Trabajo Práctico №1: Wiretapping

Alvaro Jose Fernando, Barbeito Nicolás, Brum Raúl, Nievas Yésica

***Resumen*—En este trabajo práctico utilizaremos técnicas provistas por la teoría de la información para analizar algunas redes de información. Estudiaremos diversos aspectos de la red analizando la entropía y probabilidad de distintas fuentes de información en las misma. Para ello utilizaremos como herramientas de manipulación y análisis de paquetes a Wireshark y Scapy.**

I. Introducción

A. Paquetes ARP

El protocolo ARP (Address Resolution Protocol) permite mapear direcciones de nivel de red a direcciones físicas. La idea de este protocolo se basa en el envío de paquetes que pueden ser de preguntas o respuestas. El emisor del paquete que pregunta por una dirección, envía un mensaje broadcast sobre la red local, siendo respondido por un mensaje unicast por aquel al que pertenece la dirección consultada. Mediante el envío de estos paquetes ARP se construyen las tablas que mapean direcciones de red con direcciones físicas.

B. Entropía de una fuente

Para poder definir la entropía de una fuente, necesitamos la definición de información que aportan los símbolos emitidos por dicha fuente. Se define información de un símbolo s como

$$\mathrm{I}\ (s) = \log(1/\mathrm{P}(s))$$

siendo $\mathrm{P}(s)$ la probabilidad de ocurrencia de dicho símbolo. De esta manera, puede calcularse la información media suministrada por una fuente de información de memoria nula (los símbolos emitidos son estadísticamente independientes) como

$$\sum_S P(s_i)I(s_i)\forall s_i \in S$$

Esta cantidad media de información por símbolo de la fuente, recibe el nombre de *entropía* H(S) de la fuente de memoria nula.

$$H(S) = \sum_S P(s_i)log(1/P(s_i))$$

Debido a que la entropía de una fuente depende de la probabilidad de los diferentes símbolos que la componen, se puede demostrar que para una fuente de información de memoria nula con un alfabeto de q símbolos, el valor máximo de la entropía es precisamente log q, alcanzándose solamente si todos los símbolos son equiprobables.

Alvaro Jose Fernando, LU: 89/10, email: fer1578@gmail.com
Barbeito Nicolás, LU: 147/10, email: nicolasbarbeiton@gmail.com
Brum Raúl, LU: 199/98, email: brumraul@gmail.com
Nievas Yésica, LU: 340/05, email: yesica.nievas@gmail.com

C. Fuente S

Sea P la fuente de información generada a partir de todos los paquetes Ethernet que se transmiten en una determinada red entre los instantes de tiempo $[t_i, t_f]$:

$P_{ti,tf} = \{p_1 ... p_n\}$

siendo p_i el i-ésimo paquete transmitido en la red entre los instantes de tiempo $[t_i, t_f]$.

Los paquetes p_i pertenecientes a P encapsulan diferentes protocolos, que se pueden identificar a través del campo type del frame de capa 2 (p.type en Scapy). Por lo tanto, con el objetivo de distinguir los protocolos utilizados en una red, se define otra fuente de información S de la siguiente manera:

$S_{ti,tf} = \{s_1 ... s_n\}$

siendo $\mathrm{s}_i = \mathrm{p}_i.\mathrm{type}$ /p_i perteneciente a P entre los instantes de tiempo $[t_i, t_f]$.

D. Propuesta de una nueva fuente S_1

Para analizar la entropía de la red en base a los paquetes ARP observados realizamos una nueva tool en base a la anterior que nos permitiera obtener datos de los campos de dichos paquetes. Se propone como nueva fuente S1 el conjunto de símbolos conformado por las distintas direcciones IP destino:

$\mathrm{S}_1 = \{\mathrm{s}_{11} ... \mathrm{s}_{1n}\}$ siendo s_{1i} el valor del campo *pdst* correspondiente a la ip destino del paquete

Al igual que para la fuente S, realizamos el cálculo de la entropía como fue requerido, como la probabilidad e información de sus símbolos.

II. Métodos y condiciones de cada experimento

En esta sección describiremos brevemente las redes elegidas para realizar las escuchas mediante las herramientas indicadas en el enunciado del trabajo práctico.

A. Hipótesis

Basándonos en la Teoría de la Información, propondremos distintas hipótesis a probar o descartar en función de los experimentos que se expondrán a continuación.
Nuestra primer hipótesis será sobre los valores de entropía resultantes de la fuente S y S_1, que se presentaron anteriormente, respecto a las capturas de redes controladas y redes abiertas. Dado que para redes controladas los paquetes se envían entre dispositivos en su mayoría conocidos entre sí, y los cambios que puede sufrir la red son mucho menores que para una red abierta, se espera que los valores de entropía de las fuentes sea menor que para las capturas de redes abiertas. Está hipótesis se basa en el hecho de que el

conocimiento planteado sobre la red y sus paquetes, debería disminuir la incertidumbre sobre los valores resultantes de la fuente. Si disminuye la incertidumbre, disminuye la entropía.
Plantearemos como segunda hipótesis que los valores de entropía de la fuente S serán menores a los valores de entropía de la fuente S1. Nos basamos para esta hipótesis en el hecho de que la fuente S tendrá menos símbolos respecto a la fuente S1 para los paquetes enviados por las redes. Dado que la fuente S1 presentará una cantidad mayor considerable de símbolos, podríamos decir que la proporción de aparición de estos estará más distribuida. Si considerando la definición de información, es válido plantear que el valor medio ponderado de la cantidad de información por símbolo será mayor en la fuente S1 que en la fuente S. Es decir, la entropía de la fuente S1 será mayor a la entropía de la fuente S. Otra definición que nos puede ayudar con el planteo de este hipótesis, es que la entropía de la fuente es menor o igual a la longitud de la fuente, entonces el valor de la cota para la entropía de S es menor respecto de S1.

B. Home LAN

Esta medición fue realizada en la LAN de una casa por un intervalo de 2 horas. La misma cuenta con una computadora corriendo un sistema operativo Linux la cual es el router de la LAN y provee a las demás computadoras de acceso a internet además de otros servicios de red (proxy, DNS, DHCPCD, etc). A la misma se encuentran conectadas mediante un switch 4 computadoras cableadas y 2 access point inalámbricos. A estos últimos se encontraban conectados al momento de la medición una notebook y varios celulares. Además una de las computadoras cableadas corre una máquina virtual con IP propia independiente y otra de las computadoras cableadas posee otro ISP para conectarse a internet por lo que no utiliza a la primera computadora como gateway.

C. Labos DC

Esta medición fue realizada en los laboratorios del Departamento de Computación por un intervalo de 30 minutos. Se desconoce la cantidad de computadoras y celulares conectados al momento de realizar la captura.

D. Zona pública (parque)

Se realizó una medición en un parque con wifi abierto durante 15 minutos. Debido a que la medición fue realizada en un espacio público se espera que varios dispositivos móbiles, como celulares y notebooks, se conecten a la misma sin poder precisar cuántos a priori.

E. Universidad (Universidad de la Matanza)

Esta medición se realizó por aproximadamente 10 minutos en la red wifi de la Universidad de La Matanza. Es una red abierta y desconocemos la estructura de ella.

III. Resultados y Análisis

En esta sección mostraremos y analizaremos los resultados obtenidos en las mediciones realizadas en las distintas redes tanto para la fuente S y S_1 descriptas en la sección I C y D. Para cada una de las redes veremos sus protocolos distinguidos, la proporción de paquetes ARP en el tráfico de la red y los nodos (representados por IP) distinguidos. Cabe aclarar que para la fuente S_1 podemos ver todos los pedidos "who has" de ARP ya que todas las máquinas están en el mismo dominio de broadcast pero sólo podremos ver las respuestas "is at" de aquellas máquinas que se encuentren en igual dominio de colisión (generalmente las conectadas por wireless ya que las cableadas, al estar conectadas por switches, no tienen colisión y sólo el destinatario recibirá la respuesta).

A. Home LAN

A.1 Fuente S

Los resultados obtenidos para la fuente S fueron:

Protocolo	Información	Probabilidad
EAPOL	12.31	0.01%
ARP	5.11	2.88%
IPv6	4.22	5.33%
IPv4	0.12	91.76%

TABLE I
Home LAN - Protocolos

Como puede observarse en la figura 1 el protocolo más utilizado es IPv4 en un 91.76% mientras que IPv6 y ARP sólo son utilizados en un 5.33% y 2.88% respectivamente. EAPOL (autenticación wireless) no tiene prácticamente incidencia. Observamos además que el protocolo ARP tiene sólo una incidencia del 2.88% en el tráfico total de la fuente, lo que hace que el overhead aportado por el mismo no sea significativo.

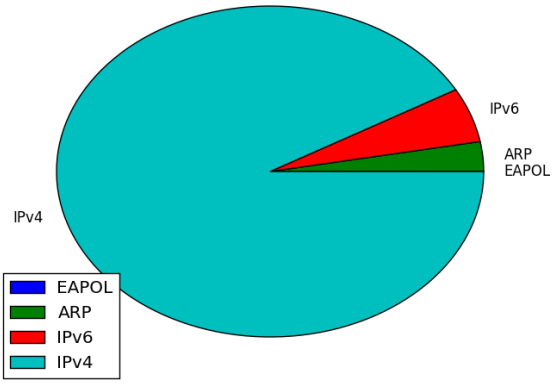

Figura 1. Home LAN - Probabilidades

Por otro lado, la entropía de la fuente S fue de 0.4892 reflejando que los símbolos emitidos por la fuente S son

muy previsibles. Esto podemos notarlo en la figura 2 donde observamos que el protocolo con mayor porcentaje de aparición, IPv4, es el que menos información aporta a la fuente. La información aportada por este se encuentra por debajo de la entropía de S. Como contraste observamos en la figura 2 que el protocolo EAPOL es el que más información aporta ya que según lo observado en la figura 1 tiene una probabilidad muy baja provocando que no incida en la entropía.

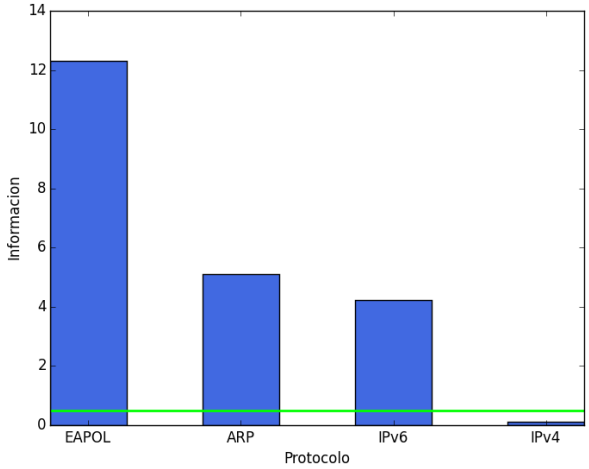

Figura 2. Home LAN - Información

A.2 Fuente S_1

Para la fuente S_1 expondremos las IPs de destino. Los resultados fueron:

Protocolo	Información	Probabilidad
192.168.10.10	7.68	0.48 %
192.168.10.27	6.41	1.16 %
192.168.10.11	6.30	1.26 %
192.168.10.13	5.09	2.92 %
192.168.10.7	4.33	4.97 %
192.168.10.1	4.00	6.23 %
192.168.10.4	2.54	17.15 %
192.168.10.6	2.11	23.09 %
192.168.10.9	1.22	42.69 %

TABLE II
Home LAN - Nodos

Como puede observarse en la figura 3 las IP que aparecen con mayor frecuencia como destino en los paquetes ARP son la "192.168.10.9" en un 42.69 %, la "192.168.10.6" en un 23.09 % y la "192.168.10.4" en un 17.15 %, por lo que estas serán los nodos distinguidos de la fuente y las que menos información aporten.

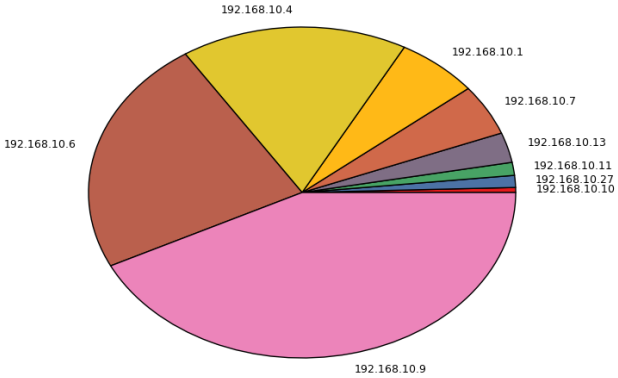

Figura 3. Home LAN - Probabilidades

La entropía de la fuente S_1 fue de 2.25. En la figura 4 se observa la información aportada por los distintos nodos y como dos de ellos quedan por debajo de la entropía de la fuente y el otro apenas por arriba.

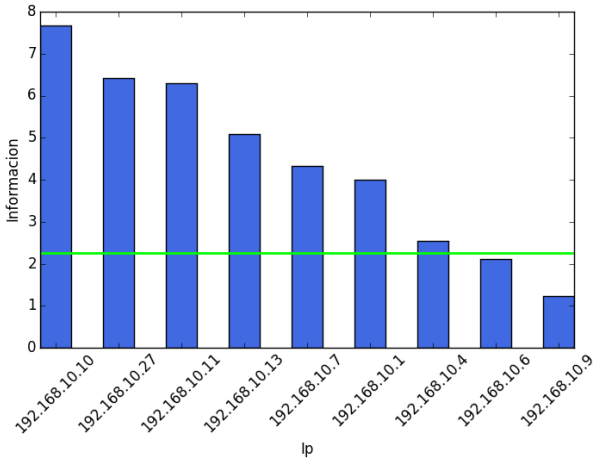

Figura 4. Home LAN - Información

En el análisis de la fuente S_1 vemos que ninguno de los tres nodos distinguidos pertenecen al router de la red, la IP "192.168.10.1". De los nodos distinguidos las IPs "192.168.10.6" y "192.168.10.4" se explican, ya que la primera es la notebook que realizó la captura en la red y la segunda es la computadora que la controlaba remotamente mediante SSH. La IP "192.168.10.9" corresponde a la IP de una consola de video juegos, la cual se encontraba apagada al momento de realizar la captura; en la IP "192.168.10.4" hay un daemon ejecutándose que se comunica con la consola de video juegos. Al estar esta apagada hay muchos pedidos "who has 192.168.10.9" y ninguna respuesta por lo que los pedidos son reiterados varias veces incrementado su incidencia.

B. Labos DC

B.1 Fuente S

Los resultados obtenidos para la fuente S fueron:

Protocolo	Información	Probabilidad
ARP	8.092	0.37 %
IPv4	0.005	99.63 %

TABLE III
LABOS DC - PROTOCOLOS

Como observamos en la figura 5 ampliamente el protocolo más utilizado es IPv4 en un 99.63 % mientras que ARP sólo es utilizado en un 0.37 %. El overhead aportado por el protocolo ARP en el tráfico de la red es despreciable. Notamos también la ausencia de protocolo para wireless aunque habían celulares conectados a la red durante la captura. Creemos que esto es debido a que la red wifi provista por los labos del Departamento de Computación es abierta y al no estar encriptadas no aparece el protocolo EAPOL utilizado durante la autenticación.

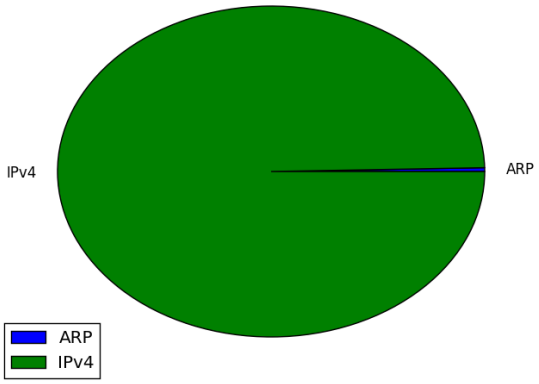

Figura 5. Labos DC - Probabilidades

La entropía de la fuente S fue de 0.034922 lo que muestra que los símbolos emitidos por la fuente S son muy previsibles como podemos notarlo en la figura 6. En la misma podemos observar que la información aportada por el protocolo IPv4 se encuentra por debajo de la entropía de S. El protocolo ARP aporta información muy por arriba de la entropía como se observa en la figura 6 correspondiéndose con lo observado en la figura 5, donde se muestra que tiene una probabilidad demasiado baja como para incidir en la entropía.

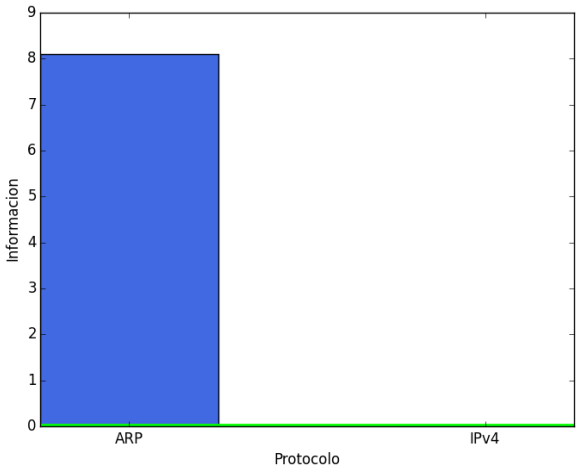

Figura 6. Labos DC - Información

B.2 Fuente S_1

Para la fuente S_1 expondremos las IPs de destino. Dada la cantidad de IPs capturadas, para facilitar la lectura de los gráficos y tablas, expondremos dos bloques (A y B) en representación de IPs destino que tuvieron igual aporte de información en la captura. El bloque A se compone de 19 IPs con igual aporte de información y el bloque B de 7 IPs. Los resultados fueron los siguientes:

Protocolo	Información	Probabilidad
A(19 IPs)	8.40	5.60 %
B(7 IPs)	7.40	4.13 %
10.2.203.144	6.82	0.88 %
10.2.202.5	6.82	0.88 %
10.2.202.228	6.82	0.88 %
10.2.203.122	6.40	1.17 %
10.2.203.117	6.40	1.17 %
10.2.203.254	4.82	3.53 %
10.2.201.214	4.23	5.30 %
10.2.202.132	4.08	5.89 %
10.2.200.217	4.01	6.19 %
10.2.202.108	2.33	19.76 %
10.2.202.100	2.33	19.76 %
10.2.202.47	2.01	24.77 %

TABLE IV
LABOS DC - NODOS

Como puede observarse en la figura 7 las IPs que aparecen con mayor frecuencia como destino en los paquetes ARP son la “10.2.202.47” en un 24.77 %, la “10.2.202.100” en un 19.76 % y la “10.2.202.108” en un 19.76 % por lo que estas serán los nodos distinguidos de la fuente y las que menos información aporten.

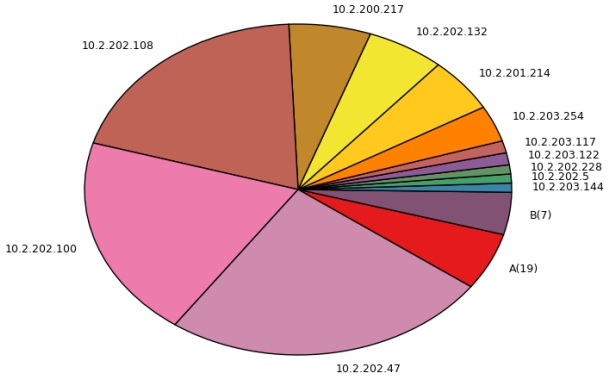

Figura 7. Labos DC - Probabilidades

La entropía de la fuente S_1 fue de 3.417440. En la figura 8 se observa la información aportada por los distintos nodos y como los tres nodos distinguidos quedan por debajo de la entropía de la fuente.

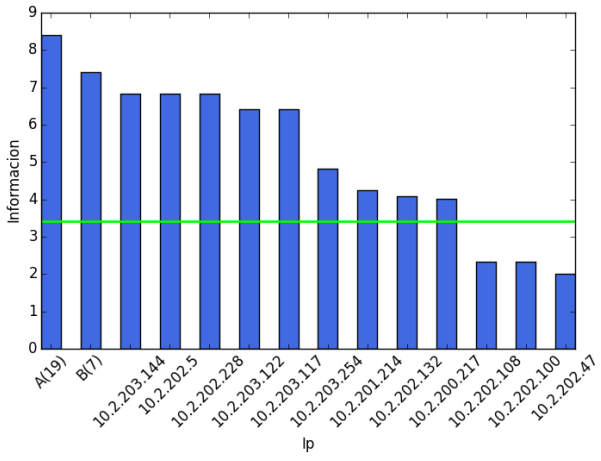

Figura 8. Labos DC - Información

En el análisis de la fuente S_1 vemos que ninguno de los nodos distinguidos pertenecen al router de la red, la IP "10.2.203.254", ni tampoco a la IP de la máquina que realizó la captura "10.2.202.132". No tenemos información de la red para precisar a que pertenecen las IPs de los nodos distinguidos.

C. Parque

C.1 Fuente S

Los resultados obtenidos para la fuente S fueron:

Protocolo	Información	Probabilidad
IPv6	1.763	29.45 %
ARP	1.646	31.95 %
IPv4	1.373	38.59 %

TABLE V
PARQUE - PROTOCOLOS

Podemos notar en la figura 9 que el protocolo más utilizado es IPv4 (38.5919 %) seguido por ARP (31.9503 %) e IPv6 (29.4578 %). En comparación con las redes analizadas en secciones anteriores, los paquetes ARP son muy frecuentes. Otra peculiaridad es que el protocolo IPv4 y el IPv6 son usados en casi la misma frecuencia, lo que no sucedió en las redes anteriores.

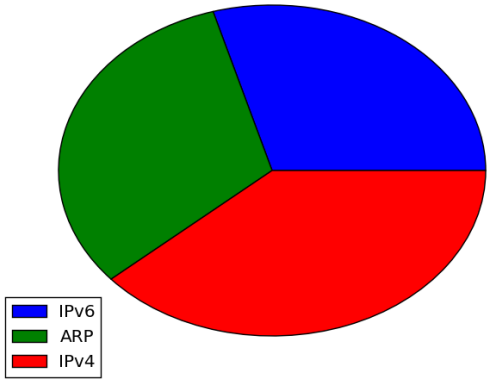

Figura 9. Parque - Probabilidades

En cuanto a la entropía de la fuente S, fue de 1,575466, lo que nos dice que los símbolos emitidos por la fuente no son previsibles, ya que ningún protocolo otorga mucha más información que los demás (ver figura 10). No podemos decir que para esta fuente haya un símbolo distinguido.

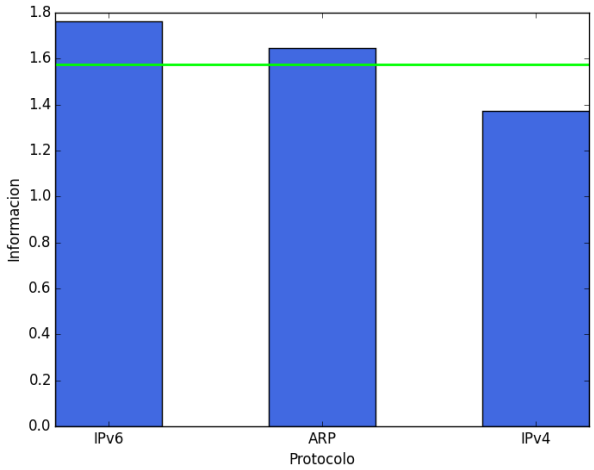

Figura 10. Parque - Informacion

C.2 Fuente S_1

Para la fuente S_1 expondremos las IPs de destino. Dada la cantidad de IPs capturadas, para facilitar la lectura de los gráficos y tablas, se agruparon las direcciones IP que tuvieron igual aporte de información en la captura.

Protocolo	Información	Probabilidad
B1(171 IPs)	14.105417	0.97%
B2(227 IPs)	13.105417	2.56%
B3(348 IPs)	12.422595	6.38%
B4(253 IPs)	11.601712	8.25%
B5(40 IPs)	10.649047	2.53%
B6(16 IPs)	9.504669	2.25%
B7(4 IPs)	8.569523	1.06%
B8(5 IPs)	7.543969	2.69%
B9(4 IPs)	6.495808	4.49%
169.254.255.255	5.380904	2.39%
10.249.147.254	4.970991	3.18%
10.248.79.254	4.908201	3.33%
10.248.0.100	4.340546	4.93%
10.248.0.68	4.277281	5.15%
10.249.111.254	4.129570	5.71%
10.248.255.254	3.972275	6.37%
169.254.188.151	3.521395	8.70%
10.249.107.249	3.278075	10.30%
10.250.127.254	2.422862	18.64%

TABLE VI
PARQUE - NODOS

La entropía de la fuente fue de 5,757719, lo que indica que la fuente tiene un alto nivel de incertidumbre. Podemos ver que los bloques de direcciones de B1 a B9 son nodos diferenciados debido a la alta cantidad de información aportada, es decir, su baja probabilidad de aparecer como símbolo en la fuente; de la misma manera, las direcciones IP `10.249.107.249` y `10.250.127.254` también se diferencian pero por la razón contraria, aportan muy poca información, por tener una frecuencia relativa de aparición alta durante las capturas en relación a los otros símbolos.

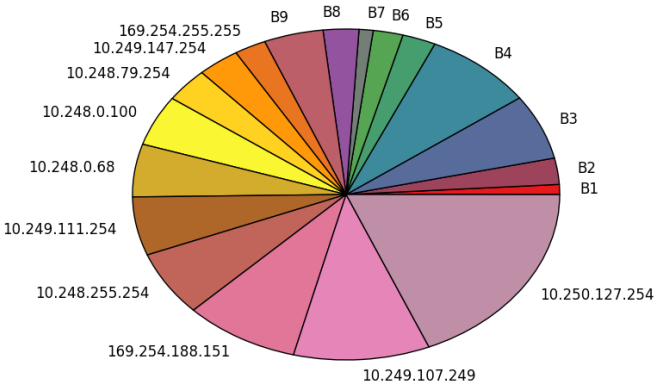

Figura 11. Parque S1 - Probabilidades

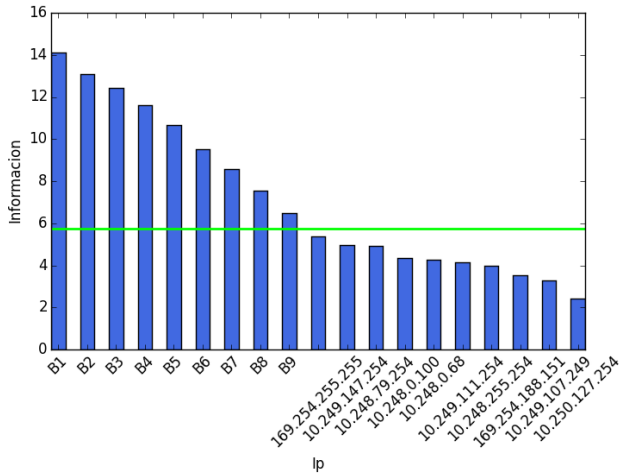

Figura 12. Parque - Información

Como particularidad en esta red observamos la presencia de nodos con IPs fuera del rango usado por la red como ser la IP 169.255.188.151. Estas IPs pertenecen a nodos que no tienen una IP asignada en la red y están usando como IP la asignada por "ipv4ll" mediante "zeroconf". Esta tecnología asigna una IP automáticamente cuando el nodo no puede obtener una IP a través de DHCPCD o no tiene configurada una IP fija. Las IPs asignadas de esta manera pertenecen al bloque de IPs 169.254.0.0/16.

D. Universidad de La Matanza

D.1 Fuente S

Los resultados obtenidos para la fuente S fueron:

Protocolo	Información	Probabilidad
IPv6	6.753	0.009 %
ARP	1.269	0.415 %
IPv4	0.796	0.576 %

TABLE VII
Universidad de La Matanza - Protocolos

Se observa en la figura 13 que los protocolos más utilizados son IPv4 en un 57.60 % y ARP con un 41.50 %, el uso del protocolo IPv6 es de apenas 0.9 %. La cantidad de mensajes ARP es bastante elevado para esta red, creemos que esto podría indicar una gran variabilidad de los componentes conectados (dispositivos que se conectan por poco tiempo). El uso despreciable del protocolo IPv6 creemos que podría justificarse por un tema de configuración propia de la red.

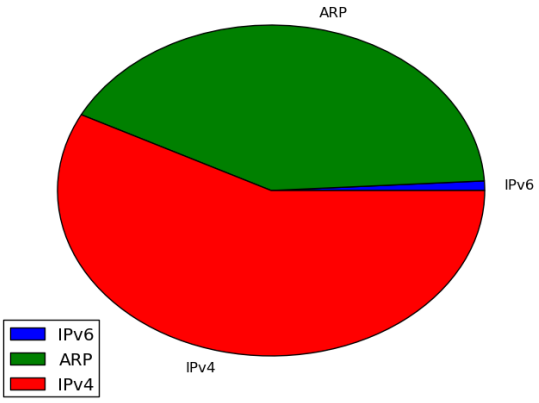

Figura 13. Universidad de La Matanza - Probabilidades

La entropía de la fuente S fue de 1.047737 reflejando como en el caso anterior que los símbolos emitidos por la fuente S sean (POCO) muy imprevisibles, como podemos notarlo en la figura 14. En la misma podemos observar que la información aportada por el protocolo IPv4 se parece a la del protocolo ARP, ambos muy cerca de la entropía de S. El protocolo IPv6 aporta información muy por arriba de la entropía según la figura 14, y así también lo observamos en la figura 13 que nos permite ver que tiene una probabilidad extremadamente baja.

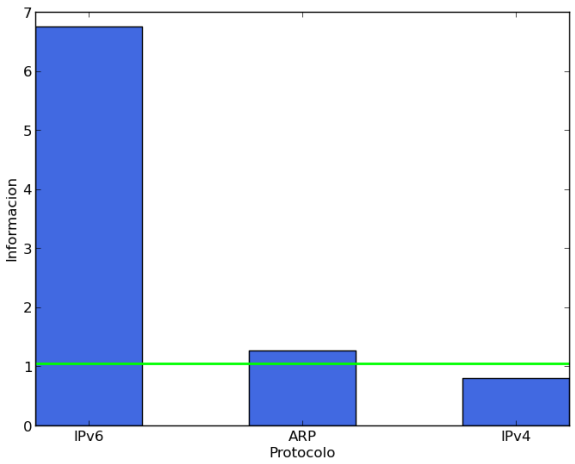

Figura 14. Universidad de La Matanza - Información

D.2 Fuente S_1

Para la fuente S_1 expondremos las IPs de destino. Dada la cantidad de IPs capturadas, para facilitar la lectura de los gráficos y tablas expondremos varios bloques de (A hasta I), en representación de IPs de destino que tuvieron un aporte de información parecido en la captura. Como había demasiadas IPs, para las que estaban muy por encima de la entropía se utilizó un promedio para darle un valor para el gráfico. Los resultados fueron :

Protocolo	Información	Probabilidad
A(131 IPs)	13.450309	1.16 %
B(74 IPs)	12.450309	1.32 %
C(155 IPs)	11.540077	5.31 %
D(77 IPs)	10.565611	5.16 %
E(103 IPs)	9.564986	13.86 %
F(38 IPs)	8.755480	8.83 %
G(38 IPs)	8.271055	12.36 %
H(28 IPs)	7.732110	13.23 %
I(20 IPs)	7.272395	12.99 %
10.5.83.154	6.990878	0.78 %
10.5.73.140	6.974576	0.79 %
10.5.78.34	6.942515	0.81 %
169.254.255.255	6.895720	0.83 %
10.5.77.136	6.865347	0.85 %
10.5.82.0	6.668949	0.98 %
10.5.84.63	6.579944	1.04 %
10.5.70.239	6.543419	1.07 %
10.5.70.216	6.119392	1.43 %
10.5.80.111	6.024044	1.53 %
10.5.200.13	5.911150	1.66 %
10.5.77.177	5.784973	1.81 %
10.5.0.1	5.531446	2.16 %
10.5.80.17	5.455956	2.27 %
10.5.103.253	3.707158	7.65 %

TABLE VIII
UNIVERSIDAD DE LA MATANZA - NODOS

Como puede observarse en la figura 15 las IPs que aparecen con mayor frecuencia como destino en los paquetes ARP son la "10.5.103.253" en un 7.65 %, la "10.5.80.17" en un 2.27 %, "10.5.0.1" en un 2.16 %, la "10.5.77.177" en un 1.81 %, la "10.5.200.13" en un 1.66 %, la "10.5.70.216" en un 1.43 % y varias más, por lo que estas serán los nodos distinguidos de la fuente y las que menos información aporten.

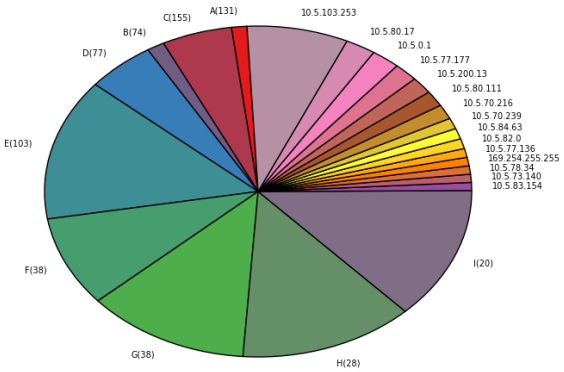

Figura 15. Universidad de La Matanza - Probabilidades

La entropía de la fuente S_1 fue de 7.951183. En la figura 16 y la figura 15 se observa la información aportada por los distintos nodos y como los cuatro nodos distinguidos quedan por debajo de la entropía de la fuente. En este caso había muchas IPs por debajo de la entropía (casi 30 IPs), así que tomamos como punto de corte los que tengan una probabilidad mayor a 1 %.

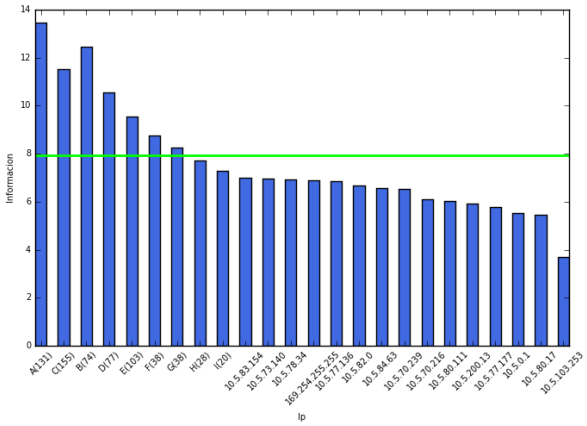

Figura 16. Universidad de La Matanza - Información

En el análisis de la fuente S_1 vemos que la IP "10.5.0.1" podría corresponderse a algún router y de los otros tres nodos no disponemos de suficiente información sobre la red para saber cuál es el rol que cumplen.

IV. Conclusión

En este trabajo se analizaron 4 redes distintas desde el punto de vista de la teoría de la información. Este enfoque nos permitió distinguir y cuantificar la importancia de distintos entes dentro de las redes , ya sean protocolos y su uso relativo en las redes LAN analizadas o nodos que las conforman.

En cuanto a la fuente S, se notó que en todas las redes el protocolo más usado fue IPv4. Otro patrón interesante descubierto es que las redes en donde hay menor cantidad de tráfico de personas, es decir, dispositivos entrando y saliendo del alcance de la red, parecen tener una menor entropía. En efecto, las redes en las cuáles se midió la menor entropía fueron Labos DC y la red hogareña, con 0,034922 y 0,4892 respectivamente.

En cuanto a la fuente S_1 se confirmó que hay una correlación entre ser o no una red controlada y la entropía medida. El patrón es claro al enumerar las redes con sus entropías: red hogareña (2,25), Labos DC (3,417440), Parque (5,757719), Universidad La Matanza (7,951183); donde las primeras dos son controladas y las últimas dos no lo son.

Es interesante observar que la entropía de la fuente S parece ser menor que la de S_1, para cada red. Esto indica que es más predecible el protocolo de un paquete que la dirección destino del mismo. Esto se esperaba, ya que, como ya se mencionó, en todas las capturas, el protocolo

IPv4 prevaleció entre los demás (en mayor o menor medida). Por último, concluímos como se puede ver, que los experimentos realizados probaron las hipótesis planteadas previamente basadas en la Teoría de la Información.

V. Referencias

- N. Abramson, Teoría de la Información y Codificación 5ta edición