REVISTA DIGITAL DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES
TELEM@TICA
Revista Telem@tica. Vol. 13. No. 1, enero-abril, 2014, p. 72-81 ISSN 1729-3804

ADICIÓN DE HEURÍSTICA BASADA EN LA ESTRUCTURA DE LOS FICHEROS PE A SEGURMÁTICA ANTIVIRUS

ADDITION OF HEURISTIC BASED ON PE FILE STRUCTURE TO SEGURMATICA ANTIVIRUS

Walber Zaldivar Herrera1, Alejandro Rivero Pérez2
1 Segurmática. Empresa de Consultoría y Seguridad Informática, Cuba, walber@segurmatica.cu,
Zanja No. 651 Esq. a Soledad, Centro Habana, La Habana
1 Segurmática. Empresa de Consultoría y Seguridad Informática, Cuba, alex@segurmatica.cu

RESUMEN: La mayor parte de los programas malignos son ficheros PE, el formato de los ejecutables, dlls y otros archivos de Windows. La variedad y rápida difusión de este tipo de programas, es uno de los retos para los programadores de los antivirus, se hace necesario adicionar métodos heurísticos, que permitan la identificación de programas malignos que no se encuentren en sus bases de datos.
El análisis de la información contenida en los encabezados de este tipo de ficheros, puede ser utilizado para encontrar aquellos modificados y potencialmente peligrosos. Mediante el desarrollo de herramientas que realizan la revisión y extracción de los datos de interés presentes en la estructura de muestras de ficheros PE, malignas o no, y la utilización de sistemas de bases de datos para guardar la información obtenida; es posible realizar búsquedas que permitan encontrar características que sirven para distinguir programas malignos o potencialmente peligrosos.
Como resultado de este proceso, la identificación de anomalías comúnmente presentes en los programas malignos, ha hecho posible la creación de los algoritmos utilizados para adicionar a Segurmática Antivirus, métodos heurísticos basados en la estructura de los ficheros PE; significando una mejora en su capacidad de detección.
Palabras Clave: antivirus, heurística, programas malignos, programas sospechosos, ficheros PE

ABSTRACT: Most of malware are PE files, the format of the executables, DLLs and other Windows files. The diversity and rapid dissemination of this type of programs, is one of the challenges for antivirus developers, be-comes necessary to add heuristic methods that allow identification of malicious programs that are not in their databases.
The analysis of the information contained in the headers of such files, can be used to find those modified and potentially dangerous. By developing tools which perform the review and extraction of interest data present in the structure of PE files samples, malware or not, and utilization of database systems to store the obtained in-formation, it is possible to realize searches allowing to find characteristics that serve to distinguish malware or potentially dangerous programs.
As a result of this process, the identification of anomalies commonly present in malware, has made possible the creation of the algorithms used to add to Segurmática Antivirus, heuristic methods based on the structure of PE files, meaning an improvement in its detection capability.
KeyWords: antivirus, heuristic, malware, suspicious software, PE files

INTRODUCCIÓN

El uso de las Tecnologías de la Información se ha vuelto imprescindible para el desarrollo. La utilización de las computadoras se ha extendido a todas las esferas, no sólo en la academia militar o industrial, sino en el sector empresarial, en la salud, la educación, la industria del entretenimiento y el uso doméstico.

Al mismo tiempo se han extendido los programas malignos, su cantidad y complejidad aumentan constantemente. Las causas de su creación han dejado de ser la superación personal, la demostración de superioridad o el deseo de reconocimiento; para convertirse en una industria de desarrollo de software que responde a intereses de grupos organizados. Los métodos de ataque y su objetivos se diversifican [1], volviendo cada vez más compleja la lucha contra los creadores de programas malignos.

Es en este escenario que toman importancia métodos heurísticos de detección, que permitan identificar programas que contengan código maligno sin la necesidad de que estos se encuentren en las bases de datos de los antivirus.

Los ficheros PE son los ejecutables, dlls y otros ficheros del sistema de Microsoft Windows, su estructura está establecida y es conocida [2], pero la implementación de su cargador en Windows permite utilizar ficheros que no cumplen completamente con esta especificación [3],[5]. Las variaciones respecto a la especificación dificultan el análisis de estos ficheros mediante herramientas de seguridad y pueden resultar errores en el procesamiento de estos ficheros, con implicaciones en la seguridad, difíciles de detectar [5].

Realizar un análisis de estos ficheros a partir de la información contenida en sus encabezados, permite encontrar aquellos que han sido modificados [6] y por lo tanto, resultan potencialmente peligrosos. El desarrollo de herramientas que realizan la revisión y extracción de los datos de interés presentes en la estructura de muestras de ficheros PE, malignas o no, y la utilización de sistemas de bases de datos para guardar la información obtenida; hace posible realizar búsquedas que permitan encontrar características que sirven para distinguir programas malignos o potencialmente peligrosos y elaborar métodos heurísticos sobre la base de estos resultados.

En este trabajo discutiremos la infraestructura y la metodología utilizadas para adicionar métodos heurísticos basados en la estructura de los ficheros PE a Segurmática Antivirus.

METODOLOGÍA Y DISCUSIÓN

El proceso de adición de métodos heurísticos basados en la estructura de los ficheros PE a Segurmática Antivirus se puede dividir en los siguientes elementos:

- Determinación de las características de los ficheros PE que se quieren analizar
- Diseño e implementación de una base de datos con información de las características a tener en cuenta.
- Diseño e implementación de una aplicación para analizar las muestras y guardar los datos necesarios.
- Determinación e implementación de algoritmos de métodos heurísticos
- Prueba de métodos heurísticos

- Puesta en producción de métodos heurísticos

A continuación se hará una breve descripción de cada uno de estos elementos.

Características a analizar de los ficheros PE

La estructura de los ficheros PE puede ser resumida de modo gráfico como se muestra en la figura 1 y es descrita con detalle en su especificación [2].

Figura 1: Estructura simplificada de los ficheros PE

-Encabezado DOS: tiene como objetivo mantener la compatibilidad con MSDOS.

- Encabezado PE: contiene información relativa a la estructura y contenido del fichero, localización de las tablas de exportación e importación, tipo de sistema operativo, procesador, entre otros.

- Tabla de secciones: Información relativa a las secciones, nombre, tamaño, características, localización.

- Secciones: código, datos y recursos del programa, tablas de exportación e importación.

- Overlay: datos adicionados al final del fichero.

Estudios realizados con anterioridad [3]-[5], indican algunas modificaciones que se pueden realizar sobre los ficheros PE, de modo que sigan resultando válidos para el cargador de Windows. Otros estudios, muestran la posibilidad de realizar análisis mediante técnicas de minería de datos llamadas a las funciones importadas y otras informaciones obtenidas de las muestras de ficheros PE malignas, con el objetivo de lograr sistemas inteligentes, capaces de detectar muestras de programas malignos desconocidos[7]-[11].

Se determinó extraer información que permitiera realizar la búsqueda de anomalías presentes en los programas malignos respecto a la especificación y que permitieran diferenciarlos de los programas no malignos. También se obtiene información que permita determinar el origen de la muestra para análisis posteriores y otras informaciones que requieren ser calculadas como la entropía de los ficheros y de las secciones. En la tabla 1 se muestran los datos utilizados.

Tabla 1. Datos a extraer

Elemento	Datos
Fichero	- ubicación - tamaño - dirección del punto de entrada - contiene TLS - PE de 32 o 64 bits - entropía - MD5 - tamaño de overlay - características - tipo de procesador
Funciones en la tabla de exportación	- nombre - ordinal
Funciones en la tabla de importación	- nombre de la dll - nombre - ordinal
Secciones	- nombre - posición en la tabla de secciones - posición física en el fichero - tamaño físico - posición en RVA - tamaño en RVA - características - entropía

Diseño e implementación de la base de datos

Determinados los datos a tener en cuenta, se diseñó una base de datos para guardar la información recolectada de las muestras de ficheros PE, malignas y no malignas. Esta base de datos tiene el objetivo de realizar posteriormente la búsqueda de anomalías que permitan identificar las características que distinguen a los programas malignos respecto a las muestras no malignas. Estas características son utilizadas como base para la elaboración de los algoritmos heurísticos. El diagrama simplificado de la estructura de datos se muestra en la figura 2. La base de datos fue implementada utilizando Microsoft SQL Server.

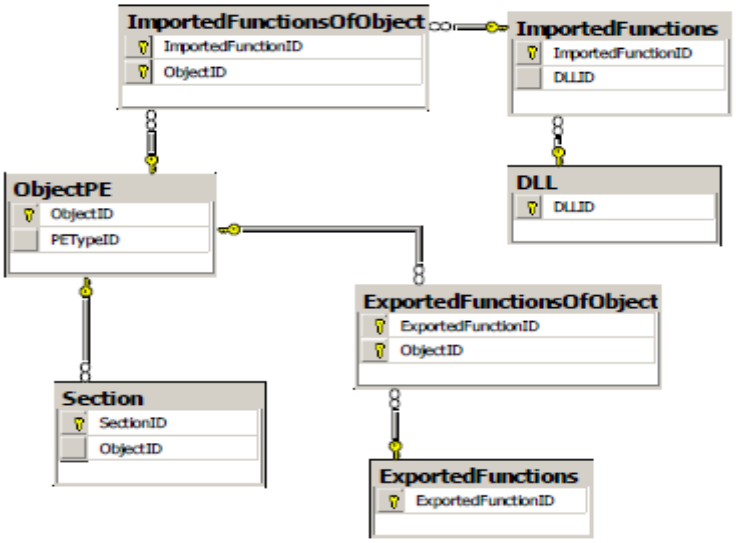

Figura 2. Estructura simplificada de la base de datos

Aplicación para el análisis y extracción de datos de los ficheros PE

Para el análisis y extracción de datos de los ficheros PE se diseñó e implementó la aplicación con interfaz de línea de comandos pe_detective.

Esta aplicación escanea un origen de muestras en busca de aquellas que son ficheros PE y realiza las siguientes acciones sobre estas muestras:

- extracción de los datos de interés de los ficheros PE
- análisis e identificación de anomalías presentes en las muestras
- guarda en la base de datos los resultados obtenidos

Entre los parámetros que se le indican a la aplicación se le indica si el origen del que se toman las muestras, es o no, un repositorio de programas malignos. Esto permite que sean identificadas correctamente en el interior de la base de datos.

Entre los objetivos tenidos en cuenta en el diseño de pe_detective, se encuentran que sea una aplicación modular, extensible y fácil de mantener.

Para lograr estos objetivos se diseñó la clase pe_context, que realiza la función de contenedor de los datos que son extraídos de las muestras de los ficheros PE y los resultados de los análisis realizados. Sirve de vínculo entre las diferentes etapas en el procesamiento de los ficheros PE hasta que los datos recolectados son enviados a la base de datos.

El primer paso en el análisis de una muestra es determinar si se trata de un fichero PE, en el caso afirmativo esta muestra pasa a la clase extractors_sequence que es la encargada de llamar a cada uno de los extractores que operan sobre los ficheros PE. Los datos obtenidos son enviados a pe_context. La figura 3, muestra el diseño simplificado de los extractores en pe_detective.

Cada uno de los extractores está diseñado como un módulo independiente encargado de tomar un grupo determinado de datos de la muestra, por ejemplo datos los datos de la tabla de exportación, la tabla de importación o los datos relativos a las secciones. Este diseño permite que cada uno de los extractores esté encargado de obtener datos específicos y para adicionar nuevos extractores o desactivar algún extractor existente solamente sea necesario adicionarlo o quitarlo de la secuencia de extractores vinculados a extractors_sequence.

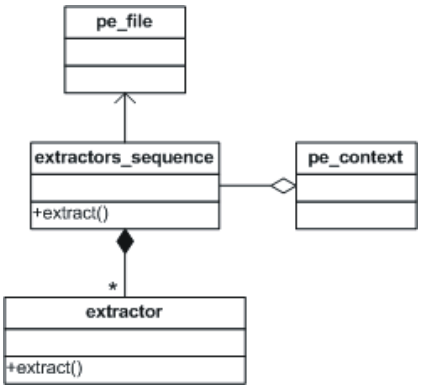

Figura 3. Extractores

El próximo paso en la secuencia de operaciones realizadas sobre una muestra, es pasarla a la clase analyzers_sequence. Esta clase es la encargada de llamar los distintos análisis que se efectúan en los datos de la muestra, los resultados de esos análisis son enviados a pe_context.

Mientras los extractores son los encargados de extraer directamente datos de la muestra, los analizadores procesan la información obtenida por medio de los extractores y que se encuentra previamente en pe_context.

De manera similar a los extractores, cada analizador realiza un análisis específico, como por ejemplo, la existencia de espacios entre las secciones en el fichero PE o la presencia de secciones sin nombre. La figura 4, muestra el diseño simplificado de los analizadores en pe_detective.

El diseño similar al de los extractores, permite que la adición de nuevos analizadores o la desactivación de analizadores existentes, sea realizada mediante su adición o eliminación de la secuencia de analizadores en analyzers_sequence.

Otra ventaja de este diseño, es que permite adicionar versiones preliminares de los métodos heurísticos, para realizar pruebas de detección.

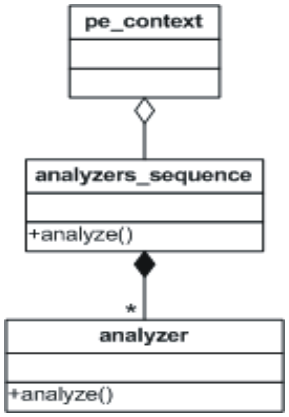

Figura 4. Analizadores

El último paso es almacenar en la base de datos los datos contenidos en pe_context, que han sido obtenidos mediante los extractores y los analizadores. Esto es realizado por la clase database_writer, que se muestra de forma simplificada en la figura 5. La ejecución de este paso puede ser desactivada desde las opciones en la línea de comandos.

Esta clase es la encargada de que los datos almacenados para cada muestra de ficheros PE sean únicos. Adiciona los datos para las nuevas muestras y actualiza los datos de las muestras ya almacenadas en la base de datos, en caso de que se amplíen los datos a extraer de las muestras o los análisis realizados.

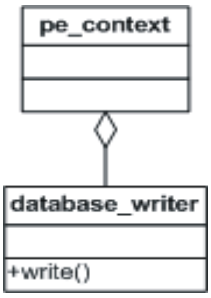

Figura 5. Database_writer

Determinación de algoritmos heurísticos

Llegado este punto, las muestras de ficheros PE, han sido tomadas de dos orígenes de datos que han sido identificados previamente como repositorios de programas malignos y no malignos. Estos dos orígenes de muestras le son entregados a pe_detective, que se encarga de extraer los datos de cada una de las muestras y guarda toda esa información en la base de datos.

Comienza entonces, por parte de los especialistas de Segurmática, la búsqueda patrones en los datos obtenidos que permitan separar las muestras malignas de las que no lo son.

Entre los patrones encontrados, se encuentran cadenas de texto ilegibles en zonas de los ficheros PE que de acuerdo con su estructura, deben contener cadenas de texto. Por ejemplo, nombres de secciones y nombres de dlls en la tabla de importación.

También se encuentran en estas zonas, cadenas de texto que con referencia sexuales, a la pornografía, al hacking, o que simplemente son palabras obscenas.

Es común encontrar entre las muestras malignas algunas que en su encabezado PE contienen referencia a la presencia dentro del fichero de la tabla de exportación o la tabla de importación y que estas no se encuentren en el fichero o los datos en las ubicaciones referenciadas no se correspondan con el formato de estas tablas.

Otro elemento común es la presencia de funciones importadas cuyos nombres son direcciones de memoria o funciones importadas mediante ordinales cuyos valores superan en gran medida los valores de los ordinales denciones de las funciones importadas en las muestras no malignas.

Teniendo en cuenta estos datos y otros de interés se procede a elaborar algoritmos heurísticos de detección. Estos algoritmos son diseñados de manera que cada uno tenga varios niveles de detección: normal, medio y alto.

Estos algoritmos se basan en las anomalías detectadas durante estos análisis. Como son algoritmos basados en la estructura de los ficheros PE, los ficheros que son detectados son clasificados como sospechosos. Esta misma característica, permite que sean algoritmos de rápida ejecución, debido a la facilidad con que son extraídos los datos necesarios de los ficheros procesados.

Prueba de los métodos heurísticos y puesta en producción

Antes de pasar a formar parte de Segurmática Antivirus, los métodos heurísticos son probados utilizando pe_detective. Con esta aplicación se realizan las primeras pruebas para asegurar que

solamente detecta muestra del repositorio de muestras de programas malignos y no detecta ninguna muestra de las que se conoce que no son malignas.

Una de las pruebas realizadas es el análisis de muestras que se conoce, son malignas, pero que no fueron tomadas en cuenta en el desarrollo de los algoritmos porque no se contaba con estas muestras. La tabla II, muestra el resultado de una de las pruebas realizadas contra muestras conocidas y desconocidas.

Tabla II: Resultado de prueba de detección

Origen	Muestras malignas	Detectadas	Porciento
Conocidas por Segav	505191	28553	5.7%
Desconocidas por Segav	233249	14919	6.4%

Segav: Segurmática Antivirus

No se tienen en cuenta muestras de programas no malignos porque estas pruebas se realizan cuando los algoritmos no detectan ninguna muestra no maligna en el repositorio.

Los métodos heurísticos ya probados, son adicionados a Segurmática Antivirus donde pasan por el control de calidad del producto y se comprueba nuevamente que no detectan ninguna de las muestras de programas no malignos.

En caso de detectar alguna muestra no maligna, se refina el algoritmo hasta que pase todos los controles.

Luego de que los algoritmos son adicionados a la actualización de Segurmática Antivirus, el proceso de refinamiento de los algoritmos continúa a partir de las muestras de programas no malignos que son detectados por los algoritmos en las computadoras de los clientes finales.

CONCLUSIONES

Los métodos heurísticos basados en la estructura de los ficheros PE, permiten encontrar gran cantidad de ficheros sospechosos de ser programas malignos, de forma rápida y mediante métodos relativamente simple manteniendo una tasa baja de detección de programas no malignos.

Mediante este trabajo:

• Se creó la infraestructura que sirve de base para la obtención de algoritmos heurísticos basados en la estructura de los ficheros PE de Segurmática Antivirus.

• Se obtuvieron varios algoritmos heurísticos que han sido adicionados a Segurmática Antivirus.

• Se Mejoró la capacidad de respuesta de Segurmática Antivirus ante amenazas desconocidas.

Esta es una labor que recién comienza, todavía no se han explorado todas las posibilidades que ofrece. Se hace necesaria la investigación de más características que lleven a la creación de nuevos algoritmos heurísticos. Tener en cuenta las anomalías de en los ficheros PE descritos en [5].

La búsqueda de características distintivas de los programas malignos ha sido realizada directamente por los especialistas de Segurmática. Se hace necesario realizar investigaciones con el objetivo de utilizar métodos de minería de datos para la búsqueda de características que conduzcan a nuevos algoritmos heurísticos.

REFERENCIAS

1. Gostev A. y C. Raiu: "Malware Evolution 2011", Kaspersky Security Bulletin, Marzo 2012.

2. Microsoft Portable Executable and Common Object File Format Specification, Rev. 8.2, 2010.

3. Sotirov, A.: Tiny PE, 2006.

4. Liskin, A.: PE: specification vs. loader, 2010.

5. Vuksan, M y T. Pericin: Undocumented PECOFF, BlackHat USA, 2011.

6. Tevis, J. E. J. and J. A. Hamilton Jr: "Static analysis of anomalies and security vulnerabilities in executable files", Proceedings of the 44th annual Southeast regional conference. ACM, pp. 560 - 565, 2006.

7. Schultz, M. G.; E. Eskin; F. Zadok and S. J. Stolfo: "Data mining methods for detection of new malicious executables", Proceedings 2001 IEEE Symposium, pp. 38 - 49, 2001.

8. Wang, J. H.; P. S. Deng; Y. S. Fan; L. J. Jaw; y Y. C. Liu: "Virus detection using data mining techinques", Proceedings IEEE 37th Annual 2003 International Carnahan Conference, pp. 71 - 76, 2003.

9. Ye, Y.; D. Wang; T. Li; D. Ye and Q. Jiang: "An intelligent PE-malware detection system based on association mining", Journal in computer virolo-gy, Vol.4, No.4, pp. 323 - 334, 2008.

10. Siddiqui, M. A.: Data mining methods for malware detection, ProQuest, Chicago, 2008.

11. Sami, A.; B. Yadegari; H. Rahimi; N. Peiravian; S. Hashemi; y A. Hamze: "Malware detection based on mining API calls", Proceedings of the 2010 ACM Symposium on Applied Compu-ting. ACM, pp. 1020 - 1025, 2010.