di Giuseppe Di Ieva e Gianpaolo Zambonini

DAL SURFACE AL DARK WEB, PASSANDO PER IL DEEP WEB (II PARTE)

Giuseppe DI IEVA ha conseguito il Master "Forensic Computing and Cybercrime Investigation" presso l'University College Dublin, Ireland. Presta servizio presso la Polizia di Stato con ruoli di responsabilità e coordinamento nelle attività di indagini tecniche della Polizia Scientifica, in materia di Computer e Cellular Forensics, Lawful Interception e analisi delle informazioni.

Gianpaolo ZAMBONINI, Primo Dirigente Ingegnere della Polizia di Stato, è Direttore della IV Divisione del Servizio Polizia Scientifica, nonchè Direttore della Sezione Indagini Elettroniche, presso il Dipartimento di Pubblica Sicurezza del Ministero dell'Interno.

PRIMA PARTE (nel precedente numero): 1. Introduzione, 2. Surface Web, 2.1. Raccolta/Indicizzazione/Ordinamento, 2.2. Link Popularity vs PageRank, 2.3. I motori di ricerca del Surface Web, 3. Deep Web, 3.1. Alla scoperta del Deep Web
SECONDA PARTE (in qesto numero): 3.2. Anatomia di una query, 3.3. L'Albo pretorio on-line, 3.4. I contenuti del Deep Web, 3.5. La rete Usenet come Deep Web, 4. Dark Web, 4.1. L'anonimato, 4.2. I Proxy, 4.3. Tor, 4.4. La rete Tor, 4.5. Hidden Service, 4.6. Navigare su un sito onion, 4.7. Basta un click. **TERZA PARTE**: 4.8. La valigetta degli attrezzi per il Dark Web, 4.9. ACTIVE o INACTIVE onion site, 4.10. I motori di ricerca del Dark Web, 5. Conclusioni.

3.2. Anatomia di una query

Prima di addentrarci nell'illustrare aspetti concettuali e servizi del Deep Web, è importante parlare delle possibilità che offrono gli operatori booleani e gli operatori di ricerca avanzati di rovistare più in profondità gli aspetti specifici di ciascuna SERP (Search Engine Results Page - pagina dei risultati del motore di ricerca). La necessità dell'utilizzo di queste tecniche nasce dallo studio del posizionamento del proprio sito web nella *top search* di un motore di ricerca.
Sicuramente, utilizzando queste sintassi, è possibile scendere in basso nel Surface Web, tanto da toccare anche il Deep Web e alcune di quelle categorie che ne fanno parte. Ricordiamo sempre, che non è detto che un motore di ricerca riesca a trovare tutte le pagine che ci interessano. Ciò può essere dovuto sostanzialmente a due ragioni:

- la prima, banale, è che la pagina non è ancora stata indicizzata dal motore o il sito non è stato segnalato;
- la seconda, invece, dipende dal gestore delle pagine il quale può desiderare che esse non vengano indicizzate per qualche ragione, negando quindi l'accesso a parti di un sito.

Per giocare e capire gli operatori booleani è possibile visitare il sito web (http://www.kidzsearch.com/boolify/) molto istruttivo e che aiuta l'utente a comprendere le ricerche web mostrando visivamente la logica della ricerca. Quando l'utente aggiunge o toglie i pezzi con l'operatore "and", "or" e "not" i cambiamenti dei risultati nella ricerca appaiono in fondo alla schermata della ricerca. Nel sito è possibile combinare anche esempi con gli operatori di ricerca avanzati. Di questi non tutti i motori ne fanno uso, infatti nella tabella seguente sono mostrati alcuni dei più usati e la loro corrispondenza all'interno di Google, Yahoo e Bing.

Per capire meglio gli operatori di ricerca avanzati dobbiamo tener presente sempre che la loro sintassi è costituita da:

operatore:termine_di_ricerca

Per fare qualche esempio proviamo a segmentare un url e vediamo che:

1- rappresenta la ricerca avanzata **site:**. Tutte le pagine all'interno di un determinato dominio e tutti i suoi sottodomini, se si cerca senza il www.

Per questa molti operatori hanno bisogno del http, mentre altri, come Google, vogliono solo la sintassi: **site:poliziadistato.it** o **site:www.poliziadistato.it.**

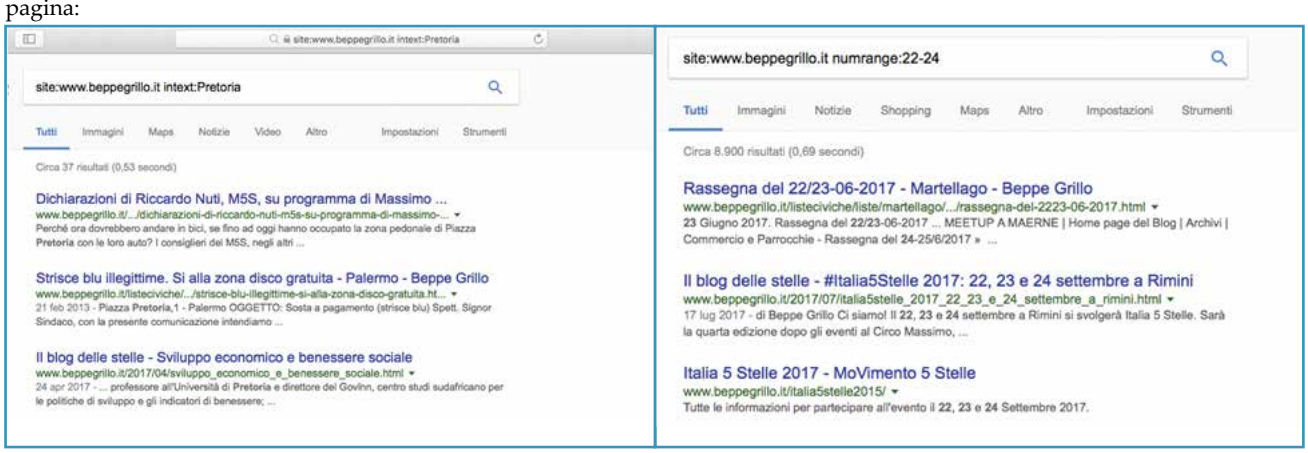
2- rappresenta la ricerca avanzata di **inurl:**. Pagine che contengono la parola chiave nel URL.

3- rappresenta la ricerca avanzata **filetype:.** Pagine di uno specifico tipo di file che contengono la parola chiave.

Google	Yahoo	Bing	Termini di ricerca	Risultati della ricerca
site:	site:	site:	url	Tutte le pagine all'interno di un determinato dominio e tutti i suoi sottodomini, se si cerca senza il www.
cache:			url	La versione cache di una pagina.
correlate:			url	Pagine che sono "simili" a un URL specificato
link:	link:	link:	url	Pagine che puntano a un determinato URL (Nota: in yahoo, debbono usare http://.)
filetype: o ext:	originurlextension:	filetype:	filetype parola chiave	Pagine di uno specifico tipo di file che contengono la parola chiave.
intitle: title:	intitle: title:	intitle:	parola chiave	Pagine che contengono la parola chiave nel tag title.
allinurl:			parola chiave parola chiave	Pagine che contengono tutte le parole chiave nel URL.
inurl:	inurl:	url:	parola chiave	Pagine che contengono la parola chiave nel URL.
allintext:			parola chiave parola chiave	Pagine che hanno il corpo del testo che contiene le tutte le parole chiave.
intext:		inbody:	parola chiave	Pagine che hanno il testo che contiene la parola chiave.

Tabella 1

Se si vuole cercare quale sito ha postato una determinata informazione si può effettuare direttamente ricerche all'interno della pagina:



1- rappresenta la ricerca **intext:** Pagine che hanno il testo che contiene la parola chiave. Questo può essere utilizzato da solo o per esempio, abbinato a **site: "site:www.beppegrillo.it intext:Pretoria",**

2- oppure utilizzando la *query* **numrange:,** è possibile la ricerca *range* di numeri che siano date o valori non importa.

Queste query sono chiamate *dork*: "testi di ricerca preparati ad hoc composte da diverse *keywods"*. Questi *testi se* vengono immesse in un motore di ricerca è possibile ottenere dei risultati specifici.

Nella tabella 1 non sono presenti tutte le *dorks*. La guida ufficiale sulle è possibile reperirla al link: https://www.google.com/support/enterprise/static/gsa/docs/admin/72/gsa_doc_set/xml_reference/request_format.html#1086546.

Un'altra alternativa è utilizzare il motore di ricerca avanzata di Google presente al link https://www.google.it/advanced_search, oppure per chi invece vuole addentrarsi in *query* molto più spinte è possibile effettuare delle ricerche su:
https://www.exploit-db.com/google-hacking-database/

3.3. L'Albo pretorio on-line

Il concetto e gli esempi del Deep Web, non sono poi così *profondi*. Senza andare troppo lontano, dobbiamo pensare all'"Albo pretorio on-line". Se accediamo al sito on-line del nostro Comune, è possibile osservare che esso contiene una sezione chiamata "Albo pretorio on-line".

Questo Albo è stato istituito con legge del 18 giugno 2009 n.69, che ha previsto che "… *gli obblighi di pubblicazione di atti e provvedimenti amministrativi aventi effetto di pubblicità legale si intendono assolti con la pubblicazione nei propri siti …"*. L'"Albo pretorio on-line" rappresenta uno spazio informatico accessibile senza formalità da chiunque, ma il Garante per la protezione dei dati personali, ha stabilito con le sue *linee guida* che è necessario evitare l'indicizzazione dei documenti non pubblicati per finalità di trasparenza tramite motori di ricerca generalisti, privilegiando funzionalità di ricerca interne ai siti web delle amministrazioni.

Quindi, tali documenti, pur essendo esposti su siti accessibili a chiunque, ma raggiungibili solo tramite un motore di ricerca interno al sito stesso, possono essere considerati parte del Deep web.

3.4. I contenuti del Deep Web

Nel Deep Web fanno parte sicuramente anche i contenuti non testuali. Tra questi bisogna fare una distinzione tra quei contenuti direttamente fruibili attraverso un comune browser e quelli che richiedono uno specifico programma per essere utilizzati.
Nella prima categoria troviamo i file multimediali, che sono privi di *tag* e di riferimenti testuali in grado di identificarne il cotenuto e quindi non possono essere correttamente indicizzati. Dobbiamo immaginare che in realtà Google non è in grado di comprendere correttamente il contenuto di un'immagine, ma lavora per somiglianze e similitudini. Il motore di ricerca, in questo modo, è in grado di raggruppare immagini con contenuti simili, poi se una di queste contiene un *tag* che la descrive, sarà in grado di indovinarne il contenuto.
Della seconda categoria fanno parte i giochi on-line, il *peer to peer* (P2P), i contenuti condivisi attraverso *Usenet*, i contenuti condivisi attraverso le chat o come abbiamo visto i contenuti presenti all'interno di server ftp. Chi tra queste categorie offre un gran numero di informazioni, oltre al mondo P2P, è sicuramente quello della *Usenet*.

3.5. La rete Usenet come Deep Web

La rete *Usenet* rappresenta il panorama del materiale dei gruppi di discussione. Usenet nasce dall'unione dei due termini inglesi User e network.
Questa rete mondiale, formata da migliaia di server interconnessi tra loro, raccoglie articoli testuali, news e i file, che le persone che hanno accesso alla rete, inviano in un determinato gruppo di discussione, sul quale viene aperto un *topic* che rappresenta il tema dello specifico gruppo di discussione.
Una caratteristica che arricchisce questa base dati e che, tutti i messaggi inviati a un server si trovano duplicati su tutti gli altri server anche se, per motivi di economia/spazio, non tutti i server contengono gli stessi *news group*.
Molti server hanno delle *newsgroup* private e anche per questo aspetto la rete Usenet si può definire parte integrante del Deep Web: il contenuto dei gruppi sfugge alle indicizzazioni e esistono gruppi di discussione segreti, la cui esistenza è nota esclusivamente a coloro i quali ne conoscono ubicazione e password di accesso.
I gruppi di discussione all'interno della rete sono divisi in gruppi testuali, supportano solo messaggi di testo, e gruppi che invece consentono di inviare anche degli allegati, denominati *Binaries.*

In questi gruppi è assai difficile rilevare la stima della quantità di materiale presente sui loro server, ma secondo il sito www.binsearch.info attualmente questi gruppi contengono circa 22 PetaByte di informazioni, escludendo i gruppi ospitati su server segreti.
La caratteristica principale di questi gruppi che fanno parte integrante del Deep Web, è, appunto, la non indicizzazione e l'accesso con appositi programmi. Questi programmi sono noti come *news reader*, ma per poter accedere a questi gruppi è possibile anche utilizzare il web: *web front-ends* (web2news), che solitamente sono messe a disposizione da EasyNews, Giganews, Astraweb, oltre a Google Group.

Il *newsgroup 2.0* è sicuramente rappresentato dal sito Reddit (https://www.reddit.com/). Il portale è la fusione delle parole inglesi *read* e *edit*, che riportano all'assonanza con *read it*. Oltre a dar spazio ai post degli utenti e alla loro interazione chiede anche di votare gli articoli inseriti. Il sito di social news condivide ogni giorno milioni di notizie e si descrive come un motore per creare community. Il suo "segreto" risiede nel voto e nel commento, che sono i due meccanismi che permettono che si creino community e readership riguardo a determinati argomenti.
Gli utenti iscritti (redditor), oltre a postare le loro storie sono i fautori della discesa o salita di un post nella homepage, tramite le funzioni di *upvote* o *downvote*.

4. Dark Web

Terminata questa carrellata tra Surface e Deep web, nella quale abbiamo visto che i due mondi molte volte si trovano a contatto, di tutt'altra pasta è fatto il Dark Web. La più importante caratterista di questa rete è sicuramente data dal suo aggettivo presente nella sua definizione, "intenzionalmente invisibili":

> insieme di dati e contenuti resi internazionalmente invisibili e non accessibili dai comuni browser, ma da software per la comunicazione anonima.

4.1. L'anonimato

Premettiamo che l'anonimato non è un crimine, ma un diritto come stabilito dall'articolo 1 del D.Lgs 196/03, Codice in materia di protezione dei dati personali: *"…Chiunque ha diritto alla protezione dei dati personali che lo riguardano…"*.
Anche la Carta dei diritti fondamentali dell'Unione Europea (2007/C 303/01) all'articolo 8 interviene prevedendo che "... *ogni persona ha diritto alla protezione dei dati di carattere personale che la riguardano e che tali dati devono essere trattati secondo il principio di lealtà per finalità determinate e in base al consenso della persona …"*
Ad onore del vero: se un individuo non ha nulla da nascondere allora, non ha nulla da temere. Ma è anche vero che tutti noi possiamo avere la necessità di tenere riservato e lontano dalla curiosità degli altri, un'informazione senza che questa sia illegale, vergognosa o imbarazzante. Quindi: se si tratta di una debolezza, o di un'opinione contraria a un regime non democratico, la riservatezza rappresenta un diritto di ogni essere umano.
Sicuramente, però, non è facile tracciare una linea netta di demarcazione tra una parte *buona* ed una *cattiva* del Dark Web: la maggior parte dei servizi si presta tanto a un uso lecito tanto quanto ad un uso illecito.

4.2. I Proxy

Se si parla di software per comunicazione anonima è d'obbligo introdurre il concetto di proxy. Ad ogni utente che si collega in rete viene assegnato uno specifico indirizzo IP, mediante il quale saranno instaurate le richieste di navigazione per richiedere il sito d'interesse che vuole visualizzare. Il proxy è un intermediario tra il computer di origine e quello di destinazione, un server che si interpone tra un computer e la sua destinazione, facendo da tramite tra i due sistemi e inoltrando le richieste e le risposte dall'uno all'altro. Esistono varie tipologie di Proxy:

- Trasparent Proxy;
- Anonymous Proxy;
- Highly Anonymous Proxy;
- Proxy CGI;
- Tor Onion Proxy.

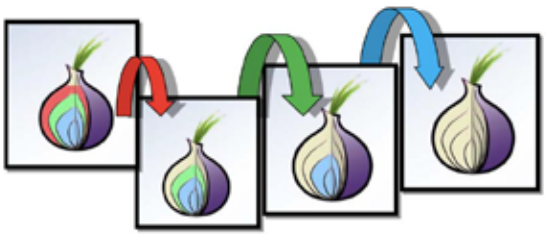
I **Trasparent Proxy** sono tipici degli ISP (Internet Service Provider), intercettano le normali comunicazioni senza particolari configurazioni e vengono usati anche per carenze di indirizzi IP v4.
Gli **Anonymous Proxy** non trasmettono l'IP del richiedente, ma modificano o aggiungono alcuni *header*. Da una analisi delle comunicazioni di rete è possibile riconoscere questi tipi di proxy quindi sono facilmente contrastabili; per evitare questo, molti proxy trasmetto un IP casuale senza modificare l'*header*.
Gli **Highly Anonymous Proxy** non trasmetto nessun IP e non cambiano gli *header* ed è per questo che sono molto difficili da contrastare ed individuare nell'analisi del traffico, mentre ai **Proxy CGI** è possibile accedervi attraverso un'interfaccia web e consentono di visitare altri siti in modo anonimo.
In ultimo abbiamo i **Tor Onion Proxy** che rappresentano una vera e propria catena di proxy, salvaguardando l'anonimato del cibernauta.

4.3. Tor

Tor e l'acronimo di "The Onion Router", sistema di comunicazione anonima per Internet basato sul protocollo *onion routing*. Il protocollo è stato sviluppato negli anni Novanta dal matematico Paul Syverson e da Michel Reed per la *US Naval Research Laboratory*, con lo scopo di proteggere le comunicazioni dei servizi segreti americani. Dopo vari enti che hanno ereditato e continuato a sviluppare il protocollo, si è giunti nel 2006, anno in cui viene fondata "*The Tor Project*", un'associazione senza scopo di lucro responsabile dello sviluppo di Tor.
Tor è un insieme di server che ospitano gli *hidden services*: servizi raggiungibili dagli utenti non localizzabili nella rete. Un servizio nascoso può essere ospitato da qualsiasi nodo Tor, sia esso un router o un client, ma per poter accedere a tale servizio occorre l'uso di Tor. A questi servizi si accede utilizzando uno pseudo-dominio di primo livello identificato dal suffisso .**onion**.

4.4. La rete Tor

Andiamo per ordine: l'idea della rete Tor si basa sulla costruzione di un percorso tortuoso e difficile da seguire e da decodificare. Per creare un percorso "tortuoso", a prova di "inseguimento", il software crea incrementalmente un circuito di connessioni cifrate attraverso i *relay* della rete. Il circuito viene esteso un salto alla volta e ogni *relay*, lungo il percorso, conosce solo quale *relay* gli ha dato le informazioni e verso quale *relay* inoltrarle. Nessun *relay* conosce il percorso completo che il pacchetto ha intrapreso.
Per creare l'affidabilità dell'intellegibilità dalle informazioni che passano nella rete, il software negozia un nuovo insieme di chiavi crittografiche per ogni salto lungo il circuito, così da assicurarsi che ciascun nodo non possa tracciare queste connessioni durante il passaggio dei messaggi.
Una volta che un circuito è stato stabilito, si possono scambiare diversi tipi di dati e usare molti tipi di applicazioni attraverso una rete Tor.
Il c.d. circuito Tor è il cammino attraverso la rete, scelto dai client denominati Onion Proxy (OP). Ogni nodo del circuito, definito anche come Onion Router, conosce soltanto il suo predecessore ed il suo successore.
Un circuito è generalmente un percorso di tre OR. Il primo OR su un circuito è chiamato *entrance router*, il secondo OR è detto *mix router* e l'ultimo hop è l'*exit router*.



Gli OR (Onion Router) hanno la capacità di mantenere una connessione TLS con agli altri OR. Il protocollo TLS viene utilizzato per creare chiavi temporanee (link key) per cifrare la comunicazione tra gli OR della rete. Le chiavi di breve durata vengono cambiate periodicamente ed indipendentemente per limitare l'impatto di chiavi compromesse.
Le chiavi pubblica/privata associate ad ogni Onion Router si distinguono in:

- identity key a lungo termine, utilizzate per firmare i certificati TLS, l'*onion router descriptor* e, in caso di *directory server*, per firmare le *directory*.
- onion key a breve termine, utilizzate per decifrare i messaggi inviati dall'utente per costruire un circuito e negoziare la chiave effimera.

4.5. Hidden Service

Usando la rete Tor, è possibile ospitare server con dei servizi denominati *hidden service*, in modo che la loro localizzazione nella rete sia sconosciuta. Il server verrà configurato in modo tale da non fornire ad eventuali utenti nessuna informazione relativamente

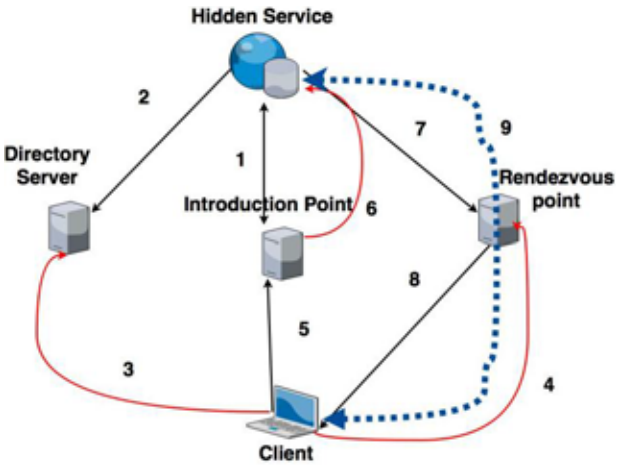
alla macchina su cui viene avviato. A tal fine occorre che il server non possa essere raggiunto direttamente fuori dal circuito Tor e che questo sia installato su una macchina virtuale per non fornire eventuali caratteristiche legate a un Sistema Operativo di uso comune dall'utente.
Un servizio nascosto può essere ospitato da qualsiasi nodo della rete Tor, non importa che esso sia un *relay* o solo un client; per accedere ad un servizio nascosto, però, è necessario l'uso di Tor da parte del client.
È possibile offrire un servizio (ad esempio un sito web) in modo totalmente anonimo, come *hidden service Tor:* ipotizziamo di essere un client all'interno della rete ed abbiamo avviato Tor sulla nostra macchina. Questo, al suo avvio, genera una coppia di chiavi crittografiche dedicate al servizio nascosto che vogliamo avviare.
Tor andrà poi a generare un documento chiamato *hostname*, contenente la chiave pubblica del server e l'indirizzo pubblico dello stesso. L'indirizzo apparirà come 1Y5Z.onion e sarà utilizzato per accedere ai contenuti che saranno pubblicati. Successivamente un *hidden service* deve rendere nota la sua esistenza nella rete Tor. Per questo il servizio sceglie alcuni *relay* a caso, stabilisce dei circuiti verso di essi e chiede loro di fungere da *introduction point* comunicandogli la sua chiave pubblica. Creando questa connessione è difficile stabilire se gli *introduction point* sono associati a qualche *hidden services*. Gli *introduction point* non conoscono l'IP del *hidden service* ma solo la loro chiave pubblica.

A questo punto, affinché i *client* possano trovare il nuovo servizio nascosto, è necessario comunicare la sua esistenza all'interno della rete Tor attraverso la pubblicazione di uno specifico *descriptor*, ovvero un pacchetto contenente la sua chiave pubblica ed un sommario degli *introduction point*, firmato con la sua chiave privata.
Il *descriptor* sarà inviato ad un gruppo di *directory server*, usando sempre il circuito Tor. Successivamente, il *descriptor* verrà trovato dai *client* che vogliono accedere al nuovo servizio mediante l'indirizzo 1Y5Z.onion.
Gli indirzzi .onion, sono composti da un nome di 16 caratteri derivato in modo unico dalla chiave pubblica dell'*hidden service*. Dopo questo passo, l'*hidden service* è attivo.

4.6. Navigare su un sito onion

Quando un client desidera contattare un *hidden service*, deve conoscere prima il suo indirizzo onion. Dopodiché il *client* può iniziare a stabilire la connessione scaricandone il *descriptor* dal *directory server*.
Supponiamo ad esempio di voler navigare su http://deepdot35wvmeyd5.onion (sito del dark web sui *market place*), questo avrà il suo *descriptor* pubblicato (linea 2) sul directory server e avrà scelto il relativo introduction point (linea 1). Il client leggerà il descriptor (linea 3), che lo aiuterà a conoscere anche il gruppo di *introduction point* e la corretta chiave pubblica.

Il *client* crea un circuito verso un altro *relay* scelto a caso e gli chiede di fungere da *rendezvous point* (linea 4), procedura che inoltra semplicemente i messaggi dal *client* al *service* e viceversa) comunicandogli un *one-time secret*. In questo modo il *client* apre un canale comunicativo *cifrato* per dialogare con il *service*.
Il *service* comunicherà il *descriptor* ed il *client* costruirà un *introduce message* (messaggio cifrato con la chiave pubblica dell'*hidden service*) contenente l'indirizzo del *rendezvous point* ed il one-time secret.
Il *client* invia questo messaggio a uno degli *introduction point* (linea 5/6), chiedendo che venga consegnato all'*hidden service*. Ogni scambio di messaggi avviene sempre tramite un circuito Tor.
L'*hidden service* decifra l'*introduce message* del *client* e scopre l'indirizzo del *rendezvous point* (linea 7/8) ed il *one-time secret* contenuto.
Il *service* crea un circuito verso il *rendezvous point* e gli invia il *one-time secret* in un *rendezvous message*. Il *rendezvous point* notifica al client che la connessione è stata stabilita con successo (linea 9). Il sito sarà fruito dal *client*.



4.7. Basta un click

Per entrare nel Dark Web può bastare un click, o meglio due. Il primo per scaricare dal sito https://www.torproject.org l'applicativo TorBrowser e il successivo per avviarlo. TorBrowser è un applicativo stand-alone che con un semplice click instaura il circuito Tor sul nostro PC, garantendo la massima sicurezza e anonimato nella navigazione e l'accesso al Dark Web. Una volta scaricato il file, ed eseguiti i passi di installazione, si aprirà una versione modificata di FireFox adattata a questo genere di navigazione.

Utilizzando questa applicazione crediamo di essere nell'anonimato più completo, ma se proviamo a massimizzare le dimensioni della finestra TroBrowser è possibile perdere parte dell'anonimato, dato che i siti web visitati possono imporre le dimensioni dello schermo.
In determinate condizioni, anche quando siamo collegati alla rete anonima, il Sistema Operativo del nostro computer continua ad utilizzare i server DNS di default invece dei server DNS anonimi assegnati dalla rete anonima. Si possono avere così delle perdite di DNS o IP che prendono il nome di "DNS leaks" o "IP leaks", creando una grave minaccia all'anonimato. ©