DARK WEB Y DEEP WEB COMO FUENTES DE CIBERINTELIGENCIA UTILIZANDO MINERÍA DE DATOS

EVA MARTÍN IBÁÑEZ

DOCTORA EN CIENCIAS DE LA INFORMACIÓN

RESUMEN

Deep Web es la parte más grande de Internet, cuyos contenidos no pueden indexar los buscadores web convencionales; puede servir para actividades legales e ilegales. Dark Web ocupa las zonas más oscuras de Deep Web y requiere de herramientas específicas de acceso. Ambas constituyen una importante fuente de ciberinteligencia, especialmente sobre amenazas, vulnerabilidades y riesgos. La minería de datos, en sentido amplio, puede ayudar a encontrar sentido a las ingentes cantidades de datos existentes en Deep Web y Dark Web. Esas técnicas permiten el análisis cuasi automático de conjuntos de datos enormes y complejos para desvelar patrones e identificar tendencias.

Palabras clave: Deep Web, Dark Web, ciberinteligencia, darknet, minería de datos, knowledge discovery

ABSTRACT

The Deep Web is the largest part of the Internet, whose contents are not indexed by standard search engines; it can be used for legal and illegal activities. The Dark Web takes up the darkest corners in the Deep Web and requires specialized tools to access. Both represent an important source of cyber intelligence, mainly in regards to threats, vulnerabilities and risks. Data mining can help to make sense of massive amounts of data from the Deep Web and the Dark Web. Those techniques allow quasi-automatic analysis of large and complex data sets to unveil patterns and identify trends.

Keywords: Deep Web, Dark Web, cyber intelligence, darknet, data mining, knowledge discovery

1. INTRODUCCIÓN

La minería de datos (*data mining*), en sentido amplio, puede ayudar a mejorar las capacidades de lucha contra el crimen organizado y el terrorismo, porque contribuye a reducir la sobrecarga informativa y cognitiva de las personas. Ofrece un gran potencial a la hora de extraer conocimiento implícito en los datos. Miembros de los cuerpos y fuerzas de seguridad y de la comunidad de la inteligencia pueden beneficiarse de estos métodos y técnicas para encontrar sentido a los datos y para presentar los resultados eficazmente a los decisores.

La tremenda cantidad de datos, además en rápido crecimiento, excede la capacidad de comprensión humana. Eso conduce a la situación actual, rica en datos pero pobre en información. Esa brecha entre datos e información, que está en constante

ensanchamiento, requiere el desarrollo de herramientas potentes que puedan convertir los datos en perlas de conocimiento (Han, Kamber y Pei, 2012, p. 5).

Deep Web y Dark Web constituyen unas fuentes muy relevantes para la labor de la ciberinteligencia. Allí es posible encontrar información valiosa sobre amenazas, vulnerabilidades y riesgos. El problema es que las cantidades de datos que hay que recopilar y analizar son tan enormes que sobrepasan los métodos tradicionales de análisis y vigilancia.

Una posible solución a la sobrecarga informativa es recurrir a herramientas como la minería de datos o el Knowledge Discovery in Database (KDD). Esas técnicas permiten el análisis cuasi automático de conjuntos de datos enormes y complejos para identificar tendencias y patrones previamente desconocidos. El objetivo final es generar información válida para tomar decisiones o para aportar como prueba en un procedimiento judicial.

Las páginas siguientes están estructuradas en varios apartados. Para empezar, se aclaran las definiciones de Deep Web, Dark Web, Surface Web, ciberespacio y minería de datos. Seguidamente, se comenta por qué interesa analizar Deep Web y Dark Web desde el punto de vista de la seguridad. Después se explican qué dificultades presentan estos ámbitos para la minería de datos. A continuación figuran ejemplos de uso, clasificados en tres grandes áreas: las arañas (*crawlers*) para Deep Web, los sistemas de detección y prevención de intrusiones (IDPS) y la detección de comunidades. Las conclusiones ponen el cierre.

2. DEFINICIONES

El ciberespacio designa el “dominio global y dinámico compuesto por las infraestructuras de tecnología de la información –incluida Internet–, las redes y los sistemas de información y de telecomunicaciones” (Gobierno de España, 2013, p. 9). Dentro de Internet, existe una cierta confusión de términos. A menudo Deep Web (Internet Profunda) y Dark Web (Internet Oscura) se usan indistintamente. Sin embargo, es conveniente diferenciarlos.

Deep Web es aquella parte de Internet que no es accesible a los motores de búsqueda basados en enlaces como Google. La única manera de acceder a ella es introducir una consulta directa en un formulario de búsqueda web. De esa forma, se pueden recuperar contenidos dentro de una base de datos que no está enlazada (Pederson, 2013, p. 2). En cambio, Surface Web (Internet Superficial) sí es accesible a través de técnicas de rastreo web basadas en enlaces, que conducen a datos localizables vía hiperenlaces desde la página principal de un dominio. Buscadores como Google, Bing o Yahoo pueden encontrar esos datos en la Internet Superficial (Pederson, 2013, p. 2).

Deep Web se refiere a cualquier contenido de Internet que, por diversos motivos, no puede ser indexado por los buscadores. Incluye páginas web dinámicas, sitios bloqueados (como los que requieren responder un CAPTCHA para acceder), sitios no enlazados, sitios privados (que necesitan credenciales para entrar), contenidos que no son HTML, contextuales o con scripts, y redes de acceso limitado. Las redes de acceso limitado están formadas por nombres de dominio registrados en sistemas de nombres de dominio (DNS) no gestionados por ICANN (Internet Corporation for Assigned Names and Numbers) y por direcciones URL (Uniform Resource Locator) con dominios de

primer nivel o TLD (Top-Level Domains) no estandarizados que generalmente requieren un servidor DNS específico para resolver correctamente. Un ejemplo de redes de acceso limitado son los sitios con dominios registrados en sistemas distintos del estándar DNS, como los .BIT. Esos sitios no solamente escapan de las regulaciones impuestas por la ICANN, sino que, debido a su naturaleza descentralizada, son muy complicados de desviar a un sumidero. Bajo la categoría de redes de acceso limitado también se encuentran las Darknets o sitios alojados en infraestructuras que requieren el uso de software específico como Tor para acceder. Precisamente la mayor parte de las actividades de interés público dentro de Deep Web ocurren dentro de las Darknets (Ciancaglini, Balduzzi, McArdle y Rösler, 2015, p. 5).

Un estudio de 2001, cuando solo había unos tres millones de dominios en Internet, estima que el tamaño de Deep Web es aproximadamente entre 400 y 550 veces mayor que el de Surface Web. Por aquella época, Deep Web contenía 750 Terabytes de información frente a los 19 Terabytes de Surface Web, y el 95% de Deep Web era públicamente accesible, en el sentido de no requerir cuotas de suscripción (Bergman, 2001, p. 1).

Deep Web es la parte más grande de Internet y puede usarse para el bien y para el mal, para actividades legales y para actividades ilegales. Conviene saber que no todo es malo. Hay muchos aspectos buenos en Deep Web, incluyendo el derecho a la privacidad cuando se navega por Internet (Hawkins, 2016, p. 17).

Dark Web no es lo mismo que Deep Web. Dark Web se refiere a cualquier página web que se oculta a plena vista o que reside dentro de una capa pública pero separada de la Internet estándar. Por ejemplo, una página web que carece de enlaces de entrada, de manera que ni los usuarios ni los motores de búsqueda pueden localizarla (Pederson, 2013, p. 3). En definitiva, Dark Web es una parte de Deep Web. Si se adopta la metáfora de los túneles de una mina, Dark Web ocuparía las zonas más profundas de Deep Web que requieren herramientas o equipamiento altamente especializados para acceder a ellas. Reside en los subterráneos más profundos, y los dueños de los sitios tienen más razones para mantener sus contenidos ocultos (Ciancaglini et al., 2015, p. 6).

Las Darknets modernas necesitan software específico para usar la red distribuida. Hoy en día los ejemplos más notables son Tor, I2P (Invisible Internet Project) y Freenet. La arquitectura fluida de estas redes complica estimar su tamaño, pero parece que Tor es la más grande con I2P a bastante distancia. El resto son mucho más pequeñas en alcance y popularidad (Moore y Rid, 2016, p. 15).

Otros conceptos relacionados son Dark Net o Darknet, que frecuentemente se utilizan como equivalentes de Dark Web. En realidad, hay varias definiciones de Darknet (Fachkha y Debbabi, 2016, p. 1198). La primera es cualquier sistema de comunicación que opera furtivamente y oculta la identidad de sus usuarios, como pueden ser Freenet y BitTorrent. La segunda se refiere a servidores y programas que sirven para distribuir ilegalmente material protegido por derechos de autor, como las tecnologías P2P (Peer-to- Peer). La tercera definición remite a servidores configurados para atrapar adversarios y recopilar datos sospechosos. Este último tipo de Darknet funciona en modo pasivo sin interactuar con los atacantes; corresponde a dispositivos y servidores sin utilizar; también se conoce como Darkspace o como direcciones IP sin usar. En resumen, Darknet, según esa tercera acepción, es un sistema de monitorización con trampas que funciona de modo pasivo. Su tecnología está diseñada para inferir actividades y amenazas en Internet (Fachkha y Debbabi, 2016, p. 1223).

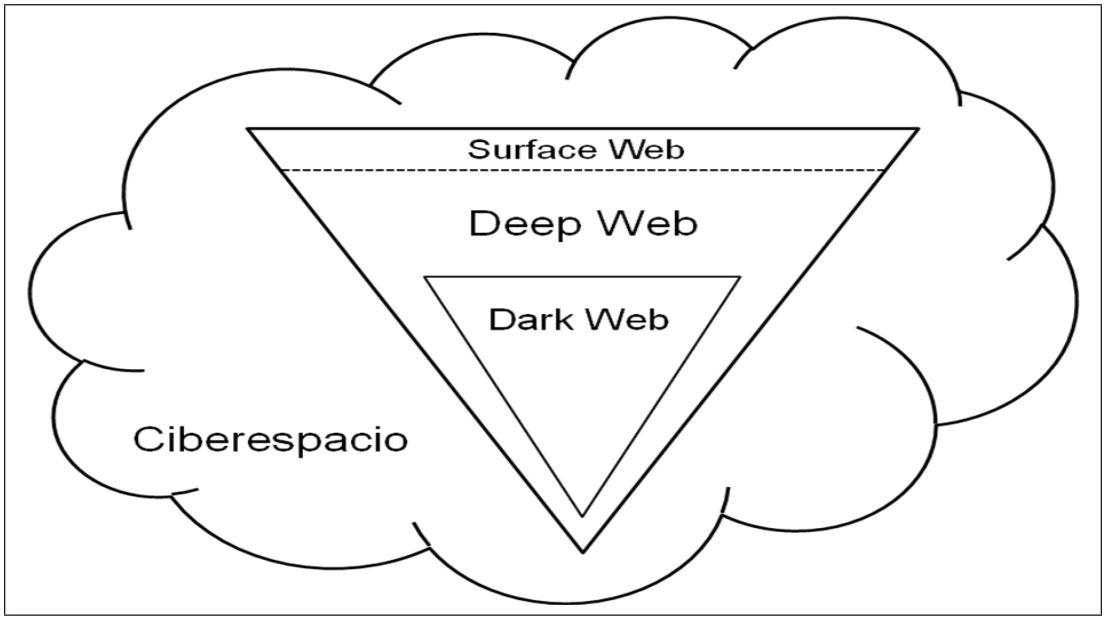

Figura 1. Ciberespacio, Deep Web y Dark Web. Fuente: Elaboración propia.

El tamaño de Darknet se estima en 270.000 direcciones IP (Internet Protocol) a mayo de 2015 (Nakao, 2016, p. 4). Los ataques en Darknet a través del puerto 23, que corresponde el servicio Telnet, se han disparado en 2015, según el NICT japonés (National Institut of Information and Communications Technology). El principal origen de dichos ataques son dispositivos IoT (Internet de las Cosas), con más de 150.000 direcciones IP atacantes y 361 modelos de dispositivos IoT observados en sólo cuatro meses (Nakao, 2016, p. 5).

Existen muchos motivos, aparte de comprar drogas, por los que la gente quiere mantenerse en el anonimato o instalar sitios en línea que no se puedan vincular con un lugar físico o con una entidad (Ciancaglini et al., 2015, p. 6). Por ejemplo, informantes o disidentes políticos pueden estar interesados en proteger sus comunicaciones, sobre todo en países con regímenes políticos represivos. Dark Web también es útil para periodistas y para activistas pro derechos humanos que en ciertos lugares del mundo pueden sufrir amenazas de cárcel o censura. Aunque se busquen soluciones para combatir las actividades ilegales y nefarias en Dark Web, eso no debería perjudicar las actividades legales y legítimas de libertad de expresión (Weimann, 2016, p. 43).

Finalmente, el término minería de datos a menudo se usa para referirse a todo el proceso de Knowledge Discovery in Database (KDD)[1]. Así, la minería de datos, entendida en sentido amplio, sería el proceso de descubrir conocimientos y patrones interesantes en grandes cantidades de datos. Las fuentes de datos pueden ser bases de datos, almacenes de datos, la web, otros repositorios de información o

1 Knowledge Discovery in Database (KDD) incluye varias etapas: Limpieza de datos (para eliminar datos inconsistentes y ruidosos); integración de datos (donde se pueden combinar varias fuentes); selección de datos (se recuperan de la base de datos los que son relevantes para el análisis); transformación de datos (los datos se agregan o se consolidan en formas adecuadas para realizar después el minado); minería de datos (proceso esencial en el que se aplican métodos de inteligencia para extraer patrones de datos); evaluación de patrones (para identificar aquellos patrones que son realmente interesantes y que suponen conocimiento); y presentación del conocimiento (mediante técnicas de visualización y representación para mostrar el conocimiento extraído a los usuarios) (Han et al, 2012, pp. 7-8).

datos transmitidos al sistema de forma dinámica (Han et al., 2012, p. 7). Abarca una amplia variedad de técnicas de diversos campos como la estadística, el aprendizaje automático (*machine learning*), el reconocimiento de patrones, las bases de datos, la recuperación de información, la visualización, los algoritmos o la computación de alto rendimiento, entre otros (Han et al., 2012, p. 23).

Muchas de esas técnicas de Knowledge Discovery se pueden aplicar a los estudios de seguridad, teniendo en cuenta sus peculiaridades. Las más empleadas se pueden clasificar en estas categorías: compartición de información, análisis de asociaciones criminales, clasificación y clusterizado del crimen, análisis de inteligencia y análisis espacio-temporal de delitos (Chen, 2012, p. 26).

3. POR QUÉ INTERESA

En los ámbitos de la seguridad y la defensa, Deep Web en general y Dark Web en particular constituyen una importante de fuente de ciberinteligencia, especialmente sobre amenazas, vulnerabilidades y riesgos.

El uso de Deep Web puede dividirse en dos clases: actividades legales y actividades ilegales. Con independencia de que el uso sea legal o ilegal, el acto de acceder a Deep Web siempre implica una acción deliberada (Hawkins, 2016, p. 7).

Aunque no lo parezca, hay muchas actividades legales que se desarrollan en Deep Web. Es un recurso útil para multitud de información. Por ejemplo, hay bases de datos con librerías académicas virtuales y con versiones antiguas de páginas web (Hawkins, 2016, p. 7). Junto a estos contenidos legales hay otros que son irregulares o que son completamente ilícitos.

Una muestra de los productos y servicios ilegales que están disponibles dentro de las Darknets abarca los siguientes: contenidos pirateados; drogas; dinero falsificado; productos de lujo robados; tarjetas de crédito y cuentas bancarias; robo de identidad; pasaportes y otros documentos oficiales; armas, munición y explosivos; servicios de mercenarios y asesinos a sueldo; contenidos de abuso sexual a menores; tráfico de seres humanos (adultos y menores); y tráfico de órganos (Goodman, 2015, pp. 201-204).

Un reciente estudio realizado con 300.000 direcciones de la red Tor revela que solamente algo más de la mitad (el 52,3%) están activas. Entre los sitios activos, el 56,8% están dedicados a actividades ilícitas (Moore y Rid, 2016, p. 21). Los contenidos se pueden clasificar en 13 categorías: armas, drogas, extremismos, finanzas, hacking, pornografía ilegal, nexos, otros ilícitos, sociales, violencia, otros lícitos, ninguno y desconocido.

- Armas: Armas y municiones.
- Drogas: Drogas ilegales o medicamentos ilícitos.
- Extremismos: Ideologías extremistas, incluyendo expresiones de apoyo al terrorismo, guías prácticas militantes y foros extremistas.
- Finanzas: Blanqueo de dinero, moneda falsificada y venta de cuentas y tarjetas de crédito robadas.
- Hacking: Hackers de alquiler y distribución de malware o ataques DDoS.

- Pornografía ilegal: Material pornográfico que involucra menores, violencia, animales o materiales obtenidos sin el consentimiento de los participantes.
- Nexos: Dedicados a enlaces con otros sitios ilícitos y recursos dentro de Darknets.
- Otros ilícitos: Materiales que no entran en las categorías anteriores pero que son problemáticos, como la venta de carnets y pasaportes falsificados.
- Sociales: Comunidades en línea para compartir material ilícito en forma de foros, redes sociales y tablones de mensajes.
- Violencia: Asesinos a sueldo e instrucciones sobre cómo realizar ataques violentos.
- Otros lícitos: Servicios legítimos, como contenidos de tipo ideológico o político, puntos seguros de entrega y recogida y repositorios de información.
- Ninguno: Sitios que son completamente inaccesibles o carecen de contenidos visibles, incluyendo aquellos que están en pruebas.
- Desconocido: Su naturaleza es difícil de determinar por ser contenidos ilegibles o dispersos.

Hay una manera de diferenciar entre sitios lícitos e ilícitos en Dark Web. Los sitios legítimos casi siempre identifican a sus operadores, mientras que los ilícitos los esconden. Los proveedores de servicios ilícitos se esconden detrás del anonimato o se aprovechan de las ventajas de seguridad de la plataforma (Moore y Rid, 2016, pp. 24-25).

La Figura 2 muestra la distribución de los contenidos en la red Tor.

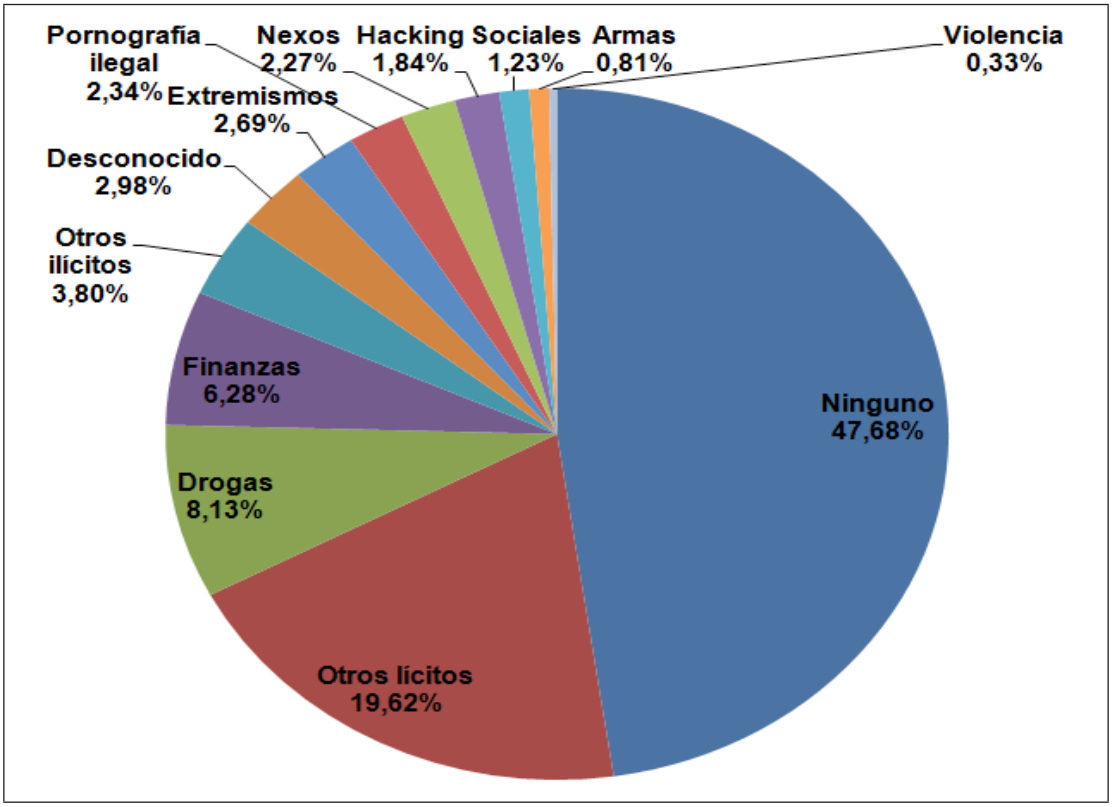

Figura 2. Contenidos en la red Tor. Fuente: Elaboración propia a partir de Moore y Rid, 2016, pp. 20-21.

En Dark Web, igual que en el mundo real, los delincuentes necesitan ser presentados y sus referencias deben ser comprobadas antes de poder realizar transacciones. La distribución de bienes y servicios está organizada alrededor de miles de salas de chat ilícitas y de foros mediante previa invitación. Los sitios ilícitos más exclusivos requieren contar con direcciones alfanuméricas secretas, que no están listadas en línea, sino que se pasan de persona a persona. Ciertos foros criminales impiden que los candidatos aspirantes entren en sus mundos clandestinos sin haber recibido la aprobación unánime por parte de los miembros más antiguos de la organización y transcurrido un periodo de espera de más de una semana (Goodman, 2015, p. 202).

Casi la mitad de los Estados miembros de la Unión Europea ha investigado actividades relacionadas con drogas o con pagos fraudulentos con tarjetas en Darknet, y más de un tercio de los países europeos ha indagado en actividades relacionadas con propiedad intelectual, tráfico de armas o cuentas bancarias comprometidas. Casi un tercio de los cuerpos policiales de la Unión Europea monitoriza activamente los mercados en Darknet, sobre todo para operaciones específicas en lugar de para recopilar inteligencia con carácter general. Una pequeña parte de los delincuentes activos en la Internet Oscura consigue explotar con éxito negocios que generan pingües beneficios. Es un mercado muy concentrado, donde el 1% de los vendedores acapara el 51,5% del total de las transacciones en Darknet (Europol, 2015, p. 52).

Dark Web ofrece recursos a los criminales para acceder a tutoriales, donde pueden adquirir instrucciones y comprar las herramientas necesarias para hackear ordenadores y cometer delitos, con un anonimato virtual. Esa parte del ciberespacio puede crear una desconexión mental entre el criminal, el delito y el mundo real, porque las víctimas no tienen rostro y no se emplea violencia directa. Ya no hace falta una banda, unas pistolas y un coche para robar un banco; cualquier individuo puede cometer el crimen desde su habitación con un simple ordenador portátil (UK Government Office of Science, 2015, p. 57).

Actividades delictivas, como el crimen organizado, el terrorismo o el espionaje, se desarrollan cada vez con mayor frecuencia en el ciberespacio a través de sofisticados procedimientos técnicos y operativos. Un ejemplo es “la internet oculta (deep web), donde se realizan actividades ilegales de todo tipo favorecidas por el anonimato del usuario” (Gobierno de España, 2016, p. 54).

Dark Web ha crecido significativamente en la última década. Los contenidos almacenados en las redes oscuras requieren software específico para acceder a ellas y, por tanto, quedan ocultas para la mayoría de los internautas y facilitan la navegación y las comunicaciones anónimas. Dark Web se ha convertido en una cadena de suministro madura que sostiene actividades cibercriminales, que incluye la venta de todo lo que un ciberdelincuente necesita para embarcarse en una actividad maliciosa. Además, esa misma cadena de suministro hace posible que el delincuente pueda vender lo que ha robado (UK Government Office of Science, 2015, p. 75).

Esas complejas cadenas de suministro para el cibercrimen pueden extenderse por todo el mundo, como Silk Road, y usar tecnologías anonimizadoras como Tor, que permiten que cualquier persona pueda comprar cualquier producto o servicio, desde software malicioso a tiempo de alquiler de una botnet (UK Government Office of Science, 2015, p. 76).

Los mercados ocultos o criptomercados son unas plataformas comerciales en línea que reúnen múltiples vendedores que ofrecen bienes y servicios generalmente ilegales. En ellos predomina la venta de estupefacientes y sustancias ilegales (Aldridge, Judith y Décary-Hétu, 2016, p. 1).

A principios de octubre de 2013, el principal mercado negro en línea, Silk Road, era desmantelado por el FBI. En pocas semanas sus vendedores se pasaban a los de la competencia o abrían sus propios negocios anónimos. A principios de noviembre de 2013 se inauguraba Silk Road 2.0, que acabó cerrando un año después. En pocos meses muchos mercados anónimos nuevos aparecieron, con distintos grados de sofisticación, duración y especialización. A la vez otros desaparecieron, ya fuera debido a detenciones o voluntariamente. El ecosistema de los mercados anónimos en línea ha evolucionado significativamente, comparado con la primera época cuando Silk Road casi era un monopolio. La facturación estimada de Silk Road oscilaba entre 1,1 y 1,2 millones de dólares al año. Por su parte, Silk Road 2.0 vendía unos ocho millones de dólares al mes antes de su clausura en noviembre de 2014 (Soska y Christin, 2015, p. 33 y 46).

Los mercados negros en Deep Web se sustentan en tres elementos tecnológicos. Criptodivisas como Bitcoin, que funcionan igual que el dinero en efectivo; la red Tor anonimiza el tráfico web; y los programas de cifrado PGP (Pretty Good Privacy) blindan los datos dentro de los mensajes de correo electrónico. El anonimato es esencial. La verdadera identidad de compradores y vendedores permanece oculta, y solo se les conoce por sus nombres de usuario (Hardy y Norgaard, 2015, p. 2).

Los mercados ocultos representan un nuevo canal de distribución de drogas. Su crecimiento y su gran resiliencia frente a operaciones policiales y a fraudes dentro del propio mercado sugiere que su importancia va a aumentar en los próximos años. Hacen posible que traficantes minoristas de droga transciendan los límites geográficos, lo que aumenta potencialmente la difusión de drogas en lugares donde antes no estaban disponibles o cuya disponibilidad era limitada (Aldridge et al., 2016, p. 6).

Dentro de las prioridades operativas, Europol destaca la infiltración legítima y el cierre de comunidades en línea que fomenten la producción de materiales de pornografía infantil, sobre todo en Darknet. Asimismo considera que Darknet es un facilitador transversal del crimen. Por eso, otra de sus prioridades es combatir los sitios de comercio ilegal en la Internet Oscura. Aparte, Europol (2015, p. 15) recomienda implantar acciones que fomenten la compartición de inteligencia y los análisis tácticos, especialmente sobre esas cuestiones preferentes.

En el caso de la explotación sexual de menores, los criminales que se mueven en Darknet parecen estar más cómodos a la hora de cometer abusos y discutir sus preferencias sexuales por menores que aquellos que usan la Surface Web. Un mayor anonimato y relaciones de camaradería pueden favorecer sus impulsos sexuales, que podrían no manifestarse en ningún otro ambiente que careciera de esas características (Europol, 2015, p. 30).

Los delincuentes relacionados con la explotación sexual a menores siguen aprovechando las redes anónimas para ocultar sus actividades de los cuerpos de seguridad. Las comunidades de delincuentes evolucionan y aprenden de los errores cometidos por aquellos que han sido detenidos por la policía, lo que dificulta la infiltración. Los

usuarios de Darknets están continuamente desarrollando relaciones de confianza y comparten experiencias técnicas. Sensaciones de seguridad, reforzadas por la percepción de anonimato y por un fuerte apoyo por parte de una comunidad de ideas similares, influyen en el comportamiento de los individuos haciéndolos más propicios a cometer delitos (Europol, 2015, p. 31).

Los pagos entre criminales también son más sencillos en Dark Web. Servicios ocultos como Agora o el desaparecido Evolution están dedicados casi en exclusiva a pagos con Bitcoin, con mecanismos de gestión y funciones de fideicomiso construidas en interfaces de mercado. Actualmente, Bitcoin es clave en muchas investigaciones policiales dentro de la Unión Europea, ya que representa más del 40% de los pagos entre criminales detectados (Europol, 2015, p. 46).

La Internet Oscura favorece igualmente las comunicaciones entre delincuentes. El uso de foros en Deep Web o en Dark Web son muy comunes. Dichos foros son puntos de encuentro y de intercambio para criminales que hacen negocios y establecen relaciones con individuos de ideas afines. En julio de 2015, tuvo lugar la Operación Bugbite que logró acabar con Darkode, el foro cibercriminal en inglés más prolífico hasta la fecha. Ese foro abarcaba una amplia variedad de productos y servicios relacionados con el cibercrimen, incluyendo malware, exploits de Día Cero, hacking, robo de credenciales y de tarjetas bancarias, botnets en alquiler y ataques de denegación de servicio distribuidos (DDoS) (Europol, 2015, p. 50).

El uso de Dark Net en el tráfico de drogas ha aumentado en los últimos años, y los beneficios de ese tráfico presentan un fuerte potencial para financiar el terrorismo y los extremismos violentos. El suministro de drogas vía Internet, incluyendo mercados en línea anónimos en Dark Net, es cada vez más relevante. El potencial de Dark Net para atraer nuevos grupos de consumidores preocupa a la Oficina de las Naciones Unidas contra la Droga (UNODC), debido a que facilita el acceso a las drogas tanto en países desarrollados como en desarrollo (UNODC, 2016, pp. v y xiii).

Los cuerpos de seguridad y los sistemas de justicia penal de muchos países todavía no se encuentran capacitados para tratar eficazmente con los mercados en línea anónimos en Dark Net. Aparte de problemas prácticos, hay otras dificultades legales que deben solucionarse, como la identificación de la jurisdicción responsable y la rutina internacional de compartir información, especialmente cuando la localización física de compradores y vendedores es desconocida (UNODC, 2016, p. xv).

La compra de drogas vía Dark Net está aumentado. No solo preocupa en términos de atraer a nuevos consumidores, sino además porque los usuarios pueden evitar el contacto directo con criminales y con la policía. Los buscadores web tradicionales no permiten acceder a Dark Net. Así compradores y vendedores suelen utilizar Tor para intentar ocultar su identidad. Los productos se suelen pagar con bitcoins u otras criptomonedas y se suelen despachar vía servicios postales (UNODC, 2016, p. 24).

Según una reciente encuesta mundial, la proporción de internautas que usan Dark Net para comprar drogas ha crecido, alcanzando un 6,4% en 2014. Ese porcentaje todavía es mayor entre los nuevos consumidores de estupefacientes, con un aumento del 25% entre 2013 y 2014. Los entrevistados mencionaban varias ventajas de comprar drogas en Dark Net. Una es el propio producto, que suele ser de mayor calidad y estar más disponible. Otra es el hecho de que las interacciones del comprador sean

virtuales, lo que reduce el riesgo de seguridad personal durante las transacciones, porque evita la exposición a violencia física. A esto se añade que reduce la sensación de peligro de ser detenido por la policía. Todos estos factores ayudan a explicar por qué los compradores de drogas en Dark Net suelen estar dispuestos a pagar precios más altos y por qué esos mercados atraen a nuevos clientes. Casi un 4% de los encuestados manifestaba que nunca había consumido drogas antes de empezar a entrar en la Internet Oscura. Por otro lado, el 30% de los compradores de estupefacientes vía Dark Net señalaba haber consumido una mayor variedad de drogas que antes de empezar a aprovisionarse en esos mercados ocultos (UNODC, 2016, pp. 24-25).

El terrorismo se ha mudado a Deep Web. La Surface Web convencional se ha hecho demasiado peligrosa para los terroristas que buscan el anonimato; allí se les puede monitorizar, rastrear y localizar. Por el contrario, en Dark Web, las redes descentralizadas y anónimas ayudan a evitar arrestos y el cierre de plataformas terroristas. La tendencia reciente es que los terroristas usen Dark Web para comunicarse, conseguir financiación y almacenar información y otros materiales en línea (Weimann, 2016, p. 40).

La seguridad pública no es una misión exclusiva de la policía. Hace falta un enfoque integral con muy diversos actores implicados. Los cuerpos policiales no van a ser capaces de afrontar el problema en solitario. Las empresas, las agencias, los departamentos gubernamentales, la industria y las universidades -cualquiera que tenga experiencia en el campo de lo ciber aunque no esté relacionado con la persecución de delitos- tienen un papel relevante para mantener la seguridad pública (UK Government Office of Science, 2015, p. 57).

En su último informe sobre crimen organizado en Internet (IOCTA), Europol sugiere varias recomendaciones relacionadas con Darknets. Una de ellas es que los cuerpos de seguridad deberían recopilar proactivamente inteligencia relacionada con los servicios ocultos. Cada Estado miembro de la Unión Europea debería proporcionar inteligencia sobre estos servicios ocultos a Europol. Para ello hace falta un mayor compromiso por parte las fuerzas policiales no dedicadas específicamente al cibercrimen, porque la venta de drogas y la de armas en esos mercados de la Internet oculta representan una parte muy relevante de ese tipo de tráficos. Otra recomendación insiste en la necesidad de colaboración entre los cuerpos de seguridad, el sector privado y la universidad para explorar investigaciones relacionadas con tecnologías emergentes en Dark Web como los mercados descentralizados como OpenBazaar (Europol, 2015, p. 53).

4. DIFICULTADES DE EXPLOTACIÓN

Utilizar Deep Web como fuente de ciberinteligencia no es una tarea sencilla. Esta parte de Internet tiene una serie de peculiaridades que limitan su explotación. El gran escollo es recopilar los datos. La información no está directamente accesible como páginas web, porque suele estar detrás de formularios web.

Generalmente, los formularios web se presentan como una colección de campos de entrada, casillas de verificación, listas desplegables y otros elementos de selección, algunos de los cuales pueden ser obligatorios. Actúan como una interfaz que especifica todos los posibles patrones de acceso subyacentes en los datos, y los protege de accesos no deseados (Bienvenu, Deutch, Martinenghi, Senellart y Suchanek, 2012, p. 2).

Además, acceder a los datos de Deep Web es costoso, debido a la latencia de acceso en la red. Diversas limitaciones entorpecen el acceso a los datos, lo que todavía encarece más las consultas. Algunas fuentes solo proporcionan sus registros en lotes de tamaño fijo. Otras solo conducen a los registros superiores, en función de un ránking, dejando fuera el resto. En otros casos, solo admiten una cantidad limitada de accesos en un periodo concreto. Ciertas fuentes solo dan información incompleta, por ejemplo, presentando únicamente un subconjunto de los campos subyacentes, o bien distribuyen los datos con un nivel de granularidad diferente al que requieren otras fuentes, como meses frente a fechas completas (Bienvenu et al., 2012, p. 2).

Otro reto importante es la heterogeneidad de los datos. Conviene considerarlo a la hora de diseñar estrategias de obtención de información y de seleccionar las técnicas de Knowledge Discovery adecuadas para cada caso. En Deep Web pueden encontrarse tres tipos de contenidos:

- Datos dinámicos. Solo son accesibles a través de su interfaz de consulta. Esas interfaces pueden estar basadas en atributos de entrada y una consulta de un usuario puede involucrar valores específicos
- Contenidos sin enlazar. Los datos no están disponibles durante los análisis realizados por las arañas web tradicionales.
- Contenidos que no son texto. Diversos formatos de archivos multimedia, PDF y documentos que no son HTML (Khurana y Chandark, 2016, p. 209).

Básicamente hay tres etapas que hay que completar para poder acceder a los contenidos en Deep Web. En primer lugar, encontrar las fuentes de datos. En segundo, seleccionarlas. Por último, enviar las fuentes de datos elegidas al sistema de integración de datos.

Dependiendo de cómo sea el sistema de integración, se pueden añadir varias fuentes de datos. Sin embargo, no deberían incluirse todas dentro del sistema, por estos motivos:

- Podrían añadirse datos redundantes.
- Los datos irrelevantes podrían reducir la calidad global del sistema.
- Podrían introducirse datos de baja calidad.
- Aumentarían los costes de adquisición de datos (Khurana y Chandark, 2016, p. 210).

Suelen emplearse tres tipos de técnicas para acceder a los datos en Deep Web: Procesado de formularios y consultas a bases de datos web; correspondencia de esquema; y otras técnicas de extracción (Khurana y Chandark, 2016, p. 210).

1. Procesado de formularios y consultas a bases de datos web

Dada la enormidad de datos almacenados en la web oculta, para poder acceder a ellos hace falta rellenar y enviar formularios para conseguir la información de las bases de datos. Las arañas (*crawlers*) para Deep Web son la técnica más empleada. Se puede distinguir entre los genéricos y los verticales. Los genéricos realizan búsquedas en anchura, mientras que los verticales hacen búsquedas en profundidad concentrándose en un tema específico.

2. Correspondencia de esquema

La correspondencia de esquema (*matching schema*) es el proceso de identificar dos objetos semánticamente relacionados. En lugar de rellenar el formulario en el sitio Deep Web y luego extraer los datos para comprobar si son relevantes, se prepara un esquema de los datos requeridos. Eso reduce los costes de extracción y procesado.

3. Otras técnicas de extracción en búsquedas Deep Web.

Entre ellas figuran la minería de datos, las arañas web basadas en ontología, el clustering o la extracción visual de datos. Tienen en común que, en lugar de extraer la información completa y luego parsearla, solo capturan la sección que contiene la información relevante (Khurana y Chandark, 2016, 210, pp. 415-416).

Antes de poder extraer los datos almacenados en Deep Web es necesario establecer un conjunto de normas que determinen la información de interés (ejemplos positivos) y descarten los datos espurios (ejemplos negativos). Existen diversas propuestas para elaborar esas reglas de aprendizaje, que además deben ser adaptables, porque la web evoluciona rápidamente (Jiménez y Corchuelo, 2015, p. 140). Por ejemplo, es posible adoptar un enfoque de arriba-abajo, que empieza con la regla más general y va añadiendo iterativamente condiciones basadas en las características del catálogo hasta que la regla ya no encuentra ningún ejemplo negativo. El proceso finaliza cuando todos los ejemplos positivos coinciden. En caso contrario, el proceso continúa aprendiendo nuevas reglas. El sistema incluye mecanismos para evitar que se generen reglas demasiado complejas o excesivamente específicas. Esta propuesta de aprendizaje de reglas permite extraer información de interés desde Deep Web de forma automática, para que pueda ser procesada posteriormente por agentes de software. Para disminuir el coste de las búsquedas, se incluye una técnica que reduce los ejemplos negativos (Jiménez y Corchuelo, 2015, pp. 141 y 149).

Automatizar procesos es otro escollo complicado de superar. Uno de los primeros procesos que tiene que ser automático es la identificación de la interfaz de búsqueda en Deep Web. En un entorno de aprendizaje automático, hace falta un clasificador binario que diferencie entre interfaces buscables y no buscables. Se pueden usar varios métodos como árboles de decisión o redes de neuronas artificiales, entre otros. Sin embargo, independientemente del algoritmo de aprendizaje utilizado, suelen ser técnicas supervisadas que deben enfrentarse al problema de la escasez de datos etiquetados (Wang, Xu y Zhou, 2014, p. 635).

Por otro lado, en Deep Web es habitual que las páginas de resultados de consultas se generen dinámicamente desde las bases de datos en respuesta a las consultas enviadas por los usuarios. Extraer automáticamente datos estructurados de dichos resultados es un problema complicado, porque la estructura de los datos no está explícitamente representada (Anderson y Hong, 2013, p. 1233).

Los servicios en Deep Web presentan interdependencias, sobre todo cuando un mismo dato está disponible por varias vías. Es aconsejable tenerlo en cuenta para planificar y optimizar las consultas. Por eso, la diversificación y la eliminación de duplicados son aspectos que no deberían olvidarse (Bienvenu et al., 2012, p. 2). Así, identificar y deshacerse de los registros duplicados es otra tarea clave a la hora de preprocesar datos de Deep Web, sobre todo cuando se trata de integrarlos desde

múltiples orígenes. Las técnicas de aprendizaje automático pueden ayudar en esta operación y contribuir a reducir los costes de etiquetado. Y es que las técnicas estándar de detección de duplicados no funcionan bien en este ámbito, como, por ejemplo, escoger aleatoriamente parejas de registros o usar otras distribuciones. El objetivo final es identificar los registros duplicados de la misma o de distintas bases de datos, incluso si los registros no son idénticos (Zhao, Xin, Xian y Cui, 2014, pp. 125-127).

En muchos casos, encontrar instancias raras o atípicas (*outliers*) puede ser mucho más interesante que hallar patrones. Un valor atípico es aquel que se desvía tanto de otras observaciones que despierta la sospecha de que ha sido originado por un mecanismo distinto. Desde el punto de vista de la seguridad, hay muchos escenarios en los que resulta muy útil descubrir comportamientos que se salen de la norma. La cuestión es que los métodos habituales de minería de datos son inaplicables en Deep Web, porque se necesita conocer la distribución subyacente a los datos, algo que es impracticable en el contexto de la Internet Profunda (Xian, Zhao, Sheng, Fang, Gu, Yang y Cui, 2016, p. 1).

Entre las posibles soluciones, la más simplista, pero muy costosa, es descargar todos los registros de la base de datos subordinada y minar los atípicos con técnicas tradicionales (métricas de distancia y de densidad, o clustering, por ejemplo). La segunda alternativa es efectuar un muestreo aleatorio en la base de datos subordinada existente en Deep Web. Eso requeriría gran cantidad de muestras, lo que encarecería demasiado el proceso (Xian et al., 2016, p. 2).

Otra posibilidad consiste en desglosar la detección de valores atípicos en Deep Web en tres fases: estratificación, muestreo por vecindad y muestreo por incertidumbre. Primero se crea un esquema de estratificación a través de un árbol jerárquico que modele la relación entre los atributos de entrada y los de salida. Luego, en lugar de realizar un muestreo aleatorio por todo el estrato, se puede aplicar un esquema de muestreo por vecindad para recopilar más valores atípicos. Seguidamente, un algoritmo de muestreo por incertidumbre se ocupa de verificar las instancias dudosas para mejorar el proceso de detección (Xian et al., 2016, p. 12).

Por su parte, los cuerpos de seguridad se enfrentan a tres retos adicionales en Deep Web: el cifrado, la atribución y la fluctuación.

1. Cifrado: Todo lo que hay en Deep Web y Dark web está cifrado. Eso significa que los delincuentes son mucho más conscientes de estar vigilados y de la posibilidad de ser atrapados. El cifrado es la primera contramedida para evitar la detección.

2. Atribución: En Deep Web todavía es mucho más complicado determinar la atribución que en Surface Web. Todo sucede en dominios como los .onion (Tor). El enrutado a esos dominios tampoco está claro.

3. Fluctuación: Deep Web es un lugar muy dinámico. Un foro en línea puede estar en una dirección URL un día y otro en otra. Los esquemas de nombres y direcciones a menudo cambian. Eso significa que la información recopilada hace un par de semanas hoy deja de ser relevante. Eso tiene consecuencias a la hora de conseguir pruebas de delitos. Si se tienen en cuenta los plazos que tardan los procedimientos judiciales penales, los cuerpos de seguridad deben ser capaces de documentar rigurosamente cualquier actividad criminal en línea mediante

capturas de pantalla con sellos de tiempo para evitar que sus casos sean invalidados (Ciancaglini et al., 2015, p. 38).

5. EJEMPLOS DE USO

El uso de técnicas minería de datos y Knowledge Discovery en Deep Web y Dark Web está extendido, aunque en constante evolución. Las áreas que captan una mayor atención por parte de los investigadores son las arañas (*crawlers*) para Deep Web, los sistemas de detección y prevención de intrusiones (IDPS) y la detección de comunidades virtuales.

5.1. ARAÑAS PARA DEEP WEB

Deep Web sigue creciendo a un ritmo muy rápido. Esto aumenta el interés en desarrollar técnicas eficientes de localizar recursos. Aparte es crucial idear estrategias de rastreo en la Internet Profunda para descubrir rápidamente fuentes con contenidos relevantes. Para recolectar información en Deep Web hace falta un enfoque que proporcione una amplia cobertura pero que además mantenga una alta eficiencia de rastreo (Zhao, Zhou, Nie, Huang y Jin, 2015, pp. 1-2).

Explorar la web oculta implica dos tareas: descubrir recursos y extraer contenidos. La primera se ocupa de encontrar automáticamente sitios web que contienen interfaces con formularios de búsqueda. La segunda trata de obtener información de esos filtros filtrando los formularios mediante consultas o palabras clave relevantes (Gupta y Bhatia, 2014, p. 112).

Para completar esas tareas, una araña para Deep Web debe simular las operaciones del navegador del usuario, por ejemplo, rellenar formularios o hacer clic en el botón de aceptar (Yu, Guo, Yu, Xian y Yan, 2014, p. 5050).

En un escenario real de Deep Web, normalmente es imposible aplicar un algoritmo que cubra todos los documentos. El algoritmo todavía los desconoce. Además, el volumen de datos suele ser tan grande que ni siquiera los algoritmos de carácter aproximativo lo pueden manejar. La única opción es ejecutar un algoritmo sobre una muestra con un subconjunto de los datos (Wang, Lu y Chen, 2014, p. 199).

Los materiales en Dark Web tienen importantes implicaciones para la ciberinteligencia y la ciberseguridad. La recopilación de dichos contenidos también es relevante para estudiar diversos puntos de vista sociales y políticos presentes en esas comunidades virtuales. En concreto, los foros en Dark Web muestran una problemática faceta que está asociada con el cibercrimen, el odio y los extremismos. La naturaleza encubierta de esa parte de Internet hace que las técnicas tradicionales de rastreo web sean insuficientes para capturar tales contenidos. El sistema suele estar asistido por humanos, que se encargan de registrarse como miembros. Después entran en acción las arañas, que localizan, recopilan e indexan la información según parámetros predeterminados (Fu, Abbasi y Chen, 2010, pp. 1213-1214).

Las arañas para rastrear foros en Dark Web tienen tres dificultades de diseño. La primera es la accesibilidad; suelen requerir registrarse como miembro, a veces incluso por invitación. En segundo lugar, son multilingües. En tercer lugar, incorporan

contenidos multimedia en muy diversos formatos (fotos, vídeos y audios) que, al no ser de texto, resultan complicados de indexar (Fu, Abbasi, A. y Chen, 2010, p. 1214).

5.2. SISTEMAS DE DETECCIÓN Y PREVENCIÓN DE INTRUSIONES (IDPS)

Los sistemas de detección y prevención constituyen un área muy relevante donde aplicar minería de datos en el campo de la lucha contra el cibercrimen. Los avances en Tecnologías de Información y Comunicación (TIC) están propiciando que los delincuentes usen el ciberespacio para cometer ciberdelitos. Las ciberinfraestructuras son altamente vulnerables a intrusiones y otras amenazas. Dispositivos físicos como sensores y detectores no son suficientes para monitorizar y proteger las infraestructuras. Hacen falta sistemas de ciberdefensa flexibles, adaptables y robustos, capaces de detectar una amplia variedad de amenazas y de tomar decisiones inteligentes en tiempo real. En este contexto son de utilidad agentes semiautónomos inteligentes capaces de detectar, evaluar y responder a los ciberataques (Dilek, Çakır y Aydın, 2015, p. 21).

Diversas técnicas relacionadas con la minería de datos están ganando importancia en los sistemas de detección y prevención de intrusiones (IDPS). Es el caso de las redes de neuronas artificiales (ANN), los agentes inteligentes, los sistemas inmunes artificiales (AIS), los algoritmos genéticos o los conjuntos difusos, entre otras.

La Tabla 1 resume las ventajas que esas técnicas de minería de datos aportan a los sistemas de detección y prevención de intrusiones (IDPS).

Redes de neuronas artificiales (ANN)	Procesado de información en paralelo. Aprendizaje mediante ejemplos. Capaces de manejar complejas funciones no lineales. Resiliencia al ruido y a datos incompletos. Modelos de aprendizaje versátiles y flexibles. Intuitivos, al ser una abstracción de redes neurales biológicas.
Agentes inteligentes	Movilidad. Buena disposición para intentar completar una tarea incluso con objetivos contradictorios. Racionalidad al lograr sus objetivos. Adaptabilidad al entorno y a las preferencias del usuario. Colaboración con el usuario para comprobar inconsistencias en los datos.
Sistemas inmunes artificiales (AIS)	Estructura dinámica y robustez. Aprendizaje distribuido y en paralelo. Autoadaptable para actualizar las marcas de intrusión sin intervención humana. Respuesta selectiva y optimización de recursos. No dependiente de un único componente que puede ser fácilmente sustituido por otro.
Algoritmos genéticos	Robustez. Adaptabilidad al entorno. Optimización de soluciones a problemas computacionales complejos. Evaluación de múltiples esquemas en paralelo. Búsquedas flexibles y globales.
Conjuntos difusos	Mecanismo de razonamiento interpolativo robusto; Interoperabilidad; y Amigables para usuarios humanos.

Tabla 1. Ventajas de algunas técnicas de minería de datos para los sistemas de detección y prevención de intrusiones. Fuente: Elaboración propia a partir de Dilek, Çakır y Aydın, 2015, pp. 32-33.

Por otro lado, los sistemas actuales de detección y prevención de intrusiones (IDPS) permiten reconocer anomalías y ataques desconocidos previamente, pero también presentan importantes limitaciones (Dilek et al, 2015, p. 33):

1. La principal es construir un modelo sólido sobre lo que es un comportamiento aceptable y lo que es un ataque. Se puede producir un elevado número de falsos positivos, causados por un comportamiento atípico que realmente es normal y está autorizado.
2. Estos sistemas deben ser capaces de caracterizar patrones normales y para crear un modelo de comportamiento normal necesitan amplios conjuntos de datos de entrenamiento. Cualquier cambio en los patrones normales requieren actualizar la base de conocimiento del sistema.
3. Si el sistema clasifica incorrectamente una actividad legítima como maliciosa, el resultado puede ser un intento de parar esa actividad o cambiarla.
4. Cualquier sistema de detección, sin importar lo eficiente que sea, puede ser desactivado por los atacantes si averiguan cómo funciona.
5. En entornos heterogéneos también está la cuestión de integrar la información procedente de distintos sitios.
6. Los sistema asimismo deben ser diseñados de forma que cumplan las normas legales, los requisitos de seguridad y los acuerdos de niveles de servicio correspondientes.

Ciertos sistemas IDPS se ocupan de analizar el comportamiento de las redes. Examinan el tráfico de red para identificar amenazas que generan flujos de tráfico inusuales, como pueden ser los ataques de denegación de servicio. Se pueden utilizar sistemas de monitorización a gran escala para detectar ataques DDoS a partir de datos de tráfico en Darknet. Por ejemplo, algunos son adaptativos y emplean un modelo de aprendizaje supervisado a través de máquinas de soporte vectorial (SVM). Trabajan con los paquetes observados en Darknet; algunos son fácilmente distinguibles a partir de los números de los puertos de origen y destino, y de las banderas (*flags*), pero otros no. Para detectar los paquetes complicados, el sistema extrae características determinadas basándose en estadísticas y las clasifica usando un modelo SVM. Además, para tratar los cambios en los patrones de actividad, se aplica un aprendizaje incremental (Furutani et al., 2015, p. 382).

5.3. DETECCIÓN DE COMUNIDADES

La identificación de comunidades virtuales en Dark Web resulta de gran utilidad no solo en el crimen organizado, sino también en la lucha contra el terrorismo y los extremismos violentos. Para estudiarlas se pueden combinar técnicas de análisis de redes sociales (SNA) con minería de datos. Dark Web ofrece un inagotable potencial para lograr la coordinación, la distribución de propaganda y otras interacciones no deseadas entre grupos extremistas, terroristas y ciberdelincuentes. El reto está en identificar esas comunidades y a sus líderes.

Las comunidades virtuales en Dark Web reúnen a miembros que comparten intereses sobre determinados temas. Por eso, para comprenderlas resulta fundamental conocer cuáles son los principales intereses en cada una. A partir de ahí es posible identificar a sus miembros clave, por ejemplo, los líderes de opinión. Un miembro clave sería una persona totalmente alineada con las metas y los temas de la comunidad que produce contenidos que son muy relevantes para satisfacer los intereses del resto de los miembros. Los miembros clave pueden o no estar altamente radicalizados, pero lo que siempre sucede es que aumentan las interacciones en la comunidad gracias a sus mensajes, que producen réplicas de miembros de distintos niveles (L'Huillier, Alvarez, Ríos y Aguilera, 2010, pp. 66-67).

En una comunidad virtual hay metas diferentes asociadas con los objetivos de sus miembros. El apoyo de la comunidad en un foro en Dark Web donde reina el anonimato, la ubicuidad y la libertad de expresión es el ambiente perfecto para compartir propaganda fundamentalista y terrorista. Un método para reconocer los objetivos subyacentes de los miembros requiere identificar amenazas o cuestiones de seguridad (L'Huillier et al., 2010, p. 67).

La topología de las Darknets comparte propiedades con otros tipos de redes sociales, donde las estructuras de mundo pequeño están determinadas por las propiedades del flujo de información, y caracterizadas por un camino medio corto y por un alto coeficiente de clustering. Se pueden emplear diferentes medidas de centralidad, como el grado, la intermediación (*betweenness*) y la cercanía, para identificar a los miembros clave de una comunidad. Este análisis de redes sociales se puede completar con minería de datos de texto mediante análisis semántico latente. Generalmente se elabora un modelo de evaluación y selección que mejora la clasificación de los mensajes que contienen información sensible sobre las opiniones y sentimientos de los extremistas. Además el análisis de autoría de las tendencias del grupo debe lidiar con el problema del anonimato asociado a este tipo de comunidades virtuales (L'Huillier et al., 2010, p. 67).

6. CONCLUSIONES

La minería de datos, entendida en sentido amplio, puede ser un gran aliado en entornos fluctuantes y dinámicos como Deep Web y Dark Web. Sus técnicas que pueden ayudar a encontrar sentido a cantidades ingentes de datos. También pueden contribuir a reducir la sobrecarga informativa y cognitiva de los miembros de los cuerpos y fuerzas de seguridad y de la comunidad de la inteligencia.

En Deep Web y Dark Web se puede encontrar información valiosa sobre amenazas, vulnerabilidades y riesgos. Actualmente son fuentes muy importantes para la ciberinteligencia. Las potentes técnicas de minería de datos permiten convertir datos en conocimiento. Por ejemplo, sirven para desvelar patrones e identificar tendencias.

Es esencial tener claros los objetivos y la estrategia desde el principio. Así será posible seleccionar las técnicas de minería de datos y de Knowledge Discovery adecuadas para cada caso. A menudo va a ser necesario adaptarlas a las peculiaridades del campo de la seguridad. Tampoco hay que olvidar la importancia de interpretar y evaluar los resultados antes de presentarlos a los decisores. Además, el contexto lo cambia todo. Algo que funciona en un contexto y en un momento determinados puede no hacerlo en otros.

La minería de datos no está libre de limitaciones. Una de ellas es la buena calidad de los datos, que no es fácil de lograr en Deep Web, como antes se ha reseñado, por dificultades técnicas y de costes. Otra limitación está relacionada con los falsos positivos y los falsos negativos, esto es, la precisión y la sensibilidad de los modelos generados. En tercer lugar, está la cuestión de la rareza, en el sentido de poca frecuencia. Los hechos delictivos son eventos poco frecuentes, lo que hace poco fiable la extrapolación de los modelos.

En definitiva, la minería de datos y el Knowledge Discovery son útiles herramientas que conviene manejar con sabiduría y prudencia, siendo conscientes de sus limitaciones.

AGRADECIMIENTOS

A Ramón Fuentes por leer el borrador de este artículo.

REFERENCIAS BIBLIOGRÁFICAS

Aldridge, J. y Décary-Hétu, D. (2016). Hidden wholesale: The drug diffusing capacity of online drug cryptomarkets. *International Journal of Drug Policy* (en prensa).

Anderson, N. y Hong, J. (2013). Visually Extracting Data Records from the Deep Web. *WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web*, 1233-1238.

Bienvenu, M., Deutch, D., Martinenghi, D., Senellart, P. y Suchanek, F. (2012). Dealing with the Deep Web and all its Quirks. En M. Brambilla, S. Ceri, T. Furche, & G. Gottlob (Eds.), *VLDS 2012: Very Large Data Search* (pp. 21-24). Aachen: CEUR.

Bergman, M. (2001). *The Deep Web: Surfacing Hidden Value*. BrightPlanet. Disponible en http://brightplanet.com/wp-content/uploads/2012/03/12550176481-deep-webwhitepaper1.pdf

Chen, H. (2012). *Dark Web: Exploring and Data Mining the Dark Side of the Web*. Nueva York: Springer.

Ciancaglini, V., Balduzzi, M., McArdle, R. y Rösler, M. (2015). *Below the Surface: Exploring the Deep Web*. Trend Micro. Disponible en https://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp_below_the_surface.pdf

Dilek, S., Çakır, H. y Aydın, M. (2015). Applications of Artificial Intelligence Techniques to Combating Cyber Crimes: A Review. *International Journal of Artificial Intelligence & Applications (IJAIA), 6*(1), enero 2015, 21-39.

Europol. (2015). *The Internet Organised Crime Threat Assessment (IOCTA) 2015*. La Haya: Europol.

Fachkha, C. y Debbabi, M. (2016). Darknet as a Source of Cyber Intelligence: Survey, Taxonomy, and Characterization. *IEEE Communications Surveys & Tutorials, 18*(2), 1197-1227.

Fu, T., Abbasi, A. y Chen, H. (2010). A Focused Crawler for Dark Web Forums. *Journal of the American Society for Information Science and Technology, 61*(6), 1213-1231.

Furutani, N., Kitazono, J., Ozawa, S., Ban, T., Nakazato, J. y Shimamura, J. (2015). En Arik, Sabri, Huang, Tingwen, Lai, Weng Kin y Liu, Qingshan (Eds.), *Neural Information Processing, 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9–12, 2015 Proceedings, Part IV* (pp. 376-383). Cham: Springer.

Gobierno de España. (2013). *Estrategia de Ciberseguridad Nacional 2013*.

Gobierno de España. (2016). *Informe Anual de Seguridad Nacional 2015*.

Goodman, M. (2015). *Future Crimes: A Journey to the Dark Side of Technology - and How to Survive it*. Londres: Transworld Publishers.

Gupta, S. y Bhatia, K. K. (2014). A Comparative Study of Hidden Web Crawlers. *International Journal of Computer Trends and Technology, 12*(3), 66, 111-118.

Han, J., Kamber, M. y Pei, J. (2012). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

Hardy, R. A. y Norgaard, J. R. (2015, 4 de noviembre). Reputation in the Internet black market: an empirical and theoretical analysis of the Deep Web. *Journal of Institutional Economics*, 1-25. doi: 10.1017/S1744137415000454

Hawkins, B. (2016). *Under The Ocean of the Internet - The Deep Web*. SANS Institute Reading Room. Disponible en https://www.sans.org/reading-room/whitepapers/covert/ocean-internet-deep-web_37012

Jiménez, P. y Corchuelo, R. (2015). On Extracting Information from Semi-structured Deep Web Documents. En Abramowicz, W. (Ed.), *Business Information Systems, 18th International Conference, BIS 2015, Poznań, Poland, June 24-26, 2015, Proceedings* (pp. 140-151). Cham: Springer.

Khurana, K. y Chandak, M. B. (2016). Survey of Techniques for Deep Web Source Selection and Surfacing the Hidden Web Content. *International Journal of Advanced Computer Science and Applications, 7*(5), 409-418.

L'Huillier, G., Alvarez, H., Ríos, S. A. y Aguilera, F. (2010). Topic-Based Social Network Analysis for Virtual Communities of Interests in the Dark Web. *SIGKDD Explorations, 12*(2), 66-73.

Moore, D. y Rid, T. (2016) Cryptopolitik and the Darknet. *Survival, 58*(1), 7-38.

Nakao, K. (2016). IoTSecurity issues related to the future Networked Car. *Symposium on The Future Networked Car, Geneva, Switzerland, 3 de marzo de 2016*. Disponible en https://www.itu.int/en/fnc/2016/Documents/Presentations/Koji-Nakao.pdf

Pederson, S. (2013, marzo). *Understanding the Deep Web in 10 Minutes*. BrightPlanet. Disponible en http://bigdata2.brightplanet.com/whitepaper-understanding-the-deep-web-in-10-minutes

UK Government Office of Science. (2015). *Annual Report of the Government Chief Scientific Adviser 2015: Forensic Science and Beyond: Authenticity, Provenance and Assurance. Evidence and Case Studies*.

Soska, K. y Christin, N. (2015). Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. *Proceedings of the 24th USENIX Security*

Symposium, August 12–14, 2015, Washington, D.C, 33-48.

United Nations Office on Drugs and Crime (UNODC). (2016). *World Drug Report 2016*.

Wang, H., Xu, Q. y Zhou, L. (2014). Deep Web Search Interface Identification: A Semi-Supervised Ensemble Approach. *Information, 5*, 634-651.

Wang, Y., Lu, J. y Chen, J. (2014). TS-IDS Algorithm for Query Selection in the Deep Web Crawling. En Chen, L., Jia, Y., Sellis, T. y Liu, G. (Eds.), *Web Technologies and Applications, 16th Asia-PacificWeb Conference, APWeb 2014, Changsha, China, September 5-7, 2014 Proceedings* (pp. 189-200). Cham: Springer.

Weimann, G. (2016). Terrorist Migration to the Dark Web. *Perspectives on Terrorism, 10*(3), 40-44.

Xian, X., Zhao, P., Sheng, V. S., Fang, L., Gu, C., Yang, Y. y Cui, Z. (2016). Stratification-Based Outlier Detection over the Deep Web. *Computational Intelligence and Neuroscience, 2016*, 1-13.

Yu, H., Guo, J., Yu, Z., Xian, Y. y Yan, X. (2014). A Novel Method for Extracting Entity Data from Deep Web Precisely. *26th Chinese Control and Decision Conference (CCDC),* 5049-5053.

Zhao, F., Zhou, J., Nie, C., Huang, H. y Jin, H. (2015). SmartCrawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces. *IEEE Transactions on Services Computing* (en prensa).

Zhao, P., Xin, J., Xian, X. y Cui, Z. (2014). Active Learning for Duplicate Record Identification in Deep Web. En Wen, Z, y Li, T. (Eds.), *Foundations of Intelligent Systems, Advances in Intelligent Systems and Computing 277* (pp. 125-134). Berlín: Springer.

Fecha de recepción: 21/09/2016. Fecha de aceptación: 20/12/2016