Entregable D1.
Informe de requisitos del entorno de experimentación y pruebas.

Tarea: T2. Definición de requisitos
Investigador Responsable: Gabriel Maciá Fernández
Grupo de trabajo (% dedicación): GMF (65), PGT (50), RTS (50), JCP (50), RM (25), PCS (25)
Requisitos previos: N/A.
Riesgos y contingencia: R1 – C1.a y/o C1.b
Duración: 5 meses
Intensidad de trabajo (hombre-mes): 13,25
Objetivo Específico (IP Responsable): OE5 (Gabriel Maciá Fernández)
Entregables: D1,D2.

Índice:

1. Requisitos de la Experimentación

Para abordar adecuadamente la definición de los requisitos del entorno de experimentación y pruebas, hemos de partir necesariamente de las **hipótesis de partida** y de los **objetivos técnicos principales del proyecto**. A partir de hipótesis y objetivos, se definirán los requisitos de la experimentación y, a partir de los mismos, los del entorno.

Las **hipótesis de partida**, según se especifica en el documento del proyecto, son:

H1: Las técnicas multivariantes desarrolladas por el equipo de investigación, en combinación con las adecuadas técnicas de caracterización de la información y de análisis visual interactivo, constituyen una metodología de altas prestaciones para la detección y diagnóstico de anomalías en un SIEM.

H2: Extensiones jerárquicas del análisis multivariante permiten desarrollar un sistema SIEM distribuido de detección de anomalías, donde (a) las prestaciones de detección y diagnóstico son equivalentes al sistema no distribuido y (b) se reduce el volumen de tráfico de seguridad en red.

H3: La reducción del volumen de tráfico de seguridad en la red supone una ventaja competitiva en sistemas SIEM.

H4: El esquema SIEM distribuido permite desacoplar la detección del diagnóstico de anomalías, de forma que (a) se mantienen las prestaciones originales y (b) se mejora la privacidad del sistema.

H5: El desacoplo de detección y diagnóstico supone una ventaja competitiva en el sistema SIEM resultante.

Relacionando estas **hipótesis con requisitos de la experimentación**, tenemos los siguientes requerimientos:

R1: H1 queda parcialmente demostrado con trabajo previo. No obstante, podría ser adecuado comparar la metodología multivariante, denominada de ahora en adelante Principal Component Analysis - Multivariate Statistical Network Monitoring (PCA-MSNM), con técnicas representativas del estado del arte.

R2: Previamente a dicha comparativa, es necesario definir una serie de buenas prácticas en nuestra propuesta PCA-MSNM, que siguen siendo materia de controversia en la literatura:

- Selección automática de PCs para PCA-MSNM: validación cruzada, técnicas estadísticas, teoría matricial, heurísticas,etc.
- Evaluación de distintas estrategias de *feature engineering*: Contadores, Entropía, K-L.
- Comparación de métodos de diagnóstico: oMEDA, CDC, RBC, AC.
- Evaluación de distintas metodologías de monitorización: Q-st, PCA con D&Q-st, PCA con índice combinado.

R3: Adicionalmente, es necesario contemplar en dicha comparativa un listado de fuentes o sensores de datos lo más completo posible, incluyendo fuentes habituales (típicamente disponibles en redes comunes) y menos habituales, y estudiar su valor informativo para la seguridad y su carga en la red. Las fuentes a considerar son:

- Firewall logs
- IDS logs
- Netflow
- Syslogs / event viewer
- AV logs
- SNMP info
- CPU, memory, IO, processes (big brother)
- Clasificador de tráfico
- Reputaciones

R4: Adicionalmente, es necesario contemplar en dicha comparativa un listado de ataques lo más completo en tipología posible:

- Network scan
- DoS dirigido a un nodo monitorizado
- Botnet C&C con puertos diferentes (IRC)
- Local infection
    - Infección drive-by download en un solo nodo (exe, pdf, java)
    - Infección local con malware mediante campaña (social engineering)
- Remote attacks to services
    - Ataque MS-08-067 a Windows en un nodo

R5: Adicionalmente, es necesario contemplar en dicha comparativa un listado de nodos lo más completo posible:

- Windows
- Linux
- Routers
- Switches
- Nodos Wireless (finales y switches)
- Android (+ adelante?)
- Cloud (+ adelante?)

R6: H2 puede demostrarse comparando un PCA-MSNM plano con un PCA-MSNM jerárquico en términos de capacidad de detección y tráfico generado. Para ello, existe una serie de puntos a tratar:

- Estudiar el efecto de los sensores intrusivos, que a la vez que capturan datos tienen un efecto en la red, típicamente bloqueando acciones potencialmente maliciosas. Es el caso de firewalls, antivirus y algunas configuraciones de IDS (IRS).
- Estudiar el rendimiento y carga en la red en función de la metodología de monitorización, el tiempo de *pooling*, el número de ataques, etc.

R7: Aunque H3 parece clara, ya que cualquier reducción en tráfico es a-priori interesante, una experimentación relevante es la relacionada con los límites de rendimiento de la propuesta jerárquica al permitir la compresión de información, básica para definir potenciales entornos de aplicación, incluyendo:

- Número máximo de fuentes de información/variables medidas en global.
- Número máximo de niveles de jerarquía.
- Número máximo de fuentes por nivel jerárquico.
- Período de muestreo/*pooling* mínimo.

R8: H4 y H5 requieren comparar el sistema PCA-MSNM acoplado, donde detección y diagnosis se hacen en el SIEM, con el sistema desacoplado, donde la detección se hace en cualquier parte de la jerarquía pero la diagnosis se hace en el sistema final. Esto incluye:

- Evaluar la técnica de diagnosis más adecuada para esta filosofía.
- Evaluar el grado de privacidad obtenido con los estadísticos PCA-MSNM.

Los **objetivos técnicos principales del proyecto**, según se especifica en el documento del proyecto, son:

OE1. Diseñar técnicas optimizadas de recolección de datos e identificación de características de las fuentes de datos monitorizadas.

OE2. Desarrollar algoritmos jerárquicos basados en el análisis multivariante para la detección y diagnóstico de anomalías en red que permitan la reducción de tráfico de seguridad y proteger la privacidad de la información.

OE3. Diseñar una estructura distribuida para la recolección, procesado y transmisión de información de seguridad en un SIEM con las técnicas desarrolladas.

OE4. Desarrollar soluciones visuales interactivas para el análisis de eventos en red que permitan alta interactividad al analista.

OE5. Evaluar las técnicas desarrolladas en entornos virtualizados y con datos reales.

OE6. Evaluar de forma preliminar la potencial aplicación de la solución propuesta en redes y modelos de mercado modernos.

Todos estos objetivos no aportan ningún requisito experimental adicional con la excepción del siguiente:

R9: OE5 implica el uso de virtualización y datos reales.

2. Requisitos del entorno de experimentación

R9 nos define la necesidad de establecer un entorno virtual que permita la realización de la mayor parte de la experimentación y un entorno real que permita validar los puntos o resultados más relevantes.

A partir de los requisitos de experimentación, podemos definir el siguiente listado de **requisitos del entorno virtual (EV) de experimentación**:

REV1: El EV debe permitir la instalación de los sistemas de detección listados R1.

REV2: El EV debe permitir conocer cuándo existe una anomalía de seguridad (*ground-truth* accesible) con la mayor precisión temporal posible, de cara a permitir la medición de las estadísticas de rendimiento más utilizadas en el contexto del control estadístico: Overall Type I (OTI) Error, Overall Type II (OTII) Error, Average Run Length (ARL). Esto es necesario para realizar convenientemente las comparativas en R1, R2, R6 y R8.

REV3: El EV debe permitir la instalación de las fuentes listadas en R3.

REV4: El EV debe permitir la realización de los ataques listados en R4.

REV5: El EV debe proveer de flexibilidad de topología para la interconexión de los nodos listados en R5.

REV6: El EV debe proveer de flexibilidad de localización de sensores en dicha topología para permitir R6.

REV7: El EV debe ser escalable en dispositivos y tráfico para permitir R7.

A partir de los requisitos de experimentación, podemos definir el siguiente listado de **requisitos del entorno real (ER) de experimentación**:

RER1: El ER debe permitir la instalación de los sistemas de detección listados R1.

RER2: El ER debe permitir la instalación de las fuentes más relevantes en R3.

RER3: El ER debe permitir la realización de los ataques más relevantes en R4.

RER4: El ER debe tener varios niveles de jerarquía, y un nivel de agregación en cada jerarquía significativo.

RER5: El ER debe permitir la localización de determina dos sensores en dicha topología.