algunos ejemplos de subconjuntos de Internet que se pueden definir como chilenos, los cuales han sido usados para escaneos ya existentes, o se usarán en las nuevas estrategias a proponer.

1.3.1. Según Sistemas Autónomos

Como se mencionó brevemente en la introducción, el RFC 1930 [58] define un Sistema Autónomo (*Autonomous System o AS*) como un grupo conectado de uno o más prefijos IP, mantenidos por uno o más operadores de red, el cual tiene una política de ruteo únicamente definida. La IANA asigna a cada AS un número único denominado *Autonomous System Number* o ASN al momento de registrarlo. Además del número, la IANA registra datos como

Fecha obtención	$N°IPs$
2019-08-25	8.593.920
2019-09-01	8.594.944
2019-09-08	8.577.792
2019-09-15	8.574.208
2019-09-22	8.574.976
2019-09-29	8.575.232
2019-10-06	8.574.464
2019-10-13	8.575.488
2019-10-20	8.573.184
2019-10-27	8.577.024
2019-11-03	8.576.768
2019-11-10	8.577.536

Tabla 1.1: Número de IPs asociadas a Sistemas Autónomos chilenos.

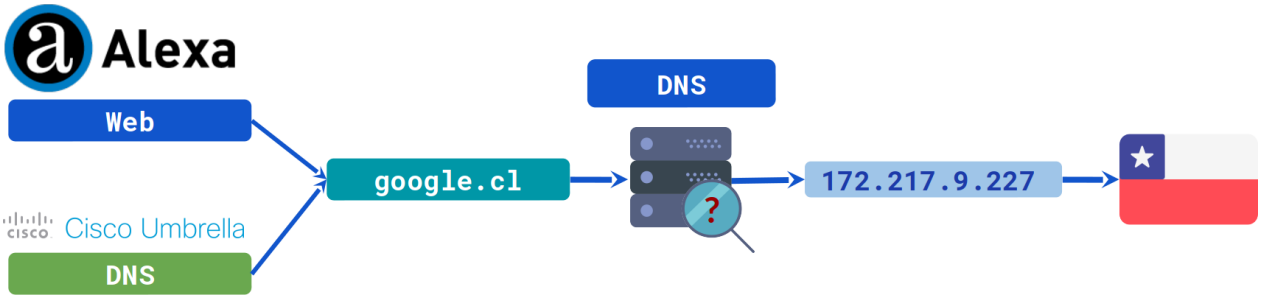


Figura 1.5: Este ejemplo muestra el uso de datos de rankings de dominios en Chile para determinar las IPs consideradas como chilenas. Este enfoque requiere tener en consideración el tipo de servicio asociado al ranking ranking usado. En el ejemplo, el dominio `google.cl` *es considerable como chileno debido a su masividad de uso en el país.*

contacto técnico del AS, dirección y país. Estos datos son consultables a través del protocolo RDAP [66], el cual permite consultar a los 5 RIRs existentes información de tipo WHOIS.

Entonces, como se muestra en la figura 1.4, una forma de definir la Internet Chilena puede ser el considerar solamente las IPs que se encuentren en subredes asociadas a Sistemas Autónomos registrados como chilenos. Esto tiene como requisito el confiar en que los datos de contacto se mantienen actualizados, y que los datos de dirección del administrador coinciden en país con la ubicación de los equipos conectados a Internet a través de esta ASN. Debido a lo anterior, es importante mencionar que organizaciones como *Team Cymru* no recomiendan usar esta información para geolocalizar IPs [21].

Como referencia, la tabla 1.1 muestra la cantidad de IPs asociadas a Sistemas Autónomos chilenos entre los meses de Agosto y Noviembre del 2019.

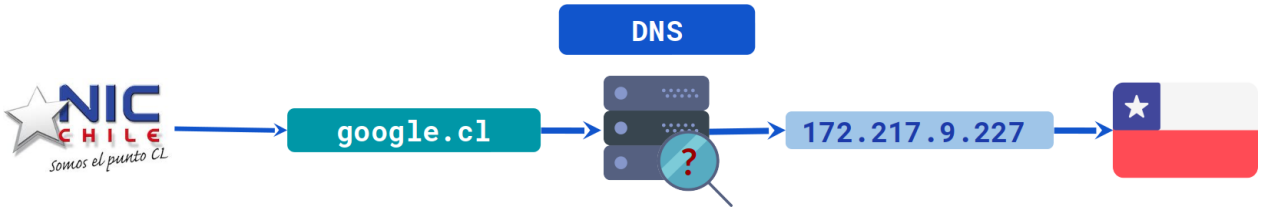


Figura 1.6: Esta categorización considera como IPs chilenas el conjunto de éstas asociadas a un nombre de dominio administrado por NIC Chile, es decir, aquellos dominios que terminan en `.cl`*. En el ejemplo de la imagen, una IP asociada al dominio* `google.cl` *es considerada como chilena, porque el dominio* `google.cl` *es un dominio chileno.*

1.3.2. Según relevancia para usuarios chilenos

Esta definición depende de contar con una lista de servicios más utilizados por los usuarios chilenos en la Internet. Lo anterior tiene como resultado que una gran cantidad de plataformas internacionales, como redes sociales o portales de compra y venta, fuesen considerados como "Internet Chilena" en ciertos casos.

Sin embargo, es complicado conseguir esta información de forma directa de parte de los ISPs. Sin embargo, existen plataformas privadas (como *Amazon* a través del servicio web *Alexa* [33] y *Cisco* a través de su servicio *Umbrella 1 Million* [15]) que entregan estimaciones de rankings de uso segregados por países. Un ejemplo de cómo usar estos datos se encuentra graficado en la figura 1.5.

Cabe destacar que estos rankings suelen considerar solamente un tipo de servicio. Por ejemplo, el ranking de *Amazon* considera visitas a sitios web detectadas por usuarios que mantienen instalada su extensión de navegador web, mientras que el ranking de *Cisco* considera las más de 100 mil millones de consultas DNS diarias a sus sistemas, por lo que usar cada ranking en servicios distintos a los usados por las fuentes puede entregar información equivocada. Además, esta perspectiva considera el escaneo de máquinas que, si bien son importantes para los usuarios de nuestro país, no es claro que debiesen ser consideradas en el espacio de dispositivos de los cuales el CLCERT debiese estar preocupado.

Como referencia, entre Agosto y Noviembre de 2019, la cantidad de dominios chilenos en el ranking Alexa Top 1M varió entre 641 y 1.877 entradas.

1.3.3. Según servicios asociados a dominios .CL

Al contrario de otros ccTLD como .co (Colombia), .io (Territorio británico del océano índico) y .fm (Estados federados de Micronesia), los cuales se suelen usar en la práctica para empresas comerciales, *startups*, y radioemisoras respectivamente, el TLD chileno suele ser usado fundamentalmente para servicios dirigidos a usuarios de nuestro país. Esto motiva a considerar como parte de la "Internet Chilena" a los servicios que corren en máquinas asociadas a FQDNs con dominio chileno. El flujo de uso de esta estrategia se ve representado en la figura 1.6.

Cabe destacar que esta lista no está pensada para incluir tanto servicios orientados a Chile con dominios distintos a .cl, como máquinas que ofrecen servicios orientados a Chile, pero

Figura 1.7: En este caso, se consideran como IPs chilenas las IPs reportadas como tal por algún servicio de geolocalización de IPs confiable.

que no tienen ningún FQDN chileno asociado.

Un problema de este enfoque es que la lista de dominios chilenos registrados no es pública. Sin embargo, NIC Chile publica tanto dominios registrados como vencidos en los últimos 30 días. Por lo tanto, es posible llevar registro de todos los dominios activos registrados desde la fecha en que se empiezan a recopilar estas listas. Pero incluso recopilando los dominios de esta forma, no existe un método directo y general para obtener sitios de la internet chilena asociados a subdominios de estos dominios.

Como referencia, la cantidad de IPs asociadas a los dominios chilenos que se revisarán en el capítulo 3, ya sea a través de registros A, MX o NS, es de 90.679 IPs.

1.3.4. Según información de proveedores externos

Existen algunos proveedores, tanto pagados como gratuitos, que entregan información de ubicación geográfica de IPs y subredes específicas, e incluso en algunos casos, entregan la ciudad en la que potencialmente se encuentra ubicada la máquina asociada a la IP consultada. A continuación, se mencionan brevemente estos servicios:

- **Maxmind GeoLite2 y GeoIP2:** *(Versiones Gratuita y Pagada)* La empresa *Maxmind* cuenta con el servicio pagado de listas de IPs geolocalizadas denominado *GeoIP2*, el cual cubre el 99,9999 % de las IPs en uso y argumenta poseer una confiabilidad del 99,8 % al determinar el país de una IP.[40]. Además, cuentan con una versión gratuita de la base de datos denominada *GeoLite2* [39], la cual según la misma fuente argumenta poseer una exactitud de país cercana a la de la versión pagada.
- **Neustar IP Geopoint:** *(Pagada)* Alternativa manejada por la empresa *Neustar* para relacionar IPs con ubicaciones geográficas. Sin embargo, no existen de forma pública los datos de confiabilidad ni de precio de las bases de datos de esta fuente.
- **IP2Location:** *(Pagada y Gratuita)* Servicio ofrecido por la empresa homónima que ent rega datos de ubicaciones geográficas según IP, además de datos extra como ISP, coordenadas de ciudad, código postal, entre otros [35]. El costo de las bases de datos ronda entre 49 y 1549 dólares al año, y posee una confiabilidad de 99,5 % [34]. También ofrecen una versión gratuita de la base de datos, actualizada mensualmente, y que entrega datos de rangos de IP y países con una confiabilidad de 98 %.
- **API RIPEstat de RIPE NCC:** *(Gratuita)* RIPE NCC *(Network Coordination Centre)* es una organización que entrega soporte al RIR RIPE, el cual está encargado

Fecha obtención	N^oIPs
2019-08-18	10.116.219
2019-08-25	10.115.875
2019-09-01	10.116.371
2019-09-08	10.114.325
2019-09-15	10.114.325
2019-09-22	10.119.693
2019-09-29	10.117.645
2019-10-06	10.117.645
2019-10-13	10.117.645
2019-10-20	10.119.941
2019-10-27	10.120.709
2019-11-03	10.120.709
2019-11-10	10.121.221

Tabla 1.2: Número de IPs chilenas según la base de datos Maxmind GeoLite2.

de Europa el este de Asia y la antigua Unión Soviética. Al mismo tiempo, mantiene una API de consulta de datos de IPs y ASNs, *hostnames* y países [44]. Los datos de geolocalización proveídos por esta API están basados en la información de la base de datos GeoLite2, mencionada anteriormente [45].

La figura 1.7 muestra en términos simples el flujo de consulta en este caso.

La mayor desventaja de esta estrategia es que requiere tener confianza en alguna de estas fuentes, dado que ninguna de ellas revela la forma en que obtienen estos datos, dificultando la replicabilidad y verificación se los resultados. Sin embargo, una gran cantidad de servicios externos confían en al menos una de estas fuentes, como es por ejemplo el caso ya mencionado de *RIPE NCC*.

Como referencia, al menos desde el año 2017 el tamaño del espacio de IPs chilena indicada por la fuente MaxMind Geolite2 es de alrededor de 10 millones de entradas, manteniendo una tendencia al crecimiento durante los últimos años, llegando a un total de 10.121.221 en el mes de noviembre de 2019. La variación en cantidad de IPs durante los años en que se realizaron las mediciones de este trabajo no superó el 1 %, lo cual fue comprobado al revisar la cantidad de IPs chilenas según los resultados almacenados por *Wayback Machine*[6]. Es posible revisar en la tabla 1.2 la cantidad de IPs chilenas por semana entre Agosto y Noviembre de 2019.

1.3.5. Definiciones a usar

Dada la variedad de definiciones de red chilena, cada una con sus ventajas y desventajas, se decidió justificar en cada estrategia el uso de cada definición por separado. Por lo tanto, se hace necesario mantener especial atención en el universo de redes estudiadas al momento de comparar, de forma de evitar conclusiones desacertadas.

Capítulo 2

Revisión Preliminar de Datos de Escaneo manejados

El objetivo de este capítulo es explicar qué es el CLCERT y cuál es su rol en temas de seguridad computacional en Chile, además de exponer los datos de escaneos recopilados por el CLCERT a la fecha de Noviembre de 2019, actualizando los resultados obtenidos por Eduardo Acha hasta fines del año 2016. Posteriormente, se analizan de forma paralela los resultados de escaneos del mismo tipo realizados por una fuente externa, con el objetivo de determinar en primera instancia qué tan parecidos son. Finalmente, el capítulo ahonda en describir las dificultades apreciables sobre los resultados del CLCERT, los resultados de la fuente externa y las diferencias entre ambos resultados.

2.1. El CLCERT

El CLCERT es un grupo de investigación de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, el cual está dedicado al monitoreo y análisis de los problemas de seguridad de los sistemas computacionales de nuestro país, y a la generación tanto del conocimiento como el recurso humano especializado para asegurar dichos sistemas [16]. Para cumplir el objetivo anterior, el centro recopila, tanto a través de mecanismos internos como por reportes de organizaciones externas asociadas, un gran volumen de datos de escaneos de puertos y vulnerabilidades. Estos reportes suelen ser entregados a las instituciones que los requieren.

2.1.1. Datos Manejados

Actualmente, el CLCERT maneja los los siguientes datos de forma periódica.

- **Escaneos de Puertos:** *(Semanales, propios y externos)* Corresponde a un escaneo de puertos abiertos, además de metainformación relacionada con el servicio corriendo en cada puerto. Se escanean más de 20 puertos distintos cada semana, repartidos durante varios días, en busca de determinar qué dispositivos los tienen abiertos, y por lo tanto, pudiesen estar entregando algún servicio a través de ellos. Algunos puertos escaneados

(y los protocolos asociados a ellos) son:

- 21 (FTP)
- 22 (SSH)
- 23 (Telnet)
- 25 (SMTP)
- 53 (DNS)
- 80 (HTTP)
- 110 (POP3)
- 123 (NTP)
- 143 (IMAP)
- 443 (HTTPS)
- 465 y 587 (SMTP encriptado)
- 631 (CUPS)
- 993 (IMAP encriptado)
- 995 (POP3 Encriptado)
- 1433 (Base de datos MSSQL)
- 3306 (Base de datos MySQL)
- 5432 (Base de datos PostgreSQL)
- 5800 y 5900 (VNC)
- 8080 (HTTP en desarrollo)
- 8443 (HTTPS en desarrollo)
- 27018 (Base de Datos MongoDB)

- **Escaneos de Protocolos:** *(Semanal, propio y externo)* Corresponden a escaneos de protocolos realizados localmente y recibidos de organizaciones externas asociadas. Algunos de los protocolos de los cuales se tienen información son los siguientes:
  - Páginas web sobre el protocolo HTTP y HTTPS.
  - Certificados sobre protocolos HTTPS, IMAP encriptado y POP3 Encriptado.
  - *Banners* SSH, SMTP, IMAP.
- **Paquetes de la Darknet del CLCERT:** *(Diarios y propios)* Desde el año 2007, el CLCERT cuenta con un subespacio de red de 255 IPs, el cual escucha constantemente todas las conexiones entrantes pero no posee ningún servicio corriendo. Esta información es guardada en archivos PCAP que la almacenan temporalmente, debido a su gran tamaño (aproximadamente, de 1 GB por hora guardada), por lo que no se cuenta con datos históricos de esta información de más de un mes de antiguedad.
- **Reportes de malware y vulnerabilidades** *(Diarios y Semanales, propios y externos)* Corresponden a reportes que indican máquinas de la red chilena infectadas con cierto malware o vulnerabilidad. Algunos de estos reportes provienen de escaneos propios, mientras que otros provienen de fuentes externas reservadas, las cuales requieren no publicar su autoría para el uso de los datos. Los reportes trabajados actualmente son los siguientes:
  - Heartbleed
  - BlueKeep
  - Botnets (como *Mirai*)
  - Servidores de *Comando y Control*
  - Servidores con riesgo de ser explotados para *DDOS* a través de *Ataques de Amplificación*
  - Dispositivos detectados por *darknets* externas.
  - Proxies y resolvers DNS abiertos.
  - Malware en sitios web de phishing

Actualmente, esta información es recopilada de forma automática. Sin embargo, el procesamiento de estos datos es realizado de forma manual y a pedido, específicamente cuando

es necesario hacer entrega de reportes periódicos a las organizaciones que los solicitan. Solo una parte de las estadísticas de conteo sobre estos datos es publicada en el sitio web del CLCERT de forma automática. En el capítulo 5, se abordará este problema y se propondrá una solución a partir de un sistema que programe, procese, transforme y agregue los datos recopilados.

2.2. Análisis Preliminar de Datos Manejados

El trabajo realizado por Eduardo Acha como parte de su tesis [1] utiliza escaneos activos de 15 puertos y 7 protocolos distintos, recopilados durante los meses de enero y octubre de 2016. Entre otros protocolos, se utilizaron cabeceras y versiones de servicios HTTP, HTTPS, SMTP, IMAP, POP3, SSH y FTP. Además, en el caso de los protocolos seguros recién mencionados, se realizó un estudio detallado de la cantidad, metadatos y tipos de certificados SSL/TLS usados en las IPs chilenas. Este documento finaliza con la definición de una serie de recomendaciones, puntuando su cumplimiento para clasificar de manera cuantitativa el estado de seguridad de la Red Chilena. En contraste con el aporte de Acha, el objetivo del trabajo actual es extender el estudio de estos datos, aprovechando que se cuenta con casi dos años más de información y las ventajas de automatización y procesamiento entregadas por la nueva plataforma de procesamiento de información, diseñada y explicada en el capítulo 5 de este documento.

Los datos recopilados periódicamente por el CLCERT a la fecha provienen de tres fuentes distintas, entre las cuales se encuentran una fuente interna (denominada CLCERT de ahora en adelante) y dos fuentes externas (denominadas *Fuente Reservada de Malware #1*, *Fuente Reservada de Malware #2*). Asímismo, con el objetivo de entender mejor la replicabilidad de las estrategias basadas en escaneos de puertos y protocolos entre distintas organizaciones, en esta ocasión se usó una cuarta fuente de datos proveniente de *Censys*, la cual se introduce preliminarmente también en esta sección.

Todos los datos presentados en esta sección fueron analizados e importados utilizando un sistema ideado y desarrollado en el marco de este trabajo, denominado *OSR* (sigla de Observatorio de Seguridad de la Red Chilena). El diseño, la programación y la puesta en marcha de este sistema se explica en el Capítulo 5 de este trabajo de tesis.

A continuación, se describen los conjuntos de datos manejados por las cuatro fuentes, y se destacan algunas observaciones y comparaciones interesantes de ellos.

2.2.1. Datos del CLCERT

Esta fuente corresponde a datos recopilados internamente con infraestructura del CLCERT. La principal característica que diferencia los datos de esta fuente de los demás es que internamente se manejan detalles metodológicos acerca de su obtención, tales como herramientas utilizadas y periodicidad de los escaneos.

Los escaneos del CLCERT son realizados sobre la red chilena reportada por *RIPE NCC*, fuente de ubicación geográfica que entrega esta información a partir de los datos de la fuente *MaxMind Geolite 2*. Por lo tanto, se puede considerar la segunda como fuente oficial de

geolocalización de estos datos.

A continuación se presentan, por categoría, los datos recopilados por esta fuente.

Escaneos de Puertos

Obtenidos semanalmente entre el 27 de junio de 2016 y 18 de noviembre de 2019

Los escaneos de puertos del CLCERT se obtienen a partir de datos de la ejecución de la herramienta ZMap en el universo de IPs chilenas según la base de datos GeoLite2, actualizada de forma semanal. Los escaneos se realizaron una vez a la semana, y fueron distribuidos entre los días lunes y viernes.

La tabla 2.1 y la figura 2.1 muestran estadísticas por puerto de rango de escaneos, semanas escaneadas y relación de cobertura entre semanas y escaneos. Todos los escaneos tienen una cobertura de al menos un 85 %, y se destaca un valor alto en la cantidad promedio de IPs detectadas en los puertos 80/TCP (protocolo HTTP), 631/TCP (protocolo CUPS), 443/TCP (protocolo HTTPS), 23/TCP (protocolo Telnet) y 22/TCP (protocolo SSH). Con respecto al número de IPs únicas existentes por puerto en el rango de tiempo estudiado, destacan los mismos mencionados anteriormente, además de los puertos 8000/TCP y 8080/TCP (HTTP, usado generalmente para servidores en desarrollo).

Los gráficos en 2.3 muestran, entre otras cosas, el historial de máquinas encontradas durante el periodo de escaneo de 9 puertos escaneados. A partir de estas imágenes, es posible notar que los resultados de escaneos mantienen cifras bastante similares hasta mitad del año 2018, momento en el que estos empiezan a mostrar una alta variabilidad entre semanas. Sin embargo, estas cantidades no suelen bajar de la cantidad de IPs promedio previos a fines de julio de 2018. Las razones posibles de esta diferencia son variadas. Una posible explicación es que se deben a problemas de configuración de la red de la Universidad o actualizaciones de los sistemas involucrados, las cuales permitieron encontrar más resultados que los visibles anteriormente.

Otro detalle importante rescatable de los gráficos es la inexistencia de datos entre diciembre 2018 y marzo 2019, la cual se debe a problemas de almacenamiento de las máquinas de escaneo que no se detectaron por razones explicadas en el capítulo 5 de este documento. Es necesario tener en cuenta esta situación al momento de utilizar estos datos a lo largo de este trabajo.

Escaneos de Protocolos, Software y Versiones

Obtenidos semanalmente entre el 27 de junio de 2016 y 18 de noviembre de 2019

Los escaneos de protocolos del CLCERT se obtienen a partir de la ejecución de la herramienta *Mercury* sobre las máquinas con puertos abiertos detectadas por ZMap. Posteriormente, utilizando estos datos, el sistema OSR realiza un proceso de validación sobre los *banner* o cabeceras recogidas, en el que se considera como válida toda entrada que tenga el formato esperado para el protocolo analizado. Paralelamente a la revisión de banners, el sistema OSR intenta identificar el software y la versión usada por el servidor para servir cada protocolo, a partir de un conjunto de expresiones regulares usadas sobre los banners recolectados.

La tabla 2.2 y la figura 2.2 resumen las estadísticas sobre los 9 puertos escaneados, consistentes en 6 protocolos distintos. Al igual que en el caso del escaneo de puertos, los más populares según los datos fueron el HTTP vía 80/TCP, SSH en el puerto 22/TCP y HTTPS en el puerto 443/TCP, a los cuales se le suman HTTP en puerto 8000/TCP y 8080/TCP si se consideran los protocolos con más IPs únicas en el intervalo escaneado. L

El bajo promedio y alta variabilidad de los resultados de los puertos 8000/TCP y 8080/TCP se justifica justamente por el hecho de ser usados para servidores de prueba, por lo que su existencia en Internet suele ser temporal ya sea por que se dan de baja o porque se configuran adecuadamente firewalls para que solo sean accesibles internamente.

Los gráficos en 2.3 (páginas 40, 41 y 42) muestran la cantidad de máquinas encontradas con un protocolo activo, en comparación con la cantidad de máquinas encontradas con el respectivo puerto abierto. Recordando que los escaneos de protocolos se hicieron sobre las IPs reportadas por ZMap en el escaneo de puertos abiertos, se entiende que el conjunto de datos de protocolos activos es un subconjunto del conjunto de datos de puertos abiertos. Además, es posible ver que en el último tiempo (desde mayo de 2019), la mayoría de los escaneos ha entregado un número bastante similar de resultados positivos, lo que contrasta bastante con los resultados de los escaneos de puertos, y refuerza la importancia de tener buenos escaneos de protocolos por sobre los escaneos de puertos.

Otro detalle destacable de los mismos gráficos es que en todos los casos, aproximadamente desde mayo de 2018, los resultados de escaneos de protocolos se vuelven bastante variables, pero mayores que los obtenidos anteriormente. Se conjetura que esto puede ocurrir debido a cambios de configuración en la red interna de la universidad, o incluso a cambios en la forma en que un conjunto no menor de routers conectados a la red chilena contestan paquetes de escaneo.

Puerto	Desde	Hasta	Escaneos	Semanas	Cobertura	$min(\#IPs)$	$max(\#IPs)$	$\bar{X}(\#IPs)$	$\#uniq(IPs)$
21	2016-06-27	2019-11-18	156	178	0,88	2.682	21.121	11.184,87	192.117
22	2016-06-27	2019-11-18	156	178	0,88	3.287	57.753	39.947,55	1.033.207
25	2016-07-04	2019-11-18	154	177	0,87	18.766	58.826	32.523,26	202.458
80	2016-07-04	2019-11-18	156	177	0,88	16.556	136.639	78.527,80	1.670.447
110	2016-07-04	2019-11-18	154	177	0,87	1.521	18.165	12.750,08	60.213
143	2016-07-04	2019-11-18	153	177	0,86	699	16.678	12.193,12	59.764
443	2016-07-04	2019-11-18	155	177	0,88	22.844	53.067	39.383,46	1.002.623
8000	2016-07-04	2019-11-18	153	177	0,86	3.074	47.996	9.564,01	941.499
8080	2016-07-04	2019-11-18	152	177	0,86	8.041	38.134	20.654,46	612.852

Tabla 2.2: Tabla que muestra los rangos de tiempo en que se ejecutaron los escaneos de protocolos del CLCERT de cada puerto. La tabla también muestra la cantidad de escaneos realizados en ese rango temporal, y el promedio de escaneos por semana. La columna Cobertura representa la proporción entre las semanas efectivamente escaneadas y el rango de semanas del protocolo.

Certificados SSL/TLS

Obtenidos semanalmente entre el 4 de julio de 2016 y el 19 de noviembre de 2019.

La herramienta *Mercury* se encuentra configurada para recopilar, de forma paralela y en los protocolos seguros, la cadena de certificados SSL/TLS utilizados en la interacción con

Puerto	Desde	Hasta	Escaneos	Semanas	Cobertura	$min(\#IPs)$	$max(\#IPs)$	$\bar{X}(\#IPs)$	$\#uniq(IPs)$
21	2016-06-27	2019-11-18	156	178	0,88	23.252	219.555	58.057,37	578.020
22	2016-06-27	2019-11-18	156	178	0,88	54.295	205.186	89.231,85	1.434.443
23	2016-06-27	2019-11-18	151	178	0,85	67.234	221.840	98.038,60	1.245.220
25	2016-07-04	2019-11-18	154	177	0,87	22.680	209.000	65.129,84	449.615
53	2016-07-04	2019-11-18	157	177	0,89	14.217	20.330	17.553,90	138.636
80	2016-07-04	2019-11-18	156	177	0,88	137.984	354.704	181.026,55	2.646.921
110	2016-07-04	2019-11-18	154	177	0,87	12.986	211.353	45.106,47	306.435
119	2016-07-04	2019-11-18	157	177	0,89	801	197.718	31.323,46	244.888
123	2016-07-04	2019-11-18	157	177	0,89	79	5.139	2.617,14	124.278
143	2016-07-04	2019-11-18	153	177	0,86	12.353	210.875	44.234,89	304.642
179	2016-07-04	2019-11-18	157	177	0,89	3.839	202.964	39.604,62	285.040
443	2016-07-04	2019-11-18	155	177	0,88	71.111	258.493	110.565,73	1.747.635
445	2017-05-15	2019-11-18	115	132	0,87	5.034	161.717	36.213,46	261.807
465	2016-07-04	2019-11-18	154	177	0,87	9.517	208.527	40.394,58	279.086
520	2016-10-10	2019-11-18	145	163	0,89	3	23	11,84	197
623	2017-05-01	2019-11-18	117	134	0,87	928	197.864	34.708,61	205.972
631	2016-06-27	2019-11-18	156	178	0,88	85.343	287.887	168.978,22	2.028.565
664	2017-05-01	2019-11-18	117	134	0,87	690	197.590	35.176,91	205.166
993	2016-07-04	2019-11-18	153	177	0,86	10.735	210.556	42.671,33	294.763
995	2016-07-04	2019-11-18	154	177	0,87	10.472	210.587	42.941,22	287.100
1433	2016-06-27	2019-11-18	151	178	0,85	3.208	190.464	29.174,62	238.547
1699	2017-05-01	2019-11-18	117	134	0,87	726	197.802	53.040,00	206.341
1911	2016-10-24	2019-10-28	136	158	0,86	685	196.605	29.622,71	242.479
2701	2016-06-27	2019-11-18	151	178	0,85	616	197.274	32.229,50	243.358
3306	2016-06-27	2019-11-18	156	178	0,88	11.753	210.033	43.691,04	357.300
3389	2017-06-05	2019-11-18	111	129	0,86	12.673	207.653	47.344,20	344.609
4911	2016-10-24	2019-10-28	136	158	0,86	684	196.598	28.778,63	241.142
5432	2016-06-27	2019-11-18	156	178	0,88	17	199.043	32.094,13	253.316
6667	2016-07-04	2019-11-18	157	177	0,89	760	158.793	30.206,31	245.020
7000	2016-07-04	2019-11-18	157	177	0,89	2.701	160.013	31.675,53	323.438
8000	2016-07-04	2019-11-18	153	177	0,86	17.330	182.140	50.395,41	1.494.623
8080	2016-07-04	2019-11-18	152	177	0,86	22.651	179.277	65.421,00	1.256.085
8333	2017-07-10	2019-11-18	105	124	0,85	623	197.731	34.459,61	205.510

Tabla 2.1: Tabla que muestra los rangos de tiempo en que se ejecutaron los escaneos de puertos del CLCERT de cada puerto. La tabla también muestra la cantidad de escaneos realizados en ese rango temporal, y el promedio de escaneos por semana, tratando de notar a simple vista la completitud de cada conjunto de datos. La columna Cobertura representa la proporción entre las semanas efectivamente escaneadas y el rango de semanas del protocolo. Todos los escaneos son sobre el protocolo TCP, salvo el puerto 53 (DNS) 123 (NTP), 520 (RIP) y 623 (ASF-RCMP).

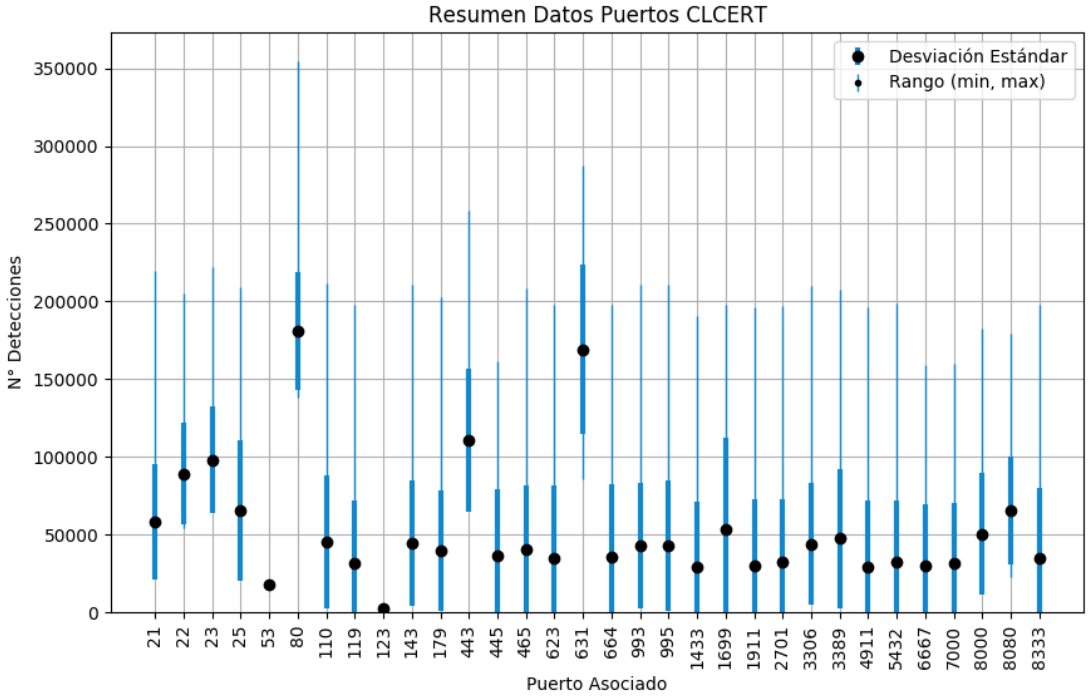


Figura 2.1: Resumen gráfico de los resultados de número IPs mínimas, máximas, promedio y desviación estándar del histórico de escaneos de puertos del CLCERT. Para mejorar la visibilidad del gráfico, se consideraron solamente los escaneos con más de mil IPs en promedio.

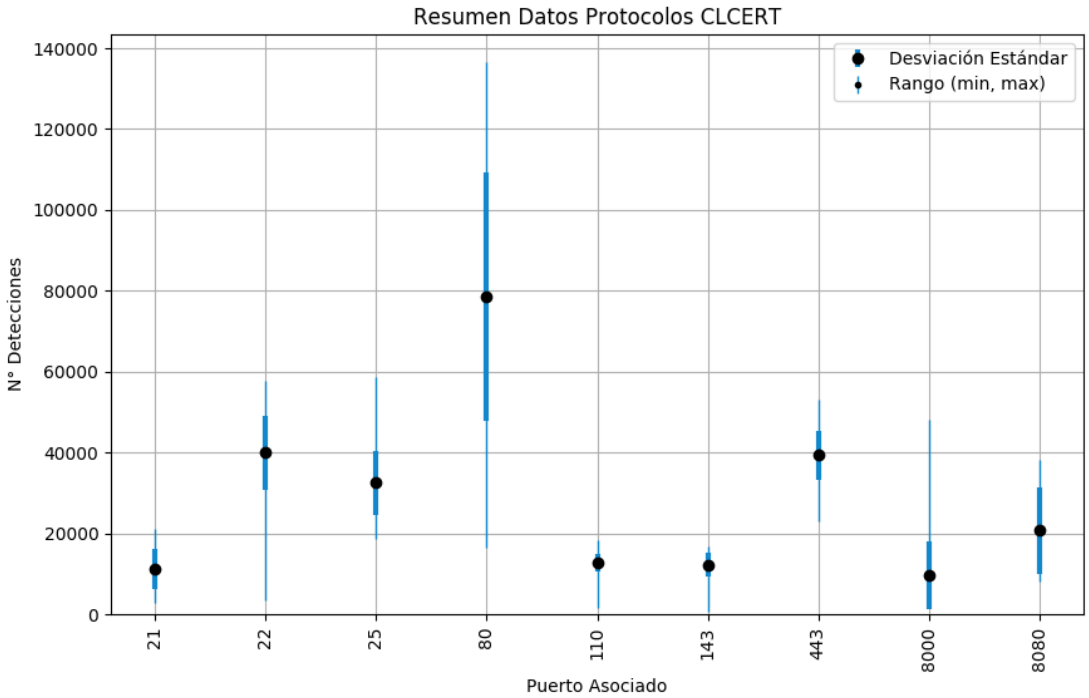
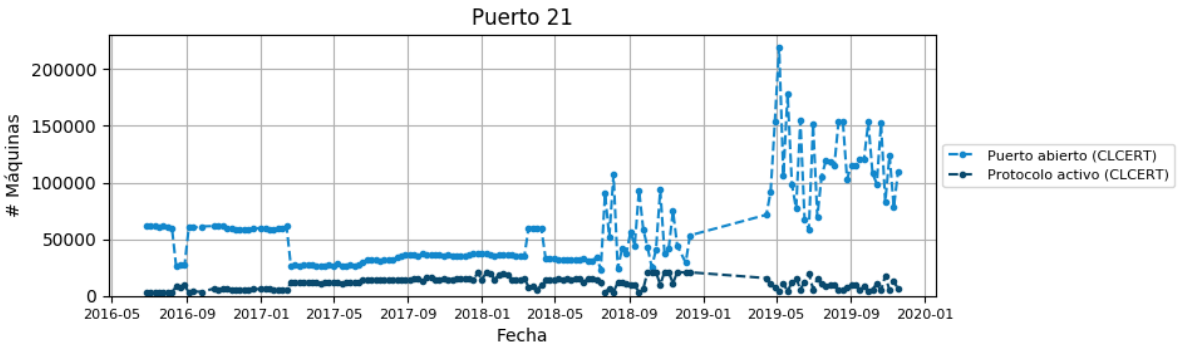
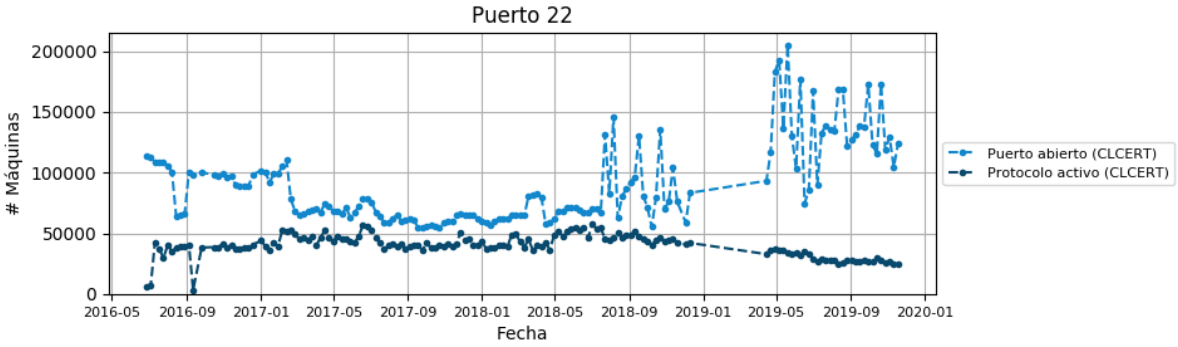


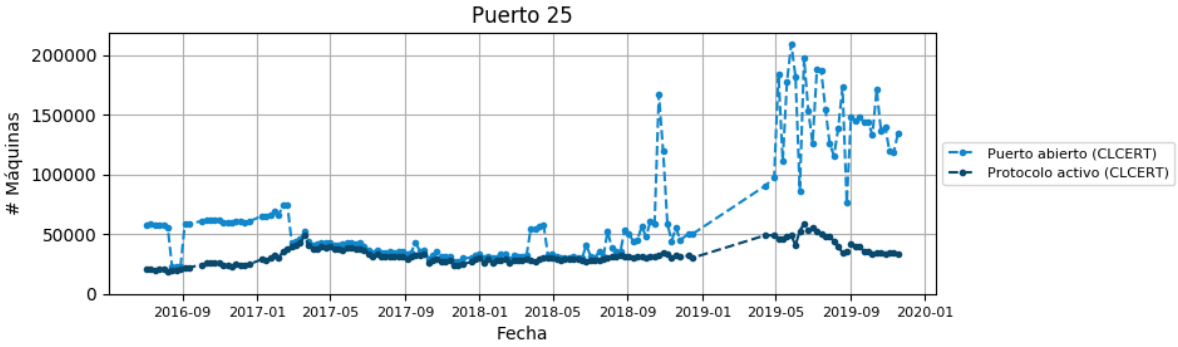
Figura 2.2: Resumen gráfico de los resultados de número IPs mínimas, máximas, promedio y desviación estándar del histórico de escaneos de protocolos del CLCERT. Para mejorar la visibilidad del gráfico, se consideraron solamente los escaneos con más de mil IPs en promedio.



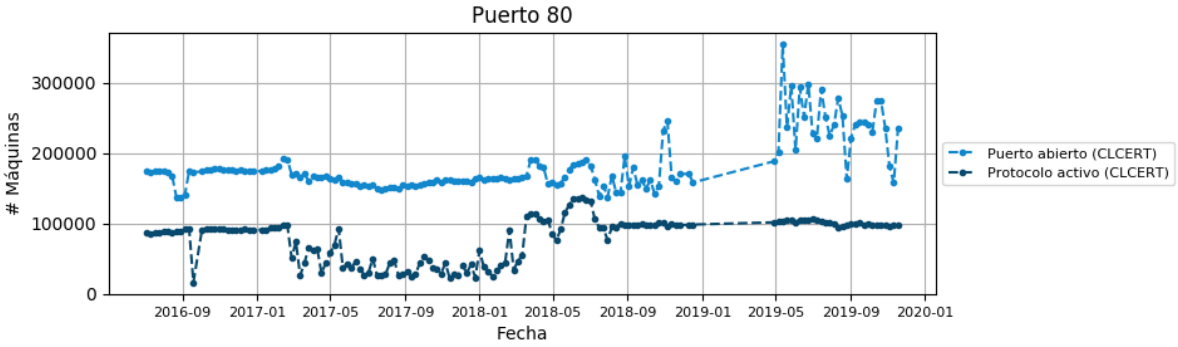
(a) Puerto 21



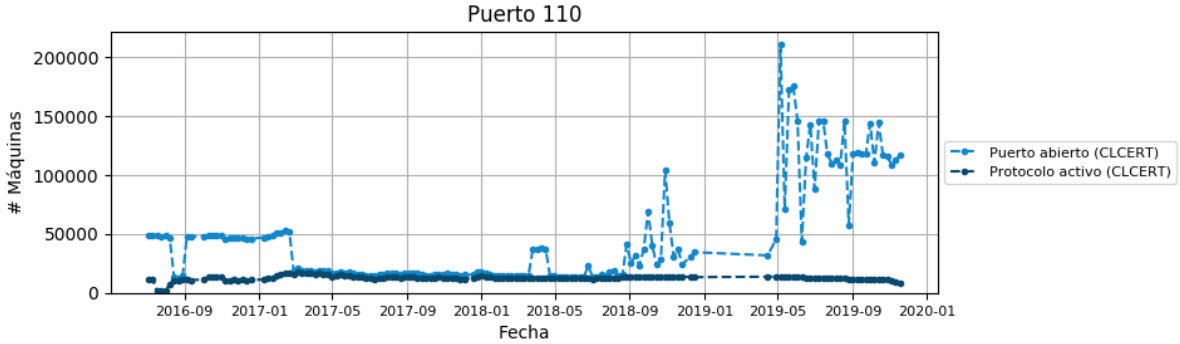
(b) Puerto 22



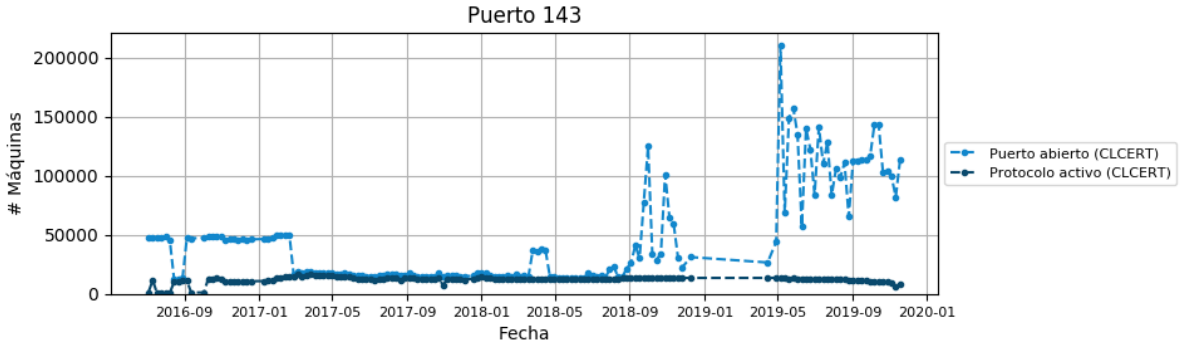
(c) Puerto 25



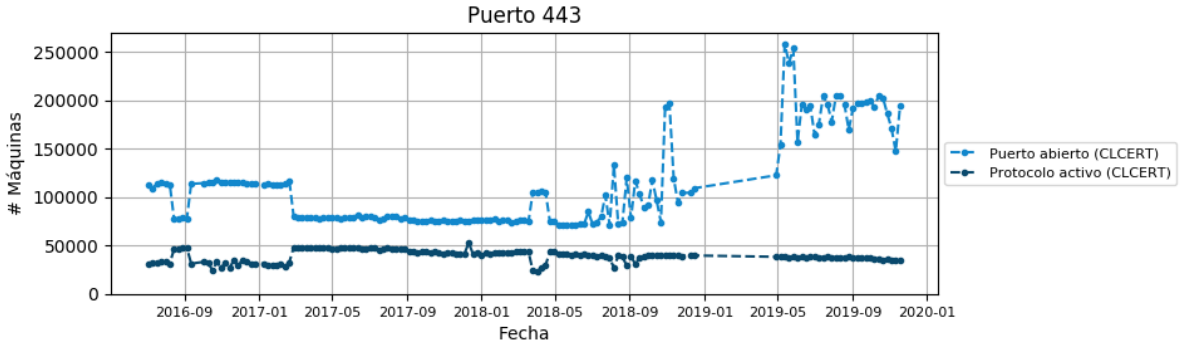
(d) Puerto 80



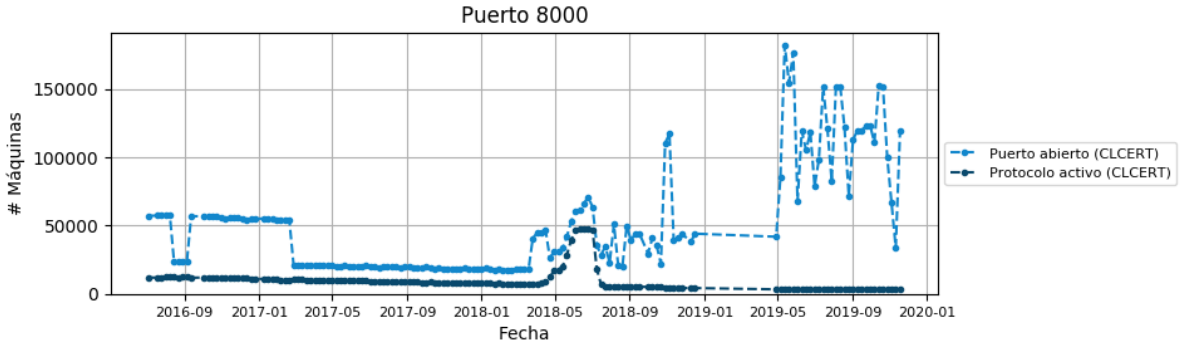
(e) Puerto 110



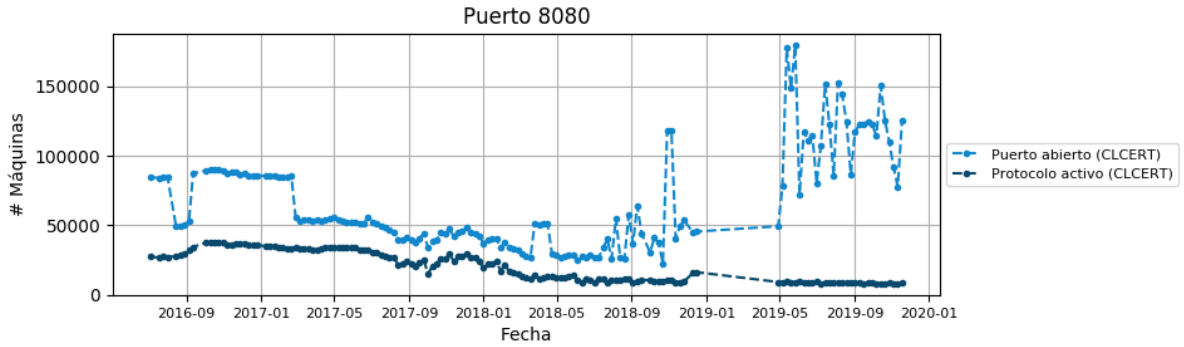
(f) Puerto 143



(g) Puerto 443



(h) Puerto 8000



(i) Puerto 8080

Figura 2.3: Comparación entre la cantidad de máquinas corriendo un servicio asociado a un puerto y cantidad de máquinas que tienen abierto ese mismo puerto según información del CLCERT.

el protocolo. El trabajo de Acha usó esta información como suministro importante para la ponderación del estado de seguridad de los sitios escaneados.

A partir de estos datos, el sistema OSR ingresa a su base de datos información importante relacionada con estos certificados, como por ejemplo, estado (vencido o por expirar, autofirmado, inválido), algoritmo de hash, protocolo TLS, largo de llave, entre otros datos. Esta información se puede ver resumida y agrupada por puerto estudiado en la figura 2.4 (páginas 44 y 45), las cuales muestran una línea de tiempo con certificados por protocolo, así como también cantidad de certificados que no cumplen ciertos requisitos mínimos de seguridad, según lo planteado en el trabajo de Acha:

- Un certificado es considerado como **inválido** si falla en validar la cadena que adjunta o es autofirmado.
- Un certificado es considerado como **autofirmado** si el campo *issuer* de éste es igual al campo *subject*.
- Un certificado es considerado **expirado** si su periodo de validez no incluye la fecha en la cual se obtuvo. Es importante notar que esta definición incluye en la misma situación si el certificado venció y si todavía no es válido.
- Los certificados **TLS Inseguros** en el gráfico son los que no usan protocolo TLS 1.2 o TLS 1.3.
- Los certificados con **Algoritmo de firma insegura** son los que usan como algoritmo de hashing MD5 o SHA1.
- Los certificados con **Tamaño de llave insegura** son aquellos cuya llave tiene una longitud menor a 2048 bits. Esta medida se tomó del trabajo de Eduardo Acha. Esta medida se considera relevante debido a que *Mercury* ha capturado históricamente solamente certificados de tipo RSA.

Se puede notar en todos gráficos que la cantidad de certficados firmados con protocolos TLS "inseguros" ha ido disminuyendo desde el 2006. Sin embargo, sigue siendo un valor bastante alto. Al mismo tiempo, se observa como en todos los casos también hay una alta cantidad de certificados inválidos, los cuales probablemente están asociados a servicios no pensados para ser usados de forma pública. Por otro lado, en todos los gráficos se cumple

que los certificados con algoritmo de firma inseguro, tamaño de llave inseguro y expirados no son tantos en comparación al conjunto total.

Darknet del CLCERT

Escaneos en tiempo real entre el 24 de mayo de 2019 y el 9 de julio de 2019.

El funcionamiento de las *Darknet* es explicado en la introducción de este trabajo, así como también la existencia de una en la infraestructura de escaneos del CLCERT es explicada en el capítulo 1. Como ya se ha mencionado, el tamaño de los datos de la darknet es muy grande como para almacenar todo lo recolectado, por lo que se decidió recolectar solamente las cabeceres TCP en una primera fase del proyecto. Sin embargo, durante el desarrollo del sistema de importación de datos de la *Darknet*, el servidor encargado de recibir su información dejó de funcionar, por lo que solamente se recopiló un mes y medio de datos. En este periodo, se capturaron mensajes de 750.090 IP distintas del universo total de IPv4.

Debido a estos problemas, se decidió no utilizar la información obtenida por la *darknet* del CLCERT para el desarrollo de estas estrategias, quedando propuesto como trabajo futuro su revisión, a la espera de que se puedan conseguir los recursos necesarios para reiniciar ese proyecto.

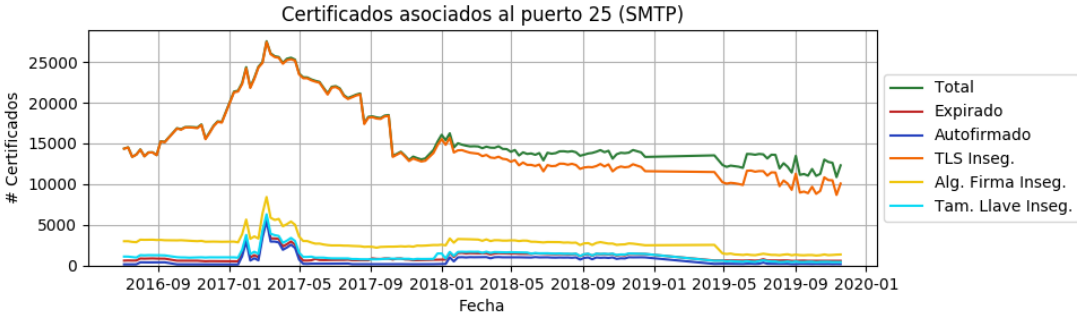
2.2.2. Datos de Censys

Una fuente no usada de forma regular por el CLCERT, pero sí utilizada para este trabajo, corresponde a datos de escaneo de puertos y protocolos de la organización Censys, mencionada en la introducción de este trabajo. La razón por la que los datos de Censys no son utilizados regularmente es que su agregación para la entrega de reportes a terceros requiere de la adquisición de una licencia comercial especial, la cual involucra un costo monetario. Sin embargo, debido a que este trabajo en específico no involucra reportes a terceros, no se transgrede la licencia en este caso.

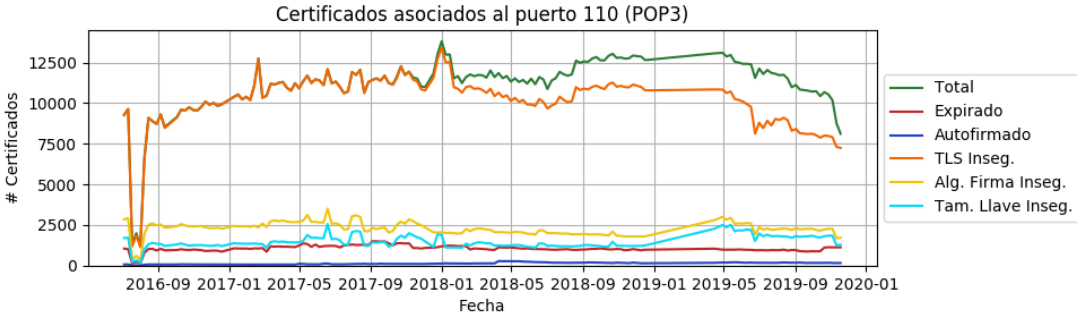
Como se explicó en el capítulo 1, *Censys* es una empresa fundada por los mismos investigadores que diseñaron y desarrollaron la suite de herramientas *ZMap*. A través de su utilización en sus propios servidores, ofrecen el servicio de monitoreo de redes a través de escaneos a quienes estén interesados.

En su trabajo, ellos recolectan periódicamente información de escaneo de puertos sobre todas las IPv4 de Internet y certificados SSL sobre dominios populares según el Ranking Alexa. Los datos puros pueden ser solicitados en caso de querer usarlos para proyectos de investigación, siempre y cuando esta utilización sea sin fines de lucro y sin intenciones de entregarla a terceras personas. Lo anterior limita la utilización de este conjunto de información a la presente investigación. Sin embargo, se considera de gran utilidad debido a que permitirá contrastar los datos obtenidos por Censys con los obtenidos por CLCERT, validando sus resultados.

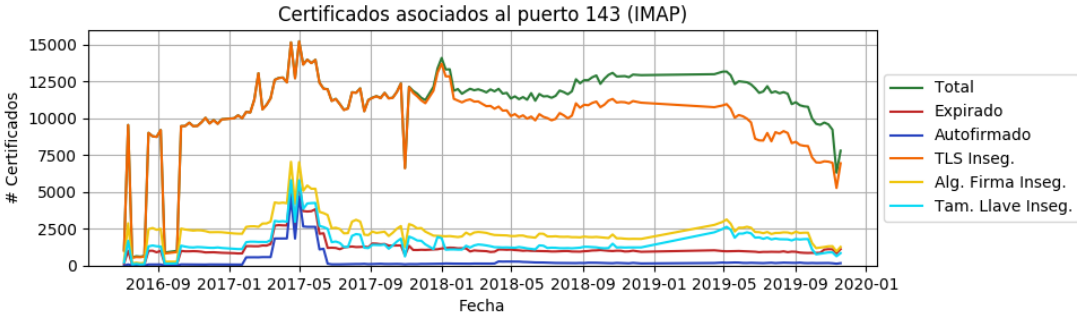
La definición de "red chilena" de Censys es, según su sitio web [10], la misma utilizada por *MaxMind Geolite2*, por lo que esta definición es compatible con la usada por el CLCERT en sus escaneos.



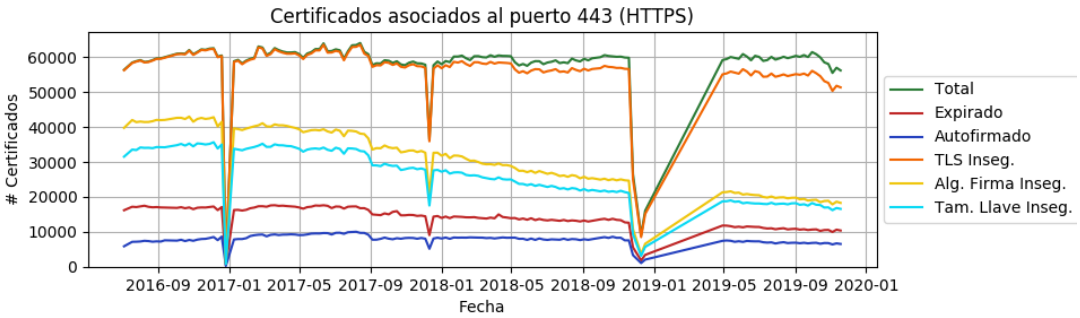
(a) Puerto 25



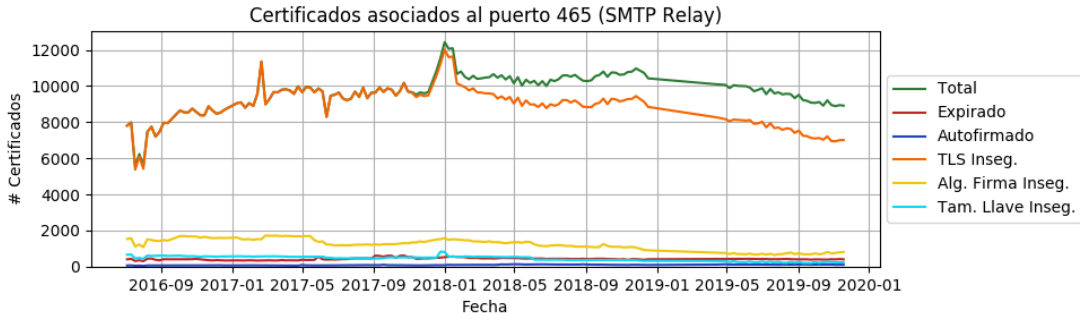
(b) Puerto 110



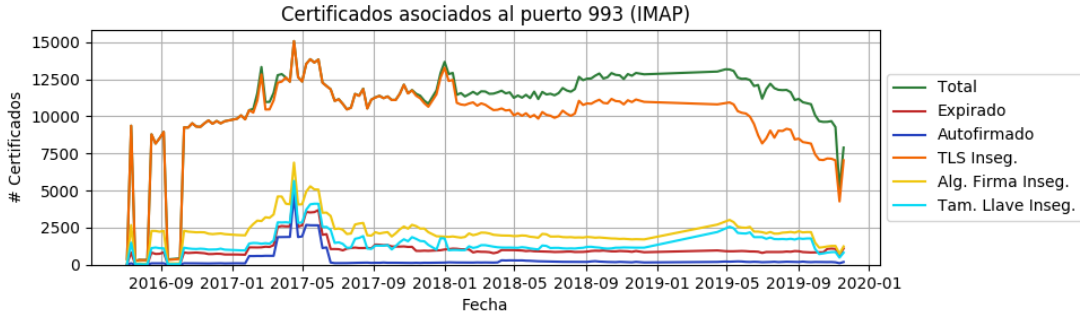
(c) Puerto 143



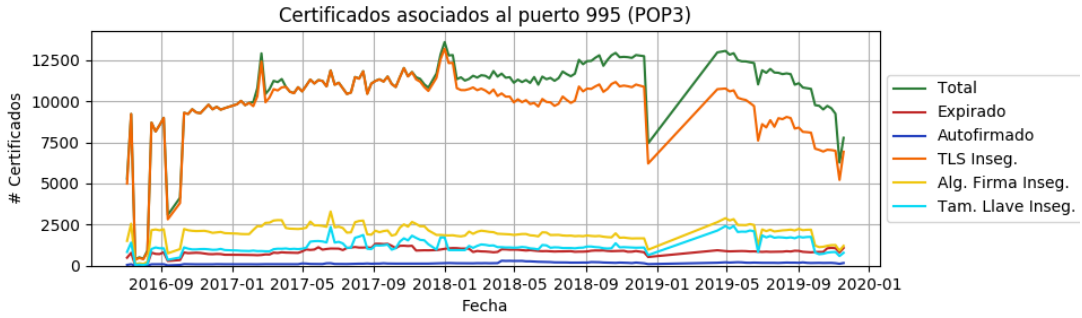
(d) Puerto 443



(e) Puerto 465



(f) Puerto 993



(g) Puerto 995

Figura 2.4: Historial con cantidad de máquinas detectadas con certificado SSL y TLS usado en el protocolo de cada puerto estudiado.

Al contrario del CLCERT, Censys pone a disposición solamente los escaneos de máquinas con puertos abiertos y protocolos activos en la Internet, por lo cual no se cuenta con información de escaneos de puertos sin protocolos para contrastar.

2.2.3. Escaneo de Protocolos, Software y Versiones

Obtenidos entre el 21 de noviembre de 2016 y el 18 de noviembre de 2019

Censys escanea una veintena de protocolos distintos, cuyo detalle se puede encontrar en su página web [10]. La información de la tabla 2.3 y la figura 2.5 muestra datos estadísticos generales de cada puerto revisado por Censys. Al comparar esta tabla con la equivalente del CLCERT, se puede ver que los puertos más populares son aproximadamente los mismos (80, 631, 22, 443, 8080), sin embargo, Censys no revisa el puerto 8000.

Censys aclara en su página de preguntas frecuentes que los resultados de escaneos de protocolos son realizados *al menos una vez a la semana*. Debido a lo anterior y a que los datos de Censys están sobre Google BigQuery (sistema que requiere del pago de los recursos computacionales usados para la exportación), se decidió recopilar la información de escaneos publicada el lunes de cada semana. La sección de Preguntas Frecuentes también menciona que ellos clasifican sus resultados geográficamente según los datos de la base de datos Geolite2 de MaxMind. Teniendo en cuenta que la actualización de los datos geográficos de esta fuente es semanal, el universo de IPs entre los escaneos del CLCERT y Censys debiese ser el mismo.

Los datos de Censys ya mencionados fueron importados a la base de datos del CLCERT a través del sistema OSR. Este sistema realizó el mismo procedimiento que en el caso de los datos del CLCERT para obtener información de software y versiones válidas.

La comparación directa de estos datos con los del CLCERT requiere establecer algunos supuestos, los cuales se justificarán más adelante. Uno de estos supuestos corresponde a que los resultados de escaneo de protocolos no varían considerablemente dentro de una misma semana, ya que los escaneos de ambas fuentes son realizados en días y horas distintos, y se usará al analizar los gráficos de la figura 2.6, los que muestran una comparación preliminar entre la cantidad de datos del CLCERT con la cantidad de datos de Censys.

En los gráficos mencionados, se observa que en muy pocas ocasiones los resultados obtenidos son similares.En los gráficos 2.6b, 2.6c, 2.6e y 2.6f, los escaneos del CLCERT suelen obtener más resultados, mientras que en los gráficos 2.6g y 2.6h, Censys consigue un mayor número de máquinas con el protocolo activo.

La figura 2.6a muestra que la fuente CLCERT obtiene resultados muy variables desde Enero de 2018, mientras que Censys se ha mantenido en valores relativamente estables. Por otro lado, también se nota que el puerto 80 es el único que encuentra resultados parecidos, desde mayo de 2017. Se propone que la disminución de máquinas detectadas que se observa en el gráfico 2.6d se debe a un problema de configuración en los escaneos de Censys, el cual considera temporalmente un conjunto mayor al real de IPs asociadas a Chile. En las mediciones posteriores, se nota un número menor de resultados de parte del CLCERT hasta mayo de 2018, momento en el cual estos se asemejan bastante al número de resultados del Censys.

Puerto	Desde	Hasta	Escaneos	Semanas	Cobertura	$min(\#IPs)$	$max(\#IPs)$	$\bar{X}\#IPs$	$\#uniq(IPs)$
21	2016-11-21	2019-11-18	142	157	0,90	1	19.886	14.894,23	198.761
22	2017-09-04	2019-05-27	87	91	0,96	1.919	34.996	23.655,26	436.636
23	2017-12-18	2019-11-18	100	101	0,99	21.489	32.221	26.611,60	234.052
25	2016-11-21	2019-11-11	143	156	0,92	6.564	21.554	10.759,40	51.674
53	2017-12-18	2019-11-18	95	101	0,94	14.258	23.223	17.393,37	117.361
80	2016-11-21	2019-11-18	147	157	0,94	189	576.607	151.235,80	1.902.959
102	2017-12-18	2019-11-18	98	101	0,97	1	3	1,15	9
110	2016-11-21	2019-11-11	144	156	0,92	5.075	9.358	7.452,80	23.640
143	2016-11-21	2019-11-11	146	156	0,94	3.143	9.172	6.738,46	23.081
443	2018-10-15	2019-11-18	57	58	0,98	23.451	54.876	45.840,09	402.497
445	2017-12-18	2019-11-18	100	101	0,99	559	8.765	3.090,36	28.438
465	2018-12-10	2019-11-18	50	50	1,00	208	6.964	4.815,76	10.469
502	2017-12-18	2019-11-18	100	101	0,99	73	137	94,51	5.263
587	2018-01-08	2019-11-18	98	98	1,00	2.235	10.726	5.289,37	17.323
623	2018-12-10	2019-11-18	50	50	1,00	15	23	18,84	36
631	2018-09-17	2019-11-18	62	62	1,00	4.179	273.502	34.907,68	955.853
993	2016-11-21	2019-11-11	143	156	0,92	43	28.604	6.544,18	41.594
995	2016-11-21	2019-11-11	145	156	0,93	162	8.262	6.599,36	19.705
1433	2018-05-21	2019-11-18	79	79	1,00	2.100	3.868	2.444,22	15.698
1521	2018-05-21	2019-11-18	77	79	0,97	134	316	187,90	1.457
1883	2019-03-18	2019-11-18	36	36	1,00	75	98	84,83	410
1900	2017-12-18	2018-08-27	36	37	0,97	7.217	7.217	7.217,00	7.217
1911	2017-12-18	2019-11-18	86	101	0,85	1	6	2,08	96
2323	2017-12-18	2019-11-18	100	101	0,99	45	230	93,95	534
3306	2018-05-21	2019-11-18	79	79	1,00	4	11.013	6.188,76	40.553
3389	2018-12-17	2019-11-18	49	49	1,00	7.517	9.263	8.523,45	50.836
5432	2018-06-04	2019-11-18	77	77	1,00	655	1.117	785,00	3.702
5632	2019-01-14	2019-11-18	38	45	0,84	7	22	11,84	69
5672	2019-01-21	2019-11-18	36	44	0,82	56	73	63,42	184
5900	2018-11-19	2019-11-18	53	53	1,00	897	1.576	1.138,87	13.631
5901	2018-11-19	2019-11-18	53	53	1,00	170	236	201,11	1.440
5902	2018-11-19	2019-11-18	53	53	1,00	108	134	122,91	740
5903	2018-11-19	2019-11-18	52	53	0,98	45	65	54,50	526
6443	2019-06-17	2019-11-18	23	23	1,00	4	10	5,78	12
7547	2017-12-18	2019-11-18	100	101	0,99	5.501	26.164	12.010,23	143.302
8080	2017-09-11	2019-11-11	113	114	0,99	7.134	37.159	17.987,39	543.020
8883	2019-03-18	2019-11-18	36	36	1,00	32	44	35,83	323
8888	2017-12-18	2019-11-18	100	101	0,99	457	4.709	1.651,71	13.152
9090	2019-06-17	2019-11-18	23	23	1,00	3	7	4,91	11
9200	2018-12-03	2019-11-18	37	51	0,73	5	66	18,97	88
16992	2018-11-26	2019-11-18	51	52	0,98	18	46	25,78	212
16993	2018-12-17	2019-11-18	49	49	1,00	2	9	4,31	59
20000	2017-12-18	2019-11-04	90	99	0,91	1	5	1,86	125
27017	2018-09-17	2019-11-18	62	62	1,00	38	58	48,15	241
47808	2017-12-18	2019-11-18	100	101	0,99	4	9	6,69	27

Tabla 2.3: Tabla que muestra los rangos de tiempo que se ejecutaron los escaneos de protocolos de Censys de cada puerto. La tabla también muestra la cantidad de escaneos realizados en ese rango temporal, y el promedio de escaneos por semana. La columna Cobertura representa la proporción entre las semanas efectivamente escaneadas y el rango de semanas del protocolo.

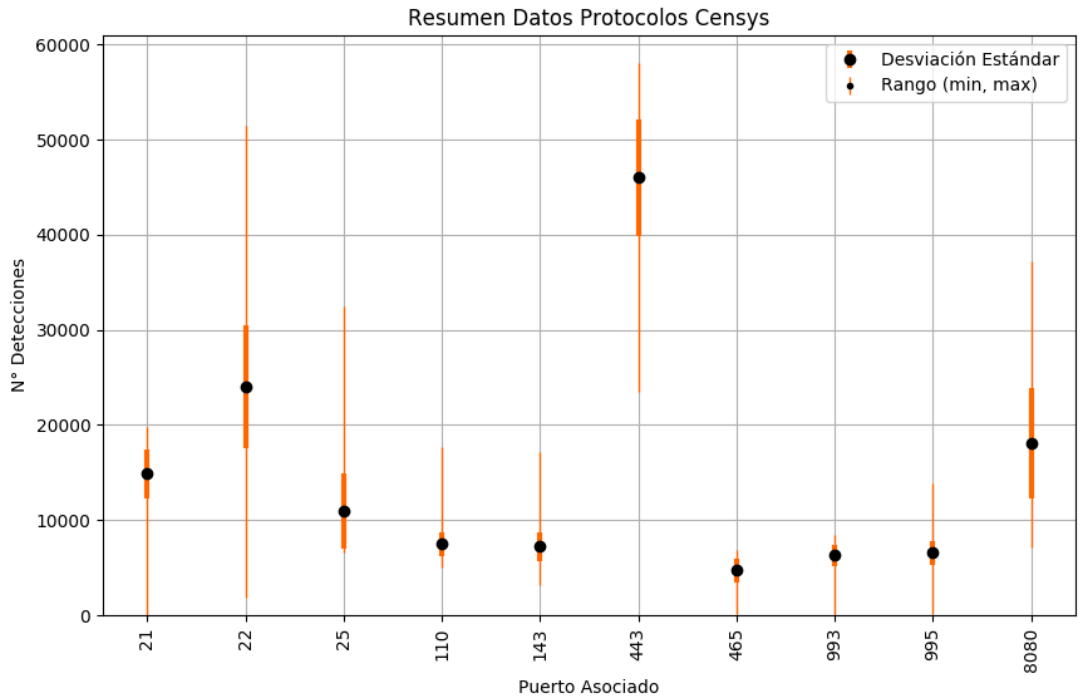


Figura 2.5: Resumen gráfico de los resultados de número IPs mínimas, máximas, promedio y desviación estándar del histórico de escaneos de puertos de Censys. Para mejorar la visibilidad del gráfico, se consideraron solamente los escaneos con más de mil IPs en promedio.

2.2.4. Datos de Malware Fuentes Reservadas

Desde hace 2 años, CLCERT recibe información periódica de 2 fuentes que recopilan datos de malware en la Internet, la cual es puesta a disposición de los administradores de redes que la soliciten. Sin embargo, ambas fuentes requieren la no publicación de sus nombres para poder publicar los resultados obtenidos por ellas. Debido a lo anterior, en este documento serán denominadas como *Fuente Reservada #1* y *Fuente Reservada #2*.

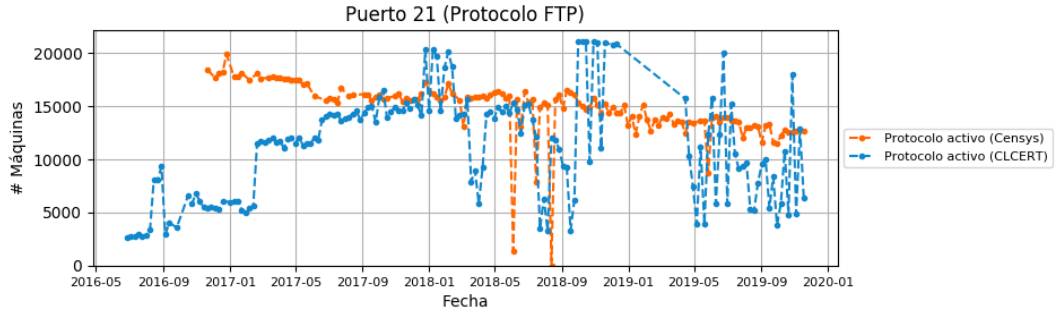
Con respecto al concepto de "red chilena" de estas dos fuentes, ninguna de ellas explicita cómo determina que los datos recopilados corresponden o no a nuestro país. Por lo tanto, se recomienda que el uso de estos datos se realice de forma complementaria a otra fuente cuya ubicación sí se conozca, como por ejemplo, intersectando estos datos con una definición de Red Chilena más confiable o transparente.

Datos de Malware de Fuente Reservada #1

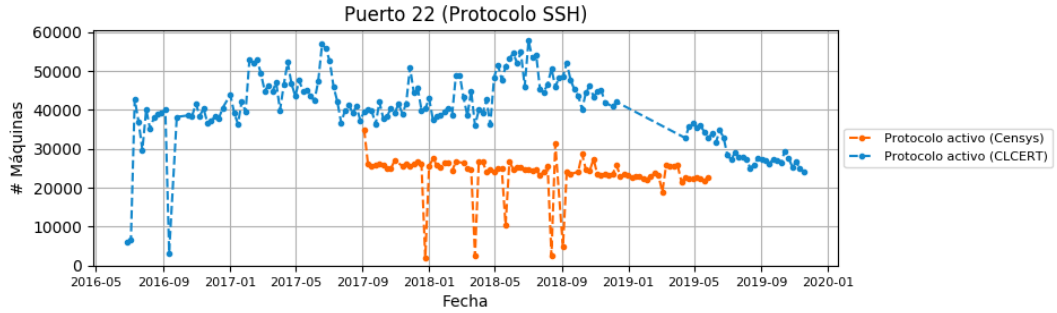
Entre el 5 de septiembre de 2017 y el 22 de noviembre del 2019

La Fuente Reservada #1 entrega de manera diaria datos de reportes de malware en 7 categorías distintas, que incluyen momento de detección, IP involucrada, una categoría de amenaza y un banner con un formato bastante poco estructurado, el cual cuenta a veces con información extra sobre la amenaza detectada.

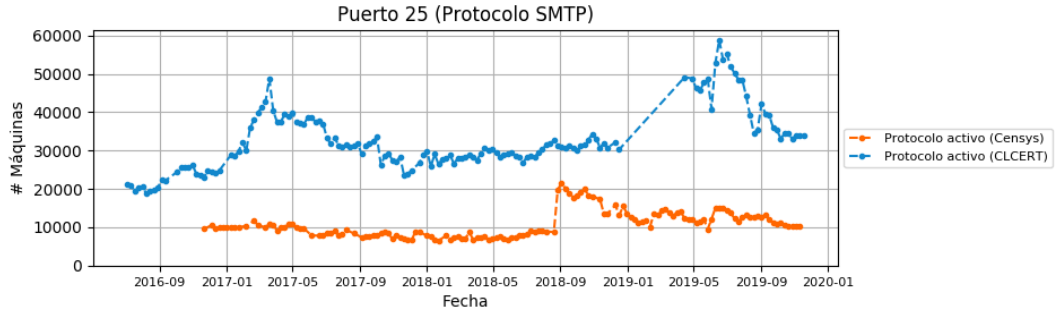
Las categorías usadas por esta fuente son las siguientes:



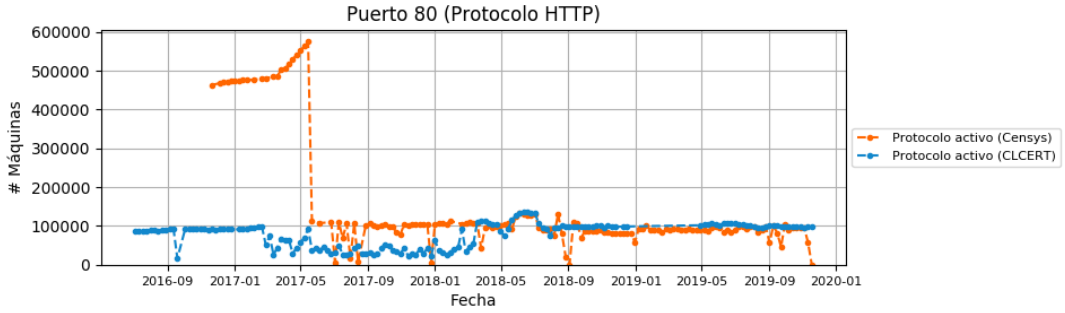
(a) Puerto 21



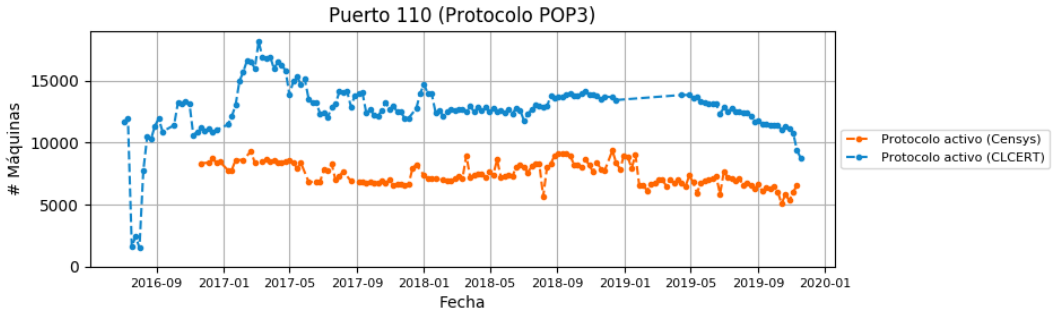
(b) Puerto 22



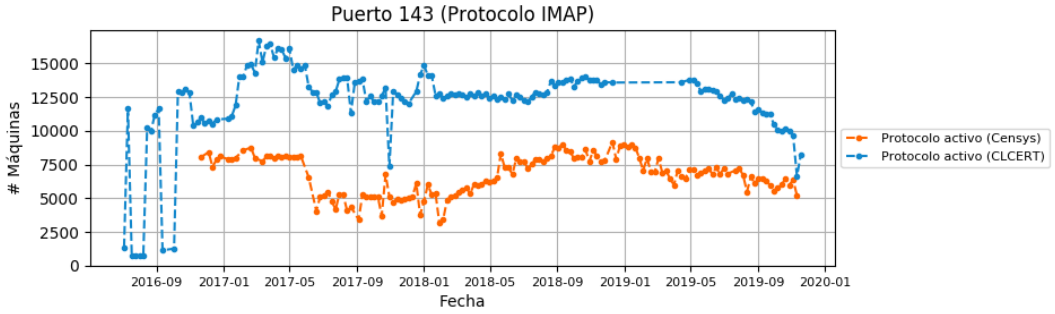
(c) Puerto 25



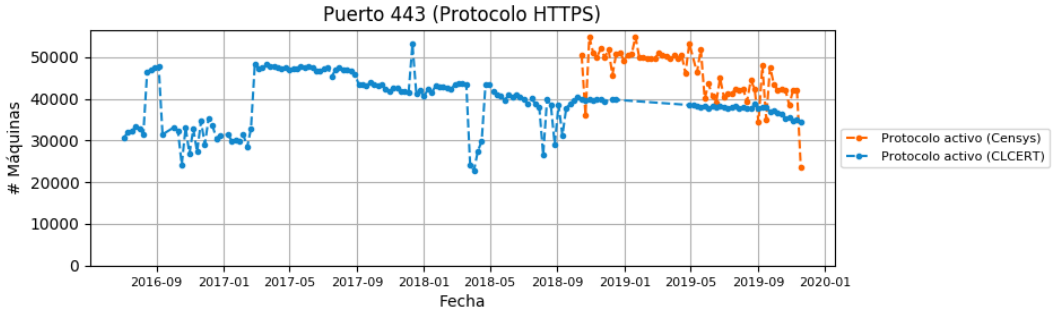
(d) Puerto 80



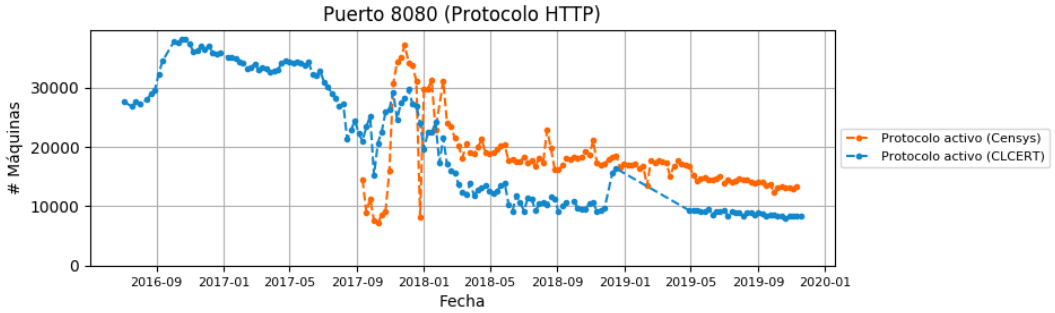
(e) Puerto 110



(f) Puerto 143



(g) Puerto 443



(h) Puerto 8080

Figura 2.6: Comparación de resultados de CLCERT y Censys que muestra cuántas máquinas de la red chilena poseen corriendo un servicio asociado a cada puerto.

- **Bot**, es decir, infectada con algún malware que la hace actuar como bots conocidos.
- **Fuerza Bruta,** o detectadas realizando ataques de fuerza bruta sobre algún servicio.
- **Controlador C&C,** o bajo sospecha de controlar Botnets.
- **Detectada por Darknet,** es decir, la IP fue detectada enviando paquetes a una subred controlada por una *darknet*.
- **Detectada por Honeypot,** es decir, la IP fue detectada enviando paquetes a un servidor *honeypot*.
- **Resolvers DNS abiertos,** o máquinas habilitadas como resolvers DNS para uso de cualquiera.
- **Phishing,** o máquinas detectadas como hospedaje de sitios de *phishing*.
- **Proxy,** o máquinas corriendo un proxy sin credenciales.
- **SPAM,** o máquinas enviando correo electrónico no deseado.

La fuente reservada #1 envía datos correspondientes a la subred administrada. En el caso del CLCERT, esto corresponde a todo el conjunto de IPs chilenas. Sin embargo, no especifican de qué forma determinan la ubicación geográfica de los resultados. En pruebas realizadas sobre los datos obtenidos, se determinó que las IPs declaradas como chilenas no se parecen a las declaradas por la base de datos Geolite. Al mismo tiempo, esta fuente reservada no da detalles de la metodología con la que se obtienen estos datos, haciendo imposible su replicación de forma certera.

Un resumen de los datos manejados similar a los ya mostrados para los datos de CLCERT se puede encontrar en la tabla 2.4, la cual revela tanto la importancia como la recurrencia de estos reportes diarios. Destaca de estos resultados que el reporte de malware de tipo *Command and Control* (C2) no entrega más de una decena de valores por día, y solamente el 15 % de los días transcurridos desde el primer hasta el último escaneo recibido. Una situación similar ocurre con el reporte *phishing*, el cual ha entregado uno o más valores el 43 % de los días escaneados, con un máximo de solamente 20 coincidencias. En contraposición, reportes como *bot, bruteforce y darknet* han sido recopilados el 96 % de los días, con un promedio de entre 1.804 y 29.017 valores.

Los gráficos en 2.7 (páginas 53 y 54) muestran los números históricos diarios para cada tipo de amenaza. Se puede observar que un gran número de estos se comporta de forma bastante irregular. Dentro del grupo anterior, destacan los resultados de *fuerza bruta* en 2.7b, debido al alza abrupta de 60 mil dispositivos entre abril y mayo de 2019. También llaman la atención los resultados de 2.7c, debido a la inexistencia de valores entre mayo de 2018 y fines de marzo de 2019.

El gráfico de las máquinas *Command and Control* detectadas mostrado en 2.7f muestra un claro patrón de altos periódicos, el cual puede tener que ver con la forma en que los mismos datos son reportados, debido a que no corresponden a máquinas infectadas. Será necesario revisar en mayor profundidad estos valores al momento de utilizarlos, de forma de entender su comportamiento.

Por último, cabe destacar que varias categorías cuentan con menos de 300 valores de forma regular, algunas incluso teniendo como valores máximos no más de 40 resultados, como es el

caso de *Command and Control* y *Phishing*.

Tipo	Desde	Hasta	Escaneos	Días	Cobertura	$min(\#Hits)$	$max(\#Hits)$	$\bar{X}(\#Hits)$
bot	2017-05-10	2019-11-23	903	927	0,97	186	18.588	9.697,49
bruteforce	2017-05-10	2019-11-21	879	925	0,95	1	60.809	1.804,57
c2	2018-04-30	2019-11-21	85	570	0,15	1	34	8,78
darknet	2018-04-06	2019-11-21	569	594	0,96	537	64.478	29.017,95
honeypot	2018-04-06	2019-11-21	567	594	0,95	1	1.939	155,21
dnsresolver	2017-05-10	2019-11-21	886	925	0,96	6	1.332	65,53
phishing	2017-05-11	2019-11-18	396	921	0,43	1	20	2,97
proxy	2017-05-11	2019-11-23	802	926	0,87	2	356	146,09
spam	2017-11-09	2019-11-21	574	742	0,77	1	712	94,16

Tabla 2.4: Tabla que muestra los rangos de tiempo que se ejecutaron los escaneos de malware de la fuente reservada #1. La tabla también muestra la cantidad de escaneos realizados en ese rango temporal, y mínimo, máximo, promedio y desviación estándar de escaneos por semana, tratando de notar a simple vista la completitud de cada conjunto de datos.

Tipo	Desde	Hasta	Escaneos	Días	Cobertura	$min(\#Hits)$	$max(\#Hits)$	$\bar{X}(\#Hits)$	$\#uniq(Hits)$
bot	2018-03-09	2019-11-21	609	622	0,98	117	11.562	6.658,71	1.953,44
bruteforce	2018-04-04	2019-11-21	578	596	0,97	2	2.882	646,75	565,49
darknet	2018-10-30	2019-04-27	163	179	0,91	8	1.060	563,94	187,36

Tabla 2.5: Tabla que muestra los rangos de tiempo que se ejecutaron los escaneos de malware de la fuente reservada #2. La tabla también muestra la cantidad de escaneos realizados en ese rango temporal, y mínimo, máximo, promedio y desviación estándar de escaneos por semana, tratando de notar a simple vista la completitud de cada conjunto de datos.

Datos de Malware de Fuente Reservada #2

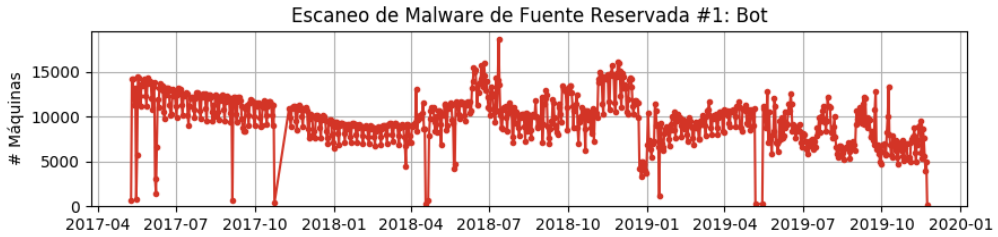
De forma diaria o semanal, entre el 8 de marzo de 2018 y el 21 de noviembre del 2019

Esta fuente entrega reportes sobre la red chilena mucho más variados y estructurados que la fuente anterior, los cuales suman casi 100 informes distintos. Sin embargo, la estructura de cada uno de ellos es bastante distinta, por lo que es necesario desarrollar un modelo de datos especial para cada reporte si es que se quiere aprovechar de la mejor manera.

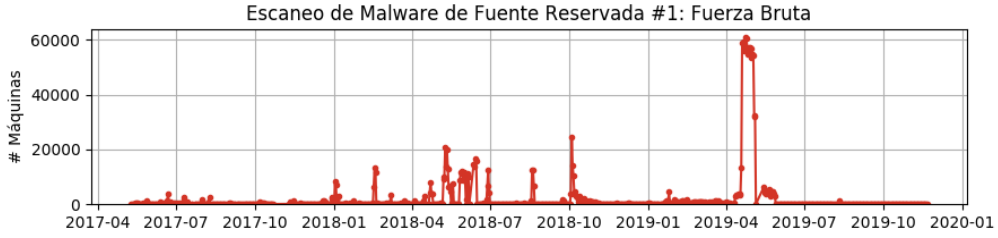
Debido a lo anterior, se decidió elegir un subconjunto de estos reportes para procesar en el sistema OSR, los cuales se clasificaron en categorías similares a las de los reportes de la fuente reservada #1: **Bot**, **Fuerza Bruta** y **Darknet**.

El mapeo de estos reportes es el siguiente:

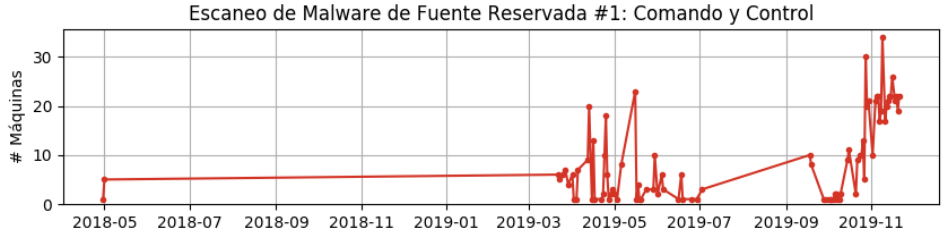
- El reporte *botnet* reúne reportes de botnets obtenidos por esta fuente, y fue lógicamente asociado a la categoría *botnet*.
- El reporte *brute-force* reúne reportes de ataques de fuerza bruta obtenidos por esta fuente, y como es de esperar fue asociado a la categoría *bruteforce*.
- El reporte *sinkhole-http* contiene IPs de máquinas que solicitaron un dominio usado generalmente por bots a un *sinkhole* de la fuente, por lo cual estas máquinas fueron asociadas a la categoría *botnet*



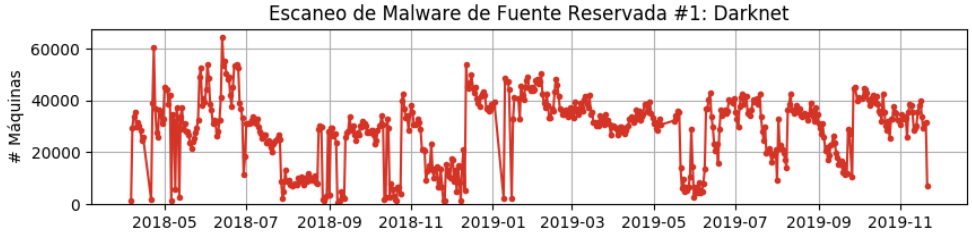
(a) bot



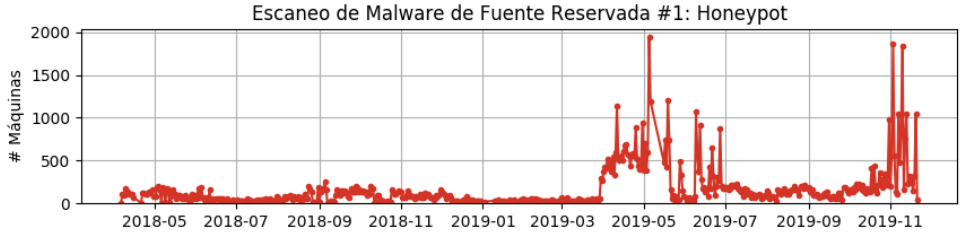
(b) bruteforce



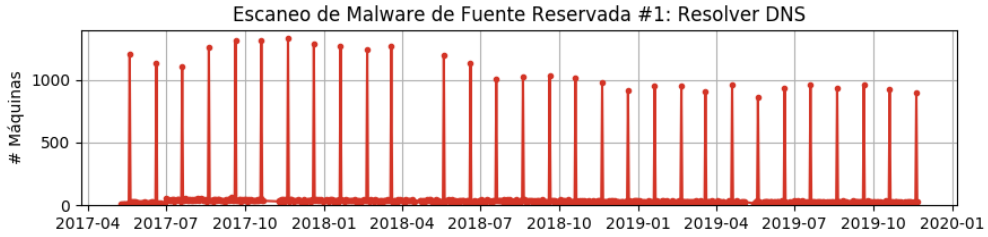
(c) c2



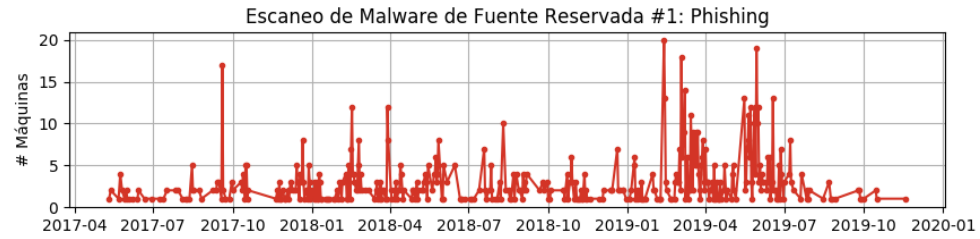
(d) darknet



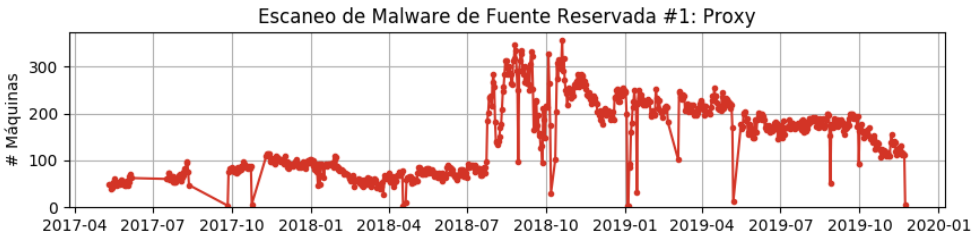
(e) honeypot



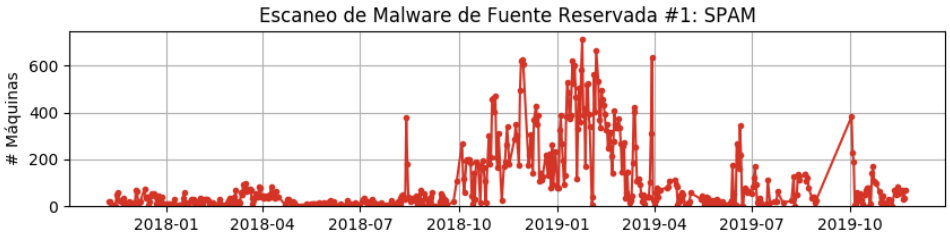
(f) dnsresolver



(g) phishing

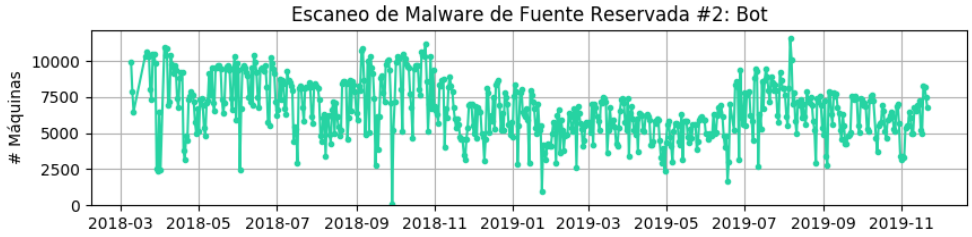


(h) proxy

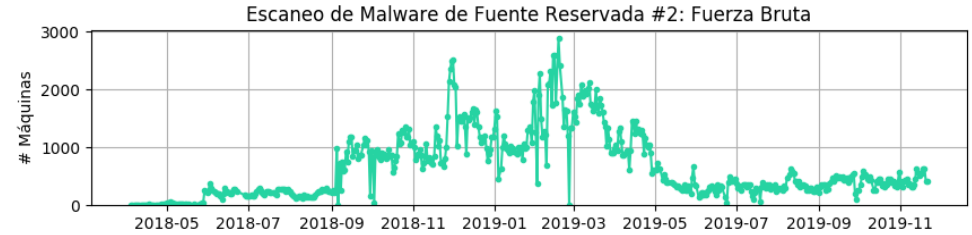


(i) spam

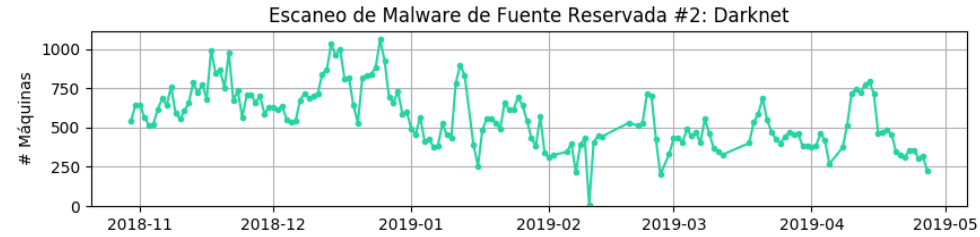
Figura 2.7: Historial con cantidad de máquinas detectadas por escaneo realizado por la Fuente Reservada #1, divididos por tipo.



(a) Tipo bot .



(b) Tipo bruteforce .



(c) Tipo darknet .

Figura 2.8: Historial con cantidad de máquinas detectadas por escaneo realizado por la Fuente Reservada #2 de cada tipo.

- El reporte *sinkhole-microsoft* contiene IPs de máquinas detectadas por un *sinkhole* de Microsoft, por lo cual estas máquinas fueron asociadas a la categoría *botnet*
- El reporte *darknet* contiene reportes de IPs detectadas por la *darknet* de la fuente, por lo que estas fueron clasificadas en *darknet*.

Gran parte de los otros reportes de esta fuente corresponden a servicios abiertos, y no se revisarán en este trabajo.

La tabla 2.5 muestra un resumen numérico de los valores agrupados en las tres categorías usadas. Esta tabla muestra que la cantidad de máquinas encontradas no es tan grande como la reportada por la fuente externa #1, sin embargo, la cobertura es del mismo orden que la otra fuente, oscilando entre el 91 y el 98 %. La cantidad promedio de IPs aportadas por día oscila también entre 563,94 y 6.658,71.

A continuación, se revisará el conjunto de gráficos 2.8. Se observa en la figura 2.8a una cantidad alta y constante de reportes diarios de tipo *bot*, variando harto entre semanas pero manteniéndose relativamente constante en promedio. En el caso de los reportes *bruteforce* de la figura 2.8b, se observa un aumento de resultados cercano a los mil reportes entre septiembre de 2018 y mayo de 2019, teniendo como peak el periodo de fines de febrero de 2019, en un valor de 2.882 máquinas. El reporte de tipo *darknet* de la figura 2.8c muestra resultados desde noviembre de 2018 solamente, con valores más altos hacia finales del 2018 y peaks periódicos aproximadamente 2 veces al mes.

Es importante mencionar que, al igual que en la fuente anterior, esta fuente no explicita la forma en la que determina qué reportes son considerados como parte de la red chilena y cuales no, por lo que las comparaciones realizadas entre ambos conjuntos tendrán que hacerse teniendo esta situación en cuenta.

2.3. Dificultades

A partir de la revisión preliminar de datos del CLCERT y de Censys, es posible reconocer las siguientes dificultades relacionadas con la calidad de las fuentes de datos.

2.3.1. Existencia de lagunas de datos en algunos intervalos de tiempo

Como ya se vio, algunas fuentes poseen "lagunas de tiempo" o periodos largos sin información debido a diversos motivos. En el caso de los escaneos del CLCERT, es posible observar en todos los gráficos la ausencia de datos entre diciembre del año 2018 y abril del año 2019, la cual es debido a problemas de almacenamiento en el sistema antiguo usado por el servidor de escaneo. Una propuesta de solución de infraestructura a este problema para futuros escaneos se puede leer en el Capítulo 5 de este trabajo.

2.3.2. Existencia de datos con alta varianza en cortos intervalos de tiempo

Se nota especialmente en el caso de los escaneos de puertos, pero también en algunos escaneos de protocolos, que algunos resultados entre semanas varían en grandes proporciones de forma bastante breve.

En el caso de escaneos de puertos, su alta variabilidad y la inexistencia de otra fuente con la cual corroborar motivan a que estos no se usen como fuente de información primaria en las estrategias a proponer.

En el caso de escaneos de protocolos, el trabajo de Eduardo Acha mostró que estas variaciones pueden deberse, en algunos casos, a cambios en configuraciones de red de Sistemas Autónomos grandes, por lo que se continuará con este supuesto a lo largo de este trabajo.

Una revisión más en detalle de este problema se dará como parte de una estrategia de uso de datos presentada en el Capítulo 4.2 de este trabajo.

2.3.3. Diferencias de resultados entre datos de distintas fuentes

Como se vio en el gráfico 2.6, el número de escaneos de protocolos varía ampliamente entre ambas fuentes para una misma semana en algunos protocolos, lo cual puede deberse a varias razones, como por ejemplo, a diferencias de tiempo en que estos se escanean (las cuales están contenidas en el rango de la semana); a diferencias de metodología, implementación y configuraciones en el software utilizado para realizar los escaneos; a situaciones de IPs en listas negras que bloquean el tráfico en algunos servicios o a respuestas diferentes desde distintas redes y puntos geográficos debido a bloqueos o a la existencia de *anycast*, configuraciones distintas en términos de conectividad de parte de los sistemas autónomos. A continuación, se analizan estas cuatro posibilidades.

Escaneos de protocolos en distintos días dentro de una misma semana

Los escaneos del CLCERT son realizados todas las semanas, los mismos días, y una vez a la semana. En específico, los días lunes se escanean los protocolos relacionados con servicios web (80, 443, 8000, 8080), los días martes los protocolos relacionados con servicios de correo electrónico (25, 110, 143, 993, 995) y los días viernes los protocolos FTP y SSH (21, 22).

Con respecto a los escaneos de Censys, según la indicación de la página de preguntas frecuentes de su sitio web, este conjunto de datos ejecuta sus escaneos de protocolos *al menos una vez a la semana*. Esto contradice los resultados obtenidos a través de la revisión manual de sus datos, ya que se observa que esto no se cumple para algunos puertos, los cuales en algunos casos no reciben nuevos valores desde hace más de dos meses (ver figura 2.6b). En el otro extremo, hay protocolos que se escanean de forma diaria, como los puertos 21, 443 y 8080. La mayor cantidad de escaneos a la semana es razón bastante probable del mayor número de IPs únicas en Censys con respecto a CLCERT.

Se decidió revisar qué escaneos de ambas fuentes se habían realizado los mismos días, encontrándose los 97 casos mostrados en la tabla 2.6. En el caso del puerto 443, se observa

que la cantidad de resultados obtenidos en una sola sesión de escaneo es consistentemente menor a la obtenida por CLCERT, con porcentajes que van entre el 51,05 % y el 63,38 % de la cantidad de resultados obtenidos por el CLCERT. Esto muestra que incluso manteniendo el día de escaneos, los resultados en número y valor de IPs detectadas entre ambas fuentes pueden variar.

Basándose en lo anterior, se asumirá que el día de semana en que se realiza el escaneo no ejerce una diferencia considerable en el número de resultados obtenidos para los protocolos estudiados. En contraparte, se asumirá que la unión de las IPs detectadas a lo largo de muchos escaneos dentro de una misma semana sí tiene un impacto en la cantidad de resultados obtenidos. Esto último se justifica a partir de revisar cuántos días distintos a la semana realiza Censys escaneos de IP por cada protocolo. La tabla 2.7 muestra esta situación, contando la cantidad de escaneos promedio por semana. Se observa que en los 4 puertos en que Censys obtiene más resultados que el CLCERT hay al menos casi el doble de escaneos que los realizados por el CLCERT. Esta situación eso sí no se repite con los puertos relacionados con correo electrónico ni con el puerto 22, lo que potencialmente representa a un bloqueo de acceso regional.

Impacto en metodología, tipo de software y configuraciones usadas

Otro factor importante en los resultados obtenidos es la metodología usada para realizar cada escaneo, así como también las herramientas que detectan los puertos y capturan los banners. En este caso concreto, ambas fuentes utilizan *ZMap* para detectar puertos abiertos, sin embargo, Censys utiliza la herramienta *ZGrab*, mientras que el CLCERT usa *Mercury*.

Con respecto a la metodología, no hay detalles de la cantidad, tipo de máquinas e IPs usadas por éstas de parte de Censys para obtener sus resultados, por lo cual se hace difícil una posible repetición de los experimentos que ellos ejecutan. Esta decisión es comprensible desde un punto de vista comercial, pero limita las posibilidades de trabajo con estos datos, teniendo mayor utilidad como apoyo que como resultados exactos.

Sobre el software y configuraciones, si bien se sabe qué programa de escaneo usa cada fuente, no hay detalles en el caso de Censys sobre qué configuración se está usando tanto en la máquina como en el mismo software. Lo anterior provoca que no sea suficiente la repetición de los escaneos con *ZGrab* para verificar si las diferencias se deben en parte a las distintas implementaciones, por lo que no se puede conocer por completo en estas condiciones cuál es el impacto de estas variables en el resultado final. La revisión de esta variable, por lo tanto, quedará propuesta como trabajo futuro.

IPs de escaneo en listas negras

Las listas negras o *blacklists* de IP nacen con el objetivo de advertir a administradores de sistemas de IPs que generan o han generado en el pasado tráfico malicioso, de forma de bloquear cualquier solicitud de ellas tratando de prevenir posibles problemas de seguridad.

Algunos servicios que mantienen listas negras de IPs son *Spamhaus* [82] y *SORBS* [81], a través de contar con infraestructura similar a *darknets* que agrega a la lista a todas las IP que se intentan comunicar con sistemas de monitoreo esparcidos por la Internet. Asimismo,

Fecha	Puerto	N^o_{CLCERT}	N^o_{Censys}	%
2017-01-13	21	6.059	17.701	292,14
2017-01-20	21	6.043	17.701	292,92
2017-03-17	21	11.804	16.212	137,34
2018-06-08	22	54.584	20.479	37,52
2018-09-07	22	48.598	3.830	7,88
2016-12-20	25	24.817	9.905	39,91
2017-02-07	25	29.945	10.394	34,71
2017-02-28	25	39.890	10.565	26,49
2017-03-21	25	48.794	10.914	22,37
2017-03-28	25	40.446	10.637	26,30
2017-04-04	25	37.422	9.120	24,37
2017-04-11	25	37.618	10.041	26,69
2017-04-18	25	39.396	10.085	25,60
2017-04-25	25	38.916	10.719	27,54
2017-05-02	25	39.854	10.735	26,94
2017-05-09	25	37.390	10.105	27,03
2017-05-16	25	37.121	9.641	25,97
2017-05-23	25	36.850	9.641	26,16
2018-02-06	25	28.000	1.331	4,75
2018-11-13	25	30.788	15.568	50,57
2018-12-11	25	32.147	12.182	37,89
2016-11-21	80	91.154	463.598	508,59
2016-12-05	80	91.359	467.970	512,23
2017-01-23	80	93.669	475.783	507,94
2017-10-16	80	37.503	5.090	13,57
2017-10-23	80	35.050	4.114	11,74
2018-08-13	80	94.734	18.747	19,79
2018-09-10	80	98.655	52.003	52,71
2019-09-30	80	99.139	64.385	64,94
2016-12-20	110	11.005	8.347	75,85
2017-02-07	110	15.709	8.570	54,55
2017-02-28	110	15.991	8.352	52,23
2017-03-21	110	16.774	8.596	51,25
2017-03-28	110	16.920	8.482	50,13
2017-04-04	110	15.948	8.508	53,35
2017-04-11	110	16.511	8.364	50,66
2017-04-18	110	16.215	8.377	51,66
2017-04-25	110	15.828	8.441	53,33
2017-05-02	110	13.844	8.559	61,82
2017-05-09	110	14.943	8.381	56,09
2017-05-16	110	15.299	7.918	51,76
2017-05-23	110	14.689	8.329	56,70
2018-12-11	110	13.655	7.716	56,51
2016-12-20	143	10.841	7.831	72,24
2017-02-07	143	14.030	8.535	60,83
2017-02-28	143	14.315	7.991	55,82
2017-03-21	143	16.291	8.092	49,67
2017-03-28	143	16.436	8.149	49,58
2017-04-04	143	15.435	7.960	51,57
2017-04-11	143	16.091	8.093	50,30
2017-04-18	143	16.029	8.018	50,02
2017-04-25	143	15.386	8.159	53,03
2017-05-02	143	16.087	8.053	50,06
2017-05-09	143	14.535	8.073	55,54
2017-05-16	143	14.854	8.030	54,06
2017-05-23	143	14.627	8.089	55,30
2018-12-11	143	13.581	7.749	57,06
2018-10-15	443	39.862	6.665	16,72
2018-10-22	443	39.740	3.313	8,34
2018-10-29	443	39.762	23.421	58,90
2018-11-05	443	39.524	22.993	58,17
2018-11-12	443	39.876	22.728	57,00
2018-11-19	443	39.842	22.906	57,49
2018-11-26	443	39.350	22.579	57,38
2018-12-10	443	39.883	13.582	34,05
2018-12-17	443	39.755	23.110	58,13
2019-04-29	443	38.502	22.223	57,72
2019-05-13	443	38.274	20.849	54,47
2019-05-20	443	38.034	22.041	57,95
2019-05-27	443	38.353	19.581	51,05
2019-06-03	443	37.725	22.715	60,21
2019-06-10	443	38.132	22.349	58,61
2019-06-17	443	37.984	20.468	53,89
2019-06-24	443	38.208	22.710	59,44
2019-07-01	443	38.093	22.133	58,10
2019-07-08	443	37.737	20.650	54,72
2019-07-15	443	37.971	21.667	57,06
2019-07-22	443	38.311	22.460	58,63
2019-07-29	443	37.639	22.302	59,25
2019-08-05	443	37.988	22.238	58,54
2019-08-12	443	37.778	22.177	58,70
2019-08-19	443	37.666	21.972	58,33
2019-08-26	443	38.804	22.283	57,42
2019-09-02	443	37.804	21.791	57,64
2019-09-09	443	37.886	22.460	59,28
2019-09-16	443	37.914	22.248	58,68
2019-09-23	443	36.932	22.276	60,32
2019-09-30	443	37.197	22.910	61,59
2019-10-07	443	36.594	21.725	59,37
2019-10-14	443	36.261	22.493	62,03
2019-10-21	443	35.362	21.774	61,57
2019-10-28	443	35.514	20.383	57,39
2019-11-04	443	34.620	22.286	64,37
2019-11-11	443	34.990	22.089	63,13
2019-11-18	443	34.443	21.831	63,38
2017-09-11	8.080	20.948	2.097	10,01

Tabla 2.6: Tabla que muestra los rangos de tiempo en los cuales se ejecutaron escaneos de protocolos el mismo día, tanto en la fuente Censys como CLCERT.

Puerto	Prom. Semanas Esc.
443	3.86
80	2.95
8080	2.08
21	1.95
465	1.92
143	1.92
22	1.48
25	1.47
110	1,45
995	1,44

Tabla 2.7: Tabla que muestra cantidad de escaneos promedio por semana en protocolos de Censys más populares.

estas organizaciones mantienen formularios que permiten eliminar una IP de una lista negra, en el caso que haya sido agregada por error.

Estas listas negras suelen estar orientadas a servicios específicos, por lo que una máquina puede salir en una sin salir en otra. Esto explica en parte que las diferencias entre ambas fuentes sean de distinta envergadura para cada protocolo, según lo muestra la figura 2.3.

Tanto para evitar ser bloqueados como para evitar que actores maliciosos contesten de forma distinta, Censys no publica las IP que usan sus máquinas para realizar escaneos. Es más, también censuran este valor de los banners que recopilan en los casos en que el servidor del protocolo contesta con un mensaje que incluye la IP del cliente. Por lo tanto, no es posible verificar directamente si la o las IP usadas por Censys están bloqueadas por alguna lista negra. Sin embargo, sí se puede apreciar en varios banners recopilados que las IPs de Censys se encuentran actualmente bloqueadas por algunas de estas listas, ya que existen respuestas que mencionan que no se prestará el servicio solicitado al escáner por tratarse de una IP bloqueada.

En el caso del CLCERT, se usaron 2 IPs en el transcurso de los escaneos. La primera desde inicios de 2016 hasta el 25 de diciembre de 2018, y la segunda desde el 26 de diciembre de 2018 hasta la fecha. Ninguna de las dos IPs aparece marcada en algún filtro de SPAM a la fecha.

Este trabajo postula que la razón por la cual las IPs de CLCERT no se encuentran en lista negra es que los escaneos realizados por la organización se limitan a la red chilena, al contrario de los de Censys que son sobre toda la Internet. Considerando la cantidad de años que CLCERT lleva realizando escaneos y la poca cantidad de consultas sobre la práctica o solicitudes de exclusión de ella (no más de tres solicitudes de exclusión en tres años), se asume que en general, las instituciones que administran sistemas en la red chilena no están pendientes de los escaneos sobre sus máquinas. Al mismo tiempo, esto permite ver que las mismas listas negras conocidas no cuentan con máquinas de monitoreo en la red chilena, dado que en caso contrario las IPs de CLCERT debiesen encontrarse listadas en ellas.

Efecto en red de salida y ubicación geográfica

Otro factor influyente en los resultados puede ser el tipo de enlace con el que se cuenta en cada caso, así como también su ubicación geográfica y el Sistema Autónomo al cual pertenece.

En capítulos anteriores se ha discutido el impacto que puede tener el realizar escaneos desde distintas partes del mundo con respecto a los resultados que se pueden recibir. Algunos sistemas filtran por IP del cliente la respuesta que entregan, como por ejemplo, los servicios de streaming de video con el objetivo de limitar la distribución de sus contenidos de forma geográfica. En otros casos, se utilizan configuraciones de tipo *anycast* para mejorar el tiempo de respuesta de algún servicio en específico. También pueden afectar problemas de conectividad debido a la calidad del enlace o a la tolerancia del servicio frente a paquetes perdidos.

Los datos actuales con los que se cuenta de ambas fuentes no permiten establecer ninguna conclusión sobre este punto. Esto motiva una estrategia de verificación que será revisada en el Capítulo 4, la cual consiste en replicar estos escaneos desde ubicaciones remotas, a partir del uso de un servidor virtual extranjero.

2.4. Conclusiones

La revisión preliminar de los datos de escaneos manejados actualmente por el CLCERT permite una comprensión más realista de la magnitud y configuración básica de los servicios prestados en Internet. Sin embargo, se observa desde ya que la variabilidad de los resultados entre semanas dificulta un uso periódico y consistente de estos datos para la definición de grupos de equipos de interés. Al mismo tiempo, considerando los datos obtenidos localmente comparados con fuentes externas, se observa que existen diferencias de resultados entre ambos conjuntos para los mismos periodos de tiempo, por lo cual se hace necesario entender las razones que determinan estas diferencias y sus implicancias.

Los próximos capítulos experimentan con estas definiciones de grupos de interés crítico de dos formas distintas. En el capítulo 3 se revisarán máquinas críticas según su asociación con dominios del ccTLD chileno (`.cl`), así como también clasificando estos dominios en categorías de interés general. En contraparte, el capítulo 4 de este trabajo utiliza los mismos datos de escaneo históricos recopilados y busca entender de mejor forma sus diferencias temporales y entre fuentes, para así proponer subconjuntos de interés según distintos criterios.

Capítulo 3

Análisis de Concentración de Servicios Dependientes del ccTLD chileno

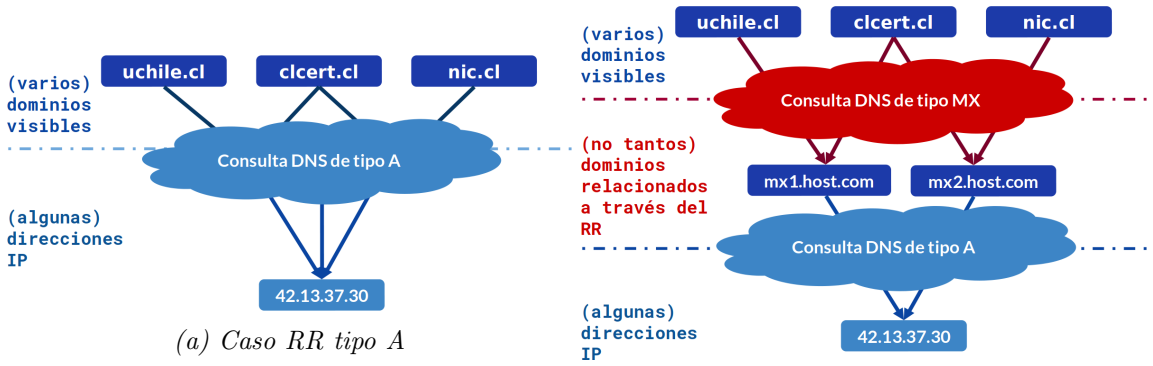
En este capítulo se presenta una nueva perspectiva para analizar y validar el estado de la red de un país a través del uso de los dominios *ccTLD* asignados al territorio. Al contrario de lo que pudiese parecer a primera vista, esto no restringe el concepto de red a solo *World Wide Web*, ya que los dominios también son utilizados en protocolos como *SMTP/IMAP*, *SSH* y *FTP*, entre otros.

Concretamente, el capítulo presente analiza un nuevo enfoque donde se evalúa la robustez y seguridad de esta definición de la *red chilena* (dominios chilenos) a partir de la distribución y concentración de estos servicios en distintas IPs, sistemas autónomos y países. El capítulo concluye con algunas observaciones obtenidas sobre estos datos y la utilidad de obtenerlos periódicamente.

3.1. Antecedentes del estudio

Desde la creación y masificación de la internet, los servicios de *HTTP*, correo electrónico y *DNS* han mantenido su importancia tanto por su popularidad como por su utilidad en las actividades diarias de las personas. Una componente común de los tres servicios recién mencionados es su dependencia fuerte en el uso de nombres de dominio, debido a razones técnicas (especificaciones que requieren obligatoriamente el uso de FQDNs para funcionar, como los registros MX en el sistema de correo electrónico y los *virtual hosts* en HTTP) y de usabilidad (es más fácil recordar las palabras que componen el FQDN que un número IP).

El rol crucial de los dominios en los servicios más usados de Internet motiva al estudio de las condiciones en que se presta este servicio en términos generales. Sin embargo, tal como muestra la figura 3.1, existe una o más capas de indirección entre los dominios y las direcciones de las máquinas que los ofrecen. Esta situación se observa, por ejemplo, en que es fácil pensar que dos páginas web en dos dominios distintos usan infraestructura distinta para proveer sus servicios. Sin embargo, es necesario realizar una resolución de la o las IP asociadas al dominio para verificar si existe o no alguna relación, por ejemplo, a través de su



(a) Caso RR tipo A

(b) Caso RR tipo MX y NS

Figura 3.1: Imagen representativa del mapeo de dominios a IPs, y de IPs a sistemas autónomos, tanto para RRs de tipo A como MX, ejemplificando que la indirección agregada por los dominios dificulta la visualización de concentración de servicios.

IP, subred, Sistema Autónomo o incluso país. La dificultad anterior se hace más notoria con registros DNS que apuntan a otros dominios, ya que es necesario realizar dos resoluciones consecutivas para tener una visión clara de relación y dependencia. En el caso del servicio NS, la necesidad de diversidad en infraestructura se ve reflejada en las recomendaciones realizadas por la *Internet Engineering Task Force*, a través de los RFCs 1034 [53] y 2182 [59]. Estos documentos recomiendan diversificar la cantidad de valores y ubicaciones geográficas de los sistemas que proveen este servicio.

En la práctica, la razón más importante para estudiar la concentración de dependencias entre los dominios de un país es la posibilidad de mejorar la comprensión sobre la infraestructura real que sustenta su funcionamiento. Tal comprensión permite entregar recomendaciones a los administradores de dominios, lo que los ayuda a minimizar los riesgos de funcionamiento erróneo, diversificando sus proveedores cuando sea posible. Claramente, es imposible elaborar recomendaciones sólidas sin entender el impacto que pudiese tener la baja de algún proveedor en el funcionamiento de los dominios del país, es decir, cuántos y qué servicios se detendrían en caso que el proveedor dejase de funcionar.

Dados los datos con los que se cuentan y los objetivos mencionados en el capítulo 1, en este capítulo se busca esclarecer el estado de concentración y diversidad de la internet chilena desde la perspectiva de servicios ofrecidos en el dominio `.cl`. Para ello, se realizó un escaneo completo de servicios presentes en todos los dominios `.cl` habilitados al 11 de enero de 2019. Las medidas de concentración se mostrarán revisando los valores de dirección IP, Sistemas autónomos y países, tanto para todos los dominios `.cl` como para un ejemplo de sub categoría de estos (dominios gubernamentales). Finalmente, a partir de los datos obtenidos, se entregarán algunas observaciones y recomendaciones, las cuales buscan disminuir los puntos únicos de falla en el espacio de dominios chileno.

3.1.1. NIC Chile

Un primer antecedente importante es el que explica la procedencia y manejo de los dominios `.cl`, y el por qué conviene hacer estudios desde esta perspectiva de red chilena. Para ello, es necesario explicar brevemente la misión y el rol de NIC Chile [14].

NIC Chile es un centro perteneciente a la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile y está encargado de la administración del Registro de Nombres de Dominio `.cl`. Este dominio de nivel superior identifica a Chile en la red Internet desde el año 1987.

A Enero de 2019, hay más de 450 mil dominios `.cl` registrados, y según datos del centro, en el año 2012 el 94 % de los dominios registrados en el país correspondían a sitios `.cl` [13]. Si bien la cantidad de gTLDs[1] existentes desde ese año ha aumentado considerablemente debido a la apertura de registro de estos a empresas privadas [7], a enero de 2019 no existe alternativa de dominio orientado exclusivamente a personas y empresas relacionadas geográficamente con Chile, por lo que se considera prudente asumir su liderazgo en uso en asuntos de interés nacional.

3.1.2. Servicios a estudiar

El estudio se enfoca en tres tipos de servicios relacionados con dominios `.cl`, cuyas IP son revisables a través de consultas DNS a los mismos:

- **Páginas web**, tanto en HTTP como HTTPS. En este caso se revisarán los RRs de tipo A.
- **Servicios de correo electrónico**, tanto entrante como saliente. En este caso se revisarán los RRs de tipo MX, además de los RRs de tipo A de los valores obtenidos en el paso anterior.
- **Servidores de nombre de dominio**, o servidores DNS. En este caso se revisarán los RRs de tipo MX, además de los RRs de tipo A de los valores obtenidos en el paso anterior.

Además de consultas DNS por RRs relacionados con los servicios a estudiar, se revisará que los puertos relacionados con el servicio de las direcciones IP que lo proveen se encuentren abiertos:

- En el caso de páginas web, se revisará que las direcciones IP tengan abierto al menos uno de los siguientes puertos: 80, 443.
- En el caso de los servicios de correo, se revisará que al menos uno de los siguientes puertos esté abierto en las direcciones IP: 25, 110, 143, 465, 587, 993 y 995.
- En el caso de servicios DNS, se revisará que el puerto 53 esté habilitado en las direcciones IP.

Como referencia, la figura 3.2 resume los valores a revisar para cada servicio analizado.

3.1.3. Datos a utilizar y recopilar

Todos los conjuntos de datos discutidos a continuación fueron recopilados entre los días 11 y 14 de enero del año 2019.

[1]Dominio de Primer Nivel de uso General o gTLD corresponde al conjunto de dominios no asociados a un territorio político, por ejemplo, `.com`, `.org` y `.net`

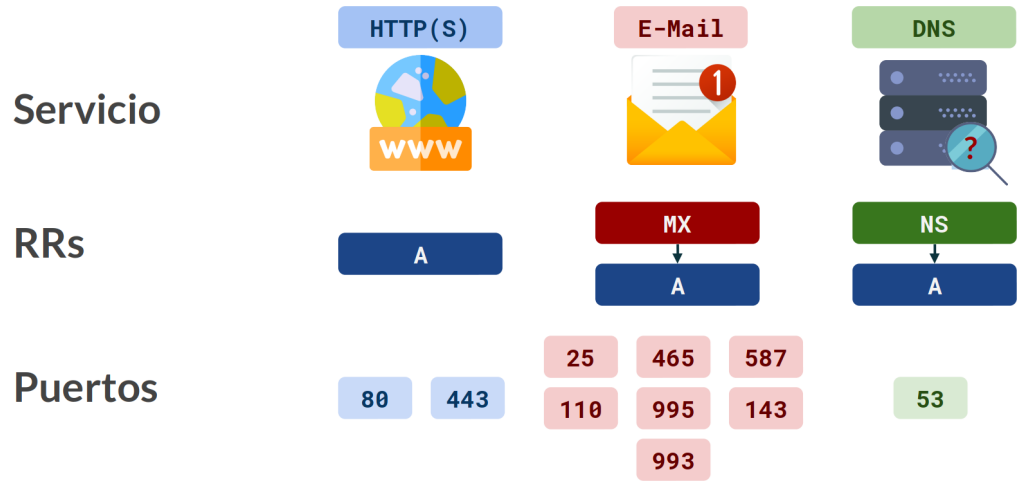


Figura 3.2: Resumen de los tres servicios a analizar, los RRs relacionados a revisar y los puertos abiertos a comprobar en cada caso.

Espacio de dominios chileno

El conjunto de datos más importante utilizado en esta investigación es una lista con 453.401 dominios del TLD chileno, obtenidos de parte de NIC Chile hasta la fecha ya mencionada.

Escaneos de RR de dominios .cl

Teniendo la lista de dominios completa a escanear, se extrajeron los RRs de tipo A, MX y NS de todos los dominios .CL analizados. En el caso de los registros A, se consideraron tanto los *Fully Qualified Domain Names* (FQDNs) de la lista ya mencionada, como los generados al concatenar la cadena de texto *"www."* a cada uno de los dominios de esta lista.

Datos geográficos y de sistemas autónomos

Para realizar una clasificación por Número de Sistema Autónomo (ASN) y país, se obtuvieron datos geográficos relacionados con los países asociados a distintas subredes, además de las AS encargadas de cada subred, a partir de la base de datos gratuita Geolite2 [39], actualizada semanalmente. Se optó por usar esta base de datos debido al carácter abierto de ella, permitiendo su uso y replicación en el futuro sin suponer un costo monetario extra para la realización del estudio, además de nivel de exactitud, el cual se considera el adecuado para este caso.

Lista de dominios gubernamentales

Se obtuvo una lista de dominios gubernamentales del directorio de organismos del Gobierno de Chile [12], con la intención de analizar por separado el comportamiento de estos dominios y compararlo con respecto al comportamiento general. Esta lista consiste en 588 dominios distintos.

3.2. Herramientas usadas y desarrolladas

Los datos de DNS mencionados en la sección anterior fueron obtenidos en tres fases, escaneo, importación y procesamiento. El detalle de las herramientas usadas para esto se da a continuación.

3.2.1. Escaneo usando ZMap y Mercury

Para obtener los datos de los escaneos de RR de dominios .cl, se utilizó la herramienta *Mercury* desarrollada por Eduardo Acha [1], utilizando el plugin de resolución DNS incluido con la herramienta.

Mercury fue configurado para consultar los registros A, MX y NS de la lista de dominios ya mencionada, a una velocidad de 1000 dominios por minuto. En el caso de los RRs de tipo MX y NS, se utilizaron solamente los dominios .cl obtenidos directamente de NIC Chile y de la página del Gobierno de Chile. En el caso de los RRs de tipo A, se consideraron como parte de un dominio tanto sus RRs de tipo A, como los RRs del dominio formado por la concatenación del prefijo "www." con el mismo dominio, debido a que el estudio de este registro se encuentra enfocado al uso de los dominios en el servicio de páginas web.

Además, en el caso de dominios con registro CNAME, se consideraron como propios los valores obtenidos a través del dominio canónico enlazado. Por ejemplo, si el dominio *a.cl* posee un registro de tipo CNAME hacia *b.cl*, todos los registros asociados a *b.cl* se consideraron como si hubiesen estado declarados en *a.cl*. En la práctica, los registros CNAME funcionan de esta forma.

Por último, considerando el enfoque de servicios tomado por este estudio, se configuró la herramienta ZMap [25, 2] para realizar un escaneo de los puertos habilitados para cada protocolo (mencionados en la sección anterior), determinándose si los registros escaneados por Mercury eran accesibles, de forma de usar solamente aquellos que lo fueran. Ambos escaneos se ejecutaron de forma secuencial utilizando la herramienta OSR, cuyo diseño fue explicado en el capítulo anterior.

3.2.2. Importación a OSR

Como se mencionó en la sección anterior, se desarrolló un módulo de importación para OSR, el cual está encargado de subir los datos generados por los escaneos de RRs y de puertos a la base de datos central del sistema. Este módulo se ejecuta inmediatamente después de terminado el proceso de escaneo, y demora 20 minutos en subir toda la información necesaria. Además, se encarga de validar los valores de los registros obtenidos, subiendo solamente aquellos que son válidos según las reglas de cada RR:

- En el caso de los RRs de tipo A, se consideraron solo las IPs bien formadas y no incluidas en rangos de IP reservados.
- En el caso de los RRs de tipo MX y NS, se consideraron solo los FQDN bien formados, con al menos dos niveles.

RR	N^o dominios c/RR	% c.r. dominios activos	% c.r. dominios escaneados
A	348.916	93,76	76,96
MX	315.431	84,77	69,57
NS	359.278	96,55	79,24
Activos	372.102	100,00	82,06

Tabla 3.1: Número de dominios válidos y accesibles con al menos un RR de cada tipo, comparando esta cantidad con la cantidad de dominios con al menos un RR de tipo A, MX o NS, y con la cantidad de dominios totales escaneados.

3.2.3. Procesamiento de datos

Luego de finalizar la importación de los datos, se ejecuta una batería de *scripts* que realizan algunas consultas agregadas sobre la información obtenida. Estas consultas se repiten tanto sobre el conjunto total de dominios, y aisladamente sobre los dominios de tipo gubernamental obtenidos del directorio web del Gobierno de Chile. Las consultas son las siguientes:

1. Lista de dominios con al menos un registro de tipo A, MX, NS.
2. Cantidad de dominios con cierta cantidad de registros, IPs, Sistemas Autónomos y países, en sus registros A, MX y NS.
3. Ranking de las IPs, Sistemas autónomos y países con más dominios asociados en A, MX, NS.
4. Cantidad de dominios con ASN cuyo nombre contiene el substring *amazon* o *aws*, *google*, *wix*, *digitalocean* y *cloudflare*.
5. Dominios con servidores compartidos entre servicio de correo y web, y entre servicio de correo y DNS.

3.3. Análisis de datos recopilados

Luego de escanear los 453.401 dominios, se obtuvieron 2.925.864 RRs distintos de 372.102 dominios válidos y accesibles, los cuales cuentan con al menos un RR de tipo A, MX, NS. Específicamente, 348.916 dominios poseen al menos un registro de uso web (A) válido y accesible (con IPs en rangos válidos y al menos un puerto HTTP/HTTPS abierto), 315.431 dominios poseen un registro de uso de correo electrónico (MX) válido y accesible (apuntando a otro FQDN que esté asociado a al menos una IP en rangos válidos, y con al menos un puerto de correo electrónico abierto) y 359.278 dominios poseen un registro de uso de servidor de nombre de dominio (NS) válido y accesible (apuntando a otro FQDN asociado a al menos una IP en rangos válidos, y con puerto del protocolo DNS abierto).

La tabla 3.1 muestra un resumen de las cantidades anteriormente mencionadas y las compara con el total de dominios escaneados y con el total de dominios con al menos un registro escaneado. La gran diferencia entre ambos valores porcentuales se puede deber a que una parte importante de los dominios escaneados no se encuentran configurados completamente o se encuentran asociados a servidores de dominio actualmente inexistentes, manteniéndose registrados o renovados continuamente.

El análisis de todos estos datos recopilados es realizado usando dos conjuntos de datos: *Análisis general de dominios chilenos* y *Análisis de FQDNs gubernamentales*, sobre las que se realizaron las siguientes observaciones:

1. **Distribución del número de RRs, IPs, sistemas autónomos y países asociados al conjunto estudiado**: El objetivo de este análisis es tener una primera impresión de la distribución en número de valores de RR, IP, ASN y países asociados a cada tipo de RR de los dominios estudiados. Lo anterior es motivado porque se plantea que a mayor cantidad de dominios con una alta cantidad de valores en cada una de las dimensiones anteriores existe un nivel mayor de "resiliencia" frente a situaciones de mal funcionamiento de sistemas en varias escalas. Por ejemplo, un dominio asociado a solo una IP en algún tipo de RR específico se encuentra más vulnerable a caídas que un dominio asociado a tres o cuatro IPs distintas, dado que si en ambos casos se dificulta la comunicación a una IP asociada a los dominios, el dominio con solo una IP asociada deja de funcionar, mientras que el que tiene cuatro IPs distintas asociadas y configuradas correctamente no debiese de tener problemas en seguir operando. El mismo razonamiento se puede aplicar en términos de cantidad de sistemas autónomos y países desde los que se entrega un servicio asociado a un dominio del TLD estudiado.

2. **Concentración del conjunto estudiado sobre IPs, sistemas autónomos y Países**: En esta sección se busca entender mejor qué recursos (IPs, sistemas autónomos, países) asociados a cada tipo de RR estudiado son los más importantes debido a la cantidad de dominios a las que están relacionados, además de entregar una medida numérica sobre esa importancia. Lo anterior es motivado por el supuesto de que a mayor concentración de dominios en estos recursos los efectos negativos del mal funcionamiento del recurso concentrado pueden tener un mayor alcance. Se revisarán también los 5 recursos más utilizados en cada servicio analizado, en el caso de sistemas autónomos y países.

3. **Uso de proveedores de servicios de infraestructura de Internet conocidos**: Se revisará a través de los nombres de sistemas autónomos asociadas a IPs escaneadas la cantidad de dominios relacionados proveedores de servicios de Infraestructura de internet conocidos, como por ejemplo *Google* (Correo electrónico), *Wix* (Hospedaje web), *Cloudflare* (DNS), *Amazon AWS* (Varios servicios) y *DigitalOcean* (Varios Servicios).

4. **Infraestructura compartida entre distintos servicios**: En este caso, se observará cuántos dominios comparten la infraestructura que usan para servir sus páginas web con la infraestructura de correo electrónico o DNS, con el objetivo de entender mejor el impacto que puede tener la caída de proveedores de un servicio en otros servicios.

Cada análisis termina mencionando algunas consideraciones y recomendaciones desprendidas de los resultados revisados.

3.3.1. Análisis sobre todos los dominios chilenos

El análisis general considera la revisión del total de dominios .CL escaneados, con resultados divididos en servicios web, correo electrónico y Servidores DNS.

Prioridad	Time to Live	Valor
1	3600	ASPMX.L.GOOGLE.COM.
5	3600	ALT1.ASPMX.L.GOOGLE.COM.
5	3600	ALT2.ASPMX.L.GOOGLE.COM.
10	3600	ALT3.ASPMX.L.GOOGLE.COM.
10	3600	ALT4.ASPMX.L.GOOGLE.COM.

Tabla 3.2: Valores MX recomendados en la actualidad por página de ayuda de GSuite.

Distribución General de Dominios Chilenos

La figura 3.3 muestra la variedad en cantidad de recursos asociados a dominios con registros de servicios web (tipo A, en azul), de correo electrónico (tipo MX, de color rojo) y de DNS (tipo NS, de color verde) tanto en número de RRs, IPs, sistemas autónomos y países asociados.

Al agrupar los dominios revisados por su cantidad de RRs (figura 3.3a), es posible notar una gran cantidad de dominios con solo un RR, tanto en los RRs de tipo A como MX, llegando a un 88,78 % y un 74,45 % respectivamente. Sin embargo, en el caso de los registros de tipo NS, un 98,39 % de los dominios con registros NS validos y accesibles está asociado a al menos 2 distintos de estos. Tan alto valor se puede deber a la influencia de las especificaciones del RFC 1034 [53], en las que se recomienda tener al menos dos valores de tipo NS distintos para cada dominio.

Dentro de los registros de tipo MX, se observa que un 15.72 % de los dominios revisados está asociado a 5 RRs distintos. Como explicación al fenómeno anterior, se plantea la posibilidad de que una gran cantidad de estos dominios hayan delegado la administración del servicio de correo a la plataforma *GSuite*. Este servicio recomienda configurar al menos 5 registros distintos de tipo MX, cuyos valores se muestran en la tabla 3.2 y fueron obtenidos de [30]. Revisando los datos obtenidos, se confirma la suposición al observar que 51.219 dominios revisados contienen como valor las cadenas `google` o `googlemail` (versión antigua del registro MX recomendado), esto equivale al 99,39 % de los dominios con 5 o más registros.

La figura 3.3b muestra la cantidad de dominios con una o más IPs distintas asociadas, agrupados por número de éstas. En el caso de registros de tipo A, este valor del gráfico es idéntico a la cantidad de valores de RR distintos debido a que un RR en este tipo es igual a una IP. Paralelamente, en el caso del registro MX se observa un porcentaje ligeramente mayor de dominios con 2 o más IPs distintas al compararlo con la cantidad de dominios con 2 o más RRs distintos. Además, se puede apreciar que existe una menor cantidad relativa de dominios con solo una IP asociada comparada con la cantidad de dominios con solo un valor de RR asociado, es decir, existe una pequeña cantidad de dominios que tiene solo un RR de tipo MX. También es posible notar en el mismo RR que el porcentaje de dominios con 3 IPs distintas se triplica al compararse con el porcentaje de dominios con 3 RRs distintos, pero que la cantidad de dominios con 5 o más IPs distintas es menor a los que tienen 5 o más RRs distintos.

Al observar el registro NS y compararlo con el gráfico 3.3a, se observa que el porcentaje de dominios con solo una IP aumenta casi 10 veces al compararse con el porcentaje de dominios

con solo un RR. Esto se explica debido a que nada evita que un dominio configure 2 valores de RR distintos de tipo NS que apuntan a la misma IP, lo cual muestra que la recomendación del RFC [53] no es globalmente seguida. A pesar de lo anterior, todavía el 83,50 % de los dominios estudiados está asociado a 2 o más IPs distintas.

El estado de cantidad de dominios asociados a distintas cantidades de Sistemas Autónomos se puede observar en la figura 3.3c. En el gráfico mencionado se observa que un 93,31 % y un 97,26 % de los dominios con registro A y MX respectivamente está asociado a solamente un Sistema Autónomo (diferenciado por su número de identificación). En el caso de los registros de tipo NS, y a pesar de las recomendaciones del RFC [59], solo un 28,65 % de los dominios está asociado a 2 o más Sistemas Autónomos.

La figura 3.3d muestra naturalmente una situación aún más concentrada, reflejando que el 98,57 % de los registros A y MX de los dominios estudiados poseen infraestructura en un solo país, mientras que un 22,58 % de los registros NS de estos dominios muestran un nivel de distribución geográfico mayor, usando dos o más países distintos para sus propósitos.

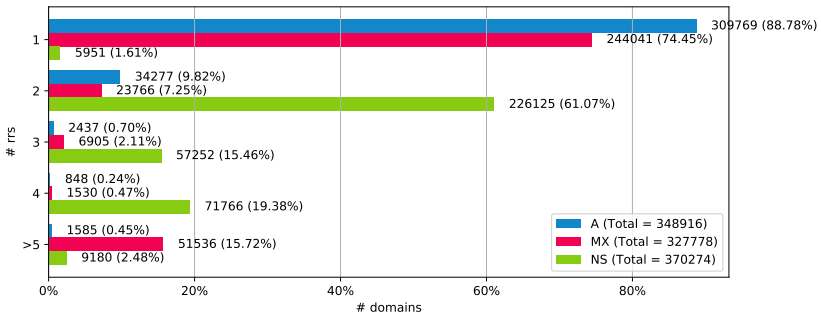
Concentración General de Dominios Chilenos

Si bien observar la cantidad de recursos distintos que los dominios tienen asociados sirve para obtener una imagen preliminar del nivel de robustez de la internet chilena, no entrega información acerca de la concentración de los dominios en conjuntos de recursos específicos, es decir, cuál es la infraestructura crítica que sostiene la red chilena. El análisis presentado en esta sección busca ponderar la importancia de cada grupo de recursos según la cantidad de dominios a los que están asociados, obteniéndose así una lista de recursos críticos en términos de cantidad de dominios relacionados.

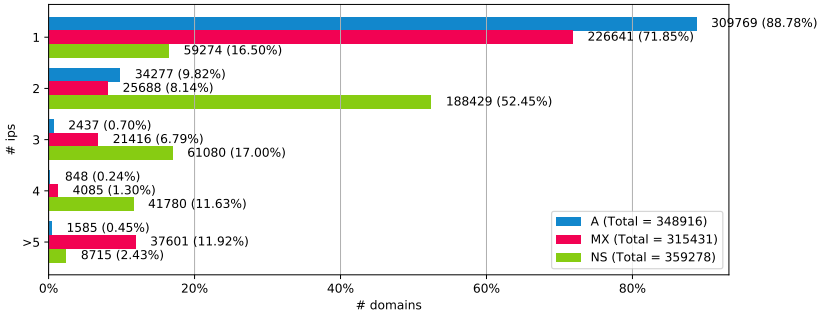
Para realizar este estudio, se consideraron todas las tuplas de valores distintos (RR, IP, ASN, país) existentes, y se contaron los dominios que estaban relacionados a ellas. Luego, se ordenaron de mayor a menor y se contó cuántos dominios dejarían de funcionar en caso de deshabilitar todos los recursos asociados a los valores de la tupla. En caso que alguna tupla todavía no revisada se bajase debido a que es un subconjunto del conjunto de valores ya revisados, la cantidad de dominios de ésta era sumada a la cantidad de dominios deshabilitados por el recurso en revisión. Esta estrategia avara ("*greedy*"), si bien no es óptima, maximiza la cantidad de dominios deshabilitados luego de revisar cada tupla, entregando una aproximación de importancia y concentración de los recursos.

Las figuras en 3.4 muestran gráficos de concentración en escala logarítmica para los recursos más populares de tipo RR, IP y ASN utilizando la metodología recién explicada. En el caso de la figura 3.4a, se observa un comportamiento similar de concentración de RRs tanto para registros de tipo A como NS, ya que en ambos se tiene que cerca del 80 % de los dominios que poseen estos servicios se encuentra sustentado sobre el 10 % de las tuplas escaneadas más importantes.

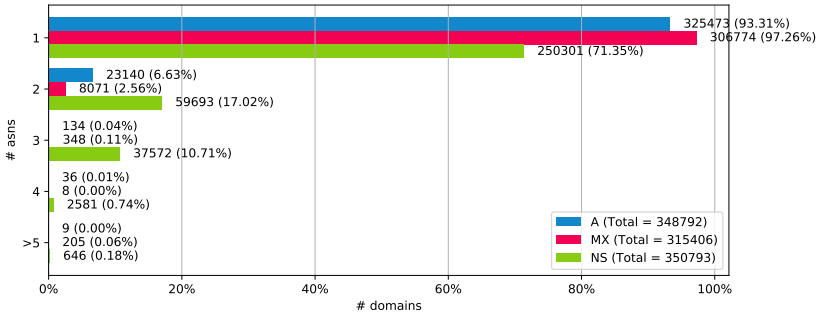
Específicamente, en el caso de los registros de tipo A, 82 valores de IP distintos, distribuidos en 79 tuplas y equivalentes a el 0,14 % de todos los valores observados para el registro, controlan el 20 % del universo de dominios relacionados a servicios web. Al mismo tiempo, en caso de los registros de tipo NS, el 20 % del universo de dominios con registros de tipo NS



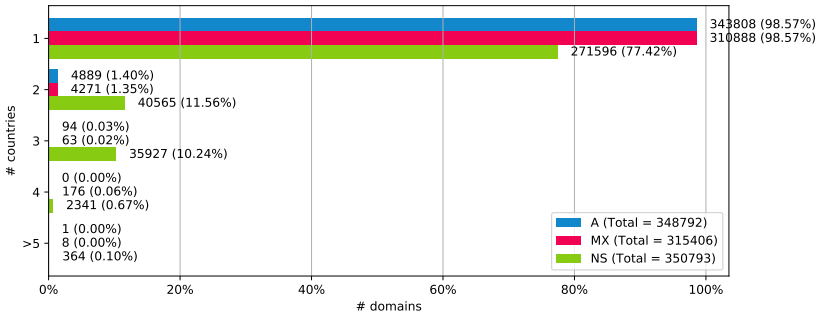
(a) Número de RRs distintos



(b) Número de IPs distintas



(c) Número de sistemas autónomos distintos



(d) Número de países distintos

Figura 3.3: Gráfico con la distribución de cantidad de RRs, IPs, sistemas autónomos y países distintos en el total de dominios escaneados con al menos un RR de tipo A, MX y NS.

es controlado por 28 valores de FQDN distintos distribuidos en 12 tuplas, lo que representa el 0,07 % del total de valores escaneados en ese registro.

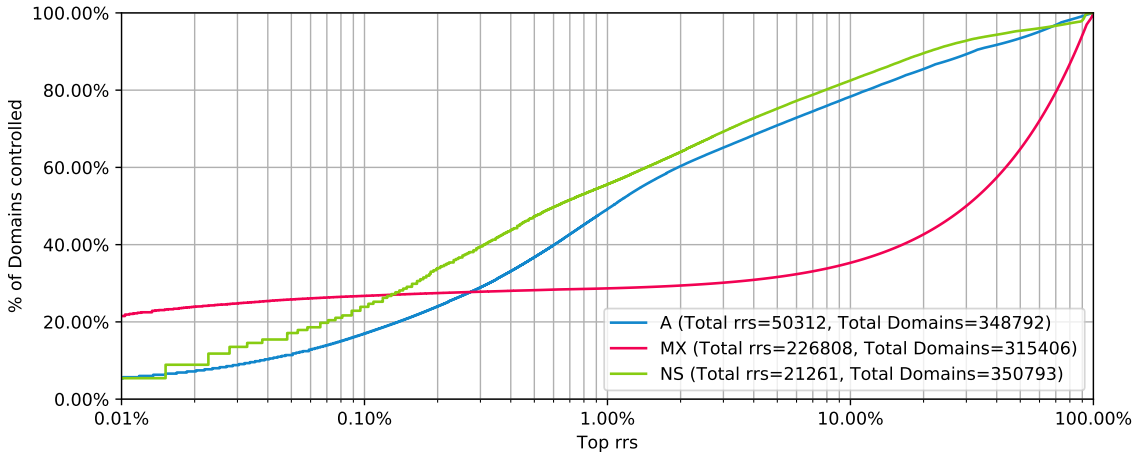
El caso de concentración de dominios relacionados a RRs de tipo MX es bastante distinto, ya que si bien 19 valores de FQDN distintos distribuidos en 9 tuplas controlan el 20,37 % de los dominios con este servicio, no más del 40 % de los dominios es controlado por el 10 % de valores más populares. Se plantea que esta diferencia se debe a que algunos servicios de hospedaje web que además incluyen servicio de manejo de correo electrónico configuran automáticamente registros de tipo MX del dominio usado de tal forma que apuntan al servidor de correo respectivo, por lo que aparecen con RRs distintos a pesar de usar el mismo servidor.

Al comparar los resultados de concentración de RRs ya mencionados con los de concentración de IPs que se observan en la figura 3.4b, se pueden apreciar diferencias bastante notorias entre concentración de valores de RR de tipo MX y concentración de IPs relacionadas al tipo MX. Estas diferencias confirman la suposición del párrafo anterior, es decir, se deben a que a pesar que los dominios poseen valores distintos como registro MX, estos valores apuntan a un conjunto acotado de IPs correspondiente a los servidores de correo que usa el servicio de hosting y que comparte con todos sus clientes. El comportamiento en los primeros valores sigue siendo parecido al anterior, ya que 14 IPs distribuidas en 9 tuplas y correspondientes al 0,07 % del total de IPs revisadas en este registro controlan el 20,03 % de los dominios escaneados. Pero a pesar de que la curva sea similar en los primeros valores, se diferencia muchísimo al ir avanzando hasta el 10 % del gráfico de valores de RR, pareciéndose mucho más a las curvas de los registros de tipo A o NS.

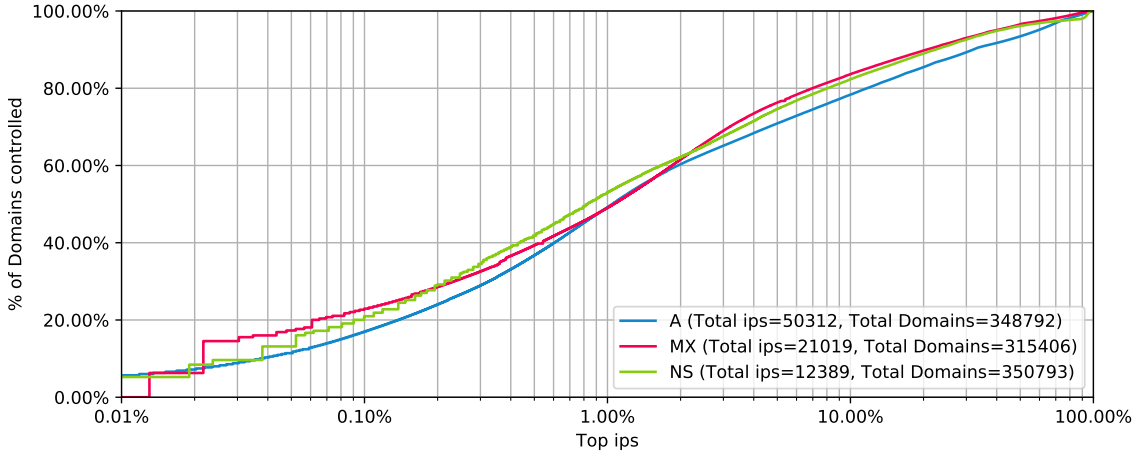
No está de más recordar que la curva asociada a los valores de registro de tipo A en los gráficos 3.4a y 3.4b son exactamente las mismas ya que los RRs del registro A son IPs, lo que explica que su comportamiento sea idéntico. Además, en el caso de la curva NS, a pesar de ser valores distintos, el comportamiento de ambas curvas relacionadas es muy similar en todo momento, ya que 19 dominios distribuidos en 10 tuplas y equivalentes al 0,09 % del universo de IPs encontradas controlan el 20,04 % de los dominios de este servicio.

Con respecto a los sistemas autónomos, el gráfico de concentración de sistemas autónomos visible en la figura 3.4c muestra una alta concentración visible en esta perspectiva en todos los registros por igual, ya que el 2 % de los sistemas autónomos más populares observados controla un número cercano al 80 % de los dominios en todos los casos. Este porcentaje llega a un valor cercano al 97 % de los dominios al revisar el 10 % de los sistemas autónomos ordenados por popularidad.

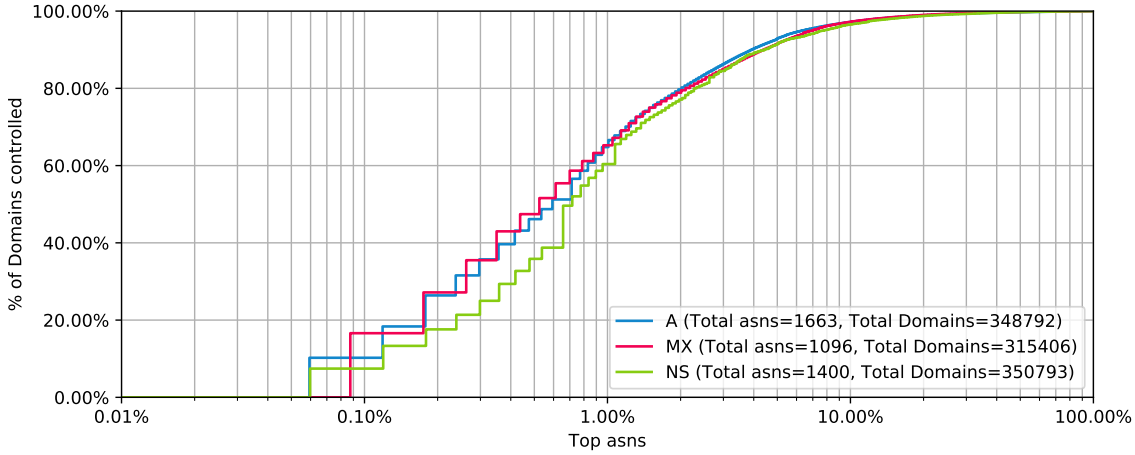
Al revisar en más detalle los 5 Sistemas Autónomos más populares para cada RR en el grupo de tablas 3.3, se puede observar que en los 3 casos, las 5 tuplas de sistemas autónomos más populares contienen solo un elemento, y controlan más de un cuarto de los servicios ofrecidos. En el caso de los RR de tipo A, visibles en la tabla 3.3a, se nota una gran predominancia de sistemas autónomos chilenos en estos 5 puestos. Incluso se observa en la posición 4 un Sistema Autónomo relacionado también con servicios de internet doméstico y para empresas. La situación para los servicios de tipo MX de la tabla 3.3b es aún más concentrada, con un 47,43 % de los dominios concentrados en solo 5 sistemas autónomos y con *Gmail* asociado exclusivamente a un 16,61 % de los dominios. Finalmente, en el caso de servicio de nombres de dominio visible en la tabla 3.3c, se observan casi los mismos nombres ya vistos en los



(a) Concentración de RRs



(b) Concentración de IPs



(c) Concentración de sistemas autónomos

Figura 3.4: Gráfico en escala logarítmica que muestra la cantidad de dominios totales que dejarían de funcionar (eje y) en el caso que cierta cantidad de IPs ordenadas según importancia (eje x) dejasen de funcionar.

N^o	**Tupla ASs**	**Tupla País**	**Dominios**	**%**	**% Acc.**
1	ZAM LTDA.	Chile	35721	10,24	10,24
2	SOC. COMERCIAL WIRENET CHILE LTDA.	Chile	28290	8,11	18,35
3	UNIFIEDLAYER-AS-1 - Unified Layer	Estados Unidos	28031	8,04	26,39
4	Gtd Internet S.A.	Chile	18071	5,18	31,57
5	HOSTING.	Chile	14481	4,15	35,72

(a) RR de tipo A

N^o	**Tupla ASs**	**Tupla País**	**Dominios**	**%**	**% Acc.**
1	GOOGLE - Google LLC	Estados Unidos	52377	16,61	16,61
2	ZAM LTDA.	Chile	33301	10,56	27,16
3	SOC. COMERCIAL WIRENET CHILE LTDA.	Chile	26291	8,34	35,50
4	UNIFIEDLAYER-AS-1 - Unified Layer	Estados Unidos	23578	7,48	42,98
5	Gtd Internet S.A.	Chile	14042	4,45	47,43

(b) RR de tipo MX

N^o	**Tupla ASs**	**Tupla País**	**Dominios**	**%**	**% Acc.**
1	SOC. COMERCIAL WIRENET CHILE LTDA.	Chile	26101	7,44	7,44
2	UNIFIEDLAYER-AS-1 - Unified Layer	Estados Unidos	20634	5,88	13,32
3	GOOGLE - Google LLC	Estados Unidos	14994	4,27	17,60
4	Gtd Internet S.A.	Chile	12714	3,62	24,99
5	HOSTING.	Chile	15334	4,37	29,36

(c) RR de tipo NS

Tabla 3.3: Ranking de 5 sistemas autónomos con más dominios asociados a través de valor del RR a alguna IP perteneciente a ellos.

primeros puestos, pero con una menor concentración comparada con la de los otros *Resource Records* y una mayor predominancia de servicios chilenos que en el caso del registro MX.

Para evaluar concentración de países en cada tipo de RR, se prefirió usar las tablas del grupo 3.4 para mostrar los resultados obtenidos. La tabla 3.4a muestra que la mayoría de los sitios web asociados a los dominios estudiados está hospedado en IPs chilenas, pero que Estados Unidos concentra un número bastante similar de dominios con servicio de página web. Además, los dos países ya mencionados concentran el 88,78 % de los dominios chilenos con al menos un RR de tipo A. En cambio, en las tablas 3.4b y 3.4c si bien la diferencia entre la cantidad de dominios asociados a Chile y Estados Unidos es pequeña, el segundo país se ubica en primer lugar en ambos casos.

Otro detalle rescatable es que, en el caso de la tabla 3.4c, es posible apreciar que los porcentajes de concentración son ligeramente más bajos que en las otras dos tablas del grupo. Esto puede deberse a que existe una mayor conciencia de la necesidad de replicación del servicio DNS en distintos puntos geográficos y topológicos de la red, alineándose parcial pero no completamente con las recomendaciones de los RFC mencionados al inicio del capítulo.

Uso General de Proveedores Conocidos

Como análisis complementario, se decidió revisar la cantidad de dominios asociados a proveedores de servicios web, de correo y de DNS conocidos (*Google, Amazon, Cloudflare, DigitalOcean, Wix*), de forma de conocer cuántos dominios están relacionados a ellos según las IPs que usan y las ASs responsables de sus direcciones.

La tabla 3.5 muestra un resumen de los resultados obtenidos al revisar todos los dominios. Es posible observar que *Google* tiene predominancia relativa entre los proveedores escaneados en los tres RR estudiados, siendo ésta mucho más pronunciada en el servicio de correo electrónico. Al mismo tiempo, se observa sin sorpresa que tanto *Cloudflare* como *Wix* no tienen presencia en servicios de correo electrónico.

Como conclusión de esta parte del estudio, al comparar todos los valores obtenidos relacionados con proveedores internacionales conocidos, salvo en el caso de correo electrónico y Google, no se ve una alta concentración de dominios basados en estos servicios. Por lo tanto, se entrega como recomendación de trabajo futuro la posibilidad de usar los datos obtenidos en el *Análisis de Concentración de Dominios Chilenos* realizado anteriormente, de forma de enfocarse en la revisión más profunda de los Sistemas Autónomos más repetidos.

Infraestructura compartida

La tabla 3.6 muestra el número total de dominios que comparten al menos una IP en los servicios de tipo MX y NS con la IP asociada al servicio de tipo A, además del porcentaje relativo a la cantidad de dominios con 1 o más RRs del tipo definido por la primera columna.

Lo primero que destaca en estos datos es que un 64,63 % de los dominios con al menos un registro A tiene también un registro MX que apunta a una o más de las IPs a las que apunta con A. Esto quiere decir que, al menos 6 de cada 10 dominios usan uno de los valores de IP registrados tanto en el servicio de correo electrónico como en el servicio web. Asimismo, un

N	Tupla País	Dominios	%	% Acc.
1	Chile	164826	47,26	47,26
2	Estados Unidos	144848	41,53	88,78
3	Canadá	12481	3,58	92,36
4	Alemania	4794	1,37	93,74
5	Brasil	5262	1,51	95,25

(a) RR de tipo A

N	Tupla País	Dominios	%	% Acc.
1	Estados Unidos	146712	46,52	46,52
2	Chile	141149	44,75	91,27
3	Canadá	11672	3,70	94,97
4	España	3237	1,03	95,99
5	Argentina	2831	0,90	96,89

(b) RR de tipo MX

N	Tupla País	Dominios	%	% Acc.
1	Estados Unidos	139941	39,89	39,89
2	Chile	137111	39,09	78,98
3	Estados Unidos, Canadá, Chile	40066	11,42	90,40
4	Argentina	4211	1,20	91,60
5	Alemania	5028	1,43	93,03

(c) RR de tipo NS

Tabla 3.4: Ranking de 5 países con más dominios asociados a través de valor del RR a alguna IP perteneciente a ellos.

	A	MX	NS
Amazon	12.993	2.176	13.350
Cloudflare	6.873	0	11.028
DigitalOcean	4.035	706	2.988
Google	18.687	56.326	17.862
Wix	16.019	0	1.854

Tabla 3.5: Número de dominios en cada RR asociados a al menos un AS cuyo nombre contiene alguna sub cadena de texto relacionada con un proveedor conocido.

RR	Comparte IP con A	Comparte Valor Dominio
MX	203.861 (64,63 %)	227.984 (72,28 %)
NS	71.326 (19,85 %)	12.267 (3,41 %)

Tabla 3.6: Número de dominios en cada RR que comparten IP o valor de dominio entre A y MX/NS. El porcentaje es relativo a la cantidad de dominios con 1 o más RRs del tipo indicado por la primera columna.

72,27 % de los dominios con RR de tipo A tiene al menos un RR de tipo MX cuyo valor es igual o es un subdominio del mismo dominio.

En el caso del registro NS, estas cifras varían notoriamente, ya que un 19,08 % de los dominios comparte infraestructura DNS con infraestructura web desde el punto de vista de las IPs, mientras que un 3,28 % de los dominios apunta sus registros NS a valores que son subdominio de los dominios estudiados. Esto quiere decir que la mayoría de los servicios de nombre de dominio de los dominios chilenos apuntan a FQDNs de otras zonas (probablemente, de zonas de los mismos proveedores del servicio de hosting).

Consideraciones Generales

El análisis completo de los dominios .cl nos muestra características importantes de concentración y resiliencia de los servicios asociados a ellos, pero se enfrenta a algunas limitaciones derivadas de considerar a todos los dominios del espacio como iguales. Como revela la existencia de rankings de visita de dominios, no todos los dominios son visitados en la misma cantidad y frecuencia, por lo que una alternativa para ponderar la importancia de cada dominio se puede relacionar con su ubicación en algún intervalo de algún ranking de dominios. Otra medida de importancia propuesta es la realización de una categorización de los dominios en grupos según su rubro o utilidad, ya que por ejemplo, un dominio de una página de tienda electrónica no debiese tener la misma importancia que uno usado para un trámite importante de gobierno.

Con el objetivo de probar el último enfoque mencionado, se realizó un estudio similar al realizado con el grupo general de los dominios, pero usando un subconjunto de dominios clasificados como de tipo gubernamental. Más adelante se discutirá sobre la posibilidad de extender este estudio a un número mayor de categorías, entregando un método público y transparente para determinarlas. También se planteará al final de este capítulo la posibilidad de usar uno o más rankings de dominios por cada servicio analizado, mencionándose las dificultades y desafíos que este enfoque pudiese traer según trabajos relacionados.

3.3.2. Análisis de FQDNs Gubernamentales

El objetivo de este análisis es dar un ejemplo de aplicación de los estudios vistos en el análisis anterior a un subconjunto de dominios definido. Como ya se mencionó, la lista categorizada de dominios se obtuvo del directorio en línea de instituciones del Gobierno de Chile [12], realizando *scraping* de los enlaces presentes a la fecha de Enero de 2019. Debido a que una parte importante de los FQDN gubernamentales obtenidos corresponden a subdominios de los dominios *gob.cl* y *gov.cl*, y estos no están presentes en la lista de

dominios de *nic.cl* facilitada para el análisis general, los resultados de este análisis no deben ser considerados como un subconjunto de los resultados anteriores.

Dentro de los dominios gubernamentales, se consideran sitios municipales, regionales, de subsecretarías y ministeriales, además de páginas relacionadas con empresas y organizaciones de tipo estatal, como *TVN, Codelco* u hospitales públicos. En total se cuenta con 588 distintos FQDNs en esta categoría.

Tanto el análisis como los resultados de este subconjunto de dominios se pueden encontrar en el Anexo A del presente documento.

3.4. Consideraciones

Para finalizar este capítulo, se comentarán algunas consideraciones relacionadas con los resultados obtenidos, las limitaciones apreciadas en el trabajo realizado y se darán ideas para extender el trabajo presentado en el futuro.

3.4.1. Recomendaciones

Los datos obtenidos y presentados en este capítulo sirven como motivación para la proposición de las siguientes medidas que buscan prevenir posibles problemas de disponibilidad de los servicios prestados por máquinas asociadas a dominios chilenos.

Detección de altas concentraciones de servicios y dominios

Los datos en bruto usados para generar los gráficos de disponibilidad y concentración permiten enviar avisos personalizados a los administradores de dominios que no cumplen con ciertos criterios de disponibilidad básicos determinados por NIC o recomendados por algún RFC. Por ejemplo, se propone el envío de un aviso recomendando a todos los dominios con solo una IP asociada como registro NS que activen el DNS secundario de NIC Chile o que soliciten a su proveedor de DNS una mayor redundancia en la prestación del servicio.

De la misma forma, es posible enfocarse en monitorear la disponibilidad, latencia y seguridad de las máquinas con mayor cantidad de dominios asociados. Otra idea interesante es utilizar estos datos para avisar a futuros clientes que quieran usar una IP popular que esa misma IP ya está siendo usada por una gran cantidad de dominios, lo que puede ser síntoma de problemas de disponibilidad por sobrecarga.

Mayor difusión de DNS Secundario NIC Chile

Desde hace varios años, NIC Chile ofrece a sus clientes un servicio de DNS secundario, en el cual actúan replicando los valores de su servidor DNS primario asociado, en caso que este no esté disponible. Esto permite aumentar la cantidad de IPs y sistemas autónomos que proveen el servicio de nombre de dominio en esa zona. Los datos obtenidos revelan que el servidor DNS secundario de NIC Chile no es tan popular como pudiese ser. Una propuesta a partir de estos datos es incentivar a los dominios a optar por el uso de este servidor, para así asegurar una mejor disponibilidad en la resolución de estos nombres de dominio.

Infraestructura de Internet para Gobierno

El estudio sobre los dominios públicos de gobierno obtenidos entrega una imagen en gran escala de la infraestructura usada en distintos organismos de gobierno. Considerando que muchas veces estos organismos guardan información sensible, es importante tener en cuenta cómo y dónde esta información terminaría guardada, revisando integralmente la privacidad, seguridad y redundancia de los proveedores de estos servicios.

Como caso concreto revisado en la actualidad, el hecho que casi el 30 % de los dominios utilice servicios de correo electrónico de *Google* se puede ver tanto como una ventaja (*Google* es reconocida como una empresa con tecnología de punta en términos de seguridad de la información), como una desventaja (*Google* es una empresa que opera en territorio estadounidense con críticas recurrentes sobre el manejo de datos para fines de marketing, por lo que es complicado hacerla responder por cualquier problema que pudiese ocurrir a partir del mal uso o filtración de estos datos).

Discusiones similares se pueden dar al considerar proveedores más pequeños pero de origen nacional, dado que si bien por un lado no hay garantías de que los datos se almacenen de forma correcta ni existe legislación que lo fomente, el hecho que las empresas operen en territorio nacional facilita la creación de normativas que exijan a esta organización el tratamiento adecuado de los datos.

Concretamente, este estudio recomienda continuar con la tendencia de administración propia de recursos web de parte del mismo gobierno, a través del Ministerio del Interior. De esta forma, la responsabilidad de manejo correcto de los datos no dependería tanto de factores externos, por lo que contando con equipos de personas capacitados y recursos suficientes para asegurar servicios confiables, seguros y privados, esta alternativa se ve como la menos riesgosa.

3.4.2. Limitaciones

A continuación se mencionarán algunas limitaciones de los estudios realizados en los dos grupos de dominios, y se darán ideas de cómo superarlas en futuros desarrollos y escaneos.

Dominios sin registros

Lo primero que llama la atención al comenzar la revisión de los datos escaneados es que un número no menor de dominios no posee valor alguno de RR. La información recopilada en este trabajo no da ninguna razón certera de por qué ocurre esto, pero se cree que puede deberse a que estos dominios son comprados con el objetivo de reventa o solamente para mantenerlos ocupados en caso de requerirse en el futuro, lo cual es una situación recurrente en el caso de marcas comerciales. Otra razón barajada, y reforzada a partir de los datos, es que es posible que parte de la infraestructura DNS asociada al dominio se encuentre abajo debido a encontrarse asociada con un plan de hosting vencido, el cual solo se podía adquirir por separado del nombre de dominio hasta hace pocos años. Los datos obtenidos al revisar concentración de servicios NS y MX parecen respaldar esta postura.

Anycast y Balanceadores de Carga

Como se mencionó brevemente en la introducción, *anycast* corresponde a una técnica para aumentar la disponibilidad y resiliencia de algunos servicios sin estado, ya que permite contar con varias máquinas sincronizadas, distribuidas geográficamente en el mundo y asociadas a la misma IP. Lo anterior provoca que, en el caso de DNS, el servidor activo que esté más cerca del cliente contestará la consulta realizada.

El uso de *anycast* no se limita solo al servicio DNS, ya que se ha reportado que algunas *Content Delivery Networks* (CDNs) como Cloudflare lo usan para acelerar la entrega de contenidos multimedia y evitar ataques de denegación de servicio. [17]

Al mismo tiempo, existe la posibilidad de que algunos servicios asociados a una sola IP estén siendo entregados por un gran número de máquinas asociadas a un balanceador de carga que se encarga de distribuir las consultas entre ellas, mejorando de esta forma la disponibilidad del servicio.

Teniendo lo anterior en cuenta, es necesario mencionar que la técnica de medición activa propuesta en este capítulo no es capaz de notar los casos anteriores. En el caso de estrategias para detectar el uso de *anycast*, es posible realizar escaneos desde distintas partes del mundo. En el caso de balanceadores de carga, se plantea que se mantiene todavía el punto de falla de conectividad a internet, incluso contando con uno de estos, por lo que los resultados siguen siendo relevantes a pesar de la existencia de esta configuración.

3.4.3. Trabajo Futuro

A continuación, se presentan algunas ideas que permiten extender este trabajo para obtener mejores recomendaciones.

Replicación de este estudio sobre otros conjuntos de dominios

Si bien este estudio se realizó sobre los dominios chilenos, nada impide su realización sobre otros conjuntos de dominios tanto chilenos como de otras partes del mundo. Por lo tanto, se propone la idea de la repetición de estos experimentos sobre otros dominios. Esto amerita contar con una lista de dominios completa para cada zona que se quisiese revisar. Estas listas no tienen que ser necesariamente TLDs, y pueden ser grupos o categorías de dominios importantes para grupos humanos específicos.

Otra ventaja del uso de categorías bien definidas es que es posible ponderar la importancia de los resultados para cada una de las categorías revisadas. Como se mencionó anteriormente, no es lo mismo la carga en infraestructura de una página correspondiente a un servicio de infraestructura crítica de un país con la de una tienda local de café. Por lo tanto, las caracterizaciones tienen como objetivo la obtención de un mejor entendimiento del impacto real que puede tener la caída de un conjunto de máquinas en específico.

Una última idea en este grupo es considerar el uso de rankings de posición específicos para distintos servicios, con el objetivo de ponderar la importancia del resguardo de disponibilidad de ciertos servicios según su popularidad. Es importante recalcar nuevamente que el uso de

estos rankings limita el alcance del estudio solamente a servicios incluidos por estos mismos rankings.

Escaneos Regulares y múltiples

La facilidad de ejecución de este estudio mediante las herramientas desarrolladas motiva a programar la realización de estos escaneos de forma periódica, de existir una forma de contar con una lista de dominios actualizada. NIC.cl publica periódicamente una lista de dominios registrados y eliminados en su página web, accesibles por cualquier persona. Para recopilar estos dominios y como ya se comentó en el capítulo anterior, se agregó una tarea al OSR que todos los días se encarga de descargar estas listas, registrar los dominios nuevos y marcar como inexistentes los eliminados.

Otra razón por la que se recomienda realizar múltiples escaneos de forma regular es que existen situaciones en las que los resultados entre distintas consultas DNS varían de forma intencional, con el objetivo de balancear la carga entre los distintos valores de RR. Al mismo tiempo y como ya se mencionó en limitaciones, hay que tener en cuenta que existen servidores DNS que entregan respuestas distintas según la ubicación geográfica de la que se consulte el recurso, de forma de, por ejemplo, entregar IPs más cercanas al cliente que realiza la consulta.

Ponderar servidores según calidad o confiabilidad

Así como se propone la ponderación de importancia de los dominios según categoría y ranking, también es necesario ponderar a los proveedores según calidad de servicio o confiabilidad al momento de almacenar los datos entregados.

La idea es entender calidad de servicio según factores como disponibilidad o *uptime*, latencia y métricas de seguridad según versiones de sistema operativo o configuraciones de *firewall*. En este caso, una concentración alta en un servicio con alta calidad no es tan preocupante como una concentración alta en un servicio de muy baja calidad.

Por otro lado, se quiere entender confiabilidad de servicio como políticas de protección de datos personales y de la información hospedada o manejada por los proveedores, así como también los mecanismos de seguridad que existen para acceder a las cuentas administrativas o realizar cambios a través de tickets de soporte. Otra dimensión importante en esta categoría es el nivel de confianza en la empresa que presta el servicio, el cual se determina tanto por su reputación, su procedencia y su historia en casos similares.

3.4.4. Conclusiones

Este capítulo dio a conocer una nueva estrategia para la evaluación de la resiliencia y concentración de los servicios asociados a un tipo de red chilena: los dominios .cl, además de análisis y observaciones a partir de una ejecución de ella. A partir de los resultados obtenidos, y considerando la posibilidad de difusión de recomendaciones para mejorar la disponibilidad de los servicios ofrecidos a las partes interesadas, se considera que esta estrategia tiene el potencial de mejorar el estado de la red escaneada y robustecer en general los servicios orientados a habitantes de nuestro país, fomentando así el desarrollo tecnológico local y

reduciendo la dependencia de servicios administrados por externos en los sistemas críticos para nuestro país.

Capítulo 4

Estrategias de análisis de datos de Escaneo usando múltiples fuentes

Este capítulo presenta y discute tres conjuntos de técnicas y estrategias diseñados para la clasificación de resultados a partir de datos de monitoreo de red, las cuales están agrupadas sobre el concepto de uso de múltiples fuentes de datos de escaneo. El concepto de múltiples fuentes usado es amplio, ya que considera la organización que realiza los escaneos, el tipo de estos datos y las fechas en las que se recopilan.

Para poner a prueba estas estrategias, se usarán los datos históricos obtenidos de escaneos de la red chilena durante los últimos tres años del CLCERT, presentados en el capítulo 2 de este trabajo. Al mismo tiempo, también se usarán de forma exclusiva los datos de escaneo de Censys, presentados y comparados en el capítulo 2 de este trabajo.

El capítulo se organiza en cinco secciones. En la primera sección, se presenta la motivación para realizar este trabajo y se define concretamente el concepto de "red chilena" a utilizar y se describen brevemente las ideas de las estrategias a proponer. Posteriormente, las secciones dos, tres y cuatro corresponden al detalle y la ejecución de estas estrategias. Finalmente, la última sección discute sobre la utilidad de cada una de las estrategias presentadas, resumiendo los problemas encontrados y entregando posibles soluciones y mejoras.

4.1. Antecedentes

Como antecedentes para el desarrollo de este capítulo, se describirán tanto la motivación para desarrollar estrategias de análisis sobre estos conjuntos de datos, como las estrategias propuestas, las cuales se ahondarán en capítulos posteriores. Finalmente, se aprovechará de aclarar y justificar explícitamente la definición de "red chilena" a usar en estas estrategias.

4.1.1. Motivación

Las organizaciones y los grupos de investigación relacionados con escaneos de puertos y protocolos del capítulo 1 presentan sus resultados periódicos como una "foto" que representa

el estado de la Internet en el momento en que realizan el proceso ya mencionado. Sin embargo, tanto en el mismo trabajo relacionado como en consideraciones de capítulos previos de este trabajo se observa que la correcta enumeración de dispositivos en Internet depende de una gran cantidad de factores, como por ejemplo, los dispositivos a enumerar y la infraestructura de *hardware* y *software* utilizada para el trabajo. Esto puede provocar diferencias en los resultados obtenidos y visibles con respecto a la "situación real" en ese momento.

Cabe destacar que, si bien el objetivo de los escaneos de red para uso en procesos de seguridad informática no requiere tener una imagen exacta del estado de la subred de Internet estudiada, una mayor claridad frente a estos resultados entrega la oportunidad para una mejor clasificación y jerarquía de las máquinas revisadas, disminuyendo las posibilidades de que se escape una gran cantidad de casos del conjunto de datos revisado.

Lo anterior motiva el desarrollo de estrategias que permitan mejorar los resultados obtenidos a partir de escaneos, a partir de la utilización de información de múltiples fuentes. En este caso, el concepto "fuente" no significa solamente organización de origen de los datos, sino que también aplica a tipo de escaneo utilizado y periodo temporal en el cual fue obtenido. Este trabajo de investigación plantea que a medida se usan más fuertes, se puede detectar un mayor conjunto de máquinas potencialmente importantes para algun tipo de escaneo, las cuales posteriormente se pueden filtrar o considerar según cantidad o calidad de detecciones.

Al mismo tiempo, este trabajo busca entender mejor la correspondencia de los resultados de escaneos realizados en redes externas con escaneos realizados internamente, bajo el supuesto de resultados distintos en ambos casos, el cual proviene de la información preliminar derivada de la comparación realizada en el Capítulo 2. Además, se quiere entender mejor la relación entre escaneos de distintas fechas, y cómo los resultados obtenidos van variando en valores de IP detectadas.

Se plantea que la profundización en las áreas ya mencionadas puede permitir un mejor dominio de los datos actualmente recopilados, así como también una mayor claridad en el tipo de escaneos que se deben desarrollar y ejecutar a futuro.

4.1.2. Red Chilena en este contexto

Considerando los supuestos geográficos tomados por cada una de las fuentes anteriores, los cuales fueron expuestos en el Capítulo 2 de este documento, se decide tomar como referencia de red chilena para este capítulo la dictada por la base de datos GeoLite2 de MaxMind, ya que es usada tanto por los escaneos de CLCERT como los de Censys.

Además, debido a que las fuentes externas no entregan información acerca de la metodología utilizada para determinar el país de los resultados, los datos de estas fuentes serán usados solamente complemento de otros datos de otras fuentes, sobre las cuales sí se tiene certeza que sus resultados se encuentran enmarcados en la "red chilena" elegida.

4.1.3. Estrategias Propuestas

Como ya se ha mencionado, el punto en común de las estrategias propuestas en este capítulo es la utilización de "múltiples fuentes de datos" para su ejecución. A continuación,

se explican brevemente sus ideas principales y objetivos.

Comparación histórica de datos de escaneo de protocolos

(Usando datos del mismo proveedor, pero de distintos periodos temporales)

Como se mencionó en la motivación, la mayoría de los estudios de escaneos de internet consideran cada instancia de escaneo como una "foto" del estado de la internet en el momento en que el escaneo se realiza. Sin embargo, una "foto" no es tan representativa del estado general de una red, dado que en ella pueden tanto aparecer equipos sirviendo un protocolo de forma temporal, como no aparecer equipos que corren un servicio de forma recurrente. El objetivo de esta métrica es tener en consideración el uso en conjunto de información histórica de escaneos de protocolos del mismo tipo y fuente, de forma de entender mejor la evolución temporal de los grupos de máquinas que participan constantemente de la red chilena y poder definir subconjuntos críticos de máquinas a escanear según la periodicidad de su aparición en los resultados de escaneos previos., con el objetivo de que la "foto" revisada tenga una validez mayor en términos de extensión temporal

Comparación de datos de escaneo de distintos proveedores

(Usando datos de distintos proveedores, pero de similares periodos temporales)

La comparación realizada en el Capítulo 2 de este documento permite observar que, en algunos casos, los resultados obtenidos por fuentes distintas pueden diferir bastante. Como ya se ha tratado en otros capítulos, esto se puede dar por diversos motivos, como el uso de Anycast, el bloqueo de IP de parte de los servidores escaneados o el tipo de software usado para realizar los escaneos. El objetivo de esta estrategia es la revisión en detalle de los motivos que pueden llevar a las diferencias de resultados de estas fuentes, así como también revisar en detalle el factor geográfico de estas diferencias, a partir de la repetición de escaneos con el mismo software desde servidores en distintas ubicaciones pero con configuración idéntica.

Detección de máquinas vulnerables a través del uso de datos de abandono

(Usando datos de tipos y proveedores distintos)

Trabajos previos ya mencionados, como el de Acha y el de O'Hare, intentan obtener métricas de vulnerabilidad a partir de datos como versiones de software o configuración de parámetros de los protocolos estudiados. El trabajo actual se enfocará en la proposición de una métrica distinta, denominada como *abandono*, la cual consiste en considerar factores que permiten estimar hace cuánto tiempo no se actualiza una máquina. Este factor se usará en conjunto con datos de máquinas vulnerables de fuentes externas para asignar un grado de vulnerabilidad a las máquinas encontradas en escaneos. Las fuentes consideradas son el uso de versiones de software obsoletas, certificados SSL/TLS vencidos y cantidad de reportes de malware obtenidos desde múltiples fuentes. El objetivo de esta estrategia es verificar si el factor de abandono tiene alguna influencia en los resultados de los reportes de malware entregados por fuentes externas.

4.2. Comparación histórica de datos de escaneo de protocolos

Esta sección detalla el trabajo de comparación histórica de los datos obtenidos, la cual considera los resultados de escaneos de protocolos de todo el periodo de funcionamiento de cada fuente revisada. En concreto, para entregar un conjunto de IPs "importantes" en el tiempo, se revisarán y compararán los datos históricos obtenidos por cada fuente por separado y conjuntamente, de forma de determinar subconjuntos de IPs más repetidos y el recambio que ocurre de estos datos en distintas escalas temporales.

Para cada fuente (CLCERT, Censys), se realizan dos revisiones distintas. La primera consiste en considerar cuántas veces se repiten las IP escaneadas en el conjunto total de escaneos manejado para cada puerto revisado, mientras que la segunda consiste en comparar los conjuntos de IPs de cada semana con el mismo conjunto en un periodo previo, de forma de entender mejor el recambio de estas IP y clasificarlas según antiguedad.

4.2.1. Datos históricos del CLCERT

Durante los más de 3 años de operación, los escaneos de protocolos del CLCERT han detectado 4.213.043 IPs distintas sirviendo adecuadamente algún servicio relacionado con los puertos escaneados. La tabla 2.2 muestra en su última columna la cantidad de IPs distintas encontradas por puerto escaneado. Se puede notar en esta tabla que protocolos como SSH, HTTP y HTTPS en sus puertos convencionales han estado siendo servidos por más de un millón de IPs distintas, lo que consiste en aproximadamente el 10 % del universo de IPs chilenas según Geolite2. Asímismo, se observa una cantidad baja de IPs únicas relacionadas con servidores de correo electrónico (aproximadamente 60 mil para protocolos IMAP y POP3, y 202 mil para protocolo SMTP entre servidores).

Análisis global de IPs encontradas

La figura 4.1 permite apreciar con qué frecuencia aparecieron las IP del universo completo de IPs encontradas para cada protocolo revisado. En cada gráfico, se pueden destacar 2 detalles importantes. El primero es el número en el que parte cada curva (con Número de veces escaneado igual a 1), el cual indica cuántas IP son vistas solamente una vez durante los 3 años de escaneos. El segundo detalle corresponde a la concavidad de la función, ya que un acercamiento más rápido al máximo número de IPs del protocolo significa que una cantidad menor de IPs se ven recurrentemente en una gran cantidad de escaneos, aumentando así la variedad de resultados entre cada sesión de escaneo.

En concreto, la figura 4.1a muestra que aproximadamente 100 mil IPs detectadas en algún momento utilizando el protocolo FTP fueron vistas solamente una vez ejecutando servicios relacionados con este protocolo. Situación proporcionalmente similar se puede observar en los casos del protocolo SSH, visible en la figura 4.1b, con un poco menos de 600 mil IPs vistas una única vez, del protocolo 80 (Figura 4.1d), 443 (Figura 4.1g) y 8080 (Figura 4.1h). En contraste, la rapidez con la que cada función llega al número máximo de IPs determina que son muy pocas las que se mantienen constantemente entre escaneos.

En el caso de los protocolos relacionados con el envío y recepción de correo electrónico (asociados al puerto 25 en la figura 4.1c, 110 en la figura 4.1e y 143 en la figura 4.1f), se observa que la cantidad de IPs encontradas una sola vez es considerablemente más pequeña. Sin embargo, al llegar de forma más lenta al número máximo de IPs por protocolo, se observa que una proporción mayor de éstas se mantiene durante semanas.

Otro detalle importante percibible en estos gráficos surge al comparar la cantidad de escaneos realizados por protocolo según la tabla 2.2 con el número máximo de veces que se encuentra una IP (límite derecho del eje X) de cada gráfico. En el caso del puerto 21, se observa que las IPs más veces encontradas fueron vistas no más de 100 veces, mientras que se realizaron un total de 178 escaneos en el periodo revisado. En el caso de todos los otros puertos, las IPs más revisadas fueron encontradas entre 130 y 160 veces.

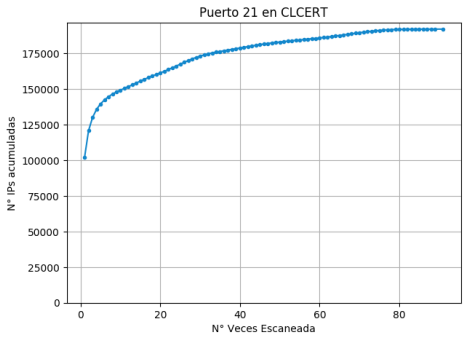
Una forma de explicar la diferencia entre los resultados de los protocolos de correo y los otros protocolos (tanto en número de IPs como variabilidad de estas) es que los servicios de correo suelen ser hospedados por empresas dedicadas a ello en pocas máquinas, mientras que servicios como SSH, FTP y Web suelen ser hospedados en máquinas comunes con IPs dedicadas para ello.

Continuidad de las IPs por protocolo

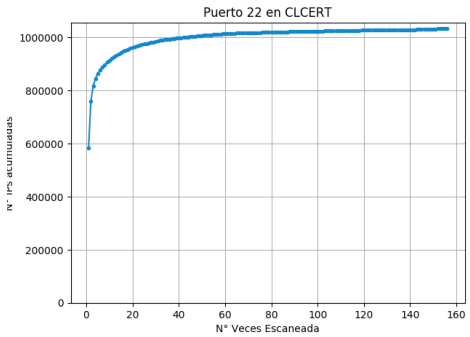
Con respecto a continuidad de existencia de las IP estudiadas, el caso del CLCERT se puede observar en las figuras del grupo 4.2 (páginas 89, 90 y 91). Estas figuras muestran el historial de número de máquinas encontradas por cada puerto, acompañado de otras líneas representando cuántas de estas IP revisadas se habían visto también anteriormente. Con respecto al protocolo FTP, se puede observar en la figura 4.2a la alta variabilidad que posee la detección de este protocolo, tanto en términos de diferencia entre semanas como al revisar los datos históricos. Específicamente en el rango de marzo 2017 a marzo 2018, se observa que la cantidad de IPs comunes con la semana y mes anterior se parecen bastante. Sin embargo, la gran diferencia entre ambas curvas y la cantidad total de IPs encontradas en este intervalo muestra de nuevo una alto número de IPs cambiantes en este escaneo. Además, la situación desde julio de 2019 se desordena bastante, en gran parte debido a la alta variabilidad obtenida en los resultados desde esa fecha, y también por el vacío de tres semanas existente en los datos registrados de esta fuente.

En un contraste bastante notorio, se observa cómo el protocolo SSH en la figura 4.2b posee un comportamiento más regular en términos de IPs repetidas en periodos anteriores. Como es de esperar, a medida la comparación engloba un intervalo más extenso, el número de IPs repetidas es mucho menor. Destaca al igual que en el caso anterior que los resultados de 1 semana y de un mes se comportan de forma bastante similar, encontrándose en ambas curvas entre tres cuartos y la mitad de las IPs totales escaneadas. Retrocediendo a la curva de IPs comunes en resultados de un año antes, la cantidad baja a entre un cuarto y un tercio de éstas en el intervalo de tiempo previo a Enero de 2019.

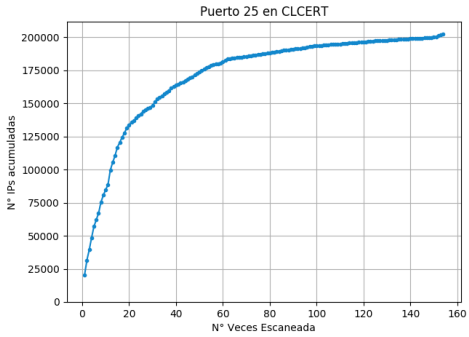
En el caso del protocolo SMTP visto en la figura 4.2c, tanto los escaneos de una semana antes como de un mes antes se comportan de una manera bastante similar, englobando entre el 80 y 90 % de las IPs del momento actual. La línea de los tres meses se empieza a separar de forma más notoria, así como también a tener menos cambios bruscos, mientras que la



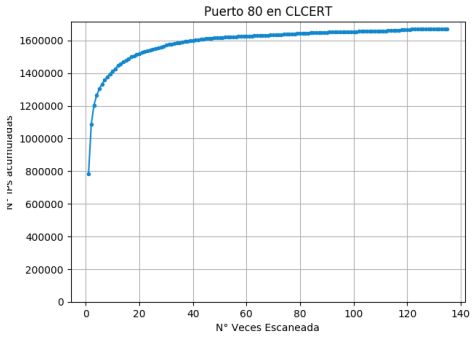
(a) Puerto 21



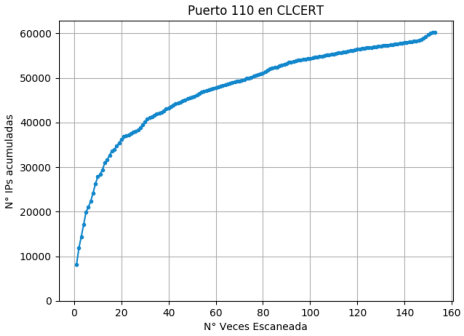
(b) Puerto 22



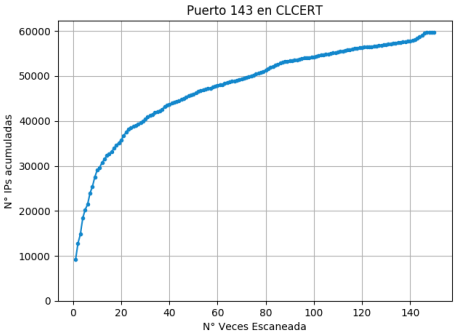
(c) Puerto 25



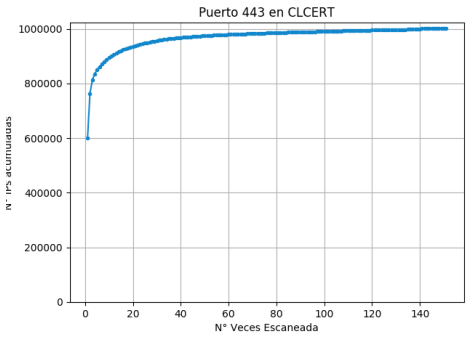
(d) Puerto 80



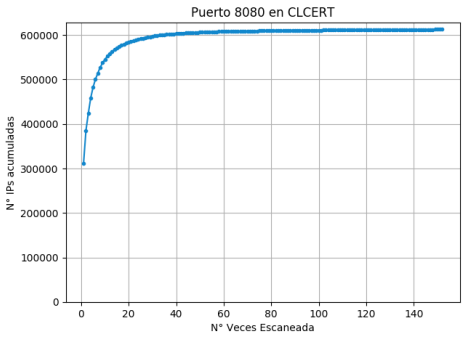
(e) Puerto 110



(f) Puerto 143

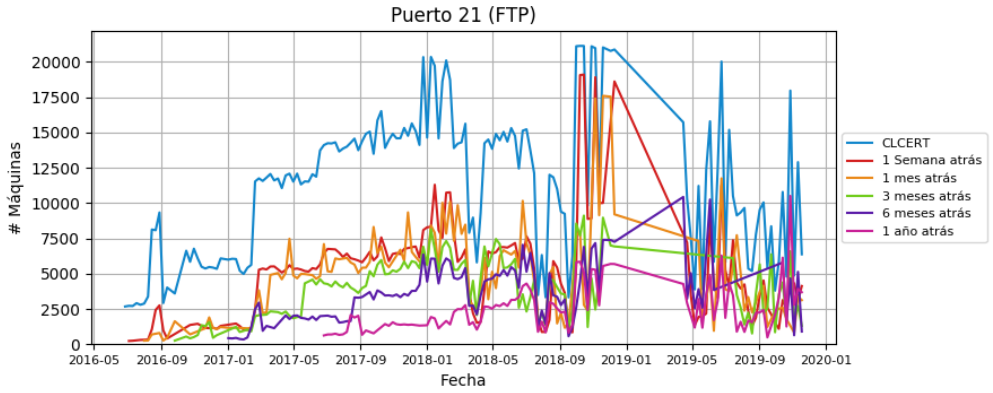


(g) Puerto 443

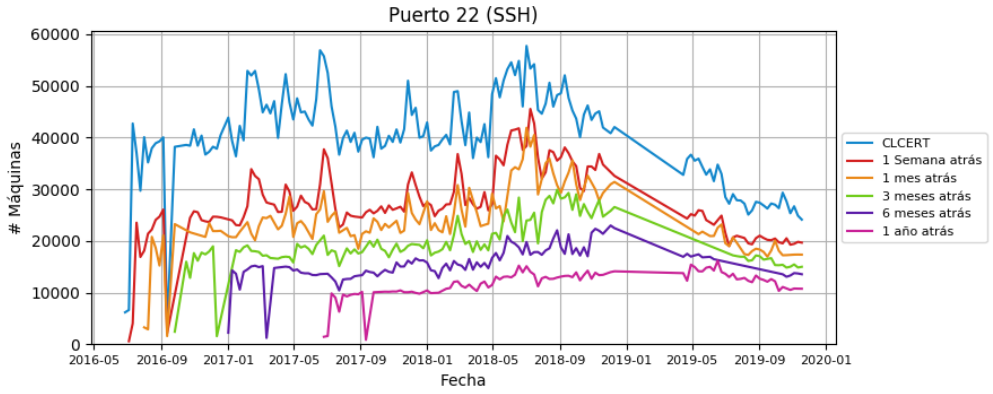


(h) Puerto 8080

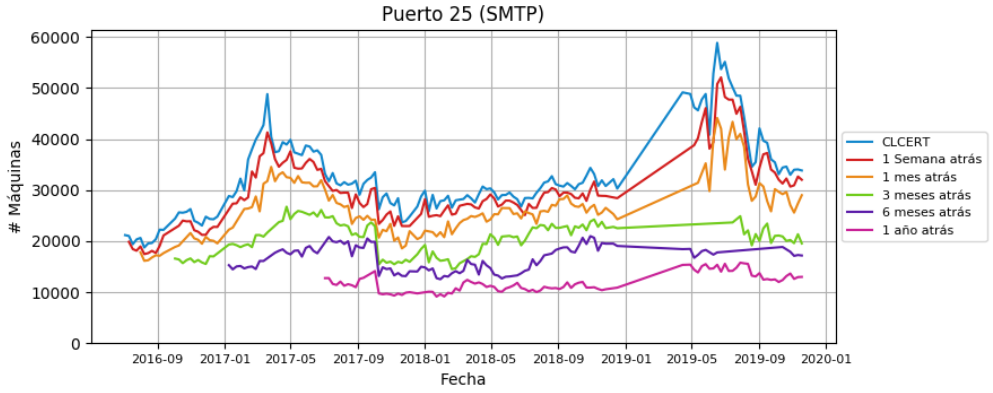
Figura 4.1: Gráfico que muestra la cantidad de IPs (eje y) que se vio al menos cierta cantidad de veces (eje x) en los resultados de escaneos del CLCERT.



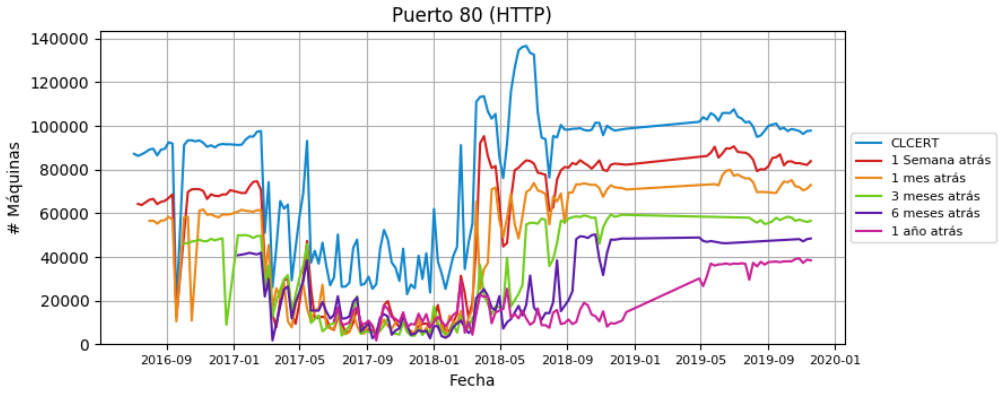
(a) Puerto 21 (FTP)



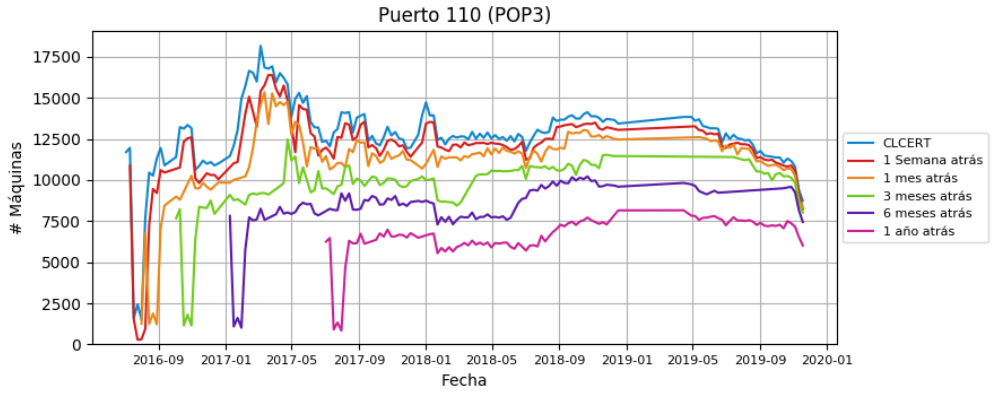
(b) Puerto 22 (SSH)



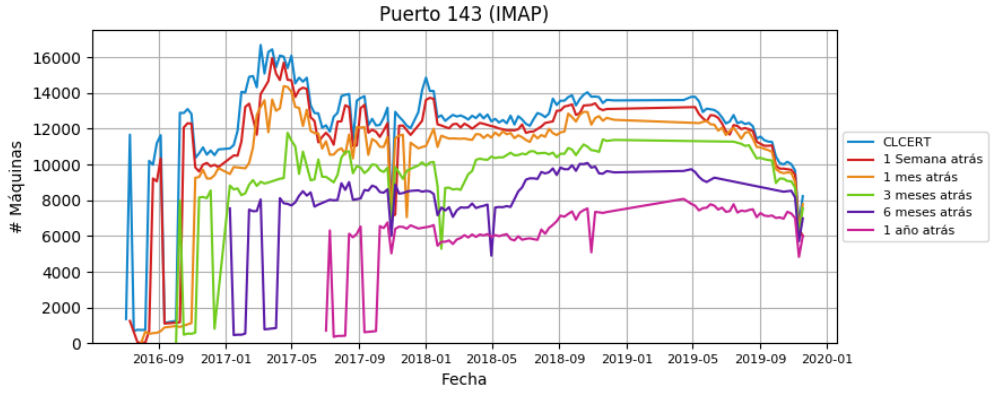
(c) Puerto 25 (SMTP)



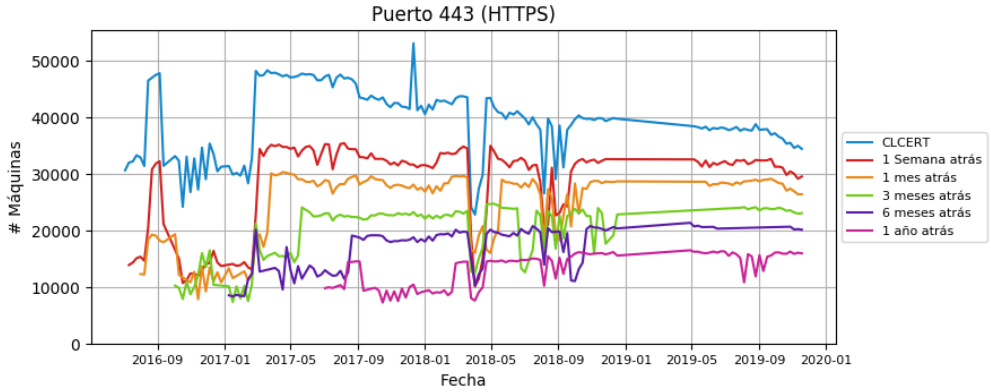
(d) Puerto 80 (HTTP)



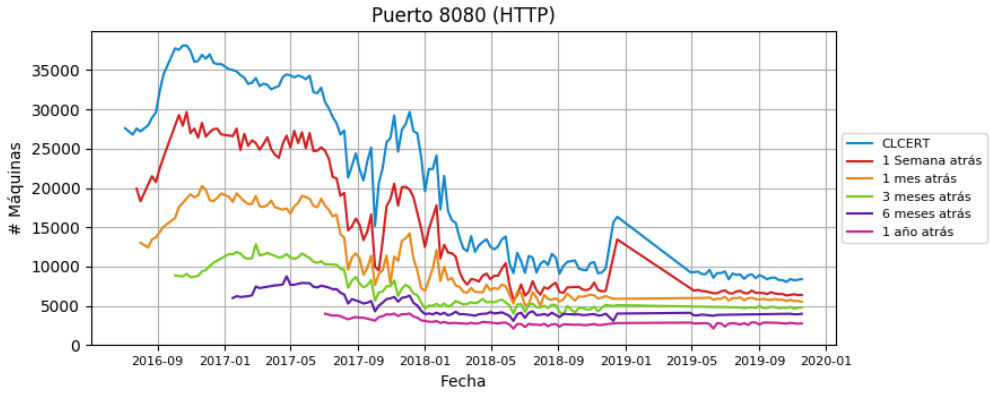
(e) Puerto 110 (POP3)



(f) Puerto 143 (IMAP)



(g) Puerto 443 (HTTPS)



(h) Puerto 8080 (HTTP)

Figura 4.2: Comparación histórica de la cantidad de IPs encontradas por cada protocolo en los escaneos de CLCERT.

línea del año muestra un comportamiento bastante estable y no tan marcado por las grandes diferencias de los escaneos de cada fecha.

El comportamiento del protocolo HTTP visible en la figura 4.2d destaca por una baja pero similar cantidad de IPs repetidas en el periodo marzo 2017 y marzo 2018, en el cual las 5 líneas de periodos anteriores se comportan de manera parecida. Además, desde septiembre de 2018 y descartando el periodo sin datos de inicios de 2019, los valores se han mantenido bastante estables. Desde ese momento, la línea que representa las IPs repetidas con la medición de hace 6 meses se mantiene relativamente estable, cubriendo alrededor de un 50 % de las IP activas en ese momento.

Las figuras 4.2e y 4.2f, correspondientes a los protocolos POP3 y IMAP respectivamente, se comportan bastante parecido en las 6 curvas mostradas, tanto en valores como en épocas de mayor o menor preponderacia. Se observa que la variación de IPs en estos casos es bastante más lenta que en los ya revisados, debido a que, por ejemplo, la cantidad de IPs repetidas de un año antes es más del 50 % de las IPs encontradas en la actualidad.

El gráfico 4.2g muestra el comportamiento del protocolo HTTPS, el cual resulta bastante estable durante el tiempo, aunque con una ligera tendencia a la baja. Salvo por un periodo entre marzo y diciembre de 2018, las 6 líneas se distinguen separadamente en distintos niveles, notándose que un poco menos de un 50 % de las IPs vistas hace un año se mantienen en un escaneo actual.

Por último, se observa una altísima variabilidad de IPs en la figura 4.2h, la cual que muestra el protocolo HTTP sobre el puerto 8080. Esto que se explica por el carácter de uso en sistemas en desarrollo de este puerto. Sin embargo, la curva que muestra las IPs en común de hace un año se mantiene en valores bastante constantes, lo que muestra una cantidad no menor (aproximadamente un tercio del total hacia fines de 2019) de servicios que usan este puerto de forma permanente.

4.2.2. Datos Históricos de Censys

Se realizó un análisis similar al anterior con los datos de Censys, el cual se puede revisar en el Anexo C de este documento. Este análisis compara los resultados obtenidos con los resultados del CLCERT, lo cual si bien no es parte de la estrategia definida, permite entender aún mejor la diferencia de ambas fuentes.

4.2.3. Conclusiones de Estrategia

Esta estrategia consideró dos mediciones distintas que permiten entender mejor las diferencias históricas de IPs en los resultados de escaneos de protocolos.

Con respecto a la estrategia de observar globalmente la cantidad de veces que se repite una IP en el universo total de escaneos, ésta permitió notar fácilmente y de primera mano que existe un número no menor de IPs en los escaneos realizados sobre puertos no relacionados con correo electrónico que no se vuelven a ver más en futuros escaneos. Con el objetivo de focalizar el trabajo de investigación, se recomienda ignorar las IP con pocas apariciones en caso de querer tener una mejor idea de la infraestructura más importante de algún servicio

dentro de la red estudiada. Por otro lado, al mismo tiempo puede ser interesante fijarse aún más en las IPs que no se vuelven a ver, debido a que el carácter efímero del servicio prestado puede indicar alguna anomalía en la configuración de la máquina.

Con respecto a la estrategia de revisión de continuidad de las IPs integrantes de cada escaneo, ésta refuerza la evidencia que muestra que el recambio de IPs en algunos protocolos es bastante común, lo cual es explicable debido a la asignación dinámica de IPs. También da paso a la creación de grupos de IPs que pueden tener mayor relevancia debido a aparecer desde hace más tiempo, de forma similar a la estrategia anterior. La revisión de resultados históricos también sirve para eliminar las grandes variaciones de número de IP entre semanas, entregando un número más conservador para el tamaño de "la internet chilena" desde el punto de vista de cada servicio ofrecido.

4.3. Comparación de datos de escaneo de protocolos de múltiples proveedores

Para entender mejor las diferencias de resultados de escaneo entre fuentes debido a razones geográficas (esto es, la ubicación de la entidad que realiza el escaneo), se propone una revisión más en detalle del contraste de los datos obtenidos tanto de CLCERT como de Censys, comparando los resultados de las mismas semanas con un método similar al del caso histórico, de forma de obtener un subconjunto de IPs con mayor trascendencia en análisis de largo plazo.

Además, se propone comparar los resultados obtenidos entre las dos fuentes estudiadas, intentando aislar cuáles fueron detectados en común y cuáles son únicos de cada una de ellas, de modo de obtener un subconjunto de IPs más validado.

Finalmente, se replicarán los escaneos de protocolos desde una máquina externa usando el mismo software que los escaneos del CLCERT, buscando entender mejor el impacto geográfico en los resultados al descartar el factor de implementación del programa de escaneo.

4.3.1. Comparación entre ambas fuentes

Queriendo entender mejor la razón de las diferencias en los resultados de ambas fuentes, se propone determinar exactamente las IPs en que difieren ambas fuentes de escaneo, creando cuatro grupos de IPs distintos para cada semana, los cuales corresponden a las detectadas por CLCERT y no por Censys, las detectadas por Censys y no por CLCERT, las detectadas por ambas fuentes a la vez y las detectadas por al menos una de las dos fuentes.

La figura 4.3 agrupa las comparaciones de los ocho protocolos estudiados en esta sección. Al revisar el caso del puerto 21 en la figura 4.3a, se observa que la cantidad de IPs en común es en general menos de la mitad que la unión de las IPs encontrada en cada fecha de escaneo, lo que da a entender que una gran cantidad de máquinas es encontrada por solo una fuente y no la otra, lo que da a pensar que factores geográficos o de software pueden estar influyendo en estos resultados. Al mismo tiempo, se observa como razonablemente, la línea que representa las IPs en común se comporta irregularmente si al menos una de las dos fuentes tiene una alta variabilidad en un corto periodo de tiempo.

Revisando los resultados para el protocolo SSH, observables en la figura 4.3b, destaca que la cantidad de IPs en común en el breve periodo en el que ambas líneas registran escaneos es relativamente baja, pero bastante estable, imitando el comportamiento de la fuente Censys. En contraste, la cantidad de IPs detectadas por la unión de ambas fuentes es alta comparada con su intersección, lo que hace pensar que la situación es similar a la vista en el caso del puerto 21, pero con resultados más estables. Además, no es posible seguir comparando después de mayo de 2019, debido a la falta de datos de la fuente Censys.

El resultado para el puerto 25 visible en 4.3c llama la atención debido a que, al menos hasta agosto de 2018 y desde mayo de 2019, los valores obtenidos en la intersección de ambas fuentes y los valores obtenidos por Censys son prácticamente idénticos, tanto en comportamiento como magnitud. Esto indica que en esos intervalos, Censys actúa como un subconjunto de los resultados del CLCERT. Esta situación se parece bastante a la observable en otros puertos relacionados con el protocolo de correo electrónico, como el 110 (figura 4.3e) y 143 (figura 4.3f).

En el caso del puerto 80, revisable en 4.3d, se nota bastante que a pesar que ambas curvas se comportan de forma similar desde abril de 2018, al menos la mitad de las IP detectadas por cada fuente no fueron detectadas por la otra, lo que refuerza el punto postulado al inicio de este capítulo. Esto también se puede observar, aunque brevemente debido a la calidad de los datos capturados, en los resultados del puerto 443, visibles en la figura 4.3g.

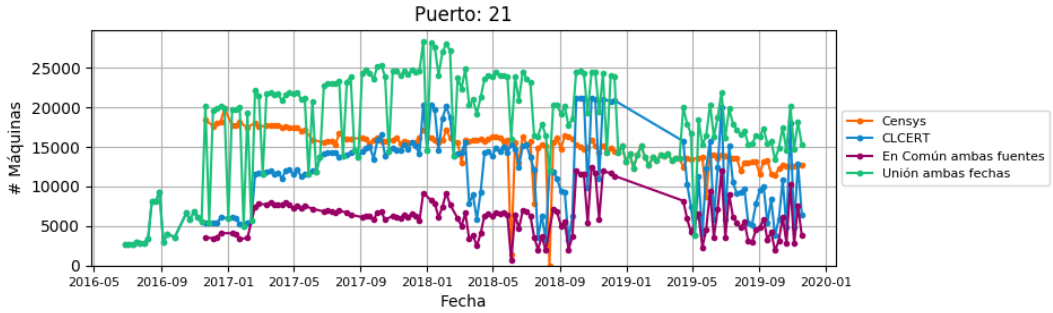
Por último, en el caso del puerto 8080 (figura 4.3h), se aprecia que gran parte de las IPs reconocidas por el escáner del CLCERT son reportadas también por Censys, haciendo que el conjunto unión no varíe tanto en número al compararlo con esta fuente.

4.3.2. Uso de Tercera Fuente

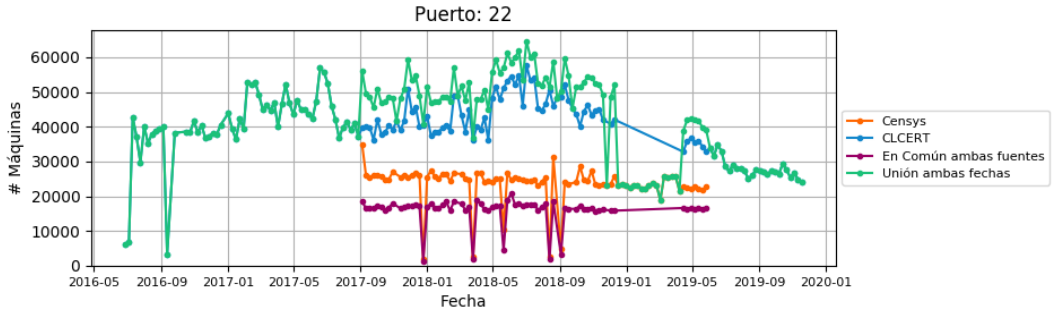
Las comparaciones anteriores ayudan a entender mejor las diferencias y similitudes entre las distintas fuentes de datos, pero no permiten asegurar concretamente las razones de estas discrepancias. Para descartar o confirmar la variable geográfica como causa de la diferencia de resultados, se propone realizar un escaneo utilizando las mismas herramientas que las usadas en la infraestructura interna del CLCERT (llamada *Scan*), pero desde una máquina de tipo *VPS* de *Digital Ocean* (llamada *Scan2*), la cual cuenta con 8 GB de RAM y está ubicada geográficamente en Alemania.

Los escaneos se realizaron a la misma hora durante dos semanas tanto en *Scan* como en *Scan2*, utilizando exactamente las mismas herramientas, sistema operativo (CentOS 7) y configuraciones. La tabla 4.1 muestra los resultados obtenidos en cada una de las dos semanas revisadas.

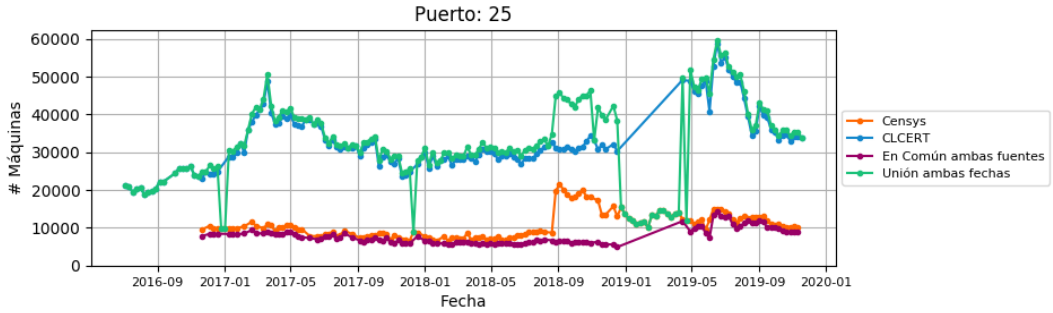
En el caso del puerto 21, la diferencia entre ambas ubicaciones geográficas luego de las dos mediciones tomadas fluctúa entre un 19 % y un 24 % de IPs adicionales detectadas por *Scan2*, mientras que la comparación de cada fuente entre las dos semanas, esta diferencia es casi nula en el caso de *scan2* (0,59 %), y cercana a un 7 % en el caso de *Scan*. Lo anterior hace pensar que, al menos en el caso del puerto 21, la razón de la diferencia en valores debe ser geográfica. Esta suposición se potencia al observar que la cantidad de IPs detectadas por solo una máquina de escaneo son es por lo general mayor a la cantidad de IPs comunes.



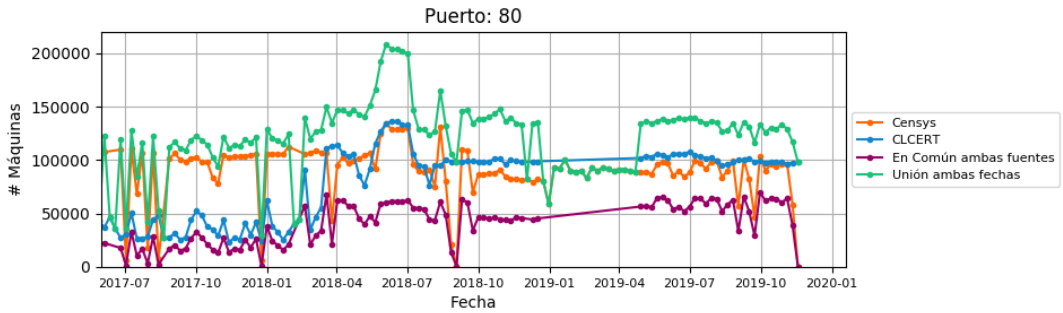
(a) Puerto 21



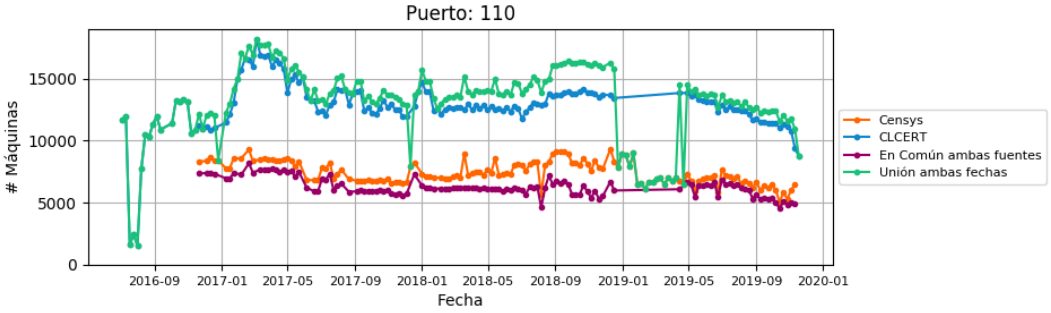
(b) Puerto 22



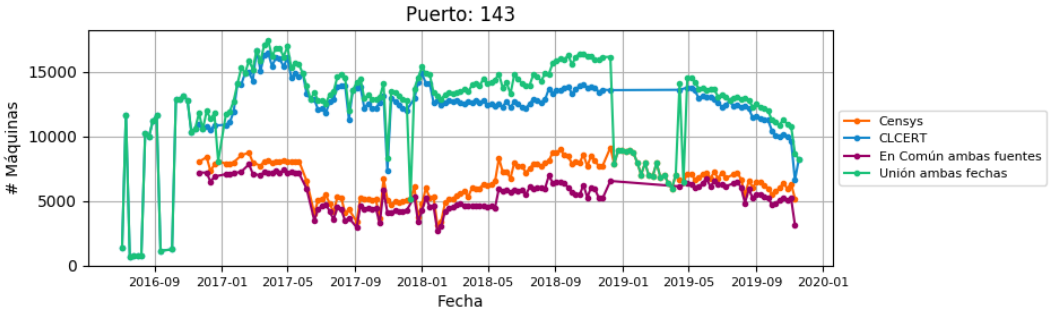
(c) Puerto 25



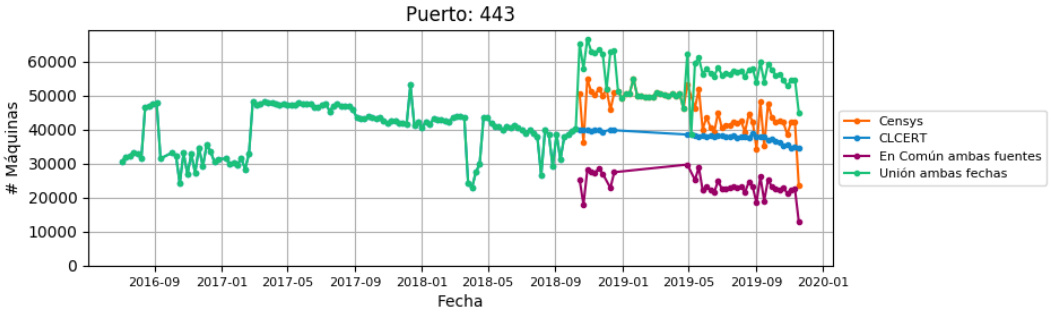
(d) Puerto 80



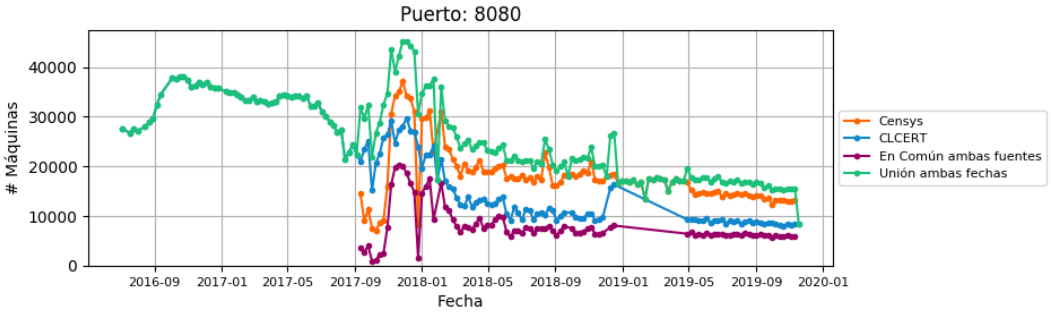
(e) Puerto 110



(f) Puerto 143



(g) Puerto 443



(h) Puerto 8080

Figura 4.3: Comparación entre los servicios encontradas por CLCERT, los encontrados por Censys y los encontrados por ambas fuentes en cada puerto.

Puerto	Fecha	Comunes	$Uniq_{Scan}$	$Totales_{Scan}$	$UniqScan2$	$Totales_{Scan2}$	$Diff_{Scan,Scan2}$	$Diff_{Scan}$	$Diff_{Scan2}$
21	2019-11-18	2.555	3.281	5.836	4.711	7.266	124,50 %	107,38 %	100,59 %
	2019-11-25	2.580	3.523	6.103	4.729	7.309	119,76 %		
22	2019-11-18	22.011	2.135	24.146	1.552	23.563	97,59 %	102,10 %	104,01 %
	2019-11-25	23.147	1.505	24.652	1.362	24.509	99,42 %		
25	2019-11-18	33.324	518	33.842	1.342	34.666	102,43 %	97,89 %	95,82 %
	2019-11-25	32.055	1.074	33.129	1.163	33.218	100,27 %		
80	2019-11-18	96.601	1.221	97.822	2.769	99.370	101,58 %	99,57 %	92,40 %
	2019-11-25	90.141	7.265	97.406	1.677	91.818	94,26 %		
110	2019-11-18	8.218	520	8.738	2.189	10.407	119,10 %	95,00 %	92,80 %
	2019-11-25	7.488	893	8.381	2.170	9.658	115,24 %		
143	2019-11-18	8.021	209	8.230	2.081	10.102	122,75 %	95,43 %	90,92 %
	2019-11-25	7.279	575	7.854	1.906	9.185	116,95 %		
443	2019-11-18	32.196	2.247	34.443	12.953	45.149	131,08 %	100,77 %	91,26 %
	2019-11-25	30.353	4.354	34.707	11.097	41.450	119,43 %		
8000	2019-11-18	2.981	107	3.088	45	3.026	97,99 %	101,46 %	98,65 %
	2019-11-25	2.961	172	3.133	24	2.985	95,28 %		
8080	2019-11-18	8.061	349	8.410	161	8.222	97,76 %	99,52 %	89,21 %
	2019-11-25	7.270	1.100	8.370	65	7.335	87,63 %		

Tabla 4.1: Comparación entre los resultados obtenidos durante dos sesiones de escaneo, la primera realizada por la infraestructura tradicional del CLCERT (Scan), mientras que la segunda realizada por una máquina hospedada en Digital Ocean y ubicada en Alemania (Scan2). Las columnas $Diff_{Scan}$ y $Diff_{Scan2}$ comparan los resultados de cada infraestructura entre las dos semanas estudiadas.

Otros casos que destacan por altas diferencias son los puertos 110, 143 y 443, los cuales tienen diferencias de al menos un 16 % entre fuentes, manteniendo resultados relativamente consistentes entre semanas, con una variabilidad no mayor al 10 %.

En el caso del puerto 22, la fuente *Scan* obtuvo más resultados en ambas ocasiones, con una diferencia de entre un 0,58 % y un 2,41 % con respecto a *Scan2*. Las diferencias entre las dos sesiones de escaneo de cada fuente van entre un 2 % y un 4 % aproximadamente, por lo que se observa con ambos datos que los resultados son bastante similares para este protocolo, desde cualquier ubicación. El caso del puerto 25, 80, 8000 y 8080 es bastante similar en magnitud porcentual.

4.3.3. Conclusiones de Estrategia

Esta estrategia propuesta trae dos beneficios claros en el estudio de datos de escaneos de protocolos. En primer lugar, permite la determinación de dos nuevos grupos de IPs por protocolo a estudiar a partir de 2 fuentes distintas. Uno correspondiente a las IPs visibles desde ambas fuentes, y otro correspondiente a las IPs visibles desde al menos una fuente. Estas dos medidas permiten dar estimaciones tanto conservadoras como más integrales del tamaño de la internet chilena con respecto a un servicio estudiado. La segunda enseñanza de esta estrategia es que permite el descarte de factores externos que provocan diferencias entre los resultados. Por ejemplo, la tabla 4.1 permite notar que en algunos servicios, el impacto geográfico es mucho mayor que el impacto del software usado.

Sin embargo, todavía quedan interrogantes acerca de las diferencias de los resultados obtenidos entre Censys y CLCERT, las cuales si bien pueden estar ocasionadas por diferencias en implementaciones, es necesario realizar más pruebas para establecer conclusiones convin-

centes.

4.4. Uso de datos de abandono en estimación de máquinas vulnerables

Esta sección propone una estrategia para detectar una nueva categoría de máquinas vulnerables, no recibida en estos momentos por ninguna fuente externa y basada en el "abandono" de algunas máquinas según información extraíble de los protocolos ejecutados. El objetivo de esta sección es explicar la metodología para definir a una máquina como vulnerable según abandono, para luego contrastar los datos de máquinas detectadas usando este procedimiento con los resultados de otro tipo de vulnerabilidades.

4.4.1. Definición de Abandono

Este trabajo define "abandono" como *el conjunto de características que hacen notar que una máquina con servicios públicos en Internet no está siendo mantenida de forma regular*, y propone la utilidad de esta clasificación para la detección de grupos de riesgo en los cuales es más probable encontrar máquinas vulnerables.

Para esta sección se considerarán dos técnicas de determinación de abandono. La primera forma se considera directa y consiste en la revisión de la versión y tipo del software que corre el servicio en cada máquina. La segunda forma se considera indirecta y consiste en la revisión de metadatos que hacen pensar que la máquina no ha sido actualizada en mucho tiempo. En el caso indirecto, este trabajo usará exclusivamente la fecha de vencimiento de certificados SSL/TLS.

La forma directa es la más certera para identificar casos de máquinas vulnerables por abandono, ya que de hacerse correctamente, se puede obtener exactamente el historial de actualizaciones del software revisado. Sin embargo, es más difícil de poner en práctica, debido a que se requieren estrategias para hacer un correcto cruce entre versiones mostradas por los *banners* de los servicios corridos y el historial de versiones existente para cada programa. Lo anterior se complica porque varios servicios, por temas de seguridad, ocultan intencionalmente las versiones de software que corren en los banners, justamente para evitar ser blancos de ataques dirigidos.

En contraste, la forma indirecta es utilizable cada vez que se escanea un protocolo con certificado SSL/TLS, y se justifica en el supuesto de que una deficiente actualización de estos certificados implica la existencia de falta de mantención y olvido sobre el sistema. Esta estrategia es más fácil de implementar, ya que solo requiere recopilar certificados y comparar la fecha de vencimiento con la fecha en que se consiguió.

En este trabajo se implementarán y ejecutarán ambas estrategias de abandono propuestas sobre los datos recopilados, con dos objetivos principales. En primer lugar, se propone que esta clasificación serviría para establecer un nuevo grupo de máquinas importantes de revisar en la red monitoreada. En segundo lugar, se desea conocer si los resultados de malware manejados por el CLCERT son más o menos concentrados que en los conjuntos generales en estos subconjuntos marcados como "abandonados", de forma de determinar qué impacto

Tipo	$\#IPs_{C/Cert}$	$\#IPs_{Vencido}$
Certificados Web	1.198.454	669.644
Certificados E-Mail	81.466	17.342
Total	**1.241.935**	**683.602**

Tabla 4.2: Tabla con cantidad de máquinas totales y con un certificado vencido para cada tipo.

puede tener el "abandono" en los resultados obtenidos por metodologías desconocidas de fuentes externas.

Para ejecutar la estrategia de esta sección, en primer lugar se obtendrán los subconjuntos de abandono de cada categoría ya mencionada: certificados y software obsoleto. Posteriormente, se compararán los resultados de las dos métricas de abandono propuestas con las IPs encontradas por fecha con algún malware en las categorías *bot*, *bruteforce* y *darknet*. Esto determina un total de 1.441.098 IPs distintas marcadas como malware de los tipos ya mencionados entre septiembre de 2017 y noviembre de 2019, comparadas con los valores de abandono encontradas por fechas respectivas.

4.4.2. Abandono por Certificados

Como ya se mencionó, una forma de definir la métrica de abandono es la revisión de las IPs que están asociadas a certificados SSL/TLS vencidos.

Entre julio de 2016 y noviembre de 2019, el CLCERT posee un registro total de 1.241.935 IPs asociadas a al menos un certificado, de las cuales 683.602 de ellas se se encontraban asociadas a al menos un certificado vencido durante el periodo completo de escaneos, lo que corresponde a más de la mitad de las IPs escaneadas con certificado.

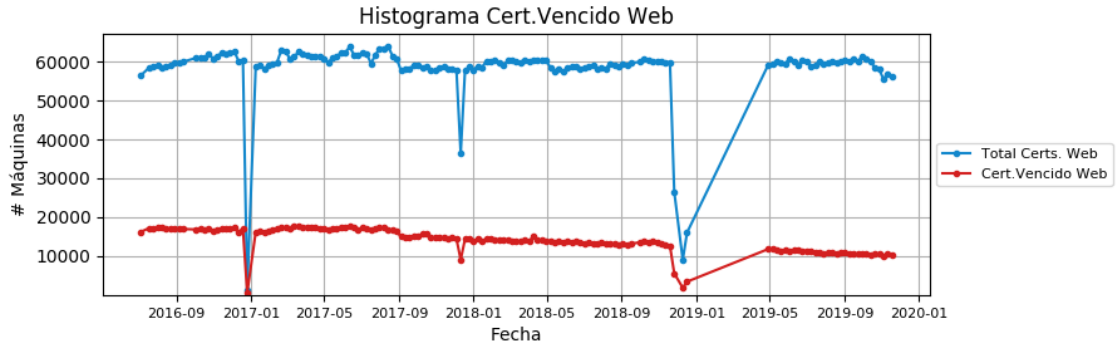
Los certificados recopilados están relacionados a los siguientes dos protocolos de capa de aplicación:

- **Protocolos Web**: Para el cual se recopilaron certificados en el puerto 443. En este conjunto hay 1.198.454 IPs distintas con certificados, de las cuales 669.644 están asociadas a al menos un certificado vencido en la historia del escaneo.
- **Protocolos Relacionados con E-Mail**: para los cuales se recopilaron certificados de los protocolos SMTP (puertos 25/TCP y 465/TCP), IMAP (puertos 143/TCP y 993/TCP) y POP3 (puertos 110/TCP y 995/TCP). En este conjunto hay 81.466 IPs distintas con certificados, de las cuales 17.342 están asociadas a al menos un certificado vencido en la historia del escaneo.

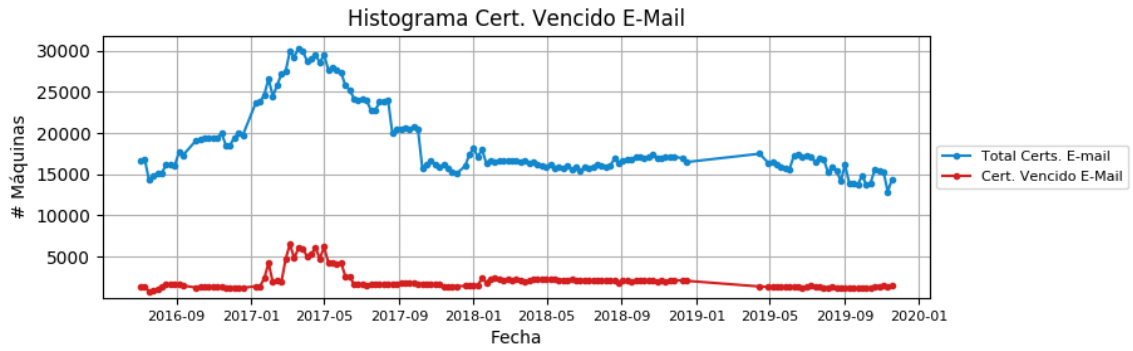
Datos históricos

El desglose numérico de IPs únicas en ambas categorías se puede observar en la tabla 4.2. En esta tabla se muestra la diferencia entre IPs asociadas a certificados relacionados con sitios web e IPs asociadas a certificados usados por servicios de correo electrónico. Así como también la cantidad total de IPs asociadas a certificados vencidos en cada caso.

El análisis partirá entendiendo mejor qué proporción del conjunto completo de certificados



(a) Tipo Web



(b) Tipo de E-Mail

Figura 4.4: Histogramas de cantidad de máquinas con un certificado vencido del tipo correspondiente.

para cada categoría se considera vencida. El grupo de figuras 4.4 muestra una comparación entre número total de IPs con al menos un certificado de cada tipo y número total de IPs que se encontraron con certificados vencidos.

En el caso de la figura 4.4a, se puede apreciar que la cantidad de certificados vencidos en servicios Web ha disminuido desde aproximadamente un tercio de la cantidad total de certificados hasta un sexto de ésta. Por otro lado, el gráfico 4.4b refuerza visualmente que la cantidad de certificados vencidos en servicios de E-mail con respecto al total es bastante baja. También se puede notar al contrastar con la tabla 4.2 que la cantidad de IPs únicas en certificados web es muchísimo mayor a la cantidad de IPs relacionadas con certificados de e-mail. Sin embargo, los gráficos muestran que la diferencia en IPs escaneadas por día no es tan grande, lo que confirma una alta rotación de IPs en el periodo completo de escaneo.

Comparación con malware

En esta sección se revisará si los conjuntos de IPs con certificados vencidos poseen una mayor concentración de IPs marcadas como malware según fuentes externas que el conjunto completo de IPs con certificados.

Los grupos de figuras 4.5, 4.6 y 4.7 muestran comparaciones históricas tanto absolutas como porcentuales, sobre los datos en común entre los subconjuntos de malware, el universo completo de IPs con certificados por categoría y solamente los certificados vencidos por

categoría. En todos los casos, la primera figura muestra en escala y logarítmica la cantidad de IPs encontradas en común entre el conjunto completo de certificados de cada categoría y el conjunto de malware correspondiente, mientras que la segunda categoría muestra los mismos datos, pero en relaciones porcentuales con el universo de IPs con certificados por categoría. La tercera y cuarta figura de cada grupo son análogas a las ya explicadas, pero utilizan el conjunto de certificados vencidos para cada categoría en vez del conjunto completo.

El grupo de figuras 4.5 muestra comparaciones entre los conjuntos de certificados y el de malware de tipo *bot*. Con respecto a 4.5a, se puede ver que la cantidad de IPs intersectadas con respecto al número total de IPs de malware no es muy alta, tanto en el caso Web (entre 100 y 900) como E-mail (entre 10 y 100). Revisando esto en términos porcentuales con ayuda de la figura 4.5b, se observa que las IPs intersectadas no corresponden a mucho más del 1,25 % en el mejor de los casos del conjunto total de IPs con certificado por categoría. Al trabajar solamente las IPs con certificados vencidos, se puede ver en la figura 4.5c que la cantidad de IPs intersectadas baja considerablemente, oscilando entre valores cercanos 10 y 200 en el caso Web, y entre valores cercanos a 0 y 10 en el caso E-mail. Sin embargo, los resultados visibles en la figura 4.5d muestran que el comportamiento relativo es bastante similar en el caso Web, mientras que en el caso de E-mail la baja cantidad de valores dificulta bastante el uso de esta cantidad de valores para establecer una conclusión determinante. Esto permite conjeturar que el subconjunto de IPs con certificados vencidos no tiene ninguna característica especial de vulnerabilidad sobre el conjunto de IPs con todos los certificados, con respecto a los reportes de malware de tipo *bots* recibidos por las fuentes externas.

El grupo de figuras 4.6 agrupa los mismos resultados revisados anteriormente, pero al comparar los conjuntos de certificados con IPs de malware de tipo *fuerza bruta*. Sin embargo, se puede observar en la figura 4.6a que la cantidad de IPs con certificados intersectadas con el número de IPs de malware no supera las 30 en el caso Web, y las 20 en el caso E-mail. La situación es aún menos representativa en el caso de la figura 4.6c, en la cual la cantidad de IPs intersectadas apenas supera las 20 en Web, y las 4 en E-mail. Esta falta de datos de intersección apunta a que no existe una relación significativa entre IPs con certificados y datos de ataques de fuerza bruta, por lo que mucho menos hay relación dentro del conjunto de las IPs con certificados vencidos.

Finalmente, al revisar el conjunto de figuras 4.7, la cual compara los datos ya revisados con conjuntos de IPs vulnerables según la categoría *darknet*, se observa en el gráfico 4.7a que la cantidad de IPs intersectadas del universo de certificados Web se encuentra en el rango de entre 50 y 200, mientras que en el caso de certificados de E-mail y 8 y 20, respectivamente. Lo último vuelve a descartar el uso de los resultados de certificados de e-mail debido a la baja cantidad de estos. Por otro lado, en el caso de la figura 4.7c, se puede ver que la cantidad de IPs con certificados vencidos de tipo web que también aparecen en reportes *darknet* oscila entre valores del orden de 10 y 100 coincidencias, mientras que se confirma la sospecha del caso de certificados E-mail, donde la cantidad de IPs no supera las 7. Finalmente, se puede apreciar en los gráficos 4.7d y 4.7b que los intervalos de porcentaje para Web son bastante similares, oscilando entre 0,1 % y 0,6 %.

En resumen, para los 3 casos, no se encontró evidencia suficiente que permitiese determinar que los subconjuntos de certificados vencidos estuviesen relacionados con un aumento relativo

en la cantidad de IPs con malware reportado por las fuentes manejadas.

4.4.3. Abandono por Software Obsoleto

Como ya ha sido mencionado, otra forma de definir abandono corresponde al uso de las versiones del software encargado de servir el protocolo escaneado que está corriendo en la máquina. En caso que esta versión no sea la más reciente existente hasta ese momento, la máquina se considera en estado de potencial vulnerabilidad. Lo anterior se basaría en el supuesto de que una gran cantidad de actualizaciones de software incluye parches de seguridad sobre vulnerabilidades descubiertas en versiones anteriores.

Para la realización este análisis se utilizaron los resultados obtenidos de los escaneos de protocolos del CLCERT. desde los cuales como se mencionó anteriormente se extrajo información de variante de software y versión utilizando expresiones regulares definidas como parte de un proceso del sistema OSR.

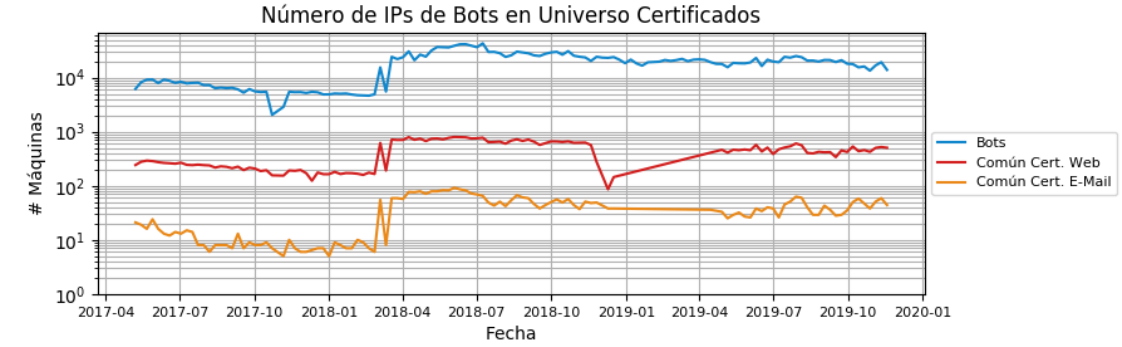
Para poder implementar esta estrategia, se seleccionaron a mano las variedades de software más detectadas a partir de los escaneos históricos, y se recopilaron sus versiones usando archivos CHANGELOG o información de *tags* y *releases* de repositorios Git en el caso del software libre. Lo anterior permitió armar una lista con más de 2.000 versiones de software y sus fechas de lanzamiento, la cual se puede ver agregada en la tabla 4.3, junto con la cantidad de IPs únicas detectadas con ese software a lo largo de los escaneos. El software recopilado se clasificó según una de las siguientes categorías: *FTP*, *SSH*, *Web* y *E-mail*.

Se decidió usar la fecha de lanzamiento del software en vez del número de versión para verificar el estado de obsolescencia, debido a que la mayoría del análisis se hace sobre datos históricos, para los cuales es necesario contar con esa información para determinar si a la fecha del escaneo existen versiones publicadas más recientes. Si bien esta metodología generaliza que las nuevas versiones de software siempre parchan problemas existentes en todas las versiones más antiguas, se plantea que es una buena aproximación al concepto buscado, debido a la baja cantidad de situaciones en las que no ocurre esto. Además, esta metodología permite discriminar entre distintos periodos de desactualización, con el objetivo de calibrar el parámetro"obsolescencia" según una cantidad de tiempo razonable para el conjunto de datos estudiado.

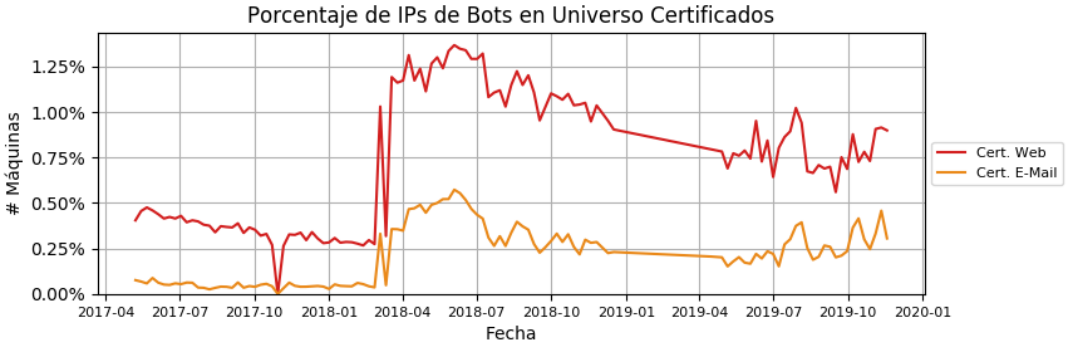
Revisión preliminar

En primer lugar, se revisarán los datos de obsolescencia, comparando la cantidad de software "obsoleto" por servicio según la cantidad de semanas que lleva este servicio sin ser actualizado a una versión posterior, con el objetivo de entender qué valor umbral de "obsolescencia" es el que entrega mejores resultados en la realidad de los servicios del país. Para poder hacer esto, el grupo 4.8 muestra la cantidad de máquinas obsoletas para cada servicio, según cantidad de semanas de diferencia entre la fecha de escaneo y el lanzamiento de una versión superior a la manejada por ellas. Se revisan en intervalos de una semana, un mes, tres meses, seis meses, un año, dos años, cuatro años y ocho años.

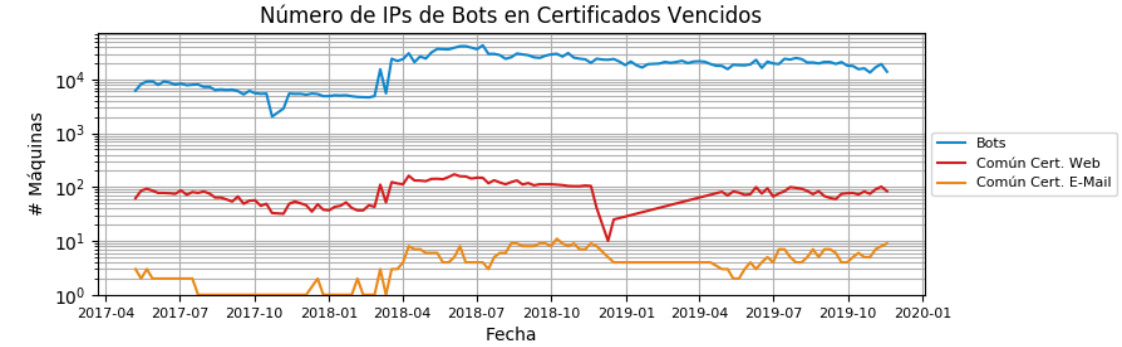
En el caso de la figura 4.8a, referida a software web, se observa que la mayoría del software obsoleto lleva un año o más sin actualizarse a una versión actualizada. Además, una cantidad



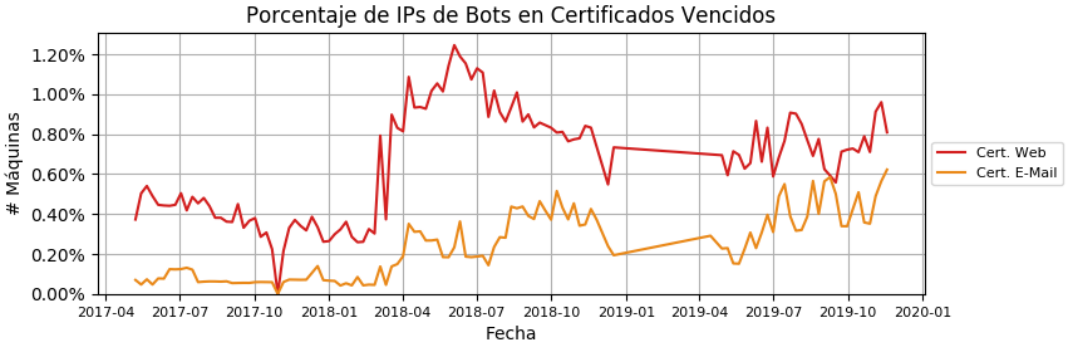
(a) Universo Certificados, absoluto



(b) Universo certificados, porcentual

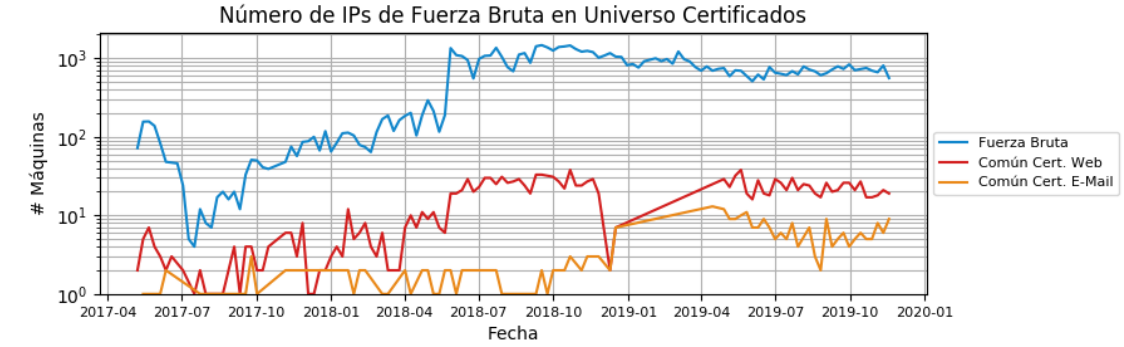


(c) Certificados vencidos, absoluto

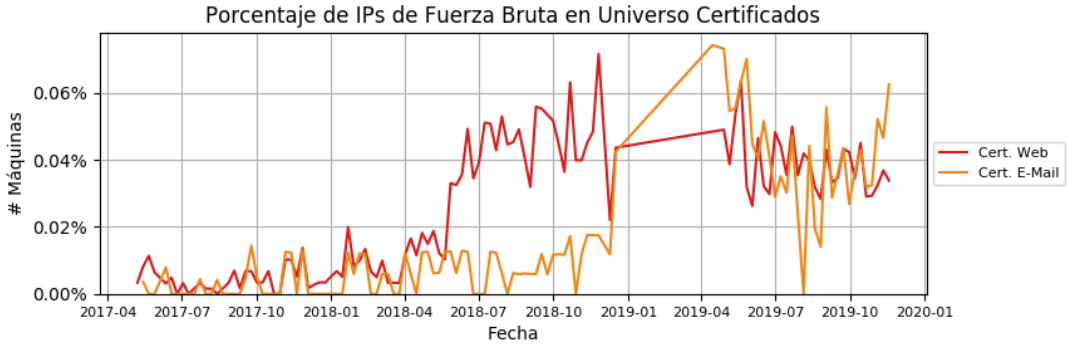


(d) Certificados vencidos, porcentual

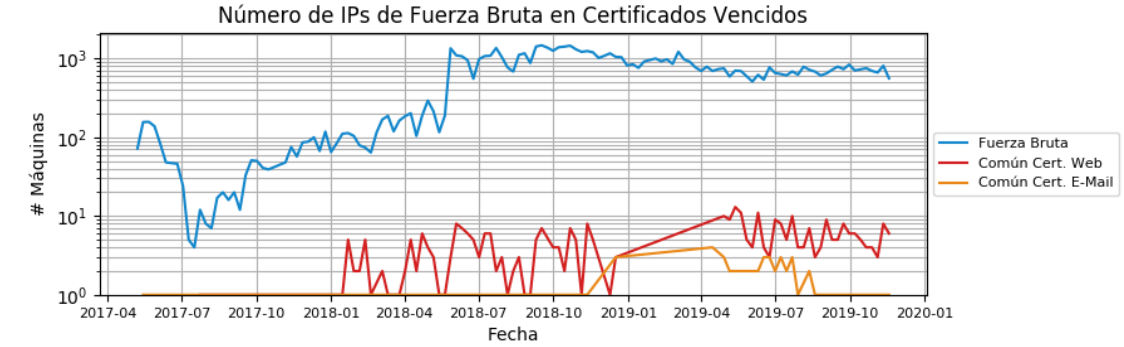
Figura 4.5: Comparaciones entre IPs de certificados (tanto vencidos como universo completo) y datos de malware de tipo bots.



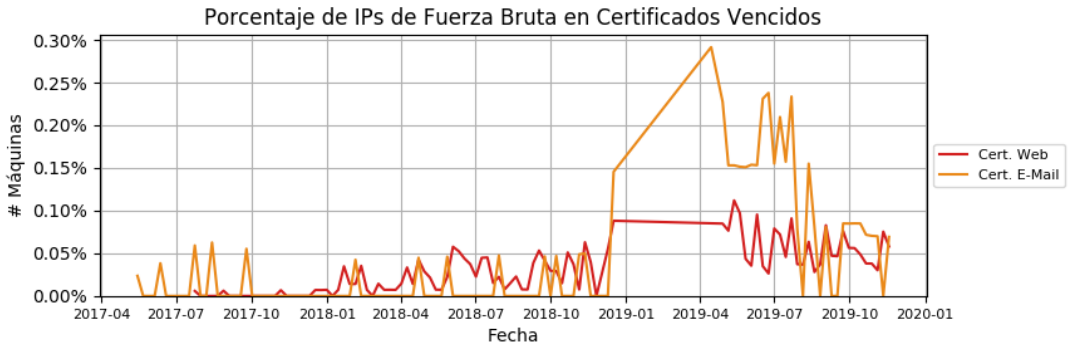
(a) Universo Certificados, absoluto



(b) Universo certificados, porcentual

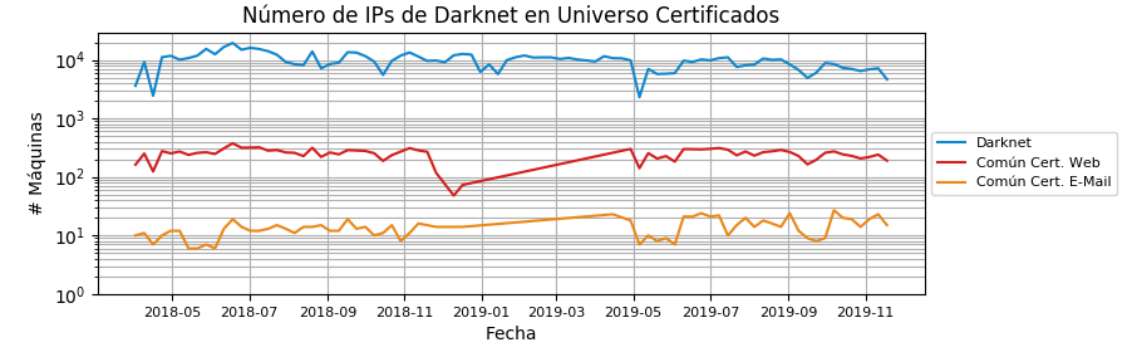


(c) Certificados vencidos, absoluto

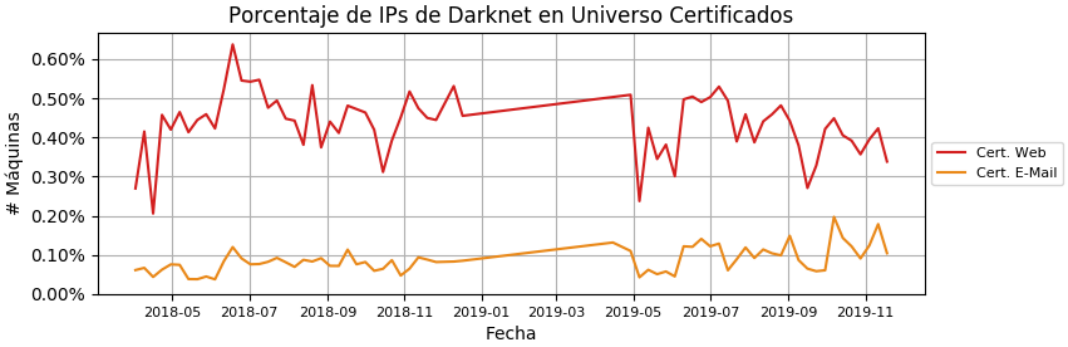


(d) Certificados vencidos, porcentual

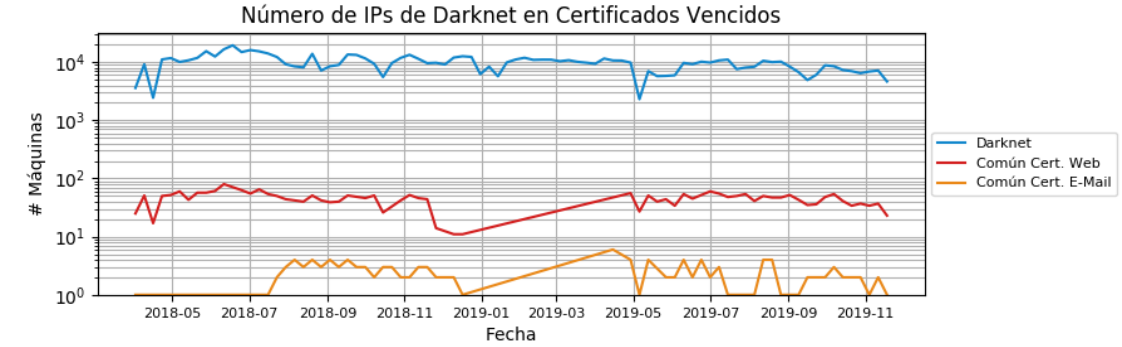
Figura 4.6: Comparaciones entre IPs de certificados (tanto vencidos como universo completo) y datos de malware de tipo bruteforce.



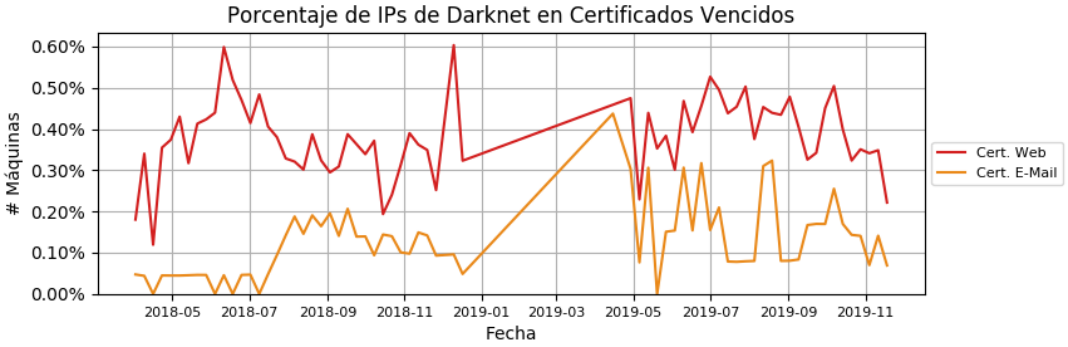
(a) Universo Certificados, absoluto



(b) Universo certificados, porcentual



(c) Certificados vencidos, absoluto



(d) Certificados vencidos, porcentual

Figura 4.7: Comparaciones entre IPs de certificados (tanto vencidos como universo completo) y datos de malware de tipo darknet.

Tipo	Nombre de Software	N^o Versiones	Menor Versión	Mayor Versión	#IPs
Web	Apache HTTP Server	219	1996-12-01	2019-08-08	300.242
	Lighttpd	61	2005-02-19	2019-05-26	119.508
	NGINX	520	2004-10-03	2019-10-21	56.712
E-Mail	Postfix	188	2008-01-23	2019-09-21	81.150
	Sendmail	85	1993-06-06	2015-07-02	63.831
	Exim	136	2005-02-16	2019-11-09	31.619
	Cyrus	186	1996-10-10	2019-11-14	858
	Kerio Connect	89	2010-02-08	2019-07-31	419
	MDaemon	203	2000-02-28	2019-07-15	278
SSH	Dropbear	57	2003-04-05	2019-03-26	612.238
	OpenSSH	80	2000-03-04	2019-10-08	348.647
FTP	PureFTPd	18	2011-04-30	2019-04-02	26.027
	Serv-U	9	2012-12-30	2019-04-24	10.107
	ProFTPd	72	1999-10-26	2019-10-18	8.657
	Filezilla Server	37	2008-07-12	2017-02-07	7.109
Total		**2009**			**1.212.902**

Tabla 4.3: Tabla que muestra los nombres, rangos de fecha, cantidad de versiones de software y de cantidad de IPs distintas que se encontraron históricamente en el software escaneado, utilizando algunos de estos programas. El valor total de la columna #IPs corresponde al total de IPs distintas, por lo que es menor que la suma de todos los valores en ella.

no despreciable se mantiene en funcionamiento ocho años o más desde que se lanzó la versión más reciente.

El caso del software de e-mail, visible en la figura 4.8b, muestra que éste tipo de software se mantiene más actualizado que en el caso anterior. Además, llama la atención el aumento abrupto de software obsoleto en algunas líneas graficadas, el cual se puede deber a máquinas con algún software popular para el cual ha salido alguna versión de software nueva hace poco tiempo, y cuya actualización no se ha realizado de forma masiva. Destaca también durante 2018 y 2019 un número no menor de servidores con software de más de ocho años sin actualización.

La figura 4.8c muestra la situación para el software SSH. Se observa al igual que en los casos anteriores una cantidad importante de máquinas con versiones obsoletas de hace más de ocho años, mientras que la mayoría de las máquinas detectadas se encuentran, en general, en el caso de un año de software desactualizado.

Finalmente, en la figura 4.8d se observa como si bien antes de mayo de 2019 se detectaba un gran número de software no actualizado hace un año, desde esa fecha una gran parte del software detectado lleva 2 años o más sin actualizarse.

A partir de los datos anteriores, y con el objetivo de utilizar un volumen de datos no muy pequeño proporcionalmente al total de los dispositivos obsoletos identificados, pero que indudablemente presenta software con una versión lo suficientemente antigua como para considerarse insegura, se decidió tomar el subconjunto de datos obsoletos **mayor o igual a**

Tipo	$\#IPs_{Activas}$	$\#IPs_{C/Software}$	$\#IPs_{C/Ver.}$	$\#IPs_{Obsoletas}$
Web	2.302.484	440.320	311.504	192.074
E-Mail	204.091	156.211	47.816	17.251
SSH	1.033.207	934.271	933.117	653.944
FTP	192.117	50.573	25.251	5.830
Total	**2.712.104**	**1.212.902**	**1.129.720**	**820.556**

Tabla 4.4: Tabla que muestra el universo de IPs en cada categoría de software revisada. La primera columna numérica muestra el universo de IPs activas en esa categoría, la segunda muestra la cantidad de IPs con nombre de software, la tercera muestra la cantidad de IPs con nombre de software y versión y la última muestra el número total de IPs obsoletas (software potencialmente actualizable hace 4 años o más) encontradas. El valor total de cada columna numérica corresponde al total de IPs distintas de esa columna, por lo que es menor que la suma de todos los valores en ella.

4 años para los cuatro servicios revisados.

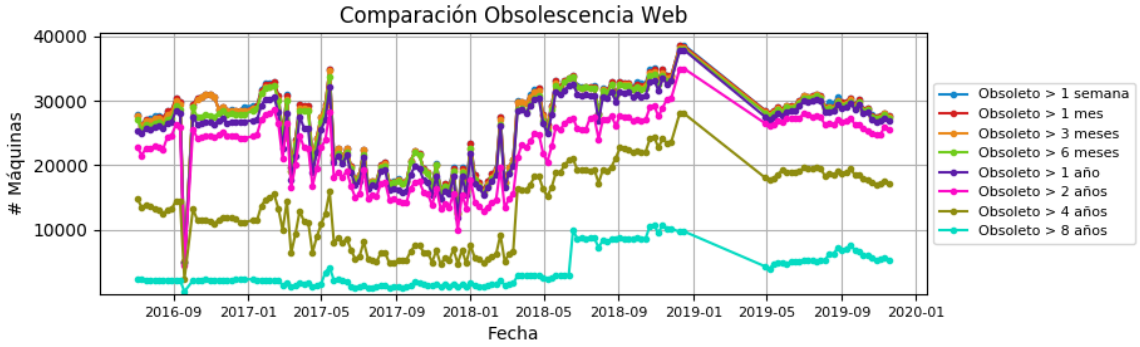
La cantidad de IPs únicas encontradas para cada categoría de software revisado se puede ver en la tabla 4.4. Esta tabla muestra cuatro columnas distintas para cada categoría. Cada columna simboliza la cantidad de IPs únicas encontradas según algún criterio, La primera columna numérica contiene la cantidad de IPs distintas activas en los protocolos relacionados con la categoría. La segunda columna contiene la cantidad de IPs a las que se les pudo asignar uno de los software recopilados en la tabla 4.3. La tercera contiene la cantidad de IPs a la que se le pudo asignar un software y una versión, y la cuarta contiene la cantidad de IPs obsoletas detectadas según el criterio explicado con anterioridad.

Varios datos de esta tabla llaman la atención. En primer lugar, se observa que el software elegido para detectar representa menos de un quinto de la cantidad de IPs activas con protocolo web. Sin embargo, de ese conjunto, se detectó en un 70,7 % algún valor usable como versión del software usado. Además, aproximadamente el 61,7 % de estas máquinas se considera como usando software obsoleto.

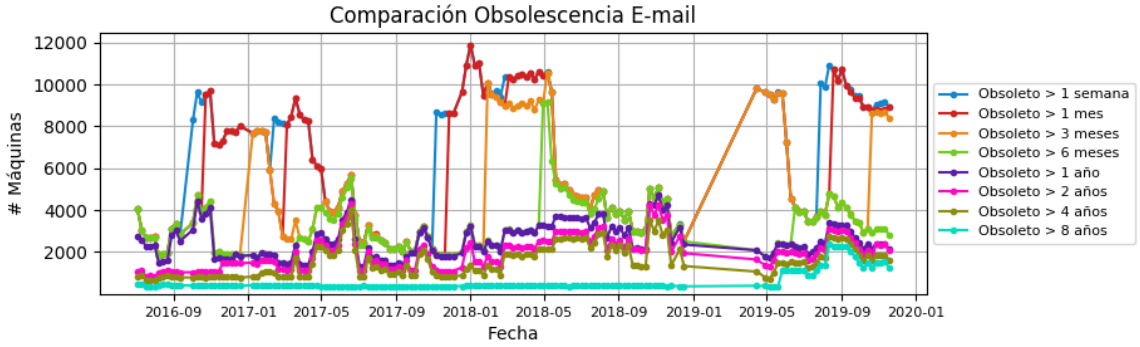
Con respecto al caso de E-mail, se pudo adivinar el software utilizado en un 76,5 % de los casos. Sin embargo, no se pudo detectar la versión de éste en más de un 30,6 % de las IPs únicas con software detectado. De este conjunto, un 36,1 % de las IPs cuentan con software obsoleto.

El caso de SSH es más exitoso, dado que un 90 % de las IPs activas en el protocolo pudieron ser relacionadas con alguno de los software estudiados. De este número, se pudo asignar una versión al 99,8 % de los servidores. Esto se debe en parte a lo estandarizado de los banners revisados. Finalmente, un 70,1 % de las IPs del grupo anterior se encontraban asociadas a programas obsoletos.

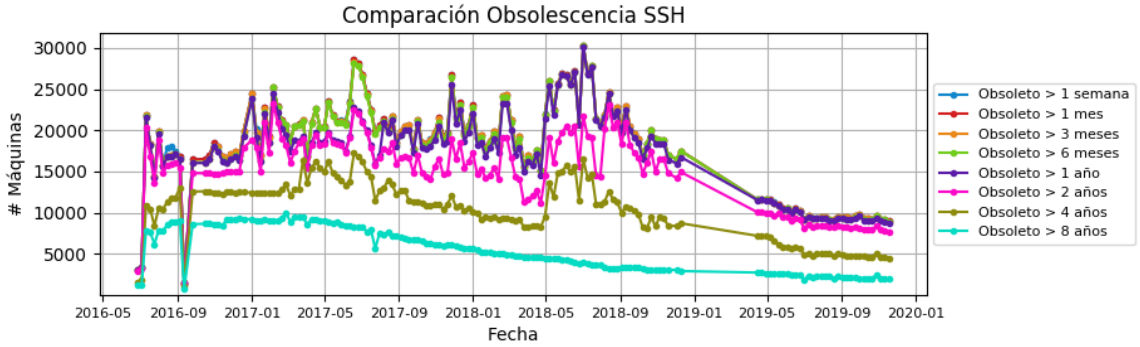
Para el caso de FTP, los programas elegidos corresponden a un total de 26,2 % del total de servidores FTP activos. De este conjunto, un 49,9 % de los resultados tiene versión asignada, y un 23,1 % del conjunto anterior se considera obsoleta. Esto puede deberse en parte a la poca actualización de los programas de FTP revisados.



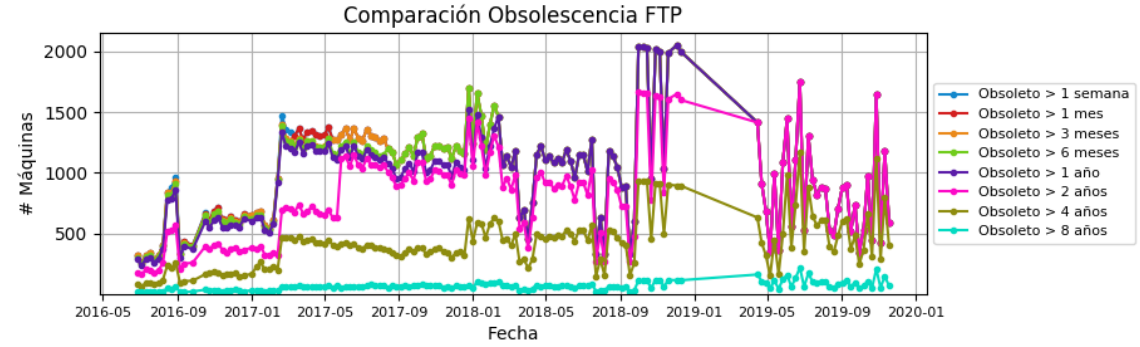
(a) Software Web Obsoleto



(b) Software de E-Mail Obsoleto



(c) Software SSH Obsoleto



(d) Software FTP Obsoleto

Figura 4.8: Comparación entre cantidad de máquinas encontradas frente a distintos umbrales de obsolescencia, aplicados a los protocolos estudiados.

Datos históricos

Hay que tener en consideración que las métricas de la tabla 4.4 indican la cantidad de IPs que **alguna vez se encontraron usando software obsoleto.**, lo cual no aclara la situación semana a semana con respecto al número total de dispositivos encontrados. Por lo tanto, para entender mejor el comportamiento de software obsoleto en los escaneos, es necesario comparar esta información con la cantidad de software activo, lo cual se puede ver en el conjunto de gráficos del grupo de figuras 4.9. Las figuras en este grupo muestran un histograma con la cantidad de servidores con software obsoleto para cada uno de estos tipos en el intervalo completo de los escaneos realizados, comparados con las otras métricas vistas en la tabla a lo largo del tiempo.

Con respecto al caso web mostrado en 4.9a, En primer lugar se observa que, a pesar de contener IPs de tres puertos distintos, la forma y las cantidades de IPs de la curva de activos web es muy similar a la curva del puerto 80 del CLCERT visible en 4.2d. La cantidad de IPs con nombre de servicio suele ser generalmente un tercio de éstas mientras que la cantidad de IPs con versión y la cantidad de IPs obsoletas tienen valores relativamente similares, en especial desde septiembre de 2018, siendo aproximadamente un quinto del total de activos.

Con respecto a 4.9b, se observa un comportamiento del conjunto total de servicios activos bastante similar al puerto 25, visto en 4.2c. En este caso, se puede observar cómo una mayor cantidad de software logró ser clasificado en un nombre de los servicios estudiados. En el caso del número de IPs relacionadas a software obsoleto, se observa que ésta está mucho más separada de la cantidad de software con versión identificada, lo que se debe en parte a que los servicios de E-Mail se actualizan con más frecuencia, tal como se vio en el gráfico 4.8b.

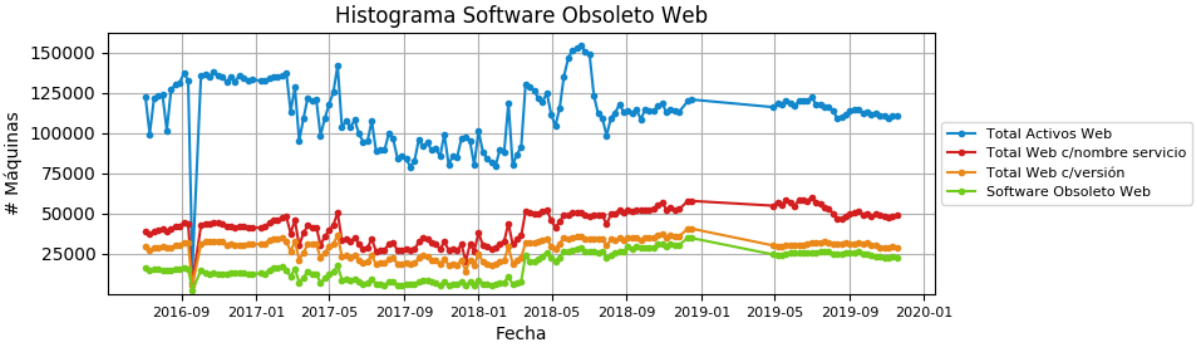
El gráfico 4.9c muestra nuevamente que a una gran cantidad de software se le logró determinar la versión. Al mismo tiempo, se muestra que entre un tercio y un quinto de los resultados en cada momento son considerados como IPs con software obsoleto. Los comportamientos de todas las curvas parecen ir al ritmo de la del total de máquinas activas, sin mostrar mucha información adicional.

Finalmente, el histograma de FTP se puede ver en la figura 4.9d. En esta figura llama la atención que aproximadamente en junio de 2017 aumentó considerablemente la cantidad de máquinas de las cuales se pudo determinar versión, para luego reducirse al número anterior en junio de 2018. Con respecto a la cantidad de software obsoleto detectado, esta se mantiene en valores bastante bajos, lo cual se debe tanto al umbral de obsolescencia (4 años) como a la baja cantidad de IPs con servicio FTP.

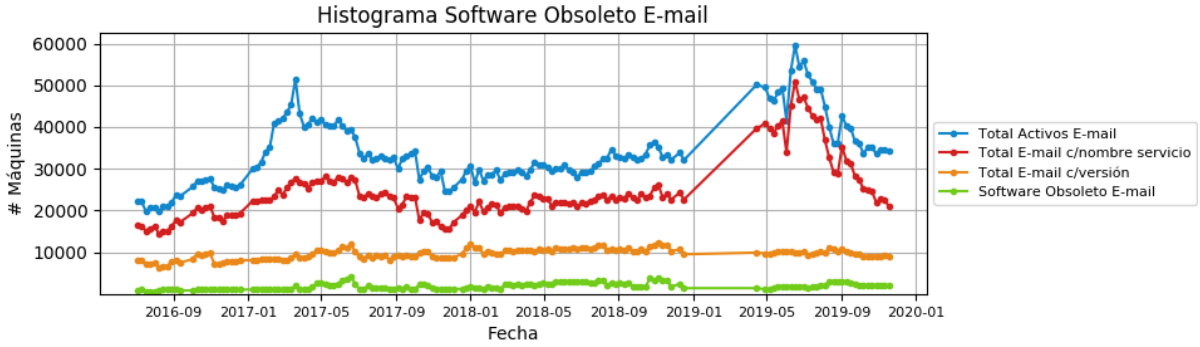
Comparación con malware

Al igual que en el caso de los certificados, los grupos de figuras 4.10, 4.11 y 4.12 agrupan las comparaciones de los conjuntos de IPs detectados por las fuentes externas como asociados a malware y los datos de máquinas activas y obsoletas determinados con información del CLCERT.

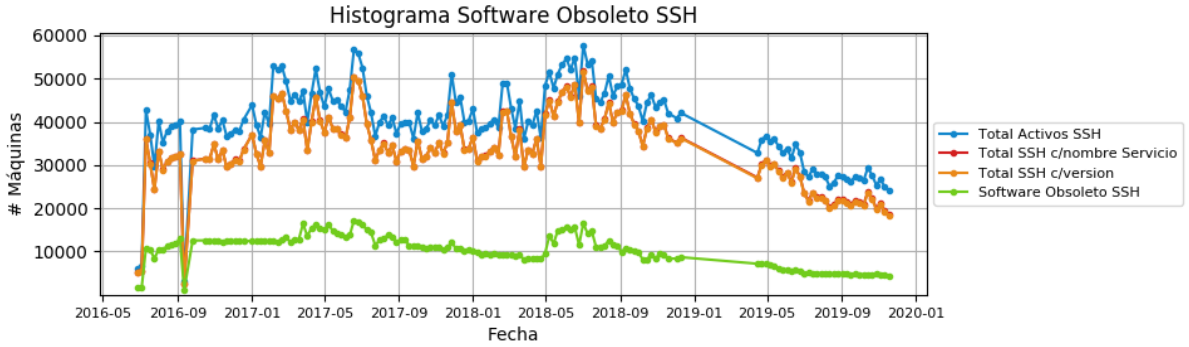
En el caso del grupo 4.10, el cual compara las categorías de software trabajadas con datos de malware de la categoría *bot*, se puede observar en la figura 4.10a que un número



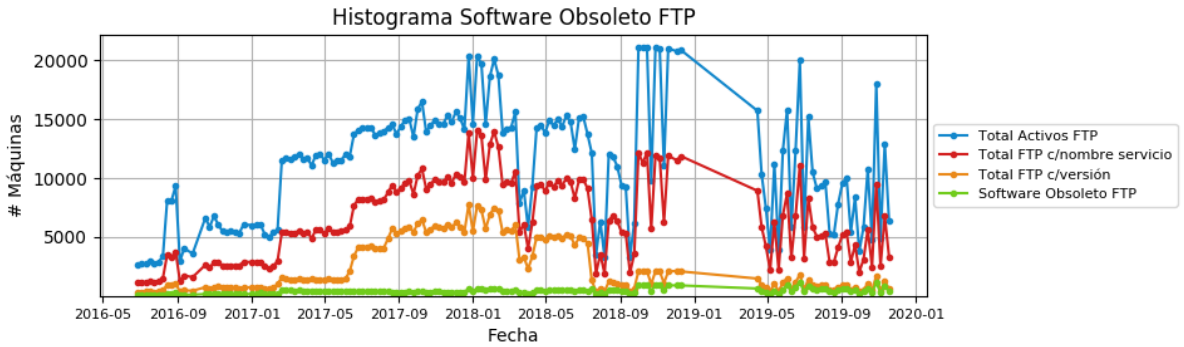
(a) Software Web Obsoleto



(b) Software de E-Mail Obsoleto



(c) Software SSH Obsoleto



(d) Software FTP Obsoleto

Figura 4.9: Comparación entre cantidad de máquinas con un software obsoleto del tipo correspondiente y el universo total de máquinas relacionadas al servicio.

de entre cien y dos mil máquinas del universo de IPs corriendo servicios SSH y Web están asociados a IPs que también se encontraron en ese periodo en la lista de malware de las fuentes externas. En el caso de los servicios FTP y SSH, este número es mucho menor, ubicándose en un intervalo entre 10 y 200. Proporcionalmente, la figura 4.10b muestra que esto equivale a entre el 0,25 % y 1,5 % del total de IPs de estas categorías, al menos en el caso FTP, SSH y Web. En el caso de servicios de E-mail visible en 4.10c, la cantidad de IPs relacionadas no supera el 0,25 %. En el caso de IPs obsoletas, los valores de intersección bajan considerablemente, ubicándose entre 10 y 200 en el caso de servicios Web y SSH, entre 1 y 50 en servicios de E-mail y entre 1 y 20 en servicios FTP. La baja cantidad de resultados explica la alta variabilidad porcentual de coincidencias mostrada en la figura 4.10d, en especial en los resultados FTP y E-mail. En el caso web, el porcentaje de coincidencias en grupos obsoletos suele mantenerse como menor que en el caso general, y en el caso SSH, éste varía entre menor y mayor según periodo, no mostrando patrón reconocible.

El grupo 4.11 se comporta de forma similar a la vista en el caso de certificados obsoletos al intersectarse con los datos de *fuerza bruta*. Por ejemplo, en la figura 4.11a se observa que los servicios de E-mail no superan las 15 coincidencias en todo el periodo, mientras que en los casos FTP, E-mail y SSH, no se superan las 50, 200 y 100 coincidencias respectivamente. Esto está condicionado por la baja cantidad de resultados totales de máquinas detectadas en la categoría *fuerza bruta*, que oscilan entre las 5 y 1000 coincidencias durante el periodo completo. Con respecto a la figura 4.11c, se observa que desde el inicio del periodo revisado hasta aproximadamente mayo de 2018, la cantidad de IPs detectadas en los casos web y SSH es bastante similar a las de la figura 4.11a, siendo varias veces menor en el futuro. En ambos casos, el porcentaje de IPs de malware en los subconjuntos de software obsoleto aumentó bastantes veces, pasando de valores cercanos al 0,02 % (visibles en la figura 4.11b) a valores cercanos al 0,1 % (visibles en la figura 4.11d) Con respecto a los casos FTP y SSH, no se pueden realizar conclusiones debido a que el número de IPs intersectadas en ambos casos no supera las 5 coincidencias.

Finalmente, en el caso del grupo 4.12, la cual muestra las intersecciones con la categoría de malware *darknet*, se puede observar en la figura 4.12a un mayor número de coincidencias, derivado de un mayor número de IPs en la categoría de malware. Los valores de IPs coincidentes en la categoría Web se encuentran entre los 300 y 1000 coincidencias, mientras que en el caso SSH se encuentran entre 100 y 300 coincidencias. El caso de IPs en la categoría FTP fluctúa entre las 20 y 150 IPs, mientras que en la categoría de correo electrónico no se superan las 20 coincidencias. La figura 4.12b muestra que los porcentajes de coincidencia en el caso general se ubican entre el 0,2 % y el 1 % del total de IPs en cada categoría para los casos SSH, FTP y Web. Sin embargo, en el caso E-mail, las coincidencias no superan el 0,1 % del conjunto completo. Con respecto a la intersección con el software obsoleto, visible en 4.10c, se nota de inmediato que no más de una decena de IPs coincidentes se encontró por jornada de revisión en los servicios de FTP y E-mail, por lo cual tampoco es posible realizar conclusiones categóricas sobre los resultados obtenidos. En el caso de servicio Web, el número de IPs obsoletas intersectadas con resultados del grupo *darknet* se mueve entre 20 y 100 coincidencias, mientras que en el caso de servicio SSH, este varía entre 10 y 100 casos. Al revisar los porcentajes de la figura 4.10d para los servicios Web y SSH, se observan situaciones similares a las vistas en el caso de la categoría *bots*, ocurriendo que los resultados de servicios Web son menores en la categoría de obsoletos que en la categoría general, y los

resultados de SSH varían entre valores mayores y menores en distintos periodos temporales.

A partir de lo anterior, se desprende que, debido a la poca cantidad de datos coincidentes en general, no es posible determinar dependencia o relación alguna entre los datos de los conjuntos de malware y los datos de abandono a través de software, más allá de la cantidad de resultados en proporción a la cantidad de malware detectado.

4.4.4. Conclusiones de Estrategia

Con respecto a los resultados obtenidos por el uso de certificados vencidos como medida de abandono, se puede observar que este subconjunto no presenta una significante mayor cantidad de coincidencias con los resultados de reportes de malware externos recibidos por el CLCERT. Esto puede deberse en parte a que si bien un certificado vencido es un riesgo para el usuario del servicio en términos de que la conexión encriptada puede no ser tan segura, su estado de vencido no significa directamente la creación de ninguna brecha de ingreso en el sistema en el cual se usa.

Sin embargo, la diferenciación de esta métrica frente a las estudiadas permite considerarla como un aporte en lo que es la selección de conjuntos de IP especiales para su revisión y análisis. Por ejemplo, el monitoreo constante de certificados vencidos se puede acompañar de notificaciones a los administradores de sistemas unos días antes de que esto ocurriese, de forma que tuviesen tiempo de actualizarlos, evitando generar daños de imágen y seguridad de datos al sitio manejado por el administrador contactado.

Por otro lado, con respecto al uso de versiones de software sobre los dispositivos de la internet chilena, se observa que la estrategia tal como está planteada requiere de la determinación certera de un umbral de obsolescencia, el cual difícilmente se puede tomar sin recurrir a decisiones arbitrarias que pueden afectar los resultados de intersección. No obstante, la métrica presenta resultados prometedores a través de cierta información que permite visualizar de forma más directa, como fechas en que se liberan nuevas versiones de software.

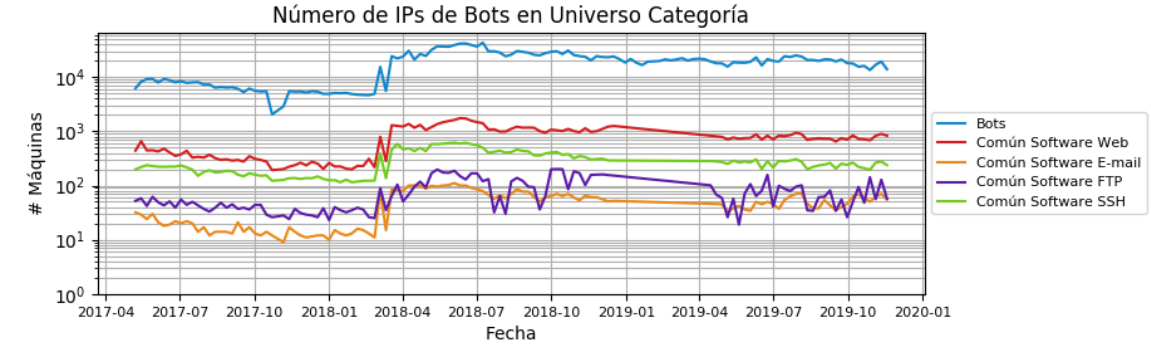
En resumen, ninguna de estas estrategias tiene potencial para identificar fuentes de vulnerabilidad sobre los datos reporte de malware similar a los ya existentes, pero sí pareciesen tener futuro como criterios utilizables para establecer máquinas prioritarias de revisión de escaneo, notificación y actualización.

4.5. Conclusiones

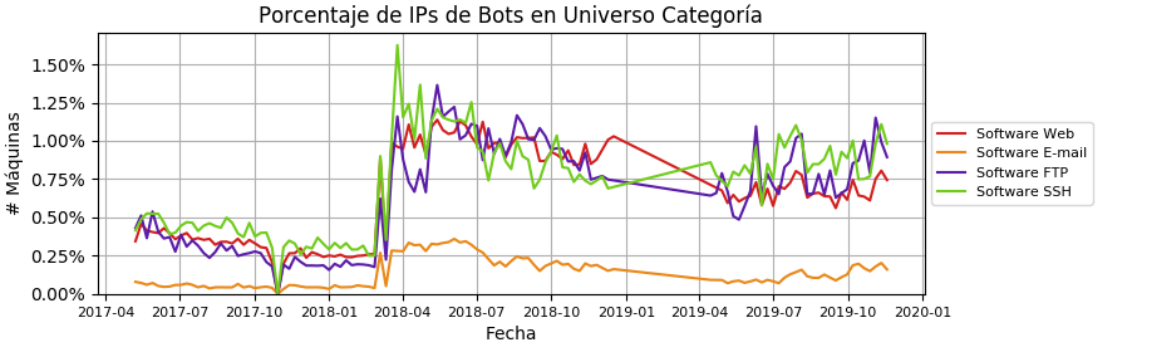
Para terminar este capítulo, se discutirán las limitaciones del trabajo realizado, se presentará trabajo futuro propuesto para continuar con esta línea de investigación y se detallarán las conclusiones generales obtenidas.

4.5.1. Limitaciones del trabajo

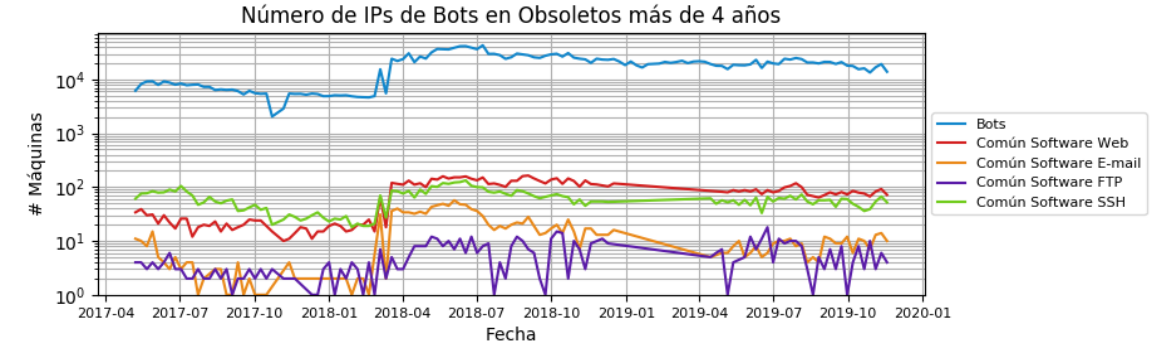
La mayoría de las limitaciones de este trabajo están dadas por la calidad de los datos manejados y la dificultad de replicar experimentos realizados por externos, junto con algunas recomendaciones para superarlas. A continuación se detallan ambas dificultades.



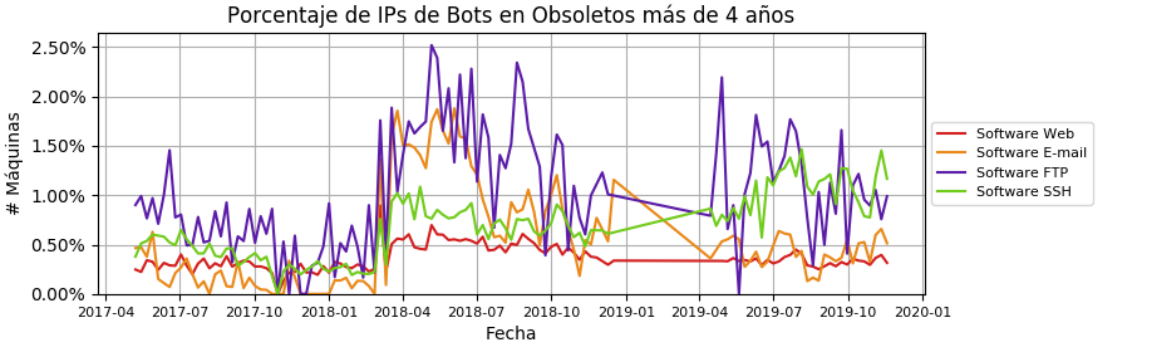
(a) Universo IPs por servicio, absoluto



(b) Universo IPs por servicio, porcentual

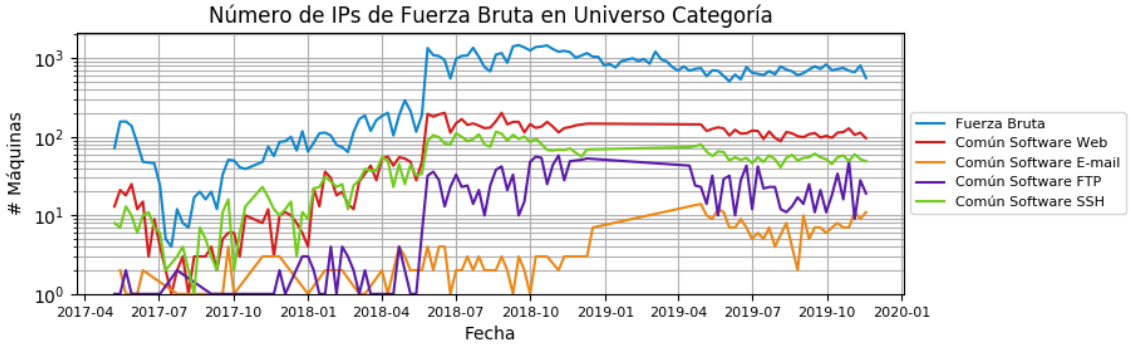


(c) IPs con software obsoleto, absoluto

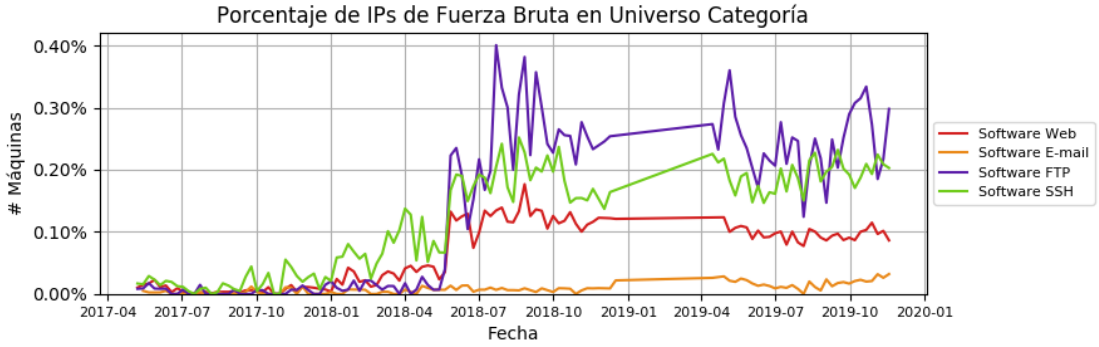


(d) IPs con software obsoleto, porcentual

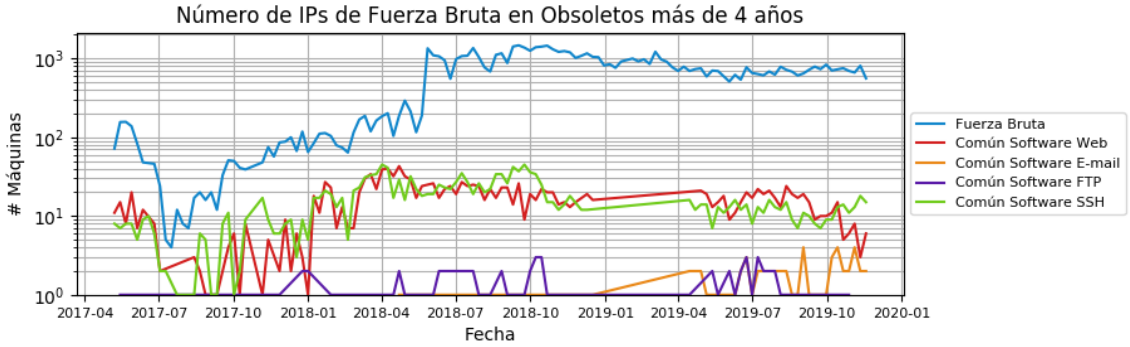
Figura 4.10: Comparaciones entre IPs de certificados (tanto vencidos como universo completo) y datos de malware de tipo bots.



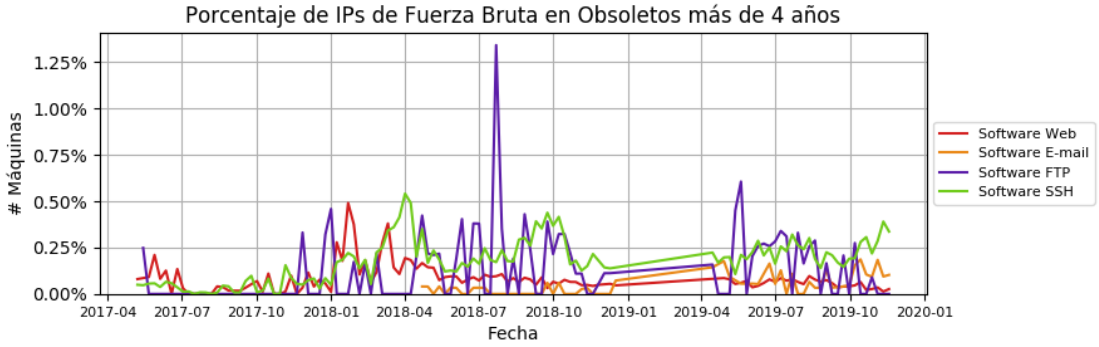
(a) Universo IPs por servicio, absoluto



(b) Universo IPs por servicio, porcentual

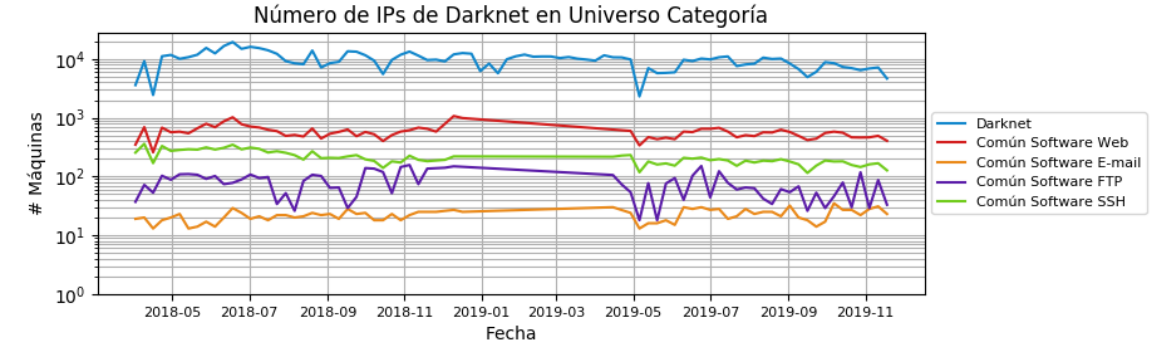


(c) IPs con software obsoleto, absoluto

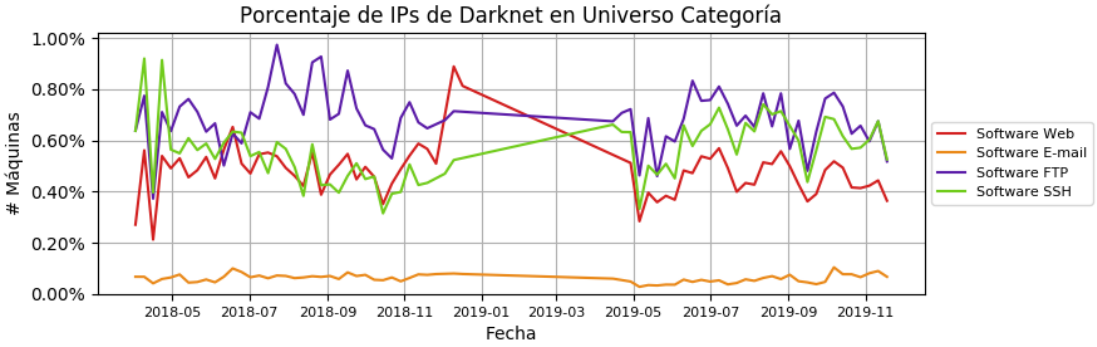


(d) IPs con software obsoleto, porcentual

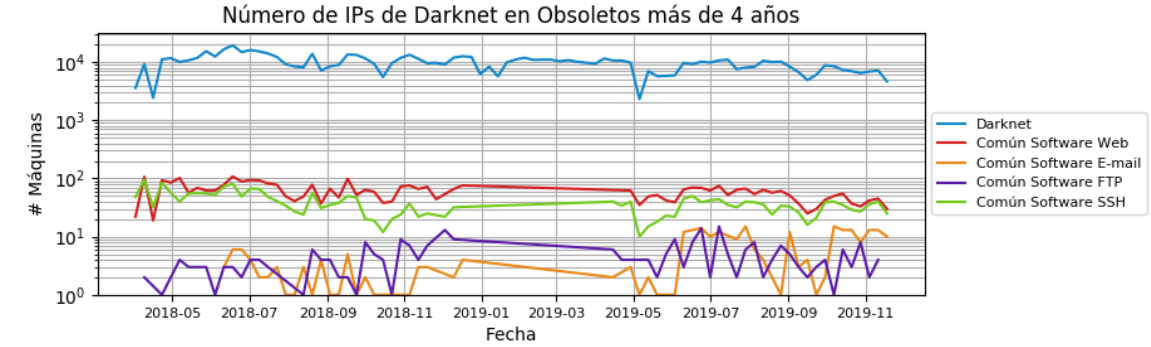
Figura 4.11: Comparaciones entre IPs de certificados (tanto vencidos como universo completo) y datos de malware de tipo bruteforce.



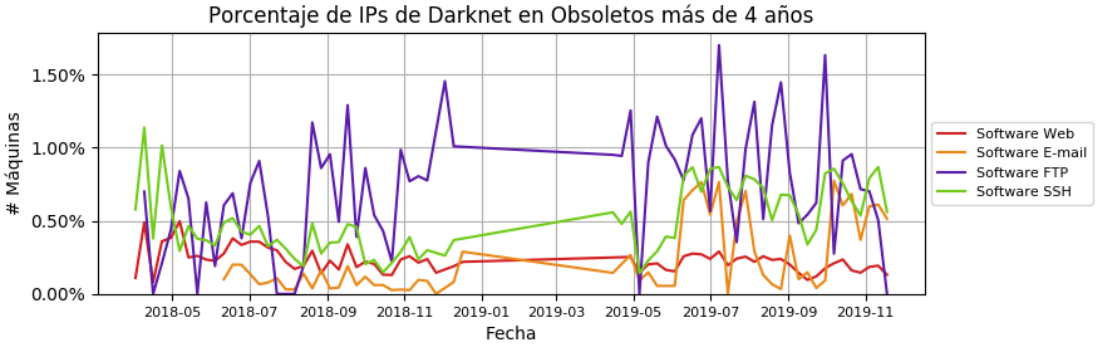
(a) Universo IPs por servicio, absoluto



(b) Universo IPs por servicio, porcentual



(c) IPs con software obsoleto, absoluto



(d) IPs con software obsoleto, porcentual

Figura 4.12: Comparaciones entre IPs de certificados (tanto vencidos como universo completo) y datos de malware de tipo darknet.

Fuentes de datos con problemas de continuidad

Como ya se explicó varias veces, ambas fuentes de datos poseen, en algunos casos, lagunas de tiempo sin información. Esto dificulta la comparación de las fuentes tanto consigo mismas como con otras en esos periodos debido a irregularidades que aparecen al comparar estos datos. Mejorar la resiliencia de las herramientas usadas para realizar estos escaneos puede ayudar a evitar estos problemas, y es en parte lo que busca el capítulo 5 de este trabajo.

Falta de transparencia en información de cómo se hacen escaneos externos

Una réplica fidedigna de escaneos realizados por entes externos para determinar razones que llevan a la diferencia de resultados no es posible sin la colaboración de estas organizaciones, por lo que es necesario aumentar la coordinación entre grupos que realizan escaneos de red de forma regular, desde distintas partes del mundo.

Una IP no siempre es una sola máquina

Al igual que en la estrategia del capítulo 3, estas estrategias tienen la limitante de que consideran que una IP representa unívocamente una máquina. En este caso, el riesgo de equivocarse con ese supuesto es mucho mayor, debido a que a lo largo del tiempo, ISPs suelen hacer reasignaciones de sus IPs entre sus propias máquinas, lo que provoca parte del ruido obtenido por los resultados de escaneos.

Determinación clara de "obsolescencia"

Definir exactamente cuánto tiempo debe pasar para considerarse una máquina obsoleta es complicado, debido a los ritmos de actualización propios de cada subred y de cada tipo de software. Si bien en este trabajo se consideró que el uso del valor "4 años o más" es el mejor umbral, no se encontraron medios para justificar un valor de antiguedad en específico, por lo que es necesario trabajar en el desarrollo de una mejor definición de esta característica.

4.5.2. Trabajo Futuro

A continuación se presentan algunas ideas que se consideraron con bastante potencial para extender el trabajo, pero que no se alcanzaron a implementar debido a limitaciones de tiempo.

Extender los protocolos revisados según popularidad de protocolos activos

Organizaciones como Censys empezaron a considerar datos de escaneos de iteraciones anteriores para determinar, 4 veces al año, qué protocolos seguir revisando y cuáles ignorar, teniendo en cuenta su popularidad entre otros motivos. Se propone que la revisión continua de los datos de escaneo recopilados hace más fácil la decisión de qué protocolos nuevos escanear en el sistema durante el tiempo y cuáles dejar de revisar.

Comparar resultados de herramientas de captura de banners distintas

En estas estrategias, no se analizaron en profundidad los banners recopilados por los protocolos, más allá de la determinación del software y versión corridos por la máquina. Una

idea a futuro consiste en utilizar también información como respuestas obtenidas en estos banners, para determinar posibles errores de configuración detectables al momento de iniciar conversaciónes con los servidores.

Intentar usar datos de escaneo de puertos para otros fines

Se propone que, en caso que la calidad de los datos de escaneos de puertos mejore, es posible utilizar la información de puertos abiertos para determinar un nuevo grupo de máquinas potencialmente vulnerables. Esto se ejecuta bajo el supuesto de que una mayor cantidad de puertos abiertos implica una mayor cantidad de servicios con fallas potenciales, o incluso una mala configuración a nivel de *firewall* de la máquina.

Cruzar información de subgrupos más importantes con información de ASNs relacionadas

Se propone repetir algunas de las ideas presentadas en el capítulo 3, relacionadas a tomar gráficos similares a los de este capítulo, pero comparando Sistemas Autónomos en vez de IPs. Al mismo tiempo, se propone agrupar las IPs diferentes encontradas en cada escaneo en sus sistemas autónomos, de forma de reconocer patrones relacionados con la topología de la red en estos resultados.

Ponderar de forma no binaria datos de abandono

Una forma de mejorar las métricas de abandono presentadas en este trabajo es considerarlas como datos no binarios, es decir, lograr matizar con un puntaje distintos niveles de abandono, según la cantidad de signos de abandono presentados por las máquinas y la ponderación proporcional de abandono según la diferencia de tiempo que lo vuelve inseguro (por ejemplo, en el caso de los certificados, desde hace cuánto tiempo que está vencido; mientras que en el caso de las versiones, cuántas versiones atrasado está el software o cuánto tiempo ha pasado desde la obsolescencia).

4.5.3. Conclusiones

A estas alturas, no hay dudas de las complejidades involucradas en la realización de escaneos de protocolos consistentes y con información representativa del estado de la Internet chilena en un momento determinado. Sin embargo, las tres recomendaciones incluidas en esta estrategia entregan una mejor visualización de los resultados según las necesidades que se tengan sobre estos.

En caso de necesitar consistencia durante largos periodos de tiempo, se recomienda tomar una intersección de las máquinas recopiladas en hartos escaneos. Esta recomendación es más necesaria en protocolos en los cuales se observa alta rotación de IPs, entre los cuales se encuentran HTTP y HTTPS tanto en puertos de producción como desarrollo. Al mismo tiempo, es posible tomar mayor énfasis en las máquinas asociadas a IPs con mayor historial de presencia, dado que su mayor antiguedad puede estar relacionada a una mayor cantidad histórica de datos manejados y almacenados.

En el caso de necesitar consistencia entre distintos escaneos, se pueden recomendar dos

estrategias: una conservadora, equivalente a la intersección de las fuentes revisadas; y una integral, consistente en el uso de todas las IP de máquinas detectadas por los escaneos revisados. En los casos en que las diferencias entre número de IPs en estrategia conservadora e integral sean muy grandes, es posible que los escaneos estén sufriendo de problemas por ubicaciones geográficas, o incluso por implementaciones de software de escaneo en caso que este sea distinto.

Por último, en caso de las estrategias propuestas relacionadas con abandono de las máquinas, se observa que en el caso de la revisión de certificados vencidos, no existen correlaciones con los datos obtenidos por las fuentes externas estudiadas. Sin embargo, la existencia de estos grupos aislados de IPs en consideración especial puede ayudar a la disminución de cantidad de certificados que vencen inesperadamente en sitios web y servidores de correo de la internet chilena. Con respecto al uso de versiones de software, tampoco se encontraron correlaciones interesantes con los datos de malware manejados. Esto puede estar relacionado, entre otros motivos, al uso de un umbral arbitrario de obsolescencia o a la calidad de los datos reportados por fuentes externas. Por lo tanto, si bien la estrategia de clasificación se considera prometedora, se hace necesario ajustar ciertos parámetros que determinan si una máquina corre o no software obsoleto, o el uso de distintos niveles de grados de obsolescencia. Contando con una herramienta de advertencia de este estilo corriendo sobre la red chilena, se abre la posibilidad de automatizar avisos de parches de seguridad a todos los administradores de sistemas de un país.

En síntesis, se propone seguir estudiando y perfeccionando todas las propuestas presentadas en este capítulo, con el objetivo de poder integrarlas en un futuro con sistemas de escaneo automatizados a nivel nacional, fomentando una mejora en la seguridad de la infraestructura de internet de Chile.

Capítulo 5

OSR: Un Observatorio de Seguridad para la Red Chilena

Como se vio a lo largo del trabajo investigativo descrito en este documento, si bien el la recolección automática de datos de escaneos activos y pasivos entrega un primer acercamiento a mediciones de seguridad de la red a partir de su revisión aislada, a medida que se acumula más información y se realizan más procedimientos (tanto en tipo de escaneos como en fechas en que estos se realizan) es necesario contar con herramientas que faciliten su procesamiento, agregación y visualización, disminuyendo a su vez la posibilidad de errores en la toma de estas mediciones.

El objetivo de este capítulo es describir el trabajo de ingeniería realizado en la elaboración de la herramienta que permitió obtener los resultados ya presentados. Para ello, se inicia describiendo cómo el CLCERT de la Universidad de Chile realizaba antiguamente el manejo de datos de escaneo de la Red Chilena y los problemas enfrentados en este proceso. Posteriormente, se muestra el trabajo de diseño, desarrollo y uso de la nueva plataforma de código abierto que maneja el escaneo, la importación y el tratamiento de los datos de escaneos activos manejados. El capítulo añade la descripción de algunas mediciones de rendimiento generales, que permiten predecir el rendimiento del sistema a medida que se agreguen más datos en el futuro. Finalmente, se comentan los desafíos que presenta el desarrollo y uso de esta herramienta a futuro, además de mencionar algunas posibilidades de nuevas estrategias de escaneo que no están consideradas en este trabajo pero son factibles con esta plataforma.

5.1. El Sistema Actual

Como ya se ha descrito en capítulos anteriores, el CLCERT trabaja con datos de escaneos pasivos, escaneos activos y reportes de *malware*que provienen de medios propios y de organizaciones externas. Tanto los escaneos como las rutinas de recolección y publicación de datos se distribuyen en varias máquinas manejadas por el grupo de investigación. A continuación, se detallarán las máquinas utilizadas y los escaneos o procesos asignados a cada una de ellas.

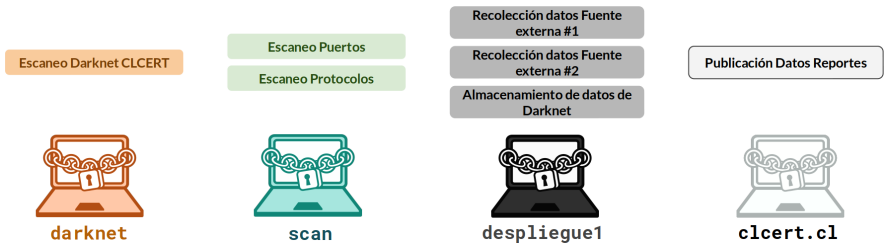


Figura 5.1: Representación gráfica de las máquinas manejadas por el CLCERT con relevancia en los procesos de escaneo de red, además de qué tareas ejecuta cada una de ellas.

5.1.1. Infraestructura y Procesos

Las máquinas virtuales del CLCERT operan en un servidor con procesador Intel Xeon ES-2670 de 8 núcleos, a 2.5 GHz. Cada una maneja una cantidad separada de RAM y de espacio de almacenamiento en disco. Dependiendo de la necesidad de cada máquina, algunas de ellas tienen asignada una IP pública, mientras que otras solamente tienen asignada una IP privada. Las máquinas usadas a la fecha de noviembre de 2019 aparecen descritas a continuación y se pueden ver representadas en la figura 5.1:

- `scan`: Máquina *CentOS 7.6* destinada a realizar los escaneos activos propios del CLCERT sobre puertos y protocolos. Posee 500 GB de espacio en disco para almacenar estos escaneos y 10 GB de RAM disponibles para su uso en los procesos de escaneo, además de una IP pública única para que los dispositivos que son escaneados puedan identificarla usando `nslookup`, encontrando una página con preguntas frecuentes e información para darse de baja de los escaneos.
- `despliegue1`: Máquina *CentOS 7.6* destinada a almacenar los escaneos, reportes antiguos y datos de la *Darknet* manejados por el CLCERT. Posee 1.5 TB de espacio en disco para almacenar esta información y 16 GB de RAM disponibles.
- `CLCERT.cl`: Máquina *CentOS 7.6* destinada a hospedar los sitios web del CLCERT, entre los cuales se encuentra un portal con datos agregados de algunos reportes recibidos. Posee 400 GB de espacio en disco para almacenar esta información y 16 GB de RAM disponibles, además de una IP pública necesaria para prestar el servicio de hospedaje de páginas web.
- `darknet`: Máquina *OpenBSD 4.7* utilizada para la realización de escaneos pasivos a través de una *Darknet*. Cuenta con dos tarjetas de red dedicadas. Una de ellas está asociada a una IP de la red interna y permite conectarse a ella sin interferir en los datos de escaneo. La otra recibe paquetes enviados a una subred de 256 IPs externas. Su única función es recibir los paquetes enviados al rango de IPs y guardarlos en un archivo *pcap*, el cual posteriormente es copiado a una máquina con más capacidad de almacenamiento, ya que esta posee solamente 80 GB de espacio en disco.

Cada una de las máquinas anteriores tiene una o más entradas en su *crontab*, encargadas de la ejecución algunos de los procesos de escaneo, procesamiento, recopilación o mantenimiento de los sistemas. Las entradas de *crontab* se encuentran distribuidas en varias cuentas de usuario, según la persona encargada del desarrollo o uso de la tarea. También existen procesos que se ejecutan manualmente en casos puntuales, como cuando alguna organización solicita datos relacionados con máquinas manejadas por ella. Ambas situaciones se representan en la

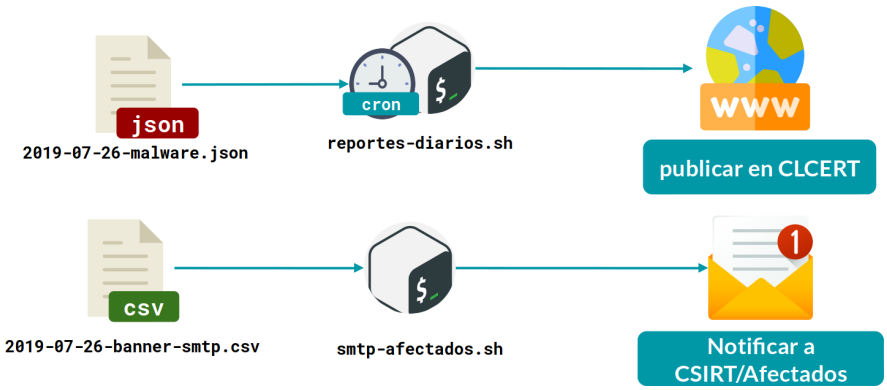


Figura 5.2: Representación gráfica de los dos tipos de procesos ejecutados por el CLCERT. Algunos se encuentran asociados a entradas en el `crontab` *de la máquina usada, por lo que se ejecutan regularmente. Otros, requieren de ejecución manual cuando se solicita.*

figura 5.2.

5.1.2. Problemas

La implementación actual del sistema de escaneo del CLCERT posee varios problemas, los cuales se detallan a continuación.

Monitoreo de Servidores y reporte de errores

El sistema actual no contempla un mecanismo para monitorear el estado de los servidores utilizados en escaneos. Lo anterior toma importancia en los casos en que los procesos pudiesen fallar debido a problemas de conectividad, memoria, almacenamiento o energía eléctrica. Si bien es posible establecer que cada proceso individualmente reporte de alguna forma su estado (por ejemplo, enviando un correo electrónico), es una solución ineficiente, ya que es un incentivo a la duplicación de código en cada proceso de escaneo y complicaría la actualización de la configuración de estos servicios de forma consistente (por ejemplo, en caso de necesitar actualizar el correo de notificación).

Revisión de escaneos habilitados

Debido a que los procesos se encuentran distribuidos en varias cuentas y varias máquinas, resulta complejo revisar qué tareas se están realizando y cuáles no, así como también qué procesos se encuentran ya implementados. Esta situación ha provocado que algunos procesos que se creían desactivados se siguiesen realizando en otra cuenta o máquina, generando problemas en la aplicación consistente de listas negras de escaneo. También es bastante común ver que algunas tareas queden programadas en cuentas de ex integrantes del proyecto de escaneo de red chilena, sin que nadie más se enterase de su existencia, o tareas que se implementan más de una vez por desconocimiento de qué tareas ya se encuentran implementadas y agendadas.

Sincronización de Datos

Existen procesos en distintas máquinas que utilizan los mismos archivos, los cuales además se suelen actualizar constantemente. Por ejemplo, la realización de escaneos requiere de una

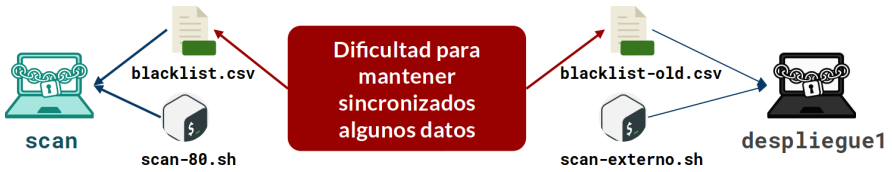


Figura 5.3: Representación gráfica del uso de los mismos archivos en más de un proceso en máquinas distintas. Los archivos no quedan automáticamente sincronizados, lo que ocasiona errores al momento de cruzar los datos, ya que estos usan fuentes distintas.

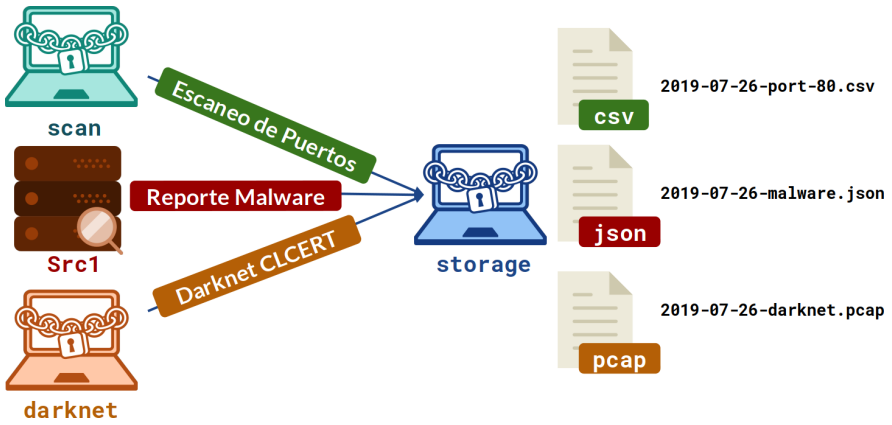


Figura 5.4: Representación gráfica de los archivos enviados por distintas fuentes, los cuales poseen distintos formatos

lista de IPs a escanear (en nuestro caso "la red chilena" según MaxMind) y una lista de IPs a evitar, las cuales deben mantenerse sincronizadas para asegurar el obtener resultados adecuados. Sin embargo, la forma actual en la que el sistema opera dificulta la sincronización de estos archivos, requiriéndose realizar copias manuales de estos entre máquinas, o editar los scripts *shell* de forma que usen una nueva versión de estos archivos. Esta situación se ve representada en la figura 5.3

Agregación de datos de escaneo

Tanto los reportes como los resultados de escaneos se suelen almacenar en archivos separados por fecha. En algunos casos, estos archivos guardan la información en formato de texto (`CSV` y `JSON`), pero en otros suele ser en formato binario (`PCAP`). En todos los casos, se dificulta enormemente la agregación y combinación de estos datos de forma directa y eficiente. Esta situación se puede ver gráficamente en la figura 5.4.

5.2. Diseño del nuevo sistema

Los problemas mencionados anteriormente limitan la capacidad del sistema actual de expandirse según los objetivos requeridos para este trabajo, en especial en la necesidad de mantener la ejecución de forma periódica confiable y resiliente los procesos desarrollados más adelante.

Por lo tanto, se hace necesario el diseño e implementación de un nuevo sistema que evite estas complicaciones y entregue herramientas que permitan la ejecución periódica de las

estrategias propuestas de forma simple y continua.

5.2.1. Requisitos

Antes de definir la estructura del sistema, es necesario especificar las tareas que se desean realizar en él. En esta sección se explicarán brevemente los objetivos que se desea que cumpla el sistema, teniendo en consideración además el poder corregir los problemas de la situación actual mencionados en la sección anterior.

Almacenamiento de datos

En primer lugar, el sistema debe ser capaz de almacenar como mínimo los datos recopilados en la situación actual. Además, es necesario guardar otros datos utilizados tanto en los escaneos como en el procesamiento posterior de la información, como por ejemplo, listas de IPs a escanear/evitar, nombres de los Sistemas Autónomos, y relaciones entre subredes y Sistemas Autónomos y/o países.

Otro requisito importante, no existente en el sistema actual, es poder guardar de forma ordenada la información procesada en el sistema, de forma de poder consultarla en el futuro o incluirla en otras tareas de procesamiento. También el sistema debe permitir extender fácilmente los tipos de datos almacenados, en caso de agregar nuevos tipos de escaneos o rutinas de procesamiento en el futuro.

Importación y exportación de datos

Los datos de escaneo pueden provenir de diversas fuentes, tales como páginas web, correo electrónico, otros servidores de la red interna o desde el mismo servidor luego de ejecutar un comando de escaneo. Además, estos datos se pueden guardar en diversos sistemas y formatos. Antes del desarrollo del sistema, la mayoría de los datos se almacenan en formato de texto CSV, JSON o binario (PCAP).

Entonces, una característica deseable para este sistema es el permitir el uso de datos de distintas fuentes y tipos de forma fácil, además de permitir el guardado de estos datos en distintos formatos con la misma simplicidad.

Procesamiento de datos

El sistema debiese permitir la realización de las consultas que se pueden ejecutar en la actualidad, además de consultas que permitan agregar estos datos con información adicional obtenida o según series de tiempo definidas. En específico y en relación a los escaneos, es necesario que el sistema permita agrupar los resultados en las siguientes categorías:

- IP/Subred
- ASN
- País (considerando los resguardos mencionados en los capítulos 1 y 2)
- Día
- Puerto (en caso de escaneo de puertos)

- Protocolo (en caso de escaneo de protocolos)
- Fuente (para comparar, distinguir y considerar de distinta forma las distintas fuentes manejadas)

Con el objetivo de aumentar la eficiencia y disminuir los requerimientos de memoria primaria, el procesamiento debiese realizarse en forma asíncrona, siempre que la tarea ejecutada lo permita.

Programación y ejecución simultánea de procesos

Dado que todos los procesos realizados por el CLCERT son periódicos, es necesario contar con un sistema que permita agendar la ejecución automática de estos procesos de forma simple. Para evitar el olvido de algunas rutinas programadas, se necesita disponer de un solo panel que muestre todos los escaneos agendados. De esta forma se evita olvidar la existencia de escaneos de los que ya no se sabe a futuro.

Otro requisito importante del sistema es que éste permita la ejecución simultánea de diferentes tipos de procesos, sin que esto genere una pérdida o corrupción de la información manejada.

Notificación de Eventos

Dado que los procesos pueden fallar en varios puntos de su ejecución, y que el problema de notar la caída de un proceso es recurrente en todos los procesos ejecutados, es necesario contar con información de salud de un proceso a medida que éste se va ejecutando, y en caso de detectar un cambio, enviar una alerta a los responsables. En el mismo contexto, notificaciones de finalización de procesos pueden ser útiles si se necesita contar con los datos procesados lo antes posible.

Resiliencia y Confiabilidad

El sistema debe ser capaz de soportar caídas en todas las etapas de su proceso, intentando que estos problemas en su ejecución repercutan lo menos posible en los procesos a ejecutar en el futuro. También es una característica deseable que la ejecución de los procesos del sistema sea idempotente, de modo que si es necesario repetir ejecuciones con los mismos parámetros, estas no generen más cambios que si se hubiesen ejecutado solo una vez de forma correcta.

Proyección a largo plazo

El sistema debiese de tener una proyección a largo plazo en su utilización, es decir, debe estar pensado para disminuir las posibles razones que lo lleven a dejar de usarse debido a complejidad, falta de documentación u obsolescencia.

5.2.2. Decisiones de Diseño

Con el objetivo de solucionar los problemas de la implementación actual y de cumplir los requisitos propuestos en la sección anterior, se tomaron las siguientes decisiones de diseño.

Sistema Centralizado, pero con posibilidad de usar distintas máquinas para cada proceso

Esta decisión consiste en que todos los procesos del sistema serán iniciados y programados desde una sola máquina, aunque su ejecución pueda realizarse en cualquier otra manejada por el CLCERT. La máquina principal (la cual llamaremos `OSR`) se debe encargar de proveer a las máquinas satélites de los archivos necesarios para la ejecución de sus tareas, así como también de publicar los resultados que se deseen publicar en las plataformas manejadas por el CLCERT. Esto permite centralizar el agendamiento de los procesos realizados, además de la revisión de errores en sus resultados, los cuales podrán ser capturados en esta máquina principal. Al mismo tiempo, esta decisión elimina el problema de sincronización de archivos comentado en secciones anteriores, asumiendo que en cada ejecución `OSR` entregará a la máquina satélite respectiva los últimos archivos de configuración existentes.

Desarrollo en lenguaje de programación compilado

Si bien lenguajes de programación como Python son más simples de utilizar y cuentan con una mayor cantidad de librerías debido a su popularidad, su rendimiento al procesar grandes volúmenes de datos no es el deseable para este tipo de proyectos. Se preferirá el uso de un lenguaje de programación compilado por sobre uno interpretado por esta razón.

Separación entre fuentes de datos, procesos y salidas

Se define como un requisito importante para el sistema la separación de las tareas de preparación de archivos de entrada, su procesamiento y su salida, buscando así facilitar la importación de datos desde distintas fuentes y su exportación en distintos medios. Para ello, se buscarán definir interfaces claras que permitan usar distintos componentes de entrada y salida de datos de forma intercambiable en cada proceso diseñado.

Uso eficiente de recursos usando *streams* de datos

Aparte del uso de un lenguaje de programación eficiente, es requerido que los algoritmos implementados en el sistema se diseñen pensando en un uso de recursos eficiente. Por ejemplo, para evitar que el sistema consuma más recursos de los necesarios a medida aumenta la cantidad de datos a procesar, una decisión de diseño tomada es privilegiar el uso de *streams* de datos de entrada y salida, provocando que el uso de recursos sea más o menos constante en el tiempo.

Almacenamiento de datos importantes en base de datos relacional

Se decidió el uso de bases de datos relacionales para el almacenamiento de datos de reportes y escaneos, debido a que esto permite realizar consultas sobre ellos de forma fácil y que requiere escaso aprendizaje extra. Al mismo tiempo, esto permite el cruce de información de distintas fuentes en distintos formatos, como por ejemplo, comparar metainformación de un archivo `·.pcap` con información de un archivo `.csv`, a través de la creación de tablas compatibles que guarden estos datos.

Sistema de notificaciones interno

La existencia de un sistema de notificaciones que informe del estado de las tareas ejecutadas a los responsables del CLCERT y la existencia de errores en ellas permite el uso de datos y corrección de errores de forma eficiente en tiempo.

Sistema de Código Abierto

Se decidió que la implementación de este sistema se pudiese publicar como código abierto en la plataforma *GitHub*, con el objetivo de fomentar a otras organizaciones a mantener sus propias bases de datos con información de escaneos realizados por ellas y a aportar al desarrollo de la herramienta.

5.2.3. Limitaciones

A continuación, se mencionan limitaciones derivadas tanto de las decisiones de diseño tomadas en la sección anterior como de otros problemas que escapan del alcance de este trabajo. Además, se mencionarán algunas posibles soluciones o consideraciones que se pueden tener en cuenta a futuro para contener el impacto de ellas.

Un único punto de fallo

Dado que el nuevo sistema considera el uso de una máquina central encargada de coordinar y ejecutar los procesos de importación y procesamiento de datos, la caída de esta misma máquina es incapaz de alertar este problema. Una posible solucion a considerar es disponer de otro servidor que revise periódicamente el estado de respuesta del servidor central.

Dependencia de funcionamiento de red interna de la Universidad

Dado que todas las máquinas utilizadas por el CLCERT están ubicadas en el mismo lugar y conectadas a Internet a través de un solo enlace del mismo proveedor, el sistema depende del funcionamiento de la red de la Facultad de Ciencias Físicas y Matemáticas para operar correctamente. Esta dependencia será mucho mayor contando con un sistema centralizado, ya que es necesario que la máquina `OSR` sea capaz de conectarse en todo momento a las máquinas satélites, requisito que no existe en la situación actual. Cabe destacar que mientras no existan los recursos como para contar con conectividad redundante de los servidores, no es mucho lo que se puede hacer para solucionar este problema.

Dificultad de aprendizaje de uso del sistema

Un nuevo sistema requiere capacitar a las personas que se encargarán de el en su forma de uso, más aún si el cambio en la forma de uso es demasiado grande. Para este sistema, se planea documentar las configuraciones y funcionalidades desarrolladas en la *Wiki* del proyecto de GitHub respectivo.

5.3. Implementación del Nuevo Sistema

Los detalles de implementación del nuevo sistema, consistentes en la explicación de los módulos que lo componen y su funcionamiento, se pueden encontrar en el Anexo B de este documento.

5.4. Funcionamiento del nuevo sistema

A continuación, se detallará el trabajo realizado con el objetivo de poner en marcha del nuevo sistema. Este trabajo consistió en la importación de datos antiguos y la creación de archivos de configuración para la recolección de datos futuros. Además, se tomaron algunas métricas de rendimiento sobre estas ejecuciones, con la intención de evaluar la carga adicional que ejerce el nuevo sistema.

5.4.1. Máquina Principal

Se destinó una máquina virtual distinta a las que procesan los datos y ejecutan los escaneos, la cual se denominará `OSR`. Esta máquina cuenta con 16 GB de memoria RAM y 1.5 TB de espacio de almacenamiento para guardar los datos del sistema.

En esta máquina se instaló *Go 1.12.7* y *Docker*, de forma de tener la base de datos *PostgreSQL 11* corriendo en un contenedor del sistema.

Además, en `OSR` se creó una cuenta especial llamada *osr*, la cual ejecutaría oficialmente todos los procesos del sistema desarrollado. En esta cuenta también se tendrían todos trabajos de *crontab* necesarios para el monitoreo de parte del CLCERT.

5.4.2. Archivos de Configuración

La migración al nuevo sistema requiere de la creación de archivos de configuración específicos para cada tarea realizada hasta este momento, además de archivos que permitan la realización de tareas futuras. En esta sección se detallarán las configuraciones fijadas para el uso del sistema en producción.

Configuración en tareas de importación iniciales

En esta categoría se tienen las configuraciones para importar datos históricos (almacenados directamente en alguna máquina del CLCERT), las cuales debiesen correrse solo una vez.

Ya que son procesos de importación, la salida por defecto de todos estos procesos es **base de datos**. Actualmente, se encuentran definidas las siguientes configuraciones de este tipo:

- **Reportes archivados de Fuentes Externas:** Debido a que las fuentes externas almacenan los datos reportados solo temporalmente, esta información se ha ido guardando en su formato original en la máquina `despliegue1`. Dado que lo único que cambia entre los datos archivados y los datos de la fuente directa es la ubicación de éstos, fue posible diseñar un solo proceso de importación, intercambiando en cada caso la configuración de entrada usada. En este caso, la configuración de entrada es de tipo

archivo remoto SFTP, la cual busca de manera recursiva en la carpeta en que todos los reportes son descargados los archivos terminados en `csv`.

- **Reportes archivados de la Darknet del CLCERT:** Al igual que en el caso de las fuentes externas, los escaneos realizados por la máquina `darknet` son almacenados en la máquina `despliegue1` a medida se van realizando. Por lo tanto, es necesario subir estos reportes al sistema una sola vez. La configuración usada es también **archivo remoto SFTP**, apuntando a la carpeta en que se encuentran estos archivos y buscando en forma recursiva los terminados en `.pcap` y `.gz` (reportes comprimidos).
- **Reportes archivados de puertos y protocolos abiertos:** Se cuenta con aproximadamente tres años de escaneos de puertos semanales, almacenados en la máquina *scan* y obtenidos con una entrada de tipo **archivo remoto SFTP**. Estos archivos poseen una ruta en el formato *[fecha_escaneo]/port-[puerto].csv* y *[fecha_escaneo]/grabber-port-[puerto].json*, la cual se usa para determinar la fecha y el puerto del reporte al insertarlo en la base de datos.

Configuración en tareas de importación recurrentes

Las siguientes tareas serán ejecutados de forma recurrente durante el funcionamiento del sistema. Al igual que en el caso anterior, las salidas por defecto de estos procesos es **base de datos**. Las configuraciones en esta categoría son:

- **Reportes de Fuentes Externas:** Ambas fuentes externas suben los nuevos datos a páginas web, listando los enlaces con los reportes en ellas. En ambos casos, se usa el mismo proceso respectivo para importar los datos, pero con una entrada de tipo **HTTP** recursiva con filtro por archivos que terminen en `csv`. Una de las fuentes externas requiere inicio de sesión a través del llenado de un formulario `POST`, mientras que la otra usa autenticación `BASIC`.
- **Datos de ubicación geográfica de MaxMind:** MaxMind provee de un enlace directo del cual se pueden descargar las bases de datos *Geolite2* en formato CSV. Estas bases de datos consisten en una lista de países con su código *GeoName* asociado y listas que asocian subredes con Sistemas Autónomos y Países. Estos archivos son importados por un proceso especial para este trabajo con una entrada de tipo **HTTP**.
- **Rankings de dominios según Alexa:** *Alexa* provee de un enlace de actualización diaria que contiene los datos de rankings generados por ellos. Estos datos son importados con un proceso especial para ello y una entrada de tipo **HTTP**.
- **Reportes de la Darknet del CLCERT:** Debido al gran peso de los archivos de la *Darknet* (alrededor de 1GB por hora comprimidos), se decide utilizar la entrada de tipo **comando remoto** para la ejecución de este archivo de importación. De esta forma, se puede dejar indefinidamente corriendo el proceso importador, y en caso de caerse, el sistema notificará de este problema.
- **Reportes de puertos y protocolos abiertos:** Se planea ejecutar directamente a través de una entrada de tipo **comando remoto** el escaneo de puertos abiertos. Estos resultados son luego importados por un proceso que toma una lista de IPs y les asigna un puerto obtenido de la configuración del archivo, mientras que la fecha asignada es la fecha en que se ejecuta el proceso.

Proceso	Rango (Aprox)	Filas (Aprox)	Peso (Aprox)	MB/Día
Ranking Alexa (solo Chile)	1 día	899	176 KB	0,18
Sistemas Autónomos[1]	Permanente	92.261	8456 KB	0
Reportes RRs	1 mes	2.925.864	698 MB	20
Darknet	1 mes	18.184.220	4387 MB	146
Escaneo de Puertos y protocolos[2]	3 años	300.313.760	141 GB	51,14
Metadatos de Certificados	3 años	18.645.320	9783 MB	8,93
Base de datos MaxMind	2 semanas	766.501	100 MB	14,29
Dominios	1 día	565.980	87 MB	0,24
Reportes (General)	1 año	25.046.640	9966 MB	27,30
	Total MB/Día			**268,08**

Tabla 5.1: Tabla que representa el uso de espacio en disco de los procesos de importación desarrollados. La columna MB/Día (y la fila total MB/Día) representa la cantidad de espacio diario extra que usa el sistema para guardar nuevos datos de cada tipo. Se omitieron algunos modelos debido a la poca relevancia de estos en el uso de espacio total del sistema.

5.4.3. Rendimiento

Al importar los datos iniciales del sistema, se tomaron medidas básicas de rendimiento para determinar el impacto de ellas en la máquina principal. Cabe destacar que el rendimiento completo de la aplicación dependerá también del tiempo que demore la ejecución de los escaneos. Sin embargo, esta sección busca mostrar solamente que el tiempo que agrega la ejecución del OSR al proceso actual es despreciable.

Memoria RAM

El uso de memoria RAM en todos los procesos fue constante durante su ejecución, ubicándose entre 10 y (en casos muy extremos) 40 MB, dependiendo del proceso. Un caso especial es el proceso *Darknet*, el cual soporta ejecutarse en varios *threads* debido al alto nivel de procesamiento que requiere. Este proceso ejecutado con 4 hilos paralelos consume en forma constante no más de 40 MB.

Paralelamente, el proceso de *PostgreSQL* consume una cantidad mayor de RAM (del orden de 2GB mientras se encuentra corriendo). Esto ocurre debido a que este proceso está configurado para aprovechar hasta 8GB de RAM del sistema, con el objetivo de mejorar la eficiencia de las consultas ejecutadas. Teniendo en cuenta que la máquina elegida para ejecutar el OSR cuenta con 16GB de RAM, y que los procesos de escaneo ocurren en máquinas diferentes, esto no debiese traer problemas.

Base de Datos y Almacenamiento en Disco actual

[1]La cantidad de sistemas autónomos nuevos cada día es tan baja que no genera carga considerable en el sistema.

[2]El peso mostrado considera la importación de datos de escaneo de Censys, los cuales se usaron solamente para este trabajo de Tesis y no se consideran como datos del sistema generalmente, por lo que el peso total del sistema funcionando debiese ser más bajo. Para más información, se recomienda consultar el Capítulo 4.

Proceso	Rango	Duración	Segundos/día	Comentarios
Ranking Alexa	1 día	7 segundos	7	Ranking actualizado diariamente.
Sist. Autónomos	1 semana	12 segundos	1,7	
Reportes RRs	1 mes	12 minutos	24	Rango debido al tiempo que demora realizar el escaneo en condiciones especificadas en cap. 3.
Darknet	1 hora	20 minutos	28.800	
Puertos, Protocolos y Certificados	1 día	2,23 minutos	133	
B. de D. MaxMind	2 semanas	70 segundos	10	
Dominios	1 día	1 segundo	1	
Fuente 1 Reportes	10 días	1 minuto	6	
Fuente 2 Reportes	4 meses	7 minutos	3,5	Realizado sobre 5 de los más de 30 reportes de esta fuente.
Total Segundos/Día			**28986,2** *8 horas aprox.*	[3]

Tabla 5.2: Tabla que representa la duración de los procesos de importación desarrollados. La columna Segundos/día representa la cantidad de minutos del día que se usan en la importación de un día de datos en promedio. Se omitieron algunos modelos debido a no estar asociados al ingreso masivo de datos de forma periódica.

La tabla 5.1 muestra el espacio utilizado por el sistema en cada tabla, además de la cantidad de líneas almacenadas en distintos periodos de tiempo para cada dato. Un dato interesante es el espacio usado promedio por día de ejecución de escaneos, el cual en estos momentos se encuentra en el orden de los 250 MB. Cabe destacar que cerca de la mitad de ese espacio es información aportada por una sola fuente (la *Darknet*).

Tiempo usado en importación

Las tareas de importación de datos de un día en específico tienen duraciones desde unos cuantos segundos hasta unas cuantas horas. La tabla 5.2 muestra duraciones promedio de estas tareas, el rango de tiempo que importan y la proporción de duración versus ese rango de tiempo en porcentaje. Al mismo tiempo, es importante notar que en este momento, el tiempo total usado en promedio al día en tareas de importación es de 8 horas aproximadamente, siendo la tarea que aporta más tiempo la que importa información de la *Darknet*. Sin considerar esa fuente, el tiempo total en procesar la información es de solamente 3 minutos.

Como se mencionó anteriormente, estos tiempos no tienen en consideración el tiempo que demora la realización de los escaneos. Sin embargo, cabe destacar que según la tesis de Eduardo Acha, el tiempo que demora *Mercury* en realizar las mediciones actuales permite su ejecución completa de forma diaria sin problemas.

[3] Es importante considerar que muchas de estas tareas pueden correr en paralelo sin afectar considerable-

5.5. Trabajo Futuro

Este trabajo contempla solo el inicio del software usado para operar el Observatorio de Seguridad de Red Chilena. A continuación, se mencionan algunas extensiones que pudiesen desarrollarse a futuro para este sistema.

5.5.1. Nuevos procesos

Los siguientes son procesos que requieren todavía ser implementados para la importación completa de los datos manejados por el CLCERT.

- **Payloads de Darknet:** Actualmente y como se mencionó en secciones anteriores, solamente se están importando datos de las cabeceras `TCP/IP` de los paquetes capturados por la *darknet*. Sin embargo, se pierde información importante al no considerar los datos contenidos en la capa de aplicación. Se hace necesario explorar los datos comunes recopilados en esta capa, de modo de crear un modelo acorde en la base de datos que permita guardar esta información.
- **Versiones de software IoT:** Actualmente, Mercury recopila información de versiones de software de equipos IoT, a través del análisis de los *headers* HTTP de las respuestas escaneadas. Un trabajo más elaborado a futuro puede requerir del uso de estas versiones de software para alertar en caso que queden obsoletas debido a la aparición de una vulnerabilidad.

5.5.2. Nuevas entradas y salidas

A medida se desarrollaba el sistema del Observatorio de Seguridad de la Red, surgieron algunas ideas de entradas y salidas cuya importancia no ameritaba su implementación de forma urgente, sin embargo, pueden facilitar la creación de nuevos tipos de procesos y escaneos. Estas ideas son las siguientes:

- **Adjuntos de Correos Electrónicos como entradas:** Algunos reportes externos suelen ser enviados como adjuntos a un correo específico. Es posible extender el sistema para conectarse vía el protocolo `IMAP` a este correo y descargar esos adjuntos.
- **Salidas que actúen como entradas de otros procesos:** El sistema actual no permite una ejecución de tareas como cadena de procesos, en las cuales la salida de un proceso es recibida directamente como entrada de otro. La implementación de una estructura que cumpla las interfaces de entrada y salida facilitaría este trabajo., facilitando la creación y ejecución de rutinas de procesamiento de escaneos más sofisticadas.
- **Salidas de archivos en otros formatos:** Actualmente, la implementación de salida como archivo remoto solo considera la exportación como archivo CSV. La exportación en otros formatos, tales como *JSON* o *XML*, entregaría mayor flexibilidad en la presentación de información al momento de ejecutar estos procesos.

mente su rendimiento.

5.5.3. Scheduler interno de tareas

Si bien se solucionó el problema de incertidumbre sobre procesos activos al centralizar el sistema de escaneos, requiriendo actualmente la revisión del *crontab* del servidor `OSR`, en este trabajo no se alcanzó a desarrollar un *scheduler* propio de OSR para agendar las tareas de importación y procesamiento. El contar con una utilidad propia facilitaría la programación y cancelación de tareas, aislándolas de otros procesos del sistema. Además, es necesario que sea posible implementar funcionalidades no existentes en `cron`, como ejecutar una tarea lo antes posible si es que la máquina estaba apagada en el momento en que le correspondía ejecutarla.

5.5.4. Sitio Web con Datos Agregados

El objetivo final del Observatorio de Seguridad de la Red es publicar reportes periódicos, consistentes en datos agregados a partir de la información manejada por el sistema. Se planea agendar una tarea periódica que exporte estos datos en formato CSV de forma periódica, con el objetivo que una página web dinámica cargue estos datos y los muestre de forma amigable.

5.6. Conclusiones

Este capítulo permitió comprender la complejidad del problema de procesar los datos manejados por el CLCERT y justificar el la necesidad de iniciar el desarrollo de una herramienta como la desarrollada en el contexto de esta tesis de investigación, en especial para asegurar el procesamiento periódico de la información. Al mismo tiempo, el trabajo a futuro muestra que para los objetivos del CLCERT, existe todavía mucho potencial en el desarrollo y extensión de esta herramienta, el cual da paso a la creación de otras estrategias de análisis y uso de los datos recopilados, sirviendo como punto de inicio para la extensión futura de las ideas propuestas en el trabajo investigativo presentado.

Conclusión

El desarrollo de este trabajo se produjo en un periodo de la historia de Chile en el que el concepto de ciberseguridad se hizo presente como nunca antes, tanto en términos legislativos, comerciales, sociales y técnicos. Eventos como *Hackeos* a bancos [77, 9], filtración de datos personales [51], phishing [19] y otras amenazas ponen en riesgo la seguridad digital y real de un gran número de habitantes de nuestro país.

Al mismo tiempo, en términos internacionales, amenazas como *botnets*, *ransomware* y *phishing* tienen un impacto cada vez mayor, lo cual ocurre debido entre otras razones a un aumento en las capacidades técnicas de los atacantes, los cuales cuentan con herramientas cada vez más sofisticadas y conocimientos cada vez más avanzados que les facilitan la realización de ataques sobre infraestructura digital importante. Al contrastar lo anterior con la poca preocupación general que hay sobre temas de seguridad en círculos en los que no se tiene mucha información sobre ella, las posibilidades de problemas mayores en el futuro frente a ataques informáticos son bastante altas.

El contexto anterior motivó la realización de este trabajo, en el sentido de la necesidad de contar con herramientas y estrategias al servicio de grupos de investigación y entidades que busquen proteger a los usuarios de una red determinada. Estas herramientas deben permitir detectar a tiempo anomalías varias en la red, de forma de poder reaccionar antes que estos ataques resulten ser efectivos. Si bien es verdad que nada impide que estas herramientas sean ocupadas por atacantes, mientras antes se trabaje en su uso, estandarización y masificación, será posible encontrarse en una situación de ventaja con respecto a no contar con ellas. Con la idea anterior en mente, este trabajo intenta avanzar a lo largo de sus capítulos con nuevas ideas y herramientas para el escaneo y monitoreo de subredes.

Con respecto al contenido desarrollado por este trabajo, el primer capítulo resumió el estado actual de investigaciones, organizaciones y herramientas relacionadas con el monitoreo y escaneo de Internet, enfocándose en aquellas que tienen utilidad desde la perspectiva de la seguridad informática según las estrategias desarrolladas. Además, se inició la discusión acerca de qué se debiese entender como internet chilena, debido a las complicaciones que existen al intentar asignar espacio geográfico exacto a recursos conectados a internet. Lo anterior determina que muchas de las definiciones de esta red no puedan ser absolutas, y deban adaptarse al contexto en que se usan y a las necesidades propias de los conjuntos de datos a analizar.

El segundo capítulo del trabajo mostró con qué tipo de datos trabaja la organización en la cual se enmarcó este trabajo (el CLCERT de la Universidad de Chile), correspondiente a un

laboratorio enfocado en temas de Seguridad y Criptografía Aplicada. Además, actualizó los datos recopilados hasta el momento por las herramientas de escaneo de Acha, con el objetivo de entender mejor los cambios históricos de ellos y sus diferencias. Además, se explicitaron las dificultades relacionadas con el trabajo de estos datos, las cuales sirvieron como base y motivación de las estrategias detalladas en los capítulos posteriores.

El tercer capítulo presentó la primera categoría de estrategias de uso de datos de escaneo de red propuesta y analizada en este documento. Esta categoría considera el uso de datos de escaneo DNS de dominios de un TLD de un país en específico, con el objetivo de entender mejor las relaciones de dependencia de recursos utilizados por estos dominios y apreciar el verdadero tamaño de infraestructura de los servicios más importantes de esta red chilena. Los resultados obtenidos demostraron lo frágil que es esta infraestructura en términos de cantidad de dispositivos encargados de actuar como respaldo en caso de un fallo generalizado, así como también la dependencia de estos servicios en infraestructura informática extranjera. Esta estrategia también dejó mucho espacio para mejora, ya sea encontrando nuevos subconjuntos de dominios de interés para escanear, o proponiendo nuevas formas de analizar y ponderar la disponibilidad de los servicios escaneados.

El cuarto capítulo se encargó de desarrollar la estrategia de uso de datos de múltiples fuentes (en el sentido amplio de la palabra) para su comparación, determinando de esta forma nuevos conjuntos de valores que pueden ser utilizados con mayor confianza en caso de querer datos más o menos conservadores relacionados con el tamaño de otros tipos de red chilena. Además, esta sección actualizó algunos de los resultados mostrados en la tesis de Eduardo Acha, a partir de los datos recolectados y almacenados desde septiembre de 2016. La primera estrategia de múltiples fuentes presentada consiste en el uso de datos históricos de escaneos de protocolos para la determinación de rangos de IP más importantes a partir de longevidad en la tarea de prestar su servicio. Esta estrategia permitió determinar la cantidad de IPs de máquinas que variaban constantemente, y la cantidad que se mantenían estáticas. La segunda estrategia de este conjunto permite la comparación de dos fuentes distintas de escaneo de protocolos a partir del uso de una tercera fuente, de forma de determinar si las diferencias de estos escaneos se deben a la diferencia en implementaciones o a otros factores, como los geográficos o de bloqueo de IPs. Esta estrategia entregó información que hace pensar que el factor geográfico es mucho más importante de lo que parece, motivando a la realización de escaneos de red desde distintas partes del mundo. La última estrategia consideró la definición del concepto de"abandono" sobre máquinas escaneadas, basado en certificados SSL/TLS vencidos o versiones de software obsoletas, de manera de intersectar estas máquinas con las reportadas por algún tipo de malware o bot, intentando entender si el estado de abandono tal como se define tiene algún efecto en la aparición de cierto malware. Si bien la última estrategia entregó nuevas clasificaciones interesantes para subconjuntos de IP que pueden tener relevancia al determinar IPs más importantes, se determinó que no existen los antecedentes suficientes para establecer alguna correspondencia entre ambos conjuntos de IPs.

El quinto y último capítulo plantea el diseño, desarrollo y operación de un sistema resiliente para la importación, transformación y exportación de datos de escaneos activos y pasivos. Esta herramienta busca ser una pieza fundamental en el sistema de monitoreo del CLCERT, controlando qué, cuándo y cómo se escanea, y avisando en caso de cualquier problema interno

de la máquina. Luego de hacerse pruebas de rendimiento y consumo de memoria a la ejecución en alta carga de esta herramienta, se obtuvo como resultado un bajo consumo de memoria y la habilidad de aprovecharse de procesadores de múltiples núcleos para aumentar efectividad.

Como conclusiones generales relacionadas al desarrollo de la tesis, se considera que se cumplieron adecuadamente los objetivos planteados como propuesta al inicio del trabajo, entregándose al autor una oportunidad única de combinar conocimientos de ingeniería e investigación para desarrollar un sistema a medida, el cual permitiese facilitar la recopilación y procesamiento de los datos que posteriormente se iban a analizar como parte investigativa del trabajo.

Finalmente, es necesario enfatizar que queda mucho potencial trabajo futuro en el área en la que se engloba este trabajo. Sin embargo, el inminente desarrollo y masificación de IPv6 dificultará su ejecución debido al cambio de contexto en infraestructura de red mundial. Si bien muy probablemente estos avances no puedan heredarse de forma directa a IPv6 cuando se realice la migración masiva, se espera que de todas formas este trabajo sirva de inspiración para el desarrollo de las técnicas de escaneo de red del futuro, manteniendo el mismo objetivo de mejorar la seguridad en la red en busca de asegurar mayor seguridad para los habitantes de nuestro país.

Bibliografía

[1] Eduardo Acha. «Monitoreo Activo de seguridad sobre la Red Chilena». Tesis de mtría. Universidad de Chile, jul. de 2017. URL: http://repositorio.uchile.cl/bitstream/handle/2250/148367/Monitoreo-activo-de-seguridad-sobre-la-red-chilena.pdf.

[2] David Adrian y col. «Zippier ZMap: Internet-Wide Scanning at 10 Gbps.» En: *WOOT*. 2014.

[3] Hugo Salgado et al. *Observatorio del DNS Latinoamericano*. 2016. URL: http://observatoriolac.nic.cl/trimestral/.

[4] Haneen Al-Alami, Ali Hadi y Hussein Al-Bahadili. «Vulnerability scanning of IoT devices in Jordan using Shodan». En: *2017 2nd International Conference on the Applications of Information Technology in Developing Renewable Energy Processes and Systems (IT-DREPS)* (2017), págs. 1-6.

[5] Mark Allman. «Comments on DNS Robustness». En: *IMC*. 2018.

[6] The Internet Archive. *Wayback Machine*. 2019. URL: https://archive.org/web/web.php.

[7] Internet Corporation for Assignment of Names y Numbers (ICANN). *New gTLD Domains: About the program*. 2012. URL: https://newgtlds.icann.org/en/about/program.

[8] Michael Bailey y col. «Practical darknet measurement». En: *Information Sciences and Systems, 2006 40th Annual Conference on*. IEEE. 2006, págs. 1496-1501.

[9] Radio Bío Bío. *Banco Consorcio pierde US$2 millones tras ciberataque y enciende alarmas en la industria financiera*. 2018. URL: https://www.biobiochile.cl/noticias/economia/negocios-y-empresas/2018/11/09/banco-consorcio-pierde-us2-millones-tras-ciberataque-y-enciende-alarmas-en-la-industria-financiera.shtml.

[10] Censys. *Censys Overview*. 2019. URL: https://censys.io/overview.

[11] CLCERT Universidad de Chile. *Al menos 2.933 dispositivos afectados por Bluekeep expuestos en la Red Chilena*. 2019. URL: https://www.clcert.cl/2019/05/31/bluekeep.html.

[12] Gobierno de Chile. *Instituciones de Gobierno*. 2019. URL: https://www.gob.cl/instituciones/.

[13] NIC Chile. *25 años de NIC Chile*. 2012. URL: https://www.nic.cl/acerca/memoria25/index.html.

[14] NIC Chile. *Página principal*. 2019. URL: https://www.nic.cl/.

[15] Cisco. *Cisco Umbrella 1 Million*. 2019. URL: https://umbrella.cisco.com/blog/2016/12/14/cisco-umbrella-1-million/.

[16] CLCERT. *Página Web CLCERT Universidad de Chile*. Ene. de 2019. URL: http://www.clcert.cl (visitado 01-01-2019).
[17] Cloudflare. *What is Anycast? How does Anycast Work?* 2019. URL: https://www.cloudflare.com/learning/cdn/glossary/anycast-network/.
[18] The MITRE Corporation. *CVE-2019-0708*. 2019. URL: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-0708.
[19] Alertas CSIRT. *CSIRT Gobierno de Chile*. 2019. URL: https://csirt.gob.cl/alertas/.
[20] Y. Cui y col. «Innovating Transport with QUIC: Design Approaches and Research Challenges». En: *IEEE Internet Computing* 21.2 (mar. de 2017), págs. 72-76. ISSN: 1089-7801. DOI: 10.1109/MIC.2017.44.
[21] Team Cymru. *IP to ASN service*. 2019. URL: https://www.team-cymru.com/IP-ASN-mapping.html.
[22] Team Cymru. *Team Cymru Homepage*. 2019. URL: https://www.team-cymru.com/index.html.
[23] Peter B. Danzig, Katia Obraczka y Anant Kumar. «An Analysis of Wide-Area Name Server Traffic: A Study of the Internet Domain Name System». En: *SIGCOMM*. 1992.
[24] Zakir Durumeric y col. «The matter of heartbleed». En: *Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM. 2014, págs. 475-488.
[25] Zakir Durumeric, Eric Wustrow y J Alex Halderman. «ZMap: Fast Internet-wide Scanning and Its Security Applications.» En: *USENIX Security Symposium*. Vol. 8. 2013, págs. 47-53.
[26] Zakir Durumeric y col. «A Search Engine Backed by Internet-Wide Scanning». En: *22nd ACM Conference on Computer and Communications Security*. Oct. de 2015.
[27] Zakir Durumeric y col. «Analysis of the HTTPS certificate ecosystem». En: *IMC '13*. 2013.
[28] Pawel Foremski, Oliver Gasser y Giovane C. M. Moura. «DNS Observatory: The Big Picture of the DNS». En: *IMC '19*. 2019.
[29] ShadowServer Foundation. *ShadowServer Homepage*. 2019. URL: http://shadowserver.org/.
[30] Google. *G Suite MX record values*. 2019. URL: https://support.google.com/a/answer/174125?hl=en.
[31] Robert Graham. *Masscan*. 2013. URL: https://github.com/robertdavidgraham/masscan (visitado 19-06-2018).
[32] The PostgreSQL GLobal Development Group. *PostgreSQL: The world's Most Advanced Open Source Relational Database*. 2019. URL: https://www.postgresql.org/.
[33] Alexa Internet. *Alexa Ranking*. 2019. URL: https://www.alexa.com.
[34] IP2Location. *IP2Location Edition Comparison*. 2019. URL: https://lite.ip2location.com/edition-comparison.
[35] IP2Location. *IP2Location Homepage*. 2019. URL: https://ip2locatioz.com/.
[36] Stratosphere Lab. *Stratosphere Intrusion Prevention Systems*. 2019. URL: https://www.stratosphereips.org/.
[37] Secitgo Limited. *crt.sh*. 2019. URL: https://crt.sh/.
[38] Gordon Fyodor Lyon. *Nmap*. 1997. URL: http://nmap.org (visitado 19-06-2018).
[39] MaxMind. *GeoLite2 Free Downloadable Databases*. 2019. URL: https://dev.maxmind.com/geoip/geoip2/geolite2/.

[40] Maxmind. *GeoIP2 city accuracy*. 2019. URL: https://www.maxmind.com/en/geoip2-city-accuracy-comparison?country=chile&resolution=250.
[41] Vladimir Mihailenco. *go-pg Library*. 2019. URL: https://github.com/go-pg/pg.
[42] Austin Murdock y col. «Target generation for internet-wide IPv6 scanning». En: *Proceedings of the 2017 Internet Measurement Conference*. ACM. 2017, págs. 242-253.
[43] RIPE NCC. *RIPE Atlas Project*. 2019. URL: https://atlas.ripe.net/.
[44] RIPE NCC. *RIPE NCC Stat API*. 2019. URL: https://stat.ripe.net.
[45] RIPE NCC. *RIPE NCC Stat Data Sources*. 2019. URL: https://stat.ripe.net/data-sources.
[46] Mozilla Developer Network. *Ajax*. 2019. URL: https://developer.mozilla.org/en-US/docs/Web/Guide/AJAX.
[47] Mozilla Developer Network. *Javascript*. 2019. URL: https://developer.mozilla.org/en-US/docs/Web/javascript.
[48] Mark Nottingham. *Identifying our Deliverables on IETF Mail Archive*. 2019. URL: https://mailarchive.ietf.org/arch/msg/quic/RLRs4nB1lwFCZ_7k0iuz0ZBa35s.
[49] Jamie O'Hare, Rich Macfarlane y Owen Lo. «Identifying Vulnerabilities Using Internet-Wide Scanning Data». En: *2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)* (2019), págs. 1-10.
[50] OpenINTEL. *OpenINTEL: Active DNS Measurement*. 2019. URL: https://www.openintel.nl/.
[51] BREACH: Data Leak Exposes Personal ID's of Over 14 Million Chilean Citizens. *WizCase*. 2019. URL: https://www.wizcase.com/blog/chile-leak-research/.
[52] Rapid7. *Project Sonar*. 2019. URL: https://www.rapid7.com/research/project-sonar/.
[53] P.V. Mockapetris. *Domain names - concepts and facilities*. RFC 1034 (Internet Standard). RFC. Updated by RFCs 1101, 1183, 1348, 1876, 1982, 2065, 2181, 2308, 2535, 4033, 4034, 4035, 4343, 4035, 4592, 5936, 8020, 8482. Fremont, CA, USA: RFC Editor, nov. de 1987. DOI: 10.17487/RFC1034. URL: https://www.rfc-editor.org/rfc/rfc1034.txt.
[54] P.V. Mockapetris. *Domain names - implementation and specification*. RFC 1035 (Internet Standard). RFC. Updated by RFCs 1101, 1183, 1348, 1876, 1982, 1995, 1996, 2065, 2136, 2181, 2137, 2308, 2535, 2673, 2845, 3425, 3658, 4033, 4034, 4035, 4343, 5936, 5966, 6604, 7766, 8482, 8490. Fremont, CA, USA: RFC Editor, nov. de 1987. DOI: 10.17487/RFC1035. URL: https://www.rfc-editor.org/rfc/rfc1035.txt.
[55] R. Braden (Ed.) *Requirements for Internet Hosts - Communication Layers*. RFC 1122 (Internet Standard). RFC. Updated by RFCs 1349, 4379, 5884, 6093, 6298, 6633, 6864, 8029. Fremont, CA, USA: RFC Editor, oct. de 1989. DOI: 10.17487/RFC1122. URL: https://www.rfc-editor.org/rfc/rfc1122.txt.
[56] M.T. Rose. *Post Office Protocol: Version 3*. RFC 1225 (Draft Standard). RFC. Obsoleted by RFC 1460. Fremont, CA, USA: RFC Editor, mayo de 1991. DOI: 10.17487/RFC1225. URL: https://www.rfc-editor.org/rfc/rfc1225.txt.
[57] M. Crispin. *Internet Message Access Protocol - Version 4*. RFC 1730 (Proposed Standard). RFC. Obsoleted by RFCs 2060, 2061. Fremont, CA, USA: RFC Editor, dic. de 1994. DOI: 10.17487/RFC1730. URL: https://www.rfc-editor.org/rfc/rfc1730.txt.
[58] J. Hawkinson y T. Bates. *Guidelines for creation, selection, and registration of an Autonomous System (AS)*. RFC 1930 (Best Current Practice). RFC. Updated by RFCs

6996, 7300. Fremont, CA, USA: RFC Editor, mar. de 1996. DOI: 10.17487/RFC1930. URL: https://www.rfc-editor.org/rfc/rfc1930.txt.

[59] R. Elz y col. *Selection and Operation of Secondary DNS Servers*. RFC 2182 (Best Current Practice). RFC. Fremont, CA, USA: RFC Editor, jul. de 1997. DOI: 10.17487/RFC2182. URL: https://www.rfc-editor.org/rfc/rfc2182.txt.

[60] K. Nichols y col. *Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers*. RFC 2474 (Proposed Standard). RFC. Updated by RFCs 3168, 3260, 8436. Fremont, CA, USA: RFC Editor, dic. de 1998. DOI: 10.17487/RFC2474. URL: https://www.rfc-editor.org/rfc/rfc2474.txt.

[61] R. Fielding y col. *Hypertext Transfer Protocol – HTTP/1.1*. RFC 2616 (Draft Standard). RFC. Obsoleted by RFCs 7230, 7231, 7232, 7233, 7234, 7235, updated by RFCs 2817, 5785, 6266, 6585. Fremont, CA, USA: RFC Editor, jun. de 1999. DOI: 10.17487/RFC2616. URL: https://www.rfc-editor.org/rfc/rfc2616.txt.

[62] J. Klensin (Ed.) *Simple Mail Transfer Protocol*. RFC 2821 (Proposed Standard). RFC. Obsoleted by RFC 5321, updated by RFC 5336. Fremont, CA, USA: RFC Editor, abr. de 2001. DOI: 10.17487/RFC2821. URL: https://www.rfc-editor.org/rfc/rfc2821.txt.

[63] K. Ramakrishnan, S. Floyd y D. Black. *The Addition of Explicit Congestion Notification (ECN) to IP*. RFC 3168 (Proposed Standard). RFC. Updated by RFCs 4301, 6040, 8311. Fremont, CA, USA: RFC Editor, sep. de 2001. DOI: 10.17487/RFC3168. URL: https://www.rfc-editor.org/rfc/rfc3168.txt.

[64] Y. Rekhter (Ed.), T. Li (Ed.) y S. Hares (Ed.) *A Border Gateway Protocol 4 (BGP-4)*. RFC 4271 (Draft Standard). RFC. Updated by RFCs 6286, 6608, 6793, 7606, 7607, 7705, 8212. Fremont, CA, USA: RFC Editor, ene. de 2006. DOI: 10.17487/RFC4271. URL: https://www.rfc-editor.org/rfc/rfc4271.txt.

[65] J. Klensin. *Simple Mail Transfer Protocol*. RFC 5321 (Draft Standard). RFC. Updated by RFC 7504. Fremont, CA, USA: RFC Editor, oct. de 2008. DOI: 10.17487/RFC5321. URL: https://www.rfc-editor.org/rfc/rfc5321.txt.

[66] A. Newton y S. Hollenbeck. *Registration Data Access Protocol (RDAP) Query Format*. RFC 7482 (Proposed Standard). RFC. Fremont, CA, USA: RFC Editor, mar. de 2015. DOI: 10.17487/RFC7482. URL: https://www.rfc-editor.org/rfc/rfc7482.txt.

[67] M. Belshe, R. Peon y M. Thomson (Ed.) *Hypertext Transfer Protocol Version 2 (HTTP/2)*. RFC 7540 (Proposed Standard). RFC. Fremont, CA, USA: RFC Editor, mayo de 2015. DOI: 10.17487/RFC7540. URL: https://www.rfc-editor.org/rfc/rfc7540.txt.

[68] J. Reschke. *The 'Basic' HTTP Authentication Scheme*. RFC 7617 (Proposed Standard). RFC. Fremont, CA, USA: RFC Editor, sep. de 2015. DOI: 10.17487/RFC7617. URL: https://www.rfc-editor.org/rfc/rfc7617.txt.

[69] J. Postel. *User Datagram Protocol*. RFC 768 (Internet Standard). RFC. Fremont, CA, USA: RFC Editor, ago. de 1980. DOI: 10.17487/RFC0768. URL: https://www.rfc-editor.org/rfc/rfc768.txt.

[70] J. Postel. *Internet Protocol*. RFC 791 (Internet Standard). RFC. Updated by RFCs 1349, 2474, 6864. Fremont, CA, USA: RFC Editor, sep. de 1981. DOI: 10.17487/RFC0791. URL: https://www.rfc-editor.org/rfc/rfc791.txt.

[71] J. Postel. *Internet Control Message Protocol*. RFC 792 (Internet Standard). RFC. Updated by RFCs 950, 4884, 6633, 6918. Fremont, CA, USA: RFC Editor, sep. de 1981. DOI: 10.17487/RFC0792. URL: https://www.rfc-editor.org/rfc/rfc792.txt.

[72] J. Postel. *Transmission Control Protocol*. RFC 793 (Internet Standard). RFC. Updated by RFCs 1122, 3168, 6093, 6528. Fremont, CA, USA: RFC Editor, sep. de 1981. DOI: 10.17487/RFC0793. URL: https://www.rfc-editor.org/rfc/rfc793.txt.

[73] J. Postel. *Simple Mail Transfer Protocol*. RFC 821 (Internet Standard). RFC. Obsoleted by RFC 2821. Fremont, CA, USA: RFC Editor, ago. de 1982. DOI: 10.17487/RFC0821. URL: https://www.rfc-editor.org/rfc/rfc821.txt.

[74] C. Partridge. *Mail routing and the domain system*. RFC 974 (Historic). RFC. Obsoleted by RFC 2821. Fremont, CA, USA: RFC Editor, ene. de 1986. DOI: 10.17487/RFC0974. URL: https://www.rfc-editor.org/rfc/rfc974.txt.

[75] Roland van Rijswijk-Deij y col. «A High-Performance, Scalable Infrastructure for Large-Scale Active DNS Measurements». En: *IEEE Journal on Selected Areas in Communications* 34 (2016), págs. 1877-1888.

[76] Sagar Samtani y col. «Identifying SCADA vulnerabilities using passive and active vulnerability assessment techniques». En: *2016 IEEE Conference on Intelligence and Security Informatics (ISI)* (2016), págs. 25-30.

[77] Bank Info Security. *Banco de Chile Loses $10 Million in SWIFT-Related Attack*. 2018. URL: https://www.bankinfosecurity.com/banco-de-chile-loses-10-million-in-swift-related-attack-a-11075.

[78] Yoichi Shinoda, Ko Ikai y Motomu Itoh. «Vulnerabilities of Passive Internet Threat Monitors». En: *USENIX Security Symposium*. 2005.

[79] Shodan. *Shodan Homepage*. 2019. URL: https://enterprise.shodan.io/platform.

[80] Tomas Sochor y Matej Zuzcak. «Study of internet threats and attack methods using honeypots and honeynets». En: *International Conference on Computer Networks*. Springer. 2014, págs. 118-127.

[81] Sorbs. *Spam and Open Relay Blocking System*. Oct. de 2019. URL: https://www.sorbs.net/ (visitado 18-06-2018).

[82] Spamhaus. *The Spamhaus project*. Oct. de 2019. URL: https://www.spamhaus.org/ (visitado 18-06-2018).

[83] International Organization of Standardization. *Country Codes - ISO 3166*. 2019. URL: https://www.iso.org/iso-3166-country-codes.html.

[84] Geonames Team. *Geonames Geographical Database*. 2012. URL: http://www.geonames.org/.

[85] Golang Team. *The Go Programming Language*. 2019. URL: https://golang.org/.

[86] The ZMap Team. *The ZMap Project*. 2019. URL: https://zmap.io.

[87] Geoff Huston Tony Bates Phillip Smith. *The CIDR Report*. 2019. URL: https://www.cidr-report.org/as2.0/.

[88] José Urzúa. «Estudio del estado del DNS en nombres de dominio .CL». En: *Jornadas Chilenas de la Computación*. 2005.

[89] Benjamin VanderSloot y col. «Towards a Complete View of the Certificate Ecosystem». En: *IMC '16*. 2016.

[90] Zoomeye. *Zoomeye - Cyberspace Search Engine*. 2019. URL: https://enterprise.shodan.io/platform.

Anexo A

Análisis de ccTLD sobre dominios de gobierno

A.1. Distribución de dominios gubernamentales

El gráfico A.1 nuestra el número de dominios gubernamentales con distintas cantidades de RRs, IPs, sistemas autónomos y países. Al comparar los valores de RR visibles en A.1a con los vistos en el caso general en el gráfico 3.3a, se observa que hay una cantidad relativa un poco mayor de dominios con solo un valor en el registro A en el caso gubernamental, la cual es compensada por una cantidad menor de dominios con 2 valores. El comportamiento en el caso del registro NS es bastante similar, mientras que en el caso del registro MX se observa una gran diferencia en cantidad de dominios de gobierno con 5 RRs o más, duplicando el valor del caso general. Más adelante se verá que esto se debe a una concentración especial del servicio de correo en ciertos proveedores externos.

Al comparar la cantidad de IPs distintas entre los gráficos 3.3b y A.1b, se observa que en el caso del RR de tipo MX de dominios gubernamentales, el 49,47 % de los dominios tienen 2 o más IPs distintas asociadas, mientras que en el caso general, esta condición es cumplida solo por el 28,15 % de los dominios. En el caso de los registros de tipo NS, se nota el mismo aumento de dominios con una sola IP asociada que se ve en el caso general, explicado por la misma razón que en el caso general. Además, al comparar el gráfico de NS en esta tabla con la tabla A.1a, es posible notar que se repiten algunos comportamientos del caso general, como que, por ejemplo, un número mucho mayor de dominios está asociado a una sola IP, con respecto a la cantidad de estos relacionado con un solo RR.

En el caso del registro NS y el número de sistemas autónomos, la situación de número de éstos es distinta en los dominios gubernamentales con respecto a la existente en el grupo de dominios general. El gráfico A.1c muestra que un 39,57 % de los dominios poseen dos o más IPs distintas, mientras que el gráfico 3.3c muestra que en en el caso general solo un 28,65 % de los dominios cumple esta propiedad. En el caso de los RRs de tipo MX y A, los valores son bastante similares entre el conjunto de dominios totales y el conjunto de dominios gubernamentales.

En el caso de número de países asociados a los dominios gubernamentales, el gráfico A.1d muestra que para el caso de los registros A y MX, la situación con respecto al caso general mostrado en 3.3d es bastante similar, mientras que en el caso del registro NS, se observa un nivel ligeramente mayor de concentración de dominios asociados a solo un país al comparar el 77,42 % de los dominios del caso general en esta situación con el 83,86 % de los dominios gubernamentales en la misma condición.

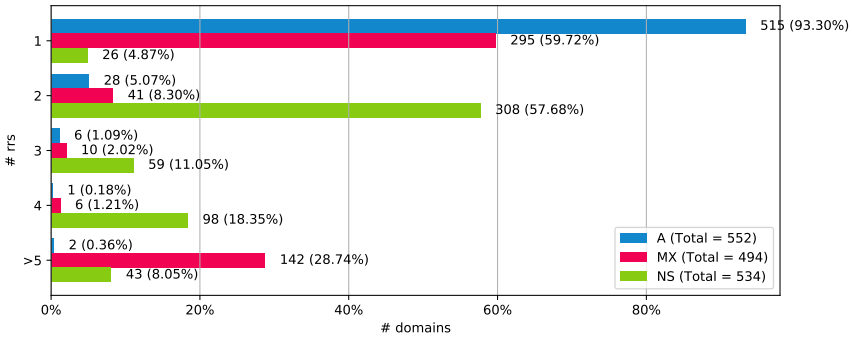
A.2. Concentración de dominios gubernamentales

Los gráficos agrupados en la figura A.2 permiten apreciar que la situación general tanto en el caso de IPs como sistemas autónomos y países varía considerablemente al compararla con la situación general de los dominios chilenos. Por ejemplo, en el caso de A.2a, se nota que la concentración en los registros de tipo A y NS no es tan alta como en el gráfico 3.4a, dado que se asemeja más a la del RR de tipo MX del último gráfico mencionado, lo que denota una mayor variedad de valores. Esto se comprueba al notar que, por ejemplo, para 552 dominios asociados a al menos un registro de tipo A escaneados, hay 465 IPs distintas encontradas en ellos, y que el comportamiento se repite en el caso de los otros tipos de RR. En el caso de MX, lo más llamativo es que el 24,41 % de los dominios es manejado por 2 tuplas de valores relacionadas con Gmail, las cuales en total suman 7 valores distintos.

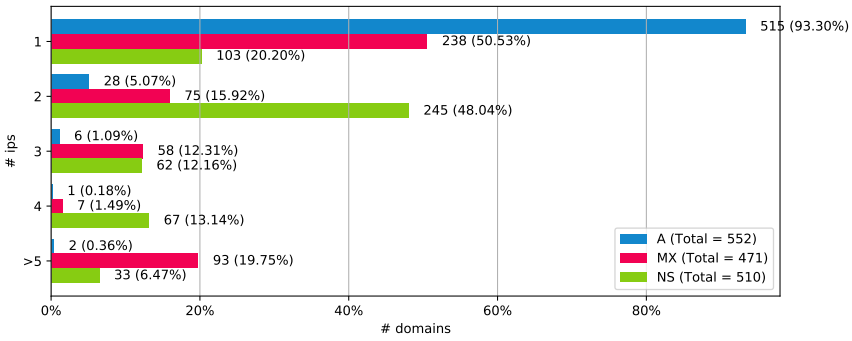
En el caso del gráfico A.2b, se observa un comportamiento bastante similar para el RR de tipo NS al del gráfico anterior. lo que se confirma con que la cantidad de tuplas de IPs distintas analizadas en este tipo de RR es muy parecida a la cantidad de tuplas de RR distintos analizados, por lo que la concentración no varía considerablemente. Sobre el comportamiento de las IPs asociadas al registro MX, se observa una menor variedad que la vista en el caso de los RRs para el mismo tipo de RR, la cual se debe a razones similares a las vistas en el caso general (dominios que apuntan a la misma IP).

El escalonamiento del gráfico A.2c hace notar que la cantidad de sistemas autónomos totales es bastante baja, a pesar de que es mayor proporcionalmente a la cantidad de valores de RR e IP distintos a la vista en el gráfico 3.4c. Sin embargo, el grupo de tablas A.1 muestra que los 5 primeros sistemas autónomos del ranking de concentración para cada tipo de RR acumulan una cantidad parecida del porcentaje de dominios al compararse con la tabla 3.3. El Sistema Autónomo más importante en los RR de tipo A y NS es “Ministerio del Interior y de Seguridad Pública - Gobierno de Chile”, presente en el 17,93 % y 10,83 % de los dominios respectivamente. En el caso del registro de correo electrónico, Google controla el 28,45 % de los dominios del grupo revisado. Las la mayoría de los sistemas autónomos mostrados en los 3 tipos de RR ya se han visto en el análisis general.

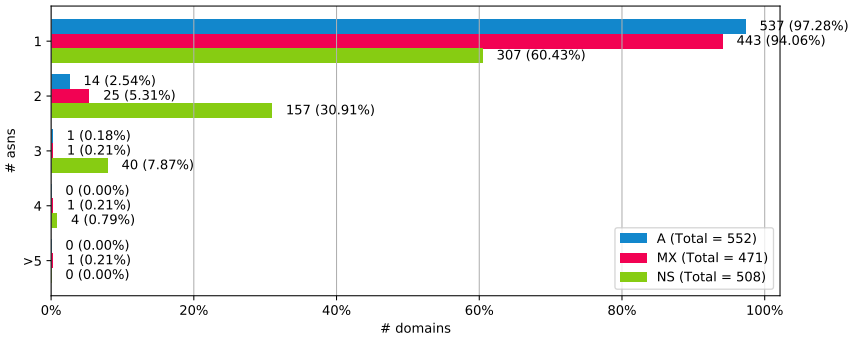
En el caso de los países asociados a dominios gubernamentales, la tabla A.2 muestra que Chile domina en los tres tipos de RR. En el caso del RR de tipo A, el 84,60 % de los dominios asociados a servicios web son servidos por IPs chilenas, mientras que en el caso del RR de tipo MX y NS el 54,14 % y 67,52 % respectivamente de los dominios cumplen esa condición. El único caso en que Chile no posee una posición tan dominante es en el RR de tipo MX, en el cual Estados Unidos concentra el 44,16 % de los dominios.



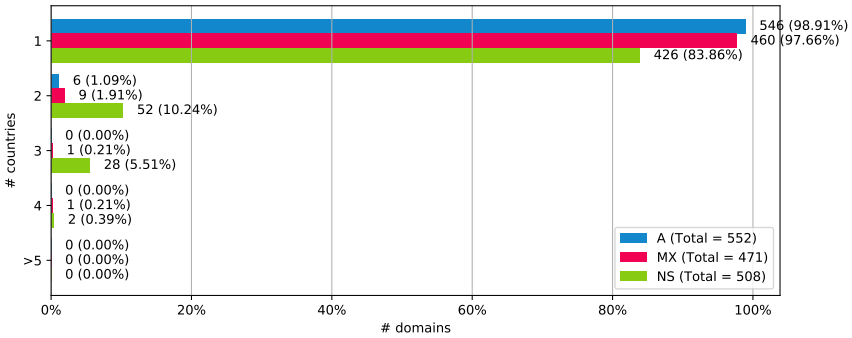
(a) Número de RRs distintos



(b) Número de IPs distintas

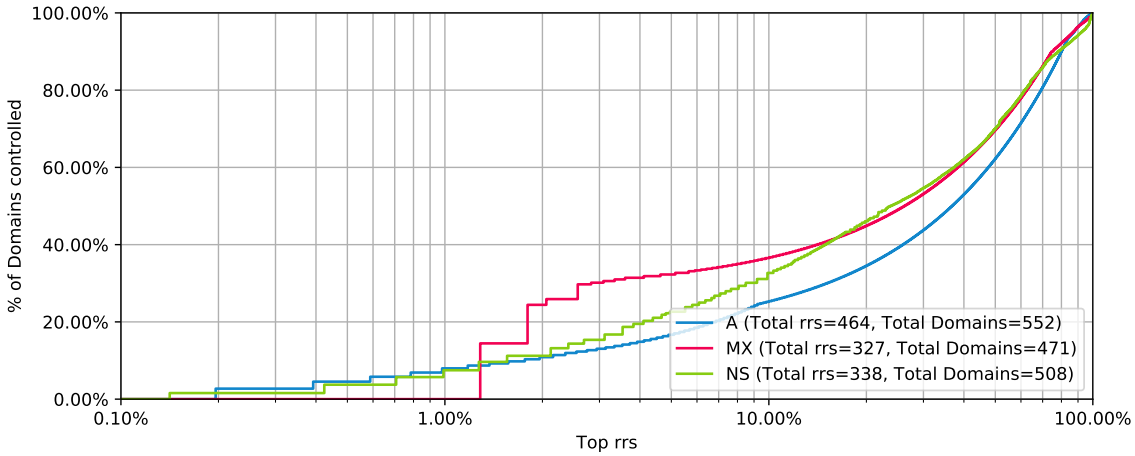


(c) Número de sistemas autónomos distintos

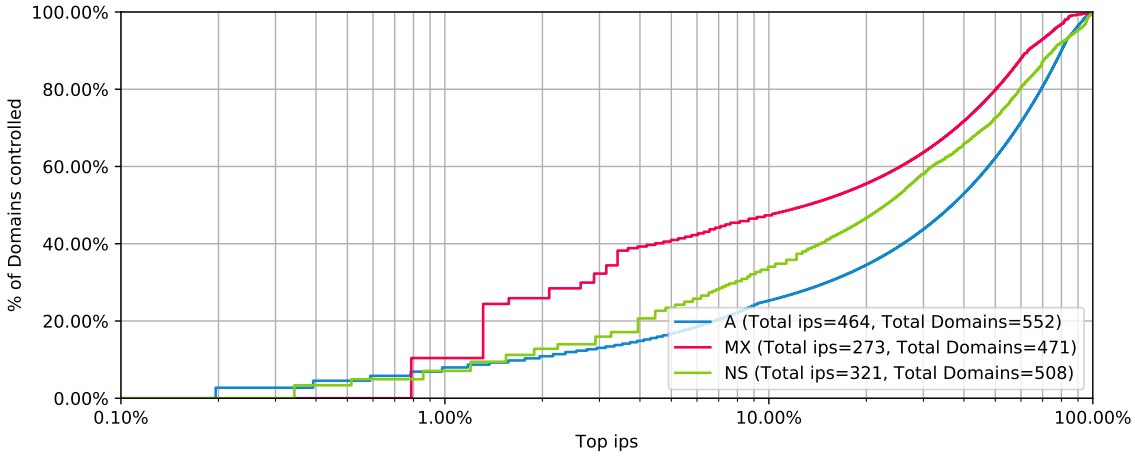


(d) Número de países distintos

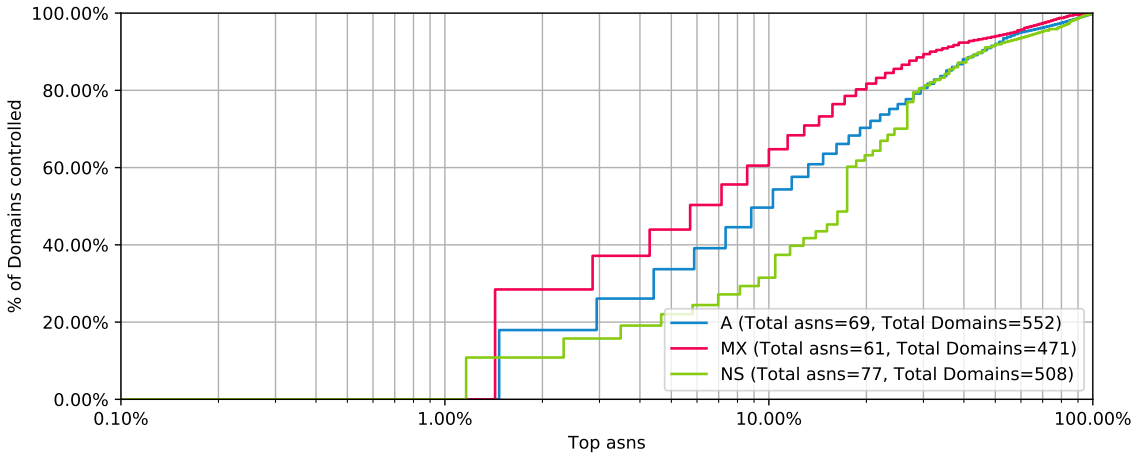
Figura A.1: Gráfico con la distribución de cantidad de RRs, IPs, sistemas autónomos y países distintos sobre los dominios gubernamentales escaneados con al menos un RR de tipo A, MX y NS.



(a) Concentración de RRs



(b) Concentración de IPs



(c) Concentración de sistemas autónomos

Figura A.2: Gráfico en escala logarítmica que muestra la cantidad de dominios gubernamentales que dejarían de funcionar (eje y) en el caso que cierta cantidad de IPs ordenadas según importancia (eje x) dejasen de funcionar.

N^o	**Tupla ASs**	**Tupla País**	**Dominios**	**%**	**% Acc.**
1	Ministerio del Interior y de Seguridad Publica - Gobierno de Chile	Chile	99	17,93	17,93
2	ENTEL CHILE S.A.	Chile	45	8,15	26,09
3	Gtd Internet S.A.	Chile	42	7,61	33,70
4	ZAM LTDA.	Chile	30	5,43	39,13
5	CTC. CORP S.A. (TELEFONICA EMPRESAS)	Chile	30	5,43	44,57

(a) RR de tipo A

N^o	**Tupla ASs**	**Tupla País**	**Dominios**	**%**	**% Acc.**
1	GOOGLE - Google LLC	Estados Unidos	134	28,45	28,45
2	MICROSOFT-CORP-MSN-AS-BLOCK - Microsoft Corporation	Estados Unidos	41	8,70	37,15
3	Gtd Internet S.A.	Chile	32	6,79	43,95
4	Ministerio del Interior y de Seguridad Publica - Gobierno de Chile	Chile	30	6,37	50,32
5	ZAM LTDA.	Chile	25	5,31	55,63

(b) RR de tipo MX

N^o	**Tupla ASs**	**Tupla País**	**Dominios**	**%**	**% Acc.**
1	Ministerio del Interior y de Seguridad Publica - Gobierno de Chile	Chile	55	10,83	10,83
2	Gtd Internet S.A.	Chile	25	4,92	15,75
3	Telmex Chile Internet S.A.	Chile	17	3,35	19,09
4	Servicios Informáticos Hostname Ltda	Chile	12	2,36	24,41
5	HOSTING.	Chile	14	2,76	27,17

(c) RR de tipo NS

Tabla A.1: Ranking de 5 sistemas autónomos con más dominios asociados a través de valor del RR a alguna IP perteneciente a ellos en dominios gubernamentales.

N	Tupla País	Dominios	%	% Acc.
1	Chile	467	84,60	84,60
2	Estados Unidos	68	12,32	96,92
3	Canadá	8	1,45	98,37
4	Argentina	4	0,72	99,09
5	Brasil	2	0,36	99,46

(a) RR de tipo A

N	Tupla País	Dominios	%	% Acc.
1	Chile	255	54,14	54,14
2	Estados Unidos	208	44,16	98,30
3	Canadá	2	0,42	98,73
4	Suiza	1	0,21	98,94
5	España	1	0,21	99,15

(b) RR de tipo MX

N	Tupla País	Dominios	%	% Acc.
1	Chile	343	67,52	67,52
2	Estados Unidos	120	23,62	91,14
3	Estados Unidos, Canadá, Chile	24	4,72	95,87
4	Argentina	4	0,79	96,65
5	Estados Unidos, Alemania, Europa	5	0,98	97,64

(c) RR de tipo NS

Tabla A.2: Ranking de 5 países con más dominios asociados a través de valor del RR a alguna IP perteneciente a ellos en dominios gubernamentales.

	A	MX	NS
Amazon	5	2	23
Cloudflare	9	0	10
DigitalOcean	0	0	3
Google	5	142	13
Wix	0	0	0

Tabla A.3: Número de dominios gubernamentales en cada RR asociados a al menos un AS cuyo nombre contiene alguna sub cadena de texto relacionada con un proveedor conocido.

RR	Comparte IP con A	valor dominio
MX	162 (34,39 %)	282 (59,87 %)
NS	119 (23,43 %)	186 (36,61 %)

Tabla A.4: Número de dominios gubernamentales en cada RR que comparten IP o valor de dominio entre A y MX/NS. El porcentaje es relativo a la cantidad de dominios con 1 o más RRs del tipo indicado por la primera columna.

A.3. Uso en dominios gubernamentales de Proveedores Conocidos

La tabla A.3 muestra cambios considerables en la cantidad de dominios gubernamentales asociados a proveedores conocidos en la mayoría de los casos, los que son respaldados por la cantidad de dominios cuya infraestructura radica en Chile según lo determinado en la sección anterior. Sin embargo, *Google* toma gran importancia como proveedor de correo electrónico al estar asociado al 30,14 % de los dominios gubernamentales estudiados. Otro detalle interesante es que *Wix* no aparece de ninguna forma como proveedor de servicios web, a pesar de su popularidad relativa en el caso general.

A.4. Infraestructura compartidas en dominios gubernamentales

Al revisar la infraestructura compartida por IPs, se observa en la tabla A.4 que en el caso de MX, un poco más de un tercio de los dominios comparte infraestructura web con infraestructura de correo electrónico, mientras que en el caso de los registros de tipo NS, un poco menos de un cuarto de los dominios con al menos un registro A y al menos un registro NS comparten ambas infraestructuras.

Sin embargo, en el caso del conteo de cantidad de FQDN escaneados con registros NS y MX en el mismo dominio, este análisis tiene un significado distinto al hablar de dominios gubernamentales, ya que un número no menor de los FQDN escaneados comparte el mismo SLD (gob.cl o gov.cl). En este caso, se observa que un 59,87 % de los dominios comparte al menos el SLD en al menos un registro MX asociado a él, mientras que en el caso del registro NS, esta situación afecta a un poco más de un tercio de los dominios revisados.

A.5. Consideraciones sobre dominios gubernamentales

Los datos anteriores nos permiten comprender de mejor forma el espacio de dominios gubernamentales chilenos, desde la perspectiva de concentración y variedad de la infraestructura de servicios ofrecidos por ellos. Sin embargo, es necesario tener en cuenta que la mayoría de los análisis proporcionales se hacen sobre el universo de dominios obtenidos, los cuales corresponden a dominios públicos. Al mismo tiempo, hay que tener en cuenta las proporciones de universos de dominios revisados en cada caso. dado que el conjunto de dominios gubernamentales es aproximadamente mil veces más pequeño que el conjunto de dominios chilenos estudiados en el análisis general.

En el caso de variedad, se observa una mayor cantidad de valores de respaldo al compararse con el conjunto completo de dominios, mientra que al revisar la concentración, se nota que una gran parte de la infraestructura web es auto hospedada por el Ministerio del Interior. Además, se puede apreciar que el uso de Gmail en los dominios con servicio de correo electrónico casi duplica el del caso general, y que a pesar de ver más variedad de valores que en el caso general en los registros de tipo NS, el servidor de nombres de dominio más utilizado es el Ministerio del Interior. Lo anterior se puede deber a que si bien este ministerio está intentando centralizar las comunicaciones de todos los servicios de dependencia estatal, todavía quedan muchos que utilizan servicios contratados individualmente para sustentar su propia infraestructura de servicios de internet.

Anexo B

Detalles de Implementación del Sistema OSR

El sistema OSR se implementó tomando las decisiones de herramientas y tecnologías explicadas en esta sección. El código fuente de esta herramienta se encuentra en GitHub de forma pública, en la URL `https://github.com/clcert/osr`.

B.1. Herramientas a utilizar

En primer lugar, como método de almacenamiento se utilizará una base de datos relacional. Estas bases de datos permiten la agregación de forma simple de los resultados obtenidos, usando el lenguaje SQL para la realización de estas consultas. La implementación de base de datos relacional utilizada es PostgreSQL versión 11 [32], debido a la estabilidad y madurez que posee el proyecto.

Con respecto al lenguaje de programación utilizado en el desarrollo de la herramienta, se eligió Go, en su versión 1.12.7.Go [85] es un lenguaje creado por Robert Griesemer, Rob Pike y Kem Thompson el año 2009, y se caracteriza por su eficiencia y facilidad de uso, además de contar buenas con librerías existentes relacionadas con los procesos que se desean realizar en este trabajo.

B.2. Módulos implementados

Los módulos implementados para el sistema se pueden dividir en dos grupos. El primer grupo incluye tanto las funcionalidades de apoyo del sistema, las cuales son utilizadas de forma transversal en operaciones no relacionadas con el procesamiento de datos de escaneo,como los sistemas de **comandos**, **conexiones remotas**, **consultas SQL** y **notificaciones**. El otro grupo incluye las funcionalidades relacionadas con el procesamiento de información de escaneos, y está compuesto por los módulos **modelo de datos** y **tareas**. Este último se divide en tres submódulos: **entradas**, **procesos** y **salidas**.

La figura B.1 muestra la interacción entre los módulos recién mencionados en el proce-

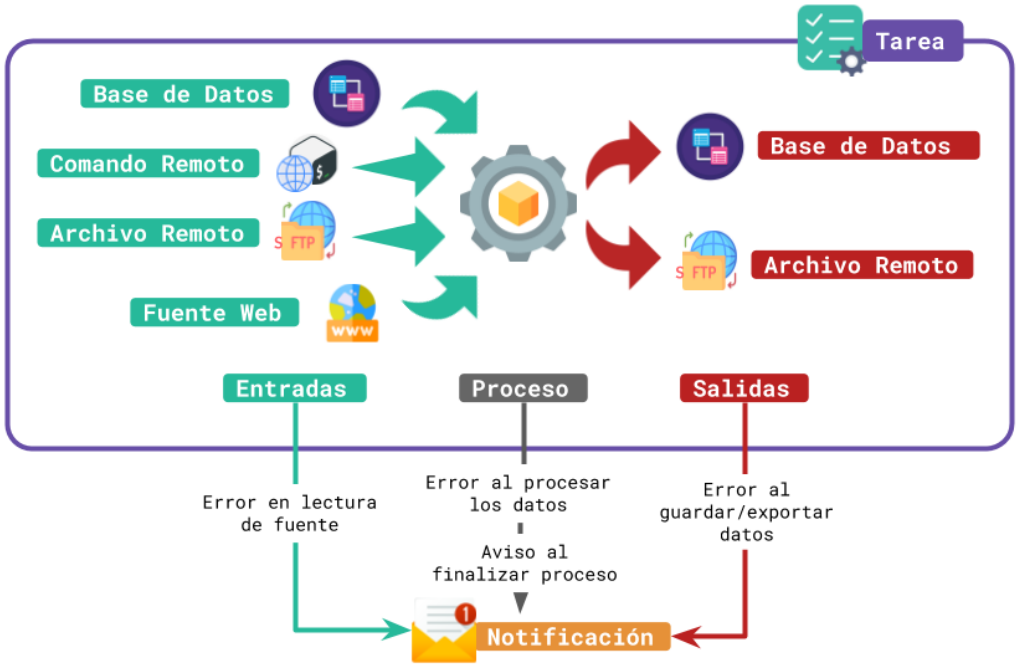


Figura B.1: Diagrama que muestra las interacciones entre gran parte de los módulos desarrollados en el sistema. Una tarea está compuesta por uno o más procesos, los cuales tienen asociados una o más entradas y una o más salidas. Tanto las entradas como los procesos y salidas notifican mediante el módulo de notificaciones del sistema cuando hay problemas de ejecución. Algunas entradas y salidas usan internamente el módulo de conexiones remotas para extraer los datos a procesar o exportar los datos ya procesados.

samiento de datos. El módulo de conexiones remotas no se muestra explícitamente, pero es usado de forma interna en el sistema de entradas y salidas.

B.2.1. Comandos

Para operar el sistema, se usan comandos de terminal asociados a la ejecución del binario `osr`. Estos comandos agrupan las distintas funcionalidades que posee la herramienta, y fueron implementados usando la librería de Go *Cobra Commander*.

Los comandos implementados en estos momentos son:

- `version`: Entrega la versión del programa
- `help`: Entrega ayuda sobre el uso del programa
- `init`: Contiene rutinas para inicializar la base de datos, creando los usuarios lectores y escritores y guardando la configuración en el archivo de configuración del sistema.
- `mailer`: Permite probar el sistema de notificaciones
- `models`: Permite inicializar las tablas de base de datos PostgreSQL de los modelos creados.
- `query`: Utilidad pequeña que permite realizar consultas SQL sobre la base de datos, exportándolas en archivos CSV.
- `remote`: Utilidad que permite configurar las conexiones remotas.

- `task`: Utilidad que permite ejecutar una tarea, siempre que se cuente con el archivo de configuración que la define.

Cada funcionalidad se explicará más en detalle en las próximas secciones, en los capítulos correspondientes a cada una de ellas.

B.2.2. Conexiones Remotas

Debido a que gran parte de los escaneos y recolección de datos se realiza en máquinas distintas, el sistema considera la posibilidad de ejecutar y usar programas y datos de servidores remotos, comunicándolos a través del uso de los protocolos SSH y SFTP respectivamente. Para esto, es requerido contar con un usuario especial en cada máquina, el cual pueda ser usado para iniciar sesión usando llaves SSH. El sistema es capaz de crear automáticamente estas llaves, manejando las conexiones a las máquinas de forma interna, a través de la asignación de un nombre único a cada una de ellas, el cual puede ser usado en otros componentes del sistema como alias.

En el caso de requerirse leer o escribir archivos remotos, el sistema inicia una conexión SFTP entre el servidor central y la máquina en la cual se desea escribir o leer. Se detallará en la sección de entradas y salidas la forma en la que estas conexiones se manejan.

B.2.3. Consultas SQL

Se definió una estructura específica para declarar consultas SQL recurrentes en el sistema. Estas consultas pueden ser ejecutadas directamente o ser usadas en los procesos de importación, y se almacenan en una carpeta en específico del sistema para poder ser encontradas con mayor facilidad. Un ejemplo de archivo con consultas SQL se puede encontrar en el extracto de código B.1. Cada entrada en un archivo de consultas SQL posee los siguientes campos:

- **name:** Nombre de la consulta SQL. Debe ser único dentro de cada archivo. Es usado para seleccionar y filtrar la consulta y para generar un archivo de salida con ella en caso de ser necesario.
- **description:** Descripción corta de la consulta.
- **query:** La consulta SQL representada por la entrada.

```
queries:
    - name: chilean-subnets
    description: subredes chilenas usando la fuente MaxMind
    query: -
        SELECT subnet from subnet_countries
        where
            source_id = 5 and
            country_geoname_id = 3895114 and
            task_id =
                (select MAX(task_id) from subnet_countries
                where
                    source_id = 5
                LIMIT 1)

    - name: blacklist
    description: "Redes que no hay que escanear"
```

```
17          query: -
18            SELECT subnet from blacklisted_subnets
```

Extracto de Código B.1: Ejemplo de archivo con consultas SQL.

B.2.4. Notificaciones

Se implementó un pequeño sistema de notificaciones a través de correo electrónico, que envía un aviso con los *logs* de los procesos al momento de finalizar. Además, el sistema de notificaciones está configurado para enviar un mensaje de correo electrónico a las personas suscritas cuando algún comando de `OSR` falla inesperadamente.

A modo de referencia, se muestra a continuación un extracto de un correo electrónico enviado por el sistema:

```
Hola,

Te informamos que se terminaron de ejecutar uno o más procesos
incluidos en una tarea.

Nombre Tarea:             Dominios NIC Chile
Número Tarea:             61
Procesos satisfactorios:  nic-chile/domains
Procesos con error:       Ninguno
Fecha Inicio:             13-08-2019 00:01:01 -0400
Fecha Fin:                13-08-2019 00:01:03 -0400
Estado Final:             Éxito


Adjuntamos también los archivos de log asociados a esta sesión
de importación.

Saludos!

Observatorio de Seguridad de la Red
CLCERT Universidad de Chile.
```

B.2.5. Modelo de datos

El componente de *modelo de datos* reúne todas las definiciones de datos recopiladas por el sistema. Este componente utiliza la librería *go-pg* [41], la cual permite definir tablas SQL a partir de estructuras de Go de forma simple. Lo anterior facilita bastante la extensión de las tablas utilizadas en presencia de nuevos procesos de escaneo o agregación de datos implementados en el sistema.

El extracto de código B.2 muestra un ejemplo de definición de modelo a través de una estructura de Go. Las anotaciones al lado de cada campo son usados por la librería *go-pg*

para determinar propiedades especiales de la tabla, como llaves, tipos y si el campo puede o no tener un valor nulo.

```
1
2 type PortScan struct {
3     TaskID     int           `sql:",type:bigint"`
4     Task       *Task
5     SourceID   DataSourceID `sql:",pk,notnull,type:bigint"`
6     Source     *Source
7     Date       time.Time     `sql:",pk,notnull"`
8     ScanIP     net.IP        `sql:",pk"`
9     IP         net.IP        `sql:",pk"`
10    PortNumber uint16        `sql:",pk,type:smallint"`
11    Protocol   PortProtocol `sql:",pk,type:smallint,notnull"`
12    Port       *Port
13 }
14
```

Extracto de Código B.2: Ejemplo de definición de modelo en Go.

Los modelos de datos implementados actualmente son los siguientes:

- **Sistemas Autónomos:** Utilizado para guardar la lista de sistemas autónomos existentes.
- **Lista de Bloqueo:** Utilizado para guardar la lista de subredes que no deben ser escaneadas por los sistemas del CLCERT, generalmente por petición expresa de los administradores de dichas subredes.
- **Países:** Lista de países del mundo, con sus nombres e identificadores geográficos según ISO 3166 [83] y GeoNames [84].
- **Paquetes de Darknet:** Tabla que almacena las cabeceras TCP/IP de los paquetes recibidos por la *Darknet* del CLCERT.
- **Resource Records de DNS**: Tabla que almacena los valores de los *Resource Records* escaneados por el CLCERT.
- **Dominios:** Tabla para almacenar dominios conocidos y de interés.
- **Categorías de Dominios:** Tabla usada para guardar las categorías a los dominios almacenados. Al mismo tiempo, existe una tabla que permite establecer una relación *many-to-many* entre dominios y categorías.
- **Rankings de Dominios:** Tabla usada para almacenar rankings (posiciones relativas según algúna métrica) asociados a los dominios almacenados, según distintas fuentes.
- **Escaneos de Puertos y protocolos:** Tabla que guarda puertos detectados como abiertos en escaneos de protocolos realizados por el CLCERT. Además, marca si es que se pudo reconocer el funcionamiento de protocolos conocidos en el puerto, intentando extraer de los datos de escaneo información como el *software* usado para proveer el servicio, la versión de este programa y los *banners* detectados.
- **Reportes de Malware y Vulnerabilidades:** Tabla que guarda reportes de malware y vulnerabilidades generados por el CLCERT o recibidos por organizaciones externas.
- **Fuentes:** Tabla que lista todas las fuentes registradas en el sistema. Una fuente es una organización identificable que provee los datos almacenados.

- **Software y Versiones:** Tabla que listan los nombres de algunos programas usados para servir algunos protocolos. También hay una tabla que mantiene todas las versiones de cada software almacenado, y la fecha de publicación de éstas.
- **Mapeo de subredes a Sistemas Autónomos y Países:** Entradas que permiten obtener el país o sistema autónomo de una IP en específico.
- **Tareas:** Listado de las tareas ejecutadas por el sistema, con sus fechas de inicio y de término y estado (en proceso, terminada satisfactoriamente o terminada con error).
- **Contactos:** Lista de contactos encargados de conjuntos de subredes, con el objetivo de, a futuro, emitir notificaciones automáticas en caso de detectar algún problema.
- **Certificados:** Lista de metadatos de certificados asociados a servicios de internet. Entre los metadatos recopilados, se encuentran fechas de vencimiento de certificados, entidades que los entregan, si son o no autofirmados y los métodos de cifrado usados.

Es posible ver las relaciones entre la mayoría de estos modelos en la figura B.2.

B.2.6. Tareas

Este sistema define como "Tarea" a un conjunto de **procesos** a ejecutarse de forma secuencial o paralela, definida por un archivo de configuración especial. El objetivo de estos archivos de configuración es agrupar procesos que deben ser ejecutados de forma secuencial.

El formato de los archivos de configuración se puede observar en el extracto de código B.3 y se explica a continuación:

- **name:** Nombre de la tarea, utilizado en *logs* y notificaciones.
- **description:** Descripción de la tarea, utilizada en *logs* y notificaciones.
- **abortOnError:** Detiene la ejecución de todas las tareas si es que esta tarea termina con error.
- **parallel:** Ejecuta esta tarea en paralelo con la siguiente definida en el archivo de configuración.
- **processes:** Lista de procesos a ejecutar.
- **params:** Parámetros globales de la forma `llave:valor` que serán entregados a todos los procesos ejecutados.

Los demás campos se explicarán en las siguientes secciones.

```
1
2     name: "Ranking Alexa"
3     description: "Esta tarea importa rankings de Alexa"
4     abortOnError: true
5     parallel: false
6     processes:
7         - command: "import/alexa-rankings"
8         params:
9             tlds: "cl"
10        sources:
11            - http:
12                url: "http://s3.amazonaws.com/alexa-static/top-1m.csv.zip"
13                method: "GET"
```

```
14          savers:
15              - postgres:
16                  insertconfig:
17                  DomainRanking:
18                      onconflict: "do nothing"
19
```

Extracto de Código B.3: Ejemplo de definición de tarea en el OSR

B.2.7. Procesos

El sistema define como **proceso** a una rutina de código específica que recibe una o más **entradas** y envía la información procesada a una o más **salidas**. El proceso es desarrollado como extensión al código del sistema OSR, y se le designa un nombre único para poder referirse a él.

Con el objetivo de identificar rápidamente el proveedor de los datos con los que trabaja el proceso, se decidió que este nombre siguiera una sintaxis que permitiese reflejar tanto el tipo de proceso como la fuente y un nombre asociado a su función. El formato de este nombre no tiene impacto directo en cómo los procesos se ejecutan, pero su estandarización facilita la memorización de este nombre de parte de las personas encargadas de crear y ejecutar los procesos.

Los procesos también permiten definir *parámetros* especiales para ellos, los cuales tienen preferencia por sobre los parámetros definidos a nivel de tarea. Estos parámetros pueden ser usados por el código de los mismos procesos, como también se pueden usar como variables en los campos *sources* y *savers* del archivo de configuración del proceso.

Los procesos actualmente definidos en el sistema son los siguientes:

- **Importación de Rankings de Alexa Internet:** Este proceso importa los archivos generados por Alexa Internet correspondientes a Rankings de Dominios. Al mismo tiempo, puede importar todos los TLDs o solamente un subgrupo de estos, usando parámetros especiales.
- **Sistemas Autónomos existentes de CIDR Report:** Este proceso importa las listas de Sistemas Autónomos publicadas por CIDR Report [87].
- **Reportes de RRs en Dominios Chilenos:** Este proceso importa reportes de RRs en dominios chilenos producidos por escaneos con la herramienta *Mercury*.
- **Darknet:** Este proceso, desarrollado por el estudiante de cuarto año *Gabriel Norambuena*, importa las cabeceras TCP/IP de los paquetes de la *Darknet* del CLCERT.
- **Escaneo de Puertos**: Este proceso importa una lista de IPs, registrando que tuvieron algún puerto específico abierto en alguna fecha determinada.
- **Base de datos MaxMind Geolite2:** Este proceso importa la base de datos de IPs a Sistemas autónomos y de IPs a países GeoLite2 de MaxMind.
- **Dominios:** Este proceso importa una lista de dominios, uno en cada línea.
- **Protocolos y Certificados:** Este proceso importa datos de escaneo de protocolos, así como también información de certificados usados en las versiones seguras de estos protocolos.

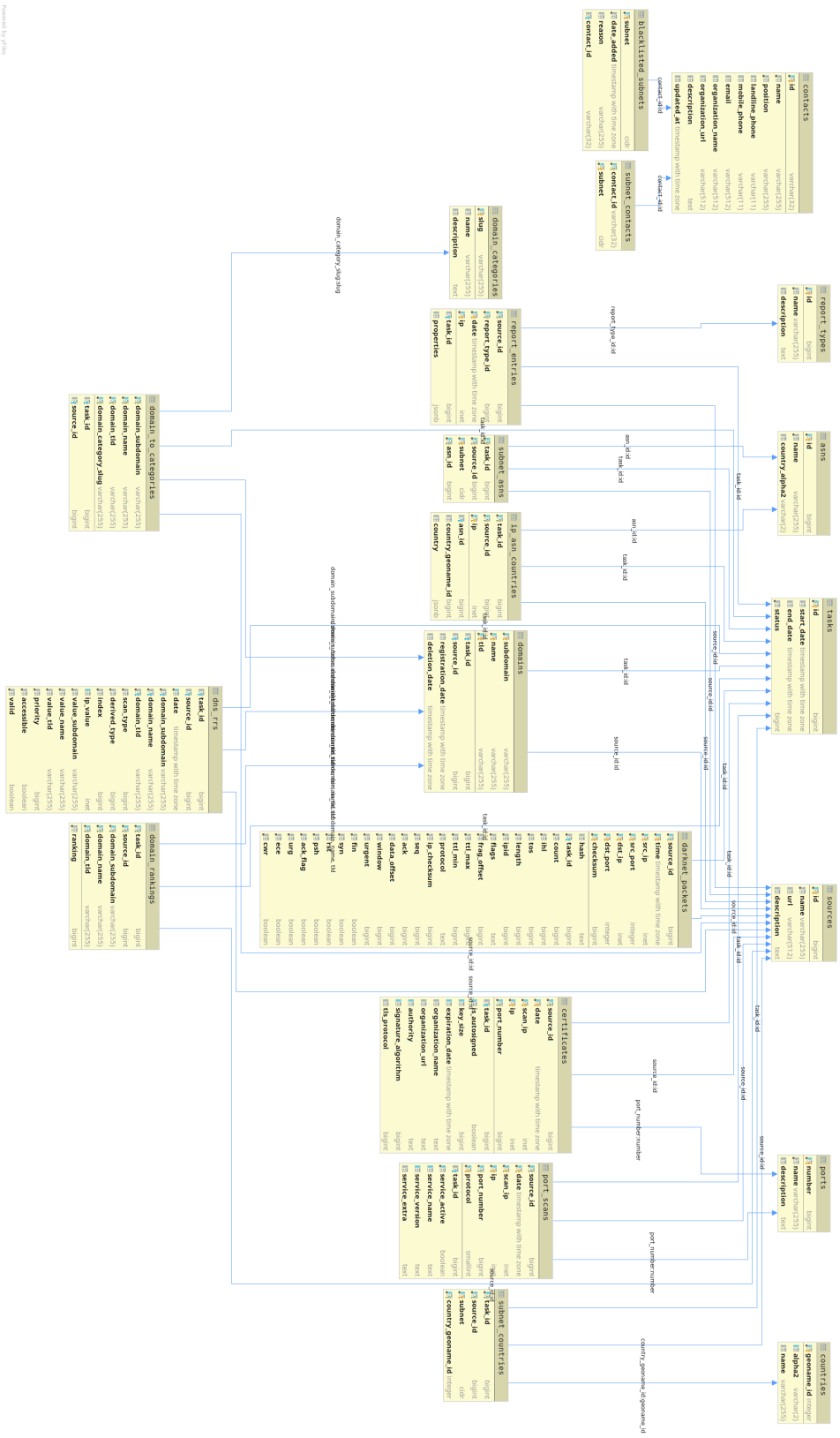


Figura B.2: Diagrama que muestra las tablas usadas para guardar los datos de los modelos desarrollados.

- **Fuentes 1 y 2 de Reportes de Malware y Vulnerabilidades:** Este conjunto de procesos importa datos de fuentes externas de reportes.
- **Diferencia entre dos Subredes o conjuntos de IPs:** Esta tarea calcula la diferencia de IPs entre dos subredes entregadas como entradas. Como diferencia se consideran las subredes/IPs añadidas, eliminadas, y en común.

B.2.8. Entradas

Una entrada es un *stream* o flujo de "archivos" recibido por el proceso desde alguna fuente en específico. Estos archivos son a su vez *streams* de datos que se pueden leer de forma secuencial. En términos del lenguaje Go, los archivos implementan la interfaz `io.Reader`.

Un proceso puede estar asociado a una o más fuentes, las cuales pueden manejar uno o más "archivos". Estos archivos son identificables con un *nombre* y una *ruta*, valores que pueden tener distintos significados en cada proceso. Por ejemplo, si este "archivo" representa un archivo real en alguna máquina, su nombre y ruta son respectivamente el nombre y ruta de éste. Sin embargo, si la entrada no representa un archivo real, estos valores pueden ser definidos en el archivo de configuración de la entrada correspondiente con el objetivo de emular un sistema de archivos.

Las entradas implementadas en el sistema a la fecha son las siguientes:

- **Archivo Remoto SFTP:** Esta entrada representa un *stream* de archivos ubicados en alguna carpeta de una máquina remota registrada en el sistema. Los archivos a ser seleccionados para su envío en *stream* pueden ser definidos a partir de su calce con una lista de expresiones regulares. Además, es posible indicar si la entrada debe buscar archivos de forma recursiva o solamente en la carpeta indicada. El orden de los archivos no está garantizado.
- **Página Web:** Esta entrada representa uno o más cuerpos de páginas web en una URL en específico. En caso de indicar una búsqueda recursiva, esta entrada extrae todos los links dentro de la página raíz entregada, creando archivos virtuales a partir de ellos. En caso de no indicar una búsqueda recursiva, la entrada contiene un solo archivo (la página raíz). Al igual que en el caso anterior, la búsqueda recursiva permite la utilización de filtros por expresiones regulares. El módulo soporta páginas web que requieren enviar parámetros `POST` o `GET` para ser vistas (como un formulario de inicio de sesión), además de páginas con autenticación de tipo `BASIC` [68].
- **Comando Remoto:** Esta entrada corresponde a un único "archivo virtual", cuyo contenido corresponde a la salida de un comando ejecutado en algún servidor remoto ejecutado. Tanto el nombre como la ruta de este archivo virtual son definibles a través del archivo de configuración.
- **Consultas SQL sobre base de datos:** Esta entrada corresponde a uno o más "archivos virtuales", cuyo contenido es el resultado en formato CSV de una consulta SQL contenida en un archivo manejado por el módulo *consultas SQL*. Es posible filtrar las consultas a realizar declarando explícitamente sus nombres.

B.2.9. Salidas

Una salida es una abstracción de un servicio de almacenamiento de los datos procesados. Una salida debe ser capaz de trabajar con una *struct* de Golang genérica, con un *map*, o con una estructura de tipo `Savable`, la cual adjunta metadatos que permiten al proceso de salida guardar la estructura de forma adecuada. La salida además debe asegurarse que todos los datos entregados a ella se guarden correctamente, o entregar un error en caso contrario.

Existen dos tipos de salidas implementadas:

- **Modelo de Base de Datos:** Permite guardar en la base de datos estructuras de tipo *modelo*, las cuales deben tener creadas bases de datos en el sistema. Las estructuras se guardan de a grupos de tamaño definido por el archivo de configuración, o luego de transcurridos algunos segundos, con el objetivo de evitar pérdida de datos en caso de error.
- **Archivo CSV en servidor remoto:** Almacena una estructura en formato CSV en un servidor remoto. El archivo de configuración requiere definir las columnas a guardar, así como también el nombre y la ruta de los archivos a exportar según el nombre de la estructura o un ID definido.

Actualmente, una estructura de tipo `Savable` puede definir propiedades especiales para el objeto contenido por ella. En estos momentos solo se encuentra definida la siguiente propiedad:

- `outID`: Define un canal de salida específico para una estructura. En el caso de modelo de Base de Datos, permite agrupar las inserciones SQL de forma de no contener elementos conflictivos. ya que insertar 2 veces el mismo elemento en una misma consulta (porque, por ejemplo, la fuente lo repite en sus reportes), se generan problemas. En el caso de archivos CSV, permite definir la configuración a utilizar para cada estructura.

Anexo C

Validación histórica de datos de escaneo de protocolos de Censys

Este capítulo repite la estrategia de comparación presentada en el capítulo 4.2 de este documento, pero sobre los datos de escaneo de protocolos de Censys.

C.1. Datos históricos de Censys

Durante los casi 3 años de operación, el escaneo de Censys detectó 2.848.242 IPs chilenas distintas sirviendo adecuadamente algún servicio relacionado con los puertos escaneados. Al igual que en el caso de CLCERT, la última columna de la tabla 2.3 muestra la cantidad de IPs únicas por tipo de escaneo manejado por ellos, lo cual muestra números de la misma magnitud que en el caso del CLCERT en los puertos 21, 80 y 8080, pero bastante distintos en los puertos 22, 25, 110, 143. Es importante mencionar que los escaneos de Censys parten en algunos casos mucho después que los del CLCERT (como por ejemplo, en el caso de HTTPS), e incluso algunos terminan antes (como con SSH, del cual no se sabe más desde mayo de 2019), lo que puede explicar gran parte de las razones por las que las estadísticas agregadas varían.

C.1.1. Análisis global de IPs encontradas

Revisando las figuras incluidas en C.1, se pueden notar comportamientos bastante similares en el puerto 21 (figura C.1a). Sin embargo, las IPs que han sido más veces escaneadas llegan a ser casi el doble del caso de CLCERT, lo que es más destacable considerando la cantidad de escaneos realizados por esta fuente (156 semanas). En contraste, el escaneo del puerto 22 (figura C.1b) ha encontrado solamete 85 de las 156 veces una misma IP, y en este caso los escaneos del CLCERT lo duplican.

En el caso del puerto 25 (figura C.1c), se ven comportamientos bastante similares en forma, aunque en escalas distintas, dado que CLCERT logra encontrar 4 veces más IPs sirviendo el protocolo SMTP que Censys. La misma situación proporcional ocurre en los protocolos 110 (figura C.1e) y 143 (figura C.1f). Además, el puerto 8080 (figura C.1h) tiene un

comportamiento y número bastante similar, aunque con un número menor de coincidencias, fomentado por la cantidad total de escaneos a este puerto (113).

Por último, con respecto a los dos puertos oficiales de contenido web, se puede notar que el comportamiento del puerto 443 (figura C.1g) difiere bastante del caso del CLCERT, lo que muy probablemente se explica por la poca cantidad de escaneos asociado a él, ya que viene recopilándose desde octubre de 2018. Mientras que por otro lado, el comportamiento del puerto HTTP es bastante peculiar, ya que se observa un salto grande de IPs reconocidas al menos en 20 ocasiones. Esta anomalía muy seguramente está relacionada con el salto que se puede apreciar en la figura 2.6d, alrededor de mayo de 2017, y que como ya se explicó, una de sus causas posibles es una mala configuración con respecto a asignación de IPs a países de parte de Censys.

C.1.2. Continuidad de las IPs por protocolo

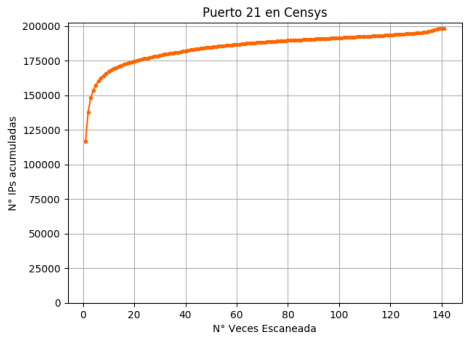
La continuidad de las IPs por protocolo de Censys se puede observar en la figura C.2 (páginas 162, 163 y 164), la cual intenta graficar el mismo concepto discutido en el caso de CLCERT, pero sobre los datos de esta fuente de escaneos.

Los resultados obtenidos sobre el protocolo FTP y observables en la figura C.2a se muestran bastante menos caóticos que los obtenidos por CLCERT. Dejando de lado unas bajas en medición producidas en junio y septiembre de 2018, se nota que las líneas de IPs comunes por periodo se mantienen separadas a distancias bastante similares. Un detalle bastante destacable es que la cantidad de IPs mantenidas durante un año hacia finales de 2019 es un poco menor al 60 %, un número bastante alto cosiderando los resultados generales de puertos distintos a los de correo electrónico.

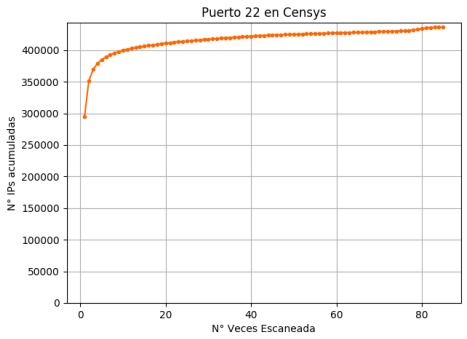
El caso del protocolo SSH visible en la figura C.2b entrega menos información debido probablemente a que se poseen una menor cantidad de escaneos que en el caso del CLCERT. En estos gráficos se puede apreciar la existencia de errores en medición, la cual afecta a la comparación histórica constantemente. Sin embargo, en promedio, las líneas se mantienen en posiciones bastante definidas, notándose que menos de un 50 % de las IPs se mantienen durante un año.

El caso del protocolo SMTP observable en la figura C.2c es bastante curioso, debido a que hacia finales de abril se nota una diferencia en IPs bastante grande entre los valores de esa fecha y los de un mes antes. Estas diferencias se vuelven a observar en distintas curvas de IPs repetidas de periodos pasados. Al mismo tiempo, la curva de 6 meses muestra una gran diferencia en el intervalo de septiembre 2018 y enero 2019, la cual puede estar motivada por la gran alza de dispositivos detectados alrededor de septiembre de 2018, la cual creció más de dos veces en una sola semana.

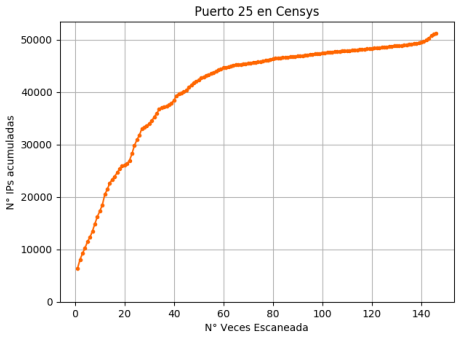
Los resultados de Censys en el protocolo HTTP, mostrados en la figura C.2d, son bastante más caóticos que su contraparte del CLCERT. Esta situación, al igual que las anteriores, se puede deber a errores internos. Sin embargo, ignorando estas medidas, los resultados se muestran bastante estables durante los 2 años revisados, así como también son estables la cantidad de IPs repetidas en los distintos periodos analizados. Es importante mencionar que en este caso, se analizaron los resultados desde mayo de 2017, debido a la anomalía en



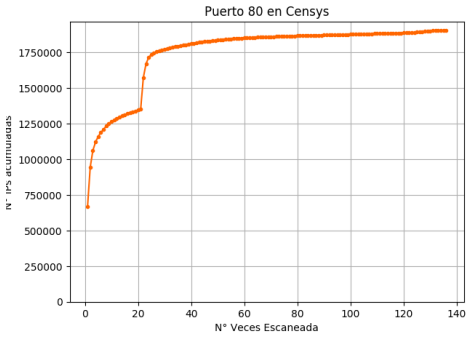
(a) Puerto 21



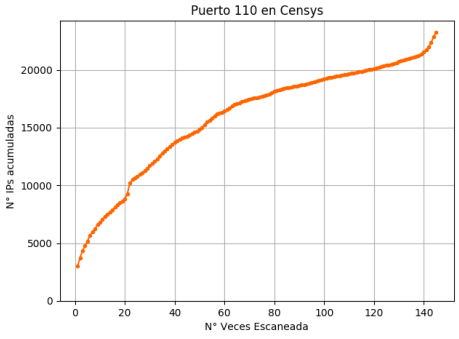
(b) Puerto 22



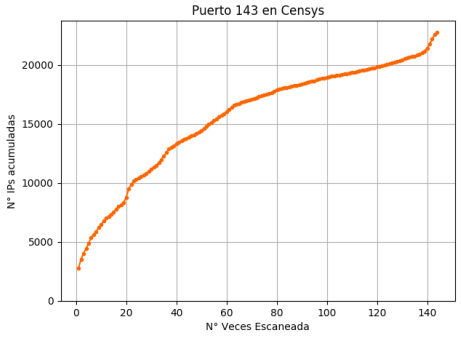
(c) Puerto 25



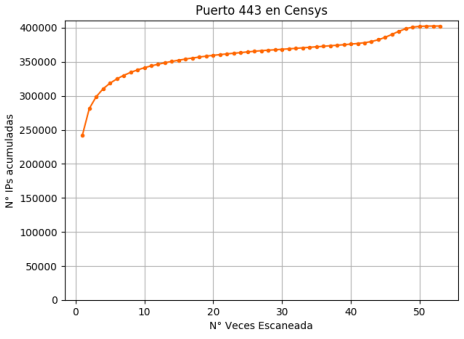
(d) Puerto 80



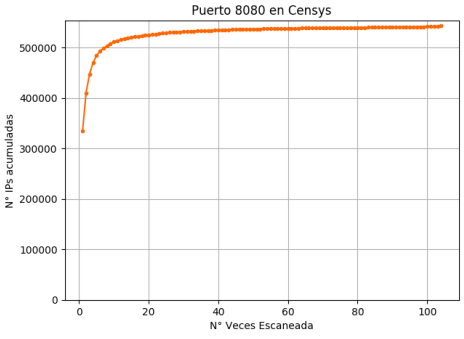
(e) Puerto 110



(f) Puerto 143

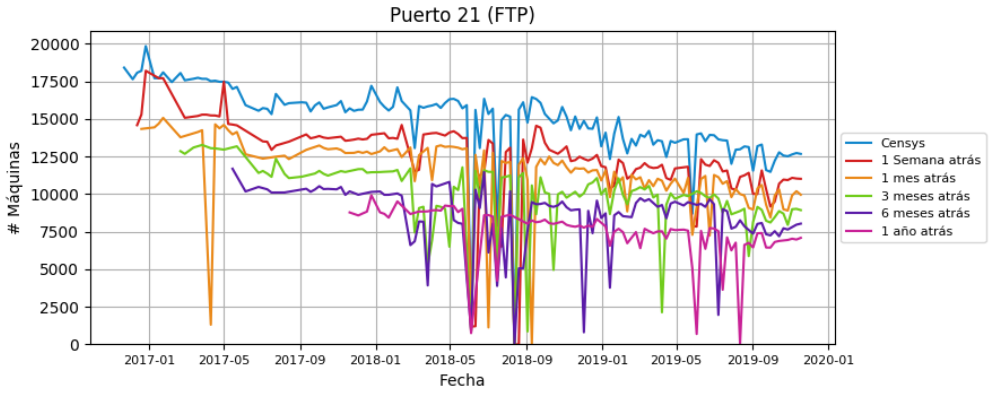


(g) Puerto 443

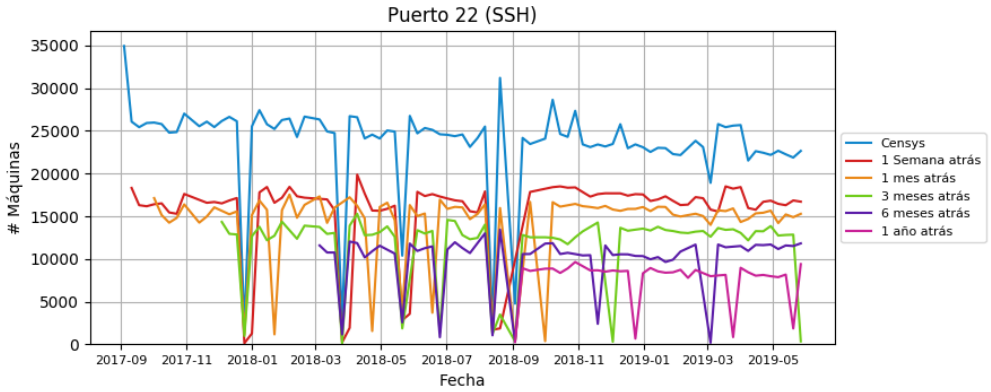


(h) Puerto 8080

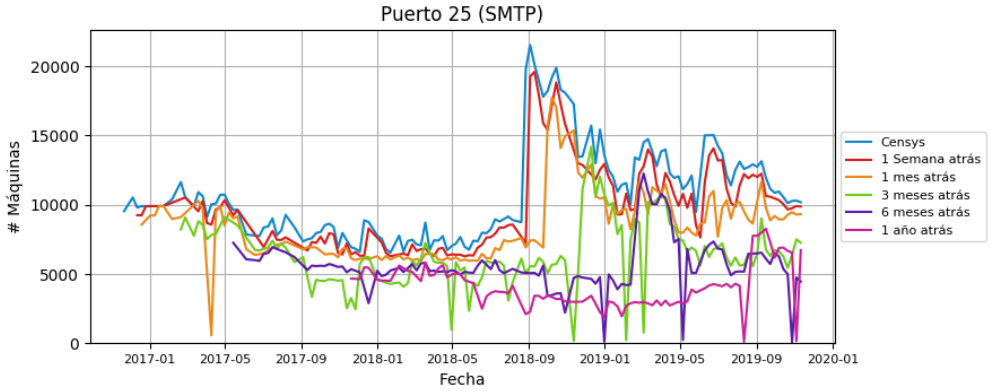
Figura C.1: Gráfico que muestra la cantidad de IPs (eje y) que se vio al menos cierta cantidad de veces (eje x) en los resultados de escaneos del Censys.



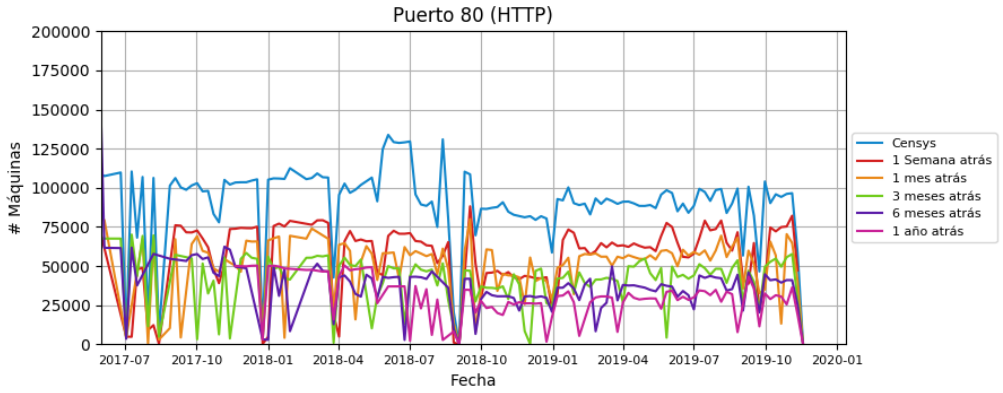
(a) Puerto 21 (FTP)



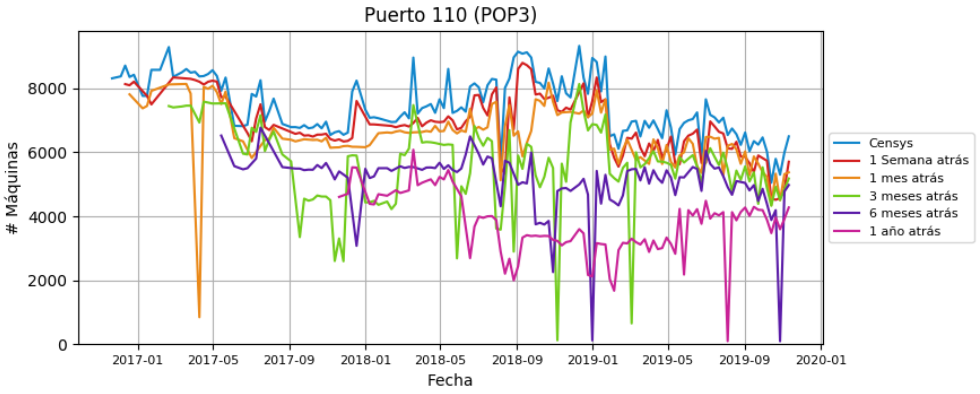
(b) Puerto 22 (SSH)



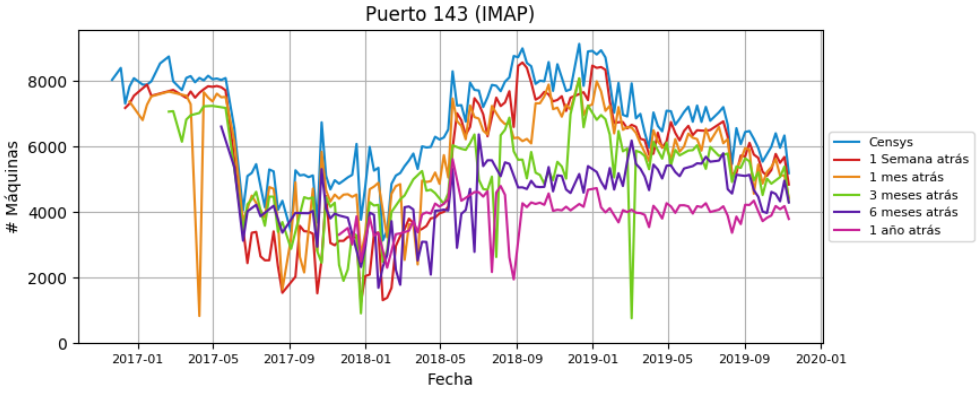
(c) Puerto 25 (SMTP)



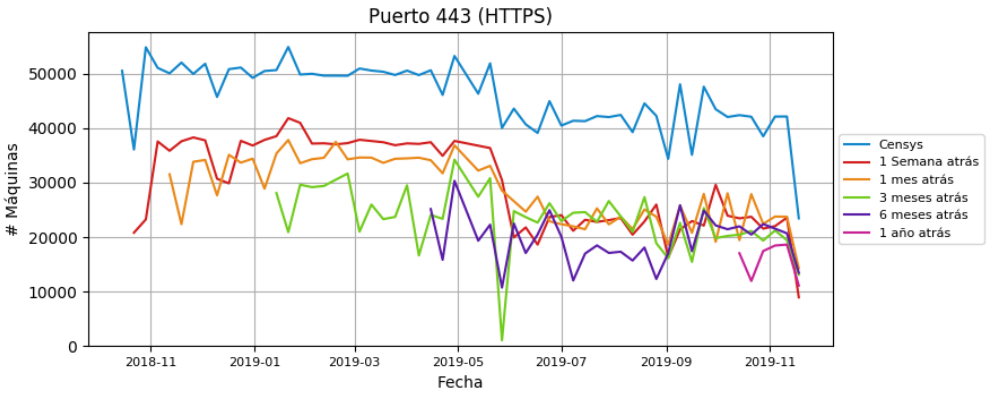
(d) Puerto 80 (HTTP)



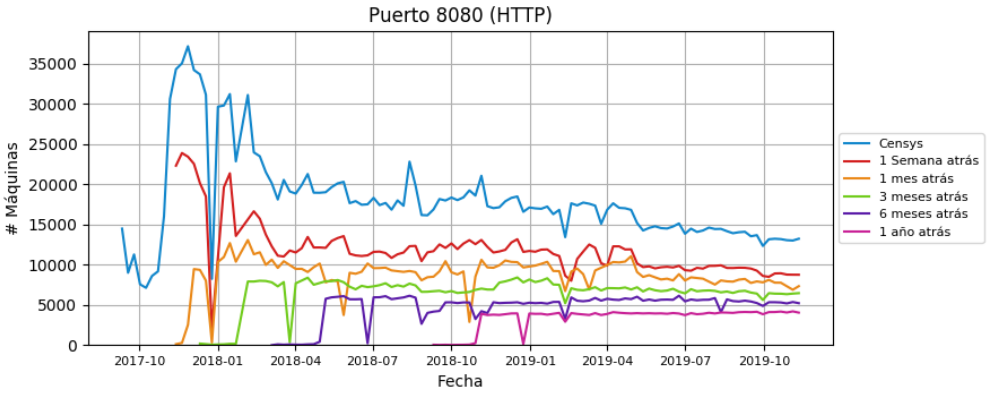
(e) Puerto 110 (POP3)



(f) Puerto 143 (IMAP)



(g) Puerto 443 (HTTPS)



(h) Puerto 8080 (HTTP)

Figura C.2: Comparación histórica de la cantidad de IPs encontradas por cada protocolo en los escaneos de Censys.

resultados de este puerto en periodos anteriores a esta fecha.

En los escaneos de Censys, los resultados visibles en las figuras C.2e y C.2f, correspondientes a los protocolos POP3 e IMAP respectivamente, no se parecen tanto como en el caso del CLCERT. La mayor diferencia se observa en el intervalo de junio de 2017 y mayo de 2018, momento en el que en el caso del puerto 143 las líneas de todos los periodos se cruzan bastante. Hacia fines del año 2019, los resultados en todos los periodos temporales se parecen bastante, ignorando los valores *outliers* que llegan a números cercanos a cero.

La poca cantidad de escaneos sobre el protocolo HTTPS visibles en la figura C.2g no permite obtener muchos resultados contrastantes, en especial en las mediciones de 6 meses y un año atrás. Al contrario de la medición del CLCERT, hacia fines de noviembre de 2019, la cantidad de IPs repetidas de periodos pasados es bastante similar en esta medición.

Por último, se observa que nuevamente los resultados del protocolo HTTP sobre el puerto 8080 visibles en la figura C.2h son bastante similares a los obtenidos por CLCERT, mostrando la misma base de IPs en común existentes hace un año, la cual cuenta con un poco menos de 5000 IPS únicas.