TP1: Wiretapping

Teoría de las Comunicaciones

Departamento de Computación
FCEN - UBA

06.09.2017

1. Introducción

El objetivo de este trabajo es utilizar técnicas provistas por la teoría de la información para estudiar diversos aspectos de una red de manera analítica. Además, sugerimos el uso de algunas herramientas de manipulación y análisis de paquetes frecuentemente usadas en el dominio de las redes de computadoras: Wireshark [1], Tcpdump [2], WinDump [3] y Scapy [4].

2. Normativa

- Fecha de entrega: 09-10-2017.
- El informe debe tener como máximo 3500 palabras.
- El trabajo práctico se deberá enviar por correo electrónico con el siguiente formato:
  to: tdc-doc at dc uba ar
  subject: debe tener el prefijo [tdc-wiretapping] y contener número de grupo
  body: nombres de los integrantes y las respectivas direcciones de correo electrónico
  attachments: el informe en formato pdf + el código fuente en formato zip.
- No esperar confirmación a menos que reciban una respuesta indicando explícitamente que el mail fue rechazado. Notar que los avisos por exceso de tamaño no son rechazos.

3. Enunciado

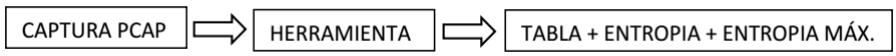
3.1. Introducción

Sean $p_1..p_n$ las tramas de capa 2 que se capturan en una red local. Se pueden modelar las tramas capturadas como una fuente de información de memoria nula S1 = $\{s_1, s_2, ..., s_q\}$, donde cada s_i está formado por la combinación entre el tipo de destino de la trama (unicast o broadcast) y el protocolo de la capa inmediata superior encapsulado en la misma. Por ejemplo, $s_i =< broadcast, ARP >$.

3.2. Primera consigna: capturando y modelando tráfico

Realizar n capturas de tráfico en formato pcap [5] (utilizando Scapy, Wireshark, Tcpdump o WinDump, a elección). **n está definido como la cantidad de miembros tenga el grupo.** Las capturas deben ser lo más extensas posibles (>10.000 tramas) y cada una de ellas en una red distinta. Al menos una de las capturas debe ser sobre una red mediana/grande no controlada (laboratorios, red de trabajo, shopping, etc.). Además, debe haber capturas tanto desde enlaces cableados como inalámbricos.

1. Implementar una herramienta que tome como entrada una captura en formato pcap y muestre representativamente la fuente modelada S1. Sugerimos que usen Scapy. La salida debe consistir en una tabla que muestre la probabilidad e información de cada símbolo, la entropía muestral de la fuente y su entropía máxima.



2. Proponga un modelo de fuente de información de memoria nula S2 con el objeto de $distinguir$ los hosts de cada red. La distinción de S2 debe estar basada únicamente en las direcciones IP dentro de los paquetes ARP [6]. El criterio para el modelado lo deberá establecer cada grupo utilizando las herramientas teóricas provistas por la teoría de la información. Se puede pensar que un símbolo es $distinguido$ cuando sobresale del resto en términos de la información que provee.

3. Adapte o extienda la herramienta anterior para que funcione con la nueva fuente.

La funcionalidad de ambas herramientas será testeada por los docentes utilizando un archivo pcap que produce una salida conocida por ellos. Los entregables deben venir acompañados por instrucciones precisas para (eventualmente compilar y) ejecutar las herramientas.

3.3. Segunda consigna: gráficos y análisis

Utilizando estas herramientas realizar experimentos para analizar los símbolos $distinguidos$ en cada una de las fuentes. Los análisis deben estar basados en conceptos formales de la teoría de la información. Específicamente, se debe analizar qué símbolos son significativos en cada red, observando la diferencia entre su información y la entropía de la fuente.

El informe debe seguir la siguiente estructura [7]: resumen; introducción; métodos, condiciones y resultados de cada experimento; resultados globales y conclusiones. Entre los métodos y condiciones de cada experimento se debe detallar la descripción de la red -tipo, tamaño, modo de acceso, etc.-, las características de la muestra -tamaño, horario, día de la semana, etc.- y la justificación de la elección del modelo de la fuente, entre otras.

La presentación de los resultados debe efectuarse **para cada red** mediante, al menos, los gráficos sugeridos a continuación:

1. Para S1: Mostrar la cantidad de infomación de cada símbolo comparando con la entropía de la fuente y la entropía máxima. Mostrar el porcentaje de tráfico broadcast sobre el tráfico total. Mostrar el porcentaje de aparición de cada protocolo encontrado.

2. Para S2: Mostrar la cantidad de información de cada símbolo comparando con la entropía de la fuente y la entropía máxima. Dados los paquetes ARP, mostrar mediante un grafo, la red de mensajes ARP subyacente *(de ser necesario, agrupar adecuadamente varios nodos en uno para mejorar la visualización).*

A su vez los resultados por experimento deben responder **para cada red**, algunas de las preguntas descriptas a continuación. *(No es necesario contestarlas todas, ni transcribirlas en el informe. Se valorará significativamente el planteo de nuevas preguntas)*:

1. Para S1: ¿Considera significativa la cantidad de tráfico broadcast sobre el tráfico total? ¿Cuál es la función de cada uno de los protocolos encontrados? ¿Cuáles son de control y cuáles transportan datos de usuario? ¿Se han encontrado símbolos $distinguidos$? ¿Les otorga algún significado? ¿La entropía de la fuente es máxima? ¿Bajo qué condiciones la entropía sería máxima? ¿Ha encontrado protocolos no esperados? ¿Puede describirlos?

2. Para S2: ¿La entropía de la fuente es máxima? ¿Qué sugiere esto acerca de la red? ¿Bajo qué condiciones la entropía sería máxima? ¿Se pueden $distinguir$ nodos? ¿Se les puede adjudicar alguna función específica? ¿Hay evidencia parcial que sugiera que algún nodo funciona de forma anómala y/o no esperada? ¿Existe una correspondencia entre lo que se conoce de la red y los nodos distinguidos detectados por la herramienta? ¿Ha encontrado paquetes ARP no esperados? ¿Cuál es su función?

A continuación, se sugieren algunas preguntas para responder a la hora de realizar un análisis global en las conclusiones. (*No es necesario contestarlas todas, ni transcribirlas en el informe. Se valorará significativamente el planteo de nuevas preguntas*):

1. De haber diferentes tamaños de redes, ¿Aprecia alguna diferencia desde el punto de vista de las fuentes de información analizadas?
2. ¿Ha notado alguna diferencia durante la captura de datos entre el acceso a la red mediante WiFi y el acceso mediante cable? ¿A qué se lo atribuye?
3. ¿Considera que las muestras obtenidas analizadas son representativas del comportamiento general de la red?
4. El modelo de fuente de memoria nula utilizado supone que las probabilidades de los símbolos son independientes. ¿Es esto verdadero para ambas fuentes? ¿Por qué? ¿Qué consecuencia tiene esto en los experimentos realizados?
5. ¿Qué otra herramienta matemática distinta a la teoría de la información podría utilizar para detectar elementos *distinguidos* en una muestra de datos?

Referencias

[1] Wireshark `http://www.wireshark.org`

[2] Tcpdump `http://www.tcpdump.org`

[3] WinDump `https://www.winpcap.org/windump`

[4] Scapy `http://www.secdev.org/projects/scapy`

[5] Ejemplos de capturas en formato pcap `http://tcpreplay.appneta.com/wiki/captures.html`

[6] RFC 826 (ARP) `http://tools.ietf.org/html/rfc826`

[7] Formatting IEEE Papers `http://mocha-java.uccs.edu/ieee/`