

Heurística Antivirus: detección proactiva de malware

Autor: Sebastián Bortnik, Analista en Seguridad de ESET para Latinoamérica

Fecha: lunes 1 de febrero de 2010

Índice

Introducción: ¿qué hace un antivirus?	3
Detección reactiva: base de firmas	4
Detección proactiva: heurística antivirus	5
Definición	6
Funcionamiento.....	7
Tipos de heurística	9
Precauciones	9
Evaluación de la detección heurística	10
Conclusión: heurística, un paso adelante	11

Introducción: ¿qué hace un antivirus?

Un **antivirus** es un software que posee la función de detectar códigos maliciosos. Aunque su nombre está relacionado con los virus informáticos, en la actualidad estos programas son soluciones **antimalware** que poseen protección contra gusanos, troyanos, rootkit, spyware y otros elementos dañinos; es decir, todo tipo de códigos maliciosos.

Además, un antivirus tiene como función identificar una amenaza. Esto se refiere a la capacidad de la aplicación no sólo de detectar un malware, sino también de describir de qué amenaza se trata, tanto por su tipo (virus, troyano, gusano, etc.) como su nombre (por ejemplo *Michelangelo*, *Conficker*, *QHost*, *Nuwar*, etc.).

Finalmente, una vez detectada e identificada cierta amenaza, un antivirus debe prevenir o eliminar la misma del sistema. En el primer caso se trata de un código malicioso que es detectado al momento de intentar infectar un sistema, por lo tanto el antivirus bloqueará su acceso y prevendrá la infección. En el otro caso, cuando se descubre el malware en un sistema que ya está infectado, el antivirus debe eliminar (o desinfectar) la amenaza.

Sin embargo, tal como se describe en este sencillo proceso de funcionamiento de un antivirus, el primer paso es la detección de un código malicioso. Para este fin el antivirus analiza los archivos (puede ser en tiempo real o a petición del usuario) en búsqueda de malware. En su visión simplificada, el antivirus examina cada archivo respondiendo a la pregunta: **¿es un código malicioso?**

El presente texto estudia cómo han evolucionado los sistemas de detección empleados para identificar si un archivo es o no una amenaza, complementando los clásicos procedimientos de detección reactivos, basados en firmas, con nuevas técnicas de detección proactivas, basadas en heurística.

Estas últimas permiten a los antivirus detectar malware nuevo o desconocido, y se explicará a lo largo del texto la necesidad de métodos de detección proactivos, sus principales características, funcionamiento, ventajas y desventajas.

El objetivo de este artículo es reducir las confusiones en torno al funcionamiento de la tecnología antimalware y clarificar qué es lo que realmente debe esperarse de una protección de este tipo, particularmente aquellas que cuentan con análisis heurístico.

Detección reactiva: base de firmas

Desde sus orígenes los antivirus cuentan con un **método de detección basado en firmas** (también llamadas vacunas). Este emplea una base de datos generada por el fabricante que permite determinar al software si un archivo es o no una amenaza. El sistema es sencillo: se coteja cada archivo a analizar con la base de datos y, si existe coincidencia (es decir, existe en la base una firma que se corresponde con el archivo), se identifica el archivo como código malicioso.

El proceso de generación de firmas se compone de los siguientes pasos:

1. Aparece un nuevo código malicioso
2. El laboratorio de la empresa antivirus recibe una muestra de ese código
3. Se crea la firma para el nuevo código malicioso
4. El usuario actualiza el producto con la nueva base de firmas y comienza a detectar el malware

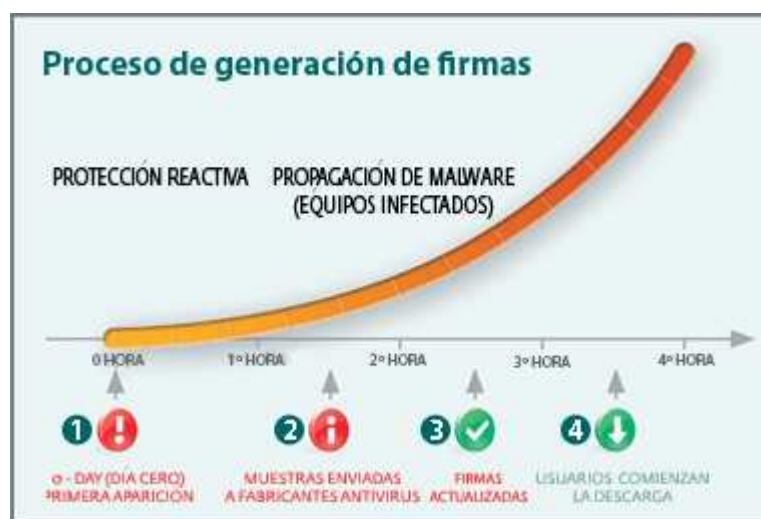


Imagen 1 – Proceso de generación de firmas

Recién a partir del último paso, el sistema estará protegido contra esta amenaza. Aquí radica la importancia de tener actualizado el antivirus: si la firma ya ha sido creada por el fabricante, pero no ha sido descargada en el sistema del usuario, el mismo no estará protegido contra esa amenaza en particular.

Además de la necesidad de mantener actualizada la base de datos, este método posee otras dos desventajas:

- El programa no puede detectar malware que no se encuentre en la base de datos
- El sistema debe contar con una firma por cada variante de un mismo código malicioso

La demora necesaria para generar una firma es variable, y depende del tiempo que tarde el malware en ser descubierto por el laboratorio, de las características del código malicioso y de la dificultad para generar la firma. De una u otra forma, se puede considerar que la demora puede oscilar entre las 2 y las 10 horas; aunque existen casos y excepciones que se escapan de este rango en ambos límites.

En conclusión, **la detección por firmas es un método de protección reactivo**: primero se debe conocer el malware para que luego sea detectado.

Sin embargo, debido a la alta velocidad de propagación de nuevos códigos maliciosos¹, y la gran cantidad de nuevas variantes que aparecen día a día², este método se volvió, con el pasar de los años, lento e insuficiente. De manera similar a lo que sucede durante las epidemias del campo biológico, en el caso del malware también existen probabilidades de que se produzca una infección antes de que aparezca la cura para dicha amenaza (la firma).

Un antivirus que utilice sólo métodos reactivos de detección estará protegiendo a sus usuarios sólo de aquellos códigos maliciosos que han sido incorporados a la base de datos, dejando siempre desprotegido al usuario frente a todas las variantes que sean desconocidas por el laboratorio del fabricante, o que aún no posean una firma.

DetECCIÓN PROACTIVA: HEURÍSTICA ANTIVIRUS

Para dar solución a esta problemática aparecen los métodos de detección proactivos basados en heurística, como complemento de la detección basada en firmas. Esto quiere decir que la detección proactiva es un agregado a la detección por firmas y para una óptima protección son necesarios ambos métodos, tal como trabajan las soluciones antimalware en la actualidad.

¹ Gusanos como Slammer han infectado miles de equipos en su primera hora de vida.

² Durante 2009, las compañías antivirus intercambiaron más de 2.500.000 muestras de malware

El objetivo esencial de los algoritmos heurísticos es dar respuestas en aquellas situaciones en donde los métodos reactivos no pueden darla: la capacidad de detectar un archivo malicioso aunque una muestra de éste no haya llegado al laboratorio antivirus, y que aún no se posea la firma correspondiente.

Definición

La etimología de la palabra **heurística** proviene del griego "*heurísko*", uno de cuyos significados es **encontrar**³.

La Real Academia Española la define como "*técnica de la indagación y del descubrimiento*"⁴. También otorga una segunda definición: "*En algunas ciencias, manera de buscar la solución de un problema mediante métodos no rigurosos, como por tanteo, reglas empíricas, etc.*".

Esta última definición es la que mejor se aplica a la utilización de heurística en tecnologías antivirus.

Por lo general la programación heurística es considerada como una de las aplicaciones de la inteligencia artificial y como herramienta para la resolución de problemas. Tal como es utilizada en sistemas expertos, la heurística se construye bajo reglas extraídas de la experiencia, y las respuestas generadas por tal sistema mejoran en la medida en que "aprende" a través del uso y aumenta su base de conocimiento.

La heurística siempre es aplicada cuando no puedan satisfacerse demandas de completitud que permitan obtener una solución por métodos más específicos (por ejemplo la creación de una firma para un malware determinado).

A manera de ejemplo, puede suponerse que un responsable de Recursos Humanos desea contratar un graduado de cierta carrera y se conecta con la universidad. La institución le ofrece un listado de 300 alumnos que se graduaron en los últimos años y él debe seleccionar a uno para su contratación. Su capacidad para realizar entrevistas es de 20 personas, por lo que debe tomar alguna decisión que le permita encontrar al candidato indicado. Una decisión heurística podría ser que se seleccione a los 20 alumnos con mejor promedio, lo cual probablemente le permita acercarse a los mejores candidatos. Sin embargo, lo ideal para el responsable de Recursos Humanos sería entrevistar a todos, ya que es probable que haya excelentes candidatos con promedios inferiores. Sin embargo, ante una limitación de completitud, las decisiones heurísticas permiten acercarse al resultado ideal.

³ De la misma palabra griega proviene el famoso 'eureka' ("lo encontré") mencionado por Arquímedes al encontrar la solución al problema de comprobar que una corona era realmente de oro.

⁴ <http://www.elpais.com/diccionarios/castellano/heur%C3%ADstica>

Si los alumnos fueran 20, sería posible entrevistar a todos y elegir sin lugar a dudas el que mejor haya pasado la entrevista. Sin embargo, en este caso es imposible entrevistar a los 300 alumnos y es por ello que se aplican métodos heurísticos.

De igual forma, con las tecnologías reactivas es imposible cubrir la protección necesaria para las condiciones actuales de evolución de amenazas, ya que no es posible contar con todos los códigos maliciosos que circulan por Internet, y tampoco se puede disminuir los tiempos de creación de firmas lo suficiente para asegurar protección total al usuario.

Ante estas imposibilidades la aplicación de heurística en tecnologías antivirus ofrece cobertura y protección inteligente ante códigos maliciosos.

Funcionamiento

Los algoritmos heurísticos son la base de la mayor parte de métodos de detección de malware proactivos.

El análisis heurístico posee un comportamiento basado en reglas para diagnosticar si un archivo es potencialmente ofensivo. El motor analítico trabaja a través de su base de reglas, comparando el contenido del archivo con criterios que indican un posible malware, y se asigna cierto puntaje cuando se localiza una semejanza. Si el puntaje iguala o supera un umbral determinado, el archivo es señalado como amenaza y procesado de acuerdo con ello.

De igual modo que un analista de malware intentaría determinar, trabajando en el laboratorio, la peligrosidad de un determinado programa, analizando sus acciones y características (por ejemplo: modifica el registro, se carga al inicio de sesión, elimina archivos, etc.), el análisis heurístico realiza el mismo proceso de toma de decisiones inteligentes, actuando como un investigador virtual de malware.

Existen diferentes métodos heurísticos (ver sección siguiente) que utilizan distintas reglas para determinar si un archivo es o no un código malicioso. Asimismo, un algoritmo de este tipo posee diferentes niveles de rigurosidad para determinar si un archivo es o no dañino. A mayor rigurosidad, mayor es la probabilidad de que se cometa un error en la detección, así como también mayor es la carga de procesamiento al momento del análisis (ver sección *Precauciones*). En el caso de los productos de ESET, se le permite al usuario seleccionar los métodos de detección y configurarlos según sus requerimientos:

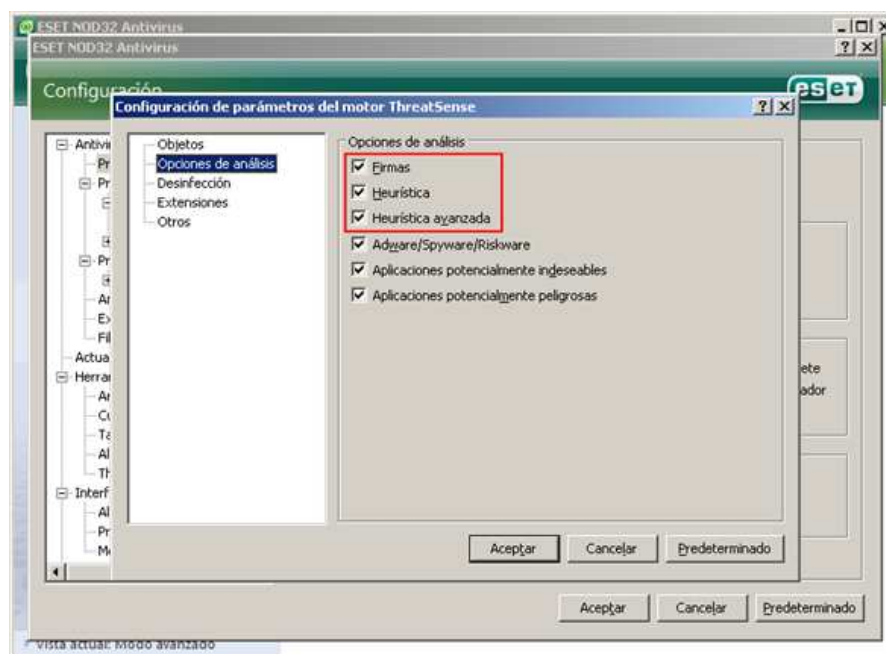


Imagen 2 – Configuración de detección en [ESET NOD32 Antivirus 4](#)

Mientras que la identificación de una amenaza realizada por medio de una detección reactiva basada en firmas posee la previa legitimación de una persona del laboratorio, la detección proactiva a través de métodos heurísticos no incluye la intervención humana, y en la detección posee un suficiente grado de certeza al respecto como para afirmar que un archivo es una amenaza. A pesar de esta aparente “desventaja”, los algoritmos heurísticos ofrecen protección donde la exploración por firmas no puede darla.

Aunque la detección proactiva no depende de la actualización de la base de firmas, sí debe mantenerse actualizado el programas antivirus, a fin de contar con los últimos algoritmos de detección heurística.

Tipos de heurística

Los algoritmos heurísticos, como su pluralidad lo indica, son distintas metodologías de análisis proactivo de amenazas. Se definen a continuación las tres variantes más comunes que son utilizadas en este tipo de análisis:

- **Heurística genérica:** se analiza cuán similar es un objeto a otro, que ya se conoce como malicioso. Si un archivo es lo suficientemente similar a un código malicioso previamente identificado, este será detectado como “*una variante de...*”.
- **Heurística pasiva:** ese explora el archivo tratando de determinar qué es lo que el programa intentará hacer. Si se observan acciones sospechosas, éste se detecta como malicioso.
- **Heurística activa:** se trata de crear un entorno seguro y ejecutar el código de forma tal que se pueda conocer cuál es el comportamiento del código. Otros nombres para la misma técnica son “*sandbox*”, “*virtualización*” o “*emulación*”.

Asimismo, los algoritmos de detección proactiva de amenazas contienen instrucciones que le permiten sortear diversos mecanismos que poseen los códigos maliciosos para ocultar su comportamiento, especialmente el empaquetamiento y el cifrado.

Precauciones

Al estar basados los algoritmos heurísticos en inteligencia artificial, poseen dos relativas desventajas, que deben ser eliminadas para hacer su funcionamiento más eficiente.

En primer lugar, al utilizar algoritmos inteligentes complejos, la carga de trabajo que posee el antivirus puede ser mayor que cuando se emplea el método basado en firmas (una simple exploración en una base de datos). Por lo tanto, es importante que los algoritmos de detección proactiva basados en heurística estén optimizados, a fin de que el rendimiento de la solución sea el máximo posible. En tal caso, incluso pueden ser más rápidos que una exploración por firma a medida que la base de datos del método reactivo vaya creciendo.

El otro factor de riesgo para los algoritmos de detección proactiva está constituido por los falsos positivos: archivos que no son códigos maliciosos, y son detectados como tales. Así como cualquier antivirus trabaja para minimizar los falsos negativos (es decir, amenazas que no son detectadas), en el caso de la heurística es necesario minimizar también los falsos positivos. El motivo de tal calificación incorrecta resulta comprensible: al trabajar estos algoritmos con grados de certeza, y análisis inteligente, puede ocurrir que se haga una detección errónea de un archivo. Con la base de firmas esto ocurre en un nivel muy bajo,

porque sólo se detectan amenazas conocidas que ya han sido catalogadas como tales por el laboratorio. Es indispensable, entonces, que los algoritmos de detección proactiva posean optimización, para minimizar la tasa de falsos positivos, ya que estos son altamente perjudiciales para los usuarios.

Por lo tanto, los algoritmos heurísticos de los antivirus deben ser evaluados, no sólo por sus capacidades de detección, sino también por su rendimiento y la cantidad de falsos positivos detectados. **La mejor heurística es aquella que combina los niveles de detección con bajos (o nulos) falsos positivos.**

Evaluación de la detección heurística

Así como las capacidades de una solución antivirus respecto a sus bases de firmas son evaluadas por diversas entidades certificadoras, las capacidades de detección proactivas también son sometidas a diversas evaluaciones para medir sus rendimientos.

Para medir las capacidades de detección proactivas, se utiliza como metodología la deshabilitación de actualizaciones de la base de firmas, y se evalúa una serie de códigos maliciosos para verificar si estos pueden ser detectados sólo por los algoritmos heurísticos.

Entre las entidades certificadoras, [Av-Comparatives](http://www.av-comparatives.org/)⁵ es la que mayor experiencia y prestigio posee en lo que respecta a evaluaciones de este tipo. [Av-Test](http://www.av-test.org/)⁶ también posee este tipo de certificaciones y [Virus Bulletin](http://www.virusbtn.com/)⁷ comenzará a realizarlas durante el año 2010.

⁵ <http://www.av-comparatives.org/>

⁶ <http://www.av-test.org/>

⁷ <http://www.virusbtn.com/>

Conclusión: heurística, un paso adelante

Desde la aparición de los primeros virus informáticos se ha descubierto una gran variedad de códigos maliciosos, cada vez en mayor cantidad y calidad (nuevos alcances y nuevas metodologías). En un principio la tecnología de detección de códigos maliciosos ha sido principalmente reactiva. Sin embargo, la evolución del desarrollo de malware hace imperiosa la necesidad de contar con soluciones antivirus con capacidades proactivas de detección.

La heurística antivirus permite disminuir la cantidad de infecciones en sistemas y redes, detectando malware nuevo (en la "ventana de tiempo" existente para liberar una firma) como así también variantes desconocidas por el fabricante.

Un antivirus debe contar con los métodos reactivos combinados con la heurística antivirus, que a su vez debe proporcionar buenos índices de rendimiento y una baja tasa de falsos positivos. La combinación de todas estas variantes permitirá al usuario confiar en una solución efectiva para las necesidades actuales en materia de protección contra malware.

Mientras que en los métodos reactivos el antivirus va un paso detrás de las amenazas, creando la firma luego de su descubrimiento, la heurística pretende situarse un paso por delante de los desarrolladores de malware. Este es el desafío al mantener algoritmos de detección proactivas: poder adelantarse a los creadores de códigos maliciosos, que están en constante movimiento y evolución.