DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

TP1: Wiretapping

15 de julio de 2016 — Teoría de las Comunicaciones

Integrante	LU	Correo electrónico
Benitti, Raúl	592/08	raulbenitti@gmail.com
Castro, Damián	326/11	ltdicai@gmail.com
Lizana, Helen	118/08	hsle.22@gmail.com
Grenier, Michelle	418/10	michelle.grenier@hotmail.com

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

Índice

1. Introducción

Las redes de computadoras han dejado de ser una tecnología reservada a ciertos ámbitos científicos y militares para convertirse en piezas fundamentales en el desarrollo de casi cualquier actividad, a tal grado que las relaciones humanas, desde el comercio hasta las guerras, han sido profundamente transformadas por la conectividad alcanzada en los últimos años. Es por esto que analizar los distintos aspectos de una red puede proveer información útil para comprender el uso que se le está dando a la red, información que sirve tanto para modificar la infraestructura y los protocolos utilizados a fin de mejorar la calidad del servicio como, incluso, manipular las actividades que se esten llevando sobre ella.

En el presente trabajo experimentaremos sobre sistemas basados en dos de las tecnologías de redes más difundidas, Ethernet (802.3) y WiFi (802.11), y analizaremos los datos obtenidos utilizando dos modelos de fuente de información para extraer conclusiones sobre el uso y la configuración de las redes. Los conceptos teóricos sobre los que basaremos el análisis se presentan a continuación.

1.1. Información y Fuente de información

Una fuente de información es todo aquello que emite mensajes de acuerdo a una ley de probabilidad fija. Los mensajes pertenecen a un conjunto finito de símbolos $S = s_1, ..., s_n$, conocido como el alfabeto de la fuente. La emisión de un símbolo s_i por parte de la fuente S representa un evento que tiene asociada una probabilidad fija $P_S(s_i)$ de ocurrir.

Dado un evento e con probabilidad $P(e)$, se define la **información del evento** e como

$$I(e) = -\log P(e)$$

$I(e)$ es una medida de la cantidad de información que obtenemos por la ocurrencia de E: mientras más improbable sea E, mayor será la información brindada por su ocurrencia (menor será la incertidumbre sobre el hecho observado). Dicho de otra manera, si sabemos que un evento E tiene alta probabilidad de ocurrir, entonces su ocurrencia no aportará mucha información sobre lo que se está observando.

1.2. Entropía

Dada una fuente de información $S = s_1, ..., s_n$, se define la entropía de S, $H(S)$, como la suma ponderada de la información de cada símbolo de S

$$H(S) = \sum_{i=1}^{n} P(s_i) * I(s_i)$$

La entropía de una fuente de información mide la cantidad de información esperada al observar la emisión de un nuevo símbolo por parte de la fuente.

1.3. ARP

Para poder realizar envío de paquetes de capa de red utilizando los servicios de capa de enlace es necesario poder realizar un mapeo entre las direcciones de ambas capas. ARP (*Address Resolution Protocol*) es un protocolo de control que surge como respuesta a esta necesidad. Cada host y switch de

una red mantiene una tabla donde se relaciona una dirección lógica d con la direccion física f a la que debe entregarse cualquier paquete destinado a d (el host con direccion física f no es necesariamente el destinatario de la dirección d: puede ser un intermediario que sabe cómo hacer llegar el paquete a d). En el caso de redes IP sobre Ethernet, el protocolo ARP es utilizado para mapear direcciones IP con direcciones MAC. La configuración de estas tablas ARP se realiza dinámicamente siguiendo un protocolo que consiste básicamente en los siguientes pasos:

1. Un host **A** desea enviar un paquete a una determinada IP. Si **A** conoce la dirección MAC a la que debe enviar los paquetes destinados a esa IP, entonces utiliza esa dirección física. Si no, envía un mensaje *broadcast*, o sea a todos los hosts dentro de la red, y aguarda la respuesta. Este mensaje se conoce como **ARP request** (WHO_HAS), y lleva la siguiente información:

   IP origen: IP de A

   IP destino: IP a la que se desea enviar un paquete

   MAC origen: MAC de A

   MAC destino: dirección broadcast de Ethernet (FF:FF:FF:FF)

2. Si dentro de la red existe un host **B** que sabe cómo direccionar a la dirección IP requerida, entonces responde al mensaje ARP request con un mensaje **ARP reply** (IS_AT) indicando su dirección física MAC. Este host puede ser el dueño de la dirección IP, o un host intermediario (como un router). Además, extrae las direcciones IP origen y MAC origen del paquete ARP request, y actualiza su tabla ARP para relacionarlas. El paquete ARP reply contiene la siguiente información:

   IP origen: IP de B

   IP destino: IP de A

   MAC origen: MAC de B

   MAC destino: MAC de A

3. **A** recibe el ARP reply de **B**, actualiza su tabla ARP y envía el paquete original utilizando la dirección física de **B**.

Además, cada entrada de las tablas ARP tiene seteado un tiempo de vida. Una vez agotado ese tiempo, la entrada se descarta y debe volver a aprenderse.

2. Experimentos

Como mencionamos anteriormente, el análisis de paquetes de una red puede utilizarse para inferir información sobre la actividad y topología de la red. En este trabajo aprovecharemos esta capacidad para dilucidar qué protocolos se distinguen del resto, cuál es la incidencia de los paquetes ARP y cuáles son los nodos destacados de las redes. Realizamos cuatro experimentos para obtener datos, uno sobre cada una de las siguientes redes:

- Red1: Red WiFi de un laboratorio del DC
- Red2: Red Wifi de un bar Starbucks
- Red3: Red Ethernet en un ámbito laboral
- Red4: Red Ethernet en un ámbito laboral

Modelamos estas redes como dos fuentes de información distintas:

1. S: este modelo fue dado por la cátedra. El alfabeto se define como los protocolos enviados dentro de los paquetes Ethernet capturados durante el experimento. Así mismo, consideramos como función de probabilidad a la frecuencia de cada símbolo dentro del experimento, donde marcamos como ocurrencia de un evento a la observación de un protocolo al capturar un paquete.
2. $S1$: con este modelo deseamos poder distinguir los nodos relevantes de una red data. Para ello definimos el alfabeto de $S1$ como las direcciones IP destino de los paquetes WHO_HAS del protocolo ARP.

   La decisión de utilizar las direcciones IP es porque nos interesa saber cuál es la dirección IP del nodo que es más requerido en la red, lo que implicaría que es un nodo central a la red. Podría ser algún router, un servidor o una impresora de red, por nombrar algunos.

2.1. Herramientas de *sniffing*

Para capturar y procesar la información, utilizamos tanto el programa *Wireshark* como dos script (`capturar.py`, `identificar.py`), escritos en Python, utilizando la librería para análisis de redes `scapy`. Ambas herramientas hacen uso del modo promiscuo de la placa de red, en el cual se capturan no solo los paquetes dirigidos a el host que esta capturando, sino todos los paquetes que se envíen por el medio.

2.1.1. Implementación de S: `capturar.py`

En su forma de ejecución básica, el script muestra por pantalla cada paquete que captura hasta que sea detenido con una interrupción (`CTRL+C`). Al finalizar, se muestra

1. el total de paquetes capturados
2. los protocolos observados (junto con la cantidad de veces que se observó cada uno)
3. la entropía correspondiente modelo S.

Si bien incorporamos varias opciones de ejecución (ejecutar el comando con la opción `-h`.), la forma más sencilla corresponde a

```
sudo python capturar.py -i <interfaz_de_captura>
```

2.1.2. Implementación de $S1$: `identificar.py`

Este script es similar a `capturar.py`, pero en lugar de analizar los protocolos de cada paquete, filtra sólo los paquetes ARP e implementa el modelo de fuente $S1$. Al igual que `capturar.py`, el script muestra por pantalla cada paquete que captura hasta que sea detenido con una interrupción (`CTRL+C`). Al finalizar, devuelve

1. un diccionario donde se mapea direcciones MAC con direcciones IP
2. un diccionario de direcciones IP observadas, y su cantidad de veces que fueron observadas
3. las direcciones MAC observadas (junto con la cantidad de veces que se observo cada una)
4. la entropía correspondiente modelo $S1$.

Para ver opciones de ejecución, ejecutar el comando con `-h`.). la forma más sencilla corresponde a

```
sudo python identificar.py -i <interfaz_de_captura>
```

3. Resultados y discusión

3.1. Experimentos de captura de protocolos

El objetivo de este experimento fue modelar el tráfico de la red como una fuente de información cuyos símbolos son los protocolos que están contenidos dentro de los paquetes Ethernet, a la que llamamos S.

3.1.1. Laboratorio de universidad

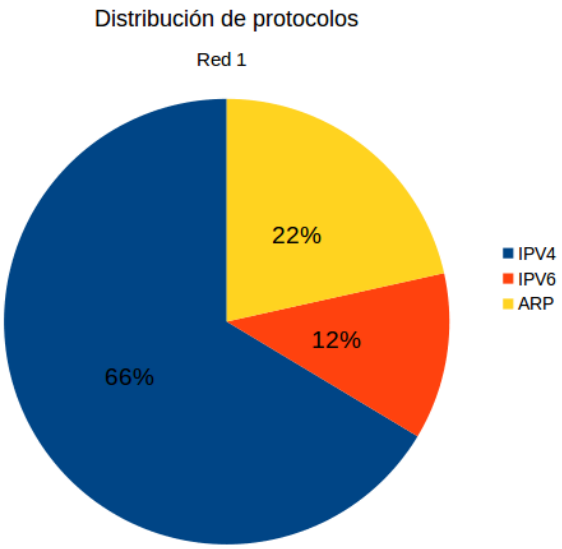

Figura 1: Distribución de protocolos de la red 1

Como podemos ver en la figura 1 hay un protocolo que tiene más presencia a la red: IPv4. Este protocolo es la piedra fundamental de la Internet, y es el que se usa para enviar paquetes de una red a otra a nivel mundial. IPv6 también se puede utilizar, pero todavía no tiene la suficiente adpoción como para reemplazar totalmente a IPv4, pero es una cuestión de tiempo ya que IPv4 está encontrando limitaciones técnicas que IPv6 soluciona. También aparecen en la red paquetes que usan el protocolo ARP descripto anteriormente que permite a los nodos de la red conocer la ubicación de otros nodos. Más adelante veremos el por qué de las apariciones de paquetes de protocolo ARP para esta red.

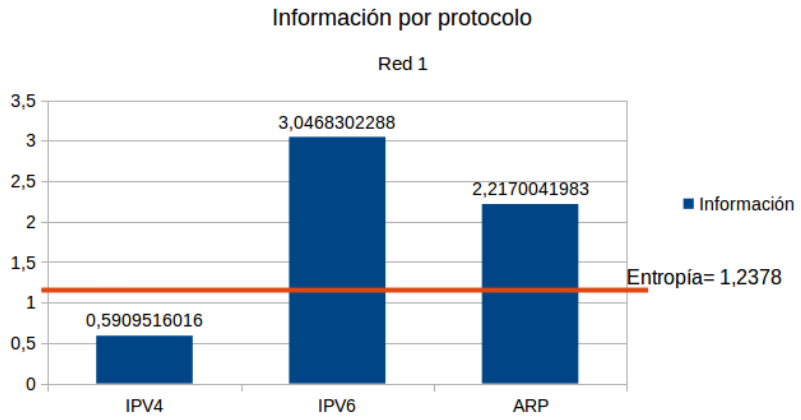

Figura 2: Cantidad de información que aporta cada protocolo, junto a la entropía de la fuente

Se puede apreciar en la figura 2 como vale la relación entre mayor cantidad de apariciones en la red implican un menor aporte de información. Asimismo vale la inversa, protocolos con pocas apariciones aportan mucha más información acerca de la fuente. La entropía viene a hacer el papel de un valor de información medio esperado por la fuente. Es decir, en cualquier momento se espera que la fuente arroje en promedio símbolos con información igual a la entropía. Un valor alto de entropía indicaría que estamos en presencia de una fuente de la cual no podemos decir con cierta seguridad qué símbolo emitirá la fuente. En este caso el protocolo IPv6 es el que más información debido a su baja frecuencia de aparición.

3.1.2. Cafetería

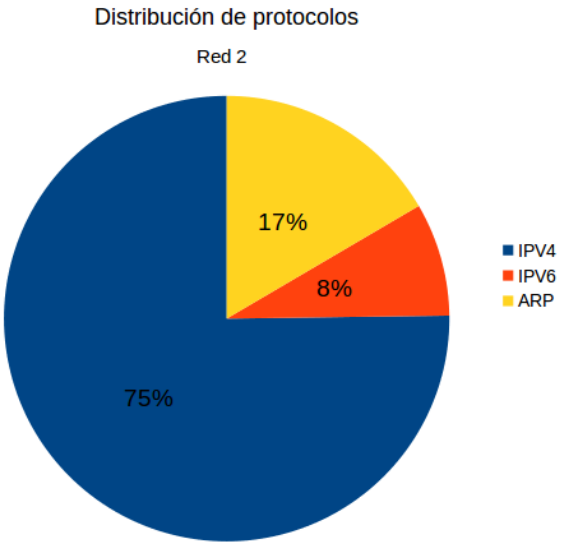

Figura 3: Distribución de protocolos de la red 2

Vemos ahora en la figura 3 que nuevamente IPv4 sobresale del resto en términos de frecuencia de aparición. Tiene sentido ya que esta red es una red de acceso público que suelen usarse principalmente para acceder a Internet para navegar o enviar o recibir mails o mensajes. Luego le siguen los paquetes del protocolo ARP, cuya frecuencia de aparición elevada se puede atribuir a que en una red de acceso

público inalámbrica es común que constantemente se vayan agregando o retirando nodos de la red. Cada vez que un dispositivo nuevo se conecta a la red tiene que saber a qué nodo mandar los paquetes para poder acceder a Internet.

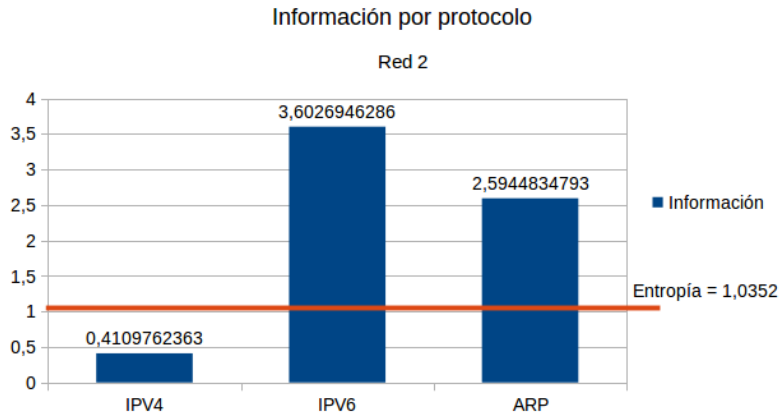

Figura 4: Cantidad de información que aporta cada protocolo, junto a la entropía de la fuente

En el caso de la figura 4 los paquetes con protocolo IPv6 son aún menos frecuentes que los anteriores ya que es una red de acceso público en la cuál se conecta toda variedad de dispositivos comerciales que sólo soportan el protocolo IPv4. Es por eso que la cantidad de información que aporta es muy alta. La entropía disminuye en medida de que un protocolo se vuelve cada vez más frecuente, como sucede en este caso en comparación a la red de la universidad.

3.1.3. Ambientes laborales

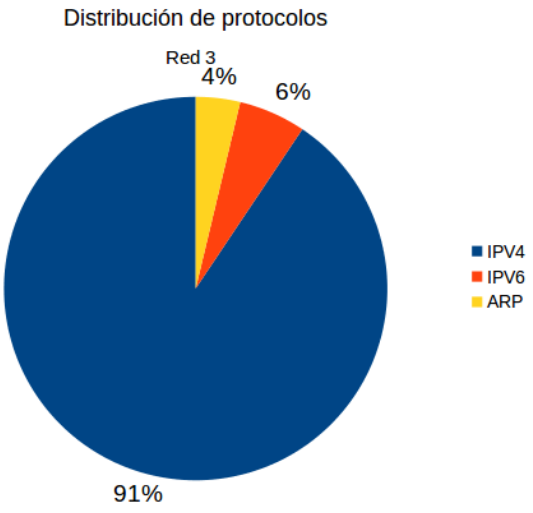

Figura 5: Distribución de protocolos de la red 3

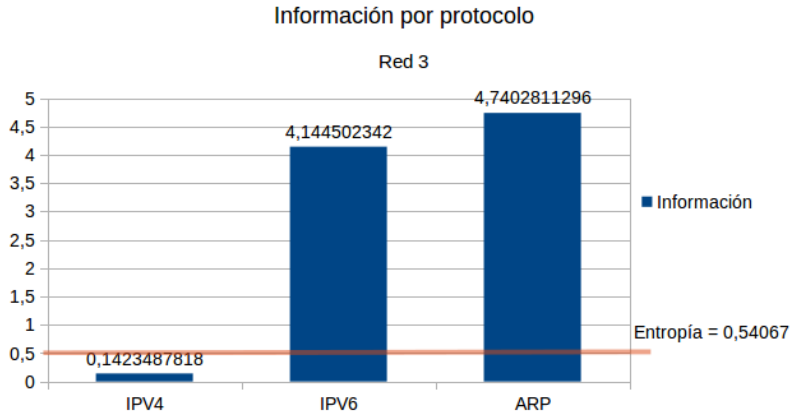

Figura 6: Cantidad de información que aporta cada protocolo, junto a la entropía de la fuente

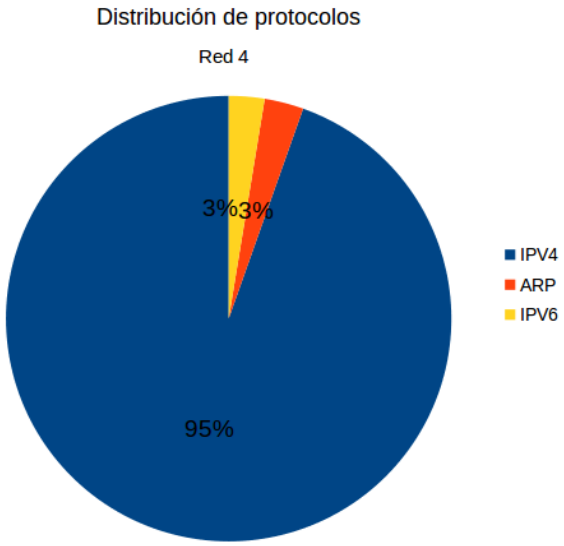

Figura 7: Distribución de protocolos de la red 4

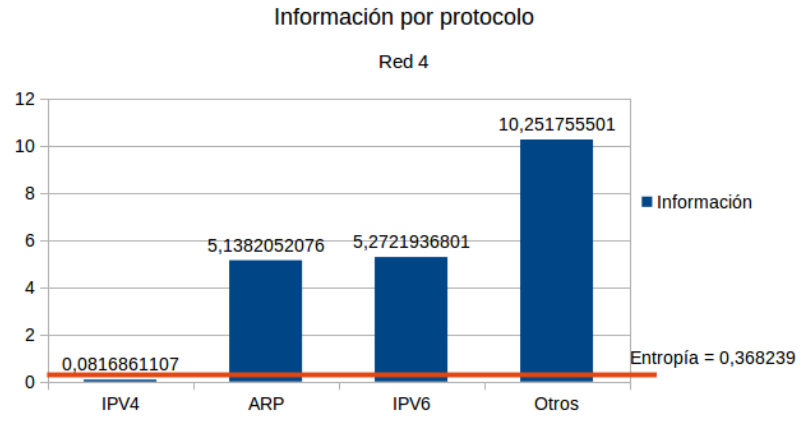

Figura 8: Cantidad de información que aporta cada protocolo, junto a la entropía de la fuente

Ahora vemos el caso de redes de ambientes laborales de las figuras 5 y 7 en donde se espera que sea estable en cantidad de nodos de que se conectan o desconectan a la red. En ambos casos el protocolo IPv4 es el más frecuente de todos y mucho más en comparación que las redes anteriores. Si los dispositivos se conectan a un servidor remoto en otra red (u otro país) es esperable ver muchos paquetes IP yendo a través de la red.

También podemos observar tanto en la figura 6 y 8 que, al haber un protocolo tan frecuente, disminuye la entropía en comparación con las otras 2 redes anteriores. Asimismo aumenta la cantidad de información que aportan los protocolos IPv6 y ARP.

3.2. Experimento de identificación de nodos

Veamos ahora el resultado de correr la herramienta `identificar.py` sobre las mismas capturas de las redes anteriores pero considerando a la fuente de información S_1 como aquella que emite direcciones de IP que aparecen como destino en los paquetes WHO_HAS.

3.2.1. Laboratorio de universidad

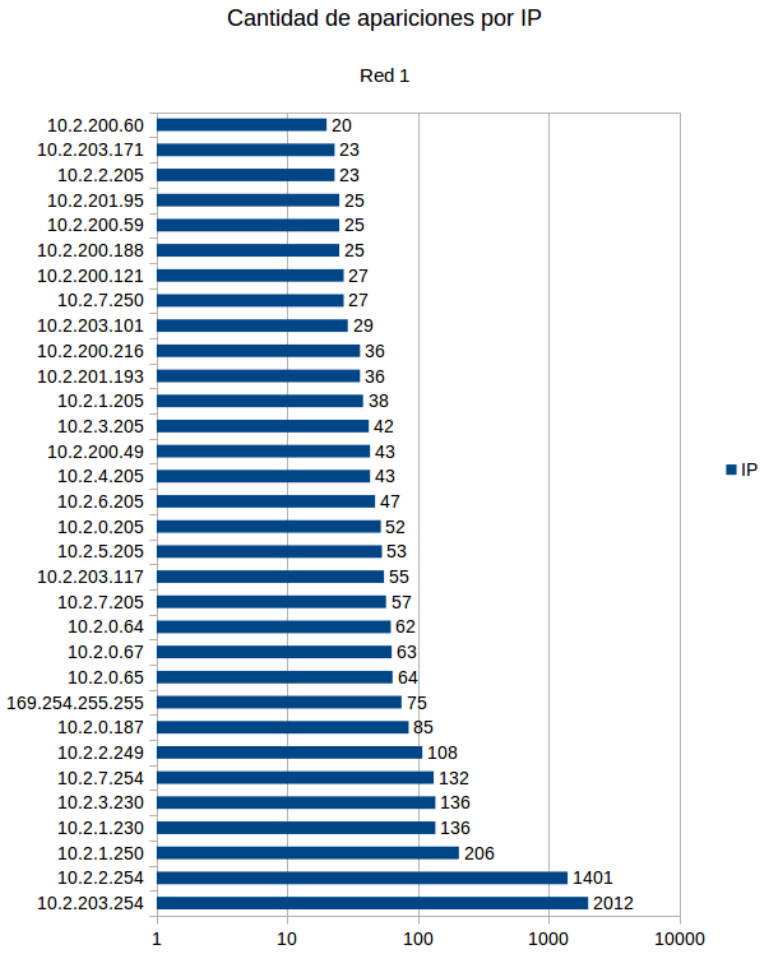

Figura 9: Cantidad de apariciones de direcciones IP en los paquetes ARP de la red 1

En la figura 9 podemos ver la cantidad de apariciones de cada dirección IP en los paquetes ARP de la red de la facultad. En realidad hay más direcciones IP en la lista, pero eran demasiadas para graficar, así que nos quedamos con las que más veces aparecen. Analizando la captura de los datos podemos ver que muchas de esas apariciones son de paquetes WHO_HAS repetidos, donde un mismo nodo pregunta varias veces por esa IP. Esto es algo normal en una red WiFi porque es un medio poco confiable y suelen perderse los paquetes, más en una red con mucho tráfico. Esta es la razón también de porque la presencia de paquetes ARP en esta red era bastante grande, en comparación con otras redes. La cantidad de información que agregan paquetes repetidos claramente es baja, ya que más de un paquete no nos dice nada nuevo sobre la topología de la red.

3.2.2. Cafetería

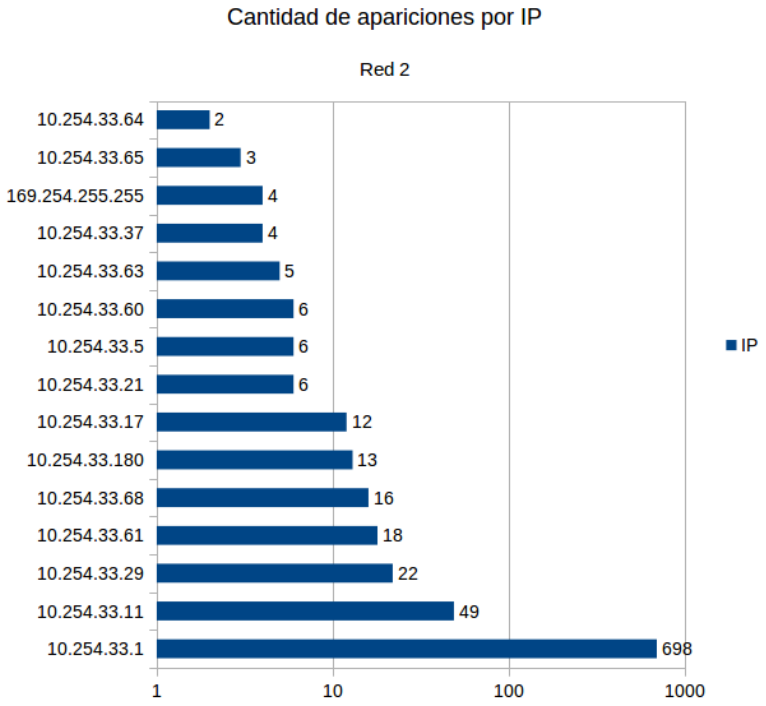

Figura 10: Cantidad de apariciones de direcciones IP en los paquetes ARP de la red 2

En un caso más normal, en la figura 10 podemos ver que el nodo con más apariciones es la primera dirección IP de la red `10.254.33.124` que usualmente es el que se le asigna al router que se encargará de recibir y enviar paquetes entre la Internet y la red propiamente dicha. Un dispositivo cuando se conecta a la red WiFi necesita saber a qué dispositivo mandar los paquetes hacia Internet, así que realiza un envío de un paquete ARP preguntando por la dirección física de la IP `10.254.33.1`. Como los dispositivos en una cafetería van y vienen, es esperable la aparición de muchos de estos paquetes.

3.2.3. Ambientes laborales

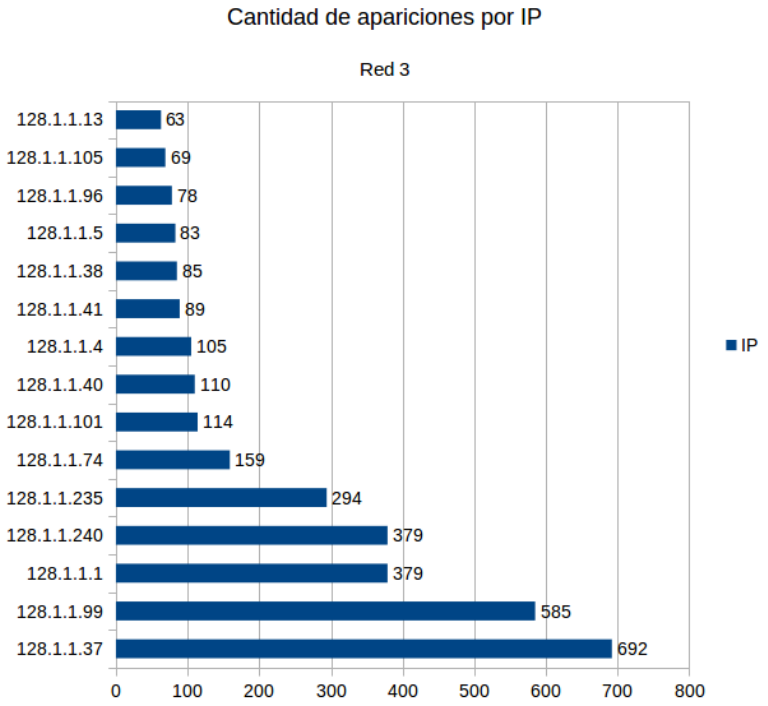

Figura 11: Cantidad de apariciones de direcciones IP en los paquetes ARP de la red 3

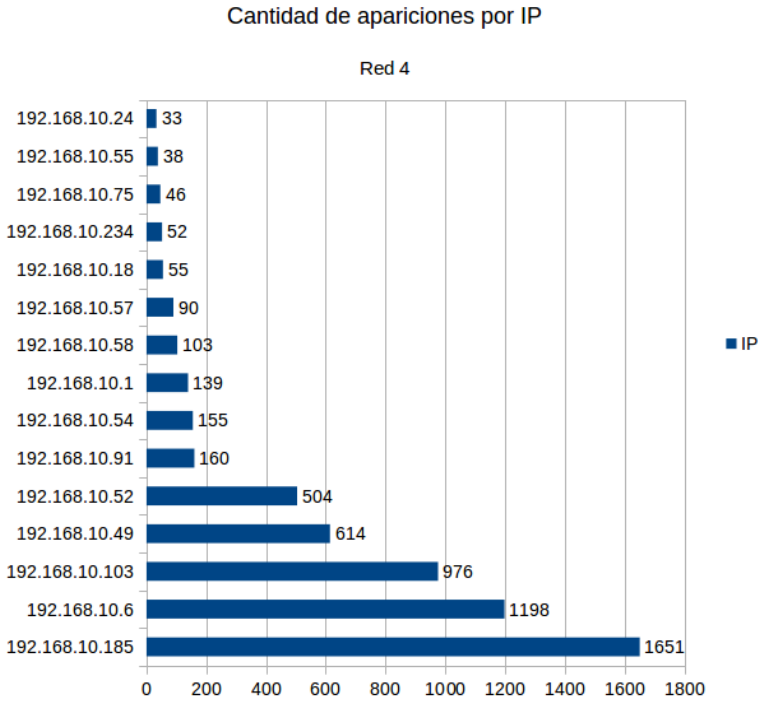

Figura 12: Cantidad de apariciones de direcciones IP en los paquetes ARP de la red 4

En las redes de ambientes de trabajo de las figuras 11 y 12 la dirección IP `128.1.1.1` no es la que más aparece, sino que hay otras. Posiblemente estas IP que más aparecen correspondan con nodos distinguidos en la red como un servidor, una base de datos, una impresora o algún otro recurso compartido por red.

3.3. Conclusiones

Para cerrar este informe queremos plantear nuestras conclusiones provenientes de la expermientación de modelar una fuente de información usando datos de capturas de paquetes de red, analizando la frecuencia con la cual aparecía cada protocolo o dirección IP. Observamos que existe una relación entre frecuencia de aparición y cantidad de información aportada por el evento que consiste en ver ese paquete en la red. Esto se debe a que nostros calculamos la probabilidad de que suceda un evento en base a las frecuencias relativas de aparición de los paquetes en la red, y como el cálculo de información aportada por un evento considera la inversa de esta probabilidad crea una relación inversamente proporcional entre ambas métricas.

Primero analizando la aparición de paquetes por protocolo podemos apreciar que mucho del tráfico generado en una red se debe principalmente al intercambio de paquetes IPv4 entre nodos de la red o hacia Internet, mientras que el protocolo ARP sólo representaba una pequeña porción. Sin embargo, esos paquetes ARP son los que más información aportan sobre la red y el cálculo de información aportada está de acuerdo con ello. Luego podemos ver cómo cambia la entropía dependiendo del tráfico en la red. Si hay muchos paquetes del mismo tipo, esto implica que la probabilidad el próximo paquete que aparece en la red sea de ese tipo va aumentando y por lo tanto sabemos más sobre la red. Esto se ve reflejado en la entropía, porque cuando un paquete era más frecuente la entropía disminuía porque la incertidumbre del tipo del próximo paquete era más baja. Adicionalmente podemos ver la poca presencia que tienen los paquetes IPv6 en las redes hoy en día, pero es cuestión de tiempo hasta que esto se vaya revirtiendo ya que IPv6 puede reemplazar a IPv4 por la creciente demanda de direcciones IP a nivel mundial.

Segundo analizamos la topología de la redes también modelando el tráfico como una fuente de información que emite direcciones IP provenientes de paquetes ARP de la red. Con esta técnica podemos apreciar qué nodos sobresalen del resto, pero a veces por cuestiones distintas. En la red de la universidad podíamos ver, analizando el tráfico, que había mucho reenvío de paquetes y por lo tanto las probabilidades de que aparezca una dirección IP aumentaban en cuanto a la frecuencia relativa. En este caso también podía deberse a que la red estaba sobrecargada y si bien los nodos probablemente respondían con un paquete IS_AT, éste nunca llegaba. En redes más controladas, pudimos ver qué un nodo principal correspondía con la dirección que generalmente tienen los routers. Como era una red de acceso público en una cafetería tiene sentido que este sea el nodo distinguido ya que dificilmente dos dispositivos en una red pública quieran comunicarse entre sí, ya que el uso principal que se les da es para acceder a Internet. Además en una red pública es esperable que nuevos dispositivos se conecten a la red en cualquier momento, entonces siempre van a necesitar saber la dirección física del router mediante el uso de paquetes ARP. Finalmente en las redes de ambientes laborales podemos ver que no es el router el nodo más distinguido ya que es posible que haya otro recursos compartidos dentro de la red del trabajo que son más requeridos que el acceso a Internet.