

Chapitre 1 Des darknets au Darknet

Qu'est-ce que le Darknet ? La confusion règne dans un espace médiatique où l'on mêle allègrement des réalités aussi différentes que le *deep web*, les échanges pair-à-pair qui effraient tant les maisons de disques, ou encore ce sombre réseau *Tor*, si sulfureux qu'il serait à l'origine de toutes les dérives de l'Internet contemporain.

« [Le darknet est] un ensemble de réseaux et de technologies utilisés pour partager du contenu numérique. Le darknet n'est pas un réseau physiquement distinct, mais bien des protocoles de transmission qui fonctionnent au sein des réseaux existants. » [Biddle *et al.*, 2003].

Il suppose [Mansfield-Devine, 2009] :

- L'usage de l'infrastructure internet.
- L'existence d'un protocole spécifique qui permet la constitution d'un sous-réseau.
- Une architecture décentralisée de type pair-à-pair.

Techniquement, selon ces auteurs, le Darknet est donc un sous-réseau pair-à-pair utilisant des protocoles spécifiques. En ce sens, il n'existe pas un Darknet, mais bien un ensemble de darknets, ou de sous-réseaux, étanches les uns aux autres. Un utilisateur de Freenet ne pourra pas se connecter directement au réseau *Tor*, de même qu'on ne peut pas utiliser un client BitTorrent pour se connecter à Shareaza.

La relation entre sous-réseaux, protocoles pair-à-pair et Darknet est certaine, mais ils ne sauraient se résumer les uns aux autres. Interpréter le Darknet comme simple réalité technique ne permet pas d'en saisir l'essence. Il est avant tout un fait social. Certains le réduisent à « la partie du *deep web* où se déroulent les opérations illégales¹ », mais, pour reprendre Philippe Davadie, le Darknet est un réseau neutre qui n'est pas substantiellement malfaisant [Davadie, 2015]. Il est vaste et divers et on ne peut le circonscrire à un espace plus ou moins caché et dédié aux opérations illicites. Il est bien plus que cela.

Socialement, le Darknet s'incarne dans la quête de l'anonymat et de la confidentialité. C'est elle qui marque sa spécificité, c'est l'usage social d'instruments techniques qui fait particularité.

1. [McCormick, 2013].

On définira donc un darknet comme *un sous-réseau d'internet utilisant des protocoles spécifiques intégrant nativement des fonctions d'anonymisation*. Le Darknet est alors l'écosystème formé par l'ensemble des darknets et des outils associés de préservation de la confidentialité. De même qu'on utilise Internet (avec une majuscule) pour représenter l'ensemble des ressources disponibles sur internet (le réseau), nous utiliserons Darknet pour représenter l'ensemble des darknets et l'écosystème associé. Par opposition, on parlera de *Cleartnet* pour qualifier l'Internet classique, ouvert. Certains darknets sont plus particulièrement orientés vers le partage de fichiers, le chat ou la messagerie électronique, mais les plus connus visent d'abord le web, ce sont eux qui forment le fameux *Darkweb*.

Dans ce contexte, Tor ou I2P sont des darknets. Le système d'échange d'e-mails chiffrés PGP fait partie de l'écosystème Darknet, tout comme la messagerie instantanée Cryptocat. En revanche, BitTorrent, Gnutella ou *comment-pirater-son-voisin.com* sont des outils classiques de partage de ressources et d'informations, ce ne sont pas des darknets. Cette approche peut être vue comme restrictive, mais elle veut marquer le haut degré de spécificité des écosystèmes anonymes.

Anonymat, confidentialité et vie privée

L'anonymat consiste à dissimuler son identité. Les actions peuvent être connues, mais on ne souhaite pas qu'il soit possible de les relier à une identité. C'est la situation classique des lanceurs d'alerte, ils ne craignent pas les rétorsions s'ils peuvent rester cachés.

La confidentialité consiste à interdire aux tiers l'accès à l'information. Elle est essentiellement basée sur les procédures de chiffrement. On peut avoir des échanges confidentiels sans qu'ils soient anonymes, c'est le cas des communications sensibles en entreprise par exemple.

La préservation de la vie privée signifie simplement que l'on ne souhaite pas que certaines de ses activités soient observées. Les hommes ont toujours cherché à préserver des espaces d'intimité, cela n'a rien à voir avec la légalité de leurs actions. Pour Edward Snowden, « Répondre "*je n'ai rien à cacher*" en matière de vie privée revient à affirmer que l'on se fiche de la liberté d'expression parce que l'on n'a rien à dire. » Sans aller aussi loin, la perte progressive de la maîtrise de la vie privée sur Internet interroge. Lors d'une expérience, une sociologue américaine n'a pas eu d'autres choix qu'utiliser Tor pour cacher sa grossesse aux opérateurs publicitaires^a. Démonstration claire, s'il en était besoin, du fait qu'il est difficile de choisir ce qu'on partage.

a. time.com/83200/privacy-internet-big-data-opt-out, vu le 4/5/2018.

Nous allons d'abord revenir sur la notion de deep web afin de lever ce malentendu si répandu qui le confond avec le Darknet ; confusion qui atteint la caricature avec ces innombrables représentations graphiques qui contribuent à la diabolisation

en montrant un Darknet infiniment plus vaste que l'Internet ouvert (figure 1.1 page ci-contre).

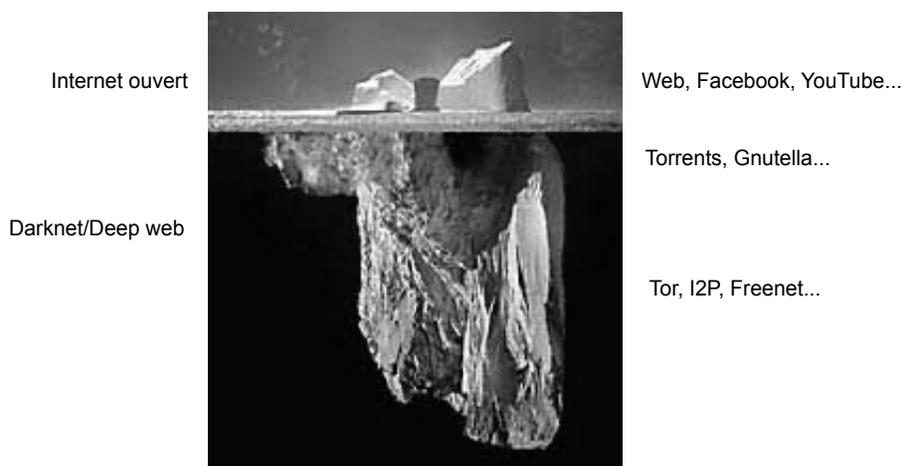


Figure 1.1 – Deep web et Darknet, la confusion

Nous verrons ensuite ce que l'on entend par sous-réseau et présenterons quelques-uns des grands systèmes pair-à-pair, afin d'introduire à cette technologie si importante pour le Darknet.

1. Web et deep web

Quand on parle d'Internet, c'est généralement d'abord au World Wide Web que l'on fait référence. À cette immense toile où l'on a pris l'habitude de « surfer » et qui forme le *Web*, où la majuscule symbolise l'unité de cet ensemble que forment les sites web.

Le Web a été conçu comme un outil d'accès à l'information : « Cette proposition concerne la gestion de l'information [...] au CERN. Elle [est] basée sur un système hypertexte distribué. » [Berners-Lee, 1989]. L'hypertexte y représente « des informations lisibles par l'homme et reliées entre elles de manière non contrainte. »

Chacun connaît le succès de la proposition de Tim Berners-Lee (figure 1.2 page suivante)². Aujourd'hui, des millions de sites web sont en ligne, reliés entre eux par ces liens hypertextes qui tissent la trame même du Web.

Pour accéder à cette information, des outils de recherche spécifiques ont dû être élaborés. Le premier, *Archie* est sorti en 1990, mais c'est à partir de 1994 avec *Lycos*, puis 1995 avec *Excite*, *Yahoo!*, *AltaVista*, suivis d'innombrables autres, toujours actifs

2. <http://line-mode.cern.ch/www/hypertext/WWW/TheProject.html>



Figure 1.2 – Le premier site web

ou déjà oubliés, que l'on est entré dans la phase moderne de la recherche, dont le point culminant reste l'arrivée de *Google* en 1998.

Les moteurs de recherche sont maintenant notre point d'entrée sur le Web. Google est très certainement la page d'accueil la plus courante en Occident. Pourtant, ces moteurs restent limités en ce sens qu'ils n'indexent que ce qui est directement accessible. Dès 1998, on a pointé le fait que les moteurs ne peuvent accéder qu'au « Web indexable » [Lawrence et Lee Giles, 1998].

Bing, Yahoo! ou Google indexent ce qu'ils peuvent atteindre en suivant les liens hypertextes. Leur échappent notamment :

- Les pages générées dynamiquement par l'interrogation d'une base de données à travers un formulaire.
- Les sites auxquels on ne peut accéder qu'après s'être identifié.
- les sites dont l'accès est limité (CAPTCHA, etc.).
- Les pages interdites aux robots de recherche³.
- Les pages qui n'ont pas de liens hypertextes avec d'autres.

C'est cette partie, que ne peuvent indexer les moteurs de recherche, qui constitue le *deep web* (ou *Hidden Web* ou encore *Invisible Web*), par opposition au « Web de surface » (*Surface Web*). Il est très difficile d'en évaluer précisément la taille. On l'a estimée au début des années 2000 a au moins 400 à 500 fois celle du Web de surface [Bergman, 2001]. En 2012, une étude a évoqué un rapport de 1 à 4000 ou 5000⁴. Ces écarts montrent bien les difficultés d'estimation du volume du deep web ; difficultés d'autant plus grandes que les modes d'accès évoluent et que sa croissance est exponentielle [Finklea, 2017]. Le point important pour nous est que sa spécificité

3. Il existe une norme qui permet entre autres de dire aux robots de ne pas indexer certaines pages.

4. Bright Planet, Deep Web: A Primer, <http://www.brightplanet.com/deep-web-university-2/deep-web-a-primer>, vu le 15/02/2018.

vient de son mode d'accès, non de son contenu qui, pour l'essentiel, est parfaitement légal.

L'accès au deep web ou plutôt la capacité à remonter ces informations au niveau du web de surface⁵ est devenu un enjeu majeur pour les grands moteurs de recherche.

Les grands acteurs travaillent sur le sujet depuis longtemps. Des chercheurs de Cornell et de Google [Madhavan *et al.*, 2008] ont par exemple montré comment il pouvait être possible de gérer les pages accessibles à travers un formulaire. Leur recherche vise à limiter l'exploration de l'espace de recherche aux données pertinentes. Il y a 5 critères différents sur le site *cars.com* qui mènent en théorie à 240 millions de combinaisons (donc d'URL) différentes, soit infiniment plus qu'il n'y a d'offres. Ces travaux montrent clairement l'intérêt que peuvent trouver les moteurs à « surfacer » le deep web.

Au-delà de la dimension commerciale, les services de sécurité s'intéressent également au deep web où peuvent se trouver des informations criminelles. La Darpa a développé un moteur dédié (Memex) qui a notamment permis de démanteler un réseau de trafiquants d'êtres humains [Berthier, 2015].

Le deep web n'est pas le Darknet [Pederson, 2013], il est l'ensemble des informations auxquelles on ne peut accéder directement par indexation. Si vous consultez certaines séries statistiques sur le site de l'INSEE, vous êtes dans le deep web, mais il ne viendrait à l'esprit de personne de considérer que vous êtes sur le Darknet. Le Darknet peut pour partie être considéré comme appartenant au deep web, en ce sens qu'il n'est pas indexé par les grands moteurs de recherche, qu'il leur est même globalement inaccessible, mais il n'en est qu'une infime fraction. Les représentations habituelles illustrent bien les fantasmes qu'il engendre. Les volumes en jeu dans les bases de données du deep web sont gigantesques et sans rapport aucun avec ce que peuvent générer les activités anonymes, légales ou non, même florissantes (figure 1.3 page suivante).

Deep web et Darknet sont fondamentalement différents. Les données auxquelles on peut accéder à travers les protocoles standards, mais qui sont inaccessibles aux robots du fait de diverses contraintes opérationnelles, ne relèvent pas de la même logique que celles qui reposent volontairement sur l'usage de protocoles spécifiques, anonymisants, non connus des robots habituels.

2. Internet et réseaux pair-à-pair

Les darknets sont des sous-réseaux d'internet. En informatique, un réseau est un ensemble d'équipements interconnectés. Ces connexions peuvent être aussi bien physiques (les câbles réseaux) que radio (le Wi-Fi, le Bluetooth, etc.), voire utiliser la

5. Ce que les anglo-saxons appellent « surfacing the deep web ».

PageRank

Recenser et indexer n'est pas suffisant pour faire un bon moteur de recherche et d'énormes efforts sont consacrés au développement d'algorithmes de classement efficaces, c'est-à-dire aptes à mettre en évidence les sites correspondant au mieux à la recherche en cours. Google doit son succès initial au fameux *PageRank*.

Sergei Brin et Larry Page ont proposé leur célèbre algorithme lors d'une conférence en Australie en 1998 [Brin et Page, 1998]. L'objectif déclaré était « d'améliorer la qualité des recherches ». Pour eux, l'indexation seule n'est pas suffisante, les résultats pertinents étant souvent noyés sous une avalanche de résultats sans valeur (*Junk results*). Pour preuve, en 1997, un seul des quatre grands moteurs commerciaux retournait son adresse dans les 10 premières réponses à une recherche sur leur propre nom !

Pour améliorer le classement des résultats, S. Brin et L. Page ont décidé « d'utiliser la structure des liens au sein du Web », le principe étant que plus un site est cité, plus il doit être pertinent. Ils se sont inspirés ici de la pratique habituelle des milieux académiques, où le nombre de citations d'un article est considéré comme un bon indicateur de sa qualité.

Le *PageRank* d'une page est fonction du nombre et du *PageRank* des pages qui y font référence. Il correspond à la probabilité qu'un *surfeur aléatoire* suivant au hasard les liens qui lui sont proposés soit sur la page en question à l'instant t .

L'exemple ci-dessous (table 1.1)^a compare les résultats obtenus pour le mot clé *University* sur AltaVista et sur l'une des premières versions de travail de Google. Dans un cas, les réponses apparaissent plus ou moins au hasard alors que dans l'autre la cohérence est évidente.

Le succès de Google a d'abord été celui de *PageRank*. C'est bien la qualité des résultats obtenus qui a fait la différence avec les autres moteurs de recherche.

Google	Altavista
1- Stanford University Homepage	1- Optical Physics at the University of Oregon
2- University of Illinois at Urbana-Champaign	2- Carnegie Mellon University - Campus Networking
3- Indiana University	3- Wesleyan University Computer Science Group Home Page
4- University of California, Irvine	4- Keio University Shonan Fujisawa Campus (SFC)
5- University of Minnesota	5- School of Chemistry, University of Sydney
6- Iowa State University Homepage	6- Mankato State University
7- The University of Michigan	7- St. Ambrose University
8- Mississippi State University	8- University of Washington ECSEL Projects

TABLE 1.1 – L'efficacité de *PageRank*

a. D'après [Page et al., 1998].

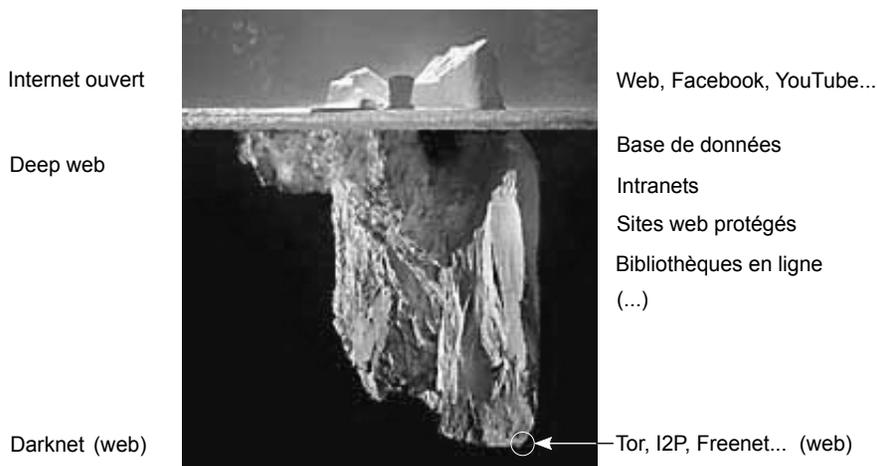


Figure 1.3 – Web, deep web et Darknet

lumière (le Li-Fi). Ces équipements sont reliés entre eux selon certaines topologies (figure 1.4 page ci-contre). Les réseaux physiques sont la base matérielle des réseaux logiques. Les logiciels réseau permettent de regrouper telle ou telle partie d'un réseau physique en un sous-ensemble logique : un sous-réseau.

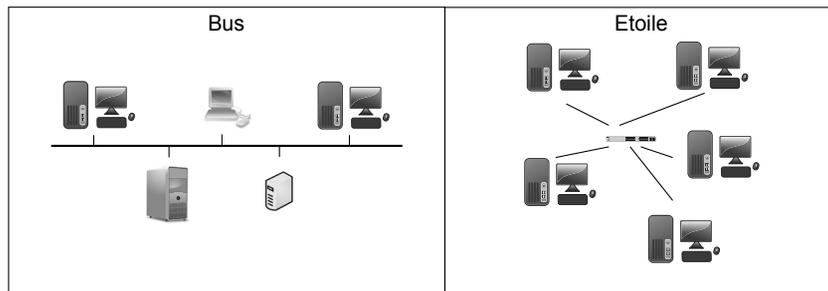


Figure 1.4 – Deux exemples de topologies physiques classiques

Internet est un réseau particulier. On le décrit habituellement comme étant le « réseau des réseaux », c'est-à-dire qu'il connecte entre eux différents réseaux de topologies diverses. La structure générale d'internet est très complexe et n'est que partiellement comprise. Au plus haut niveau, il est un rassemblement de « réseaux autonomes » (*Autonomous System ou AS*) qui sont de grandes entités comme des universités, des structures publiques ou bien sûr des fournisseurs d'accès à internet (FAI). Début 2018, il y avait environ 60.000 AS. Chacun de ces réseaux autonomes gère des milliers de *nœuds* qui sont autant de composants d'internet. La structure générale n'est pas homogène et certains nœuds concentrent la majorité des connexions. On a pu ainsi qualifier la topologie d'internet de « nœud papillon » [Broder *et al.*, 2000] où

un centre est nécessaire pour assurer les connexions globales. Autre illustration, on l'a qualifié de *réseau méduse* [Siganos *et al.*, 2006]. Autour d'un cœur, différentes couches se succèdent, marquées par une connectivité décroissante. Dans cette représentation (figure 1.5), on voit ce que les auteurs ont considéré comme les différentes couches d'internet. Le cœur est le lieu de plus forte densité des relations, la longueur des « tentacules » représente la concentration des liens directs entre membres (les voisinages de degré 1). L'immense majorité des nœuds (80 à 90 %) se trouve sur les trois premières couches.

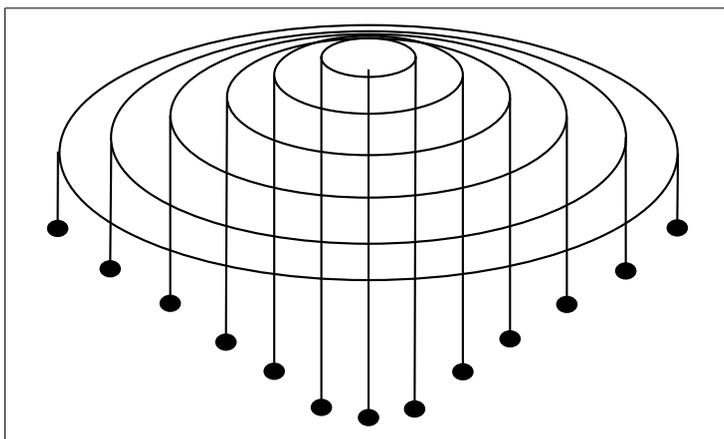


Figure 1.5 – Topologie en méduse [Siganos *et al.*, 2006]

La topologie générale d'internet est atypique et évolutive, elle ne peut se résumer aux grandes classes habituelles, mais le point important pour nous est que le réseau forme un graphe connexe, c'est-à-dire qu'il est théoriquement toujours possible de construire un chemin entre deux équipements connectés quelconques.

Pour communiquer, les différents matériels reliés au réseau utilisent des *protocoles*, des sortes de langues communes qui permettent aux équipements de se comprendre. Un protocole définit notamment le format des données et les procédures de mise en relation des éléments qui souhaitent échanger. Le plus courant est le protocole IP qui est au cœur d'internet. On lui associe généralement TCP qui le complète en garantissant notamment la correction des échanges. On parle alors de TCP/IP.

Quand on va sur Internet, outre TCP/IP, mais à un niveau différent, on utilise des protocoles célèbres comme HTTP pour naviguer sur le Web, FTP pour échanger des fichiers ou encore SMTP pour envoyer des e-mails, mais aussi bien d'autres moins connus, mais tout aussi essentiels.

HTTP est un protocole dit *client-serveur*, c'est-à-dire qu'il relie un client (vous) et un serveur (celui qui héberge votre quotidien préféré). Dans une architecture client-serveur, le client interroge un ordinateur central qui renvoie l'information demandée.