Trabajo Práctico N^o 1: Wiretapping

Leandro Ezequiel Barrios, Gonzalo Benegas, Martin Caravario, Pedro Rodriguez

Resumen —
En el presente Trabajo Práctico utilizaremos algunas de las técnicas provistas por la teoría de la información para estudiar y analizar algunas redes de información. El objetivo será distinguir diversos aspectos de la red de manera analítica. Para cumplir con nuestro objetivo, haremos uso de dos herramientas modernas de manipulación y análisis de paquetes: Wireshark y Scapy.

I. Introducción

Construimos una herramienta que hace uso de la función **''sniff''**, provista por la librería **Scapy** de Python. Esta nos permitió activar el `modo promiscuo`, o `monitor` en el caso de las placas wireless. Esto nos permitió escuchar la red durante cierto tiempo, obteniendo todos los paquetes que llegaban a nuestra placa de red. A partir de estos datos que guardamos en un archivo `pcap`, se definieron dos fuentes de información, con las cuales fuimos capaces de encontrar nodos y protocolos distinguidos en la red. Para su visualización, elegimos la realización de gráficos de torta e histogramas, ya que los consideramos los más apropiados.

Para cada una de las mediciones consideramos las siguientes fuentes:

- $S = \{s_1 \dots s_n\}$, provista por la cátedra, donde s_i es el valor del campo *type* de cada paquete de capa 2.
- $S_1 = \{s_1 \dots s_n\}$, determinada por nosotros, donde s_i es el valor del campo destino (MAC) cada paquete de capa 2 de tipo ARP.

Para entender qué es lo que se obtendrá al efectuar estas mediciones, hay que aclarar que `ARP` es un protocolo de la capa de enlace de datos, responsable de encontrar la dirección de capa 2 (`Ethernet MAC`) que corresponde a una determinada dirección IP (dirección de capa 3 de red).

Es decir, cada vez que un host quiere comunicarse con otro y su dirección `MAC` no se encuentra dentro de su tabla `ARP`, debe enviar un paquete who-has broadcast para determinar la dirección `MAC` del host destino. De este modo, todos los hosts del dominio de colisión de la máquina en la que se efectúa la medición reciben dicho paquete, siendo respondido el mismo únicamente por el host requerido, mediante un paquete *reply*.

Para distinguir *nodos* (símbolos) en este contexto, tomamos a aquellos cuya probabilidad de aparición era alta, de forma tal que la información provista por el mismo fuera menor a la entropía de la fuente a la cuál el símbolo pertenecía.

Tomamos esta decisión porque, según Shannon, el nivel de entropía de una fuente habla de la máxima compresibilidad de cada bit en un mensaje enviado con una codificación óptima ($H(F) \leq L(C)$). Luego, si la fuente presenta símbolos cuya cantidad de información está por debajo de la entropía, son símbolos que tienen mucha probabilidad de aparecer en un mensaje en comparación con los otros, y podría ser conveniente representar a estos símbolos con menos bits que al resto, para así disminuir la longitud media del código. Por ejemplo, supongamos que enviamos números binarios, y sabemos que el símbolo $X =$ "00000000" aparece en los mensajes la mitad del tiempo y que el resto de las tiras pueden aparecer todas con la misma probabilidad. Entonces, X brindaría una cantidad de información por debajo de la entropía, y convendría representar a esa tira simplemente con un 0, y al resto de las tiras prefijarlas con el 1. De esta manera, se ahorraría en promedio $7.1/2 + (-1).1/2 = 3$ bits en cada mensaje.

Entonces, también analizaremos las entropías de las fuentes en cada una de las escuchas de red realizadas, y trataremos de concluir cuál de las dos fuentes elegidas es *más compresible*.

A. Casa

Esta medición fue realizada en una red LAN hogareña, en el horario de las 15:00 hs, por un lapso de 3 hs. La herramienta utilizada fue la indicada en el enunciado del trabajo. A esta LAN hubo conectadas 5 computadoras, una impresora y 3 celulares al momento de la medición.

B. Techint

Esta medición fue realizada en la empresa Techint, en el horario de las 11:00 am, por un lapso de 30 minutos. Se utilizó la herramienta desarrollada en el ejercicio anterior. Se desconoce la cantidad de computadoras o la topología de la red medida. La conexión a la red fue efectuada mediante un cable de ethernet.

C. Hyundai

Esta medición fue realizada en la empresa Hyundai Motor Argentina, en horario laboral, durante 5 horas. El edificio en donde se realizó la medición, cuenta con unos 30 estaciones de trabajo fijas (PC Desktop), y 10 Notebooks, distribuídas a través de los pisos del edificio. La red tiene, además, un **switch de nivel 2** por sector, al que se encuentran conectados los dispositivos que pertenecen al departamento. A su vez, cada uno de estos está conectado a un **switch de nivel 3** a través de un **enlace Gigabit punto a punto**. A su vez, hay al menos un **access point** en cada uno de los sectores de la empresa, al cual se conectan la mayoría de los dispositivos móviles de los empleados. También hay diversos dispositivos, impresoras de red, lectores de códigos de barras inalámbricos, cámaras IPs, teléfonos IPS, entre otros, conectados a los **switches** o **access point** de cada sector.

A su vez, existen varios servidores conectados mediante un enlace **PPP** al switch principal, que proveen de diversos servicios a las estaciones de trabajo, por ejemplo: active directory, samba, correo electrónico, backup, acceso a bases

de datos, acceso a sistemas, telefonía IP, acceso a internet, etc.

La conexión fue realizada mediante una conexión cableada **PPP** entre el switch principal, y la computadora corriendo el sniffer, lo que nos permite suponer que pese a activar el modo promiscuo, sólo llegarán hasta la placa de red aquellos paquetes que se encuentren dentro de su dominio de broadcast.

D. Laboratorios DC

Esta medición fue realizada en los laboratorios del departamento de computación, en el horario de las 17 hs, por un lapso de 30 minutos. Se utilizó la herramienta explicada previamente. La red cuenta con 175 computadoras, y se desconoce la cantidad de celulares conectados a ella.

II. Desarrollo - Fuente S

A. Casa

Los resultados obtenidos fueron los siguientes

Protocolo	Informacion	Frecuencia
802.1X	15.38	0.00%
IPv6	9.59	0.13%
ARP	6.65	0.99%
IPv4	0.02	98.87%

TABLE I
S: Casa - Mediciones

El protocolo que más fue escuchado fue IPv4, tal como se observa en la figura 1, la cantidad de paquetes de este tipo fue del 98% mientras que la de paquetes ARP fue del 1%. Esto permite concluir que en el caso de la red hogareña, el overhead del tipo ARP en un tiempo de 3 horas es practicamente nulo, con respecto al total de paquetes de la red.

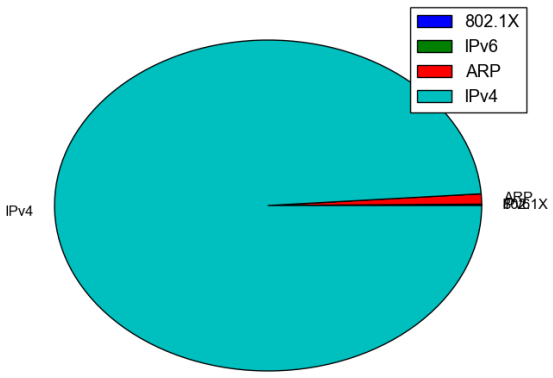

Figura 1. S: Casa - Torta

El protocolo que más fue escuchado fue IPv4, tal como se observa en la figura 1, la cantidad de paquetes de este

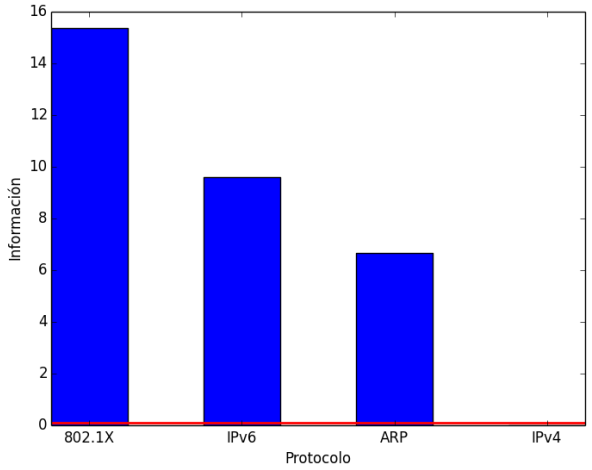

Figura 2. S: Casa - Histograma

tipo fue del 98% mientras que la de paquetes ARP fue del 1%. Esto permite concluir que en el caso de la red hogareña, el overhead del tipo ARP en un tiempo de 3 horas es practicamente nulo, con respecto al total de paquetes de la red.

La entropía de la fuente propuesta es de 0.095, lo que indica que los símbolos emitidos por la fuente son muy previsibles. Esto se puede observar en la figura 2, donde deja por debajo de ella al protocolo IPv4 que, al presentar una mayor probabilidad de aparición en la fuente, provoca que la información que aporte sea poca y lo destaque como nodo distinguido. A diferencia de IPv4, podemos encontrar al protocolo 802.1X que al tener poca probabilidad de aparición $(2{,}35 * 10^{-}5)$, aporta mucha información, siendo el protocolo que más aporta.

B. Techint

Los resultados se pueden ver en la **tablaII**.

Protocolo	Informacion	Frecuencia
ARP	3.03	12.26%
IPX	6.40	1.19%
IPv4	0.82	56.55%
802.3	2.74	14.92%
IPv6	2.74	14.93%
LLDP	9.30	0.16%

TABLE II
S: Techint - Mediciones

El protocolo que más aparece en este escenario es IPv4, tal como se observa en la figura 3, haciendo que su probabilidad sea la mayor. Esto se ve reflejado en la figura 4, en la cual se puede ver como afecta la frecuencia de aparición del protocolo a la información que este provee.

En este caso el nodo distinguido es el protocolo IPv4, pues la información que provee es menor a la entropía de

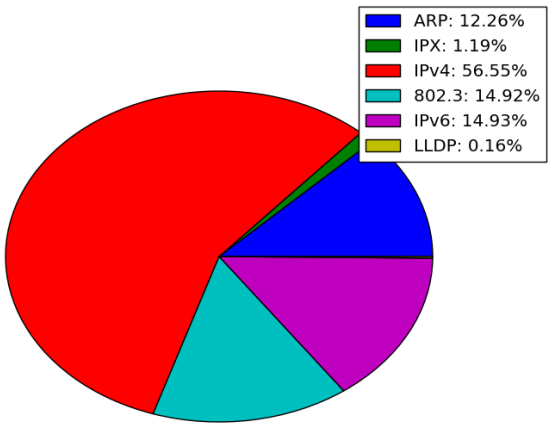

Figura 3. S: Techint - Torta

la fuente utilizada(1.74) , y además es el único que está por debajo de ella. El protocolo que más información aporta es `LLDP`, ya que su frecuencia de aparición(0.15 %) es la menor en la medición tomada. Esto genera que las pocas veces que aparece aporte más información en comparación con los protocolos que más aparecen, como por ejemplo `IPv4` cuyo porcentaje de aparición sobre el total es del 56 % tal como se observa en la figura 3.

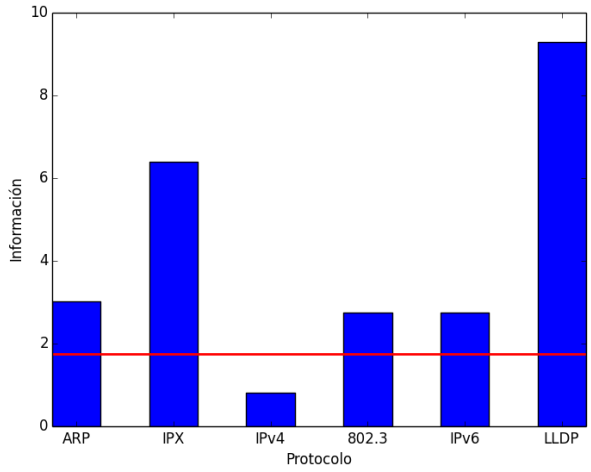

Figura 4. S: Techint - Histograma

El protocolo que mas aparece en este escenario es `IPv4`, tal como se observa en la figura 3, haciendo que su probabilidad sea la mayor. Esto se ve reflejado en la figura 4, en la cual se puede ver como afecta la frecuencia de aparición del protocolo a la información que este provee.

En este caso el nodo distinguido es el protocolo `IPv4`, pues la información que provee es menor a la entropía de la fuente utilizada(1.74) , y además es el único que está por debajo de ella. El protocolo que mas información aporta es `LLDP`, ya que su frecuencia de aparición(0.15 %) es la menor en la medición tomada. Esto genera que las pocas veces que aparece aporte mas información en comparación con los protocolos que más aparecen, como por ejemplo `IPv4` cuyo porcentaje de aparición sobre el total es del 56 % tal como se observa en la figura 3.

También se puede observar en la figura 3, la gran cantidad de paquetes de tipo `ARP` (12 %) que aparecen, en comparación con los de tipo `IPv6` (15 %) y `802.3` (14 %), que se encuentran en segundo lugar y tercer lugar respectivamente. Esto demuestra el overhead del protocolo `ARP` sobre el total de paquetes escuchados.

C. Hyundai

Protocolo	Frecuencia	Informacion
ARP	0.75 %	7.06
IPX	0.01 %	13.41
IPv4	98.82 %	0.02
IEEE_26734	0.00 %	16.21
802.3	0.15 %	9.38
IPv6	0.27 %	8.55
LLDP	0.01 %	13.71

TABLE III
S: Hyundai - Mediciones

Se pueden apreciar los resultados de las mediciones en la tabla III. Lo primero que se observa es una muy fuerte predominancia de paquetes de protocolo IPv4.

A priori, según los resultados de esta medición, parecería que `ARP` no impone un overhead considerable. Tomando este hecho, decidimos comparar esta medición con el resto, y buscar en qué cosas se diferencian, con el fin de encontrar a qué se debe esta situación de alta eficiencia. Encontramos dos grandes diferencias con el resto de las mediciones:

- El largo tiempo de medición / la alta cantidad de paquetes capturados.
- La medición, realizada mediante placa ethernet en modo promiscuo, no tiene el alcance de una medición wireless en modo monitor.
- La red switcheada, que encapsula los dominios de broadcast y colisión, reduciendo sensiblemente el tráfico de la red.

De estos tres puntos, el tiempo de medición como amortiguador del overhead de ARP resulta sumamente interesante, en parte debido a que las limitaciones de la medición en modo promiscuo eran sabidas de antemano, y son inevitables, y el beneficio de una red de switches en cascada fue explicado en clase. Además, es una conjetura fácil de poner a prueba. Basta particionar los datos de una misma captura, en intervalos de tiempo cada vez mayores, para comprobar si efectivamente al aumentar el tiempo de medición, se produce el efecto esperado.

Tomando la medición original, se la cortó en intervalos de 1 minuto, 5 minutos, 10 minutos, 20 minutos y 30 minutos. En las figuras 7, 8, 9, 10 y 11 se puede ver claramente una

progresión, en la que a medida que el tiempo de medición aumenta, el overhead de ARP disminuye.

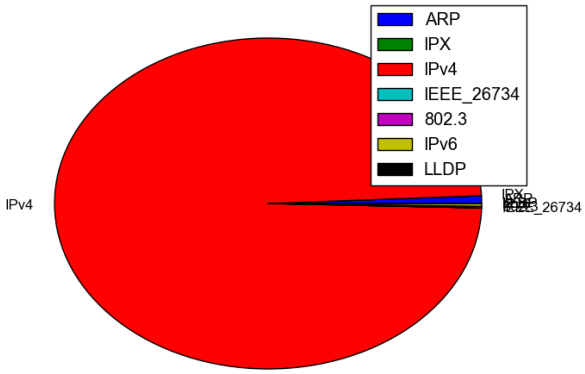

Figura 5. S: Hyundai - Torta

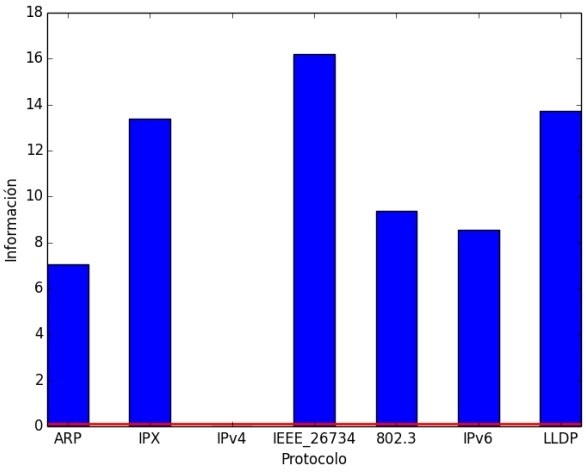

Figura 6. S: Hyundai - Histograma

Se realizó además un gráfico de histograma (*fig. 6*) y un gráfico de torta (*fig. 5*), que permiten apreciar la relación entre los distintos protocolos de una forma visual e inmediata.

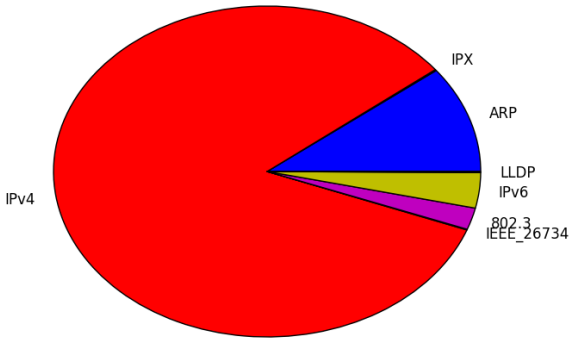

Figura 7. S: Hyundai - Torta (1 minuto)

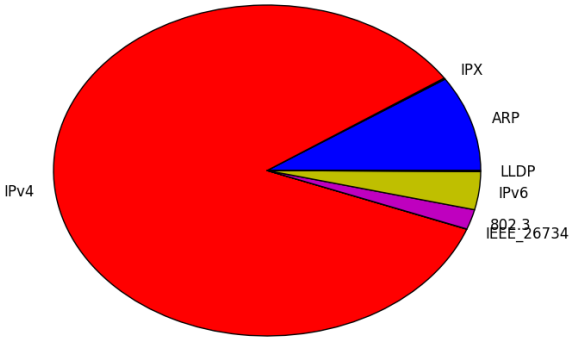

Figura 8. S: Hyundai - Torta (5 minutos)

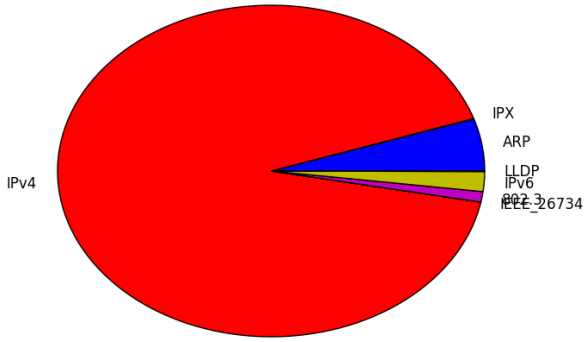

Figura 9. S: Hyundai - Torta (10 minutos)

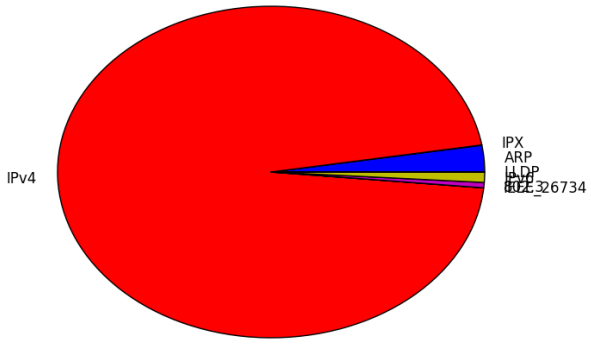

Figura 10. S: Hyundai - Torta (20 minutos)

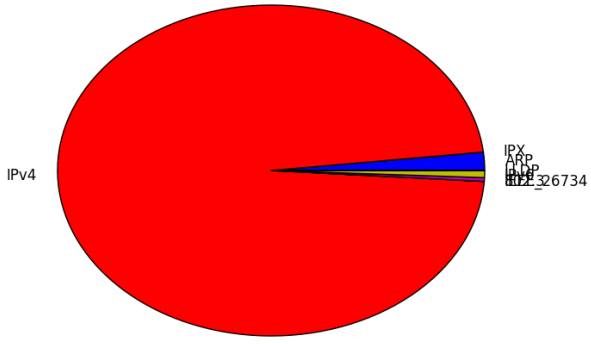

Figura 11. S: Hyundai - Torta (30 minutos)

D. Laboratorios DC

Los resultados fueron los siguientes.

Protocolo	Informacion	Frecuencia
IPv6	4.25	5.26 %
802.3	6.41	1.17 %
IPv4	0.28	82.40 %
ARP	3.16	11.17 %

TABLE IV
S: Laboratorios DC - Mediciones

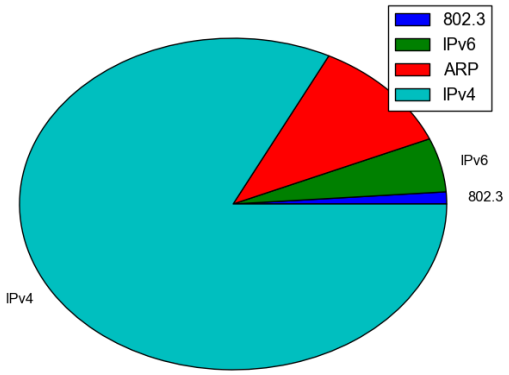

Figura 12. S: Laboratorios DC - Torta

El protocolo que más se escuchó fue IPv4, tal como se observa en la figura 12, la cantidad de paquetes de este tipo fue del 82 % mientras que la de paquetes ARP fue del 11 %. Se observa que el overhead del tipo ARP en un tiempo de 30 minutos es bajo pero no despreciable con respecto al total de paquetes de la red.

La entropía de la fuente propuesta es de 0.882, como se observa en la figura 13, dejando por debajo de ella al protocolo IPv4 que, al presentar una mayor probabilidad de aparición en la fuente, provoca que la información que aporta sea poca y se destaque como nodo distinguido. Podemos encontrar al protocolo 802.3 que al tener poca probabilidad de aparición (1 %), aporta mucha información, siendo el protocolo que más información aporta.

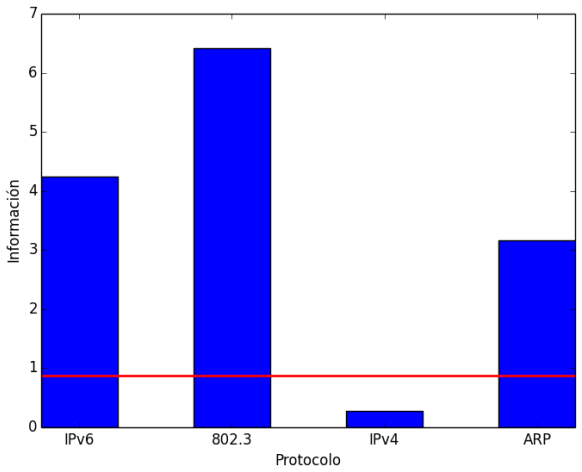

Figura 13. S: Laboratorios DC - Histograma

E. Conclusión

Notamos una fuerte tendencia del protocolo IPv4 frente al resto. Pese a que IPv6 ya se encuentra vigente hace años, es sabido el hecho de que su implementación aún no se ha extendido como se preveía. En algunas redes pudimos notar, sin embargo, una representación no negligible.

Con respecto al overhead impuesto por el protocolo ARP, notamos que no es lo suficientemente significativo como para ser motivo de preocupación, sobretodo cuando la red tiene una alta participación de otros protocolos, en donde se puede notar que mientras ARP produce una carga que se podría considerar constante o periódica, debido a la tabla caché de ARP, las cargas impuestas por otros protocoles responden a la intensidad de uso de la red. ARP en cambio, no presenta esta particularidad.

III. Desarrollo - Fuente S_1

A. Casa

Las condiciones de esta medición son las mismas que las mencionadas en la introducción, solamente se cambió la fuente utilizada y se filtraron los paquetes segun el tipo ARP.

También, con los datos escuchados, se confeccionó un grafo en el cual cada nodo es una dirección ip, y cada arista representa a un paquete enviado de una ip a otra, con un grosor que representa la cantidad de paquetes enviados entre estos. Pensamos que la confección de estos grafos permitirá tener una idea básica de una parte de la topología de la red local que escuchamos. Los resultados fueron los siguientes.

La ip que más aparece en los paquetes de tipo ARP, dentro del campo *destino*, es la 190.168.1.100, tal como se observa en la figura 14. Esta ip al ser la que más aparece dentro de las comunicaciones, será un nodo distinguido, por lo que será la que menos información aporte. Esto se ve reflejado en la figura 15, en donde la información que

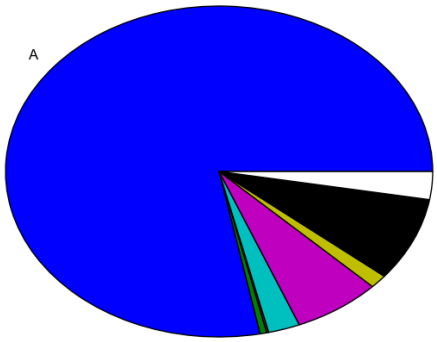

Figura 14. S_1: Casa - Torta

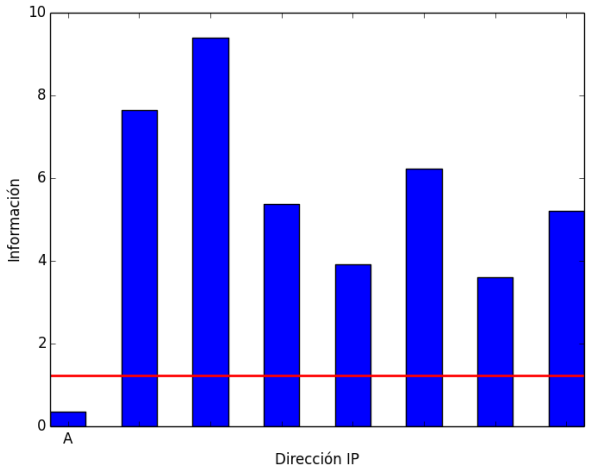

Figura 15. S_1: Casa - Histograma

brinda la ip 190.168.1.100 (representada como A) queda por debajo de la entropía de la fuente, la cual es 1.24.

La ip que más aparece en los paquetes de tipo `ARP`, dentro del campo *destino*, es la 190.168.1.100, tal como se observa en la figura 14. Esta ip al ser la que más aparece dentro de las comunicaciones, será un nodo distinguido, por lo que será la que menos información aporte. Esto se ve reflejado en la figura 15, en donde la información que brinda la ip 190.168.1.100 (representada como A) queda por debajo de la entropía de la fuente, la cual es 1.24.

Las direcciones ip que no son distinguidas tienen frecuencias y probabilidades de aparición muy similares, aportando así cantidades similares de información. Esto se puede observar en la figura 15.

En la figura 16 se observan 7 componentes conexas (hay 7 subconjuntos de ip's que intercambian información sólo entre sí y no con las demas ip's de la red). También se puede observar que hay dos pares de dos nodos (direcciones ip) que están unidos por una arista de un grosor enorme.

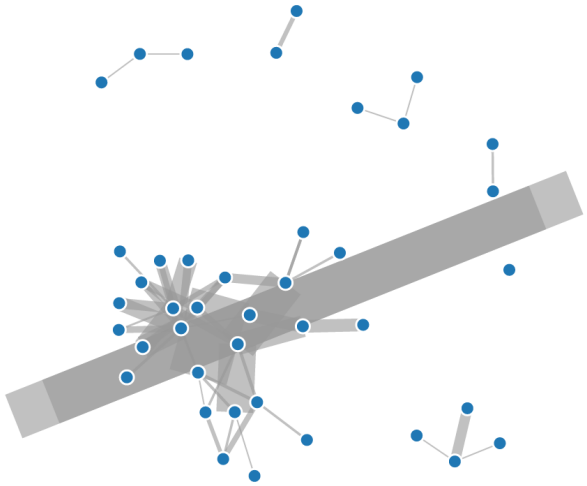

Figura 16. S_1: Casa - Grafo

Se deduce que estos pares de nodos intercambian entre sí una cantidad muy grande de paquetes. El grafo muestra nodos de la red que recibieron o emitieron paquetes, mientras que en el histograma sólo mostramos a los nodos que `recibieron` paquetes ya que los símbolos que consideramos son el campo `destino` de los paquetes. Es por esto que en el grafo aparecen representados más nodos de la red que en el histograma.

B. Techint

Las condiciones en las que se midió fueron las mismas a la detallada en la introducción, solamente se cambió la fuente utilizada y se filtraron los paquetes segun el tipo `ARP`.

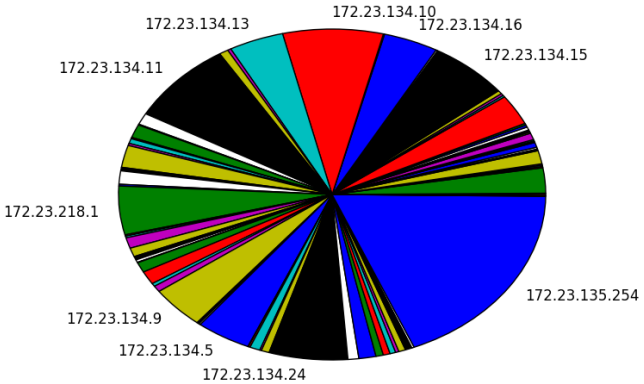

Figura 17. S_1: Techint - Torta

En este caso se encontró un subconjunto de nodos distinguidos, los cuales presentan una frecuencia significativamente mayor en comparación con el resto de los nodos. Esto se puede observar en la figura 17.

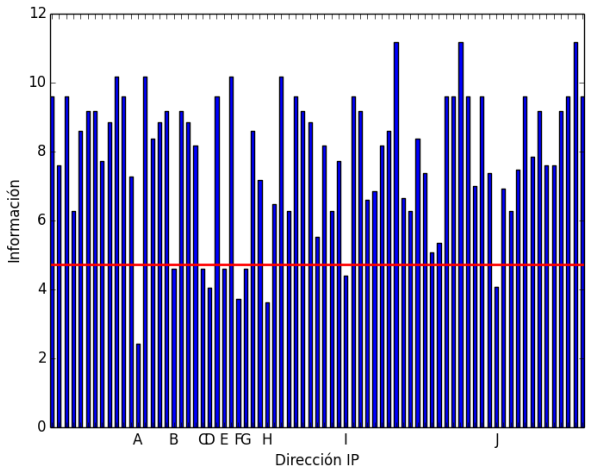

Figura 18. S_1: Techint - Histograma

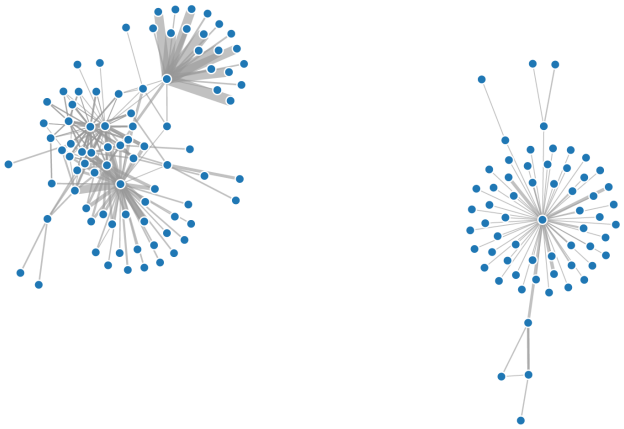

Figura 19. S_1: Techint - Grafo

Todos estos nodos aportaron un información menor a la entropía de la fuente, la cual fue de 4.73. Estos estan enumerados con letras de la A a la J, formando así un subconjunto con 10 nodos distinguidos, tal como se ve en la figura 18. Dentro de estos nodos, se observa uno en particular nombrado como A (cuya ip es 172.23.135.254), que aparece con una frecuencia mayor al resto (18 %), lo que como consecuencia aportará menor cantidad de información.

También realizamos un grafo a partir de los datos de forma similar a la hecha en la red *Casa*. En este caso se pueden observar dos componentes conexas. En ambas componentes se observa que hay uno o más `nodos especiales`, que están conectados con una cantidad grande de nodos que sólo se conectan con este. Estos nodos son nodos correspondientes a los símbolos destacados (que se encuentran por debajo de la entropía) en el histograma de la figura 18. Luego, consideramos que es esperable que estos símbolos (o al menos alguno/s) de estos se correspondan con la dirección ip de routers presentes en la red.

C. Hyundai

Se observan dos nodos distinguidos, nombrados como A y B, cuyas frecuencias de aparición son las mayores. A su vez se puede distinguir una jerarquía en la que además de estos nodos primarios existe una serie de nodos de segundo nivel en cuanto a su frecuencia, tal como se observa en la figura 21.

Esta jerarquía también es observable en la figura 22, en donde hay un nodo central bien marcado, y otros nodos importantesen su periferia más cercana.

Segun sabemos del entorno en donde fue realizada la medición, esto es posiblemente una consecuencia de la topología de la red en donde se realizó la medición, en particular pudimos corroborar que muchas de las ips correspondientes a los nodos pertenecen a servidores, que brindan servicios activos a todo el resto de las máquinas de la empresa. De esta forma es comprensible que los mismos tengan una gran cantidad de solicitudes `ARP`.

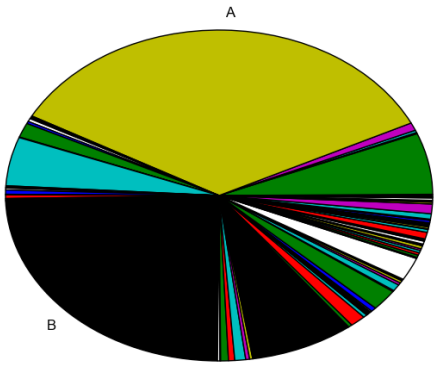

Figura 20. S_1: Hyundai - Torta

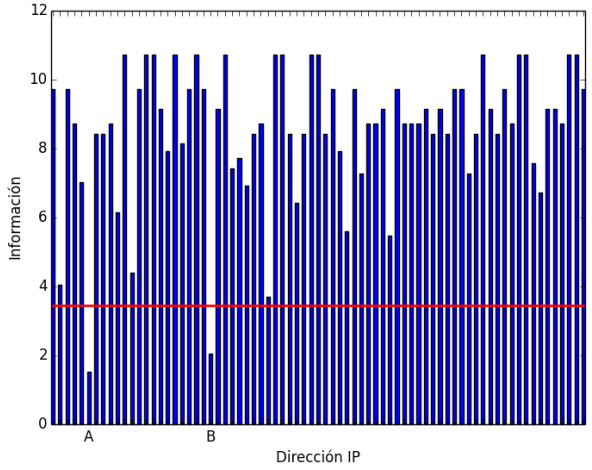

Figura 21. S_1: Hyundai - Histograma

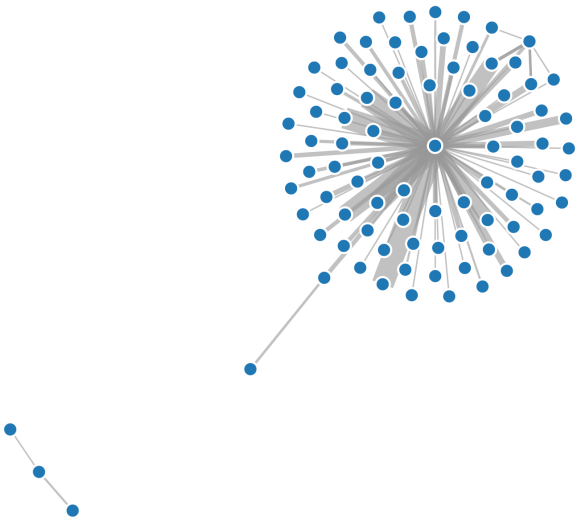

Figura 22. S_1: Grafo Hyundai

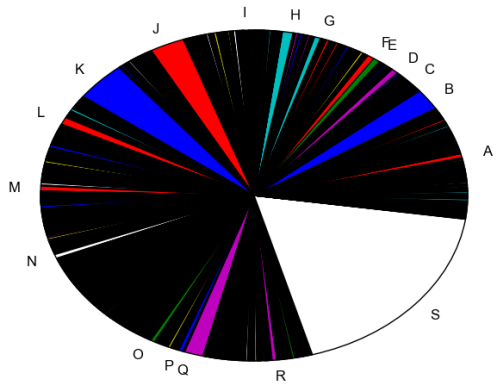

Figura 23. S_1: Laboratorios DC - Torta

Se observan dos nodos distinguidos, nombrados como A y B, cuyas frecuencias de aparición son las mayores. A su vez se puede distinguir una jerarquía en la que además de estos nodos primarios existe una serie de nodos de segundo nivel en cuanto a su frecuencia, tal como se observa en la figura 21.

Esta jerarquía también es observable en la figura 22, en donde hay un nodo central bien marcado, y otros nodos importantes en su periferia más cercana.

Segun sabemos del entorno en donde fue realizada la medición, esto es posiblemente una consecuencia de la topología de la red en donde se realizó la medición, en particular pudimos corroborar que muchas de las ips correspondientes a los nodos pertenecen a servidores, que brindan servicios activos a todo el resto de las máquinas de la empresa. De esta forma es comprensible que los mismos tengan una gran cantidad de solicitudes `ARP`.

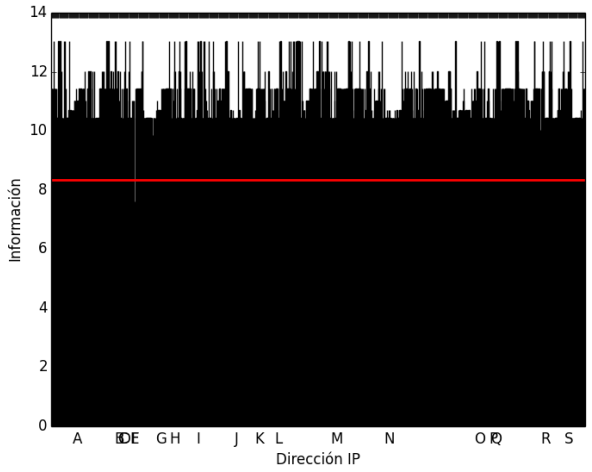

Figura 24. S_1: Laboratorios DC - Histograma

D. Laboratorios DC

Las condiciones en las que se midió fueron las mismas a la detallada en la introducción, solamente se cambió la fuente utilizada y se filtraron los paquetes segun el tipo `ARP`.

En esta medición se encontraron 20 nodos distinguidos. Dentro de este subconjunto se observó un nodo en particular, el S, como se puede observar en la figura 23. Sin embargo, en la figura 24 no se puede apreciar que aporta poca información y queda debajo de la entropía (8.33), puesto que la cantidad de símbolos graficados es muy grande, haciendo que el ancho de la columna que lo representa en el histograma no sea visible. Sí son visibles el conjunto de nodos B,C,D,E y F, que están todos por debajo de la entropía y casualmente cada uno aporta casi la misma cantidad de información.

También se observó en el histograma de la figura 24 que a simple vista hay aproximadamente 4 niveles de cantidad de información provista por los símbolos, y que hay poca variación en la cantidad de información provista por prácticamente todos ellos (excepto por el símbolo S, que aporta 2.44 de información).

Finalmente, realizamos un grafo de igual forma que en las mediciones anteriores. En este caso se observa una única componente conexa, con muchos nodos frontera conectados mediante una arista con pocos nodos centrales. Los gráficos sigieren que estos se correspondan con los símbolos que están por debajo de la entropía en el histograma de la figura 24. Al igual que en las mediciones ya analizadas, será esperable que estos símbolos sean la dirección ip de dispositivos (nodos) destacados en la red local estudiada (por ejemplo, podrían ser routers).

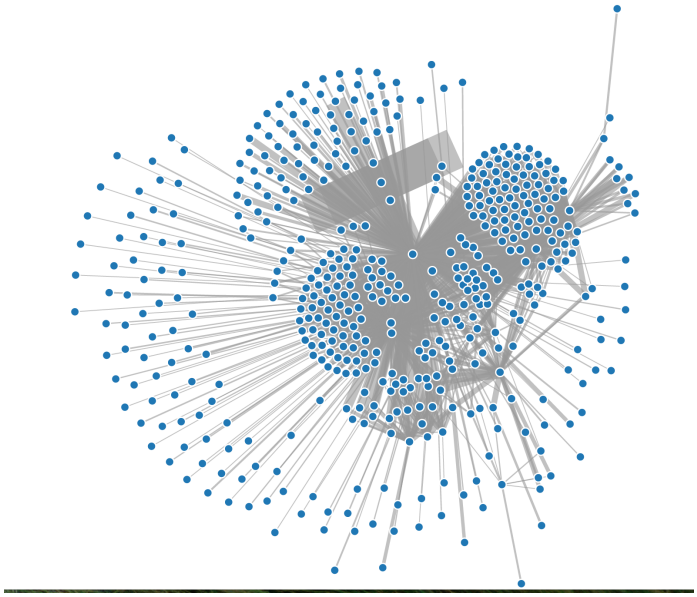

Figura 25. S_1: Laboratorios DC - Grafo

E. Conclusión de las mediciones

En todas las mediciones se encontraron indicios que parecen responder a una topología lógica con forma de estrella subyacente. Esto se condice con lo visto en clase ya que responde a las metodologías de ruteo explicadas. Por ejemplo, routers en cascada o subredes bien definidas.

A diferencia de la fuente anterior, se encontraron en todos los casos más de un símbolo distinguido. Aparentemente, la distribución de probabilidades de esta fuente tiene una estructura jerárquica.

IV. Conclusión

Para finalizar, podemos decir que nos divirtió mucho la realización de este TP, ya que fuimos capaces de internalizarnos en el funcionamiento de las redes de área local (LAN) que usamos desde temprana edad prácticamente todos los días de nuestras vidas.

Pudimos entender que para la comunicación entre un servidor y un host, se tienen que efectuar una serie de intercambios de datos entre distintas capas de datos de la red. También, utilizamos lo aprendido en clase de teoría de la información para analizar características de las redes analizadas.

Como ya fue dicho, sobre un mismo conjunto de mediciones de red, definimos dos fuentes distintas: S y S_1. Pudimos observar que las entropías presentadas en los experimentos sobre la fuente S fueron menores que las presentadas sobre la fuente S_1. A partir de esto, y aplicando lo que nos enseñó Shannon de teoría de la información, podemos concluir que la fuente S es *más compresible* que la fuente S_1. Es decir, que en teoría se podría lograr una codificación óptima de longitud media menor de la fuente S.

Además, en todos los experimentos (sobre ambas fuentes), identificamos nodos distinguidos (por debajo de la entropía de la fuente).

El trabajo práctico tuvo un componente no presente en trabajos previos que fue la toma de mediciones en el entorno. Esto nos permitió realizar inferencias sobre el mundo real.