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 aussi, et peut-être d'abord, un fait social. Certains le réduisent à « la partie du deep web où se déroulent les opérations illégales [1], » mais le Darknet est vaste et divers, 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].

Dans le cadre de cet ouvrage, on définira un darknet comme *un sous-réseau d'internet utilisant des protocoles spécifiques et 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 *Clearnet* pour qualifier l'Internet classique, ouvert.

Dans ce contexte, Tor ou I2P sont des darknets. Le système d'échange de mails chiffrés GPG 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 l'accès à l'information aux tiers, elle est essentiellement basée sur les procédures de chiffrage. 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 sa vie privée sur Internet interroge. Lors d'une expérience récente, une sociologue américaine n'a pas eu d'autres choix qu'utiliser Tor pour cacher sa grossesse aux opérateurs publicitaires[a]. Il est important que chacun puisse avoir le choix de ce qu'il partage.

a. time.com/83200/privacy-internet-big-data-opt-out, vu le 15/12/2015.

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

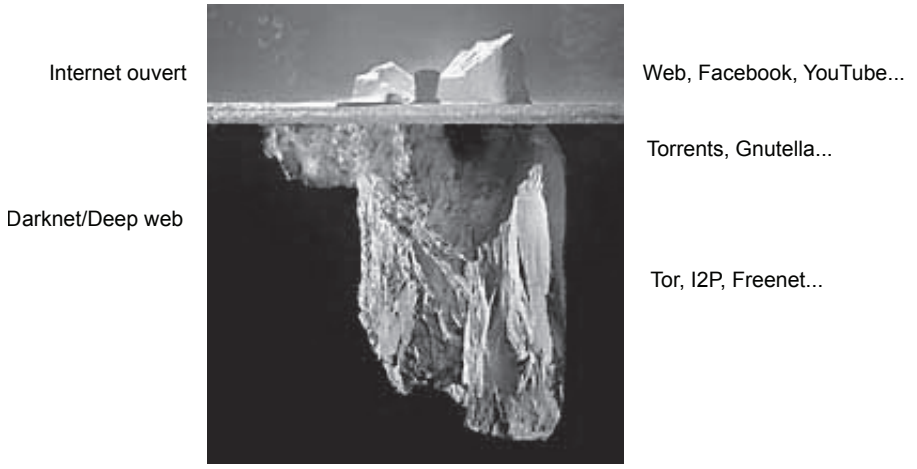

Figure 1.1 – Deep web et Darknet, la confusion

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 traite des problèmes de perte d'information dans un système évolutif complexe et propose une solution 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. Aujourd'hui, des millions de sites web sont en ligne, reliés entre eux par ces liens hypertextes qui tissent la trame même du Web.

Des outils de recherche spécifiques ont dû être élaborés pour le Web. 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 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, les chercheurs ont pointé le fait que les moteurs ne peuvent accéder qu'au « Web indexable » (*indexable Web* [2]).

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 pages interdites aux robots de recherche [3].
— 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, mais on l'estimait au début des années 2000 a au moins 400 à 500 fois celle du Web de surface [Bergman, 2001]. En ce sens l'image habituelle qui représente Web et deep web comme un iceberg n'est pas pertinente, la face cachée y est 50 fois plus importante que dans les montagnes de glace !

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 (on parle là de millier de téraoctets) et sans rapport aucun avec ce que peuvent générer les activités anonymes, légales ou non, même florissantes (figure 1.2 page 16).

Au-delà de la dimension technique, il y a une différence qualitative fondamentale entre le deep web et le Darknet. Les données auxquelles on peut accéder à travers les protocoles standards, mais qui sont inaccessibles aux robots du fait de diverses contraintes techniques, ne relèvent pas de la même logique que celles qui reposent volontairement sur l'usage de protocoles spécifiques, anonymisant, non connus des robots habituels.

2. [Lawrence et Lee Giles, 1998].
3. Il existe une norme qui permet entre autres de dire aux robots de ne pas indexer certaines pages.

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 son 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].

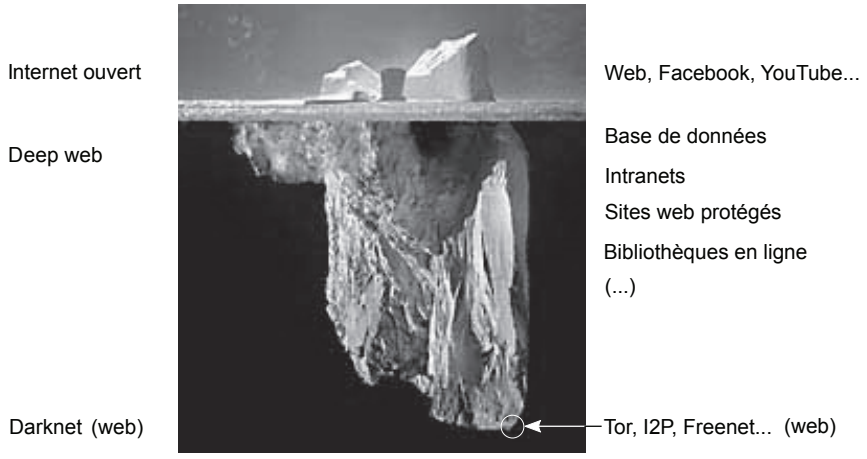

Figure 1.2 – Web, deep web et Darknet

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 lumière (le *Li-Fi*). Ces équipements sont reliés entre eux selon certaines topologies (figure 1.3). 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.

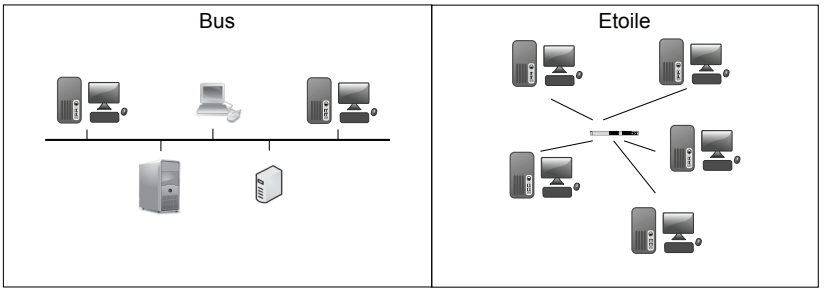

Figure 1.3 – Deux exemples de topologies physiques classiques

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 elle 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). En avril 2015, il y avait environ 50.000 AS formant plus de 550.000 routes. 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 1.4), 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.

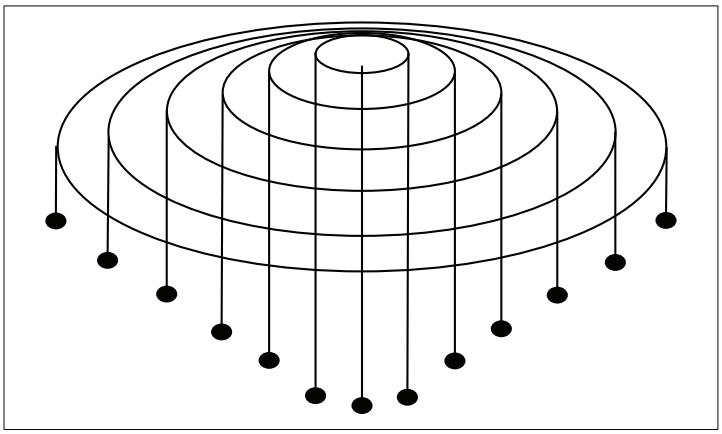

Figure 1.4 – Topologie en méduse

La topologie générale d'internet est atypique et évolutive, elle ne peut se résumer aux grandes classes habituelles, mais le point important pour nous et 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 méthodes de mise en relation et d'association 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 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. Des centaines, voire des milliers de clients peuvent être connectés simultanément au même serveur ou au même groupe de serveurs.

Internet supporte également des protocoles dits « pair-à-pair » (le P2P) où chaque client peut aussi faire office de serveur. Au lieu de l'architecture centralisée où tout est relié à un point unique, qui caractérise le modèle client-serveur, on a une architecture décentralisée où chacun est relié aux autres et où chacun contribue à la diffusion des flux de données. Cette technologie n'est pas nouvelle, puisqu'on la retrouve dans les premières connexions ARPANET.

Napster, sortie en 1999, est considéré comme le premier grand système pair-à-pair. Dédié à l'échange de fichiers musicaux (il n'autorisait que le partage des fichiers audio *mp3*), il appartenait à la catégorie des systèmes pair-à-pair *centralisés*. Le principe est le suivant (figure 1.5 page suivante) :

— Le réseau est formé par l'ensemble des utilisateurs (les clients ou les nœuds) et par un serveur central (en fait un ensemble de *serveurs index*) qui contient la liste des fichiers hébergés dans le réseau et l'adresse des clients correspondants.
— L'utilisateur qui recherche un fichier donné interroge le serveur central.
— Le serveur lui renvoie la liste des nœuds qui héberge ce fichier.
— L'utilisateur se connecte à l'un des nœuds disponibles et y télécharge directement le fichier.

Dans ce système, on a à la fois une architecture centralisée —tout le monde est relié à un serveur central— et une architecture pair-à-pair où chacun peut jouer le rôle de serveur.

Cette architecture centralisée a permis aux autorités de fermer rapidement Napster. Il a suffi de stopper les serveurs index pour bloquer l'ensemble du réseau. Les réseaux décentralisés non structurés, comme *Gnutella* sur lequel sont basés des clients aussi fameux que *Limewire* ou *Shareaza*, se sont développés en réaction à cette faiblesse.

Gnutella[4] a beaucoup évolué au cours de son histoire. La version initiale est sortie en 2000 et fonctionne selon le principe suivant (figure 1.6 page ci-contre) :

— Quand on lance une recherche, le logiciel se connecte aux clients voisins.
— Les voisins relaient la demande à leurs propres voisins (processus dit d'*inondation*) .
— De proche en proche, on identifie un nœud qui héberge le fichier recherché.

4. Contrairement aux apparences, Gnutella n'appartient pas au projet GNU ; ce dernier lui a d'ailleurs demandé de changer de nom.