Wiretapping de paquetes ARP: analizando redes locales usando Teoría de la Información

Gonzalo Ciruelos , Axel Maddonni y Federico Patané

***Resumen*—Con el objetivo de comprender diversos aspectos de una red local, nos planeteamos analizar sus paquetes ARP usando herramientas de la teoría de la información. Para lograr eso, modelamos algún aspecto de nuestra red como una fuente de inforamción, de manera de poder analizarla con las herramientas previamente dichas. Además, comparamos distintos tipos de redes para poder confirmar o refutar nuestras hipótesis generales.**

***Index Terms*—ARP, LAN, broadcast, unicast, entropía, información, Ethernet, Wi-Fi**

I. Introducción teórica

En este trabajo nos proponemos utilizar herramientas de la Teoría de la Información y los paquetes ARP para intentar comprender diferentes redes locales. La pregunta que nos intentaremos responder en este trabajo es ¿Podemos conocer la topología de la red y/o sus nodos más importantes?

La explicación de las fuentes utilizadas vendrá más adelante, pero es pertinente presentar el formato de los paquetes ARP. El protocolo de resolución de direcciones (Address Resolution Protocol) [1] es un protocolo usado para la resolución de direcciones de la capa de red a direcciones de la capa de enlace (usualmente MAC), lo que es una función crítica en redes de múltiple acceso.

Un paquete ARP tiene muchas partes, pero la que más nos interesa es el campo OPER, o sea el campo de Operación, que nos indica qué función cumple ese paquete. Los distintos valores de OPER en paquete ARP son is-at y who-has. Un paquete who-has es un paquete broadcast en el que una computadora pregunta al resto de la red quién es la que tiene una dirección de capa de red (generalmente IP) dada. La respuesta a ese paquete es un paquete is-at, que es un paquete unicast, enviado desde la computadora con la dirección de capa de red requerida hasta la computadora que emitió el who-has.

Por otro lado, con respecto a la teoría de la información, utilizaremos los conceptos de información y entropía clásicos.

II. Desarrollo

Como explicamos anteriormente, el trabajo se basará en analizar redes locales capturando paquetes ARP de redes locales y analizandolos utilizando herramientas de la teoría de la información.

Para capturar los paquetes de la red analizada, utilizamos el programa Wireshark. Luego, para postprocesar los datos y computar la inforamación pedida utilizamos la librería de Python Scapy. Scapy es una poderosa herramienta que permite capturar, decodificar, crear y envíar paquetes de manera muy sencilla.

A. Ejercicio 1

El ejercicio 1 proponía analizar los resultados modelando a los paquetes como una fuente de información muy simple: simplemente distinguir entre paquetes broadcast y paquetes unicast. La fuente consiste en dos símbolos, uno que representa a los paquetes unicast, y otro que representa a los paquetes broadcast.

Esta fuente no nos permitirá conocer muy bien quién es quién en la red, pero nos dará quizás un poco de información sobre lo que está pasando en la red local.

Que la cantidad de mensajes de broadcast sea muy grande, o sea, la información del símbolo que representa a los mensajes broadcast sea muy chica, será lo esperado. Esto se debe a que en una red normal, en la cual podemos escuchar todo lo que está pasando, lo esperable es que los dispositivos se comuniquen unos con otros en vez de estar envíando broadcasts todo el tiempo.

De alguna manera, si vemos pocos paquetes unicast, o sea que la información de este símbolo es muy alta, puede significar dos cosas:

1. No tenemos visibilidad total de la red, por ejemplo porque está switcheada, y entonces vemos solo los paquetes unicast dirigidos a nuestro host.
2. No hay comunicación efectiva entre los hosts de la red, porque los paquetes unicast de alguna manera miden cuanta comunicación de un host a otro está sucediendo.

Nuestra implementación de este ejercicio puede verse en el archivo `ejercicio1.py`.

B. Ejercicio 2

El ejercicio 2 proponía que diseñemos una fuente de memoria nula en base a los paquetes ARP, que nos permitiera distinguir a los nodos distinguidos de la red, para alguna definición que demos de nodo distinguido.

Experimentamos con varias fuentes, y terminamos seleccionando que la fuente sea el destino de los paquetes who-has, por las siguientes razones:

1. Si estamos analizando una red switcheada o muy subdividida en subredes virtuales (VLANs), entonces lo más probable es que solo recibamos paquetes who-has, dado que estos paquetes se envían en modo broadcast. Por esta razón los switches de la red local no los filtran y le llegan a todos los hosts. Quizás esta diferencia no sea notable en redes inalámbricas, pero en redes cableadas complejas, que generalmente tienen switches, puede hacer una gran diferencia.

G. Ciruelos e-mail: gonzalo.ciruelos@gmail.com
A. Maddonni, e-mail: axel.maddonni@gmail.com
F. Patané, e-mail: fedepatane20@gmail.com

2. Como justificamos anteriormente, vamos a usar paquetes who-has, ahora bien, la pregunta es porqué el destino de esos paquetes. En este caso la respuesta es más obvia: si un host es el destinatario de más paquetes ARP, entonces es más requerido por el resto de los hosts, entonces es más probable que sea un nodo distinguido, como por ejemplo un router.

Siguiendo estos simples preceptos, diseñamos nuestra fuente de información S1. Nuestra implementación de este ejercicio puede verse en el archivo `ejercicio2.py`.

C. Grafo de la red

Para los tres experimentos que realizamos, hicimos el grafo de la red que se desprende de los paquetes ARP enviados a lo largo de la captura.

El grafo fue realizado de la manera más natural. Por cada mensaje who-has capturado, el grafo tendrá una arista. Además, esa arista irá del nodo con IP igual a la IP fuente del who-has al nodo con IP destino del who-has.

Por razones de comodidad, solo mostramos aquellos nodos relevantes, es decir, los que tienen información baja con respecto a la fuente de información que S1.

Además, juntamos en uno a todos los hosts que tienen exactamente el mismo conjunto de aristas adyacentes. Esto se indicara con un [X]: si un nodo tiene [X] significa que ahí condensamos X cantidad de hosts, que tienen todos exactamente la misma conectividad que el nodo representado.

Además, marcamos con un cuadrado aquellos nodos distinguidos según la fuente S1, es decir, aquellos que tienen menos información que la entropía.

D. Conceptos generales

D.1 Gratuitous ARP

En todos los experimentos apareció un tipo de paquete ARP llamado Gratuitous ARP, con lo cual nos parece mejor introducirlo al principio para dejar en claro qué es y por qué aparece en todos los experimentos.

Gratuitous ARP puede significar tanto un reply (is-at) como un request (who-has). Gratuito en este caso quiere decir que un request o un reply no es normalmente requerido de acuerdo con la especificacion de ARP (RFC 826) [1], pero puede ser usado en algunos casos. Un request ARP gratuito es un paquete donde la IP source y destination están ambas seteadas a la IP del host que envía el paquete. Además, la MAC destino es la dirección de broadcast `ff:ff:ff:ff:ff:ff`. Ordinariamente, no habrá respuesta para tal request.

Los ARP gratuitos tienen varias utilidades:

1. Pueden ayudar a detectar conflictos de IP. Si un host recibe un paquete ARP que contiene una IP source que coincide con la suya, sabe que hay un conflicto.
2. Ayuda a actualizar las tablas ARP de los hosts de la red.
3. Cada vez que una interfaz IP se prende, el driver de la interfaz típicamente envía paquetes ARP gratuitos para precargar las tablas ARP de todos los hosts. Por eso, si un host envía muchos paquetes ARP gratuitos, podemos inferir que algo malo está sucediendo con él, por ejemplo que se está reiniciando o que su interfaz IP se reinicia continuamente porque no puede iniciarse correctamente.

D.2 Dirección 169.254.255.255

[eg an address you end up picking yourself because DHCP didn't work usually picked using some arp to figure out if it belongs to someone else]

III. Experimento 1

El primer experimento que presentamos en el trabajo fue realizado en una cafetería (Starbucks), por la tarde, mientras había aproximadamente unas 15 personas a la vista usando dispositivos móviles. La captura de paquetes duró casi una hora.

A. Resultados

Primero, la fuente S. Nuestra hipótesis sobre esta fuente, basándonos en lo que dijimos anteriormente en el trabajo, es que el símbolo de los paquetes de broadcast tendrá una información mucho mas alta que el símbolo de los paquetes unicast.

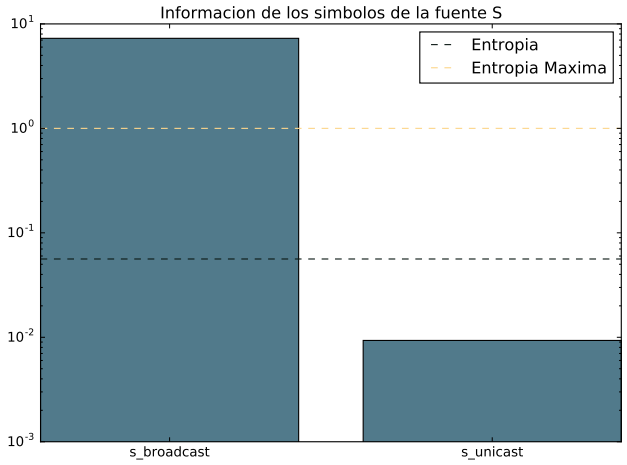

Figura 1

Nuestra hipótesis sobre la red es que habrá unos pocos dispositivos que se conectaran a un router, y quizás si el router y el AP ambos tienen IPs dentro de la red, aparezcan como dispositivos separados en la red. Analizaremos esto, y todas nuestras otras hipótesis en al discusión.

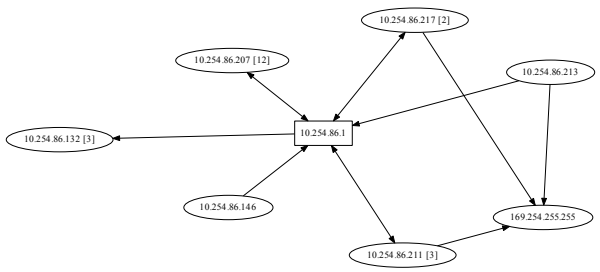

Figura 2
Grafo de conectividad de la red, inferido de los paquetes who-has. Para ver con mayor detalle, se puede hacer zoom-in en el pdf.

Sobre la fuente S1, esperaremos que haya un dispositivo que tenga menor información que el resto, y esperaremos que este sea el router. Además, esperamos que el resto de los dispositivos sean hosts, y tengan menor información si son más activos en la red.

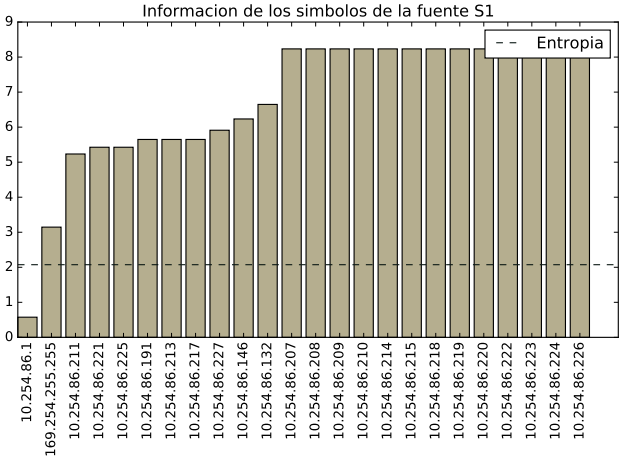

Figura 3
Información de los símbolos de la fuente S1: solamente los nodos con menor información son representados.

B. Discusión

Analicemos los resultados que vimos en la sección anterior.

Empecemos por la fuente S. Como esperabamos, la información del símbolo asociado a los paquetes broadcast es muy alta. Como dijimos al principio del trabajo, asociaremos esto a dos hechos:

1. Podemos ver todo lo que pasa en la red, en particular, podemos ver todos los paquetes que se le envian a los host distintos del nuestro. Esto en general nos va a permitir inferir o bien que la red es una red Wi-Fi, o bien que es una red Ethernet muy simple, sin switches o subnets.
2. Además, que la comunicación en la red es efectiva, o sea que los hosts están comunicando cosas unos con otros o bien con dispositivos fuera de la red. Si la mayoría de los paquetes fuera broadcast, eso nos daría la idea de que no se está dando comunicación efectiva entre los hosts, si no, que por ejemplo, todos los mensajes que se intercambian son de control.

Por esa razón, es mucho más común en este caso que un paquete sea unicast que broadcast. Por ello, entropía de la fuente S no va a ser máxima, dado que la información de los únicos dos símboos es muy diferente.

Continuando con el análisis del grafo de la red y la fuente S1, se observa que hay 24 nodos. El nodo distinguido (según la fuente S1) es lo que parece ser el router. Inferimos que es el router porque su último octeto es 1, que es el valor default para routers.

Las IP de los nodos corresponde a la red 10.254.86.0/24. La mayoría de los hosts terminaban en octetos distintos de 1, excepto el router, que como dijimos tenía a 1 como último octeto. Esto nos da la pauta de que efectivamente lo que estamos viendo es una red, que además se corresponde con la red Wi-Fi de la cafetería.

Nos parece razonable afirmar que el resto de los nodos son hosts conectados a la red Wi-Fi, dado que por lo que se veía en la cafetería, había un mínimo de 15 dispositivos conectados a la red, y además el resto se comporta de la misma manera.

Notemos que los nodos que no se conectan solamente con el router (10.254.86.1), se conectan con el nodo 169.254.255.255. El nodo que tiene esta IP ss un caso que puede ser considerado anómalo o borde y fue explicado anteriormente. Es importante decir que apareció en todos los experimentos con Wi-Fi que hicimos, con lo cual es interesante tenerlo en cuenta para futuros experimentos.

La cantidad de nodos del grafo indica que los dispositivos conectados a la red Wi-Fi son los que observamos, más algunos que no pudimos ver.

Es muy interesante que la fuente propuesta y nuestro grafo de nodos nos permitió de una manera muy precisa comprender la disposición y organización de la red. En las conclusiones haremos un análisis general de este y el resto de los experimentos.

IV. Experimento 2

El segundo experimento que mostraremos fue realizado en los laboratorios del Departamento de Computación de FCEyN, más especificamente en el laboratorio 2. La captura de paquetes duró una hora y se realizó sobre la red cableada del laboratorio, es decir, sobre Ethernet. La modalidad fue desenchufar el cable Ethernet de la computadora 3 del laboratorio 2, y enchufandolo en nuestra propia computadora, de tal manera de poder capturar paquetes en modo promiscuo.

La captura de paquetes se realizó durante la tarde, mientras había aproximadamente 10 personas utilizando las computadoras de ese laboratorio. Las mediciones fueron realizadas con la autorización de uno de los administradores de la red, quien además ayudó a verificar las teorías que extrajimos de los paquetes. Veremos todo esto más adelante.

Antes de pasar a ver los resultados, una cosa que vale la pena decir es que, por como es la red de los laboratorios, al haber utilizado una computadora distinta de la de los labos para realizar las mediciones, la red no nos asignó IP, es decir, no nos "conectó" de manera total a la red local. Sin embargo, esto no perjudicó en nada a las mediciones, y solo se vió reflejado en la fuente S.

A. Resultados

Primero analicemos la fuente S, que es la fuente de los paquetes unicast y broadcast. Como dijimos anteriormente, no teníamos dirección IP asignada. Esto sumado a que la red es switcheada, nos hace esperar que la cantidad de paquetes unicast que vamos a ver sea mucho más baja que lo normal. Esto, dicho en lenguaje de la Teoría de la Información, es que el símbolo que representa a los mensajes unicast tendrá una información más baja de la esperada.

Sin embargo, por como está configurada la red del laboratorio, sabemos que detrás de cada switch hay aproximadamente 4 o 5 computadoras, con lo cual podremos ver algunos paquetes unicast. Veamos qué pasó.

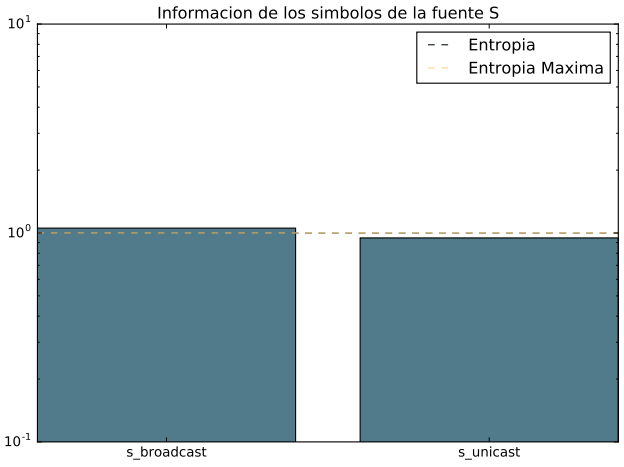

Figura 4

Ahora pasemos a ver el diagrama de conectividad de la red, que nos permitirá saber más sobre cómo está configurada.

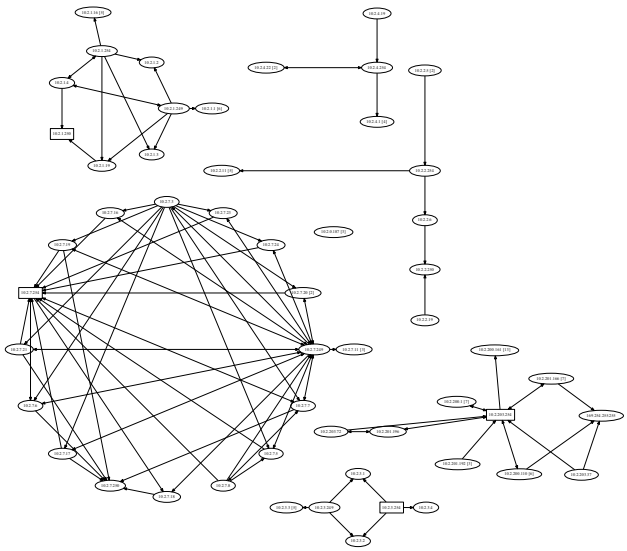

Figura 5
Grafo de conectividad de la red, inferido de los paquetes who-has. Para ver con mayor detalle, se puede hacer zoom-in en el pdf.

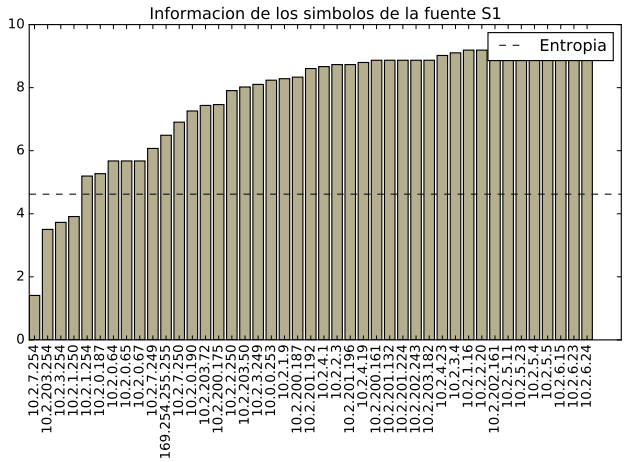

Figura 6
Información de los símbolos de la fuente S1: solamente los nodos con menor información son representados.

B. Discusión

Para empezar, se confirma lo que dijimos sobre la fuente S: la información del símbolo de los paquetes unicast es mucho más alta de lo esperada e iguala casi a la de los paquetes broadcast. Esto provoca que la entropía sea muy cercana a la máxima. Desde el punto de vista de la probabilidad, esta fuente es menos predecible que la que vimos antes.

Además, si no hubieramos sabido como estaba la configurada la red (que se ve a simple vista en los laboratorios), podríamos haber inferido que está fuertemente switcheada dada la información de estos símbolos.

Con lo cual, como dijimos al principio del trabajo, esta fuente nos permite saber, entre otras cosas, que tanta visibilidad tenemos de toda la red: a más visibilidad más paquetes unicast nos van a llegar, por lo tanto más baja va a ser la información del símbolo asociado.

Para continuar con el grafo de conectividad y la fuente S1, como vemos, hay muchos dispositivos en la red, contamos más de 100 dispositivos. Sin embargo, solo mostramos aquellos relevantes para la discusión. Esta cantidad nos dice que la red es muy compleja, y que salvo excepciones, seguramente tendrá alguna subestructuración interna, por ejemplo subnetting, y que por lo tanto será un desafío intentar comprenderla.

Yendo más a lo específico, se ve que hay un grafo muy similar a un grafo completo, que tiene un nodo distinguido con IP 10.2.7.254. Nuestra hipótesis será que este dispositivo es el router. Además, todos los nodos de ese subgrafo tienen IPs de la forma 10.2.7.XXX. Es decir, ese subgrafo representa la subred 10.2.7.0/24. Por la cantidad y el detalle de nodos observados en esta subred, podemos inferir que es la red del laboratorio 2, donde capturamos los paquetes. Sin embargo de ninguna manera podemos imaginarnos por que tantas de las computadoras se envían paquetes ARP unas a otras. Veremos la razón de esto más adelante.

Además, podemos observar que se repite el mismo patrón en distintos lugares, por lo que podemos asumir que son los distintos laboratorios. El patrón al que nos referimos es la IP de la pinta 10.2.n.XXX, donde n es el número de laboratorio y XXX es el número de host. Además, como antes, 10.2.n.254 suele ser el nodo distinguido, con lo cual podemos asumir que todos esos son los routers de cada labo.

Hay otras 2 partes de la red que nos gustaría analizar. Primero, tres nodos de IP 10.2.0.XX, que lo único que hacen es mandar Gratitious ARP. Como vimos anteriormente, un Gratuitious ARP son paquetes broadcast que actúan como paquetes is-at, pero que nunca fueron pedidos. Estos dispositivos actúan de manera anómala, pero como la red es muy compleja, podemos establecer como hipótesis que son dispositivos de control de la red.

Por último, podemos analizar la subred que se encuentra en la parte inferior izquierda de la figura. Como puede apreciarse, es bastante distinta al resto de las redes.

Esta red sigue un patrón muy similar al que observamos anteriormente en la red Wi-Fi. Además, aparece la misma IP que había aparecido antes: 169.254.255.255. Por último, esta subred tiene un nodo distinguido terminado en 254: 10.2.203.254, así que nuestra hipótesis será que esta subred representa parte de la red Wi-Fi de los laboratorios de la facultad.

Luego de plantear todas nuestras hipótesis, nos pareció interesante consultar con un administrador de la red para saber que tan exactas fueron nuestras predicciones utilizando un modelo tan simple como usamos.

Resultó que nuestras predicciones fueron muy acertadas. Como predijimos, cada labo se representa internamente con una subred. Estas subredes son las de la pinta 10.2.n.0/24. Además, los routers de cada labo son los de IP 10.2.n.254, como habíamos predicho.

Finalmente, el subgrafo que creíamos que se correspondía con la red Wi-Fi, efectivamente lo hacía. Según el administrador, esa subred corresponde a IPs de un pool dinámico de IPs, que se asignan en el rango de direcciones de 10.2.200.0 a 10.2.201.255 de forma dinámica cada vez que un dispositivo nuevo se conecta a la red. ado que el Wi-Fi de los laboratorios está configurado con un pool dinamico de direcciones, y 10.2.203.254 es el router de la red. Quizás era un poco dificil conocer toda esta información solamente con el grafo de conexiones, pero nos parece muy interesante contarlo de todas formas, porque es un patrón recurrente en redes Wi-Fi.

Además, nos aclaró que el hecho que las PCs de los laboratorios se manden paquetes unas a otras es esperado y tiene variadas causas. Primero, puede ser que dos usuarios esten usando un programa que use la red local, y entonces las computadoras se van a enviar paquetes ARP entre ellas. Otra cosa que sucede, es que las redes de los laboratorios tienen software de control y mantenimiento por temas de seguridad, y sobre todo para hacer que la red sea más fácil de mantener. Además es algo necesario para poder implementar el sistema conocido como Milagro, para poder conectarse a los labos a través de ssh.

Por último, el administrador desconocía la razón por la cual había ciertos nodos que lo único que hacían era enviar Gratuitous ARP. Nos parece interesante plantear este interrogante como trabajo futuro.

V. Experimento 1

A. Resultados

Figura 7

Figura 8
Grafo de conectividad de la red, inferido de los paquetes who-has. Para ver con mayor detalle, se puede hacer zoom-in en el pdf.

Figura 9
Información de los símbolos de la fuente S1: solamente los nodos con menor información son representados.

B. Discusión

VI. Conclusión

Referencias

[1] RFC 826 - Ethernet Address Resolution Protocol
http://tools.ietf.org/html/rfc826
C. Plummer 1982
[2] http://www.secdev.org/projects/scapy