DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 1

Taller de Wiretapping

27 de abril de 2023 — Teoría de las Comunicaciones

Integrante	LU	Correo electrónico
Alexandra Abbate	710/19	alexandra-abbate@hotmail.com
Lucas Mas Roca	122/20	lmasroca@gmail.com
Nicolás Valentini	86/21	nicolasvalentini@hotmail.com
Alan Roy Yacar	174/21	alanroyyacar@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - Pabellón I
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Argentina
Tel/Fax: (54 11) 4576-3359
http://exactas.uba.ar

1. Introducción

Este trabajo se basa en el análisis de la capa de enlace de distintas redes. Para ello se va a capturar el tráfico de distintas redes para que luego sea procesado según la teoría de la información. Para llevar a cabo la captura de paquetes, se utilizará la herramienta Wireshark, en la que se conectará una computadora a una red y se pondrá en modo promiscuo para capturar paquetes y crear un archivo de registro (dump).

Una vez que se tenga la información, se procederá a analizarla con detalle en la capa de enlace, poniendo especial énfasis en el protocolo de Ethernet. En este sentido, se diferenciarán los paquetes en función del protocolo de su capa superior y del tipo de dirección de destino, es decir, unicast y broadcast. Por último, se calculará la entropía de cada una de las redes analizadas y se proporcionará una comparación entre la cantidad de información obtenida de cada símbolo y la entropía de la red.

2. Métodos y condiciones de los experimentos

El código utilizado para el análisis es una modificación del código provisto por la cátedra, utilizando el paquete Scapy en Python para poder leer con mayor facilidad los contenidos de los archivos dump de paquetes de Wireshark. Adicionalmente, se implementaron funciones para calcular la entropía de una fuente y ver la información de un evento en una fuente. Finalmente, se trabajó con librerías de Python para realizar visualizaciones gráficas de los datos recolectados.

En cuanto a las redes sobre las cuales se realizaron las capturas, se decidió tomar una red pública y tres redes privadas. Se realizó una captura en la sala de lectura del pabellón 0+infinito, una captura en un café y las dos capturas restantes en redes domiciliarias. Tres de las capturas fueron realizadas de forma inalámbrica usando Wi-Fi, mientras que la captura restante se realizó usando Ethernet.

Para la red pública se tomaron múltiples muestras y finalmente se seleccionó la muestra que mejor representa el comportamiento típico de la red. Cabe destacar que en ninguna de las muestras se notaron resultados significativamente diferentes. La muestra seleccionada fue la de la sala de lectura del pabellón 0+infinito, tomada a las 18 horas del día martes 18 de abril. Esta red Wi-Fi funciona principalmente con el protocolo IPv4, y como dato destacable, la mayoría de los paquetes broadcast IPv4 son del protocolo DHCP. Dentro de las redes privadas, se tomaron múltiples muestras, pero finalmente se seleccionaron tres. La primera representa la gran mayoría de las veces que se analizó una red privada. En esta muestra se analizó la red Wi-Fi privada de un café el día sábado 22 de abril a las 16:45, también utilizando el protocolo IPv4. La siguiente red analizada fue una red mesh, utilizando tecnología Wi-Fi, el día domingo 23 de abril a las 14 horas. Vale la pena destacar que en este domicilio hay más de una red. Finalmente, se seleccionó una red en particular que mostraba resultados un tanto diferentes. Esta corresponde a una red Ethernet privada de Iplan que utiliza principalmente el protocolo IPv6. Cabe destacar que en el mismo domicilio, hay otra red de Fibertel. Esta muestra fue realizada el día viernes 21 de abril a las 21 horas. Es importante destacar que para tener una buena muestra del comportamiento de cada red, se tomaron alrededor de 10.000 paquetes en cada una.

Para encontrar nodos distinguidos en una red utilizando las direcciones IP de los paquetes de protocolo ARP, se decidió primero quedarse solamente con los paquetes del tipo request y proponer como símbolo del modelo $S2$ las direcciones IP fuente. El motivo de esto se debe a que durante la experimentación preliminar se notó que los paquetes de reply no proveen información relevante para el problema a analizar. En el anexo opcional se desarrollará el motivo por el cual se seleccionó como símbolo la IP fuente por encima de la IP destino.

3. Resultados

Inicialmente, se compararon dumps de las mismas redes en distintos días y distintos momentos, los que fueron nombrados en la sección 2 son los que fueron tomados para el análisis, ya que las mediciones que fueron realizadas para las mismas redes convergen a los mismos resultados.

3.1. Broadcast vs. Unicast

En todas las redes muestreadas, prevalecía el uso de comunicaciones del tipo unicast sobre el tipo broadcast. En el siguiente gráfico se evidencia que independientemente de la naturaleza de la red, es siempre mucho mayor el tráfico de tipo unicast.

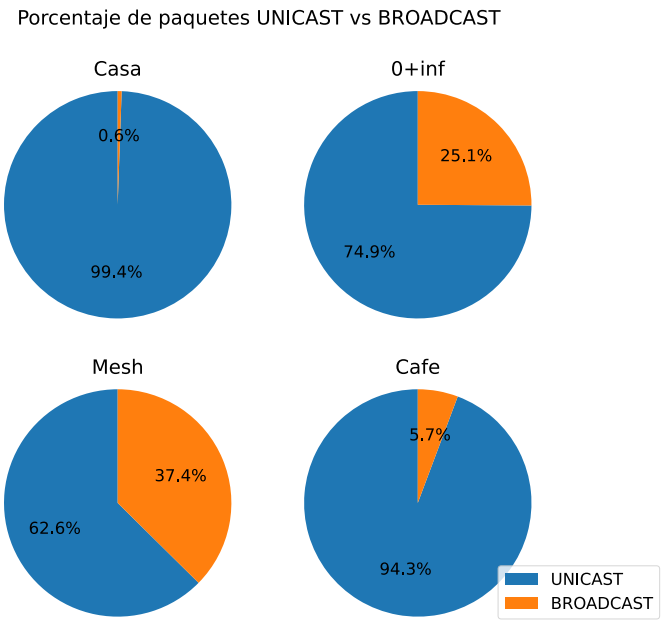

Figura 1

A pesar de que el tráfico unicast es mayor que el broadcast en todas las redes, se pueden observar algunas particularidades, especialmente en la red mesh y 0+infinito, donde el porcentaje de comunicaciones broadcast es significativamente más alto que en las otras redes. En el caso de la red mesh, esto puede deberse a que intenta conocer a todos los dispositivos todo el tiempo por razones de optimización, utilizando ampliamente el protocolo ARP para este fin. En el caso de la red 0+infinito, al ser una red pública con una gran cantidad de usuarios, se observó una alta cantidad de paquetes IP broadcast, cuyo protocolo es DHCP, lo cual es esperable para una red con una gran cantidad de usuarios. Un último fenómeno destacable es que la red que utiliza tecnología Ethernet presenta un mayor porcentaje de tráfico unicast en comparación con las otras redes.

3.2. Protocolos

Durante los experimentos se registraron múltiples protocolos de capa de redes, entre ellos se destacan IPv4, IPv6 y ARP. A continuación se proporcionan los porcentajes de apariciones de cada protocolo en cada red.

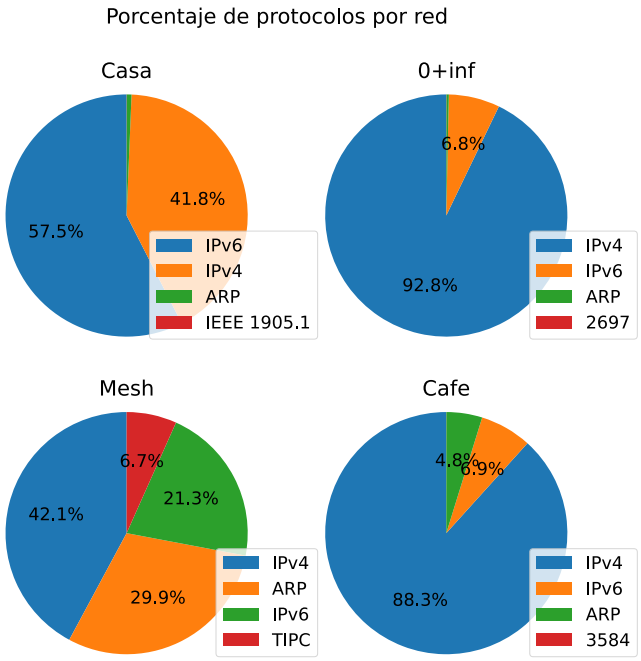

Figura 2

A partir de estos resultados, se puede observar que la gran mayoría de los paquetes enviados son a través del protocolo IP. Por un lado, todas las redes analizadas utilizaron el protocolo IP; sin embargo, una era particular de IPv6. Algo interesante para destacar es que, incluso dentro de esta red, un alto porcentaje de paquetes utilizan el protocolo IPv4. Esto puede ser un indicador de que el protocolo dominante en internet hoy en día a nivel de usuario es el protocolo IPv4. Adicionalmente, se puede notar que si bien en la mayoría de las redes el porcentaje de paquetes ARP es muy bajo, en la red mesh esos valores aumentan significativamente. Esto se debe a que el mesh constantemente pregunta por el estado de los dispositivos en su rango. Una observación importante de estos resultados es que en todas las redes analizadas hay una gran dominancia porcentual de paquetes de usuarios en comparación con paquetes de control.

3.3. Entropía e información de cada símbolo en una red

Viendo la entropía de las redes muestreadas, como se puede ver en el siguiente gráfico de barras, la red de mayor entropía fue mesh y la de menor fue la red del café.

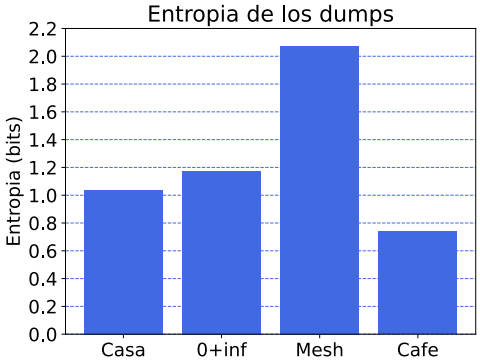

Figura 3

Tratando de explicar las diferencias de entropía entre esas redes, se pueden analizar las características de estas, en este caso se vio la relación porcentual de la distribución de paquetes entre la red con mayor y menor entropía.

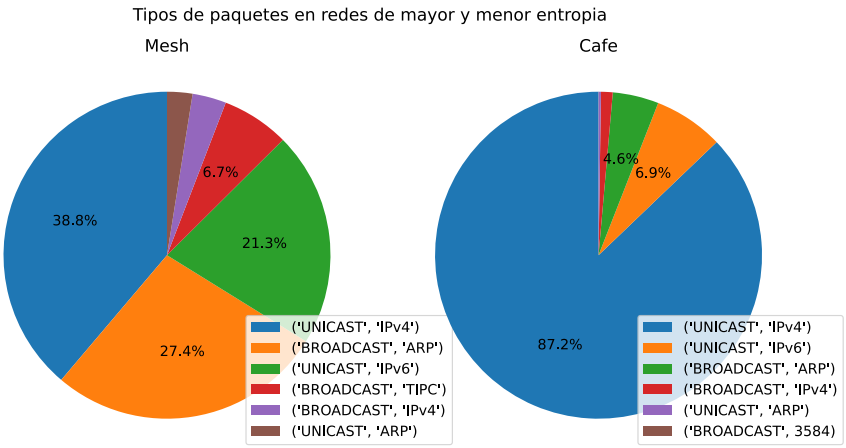

Figura 4

Observando los resultados obtenidos, se puede apreciar que la prominencia porcentual de ciertos símbolos afecta los valores de entropía. La razón de este fenómeno es que la entropía se maximiza cuando la aparición de los símbolos en una fuente de información es más equiprobable. Esto se puede explicar al observar los gráficos obtenidos y concluir que en la red del café, que es una red principalmente destinada a usuarios, la gran mayoría de los paquetes fueron del tipo IPv4, lo que redujo significativamente la entropía. Comparativamente, aunque la red mesh también está destinada a usuarios, el volumen de los mismos es menor y los mecanismos de control son mayores debido a la tecnología de la red.

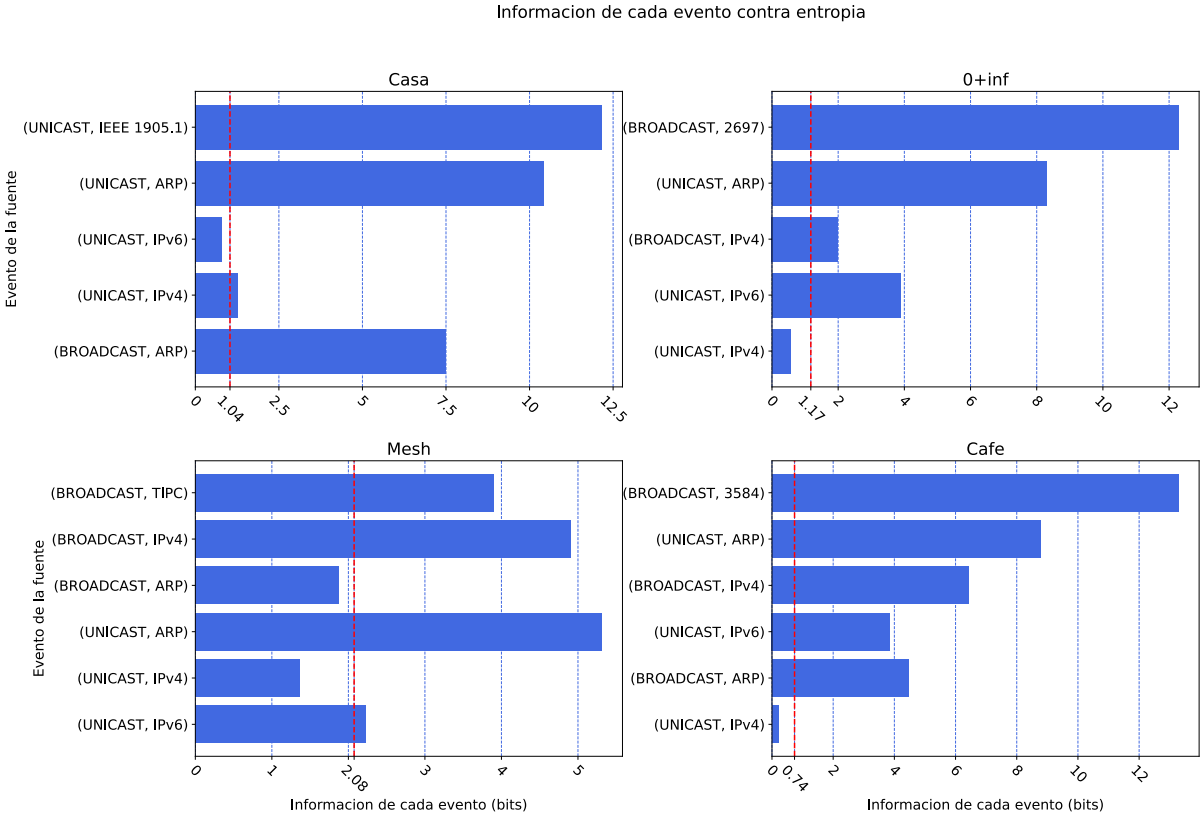

Figura 5

Algo que se puede observar en el gráfico obtenido es que, en las redes donde su entropía fue mayor, la diferencia entre el símbolo que proporciona menos información y el que más información provee se minimiza. Esto se debe, a que en fuentes de información con mayor entropía, tienen una distribución más variada de las apariciones de los símbolos, por ende la información que provee cada uno es en promedio mayor. Otro

fenómeno observado es que el símbolo que proporcionó más información estaba ubicado dentro de la red de menor entropía. Esto tiene sentido, ya que significa que tuvo muy poca aparición, y por ende tuvo poco impacto llegado el momento del cálculo de la entropía de la fuente. Siguiendo una observación análoga, el símbolo que menos información provee entre todas las fuentes, es el símbolo que está por debajo de la entropía del café.

4. Conclusiones

Como primera conclusión, parece ser que mientras mayor sea el overhead de la red (es decir, una mayor cantidad de paquetes de control para mantener una comunicación), mayor será la entropía. Esto puede ocurrir por dos motivos: una alta cantidad de usuarios y dispositivos o una alta cantidad de tecnologías que realicen mecanismos de control. Por ejemplo, en una red con muchos dispositivos, aunque haya más dispositivos conectados enviando y recibiendo paquetes de control, no necesariamente estos dispositivos se encuentran haciendo uso activo de la red todo el tiempo, por lo cual no necesariamente se observará un incremento proporcional (a la cantidad de dispositivos) en la aparición de símbolos que pertenezcan a paquetes de usuarios. En cambio, sí podemos esperar un incremento proporcional en la aparición de símbolos que pertenezcan a paquetes de control, como los de protocolo ARP o IP broadcast (DHCP por ejemplo).

En la mayoría de las redes, el tráfico broadcast no resultó una cantidad muy significativa de las redes muestreadas. Sin embargo, en el caso de la red mesh o la red 0+infinito, se puede observar más tráfico broadcast que en las otras redes. Esto probablemente se debe al aumento de paquetes de control. Por el contrario, la red que registró el menor overhead fue la red Ethernet, que con muy pocos paquetes de control, pudo realizar muchas transacciones de datos.

Protocolos esperados:

- IPv4 e IPv6: transporta mayoritariamente datos de usuario (en el caso de unicast), la mayor parte del tráfico observado en todas las redes sobre las cuales realizamos experimentos usa este protocolo. Si se usa protocolo IP broadcast en general son paquetes de control como DHCP. El resto de los protocolos son mayoritariamente de control.
- ARP: protocolo de control utilizado para saber qué dirección MAC corresponde a qué dirección IP en una red.

Protocolos no esperados:

- EAP over LAN
- Transparent Inter Process Communication (TIPC)
- Link Layer Discovery Protocol (LLDP)
- IEEE 1905.1
- 3584
- 2697

Vale la pena destacar que ninguna de las redes sobre las cuales se tomaron muestras alcanzo la entropía máxima teórica, esto implicaría que la probabilidad cada evento es la misma (los símbolos son equiprobables). En general, basta con comparar la cantidad de paquetes ARP unicast contra la cantidad de paquetes IP unicast en una red para darse cuenta la diferencia de probabilidad de esos símbolos.

Por último, se considera que las muestras recolectadas son representativas del comportamiento general de cada red. Es esperable ver que en una red domiciliaria predomine el tráfico IP unicast, ya que estos son paquetes enviados y recibidos por usuarios, mientras que en una red pública como la de 0+inf haya más necesidad de paquetes de control por la cantidad de usuarios. El caso más particular es el del mesh, ya que se encontró una gran cantidad de paquetes de control, es posible que en algunos momentos con tráfico pesado sobre esta red se vea una mayor prevalencia de paquetes de usuarios, pero teniendo en cuenta que las redes domésticas no suelen tener siempre tráfico pesado se considera que ese comportamiento representativo del comportamiento general de la red. Finalmente, algo que quedó marcadamente registrado en la red del café, es que la gran mayoría del tráfico de usuarios en internet es en IPv4, ya que habiendo una muy alta cantidad de dispositivos utilizándola, igualmente, la mayoría absoluta perteneció a IPv4, y un porcentaje mucho menor a IPv6. Esto queda todavía más en evidencia, observando la red doméstica con IPv6, la cual a pesar de que la mayoría de sus paquetes serían en IPv6, casi la mitad eran de IPv4.

5. Anexo: Nodos Distinguidos observando paquetes ARP

5.1. Elección de símbolo

Para empezar a analizar este problema, es necesario poder tener un buen modelo de fuente de información. Debido a lo mencionado en la sección 2, fue determinado que lo más interesante para empezar a analizar es si dentro de las request, se puede utilizar el símbolo de la IP de destino, o la de fuente. Para ello, se experimentó con ambas y se obtuvieron los siguientes resultados.

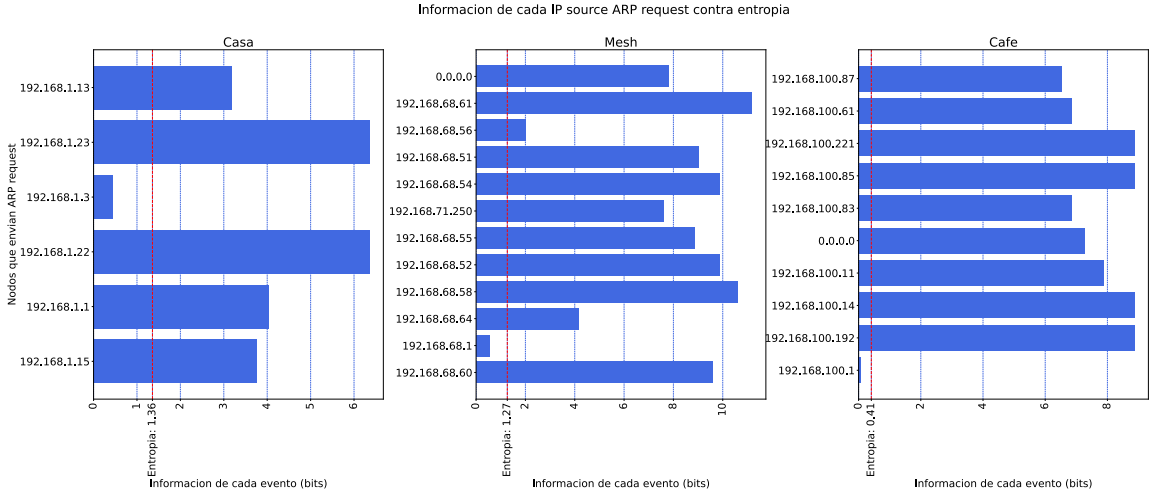

Figura 6

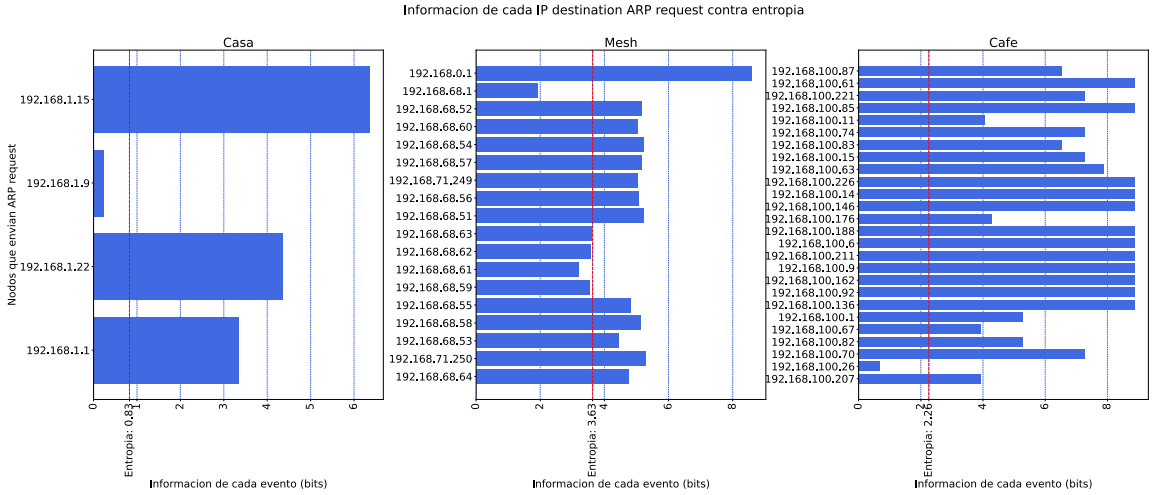

Figura 7

Un comentario necesario para entender estos gráficos, es que no se pudo utilizar la muestra obtenida de la red del 0+infinito debido a que no proporciona información para este problema. Esto se puede explicar debido a que los mecanismos de control que posee, hace que un usuario, tenga comunicación directa con su nodo (access point) a nivel control, y por ende no se pueda observar el comportamiento general de la red, a pesar de estar en modo promiscuo.

Como se puede ver en los resultados obtenidos, tomar como símbolo la IP de destino puede llegar a ser sumamente problemático, debido a que puede tener como resultado múltiples nodos como símbolos por debajo de la entropía. Por otro lado, la información se distribuye más y, por ende, también se vuelve más complicado determinar un nodo distinguido.

Por el contrario, los datos representados en la Figura 6 son más reveladores del comportamiento de la red. Por un lado, queda marcado un único nodo por debajo de la entropía y, por el otro, por lo general, la entropía es menor en comparación con tomar la IP de destino. Esto último no se cumple en el caso de la casa, ya que esta red tuvo un comportamiento no esperado. La red constantemente hacía un request al nodo distinguido en la figura 7, contaminando así el valor de la entropía obtenida. Una posible explicación posible

de este fenómeno es que el dispositivo 192.168.1.9 constantemente le era necesario ser actualizado dentro de la red. Por fuera de los resultados estadísticos, este símbolo fue elegido en base a un motivo conceptual, que es que si se quiere detectar un nodo distinguido en una red, una de sus principales funciones, es conocer quienes son los dispositivos conectados al mismo (si pensamos en el gateway de una red, este debe poder enviar los paquetes que vienen de fuera de la red al dispositivo destino dentro de la red, para esto debe conocer su dirección MAC para enviar el paquete por capa 2). Debido a esto, el nodo distinguido, se lo podría ubicar como el nodo que más veces genera la request who has broadcast dentro de su red, esto quiere decir que si se lo quisiera identificar, se puede utilizar el símbolo IP source de las ARP request.

5.2. Observaciones de lo experimentado

Un primer dato a destacar, es que con la fuente utilizada, en todos los casos, queda un nodo por debajo de la entropía. Este patrón puede ser útil para distinguir nodos, ya que significa que son el símbolo prevalente en la fuente tomada. La explicación que le damos a esto, es que estos nodos tienen una funcionalidad diferente a los otros, ya que constantemente están averiguando por las IPs de los dispositivos en su rango. Este comportamiento es esperable para los nodos que son considerados gateway. Este comportamiento en la fuente elegida claramente afecta a la entropía de la misma. La idea, en este caso, es que mientras menor sea la entropía, más fácil es reconocer al nodo distinguido, ya que es el encargado, de constantemente generar las ARP request broadcast. Por ende, en este caso, se sugiere que mientras más equiprobable sea la fuente, menos posibilidades hay de encontrar el nodo distinguido.

Se pueden encontrar nodos cuyo comportamiento es inesperado, por ejemplo en la figura 7 se puede ver que la dirección 192.168.1.9 de la red casa recibe muchas requests ARP, analizando este dump se encontró que no se recibía ninguna reply de ese dispositivo a pesar de recibir muchas requests. El motivo de esto es que la conexión es Ethernet y como los replies ARP son unicast, si el dispositivo que emitió el request se encuentra más cercano (en términos de dispositivos intermedios de la red) al dispositivo buscado, el reply no será transmitido al dispositivo que está realizando la muestra.

Se encontraron paquetes ARP de tres tipos, dos de ellos eran paquetes esperados y conocidos (reply y request), sin embargo, se encontraron también paquetes del tipo ARP announcement. Estos paquetes son muy similares a los ARP reply, con la diferencia de que los reply son pedidos con los requests mientras que los announcements se envían de forma esporádica (sin un request previo). Además, los paquetes reply son unicast mientras que los announcements son broadcast.

Por último, en las redes del mesh y del café, hubo una relación directa entre el nodo que quedo como distinguido según los muestreados, y el nodo gateway. Algo para mencionar antes de sacar más conclusiones, es que en todas las redes mirando la máscara de red y la dirección IP del host con el que se toma una muestra se puede obtener la primera dirección válida posible de una red, usualmente en la mayoría de las redes esta dirección coincide con el default gateway. En ambos casos el nodo, cuya dirección IP coincidía con la primera dirección IP posible de la red, quedó como destacado, y por ende se puede decir que la fuente tomada fue eficiente para encontrarlo. Con el fin de comprobar lo encontrado, se realizó un IPconfig en ambas redes, y quedó constatado que el nodo gateway en ambas, era quien la correspondía la primera dirección IP de la red. Un caso distinto fue el de la casa, que como se puede observar en la figura 6, quien quedó como nodo distinguido fue el 192.168.1.3. Esta IP corresponde a un repetidor dentro de la misma red, y lo más probable es que este haya trabajado en conjunto con el nodo gateway (192.168.1.1) para obtener información de las MAC y sus IPs correspondientes en su rango. Esto igualmente no quita que a pesar de que no se obtuvo el nodo gateway, el nodo distinguido encontrado igualmente es un nodo central para la red.