Teoría de las Comunicaciones
TP1

15 de julio de 2016

Integrante	LU	Correo electrónico
Martin Baigorria	575/14	martinbaigorria@gmail.com
Federico Beuter	827/13	federicobeuter@gmail.com
Mauro Cherubini	835/13	cheru.mf@gmail.com

Reservado para la cátedra

Instancia	Docente	Nota
Primera entrega		
Segunda entrega		

Índice

1. Introducción **3**

2. Desarrollo **3**
2.1. Informacion y Entropia 3
2.2. Canal elegido 4
2.3. Paquetes de red 4

3. Metodología **4**

4. Resultados **5**
4.1. Protocolos 5
4.2. Paquetes ARP 9
4.2.1. Red Laboratorios DC y Red Plaza Oeste 9
4.2.2. Red hogareña 14
4.3. Paquetes de control ARP 16

5. Conclusiones **16**

1. Introducción

La comunicación es uno de los ejes fundamentales de la humanidad, a lo largo de la historia han aparecido diferentes medios para poder satisfacer esta necesidad, sin embargo el concepto nunca se había formalizado. En 1948, el matemático e ingeniero Claude E. Shannon propone una definición formal de que es la comunicación desde una punto de vista matemático, dando origen a la *Teoría de la Información*. En este trabajo analizaremos como aplica dicha teoría a un medio de comunicación real, particularmente uno que sea altamente utilizado y tenga una alta densidad de usuarios.

2. Desarrollo

2.1. Informacion y Entropia

Como vimos en la introduccion, en 1948 el matematico Claude E. Shannon define formalmente que es la *Informacion* en su publicación *A Mathematical Theory of Communication*, junto con esta defincion tambien introduce el concepto de *Entropia* en la comunicación. Primero definimos quienes son los participes en la comunicación:

- Fuente de Informacion
- Emisor
- Receptor
- Destino de Informacion
- Ruido
- Canal

Los mismos se encuentran conectados de la siguiente forma:

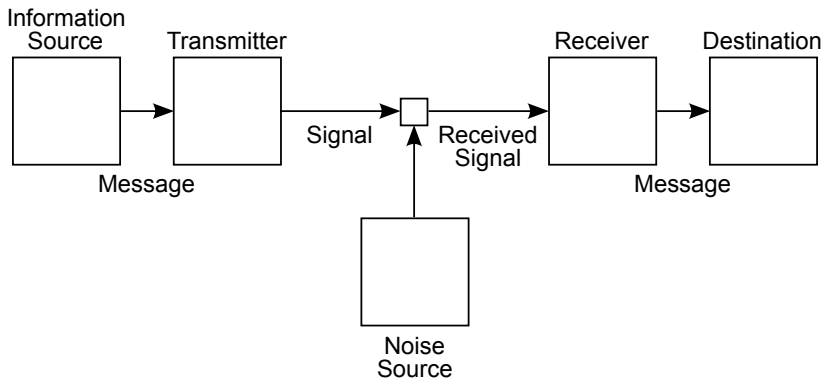

Figura 1: Diagrama de comunicación

Para poder establecer una comunicación punto a punto, es necesario que el emisor y el receptor «hablen» un lenguaje en común. Para poder satisfacer esto, se define un conjunto de símbolos $S = \{s_1, ..., s_n\}$, con una probabilidad $p(s_i)$ con $1 \leq i \leq n$ asociada a cada uno de ellos, este conjunto representa la totalidad de símbolos que pueden ser transmitidos por el canal. Una vez definido el conjunto, la información que se obtiene por símbolo es simplemente $I(s_i) = log(\frac{1}{p(s_i)})$. Se elige el logaritmo como función por cumplir con varias condiciones idóneas para calcular la informacion, de las mas interesantes tenemos que:

- Si $s \in S$, entonces $I(s) \geq 0$. Esto vale ya que la inversa de la probabilidad es siempre mayor o igual a 1
- Si $p(s_i) = 1$ para algun i, entonces $I(s_i) = 0$, ya que un evento que ocurre siempre no aporta información significativa
- $I(s_i, s_j) \leq I(s_i) + I(s_j)$, la igualdad vale unicamente si los símbolos son independientes

Como podemos apreciar, la función logaritmo cumple con todos estos puntos.

En nuestro trabajo analizaremos principalmente la entropia, esta se define como $H(S) = \sum_{i=1}^{n} p(s_i)I(s_i)$. Esta medida representa la media de información obtenida por símbolo en la comunicación, y se encuentra íntimamente relacionada con las probabilidades de cada símbolo, particularmente mientras menos aleatoria sea una red, menor es la entropia. También aplica el mismo razonamiento a la inversa, es decir, mientras mas cerca de la equiprobabilidad estén los símbolos, la entropia aumentara, maximizándose cuando los símbolos sean equiprobables.

2.2. Canal elegido

Para poder tener suficientes datos para hacer un análisis interesante, seria prudente tomar un medio que sea ampliamente usado. Por ello, elegimos utilizar una red del tipo Wi-Fi, los puntos a destacar de la red son:

- En la aplicaciones habituales, estas redes se caracterizan por tener un *nodo* central, el cual se encarga de regular el trafico en la red
- En las redes publicas, los nodos que no son el central no se suelen comunicar entre ellos, estos principalmente se conectan con servidores en Internet

2.3. Paquetes de red

En la redes la información en el canal se transmite en paquetes, estos paquetes tienen varios campos, particularmente nos interesan analizar los paquetes de capa 2 y capa 3. De los paquetes *Ethernet* de capa 2, nos interesa:

- Dirección MAC de origen
- Dirección MAC de destino
- Protocolo del payload, este depende del paquete de capa 3 que se esta transportando, puede ser IPv4, IPv6, ARP, etc.

Por otro lado, de los paquetes de capa 3 nos interesan solamente los ARP. De este tipo de paquetes nos interesan los campos:

- Tipo de operación
- Dirección MAC del emisor
- Dirección IP del emisor
- Dirección MAC del destinatario
- Dirección IP del destinatario

El fin de los paquetes ARP, es vincular la capa 2 con la capa 3, relacionando las direcciones IP con direcciones MAC fisicas. Para hacer esto el protocolo ARP cuenta con dos operaciones, estas pueden ser *who-has* o *is-at*.

Las operaciones *who-has* sirven para identificar a que dirección MAC física corresponda una dirección IP de la red, estos paquetes suelen ser de tipo *broadcast*.

En respuesta a los *who-has*, tenemos las operaciones *is-at*. Una vez que un nodo recibe un paquete de tipo *who-has*, este revisa si la dirección IP del mismo coincide con la suya, en el caso de que lo sea este enviá un paquete al nodo que genero el *who-has* para notificarle su dirección MAC física.

Para no tener que enviar paquetes ARP por cada paquete IP que se desea enviar, el emisor del *who-has* al recibir el *is-at*, guarda el resultado en una tabla para poder enviar futuros paquetes al mismo destino inmediatamente.

3. Metodología

En el trabajo se pide que analicemos dos fuentes de información, estas son S y S_1, y tienen la siguiente forma:

- S: Comprende a todos los paquetes *Ethernet* que circulan por el canal, se los diferencia por el *protocolo* del payload.
- S_1: Se limita a los paquetes *Ethernet* de tipo ARP.

Para S_1 ademas se pide establecer un criterio de diferenciación, para esto tomamos la conjunción de las direcciones IP fuente y destino de los paquetes ARP de tipo *who-has*. Elegimos estos paquetes ya que los *is-at* se dan unicamente en respuesta a algún *who-has* anterior, con lo cual estaríamos duplicando ciertos paquetes.

Las redes elegidas para *sniffear* fueron las siguientes:

- FibertelZone, en shopping Plaza Oeste Moron (60 min. de captura)
- Laboratorios-DC (30 min. de captura)
- Red hogareña (30 min. de captura)

Por ultimo, es pertinente definir que consideramos como símbolo distinguido. Siguiendo la definición de información, consideramos como símbolo distinguido aquel que provee la menor cantidad de información, si bien podríamos aplicar el mismo criterio a la inversa (es decir, considerar distinguido al nodo que mas información provea), nos interesa analizar al símbolo de menor información por las siguientes razones:

- S: En este caso los nodos que mas apariciones tengan, van a ser los protocolos que dominen el trafico de la red. Debido a que el objetivo principal de las redes publicas Wi-Fi es acceder a Internet, nuestra teoría es que los paquetes IPv4 van a ser los que mas apariciones tengan, y a su vez los que menos información provean como símbolo (si bien IPv6 sirve para acceder a Internet, no se usa en la misma medida que IPv4)
- S_1: Aquí es mas interesante estudiar los paquetes que menos información provean, ya que tenemos la teoría de que en una red Wi-Fi convencional, el nodo que corresponde al enrutador va a ser el que mas aparezca de los paquetes ARP, ya que la mayoría del trafico pasa a través de el. Con lo cual dicho nodo no solamente seria el que mas aparezca en la red, sino que ademas la información que el mismo provee como símbolo sera mucho menor que la del resto de los nodos

4. Resultados

4.1. Protocolos

Vamos a ver primero la cantidad de paquetes en las redes:

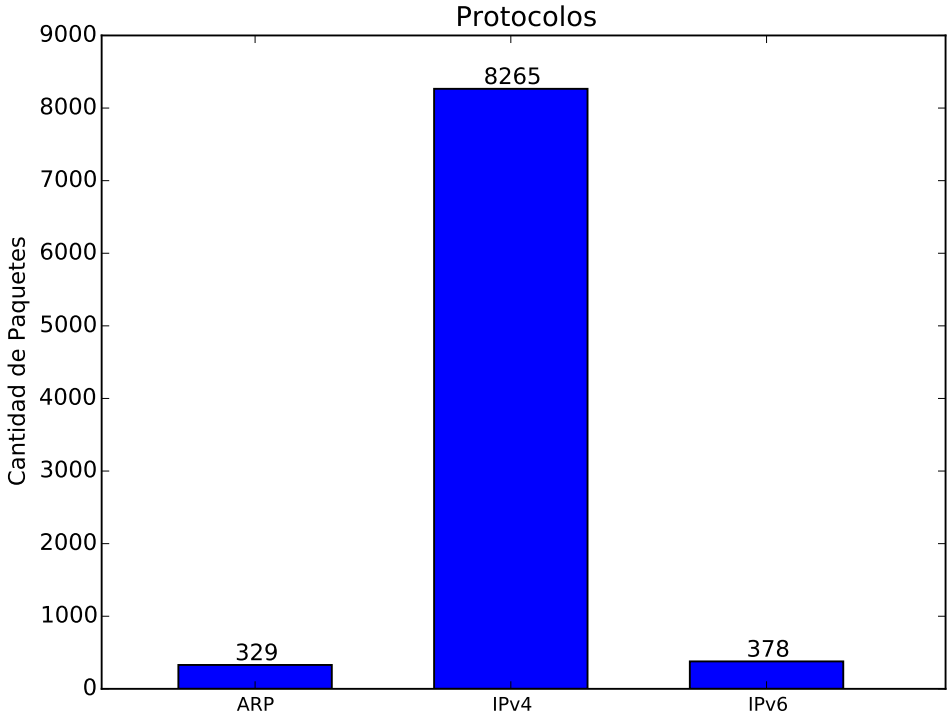

Figura 2: Red Plaza Oeste

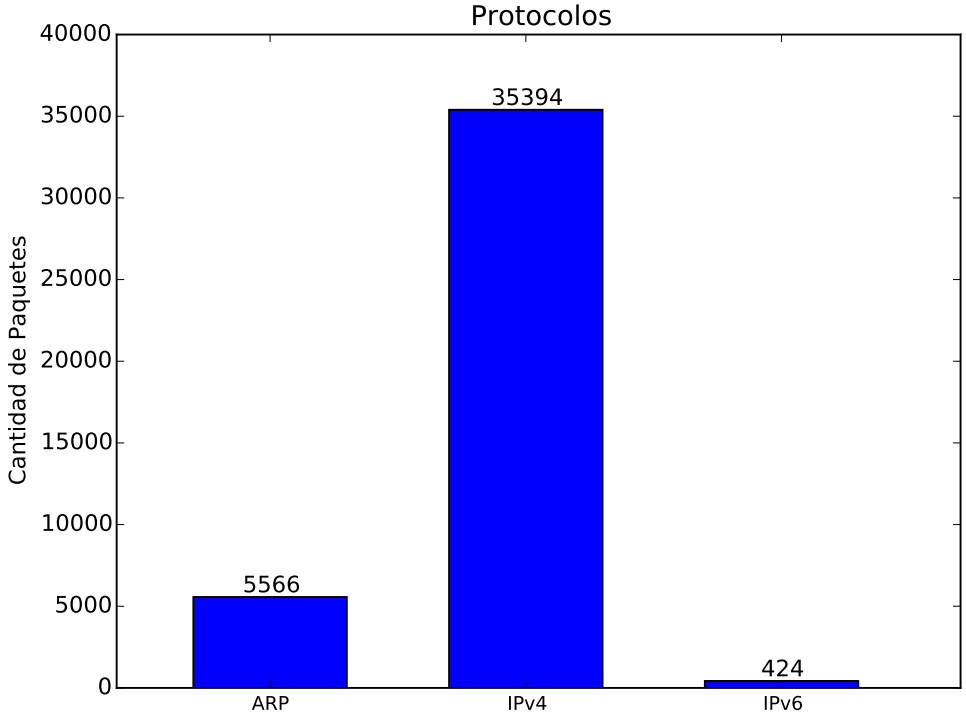

Figura 3: Red Laboratorios DC

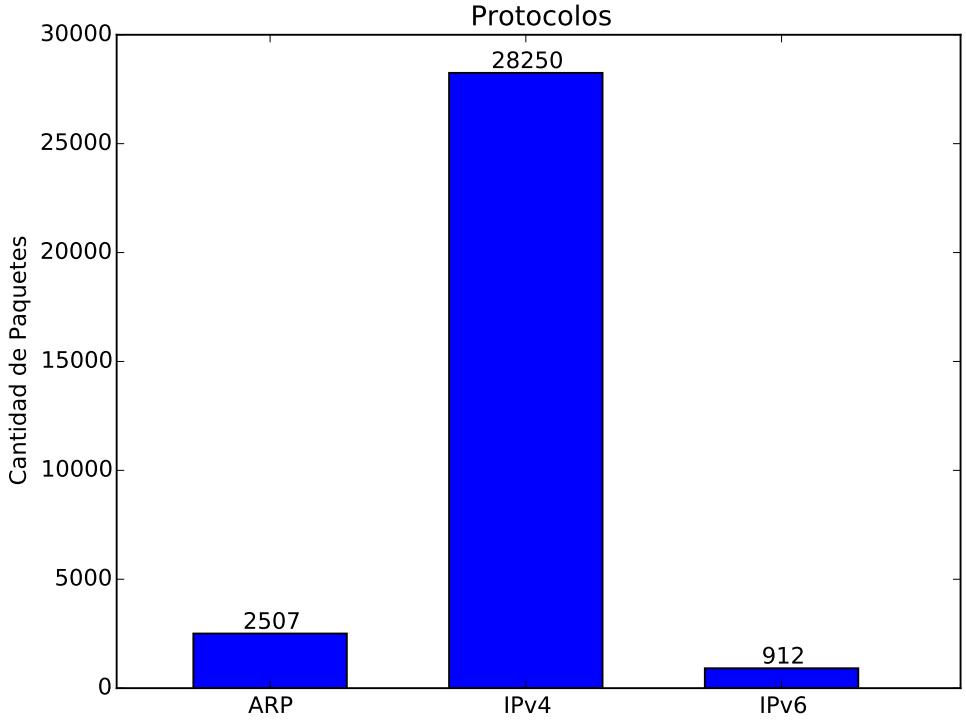

Figura 4: Red hogareña

Como podemos apreciar, los paquetes IPv4 dominan el trafico de la red. A continuación vamos a ver la información por símbolo junto con la entropia:

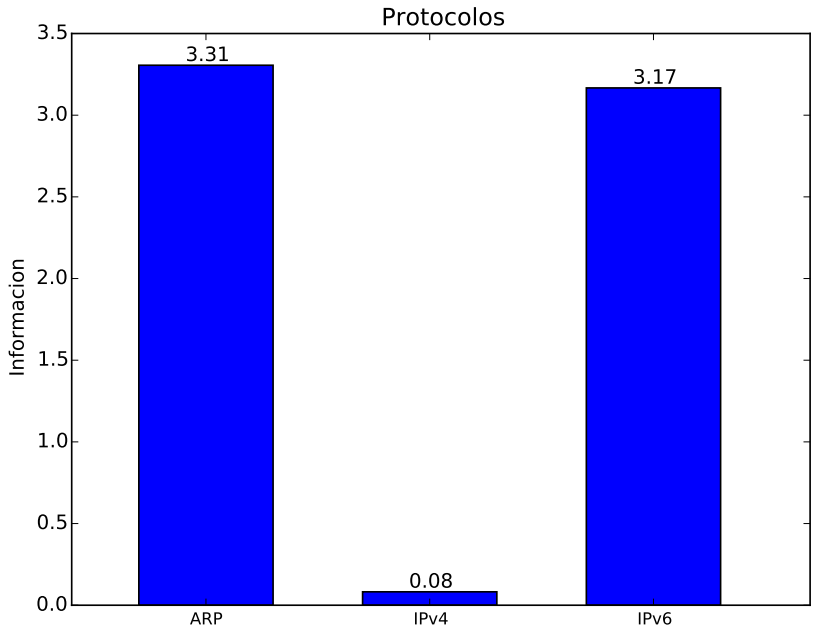

Figura 5: Red Plaza Oeste

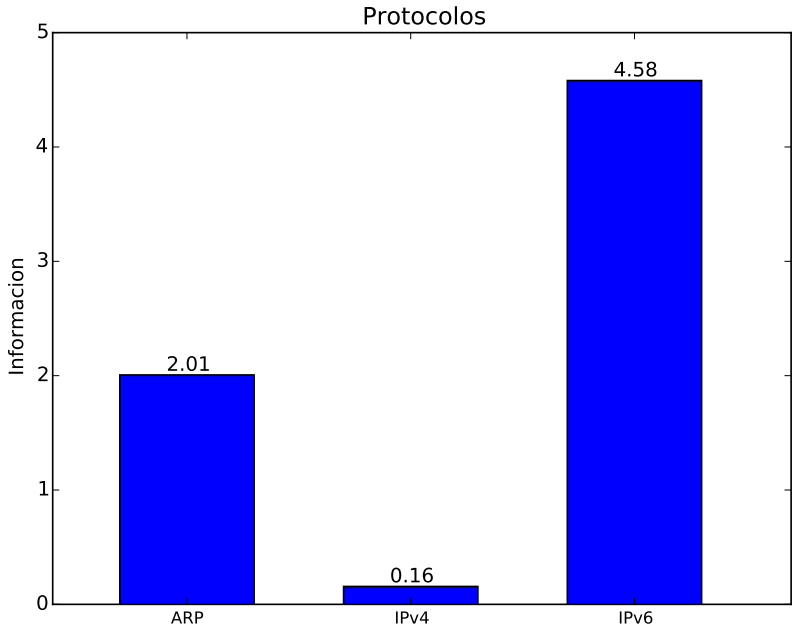

Figura 6: Red Laboratorios DC

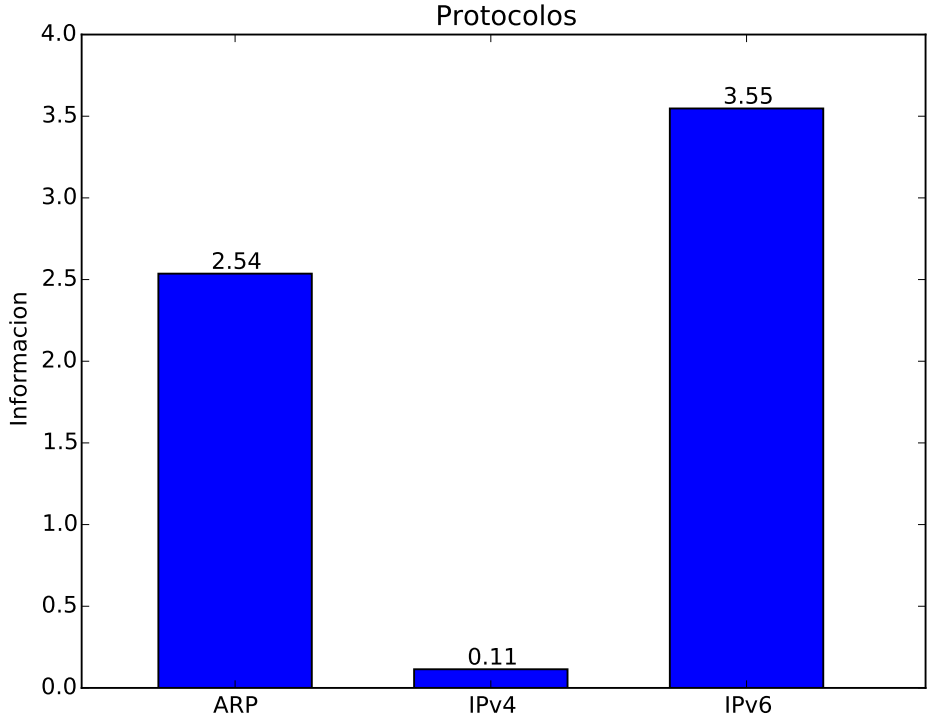

Figura 7: Red hogareña

La entropia en las redes fue:

Cuadro 1: Entropia

Red	Entropia
Plaza Oeste	0.3303
Laboratorios DC	0.4505
Hogareña	0.4048

Debido a que la cantidad de símbolos es pequeña, decidimos graficar todo el conjunto de símbolos de la fuente S.

Como podemos ver, la entropia de la red es bastante baja en ambos, esto se debe a la gran cantidad de paquetes IPv4 en ambos canales. Otro punto interesante es que en la red unicamente circulan paquetes de tipo ARP, IPv4 o IPv6, si bien esto es esperable en la red del Plaza Oeste y la red hogareña, nos sorprendió que ocurra también en la red Laboratorios DC, inicialmente creíamos que podía haber algún protocolo en la red que se use de manera casi exclusiva en algún software de índole académica. Sin embargo, el objetivo principal de las redes publicas es proveer de acceso a Internet a los usuarios de la misma, con lo cual el hecho de que solamente se encuentren este tipo de paquetes no es nada fuera de lo normal.

Nuestras predicciones indicaban que la mayoría de los paquetes iban a ser de tipo IPv4 por un margen bastante significativo, esto se corresponde con que la probabilidad de dicho paquete iba a ser mucho mas alta que el resto, con lo cual la información de dichos paquetes iba a ser considerablemente menor. Como la diferencia es tan amplia, por un margen muy significativo, siguiendo el criterio definido anteriormente, consideramos al protocolo IPv4 como símbolo distinguido.

4.2. Paquetes ARP

Como vimos en el desarrollo, definimos los símbolos de S_1 como las direcciones IP de origen y destino de los paquetes ARP. A diferencia de la fuente S, donde la cantidad de símbolos era reducida, aquí obtuvimos muchos mas resultados, por lo cual nos vimos obligados a tomar algún criterio de corte para poder graficar los datos. Particularmente elegimos graficar las direcciones que mas apariciones tuvieron, la razón de esto es que nos interesa verificar nuestra teoría expuesta en la sección anterior, para ello nos interesa la siguiente información:

- La dirección que mas apariciones tuvo
- La relación entre dicha dirección y los otros nodos que mas aparecieron en la fuente
- Poder determinar si la dirección en cuestión efectivamente se corresponde a la del enrutador

Como podemos ver, si nos limitamos a las 10 direcciones que mas apariciones tuvieron en la red, cubrimos satisfactoriamente los primeros dos puntos. Lamentablemente para cubrir el tercero, no podemos emplear una red publica, ya que no poseemos conocimiento extensivo de la misma, es por ello que tomamos la red privada para poder dar una validación básica de nuestra teoría. Primero vamos a exponer los resultados en las redes publicas, y luego los contrastaremos con los obtenidos en la red hogareña.

4.2.1. Red Laboratorios DC y Red Plaza Oeste

En este caso los resultados fueron:

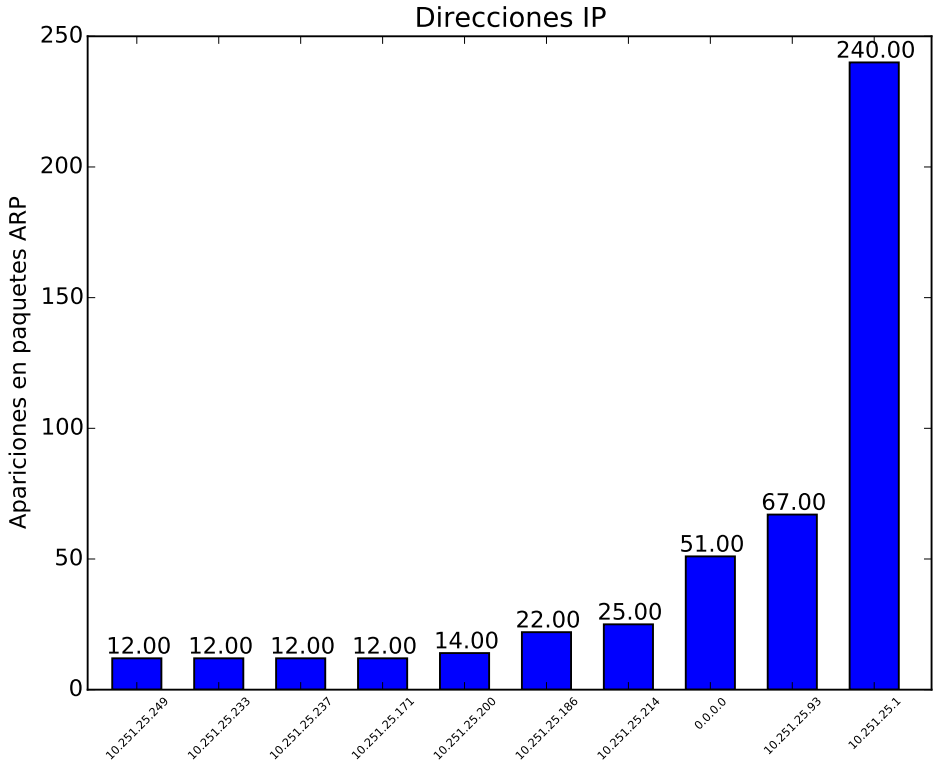

Figura 8: Red Plaza Oeste

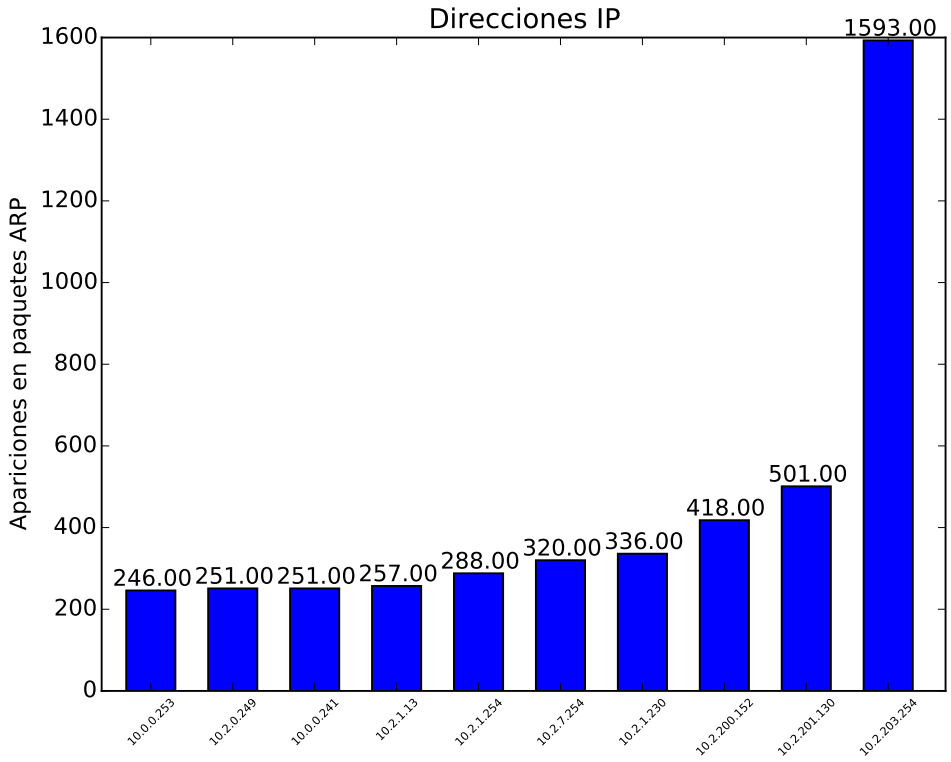

Figura 9: Red Laboratorios DC

Una de las suposiciones que teníamos, era que por el funcionamiento de la red Wi-Fi, la gran mayoría de los paquetes irían dirigidos hacia un nodo principal, el cual después se encargaría de redirigirlos al destino apropiado. Como podemos ver, esto ocurrió en ambos casos, hay un nodo el cual tiene muchas mas apariciones que el resto.

A pesar de que los resultados satisfacían nuestras expectativas, nos sorprendió que el margen de diferencia entre el nodo principal y el resto no sea mayor, con lo cual los otros nodos también están enviando paquetes ARP a nodos que no son el principal. Esto sera estudiado mas adelante, primero vamos a ver la información en ambos canales, nuevamente limitándonos a las diez direcciones que mas aparecieron:

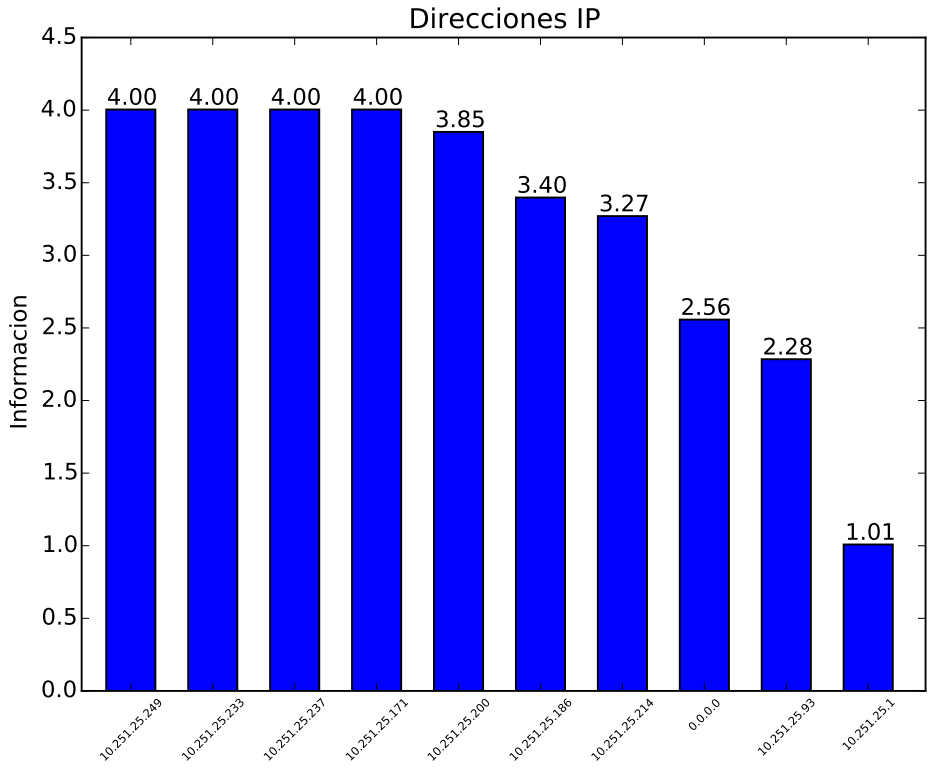

Figura 10: Red Plaza Oeste

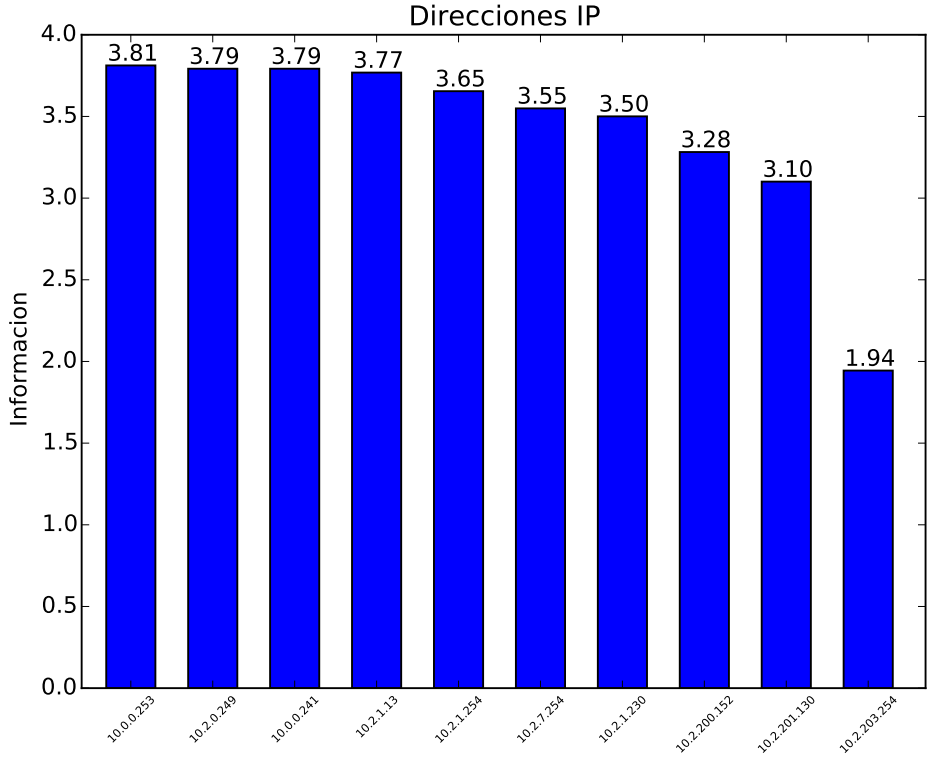

Figura 11: Red Laboratorios DC

La entropia para estas dos redes fue:

Cuadro 2: Entropia

Red	Entropia
Plaza Oeste	3.1724
Laboratorios DC	4.6459

Una cosa interesante a destacar, es que la entropia es mucho mayor tomando como símbolos las direcciones IP que los tipos de protocolos, esto ocurre en ambas redes. Esto creemos que se debe principalmente a las siguientes razones:

- La cantidad de protocolos es considerablemente menor que la cantidad de direcciones IP posibles, en la practica unicamente vimos tres protocolos en uso
- En las redes Wi-Fi, vimos que los nodos que suelen enviar paquetes de tipo *who-has* son aquellos que quieren conectarse a la red, estos suelen tener direcciones IP diferentes a las que ya se encuentran en el sistema, con lo que tenemos que agregar un nuevo símbolo
- Es posible que el tamaño de la muestra no haya sido suficiente, y que se necesiten mas tiempo de *sniffeo* para poder calcular bien la entropia

Este ultimo punto es particularmente interesante para nosotros, ya que es posible que la muestra tomada no sea significativa del trafico de la red y no sea suficiente para estimar las probabilidades, efectivamente afectando la entropia de la red. También seria interesar estudiar si las políticas de asignación de IP de la red y el uso de la misma terminan amortizando el segundo punto, ya que si el sistema reasigna direcciones siempre que puede, mientras mas grande sea la muestra la diferencia entre el nodo principal y el resto potencialmente seria mayor. Para probar esta teoría, planteamos un experimento simple que consiste en tomar parte de la captura (la mitad de la muestra original), y contrastar los resultados de dicha parte con los totales, con esto podemos estimar si el tiempo puede afectar considerablemente la entropia de la red. Los resultados fueron:

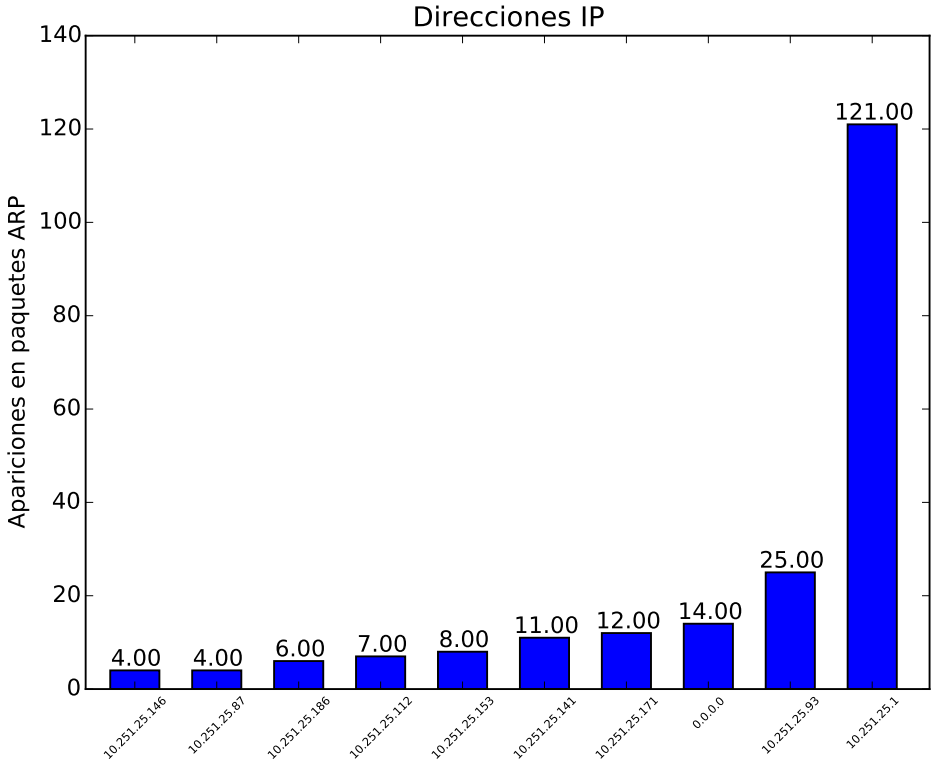

Figura 12: Red Plaza Oeste

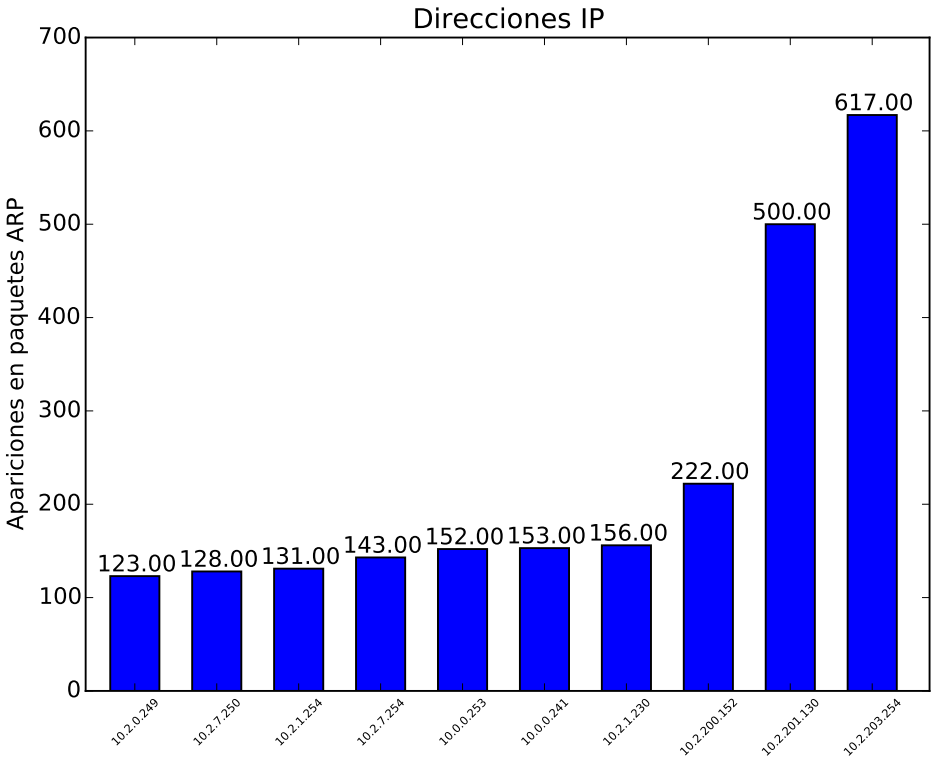

Figura 13: Red Laboratorios DC

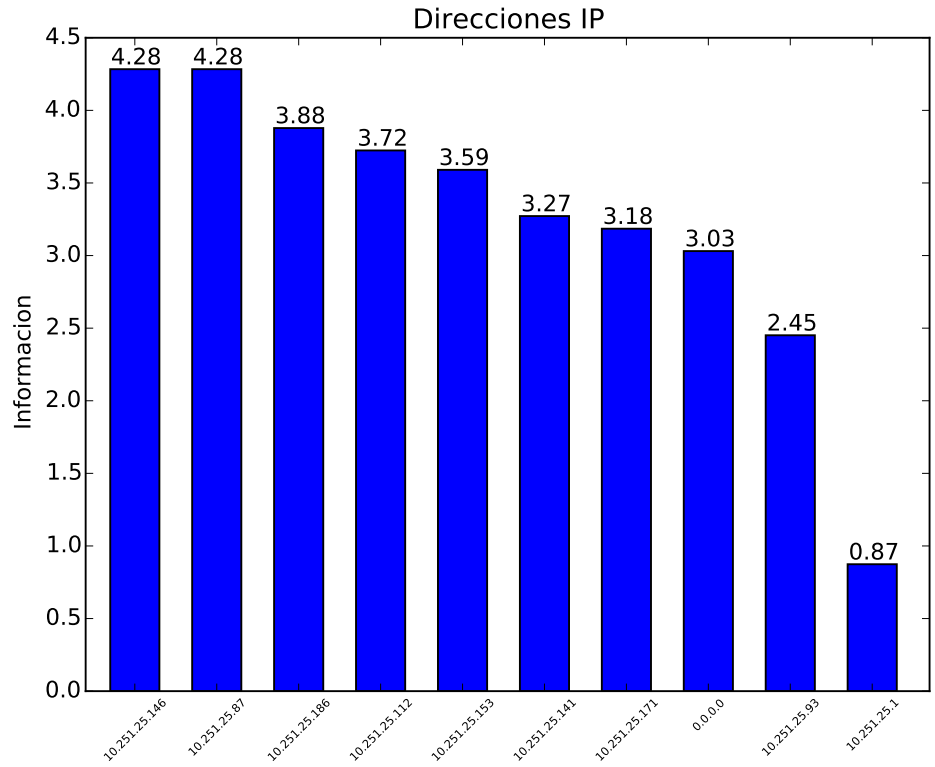

Figura 14: Red Plaza Oeste

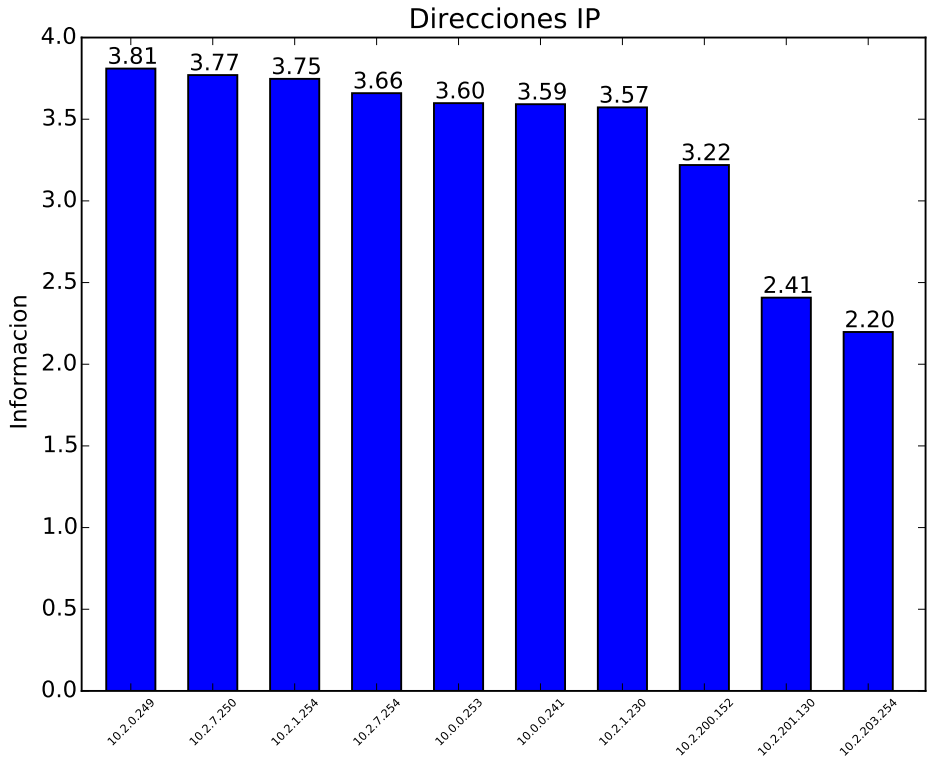

Figura 15: Red Laboratorios DC

La entropia para estas dos redes fue:

Cuadro 3: Entropia

Red	Entropia
Plaza Oeste	2.8027
Laboratorios DC	4.5887

Como podemos ver, si bien hubo diferencia entre los resultados parciales y los finales, esta diferencia no fue muy grande, esto lo podemos interpretar de dos formas. Por un lado, podríamos pensar que la captura tiene que ser mucho mas grande (del orden de días) para que las políticas de asignación de IP tengan efecto sobre la entropia, o podemos pensar que dichas políticas no van a tener peso significativo sobre los resultados finales. Particularmente nos inclinamos por la segundo opción, ya que hay muchos elementos en juego a la hora de tomar una medición, particularmente el canal puede estar mas o menos congestionado, afectando la medición, esto podemos verlo contrastando la muestra reducida de Laboratorios DC con la muestra completa, allí en la completa se ve claramente marcada la dirección `10.2.203.254` por sobre el resto, tanto en información como en apariciones, mientras que en la muestra parcial no es encuentra tan marcada.

En general, la red respondió de la manera que esperábamos, pudimos identificar un nodo principal el cual aparece en muchas mas ocasiones que el resto en ambas redes. Si bien la diferencia en cantidad de apariciones de dichos nodos respecto al resto no fue tan significativa, cumplen con el criterio de símbolo distinguido planteado anteriormente, con lo cual podemos marcarlos como símbolos distinguidos en sus respectivos canales. Por ultimo queda analizar el caso de la red hogareña y determinar si el símbolo distinguido de la misma corresponde con la dirección del enrutador.

4.2.2. Red hogareña

Como era de esperar, los resultados en la red hogareña fueron considerablemente mas reducidos que en las redes publicas, con lo cual optamos por graficar la totalidad de las direcciones. Los resultados fueron las siguientes:

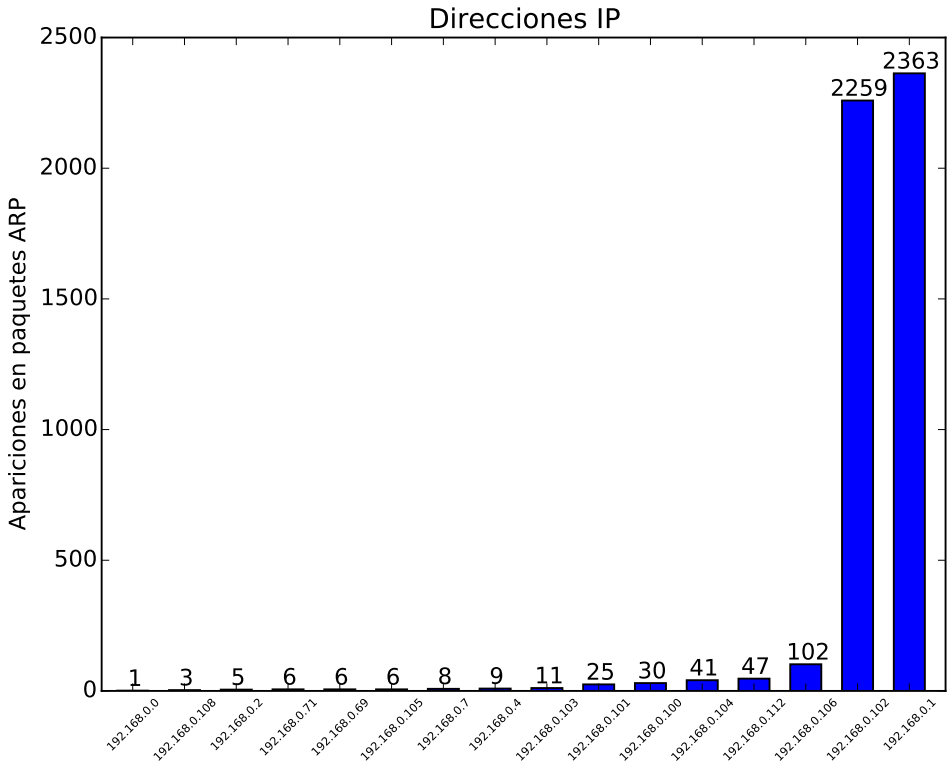

Figura 16: Red hogareña

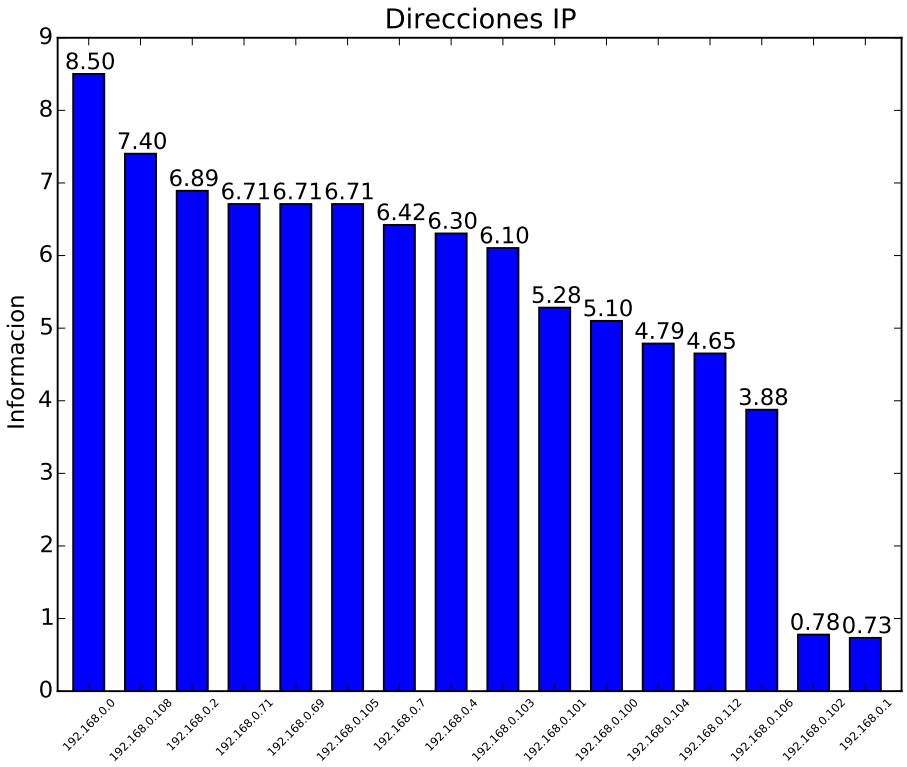

Figura 17: Red hogareña

La entropia fue la siguiente:

Cuadro 4: Entropia

Red	Entropia
Red hogareña	1.0057

Debido a que conocemos como esta formada esta red, sabemos que la dirección IP del enrutador es `192.168.0.1`. Si bien la direccion IP `192.168.0.102` tambien aparecio repetidas veces, esto se debe a que la computadora desde donde se tomo la muestra, ya que era uno de los pocos dispositivos conectados a la red activamente utilizando la misma, con lo cual su presencia fue mayor en la medición. Como podemos apreciar, utilizando el criterio expuesto anteriormente, vemos que el símbolo distinguido en la red es efectivamente la dirección IP `192.168.0.1`, con lo cual podemos dar una verificación básica de nuestra teoría, ya que el símbolo se corresponde con el enrutador.

4.3. Paquetes de control ARP

En la sección anterior hablamos de paquetes que no eran enviados hacia el nodo principal, ademas de esto, nos dimos cuenta que varios de ellos tenían una forma bastante particular, tras consultar diferentes recursos nos dimos cuenta que varios de ellos eran paquetes de control. Nos topamos con los siguientes:

- Dirección IP `0.0.0.0`: Estos paquetes se utilizan para revisar si una dirección IP se encuentra en uso por algún host, la idea es que un nodo al recibir una dirección IP para usar, envia este paquete, y si recibe un *is-at* de otro nodo quiere decir que la dirección que pretendía utilizar esta en uso
- Misma dirección IP de fuente y destino: Este es un paquete bastante particular, sirve para que los diferentes hosts en la red tengan sus tablas de dirección IP y MAC actualizadas. La idea es que al recibir el paquete y verificar que las direcciones origen y destino son iguales, el host revisa si tiene la dirección MAC en su tabla, en caso de tenerla actualiza la dirección IP almacenada si es que esta cambio por la dirección que figura en el paquete. Mantener las relaciones IP y MAC actualizadas es sumamente importante en la red, ya que eso puede ahorrarnos una cantidad significativa de tiempo a la hora de manejar el trafico.

Las apariciones de estos paquetes consideramos que terminaron quitandole bastante peso al nodo central respecto a los demás, haciendo que la entropia sea mayor.

5. Conclusiones

Es bastante interesante ver como los diferentes conceptos de teoría de la información aplican en canales reales, particularmente considerando que fueron planteados en 1948 cuando las redes modernas no existian. Desde el punto de vista de las redes Wi-Fi puntualmente, vimos como la comunicación es sumamente centralizada y dominada por IPv4, lo primero se puede ver claramente en los paquetes ARP los cuales en nuestras muestras estan en gran parte dirigidos a un único nodo, mientras que lo segundo no solamente se ve a simple vista, sino que ademas es esperable considerando que la Internet funciona sobre IPv4.

Como estudio a futuro, seria interesante hacer un *sniffeo* durante el lapso de uno o mas días, para poder ver si las proporciones obtenidas con la fuente S_1 se mantienen, o si eventualmente el nodo principal se impone respecto al resto aumentando aun mas la diferencia entre estos. También consideramos importante analizar las diferentes técnicas de ARP Spoofing, y analizar cuan factible es realizar dichos ataques y el alcance de los mismos, particularmente hacer spoofing del gateway, en redes Wi-Fi esto es sumamente interesante ya que al ser centralizada podríamos redirigir el trafico y administrar el acceso a la red.