TP1: Wiretapping

Teoría de las Comunicaciones
Departamento de Computación
FCEN - UBA

24.08.2016

1. Introducción

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

2. Normativa

- Fecha de entrega: 19-09-2016.
- El informe deberá haber sido enviado por correo para esa fecha con el siguiente formato:
  to: tdc-doc at dc uba ar
  subject: debe tener el prefijo [tdc-wiretapping] y contener en numero 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. Es indispensable para la aprobación del trabajo práctico que junto con el código fuente haya un ejecutable (puede ser un simple script, un makefile, etc.) que ejecute automáticamente todos los experimentos y los eventuales gráficos de forma automática y un archivo de texto con indicaciones para su uso. El ejecutable debe recibir, como mínimo, un parámetro que determine el archivo de capturas de una red en formato `libpcap`. **No** deben entregar el archivo de capturas de sus experimentos, los docentes evaluamos el código con nuestro propio conjunto de datos.
- No esperar confirmación. Todos los mails llegan a la lista a menos que reciban una respuesta indicando explícitamente que el mail fue rechazado. Los avisos por exceso de tamaño no son rechazos.

3. Enunciado

3.1. Introducción

Sean $p_1..p_i$ los paquetes que se capturan en un enlace. Podemos conocer los destinos a los que están apuntados los paquetes que encapsulan la información con el campo dst del frame de capa de enlace (p_i.dst en Scapy). Se pueden modelar los paquetes capturados como una fuente de información binaria de memoria nula S, definiendo el conjunto de símbolos que emite como $\{s_{BROADCAST}, s_{UNICAST}\}$. Si se captura un paquete p en un intervalo de tiempo $[t_i, t_f]$, se dice que S emite $s_{BROADCAST}$ si $p.dst == \texttt{ff:ff:ff:ff:ff:ff}$, sino emite $s_{UNICAST}$. Esta fuente distingue entre mensajes broadcast y unicast que aparecen en la red en ese intervalo.

3.2. Primera consigna: capturando tráfico

1. Implementar una herramienta que escuche pasivamente los paquetes Ethernet de la red y muestre representativamente la fuente modelada S. Además, la herramienta debe proveer la opción de tomar un archivo en formato `libpcap`, en lugar de realizar la captura.

2. Adapte la herramienta anterior proponiendo un modelo de fuente de información de memoria nula $S1$ con el objetivo de distinguir, en lugar de tipos de destinos como hace S, los nodos (hosts) de la red. La distinción de S1 debe estar basada únicamente en las direcciones IP de paquetes ARP. 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. Esta distinción no debe ser arbitraria y debe estar sustentada matemáticamente.

3.3. Segunda consigna: gráficos y análisis

Utilizando estas herramientas, realizar experimentos para analizar: i) Los paquetes broadcast de la red. ii) Los nodos distinguidos. Se deben realizar tantos experimentos como cantidad de miembros tenga el grupo. Además, las capturas deben ser lo más extensas posibles ($tf - ti > 10minutos$). En la medida de lo posible, intentar capturar en al menos una red mediana/grande que no sea controlada (trabajo, shopping, etc).

Los análisis deben estar basados en conceptos formales de la teoría de la información. O sea, se debe analizar qué símbolos son significativos en cada red, viendo la diferencia entre su información y la entropía de la fuente.

El informe debe seguir la siguiente estructura: somera introducción, métodos y condiciones de cada experimento (acá debe estar justificada la elección de modelo de fuente $S1$), resultados y conclusión. La presentación de los resultados debe efectuarse **para cada red** mediante, al menos, los gráficos sugeridos a continuación:

1. Dada la fuente binaria S, mostrar la cantidad de infomación de cada símbolo comparando con la entropía de la fuente y la entropía máxima.

2. Dados los paquetes ARP, muestre mediante un grafo, la red de mensajes ARP subyacente *(de ser necesario, agrupe adecuadamente varios nodos en uno para mejorar la visualización)*.

3. Dada la fuente $S1$, mostrar la cantidad de información de cada símbolo comparando con la entropía de la fuente.

A su vez los resultados deben responder, **para cada red**, las preguntas descriptas a continuación *(no hace falta transcribir las preguntas en el informe)*:

1. ¿La entropía de la fuente S es máxima? ¿Que sugiere esto acerca de la red? ¿Está relacionado con el overhead impuesto por la red debido a los *protocolos de control* (i.e.: ARP)?

2. ¿Cómo es el tráfico ARP en la red? ¿Se pueden distinguir nodos? ¿Cuántos? ¿Indica algo la cantidad? ¿Se les puede adjudicar alguna función específica? ¿Hay evidencia parcial que sugiera que algún nodo funciona de forma anómala y/o no esperada?

3. ¿Existe una correspondencia entre lo que se conoce de la red y los nodos distinguidos detectados por la herramienta? ¿Es posible usar el criterio de distinción propuesto como método para descubrir el/los Default Gateway/s de la red? ¿Es preciso?

A continuación, se sugieren preguntas para responder a la hora de realizar un análisis global, en la conclusión (*pueden, y se valorará significativamente, plantearse nuevas preguntas*):

1. De haber diferentes tecnologías entre las redes capturadas, ¿Aprecia alguna diferencia desde el punto de vista de las fuentes de información analizadas?

2. De haber diferentes tamaños de redes, ¿Aprecia alguna diferencia desde el punto de vista de las fuentes de información analizadas?

3. ¿Que importancia tiene la entropía en la capacidad de detectar símbolos (nodos/hosts) distinguidos?

4. ¿Hay alguna relación entre la entropía y la cantidad de nodos (distinguidos y no distinguidos)?

Referencias

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

[2] Wireshark (página web oficial) http://www.wireshark.org

[3] Scapy (página web oficial) http://www.secdev.org/projects/scapy/

[4] OUI (IEEE) http://standards.ieee.org/develop/regauth/oui/oui.txt