Análisis de tráfico de redes locales usando Teoría de la Información

Manuel Costa Mathias Gatti

Resumen —

Index Terms —**entropía, Teoría de la informacin, ARP, LAN, unicast, broadcast, sniffing**

I. Introducción

Iniciamos este trabajo con el objetivo de aprender sobre el protocolo ARP y, más en general, las redes de Internet y sus algoritmos de intercambio de paquetes.

ARP es un protocolo de resolución de direcciones, mapea direcciones IP a direcciones MAC (es decir que va del nivel de red al nivel de enlace). Se compone de dos tipos de mensajes: *who-has* e *is-at*

A lo largo de este informe aplicaremos el conocimiento adquirido sobre teoría de la información para analizar propiedades en la red y detectar símbolos destacados. A partir de esta metodología creemos que podremos entre otras cosas diferenciar routers de hosts.

Nuestra hipótesis es que los routers destacarán en el intercambio de paquetes who-has siendo los que mas aparezcan como destinatarios del mismo.

II. Métodos

Herramientas

Para la captura de tráfico se utilizó el módulo de manipulación de paquetes *Scapy* para python, el cual provee una interfaz sencilla para nuestros requerimientos puntuales. *Scapy* permite la captura y posterior guardado de paquetes en una red, para luego ser filtrados, inspeccionados o manipulados con facilidad. Además, para incrementar la cantidad de paquetes vistos por un host, y que las capturas resulten más interesantes, se sniffeo con el modo promiscuo.

Modelo de las fuentes

Fuente S1

Dado el tráfico de capa 2 obtenido en cada captura, se modeló una fuente de memoria nula S1 $= \{s_1, s_2, s_3, ..., s_n\}$ donde cada s_i es una tupla ($broadcast \| unicast$, protocolo capa 3).

Fuente S2

Para la segunda parte modelamos la fuente de memoria nula S2, la cual intenta definir sus símbolos de manera que estos permitan encontrar nodos destacados a partir de herramientas de teoría de la información y los paquetes ARP de una red.

Luego de experimentar con distintas variantes posibles de los datos provistos por los paquetes ARP nos terminamos decidiendo por utilizar el destino de los paquetes who-has ya que estos a diferencia de los paquetes is-at son broadcast por lo que se propagan a través distintas subredes y switches permitiendo recibirlos en gran número, lo cual genera mediciones más robustas. Por otro lado elegimos quedarnos solo con el destino ya que creemos que es una buena heurística para encontrar nodos destacados. Si un nodo lo fuera (por ejemplo un router con salida a internet) entonces varios hosts querrán comunicarse con él, convirtiéndolo en el destino de un paquete who-has. Notar que no tendría tanto sentido utilizar la fuente ya que, al menos para routers, no va a ser de mucha utilidad debido a que probablemente estos ya tengan en su tabla a la mayoría de las MACs de los hosts de su red y no necesiten generar tantos paquetes ARP who-has. Por esta razón este pasaría desapercibido entre los demás hosts.

Capturas

Se hicieron 3 capturas en redes diferentes durante aproximadamente 30 minutos (Todas con al menos 10.000 tramas):

- Red hogareña pequeña, con aproximadamente 10 usuarios. Se utilizó una interfaz ethernet. La medición se realizó un jueves por la noche, a las 20 horas cuando la mayoría de los usuarios de la misma estaban conectados.
- Red mediana de oficina de una PyMe. Mediciones tomadas mediante la interfaz wifi. Se capturó el tráfico un miércoles a las 12 del mediodía, cuando la red estaba levemente congestionada.
- Red grande del laboratorio de informática de la universidad, mediciones tomadas mediante una interfaz wifi. Se capturó el tráfico a las 19 horas de un miércoles en el laboratorio Turing, cuando había 8 personas más utilizando las computadoras del laboratorio.

III. Resultados y análisis

Red hogareña

Resultados fuente S1

Los paquetes broadcast representan un poco más del 5 % del total, lo cual es un valor esperable en una red que ya tiene sus tablas ARP llenas y no necesita realizar tantos broadcasts.

M. Costa, e-mail: manucos94@gmail.com
M. Gatti, e-mail: mathigatti@gmail.com

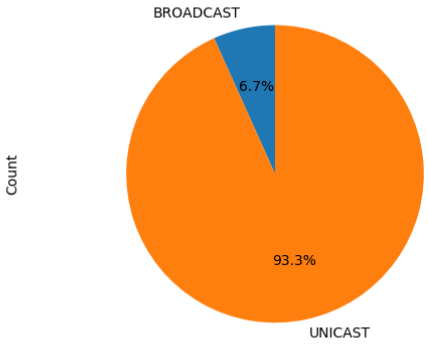

Figura 1: Proporción de paquetes unicast/broadcast en la captura

Como se puede ver en el siguiente diagrama de torta los protocolos encontrados fueron *ARP* e *IP* (tanto v4 como v6), el protocolo de internet utilizado para transmitir mensajes, por ejemplo datos de usuarios, entre redes LAN.

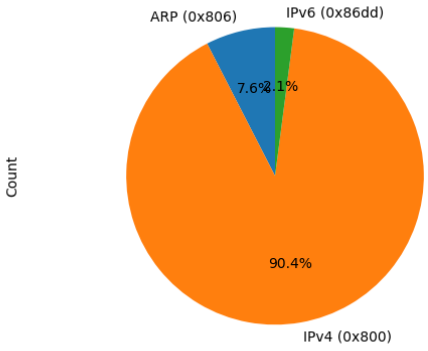

Figura 2: Proporción de protocolos en la captura

Habiendo visto el primer gráfico de torta no debería llamar la atención ver que el símbolo de menor entropía, o sea el más frecuente o de mayor probabilidad, es un unicast. En particular el protocolo con menor información es IPv4, lo que es consistente con ser el protocolo más popular en la red.

Debido a este símbolo destacado también sucede que la entropía no es máxima, aún cuando el resto de los símbolos tienen una distribución más uniforme entre sí.

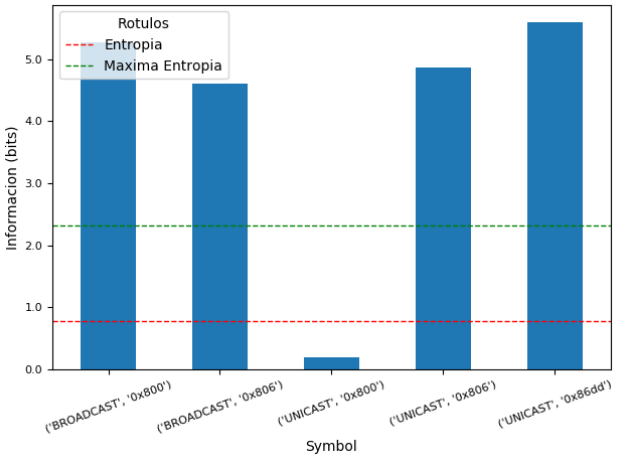

Figura 3: Información de los símbolos de la fuente S1, notando la entropía de la fuente, y la máxima entropía posible si la fuente fuera equiprobable.

Resultados fuente S2

En el siguiente gráfico se pueden ver 4 hosts con baja entropía y 5 con alta entropía. Según nuestra hipótesis será de esperarse que el router sea uno de los primeros 4, idealmente el de menor entropía.

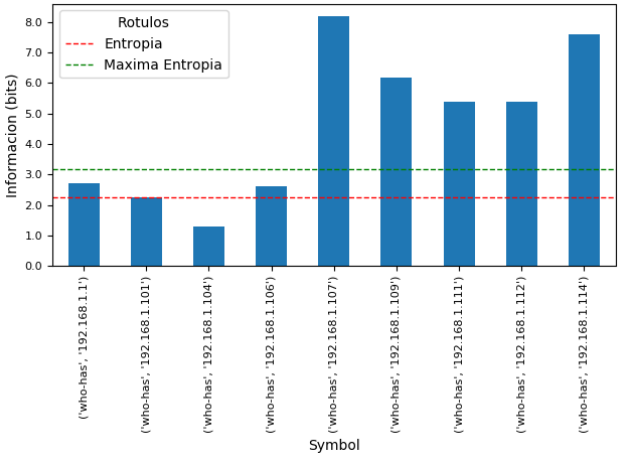

Figura 4: Información de los símbolos de la fuente S2, notando la entropía de la fuente, y la máxima entropía posible si la fuente fuera equiprobable.

A partir de los paquetes que se intercambian en la red armamos el grafo subyacente. En este cada vértice representa una IP local y cada arista va del origen al destino de un paquete who-has del protocolo ARP.

En este grafo los dos nodos con mayor grado son 192.168.1.1 y 192.168.1.112

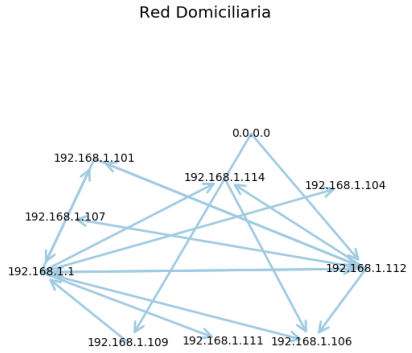

Figura 5: Grafo resultante de la red ethernet de una red domiciliaria durante la noche de un martes.

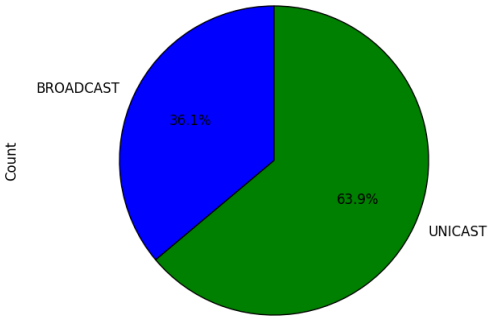

Figura 6: Proporción de paquetes unicast/broadcast en la captura

Cómo se puede ver si bien las dos técnicas utilizadas (Entropía y grafo subyacente) parecen apuntar a un conjunto pequeño de IPs no queda claro cuál sería el rol de cada IP o cual sería por ejemplo el router, el cual nosotros sabemos que es el 192.168.1.1, estos problemas probablemente se dan debido al tamaño de la red el cual no tiene una variedad suficientemente amplia de paquetes como para hacer robustas nuestras herramientas de predicción de símbolos destacados.

Red Oficina PyMe

Resultados fuente S1

Los paquetes broadcast representan casi el 25 % del total, esto es considerablemente más que en el caso anterior, una posible explicación es que la tabla de ARP de algún host se haya limpiado recientemente generando que este tenga que enviar varios who-is broadcast en la red.

La cantidad de paquetes ARP es bastante alta apoyando la hipótesis dicha anteriormente sobre la gran cantidad de broadcasts.

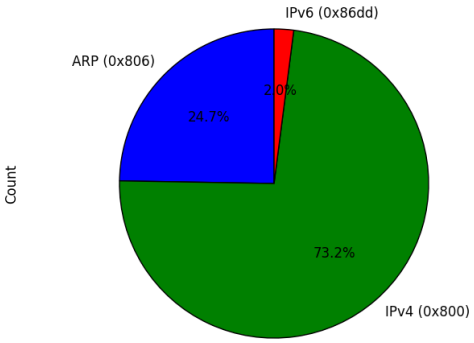

Figura 7: Proporción de protocolos en la captura

El resultado es similar al visto en la red anterior dando como símbolo destacado un $(unicast, IPv4)$.

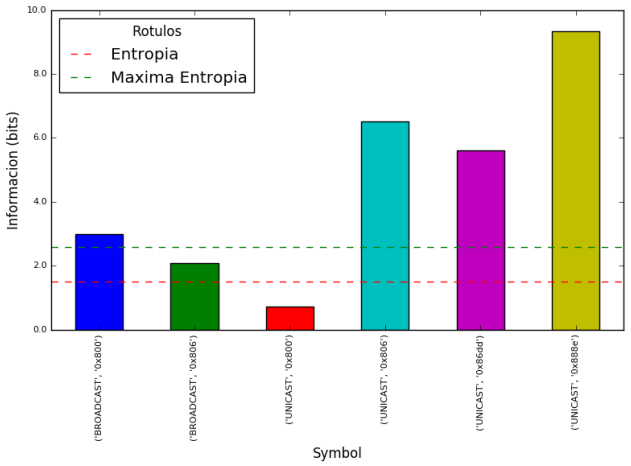

Figura 8: Información de los símbolos de la fuente S1, notando la entropía de la fuente, y la máxima entropía posible si la fuente fuera equiprobable.

Resultados fuente S2

En esta red hay un claro símbolo destacado con una entropía significativamente menor que la de cualquier otro.

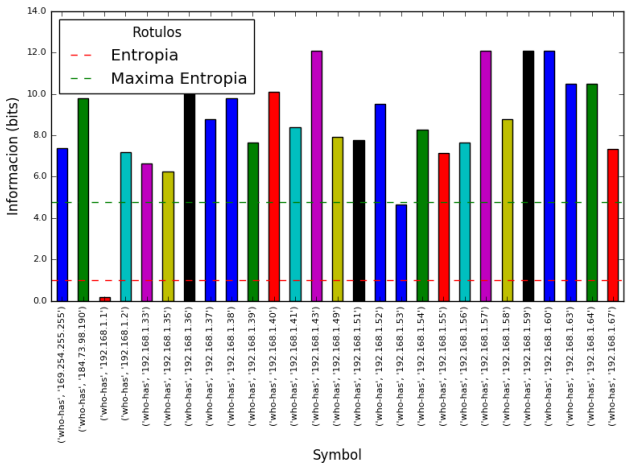

Figura 9: Información de los símbolos de la fuente S2, notando la entropía de la fuente, y la máxima entropía posible si la fuente fuera equiprobable.

Utilizando el grafo subyacente para corroborar lo visto podemos ver que efectivamente el host 192.168.1.1 destaca (En el grafo tiene un grado mucho más alto que los demás nodos).

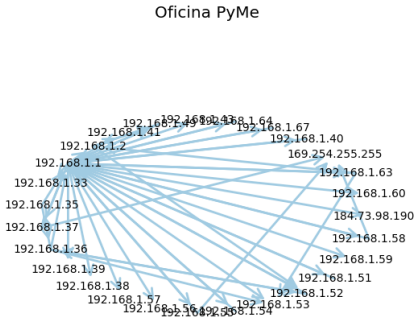

Figura 10: Grafo resultante de la red wifi de una red pública en un starbucks durante la tarde de un día de semana.

A diferencia del experimento con la red hogareña pudimos detectar correctamente el router de la red, esto apoya nuestra hipótesis de que el método estadístico funciona mejor con redes más grandes.

Red laboratorios

Resultados fuente S1

Al igual que con la primer red, los paquetes broadcast comprenden aproximadamente el 5 %.

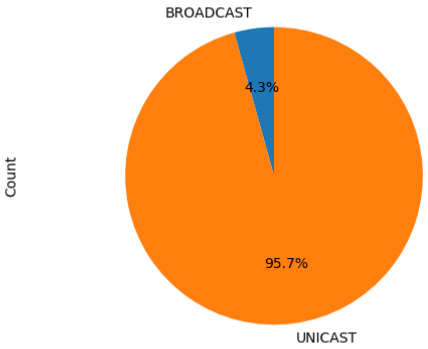

Figura 11: Proporción de paquetes unicast/broadcast en la captura

Los paquetes ARP comprenden menos del 5 % lo cual parece indicar que la red está bastante estable. También hay algunos pocos paquetes IPv6, confirmando su escasez en comparación al protocolo IPv4.

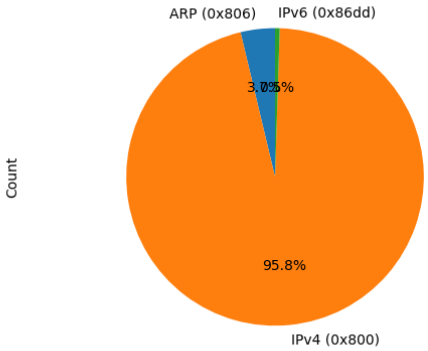

Figura 12: Proporción de protocolos en la captura

Al igual que en los casos anteriores el símbolo de menor entropía es un unicast de protocolo IPv4 (0x800), es interesante observar como dependiendo del protocolo varía si unicast o broadcast tiene mayor información. Por ejemplo con el protocolo ARP (0x806), los broadcasts tienen menor cantidad de bits, contrariamente a lo que pasa con IPv4. Esta tendencia se vió en todos los casos.

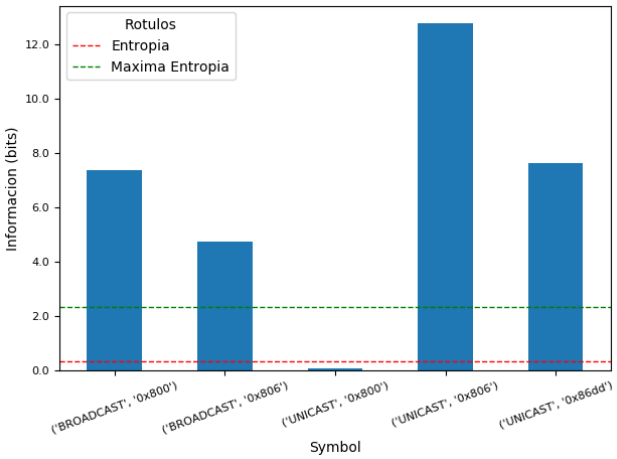

Figura 13: Información de los símbolos de la fuente S1, notando la entropía de la fuente, y la máxima entropía posible si la fuente fuera equiprobable.

Resultados fuente S2

Destaca completamente el símbolo 10.2.203.254, este parece ser un destino muy frecuente de los paquetes who-has.

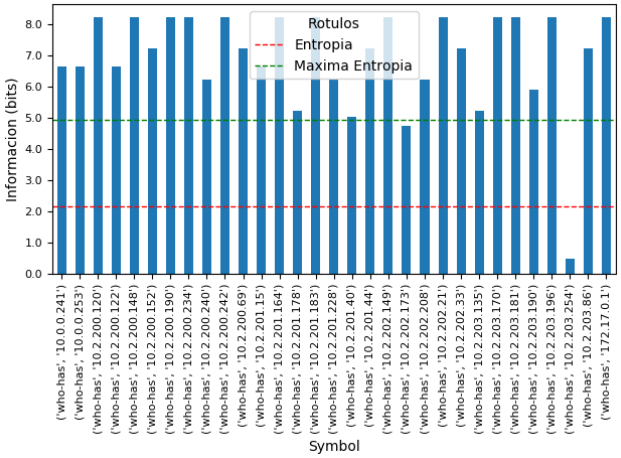

Figura 14: Información de los símbolos de la fuente S2, notando la entropía de la fuente, y la máxima entropía posible si la fuente fuera equiprobable.

En este grafo, el cual es mucho más grande que los anteriores, el nodo 10.2.203.254 es el de mayor grado corroborando lo visto en el gráfico de entropía.

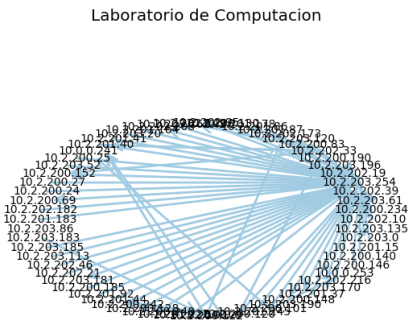

Figura 15: Grafo resultante de la red wifi del laboratorio de computación de la facultad a las 18 PM de un miércoles.

De nuevo volvemos a detectar al router gracias a nuestra técnica de buscar al símbolo de menor entropía, esta parece funcionar muy bien en redes de gran tamaño.

IV. Discusión

Como resultado final pudimos identificar símbolos destacados efectivamente en redes de tamaño mediano y grande, y con moderado éxito en redes pequeñas. Creemos que el problema de la red pequeña podría contrarrestarse tomando mediciones durante una mayor cantidad de tiempo aunque si la red es suficientemente chica (por ejemplo un

solo host y un solo router) probablemente se vuelva imposible diferenciar a un host de un router debido a la escasez de paquetes ARP.

Es importante destacar a parte de nuestro método estadístico a la utilización del grafo subyacente para identificar satisfactoriamente gateway a partir del nodo de mayor grado.

Cabe destacar finalmente que si bien los resultados fueron satisfactorios estos están condicionados al momento y lugar en que tomamos las muestras, de todas maneras por lo dicho anteriormente creemos que si la red es suficientemente grande nuestras técnicas deberían adaptarse y funcionar correctamente en ámbitos bastante diversos.

Las mediciones hechas podrían haber terminado en resultados distintos si se hubieran tomado en horarios con menor actividad, lo cual podría haber resultado en menor cantidad de paquetes y por lo tanto resultados más pobres. Por esta razón intentamos realizar las mediciones en horarios en que estas redes son normalmente utilizadas, o sea en el momento en que se puede ver su comportamiento general.