Análisis de las medidas de distancia entre sesiones para la clasificación de intrusos

Sebastián García

Si6 - CITEFA[**]
Instituto de Investigaciones Científicas y Técnicas de las Fuerzas Armadas. Argentina
sgarcia@citefa.gov.ar

Resumen Este paper es una investigación en progreso enfocada al análisis de medidas de distancia entre sesiones de diversos intrusos, para la clasificación de los mismos mediante el estudio de su comportamiento. Se utilizan datos de intrusos reales, capturados durante los años 2005, 2006 y 2007 en diversos honeypots. Se estudia el comportamiento en referencia a las acciones del intruso, a su manera de utilizar el sistema y a sus intenciones. Los datos base corresponden a las teclas presionadas capturadas por un keylogger y no a datos de red como la dirección IP origen. El objetivo de este trabajo es lograr identificar y analizar las mejores medidas de distancia entre sesiones. En una primera instancia, las medidas de distancia seleccionadas permiten una clasificación satisfactoria.

Palabras clave: Detección de intrusos, Identificación, Clasificación, Honeypots

1. Introducción

Este trabajo es parte de un proyecto más amplio del Laboratorio de Investigación en Seguridad Informática Si6[1] de CITEFA[2] qué comenzó en el año 2004. El objetivo principal del proyecto del Si6 es el desarrollo de un sistema de identificación de intrusos basado en el comportamiento de los mismos, por medio del análisis de parámetros obtenidos dentro del honeypot (patrones de keystroke dynamics, comandos utilizados, lenguaje del intruso, errores, etc.) y de parámetros obtenidos en la red (dirección IP, routers, sistema operativo, etc.)

El proyecto se enfoca en los comportamientos y no en las motivaciones o acciones.

Los dos conceptos más importantes que caracterizan esta investigación del Si6 son:

- Análisis de intrusos: Trabajar con la persona que está realizando el ataque, no con intrusiones (acciones o ataques)

[**] Juan Bautista de la Salle 4397 (B1603ALO) Villa Martelli, (5411)-4709-8100, Fax:(5411)-4709-8228

[1] http://www.citefa.gov.ar/SitioSI6_EN/si6.htm

[2] http://www.citefa.gov.ar

- Trabajar con la intencionalidad: No trabajar con usuarios comunes para tomar los datos. La *intencionalidad* del intruso permite generar patrones de comportamiento clasificables.

Dentro de este proyecto principal, el presente trabajo se enfoca en el análisis de medidas de distancia para la clasificación de intrusos analizando sus patrones de comportamiento.

2. Antecedentes

Existe una gran cantidad de trabajos enfocados a la *clasificación de intrusiones* orientados a las acciones. Joseph S. Sherif [1] realizó una extensa recopilación de trabajos en esta área. Pero ninguno está orientado a la clasificación de intrusos.

Kent E. Anderson [2] abordó la clasificación de intrusos por sus motivaciones, pero no en el foco exacto del presente trabajo.

También existen varios trabajos de Roy A. Maxion [3] enfocados en la clasificación de usuarios de sistemas tipeando comandos. Pero este tampoco es el foco preciso del presente trabajo, ya que los usuarios no tienen *intencionalidad*.

Estos antecedentes permitieron crear el marco téorico donde proponer un nuevo enfoque. El Si6 publicó en este área [4], [5] y [6].

3. Problema

Los problemas a resolver para lograr un análisis de las medidas de distancia para la clasificación de sesiones incluyen:

1. Procesamiento de los datos en crudo para extraer las características básicas de las sesiones.
2. Creación de las medidas de distancia apropiadas entre características.
3. Análisis de la aplicación de las medidas de distancia en las sesiones.

4. Propuesta

- Creación de una herramienta que permita analizar de forma automática las teclas presionadas por los intrusos y obtener las características relevantes de las sesiones.
- Preprocesamiento automatizado de los datos de los tres años para obtener las sesiones y sus características principales.
- Análisis de las características más prometedoras para definir qué medidas de distancia son las más efectivas.
- Proposición de diferentes medidas de distancia en base a los datos reales.
- Creación de una herramienta que implemente las medidas de distancia para su análisis y comparación.
- Generación y análisis de resultados

5. Herramienta para el análisis automatizado de teclas: Tcleoanalyzer

Tcleoanalyzer es la herramienta desarrollada para esta investigación que extrae las características de las sesiones. Publicada bajo licencia GPL [3] permite un análisis de las teclas y sesiones de los intrusos.

Las características más importantes de la herramienta son:

- Análisis de paquetes online desde una placa de red o desde archivos pcap.
- Reconocimiento de archivos *pcap* truncados, permitiendo su correcto análisis.
- Diversos modos de salida de datos incluyendo características de las sesiones, comandos, fechas y horas, tiempos de ejecución, modo para parsers de datos, salida para analizar con keystrokes dynamics, etc.
- Envío de mails con los comandos de los intrusos
- Sonido de alerta cuando ingresa un intruso
- Reconocimiento de carácteres especiales y su significado en sesiones. Por ej. el carácter $< SUB >$ que aparece cuando se apreta CTRL-Z.
- Trabajo con las sesiones:
  - Reconocimiento de sesiones nuevas por no haber visto nunca ese usuario en esa terminal y por tiempo de inactividad configurable
  - Reconocimiento del intento de utilización de los programas ftp, lynx, mutt, less, vi y man. El método es la búsqueda del carácter $' < NULL >'$
  - Reconocimiento de las formas de salir de los programas ftp, lynx, mutt, less, vi y man. Incluyendo CTRL-Z y CTRL-C.
  - Se guardan las teclas presionadas dentro de cada programa (ftp, lynx, mutt, less, vi y man)
  - Reconocimiento del escape de carácteres en comandos comunes y en los programas (ftp, lynx, mutt, less, vi, man)
  - Se reconocen las teclas especiales del teclado.
  - Se realiza manejo de history dentro de la sesión, siguiendo los mismos pasos del intruso.
- Información procesada de las sesiones que se obtienen:
  - Fecha y hora en milisegundos de comienzo de la sesión
  - Fecha y hora en milisegundos de finalización de la sesión
  - Duración en días, horas y segundos de la sesión
  - Cantidad de comandos en total de la sesión
  - Número de terminal utilizada para acceder a la sesión
  - UID del usuario utilizado para ingresar a la sesión
  - Primeros tres comandos de la sesión
  - Últimos tres comandos de la sesión
  - Los 5 comandos más utilizados y la cantidad de veces de cada uno
  - Los 5 directorios más utilizados y la cantidad de veces de cada uno
  - Los 5 sitios más accedidos (métodos utilizados y cantidad de veces)
  - Los modificadores más usados en los comandos más comunes (ls, rm, ps, lynx, ftp, wget, ssh)

[3] `http://www.gnu.org/licenses/old-licenses/gpl-2.0.txt`

- Los 5 comandos no básicos más comunes. No ls, rm, ps, lynx, ftp, wget, ssh, w, cat, mkdir, man , cd, less ni vi
- Las características especiales de la sesión: Si usa las flechas para el history, si usa Tab, si apretó la tecla Insert o la tecla Delete.

Ejemplo de salida del tcleoanalyzer

```
Session ID = 7
        Session Start time = Sun Apr 3 07:49:50.045222 2005
        Session End Time   = Sun Apr 3 07:52:03.874202 2005
        Session duration = 0 days, 0 hours, 2 minutes, 13 seconds (133 total seconds)
        Session Total Commands = 21
        Session Terminal = pts/0
        Session UID = 0
        Session first 3 Commands:
                Command: 'w'
                Command: 'ps ax'
                Command: 'cd /tmp'
        Session last 3 Commands:
                Command: 'ls -la'
                Command: 'wget gate.polarhome.com/~ircs/s1.tar'
                Command: 'kill -9 $$'
        Top five complete Commands:
                Command: 'ls -la' (5 times)
                Command: 'cd /tmp' (2 times)
                Command: 'cd /usr/lib/games' (1 times)
                Command: 'w' (1 times)
                Command: 'ps ax' (1 times)
        Top five programs:
                Command: 'cd' (6 times)
                Command: 'ls' (5 times)
                Command: 'locate' (2 times)
                Command: 'cat' (1 times)
                Command: 'wget' (1 times)
        Top five directories:
                Directory: '/tmp' (2 times)
                Directory: '".  "' (1 times)
                Directory: '/usr/lib/games' (1 times)
                Directory: '/usr/local/games' (1 times)
                Directory: '/usr/games' (1 times)
        Top five sites accessed:
                Site: 'gate.polarhome.com' (1 times) (wget)
        Most used parameters in common commands:
                Command: ls , Parameters: '-la'
        Top five uncommon programs (5 total):
                Command: 'locate' (2 times)
                Command: '/usr/sbin/traceroute' (1 times)
                Command: 'kill' (1 times)
                Command: 'ping' (1 times)
                Command: 'traceroute' (1 times)
        Session special characteristics:
                Characteristic: Movement, Value: Use commands history with arrows
```

6. Preprocesamiento de los datos

La característica principal de los datos utilizados radica en la definición del término Intruso y del termino Honeypot. Se define Intruso como *'(...)un individuo que ilegalmente accede o utiliza un sistema sin autorización'* [2]. Y se define Honeypot[4] como un sistema que *'no debería recibir ningún tráfico legal o activi-*

[4] `http://en.wikipedia.org/wiki/Honeypot_(computing)`

dad'. Todo el tráfico capturado se considera un ataque, independientemente del tipo de tráfico.

Los datos se extrajeron de archivos en formato *pcap*[5] conteniendo las capturas de paquetes de red de un honeypot. Estas se realizaron sin interrupción durante los años 2005 al 2007, con un total de 79 GB de capturas (28GB en 2005, 35GB en 2006 y 16GB hasta Julio del 2007).

El honeypot linux utilizado, tiene instalado el keylogger *tcleo*[6] que envía a la red, en un paquete UDP los siguientes datos: tecla presionada, terminal utilizada por el intruso, el User ID y la fecha en que la tecla fue presionada. Una descripción detallada de las técnicas implementadas en esta herramienta se puede encontrar en [7]. Detalles de la implementación de la topología se encuentran en [5].

La herramienta *tcleoanalyzer* analiza las teclas enviadas por *tcleo* para obtener información sobre las acciones del intruso.

En los tres años hubo aproximadamente 273 sesiones, un promedio de 1 intruso ingresando cada 4 días.

Las sesiones de los tres años representan la base de datos para este análisis de medidas de distancia.

El mayor inconveniente de trabajar con datos de intrusos reales es que no es posible verificar los resultados de la clasificación con datos comprobados o *ground-truth data*. Es necesario diseñar formas alternativas de verificación de resultados. Las medidas de verificación más utilizadas son: comprobación por medio de la experiencia de diferentes expertos en la clasificación de intrusos y la comprobación de coherencia con los mismos datos.

7. Medidas de distancia

Parte fundamental de las técnicas de clasificación, es la definición de las medidas de distancias entre los datos. Se necesita un análisis puntual y específico para encontrar la mejor manera de comparar las características de las sesiones.

En [8] 1.3.1, `Feature Extraction` Duda R.O. habla sobre la importancia de conocer el dominio del problema para lograr una extracción de características adecuada. También en [8] 1.3.5, `Prior Knowledge`, Duda R.O. aclara cómo el conocimiento previo del problema ayuda en el diseño del clasificador sugiriendo nuevas características.

En el caso del laboratorio Si6, fueron necesarios varios meses de estudio de los datos para llegar a obtener las características más prometedoras para la clasificación.

Se definieron 8 medidas de distancia para 8 características diferentes.

Las características a extraer y comparar entre las sesiones son:

[5] `http://en.wikipedia.org/wiki/Pcap`
[6] `http://www.citefa.gov.ar/SitioSI6_EN/tcleo.htm`

Medida de distancia
Separación temporal entre sesiones
Comparación de directorios utilizados
Comparación de los User ID
Comparación de los sitios accedidos
Comparación de los primeros tres comandos utilizados
Comparación de los ultimos tres comandos utilizados
Comparación de los cinco comandos más utilizados
Comparación de los cinco comandos menos comunes

7.1. Algoritmos de comparación

Algoritmo de comparación de todas las sesiones

1. Calcular las características para cada sesión
2. Calcular la distancia que cada sesión tiene con todas y cada una del resto de las sesiones. Se arma una matriz.
3. Analizar las distancias para reconocer diferentes grupos en las sesiones

Algoritmo de comparación entre dos sesiones

1. Calcular las distancias entre cada una de las características de las dos sesiones, armando un vector de distancias. Cada distancia varía entre 0 y 100. Siendo 0 la distancia más cercana.
2. Asignar a cada característica un valor de importancia, para poder priorizar algunas características sobre otras. Se arma un vector de pesos.
   Por ejemplo, se comprueba que la medida de distancia entre directorios utilizados por el intruso es más significativa que la distancia entre los User ID (Identificadores de Usuario al ingresar al sistema).
   Los valores del vector de pesos varían entre 0 y 1.
3. Se multiplica el vector de distancias por el vector de pesos, logrando un vector de distancias ponderadas.
4. Se calcula la distancia euclidiana de este vector:
   a) A cada valor del vector ponderado se lo eleva al cuadrado
   b) Se realiza la sumatoria de cada valor del vector resultante.
   c) Se calcula la raíz cuadrada del valor resultante

$$DistanciaEuclidiana = \sqrt{\Sigma((distanciaponderada_n)^2)}$$

5. Se calcula el rango de valores en el que oscilan las distancias para este vector de pesos. Obteniendo un `ValorMaximo` Este `ValorMaximo` siempre supera los 100.
6. Como para la comparación de resultados entre dos sesiones es conveniente que las distancias oscilen entre 0 y 100, el valor final de la comparación de dos sesiones es el **porcentaje** dentro de ese rango $[0 - ValorMaximo]$. Llamaremos a este proceso normalización.

Por ejemplo, si el vector de las distancias de las ocho características definidas anteriormente entre dos sesiones es $[0, 20, 100, 0, 10, 30, 40, 30]$ y el vector de pesos es [0,8, 0,8, 0,3, 0,6, 0,6, 0,4, 0,4, 0,4], las distancias ponderadas son $[0, 16, 30, 0, 6, 12, 16, 12]$ y la distancia euclideana es 41,66. Luego, la normalización calculada es igual a 25,99, que es la distancia final entre estas dos sesiones.

7.2. Herramienta para analizar automáticamente las distancias: Tcleoclusterer

Tcleoclusterer es la herramienta desarrollada para esta investigación que analiza las sesiones encontradas por el tcleoanalyzer para evaluar las distintas medidas de distancias propuestas. También está publicada bajo licencia GPL.

Las características más importantes de la herramienta son:

- Cálculo de la totalidad de las medidas de distancia de la matriz
- Impresión a color de la matriz, identificando cantidad de conexiones de cada sesión.
- Impresión de los detalles de cada sesión y sus conexiones.
- Reconocimiento de las distancias menores a 25, a 50 y a 75.
- Cálculo de la sesión que más se relaciona con cada sesión.
- Verificación de resultados mediante la selección de un grupo de sesiones.
- Automatización de la verificación de múltiples vectores de pesos simultaneamente.
- Cálculo del nivel de relación de cada sesión con el grupo verificado.

Ejemplo de salida del tcleoanalyzer para la sesión 0, analizando 100 sesiones

```
(...)
Session 0:
S0=0.00    S1=36.85   S2=99.55   S3=89.63   S4=89.63   S5=99.41   S6=74.07   S7=99.41
S8=99.41   S9=68.69   S10=76.76  S11=70.16  S12=96.79  S13=59.89  S14=100.00 S15=83.26
S16=68.69  S17=97.36  S18=99.41  S19=71.31  S20=60.27  S21=70.52  S22=99.41  S23=100.00
S24=99.41  S25=74.07  S26=83.26  S27=97.36  S28=99.41  S29=99.41  S30=74.07  S31=83.26
S32=80.81  S33=96.76  S34=74.07  S35=74.07  S36=80.81  S37=99.41  S38=83.97  S39=99.41
S40=48.69  S41=91.85  S42=99.41  S43=99.41  S44=83.26  S45=74.07  S46=99.41  S47=99.41
S48=65.67  S49=57.61  S50=66.70  S51=80.81  S52=59.01  S53=99.41  S54=100.00  S55=74.86
S56=80.07  S57=83.26  S58=74.07  S59=100.00  S60=59.01  S61=83.26  S62=79.37  S63=68.69
S64=74.07  S65=58.01  S66=70.47  S67=62.16  S68=83.26  S69=74.86  S70=71.31  S71=61.88
S72=93.95  S73=80.07  S74=83.26  S75=74.86  S76=74.63  S77=92.49  S78=76.76  S79=80.81
S80=60.27  S81=83.26  S82=83.26  S83=83.26  S84=83.26  S85=83.26  S86=99.41  S87=99.41
S88=100.00  S89=99.41  S90=83.26  S91=83.97  S92=83.26  S93=80.07  S94=83.26  S95=54.80
S96=80.07  S97=79.37  S98=80.07  S99=83.26  S100=83.26  S101=100.00
Min: 36.85 (1) [<25=1, <50=2, <75=32]
(...)

Session 2: (INSIDE)
Connections outside the filtered group: <25 = 0,<50 = 0,<75 = 41

Session 12: (OUTSIDE)
Connections to the filtered group: <25 = 0,<50 = 0,<75 = 1

(...)
Connection Statics: (Distance level = 75)
Weight array: [ 0.4  0.2  0.7  1.   0.6  0.6  0.7  0.7]
 (temporal | uid | 5 dir | 5 sites | first 3 | last 3 | 5 progs | 5 uncommon)
Total connection ratio (bonding): 0.13% [97 74034]
Total outside connections linking (>0 and <75) to the group: 46 from 73983 (0.06%)
```

```
Total inside connections linking (>0 and <75) outside the group: 47 from 51 (92.16%)
Total inside connections linking (>0 and <75) to themselves: 4 from 51 (7.84%)
Ratios between percentages : 0.00 and 0.09
```

8. Detalles de los cálculos de las medidas de distancia

8.1. Separación temporal entre sesiones

Dos sesiones que están próximas en el tiempo tienen más probabilidad de pertenecer al mismo intruso que dos sesiones distanciadas. Esta observación se basa en la experiencia de ver a intrusos crear usuarios del sistema y conectarse con los nuevos usuarios, o cambiar el password del usuario root y volver a conectarse como el usuario root.

Se encontraron una gran cantidad de sesiones solapadas, con una distancia entre sesiones menor a cero y una gran cantidad de sesiones con una distancia menor a 15 minutos respecto a la anterior.

8.2. Distancia en la comparación general de directorios

Las medidas de distancia entre directorios es la más prometedora según los análisis manuales. Se comparan los strings de los directorios que utilizaron los intrusos en sus sesiones. Por ejemplo `/var/db/en/dist/`

La lógica de este método es acercar a la distancia 0 dos directorios que sean iguales a un nivel de profundidad (cantidad de subdirectorios utilizados) igual a cinco. De la misma manera, la máxima distancia se encuentra en dos directorios que son diferentes con un nivel de profundidad igual a cinco.

Los valores que se restan y suman están basados en la experiencia y fueron asignados siguiendo la siguiente lógica de comparación:

- Si comparten subdirectorios, se parecen más y por lo tanto resta en la medida de distancia, acercando el valor a 0.
- Si no comparten todos los subdirectorios, se parecen menos y suma en la medida de distancia.
- El hecho de que compartan directorios comúnes como el '/', hace que se parezcan, pero no tanto como si comparten 3 subdirectorios dentro del directorio '/usr'. Por lo tanto estos últimos casos deben ser más cercanos a la medida cero.

Ejemplos de algunos casos:

- El peor caso (los directorios no tienen similitud aparente). La distancia entre `/usr/tmp/log/cache/` y `/var/db/en/dist/` es 100.
- El mejor caso (los 5 subdirectorios compartidos son iguales). La distancia entre `/var/tmp/db/test/` y `/var/tmp/db/test/` es 0. Notar que la distancia entre los directorios `/var` y `/var` no es 0 porque es más común que utilicen `/var` que `/var/tmp/db/test/`.

8.3. Distancia en la comparación de UID

Los UID son una medida de distancia con poca variación, pero se comprobó su utiliad bajo ciertas condiciones. El UID=0 es el correspondiente al usuario 'root' o administrador del sistema y es comunmente el más utilizado. La lógica de comparación es: Si los UID son diferentes, la distancia es 100, si son iguales, la distancia es 0.

Esta técnica se incluyó por la cantidad de UID diferentes que se utilizaron en los tres años.

8.4. Distancia en la comparación de sitios FQDN accedidos

Dos sitios FQDN[7] (nombres de dominio en un formato legible por humanos) se comparan según la siguiente lógica:

1. Se separan los nombres de los sitios en sus componentes del dominio. Por ej. `boogdan.home.ro` se separa en 'máquina'=`boogdan`, 'organización'=`home` y 'TLD'=`ro`. TLD es el acrónimo para 'Top Level Domain' o dominio de nivel superior.
2. Se comparan cada componente con su par del otro sitio, no se realizan comparaciones entre componenetes diferentes.
3. Cada componente puede ser igual al componente del otro sitio, o diferente. No existen valores intermedios.

Los valores de distancia para esta característica fueron calculados en base a la utilización de sitios por los intrusos.

Esta medida de distancia puede ser mejorada incorporando dos componentes más que hoy forman parte de la sesión y están incluidos en la salida del `tcleoanalyzer`: la cantidad de veces que un sitio fue usado en la sesión y el método de acceso. Estos dos datos permiten mejorar la precisión en la utilización de sitios. Un ejemplo de salida del `tcleoanalyzer` es:

```
Site: 'boogdan.home.ro' (9 times) (ftp , lynx)
Site: 'ircopzip.ifreepages.com' (1 times) (ftp)
Site: 'khaos.home.ro' (1 times) (ftp)
Site: 'boogdan.dap.ro' (1 times) (ftp)
```

8.5. Distancia en la comparación de comandos

Se toman tres distancias diferentes: comparación de los primeros tres comandos de la sesión, los últimos tres comandos de la sesión, y los comandos poco comunes utilizados por la sesión.

En estas tres distancias, individualmente, la lógica es que cuantos más comandos compartan las dos sesiones que se comparan, más se parecerán. Como los tres comandos están ordenados por cantidad de veces que se usaron, se asigna más valor cuando comparten los comandos más usados.

[7] `http://es.wikipedia.org/wiki/FQDN`

9. Resultados

Antes de realizar pruebas sobre la identificación de grupos y usuarios, se realizaron algunas verificaciones para comprobar el algoritmo y los diferentes vectores de pesos.

La verificación de los resultados no se puede realizar convenientemente porque no existen datos verificados. Los datos con los que actualmente cuenta el proyecto Paranoid del Si6 son datos de intrusos reales.

Se consideran conexiones significativas entre dos sesiones aquellas cuya distancia oscila entre 0 y 75. Cuando la distancia es mayor a 75 se considera que las sesiones no tienen relación aparente.

9.1. Primer experimento de verificación de los resultados

La primera verificación se realizó diferenciando manualmente las sesiones que parecían provenir de intrusos **rumanos** dada la experiencia en detección de intrusiones del laboratorio. Se seleccionó este grupo conocido de intrusos porque sus sesiones son lo suficientemente identificables dada la característica de conectarse y descargar software de sitios rumanos (*.ro*).

La verificación consistió en:

- Identificar las sesiones rumanas en las 273 sesiones de los datos. Se identificaron 48 sesiones.
- Analizar con el tcleoclusterer la relación de estas 48 sesiones con el resto de las sesiones, aplicando los siguientes 10 vectores de pesos.

```
1ro :  1 ,  0  ,  0  , 0 ,  0  ,  0  ,  0  ,  0  , 0
2do :  0 ,  1  ,  0  , 0 ,  0  ,  0  ,  0  ,  0  , 0
3ro :  0 ,  0  ,  1  , 0 ,  0  ,  0  ,  0  ,  0  , 0
4to :  0 ,  0  ,  0  , 1 ,  0  ,  0  ,  0  ,  0  , 0
5to :  0 ,  0  ,  0  , 0 ,  1  ,  0  ,  0  ,  0  , 0
6to :  0 ,  0  ,  0  , 0 ,  0  ,  1  ,  0  ,  0  , 0
7mo :  0 ,  0  ,  0  , 0 ,  0  ,  0  ,  1  ,  0  , 0
8vo :  0 ,  0  ,  0  , 0 ,  0  ,  0  ,  0  ,  1  , 1
9no :  1 ,  1  ,  1  , 1 ,  1  ,  1  ,  1  ,  1  , 1
10mo:0.4 , 0.2 , 0.7 , 1 , 0.6 , 0.6 , 0.7 , 0.7
```

  Se eligieron los vectores de pesos para estudiar el comportamiento del algoritmo en las siguientes situaciones:
  - Priorizando cada una de las características sobre las demás
  - sin priorizar ninguna característica (el mismo valor para cada característica)
  - Vector de pesos 'real' considerado como el mejor por los analistas de seguridad.

- Resultado parcial del tcleoclusterer

```
Connection Statics:
Weight array: [ 0  0  0  1  0  0  0  0]
 (temporal | uid | 5 dir | 5 sites | first 3 | last 3 | 5 progs | 5 uncommon)
Total connection ratio (bonding): 3.45% [2574 74529]
Total outside connections linking (>0 and <75) to the group: 135 from 61425 (0.22%)
Total inside connections linking (>0 and <75) outside the group: 135 from 13104 (1.03%)
Total inside connections linking (>0 and <75) to themselves: 2304 from 13104 (17.58%)
Ratios between percentages : 0.21 and 0.06

Connection Statics:
Weight array: [ 0.4  0.2  0.7  1.   0.6  0.6  0.7  0.7]
 (temporal | uid | 5 dir | 5 sites | first 3 | last 3 | 5 progs | 5 uncommon)
Total connection ratio (bonding): 3.21% [2395 74529]
Total outside connections linking (>0 and <75) to the group: 265 from 61425 (0.43%)
Total inside connections linking (>0 and <75) outside the group: 289 from 13104 (2.21%)
Total inside connections linking (>0 and <75) to themselves: 1841 from 13104 (14.05%)
Ratios between percentages : 0.20 and 0.16
```

- Los resultados indican que:
  - El vector de pesos que mejor se comportó es el que priorizaba los sitios accedidos. Verificando que el algoritmo funciona correctamente ya que esta fue la característica seleccionada para separar las sesiones.
  - El vector de pesos 'real' fue el segundo mejor. Comprobando que es muy útil para separar sesiones. Incluso obtuvo un buen nivel de '*bonding*' o aglutinamiento de las sesiones. Lo que significa que mantuvo un alto grado de separación de sesiones del grupo al mismo tiempo que las sesiones en total se relacionaron en gran cantidad unas con otras.
  - En las 48 sesiones rumanas, solo el 2.21 % de las conexiones significativas se relacionaron con sesiones fuera del grupo.
  - De las sesiones fuera del grupo, solo el 0.43 % de las conexiones eran con sesiones dentro del grupo (solo 3 sesiones con más de 10 conexiones significativas con el grupo).

9.2. Segundo experimento de verificación de los resultados

Se extrajeron las sesiones número 24, 108 y 209 del grupo de sesiones rumanas, para verificar cómo actúa el algoritmo cuando hay sesiones con una conexión fuerte con el grupo.

Resultados filtrados del tcleoclusterer:

```
Connection Statics:
Weight : [ 0.4  0.2  0.7  1.   0.6  0.6  0.7  0.7]
 (temporal | uid | 5 dir | 5 sites | first 3 | last 3 | 5 progs | 5 uncommon)
Total connection ratio (bonding): 3.18% [2370 74529]
Total outside connections linking (>0 and <75) to the group: 366 from 62244 (0.59%)
Total inside connections linking (>0 and <75) outside the group: 391 from 12285 (3.18%)
Total inside connections linking (>0 and <75) to themselves: 1613 from 12285 (13.13%)
Ratios between percentages : 0.18 and 0.24
```

Podemos observar como se produjo un decaimiento en los porcentajes de los resultados. Un 13.13 % donde antes obtuvimos 14.05 % y 3.18 % donde antes obtuvimos 2.21 %. Esto marca la saparación de sesiones del grupo.

Para buscar qué sesiones fuera del grupo se relacionan de manera significativa con las sesiones dentro del grupo tomamos como umbral el 0,5 % del total de conexiones de las sesiones fuera del grupo con el grupo. En el ejemplo anterior, hay 366 conexiones de sesiones fuera del grupo hacia el grupo, lo que establece el umbral en 19 conexiones. Si una sesión tienen más de 19 conexiones con el grupo, se considera parte del grupo.

En el análisis detallado de la salida del tcleoclusterer, vemos que solo las sesiones que hemos extraido tienen una cantidad de conexiones que supera el umbral propuesto de 0.5 % :

```
Session 24: (OUTSIDE)
Conections to the filtered group: <25 = 0,<50 = 1,<75 = 34

Session 108: (OUTSIDE)
Conections to the filtered group: <25 = 0,<50 = 2,<75 = 39

Session 209: (OUTSIDE)
Conections to the filtered group: <25 = 0,<50 = 0,<75 = 33
```

Lo cual confirma que el algoritmo encuentra satisfactoriamente las sesiones (y solo las sesiones) que hemos extraido del grupo.

9.3. Experimento de agrupado de sesiones

El objetivo del experimento es comenzar por una sola sesión y progresivamente ver como se comporta el algoritmo al introducir nuevas sesiones en el grupo.

1. Se seleccionó la sesión 2, donde se ve que un intruso utiliza asiduamente el comando '`kill -9 0`'
2. La respuesta del tcleoclusterer es:

   ```
   Connection Statics: (Distance level = 75)
   Weight : [ 0.4  0.2  0.7  1.   0.6  0.6  0.7  0.7]
    (temporal | uid | 5 dir | 5 sites | first 3 | last 3 | 5 progs | 5 uncommon)
   Total connection ratio (bonding): 0.11% [84 74299]
   Total outside connections linking (>0 and <75) to the group: 41 from 74256 (0.06%)
   Total inside connections linking (>0 and <75) outside the group: 42 from 43 (97.67%)
   Total inside connections linking (>0 and <75) to themselves: 1 from 43 (2.33%)
   Ratios between percentages : 0.00 and 0.02
   ```

3. Podemos apreciar que de las 43 conexiones de esta sesión, solo 1 se conecta consigo misma (2.43 %), el resto de las 42 se conectan con sesiones fuera del grupo (97.67 %). Este comportamiento es coherente.
4. Basándose en datos obtenidos por el tcleoclusterer, se selecciona la sesión 5 como la siguiente más relacionada a la sesión 2.

5. La respuesta del tcleoclusterer es:

```
Connection Statics: (Distance level = 75)
Weight : [ 0.4  0.2  0.7  1.   0.6  0.6  0.7  0.7]
 (temporal | uid | 5 dir | 5 sites | first 3 | last 3 | 5 progs | 5 uncommon)
Total connection ratio (bonding): 0.13% [97 74034]
Total outside connections linking (>0 and <75) to the group: 46 from 73983 (0.06%)
Total inside connections linking (>0 and <75) outside the group: 47 from 51 (92.16%)
Total inside connections linking (>0 and <75) to themselves: 4 from 51 (7.84%)
Ratios between percentages : 0.00 and 0.09
```

6. Podemos apreciar una mejora en los porcentajes. Un 7.84 % son conexiones con sesiones del mismo grupo, mientras que el 92.16 % se relaciona con sesiones fuera del grupo. La disminución del porcentaje de conexiones con sesiones fuera del grupo es un factor que indica qué sesiones pueden ser agrupadas por este método.
7. Si en lugar de la sesión 5, se intenta armar el grupo con la sesión 19, que no guarda parecido aparente con la sesión 2, observamos el siguiente comportamiento:
8. Respuesta del tcleoclusterer:

```
Connection Statics: (Distance level = 75)
Weight : [ 0.4  0.2  0.7  1.   0.6  0.6  0.7  0.7]
 (temporal | uid | 5 dir | 5 sites | first 3 | last 3 | 5 progs | 5 uncommon)
Total connection ratio (bonding): 0.38% [279 74124]
Total outside connections linking (>0 and <75) to the group: 138 from 73983 (0.19%)
Total inside connections linking (>0 and <75) outside the group: 139 from 141 (98.58%)
Total inside connections linking (>0 and <75) to themselves: 2 from 141 (1.42%)
Ratios between percentages : 0.00 and 0.01
```

9. Podemos apreciar que los porcentajes empeoraron respecto a agrupar la sesión 2 con la 5. Ahora menos sesiones se relacionan con el grupo (1.42 %) y muchas más se relacionan fuera del grupo (98.58 %)
10. Volviendo al grupo de las sesiones 2 y 5, se agregó una tercera sesión, la 7, que según el tcleoclusterer estaba muy relacionada con la 2 y la 5.
11. La respuesta del tcleoclusterer es:

```
Connection Statics: (Distance level = 75)
Weight : [ 0.4  0.2  0.7  1.   0.6  0.6  0.7  0.7]
 (temporal | uid | 5 dir | 5 sites | first 3 | last 3 | 5 progs | 5 uncommon)
Total connection ratio (bonding): 0.16% [115 73772]
Total outside connections linking (>0 and <75) to the group: 53 from 73710 (0.07%)
Total inside connections linking (>0 and <75) outside the group: 53 from 62 (85.48%)
Total inside connections linking (>0 and <75) to themselves: 9 from 62 (14.52%)
Ratios between percentages : 0.00 and 0.17
```

12. Comprobamos que los porcentajes siguen mejorando.

Este experimento muestra que es posible clasificar las sesiones basándose en sus características comunes.

10. Conclusiones

El análisis de los datos de los intrusos con las herramientas desarrolladas arrojó resultados importantes en el conocimiento del comportamiento de los mismos. Ahora podemos comenzar a analizar los grupos de sesiones.

A medida que se trabajó con los detalles de las sesiones se logró definir algunas de las características más importantes para clasificar intrusos y los mejores vectores de pesos para evaluarlos. Fue importante evaluar las características para descartar algunas técnicas y para potenciar otras.

El continuo análisis del comportamiento de los intrusos, incluyendo seguimientos en tiempo real posibilitó la creación de las diversas comparaciones. Contar con datos reales fue decisivo en el análisis de las distancias y la especificación de los valores. Estas medidas de distancia permitieron comprender mejor los patrones inherentes al trabajo de cada intruso y mejorar progresivamente nuestros métodos de clasificación.

Se comprobó el vector de pesos que mejor clasifica las sesiones.

Se comprobó que es posible agrupar a los intrusos analizando su comportamiento.

Se comprobó que el algoritmo encuentra sesiones *perdidas* cuando se definen grupos.

Se comprobó que es posible, bajo ciertas condiciones, ir encontrando las sesiones e ir armando los grupos paulatinamente.

Se comprobó que los porcentajes de las relaciones se pueden aplicar para la automatización del proceso de armado.

11. Trabajo futuro

Como el proyecto principal en donde se realizó esta investigación está en proceso, hace falta continuar las pruebas para mejorar los resultados.

Es necesario automatizar el proceso de selección de sesiones y armado de grupos, comprobar una cantidad de situaciones más representativa de un muestreo estadístico y aplicar técnicas de clustering para lograr armar clusters que separen las sesiones.

Referencias

1. Sherif, J.S., Dearmond, T.G.: Intrusion detection: Systems and models. In: WETICE. (2002) 115–136
2. Anderson, K.E.: International Intrusions: Motives and Patterns (1994) Statistical Science (submitted).
3. Maxion, R., Townsend, T.: Masquerade detection using truncated command lines (2002)
4. Benitez, C.: El arte del Estado. Seminario de Seguridad Informatica - Prince and Cooke (2006)
5. García, S.: E-Pollen, usa un honeypot y te dire quien eres. Workshop de Soluciones Informáticas 2006, Universidad FASTA (2006)

6. García, S.: An evening with kha0s. FIRST Security Workshop, ArCert (2005)
7. Fernandez, F.: Linux kernel hacking aplicado a Honeypots. CONSECRI 2do Congreso Nacional de Seguridad en Sistemas Teleinformáticos y Criptografía (2005)
8. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley-Interscience Publication (2000)