



TRABAJO FIN DE GRADO
INGENIERÍA EN INFORMÁTICA

Implementación en tiempo real de sistemas de identificación de tráfico de red

Subtítulo del proyecto

Autor

Álvaro Maximino Linares Herrera

Directores

Jesús Esteban Díaz Verdejo



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE
TELECOMUNICACIÓN

—
Granada, 3 de septiembre de 2017



Implementación en tiempo real de sistemas de identificación de tráfico de red

Subtitulo del proyecto

Autor

Álvaro Maximino Linares Herrera

Directores

Jesús Esteban Díaz Verdejo

Implementación en tiempo real de sistemas de identificación de tráfico de red

Álvaro Maximino Linares Herrera

Palabras clave: inspección profunda de paquetes, bro, flujos, identificación, clasificación, red, tráfico

Resumen

La identificación de tráfico en red es realmente importante para aplicaciones de ingeniería de tráfico y de seguridad.

En este trabajo se tratará la creación de un programa para un NMS (Network Monitoring System), en este caso se usará BRO, mediante el cual se pueda resolver el emparejamiento de flujos. BRO consiste en un NMS que funciona mediante el terminal en Linux o Mac, una de las peculiaridades de este programa es que para la creación de scripts que nos permitan extender la funcionalidad de la que dispone, tendremos que usar BRO como lenguaje de programación. Es un lenguaje de scripting, el cual está orientado a eventos, que se lanzan cuando ocurre algo relacionado con el control y análisis de redes, es un lenguaje que para los que vienen de C++, Java o Python, no debe de suponer un gran reto, más allá de acostumbrarse a sus sintaxis. Es un lenguaje potente que al estar orientado a redes nos permite obtener mucha información de los flujos que tenemos en la red o en el archivo que vayamos a analizar. En este trabajo mediante implementaciones offline se verificará la eficacia de esta técnica de clasificación de tráfico.

Project Title: Project Subtitle

First name, Family name (student)

Keywords: Keyword1, Keyword2, Keyword3,

Abstract

Write here the abstract in English.

Yo, **Álvaro Maximino Linares Herrera**, alumno de la titulación TITULACIÓN de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 76669401M, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Álvaro Maximino Linares Herrera

Granada a 3 de septiembre de 2017.

D. **Jesús Esteban Díaz Verdejo**, Profesor del Área de XXXX del Departamento YYYY de la Universidad de Granada.

Informa:

Que el presente trabajo, titulado *Implementación en tiempo real de sistemas de identificación de tráfico de red*, ha sido realizado bajo su supervisión por **Álvaro Maximino Linares Herrera**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a 3 de septiembre de 2017.

Los directores:

Jesús Esteban Díaz Verdejo Nombre Apellido1 Apellido2 (tu-
tor2)

Agradecimientos

Poner aquí agradecimientos...

Índice general

1. Introducción	17
1.1. Motivación y antecedentes	17
1.2. Objetivos	20
1.3. Metodología	21
1.4. Estructura de la memoria	22
2. Estado del arte	23
2.1. Identificación de tráfico	23
2.1.1. Técnicas de identificación de tráfico	23
2.1.2. Identificación de tráfico basada en flujos	24
2.1.3. DPI	25
2.2. BRO	25
2.2.1. Funcionalidades básicas de BRO	27
2.2.2. Eventos y trazas	28
2.2.3. Incorporación de funcionalidades	28
2.3. Emparejamiento de flujos	29
3. Diseño y arquitectura del sistema	31
3.1. Arquitectura del sistema	31
3.2. Módulo y funciones	32
3.3. Gestión de flujos	33
3.4. Estructuras de datos	34
4. Implementación	37
5. Evaluación y pruebas	41
6. Conclusiones y trabajo futuro	43
Bibliografía	47

Índice de figuras

1.1. Modelos de capas de Internet.	17
1.2. Diagrama de conexión de TCP y UDP.	19
1.3. Temporización en diagrama de Gantt.	22
2.1. Logo de Bro.	26
2.2. Arquitectura de Bro.	26

Capítulo 1

Introducción

1.1. Motivación y antecedentes

Dada la gran expansión de Internet desde los años 90, la seguridad se ha convertido en algo fundamental. Hoy en día el uso de Internet se ha generalizado tanto que ha llegado a todas las áreas, desde finanzas hasta el envío de pruebas médicas, pasando por compras, redes sociales y demás, por lo tanto se tratan temas de carácter muy privado y personal.

Internet se organiza mediante un modelo de capas. Hay varios modelos de capas, como son el modelo TCP/IP y el modelo OSI. [1]

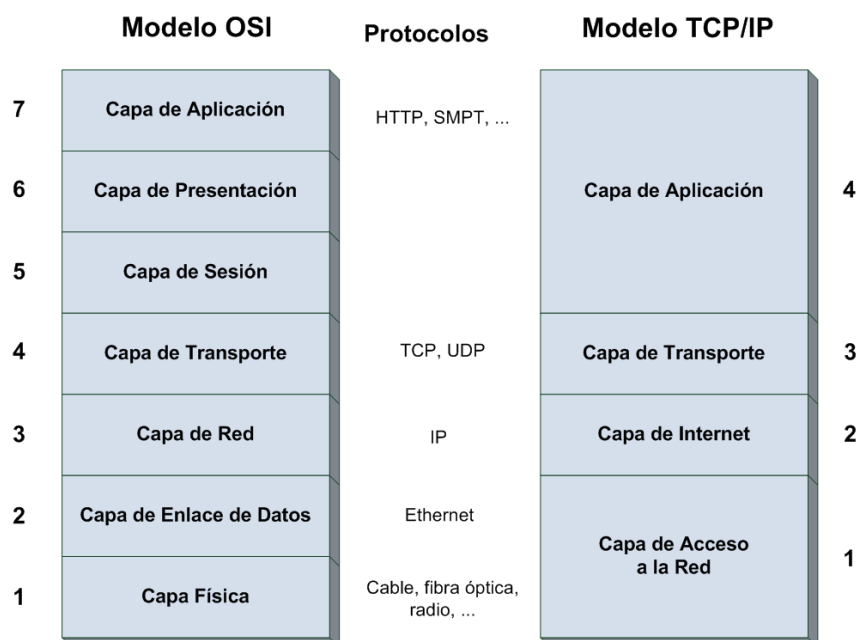


Figura 1.1: Modelos de capas de Internet.

Dentro del modelo TCP/IP, que es el que se seguirá en este proyecto, es importante tener claro las siguientes capas, pues forma parte de los conocimientos básicos para lograr entender el emparejamiento de flujos.

■ Capa de aplicación

En la capa de aplicación se encontrarán las aplicaciones de red y sus protocolos. Los protocolos que se encuentran en esta capa son HTTP, SMTP y FTP.

- El protocolo *HTTP* permite la solicitud y transferencia de documentos web.
- El protocolo *SMTP* permite la transferencia de mensajes de correo electrónico.
- El protocolo *FTP* permite la transferencia de archivos entre dos terminales.

A parte de estos protocolos también corresponde a esta capa el **DNS**, el cual se encarga de traducir los nombres de los sitios web.

Para una información más extensa sobre la capa de aplicación consulte [2]

■ Capa de transporte

La capa de transporte es la encargada de transportar los mensajes de la capa de aplicación. Los protocolos de esta capa son **TCP y UDP**. La principal diferencia y será algo que siempre se debe de tener presente cuando se trabaja con redes es que *TCP* está orientado a garantizar la conexión, mientras que *UDP* no garantiza la conexión.

Para ampliar información sobre esta capa consulte [3]

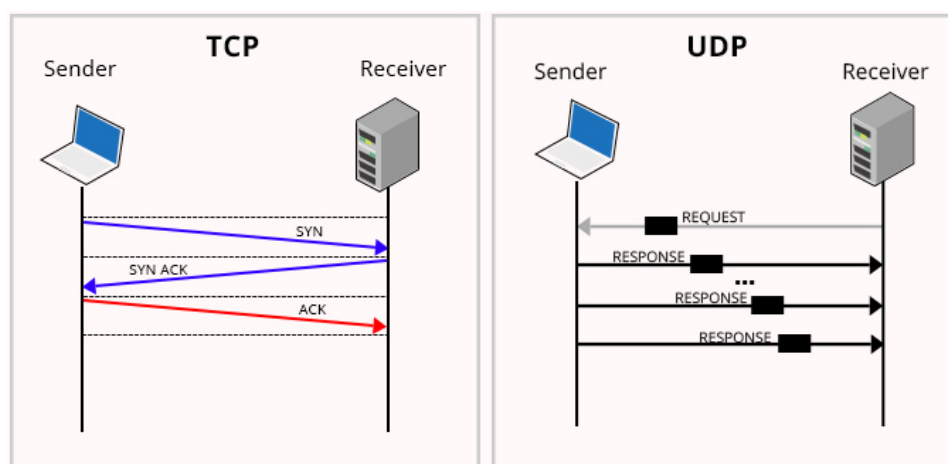


Figura 1.2: Diagrama de conexión de TCP y UDP.

■ Capa de red

En la capa de red es donde se encuentra el protocolo **IP**, el cual es el encargado de especificar el formato de envío entre los routers y los hosts y también el protocolo **ICMP**, el cual es un protocolo de control.

En esencia, esta capa es la que se encarga de transportar los *datagramas* de un *host* a otro.

Para ampliar sus conocimientos sobre la capa de red consulte [4]

En Internet la información que se envía se divide en paquetes. Estos paquetes tienen una IP de origen y otra de destino, así como un puerto de origen y un puerto de destino. Estos paquetes, que contienen parte de la información que se envía por la red, son unidos de nuevo una vez que llegan a su destino, de forma que entonces se dispondrá de la información completa. Dependiendo del tipo de protocolo se tolerarán pérdidas de paquetes o no, para ilustrar esta idea, en una llamada mediante *Skype* se tolera perder algunos paquetes, pues no se notará mucha diferencia. Si se pierden muchos paquetes la imagen se verá borrosa y la voz no se escuchará bien. Ahora en un correo electrónico, no se tolerará perder paquetes, pues entonces no se obtendrá la información completa. Cuando varios paquetes tienen el mismo origen y el mismo destino se denomina flujo. [5]

De esta forma se llega al emparejamiento de flujos, el cual es muy útil para los NMS, sistemas de monitoreo de red, *Network Monitoring System*, como podría ser BRO [6]. Gracias al uso de estos programas el administrador

de sistemas podrá prevenir ataques, como por ejemplo, los de denegación de servicio, DDoS, *Distributed Denial of Service* [7]. Un administrador de sistemas deberá de garantizar que el recurso que facilita y gestiona esté operativo siempre, por lo tanto analizará el tráfico para evitar los posibles ataques, virus, etc.

Existen otras técnicas de clasificación de tráfico, como se puede ver en el artículo del departamento [8], estas técnicas son:

- Clasificación basada en los puertos de la capa de transporte. [9]
- Clasificación basada en el contenido del paquete. [10]
- Clasificación basada en la aplicación de técnicas de aprendizaje automático sobre estadísticas de tráfico. [11]

Hoy en día, las dos primeras técnicas mencionadas no son muy eficaces, pues por ejemplo:

- La primera técnica puede ser burlada fácilmente, haciendo que las conexiones pasen por el puerto 80.
- En la segunda técnica, al tener que acceder a los paquetes, solo bastará con que estos estén encriptados para que no se pueda acceder a dicho paquete. Y también se debe de tener en cuenta que al acceder a los paquetes de una red que no es del propietario de la información es un delito y no respeta la privacidad.

Ahora bien, la tercera técnica si es eficaz y fiable, pero su principal inconveniente es que necesita cierto tiempo para aprender a clasificar de una forma correcta el tráfico.

Por lo tanto es necesario un nuevo método de clasificación, que sea funcional en cualquier circunstancia y sea eficiente desde el principio. Por ello, el método de emparejamiento planteado por el departamento [8] es muy atractivo al ser necesario saber solo las IP's y los puertos, ya que se respeta la seguridad, pues no es necesario acceder a los paquetes para conocer estos detalles. Incluso no sería necesario ser un proveedor de servicios de Internet, ISP, para conocer estos datos.

1.2. Objetivos

En este trabajo se hará uso del NMS mencionado anteriormente, llamado Bro [6], el cual cuenta con su propio lenguaje de scripting. Dicho lenguaje es de fácil uso y comprensión, además de que está completamente orientado a

trabajar con conexiones, teniendo distintos tipos de datos que son realmente útiles. Dicho esto el objetivo del trabajo es desarrollar un módulo para Bro. En este módulo irá implementada la técnica de emparejamiento de flujos, de forma que pueda ser usada en redes fuera de un entorno de laboratorio.

- Para ello será necesario conocer como Bro gestiona los flujos, desde su nacimiento a su muerte.
- También hará falta gestionar las entradas y salidas, aunque al principio se hará uso de un archivo *pcap* de prueba para controlar si todo funciona bien, pudiendo pasar luego a analizar el tráfico de una red online.
- Implementar la función de emparejamiento de flujos.
- Será necesario realizar pruebas al módulo que se pretende desarrollar, de modo que se detecten los errores y se subsanen.

Por último, y fuera de los objetivos más técnicos, el desarrollo de esta memoria también se considera un objetivo, que se tendrá en cuenta dentro de la temporización.

1.3. Metodología

Para realizar este trabajo se establecen una serie de tareas:

- Lectura del artículo del departamento. [8]
- Estudio del uso de Bro y de su lenguaje de scripting.
- Pruebas con el lenguaje propio de Bro.
- Desarrollo del módulo en Bro.
- Pruebas del módulo.
- Desarrollo de la memoria.

Cada tarea anterior se realimentará de la siguiente, el ejemplo más claro es el desarrollo del módulo, que irá cambiando por las pruebas, pues se detectarán los fallos, a su vez habrá que releer el artículo mientras se desarrolla el módulo para asegurarse de que todo se realiza de acuerdo al trabajo previo, etc.

En el siguiente diagrama de Gantt se puede ver una temporización de estas tareas.

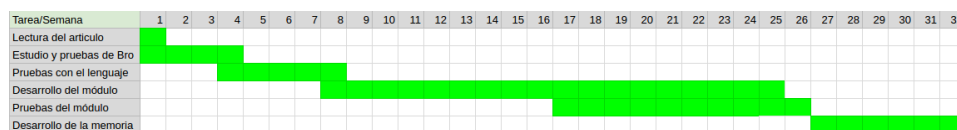


Figura 1.3: Temporización en diagrama de Gantt.

Este proyecto no necesita de ningún gasto, pues Bro [6] proporciona sus binarios de forma gratuita. El módulo será subido a GitHub [12] con licencia de software libre, por lo que cualquiera podrá usarlo o modificarlo en el futuro. El único *gasto* es el de un ordenador con Linux, o bien con Mac OS X, pues de momento no está disponible para Windows. [13]

1.4. Estructura de la memoria

Esta memoria se organizará de la siguiente forma:

- En el capítulo 2 se hablará de todos los fundamentos teóricos y tecnológicos sobre los que se basa el proyecto.
- En el capítulo 3 se contará cómo se pretende resolver el problema expuesto.
- En el capítulo 4 se encontrará detallado cómo se han implementado los diferentes módulos.
- En el capítulo 5 se realizarán las pruebas, para comprobar que todo funciona como estaba previsto, tanto a nivel funcional como a nivel de aplicación.
- En el último capítulo se hablará de las conclusiones recogidas a lo largo de este proyecto y las posibles opciones que tiene para seguir trabajando sobre él.

Capítulo 2

Estado del arte

Aquí se encuentra la parte teórica y tecnológica del proyecto. Se hablará de la identificación de tráfico y de las distintas técnicas que existen para ello. También se hablará de Bro [6], el analizador de tráfico que se usará para llevar a cabo el desarrollo del módulo. Se introducirá en que consiste, cómo gestiona los eventos y sus funcionalidades básicas, así como la posibilidad de ampliar estas.

Por último se presentará de forma teórica en que consiste el emparejamiento de flujos y de qué forma se podría usar para identificar el tráfico.

2.1. Identificación de tráfico

Hay que tener claro que identificar el tráfico no es clasificarlo. Identificar el tráfico consiste en analizarlo y obtener patrones comunes con los cuales se pueda ordenar el tráfico de forma posterior. Algunos de los patrones que se pueden tener en cuenta a la hora de identificar tráfico son los puertos, las IP's o la clase a la que pertenece.

Por lo tanto la identificación del tráfico de la red es esencial para realizar una clasificación posterior. Dicha clasificación se podrá analizar en busca de amenazas o intrusos, a nivel de seguridad y componentes lentos o defectuosos que afecten al rendimiento del servicio. Para poder identificar el tráfico lo deseable es un método que sea rápido y proteja la privacidad de los distintos usuarios.

2.1.1. Técnicas de identificación de tráfico

Existen varias técnicas de identificación de tráfico, como ya se explicó anteriormente. Estas técnicas son el paso previo a la clasificación del tráfico. Podría considerarse que son el paso intermedio a la clasificación, teniendo

en cuenta que el tráfico es capturado, identificado y clasificado. El tráfico, hasta ahora, se puede identificar de la siguiente forma.

- Por los puertos de la capa de transporte, establecido por *IANA*. [9]
- Por el contenido del paquete o *DPI*. [10]
- Por la aplicación de técnicas de aprendizaje automático sobre estadísticas de tráfico, *machine learning*. [11]

En los siguientes apartados se verán estas técnicas de una forma más detallada.

2.1.2. Identificación de tráfico basada en flujos

Dentro de este ámbito se encuentran dos de las tres técnicas que se han presentado anteriormente.

- Identificación basada en los puertos de la capa de transporte. [9]

Esta técnica consiste en identificar el tráfico dependiendo de los puertos de la capa de transporte, fue diseñada por *IANA* [14]. Esto se debe a que *IANA* es la organización que se encarga de asignar los protocolos de Internet, algunos nombres de dominios y coordinar la asignación de direcciones IP.

Por lo tanto, *IANA* es quien asigna el número de puerto a los distintos protocolos, haciendo que se pueda identificar el tráfico por el número de puerto al que se envía. Este sistema tiene el inconveniente de que ahora mismo, se están desarrollando multitud de aplicaciones web. Estas aplicaciones se conectan mediante el puerto 80, es decir, mediante el protocolo HTTP. Pero esto es solo teoría, pues se puede hacer que cualquier aplicación envíe información por el puerto 80, sin que sea HTTP, lo cual daría como resultado una mala identificación y posterior clasificación.

- Identificación basada en la aplicación de técnicas de aprendizaje automático sobre estadísticas de tráfico [11].

Esta técnica de identificación es muy sencilla. Consta de algoritmos de aprendizaje automático, los cuales se dedican a realizar análisis estadísticos. De esta forma los algoritmos aprenden que características tienen los flujos y los identifican en función de estas. Por lo tanto cuanto más tiempo pase, los patrones serán mejores y se tendrán más aciertos a la hora de la identificación.

2.1.3. DPI

- Identificación basada en el contenido del paquete o DPI. [10]

Esta técnica de identificación consiste en analizar los paquetes, más allá de su cabecera. La Inspección Profunda de Paquetes, DPI por sus siglas en inglés, *Deep Packet Inspection*, esta técnica lo que hace es analizar todos los paquetes en busca de cadenas que se repitan, además de las IP's y los puertos. Una vez que tiene una cadena, la usa para buscar paquetes que tienen ese patrón, de esta forma establece un patrón de identificación.

Esta técnica es útil para buscar virus, intrusiones y demás. Suele ser usada por los proveedores de servicios de Internet, ISP, *Internet Service Provider* y grandes empresas. Se puede decir que es una técnica que no respeta la privacidad, pues entra dentro de la propia información del paquete. Aunque en el caso de una gran empresa si puede hacerse, pues si se está en la red propia no es delito mirar la información que contienen los paquetes.

En algunos países, como Alemania, donde es ilegal el uso de *torrents*, las autoridades pueden pedir información al proveedor de servicios sobre uno de sus clientes, si cree que este está infringiendo la red, por lo que esta técnica es ideal para este tipo de cometidos.

Además de para lo descrito, esta técnica también es ideal para saber si se está siendo atacado y por quien. [15]

2.2. BRO

Bro [6], es un analizador de tráfico de red de código abierto. Al ser de código abierto permite ser usado por quien lo desee sin necesidad de pagar licencias. Funciona sobre sistemas basados en Linux y Mac OS X. Una de sus principales características es la gran cantidad de información que devuelve con un solo escaneo. Otros monitores de red devuelven menos información, teniendo que ser el propio administrador el que analice después la información obtenida por estos. Por lo tanto tardará más en resolver los posibles problemas que encuentre en la red.



Figura 2.1: Logo de Bro.

No tiene interfaz gráfica, por lo que su gestión se realiza desde la línea de comandos. Cuando se analiza tráfico, Bro generará unos registros, los cuales están divididos en función de parámetros definidos por parte del equipo que lo ha desarrollado.

Bro incorpora la posibilidad de introducir funcionalidades nuevas, mediante la programación en su propio lenguaje de *scripting*, del mismo nombre. Esto se desarrollará más adelante y en la parte de implementación se podrá ver cómo se desarrolla y algunos ejemplos de código escrito en Bro.

El lenguaje de *scripting* está orientado a trabajar con eventos. Por lo tanto a la hora de añadir una nueva funcionalidad habrá que crearla a partir del uso de los eventos que puede gestionar el programa.

Bro está estructurado de forma que todos los flujos de paquetes que analiza son procesados por el motor de eventos. Este motor convierte los flujos de paquetes en eventos, de forma que es más sencillo trabajar con ellos. Una vez que los eventos son procesados se generan los registros correspondientes, los cuales podrán ser analizados posteriormente. [16]

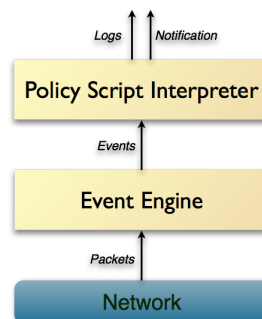


Figura 2.2: Arquitectura de Bro.

2.2.1. Funcionalidades básicas de BRO

La funcionalidad básica de Bro es la monitorización de la red en la que se ejecuta. Mientras que se encuentra en ejecución, genera registros o *logs* en texto plano que se podrán leer usando un editor de texto. Si se analiza un archivo *pcap* los *logs* no cambiarán. Sin embargo si se analiza tráfico a tiempo real, los registros se irán actualizando a medida que pase el tiempo. Algunos de los *logs* que se generarán son los siguientes.

- *dpd.log*. Consiste en un resumen de los protocolos encontrados en puertos que no son estándar.
- *dns.log*. Contendrá toda la actividad correspondiente al *DNS*.
- *ftp.log*. Un registro de la actividad a nivel de sesión de *FTP*.
- *files.log*. Un resumen con los archivos transferidos a través de una red. Incluye protocolos *HTTP*, *FTP* y *SMTP*.
- *http.log*. Registro de toda la actividad *HTTP* con sus réplicas.
- *ssl.log*. Un registro de las sesiones *SSL*, incluidos los certificados que se utilizan.
- *weird.log*. En este *log* se guarda la información correspondiente a actividad inesperada o rara a nivel de protocolo. Al analizar gran cantidad de tráfico no es muy útil, pues ocurren muchas cosas raras, pero a pequeña escala es bastante interesante para detectar intrusiones y demás.
- *conn.log*. Aquí se puede ver la información correspondiente a conexiones *TCP*, *UDP* e *ICMP*.

Para más información sobre *logs* generados por una monitorización de Bro lea [17].

Bro dispone además de varios *frameworks* que extienden su funcionalidad. Con ellos se podrá crear *scripts* muy potentes. Algunas de las utilidades de estos *frameworks* son las siguientes.

- *Geolocalización*. Se podrá encontrar la localización geográfica de una IP.
- *Análisis de ficheros*. El monitor de red tiene la capacidad de trabajar con ficheros.
- *Framework de loggins*. Con este *framework* se podrá extender los archivos de registro que se generan.

- *NetControl*. Este *framework* permitirá tener un control muy amplio y diverso del tráfico.

Para mayor conocimiento de estos *frameworks* y otros lea [18].

Ahora con toda esta información de los registros, el administrador del sistema podrá determinar si existen amenazas en la red o si hay algún componente defectuoso. Para ello deberá de hacer uso de los eventos con los que Bro trabaja, para sacar mayor partido a todo el trabajo que se realice sobre la red.

2.2.2. Eventos y trazas

La programación en Bro no es secuencial, no se escribe un *script* y se espera que se ejecute en el orden en el que está escrito. La programación aquí esta orientada a eventos. Por lo tanto el orden depende de cuando se ejecute un determinado evento en el análisis.

Un evento se da cuando Bro detecta un “comportamiento”, por ejemplo cuando detecta un paquete de respuesta *UDP*. Si se programa para que ese evento muestre un mensaje, este se mostrará cada vez que se detecte un paquete de respuesta *UDP*. Por lo tanto la programación deberá de estar pensada en función de los eventos disponibles.

Dentro de cada evento se podrá hacer lo que se desee, con la información que capture ese evento. Si el evento captura información de una conexión *TCP*, se tendrá que trabajar con esa información y las distintas variables globales que se hayan definido previamente.

Una traza es una captura del tráfico de una red. Con casi cualquier programa de diagnóstico se puede realizar capturas de red. En la propia web de Bro se encuentran algunos archivos de trazas de red, en formato *pcap*, de forma que se puedan realizar pruebas sobre ellos sin necesitar realizar una traza propia.

En el caso de Bro se podrá generar trazas de red mediante un *script* de *Python* [19], el cual se puede descargar desde el repositorio de *GitHub* de Bro. Con dicho script también se podrá trabajar sobre la traza, separando el tráfico entrante del saliente, desglosando en distintos registros el tipo de tráfico y demás.

2.2.3. Incorporación de funcionalidades

La incorporación de funcionalidades al monitoreo realizado por Bro es una característica muy llamativa. Gracias a esto se podrá realizar un análisis

muy personalizado usando un *script* creado por el administrador de redes. De esta forma podrá, por ejemplo, filtrar el tráfico de una determinada IP mientras se sigue analizando el tráfico de forma normal, con los registros que genera Bro de forma automática. Es una forma muy sencilla de comprobar si por ejemplo el servicio que administra está recibiendo demasiadas peticiones desde una misma IP. Lo cual sería un indicio de ataque de denegación de servicio.

A la hora de incorporar funcionalidades a Bro se puede hacer todo lo que se desee. Una búsqueda rápida por *GitHub* arrojará una gran cantidad de personas que contribuyen con una gran cantidad de nuevas funcionalidades [20]. Ahora lo ideal sería incorporar un módulo, de forma que si el resto de la comunidad lo desea pueda hacer uso de él de una forma sencilla.

2.3. Emparejamiento de flujos

La técnica de emparejamiento de flujos, fue planteada por el departamento de Teoría de la Señal, Telemática y Comunicaciones, *TSTC*, de la Universidad de Granada, en el año 2011 [21] [8].

En el emparejamiento de flujos, la idea de la que se parte es muy sencilla. Si dos paquetes comparten IP's de origen y destino, y puertos de origen y destino, se podrá decir que esos dos paquetes pertenecen a la misma clase. Una vez que están identificados como pertenecientes a la misma clase, lo que se debe de hacer es comparar los tiempos de esos paquetes, de forma que sean coherentes en el flujo. Si son muy lejanos en el tiempo, se podrá descartar que pertenezcan al mismo flujo. Esta idea se puede ver en [8] de una forma mucho más técnica.

En el mismo artículo [8] se encuentra la fórmula que se sigue para el emparejamiento.

$$F(x, y) = \begin{cases} G(x, y), & NIP(x, y) \geq 1 \\ -\infty, & \text{en otro caso} \end{cases}$$

Donde tenemos que $G(x, y)$ se corresponde con la siguiente función:

$$G(x, y) = |NIP(x, y) - 1| + 1 / (dp1(x, y) + k1) + 1 / (dp2(x, y) + k1) + 1 / (dt(x, y) + k2)$$

Las variables de la función son las siguientes:

- x : Primer paquete a comparar.

- y : Segundo paquete a comparar.
- $NIP(x,y)$: Es el número de paquetes analizados con las mismas características.
- $dp1(x,y)$: Se corresponde con los puertos de origen de los dos paquetes.
- $dp2(x,y)$: Serán los puertos de destino de los dos paquetes.
- $k1$: Es una variable definida previamente.
- $k2$: Es otra variable definida antes del comienzo del análisis. En trabajos anteriores esta variable y la anterior suelen estar definidas entre 1 y 10000.
- dt : Es la diferencia de tiempo existente entre los tiempos de inicio de los paquetes o *timestamps*.

Con esta fórmula se obtendrá como resultado un número, el cual será comparado con un umbral que se define previamente. Si el umbral definido es pequeño tendremos más flujos emparejados. Si el umbral es mayor serán menos los flujos que serán emparejados al no cumplir los requisitos.

El emparejamiento de flujos, por si mismo, sólo identifica el tráfico. Por lo tanto se necesitará otra técnica para clasificar el tráfico.

Capítulo 3

Diseño y arquitectura del sistema

En este capítulo se hablará cómo se va a estructurar el módulo, las diferentes funcionalidades que debe tener para que sea funcional y por último cómo Bro gestiona los flujos.

Todo esto se planteará desde la vista del diseño, se verán algunos tipos de datos interesantes para el desarrollo, aunque el código se verá más adelante en el capítulo de implementación. También se hablará de la arquitectura que va a seguir el sistema.

3.1. Arquitectura del sistema

Para describir la arquitectura del sistema hay que tener en cuenta la arquitectura que compone a Bro. Este monitor de red es un software modular, esto quiere decir que esta compuesto de diferentes módulos que al ser ejecutados funcionan como un único sistema.

Por lo tanto la arquitectura de este problema se debe de acoplar a la arquitectura propia de Bro. Para ello, se construirá un módulo que se acople con el monitor de red. Entre los requisitos del módulo a desarrollar se encuentra que sea ligero y eficiente. Por lo tanto se tendrá que usar los distintos eventos que proporciona Bro para crearlo de la forma deseada.

Se prescindirá del uso de los *frameworks* descritos en el capítulo anterior, pues su uso no aporta nada que no se pueda realizar exclusivamente con los eventos destinados a gestionar el tráfico de la capa de transporte. Se tendrá que usar los eventos para gestionar absolutamente todo lo que pasa en el módulo, a parte también se tendrá que usar una función propia. Esta función se encargará de aplicar la fórmula del emparejamiento de flujos vista en el

último apartado del capítulo anterior, esta función devolverá un número que será el que se compare con el umbral.

Por lo tanto, a modo de resumen, la arquitectura final del sistema que se plantea será un módulo compuesto por los distintos eventos destinados a la gestión del tráfico de la capa de transporte y una función auxiliar para el cálculo del umbral de emparejamiento.

3.2. Módulo y funciones

Dado que una mala programación del módulo puede dejar al monitor de red colgado con una traza sencilla. Se tratará que el módulo contenga solamente lo justo para llevar a cabo la identificación del tráfico. De esta forma se obtendrá un módulo cuya ejecución será óptima.

Las funcionalidades que se espera que tenga este módulo en esencia son dos, la detección y almacenado del tráfico y la aplicación de la fórmula para conocer si dos flujos son emparejables, a los distintos flujos que se han detectado. De una forma más amplia las funciones del módulo serán las siguientes.

- *Función que aplique la fórmula de emparejamiento.*

A esta función se le pasará dos flujos, de forma que se aplique la fórmula y devuelva un número, el cual será el que indique si los flujos son emparejables o no.

- *Funciones que detecten el tráfico.*

Esto se hará con los eventos de Bro. Los eventos detectarán el tipo de tráfico que se está analizando y aplicarán la función anterior.

Tras el uso de esta fórmula se almacenará o no el flujo que está siendo analizado. Por lo tanto será necesario el uso de algún tipo de contenedor para este cometido.

Lo que se espera es capturar el tráfico de la capa de transporte. Dicho tráfico se corresponde a los protocolos *TCP* y *UDP*. Por lo tanto será necesario ver que tipo de eventos son los que controlan el tráfico de estos dos protocolos. Esto se verá de forma más amplia en la siguiente sección.

Las entradas y salidas son de fácil gestión. Las entradas de tráfico podrán ser mediante archivos o analizando directamente el tráfico de la red. Las salidas, por su parte, serán mediante terminal. Esto puede suponer cierto inconveniente si se obtienen demasiadas salidas. También se pueden guardar las salidas en un fichero mediante el carácter *mayor que* . Con esto se guardará en un fichero en la ruta que se especifique, siendo el posterior análisis mucho más cómodo desde, por ejemplo, un editor de texto.

Se debe de tener en cuenta que siempre se podrá extender la funcionalidad del módulo. Pero de momento no resulta interesante. La posible extensión correspondería a posibles trabajos futuros.

Para realizar este módulo es necesario conocer como gestiona los flujos Bro y de que forma se mantendrán los que son emparejados y los que están activos.

3.3. Gestión de flujos

La gestión de flujos en Bro se realiza completamente con eventos. Por lo cual se tendrán que crear variables globales para el almacenamiento de los flujos que sean emparejables y los que estén activos.

El *nacimiento de un flujo* es controlado por un evento. Por lo tanto cuando se detecta un nuevo flujo se lanza un evento. Este evento se tendrá que controlar de forma que si se tiene ya un flujo activo con las mismas características, se compare y se almacene. Si por el contrario no se tiene ningún flujo con esas características se tendrá que almacenar directamente en el contenedor de flujos activos.

La *muerte de un flujo* también es controlada por un evento. Ahora lo importante es si es interesante a nivel del análisis seguir almacenando los flujos aunque estos hayan muerto. Si no se quiere tener almacenados flujos muertos se tendrá que eliminar de la estructura en la que está almacenado. Si se quiere seguir trabajando con ellos habrá que mantenerlos guardados en la estructura. Obviamente si el flujo que va a morir está emparejado con otro se borrará solo del contenedor de flujos activos. Manteniéndose en el contenedor de flujos emparejados. De lo contrario se perderá información. Todo esto habrá que decidirlo en el evento que gestiona la muerte del flujo.

Estos dos comportamientos de los flujos están controlados por eventos genéricos. Con un único evento se detecta que el flujo ha nacido y con otro evento si ha muerto, independientemente del tipo de protocolo al que pertenezca. No pasará esto con los distintos estados de los flujos que serán detectados. Pues no es lo mismo detectar un ACK de un flujo TCP que una

respuesta de un flujo UDP. Serán tratados en eventos distintos y tendrán que ser gestionados con eventos distintos, cada uno destinado a un protocolo distinto.

A continuación se verán las estructuras de datos necesarias para llevar a cabo el desarrollo del módulo.

3.4. Estructuras de datos

Las principales estructuras de datos que se necesitan para el desarrollo de este trabajo serán dos tablas, *table* [22], de vectores para el almacenamiento de los flujos activos y los emparejados. Los vectores son iguales que en cualquier otro lenguaje de programación, mientras que las tablas son parecidas a los *maps* de *C++*. Aunque esto se podrá ver con mejor detalle en el apartado de implementación.

Dichas estructuras de almacenamiento, deberán de ser capaces de estar ordenadas por las IP's y los puertos. Lo cual se obtiene juntando las tablas con los vectores para así conseguir una especie de matriz bidimensional, la cual está indexada. Por lo tanto se consigue que el acceso a los datos sea mucho más rápido. Incluso se podría prescindir de bucles, los cuales pueden acabar siendo un problema en cuanto a rendimiento si se llega a almacenar muchos flujos.

Bro proporciona cierto tipos de datos muy interesantes, los cuales además, incluyen mucha más información. Algunos de estos tipos de datos son.

- *connection*.

Este tipo de dato es el flujo en si. Por lo tanto será de vital importancia comprenderlo para poder trabajar como se desea.

Dentro del tipo de dato *connection* existe un registro llamado *id*, el cual esta compuesto por el tipo de dato *conn_id* [23]. Este dato sirve para identificar los flujos mediante una tupla formada por 4 datos. Estos datos son los que se precisan para indexar la matriz bidimensional, siendo pues las IP's y los puertos.

- *addr*. Este tipo de dato representa una IP. Reconoce tanto IPv4 como IPv6. Este tipo de dato puede ser comparado e incluso ordenado mediante operadores. [24]
- *port*. Este tipo es el usado para los puertos. Además del número de puerto también indica el protocolo de la capa de transporte que usa. Soporta la comparación y ordenación, pero no por el número, sino por el tipo de protocolo. [25]

Para obtener más información sobre el tipo de dato *connection* lea [26].

- *time*.

Este tipo de dato también es interesante. Aunque en otros lenguajes se puede obtener, en Bro es un tipo de dato por si mismo. Por lo tanto se podrá operar sobre él desde el principio, siendo una gran ventaja a la hora de calcular el tiempo de inicio de los flujos. Para leer más [27].

Es importante entender que para realizar el cálculo para el emparejamiento de flujos, se necesita el *timestamp* del primer paquete de cada flujo, pues será sobre este tiempo sobre el que se apoye el cálculo del emparejamiento.

Estos dos tipos de datos a parte de ser los más interesantes para el cálculo del emparejamiento, también serán los más utilizados junto a los contenedores para los flujos. Existen más tipos de datos e incluso los hay que extienden la información disponible sobre los flujos. Para leer más sobre esto [28].

Capítulo 4

Implementación

A continuación se contará cómo se ha implementado el módulo de Bro, resolviendo así el problema planteado.

La descripción del módulo será sin entrar en detalles de la programación. Se trata de una descripción como la descripción de una API.

Lo primero que se va a detallar es la función para calcular el emparejamiento.

```
1 function emparejamiento(c1: connection , c2: connection ):double
```

Esta función recibe como entrada dos flujos y devuelve un número de tipo *double*. En esta función lo que se hace es aplicar la fórmula de emparejamiento a los dos flujos que entran.

Al querer hacer la comparación lo que se hace es sacar las IP's de origen y destino y los puertos de origen y destino de los flujos. También se obtendrán los *timestamps* de los flujos, de esta forma se conseguirá la diferencia de tiempo.

Para poder operar con los puertos habrá que pasarlos a tipo *count*, de esta forma se elimina la terminación con el tipo de protocolo del puerto. Se comentó anteriormente que se podía operar con esta terminación pero al tener que operar con otros tipos de datos que no son puertos es mejor quitar la terminación, pues de lo contrario no se obtendrá un buen resultado.

Lo mismo que con los puertos pasa con el tipo *time*. Se puede operar con el tipo de dato *time* pero no es recomendable hacerlo ya que se va a trabajar con más tipos de datos de carácter matemático.

El número de veces que se tiene un flujo con los mismos datos o *Nip*, en la fórmula, es de fácil cálculo. Gracias a la indexación basta con buscarlo en la tabla y calcular el tamaño del vector.

A continuación se procederá a la explicación de los distintos tipos de eventos usados y para que son usados.

```
1 event new_connection(c: connection)
```

Este evento recibe como entrada un flujo nuevo. Este flujo es una nueva conexión, la cual no está identificada previamente. Esto quiere decir o que es nueva o que ha sido borrada.

Este evento no devuelve nada, por lo tanto si en el momento en el que es detectado se tendrá que crear un nuevo índice o guardarlo en las tablas.

Este tipo de evento detecta las conexiones de tipo *TCP* y *UDP*.

Ahora se verá cómo se van a gestionar la muerte de los flujos.

```
1 event connection_state_remove(c: connection)
```

Este evento se activa cuando el flujo que entra en él va a morir, o ser borrado de la memoria. Es un flujo que ya ha sido procesado por el módulo.

Lo que se realiza dentro de este evento es buscar en la tabla el índice el vector. De esta forma se borrará el primer flujo almacenado en el vector.

De ser el único flujo almacenado en el vector, se borrará el vector entero. Si hay más flujos almacenados se moverán los demás una posición hacia atrás. De esta forma se seguirá teniendo un rendimiento óptimo.

A continuación se verán los distintos eventos que van a detectar el diferente tipo de tráfico.

```
1 event connection_established(c: connection)
2
3 event connection_finished(c: connection)
4
5 event udp_request(u: connection)
6
7 event udp_reply(u: connection)
```

Los dos primeros eventos son los correspondientes al tráfico *TCP*. Los otros dos, como se puede ver en el nombre están destinados al tráfico *UDP*.

El primer evento relacionado con *TCP* se activa cuando se detecta un paquete *SYN-ACK* que responde al *handshake* que se realiza en las conexiones de este tipo.

El segundo evento detecta cuando la conexión *TCP* finaliza de forma normal.

Los dos eventos relacionados con *UDP* detectan paquetes de dos tipos distintos.

- *UDP request*. Se genera por cada paquete que es enviado por el creador del flujo.
- *UDP reply*. Este es generado por cada paquete que es enviado por el receptor del flujo.

Estos dos últimos eventos son bastante costosos, pero son absolutamente necesarios, son los únicos eventos que detectan conexiones de tipo *UDP*.

Dentro de todos estos eventos se realiza lo mismo. Primero se comprueba que los dos flujos no son el mismo, si son el mismo se termina el análisis. Si no son lo mismo se pasa a la función de emparejamiento ya descrita y se compara el número que se obtiene con el umbral que se ha definido antes de la ejecución. Si el resultado es mayor que el umbral son emparejables, por lo cual se tendrá que guardar en la tabla de emparejados y se informa mediante un mensaje en pantalla de que lo son. Si es menor que el umbral no son emparejables y mediante un mensaje en pantalla se informa de que no son emparejables.

Es necesario recordar que se puede dar que se puede estar usando dos eventos o más a la vez, por lo que los mensajes en pantalla pueden ser confusos. Se tiene que tener en cuenta que hay cierto retardo en los mensajes, por lo que lo interesante es el resultado final de las tablas.

Además de estos eventos se tienen los siguientes eventos genéricos.

```
1 event bro_init()
2
3 event bro_done()
```

El primero se lanza cuando Bro se inicia y mostrará el tiempo de inicio. El segundo se lanza cuando Bro finaliza, por lo tanto es el último evento que se lanzará y mostrará la hora de finalización. Con estos dos eventos se tendrá control de cuando fue lanzado el análisis y cuando finalizó.

Capítulo 5

Evaluación y pruebas

Capítulo 6

Conclusiones y trabajo futuro

Bibliografía

- [1] James F. Kurose and Keith W. Rose. *Redes de computadores un enfoque descendente*. Pearson, 2010.
- [2] James F. Kurose and Keith W. Rose. *Redes de computadores un enfoque descendente*, chapter 2. Pearson, 2010.
- [3] James F. Kurose and Keith W. Rose. *Redes de computadores un enfoque descendente*, chapter 3. Pearson, 2010.
- [4] James F. Kurose and Keith W. Rose. *Redes de computadores un enfoque descendente*, chapter 4. Pearson, 2010.
- [5] James F. Kurose and Keith W. Rose. *Redes de computadores un enfoque descendente*, chapter 1. Pearson, 2010.
- [6] Bro Team. Bro indice, . URL <https://www.bro.org>.
- [7] James F. Kurose and Keith W. Rose. *Redes de computadores un enfoque descendente*, page 55. Pearson, 2010.
- [8] José Camacho, Pablo Padilla, Pedro García-Teodoro, and Jesús Díaz-Verdejo. A generalizable dynamic flow pairing method for traffic classification. 2013.
- [9] Internet Assigned Numbers Authority (IANA). Service name and transport protocol port number registry. URL <https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml>.
- [10] Christian Fuchs. Implications of deep packet inspection (dpi) internet surveillance for society. URL <http://fuchs.uti.at/wp-content/uploads/DPI.pdf>.
- [11] Thuy T.T. Nguyen and Grenville Armitage. A survey of techniques for internet traffic classification using machine learning. URL <http://ieeexplore.ieee.org/document/4738466/?reload=true>.

- [12] Álvaro Maximino Linares Herrera. Repositorio del trabajo con bro. URL <https://github.com/Lynares/bro-flows>.
- [13] Bro Team. Descarga de bro, . URL <https://www.bro.org/download/index.html>.
- [14] Kim Davies. Una introducción a iana. URL <https://www.iana.org/about/presentations/davies-atlarge-iana101-paper-080929-es.pdf>.
- [15] Dr. Thomas Porter. The perils of deep packet inspection. URL <https://www.symantec.com/connect/articles/perils-deep-packet-inspection>.
- [16] Bro Team. Arquitectura de bro, . URL <https://www.bro.org/sphinx/intro/index.html#architecture>.
- [17] Bro Team. Logs de bro, . URL <https://www.bro.org/sphinx/script-reference/log-files.html>.
- [18] Bro Team. Frameworks de bro, . URL <https://www.bro.org/sphinx/frameworks/index.html>.
- [19] Bro Team. Trazas en bro, . URL <https://www.bro.org/sphinx/components/trace-summary/README.html>.
- [20] securitykitten. Finding beacons with bro. URL <https://gist.github.com/securitykitten/a7edcee0932c556d5e26>.
- [21] José Camacho, Pablo Padilla, F. Javier Salcedo-Campos, Pedro Garcia-Teodoro, and Jesús Diaz-Verdejo. Pair-wise similarity criteria for flows identification in p2p/non-p2p traffic classification. 2011.
- [22] Bro Team. Tablas en bro, . URL <https://www.bro.org/sphinx/script-reference/types.html#type-table>.
- [23] Bro Team. Tipo conn-id de bro, . URL https://www.bro.org/sphinx-git/scripts/base/init-bare.bro.html#type-conn_id.
- [24] Bro Team. Tipo addr de bro, . URL <https://www.bro.org/sphinx-git/script-reference/types.html#type-addr>.
- [25] Bro Team. Tipo port de bro, . URL <https://www.bro.org/sphinx-git/script-reference/types.html#type-port>.
- [26] Bro Team. Tipo cennnection, . URL <https://www.bro.org/sphinx/scripts/base/init-bare.bro.html#type-connection>.
- [27] Bro Team. Tipo time, . URL <https://www.bro.org/sphinx/script-reference/types.html?highlight=time#type-time>.

-
- [28] Bro Team. Tipo conn, . URL <https://www.bro.org/sphinx/scripts/base/protocols/conn/main.bro.html#type-Conn::Info>.
 - [29] Bro Team. Web de bro, . URL <https://www.bro.org/sphinx/intro/index.html>.
 - [30] Bro Team. Instalación de bro, . URL <https://www.bro.org/sphinx/install/index.html>.
 - [31] Bro Team. Función get_port_transport_proto, . URL https://www.bro.org/sphinx/scripts/base/bif/bro.bif.bro.html#id-get_port_transport_proto.
 - [32] Bro Team. Analizadores de protocolos, . URL <https://www.bro.org/sphinx/script-reference/proto-analyzers.html>.

