Mejores prácticas para la evaluación dinámica

Disclaimer

This document is a translation of the English-language AMTSO document **" Best Practices for Dynamic Testing" (version 2008-10-31)** at http://www.amtso.org/documents/doc_download/7-amtso-best-practices-for-dynamic-testing.html. It is provided in good faith in order to make the concepts clearer to a wider audience. However, we cannot guarantee that it reflects the content of the master version with complete accuracy.

In the event of any conflict or divergent interpretation, the authoritative version remains the latest approved version of the English-language master document on the AMTSO web site.

Introducción

Este documento describe las mejores prácticas para la evaluación dinámica de productos antimalware basados en el entorno del equipo donde se realiza. La evaluación dinámica significa una evaluación donde la computadora se expone a una amenaza viva (por ejemplo, tratando de ejecutar el código malicioso) como parte de la evaluación. Este tipo de evaluación es una prueba de la eficacia del producto mucho más realista que las evaluaciones estándar (por ejemplo, exploración bajo demanda), ya que imita el código malicioso que se ejecuta en la máquina de la víctima en forma directa. Además de ser la única manera de evaluar algunas tecnologías antimalware, la evaluación dinámica es apropiada como metodología de prueba para todos los tipos de productos antimalware.

Lamentablemente, este tipo de evaluación es más compleja de realizar que las evaluaciones estáticas, ya que hay muchos más aspectos para tener en cuenta. Este documento tiene como objetivo resumir las mejores prácticas y brindar una guía para ayudar a los evaluadores a lidiar con esos aspectos. Las siguientes secciones incluyen reproducibilidad, selección del producto, selección de muestras, entorno de evaluación, generación de registros y verificación, medición de aciertos, y manejo de mensajes emergentes y eventos que requieren la intervención del usuario. Al final de este documento se dan dos ejemplos de metodologías de evaluación.

Este documento es una extensión del documento “Principios fundamentales para la evaluación” de AMTSO disponible en http://www.amtso.org.

Reproducibilidad

A diferencia de las evaluaciones estáticas, las evaluaciones dinámicas son, por sus propias características, difíciles de reproducir, lo que significa que la misma evaluación ejecutada en distintos momentos puede arrojar resultados diferentes. Esta variabilidad puede estar causada por cambios en los productos de los vendedores (por ejemplo, por la actualización de las firmas de virus) y también por alteraciones en el entorno de los códigos maliciosos. Algunos códigos maliciosos sólo funcionarán bien si están disponibles ciertos recursos externos (como servidores NTP, páginas de sitios Web bancarios, sitios liberadores de programas maliciosos). Aunque es posible imitar algunos de estos factores en un entorno de evaluación, es difícil reproducirlos con una exactitud total, y el entorno de evaluación puede resultar cada vez más artificial. Esta variabilidad significa que es difícil llegar a conclusiones certeras habiendo realizado una única evaluación.

El evaluador cuenta con dos defensas contra esta variabilidad. La primera es reunir suficientes registros e información para verificar lo que ocurrió en la evaluación. Por ejemplo, para mostrar que el código malicioso “x” evaluado el día “y” sobre el producto “z” realmente llegó a actuar. La segunda es usar una cantidad suficiente de muestras y repetir las evaluaciones con el paso del tiempo (usando distintos grupos de muestras) para que las inconsistencias en el comportamiento de los códigos maliciosos tengan menos probabilidad de desviar los resultados. Por ejemplo, los códigos maliciosos

recientes pueden probarse durante un mes desde el día en que se obtuvieron, comparando el rendimiento o las tasas de detección diarios.

Selección del producto

A veces la protección antimalware ya está incorporada en paquetes de productos, mientras que otras veces se encuentran como productos independientes. El evaluador debe darse cuenta de estas diferencias y elegir los productos con detenimiento para realizar evaluaciones comparativas razonablemente apropiadas. Una forma de descubrir estas diferencias sería usar los anuncios de vendedores como base para elegir productos; por ejemplo, si los vendedores aseguran que sus productos tratan la amenaza “x”, entonces sería razonable evaluar varios productos que traten la amenaza “x”.

Selección de muestras

En cualquier evaluación, la selección de muestras es importante. Sin embargo, para las evaluaciones dinámicas, los grupos de muestras deben evaluarse según los criterios que se detallan a continuación. En líneas generales, la calidad de las muestras es más importante que la cantidad. Las muestras deben ser:

1. **Funcionales**. Para que sea una buena evaluación de la protección, las muestras de códigos maliciosos deben ser viables. Es importante elegir muestras que “funcionen” (por ejemplo, las muestras alteradas o fragmentadas no conforman material válido para una evaluación) y luego verificar que las muestras realmente hicieron algo malicioso en el entorno de la evaluación.
2. **Diversas**. Con frecuencia, las evaluaciones dinámicas se llevan a cabo con un grupo de muestras más reducido que en las demás evaluaciones. Por esta razón, es aún más importante que el grupo de muestras sea variado. La diversidad en este sentido se debe aplicar tanto a la variedad de familias de códigos maliciosos evaluados (por ejemplo, 50 variantes del RBot no sería diverso) como al comportamiento esencial de los códigos maliciosos. Por ejemplo, no tendría sentido evaluar la eficacia general usando un grupo de prueba conformado por 50 discadores (*dialers*).
3. **Relevantes**. Es importante tener en cuenta el predominio de un código malicioso determinado cuando se crea un grupo de muestras relevante.
4. **Recientes**. Un aspecto importante de todas las tecnologías es la protección del día cero (*zero day*). La mejor manera de evaluar esta protección es usando amenazas recientes y que sean relevantes en la actualidad. Por lo tanto, es necesario considerar la edad de las muestras.

Evaluación de falsos positivos

Para brindar una evaluación equilibrada de lo que experimenta el usuario, las pruebas deben incluir la búsqueda de falsos positivos evaluando los productos también con programas no maliciosos. Estos programas deben cubrir el grupo de operaciones

comunes efectuadas por el usuario en su computadora; por ejemplo, la instalación, la actualización y la ejecución de aplicaciones, la descarga de parches para el sistema operativo, y la instalación y el uso de *plugins* para exploradores. Las aplicaciones instaladas deben ser ejecutadas para asegurarse de que funcionan correctamente.

Entorno de evaluación

En las evaluaciones dinámicas, el rendimiento de los productos está altamente determinado por el comportamiento de los códigos maliciosos; y el comportamiento de los códigos maliciosos está en sí altamente determinado por el entorno donde están activos. En consecuencia, es importante crear un entorno de ejecución razonable para obtener buenos resultados. En este caso el entorno significa, por ejemplo, el sistema operativo de la máquina, si se trata de una máquina real o virtual, la conectividad de red, cómo se introduce el código malicioso, etc. No existe un entorno "correcto", por lo que los evaluadores deben tener conocimiento de los cambios que pueden ocurrir al seleccionar un entorno particular y cómo (sin darse cuenta) se podrían llegar a alterar los resultados.

Uso de máquinas virtuales
Muchos códigos maliciosos no manifiestan su espectro completo de comportamiento cuando se encuentran en entornos virtuales (por ejemplo, software de virtualización VMware o computadoras virtuales). Además, el uso de ciertos productos no es soportado en entornos virtuales. La mejor alternativa es usar máquinas reales, lo que es más complejo a nivel técnico y más difícil de automatizar. A pesar de estas dificultades, en las evaluaciones dinámicas se recomienda el uso de máquinas reales. Para minimizar las dificultades técnicas, AMTSO alienta a los miembros participantes a divulgar las herramientas de prueba útiles para que estén disponibles para todos los demás miembros. Los evaluadores siempre deben indicar qué tipo de máquinas fueron usadas en las evaluaciones.

Red
Muchas formas de códigos maliciosos y algunos productos requieren una conexión de red para ejecutarse con todas sus funcionalidades. Por lo tanto, la evaluación requiere la habilitación de un acceso a Internet desde las máquinas de prueba. No obstante, es una práctica peligrosa, ya que los códigos maliciosos pueden llegar a propagarse a otras máquinas o causar otros daños. Existen dos enfoques comunes para tratar esta dificultad. El primero es permitir la conexión de red, pero restringiendo los protocolos permitidos. Por ejemplo, uno de los arreglos más usados es permitir el acceso del tráfico http (de red) a Internet, pero bloqueando todos los demás protocolos. Normalmente esto se logra haciéndolo en la puerta de enlace en vez de en la máquina de prueba. Una alternativa es crear una red de Internet virtual (también conocida como la caja de Truman), donde se envían respuestas falsas a todas las solicitudes de red. Otra alternativa es usar un vínculo lento de red (por ejemplo, acceso por módem o ISDN en lugar de DSL).

No existe una configuración "correcta", por lo tanto los evaluadores deben hacer una elección basándose en información suficiente y documentar la configuración en la metodología de prueba.

Ejecución de los códigos maliciosos
Es posible que los productos tengan en cuenta la manera en que los códigos maliciosos infectan las máquinas y estén configurados para ser sensibles a los vectores de infección más comunes, por ejemplo, las descargas no autorizadas (*drive-by downloads*). En consecuencia, la forma en que se inicia una infección puede afectar el rendimiento del producto. Un principio que sirve como guía es ejecutar los códigos maliciosos de la misma forma en que se introducirían a la máquina durante una infección real. Por ejemplo, si un código malicioso específico se origina como una descarga de HTTP, la evaluación debe presentar el código malicioso a través de su protocolo de descarga nativo en vez de ejecutarlo en forma local. Si resulta difícil de implementar, una alternativa menos precisa es introducir el código malicioso en forma manual y variar las formas en que es ejecutado.

Registros/Verificación

Como en las evaluaciones dinámicas el comportamiento de los códigos maliciosos es un elemento crucial para la forma de funcionamiento del producto, es de suma importancia que el evaluador realice los registros y las verificaciones apropiados indicando el proceso completo de evaluación. Los datos mínimos que debe incluir son:

1. Las acciones realizadas por los códigos maliciosos en la máquina infectada/comprometida
2. Las modificaciones a archivos, registros y áreas del sistema
3. Los rastros de la actividad de red

Medición de aciertos

Como las evaluaciones dinámicas requieren la ejecución de códigos maliciosos, y los códigos maliciosos pueden tener efectos muy diferentes en un sistema (como instalación de archivos, cambios de configuración, pérdida de información, etc.), es importante que la definición de los aciertos se considere con detenimiento. Hay una gran variedad de formas para medir los aciertos, algunas de las cuales serán más relevantes en una evaluación en particular que en otras. Entre estas medidas se encuentran:

- **Detección**. ¿Detectó (informó o registró) algo el producto?
- **Eliminación**. ¿Hubo cambios en los archivos o de configuración (por ejemplo, modificaciones de registro) que permanecieron después de que el producto trató el código malicioso? Las formas en que se puede aplicar varían desde una interpretación estricta (anunciar absolutamente todos los cambios) hasta ignorar (por ejemplo) archivos basura y archivos de datos creados o agregados, opciones de configuración que no tienen una función real, etc. Algunos enfoques pueden deshabilitar los códigos maliciosos por considerarlos incapaces de seguir provocando daños, en vez de eliminarlos; por eso hay que tener esto en cuenta al juzgar los aciertos y desaciertos.
- **Persistencia (actividad/supervivencia después de reiniciar el sistema)**. ¿El código malicioso permaneció activo (ejecutándose en memoria) o quedó

configurado para sobrevivir al reinicio del sistema después de que el producto terminó de tratarlo? Este punto es un derivado más leve pero más accesible del punto anterior.

- **Daños**. ¿El código malicioso logró comprometer la máquina? Esto es especialmente relevante en los casos de códigos maliciosos que roban información, donde pueden haberse modificado, eliminado o perdido datos personales de la computadora aún cuando el código malicioso se haya bloqueado con éxito. Aunque ésta puede ser una buena medición de aciertos, en general es difícil medirlo en la práctica.

Mensajes emergentes y eventos que requieren la intervención del usuario

Muchos productos usan mensajes emergentes u otro tipo de ventanas para pedirle al usuario información sobre cómo proseguir. Esto puede provocar confusión en los resultados de la evaluación (por ejemplo, si el producto pide permiso para bloquear una amenaza y el evaluador siempre contesta que no, entonces el rendimiento del producto será muy diferente a cuando dice siempre que sí). Las recomendaciones más importantes para el manejo de mensajes emergentes y las acciones del usuario son:

1. **Política**. Los evaluadores deben decidir qué política usarán para manejar la interacción del usuario. Por ejemplo, pueden elegir responder a los mensajes emergentes de la manera más favorable para el producto, o de la menos favorable. La política utilizada debe describirse en forma explícita en el informe de la evaluación.
2. **Consistencia**. Una vez que se decidió la política que se usará, se debe aplicar en forma consistente en todas las pruebas (por ejemplo, aplicarlas tanto en evaluaciones de falsos positivos como de detección).
3. **Informes**. El evaluador debe medir e informar la cantidad de interacciones con el usuario que requiere el producto. Esto permitirá a los lectores del informe determinar de qué tipo de producto se trata (por ejemplo, es efectivo pero abre mensajes constantemente; es efectivo y menos entrometido, etc.). Hay distintas clases de mensajes emergentes, que deben ser notificados en forma separada. Por ejemplo:
    a. **Disculpa** – el producto realiza una acción sin requerir interacción del usuario; luego le informa al usuario que ya lo realizó.
    b. **Permiso** – el producto solicita permiso o la toma de una decisión antes de realizar una acción.
    c. **Notificación** – mensajes de notificaciones en general, quizá diferenciando entre las que requieren intervención del usuario y las que no.

Estilos de evaluaciones dinámicas

En esta sección se describen dos estilos comunes de evaluación dinámica. No es una lista exhaustiva y los evaluadores sólo deben usarla como fuente de inspiración.

El primer estilo es el enfoque “uno por vez”. En este caso, las máquinas se configuran con un solo producto instalado de un vendedor, se ingresa una sola muestra de código malicioso y una vez que el producto tuvo la oportunidad de detectarla y eliminarla se analiza el estado de la máquina para verificar que el código malicioso fue detectado y eliminado con éxito. Después de la prueba, la máquina se revierte a su estado original y se repite la prueba con el siguiente código malicioso. Este enfoque es preciso pero requiere mucho tiempo. El análisis debe medir, por lo menos, cualquier cambio realizado en el sistema de archivos (archivos agregados, eliminados o modificados) y en la configuración del registro como consecuencia de la ejecución del código malicioso y el hecho de haber sido tratado por el programa antimalware.

El segundo estilo es el enfoque “muchos a la vez” que presenta las mismas características generales pero ejecutando múltiples códigos maliciosos a la vez. Este enfoque es menos preciso pero más eficiente en cuanto al análisis. Una variación interesante es obtener los códigos maliciosos provocando que la máquina visite una gran cantidad de sitios Web sospechosos, con la esperanza de que se infecte, por ejemplo usando una secuencia de comandos para iniciar el explorador en forma repetida. Una vez que todos los sitios fueron visitados, se analiza la máquina para detectar la precisión con la que el producto del vendedor protegió la máquina de las infecciones. Esta evaluación es una buena simulación de un vector de infección muy común: las descargas no autorizadas (*drive-by downloads*). Un obstáculo de este enfoque es que, en parte debido a los esfuerzos de los investigadores de códigos maliciosos, es posible que los servidores maliciosos (o cualquier código malicioso) no funcionen adecuadamente si reciben solicitudes múltiples. En esta evaluación es difícil garantizar que el producto de cada vendedor quedará expuesto a las mismas amenazas, por lo que es importante usar una lista diversa de sitios sospechosos como así también repetir la evaluación varias veces para ver las tendencias de rendimiento.