Desanonimización y categorización de servicios ocultos de la red Tor

Rodríguez, Ricardo J. [1] y García de Quirós, Jorge [2].

[1] Centro Universitario de la Defensa, Academia General Militar, Carr. de Huesca s/n, 50090 Zaragoza. Correo electrónico: rjrodriguez@unizar.es (R.J. Rodríguez)

[2] Dpto. de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, María de Luna 1, 50018 Zaragoza. Correo electrónico: 680258@unizar.es (J. García de Quirós)

* Autor Principal y responsable del trabajo; Correo electrónico: rjrodriguez@unizar.es (R.J. Rodríguez)

Resumen: La red Tor garantiza el anonimato y la privacidad de sus usuarios durante la navegación por Internet, mediante el establecimiento de circuitos entre los diferentes nodos de la red. Además del uso de la red Tor con fines lícitos, existen usuarios que se amparan en la propia anonimicidad proporcionada por la red Tor para cometer actividades ilícitas. Por ejemplo, existen páginas web dentro de la red Tor como la venta de drogas, documentación falsa, armas, o pornografía infantil. Estos servicios (comúnmente conocidos como servicios de la dark web) se denominan Servicios Ocultos, según la terminología Tor, estando únicamente accesibles a través de Tor. Sin embargo, muchos de estos servicios ocultos pueden presentar alguna mala configuración o elementos suficientemente identificativos que ayuden a localizar el origen real del servicio y por tanto, se pueda proceder a su eliminación y detención del autor. En este estudio, se presenta un sistema automático que recopila diversas fuentes de direcciones .onion, accede a ellas y las analiza con el objetivo de encontrar elementos que sirvan para su desanonimización. Además, se ha realizado un estudio estadístico sobre la temática y el idioma de estos sitios.

Palabras clave: Tor, privacidad, servicios ocultos, desanonimización.

---

1. Introducción

La red Tor [1] (*The Onion Router*) es una red formada por nodos puestos a disposición de los usuarios voluntariamente para mejorar su anonimato y privacidad durante la navegación por Internet. El proyecto Tor se mantiene por una organización sin ánimo de lucro encargada de gestionar el

desarrollo de la aplicación y de los protocolos de la red Tor (o simplemente *protocolo Tor*, por simplificar), así como de controlar el estado de la red.

Originalmente la red Tor fue diseñada por el U.S. Naval Research Laboratory con la idea de proteger las comunicaciones gubernamentales. Hoy en día, se usa por una amplia comunidad de individuos y con diferentes propósitos, como por ejemplo periodistas para contactar con sus fuentes de manera anónima, activistas para denunciar abusos en zonas de conflicto, miembros y fuerzas de los cuerpos de seguridad de los estados para la llevar a cabo operaciones encubiertas o de vigilancia, e incluso propósitos militares.

La red Tor proporciona un servicio de comunicación anónima de baja latencia basado en circuitos. En concreto, se forma un circuito virtual entre un nodo origen y un nodo destino garantizando así que las organizaciones y usuarios puedan compartir información sin comprometer su privacidad. Este circuito virtual se compone de numerosos nodos dentro de la red Tor que retransmiten el tráfico entre ellos, de modo que finalmente se consiguen comunicar el nodo origen y el nodo destino, pero a través de múltiples "saltos" intermedios (cada uno de estos saltos es un nodo de la red). Estos circuitos virtuales garantizan que no hay una conexión directa entre cliente y servidor, sino a través de los diferentes nodos de la red donde cada uno conoce únicamente a su siguiente nodo en la cadena. El objetivo de estos circuitos virtuales es ocultar mediante las retransmisiones intermedias la dirección IP de los partícipes en la comunicación. Recuérdese que una dirección IP es un valor numérico que se asigna a cada dispositivo cuando se conecta a Internet, permitiendo identificarlo de manera inequívoca.

Para mantener la privacidad en la navegación, todo el tráfico que discurre por un circuito Tor se encuentra cifrado. Además, cada par de nodos que se comunican dentro del circuito se mandan también entre ellos el tráfico cifrado punto a punto. Es decir, el tráfico de red se va cifrando por capas. Es por ello que el tipo de encaminamiento que realiza la red Tor recibe el nombre de *encaminamiento de cebolla*.

La propia red Tor permite que los usuarios publiquen páginas web únicamente accesibles a través de la red Tor, y no accesibles por la red de Internet convencional. Estos sitios se conocen como *servicios ocultos* de la red Tor, comúnmente conocidos como sitios de la *dark web* [2, 3]. La dirección de dominio de estos servicios es .onion, aprobada oficialmente por la IETF/IANA en 2015.

A pesar de que la red Tor y sus servicios ocultos tienen un objetivo y usos legítimos, desgraciadamente en ocasiones se han explotado con objetivos ilegítimos. El caso más conocido es *SilkRoad*, un mercado negro online creado en 2011 y clausurado dos años más tarde por el FBI que operaba en la red Tor, mayormente conocido por la venta de drogas ilegales a través de Internet [4]. Además de mercados de drogas, en la red Tor existen otros servicios ocultos orientados hacia otras actividades ilícitas, como venta de documentos falsificados, venta de armas o radicalización terrorista. Es por tanto de interés disponer de mecanismos que permitan una rápida actuación de las fuerzas y cuerpos de seguridad del Estado (FCSE) para la identificación de los posibles criminales detrás de estos servicios ocultos.

Este trabajo presenta un sistema automático que recopila direcciones .onion de diferentes fuentes y accede a cada una de ellas para analizarlas, distinguiendo elementos para su desanonimización. Además, se ha realizado un estudio estadístico sobre la temática y el idioma de los sitios recopilados. De las 1796 direcciones recopiladas y con acceso, se ha conseguido localizar un dispositivo similar para 346 de ellas. Respecto al estudio estadístico, se han detectado servicios ocultos relacionados con

el terrorismo, la pornografía, la venta de drogas y las criptomonedas, mayoritariamente en lengua inglesa.

Trabajo relacionado. Diversos tipos de aproximaciones se han utilizado con éxito para desanonimizar un servicio oculto mediante la participación directa en el circuito virtual [5] o para identificar el servicio oculto al que se conecta un usuario mediante análisis del tráfico [6, 7]. Una categorización más exhaustiva de los servicios ocultos de Tor se presenta en [8]. Similar a este trabajo, en [9] se analizan 6426 direcciones .onion, estableciendo conexión con 1974 y con un porcentaje de éxito en la desanonimización del 5%.

Implicaciones éticas. Conviene destacar que se ha accedido a los servicios ocultos de manera automática y sólo a su página principal, descartando cualquier contenido gráfico y procesando sólo texto, evitando así la descarga de contenido ilegal y no deseado. El sistema desarrollado sirve a las FCSE a encontrar servicios ocultos con fines ilegítimos, evitando que actúen bajo el amparo del anonimato sobre la red Tor.

Este artículo se organiza de la siguiente manera. En la Sección 2 se describe el funcionamiento de los servicios ocultos en más detalle. Después, la Sección 3 describe el sistema desarrollado para la desanonimización. La discusión de los resultados de las direcciones .onion capturadas se explica en la Sección 4. Por último, la Sección 5 concluye el artículo.

2. Los servicios ocultos de la red Tor

Los servicios ocultos se introdujeron en el año 2004 como un mecanismo que permitía la anonimicidad de quién responde a las peticiones de un cliente dentro de la red Tor. Los servicios ocultos proporcionan servicios de Internet (e.g., páginas web, conexiones SSH, servidores FTP, etc.) de manera similar a un servidor convencional, pero con la peculiaridad de que los clientes no conocen la dirección IP del servidor. Para lograr este tipo de conexión, la conexión entre el servicio oculto y el cliente sucede a través de un punto de cita (*rendezvous*).

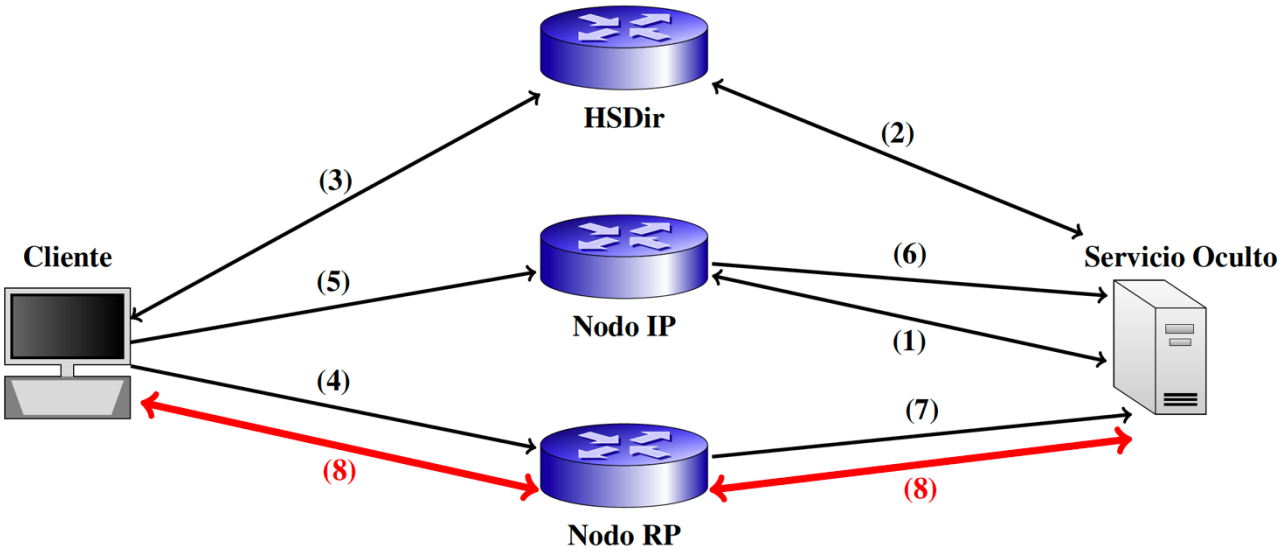


Figura 1. Funcionamiento de los servicios ocultos de la red Tor.

En concreto, la forma de conexión con los servicios ocultos se ilustra en la Figura 1. Nótese que las conexiones que se ilustran en la figura son circuitos virtuales. En primer lugar, el servicio oculto escoge aleatoriamente una serie de nodos en la red Tor, llamados puntos de introducción (*nodo IP*), y

establece un circuito virtual con ellos (paso 1). Después (paso 2), el servicio oculto comunica su descriptor y la lista de nodos IP escogidos a una lista de nodos directorio de servicio oculto (*HSDir*). El descriptor del servicio oculto está formado por una cadena no nemotécnica codificada en base 32. Para conectarse al servicio, el cliente sólo tiene que proveer el nombre de dominio del mismo, que es el identificador del servicio oculto junto con el dominio .onion. El cliente solicitará a los HSDir los nodos IP del servicio oculto (paso 3). Tras ello, escogerá de manera aleatorio un nodo de la red Tor para hacer de nodo rendezvous (*nodo RP*; paso 4) y lo comunicará a uno de los nodos IP (paso 5). El nodo IP mandará la información al servicio oculto (paso 6), quien establecerá un circuito virtual con el nodo rendezvous seleccionado anteriormente (paso 7). A partir de entonces, el cliente y el servicio oculto se comunican a través de este nodo rendezvous (paso 8).

3. Sistema desarrollado para captura de direcciones .onion y desanonimización

Esta sección explica el sistema desarrollado para la captura de direcciones .onion y los datos relevantes usados para la desanonimización. Este sistema, mostrado esquemáticamente en la Figura 2, realiza dos tareas diferenciadas:

- **Recopilación de direcciones .onion.** Las direcciones que permiten el acceso a los servicios ocultos a través de Tor no están publicadas en ningún documento de la red Tor, así que hay que obtenerlas por otros medios. Esta parte del sistema, denominada *Crawler*, establece conexiones a través de un cliente Tor (*Onion proxy*) con los servicios ocultos asociados a un repositorio de direcciones .onion dadas, almacenando información sobre ellos en una base de datos MySQL. Además, de manera automática se analiza la página principal del servicio oculto en busca de nuevas direcciones .onion que serán procesadas posteriormente.

  La información recopilada para la desanonimización han sido las cabeceras del servicio web proporcionado. Estas cabeceras contienen información adicional sobre el servicio, como información del servidor, lenguaje del contenido, y otras características técnicas del servidor. Adicionalmente, también se han recopilado los certificados SSL/TLS dado que suministran más información de interés, como la identidad del propietario del certificado, la firma de la entidad que lo expide, su fecha de validez, etc. Por último, el código HTML de la página principal del servicio oculto ha sido usado para la categorización del tipo de actividad ofertado por el servicio oculto.
- **Desanonimización y detección de actividades ilegítimas de servicios ocultos**. La otra parte del sistema se encarga del análisis de los datos recopilados anteriormente. Así, *Onion tool* extrae las características que permiten acotar el número de dispositivos conocidos asociados a un servicio oculto. A este respecto, se ha utilizado Shodan, un motor de búsqueda de metadatos (conjunto de información que da información sobre otros datos) de servicios en todo el espectro de direcciones IPv4 ubicados en puertos accesibles en Internet. Para la obtención de la combinación de cabeceras que acote un menor número de dispositivos para cada servicio oculto se ha desarrollado un algoritmo voraz que permite encontrar dispositivos en Shodan con cabeceras de sus servicios muy similares o iguales a los servicios ocultos encontrados en el punto anterior.
  Esta parte del sistema también detecta el tipo de actividades ofrecidas por los servicios ocultos analizados. Para ello, se ha definido un diccionario con un conjunto de términos relacionados con actividades ilegítimas en lengua inglesa. En concreto, se han considerado las categorías de drogas (*marijuana*, *cocaine*, *meth*, etc.), contenido sexual (*porn*, *pedophile*, *anal*, etc.), criptomonedas (*bitcoin*, *ethereum*, *litecoin*, etc.) y terrorismo

(*terrorism*, *yihad*, *IED*, etc.). El análisis realizado consiste en contar el número de apariciones de palabras de estos diccionarios contenidas en la página principal del servicio oculto. De manera adicional también se ha detectado el lenguaje utilizado en el servicio oculto usando herramientas de procesamiento de lenguaje natural (como la librería NLTK).

Toda la información extraída de los servicios ocultos puede consultarse a partir de la herramienta *Onion query*, desarrollada adicionalmente.

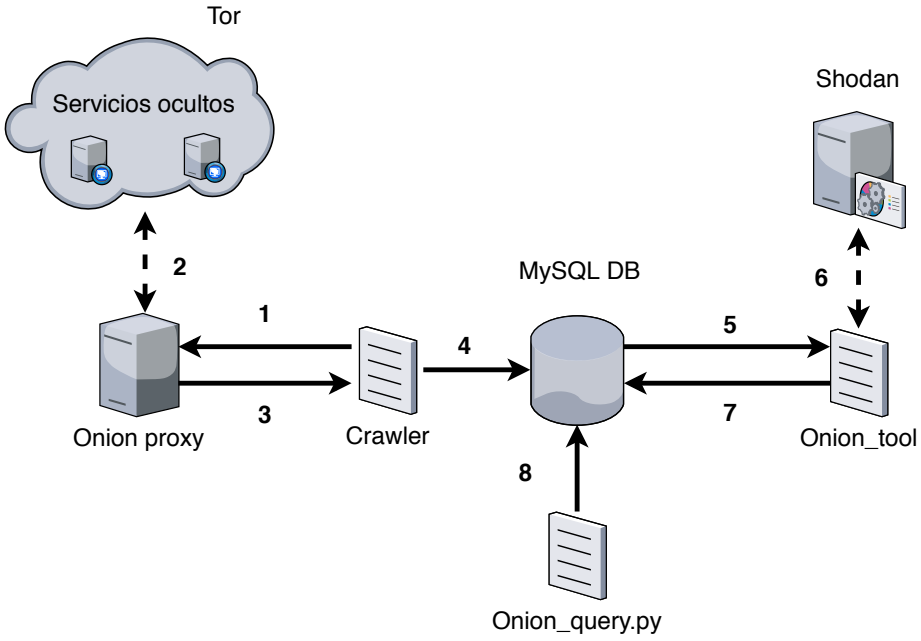


Figura 2. Esquema del sistema desarrollado.

4. Análisis y discusión de resultados obtenidos

Se han obtenido 17328 direcciones .onion (poco más de un 15% de las 110.000 direcciones reconocidas oficialmente). De todas ellas, únicamente se ha logrado establecer una conexión HTTP/HTTPS al 10%, considerando un tiempo máximo de respuesta de alrededor de 30 segundos. Este número tan bajo de conexiones puede ser debido a que los servicios ocultos permiten configurarse para solicitar una contraseña de acceso. Además, los servicios ocultos pueden ofrecer otros protocolos distintos del estudiado aquí (como conexión mediante SSH o Telnet, entre otros). Por último, cabe destacar que algunas direcciones de muchos de los servicios ocultos que se estiman se encuentran caídos o eliminados, existiendo además una alta rotación de direcciones de los servicios ocultos actuales (cada cierto tiempo cambian su dirección para evitar ser perseguidos).

De las 1796 direcciones recopiladas, con 346 de ellas se ha conseguido acotar a un sólo dispositivo con cabeceras idénticas o muy similares. En términos generales, el sistema desarrollado consigue acotar el 30% de los servicios ocultos analizados a 10 dispositivos o menos. En la Figura 3 se muestra una gráfica del número de dispositivos acotados por servicio oculto.

Nótese, sin embargo, que el sistema no garantiza una relación directa entre el servicio oculto y el dispositivo encontrado como similar. Esto no implica un error, dado que puede ocurrir que el servicio oculto sea una web diferente al servicio accesible por el Internet convencional pero la configuración del servidor sea idéntica para ambos servicios. Por ejemplo, en la Figura 4(a) se muestra un caso de identificación encontrado donde las páginas web son diferentes, pero se trata de la misma organización.

Por tanto, los resultados sí que implican que los metadatos suministrados por una cantidad considerable de los servicios ocultos accedidos pueden llevar a una desanonimización de los mismos, en caso de que se disponga de otro servicio con similar configuración en un puerto abierto a la red convencional. Este es el caso, por ejemplo, de una de las identificaciones realizadas, mostrado en la Figura 4(b).

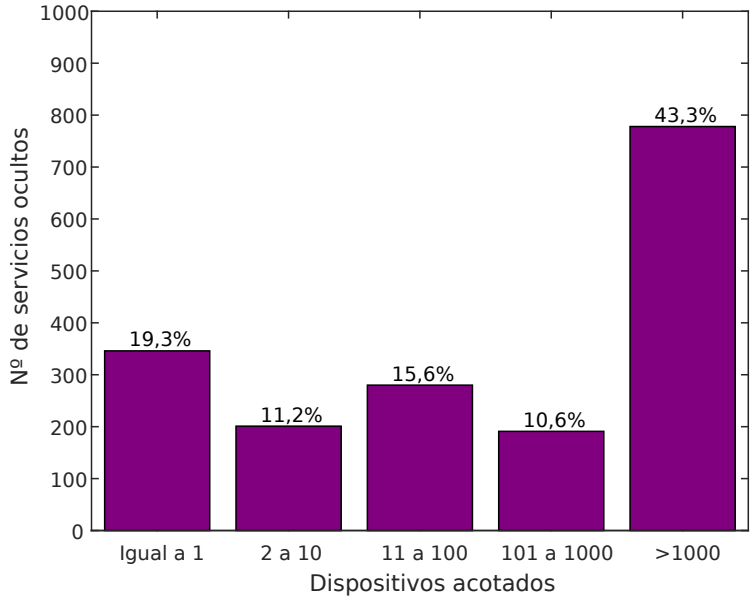


Figura 3. Dispositivos acotados por servicio oculto.

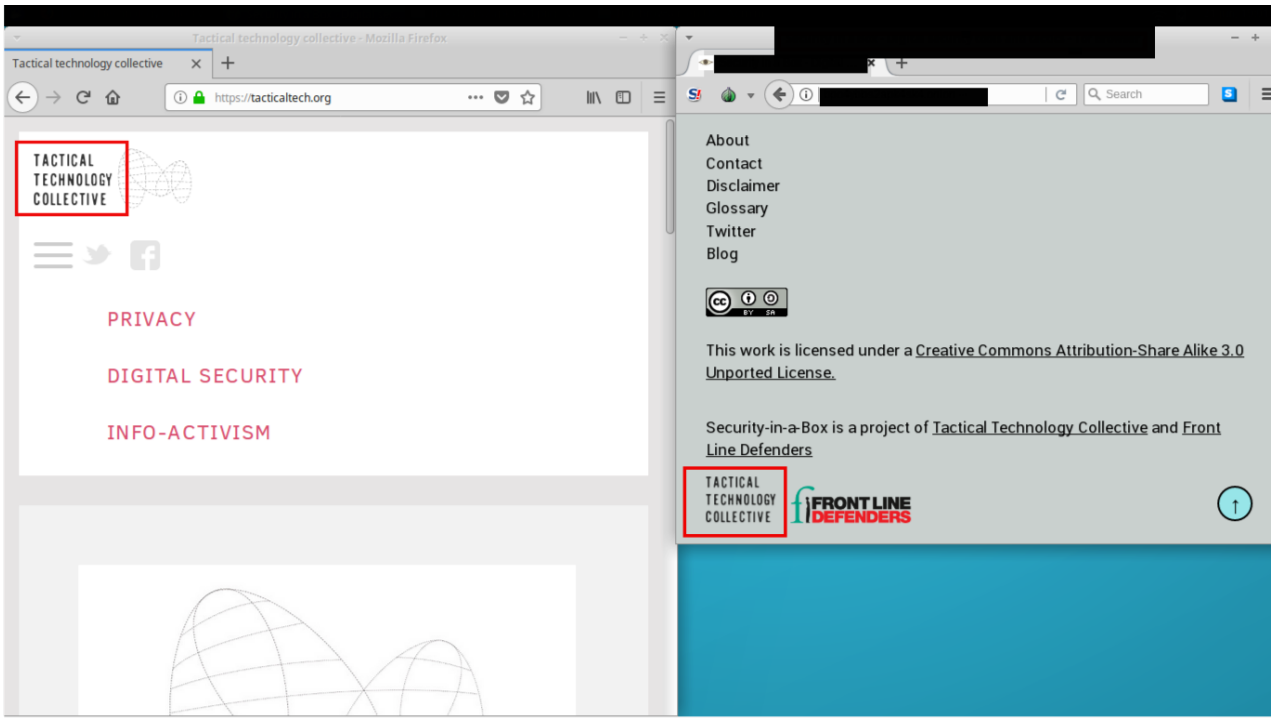
En la parte de detección de actividades ilegítimas se han detectado servicios ocultos relacionados con el terrorismo, la pornografía, la venta de drogas y las criptomonedas (orden creciente de servicios). Aunque no se ha realizado un sistema de clasificación automático, se pueden detectar servicios con una cantidad importante de palabras de estas categorías, identificándolo como sospechoso de este tipo de actividades para una posterior revisión humana. El estudio preliminar (y fácilmente ampliable) muestra que existen en Tor servicios ocultos con finalidades cuestionables.

Respecto a los idiomas, se han detectado 28 idiomas diferentes de los 55 soportados por la herramienta usada. El más utilizado es con diferencia el inglés, asociado a 1314 servicios ocultos. El idioma castellano aparece únicamente en 25 servicios.

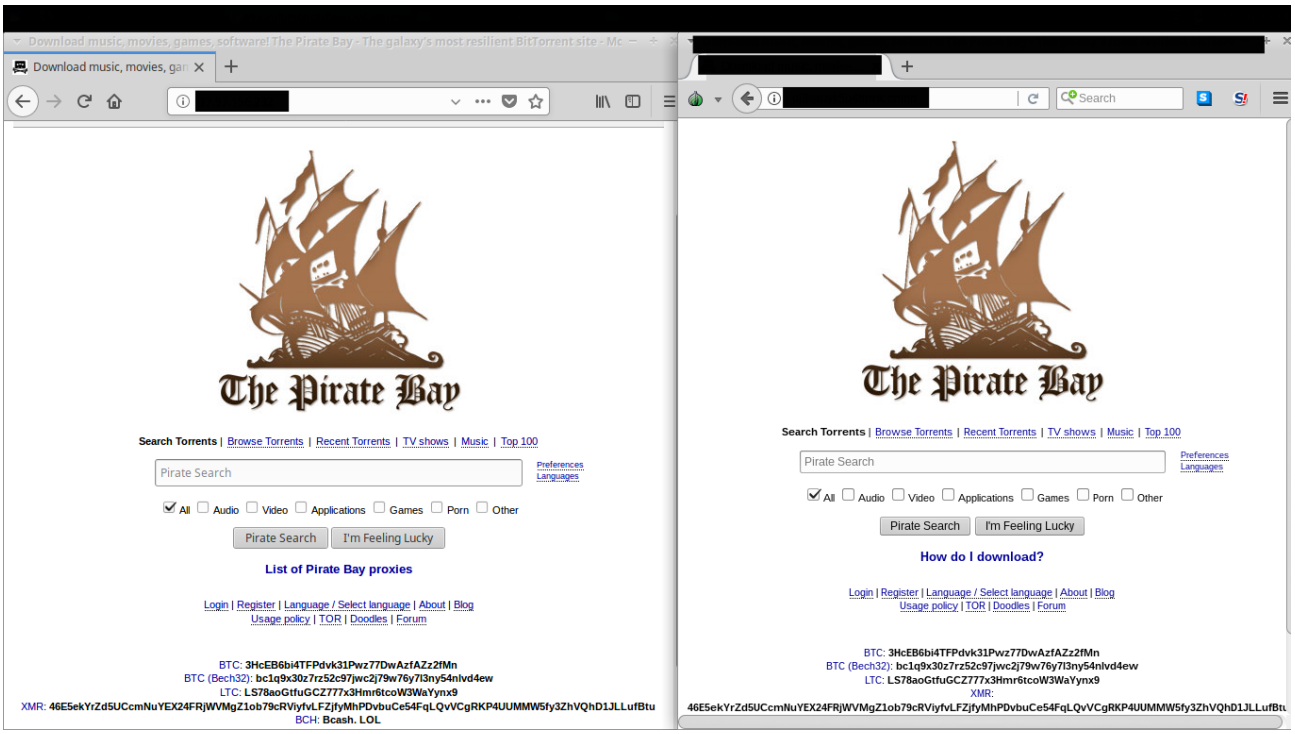
5. Conclusiones y trabajo futuro

La red Tor nació para mejorar el anonimato y la privacidad de sus usuarios durante la navegación por Internet. Basada en el establecimiento de circuitos entre los diferentes nodos de la red se garantiza la no trazabilidad de las conexiones. Además, Tor permite la creación de servicios ocultos únicamente accesibles a través de la propia red Tor y no mediante el Internet convencional, garantizando el anonimato del proveedor de servicios. Este anonimato conlleva a que proliferen sitios en la red Tor de dudosa reputación.

En este trabajo se ha desarrollado un sistema que permite analizar los servicios ocultos para desanonimizarlos. El sistema desarrollado consigue acotar la búsqueda a menos de 10 dispositivos en más de un 30% de los servicios analizados, usando simplemente los metadatos de los protocolos web HTTP/HTTPS. Además, un estudio estadístico realizado sobre la temática y el idioma ha detectado muchos servicios ocultos relacionados con la venta de drogas y las criptomonedas, mayoritariamente en lengua inglesa. Como trabajo futuro, se plantea la mejora del proceso de desanonimización y la categorización de los servicios..



(a) Misma organización



(b) Mismo contenido

Figura 4. Ejemplos de desanonimización: (a) misma organización y (b) mismo contenido.

Agradecimientos

Este trabajo ha sido subvencionado en parte por el proyecto MINECO CyCriSec (TIN2014-58457-R) y en parte por el proyecto CUD-2017-14 (financiado por el Centro Universitario de la Defensa-Zaragoza). Los autores quieren mostrar su agradecimiento a José Merseguer, de la Universidad de Zaragoza, y a Francisco Monserrat Coll, de RedIris, por toda la ayuda prestada durante la realización de este trabajo.

Referencias

1. McCoy D, Bauer K, Grunwald D, Kohno T, Sicker D. Shining Light in Dark Places: Understanding the Tor Network. Springer Berlin Heidelberg; **2008**. p. 63–76.
2. Chen H. Dark Web: Exploring and Data Mining the Dark Side of the Web. Integrated Series in Information Systems. Springer New York; **2011**.
3. Goulet D, Johnson A, Kadianakis G, Loesing K. Hidden-service statistics reported by relays. The Tor Project; **2015**. 2015-04-001.
4. Christin N. Traveling the Silk Road: A Measurement Analysis of a Large Anonymous Online Marketplace. En Proceedings of the 22nd International Conference on World Wide Web, WWW '13. ACM; **2013**. p. 213–224.
5. Overlier L, Syverson P. Locating hidden servers. En: 2006 IEEE Symposium on Security and Privacy (SP'06); **2006**. p. 15 pp.–114.
6. Panchenko A, Mitseva A, Henze M, Lanze F, Wehrle K, Engel T. Analysis of Fingerprinting Techniques for Tor Hidden Services. En: Proceedings of the 2017 on Workshop on Privacy in the Electronic Society, WPES '17. ACM; **2017**. p. 165–175.
7. Overdorf R, Juarez M, Acar G, Greenstadt R, Diaz C. How Unique is Your .Onion?: An Analysis of the Fingerprintability of Tor Onion Services. En: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17. ACM; **2017**. p. 2021–2036.
8. Owen G, Savage N. Empirical analysis of Tor Hidden Services. IET Information Security. **2016**;10(3):113–118.
9. Matic S. Active Techniques for Revealing and Analyzing the Security of Hidden Servers [phdthesis]. Università degli Studi di Milano, Milan, Italy; **2016**.