Universitat Oberta de Catalunya

MÁSTER INTERUNIVERSITARIO EN SEGURIDAD DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES

TRABAJO FIN DE MÁSTER

¿QUÉ SABE INTERNET DE NOSOTROS?

VICENT RIBAS MARCOS

ÍNDICE

PLAN DE TRABAJO 5

CAPÍTULO I: BIG DATA 7

1.1. Qué es Big Data 7

1.2. Factores que influyen en el aumento de datos masivos 10

1.3. Explorando Big Data 11

CAPÍTULO II: BUSINESS INTELLIGENCE 13

2.1. Qué es Business Intelligence 13

2.2. Explotación de una base de datos 14

CAPÍTULO III: OBTENCIÓN DE INFORMACIÓN SOBRE OBJETIVOS CONCRETOS . 17

3.1. Information Gathering 17

3.2. Buscadores 18

3.3. Herramientas de búsqueda automatizada 21

CAPÍTULO IV: LA NUEVA HUELLA DIGITAL 27

4.1. Los orígenes de la identificación de objetivos 27

4.2. Fingerprint en la Red 27

CONCLUSIONES 31

BIBLIOGRAFÍA 33

FUENTES ELECTRÓNICAS 34

ABREVIATURAS 36

PLAN DE TRABAJO

Problemática:

Internet se ha convertido en los últimos años en la fuente principal de obtención de información para las grandes empresas. Clive Humby definió “Data” como “the new oil” en 2006, vaticinando por aquel entonces que la recopilación y venta de información procedente de la Red se iba a convertir en un negocio muy lucrativo.

A raíz de la creación de los sitios Web 2.0 en 2003, el usuario puede interactuar con las páginas web, aportando información personal y alimentando notablemente Big Data durante la navegación. Esa información del usuario de Internet, que pasa a ser pública al cederla al servidor del sitio web, tiene un valor incalculable tanto para empresas como para las Fuerzas y Cuerpos de Seguridad.

Por un lado, desde el punto de vista policial, son muchos los delitos que se cometen a través de la Red y no siempre es posible identificar a la persona que está detrás del sistema; no obstante, existen maneras de crear un perfil digital según los hábitos de cada usuario de Internet (navegador que usa, horarios habituales de conexión…), y de la información que se obtiene de la explotación de fuentes abiertas.

Por otro lado, desde el punto de vista del usuario, muchas veces no se es consciente de la gran cantidad de información personal que se está exponiendo a Internet, y el mal uso de esta información por parte de terceros puede tener graves consecuencias para el propio usuario.

El **objetivo** del trabajo es analizar e investigar las siguientes **hipótesis** para llegar a una conclusión:

H1: La venta de información obtenida en forma de Big Data se ha convertido en uno de los negocios más lucrativos en el mundo.

H2: Con la información que hay en Internet, es posible crear perfiles o huellas digitales con el fin de identificar a objetivos concretos.

Con el fin de reunir toda la información posible para posteriormente poder analizarla y llegar a una conclusión coherente con la hipótesis anterior, se llevará a cabo un **método de trabajo de tipo descriptivo**. La investigación se realizará a través de la recopilación de información, y se entrevistará a personal de organismos o empresas que tengan relación con la temática del trabajo.

Universo de estudio:

A través de este trabajo de investigación se pretende analizar el gran poder que tiene la explotación de información a través de fuentes abiertas, aprovechando la ingente cantidad de datos de acceso público que se almacenan en Internet en forma de Big Data.

Líneas de investigación:

Se considera oportuno mencionar las diferentes líneas de investigación que se seguirán a la hora de realizar el trabajo:

1. Se estudiará la base del funcionamiento de Big Data y su explotación desde el punto de vista del Business Intelligence.

2. Se hará referencia a la posibilidad o no de identificar objetivos concretos mediante la explotación de Big Data y técnicas OSINT.

3. Se analizará el concepto de “huella digital” en las nuevas tecnologías conectadas a la Red, y se estudiará la posibilidad de utilizar este tipo de prueba en el proceso judicial tras la investigación de un delito.

Planificación temporal:

Se confeccionará al menos una línea de investigación, en forma de capítulo, para cada una de las entregas programadas en la plataforma de la UOC.

I. *BIG DATA*

1.1 Qué es *Big Data*

A pesar de que la Real Academia de la Lengua todavía no ha incluido el término en su diccionario, es posible definir *Big Data* como toda aquella información que, por su gran tamaño, no puede ser procesada o analizada utilizando procesos o herramientas comunes. Hablamos de Exabytes, que son 10^{18} Bytes o 1 millón de Terabytes, que las bases de datos de grandes empresas almacenan, y que requieren de un análisis rápido y preciso para dar respuesta a las peticiones que se realizan de manera casi instantánea a través de Internet[1].

IBM, empresa fundada en 1911 y líder en el ámbito de la informática hasta la llegada de Windows en 1985[2], ha sido una de las promotoras principales del término *Big Data*. Esta empresa norteamericana revela que la cantidad de datos masivos que se generan diariamente en el mundo podrían rondar el quintillón de bytes (10^{30})[3]. Y no solo eso, sino que grandes empresas como CISCO[4] relacionadas con el mundo de las TIC[5], han realizado estudios que demuestran que la creación de datos va aumentando de manera exponencial según pasan los años. Tanto es así, que desde 2011 hasta 2016 ha habido 4 veces más tráfico de datos, hasta llegar a los 110 Exabytes cada mes[6].

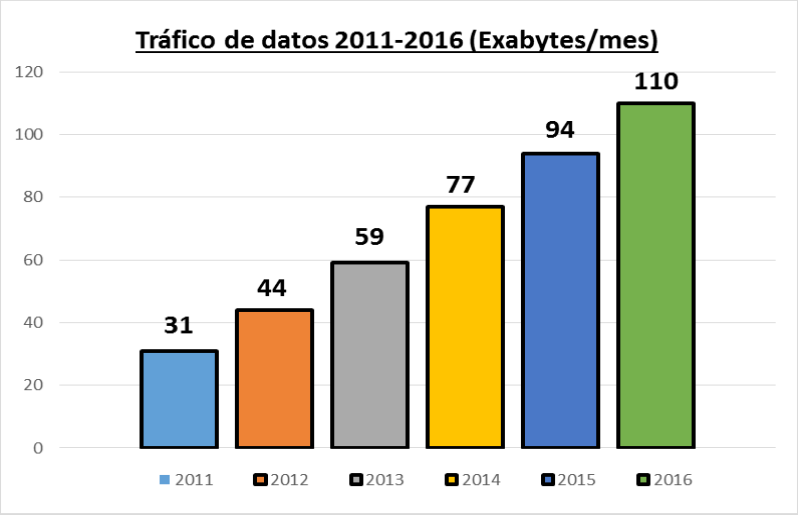

Figura 1. Tráfico de datos entre 2011 y 2016. Fuente propia. Datos de IBM.

---

1 Definición de IBM recuperada de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

2 Recuperado de http://www.actibva.com/magazine/empresas/el-nacimiento-de-windows-el-error-de-ibm-que-enriquecio-a-microsoft

3 Recuperado de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

4 http://www.cisco.com/c/en/us/solutions/data-center-virtualization/big-data/index.html

5 Tecnologías de la Información y las Comunicaciones

6 Datos extraídos de http://globalnewsroom.cisco.com/es/es/release/Internet-ser%C3%A1-cuatro-veces-m%C3%A1s-grande-en-2016-1665301

Según publicó IBM[7] en 2012, para que algo sea considerado *Big Data* ha de cumplir las 4 V's:

1. Volúmen: como podemos ver en la Figura 1, hablamos de más de 100 Exabytes al mes de volúmen de información que se transmite por la red. Una cantidad de información que imposibilita su almacenamiento y análisis de un modo tradicional.

2. Velocidad: la velocidad es un factor muy importante a la hora de discernir entre *Data* y *Big Data*; influye tanto la velocidad con la que se genera todo ese volúmen de información, como la velocidad necesaria para analizarla y reaccionar. De cara a las empresas relacionadas con el *marketing,* lo importante es digerir esa información en tiempo real, o en el menor tiempo posible. Es lo que se denomina *Analysis of Streaming Data*.

3. Variedad: la peculiaridad de los datos masivos radica en la diversidad de datos que podemos encontrar. Suele tratarse de información desestructurada el 80% de las veces[8] o semi-estructurada que debe tratarse de distinta manera en función del tipo de dato que se esté tratando (fotos, videos, mensajería, direcciones de correo electrónico, notas de voz…).

4. Veracidad: la veracidad hace referencia a la calidad de la información que se está analizando; un ejemplo de ello son los *Tweet*[9] escritos en lenguaje informal, frente a artículos periodísticos publicados en alguna web con prestigio. Ambos forman parte de *Big Data*, pero gracias a las herramientas de análisis como Hadoop, es posible distinguirlos y clasificar la información que contienen esos escritos en función de su origen.

Estas son las 4 V's de referencia para identificar los datos masivos que se encuentran en Internet; sin embargo, en Marzo de 2015 Bernard Marr[10] publica un artículo para IBM[11] en el que habla de la 5ª V:

5. Valor: Marr sostiene que las 4 V's desembocan en esta última. La información no sirve de nada si no tiene valor para el que la está explotando. Este factor va muy

---

[7] Zikopolous Paul, Deroos Dirk, Deutsch Tom, Lapis George, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill, 2012

[8] Extraído de http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters

[9] Denominación de los mensajes transmitidos a través de la red social *Twitter*. Extraído de http://www.webespacio.com/que-es-un-tweet/

[10] Experto en *BigData*, escritor y colaborador en revistas y blogs como la revista Forbes

[11] Artículo publicado en http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters

enfocado a las empresas. La información valiosa es aquella que permite a Uber[12] predecir dónde va a haber más demanda de clientes y en qué horas, para tener a los conductores cerca de esas zonas; o aquella que permite a Amazon personalizar las recomendaciones de productos para un usuario, en función de las webs que ha visitado, o los sectores en los que tiene mayor interés.

Haciendo referencia al requisito "Variedad", Dataversity.net, uno de los repositorios más grandes con contenido relacionado con el tratamiento de datos en Internet, clasifica *Big Data* en 5 tipos de datos diferenciados[13]:

1) *Web and Social Media*: esto incluye todo el contenido *World Wide Web*, y los datos subidos por usuarios de redes sociales que permiten interacción como Twitter, Facebook, LinkedIn, blogs, etc.

2) *Machine-to-Machine (M2M):* hace referencia a las tecnologías que permiten la comunicación entre dispositivos a través de cable o de manera inalámbrica. M2M se compone de un dispositivo, como un sensor, que recoge información y la envía al segundo dispositivo que es el encargado de almacenarla e interpretarla. La comunicación *Machine-to-Machine* provocó la adopción del término "*Internet of Things*"[14].

3) *Big transaction data*: incluye metadatos, registros de facturación de servicios públicos, políticas de privacidad de servicios de Internet, etc. Sobre todo se compone de datos desestructurados o semiestructurados.

4) *Biometrics*: compuesto por huellas digitales, datos de reconocimiento facial, ADN… el uso de la biometría en el mundo de la tecnología está favoreciendo la creación de este tipo de *Big Data*.

5) *Human generated data*: hace referencia a toda aquella información que genera cada persona al hacer uso de sus dispositivos y aplicaciones; esto incluye notas de voz, *emails*, mensajes de texto, documentos, etc. Esta información es sensible, pues contiene datos relacionados con comunicaciones entre usuarios, y es importante que cada empresa establezca políticas de uso adecuadas. En España

[12] Empresa que ofrece servicio de taxi privado

[13] Extraído de http://www.dataversity.net/not-your-type-big-data-matchmaker-on-five-data-types-you-need-to-explore-today/?cm_mc_uid=01090455328514540019798&cm_mc_sid_50200000=1454009917

[14] *Internet of Things* 2016 http://www.dataversity.net/internet-of-things-2016-trends-leveraging-actionable-insights/

se realiza según la L.O. 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal[15].

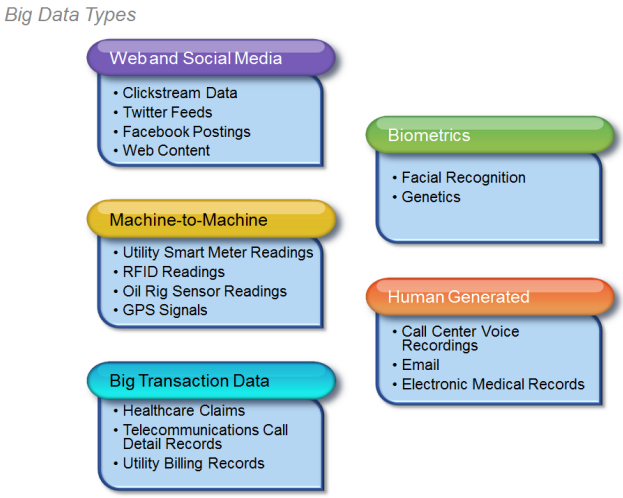

Figura 2. Tipos de datos masivos. Fuente: Dataversity.net

1.2 Factores que influyen en el aumento de datos masivos

Este crecimiento se debe a factores como la evolución del "Internet de las Cosas", que aumenta notablemente el número de dispositivos conectados a Internet. Cada vez son más los elementos tecnológicos que se conectan a la red para conseguir una interacción en tiempo real entre usuarios; desde teléfonos, tablets, neveras que se controlan el consumo de alimentos[16], coches inteligentes, hasta las ventanas de una casa con sistema domótico que se abren y cierran de manera remota. Se calculan un total de 8400 millones de dispositivos que se conectarán a finales de 2017 a Internet; es decir, algo más de 5 dispositivos por habitante en España[17].

Otro factor a tener en cuenta es el aumento del número de usuarios de Internet, directamente relacionado con el aumento de la población mundial y el desarrollo de países emergentes, sobre todo del continente africano[18]. El número de usuarios de Internet previstos por CISCO[19] para este 2017 es de 3400 millones, lo que se corresponde con un 45% de la población mundial.

---

15 BOE núm. 298, de 14/12/1999 por el que se publica la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.

16 Nevera inteligente http://www.xataka.com/otros/el-frigorifico-de-lg-te-dira-que-tienes-que-comprar-cuando-estes-en-el-supermercado

17 Datos extraídos de https://blogthinkbig.com/8-400-millones-de-dispositivos-estaran-conectados-a-internet-a-finales-de-2017

18 Extraído de http://www.tendencias21.net/La-poblacion-mundial-llegara-a-11-000-millones-de-personas-a-finales-de-este-siglo_a37182.html

19 Empresa de referencia a nivel mundial en materia de desarrollo de tecnologías de la información

El desarrollo de las tecnologías de transferencia de datos como la fibra óptica, favorecen este crecimiento, dando lugar a un aumento notable en la cantidad de minutos de video que se ven en streaming. Un ejemplo de ello son algunas plataformas como *Netflix*[20] (lanzado en España el 20 de octubre de 2015), que permiten la visualización de contenidos de video en esta modalidad, sin necesidad de descargar los videos en el ordenador personal. El tráfico IP global producido el año pasado equivale a 278 millones de personas viendo una película en alta definición en streaming de manera simultánea[21].

El tráfico de datos mediante servicios de Voz IP y videoconferencias IP va en aumento desde su implementación. Algunas empresas como *Skype* o *Whatsapp*[22] con su servicio VoIP[23], el aumento de velocidad de conexión y los dispositivos móviles con cámara integrada; hacen que, llamadas que hasta hace unos años se transmitían exclusivamente por linea telefónica, ahora se unan a ese *Big Data* que viaja a través de Internet.

1.3 Explorando *Big Data*

Una vez que las organizaciones más relevantes relacionadas con las TIC se percataron de la importancia que tiene el análisis y explotación de la información, fueron surgiendo herramientas para ordenarla o estructurarla.

La herramienta más popular para analizar *Big Data* es *Hadoop*[24]; un software *Open Source* (de código abierto) inspirado en el proyecto de *Google File System (GFS)* que consiste en dividir la programación en dos tareas (*mapper – reducer*) para manipular los datos con un alto paralelismo en el procesamiento, que para el consumidor final se traduce en una gran velocidad en el análisis de datos masivos.

Supongamos entonces que un grano de arena es información que se almacena en una playa. Toda la arena de la playa constituye *Big Data*. Si quisiéramos analizar cada componente de la arena a simple vista, nos sería imposible; existen miles de granos

[20] http://magazine.previsorabilbaina.com/te-interesa/netflix-espana-2015/
[21] http://globalnewsroom.cisco.com/es/es/release/Internet-ser%C3%A1-cuatro-veces-m%C3%A1s-grande-en-2016-1665301
[22] Datos extraídos de http://www.abc.es/tecnologia/moviles-aplicaciones/20150405/abci-whatsapp-llamadas-201504011958.html
[23] Servicio de llamadas utilizando la transmisión de datos por Internet
[24] Referencia extraída de http://hortonworks.com/hadoop/

distintos (tamaño, color, material). Lo que hace Hadoop es ejecutar su núcleo *MapReduce* de la siguiente manera[25]:

1. El proceso *map* coge un puñado de arena y lo dispone en elementos individuales separados en *tuplas* (en pares de valores).
2. El proceso intermedio denominado *shuffle*, determina qué nodo del clúster va a realizar esa tarea de reducir el puñado grande en grupos más pequeños. Un clúster es un conglomerado de ordenadores (nodos) unidos por una red de alta velocidad, que permiten realizar tareas como si se tratara de un único sistema[26]. Si cada persona que recoge arena es un ordenador, el grupo de personas encargado de analizar la arena es el clúster. El *shuffle* va diciendo a cada persona qué montón de arena tiene que analizar, para evitar que varias personas analicen el mismo puñado.
3. El proceso *reduce*, coge ese puñado de arena que le da *map* y combina las *tuplas* para formar grupitos de arena más pequeños y manejables a la hora de realizar un análisis de cada elemento.
4. Análisis de cada una de esas combinaciones de *tuplas* para extraer la información (*Data*).

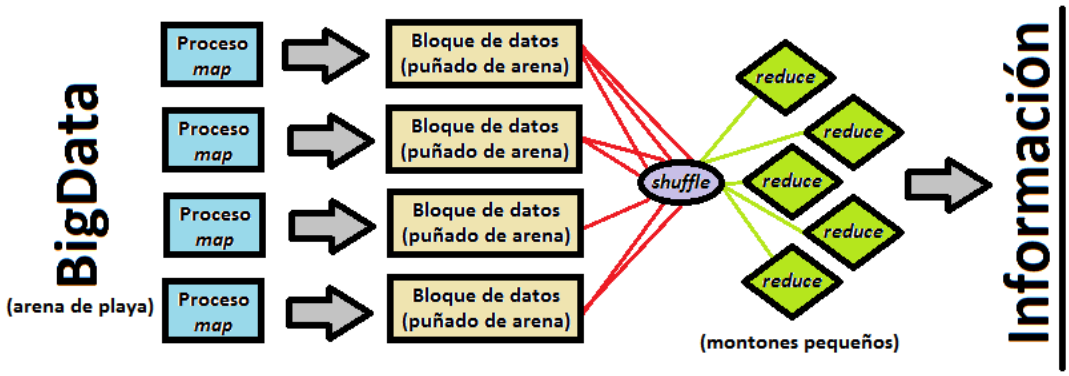

Figura 3. Fuente propia. Datos de IBM

[25] Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

[26] Definición extraída de http://www.solingest.com/blog/cluster-de-servidores-que-es-y-como-funciona

II. *BUSINESS INTELLIGENCE*

2.1 Qué es *Business Intelligence*

El término BI surge de la necesidad de explotar los datos que forman parte de *Big Data* para transformarlos en información y conocimiento, y poder así optimizar el proceso de toma de decisiones en una empresa[27].

Desde el punto de vista de las tecnologías de la información se podría definir como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada en información estructurada[28].

Toda gran organización que pretenda realizar un análisis de la información relacionada con su empresa, y que le afecta directa o indirectamente, debe asegurarse de que la captura y almacenamiento de esta información se realiza adecuadamente para facilitar el análisis posterior.

Para conseguir almacenar información de manera estructurada, es necesaria la implementación de un *Datawarehouse* o almacén de datos[29]; este almacén se conforma por una base de datos corporativa que integra y depura toda la información que llega a la base de datos desde distintas fuentes, para luego procesarla a gran velocidad, y permite la consulta y tratamiento jerarquizado de los datos almacenados.

Entre las características que definen el *Datawarehouse* destacan:

- Integración de los datos en una estructura consistente.
- Datos de temática relacionada con la necesidad de generación de conocimiento sobre la empresa o negocio. Se debe almacenar cada dato en una tabla según la temática a la que pertenece (ej. datos personales, datos de vehículos…).
- Almacenamiento según histórico de los datos; el tiempo de indexación de la información es importante de cara a obtener datos en tiempo real.
- La información ha de ser no volátil o permanente; no puede ser modificada una vez incorporada a la base de datos.

[27] Extraído de http://www.sinnexus.com/business_intelligence/
[28] Díaz, J. C. (2012). *Introduccion al business intelligence*. Editorial UOC.
[29] Extraído de http://www.dataprix.com/que-es-un-datawarehouse

Además, toda la información que se almacena en un *Datawarehouse* lleva consigo una serie de metadatos, que permiten saber en todo momento la procedencia de la información, su fiabilidad, algoritmo de cálculo, etc.

A su vez, un *Datawarehouse* se puede dividir en uno o más *Datamart*, destinados a almacenar y analizar la información que afecta a un departamento concreto. Esto permite una optimización importante en la fase de explotación de los datos[30].

2.2 Explotación de una base de datos

De forma muy básica, un ejemplo de BI con una base de datos conformada por datos estructurados puede ser la UOC. Cuando se da de alta un nuevo alumno en la base de datos, se han de rellenar una serie de campos que determinan en qué tabla y columna se va a almacenar esa información. Si, por ejemplo, se va a introducir el nombre del alumno, se rellenará el campo "nombre" que indexará esa información en la tabla correspondiente, y que guardará relación con la tabla "alumnos"; pudiéndose extraer posteriormente otros datos como la fecha de alta en la Universidad, los estudios en los que está matriculado, teléfono de contacto, etc.

Para explotar una base de datos debemos tener en cuenta 3 cosas:

- Se debe indexar toda la información y guardarla; esta información se puede extraer de Redes Sociales, buscadores, foros… este proceso se conoce como *Information Gathering* o recopilación de información sobre el objetivo.
- Los datos están desestructurados cuando realizamos este proceso de recopilación de información.
- Se han de minar esos datos mediante patrones preestablecidos. Este proceso constituye el *Data Set*, que tiene por objetivo tabular toda esa información ordenándola por filas y columnas según los patrones que se determinen.

La determinación de los patrones que van a regir el proceso de minería de datos, se apoya en el concepto de *Machine Learning*. Este proceso de "aprendizaje" es fundamental a la hora de captar de un conjunto de datos desestructurados, aquellos que realmente interesan para el caso que se esté estudiando.

[30] http://www.sinnexus.com/business_intelligence/datamart.aspx

Con este proceso de aprendizaje, mediante el estudio de algoritmos y patrones, se pretende obtener una predicción del comportamiento a tiempo real de los datos que se tienen. Muchas empresas de *online advertising* (anuncios en sitios web) explotan esta información para captar a clientes potenciales en base a su comportamiento cuando navegan por una web. Por ejemplo, es posible determinar la hora óptima para que los anuncios tengan mayor impacto; decidir cuál es la mejor hora para llamar a un cliente; predecir fallos en los equipos tecnológicos de la empresa; detectar fraude en las transacciones, etc.[31]

Aplicado a la UOC, gracias al *Machine Learning* los administradores podrían ser capaces de detectar duplicidades y fallos en la grabación de datos; simplemente se ha de programar el algoritmo para el que queremos que nos salte la alerta, y el sistema nos avisará o corregirá de manera automática ese fallo.

[31] Extraído de http://www.clevertask.com/que-es-machine-learning-big-data/

III. OBTENCIÓN DE INFORMACIÓN SOBRE OBJETIVOS EN INTERNET

3.1 ***Information Gathering***

Gracias a *Big Data*, la recopilación de información sobre usuarios de la Red en sentido genérico resulta sencilla. Basta con realizar una petición mediante una búsqueda simple como *Google Dorks*[32] a los sitios web indexados en los distintos buscadores existentes, obteniendo miles de datos como números de teléfono, direcciones postales, direcciones IP, *emails*…

En un ejemplo sencillo, podemos tratar de ver qué documentos ha escrito un alias (nombre) determinado en un foro o una web utilizando el *dork* "*site:*" y "*filetype:*":

Figura 4. Fuente propia. Extraído de *Google*.

Pero el abanico de posibilidades que ofrece este sistema permite realizar todo tipo de búsquedas más complejas que pueden aportar información relevante al proceso. En algunas ocasiones las bases de datos de los sitios web, por errores de configuración o de programación en el código, vuelcan datos que no deberían salir del servidor local[33]. Por ejemplo, los datos de registro de un determinado objetivo en un foro incluyendo correo electrónico, nombre y apellidos, número de teléfono, etc. En este segundo ejemplo, se

[32] Según el INCIBE, *Google Dorks* es una técnica de recopilación de información que permite la búsqueda de términos específicos a través del uso de caracteres determinados con el buscador de Google. https://www.certsi.es/servicios-operadores/information-gathering

[33] Rando, E. (2015). *Hacking con buscadores: Google, Bing & Shodan*. Editorial 0xworld

accede a los nombres y contraseñas de los usuarios de una base de datos de la Universidad de Vigo:

filetype:sql "# dumping data for table" "`PASSWORD` varchar"

```
#
# Table structure for table 'USUARIO'
#

# DROP TABLE IF EXISTS USUARIO;
CREATE TABLE `USUARIO` (
  `ID` bigint(20) NOT NULL AUTO_INCREMENT,
  `TIPO_USUARIO` varchar(20) DEFAULT NULL,
  `EMAIL` varchar(255) DEFAULT NULL,
  `ULTIMOACCESO` date DEFAULT NULL,
  `LOGIN` varchar(255) DEFAULT NULL,
  `FECHAALTA` date DEFAULT NULL,
  `PASSWORD` varchar(255) DEFAULT NULL,
  `VERSION` int(11) DEFAULT NULL,
  `NOMBRE` varchar(255) DEFAULT NULL,
  `APELLIDOS` varchar(255) DEFAULT NULL,
  `CODPOSTAL` varchar(255) DEFAULT NULL,
  `LOCALIDAD` varchar(255) DEFAULT NULL,
  `NIF` varchar(255) DEFAULT NULL,
  `DOMICILIO` varchar(255) DEFAULT NULL,
  `TELEFONO` varchar(255) DEFAULT NULL,
  `PROVINCIA` varchar(255) DEFAULT NULL,
  PRIMARY KEY (`ID`)
) ENGINE=MyISAM AUTO_INCREMENT=5 DEFAULT CHARSET=latin1;

#
# Dumping data for table 'USUARIO'
#

INSERT INTO USUARIO VALUES (1,'administrador'[redacted]'2011-01-24','ana','2011-01-24','MwB+aDVFLGv1afWMderst9h2TCAxAaix',1,'Ana','Pére
INSERT INTO USUARIO VALUES (2,'administrador'[redacted]'2011-01-24','eva','2011-01-24','yjP8Y4Z5UX65i6wlCI9piwdy9gHQN7eI',1,'Eva','Garc
INSERT INTO USUARIO VALUES (3,'cliente','luis[redacted]11-01-24','luis','2011-01-24','W69Hv5nCOyjnFHaadauDgLOE47gBWyum',1,'Luis','Fernár
INSERT INTO USUARIO VALUES (4,'cliente','marc[redacted]'2011-01-24','marcos','2011-01-24','5ZahAja159+WHaOWqUX+WngwHk3mn2Uy',1,'Marcos'
```

Figura 5. Fuente: Área de ciencias de la Computación de la Universidad de Vigo

En definitiva, sobre toda esta inmensa cantidad de datos que se encuentran alojados en Internet, es posible obtener información mediante técnicas como OSINT (*Open Source Intelligence*), que hace referencia al conocimiento recopilado a partir de fuentes de acceso público. Estos procesos incluyen la búsqueda, selección y adquisición de la información, así como un posterior procesado y análisis de la misma con el fin de obtener conocimiento útil y aplicable en distintos ámbitos.[34]

Sin embargo, en ocasiones interesa la obtención de información sobre objetivos concretos. El éxito de esta fase de recopilación de datos dependerá siempre de lo cuidadoso que haya sido el objetivo a la hora de ceder datos que le puedan identificar; bien sea en Redes Sociales, foros o cualquier otro sitio web que almacene datos personales y se indexen posteriormente en los buscadores.

3.2 Buscadores

Debido al *boom* de Internet a principios de Siglo, las empresas empezaron a investigar cómo aparecer entre los primeros resultados de los buscadores. A partir de este momento surgen algoritmos automáticos y la necesidad de las compañías de capturar y

[34] https://www.certsi.es/blog/osint-la-informacion-es-poder

analizar la información que circula por la Red. Cada buscador utiliza sus propios algoritmos de búsqueda en función de sus políticas u objetivos.

El *Googlebot* es el robot de rastreo web de *Google*, conocido también por "*araña*", y es quién realiza todo el proceso de búsqueda y clasificación de información por palabras clave[35]. Pero como *Google*, son muchos los buscadores que aplican sus propios algoritmos para capturar la información que más interese. Un ejemplo de ello es el buscador *Shodan*.

Shodan fue creado por el experto en seguridad John Matherly motivado por la creciente tendencia al "Internet de las Cosas". Este buscador permite indexar direcciones HTTP[36] conectadas a Internet, incluso pertenecientes a la *DeepWeb*[37] que no salen en las búsquedas de *Google* o *Bing*. Es capaz de localizar cualquier dispositivo que sea visible en la Red como cámaras, alarmas de seguridad, frigoríficos, coches e incluso centrales nucleares[38].

Para acceder al mayor número de páginas posibles, los buscadores almacenan las URL de los hipervínculos en una lista (semillas), y la "araña" se encarga de acceder a ellas posteriormente. Una vez que un robot entra en un sitio web, lee el código tratando de identificar zonas como la cabecera, el cuerpo, etc. Y dentro de éstas, intenta localizar las etiquetas clave y palabras que identifican los contenidos, estructuras, archivos, sesiones, etc. de la Web[39].

El uso de buscadores se ha convertido en una fase elemental para la recopilación de información sobre objetivos concretos. Aunque es posible la implementación de rastreadores web propios, hay que tener en cuenta el gran potencial que tienen los buscadores de las grandes empresas, y aprovechar esa fase de recogida de datos que nos da acceso directo a la información ya estructurada o semiestructurada. Tanto es así, que en algunas ocasiones una simple búsqueda en Google ha servido para identificar a un objetivo del que no se conocía prácticamente nada. Es el caso de Ross Ulbricht.

Ross William Ulbricht, bajo el pseudónimo de *Dread Pirate Roberts (DPR)*, es uno de los delincuentes relacionados con las nuevas tecnologías más conocidos del

---

35 Extraído de https://support.google.com/webmasters/answer/182072?hl=es

36 Protocolo utilizado en la transmisión de datos de sitios web

37 También conocida por "Internet profunda", donde se almacenaría entorno al 80% de la información que circula por Internet, y que no es indexada por los buscadores convencionales como Google.

38 Extraído de http://computerhoy.com/noticias/internet/shodan-buscador-favorito-hackers-32011

39 Extraído del temario del Curso Online de Posicionamiento Web en Buscadores de Solucioneduca.

mundo; fue el fundador de *Silk Road*, el mercado negro más importante alojado en la *DeepWeb* entre 2011 y 2013[40] donde era posible adquirir todo tipo de productos y servicios ilegales como drogas, armas, servicios de *hacking*, dinero y documentación falsa… Llegó a generar 80 millones de dólares en comisiones, y administraba el mercado *online* a través de la red TOR[41], dificultando enormemente la trazabilidad de los intercambios de datos y de las comunicaciones entre usuarios. Además, hacía uso de herramientas como *OpenPGP*[42] para cifrar los mensajes con vendedores de *Silk Road* y otros colaboradores. En cuanto a la actividad económica, resultó prácticamente inviable realizar una trazabilidad de los *Bitcoin*[43] con los que operaba, pues añadía medidas de anonimato como el uso de *tumbler* que enrutaba los pagos con las criptodivisas a través de cientos de transacciones falsas para "lavar" los *Bitcoin*.

Pues bien, a pesar de todas las medidas y herramientas que utilizaba Ross para evitar ser identificado, se logró determinar la identidad real de *DPR* con una simple búsqueda en *Google*:

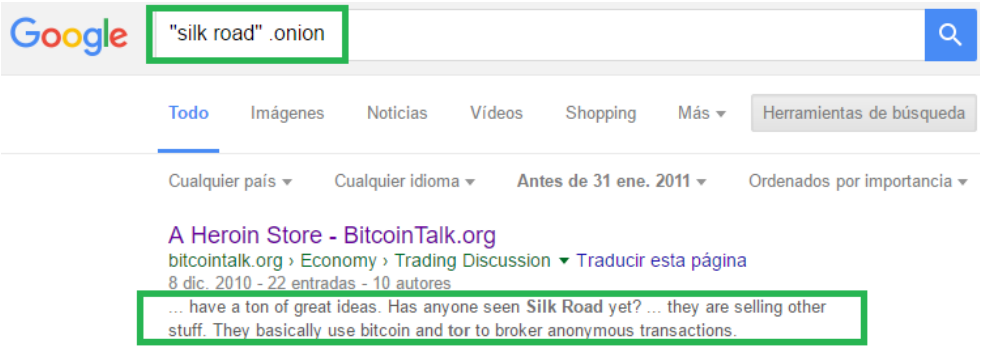

Figura 6. Fuente propia. Extraído de Bitcointalk.org

Ross Ulbricht comenzó a hacer publicidad de la web que había creado en el conocido foro *Bitcointalk*[44]; y en uno de sus mensajes, solicitaba la ayuda de algún experto en *Bitcoin* para llevar adelante la "nueva empresa"; su fallo fue que por aquel entonces no sabía que *Silk Road* fuera a llegar tan lejos y dio una dirección *gmail* con su nombre y apellidos reales.

---

40 Extraído de https://www.fayerwayer.com/tag/silk-road/
41 Extraído de https://www.fayerwayer.com/2013/10/como-el-fbi-encontro-al-fundador-de-silk-road/
42 Herramienta que permite el cifrado de comunicaciones a través de Internet mediante un sistema de clave pública y clave privada. Extraído de www.openpgp.org
43 Criptodivisa creada en 2009, que permite realizar pagos dificultando la trazabilidad a través de Internet
44 Extraído de https://bitcointalk.org/index.php?topic=47811.0

Topic: IT pro needed for venture backed bitcoin startup (Read 28831 times)

IT pro needed for venture backed bitcoin startup
October 11, 2011, 08:06:22 PM

Hello, sorry if there is another thread for this kind of post, but I couldn't find one. I'm looking for the best and bri
ideal candidate would have at least several years of web application development experience, having built applica
start-up environment is a plus, or just being super hard working, self-motivated, and creative.

Compensation can be in the form of equity or a salary, or somewhere in-between.

If interested, please send your answers to the following questions to rossulbricht at gmail dot com

1) What are your qualifications for this position?
2) What interests you about bitcoin?

From there, we can talk about things like compensation and references and I can answer your questions as well.

Figura 7. Fuente propia. Extraído de Bitcointalk.org

Este caso encaja con una de las premisas a tener en cuenta a la hora de establecer una vinculación indiciaria entre la identidad real y la/s virtual/es utilizando OSINT: nadie nace sabiendo. Existe un proceso de aprendizaje, y durante ese proceso se producen fallos que puede que podamos recuperar (pues no es sencillo destruir la información de la Red) y aprovechar estos fallos para establecer esa vinculación entre el mundo físico y el virtual.

3.3 Herramientas de búsqueda automatizada

A la hora de realizar un proceso de obtención de información sobre un objetivo concreto en Internet, es importante utilizar herramientas que faciliten dicha búsqueda. Realizar esta tarea de monitorización de forma manual, con el volumen de información que se genera y muta a cada segundo, resultaría totalmente inviable. Desde hace unos años son muy comunes este tipo de herramientas y están siendo utilizadas de forma asidua, no sólo por fuerzas policiales, sino por empresas que quieren conocer en tiempo real qué se está diciendo sobre ellas o sobre algún interés comercial que tengan.

Un ejemplo de herramienta para realizar minería de datos es *Maltego*. *Maltego* es un *software* OSINT y forense que permite recopilar información y crear relaciones entre personas, empresas, organizaciones, sitios web, direcciones IP, documentos… y poder ver toda la información extraída de Internet de manera gráfica a través de su interfaz gráfica (GUI)[45].

[45] https://www.paterva.com/web7/buy/maltego-clients/maltego-xl.php

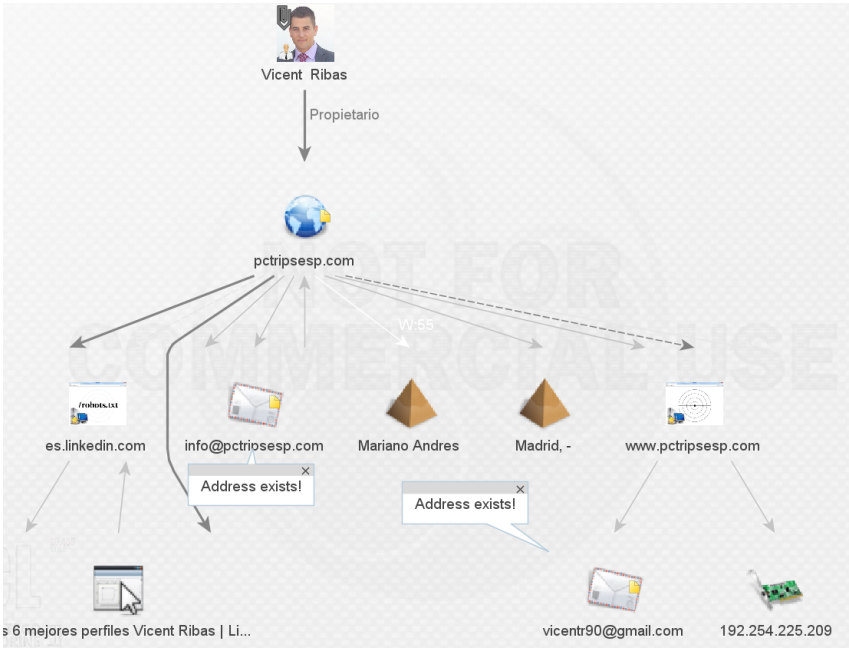

Figura 8. Fuente propia. Extraído de Maltego

La figura 8 es un ejemplo de cómo introduciendo un nombre o un pseudónimo, y ejecutando diversas transformadas de *Maltego*, es posible determinar perfiles de redes sociales de un objetivo, direcciones de correo electrónico, dirección IP de sus servidores e incluso la dirección del domicilio donde reside.

Es un hecho que las redes sociales como *Facebook*, *Twitter*, Instagram o *LinkedIn*; son una de las fuentes de información más potentes que tienen las herramientas basadas en OSINT. La gran ventaja de plataformas como *Maltego*, es que nos permite la integración de transformadas propias o de terceros para añadir datos de las APIs[46] de aquellas redes sociales que nos interesen.

[46] API es una interfaz de programación de aplicaciones, que implementan funciones que permiten al programador acceder a servicios de una aplicación a través del uso de un lenguaje de programación. Extraído de http://es.ccm.net/contents/300-lenguajes-de-programacion-ndash-api

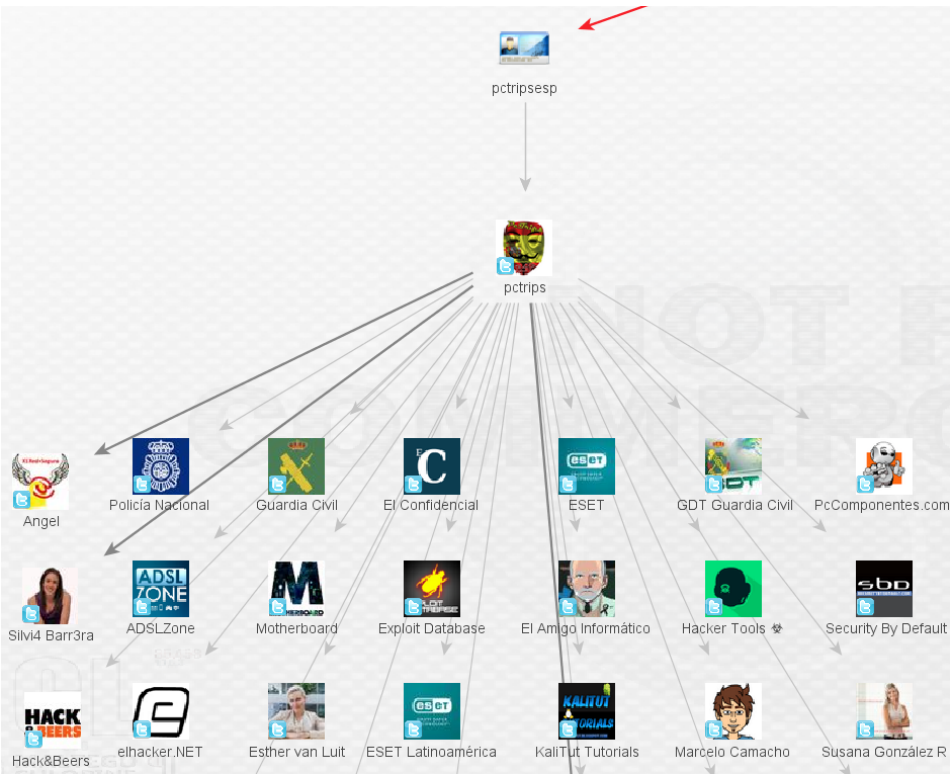

Figura 9. Fuente propia. Extraído de Maltego

En la figura 9 se pueden observar las relaciones extraídas a través de la transformada de *Twitter* que viene por defecto en la última versión del programa, y que permite establecer relaciones entre usuarios de esta red social a través del análisis del perfil dado. El potencial de esta herramienta reside en la facilidad a la hora de ir estableciendo relaciones por parte del analista entre distintos perfiles, y establecer contactos de una red delincuencial o terrorista. Así, de la misma manera que se ha realizado en el ejemplo anterior, sería posible determinar vínculos de unión entre integrantes de una célula terrorista.

Otro ejemplo de herramienta que facilita el proceso de recolección de información sobre objetivos concretos es el *OSRFramework* de Félix Brezo y Yaiza Rubio[47].

[47] Analistas en Inteligencia en el sector de las comunicaciones. Félix Brezo trabajó para el Centro de Análisis y Prospectiva de la Guardia Civil en 2012.

```
C:\            Desktop\osrframework-master\osrframework-master\osrframework>usufy.py -n pctripsesp -p twitter

                Version:      OSRFramework 0.11.0b9
                Created by:   Felix Brezo and Yaiza Rubio, (i3visio)

usufy.py Copyright (C) F. Brezo and Y. Rubio (i3visio) 2016
This program comes with ABSOLUTELY NO WARRANTY.
This is free software, and you are welcome to redistribute it under certain conditions. For additional info, visit <ht
tp://www.gnu.org/licenses/gpl-3.0.txt>.

2016-04-16 18:28:43.487000      Starting search in 1 platform(s)... Be patient!

A summary of the results obtained are shown in the following table:
Sheet Name: Profiles recovered (2016-4-16 18h28m).

,-------------------------------,-----------------,--------------------,
| http://twitter.com/pctripsesp | pctripsesp      | Twitter            |
+-------------------------------+-----------------+--------------------+

You can find all the information collected in the following files:
        -./results\profiles.xls

2016-04-16 18:28:45.917000      Finishing execution...

Total time used:        0:00:02.430000
Average seconds/query:  2.43 seconds
```

Figura 10. Fuente propia. Extraído de OSRFramework

Este *software* permite realizar búsquedas de usuarios en más de 200 plataformas web y redes sociales, almacenando los resultados obtenidos en varios formatos como *.xls* para facilitar el posterior análisis de la información recopilada, y su integración en otras herramientas enfocadas al *OSINT*.

Además, y puesto que la interfaz gráfica no es muy amigable para el analista, al ser un proyecto *Open Source*[48] es posible integrarlo en una interfaz como *Maltego* en forma de transformadas, para poder ejecutar los *scripts* directamente desde esa plataforma con interfaz gráfica. Y si se quiere añadir alguna plataforma de una red social que no venga por defecto en la herramienta, solo hay que hacerlo siguiendo los pasos de los autores en lenguaje de programación *Python*.

Son cientas las herramientas que permiten hacer esa fase de identificación de objetivos mediante la recopilación de información en Internet. Michael Bazzell, un investigador del FBI[49] en materia de cibercrimen, ha hecho una recopilación de herramientas para hacer búsquedas OSINT por temática[50]; incluyendo búsqueda de perfiles en redes sociales, direcciones de correo electrónico, documentos bajo la firma de un pseudónimo dado, números de teléfono asociados a perfiles, direcciones IP e incluso vista satelital de una dirección postal dada.

---

48 *Open Source* o Código Abierto, hace referencia al *software* cuyas líneas de programación que lo integran son públicas, de tal manera que cualquiera puede ver/modificar ese código y personalizarlo bajo licencia GNU/LINUX. Extraído de https://opensource.org

49 Federal Bureau of Investigation

50 Extraído de https://inteltechniques.com/menu.html

Figura 11. Fuente propia. Extraído de *Google*

Por último, algunos buscadores como *Google, TinEye, Bing* o *Baidu*; permiten la búsqueda inversa de fotografías y videos. Esto facilita la identificación en muchos casos de un objetivo a partir de una instantánea a pesar de que no se tengan más datos sobre este. En la figura 11 se puede ver un ejemplo de búsqueda inversa de imagen con *Google*, en la que se ha subido una fotografía de Bin Laden. El buscador responde a la petición con imágenes similares a la que se ha subido para realizar la búsqueda, y además indica que la fotografía podría estar ligada con el nombre de “Osama Bin Laden”.

IV. LA NUEVA HUELLA DIGITAL

4.1 Los orígenes de la identificación de objetivos

Desde los orígenes de la criminología, se ha mantenido una estrecha relación con la antropología a la hora de establecer relaciones entre hechos criminales e identificación de autores. En 1883, el francés Alphonse Bertillon confeccionó un estructurado sistema de mediciones corporales que, acompañadas de la reseña fotográfica y el retrato hablado, servían para filiar a un individuo con su identidad. De esta manera se conseguía evitar que no se pudiera atribuir la autoría de hechos delictivos a las personas que los cometían[51]. En 1892, el antropólogo inglés Francis Galton, demostraba la invariabilidad de las huellas dactilares a lo largo de toda la vida de un individuo, haciéndole único incluso entre hermanos gemelos. En 1896, Vucetich adopta el método de identificación galtoniano, siendo el primer policía en realizar fichas decadactilares.

En España se adopta el Sistema Oloriz en 1909 basado en el sistema Vucetich en el Cuerpo de Prisiones. Finalmente lo hace la Guardia Civil en 1914. A día de hoy, aún se sigue haciendo uso de las huellas dactilares para identificar a personas o para demostrar su participación o autoría en hechos delictivos.

Si bien es cierto que no se ha podido probar de manera fehaciente la probabilidad de choque entre dos huellas de distintas personas, Galton calculó una probabilidad de 1 entre 64 billones; superando notablemente la cifra de la población mundial actual[52].

En España, los técnicos en dactiloscopia, han considerado que doce puntos característicos son suficientes para establecer la identidad de una persona; aunque se puede afirmar la identidad establecida con un número menor, cuando así lo permita la morfología, situación y rareza de los mismos. Siendo necesarios al menos ocho cuando cuatro de ellos sean de baja frecuencia de aparición.

4.2 *Fingerprinting* en la Red

El desarrollo de las tecnologías de la información ha trasladado en muchos casos la delincuencia a Internet, lo que supone una necesidad por parte de las Fuerzas y

[51] Extraído de https://wilkhb450.wordpress.com/2013/04/29/alphonse-bertillon/

[52] Extraído de https://www.elconfidencial.com/alma-corazon-vida/2014-05-01/los-expertos-insisten-es-imposible-probar-que-no-hay-dos-huellas-dactilares-iguales_119652/

Cuerpos de Seguridad de identificar delincuentes que actúan detrás de sistemas informáticos, no siendo posible hacerlo mediante métodos convencionales como la comparación de huellas dactiloscópicas.

Surge entonces la necesidad de trasladar el concepto de "huella digital" al mundo virtual, donde no hay direcciones postales sino direcciones IP; donde no hay nombres sino alias. Sin embargo aunque pueda parecerlo, las Fuerzas y Cuerpos de Seguridad no han sido los primeros en tratar de identificar usuarios; las empresas han visto un gran negocio detrás de las huellas digitales, pues conociendo a un usuario de la Red es posible adaptar los productos a sus necesidades o gustos.

A *Google*, por ejemplo, le interesa almacenar toda la información posible relativa a búsquedas que realiza un usuario con el fin de adaptar sus anuncios. De esta manera, cuando un usuario realiza búsquedas sobre una materia determinada, *Google AdWords*[53] se encarga de relacionar esa temática con la huella digital de ese usuario, y personalizar los anuncios que va a visualizar en el futuro durante su navegación.

Este es un ejemplo de toda la información que puede almacenar *Google* cuando navegamos estando autenticados con una cuenta *Gmail*[54] u otra dirección de correo electrónico, y que forma parte de la huella digital[55]:

- Navegador de Internet
- Versión de *USER-AGENT*[56]
- Dirección IP
- Versión del sistema operativo
- Resolución de pantalla
- Tipografías utilizadas
- Elección de usar *Java* o no
- Elección de usar *Flash* o no
- Horas de conexión
- Sitios a los que se conecta
- Historial de navegación
- *Cookies*[57]
- Opciones de caché

---

53 Empresa de Google dedicada a gestionar publicidad en sitios web.

54 Gestor de correo electrónico de *Google*.

55 Conferencia del Doctor en seguridad informática Chema Alonso en las Jornadas Espacio de la Fundación Telefónica en 2014.

56 Sirve para identificar el navegador o aplicación que estamos utilizando durante la navegación en servidores para asegurar la compatibilidad de las versiones.

57 Una *Cookie* es un archivo que se almacena en el navegador de nuestro sistema al navegar por sitios web, de manera que permite al sitio web consultar la actividad previa del usuario y hábitos de navegación. Extraído de http://politicadecookies.com/cookies.php

Figura 12. Fuente propia. Extraído de whatsmyuseragent.com

Las técnicas de identificación en los navegadores se denominan *Web Browser Fingerprinting*, y consisten en tratar de averiguar quién es un determinado usuario con los datos que muestra el navegador que usa, aunque el usuario no los muestre de manera voluntaria. Lo que se consigue con este tipo de técnicas, es determinar la verdadera identidad del navegador que está en uso, aunque el usuario utilice programas para modificar los datos del *USER-AGENT*. Incluso, por un fallo de seguridad en algunos navegadores, es posible conocer la IP local del usuario[58]; permitiendo conocer, por ejemplo, qué empleado de una determinada empresa ha sido el que ha realizado determinadas operaciones en la Red.

Existen diversas herramientas que calculan la huella digital de un usuario, y muestran cuantitativamente cómo de única es la configuración de nuestro sistema en la navegación. El proyecto *Panopticlick*[59] trata de demostrar la posibilidad de identificar un usuario utilizando los datos de la configuración de su navegador. Al realizar una prueba de concepto aparece el siguiente resultado:

"Your browser fingerprint **appears to be unique** among the 137,113 tested so far. Currently, we estimate that your browser has a fingerprint that conveys **at least 17.07 bits of identifying information.**"

Es decir que, comparando la configuración de nuestro navegador con la base de datos del Proyecto, es posible determinar que esta configuración es única entre 140.000

[58] Se puede consultar la IP local en net.ipcalf.com
[59] Extraído de panopticlick.eff.org

en el mundo. Si se añade la dirección IP, se controlan las horas habituales de conexión, las webs visitadas, etc. se podría llegar a obtener una cifra realmente baja a la hora de identificar a un usuario concreto.

Si bien es cierto que no se consiguen cifras de unicidad como las de las huellas dactilares en la identificación de sujetos, sí podría utilizarse esta huella digital como prueba indiciaria no determinante en un proceso judicial; a menor probabilidad de choque, mejor podrá valorar el juez o tribunal esa identificación del objetivo de la investigación.

CONCLUSIONES

Las comunicaciones a través de Internet son cada día más frecuentes. El desarrollo de las tecnologías de la información, junto con la reciente integración del concepto del “Internet de las Cosas”, está provocando la transmisión de millones de datos cada segundo a través de la Red. El almacenamiento masivo de gran parte de esos datos es lo que hace que todo contacto deje huella en el mundo virtual. La explotación de esa información latente para elaborar inteligencia, está suponiendo una gran fuente de ingresos en el sector empresarial, y una gran fuente indiciaria para las Fuerzas y Cuerpos de Seguridad a la hora de determinar la autoría de hechos delictivos en Internet.

Se puede observar en la evolución del uso del malware, cómo cada vez son menos las infecciones en forma de virus que simplemente tratan de destruir la información de aquellos sistemas por los que pasan, y cómo cada vez son más comunes aquellas infecciones que tratan de recopilar o cifrar toda la información personal posible de los usuarios de sistemas. Información que, una vez estructurada por el ciberdelincuente, se vende en el mercado negro para ser explotada posteriormente.

El principal problema de seguridad es que el usuario no valora lo suficiente su información personal, y no es consciente de que para otros esa información sí que puede ser realmente valiosa. Tanto es así, que la mayor parte de las aplicaciones y servicios de los que somos diariamente usuarios, son gratuitos; dando peso a la máxima: "cuando no tienes que pagar por un producto, es que el producto eres tú."

A raíz de la anterior conclusión parcial, es posible afirmar la primera hipótesis planteada en la introducción de este trabajo: “la venta de información obtenida en forma de *Big Data* se ha convertido en uno de los negocios más lucrativos en el mundo.” Con un matiz: la información valiosa no es la que se adquiere en forma de *Big Data*, sino aquella que ha sido posteriormente analizada y estructurada, con el fin de transformarla en inteligencia.

En cuanto a la segunda hipótesis planteada: “es posible crear perfiles o huellas digitales con el fin de identificar a objetivos concretos.”, es cierta parcialmente. Por un

lado, la Ley Orgánica 15/1999 no permite a las empresas dar a conocer datos de carácter personal, impidiendo relacionar de manera directa a través de la explotación de fuentes abiertas a una persona física con su huella digital (hábitos de navegación, características de su sistema, etc.); entendiéndose por datos de carácter personal, cualquier información concerniente a personas físicas identificadas o identificables.

A las empresas no les interesa, como norma general, saber qué persona está detrás de un alias o usuario; sino que le basta con conocer los hábitos y preferencias del usuario que se conecta a través de ese sistema, con una dirección IP determinada, e incluso con una dirección MAC concreta; con el fin de adaptar sus productos y servicios a ese usuario. Esto permite el tratamiento de información respetando el artículo 10 de la Ley Orgánica citada anteriormente, que regula la obligación de garantizar el secreto de los datos de carácter personal.

Sin embargo, por otro lado, los ordenadores personales y los dispositivos móviles están orientados a un uso personal. Esto podría permitir que, en un momento dado, sea posible establecer una relación fehaciente entre una huella digital o usuario virtual, y una persona física con nombre y apellidos. Es decir que, desde el punto de vista objetivo, una huella digital por sí sola no constituye una prueba determinante para identificar a un objetivo; pero sí puede servir como una prueba indiciaria más en la investigación de un hecho.

BIBLIOGRAFÍA

ARDILA CAÑAS, E., GÓMEZ DÍAZ, I. C. (2015): "*Estrategias para la gestión de grandes volúmenes de datos por medio de big data en el contexto de la analítica de negocios: caso MVM ingeniería de software SAS*". Universidad de Buenaventura. Colombia.

DÍAZ, J. C. (2012): "*Introduccion al business intelligence*". Editorial UOC.

FOSTER K., NATHAN S., RAJAN D., BALLARD C. (2011): "*IBM InfoSphere Streams: Assembling Continuous Insight in the Information Revolution*". IBM RedBooks.

GARCÍA-PALOMERO I. y otros (2004): *"Guía y Norma para la realización de Trabajos Académicos Dirigidos"*.

GARCÍA PALOMERO, I., MOURE COLÓN, F. (2011): "*Metodología de la Investigación en 4D, soporte didáctico para el Diseño, Desarrollo, Defensa y Difusión de trabajos de investigación de la Guardia Civil*".

JOYANES, L. (2013). "*Big Data: Análisis de grandes volúmenes de datos en organizaciones*". Alfaomega.

LÓPEZ MUÑIZ, L. (2012): "*Big Data: orígenes y aplicación*". IBM.

MAYER-SCHÖNBERGER, V., CUKIER, K. (2013): "*Big data: la revolución de los datos masivos"*. Turner.

PRIMO, E. (1981). "*La investigación. Un problema de España".* Valencia.

RAMOS VARÓN, A. (2015): "*hacking con ingeniería social: técnicas para hackear humanos*". RA-MA.

RANDO, E. (2015): "*Hacking con buscadores: Google, Bing & Shodan*". 0xworld.

RUBIO, Y., BREZO, F. (2015): "*La utilización de herramientas de monitorización de usuarios como base para el perfilado de identidades en fuentes abiertas: OSRFramework*". Universidad Politécnica de Cataluña.

VISAUTA, B. (1989). *"Técnicas de Investigación Social y recogida de datos".* PPU. SA. Barcelona.

ZIKOPOLOUS P., DEROOS D., DEUTSCH T., LAPIS G. (2012): "*Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*". McGraw-Hill.

FUENTES ELECTRÓNICAS

https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/ Consultado el 28/9/2017
http://www.eleconomista.es/tecnologia/noticias/5578707/02/14/La-moda-del-Big-Data-En-que-consiste-en-realidad.html Consultado el 28/9/2017
http://www.actibva.com/magazine/empresas/el-nacimiento-de-windows-el-error-de-ibm-que-enriquecio-a-microsoft Consultado el 28/9/2017
http://www.cisco.com/c/en/us/solutions/data-center-virtualization/big-data/index.html Consultado el 28/9/2017
http://www.cisco.com/web/ES/about/press/2012/2012-05-30-internet-sera-cuatro-veces-mas-grande-en-2016--informe-vini-de-cisco.html?cm_mc_uid=01090455328514540019798&cm_mc_sid_50200000=1454005437 Consultado el 28/9/2017
http://www.xataka.com/otros/el-frigorifico-de-lg-te-dira-que-tienes-que-comprar-cuando-estes-en-el-supermercado Consultado el 28/9/2017
http://www.tendencias21.net/La-poblacion-mundial-llegara-a-11-000-millones-de-personas-a-finales-de-este-siglo_a37182.html Consultado el 28/9/2017
http://magazine.previsorabilbaina.com/te-interesa/netflix-espana-2015/ Consultado el 28/9/2017
http://www.dataversity.net/not-your-type-big-data-matchmaker-on-five-data-types-you-need-to-explore-today/?cm_mc_uid=01090455328514540019798&cm_mc_sid_50200000=1454009917 Consultado el 28/9/2017
http://www.solingest.com/blog/cluster-de-servidores-que-es-y-como-funciona Consultado el 30/10/2017
http://hortonworks.com/hadoop/ Consultado el 30/10/2017
http://manueldelgado.com/las-4v-del-big-data/ Consultado el 30/10/2017
http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters Consultado el 02/11/2017
http://www.sinnexus.com/business_intelligence/ Consultado el 02/11/2017
http://www.dataprix.com/que-es-un-datawarehouse Consultado el 02/11/2017
http://www.sinnexus.com/business_intelligence/datamart.aspx Consultado el 02/11/2017
http://www.clevertask.com/que-es-machine-learning-big-data/ Consultado el 02/11/2017
https://www.incibe.es/CERT/guias_estudios/guias//guia_information_gathering Consultado el 02/11/2017
http://ccia.ei.uvigo.es Consultado el 02/11/2017
https://support.google.com/webmasters/answer/182072?hl=es Consultado el 10/11/2017
https://www.shodan.io Consultado el 10/11/2017
https://www.fayerwayer.com/2013/10/como-el-fbi-encontro-al-fundador-de-silk-road/ Consultado el 10/11/2017
https://www.fayerwayer.com/tag/silk-road/ Consultado el 10/11/2017
www.openpgp.org Consultado el 10/11/2017
https://bitcointalk.org/index.php?topic=47811.0 Consultado el 10/11/2017
https://www.paterva.com/web6/products/maltego.php Consultado el 10/11/2017
http://www.internetya.co/que-es-y-para-que-sirve-una-api/ Consultado el 16/11/2017
https://opensource.org Consultado el 16/11/2017
https://inteltechniques.com Consultado el 16/11/2017
http://osintframework.com Consultado el 16/11/2017
https://sanchezlm.wordpress.com/2013/03/25/alphonse-bertillon-el-precursor-de-la-identificacion-antropometrica/ Consultado el 16/11/2017
http://laflecha.net/archivo/canales/curiosidades/noticias/hay-dos-huellas-dactilares-iguales Consultado el 16/11/2017
http://www.whatsmyua.com Consultado el 16/11/2017
http://politicadecookies.com/cookies.php Consultado el 16/11/2017
http://whatsmyuseragent.com Consultado el 17/11/2017
https://panopticlick.eff.org Consultado el 17/11/2017
http://definicion.de/web-2-0/ Consultado el 17/11/2017
http://www.webespacio.com/que-es-un-tweet/ Consultado el 17/11/2017
http://es.ccm.net/faq/1572-como-saber-tu-direccion-mac Consultado el 17/11/2017

Revistas

ARTEAGA, F. (2015): *"Ciberseguridad"*. Política Exterior. Vol. 29 págs. 148-156.
GLASSMAN, M., KANG, M. J. (2012): "*Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT)".* Computers in Human Behavior, *28*(2), 673-682.

MOYA, E. (2012): "*Las Redes Sociales como fuentes de información (OSINT)*". Boletín octubre IUISI.

BONILLA, D. N. (2009): "*Inteligencia generada por grupos terroristas: aprovechamiento de fuentes y recursos de información*". Cuadernos de estrategia, (141), 163-194. Universidad de la Rioja.

Legislación

Consejo de Europa. (2001, noviembre). Convenio sobre la ciberdelincuencia del

Consejo de Europa.

L.O. 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. Última modificación en 2011.
Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal. Última modificación en 2015.
Ley Orgánica 2/1986 de 13 de marzo, de Fuerzas y Cuerpos de Seguridad del Estado. Última modificación en 2015.
Real decreto de 14 de septiembre de 1882 por el que se aprueba la Ley de Enjuiciamiento Criminal. Última modificación en 2015.

ABREVIATURAS

API: Application Programming Interface
BI: Business Intelligence
FBI: Federal Bureau of Investigation
GNU: GNU No es Unix
GUI: Graphical User Interface
HTTP: Hypertext Transfer Protocol
IBM: International Business Machines Corporation
INCIBE: Instituto Nacional de Ciberseguridad de España
IP: Internet Protocol
LINUX: Linux Is Not Unix
M2M: Machine To Machine
MAC: Media Access Control
PGP: Pretty Good Privacy
SAID: Sistema Automático de Identificación Dactilar
TI: Tecnologías de la Información
TOR: The Onion Router
URL: Uniform Resource Locator