REVISTA DIGITAL DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES
TELEM@TICA
Revista Telem@tica. Vol. 13. No. 1, enero-abril, 2014, p. 12-19 ISSN 1729-3804

MOTOR ANTIVIRUS DE SEGURMÁTICA: PARA EL PRESENTE Y EL FUTURO
SEGURMÁTICA ANTIVIRUS ENGINE: FOR PRESENT AND FUTURE

Lic. Alejandro Rivero Pérez
Segurmática, Cuba, alex@segurmatica.cu
MSc. Jorge Lodos Vigil
Segurmática, Cuba, lodos@segurmatica.cu

RESUMEN: La industria de los Programas Malignos se ha desarrollado a pasos agigantados en los últimos años, con especímenes que van desde la molestia al usuario a través de ventanas emergentes de propaganda no deseada, pasando por robo de contraseñas de la banca y portales de juegos online, la creación de plataformas para el desarrollo rápido de variantes de programas malignos con diversas funcionalidades, entre ellas la de crear "Redes Zombis", hasta la creación de armas cibernéticas como el "Stuxnet" y el "Flame".
En este mundo salvaje los Antivirus cobran mayor importancia y más aun su componente más importante, la identificación, de cuya capacidad de actualización y mejora progresiva depende en gran parte la calidad del Antivirus en cuestión, su capacidad de estar al día con las nuevas amenazas que aparecen y la rapidez de agregar nuevos mecanismos de identificación o modificaciones de los existentes para dar respuesta rápida.
Un diseño adecuado de la identificación y todos los mecanismos adyacentes, que permita su adaptación a los tiempos cambiantes en cuanto al desarrollo de Programas Malignos y los vectores de ataques que estos utilizan puede ser la diferencia entre un tiempo de respuesta rápido y la infección masiva.

Palabras Clave: identificación, programas malignos, antivirus, diseño, escalabilidad, modularidad.

ABSTRACT: The malware industry has been developed quickly during recent years, with specimens ranging from discomfort to the user via unwanted pop-up advertising, to stealing passwords of online banking and online gaming portals, creating platforms for rapid development of variants of malware with different functions, including but not limited to create "Botnets", up to creation of cyber weapons such as "Stuxnet" and "Flame".
In this wild world Antivirus are becoming more important and more so its most important component, the identification, from which upgradeability and progressive improvement depends greatly on the quality of the Antivirus concerned, their ability to keep up with the new threats that appear and quickly adding new identification mechanisms or modifications of existing ones to quick reply.
Proper design of identification and all adjacent mechanisms that allow its adaptation to time change in malware development and used attack vectors can be the difference between a fast response time and massive infection.

KeyWords: identification, malware, antivirus, software design, scalability, modularity.

INTRODUCCIÓN

La complejidad y diversidad de las funcionalidades utilizadas por los programas malignos ha aumentado muchísimo en los últimos años, causando que la identificación de los mismos deba evolucionar en igual medida. La rapidez en la utilización de nuevos métodos de infección, propagación, ofuscación, etc… hace que el diseño de un componente de identificación que permita la modularidad y extensibilidad necesaria para estar al día en el estado del arte de la detección de comportamiento maligno sea de extrema importancia para cualquier casa Antivirus.

Teniendo en cuenta las dificultades agregadas que se ven en la identificación de comportamiento maligno (como son los diferentes formatos de ficheros utilizados, que pueden ser formatos contenedores, empaquetados, entre otros, o la reutilización de funcionalidad maligna en diferentes versiones de diferentes programas malignos, la complejidad de estas, donde algunos ejemplos pueden llegar a niveles de complejidad de Rootkits y ROP), la identificación debe ser lo suficientemente flexible para que el soporte para dicha detección pueda ser agregado, probado y puesto en producción en el menor tiempo y con la mayor exactitud posible.

El en presente trabajo se presentara el diseño y funcionamiento general del componente de identificación del Motor Antivirus de Segurmática, su arquitectura, capacidades y los resultados obtenidos.

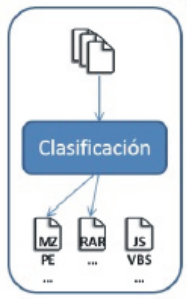
DISEÑO

Uno de los componentes más importantes de un Antivirus (AV) en especial de su Motor Antivirus es la identificación, que permite la detección de comportamiento maligno en flujos de datos, como pueden ser ficheros, conexiones de red, etc…

La identificación del motor antivirus de Segurmática consta de 4 subcomponentes principales que llevan a cabo las tareas más importantes de la misma:

- Clasificación
- Inicialización y Preprocesamiento
- Operación
- Búsqueda

Clasificación



La clasificación de los flujos de datos en los cuales se va a identificar funcionalidad maligna es muy necesaria, nos permite especializar la búsqueda y definir la información necesaria en ese tipo de formato para el escaneo.

Este componente es el encargado de, dado un flujo de datos desconocido (ej: ficheros, conexiones de red, memoria, etc…) identificar los posibles formatos a lo que pudieran referirse los datos analizados (ej: PE, ELF, JS, RAR, MSI, TCP, Shared Memory, etc…)

Entre los clasificadores más comunes se encuentran:

• Firmas: utilizan una o varias secuencias de bytes que deben estar presentes en los datos.

• Contenido Textual: realizan búsquedas de patrones textuales normalmente correspondientes al lenguaje a identificar.

• Complejos: utilizan información obtenida de formatos detectados con anterioridad.

Inicialización y Preprocesamiento

Los datos necesarios de los ficheros o flujos de datos a analizar son extraídos, precalculados o preprocesados en este subcomponente y puestos a disposición del resto del motor antivirus para su consulta, de esta forma solo se calculan 1 vez.

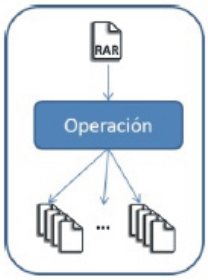
Entre las informaciones más comunes que son extraídas están:

• Sobre el formato del fichero: donde empiezan las secciones de un fichero PE, donde comienzan los tags script de un fichero html, etc…

•Precalcular información costosa necesaria: obtener la frecuencia de uso de los tipos de instrucciones en Intel x86

•Preprocesar el fichero: se genera una versión del fichero donde se ignoran las áreas que no pueden contener funcionalidad maligna.

Operación

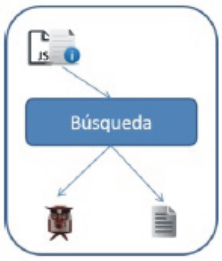


Varios tipos de ficheros tienen la posibilidad de contener datos que son identificables independientemente, por ejemplo un fichero RAR puede ser escaneado e identificado por el motor antivirus, pero también pueden serlo los ficheros comprimidos que contiene, de la misma forma existen ficheros de instalación que contienen otros ficheros.

También existen ficheros cuyo contenido real solo se obtiene al ejecutarlo (empaquetados) y es muy beneficioso escanear el contenido real, que puede obtenerse del contenido físico.

Este subcomponente puede crear un nuevo contenido que será analizado por el motor antivirus y tiene gran importancia en la detección de un mismo programa maligno ofuscado y/o protegido con numerosos mecanismos que provocan que el fichero resultante sea diferente.

Búsqueda



Este subcomponente describe como se detectará el comportamiento maligno en los datos a analizar, tiene la posibilidad de utilizar todos los elementos disponibles en la identificación para la búsqueda. La búsqueda puede tener por ejemplo:

- Inicialización y preprocesamiento o no
- Utilizar método específico de búsqueda o algunos de los disponibles
- Almacenar la información de identificación en las estructuras de datos disponibles o en alguna especifica más eficiente

Método

El mecanismo especifico de búsqueda del comportamiento maligno, puede ser desde tan sencillo como buscar una secuencia de bytes en el flujo de datos en una posición determinada, hasta tan complejo como emular un fichero ejecutable y de acuerdo a la frecuencia de uso de instrucciones Intel x86 o llamado a las APIs de Windows detectarlo.

En este subcomponente es donde se encuentran los diferentes métodos de búsqueda disponibles para los tipos de ficheros detectados por el motor antivirus. Cuando se agrega un método queda disponible para todos los tipos de ficheros a los que se pueda aplicar y permite la composición de métodos más complejos utilizando métodos existentes.

Almacén

La información de identificación son los datos más importantes y voluminosos de la identificación. De su almacenaje, salva y carga eficiente depende la velocidad y uso de memoria del motor antivirus.

La forma de guardar la información de identificación puede priorizar la rapidez de búsqueda del comportamiento maligno o la cantidad de información necesaria a almacenar, entre otras.

ARQUITECTURA

La identificación del motor antivirus tiene como entrada un flujo de datos desconocido, puede ser ficheros, conexiones de red, etc… Dado esos datos es necesario clasificarlos en los posibles formatos que cumple, utilizando el componente de clasificación, en algunos casos es posible que para la propia clasificación haga falta inicializar algunos datos, dicha inicialización es almacenada para su posterior consulta en cualquier componente de la identificación.

Por cada tipo de formato (ej: PE, DLL, HTML, etc…) se definen los métodos de búsqueda a aplicar. La cantidad de dichos métodos pueden ser desde ninguno (solo se operará o se excluirá el fichero), 1 o varios métodos, también se define la operación que se aplicara para ese tipo de formato en caso que lo requiera (ej: RAR, ZIP, MSI, etc…), si se genera nuevo contenido en la operación se inicia el proceso de identificación desde el principio con cada nuevo contenido generado.

Cada método tiene especificado la forma en la que se almacena la información que se utilizará para identificar y como se salva y carga desde la actualización.

CAPACIDADES

La modularidad y extensibilidad de la identificación permite que se pueda trabajar en distintas implementaciones de sus distintos componentes y subcomponentes, sin afectarse a las otras partes. Posibilita además que pequeñas áreas muy importantes o complejas puedan ser optimizadas y probadas independientemente y esos resultados sean aprovechados por el resto de los componentes mediante la reutilización cuando esté listo.

La facilidad de modificación y posibilidad de agregar nuevos clasificadores, operadores, etc…, hace que estar al día con las nuevas amenazas sea un poco más fácil y totalmente independiente de la mejora continua que tienen los AV día a día.

La posibilidad de dividir las identificaciones de programas malignos en tipos de formatos más específicos y de optimizar la información de identificación de la manera más beneficiosa para ese formato permiten que esté más al alcance la eficiencia que se desea y que requieren los clientes.

Los formatos de ficheros contenedores (comprimidos, instaladores, etc...) aumentan cada día, muchos de ellos no está documentados o no se corresponden exactamente con la especificación. El componente de operación, permite que nuevos formatos puedan ser agregados con pasos simples para su utilización

en la identificación y permitiría también su reutilización en herramientas externas, donde pueden llevarse a cabo las pruebas iniciales y la puesta a punto para su utilización en el motor antivirus.

RESULTADOS

Con el nuevo componente de identificación se agregó soporte para la actualización incremental, donde la información de identificación de los diferentes métodos de búsqueda se salvan en varios ficheros con un pequeño tamaño permitiendo que cuando se agreguen nuevas identificaciones se modifiquen la menor cantidad de ficheros posibles y la cantidad de información a actualizar sea lo más cercano posible a la cantidad de información nueva agregada.

Debido al diseño modular de la identificación se han podido utilizar varios subcomponentes como la clasificación de tipos de formatos y las operaciones en el proceso de generación automático de nuevas identificaciones, entre otros procesos automáticos, que sirven también como pruebas iniciales de correcto funcionamiento y posibilidad de corrección de los problemas encontrados.

Se optimizaron varias partes importantes de la identificación como el procesamiento de ficheros binarios de Windows (ficheros PE), el cual permitió una mejora en el rendimiento significativa en el procesamiento de estos ficheros (más del 30%), que son los más comunes en la búsqueda de programas malignos. También se mejoró el almacenamiento en memoria de la información de identificación de la mayoría de los métodos de identificación por cadenas, para permitir una búsqueda más eficiente, obteniéndose una mejora del 10%, sin afectarse el uso de memoria total ni el ocupado por cada nueva información de identificación.

Los subcomponentes de la identificación tienen un diseño basado en librería y con una interfaz definida, que permite ser utilizado en otros motores antivirus.

Debido a la arquitectura y a la actualización incremental de la identificación se han ido agregando nuevas funcionalidades al motor antivirus en los últimos meses, algunos ejemplos son:

- Clasificación de nuevos tipos de formatos (ej: Visual Basic, PDF, Python, Perl, etc…)
- Nuevos métodos de búsqueda de comportamiento maligno en dichos tipos de formatos agregado.
- Métodos Heurísticos
- Se mejoraron métodos de tipos de formatos populares como: JS, VBS, etc…

Uno de los resultados más importantes es la disminución en el tiempo necesario para agregar nueva funcionalidad, sobre todo funcionalidad compleja, muy necesaria para estar al día en la lucha contra los programas malignos y proveer la seguridad y protección que se merecen los clientes y que estas mejoras puedan estar disponibles en cuanto estén listas a través de la actualización incremental.

REFERENCIAS

1. Symantec Connect: "Building an Anti-Virus Engine",http://www.symantec.com/connect/articles/building-anti-virus-engine, 2002.

2. Wikipedia: "List of File Formats", http://en.wikipedia.org/wiki/List_of_file_formats.

3. Wikipedia: "List of Files Signatures", http://en.wikipedia.org/wiki/List_of_file_signatures.

4. File Format description and detection: http://filext.com/.