Wiretapping

Carolina Wright, Ignacio Rodriguez, Natalia Enrique, Nicolas Mastropasqua

Universidad De Buenos Aires, FCEyN, Dpto.Computación

19 de Septiembre de 2016

Resumen

El siguiente trabajo tiene como objetivo analizar y modelar diferentes redes basándose en conceptos de la teoría de la información. Se intentará estudiar la entropía de las fuentes y la información de sus símbolos, para poder hallar nodos distinguidos, como ***default gateways****. Además se buscara establecer relaciones entre los distintos tipos de redes evaluadas y obtener una conclusión acerca de la eficiencia del método propuesto.*

I. Introducción

A continuación mostraremos algunas nociones básicas para el entendimiento del modelo y análisis que se realizara posteriormente.

1. Definiciones

Dada la aparición de un evento E, diremos que se obtienen

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

unidades de información. De ahora en más, se tomara el logaritmo en base dos, obteniendo como unidad resultante el **bit**.

De esta forma podemos pensar a la red como una fuente de información, la cual emitirá ciertos símbolos con alguna probabilidad y por lo tanto cada uno de ellos aportara una cantidad de información en bits. La elección de los mismos dependerá de lo que se quiera modelar y será tema de discusión en la sección de métodos de dicho informe.

Finalmente, diremos que la cantidad media o esperada de información por símbolo de una fuente S de memoria nula corresponde a su entropía:

$$H(S) = \sum_{s_i \in S} P(s_i) log(1/P(s_i))$$

Notar que si definimos una variable aleatoria **X** cuya rango esta dado por los símbolos de la fuente y le asignamos como probabilidad la frecuencia relativa de las apariciones, se puede ver que la entropía es la esperanza de **I(X)**, siendo **I** la función información. Se puede ver que el valor máximo teórico de la entropía se corresponderá con el logaritmo de la cantidad de símbolos. El mismo se realiza solo cuando los valores son equiprobables. Con este marco teórico, ahora podremos abordar el estudio de las redes evaluadas, así también como el modelado de las fuentes.

II. Protocolo ARP

El protocolo ARP se utiliza para asociar una dirección de red(como una ip) a otra dirección física(como la MAC de Ethernet). Cabe mencionar que no esta limitado a un protocolo determinado, por lo que su uso es mas general. El funcionamiento básico del mismo, asumiendo a **IP** como protocolo de red y a **Ethernet** como el de enlace de datos, se puede resumir en el envío de dos tipos de mensajes. Uno de ellos, **who-has** es un mensaje **broadcast** a la red local que tiene como fin conocer la **MAC** de una ip de un host. El mismo será respondido con un mensaje **is-at** de forma **unicast**, de manera que solo quien envió el mensaje original obtenga la respuesta(es decir, la **MAC** adress del host por el que preguntó originalmente). De esta forma, se construyen las tablas que vinculan a los distintos hosts con sus ip's. Cada determinado tiempo, las mismas son actualizadas con el fin de mantener la información de manera coherente. Finalmente, cabe aclarar que dicho protocolo no provee mayores medidas de seguridad, de ahí que ataques como **ARP poisoning** tengan como objetivo aprovechar dichas vulnerabilidades

II. Métodos

Como se mencionó anteriormente, se buscará modelar a cada red como una fuente de símbolos con memoria nula. Para los experimentos se utilizaron dos construcciones distintas:

- Una fuente S_1, cuyo conjunto de símbolos emitidos esta determinado por los destinos de los paquetes **who-has**, del protocolo **ARP**. Es decir, cada vez que se envié un paquete a la red local preguntando por la **MAC** asociada a una **ip**, esta dirección será considerada un símbolo de nuestra fuente.
  La motivación de este modelo se sostiene bajo la hipótesis que tenemos acerca del comportamiento de los nodos distinguidos, mas precisamente el de un **default gateway**. Dicho nodo es "la puerta de salida"de los hosts de una red local hacia otras redes, por lo que se supone que su dirección será consultada con frecuencia por los distintos dispositivos que quieran acceder a Internet. Mas precisamente, cada vez que quieran comunicarse con alguna ip del .exterior"(que sería lo mas comun en una red publica), los hosts se veran obligados a mandar el paquete por el **default gateway** y para ello, deberan conocer su **MAC**. Luego, mandaran una petición ARP(who-has) para obtenerla. De esta forma, la "probabilidad de emisión"del símbolo asociado a dicho nodo sera elevada y, por consiguiente, la información que aporte una aparición del mismo será baja.
  Asumiendo que la interacción entre los hosts de dicha red va a ser muy baja (por ejemplo en una red publica de un bar), podemos pensar que ,además, dichas direcciones no serán consultadas con normalidad y por lo tanto su información sera relativamente alta a comparación con la de un nodo distinguido.
  Además es de esperar que la entropía de dicha fuente sea elevada en redes públicas, ya que los dispositivos que se conectan a ella son altamente impredecibles, por lo que el grado de incertidumbre que aportan sus símbolos, en general, debería ser relativamente alto. Todo lo contrario se debería esperar para redes hogareñas y pequeñas, donde el grado de conocimiento entre los hosts es alto , al igual que la previsibilidad de la información.
- Una fuente S, que define el conjunto de símbolos emitidos como $\{S_{BROADCAST}, S_{UNICAST}\}$. Durante una captura, cualquier paquete **Ethernet** será tenido en cuenta y se clasificará como broadcast si su **MAC** destino es **FF:FF:FF:FF:FF:FF**. En este caso, es de esperar que una red grande y publica tenga un mayor trafico del tipo broadcast, con mas dispositivos haciendo peticiones ARP(who-has).

Bajo el modelo anterior se realizaron mediciones en redes de distinto tamaño, con la intención de capturar el trafico de cada fuente para poder evaluar y comparar la información de sus símbolos, su entropía y encontrar nodos distinguidos. Para las mediciones seleccionamos cuatro fuentes distintas y las evaluamos por un lapso de tiempo para luego ser analizadas. Las placas se encontraban en **modo promiscuo**, permitiendo escuchar ,además del trafico **broadcast**, todo tipo de mensajes **unicast**. Los graficos de las redes con mas de diez nodos fueron agrupados por una ip representante", que tiene la misma informacion que los que fueron omitidos. Al momento de realizar las mismas, los integrantes nos encontramos con distintos escenarios. A continuación se muestran las redes intervenidas:

- Laboratorios Departamento de Computación: Realizamos una medición de 30 minutos conectados a la red Wifi de los Laboratorios. En este caso utilizamos un Sistema Operativo Windows, nos parece importante destacar este punto ya que en la sección de resultados ahondaremos en detalle cómo influyó esto en nuestros resultados.
- Red Doméstica: Realizamos una medición de 18 minutos conectados a una red local Ethernet. La misma consiste de tres computadoras(una de las cuales realizo la experiencia) conectadas por medio de un cable UTP a un modem/switch hogareño que provee el **ISP**. Dicho dispositivo también funciona como un Wireless **A**ccess**P**oint, por lo que otros dispositivos inalámbricos se conectan a Internet de esta forma(celulares por ejemplo). Sin embargo, como la computadora en la que se realizo la medición no tiene placa WIFI, no podría escuchar el trafico unicast entre los dispositivos inalámbricos, aun estando en modo promiscuo. Con respecto a las condiciones de la medición; se intento que ningún dispositivo inálambrico se en-

contrase conectado, evitando así generar ruido". En el momento de la captura, las dos computadoras restantes se encontraban reproduciendo videos de Internet, mientras que la otra, además de ejecutar **wireshark**, tuvo un uso de red "normal"(navegación en Internet).

- Red Laboral: Se realizo una medición de una hora y media, en el cual en un principio solo había seis maquinas conectadas a la red, pero al finalizar la medición había diez. Las mediciones se realizaron en un sistema operativo Windows, con **wireshark**.
- Red pública: Se realizo una medicion en el Shopping Dot. Es una red abierta y se desconoce su estructura. No tenemos datos de cuántos dispositivos estaban conectados al momento de la captura.

III. Resultados y Análisis

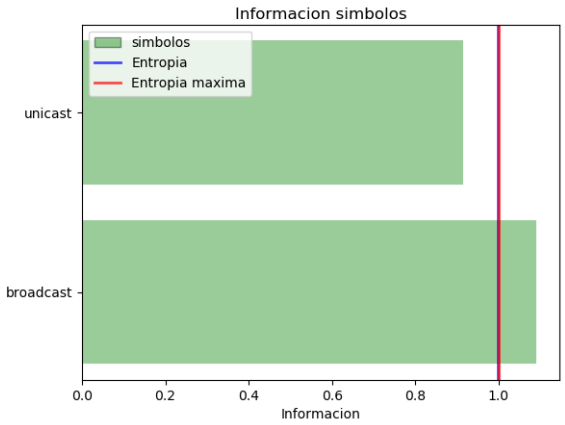
I. Resultados y Análisis Fuente S

I.1. Laboratorios Departamento de Computación

Luego de capturar los paquetes por medio de Wireshark obtuvimos los siguientes resultados:

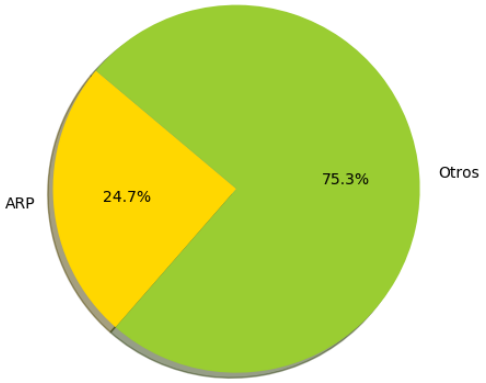
Cuadro 1: *Laboratorios DC, info. por símbolo*

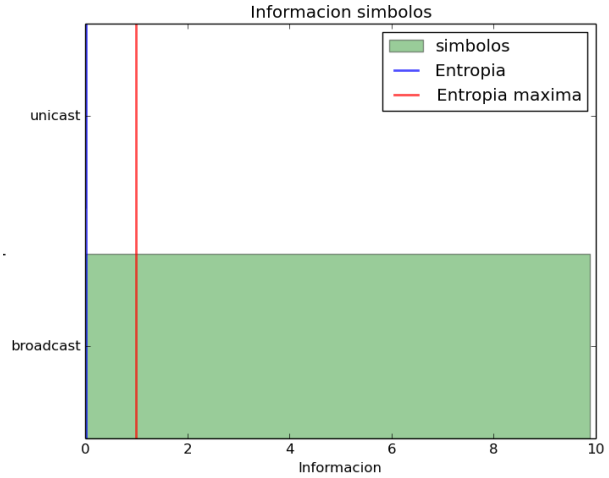
Datos fuente S		
Símbolo	Proba.	Info.
$S_{UNICAST}$	0,530646	0,914179
$S_{BROADCAST}$	0,469354	1,091251
$H(\mathrm{S}) = 0{,}997288$		
$H_{max}(S) = 1$		



Pudimos notar como la probabilidad de cada símbolo fue casi de 50 % para cada uno, al realizar una medición en una red abierta, se esperaba que la entropía de la fuente **S** se acerque a la máxima dado que suponíamos símbolos equiprobables. Luego, debido a lo mencionado anteriormente pudimos confirmar ésta hipótesis. Otro punto interesante que nos interesa destacar es que, como mencionamos anteriormente,en este caso, realizamos las mediciones a través de la herramienta Wireshark en un Sistema Operativo Windows, al notar los resultados de nuestro experimento, que si bien confirmaron nuestra hipótesis notábamos que a medida que capturábamos los paquetes los únicos de tipo unicast que podíamos ver era el de nuestro dispositivo, aún estando en modo promiscuo. Investigando, pudimos conocer que la razón de este hecho es que en Windows se encuentra la librería Winpcap, la cual no es compatible con la librería que utiliza Wireshark (libpcap)

Podemos observar que el **overhead** de los paquetes ARP, que en este caso es del 24 %. Notamos que es alto debido a las mediciones tomadas, ya que leímos más paquetes del tipo Broadcast.





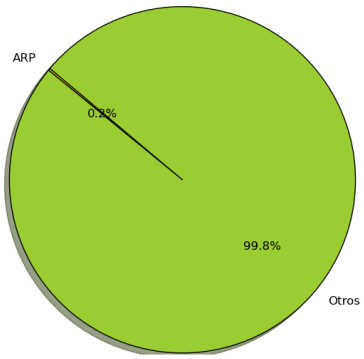
Se esperaba que la fuente **S** tuviese una baja entropía debido al tamaño de la red. La probabilidad de aparición de los símbolos debería alejarse de la equiprobabilidad , siendo mas alta en los broadcast. Sin embargo, la brecha entre ambas fue muy superior a lo previsto, traduciéndose en una entropía casi nula. Notar tambien que dicho se aleja demasiado de la entropía máxima teórica que ocurre bajo equiprobabilidad, y que, en el caso de una fuente binaria, es igual a uno.
En cuanto al overhead de paquetes ARP, los resultados nos muestran que el intercambio de los mismo es realmente muy escaso y el trafico generado por los mismos es casi despreciable comparado con el resto.

i.2. Red Doméstica

Luego de obtener la captura de esta red, se analizo el trafico bajo el modelo referido como fuente **S**, que tiene como finalidad observar el comportamiento del trafico broadcast y unicast. El script que procesó el archivo **.pcap** obtenido previamente, arrojo los siguientes datos, resumidos en la tabla y gráfico a continuación.

Cuadro 2: *Red Doméstica, info. por simbolo*

Datos fuente S		
Símbolo	Proba.	Info.
$S_{UNICAST}$	0,9989	0,0015
$S_{BROADCAST}$	0,0010	9,8941
$H(\mathrm{S}) = 0{,}011913$		
$H_{max}(S) = 1$		



Observando con mayor detenimiento la razón de lo ocurrido, notamos que la escasez de los paquetes broadcast(hecho que fuerza a que dichos símbolos

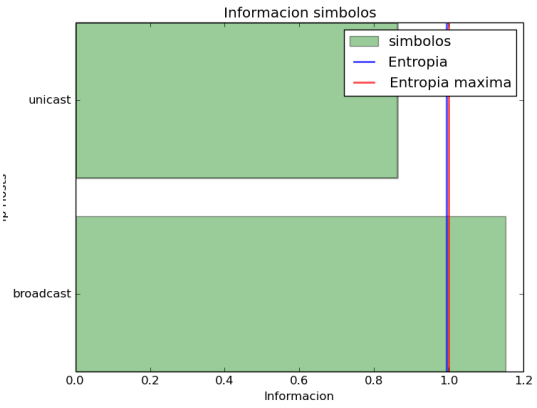
tengan una información muy alta) se debe a que , una vez que se intercambiaron los primeros who-has entro dos host dados, los dispositivos se conocen entre sí y luego los sucesivos paquetes who-has que se producían eran entre ellos(unicast). Investigando, encontramos que en ocasiones, para evitar congestionar la red de pedidos broadcast, el módem genera un tipo de **ARPing** con un dispositivo para validar la **MAC** que tiene guardada y si responde afirmativamente, conservarla. De esta manera se evita poblar la red de pedidos broadcast para actualizar las tablas ARP.

i.3. Dot Shopping

Con los datos recolectados en esta red, obtuvimos la siguiente información.

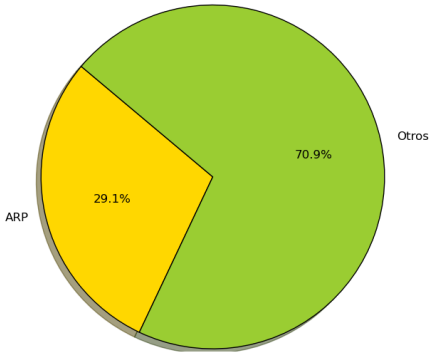
Cuadro 3: *Red pública, info. por símbolo*

Datos fuente S		
Símbolo	Proba.	Info.
$S_{UNICAST}$	0,550000	0,862496
$S_{BROADCAST}$	0,450000	1,152003
$H(\mathrm{S}) = 0{,}992774$		
$H_{max}(S) = 1$		



En este caso, para la fuente S, se esperaba una entropía alta. Esta suposición proviene de la captura sobre una red abierta, luego la incertidumbre con respecto a los símbolos debería alta, es decir, símbolos equiprobables.

En la muestra tomada, obtuvimos un overhead de paquetes ARP de un 29,1 %. Tiene sentido que este valor sea alto porque los dispositivos no se conocen entre sí, y en general, hay nuevos dispositivos conectándose a la red permanentemente.

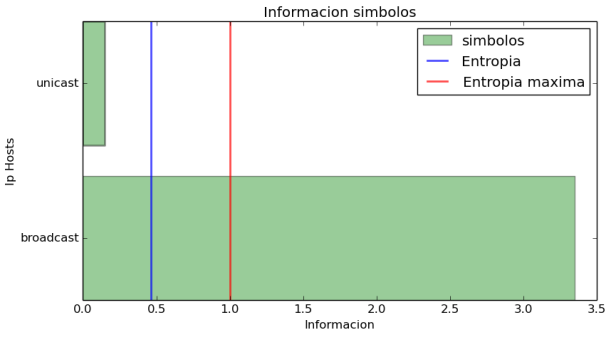


i.4. Red de Trabajo

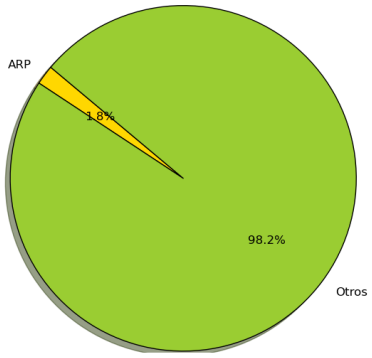
Luego de la recolección de datos, estos fueron analizados y se obtuvo la siguiente información.

Cuadro 4: *Red de trabajo, info. por simbolo*

Datos fuente S		
Símbolo	Proba.	Info.
$S_{UNICAST}$	0,901750	0,149201
$S_{BROADCAST}$	0,098250	3,347399
$H(\mathrm{S}) = 0{,}463424$		
$H_{max}(S) = 1$		



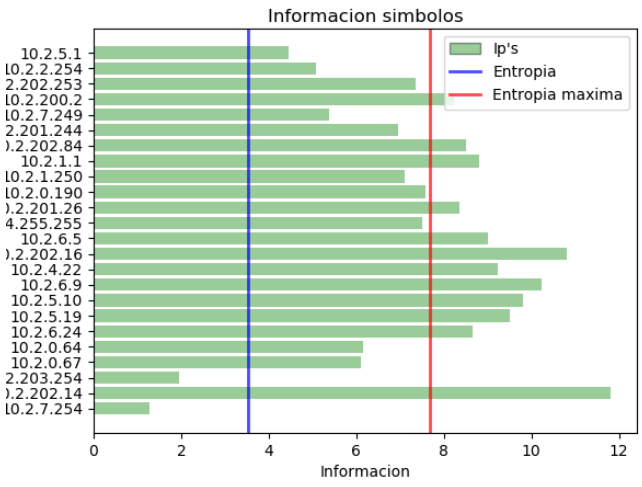
Para esta fuente,como es una red relativamente chica, era de esperar que su entropía sea baja. Sin embargo, debido a que constantemente se estaban sumando nuevas maquinas a la red la entropía es un poco mas alta de lo esperado, pero si se hubiese continuado la medición o se hubiese hecho en un momento posterior, en el cual las maquinas ya se conocían entre todas, la entropía hubiese sido mas baja.



II. Resultados y Análisis Fuente S_1

ii.1. Laboratorios Departamento de Computación:

Con la misma captura realizada, analizamos los datos para la fuente S_1.



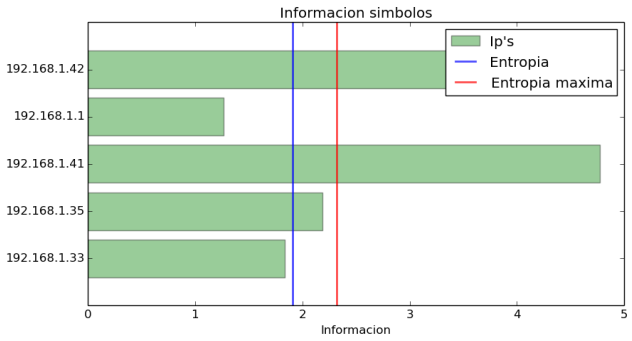
Nuestros resultados arrojan que dos de los símbolos se encuentran por debajo de la entropía que calculamos. El que correspone a la menor cantidad de información, desconocemos de que red se trata pero podemos suponer que le pertenece a algún servidor de la red, mientras que el segundo (**10.2.203.254**) corresponde al router de la red del Departamento. Bajo este resultado podemos decir que el criterio de distinción para descubrir el Default Gateway de la red no es preciso, ya que nos dio otro nodo con menor información. Otro punto para destacar, es que el nodo correspondiente a nuestra computadora no aparece en este experimento, creemos que esto puede deberse a que no se estaba utilizando en otra cosa, y el default gateway sólo preguntó por ella cuando se conectó a la red, antes de realizar las capturas. Lo que sí pudimos notar es que nuestra ip sí manda mensajes who-has al default Gateway de los Laboratorios.

Cuadro 5: *Departamento de Computación, info. por símbolo*

Datos fuente S_1		
Símbolo	Proba.	Info.
10.2.5.1	0,046022	4,441544
10.2.2.254	0,029387	5,088663
10.2.202.253	0,006099	7,357152
10.2.200.2	0,003327	8,231621
10.2.7.249	0,024120	5,373640
10.2.201.244	0,008040	6,958603
10.2.1.250	0,007208	7,116144
10.2.0.190	0,005268	7,568656
10.2.201.26	0,003050	8,357152
169.254.255.255	0,005545	7,494656
10.2.6.5	0,001941	9,009229
10.2.202.16	0,000554	10,816584
10.2.4.22	0,001663	9,231621
10.2.6.9	0,000832	10,231621
10.2.5.10	0,001109	9,816584
10.2.5.19	0,001386	9,494656
10.2.6.24	0,002495	8,646659
10.2.0.64	0,014139	6,144158
10.2.0.67	0,014416	6,116144
10.2.203.254	0,257000	1,960158
10.2.202.14	0,000277	11,816584
10.2.7.254	0,416413	1,263915
$H(S) = 3{,}519460$		
$H_{max}(S) = 7{,}679480$		

ii.2. Red Doméstica:

A continuación se muestra el gráfico obtenido por el script que recibió la captura realizada en primera instancia.



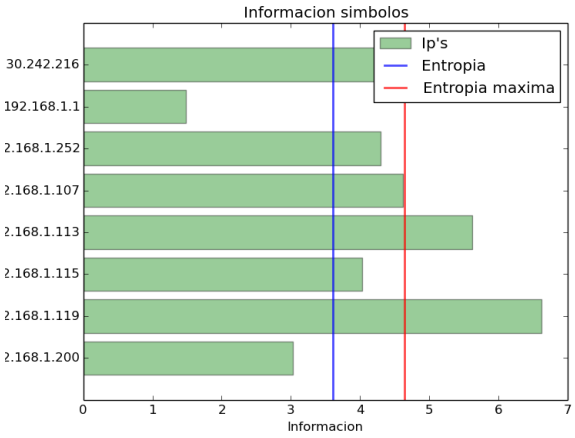
El gráfico muestra los símbolos de la fuente(las ip destino de los who-has) junto con la información que aportan. Como habíamos supuesto, se verifica que el **default gateway** coincide con el nodo de menor información(**192.168.1.1**). A su vez, otro de los nodos que se encuentran por debajo de la entropía de la fuente es de la computadora que realizo la medición(**192.168.1.33**), que solo corresponde a un host normal en la red. También observamos cierta anomalía en el funcionamiento de las peticiones ARP(who-has). El dispositivo asociado a **192.168.1.35** era un celular que, como se dijo anterioremente, no se encontraba conectado a la red. Sin embargo, en un momento dado, el default gateway comenzó a enviar paquetes who-has preguntando por dicho host. Esto ocurrió repetidas veces, al no obtener una respuesta del mismo y por lo tanto, elevo la probabilidad de aparición de este host(bajando su información). Por otro lado se ve como la entropía de esta fuente esta mas cercana al valor teórico máximo. Finalmente, si bien el método permitió distinguir nodos, cabe mencionar que en otras mediciones de prueba el comportamiento se alejaba fuertemente de lo esperado, por irregularidades similares a las que se comentaron anteriormente.

Cuadro 6: *Red Doméstica, info. por símbolo*

Datos fuente S_1		
Símbolo	Proba.	Info.
192.168.1.41	0,036	4,772
192.168.1.42	0,048	4,357
192.168.1.35	0,219	2,187
192.168.1.33	0,280	1,833
192.168.1.33	0,414	1,270
$H(S) = 1{,}908$		
$H_{max}(S) = 2{,}321$		

ii.3. Dot Shopping

Con los resultados que obtuvimos anteriormente, estudiaremos ahora la fuente S_1

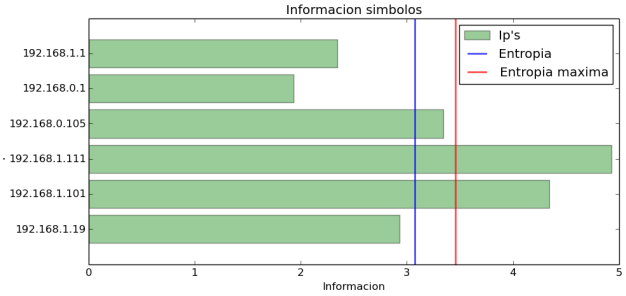


Al ser una red pública, no conocemos los dispositivos conectados a la misma ni su distribución. Sin embargo era de esperarse, que el símbolo con menor información es el de la ip 192.168.1.1, que corresponde al Default Gateway. Otro de los símbolos que nos figuran por debajo de la entropía calculada es el 192.168.1.200, que desconocemos a que dispositivo puede pertenecer. Podemos notar además que varias de nuestros siembolos se acercan a la entropía máxima teórica, esto se debe a lo mencionado anteriormente al ser una red abierta.

Cuadro 7: *Red pública, info. por simbolo*

Datos fuente S_1		
Símbolo	Proba.	Info.
192.168.1.1	0,357143	1,485427
192.168.1.119	0,010204	6,614710
192.168.1.115	0,061224	4,029747
192.168.1.113	0,020408	5,614710
181.30.242.216	0,030612	5,029747
192.168.1.200	0,122449	3,029747
192.168.1.107	0,040816	4,614710
192.168.1.252	0,051020	4,292782
$H(S) = 4{,}067$		
$H_{max}(S) = 6{,}512$		

ii.4. Red de trabajo



Cuadro 8: *Red de trabajo, info. por simbolo*

Datos fuente S_1		
Símbolo	Proba.	Info.
192.168.0.1	0,262295	1,930737
192.168.1.1	0,196721	2,345775
192.168.1.19	0,131148	2,930737
192.168.0.105	0,098361	3,345775
192.168.1.101	0,049180	4,345775
192.168.1.111	0,032787	4,930737
192.168.0.112	0,032787	4,930737
$H(S) = 3{,}073299$		
$H_{max}(S) = 3{,}459432$		

En el caso de esta red, el **default gateway** es la ip **192.168.0.1**, que a su vez es el nodo con menor información, pero es un nodo destacado. La ip del dispositivo que realizaba la medición es el **192.168.1.112**, que no fue detectado como un nodo distinguido. Esto se puede deber a que por mas de que era el que realizaba la medición no era usado para otra cosa. Tambien se noto que existe otro **default gateway** con la ip **192.168.1.1**, que es de la red de backup. A esta direccion se le hace who has varias veces pero unicamente desde el chip de ethernet. Podría entenderse que dicho nodo es parte de la redundancia de la red, y que tiene alguno de sus puertos bloqueados.

IV. CONCLUSIONES:

Luego de observar los distintos experimentos realizados sobre la fuente **S**, podemos concluir que existe una correlación positiva entre la entropía y la cantidad de hosts en una red. Notamos que si ordenando las redes por su entropia de forma creciente(Doméstica,de Trabajo,Publica(Laboratorios y Shopping)) obtenemos el mismo orden si se lo hace por su jerarquía de tamaño. Además el **overhead** por tráfico ARP nunca supero el 30 % de los paquetes Ethernet intercambiados. Dicho proporción tambien esta ligada al tamaño de la red, siendo mayor en las redes públicas.
Por otra parte, los distintos resultados de los modelos de la fuente **S1** muestran exitosamente a ,al menos, el nodo distinguido correspondiente al **default gateway**, respaldando la hipótesis que suponíamos. Es decir, que ellos iban a ser los de menor información. Sin embargo, nos resulta dificil poder concluir algo sobre la topología de las redes subyacentes. En general, la entropía de estas fuentes se encontraba a una distancia menor que el 70 % de la máxima, con excepción de la red del laboratorio, donde era levemente mayor. El método nos permitió asociar a el default gateway con los nodos que mostraban menor información con un grado de precisión que disminuye conforme lo hace el tamaño de la red. El caso extremo es el de la red hogareña, donde realizando varias capturas se obtenían datos muy dispares, hasta que se encontró lo que buscábamos. En la red de trabajo, la diferencia entre los nodos que tienen información menor a la entropía es también sutil. Por lo tanto se verifica de cierto modo dicho supuesto.

En la mayoría de los casos, los nodos cuya información es menor que la entropía, se corresponden con alguna funcionalidad particular. En las redes publicas, esto se verifica de cierto modo. En los laboratorios, sospechamos que los que se encuentran por debajo, corresponden a un default gateway y a algún servidor. Para la red laboral, dos de los tres nodos son puerta de salida. En el caso del Shopping, no podemos afirmar nada acerca del nodo que no es el default gateway y que también esta por abajo del valor de la entropía. Finalmente, para el caso de la red domestica, lo mismo se sostiene, con la salvedad del host que realizó la medición también se incluye como "destacado".