|  |  |  |
| --- | --- | --- |
| Características de los sitios web de phishing | | |
| Rami M. Mohammad  Escuela de Computación e Ingeniería  Universidad de Huddersfield  Huddersfield, Reino Unido.  [rami.mohammad@hud.ac.uk](mailto:rami.mohammad@hud.ac.uk) | Fadi Thabtah  Departamento de Comercio Electrónico  Universidad Canadiense de Dubai  Dubai, Emiratos Árabes Unidos.  [fadi@cud.ac.ae](mailto:fadi@cud.ac.ae) | Lee McCluskey  Escuela de Computación e Ingeniería  Universidad de Huddersfield  Huddersfield, Reino Unido.  [tlmccluskey@hud.ac.uk](mailto:t.l.mccluskey@hud.ac.uk) |

### Características de los sitios web de phishing

Uno de los desafíos que enfrentó nuestra investigación fue la falta de disponibilidad de conjuntos de datos de entrenamiento confiables. De hecho, este desafío lo enfrenta cualquier investigador en el campo. Sin embargo, aunque hoy en día se han difundido muchos artículos sobre la predicción de sitios web de phishing utilizando técnicas de minería de datos, no se ha publicado públicamente ningún conjunto de datos de entrenamiento confiable, tal vez porque no hay acuerdo en la literatura sobre las características definitivas que caracterizan a los sitios web de phishing, por lo que es difícil para dar forma a un conjunto de datos que cubra todas las características posibles.

En este artículo, arrojamos luz sobre las características importantes que han demostrado ser sólidas y efectivas para predecir sitios web de phishing. Además, propusimos algunas características nuevas, asignamos experimentalmente nuevas reglas a algunas características conocidas y actualizamos algunas otras características.

### Funciones basadas en la barra de direcciones

#### Usando la dirección IP \*having\_IP\_Address

Si se utiliza una dirección IP como alternativa al nombre de dominio en la URL, como " <http://125.98.3.123/fake.html>", los usuarios pueden estar seguros de que alguien está intentando robar su información personal. En ocasiones, la dirección IP incluso se transforma a código hexadecimal como se muestra en el siguiente enlace “ <http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html>”.

*Regla* : SI

#### URL larga para ocultar la parte sospechosa \*URL\_Length

Los phishers pueden utilizar URL largas para ocultar la parte dudosa en la barra de direcciones. Por ejemplo:

[http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=\_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd410](http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8) 5e8 @phishing.sitio web.html

Para garantizar la precisión de nuestro estudio, calculamos la longitud de las URL en el conjunto de datos y generamos una longitud promedio de URL. Los resultados mostraron que si la longitud de la URL es mayor o igual a 54 caracteres, entonces la URL se clasifica como phishing. Al revisar nuestro conjunto de datos, pudimos encontrar 1220 longitudes de URL equivalentes a 54 o más, lo que constituye el 48,8% del tamaño total del conjunto de datos.

*Regla: SI*

Hemos podido actualizar esta regla de función utilizando un método basado en la frecuencia y mejorando así su precisión.

#### Uso de los servicios de acortamiento de URL "TinyURL" \*Shortining\_Service

El acortamiento de URL es un método en la "World Wide Web" en el que una URL puede reducirse considerablemente y aún así conducir a la página web requerida. Esto se logra mediante un “redireccionamiento HTTP” en un nombre de dominio corto, que enlaza a la página web que tiene una URL larga. Por ejemplo, la URL "http://portal.hud.ac.uk/" se puede acortar a "bit.ly/19DXSk4".

*Regla* : SI

#### Las URL tienen el símbolo "@" \*having\_At\_Symbol

El uso del símbolo "@" en la URL hace que el navegador ignore todo lo que precede al símbolo "@" y la dirección real a menudo sigue al símbolo "@".

Regla: SI

#### Redirigir usando “//” \*double\_slash\_redirecting

La existencia de “//” dentro de la ruta URL significa que el usuario será redirigido a otro sitio web. Un ejemplo de este tipo de URL es: “http://www.legitimate.com//http://www.phishing.com”. Examinamos la ubicación donde aparece “//”. Encontramos que si la URL comienza con “HTTP”, eso significa que “//” debería aparecer en la sexta posición. Sin embargo, si la URL emplea "HTTPS", entonces "//" debería aparecer en la séptima posición.

Regla: SI

#### Agregar prefijo o sufijo separado por (-) al dominio \*Prefix\_Suffix

El símbolo de guión rara vez se utiliza en URL legítimas. Los phishers tienden a añadir prefijos o sufijos separados por (-) al nombre de dominio para que los usuarios sientan que están ante una página web legítima. Por ejemplo http://www.Confirme-paypal.com/.

Regla: SI

#### Subdominio y múltiples subdominios \*having\_Sub\_Domain

Supongamos que tenemos el siguiente enlace: http://www.hud.ac.uk/students/. Un nombre de dominio puede incluir dominios de nivel superior de código de país (ccTLD), que en nuestro ejemplo es "uk". La parte "ac" es una abreviatura de "académico", la combinación "ac.uk" se denomina dominio de segundo nivel (SLD) y "hud" es el nombre real del dominio. Para generar una regla para extraer esta característica, primero debemos omitir (www.) de la URL, que de hecho es un subdominio en sí mismo. Luego, tenemos que eliminar el (ccTLD ) si existe. Finalmente, contamos los puntos restantes. Si el número de puntos es mayor que uno, entonces la URL se clasifica como "sospechosa" ya que tiene un subdominio. Sin embargo, si los puntos son mayores que dos, se clasifica como “Phishing” ya que tendrá múltiples subdominios. De lo contrario, si la URL no tiene subdominios, asignaremos "Legítimo" a la función.

Regla: SI

#### HTTPS (Protocolo de transferencia de hipertexto con capa de sockets seguros) \*SSLfinal\_State

La existencia de HTTPS es muy importante para dar la impresión de legitimidad del sitio web, pero claramente no es suficiente. Los autores en(Mohammad, Thabtah and McCluskey 2012) (Mohammad, Thabtah and McCluskey 2013)Sugerimos verificar el certificado asignado con HTTPS, incluido el alcance del emisor del certificado de confianza y la antigüedad del certificado. Las autoridades de certificación que figuran constantemente entre los nombres más confiables incluyen: “GeoTrust, [GoDaddy](http://www.godaddy.com/gdshop/ssl/ssl.asp?isc=BESTSSL1) , Network Solutions, Thawte, Comodo, Doster y VeriSign”. Además, al probar nuestros conjuntos de datos, descubrimos que la antigüedad mínima de un certificado acreditado es de dos años.

Regla: SI

#### Longitud del registro de dominio \*Domain\_registeration\_length

Partiendo del hecho de que un sitio web de phishing dura un corto período de tiempo, creemos que los dominios confiables se pagan regularmente con varios años de anticipación. En nuestro conjunto de datos, encontramos que los dominios fraudulentos más largos se han utilizado sólo durante un año.

Regla: SI

#### Favicon \*Favicon

Un favicon es una imagen gráfica (icono) asociada a una página web específica. Muchos agentes de usuario existentes, como navegadores gráficos y lectores de noticias, muestran el favicon como un recordatorio visual de la identidad del sitio web en la barra de direcciones. Si el favicon se carga desde un dominio distinto al que se muestra en la barra de direcciones, es probable que la página web se considere un intento de phishing.

Regla: SI

#### Usando un puerto no estándar \*port

Esta característica es útil para validar si un servicio en particular (por ejemplo, HTTP) está activo o inactivo en un servidor específico. Para controlar las intrusiones, es mucho mejor simplemente abrir los puertos que necesite. Varios firewalls, servidores proxy y de traducción de direcciones de red (NAT ), bloquearán, de forma predeterminada, todos o la mayoría de los puertos y solo abrirán los seleccionados. Si todos los puertos están abiertos, los phishers pueden ejecutar casi cualquier servicio que quieran y, como resultado, la información del usuario se ve amenazada. Los puertos más importantes y su estatus preferido se muestran en la Tabla 2.

Regla: SI

Tabla Puertos comunes a verificar

|  |  |  |  |
| --- | --- | --- | --- |
| PUERTO | Servicio | Significado | Estado preferido |
| 21 | ftp | Transferir archivos de un host a otro | Cerca |
| 22 | SSH | Protocolo de transferencia segura de archivos | Cerca |
| 23 | Telnet | Proporcionar una comunicación bidireccional interactiva orientada a texto. | Cerca |
| 80 | HTTP | Protocolo de transferencia de hiperprueba | Abierto |
| 443 | HTTPS | Protocolo de transferencia de hipertexto seguro | Abierto |
| 445 | PYME | Proporcionar acceso compartido a archivos, impresoras y puertos serie. | Cerca |
| 1433 | MSSQL | Almacene y recupere datos según lo soliciten otras aplicaciones de software. | Cerca |
| 1521 | ORÁCULO | Acceda a la base de datos Oracle desde la web. | Cerca |
| 3306 | mysql | Acceda a la base de datos MySQL desde la web. | Cerca |
| 3389 | Escritorio remoto | permitir el acceso remoto y la colaboración remota | Cerca |

#### La existencia del token "HTTPS" en la parte de dominio de la URL \*HTTPS\_token

Los phishers pueden agregar el token "HTTPS" a la parte del dominio de una URL para engañar a los usuarios. Por ejemplo,   
http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.

Regla: SI

### Características basadas en anomalías \*

#### Solicitar URL \*Request\_URL

La URL de solicitud examina si los objetos externos contenidos en una página web, como imágenes, videos y sonidos, se cargan desde otro dominio. En las páginas web legítimas, la dirección de la página web y la mayoría de los objetos incrustados en la página web comparten el mismo dominio.

Regla: SI

#### URL del ancla \*URL\_of\_Anchor

Un ancla es un elemento definido por la etiqueta <a>. Esta función se trata exactamente como "Solicitar URL". Sin embargo, para esta característica examinamos:

1. Si las etiquetas <a> y el sitio web tienen nombres de dominio diferentes. Esto es similar a la función de solicitud de URL.
2. Si el ancla no enlaza a ninguna página web, por ejemplo:
3. <a href=“#”>
4. <a href=“#contenido”>
5. <a href=“#saltar”>
6. <a href=“JavaScript ::void(0)”>

*Regla* : SI

#### Enlaces en etiquetas <Meta>, <Script> y <Link> \*Links\_in\_tags

Dado que nuestra investigación cubre todos los ángulos que probablemente se utilizarán en el código fuente de la página web, encontramos que es común que los sitios web legítimos utilicen etiquetas <Meta> para ofrecer metadatos sobre el documento HTML; Etiquetas <Script> para crear un script del lado del cliente; y etiquetas <Link> para recuperar otros recursos web. Se espera que estas etiquetas estén vinculadas al mismo dominio de la página web.

Regla: SI

#### Controlador de formularios del servidor (SFH) \*SFH

Los SFH que contienen una cadena vacía o "acerca de: en blanco" se consideran dudosos porque se debe tomar una acción sobre la información enviada. Además, si el nombre de dominio en SFH es diferente del nombre de dominio de la página web, esto revela que la página web es sospechosa porque la información enviada rara vez es manejada por dominios externos.

Regla: SI

#### Envío de información al correo electrónico \*Submitting\_to\_email

El formulario web permite al usuario enviar su información personal que se dirige a un servidor para su procesamiento. Un phisher podría redirigir la información del usuario a su correo electrónico personal. Para ello, se podría utilizar un lenguaje de script del lado del servidor, como la función “mail()” en PHP. Una función más del lado del cliente que podría usarse para este propósito es la función "mailto:".

Regla: SI

#### URL anormal \*Abnormal\_URL

Esta característica se puede extraer de la base de datos WHOIS. Para un sitio web legítimo, la identidad suele ser parte de su URL.

Regla: SI

### Funciones basadas en HTML y JavaScript

#### Reenvío de sitios web \*Redirect

La delgada línea que distingue los sitios web de phishing de los legítimos es cuántas veces se ha redirigido un sitio web. En nuestro conjunto de datos, encontramos que los sitios web legítimos han sido redirigidos una vez como máximo. Por otro lado, los sitios web de phishing que contienen esta función han sido redirigidos al menos 4 veces.

Regla: SI

#### Personalización de la barra de estado \*on\_mouseover

Los phishers pueden usar JavaScript para mostrar a los usuarios una URL falsa en la barra de estado. Para extraer esta característica, debemos desenterrar el código fuente de la página web, particularmente el evento "onMouseOver", y verificar si realiza algún cambio en la barra de estado.

Regla: SI

#### Deshabilitar el clic derecho \*RightClick

Los phishers utilizan JavaScript para desactivar la función de clic derecho, de modo que los usuarios no puedan ver ni guardar el código fuente de la página web. Esta característica se trata exactamente como "Usar onMouseOver para ocultar el enlace". No obstante, para esta función, buscaremos el evento "event.button==2" en el código fuente de la página web y comprobaremos si el clic derecho está deshabilitado.

Regla: SI

#### Usando la ventana emergente \*popUpWidnow

Es inusual encontrar un sitio web legítimo que solicite a los usuarios que envíen su información personal a través de una ventana emergente. Por otro lado, esta función se ha utilizado en algunos sitios web legítimos y su objetivo principal es advertir a los usuarios sobre actividades fraudulentas o emitir un anuncio de bienvenida, aunque no se solicita que se rellene información personal a través de estas ventanas emergentes.

Regla: SI

#### Redirección de marco flotante \*Iframe

IFrame es una etiqueta HTML que se utiliza para mostrar una página web adicional a la que se muestra actualmente. Los phishers pueden utilizar la etiqueta “iframe” y hacerla invisible, es decir, sin bordes de marco. En este sentido, los phishers utilizan el atributo "frameBorder" que hace que el navegador muestre una delimitación visual.

Regla: SI

### Funciones basadas en dominio

#### Edad del dominio \*age\_of\_domain

Esta característica se puede extraer de la base de datos WHOIS (Whois 2005). La mayoría de los sitios web de phishing duran un corto período de tiempo. Al revisar nuestro conjunto de datos, encontramos que la edad mínima del dominio legítimo es de 6 meses.

Regla: SI

#### Registro DNS \*DNSRecord

Para los sitios web de phishing, la base de datos WHOIS no reconoce la identidad reclamada (Whois 2005)o no se encuentran registros para el nombre de host (Pan and Ding 2006). Si el registro DNS está vacío o no se encuentra, entonces el sitio web se clasifica como “Phishing”, de lo contrario se clasifica como “legítimo”.

Regla: SI

#### Tráfico del sitio web \*web\_traffic

Esta función mide la popularidad del sitio web determinando la cantidad de visitantes y la cantidad de páginas que visitan. Sin embargo, dado que los sitios web de phishing viven por un corto período de tiempo, es posible que la base de datos de Alexa no los reconozca (Alexa the Web Information Company., 1996). Al revisar nuestro conjunto de datos, descubrimos que, en los peores escenarios, los sitios web legítimos se ubicaban entre los 100.000 mejores. Además, si el dominio no tiene tráfico o no es reconocido por la base de datos de Alexa, se clasifica como “Phishing”. De lo contrario, se clasifica como “sospechoso”.

Regla: SI

#### Rango de página \*Page\_Rank

PageRank es un valor que va de “0” a “1”. PageRank tiene como objetivo medir la importancia de una página web en Internet. Cuanto mayor sea el valor del PageRank, más importante será la página web. En nuestros conjuntos de datos, encontramos que alrededor del 95% de las páginas web de phishing no tienen PageRank. Además, encontramos que el 5% restante de las páginas web de phishing pueden alcanzar un valor de PageRank de hasta “0,2”.

Regla: SI

#### Índice de Google \*Google\_Index

Esta función examina si un sitio web está en el índice de Google o no. Cuando Google indexa un sitio, se muestra en los resultados de búsqueda (Webmaster resources, 2014). Por lo general, solo se puede acceder a las páginas web de phishing durante un período breve y, como resultado, es posible que muchas páginas web de phishing no se encuentren en el índice de Google.

Regla: SI

#### Número de enlaces que apuntan a la página \*Links\_pointing\_to\_page

El número de enlaces que apuntan a la página web indica su nivel de legitimidad, incluso si algunos enlaces son del mismo dominio (Dean, 2014). En nuestros conjuntos de datos y debido a su corta vida útil, encontramos que el 98% de los elementos del conjunto de datos de phishing no tienen enlaces que apunten a ellos. Por otro lado, los sitios web legítimos tienen al menos 2 enlaces externos que apuntan a ellos.

Regla: SI

#### Función basada en informes estadísticos \*Statistical\_report

Varias partes, como PhishTank (PhishTank Stats, 2010-2012)y StopBadware, (StopBadware, 2010-2012)formulan numerosos informes estadísticos sobre sitios web de phishing en cada período de tiempo; algunos son mensuales y otros son trimestrales. En nuestra investigación, utilizamos 2 formas de las diez estadísticas principales de PhishTank: "10 dominios principales" y "10 IP principales" según los informes estadísticos publicados en los últimos tres años, desde enero de 2010 hasta noviembre de 2012. Mientras que para "StopBadware ”, utilizamos direcciones IP del “Top 50”.

Regla: SI