Teoría de las Comunicaciones
Trabajo Práctico 1
Wiretapping

Joaquín Arribas Martín Forte Ignacio Lebrero Jésica Vázquez

Resumen

En el presente trabajo nos proponemos experimentar sobre diversas redes WI-FI, analizando la información obtenida por dos fuentes distintas en base a qué símbolos son significativos en cada una. Para esto utilizamos herramientas formales de la teoría de la información.

Index Terms

Información, Entropía, Fuente, IP, Host

I. Introducción

En este trabajo presentamos un análisis sobre paquetes capturados en distintas redes WI-FI. Dado que existen diversos tipos de paquetes nos centraremos en los paquetes ARP, los cuales se utilizan para relacionar direccionamiento de capa 3 (en nuestro caso IP) con direccionamiento de capa 2 (en nuestro caso Mac address). En otras palabras, se utilizan para consultar que Mac tiene asociada una IP determinada en la red.

Para el análisis de la red modelaremos distintas fuentes de información como diferencias entre paquetes *Unicast* y *Broadcast*, así como también modelar otras más complejas que tenga en cuenta IP origen, IP destino y el tipo de mensaje (WHO HAS/IS AT). A partir de estas fuentes analizaremos la diferencia entre la información de cada símbolo y la entropía de la fuente. Esto nos permitirá identificar los nodos destacados según algún criterio que explicaremos luego, así como también características particulares de cada red.

II. Experimentación

A. Redes Analizadas

Para la realización de este trabajo capturamos paquetes de distintas redes de WI-FI:

1. La casa de uno de los integrantes. IPs interactuando en la red: 59, Paquetes transmitidos: 50686
2. Un café *Starbucks*. IPs interactuando en la red: 56, Paquetes transmitidos: 4275
3. Dot Baires Shopping. IPs interactuando en la red: 248, Paquetes transmitidos: 623929
4. Shopping *Village Recoleta*. IPs interactuando en la red: 41, Paquetes transmitidos: 29409

La elección de estos lugares fue a partir de que decidimos experimentar sobre redes relativamente chicas (1, 2) y dos grandes (3, 4), tomando como medida la cantidad de dispositivos que posiblemente puedan interactuar. Suponemos que en un cafe o en una red hogareña la cantidad de dispositivos será mucho menor a la de cualquier Shopping.

B. Modelado de Fuentes

Modelamos los paquetes capturados en dichas redes como dos fuentes de información distintas S y S_1:

- El conjunto de símbolos definidos en S son $\{s_{Broadcast}, s_{Unicast}\}$ correspondientes a los mensajes Broadcast y Unicast de la red. La manera de realizar dicha distinción entre paquetes fue mediante la *destination address*, que en caso de ser igual a ff:ff:ff:ff:ff:ff indica que es Broadcast, y caso contrario Unicast.
- El conjunto de símbolos definidos en S_1 son, para cada nodo (host) de la red, la cantidad de mensajes envíados a ellos preguntándole por una determinada IP. Los mensajes se dividen en dos tipos:

- **Who Has**: cuando un host quiere averiguar qué host tiene determinada dirección IP.
- **Is At**: cuando un host quiere responderle a otro que efectivamente tiene la dirección IP pedida. Devuelve su *MAC address*.

Lo que nos interesa entonces es ver la cantidad de **Who Has** que le enviaron a cada IP. Los símbolos serán $\{IP_{dest_1}, ..., IP_{dest_n}\}$

Para analizar cada fuente se tuvo en cuenta:

- La entropía de la fuente para entender qué tan distribuidos están los datos dentro de la misma.
- La información de sus símbolos para compararla con la entropía de la fuente y de este modo detectar los nodos destacados.

C. Métodos utilizados

Para realizar un análisis más riguroso respecto de los datos obtenidos utilizamos métricas propias del área de la teoría de la información:

- La *información* de un evento e que, dada la probabilidad de ocurrencia $P(e)$, se calcula como:

$$I(e) = -log_2(P(e)) \text{ bits}$$

- La *entropía* de una fuente S que, dada la probabilidad de suceso $P(s)$ de todos los símbolos $s \in S$, se calcula como:

$$H(S) = -\sum_{s \in S} P(s) * log_2(P_S(s))$$

III. Resultados

A continuación, exhibiremos para las dos fuentes modeladas los experimentemos que realizamos y sus conclusiones, utilizando la información obtenida de las distintas redes de WI-FI utilizadas.

A. Fuente S

Como mencionamos previamente, los símbolos definidos en la fuente S son $\{s_{Broadcast}, s_{Unicast}\}$. En este caso, dado que hay tan solo 2 símbolos, la entropía máxima será 1 si son equiprobables. Este resultado se deduce de las fórmulas previamente exhibidas.

A.1 Red de una casa

La cantidad total de paquetes transmitidos en 37 minutos fueron 50686 de los cuales:

	Cantidad total	**Probabilidad de ocurrencia**	**Información**
Mensajes Brodcast	1669	0.032928	4.924531
Mensajes Unicast	49017	0.967072	0.048305

La entropía de la fuente S es 0.208871.

A.2 Red de un Starbucks

La cantidad total de paquetes transmitidos en 84 minutos fueron 4275 de los cuales:

	Cantidad total	**Probabilidad de ocurrencia**	**Información**
Mensajes Brodcast	851	0.199064	2.328693
Mensajes Unicast	3424	0.800936	0.320242

La entropía de la fuente S es 0.720053.

A.3 Red del Village Recoleta

La cantidad total de paquetes transmitidos en 22 minutos fueron 29409 de los cuales:

	Cantidad total	**Probabilidad de ocurrencia**	**Información**
Mensajes Brodcast	578	0.019654	5,669044
Mensajes Unicast	28831	0,980346	0,028637

La entropía de la fuente S es 0,139493.

A.4 Red del Dot Baires Shopping

La cantidad total de paquetes transmitidos en 17 minutos fueron 623929 de los cuales:

	Cantidad total	**Probabilidad de ocurrencia**	**Información**
Mensajes Brodcast	29775	0,047722	4,389208
Mensajes Unicast	594154	0,952278	0,070545

La entropía de la fuente S es 0,276639.

A.5 Análisis

En ninguno de los casos la entropía de la fuente fue máxima ya que la distribución de los paquetes *Unicast* y *Broadcast* no fue equiprobable. Al ser más frecuentes los paquetes *Unicast* estos proporcionan menos información.

Creemos que la cantidad de paquetes *Broadcast* envíados, de los cuales un porcentaje significante son ARP, se deben a que son, por lo general, utilizados para establecer conexiones o envíar datos entre distintos dispositivos de uso común de la red, lo cual genera una carga adicional.

Como suponemos una estrecha relación entre los paquetes *Broadcast* y el overhead de la red, a mayor entropía (mejor distribución de los símbolos) mayor será el overhead en la red.

Análogamente, a menor proporción de paquetes *Broadcast*, menor overhead, y por lo tanto al ser mayor la proporción de paquetes *Unicast* menor será la entropía.

Algo que nos llamó la atención fue que el overhead en el *Starbucks* fue mucho mayor al resto y creemos que se puede deber a que en este la relación de personas que están de paso comparada a las que están fijas es mucho mayor al resto. Esto causa que haya muchos más mensajes *Broadcast* para establecer conexión, que *Unicast* para intercambiar información. Esto se puede ver de manera intuitiva al compararlo con la red doméstica, pero no tanto con los shoppings.

B. Fuente S_1

Como se mencionó previamente se modeló esta fuente como los paquetes ARP de tipo *Who Has*. Más en particular, observando la dirección destino.

Definiremos para cada red un grafo dirigido donde los nodos serán las direcciones IP y las aristas representarán los paquetes. De esta manera, un nodo A tendrá una flecha hacia B si existe algún paquete con IP origen A e IP destino B. Para cada nodo definiremos su peso como la cantidad de paquetes que recibe. Sobre cada uno de estos grafos haremos un análisis para entender la topología de cada red, y obtener datos relevantes.

Para distinguir un host se tuvo en cuenta la información que proveía y la entropía de la fuente. Diremos que un nodo es distinguido si la información que provee es menor que la entropía de la fuente. Formalmente, sea s un elemento de la fuente S, buscamos que $\mathrm{I(s)} < \mathrm{H(S)}$.

B.1 Red de una casa

La cantidad total de paquetes transmitidos en 37 minutos fueron 50686 de los cuales 1712 fueron ARP ($\mathrm{P(ARP)} = 0.033777$). En el siguiente gráfico podemos distinguir los distintos símbolos de la fuente junto con su frecuencia:

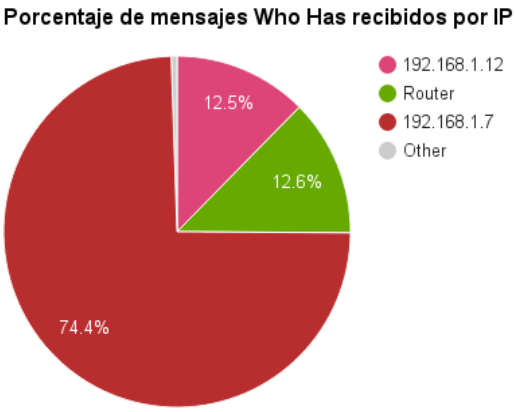

Figura 1. Símbolos por frecuencia

Son tres los nodos que nos despiertan interés:

- El Router, al cual le llegaron el 12.6 % de los mensajes *Who Has*.
- Creemos que el símbolo de mayor frecuencia fue ocasionado por un dispositivo utilizando la aplicación *whatsapp-web* lo cual causó un gran caudal de paquetes en busca del celular al que le correspondía la sesión.
- La IP 192.168.1.12 que creemos que fue algún otro dispositivo del hogar.

A continuación mostramos un gráfico que muestra para cada símbolo de la fuente su información, al mismo tiempo que exhibimos el punto de corte (la entropía) por el cual consideramos que un host es distinguido.

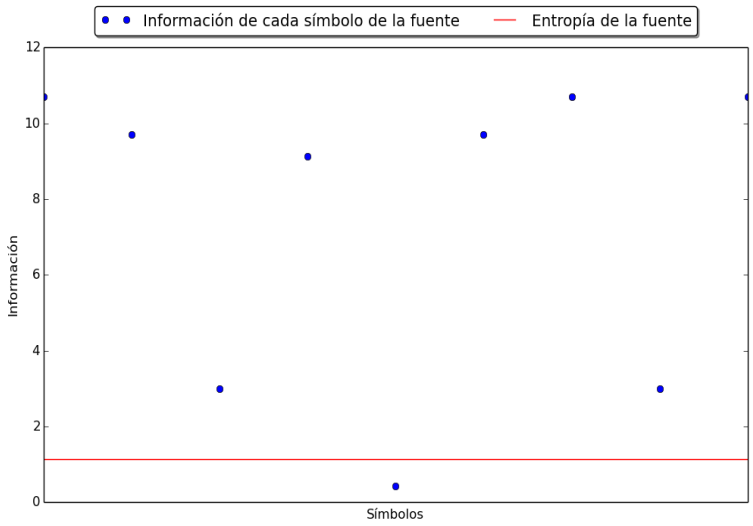

Figura 2.

El nodo distinguido (por debajo del corte) corresponde a la IP 192.168.1.7

Observamos aquí el contraste que surge entre la probabilidad de los eventos y su información, donde los símbolos con mayor probabilidad son los que menos información dan y viceversa.

A continuación mostraremos el grafo de conexiones ARP subyacente previamente mencionado:

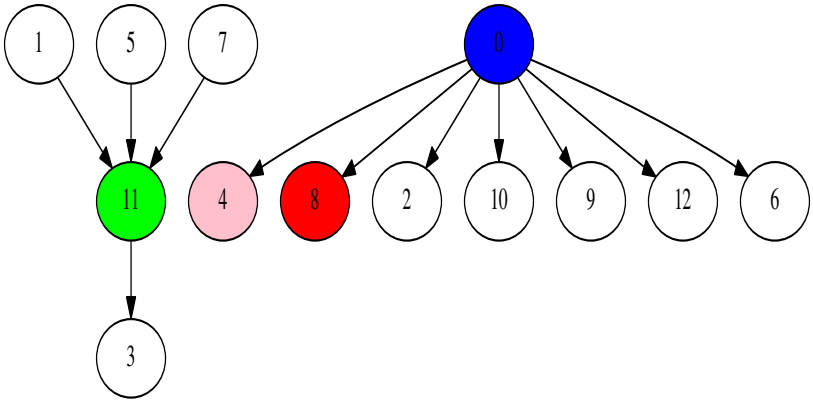

Figura 3. Grafo subyacente de la red doméstica

Coloreamos determinados nodos de manera ilustrativa, para realizar un análisis a continuación. De este gráfico y el fácil acceso a la información de la red podemos sacar las siguientes conclusiones:

- Los nodos 1, 5 y 7 son dispositivos, celulares o computadoras, que se unieron a la red. Se comunicaron únicamente con el *router* representado por el nodo 11 (marcado en verde).
- El nodo 0 (marcado en azul) es la computadora que más se comunico por la red. Fue la que constantemente preguntó por el nodo 8 (marcado en rojo), el cual fue la IP que apareció con mayor frecuencia, y por el nodo 4 (marcado en rosa) también descripto en el gráfico de tortas previo.

La entropía de la fuente fue 1.128559 mientras que la entropía máxima posible en este caso es 3.169925. La entropía es baja dado que hay un símbolo que es muy repetido en la fuente, los mensajes envíados desde el nodo 0 al nodo 8, y porque hay otros que sucedieron muy esporádicamente, sin interactuar de manera frecuente en la red.

Como tenemos un conocimiento de la red en la cual experimentamos, podemos deducir que en este caso la herramienta no fue muy efectiva, ya que no reconoció como nodo distinguido el *Default Gateway*. Creemos que al ser una red chica el comportamiento anómalo del nodo 8 pudo haber generado mucho ruido en el análisis.

B.2 Red de un Starbucks

La cantidad de paquetes transmitidos fue 4275, de los cuando 532 son ARP. La probabilidad de que sea ARP es P(ARP) = 0,124.
A continuación mostramos la red de mensajes ARP subyacente a un Starbucks (ver 4).

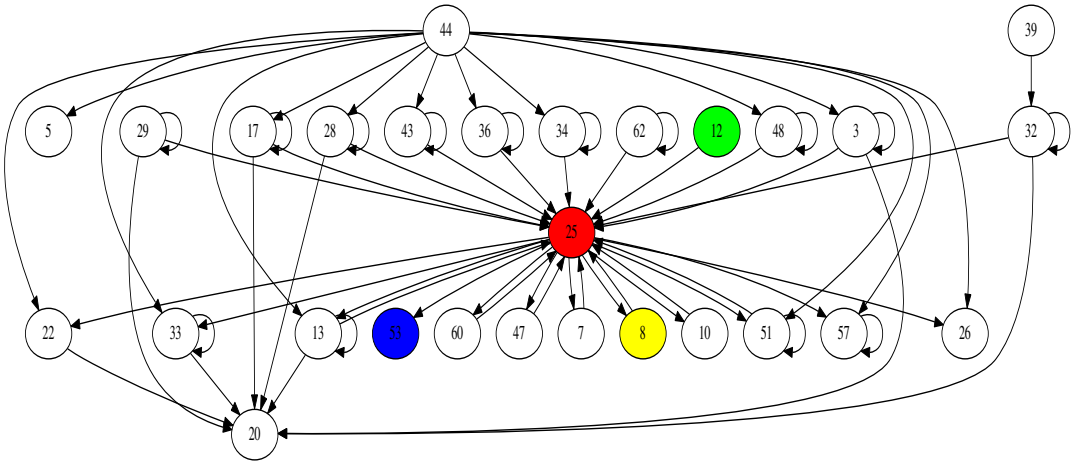

Figura 4. Red de mensajes de Starbucks

Algunas observaciones sobre este gráfico (los coloreados son algunos nodos sobre los que hacemos aclaraciones):

- La IP de nodo “25” fue la más requerida.
- La IP del nodo “44” preguntó por muchas direcciones IP distintas y ningún nodo preguntó por su IP.

- El host correspondiente al nodo “8” preguntó por la del “25” 29 veces.
- El host correspondiente al nodo “12” preguntó por la del “25” 166 veces.
- El host correspondiente al nodo “25” preguntó por la del “53” 29 veces.
- Los demás host preguntaron por IPs menos de 10 veces.
- La dirección del nodo 20 fue muy requerida.

Con la información obtenida nos pareció que el nodo 25 podría ser el *Default Gateway*, dada la cantidad de veces que fue requerida en comparación con los demás. Hubo muchos hosts que preguntaron por la dirección IP del que suponemos es el router pocas veces, y suponemos que eso puede deberse a los celulares y computadoras que estaban conectándose a la red constantemente. Otro nodo con un comportamiento diferente es el 20, ya que varios lo requieren pero este no manda ningún mensaje preguntando por una IP. El nodo 44 requiere a muchas direcciones IP y ninguna lo requiere a este. Una hipótesis que tuvimos fue que pertenece a un dispositivo que se está conectando y desconectando todo el tiempo y cada vez que se conecta hace un requerimiento de IP diferente. Creemos que el nodo 12 podría ser un dispositivo del lugar (o que lleve el suficiente tiempo necesario) ya que hizo muchos requerimientos de la IP del nodo 25, pero en nigún momento fue destinatario de un *Who Has*, (tampoco del host correspondiente al nodo 25 que creemos que es el router) con lo cual el router podría tener guardada su MAC address y entonces no necesitar mandar un *who has*.

A continuación, mostramos en un gráfico de la cantidad de información que proveen los distintos hosts. En el eje horizontal vemos la información y en el vertical la cantidad de nodos que tienen dicha información.

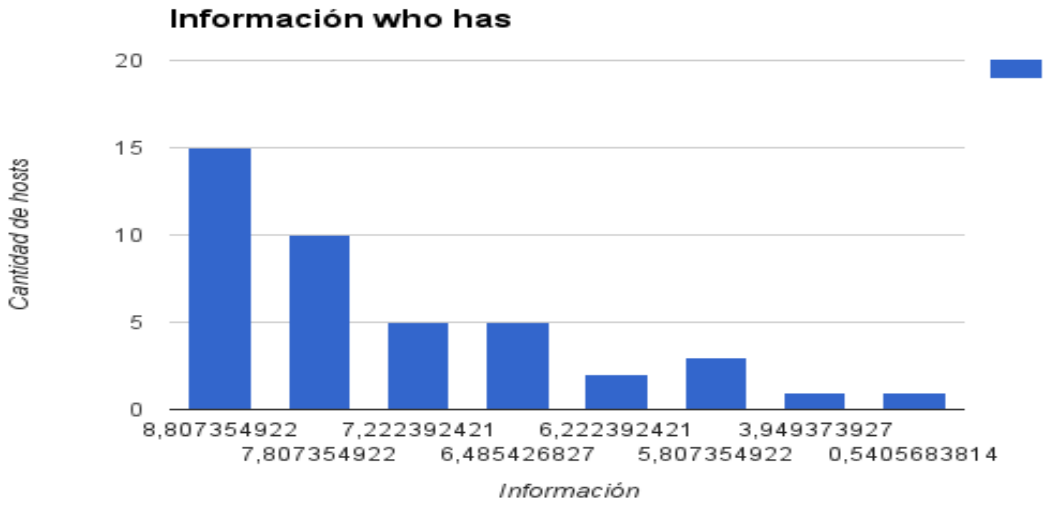

Figura 5. Cantidad de hosts para cada información en Starbucks

Como se ve en la figura 5 hay muchos nodos que proveen mucha información (que suponemos que son las computadoras y celulares que se conectan ocasionalmente).

A continuación mostraremos un gráfico, en el cual el eje horizontal representa los símbolos y el vertical la información que provee cada uno. La línea roja es la entropía de la fuente.

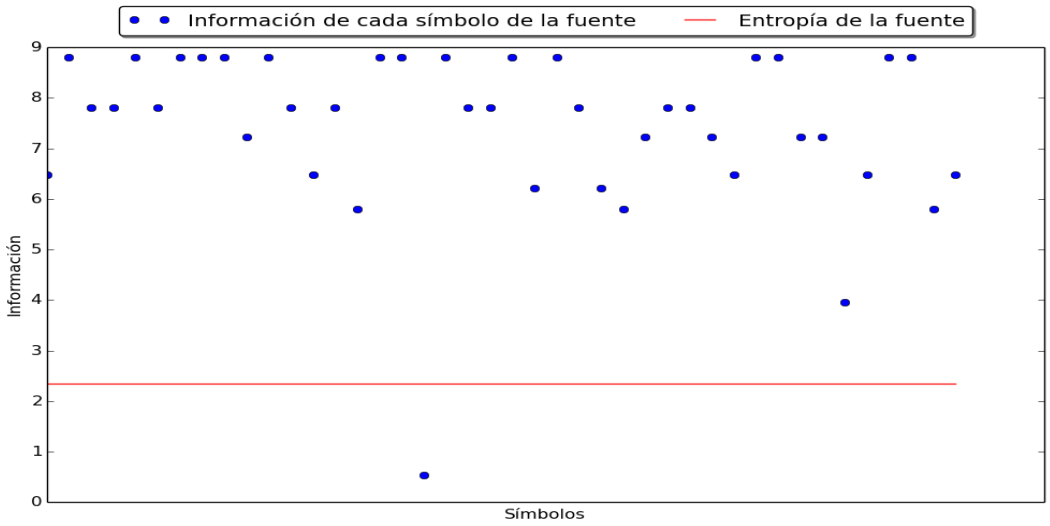

Figura 6. Información de cada host en Starbucks

La entropía de esta fuente es 2,3522. Entropía de la fuente máxima es 5,3923. El único nodo distinguido es el mismo que llamamos 25 (el rojo) en el grafo, que tiene menos información que la entropía, el cual suponemos el router.

Información de algunos nodos:

Nodo	**Información**
8	8,81
25	0,54
53	3,95
20	5,81

Como se ve en las informaciones particulares, el nodo 25 (host con IP 10.254.91.1) pertenece a aquellos que proveen poca información, en este caso, por la definición de los símbolos, es muy probable que un host haga un *who has* a esta dirección IP. Con estos datos y el grafo de red de mensajes hipotetizamos que podría ser el *default gateway*, con lo cual es un *nodo distinguido* (que aporta menos información que la entropía de la fuente y es el único con

esta característica). La información que provee es 0,54 que es por debajo de la entropía de la fuente. También detectamos que el nodo 53 brinda bastante información, y creemos que eso se debe a que, si bien nunca hizo un *who has*, fue destinario de uno, lo cual tenía poca probabilidad de ocurrir. Suponemos que algo similar podría estar ocurriendo con el nodo 8.

Debido a que no tenemos más información sobre la red, no fue posible sacar otras conclusiones ni verificar que el nodo 25 sea el router, pero creemos que tenemos evidencia significativa para afirmarlo.

B.3 Red del Village Recoleta

La cantidad de paquetes ARP recolectada fue 500 (p(ARP)= 0.017002) y la cantidad de paquetes *Who Has* fueron 459. La entropía de la fuente fue 3,518831. Esto representa un porcentaje muy elevado, suponemos que se debe a la conexión y desconexión constante de dispositivos debido a que es un shopping. Esto provoca que al momento de conectarse el dispositivo pregunte su antigua IP y esto genere un nuevo paquete.

Dentro de la red el tráfico ARP fue el 1.7 %. Esto no presenta una gran cantidad de paquetes con lo que no genera un overhead muy grande en la misma.

A continuación mostramos un gráfico que muestra para cada símbolo de la fuente su información, al mismo tiempo que exhibimos el punto de corte (la entropía) por el cual consideramos que un host es distinguido.

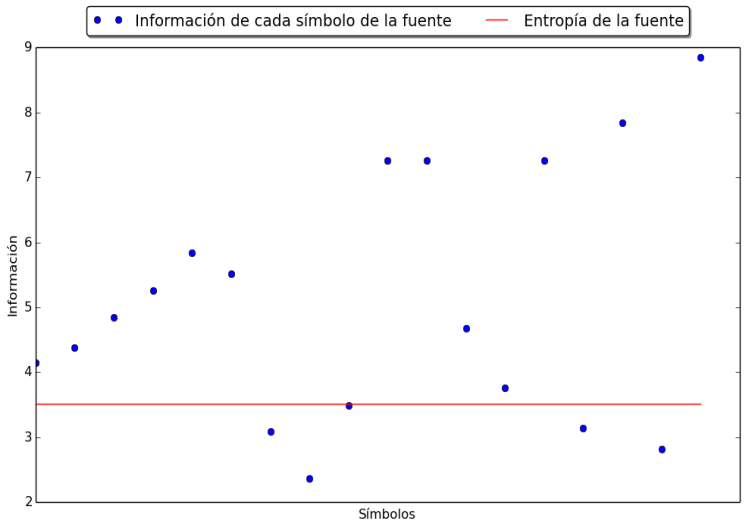

Figura 7. Información en Village

Hubieron varios nodos distinguidos los cuales creemos que son el *router* y algunos dispositivos más. Las IPs distinguidas son:

- 172.30.120.1 (el cual creemos que es el *router*)
- 172.30.120.194
- 172.30.120.128
- 172.30.120.193
- 172.30.120.196

Del resto de las IPs, al no conocer la topología de la red, no podemos concluir nada al respecto.

A continuación mostramos el grafo subyacente de la red:

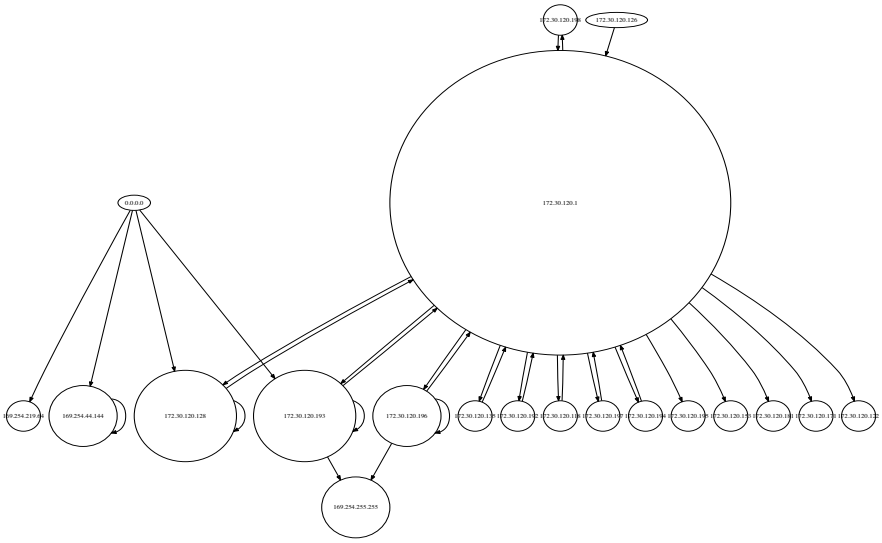

Figura 8. Grafo de conexiones Village

Efectivamente podemos observar que la dirección que proveyó menos información fue la más apuntada en el grafo. También podemos ver que hay algunos nodos que destacan en el grafico, los cuales(viendo el histograma de información) son menores a la entropía. Nuestra hipótesis es el nodo más distinguido debe ser el *default gateway* de la red.

Algo que nos llamó la atención fue la aparición de un nodo con IP 0.0.0.0. Nuestra hipótesis es que al ser una red grande, solo estamos viendo una porción de la misma ya que es probable que este subdivida por subredes y estuvieramos conectados a una de ellas. De esta manera esa sería la IP proveniente del *default gateway* de la capa superior.

Debido a la información que obtuvimos de la red podemos suponer que el nodo con menor información representa el *default gateway* de la red actual.

B.4 Red del Dot Baires Shopping

La cantidad total de paquetes transmitidos en 17 minutos fueron 623929 de los cuales 5814 fueron ARP (P(ARP) = 0.009318) de los cuales 5403 (92,93 %) fueron WHO HAS.

Utilizando estos últimos se modela la fuente S_1 cuya entropía es 3,111714.

Algo a destacar es que casi el 1 % del tráfico de la red es de paquetes ARP lo cuál genera un overhead que no tiene un gran impacto en el volumen general pero que tampoco es para despreciar.

A continuación presentaremos un diagrama de la red donde se ven los nodos que más fueron destino de los paquetes ARP WHO HAS. Por cuestiones visuales se muestran sólo los nodos que fueron solicitados más que 10 veces y el incremento de tamaño en función de la cantidad de ocurrencias es en escala logarítmica:

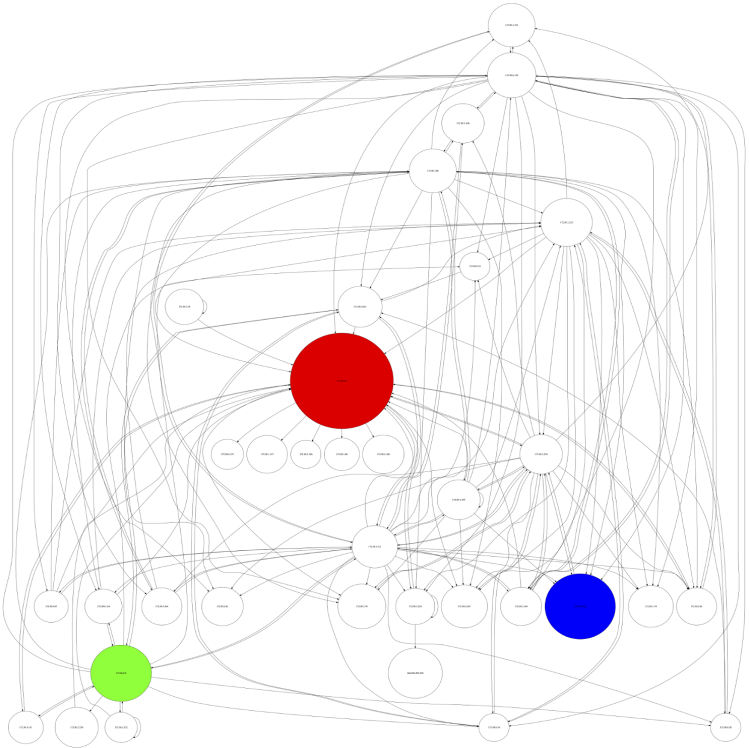

Figura 9. Grafo de conexiones DOT Baires Shopping

Visualmente el nodo 172.50.0.1 marcado en Rojo es el nodo que recibió más paquetes ARP WHO HAS. Luego le sigue el nodo 172.50.2.203, marcado en azul y en tercer lugar el nodo 172.50.0.5 marcado en Verde.

Analizando la red con el criterio teórico, previamente mencionado:

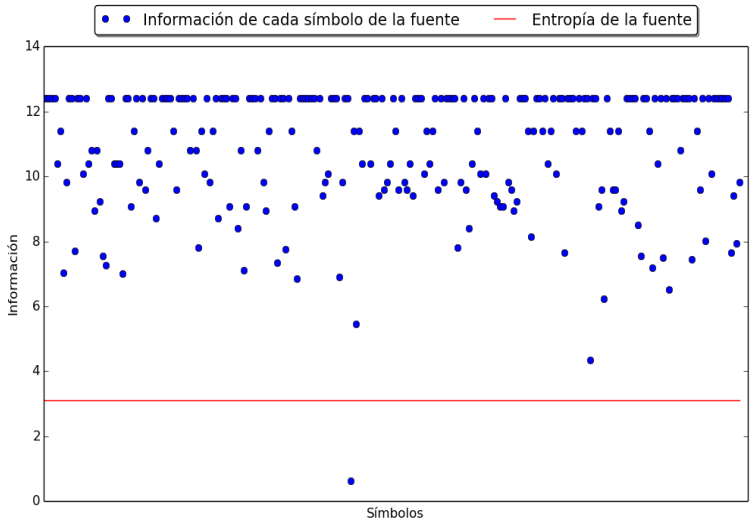

Figura 10.

Existe un único nodo destacado y es el 172.50.0.1 con información 0.61, y por su IP podemos suponer que puede ser el *Default Gateway*.

Los primeros dos nodos por arriba de la linea de corte (la entropía) son: El Azul con IP 72.50.2.203 e información 4.35 y el Verde con IP 172.50.0.5 e información 5.45. Con este enfoque se puede ver que no son distinguidos aún cuando en el grafo parecía que si. Por no conocer la infraestructura del Shopping, desafortunadamente no tenemos forma de inferir nada sobre estos nodos.

IV. Conclusión

En este trabajo presentamos por un lado, dos modelos de fuentes información con los cuales trabajamos en diversas redes WI-FI, y por otro, herramientas para analizar y extraer conclusiones con respecto a las mismas.

La fuente S presentó una buena estimación sobre el overhead de la red. Por otro lado, la fuente S1 mostró un analisis más especifico permitiendo encontrar nodos destacados, y de esta forma estimar cuál es el *default gateway*. Por otra parte vemos que esta fuente es buena pero no infalible ya que en la red hogareña pudimos apreciar cómo un comportamiento anómalo generó suficiente ruido como para evitar la detección del mismo.