Méthodologie de caractérisation socio-organisationnelle des adresses IPs appliquée à la sécurité

Camille Moriot

N° d'ordre NNT : $2024ISAL0077$

THESE de DOCTORAT DE L'INSA LYON,
membre de l'Université de Lyon

Ecole Doctorale N° 512
Informatique et Mathématiques

Spécialité/discipline de doctorat : Informatique

Soutenue publiquement le 24/09/2024, par :

Camille Moriot

Méthodologie de caractérisation socio-organisationnelle des adresses IPs appliquée à la sécurité

Devant le jury composé de :

LAURENT	Maryline	Professeure des Universités	Télécom SudParis	Rapporteure
CHRISMENT	Isabelle	Professeure des Universités	TELECOM Nancy	Rapporteure
OWEZARSKI	Philippe	Directeur de Recherche	LAAS-CNRS	Examinateur
BOUABDALLAH	Abdelmadjid	Professeur des Universités	UTC Compiègne	Examinateur
BOSSERT	Georges	Docteur	Sekoia	Invité
LESUEUR	François	Docteur	Worteks	Co-encadrant de Thèse
STOULS	Nicolas	Maître de Conférences	INSA-LYON	Co-encadrant de Thèse
VALOIS	Fabrice	Professeur des Universités	INSA-LYON	Directeur de thèse

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Département de la Formation par la Recherche et des Études Doctorales (**FEDORA**)

Bâtiment Charlotte Perriand 3e étage
17, av. des Arts
69621 Villeurbanne Cédex
fedora@insa-lyon.fr

Référence : TH1135_MORIOT Camille

L'INSA Lyon a mis en place une procédure de contrôle systématique via un outil de détection de similitudes (logiciel Compilatio). Après le dépôt du manuscrit de thèse, celui-ci est analysé par l'outil. Pour tout taux de similarité supérieur à 10%, le manuscrit est vérifié par l'équipe de FEDORA. Il s'agit notamment d'exclure les auto-citations, à condition qu'elles soient correctement référencées avec citation expresse dans le manuscrit.

Par ce document, il est attesté que ce manuscrit, dans la forme communiquée par la personne doctorante à l'INSA Lyon, satisfait aux exigences de l'Etablissement concernant le taux maximal de similitude admissible.

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
ED 206 CHIMIE	**CHIMIE DE LYON** **https ://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage **secretariat@edchimie-lyon.fr**	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne **directeur@edchimie-lyon.fr**
ED 341 E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http ://e2m2.universite-lyon.fr** Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 **secretariat.e2m2@univ-lyon1.fr**	**Mme Sandrine CHARLES** Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX **e2m2.codir@listes.univ-lyon1.fr**
ED 205 EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http ://ediss.universite-lyon.fr** Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 **secretariat.ediss@univ-lyon1.fr**	**Mme Sylvie RICARD-BLUM** Laboratoire ICBMS - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
ED 34 EDML	**MATÉRIAUX DE LYON** **http ://ed34.universite-lyon.fr** Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 **yann.de-ordenana@ec-lyon.fr**	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 **stephane.benayoun@ec-lyon.fr**
ED 160 EEA	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **https ://edeea.universite-lyon.fr** Sec. : Philomène TRECOURT Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 **secretariat.edeea@insa-lyon.fr**	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 **philippe.delachartre@insa-lyon.fr**
ED 512 INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http ://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 **infomaths@univ-lyon1.fr**	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 **direction.infomaths@listes.univ-lyon1.fr**
ED 162 MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http ://edmega.universite-lyon.fr** Sec. : Philomène TRECOURT Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon **mega@insa-lyon.fr**	**M. Etienne PARIZET** INSA Lyon Laboratoire LVA Bâtiment St. Exupéry 25 bis av. Jean Capelle 69621 Villeurbanne CEDEX **etienne.parizet@insa-lyon.fr**
ED 483 ScSo	**ScSo**[1] **https ://edsciencessociales.universite-lyon.fr** Sec. : Mélina FAVETON Tél : 04.78.69.77.79 **melina.faveton@univ-lyon2.fr**	**M. Bruno MILLY** (INSA : J.Y. TOUSSAINT) Univ. Lyon 2 Campus Berges du Rhône 18, quai Claude Bernard 69365 LYON CEDEX 07 Bureau BEL 319 **bruno.milly@univ-lyon2.fr**

1. ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Résumé

Internet est un système clé dans la société contemporaine. Il s'agit d'un système complexe réparti entre de nombreuses organisations ayant une variété de rôles et d'intérêts. Depuis leur création, les cyberattaques sont devenues des actifs précieux, car elles donnent aux rivaux des avantages, par exemple dans les domaines politique ou économique. Il est nécessaire d'analyser ces attaques, d'identifier leurs singularités et les mécanismes sur lesquels elles s'appuient afin de les contrer. Cela permettra d'établir des signatures plus précises et plus pertinentes et aidera la conception des contre-mesures. Un des aspects d'analyse des attaques sont les infrastructures utilisées par les attaquants pour générer les attaques. De nombreux outils aujourd'hui permettent de caractériser l'aspect technique des machines qui composent ces infrastructures. Mais comme les attaques ont lieu dans un environnement social, politique, économique et organisationnel, nous revendiquons qu'il est nécessaire d'évaluer ces machines d'un point de vue organisationnel.

Cette thèse propose une méthodologie originale de catégorisation des adresses IP, à l'aide de 6 étiquettes décrivant deux axes : un axe technologique et un axe organisationnel. Nous proposons également un outil d'investigation, IPSeen, qui implémente cette méthodologie, en affectant les étiquettes aux adresses IP. Il s'appuie sur différentes sources de données : Wikidata, RDAP, Onyphe, GeoIPLite. Deux versions d'IPSeen sont proposées et évaluées dans ce manuscrit. Ces deux versions se différencient par leur rapidité et leur niveau de précision. Enfin, nous appliquons notre méthodologie à un ensemble de données réelles de suivi d'infrastructure de type "command and control". L'analyse produite propose une description des infrastructures des organisations qui maintiennent les machines participant aux infrastructures d'attaques. Nous montrons que notre approche apporte un éclairage essentiel sur la compréhension des attaques, en complément des nombreuses caractérisations techniques par ailleurs disponibles.

Table des matières

Table des figures

Liste des tableaux

Glossaire

— **API** : Application Programming Interface.
— **AS** : Autonomous System.
— **ASN** : Autonomous System Number.
— **CDN** : Réseau de diffusion de contenu.
— **CSV** : comma separated values, format texte ouvert représentant des données tabulaires sous forme de valeurs séparées par des virgules.
— **FAI** : Fournisseur d'accès à Internet.
— **ICANN** : Internet Corporation for Assigned Names and Numbers, autorité de régulation d'Internet en charge de coordonner les différents identifiants d'Internet à l'échelle internationale.
— **IoT** : Internet des objets.
— **JSON** : JavaScript Object Notation, format léger d'échange de données.
— **NAT** : Network Address Translation, processus de correspondance d'adresse IP.
— **NLP** : Traitement automatique du langage naturel, domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle.
— **Qid** : identifiant unique qui désigne un résultat de Wikidata.
— **RF** : Random Forest.
— **RGPD** : Règlement général sur la protection des données.
— **RIR** : Registres Internet Régionaux.
— **SEC** : Securities and Exchange Commission des États-Unis
— **SPARQL** : SPARQL Protocol and RDF Query Language, langage de requêtes et un protocole permettant d'interroger,des données structurées en RDF.

Chapitre 1

Introduction

Depuis son apparition dans les années 60, l'infrastructure d'Internet n'a cessé de s'agrandir et de se complexifier. Il s'agit d'une infrastructure partagée, où de nombreux acteurs sont responsables d'un sous-ensemble de l'infrastructure globale.

Internet possède une infrastructure mondiale et robuste qui lui permet aujourd'hui d'être au cœur de notre société. L'infrastructure d'Internet, en sa définition technique, est une pile protocolaire complexe, basée sur une structure physique répartie à travers le monde. Depuis son apparition, l'infrastructure n'a cessé de grandir et d'accueillir de nouvelles machines, de nouveaux services et de nouvelles organisations. Si bien que l'on estime aujourd'hui à plus de 15 milliards le nombre de machines connectées à Internet. Ce nombre est en hausse constante et devrait doubler d'ici à 2030. L'infrastructure d'Internet est répartie entre de nombreux acteurs, chacun ayant des rôles différents. Pour appréhender cet écosystème complexe, il est essentiel de connaître ces divers acteurs et de saisir leurs rôles et leurs places. Cependant, comme de nombreuses technologies, le réseau Internet est aussi utilisé à des fins malveillantes. De nombreuses attaques ont vu le jour, et c'est dans un contexte de sécurité que nous souhaitons placer les travaux présentés dans cette thèse.

Dans ce chapitre d'introduction, nous discutons d'abord les raisons d'existence des attaques en étayant les motivations derrière celles-ci. Nous présentons également les différents types d'attaques, leurs fréquences ainsi qu'une introduction aux mécanismes de contre-mesures. Nous motivons par la suite la nécessité de caractériser les machines participantes aux infrastructures d'attaque. Enfin, nous présentons la problématique que nous avons décidé de traiter dans cette thèse.

1.1 Internet, un espace propice aux attaques

L'apparition des attaques informatiques n'est pas récente. Dès 1962, la première attaque informatique a été identifiée. Allan Scherr, étudiant au MIT, s'attaque pour la première fois à un système informatique en récupérant les identifiants de ses camarades pour obtenir plus de temps sur les ordinateurs, qui ont alors un accès limité en temps[1]. Les attaques ne sont donc pas un phénomène récent. Néanmoins, depuis, elles ont beaucoup évolué. Cela est dû en partie au constant jeu du chat et de la souris qui existe entre les attaquants et les personnes chargées de la protection des systèmes. Pour chaque faille ou mécanisme exploité par les attaquants, les défenseurs vont alors essayer de mettre en place le plus rapidement possible une solution afin de limiter les conséquences des attaques.

Les attaques informatiques sont une menace constante qui pèse sur de nombreux acteurs : entreprises, hôpitaux, instances gouvernementales, particuliers. Or, ces attaques ont des

conséquences avec divers degrés d'intensité. Tout d'abord, au niveau des entreprises, il s'agit principalement d'un coût financier. La paralysie de certaines machines peut entraîner une baisse de la productivité ou encore une baisse des ventes dans le cas, par exemple, d'un serveur web hébergeant un site de vente en ligne. Les attaques sur les services publics comme un hôpital peuvent avoir des conséquences graves, en empêchant par exemple le bon fonctionnement de ces infrastructures.

Ces conséquences sont en partie les sources de motivation des attaquants. On peut les classer en trois grandes catégories principales : financières, politiques et concurrences [2]. En ce qui concerne les motivations financières, les attaques peuvent agir de plusieurs manières : vols d'argent, redirection du trafic vers un autre site marchand, demande de rançon en retour de l'accès aux machines. On constate notamment une montée en puissance du nombre de rançongiciels qui s'attaquent aux entreprises. Sur le plan politique, il s'agit de s'attaquer à la liberté d'expression ou encore d'attaquer en forme de rétribution suite à une décision ou à une parole politique. C'est le cas par exemple des récentes attaques qui ont pu avoir lieu en Nouvelle-Calédonie [3], ou suite à l'affaire Floyd aux États-Unis[4]. La paralysie de l'infrastructure d'un opposant politique est aussi observée, lors d'un conflit entre une ou plusieurs nations, comme actuellement avec le conflit Ukraine/Russie [5] ou encore Palestine/Israël [6].

Il existe également des groupes d'attaquants qui lancent des attaques afin de dénoncer des décisions politiques [7]. En ce qui concerne la concurrence, il peut s'agir ici de vol d'informations confidentielles ou plus légèrement, il peut aussi y avoir des attaques dans le but de perturber les réseaux d'un joueur adverse, comme il a pu se passer à la Lyon Esport [8] ou en 2024 dans la ligue de jeu League of Legend en Corée [9].

1.2 Panorama des attaques

L'écosystème des attaques est très vaste et les attaques sont très différentes les unes des autres. Connaître et avoir une vue d'ensemble des différents types d'attaques permet de mieux comprendre ce qu'il est nécessaire d'analyser dans l'écosystème. S'il existe de nombreuses attaques, certaines sont plus récurrentes[10, 11].

1.2.1 Les différentes attaques

Premièrement, nous parcourons ici un ensemble des principales menaces qui sévissent sur Internet. Nous présentons ici les principales caractéristiques de fonctionnement de menaces courantes.

Les malwares et virus sont une autre catégorie d'attaques. Les malwares sont des logiciels créés dans le but de nuire, de perturber, d'endommager ou encore de gagner des privilèges sur un système. Il existe de nombreux types de malware, qui peuvent être classifiés en fonction du type d'action qu'ils vont effectuer sur le système victime de l'attaque. Les virus informatiques sont des logiciels capables de s'autoreproduire. Ils ont les mêmes objectifs que les malwares, c'est-à-dire de nuire à un système. Les machines infectées échangent des informations avec d'autres machines sur Internet pour remonter des données ou à leur tour devenir des relais dans un botnet. Il est donc également pertinent d'observer cet écosystème dans le but de luter contre cette menace.

Les attaques de type Homme du Milieu (MITM) désignent un type d'attaque où l'attaquant vient se placer entre deux systèmes victimes. L'objectif est d'intercepter les communications

entre deux machines sans que celles-ci puissent le voir. Ce type d'attaque repose souvent sur l'usurpation d'identité, autrement appelé spoofing. Au niveau des réseaux, il peut s'agir de l'usurpation de l'adresse IPs. Le mécanisme de spoofing peut également être utilisé pour cacher l'identité (la machine) de l'attaquant ou également être utilisé dans le cadre des attaques DDoS pour amplifier l'impact de l'attaque.

Les attaques par Déni de Service Distribué (DDoS) ont pour objectif d'entraver l'accès des utilisateurs légitimes aux services proposés par les victimes de l'attaque. L'objectif de ce type d'attaque est de tirer parti d'un rival en le privant d'accès à l'une des ressources les plus importantes aujourd'hui : l'accessibilité de ses services. On distingue deux types de catégorie d'attaques DDoS : les attaques volumétriques et les attaques applicatives. Les premières ont pour objectifs de saturer la bande passante et les ressources réseaux des machines victimes. Les secondes ont pour objectif de venir saturer les ressources applicatives des machines victimes. Les attaques DDoS s'appuient sur des failles de protocoles ou des failles applicatives. Pour avoir un effet maximum, les attaques DDoS s'appuient sur un ensemble de machines, corrompues ou non, pour générer le trafic d'attaque. Ces machines sont regroupées sous le terme de bot ou relais d'attaques, et le réseau de machine est nommé botnet. Ces machines composent une infrastructure d'attaque qu'il est pertinent d'identifier pour lutter contre les DDoS.

Les processus d'ingénierie sociale sont vastes. Dans ce cadre, l'attaquant essaye de jouer de la naïveté de la victime. Il peut se faire passer pour quelqu'un d'autre en usurpant l'identité d'une personne. Les mécanismes d'attaques liées à l'ingénierie sociale sont par exemple le phishing ou le spearphishing qui vise plus particulièrement une personne. Ils sont souvent à l'origine du déploiement des attaques. Il existe des outils disponibles aujourd'hui qui permettent de générer des campagnes et d'administrer l'infrastructure crée par les machines infectées par la suite.

Il est fréquent que les attaques ne soient pas un fait isolé et que l'on puisse observer une combinaison des attaques précédemment évoquées. Les Advanced Persistent Threats (APT) présentent souvent une combinaison des techniques précédemment citées. Il s'agit d'un terme qui désigne tant bien une menace précise que le groupe d'attaquants responsable de cette menace. Pour pouvoir être qualifié d'APT, les phénomènes observés doivent être complexes (utilisation de plusieurs mécanismes d'attaques) et persistants (menace qui dure dans le temps).

Les attaquants responsables des activités peuvent donc être isolés ou regroupés sous forme d'organisations. Il est pertinent de pouvoir traquer ces organisations afin de démanteler leurs structures et infrastructures. L'infrastructure est une composante clé dans la réalisation des attaques. Une composante qu'il est donc primordiale d'analyser.

1.2.2 La fréquence des attaques

OVH, une entreprise proposant des solutions cloud et des protections DDoS, a publié en 2018 un rapport [12] qui recense et analyse les attaques dont leurs clients ont été victimes. En 2017, ils ont détecté en moyenne 1800 attaques contre leurs infrastructures chaque jour. La fréquence des attaques a diminué un peu durant le premier quart de l'année 2019 suite à une intervention importante du FBI qui visait à démanteler les botnets, ces réseaux de machines très largement utilisées pour lancer des attaques. De nombreuses actions juridiques sont menées pour limiter les botnets [13].

Depuis, la fréquence des attaques ne semble pas ralentir. Dans un rapport de Sapio Research and Deep Instinct [14], 75% des professionnels de la sécurité interrogés ont constaté une hausse

des attaques. L'arrivée des technologies d'IA a permis en partie cet essor avec, par exemple, des outils comme WormGPT [15] qui permet d'assister les attaquants dans la génération de campagne de phishing.

1.2.3 Les moyens de lutte contre les attaques et leurs problématiques

De nombreuses initiatives de partage des informations concernant les attaques ont vu le jour. Il peut s'agir de visualisation des attaques avec, par exemple, des cartes d'attaques en temps réelle [16], mais aussi de partage d'information comme avec Malpedia [17], une encyclopédie des malwares ou encore [18] un dépôt où des solutions anti ransomware sont publiées à la suite de l'identification d'une attaque. La collaboration est donc une notion importante dans la lutte contre les attaques.

Cependant, un grand nombre de solutions sont des solutions commercialisées par des entreprises de sécurité informatique. La protection des systèmes informatiques est un enjeu clé de notre époque. De nombreuses solutions sont développées au quotidien afin de lutter contre les activités malveillantes. Cependant, il y a aussi des revers à ces solutions.

Il est nécessaire aujourd'hui pour quiconque ayant des services en ligne ou tout simplement une infrastructure numérique de les protéger. Cela représente un investissement économique pour les potentielles victimes. Le cout économique existe dans tous les cas, puisque si elles ne sont pas protégées, elles risquent des conséquences économiques des attaques. Aujourd'hui, les menaces sont si importantes qu'il n'est pas possible pour une petite entreprise ou une quelconque personne qui souhaite avoir une présence en ligne afin de maintenir une activité économique, de garantir une accessibilité à 100% de ses services sans faire appel à une protection payante. C'est le cas aussi pour lutter contre les virus et les malwares. De plus, les solutions contre les attaques ne sont pas toujours efficaces contre toutes les menaces. Dans le cas des virus, par exemple, il est parfois nécessaire qu'une signature de celui-ci ait été établie avant que les antivirus puissent protéger les machines.

De plus, les attaques, et plus particulièrement les DDoS, génèrent de plus en plus de trafics illégitimes sur les réseaux mondiaux [19]. Si le cœur du réseau Internet n'est pas perturbé par le surplus de trafic, il est tout de même préférable de réduire ce type de trafic et surtout de préserver les bandes passantes des deux extrémités de l'attaque. Aujourd'hui, nous savons que les solutions anti-DDoS proposées par les fournisseurs Cloud (Cloudflare, AWS Shield, Microsoft Azure) sont les plus répandues contre les attaques DDoS et celles-ci se placent du côté des victimes. Néanmoins, ces solutions viennent altérer la distribution d'internet, en ne plaçant au cœur des réseaux qu'un petit nombre de très gros acteurs[20]. Il est aussi important de chercher d'autres solutions afin de limiter le potentiel de pouvoir que l'on donne à ces quelques entreprises. En effet, certains spécialistes de la question essaient de prévenir et de sensibiliser quant aux potentielles dérives liées à la restructuration d'Internet[21]. Cela pose des conséquences sérieuses pour la gouvernance d'Internet.

En 2017, Cloudflare, un fournisseur de solution cloud, a interrompu son contrat avec un site web en raison de contenu non adéquat à l'image que l'entreprise souhaite donner au grand public. Ils ont par la suite justifié cet acte par la quantité importante d'attaques que ce site recevait, ce qui engendrait des complications pour accueillir le trafic DDoS d'autres clients. Après cet évènement, ils ont publié un article exprimant les risques liés à leur pouvoir de censure[22]. Malgré cet article reconnaissant les risques, l'entreprise a reçu de nombreuses critiques de la part de la communauté de cybersécurité[23, 24] suite à leur annonce.

De plus, il existe aussi des exemples de cas où les fournisseurs de protection se retrouvent devant la justice, car ils étaient à l'origine des attaques en premier lieu [25, 26] et se servaient de celles-ci pour vendre des solutions de protection.

1.3 La caractérisation des machines : sources de connaissance et de défense

Pour cela, il est pertinent de caractériser les machines et leurs utilisateurs.

La caractérisation des machines et des personnes est un enjeu majeur pour la description des systèmes qui composent Internet. Les travaux d'analyse des machines sont motivés par les contextes suivants : la sécurité, le profilage des utilisateurs, l'optimisation des infrastructures et le droit d'accès aux services. En effet, connaître les utilisateurs ou leur position, par exemple, permet d'optimiser les infrastructures et la récupération des données en sélectionnant la machine la plus proche. Au niveau des utilisateurs, établir leurs profils permet d'optimiser les services qu'ils sont amenés à utiliser. Enfin, l'accès au contenu peut être parfois limité, que cela soit pour des régions géographiques ou pour éviter la surcharge de trafic vers des machines. C'est un enjeu encore plus important lorsque nous nous plaçons dans un contexte de sécurité. Du fait de l'anonymat qui est rendu possible par les propriétés techniques sur lesquelles ont été construits Internet, nous avons vu que cela laisse la place à de nombreuses activités criminelles.

Nous nous devons donc d'agir pour la protection des infrastructures qui sont menacées chaque jour. Pour cela, il est nécessaire de développer des outils limitant ces attaques de manière efficace et pérenne. Or, les attaques, aujourd'hui, ne sont plus des faits isolés et simples, et l'écosystème à l'origine de celles-ci peut être très complexe.

De nombreuses solutions contre les attaques se concentrent sur la défense de la victime. Nous pensons néanmoins que dans de nombreux cas, une limitation de la source de l'attaque, c'est-à-dire au niveau des ressources employées par l'attaquant, pourrait s'avérer bénéfique, car cela rétablirait un accès de qualité aux ressources sans avoir besoin d'investir dans des mécanismes de sécurité chers et pas toujours efficaces.

Une façon de prévenir les attaques est de connaître les attaquants (qui ordonnent les attaques) ou/et les sources des attaques (machines compromises ou attaquantes) [27]. Dans ce cas, l'objectif est de pouvoir approfondir les connaissances actuelles des attaques par la caractérisation technique et sociale des sources pour une meilleure compréhension des stratégies de sélection de celles-ci. Et à long terme, permettre d'identifier des points stratégiques de renforcement de la sécurité. Pour cela, il est judicieux de comprendre où se situent les machines qui sont à l'origine du trafic d'attaque reçu par la victime. Par où, nous n'entendons pas seulement géographiquement, mais plutôt dans quel type de réseau, administré par qui et avec quel niveau de compétence. Ces informations nous permettraient d'imaginer des solutions à la source qui se rapprochent au plus près de la réalité du terrain.

Pour mettre en place de nouvelles solutions au plus près des sources, nous avons besoin de sources de données à analyser qui doivent être réelles, puisqu'on cherche à caractériser la réalité du terrain, et récentes, puisque nous voulons proposer une caractérisation actuelle. De plus, les réseaux et l'attribution des ressources évoluent dans le temps. En effet, avec l'émergence rapide des objets IoT durant les dernières années, les attaques, plus particulièrement les DDoS, se sont multipliées. Cependant, même si de nombreuses sources affirment que les objets IoT sont à l'origine de nombreux botnets très puissants, aucune évaluation de leur bande

passante n'a été fournie. En revanche, la publication récente d'une liste de plus de 500 000 identifiants d'objets IoT [28] semble indiquer un changement de tendance quant à la typologie des botnets, un passage des objets IoT aux serveurs cloud. En effet, les services de cloud sont très avantageux, car faciles à utiliser, et possèdent de très grandes ressources et sont très peu chers.

1.4 Qui possède quelles machines ?

Afin de caractériser les machines et les organisations responsables de ces machines, il est nécessaire de se questionner sur ce que représente une organisation. L'objectif de notre travail est d'identifier précisément les organisations qui sont administratrices des adresses IPs sur Internet. Nous définissons les organisations comme une structure qui relie des personnes dans un but précis. Dans notre cas, les regroupements peuvent varier en taille, allant d'un particulier chez lui à une entreprise multinationale. Pour qualifier les organisations, nous pouvons aborder deux aspects. Tout d'abord, quel est le type d'organisation en tant que construction sociale. Quelle est la chose qui lie et regroupe un ensemble de personnes ? Le second aspect que nous pouvons aborder est comment ce regroupement existe-t-il techniquement ? En effet, les personnes se trouvent derrière des machines qui sont reliées par une infrastructure technique.

Dans cette thèse, nous présentons une analyse technico-organisationnelle des adresses IPs. Nous proposons une classification qui s'appuie sur des critères aussi bien techniques qu'organisationnels. Si l'approche technique de l'analyse des réseaux est appropriée, nous revendiquons qu'il est nécessaire de prendre en compte une composante organisationnelle et sociale pour enrichir les analyses et prendre en compte le système dans sa globalité. Au niveau des critères techniques, nous avons décidé de nous concentrer sur les services associés à une adresse IP, sur la géolocalisation de la machine ainsi que sur la présumée taille de l'infrastructure. Sur le côté organisationnel, nous avons décidé de qualifier l'organisation propriétaire de la machine utilisant l'adresse IP en fonction de son type d'organisation, de son domaine d'activité et de sa taille humaine. Ces informations nous semblent pertinentes et utiles dans le contexte d'analyse de sécurité. Ces informations peuvent nous aider à comprendre les politiques de sélection d'appareils participant à des activités malveillantes. Ces étiquettes peuvent aussi profiter aux analystes de log de sites Web. Ils permettent d'avoir une meilleure compréhension des acteurs qui se connectent à un site. Nous choisissons d'utiliser les adresses IPs pour effectuer la relation entre machine et organisation. Les adresses IP publiques, c'est-à-dire routables sur Internet, sont attribuées par des agences régionales nommées RIR. Ces agences supervisent l'affectation des adresses et la collecte des informations sur les propriétaires de celles-ci. Les adresses sont attribuées par blocs. Les tailles des blocs varient en fonction des besoins des organisations. Il existe de nombreux types d'organisation qui possèdent des adresses IP : des entreprises au sens général du terme, des entreprises fournisseur d'accès à Internet, des universités, des associations... Notre travail commence par la définition des différentes catégories que nous souhaitons représenter ainsi que l'étude des outils qui nous permettent de qualifier les adresses IPs. Grâce à cette étude, nous construisons une caractérisation axée sur 6 labels qui nous permettent de différencier les organisations identifiées.

Pour obtenir ces labels, nous avons développé un outil, IPSeen, permettant de caractériser les adresses IPs non seulement en se concentrant sur l'aspect technique du réseau auquel elles appartiennent, mais aussi sur l'organisation qui administre l'appareil qui l'utilise. La

caractérisation se fait via l'attribution d'étiquettes aux adresses IPs. Les étiquettes sont des qualificatifs destinés à décrire des caractéristiques des réseaux et qui nous permettent de les classifier. Nous avons par la suite proposé une implémentation qui nous permet, en partant d'une adresse IP, d'obtenir ces étiquettes. Notre solution est décomposable en deux sous-parties. La première sous-partie utilise des outils bien connus de la littérature pour obtenir les étiquettes techniques qui correspondent aux adresses IP : RDAP, Onyphe et GeoIP. La seconde sous-partie se concentre sur l'obtention des étiquettes organisationnelles en s'appuyant sur deux sources de données : Wikidata et les bases RIRs accessibles via le protocole RDAP.

Au travers de cette thèse, nous proposons de venir renforcer et agréger les analyses techniques existantes des adresses IP en intégrant des composantes sociales. Celles-ci nous permettront de décrire en détail les organisations propriétaires des réseaux que nous observons. Pour cela, nous allons nous appuyer sur des outils existants qui décrivent des aspects techniques des réseaux, et nous proposons un outil technique permettant d'obtenir des informations organisationnelles à propos des différentes organisations.

1.5 Plan de la thèse

Dans cette thèse, nous présentons les travaux de Métrologie des attaques. Par métrologie, nous entendons l'analyse de l'écosystème complexe des attaques sur Internet.[1] Notre objectif est de produire de la connaissance sur les organisations qui possèdent les machines sur Internet afin de pouvoir envisager des recommandations de sécurité à implémenter au niveau des réseaux des appareils identifiés.

Pour cela, dans le chapitre 2 nous présentons l'état de l'art actuel de la recherche ainsi que des outils disponibles pour la caractérisation des machines. Cet état de l'art a également pour objectif d'introduire l'ensemble des technologies qui seront utilisées par la suite dans la thèse.

Dans le chapitre 3, nous définissons les acteurs que nous souhaitons décrire et différencier. Cette classification est par la suite confrontée à la réalité et aux informations que l'on peut obtenir par les outils. Nous en concluons donc une liste de labels à attribuer aux adresses IPs afin de différencier les acteurs évoqués. Dans la seconde partie de ce chapitre, nous présentons notre outil IPSeen, outil qui est une implémentation de notre méthodologie et qui permet d'obtenir les labels établis. Notre outil a été optimisé pour caractériser les adresses IPs de version 4 puisque nous avons obtenue en très grande majorité des ensemble de données de cette version d'IP. Il s'appuie sur différentes sources, dont Wikidata, le protocole RDAP, la base de géolocalisation des adresses IPs GeoIP et le crawler Onyphe. Nous avons développé deux versions de cet outil, une version empirique ainsi qu'une version qui s'appuie sur le machine learning et les algorithmes de NLP.

Dans le chapitre 4, nous présentons premièrement une évaluation des deux versions de notre algorithme IPSeen. Chaque étape clé de notre solution est évaluée et une comparaison entre les deux versions est proposée. À la suite de cette évaluation, nous proposons également une analyse des limites de notre solution, que celles-ci soient liées à l'implémentation ou bien aux sources de données utilisées. Nous abordons aussi les possibles travaux futurs qui pourraient permettre de pallier ses limites.

1. Ce travail a été mené avec la précieuse collaboration des entreprises Sekoia et Onyphe. Sekoia a accepté de nous fournir des données pour réaliser notre analyse présentée en Chapitre 5. Onyphe nous a offert une licence Eagle View pour accéder à son service de scan global d'Internet. Nous les remercions profondément.

Dans le chapitre 5, dernier chapitre de contribution de cette thèse, nous présentons les résultats que nous avons obtenus en appliquant notre outil à un ensemble de données d'attaques. Cet ensemble de données nous a été fourni par l'entreprise de sécurité Sekoia, et contient plus de 89 000 adresses IP uniques ainsi que 174 labels d'attaques différents. Ce chapitre propose également un retour d'expériences de la recherche de traces d'attaques, ainsi que des différentes interactions que nous avons pu avoir avec diverses industries.

Enfin, nous concluons l'ensemble de ces travaux dans le chapitre 6. Nous présentons aussi un ensemble de travaux futurs qui nous permettraient d'améliorer IPSeen et d'enrichir les différentes bases utilisées pour notre caractérisation.

Chapitre 2

État de l'art

Internet permet de relier des machines et des utilisateurs derrière ces machines. Or, le possible anonymat sur Internet n'est, dans certains cas, pas quelque chose souhaité dans le cadre des cyberattaques, même s'il est recherché par les utilisateurs [29]. Dans un contexte d'analyse d'attaques, la caractérisation des machines présente aussi un grand enjeu au niveau de la compréhension de celles-ci et du développement de mécanismes de défense face à celles-ci. Les solutions de défense contre les attaques sont développées grâce aux connaissances qui peuvent être recueillies à partir des attaques. Pour atteindre cet objectif, des outils d'analyse automatisés accélèrent l'acquisition d'informations sur tous les aspects d'une attaque. L'automatisation des outils de Cyber Threat Intelligence (CTI) est un des enjeux majeurs de la recherche scientifique[30, 31, 32].

Il y a de nombreux aspects à analyser en ce qui concerne les attaques. Plusieurs modèles visent à décrire ces aspects. La matrice ATT&CK [27] présente les différents aspects d'analyse des attaques : de la reconnaissance à l'accès aux informations d'identification en passant par l'exécution et la persistance. Parmi les caractéristiques, un aspect appelé : "développement des ressources". Le terme "développement des ressources" fait référence aux stratégies par lesquelles les ennemis produisent, acquièrent ou volent ce qui peut être utilisé pour faciliter le ciblage. L'infrastructure, les comptes et les capacités sont quelques exemples de ces ressources. La caractérisation de l'infrastructure permet donc de fournir des renseignements sur les attaques. L'infrastructure est également l'une des quatre caractéristiques du modèle Diamond [33], qui est un autre modèle de représentation des attaques. Enfin, le modèle Cyber Kill Chain [34] inclut également l'infrastructure dans l'étape *Delivery* du modèle et souligne la nécessité de prendre en compte la méthode de distribution et l'infrastructure en amont pour réagir. Ces modèles permettent de décrire et de classifier les attaques et sont très largement utilisés dans la littérature [35, 36]. Ils ont également été comparés par le passé [37]. Dans cette comparaison, les auteurs ont mis en avant les différents cas d'usage ainsi que les différents avantages et inconvénients de chacun des modèles. Ces travaux mettent en avant le modèle de la matrice ATT&CK pour sa complétude. L'étude des différentes phases de déroulement des attaques permet également d'envisager des solutions qui détectent au plus tôt les attaques [38, 39].

Dans cet état de l'art, nous parcourons les différentes propositions de la littérature qui ont pour objectif de créer de la connaissance autour des machines distantes. Nous commençons par présenter une vue générale d'Internet, de son fonctionnement et des différentes organisations qui l'administrent. Puis nous discutons des mécanismes qui se basent sur les adresses IPs. Puis, nous proposons un tour d'horizon des différents types de caractérisation des adresses IPs ainsi que des outils permettant de les réaliser. Par la suite, nous présentons l'état des connaissances actuelles des infrastructures d'attaque. Ces connaissances ont, pour une partie, été obtenues

via les outils présentés dans la section précédente. Nous discutons les différents points de défense dans l'architecture. Enfin, nous présentons des travaux menés pour la compréhension de la dimension sociale des attaques. Nous présentons aussi des travaux dans lesquels des composantes sociales se sont révélées nécessaires.

2.1 Internet : son architecture et sa pile protocolaire

Nous souhaitons caractériser des machines sur Internet, et pour cela, il est nécessaire d'avoir une vue globale sur l'infrastructure matérielle, ainsi que les différents systèmes qui en font partie. Internet s'appuie sur une pile protocolaire [40], la pile TCP/IP. Cette pile permet d'avoir des identifiants attribuables aux machines : une adresse IP [41, 42] ainsi que d'identifier des services avec des numéros de ports [43, 44]. Ces identifiants sont attribués de différentes façons. Les adresses IPs sont attribuées par bloc à une entité qui pourra par la suite attribuer une adresse par machine sur son réseau. L'adresse IP est unique, et donc semble pertinente pour identifier un appareil sur Internet. Il y a deux versions d'adresse IP : les adresses IP version 4 (IPv4) [41] et les adresses IP version 6 (IPv6) [42]. Cette dernière version répond à un épuisement des adresses IPv4. Néanmoins, son adoption est plutôt lente, même si en progrès [45].

Dans cette section, nous étudions le fonctionnement d'Internet afin d'identifier les composantes que nous pouvons exploiter afin de caractériser les machines sur Internet.

2.1.1 Les Systèmes Autonomes : répartition de l'infrastructure mondiale

Un Système Autonome (AS) est un ensemble de réseaux informatiques regroupés et généralement administrés par une même entité. Les AS sont identifiés par un numéro d'AS (ASN = Autonomous system number), numéros attribués par les Registres Internet Régionaux (RIR) [46]. Les AS sont interconnectés en de nombreux points à travers le monde appelés points de peering.

De nombreux travaux [47, 48, 49, 50] apportent des caractérisations sur les AS. Le projet *AS rank* de *CAIDA* [47] propose un suivi des différentes AS et propose un classement en fonction de leurs influences dans le monde. Des informations sur les différents points d'interconnexion ainsi que les réseaux qui y sont connectés sont aussi disponibles, comme par exemple avec la base *PeeringDB* [48]. Dans [49], les auteurs analysent et classifient les AS en termes de domaines d'activité afin de mieux de mettre en avant les possibles intérêts des entités. Afin de classifier les AS, les auteurs se basent sur deux sources d'informations : le projet *AS Rank* de *CAIDA*, et les données de la Securities and Exchange Commission des États-Unis (SEC) accessible via le système EDGAR. La SEC possède des documents et des informations sur les entreprises américaines. Le protocole WHOIS [51], qui permet d'accéder aux registres des RIRs est aussi exploité comme source de données. Plus récemment, les auteurs de [50] ont introduit ASdb, un outil de caractérisation des AS en fonction de 17 domaines d'activités et 95 sous-catégories.

2.1.2 Les différentes institutions responsables d'Internet

Il existe de nombreuses entités en charge du bon fonctionnement d'Internet. Ces entités varient en leur nature et leurs rôles. Dans cette section, nous comparons les entités ainsi que les données qu'elles mettent à dispositions.

Tout d'abord, l'ICANN est une association de droit californien en charge de coordonner les différents identifiants d'Internet à l'échelle internationale [52]. IANA est un département de l'ICANN en charge de la gestion de l'attribution des adresses IPs ainsi que des numéros d'AS, des numéros de port et des zones racine DNS[53]. IANA gère également les différents *Registres Internet Régionaux* (RIR).

En introduction de cette section, nous avons présenté le protocole IP ainsi que les adresses IP qui permettent d'identifier une source et une destination dans le réseau IP. Ces adresses sont attribuées par différentes entités régionales : Ripe NCC pour l'Europe [54], l'ARIN pour l'Amérique du Nord [55], l'APNIC pour l'Asie[56], l'AFRINIC pour l'Afrique [57] et LACNIC pour l'Amérique centrale et du sud[58]. Ces organisations sont en charge d'affecter les adresses IPs et maintiennent une base de données d'informations sur les possesseurs des adresses IPs.

2.1.3 Les points de terminaisons

Les points de terminaisons regroupent tous les types d'appareils qui sont connectés à Internet. Il en existe de nombreux types. Dans la littérature, de nombreux auteurs ont proposé des catégorisations de ces différents points de terminaisons ainsi que des solutions automatisées de classification de ceux-ci.

Il existe plusieurs classifications [59, 60, 61] plus ou moins longues, qui différencient les objets IoT ainsi que d'autres équipements de terminaisons ou intermédiaires. L'automatisation de la classification présente de nombreux avantages : la détection de trafic malveillant au niveau des objets IoT [59], la détection de vulnérabilités et l'analyse des appareils [60], la détermination des signatures des appareils IoT [61].

Afin de réaliser les analyses de ces machines, les auteurs s'appuient sur les adresses IPs pour l'identification de la machine ainsi que sur différentes caractéristiques extraites du trafic de ces machines.

2.2 Pourquoi utiliser les adresses IPs pour caractériser des machines ?

Les adresses IPs permettent d'identifier des machines sur Internet[62]. De ce fait, il existe de nombreux outils qui s'appuient sur les adresses IPs comme dénominateur commun pour estimer la qualité de la source de trafic [63, 64, 65]. Cependant, des travaux viennent questionner la pertinence de ces outils [66, 67, 68].

Avec l'apparition des technologies de NAT [66], nous ne pouvons plus faire l'association entre une machine et une adresse IPs de manière généralisée. Il est donc nécessaire de remettre en question plusieurs points : la stabilité des attributions des adresses IP ainsi que le nombre de machines qui se trouvent derrière une unique ou un pool d'adresses IP. En effet, puisque nous avons vu dans la section précédente que les adresses IPs sont à l'origine de mécanismes de filtrage, alors il est nécessaire de discuter des impacts que cela peut avoir. Nous discutons des travaux qui répondent à ces deux points dans les sections suivantes.

2.2.1 Stabilité des adresses IPs dynamiques

Dans un contexte de suivi et de caractérisation des adresses IPs, il est nécessaire de se poser la question de la stabilité des adresses IPs. En effet, si l'on étudie une adresse à un instant t, qu'en est-il à t+1 ? Et quelle est la durée de rétention d'une adresse ?

Cette question a été étudiée dans [67]. Ces travaux cherchaient à mesurer l'impact des blocklists sur les utilisateurs du fait que les adresses IPs soient partagées. L'idée est de quantifier le nombre de machines qui ont pu être victimes d'un blocklistage de leurs IPs à cause des actions malveillantes qui ont pu être entreprises par les utilisateurs d'avant. Les conséquences de la réutilisation des adresses sont importantes, avec entre 53-60% des 151 blocklistes analysées qui contiennent des adresses réutilisables. Parmi ces listes, entre 30 600 et 45 100 IPs sont des IPs utilisées par plusieurs personnes. Il est démontré que la réutilisation d'adresses peut avoir un impact sur 78 utilisateurs légitimes pendant un maximum de 44 jours pour chaque IP bloquée, sachant que 53% des 151 blocklists analysées ont au moins une IP dynamiquement attribuée et 60% de ces blocklists ont au moins une adresse IP naté.

Une autre étude [68], s'intéresse à la stabilité des affectations des adresses IPs dans le contexte du tracking web. Parmi les 34 488 adresses IP publiques étudiées, 45 % des FAI autorisent les utilisateurs à conserver la même adresse IP pendant plus de 30 jours, et 87 % des participants conservent au moins une adresse IP pendant plus d'un mois. On peut donc conclure que les impacts de filtrage des adresses IPs peuvent être importants et qu'il est nécessaire d'associer une période de validité à l'observation des adresses IPs. Une observation a donc un temps maximum de validité.

2.2.2 Nombre de machines derrière une adresse IP

Depuis de nombreuses années et avec l'apparition du NAT, la question du nombre d'appareils derrière un réseau naté a été traitée[69, 70]. Pour cela, il est nécessaire de différencier les machines au-delà de leurs adresses IPs. Une des premières solutions proposées [70] s'appuie sur l'observation des champs IP et, plus particulièrement, l'incrémentation d'un champ qui agit comme un compteur. Dans [69], les auteurs ont présenté une technique de fingerprinting des machines qui permet entre autres de connaître le nombre de machines derrière un réseau naté. Enfin, d'autres solutions s'appuient également sur des modèles de machine learning pour répondre à cette problématique [71, 72]. D'autres méthodes aussi ont été proposées, comme celle de [73] qui s'appuie sur les différents champs de TCP et d'IP. Une autre solution [74] propose une méthode extrayant des empreintes numériques sur des captures de types netflow, type de capture où toutes les informations des paquets IPs ne sont pas extraites. Cependant, ces solutions nécessitent l'accès à des captures des réseaux.

Dans [75], les auteurs ont étudié les réseaux domestiques natés. Parmi les caractéristiques, pour l'ensemble des foyers étudiés, ils ont relevé que plus de la moitié des foyers avaient au moins 5 appareils connectés, et en moyenne 7 appareils connectés. Une autre solution proposée dans la littérature [76] se concentre sur la détection des objets IoT derrière les réseaux natés.

Avec ces solutions, il est possible de connaître le nombre de machines qui se trouvent derrière un réseau natés, qu'il s'agisse d'une estimation sur certains types de réseaux, ou que cela soit fait via mesure et analyse de trafic. Néanmoins, dans ces deux cas, cela nécessite soit une analyse menée sur un certain type de réseau, soit d'avoir accès au trafic échangé avec le réseau naté. Cependant, dans le cadre de notre travail, nous n'avons pas accès à ces informations. Nous avons donc besoin d'autres indicateurs pour estimer la taille d'un réseau.

2.2.3 Usurpation des adresses IPs

Le protocole IP peut être exploité via le champ d'adresse IP source. En effet, lorsque le protocole IP a été créé, il n'y avait pas de mécanisme de sécurité qui permettait d'authentifier les machines qui communiquent. L'usurpation d'identité est donc aujourd'hui monnaie courante dans le cadre des attaques informatiques, puisqu'elle permet de dissimuler l'identité des attaquants.

Afin de lutter contre cela, un RFC [77] a été rédigé pour présenter et encourager les fournisseurs d'accès à filtrer le trafic usurpé.

Afin de lutter contre l'usurpation d'adresse IPs, différentes solutions ont été proposées dans la littérature, comme dans [78], où les auteurs ont proposé une architecture qui filtre les paquets entre les domaines en se basant sur les routes qui peuvent être construites uniquement sur la base des mises à jour BGP échangées localement.

Néanmoins, comme on peut le constater grâce au Spoofer Project [79] de CAIDA, il y a encore (le 10 mars 2024) environ 20% des blocs IPv4 qui laissent transiter du trafic spoofé. Il est donc nécessaire de prendre en compte l'existence de ce mécanisme, plus particulièrement dans le contexte des analyses d'attaques. En effet, il s'agit d'un mécanisme particulièrement abusé pour générer des amplifications DDoS (DNS :[80], NTP : [81], Memcached [82]) ou encore pour masquer l'identité de l'attaquant.

Les adresses IPs permettent d'identifier sur Internet les différentes machines. Bien sûr, les adresses IPs ne sont pas les seules informations qui permettent d'identifier des machines. On peut penser aux autres caractéristiques qui permettent de constituer des empreintes digitales des machines [69, 83]. Néanmoins, lorsque l'on fait le choix d'analyser les adresses IPs, il est nécessaire de prendre en compte les particularités exposées précédemment, à savoir l'usurpation des adresses IPs et le NAT. Dans le cadre de notre application, nous avons obtenu un set de données où les adresses IPs étaient celles des machines intermédiaires et servaient à établir un relais entre les machines infectées/bot et l'attaquant. La problématique d'usurpation d'identité n'était donc pas à intégrer dans ce contexte particulier.

2.3 Quelles sont les étiquettes qui peuvent être attribuées aux adresses IPs via des outils existants ?

Néanmoins, l'adresse IP reste une composante majeure pour l'analyse des activités malveillantes, mais également pour la caractérisation plus générale des machines sur Internet et les utilisateurs de ces dites machines.

2.3.1 Géolocalisation des adresses IPs

La géolocalisation des adresses IPs est un sujet très largement couvert dans la littérature [84], car elle possède de nombreuses applications : droit d'accès au contenu numérique (Netflix), publicité ciblée, développement d'applications (météo, événements locaux), aide à l'optimisation des réseaux, et bien sûr, analyse de sécurité. Dans cette section, nous traitons de l'obtention des localisations d'adresses IP ainsi que des bases de données de géolocalisation existantes et de leur qualité.

2.3.1.1 Méthode de géolocalisation des adresses IPs

Dans cette section, nous présentons les outils ou sources de données qui permettent de caractériser les machines via leurs adresses IPs. Les étiquettes peuvent provenir de bases de données, d'outils existants, commercialisés ou non, ainsi que de crawler de réseau. La géolocalisation des adresses IPs a de nombreuses applications :

— la personnalisation de contenus : cela peut être fait en fonction des goûts d'un utilisateur pour savoir par exemple quoi afficher, ou adapter un service à la géolocalisation de l'utilisateur, par exemple (langue, météo, évènements locaux).
— l'optimisation des services en dirigeant les utilisateurs d'une machine vers le datacenter le plus proche afin de réduire la latence, dans le cadre de réseaux CDN.
— licence pour du contenu restreint en fonction de la géolocalisation des individus

Pour répondre à ces enjeux, les informations à obtenir sur les machines ainsi que leurs utilisateurs diffèrent. Par exemple, la géolocalisation [84] permet d'adapter des informations comme la langue, la météo, les évènements locaux ou la limitation du contenu.

Les méthodes de géolocalisation des adresses IPs sont variées. Elles peuvent reposer sur divers systèmes de mesure, principalement des mesures de topologie ou de délais ou encore une combinaison des deux[85].

Les mesures par délais répondent au principe que deux hôtes proches l'un de l'autre ont des délais similaires par rapport à des points de repères connus. Afin d'obtenir des sets de données avec des localisations connues, certains travaux s'appuient sur les bases de géolocalisation connues et combinent ces géolocalisations avec des mesures de délais pour trouver les localisations manquantes[86].

En ce qui concerne les mesures par topologie, comme présenté dans [87], les mesures de l'itinéraire Internet sont converties en un ensemble de contraintes sur les emplacements inconnus des cibles et des intermédiaires, ce qui crée une topologie. La géolocalisation se fait de manière à respecter cette topologie.

Enfin, d'autres travaux s'appuient sur les retours DNS inversés des adresses IPs[88] en inspectant la méthode de construction des noms de domaines avec un modèle de machine learning pour trouver des informations sur la géolocalisation. Les auteurs ont constaté que régulièrement des informations géographiques étaient contenues dans les noms d'hôtes.

Des travaux[89] plus récents emploient, en plus des mesures vues précédemment, des algorithmes de Machine Learning supervisés (ici Naives Bayes) pour améliorer la précision des mesures.

Les classifications ont par la suite été légèrement améliorées dans [90]. Ces travaux intègrent en plus des mesures de délais et de topologie des caractéristiques sociales comme la densité de population d'une ville.

Néanmoins, ces techniques ne sont pas infaillibles, comme démontré dans [91] où les auteurs ont développé deux attaques contre les systèmes de mesure de géolocalisation par mesure de délais et de topologie. Les adversaires peuvent éviter d'être détectés en utilisant la variabilité du retard du réseau et la complexité des chemins de réseau sur Internet pour masquer leur falsification. Les résultats des mesures par topologie sont encore plus impactés par ces attaques que celles par mesure de délais.

Le point commun de toutes ces méthodes est qu'elles sont soit lentes, soit peu extensibles, c'est pourquoi, la plupart du temps, nous utilisons des bases de données géographiques précalculées.

2.3.1.2 Base de données de géolocalisation des adresses IPs

Il existe de nombreuses bases de données de géolocalisation des adresses IPs. Dans cette partie, nous effectuons un état des lieux de ces bases, puis nous présentons des évaluations de celles-ci.

De nombreuses bases de données de géolocalisation existent, car elles présentent l'avantage de recenser l'association entre adresses IPs et pays/villes/coordonnées géographiques sans avoir besoin de déployer une solution de mesure. Ici, nous listons et décrivons quelques-unes connues de la littérature.

Geolite[92] ou GeoIP[93] sont des bases de données de Maxmind. La première est gratuite et la seconde est payante. Maxmind est une société américaine spécialisée dans la localisation des adresses IPs. Les bases de données contiennent de nombreuses informations pour chaque adresse IP : la localisation au niveau de la ville, de la région, du pays, le code postal, la longitude et la latitude, le FAI pour la version GEOIP (payante) ainsi que le type de connexion. Dans la version gratuite, nous retrouvons le nom de réseau et/ou d'AS et le numéro d'AS qui sont extraits des bases de données des RIRs. IP2Location et sa version Lite (gratuite)[94] sont deux autres bases de données d'une autre entreprise, qui comme GEOIP propose de nombreuses caractéristiques associées aux adresses IPs, dont principalement leur localisation au niveau du Pays/Ville/région/FAI/code postal/nom de domaine pour la version payante.

GeoNetMap [95] est une solution de géolocalisation de l'entreprise Geobyte. La solution propose un accès via API qui permet d'obtenir la géolocalisation d'adresses IPs. Les informations récupérables sont similaires à celles précédemment citées, avec comme particularité supplémentaire le gentilé au singulier et au pluriel, la distance avec les villes voisines ainsi que la monnaie. Néanmoins, avec cet outil, il n'y a pas d'informations concernant les AS, ni les FAI. Avec le peu d'information sur l'obtention de la géolocalisation des adresses IPs dans ces bases de données, nous pouvons nous poser la question sur la qualité de celles-ci.Néanmoins, comme celles-ci sont largement utilisées, de nombreuses évaluations et comparaisons sont disponibles[96, 97, 98, 99]. Dans ces travaux de 2011 [96], les auteurs proposent une comparaison entre les informations contenues dans les bases de données comme GeoIPLite ou InfoDB et les ont comparées à la table de routage d'un routeur d'un FAI européen. Leurs conclusions confirment la qualité au niveau des géolocalisations de pays, même si certains pays comme les USA sont surreprésentés, mais déplorent la qualité de la géolocalisation au niveau des villes. Dans [97], les auteurs présentent des conclusions similaires en comparant également 5 bases de données, dont les bases de IP2Location et Maxmind, avec les localisations des POP (Point of Presence) des routeurs.

Enfin, plus récemment, des travaux[98] de comparaison de données GPS avec des bases de données ont validé la qualité de la version gratuite de la base de données de GEOIPLite. Une précision inférieure à 10 km a été trouvée pour les réseaux filaires des FAI et des universités, et une précision au niveau de la ville pour les réseaux mobiles. Néanmoins, cette étude a été menée seulement aux États-Unis.

Ces conclusions sont aussi partagées par des travaux plus récents, comme dans [99], montrant que ces outils de géolocalisation tendent vers le même résultat en ce qui concerne le pays d'origine d'une adresse IP.

Nous pouvons donc conclure que les bases de données de géolocalisations sont assez fiables au niveau de la géolocalisation des IPs dans les pays, et présentent l'avantage d'être rapides sans avoir besoin de mettre en place une méthode de mesure.

2.3.2 Port, services, systèmes d'exploitation : crawlons les réseaux

Un autre point d'analyse présent dans la littérature concerne les spécificités techniques des machines. L'objectif est le suivant : décrire les machines présentes sur un réseau en obtenant leurs caractéristiques techniques. Le scan des réseaux, et plus particulièrement des ports ouverts, est une pratique courante de sécurité, dont dans le cas de la détection de vulnérabilité ou de la prévention des attaques, car nous savons que certains services sont plus affectés par les attaques, comme, par exemple, le service DNS avec l'amplification DDoS[100].

NMAP[101], un outil développé par G.Lyon en 1999 en est un exemple parfait, puisque toujours utilisé aujourd'hui. Cet outil permet d'obtenir de nombreuses informations sur les machines qui se trouvent dans un réseau, avec, entre autres, les hôtes de celui-ci, les services ouverts, les systèmes d'exploitation des machines ainsi que leurs versions, les règles de filtrage des paquets. C'est un outil qui peut être utilisé pour faire de la découverte de réseau, des audits de sécurité, de l'inventaire de réseau. Ces informations sont obtenues en analysant les réponses obtenues suite à l'envoi de différents paquets IPs. Il s'agit donc d'une démarche active où une machine vient solliciter les hôtes d'un réseau.

Néanmoins, par la suite, d'autres outils sont apparus pour répondre à des objectifs similaires, mais cette fois-ci à une échelle beaucoup plus grande : Internet. Ces nouveaux outils ont dû répondre à de nouveaux défis à cause du passage à une échelle beaucoup plus grande, notamment en termes de vitesse de scan avec ZMAP[102], outil développé en 2013 qui a par la suite donné naissance à l'entreprise Censys et à l'outil du même nom. ZMAP permet de scanner l'entièreté de l'espace IPv4 en moins de 45 min.

Aujourd'hui, il existe plusieurs crawlers qui scannent l'ensemble d'Internet et mettent à disposition des utilisateurs les résultats. Censys[103, 104], comme nous l'avons vu précédemment, est une solution qui découle de ZMAP. Censys propose un accès gratuit pour les chercheurs. Shodan[105] est un moteur de recherche qui permet de connaître les services associés à une adresse IP. De nombreux travaux de recherche s'appuient sur Shodan comme source. Pour utiliser Shodan, il faut souscrire à un abonnement. Enfin, Onyphe[106] est un moteur de recherche qui propose des informations à propos des machines accessibles sur Internet. C'est un outil qui a été développé dans le but de réaliser des analyses de sécurité. Il permet notamment d'obtenir des informations sur les ports ouverts et les services qui se trouvent derrière les adresses IP, mais aussi les noms de domaines, la géolocalisation de l'IP ainsi que le numéro d'AS à laquelle elle appartient[1].

Les crawlers de réseau utilisent une approche proactive, ce qui signifie que les connaissances sont acquises à partir de mesures actives. Ces outils permettent d'évaluer la sécurisation des réseaux en connaissant les machines qui sont accessibles depuis Internet. Il existe plusieurs applications qui découlent de ces outils, comme la détection des vulnérabilités [107, 108, 109, 110].

D'autres travaux s'appuient sur les résultats produits par les crawleurs, facilitant la lecture des informations retournées par Censys et Shodan, dans le but de créer des analyses de sécurité dans des contextes particuliers comme l'IoT [111].

Parmi ces outils, nous pouvons nous poser la question de l'évaluation des résultats qui sont retournés. Dans [112], les auteurs proposent une comparaison de différents outils de scan, incluant Shodan et Censys. Les différentes solutions sont classées en fonction de l'accessibilité de leurs données, de leur moteur de recherche, de l'interprétabilité des résultats ainsi que

1. Dans le cadre de cette thèse, nous avons pu bénéficier d'un soutien d'Onyphe

de la complétude des résultats. Les outils sont présentés en deux catégories, les scanneurs automatiques et les scanneurs sur demande de l'utilisateur. Néanmoins, cette seconde catégorie est moins pertinente dans notre cas, car les résultats obtenus ne sont pas partagés à tout le monde.

Néanmoins, les crawleurs présentent aussi des limites. En effet, comme les résultats rendus par ceux-ci permettent de mettre en avant une certaine partie des machines d'Internet ainsi que leurs potentielles vulnérabilités, il existe aussi des solutions qui sont mises en place pour détecter le scan de port fait par ces solutions. [113] présente un détecteur de scanneur (dont Shodan et Censys) qui combine plusieurs méthodes traditionnelles pour faire la détection de crawleurs.

2.3.3 Information à propos des possesseurs des adresses IPs

Afin d'acquérir une adresse IP, les propriétaires doivent se rapprocher des RIRs : Regional Internet Registries [114, 115]. Ces organisations géographiques supervisent l'attribution des blocs d'adresses IPs. Elles possèdent des bases de données qui contiennent des informations à propos des possesseurs d'adresses IPs comme : le numéro d'AS, le pays, la description de l'AS, le nom du réseau, ainsi que des informations de contact du propriétaire comme son adresse, son numéro de téléphone, une adresse e-mail pour signaler les comportements abusifs, et parfois même l'identité d'un contact.

Ces informations sont accessibles via deux protocoles : Whois[51] et RDAP[116, 117, 118].

RDAP est le protocole le plus récent qui remplace Whois. Whois et RDAP sont des protocoles conçus pour récupérer facilement ces informations sous la forme d'un texte pour le premier et d'un JSON pour le second. Ces données peuvent être exploitées dans divers contextes : analyse de sécurité, analyse de type de réseaux, pour trouver le nom du fournisseur d'accès à Internet.

Il existe de plus en plus d'outils qui donnent des informations à propos des propriétaires des adresses IPs comme Onyphe[106], GeoIP dans sa version payante[93], mais aussi IPInfo.io[119] avec le champ org, ou encore IPInfoDB [120] qui aujourd'hui fournit l'ISP ainsi que l'usage (type de réseau, par exemple universitaire). Avec ces évolutions, nous constatons qu'il y a aujourd'hui nécessité de connaître des informations plus précises que simplement la géolocalisation sur la source du trafic que l'on reçoit. Néanmoins, ces outils sont des solutions commerciales payantes, et les méthodes d'obtention ne sont pas publiées.

2.4 Caractérisation des infrastructures d'attaques

Dans la section précédente, nous avons vu un certain nombre d'outils qui permettent d'affecter des informations techniques sur des machines. Dans cette section, nous citons des exemples d'analyses qui ont permis d'établir des connaissances sur les machines sur Internet, et plus particulièrement sur les infrastructures qui ont par le passé généré des attaques. Certaines s'appuient sur les outils que nous avons vus.

2.4.1 Identifications des infrastructures et des sources d'attaques

Dans un premier temps, intéressons-nous aux attaques par phishing. Le phishing est une attaque qui s'appuie sur l'ingénierie sociale, où l'attaquant abuse de la victime en se faisant passer pour une organisation que celle-ci connaît, et s'empare des identifiants ou autres

informations personnelles de la victime. Dans les travaux suivants[121], les auteurs proposent une caractérisation des sites de phishing et de téléchargement malveillant en se basant sur les adresses IPs. D'autres travaux [122], s'attaquent à la problématique du pharming, un type d'attaque par phishing avancée qui s'appuie sur les registres DNS. Dans [122] les auteurs proposent une solution qui s'appuie sur la comparaison des adresses IPs ainsi que les codes sources des pages web. Cette solution permet de protéger les usages contre les machines participantes aux infrastructures d'attaques.

Un des mécanismes classiques des attaques est l'utilisation de machines corrompues, appelées bots, qui ensemble créent un réseau de bots ou botnets[123]. Les auteurs de [124] ont proposé une classification des mécanismes de détection en 4 catégories : basée sur les signatures (botnet connu), basée sur les anomalies, basée sur le DNS[125] et basée sur l'exploitation d'outils de data mining du trafic de C2.

L'analyse de ces réseaux est donc très importante, puisque le démantèlement de ces réseaux permet de temporairement retirer les ressources des attaquants. De nombreux travaux traitent de l'étude de ces réseaux. Dans [126], les auteurs proposent une évaluation des similarités entre les adresses IPs et la création de clusters d'IPs qui permettent de différencier les bots du reste des machines.

Cependant, il a été montré dans [127] que l'arrivée en Europe du RGPD complique l'étude des réseaux de bots, avec en cause le fait que les adresses IPs sont considérées comme des données personnelles. Cependant, les auteurs concluent que dans certains cas, nous pouvons nous appuyer sur l'intérêt général pour justifier la nécessité de traquer les réseaux de bots.

Les analyses des attaques vont par la suite permettre de soulever des points de sécurisation. Par exemple, des analyses d'attaques passées ont révélé quels types d'appareils peuvent être utilisés de manière abusive, avec l'exemple des appareils IoT dans le cas de l'attaque DDoS Mirai en 2016 [128]. En outre, de multiples réseaux d'appareils ont été étudiés pour comprendre comment les réseaux de bots sont créés [129].

2.4.2 Type de machines

De nombreuses attaques ont pu par le passé être reliées aux objets IoTs. Nous avons par exemple abordé plus tôt le cas du botnet Mirai[128, 130]. Les auteurs de [131] ont également suivi un autre botnet nommé Mozi, et d'après leurs analyses, ce botnet IoT semble prendre les devants de Mirai à présent. Les menaces liées aux objets IoT sont liées au fait que la sécurité n'a pas été prise en compte au moment de la création et du déploiement de ces objets. Les principes de "sécurité par design" n'ont pas été respectés pour beaucoup de ces objets. Grâce aux analyses d'attaques passées qui ont attiré l'attention sur les objets IoTs[132], des chercheurs ont proposé des outils pour s'intéresser plus particulièrement à ces machines sur Internet et à leurs vulnérabilités. C'est le cas pour les auteurs des travaux [111] qui ont proposé une solution pour faciliter la lecture des informations retournées par Censys[104] et Shodan[105] pour faire de l'analyse des vulnérabilités IoT.

Des travaux [133], où les auteurs ont mis en place des pots de miel pour traquer l'évolution des réseaux de bots, montrent que le botnet Mirai est toujours actif, même si légèrement modifié, et reste dominant dans l'environnement des botnets IoT.

2.5 Analyse des composantes sociales des attaques : motivations, groupement, organisations

Si les informations techniques sont largement traitées au travers de la littérature, l'aspect socio-organisationnel est aussi analysé. Dans cette catégorie d'information, sont incluses les analyses menées sur les motivations des attaques, l'intégration des motivations dans les analyses produites ou encore l'étude des organisations d'attaquants. Ces trois aspects permettent d'approfondir les connaissances sur les attaques et de les appréhender dans leurs globalités.

2.5.1 Motivations des attaquants

Les attaquants peuvent avoir plusieurs motivations qui ne sont pas uniquement financières. Les motivations des attaques DDoS ont été étudiées dans [134]. Selon les auteurs, les attaquants peuvent être motivés par des gains financiers, la vengeance, des convictions idéologiques, des défis intellectuels ou même la cyberguerre. Les attaques ne peuvent être étudiées sans leur contexte social, économique et politique. Les auteurs de [135] concluent même qu'il est impératif de surveiller l'environnement socioculturel, économique et politique pour comprendre les motivations. Comme on l'a dit, l'étude des attaques ne peut être décorrélée de son aspect social. Mais aussi, on peut être enclin à s'interroger sur le lien entre les motivations et les moyens mis en œuvre pour créer l'attaque.

Les attaques sont également la marchandise du jour [136]. Les infrastructures et les attaques sont à vendre. Selon l'auteur, le marché de la cybercriminalité s'est considérablement développé et, à l'instar d'autres marchés en expansion, les entreprises qui le desservent ont élargi leur offre de produits. Ceci met en avant l'intérêt économique et surtout le besoin de rentabilité des attaques.

2.5.2 Inclusion de composantes sociale dans l'analyse des attaques

Des travaux comme [137] défendent la nécessité d'intégrer des composantes géopolitiques dans les analyses des attaques. Dans ces travaux, les auteurs ont repris le modèle de cyber kill chain pour intégrer des composantes d'analyses géopolitiques ainsi que des composantes stratégiques. Ils mettent en avant la nécessité d'inclure des notions comme les contextes géopolitiques ainsi que les objectifs, les effets directs des attaques et les effets indirects.

Tout d'abord, une première composante assez classique est l'origine géographique. L'objectif des analyses géographiques est de géolocaliser les machines qui participent aux activités malveillantes et de les affecter généralement à un pays. Par exemple, dans [138], les pays sont répartis en différents groupes à l'aide de l'algorithme K-means pour établir des topologies des actions de cybercriminalité en fonction des pays. D'autres travaux[139], classent les pays en fonction de leur fréquence d'apparition dans les listes connues d'incidents de cybercriminalité.

L'inclusion des motivations et de l'impact économique est également un sujet abordé dans la littérature [140]. Dans ce travail, l'auteur a mis à jour un modèle d'attaque par hameçonnage pour y inclure l'impact économique de ce type d'attaque.

Dans [141], l'utilisation d'adresses IP résidentielles par le biais de proxys pour contourner les restrictions de sécurité qui peuvent être présentes dans d'autres types de réseaux est une nouvelle pratique que les auteurs ont observée. Ces travaux montrent qu'il est nécessaire

d'étudier les aspects organisationnels et sociaux des réseaux. Ils ont utilisé les réponses Whois et les caractéristiques DNS du réseau concerné pour identifier le trafic résidentiel.

Une autre source, susceptible d'être exploitée de manière abusive, sont les machines dans le cloud. Dans [142], les auteurs ont démontré qu'il est simple et peu coûteux d'abuser des fournisseurs de Cloud pour lancer des attaques. Ici, c'est l'aspect marchand qu'il est intéressant de soulever, puisque la location de serveurs auprès des hébergeurs est relativement peu couteuse par rapport au rendement économique que peuvent générer les attaques.

2.5.3 Organisation d'attaquants et sources des attaques

Au niveau de l'analyse des organisations, les organisations malveillantes sont largement traquées afin d'être amenées devant la Justice. De plus, des analyses tentent de remonter vers les organisations qui possèdent des machines qui agissent de façon malveillante afin que ces organisations puissent agir. C'est le cas, par exemple, des travaux menés par [143], les auteurs ont pu remonter aux AS, et donc aux organisations responsables des machines effectuant de l'amplification d'attaques. Ils ont trouvé que 10 AS sont responsables de plus de 55% du trafic généré.

Derrière la réalisation technique des attaques, différents rôles sont répartis entre les membres de l'organisation malveillante[2], il est donc nécessaire de considérer les différentes formes d'organisations dans l'analyse des attaques.

Enfin, les APT sont aussi largement suivis au travers de la littérature. Dans [144], les auteurs ont analysé les cycles de vie d'Emotet, une APT, en s'intéressant aux mécanismes de C2, aux AS d'où les IPs sont originaires ainsi qu'à la répartition géographique des différentes machines de l'infrastructure.

D'autres travaux de suivi des botnets [145, 146] s'attaquent à l'étude des AS en provenance des botnets de type IoT. Les résultats ont été obtenus via l'utilisation de Honeypot afin d'observer les différentes phases de scan qui ont lieu lorsqu'un réseau cherche à s'agrandir. Les travaux montrent aussi que les blocklists d'adresse IP les plus classiques ont des difficultés à suivre les évolutions des infrastructures des botnets.

Une meilleure compréhension de l'aspect social et politique de la sélection des sources malveillantes par les attaquants peut aider à développer de nouvelles solutions. Le contexte politique peut être lié aux attaques, car celles-ci peuvent être utilisées pour faire passer un message. Tous les outils précédents se concentrent principalement sur la partie infrastructurelle du réseau, alors qu'ils ont des dimensions sociales, politiques ou économiques. Dans certains cas, les attaquants peuvent abuser d'une infrastructure spécifique, et il est nécessaire d'évaluer les phénomènes en fonction des propriétaires des adresses IP.

2.6 Positionnement des travaux réalisés dans cette thèse

À la suite de cet état de l'art, nous pouvons constater que la caractérisation des machines, des entités et des personnes est un enjeu majeur pour la description des systèmes qui composent Internet. C'est un enjeu encore plus important lorsque nous nous plaçons dans un contexte de sécurité. Du fait du pseudonymat qui est rendu possible par les propriétés techniques sur

2. https://archives.fbi.gov/archives/news/speeches/the-cyber-threat-whos-doing-what-to-whom

lesquelles ont été construits Internet, nous avons vu que cela laisse la place à de nombreuses activités criminelles. Plus particulièrement, nous pouvons agir sur la protection des infrastructures qui sont menacées chaque jour.

Pour cela, nous pouvons partir d'une donnée technique qui est l'adresse IP. Caractériser les adresses IPs, et plus particulièrement les machines qui se trouvent derrière celles-ci, permet de mieux comprendre les attaques et donc de pouvoir établir des règles et des solutions de protection des infrastructures. Malgré le recours à l'usurpation d'adresses IPs, il existe de nombreuses machines qui servent de réflexion des attaques pour protéger l'identité des attaquants. Néanmoins, si nous pouvons caractériser ces machines, alors, nous pouvons limiter le trafic d'attaque dès ce point. De plus, l'usurpation d'adresses IP est de moins en moins possible, car de plus en plus filtrée par les routeurs des fournisseurs d'accès à Internet[3].

Il existe déjà de nombreux outils qui permettent d'obtenir des informations sur les adresses IPs : géolocalisation, informations sur la machine, services ouverts, propriété. Nous pouvons constater que beaucoup de ces informations révèlent une caractéristique technique des machines. Néanmoins, nous recommandons d'inclure des composantes sociales et organisationnelles dans l'analyse des machines.

En effet, nous avons pu voir que si les attaques s'appuient sur des mécanismes techniques pour atteindre leurs objectifs, elles s'inscrivent généralement dans un contexte sociopolitique-économique. Il faut donc, pour nous, prendre en compte ces spécificités dans l'analyse des politiques de sélection des infrastructures des machines. Lorsqu'elles sont commercialisées, les attaques doivent procurer un rendement économique. Lorsqu'elles ont un objectif politique, alors ici l'attaquant se concentre plutôt sur la capacité à aboutir et l'impact créé.

De nombreuses solutions contre les attaques se concentrent sur la défense de la victime. Nous pensons néanmoins que dans de nombreux cas, une limitation de la source de l'attaque, c'est-à-dire au niveau des ressources employées par l'attaquant, pourrait s'avérer bénéfique. Pour cela, il est nécessaire de comprendre où se situent les machines qui sont à l'origine du trafic d'attaque reçu par la victime. Par où, nous n'entendons pas seulement géographiquement, mais plutôt dans quel type de réseau, administré par qui et avec quel niveau de compétence. Ces informations nous permettraient d'imaginer des solutions à la source qui se rapprocheraient au plus près de la réalité du terrain.

Pour mettre en place de nouvelles solutions au plus près des sources, nous avons besoin de sources de données à analyser qui doivent être réelles, puisqu'on cherche à caractériser la réalité du terrain, et récentes, car comme nous l'avons dit, l'attribution des adresses IP évolue au cours du temps. La collecte des données sera plus longuement discutée dans le chapitre 5.

Dans les chapitres suivants de cette thèse, nous développons une approche pour répondre à cet objectif. Nous commençons par présenter une méthodologie de caractérisation des adresses IPs ainsi qu'une implémentation de cette méthodologie sous la forme d'un outil appelé IPSeen dans le chapitre 3. Par la suite, nous présentons une évaluation de l'outil IPSeen dans le chapitre 4 et montrons les capacités et les limites de notre solution. Enfin, nous proposons une étude de trafic d'attaques (groupe d'attaques, malware, botnet) réalisée avec notre outil. Celle-ci nous permet de caractériser les infrastructures des attaquants et de comparer les infrastructures en fonction des groupes d'attaques et des types d'attaques.

3. https://spoofer.caida.org/spoofage.php

Les travaux présentés dans cette thèse ont pour objectif de caractériser précisément les machines participant à la génération de trafic d’attaque, qu’il s’agisse de trafic direct ou réfléchi, volontaire ou involontaire.

Chapitre 3

Analyse Socio-Organisationnelle des adresses IPs

Notre objectif est de caractériser les organisations qui possèdent les adresses IPs (v4 et v6) afin d'améliorer les connaissances des infrastructures d'attaques. Obtenir une caractérisation des organisations nous permet d'établir des points sensibles dans le réseau Internet qui ont besoin d'un niveau de sécurité supplémentaire. Nous définissons les organisations comme des structures qui relient des personnes dans un but précis. Dans notre cas, les regroupements peuvent varier en taille, allant d'un particulier chez lui à une énorme entreprise multinationale. Comme nous l'avons vu dans l'état de l'art (Chapitre 2), nous pouvons décomposer la caractérisation des machines en deux axes : un axe socio-organisationnel et un axe technique. Notre méthodologie propose d'aborder ces deux axes.

Dans la première partie de ce chapitre, nous commençons par discuter les étiquettes technico-organisationnelles qui sont pertinentes et que nous voulons appliquer aux adresses IPs. Si l'approche technique de l'analyse des réseaux est appropriée, nous revendiquons qu'il est nécessaire de prendre en compte une composante organisationnelle et sociale pour enrichir les analyses et prendre en compte le système dans sa globalité. En effet, les attaques sont un moyen technique utilisé dans un contexte social, qu'il soit commercial, politique ou concurrentiel. Leurs analyses doivent donc contenir un aspect technique et un aspect social. Nous choisissons d'utiliser les adresses IPs pour effectuer la relation entre machine et organisation. Comme nous l'avons vu au travers de l'état de l'art, il s'agit d'une composante identifiante des machines très largement utilisées. Nous présentons également un outil permettant de caractériser les adresses IPs non seulement en se concentrant sur l'aspect technique du réseau auquel elles appartiennent, mais aussi sur l'organisation qui administre l'appareil qui l'utilise. La caractérisation se fait via l'attribution d'étiquettes aux adresses IPs. Les étiquettes sont des qualificatifs destinés à décrire des caractéristiques des réseaux et qui nous permettent de les classifier.

Dans la seconde partie de ce chapitre, nous présentons l'outil IPSeen que nous avons développé pour répondre à cette problématique. Dans le cadre de ce travail, nous nous sommes principalement concentrés sur la caractérisation des IPv4, car les données que nous avons récupérées pour faire nos analyses ne contenait, jusqu'à tardivement que des IPV4. Dans les traces généreusement fournies par Sekoia, 2 IPv6 étaient présentes. Nous avons pu ainsi vérifier que notre outil était fonctionnel pour les IPV6, mais nous ne rentrons pas dans les détails de cette partie. Ainsi, dans la suite de ce manuscrit, nous nous concentrons sur des IPV4, sauf exception explicite.

3.1 Quelles sont les étiquettes pertinentes et réalistes que l'on peut attribuer aux adresses IPs ?

Dans cette section, nous discutons des étiquettes que nous allons attribuer aux adresses IPs. Nous allons dans un premier temps détailler notre approche et définir ce que nous souhaitons décrire. Puis, nous verrons ce qui est réalisable et quels sont les outils qui nous permettront de nous rapprocher de ce que l'on souhaite obtenir. Enfin, nous définirons les étiquettes précisément ainsi que les sources qui nous permettent de les obtenir.

3.1.1 Que souhaitons-nous décrire ?

Avant de commencer à proposer des étiquettes pour réaliser la caractérisation, nous discutons ce que nous souhaitons décrire. Quels types d'organisations souhaitons-nous afficher ? Quels sont leurs points communs ou au contraire leurs singularités ?

Pour cela, nous commençons par faire un état des lieux des différentes organisations qui jouent un rôle sur Internet et qui sont susceptibles de posséder des adresses IP.

3.1.1.1 Quelles sont les organisations présentes sur Internet ?

Afin de mettre en place notre caractérisation socio-organisationnelle, nous devons proposer une classification des différentes organisations présentes sur Internet. Pour cela, nous allons nous appuyer sur des caractérisations proposées par différents acteurs.

Tout d'abord, intéressons-nous à une classification de l'écosystème d'internet proposée par l'Internet Society[147, 148] L'Internet Society est une association de droit américain créée afin de coordonner le développement des réseaux dans le monde. L'Internet Society propose une classification de l'écosystème d'Internet en 6 catégories :

- Utilisateurs : ils se décomposent en individus, société civile, créateurs et prestataires de services, organisations et gouvernements.
- Opérations et services internationaux communs : ils se décomposent en serveurs racines, opérateurs réseaux, points d'échange Internet, créateurs et prestataires de services, domaines (gTLD, ccTLD).
- Élaboration de normes ouvertes : ici, nous trouvons les diverses organisations chargées de l'établissement des standards (W3C, IETF, IRTF, etc.)
- Nommage et adressage : ICANN, Registres Internet Régionaux (RIR), IANA, etc.
- Élaboration des politiques locales, nationales, régionales et mondiales : gouvernements, organisations intergouvernementales, entreprises, Internet Society, etc.
- Éducation : gouvernements, universités, établissement académiques, réseaux nationaux de recherche et d'éducation (NREN), Internet Society, etc.

Cette classification est similaire à celle proposée par CENTR (Counsil of European National Top Level Domain Registries)[149], où les différentes organisations sont cette fois-ci regroupées sous 5 catégories (arènes de la gouvernance de l'internet, Internet Society, utilisateurs, les organisations chargées de l'élaboration des politiques, universités et institutions académiques). CENTR propose aussi un second axe de qualification basé sur les composantes techniques d'Internet décomposé selon les catégories suivantes :

- Infrastructure : on y trouve par exemple les opérateurs, les points d'échange d'Internet, les fournisseurs d'hébergement.

Catégorie	Acteurs
Utilisateurs	Particuliers, Entreprises, Universités, Gouvernement, Société Civile
Gestion des politiques, garants du bon fonctionnement	États, Organisme de Normalisation Techniques, Organisme de Régulation, CA, IANA, RIRs, TLD
Fournisseurs de services	Auteurs de Logiciel, Opérateurs Réseaux, Hébergeurs, CDN

TABLE 3.1 – Catégorisation des différents acteurs d'Internet

— Standards et protocoles : IEEE, IETF, IRTF, W3C...
— Identifiants techniques publics : contient les différents standards et protocoles.
— Communauté de numérotation : entité chargée de l'attribution des IPs (RIRs).
— Communauté de noms : entité chargée du nommage comme l'ICANN et les registres de noms de domaine.

Ces propositions de classification sont très larges et prennent le système Internet dans sa globalité. De plus, certaines organisations peuvent se trouver dans deux catégories différentes, ce qui ne permet pas de proposer une classification exclusive.

Nous proposons un regroupement de ces différents acteurs en 3 catégories distinctes et présentées dans la Table 3.1.

Tout d'abord, nous avons la catégorie des utilisateurs. Nous regroupons sous cet ensemble les acteurs qui utilisent les infrastructures, qu'il s'agisse dans un contexte de travail, de loisir, de recherche ou de formation. Nous définissons une seconde catégorie qui comprend les entités qui ont un pouvoir politique sur les infrastructures. Nous regroupons ici tous les organismes responsables de l'établissement des normes, de l'attribution des ressources ainsi que les acteurs responsables des politiques de développement des infrastructures. Enfin, pour finir, nous regroupons tous les acteurs chargés du maintien des infrastructures sous une unique catégorie : les fournisseurs de services. Nous intégrons aussi bien les fournisseurs de logiciels que les opérateurs réseaux.

Parmi les différents acteurs, certains sont plutôt des acteurs responsables du transit des paquets sur le réseau et non des points finaux. D'autres acteurs sont plutôt garants du bon fonctionnement d'Internet. Enfin, les acteurs tels que les auteurs de logiciels sont responsables de l'utilisation qui est faite des réseaux via leurs applications.

3.1.1.2 Quels types d'acteurs souhaitons-nous décrire ?

Parmi les différents acteurs évoqués précédemment, nous choisissons d'étudier les utilisateurs. Nous cherchons à définir qui possède quels réseaux et par conséquent à un pouvoir d'action sur les machines qui le composent. Ce sont ces utilisateurs à qui sont attribuées les adresses IPs. Il va de soi que certains des acteurs de la catégorie des fournisseurs de services doivent aussi être caractérisés. Cependant, même si ces acteurs possèdent des adresses IPs, ils les mettent à disposition des utilisateurs.

Parmi les utilisateurs, il est pertinent de pouvoir différencier les profils que nous avons. Imaginons que nous sommes en train d'envisager le déploiement d'une solution de sécurité aux niveaux des différents acteurs. Le type de solution ne devrait pas être le même si on se trouve chez une grande entreprise avec un service DSI compétant ou chez M. Michu. En effet, une solution chez M. Michu devrait être complètement autonome, alors qu'une grande entreprise

Acteurs	Description
M. Michu	novice, abonné à un FAI, n'héberge aucun service
Mme Telecom	services hébergé chez soi, autonomie, ports ouverts
Gouvernements	services des états
Hébergeurs	fournissent un service à un utilisateur tiers mais sont administrateur des machines
Université	réseaux universitaire et de recherches
Associations	associations de technologie (hébergement) ou autre
Entreprises	PME à grande entreprise nationale, derrière FAI pro ou non
Multinationales	Entreprises répartie sur plusieurs territoires géographiques

TABLE 3.2 – Catégorisation visée des différents acteurs finaux

a un service DSI à disposition avec des personnes formées pour suivre le déploiement d'une solution et en capacité d'agir en réaction et de collaborer avec la solution.

Nous allons donc utiliser les différents acteurs de la catégorie Utilisateurs. Au travers des classifications proposées précédemment par CENTR et l'Internet Society, nous avions vu qu'il existait un ensemble de sous-catégories pour les utilisateurs. Nous choisissons de catégoriser les utilisateurs en fonction du type d'utilisation qu'ils font des réseaux. Ce type de classification a été proposé auparavant dans d'autres travaux. Dans [150], les auteurs ont proposé une classification des AS en fonction du type d'entreprise.

Parmi les acteurs possédant les terminaux finaux, il est intéressant de diviser certains acteurs en sous-catégories. C'est le cas pour les entreprises. Il y a des situations où les acteurs sont à la fois administrateurs et utilisateurs, tandis que d'autres sont uniquement administrateurs. Certaines entreprises utilisent leur propre matériel informatique pour leurs activités, tandis que d'autres fournissent des machines louées pour être utilisées par d'autres. Dans cette autre situation, l'entreprise possède le matériel, mais elle a moins de contrôle sur les utilisations qui en seront réalisées. Cette catégorie comprend les fournisseurs d'hébergement.

Une autre catégorie intéressante à prendre en compte est celle des clients des FAI (fournisseurs d'accès à Internet). Cette catégorie regroupe diverses catégories de personnes. Tout d'abord, M. Michu est connecté à Internet grâce à sa box à la maison. M. Michu ne possède aucune compétence en réseau et se sert simplement de son ordinateur et de son téléphone. D'un autre côté, nous avons Mme Telecom, qui possède des compétences en informatique/réseau et qui propose ses propres services chez elle.Dans cette situation, il s'agit d'un cas technique où des ports ouverts seront observés sur les machines de Mme Télécom. Enfin, dans la catégorie des clients FAI, nous avons aussi des entreprises. On peut avoir l'entreprise A, une petite entreprise d'artisanat qui a un abonnement professionnel chez un FAI grand public qui propose des abonnements pour les professionnels. Mais aussi l'entreprise G, grande entreprise qui est cliente chez un FAI qui offre seulement des abonnements aux entreprises.

La division entre entreprises de taille moyenne et grandes entreprises multinationales est également intéressante. En effet, les grandes entreprises multinationales sont réparties sur plusieurs zones géographiques mais soumises à des lois différentes selon la localisation de leurs différentes filiales.

Nous proposons une nouvelle répartition en catégories d'acteurs en utilisant les points précédemment discutés. Les nouvelles catégories sont listées dans le tableau 3.2.

3.1.1.3 Comment différencier les acteurs ?

Après l'établissement de toutes ces catégories, nous avons questionné sur les différentes typologies des réseaux de ces acteurs ainsi que les critères qui pourraient les distinguer les uns des autres. Les critères sont définis comme une composante technique ou organisationnelle d'un acteur ou de son infrastructure. La réalité de chaque critère peut varier d'un acteur à l'autre. Tout d'abord, nous n'avons pas pris en compte la disponibilité et l'accessibilité des informations. Nous avons défini cette répartition et cette analyse dans le cas où nous serions omniscients. Cela nous permet d'envisager dans un premier temps les différentes informations qui sont primordiale pour distinguer les différents acteurs.

Tout d'abord, nous avons commencé par essayer de différencier les infrastructures techniques les unes des autres. Différents travaux [151], [152] proposent une analyse des caractéristiques du trafic que l'on peut observer depuis les réseaux d'entreprises. D'autres travaux se concentrent sur la caractérisation des réseaux résidentiels[153, 154]. Plus récemment, les travaux suivants [155] ont comparé le changement de trafic résidentiel suite au confinement. Parmi ces travaux, nous pouvons identifier différents critères qui nous permettraient de différencier nos types de réseaux. Nous avons tout d'abord la répartition et le type de trafic observé via les différents protocoles employés pour gérer le trafic. Ici, par services, on entend les ports ouverts et les services disponibles derrière ces ports. Par défaut, les ports sont inaccessibles depuis une machine extérieure chez un particulier. Ce n'est pas forcément le cas pour une entreprise où les machines peuvent être utilisées pour mettre en place des services accessibles depuis l'extérieur du réseau. La bande passante varie aussi en fonction des installations, mais aussi des services fournis par les opérateurs. Une mesure du débit peut donc nous permettre de différencier les abonnés des FAI des clients haut débits.

Il est également possible que le choix des appareils ou des systèmes d'exploitation utilisés par le client final soit également lié au type de réseau. Le matériel grand public n'est pas similaire au matériel utilisé par des entreprises. Il en va de même pour ce qui est du matériel réseau.

L'analyse en temps du trafic permet également de différencier nos acteurs. Le trafic résidentiel en semaine est moindre qu'en week-end, et le trafic en journée est plus important que le trafic la nuit. On peut donc vouloir analyser les différents modèles de trafics. En entreprise, il est assez cohérent d'observer du trafic les jours de semaine sur des horaires classiques de travail, alors que ce n'est pas forcément le cas pour les autres catégories. Toujours en observant le trafic, un autre critère est l'asymétrie du trafic. Dans le foyer, on observera en majorité du volume du trafic descendant, alors que ce n'est pas le cas pour certaines des autres catégories. Le nombre de machines sur un réseau nous semble judicieux puisque nous avons l'intuition qu'un particulier chez lui [156], même avec l'arrivée des objets connectés, a moins de machines en moyenne qu'une entreprise de taille moyenne ou une grande entreprise. Enfin, une dernière information qui peut nous mettre sur la voie d'une catégorie plutôt qu'une autre est l'observation des noms de domaines qui sont associés aux adresses IPs d'un sous-réseau. Dans un sous-réseau d'hébergement, on trouvera en moyenne un plus grand nombre de sous-domaines alors que, pour des FAIs grand public, on ne trouvera en moyenne que le nom de domaine du FAI. Les noms de domaine peuvent aussi nous apporter des informations directes quant au type d'entité. Par exemple, on pourrait observer des adresses *.gouv* pour les gouvernements ou *.edu* pour les réseaux d'éducation.

Le tableau 3.3 met en avant la classification que nous proposons grâce aux critères précédemment évoqués. La classification est basée sur des hypothèses et se focalise principalement sur des critères techniques. Les critères de cellules grisées ne nous semblent pas appropriés pour le type d'acteur mentionné car plus difficiles à qualifier ou quantifier. De par l'analyse des critères techniques, nous pouvons voir que certains acteurs se différencient très bien lorsqu'une de leurs spécificités leur est unique. C'est le cas, par exemple, des noms de domaine. Les services entrants nous permettent de différencier les particuliers des autres catégories. Les services n'étant pas forcément liés à une activité, nous ne pouvons pas différencier nos acteurs en fonction des services proposés. La présence des services sortants en revanche ne nous permet absolument pas de différencier nos catégories.

Néanmoins, en observant le tableau de classification, on peut voir que certains des acteurs que nous souhaitons différencier se rapprochent énormément. Une association ou une petite entreprise vont par exemple avoir de nombreux points communs. Mais on pourra les différencier seulement au niveau des informations organisationnelles et non techniques. C'est pour cela qu'il nous faut aussi inclure des informations organisationnelles. La forme légale peut nous aider à les différencier. Un autre exemple, le nombre d'employés, peut nous aider à différencier une petite entreprise d'une grande. Le domaine d'activité d'une entreprise peut nous aider à différencier un hébergeur d'une entreprise non IT. Il est donc nécessaire de se procurer des informations organisationnelles en plus des informations techniques.

3.1.2 Quelles informations peut-on obtenir sur les machines et les organisations administrant les machines à partir de leurs adresses IPs ?

Nous avons décrit dans la partie précédente les différentes catégories que nous souhaitons obtenir ainsi que des critères technico-organisationnels qui nous permettent d'identifier la catégorie à laquelle une adresse IP appartient. Nous avons aussi abordé dans l'état de l'art 2.3 des outils qui nous permettent d'affecter des étiquettes aux adresses IPs. Dans cette section, nous confrontons notre classification à la réalité des informations que nous pouvons obtenir en partant d'une adresse IP. Notre objectif est de pouvoir classifier notre IP en obtenant les critères identifiés sur la Table 3.3. Nous discutons des critères de sélection présentés en les associant avec les sources d'informations qui nous permettent de les obtenir ou de s'en approcher au plus près.

3.1.2.1 Ports et services

Tout d'abord, intéressons-nous aux ports et aux services qui se trouvent derrière les ports. Comme vu dans l'état de l'art, de nombreux crawleurs de réseaux permettent aujourd'hui de découvrir les ports et/ou services ouverts derrière une adresse IP. Nous avons présenté et comparé 3 crawlers : Censys [104], Shodan [105] et Onyphe [106]. Nous avons vu dans l'état de l'art que ce sont trois sources pertinentes et fiables quant à l'objectif que nous voulons atteindre : connaître les ports et/ou services accessibles derrière une adresse IP.

Onyphe est une solution très complète, qui, comme les deux solutions évoquées précédemment, permet d'identifier les ports et les services accessibles via une adresse IP. Onyphe permet aussi l'accès à des fonctionnalités supplémentaires comme l'analyse des noms de domaine et des organisations. Nous favoriserons donc dans la suite l'utilisation d'Onyphe. De plus, nous avons pu bénéficier d'une licence Onyphe.

	M. Michu	*Mme Telecom*	*Gouvernements*	*Hébergeur*	*Universités*	*Associations*	*Entreprises*	*Multinationales*
Nombre de machine sur le sous-réseau	une dizaine	une dizaine		très important	entre 100 et 10000			>très important
Présence de services entrant	aucun	possible mais peu	oui	oui	oui		oui	oui
Présence de services sortants	oui	oui	oui	oui	oui	oui	oui	oui
Type d'équipements	routeur FAI	routeur FAI	équipement pro	équipement pro	équipement pro	routeur FAI	équipement pro et/ou routeur FAI	équipement pro
Modèle de trafic	plus important le week-end et en journée	plus important le week-end et en journée					journée de semaine	journée de semaine
Asymétrie du trafic	asymétrie forte	asymétrie		asymétrie très légère	asymétrie			
Noms de domaines	aucun résultat ou domaine du FAI		potentiel domaine .gouv	nombreux résultats	potentiel domaine : .edu		domaine de l'entreprise (ou alias ?)	domaine de l'entreprise(ou alias ?)
Débit	débit fibre moyen proposé par FAI	débit fibre moyen proposé par FAI		haut débit		débit fibre moyen proposé par FAI	haut débit	haut débit

TABLE 3.3 – Classification des différents acteurs en fonction des critères

3.1.2.2 Nombre de machines sur le réseau et types de machines

À présent, raisonnons sur l'obtention du nombre de machines sur un réseau. Ici, la question est plus complexe. Tout d'abord, nous pourrions avoir l'intuition que le nombre d'adresses IPs d'un sous-réseau équivaut à peu près au nombre de machines. Or, avec la démocratisation des réseaux natés et l'inégalité des affectations des adresses IPs par région, ce n'est pas le cas. En effet, si certaines régions comme l'Amérique du Nord possèdent suffisamment d'adresses IPs pour pouvoir attribuer une adresse IP à chaque machine du réseau, ce n'est pas le cas, par exemple pour l'Asie où derrière une adresse IPv4 se trouvent un grand nombre de machines. Par conséquent, nous avons besoin d'une autre estimation. Comme derrière un grand nombre de machines se trouvent des êtres humains, nous proposons de nous concentrer sur l'obtention de la taille humaine de l'organisation qui possède l'IP. Cela veut dire obtenir le nombre d'employés dans le cadre d'une entreprise, obtenir le nombre d'abonnés pour un FAI, ou obtenir le nombre d'élèves pour une université. Cela ne nous permet pas d'obtenir une information précise, mais un ordre de grandeur quant à la taille de l'organisation. L'obtention d'un ordre de grandeur est suffisante pour notre catégorisation. Bien sûr, ces informations peuvent être enrichies avec la taille des sous-réseaux qui appartiennent à une entreprise. En ce qui concerne le type d'équipement, Onyphe peut nous fournir des informations pour certaines IP sur le type de machine ainsi que la marque de la machine.

3.1.2.3 Noms de domaines

Pour les noms de domaines associés à une IP, le reverse DNS nous fournit les sites et les noms de domaines qui sont liés à une adresse IP. Nous pouvons donc observer les types ainsi que le nombre de noms de domaine derrière une IP.

3.1.2.4 Analyses du trafic

En ce qui concerne les analyses de trafic, il n'est pas possible pour nous d'obtenir ces informations. En effet, il faudrait avoir un point dans le réseau où nous pourrions observer tout le trafic qui provient d'une adresse IP. Or, nous n'avons pas accès à ces informations. Bien entendu, lorsque l'on parle ici de trafic, on s'intéresse seulement au volume échangé ainsi qu'aux adresses IPs source et destination et non au contenu du paquet IP. Par conséquent, il nous faut d'autres informations pour pouvoir départager les différentes catégories.

3.1.2.5 Obtentions d'informations organisationnelles

Nous avons vu précédemment que nous avons besoin d'informations sociales pour pouvoir départager nos catégories. Or, nous avons besoin d'une source pour obtenir ces informations. Via l'obtention du nom de l'entreprise qui possède l'adresse IP, nous pouvons facilement obtenir des informations à son propos. En effet, une simple recherche du site Web de l'organisation peut nous donner des informations sur le type d'organisation, son potentiel domaine d'activité ou encore sa taille. Malheureusement, ce n'est pas la solution la plus facile à mettre en place de manière automatisée et autonome. En effet, consulter les pages Web de chaque entité prend du temps et est difficilement automatisable, car la structure des informations est différente pour chacune d'entre elles. Nous avons donc besoin d'une source de données facilement requêtable et interprétable par une machine. Nous avons analysé plusieurs sources d'informations :

Kompass [157], LinkedIn[158] et Wikidata[159]. Ces trois sources d'informations présentent l'avantage d'avoir une API requétable. Discutons de la première source. Kompass est un catalogue d'informations requétable qui décrit les entreprises et associations d'un pays. Kompass contient les informations que nous souhaitons nous procurer : la localisation d'une entreprise, la taille de celle-ci ainsi que son domaine d'activité. En revanche, Kompass ne possède pas des informations sur tous les pays. De plus, Kompass est davantage tournée vers les filiales et les bureaux locaux d'une entreprise. Par exemple, pour la multinationale Amazon, on trouve les différents locaux, mais pas un nombre de salariés qui correspond à l'immense taille de l'entreprise. Une seconde source de données que nous avons envisagée est LinkedIn. LinkedIn est un réseau socioprofessionnel. On y trouve des pages pour les différentes entreprises ainsi que des offres d'emploi. Si LinkedIn possède de nombreuses informations concernant les entreprises, c'est moins le cas pour les écoles, universités, associations. Enfin, notre dernière source d'information est Wikidata. Wikidata est une base de données de la fondation Wikimedia, qui sert notamment de support à Wikipédia. Wikidata contient, comme une encyclopédie, des connaissances sur de nombreux sujets divers et variés. Parmi ces sujets, on retrouve des entreprises, des associations, des écoles et des universités, des pays et des gouvernements, etc. Les résultats de Wikidata sont accessibles via une API et des requêtes en langage SparQL. Wikidata est librement modifiable, c'est-à-dire que tout utilisateur peut contribuer à enrichir la base de données, y compris nous. Wikidata est très complet et contient des informations sur tous les types d'entités, nous allons donc privilégier cette source d'information par la suite. Néanmoins, pour obtenir les informations concernant une entité, il nous faut connaître le nom de celle-ci si la plage d'adresse IP qu'elle possède n'est pas renseignée dans Wikidata.

Afin d'obtenir le nom d'une entité responsable de l'adresse IP, nous pouvons interroger les différents RIRs (Regional Internet Registeries). Pour cela, il existe deux protocoles qui permettent d'interroger ces bases de données : Whois[51] et RDAP[116, 117, 118]. Whois est largement connu et étudié, et a même permis, comme on l'a vu précédemment dans l'état de l'art 2.3, d'être utilisé pour différencier des adresses résidentielles d'autres adresses. RDAP est une nouvelle version du protocole Whois. RDAP est mieux structuré et plus facilement analysable par une machine, car les informations sont disponibles sous format JSON. RDAP tend à remplacer Whois. Parmi les informations que nous renvoie RDAP sur une adresse IP, nous pouvons généralement retrouver le nom de l'entité qui possède l'adresse IP. Ce n'est bien sûr pas le cas pour les IP résidentielles, puisque les IPs ne sont pas déclarées directement aux noms des personnes les utilisant, mais nous pouvons savoir quels FAI gèrent les adresses IP et donc déduire qu'il s'agit d'un réseau grand public. Il arrive même parfois que des termes comme "résidentiel" apparaissent dans les résultats de RDAP.

Nous sommes donc, en conclusion, en capacité d'obtenir des informations sur les organisations qui se trouvent derrière les adresses IPs en obtenant via RDAP le nom de l'entité qui a renseigné ces informations au moment de l'attribution de l'adresse par les RIRs.

3.1.3 Définition des étiquettes technico-organisationelles que nous attribuons aux adresses IP

Nous souhaitons à présent mettre en relation la classification que nous avons obtenue à la suite de la première partie et les outils qui nous ont permis d'obtenir des informations sur les adresses IPs que nous venons d'étudier.

Nous définissons à présent des étiquettes basées sur les critères que nous avons définis qui nous permettent de distinguer les différents acteurs. Ces étiquettes constituent des qualificatifs que nous allons par la suite associer à des adresses IPs, moyennant un outil basé sur les sources d'informations identifiées.

Étant donné ce que nous souhaitons décrire et les données récupérables, nous proposons 6 étiquettes pour qualifier une IP. Ils se décomposent en 2 axes complémentaires : un axe technique et un axe organisationnel. L'axe technique a pour objectif de décrire précisément les caractéristiques de l'infrastructure à laquelle est attribuée l'adresse IP. Décrire l'infrastructure technique a plusieurs objectifs :

— connaître les services qui sont accessibles depuis l'extérieur sur ce réseau, et donc pouvoir identifier des vulnérabilités,
— décrire la répartition géographique et l'étendue du réseau, dans le but d'estimer si celui-ci correspond à différentes législations et si les sources de trafic peuvent être réparties entre plusieurs pays,
— estimer la taille d'une infrastructure permet d'évaluer le potentiel de menace lié à cette infrastructure.

De l'autre côté, l'axe organisationnel a pour objectif de décrire l'organisation qui gère les appareils qui possèdent les adresses IPs. Obtenir des informations sur l'organisation permet de :

— connaître le type de réseau auquel appartiennent les adresses IPs, et donc connaître de quel type d'utilisateur final, il s'agit. Nous faisons l'hypothèse ici que tous les utilisateurs d'une structure sont du même type,
— connaître le type d'activité de l'organisation, nous permet de nous apporter des réponses quant au niveau de connaissance en sécurité potentiellement ainsi que d'identifier des domaines d'activités où les machines pourraient être utilisées par des acteurs tiers,
— connaître la taille humaine de l'entité afin d'obtenir des informations sur comment celle-ci peut fonctionner.

Comme énoncé précédemment, ces deux axes sont complémentaires et l'association de plusieurs informations peut répondre à des analyses de cas particuliers. Par exemple, si l'on s'intéresse au profil de personnes qui se connectent sur un site web, on a besoin de connaître le type de réseau auquel celle-ci est reliée ainsi que potentiellement sa position géographique. Dans un contexte de sécurité, si l'on cherche à limiter un certain type d'attaque dès la source, on a besoin de savoir où il faut déployer notre solution, c'est-à-dire dans quel type de réseau, mais aussi de savoir quel type de solution déployer. Pour cela, des informations comme le domaine d'activité ou la taille de l'organisation peuvent nous aider à mieux définir quelles seront les spécificités qui faciliteront le déploiement de celle-ci (automatique, lieux, taille).

Les étiquettes que nous avons choisies d'attribuer aux adresses IPs sont les suivantes :

- Services ouverts
- Répartition géographique
- Taille du réseau

} étiquettes d'infrastructures

- Type de réseau
- Activité
- Taille humaine

} étiquettes organisationnels

Tout d'abord, nous regroupons les étiquettes décrivant une organisation. Nous voulons ici différencier une entreprise d'un client résidentiel ou d'un gouvernement. Par informations sur l'organisation, nous entendons toutes les étiquettes qui peuvent nous aider à définir l'organisation (qui peut être un groupe de personnes ou une personne seule) qui possède l'appareil correspondant à une adresse IP. En particulier, nous prendrons en considération le **type** de l'organisation, le **taille** de l'organisation ainsi que le **domaine d'activité** de l'organisation. L'étiquette de type décrira la nature de l'organisation : entreprise publique, ONG, université ou abonné d'un fournisseur d'accès résidentiel. L'étiquette du domaine d'activité concerne les entreprises et peut être utilisée pour relier l'activité à un comportement malveillant ou à des vulnérabilités. Enfin, l'étiquette de taille sera une agrégation d'informations donnant des renseignements sur la taille de l'entité. Il peut s'agir du nombre d'abonnés d'un fournisseur d'accès à Internet ou du nombre d'employés d'une entreprise. Ces informations sont récupérées à l'aide de RDAP et de Wikidata. Grâce à toutes ces informations, nous sommes en mesure de définir précisément le type de réseau auquel appartient un dispositif malveillant ou compromis.

L'objectif des étiquettes techniques est de pouvoir signaler les réseaux en fonction des particularités de leurs appareils et des risques qu'ils peuvent comporter en fonction de leurs services, ainsi que d'identifier si cet appareil peut communiquer avec nous. Nous regroupons les informations qui caractérisent principalement les appareils et les services sur le réseau sous le nom d'étiquettes techniques. Avec les étiquettes techniques, nous voulons décrire comment le réseau fonctionne, quels types d'appareils composent le réseau, quels services sont offerts par le réseau et comment le réseau est réparti géographiquement dans une région ou dans le monde pour une entreprise multinationale, par exemple. Sous l'aspect technique, nous avons défini trois autres étiquettes : la **localisation géographique**, la **taille du réseau** et les **services et types d'appareils** associés au réseau. La localisation géographique donne des informations sur l'endroit où se trouvent les appareils qui tentent de communiquer avec vous, ainsi que sur la ramification de l'infrastructure du réseau. La taille du réseau est liée à la taille de l'organisation, mais cette fois, nous essayons d'évaluer la taille potentielle de la menace, car si un appareil peut envoyer du trafic, cela peut également être le cas pour d'autres appareils. Sous cette étiquette, nous concaténons les informations relatives au nombre d'adresses IP détenues par l'organisation. Enfin, les services correspondent aux ports ouverts qui peuvent être observés ou aux services auxquels il est possible d'accéder sur l'IP.

3.2 IPSeen : un outil de caractérisation des adresses IPs

Dans cette section, nous présentons notre outil IPSeen. La figure 3.1 propose une vue d'ensemble de notre algorithme. IPSeen est un outil qui permet d'affecter les différents labels que nous avons définis dans la section 3.1.3. Tout d'abord, comme nous l'avons vu au travers de l'état

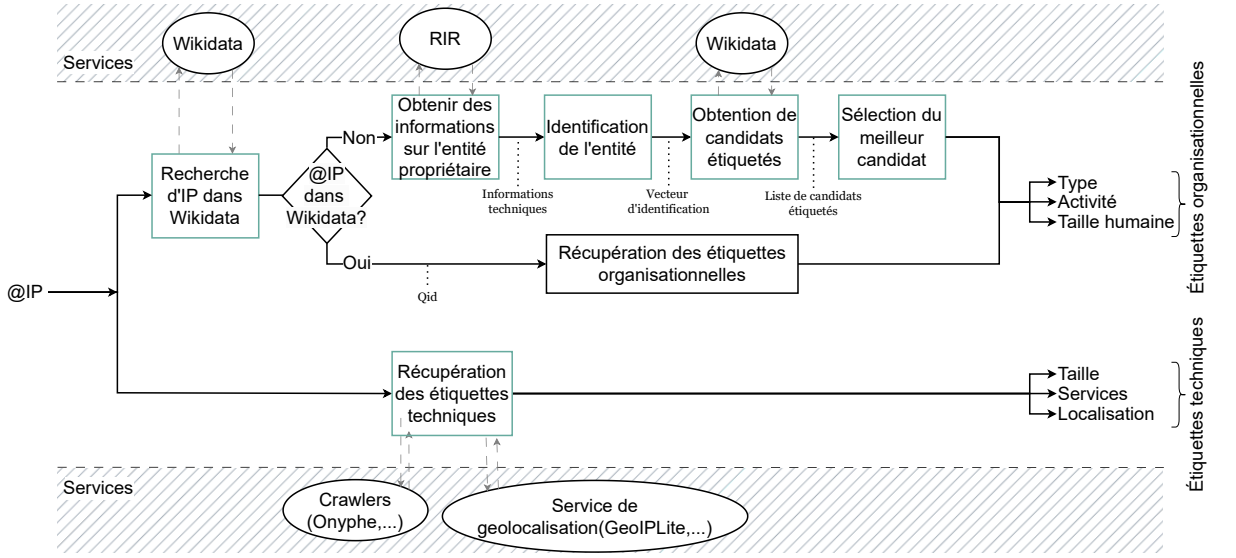


FIGURE 3.1 – Vue d'ensemble de l'algorithme d'affectation des étiquettes

de l'art, il existe de nombreux outils permettant d'obtenir des informations techniques à propos des adresses IPs. Afin d'obtenir nos étiquettes, nous avons choisi les sources d'informations suivantes :

- GeoIPLite : base de données de localisation des adresses IP. On se concentrera seulement sur la localisation au niveau des pays.
- Onyphe : pour connaître les services ouverts. Nous sommes en contact avec le fondateur de la solution. Celle-ci est plus complète que Shodan ou Censys. Néanmoins, notre solution peut être adaptée avec les autres outils. Pour ce qui concerne l'obtention des étiquettes de services, il est nécessaire d'acheter ou d'obtenir un abonnement auprès des services proposés.
- RDAP : protocole qui va nous permettre d'interroger les bases des différents RIRs.
- Wikidata : base de données pour obtenir les informations sur les organisations.

Chacune de ses sous-parties est décrite dans les sections suivantes. Nous commençons par traiter l'obtention des étiquettes techniques via des outils connus. Nous détaillons aussi notre contribution pour l'affectation des étiquettes organisationnelles. Cette contribution est basée sur les données de deux outils externes publics : RDAP et Wikidata. L'obtention des étiquettes organisationnelles est décomposée en plusieurs étapes. Nous allons en premier vérifier si l'adresse IP que nous souhaitons qualifier est incluse dans Wikidata. Sinon, alors nous allons devoir identifier l'entité propriétaire de l'adresse IP afin d'effectuer une recherche par nom d'entité sur Wikidata. Par la suite, nous obtenons une liste de candidats étiquetés que nous devons inspecter pour trouver la bonne entité. Pour certaines des étapes, plusieurs stratégies ont été envisagées. Nous les distinguons par la version d'IPSeen : **IPSeen Fast** et **IPSeen Accurate**. Les deux versions de notre outils IPSeen sont accessible ici [160].

Tout au long de cette section, nous illustrerons les différents blocs algorithmiques à l'aide d'un exemple en utilisant l'adresse IP : 134.214.58.24. Nous avons choisi cette adresse IP puisqu'il s'agit d'une adresse IP de notre campus à l'INSA et, par conséquent, nous avons connaissance des organisations ainsi que des caractéristiques techniques qui sont associées

à cette IP. Il s'agit d'une adresse utilisée par un poste. Les exemples seront identifiés par la couleur verte et sont représentés sur la vue d'ensemble (Figure 3.1) par les blocs verts.

3.2.1 Récupérations des étiquettes techniques

Dans un premier temps, nous allons décrire notre méthodologie pour récupérer des étiquettes techniques. Comme nous l'avons vu auparavant, de nombreux outils de l'état de l'art (Section 2.3) permettent de qualifier des adresses IPs, principalement en ce qui concerne les caractéristiques techniques de la machine qui l'utilise.

Comme indiqué dans la section précédente, certaines de nos étiquettes peuvent être directement agrégées à partir d'outils existants. Les étiquettes que nous pouvons attribuer en regroupant les résultats d'outils existants sont les suivantes : services ouverts, localisation géographique et taille du réseau.

Tout d'abord, les services ouverts peuvent aider à étudier la corrélation entre les services et les attaques. Comme nous le savons, certains services tels que les résolveurs DNS peuvent être utilisés de manière abusive pour générer du trafic d'attaque. De plus, connaître les services qui sont derrière les IPs permet d'avoir une vue plus générale du trafic qu'on peut voir depuis ces IPs, trafic qui peut être lié aux services proposés.

Pour obtenir les ports et les services ouverts derrière une adresse IP, nous pouvons utiliser des scanneurs de réseau (net crawlers). Nous pourrions scanner les ports nous-mêmes. Mais c'est un travail massif et des outils spécialisés sont disponibles. Les crawlers de réseau n'ont besoin que d'adresses IP comme entrées et renvoient, en fonction du crawler, une liste de ports ouverts ou une liste de services derrière les ports ouverts.

Censys, Shodan ou Onyphe sont des crawlers bien connus et notre outil peut être adapté à chacun de ces services. Censys offre une licence gratuite à des fins de recherche. Nous avons décidé d'utiliser Onyphe, car cet outil est un peu plus complet et fournit d'autres fonctionnalités telles que la localisation ou les informations DNS. Il fournit également une liste des ports ouverts et des services derrière ces ports, même s'ils se trouvent derrière un port non habituel. De plus, nous avons pu bénéficier d'une licence offerte afin de réaliser nos travaux de recherche.

En ce qui concerne l'acquisition d'informations géographiques, nous pouvons utiliser plusieurs sources. La première source d'information est déclarée par l'entreprise elle-même lorsqu'elle acquiert une plage d'adresses IP. Les informations sont stockées dans les différentes bases de données RIR. Par ailleurs, la géolocalisation des adresses IP est devenue un outil courant pour l'analyse de la sécurité, et il existe plusieurs bases de données de géolocalisation. Dans notre méthodologie, nous avons décidé de nous concentrer uniquement sur le pays associé à une adresse IP. En effet, afin d'évaluer la taille d'une infrastructure, déterminer si celle-ci est repartie sur plusieurs pays, avec potentiellement différentes législations en place, nous parait suffisante. Nous pourrions envisager plus de précision en s'intéressant au niveau des villes et voir s'il y a par exemple plusieurs sites dans un pays. Néanmoins, avec des informations comme le nombre d'employés, nous pouvons déjà différencier une petite entreprise d'une grande entreprise qui aurait plusieurs sites. Avec ce niveau de précision, la plupart des outils ont le même niveau de confiance concernant le pays d'origine, comme nous l'avons vu dans le chapitre d'état de l'art. Nous avons décidé d'utiliser la base de données Maxmind Geolite, car elle est d'accès gratuit, a fait l'objet d'un examen approfondi dans la littérature et fournit également des informations sur l'entité propriétaire de l'adresse IP. Onyphe fournit également un service de localisation. Le fait de disposer de plusieurs sources peut permettre de donner un

certain niveau de confiance à l'étiquette de localisation liée au croisement et à la validation des informations.

Enfin, en ce qui concerne la taille d'un réseau, nous avons décidé de fournir une approximation pour donner une idée de ce qui est disponible. Pour cela, nous estimons d'abord le nombre d'adresses IP disponibles pour une organisation. Pour ce faire, nous utilisons RDAP en obtenant toutes les plages d'adresses IP correspondant à une entité. Une autre façon d'estimer le nombre d'appareils et de le comparer d'une entité à l'autre consiste à obtenir le nombre d'employés d'une entreprise ainsi que le nombre d'abonnés d'un fournisseur d'accès à Internet. Pour ce faire, il faut obtenir des informations sur l'organisation qui se cache derrière une adresse IP. Ce point est abordé dans la section suivante. Nous avons choisi un ensemble d'outils connus de la littérature pour répondre à notre objectif de caractérisation de l'infrastructure technique qui est liée à une adresse IP.

En ce qui concerne notre exemple, nous obtenons pour notre adresse IP les labels présentés dans l'Exemple 1. Nous pouvons voir qu'il n'y a pas de résultat trouvé pour les services ([None]). La taille correspond à la taille du bloc IP.

Exemple 1 Étiquettes obtenues pour l'adresse IP 134.214.58.24

```
services : [None],
geographic_labels : [FR],
size : [64000]
```

3.2.2 Recherche d'IP dans Wikidata

Wikidata contient un champ, nommé *IPv4 routing prefix*, dédié aux adresses IPv4. Dans ce champ, sont déclarées les adresses IPv4 associées avec leurs entités. Par exemple, des ISP, des universités ou encore des sociétés dont le domaine d'activité est l'hébergement, peuvent avoir ce champ. Nous allons donc dans un premier temps vérifier si l'adresse IP recherchée est déjà présente dans Wikidata.

Nous nous concentrons sur les adresses IP de version 4, puisque, comme nous l'avons énoncé en introduction de ce Chapitre, les captures de trafic que nous avons obtenues et que nous analysons dans le Chapitre 5 contiennent quasiment exclusivement des adresses IPs de type 4.

Par ailleurs, la qualité des IPs contenues dans Wikidata est évaluée en Section 4.1.

Pour cela, et dans le but d'améliorer les performances, nous stockons localement les adresses IPs contenues dans Wikidata. Une requête SPARQL obtenant toutes les adresses IP est faite, et les résultats sont stockés dans un fichier CSV. Ce fichier CSV possède une durée de validité de deux semaines. En effet, les déclarations d'adresses IP ne sont pas fixées pour toujours, il nous faut donc pouvoir mettre à jour les modifications qui sont apportées dans Wikidata.

Nous vérifions ensuite si l'adresse IP que nous cherchons est contenue dans l'un des blocs de sous-réseaux de Wikidata. Pour cela, tous les blocs du fichier sont parcourus. Si nous trouvons un bloc qui englobe l'IP, alors le Qid (identifiant unique qui désigne un résultat de Wikidata) associé à ce bloc est récupéré, et une requête SPARQL est effectuée pour obtenir les étiquettes organisationnelles précédemment définies. L'Exemple 2 illustre les labels que nous obtenons lorsque nous utilisons l'adresse IP qui correspond au nom de domaine de *Wikidata.org*.

Nous pouvons voir que nous récupérons l'entièreté des labels sociaux recherchés et que nous obtenons bien une caractérisation qui correspond à Wikimédia.

Bien que cette stratégie puisse sembler naïve, nous montrons dans le Chapitre 4 que nous pouvons avoir confiance dans les résultats qui se trouvent dans Wikidata. Si néanmoins l'adresse IP n'est pas déclarée dans Wikidata, il nous faut identifier l'entité qui possède l'adresse IP par un autre moyen pour pouvoir faire une recherche textuelle sur Wikidata.

C'est ce cas que nous allons traiter à présent et qui constitue la première branche sur notre vue générale de l'algorithme (Figure 3.1). Cette branche se décompose en 4 étapes successives que nous allons à présent décrire.

Exemple 2 Résultat pour l'adresse IP de Wikidata.org

L'adresse IP de Wikidata.org est incluse dans les adresses IPv4 déclarées dans Wikidata. Nous obtenons donc une association avec l'entité possédant le label suivant : *Wikimedia Foundation* ainsi que les étiquettes suivantes :

```
industryLabel : ['wiki']
typeLabel : ['charitable organization', 'foundation']
employees : ['700'].
```

3.2.3 Obtenir des informations sur l'entité propriétaire

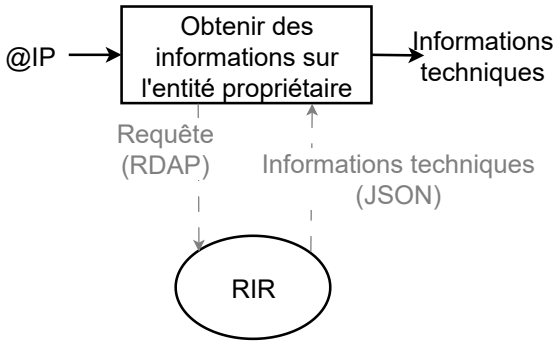


FIGURE 3.2 – Obtenir des informations sur l'entité propriétaire en utilisant RDAP

Afin de retrouver l'entité propriétaire de l'adresse IP dans la base de Wikidata, nous avons besoin de connaître le **nom de cette entité**. Pour ce faire, les registres Internet régionaux (RIR) peuvent être interrogés, car ils sont responsables de l'attribution des adresses IP. Lorsqu'une adresse IP est attribuée, un certain nombre d'informations sont fournies par le nouveau titulaire. Ces informations sont conservées par les RIR et accessibles par le biais de protocoles tels que Whois[51] ou RDAP[118, 117, 116].

Par conséquent, comme le montre la figure 3.2, la première étape de notre processus consiste à accéder aux données détenues par le RIR à l'aide du protocole RDAP en partant d'une **adresse IP**. Nous avons choisi d'utiliser RDAP plutôt que Whois car RDAP est le protocole le plus récent et il devrait tendre à remplacer Whois. De plus, RDAP propose des données structurées, alors que Whois renvoie une chaîne de caractères.

Le protocole RDAP renvoie un JSON contenant toutes les informations et un extrait peut être trouvé dans l'Exemple 3 en version tronquée et dans l'annexe A en intégralité. De toute évidence, le champ du RDAP appelé *entity*, qui vise à désigner le propriétaire de l'adresse IP, est très intéressant. Malheureusement, de nombreuses informations contenues dans ce champ ont été rendues anonymes en Europe conformément au RGPD [1].

Cependant, comme illustré dans l'Exemple 3, nous pouvons voir que d'autres champs contiennent des informations concernant l'entité que nous souhaitons identifier. Par exemple, le champ *remarks* contient ici une description de l'entité qui utilise l'adresse IP, ainsi que du type d'entité duquel il s'agit.

Intéressons nous à un second exemple pour comprendre le format des données renvoyées par RDAP. L'adresse IP obtenue via requête DNS du nom de domaine de "wikidata.org" donne le résultat visible en Annexe B. Dans cet exemple, nous pouvons voir d'autres indications de l'entité propriétaire qui est Wikimédia. Nous pouvons le voir dans les noms de domaine (ligne 14) et dans le champ *name* (ligne 38).

À travers ces deux exemples, nous comprenons qu'il est nécessaire d'analyser les résultats renvoyés par RDAP pour identifier l'entité qui possède ou utilise une adresse IP. C'est ce que nous allons faire dans l'étape suivante.

3.2.4 Identification de l'entité

L'objectif de cette étape est d'obtenir un **vecteur d'identification** de l'entité propriétaire de l'adresse IP. Le vecteur d'identification contient le nom de l'organisation et les noms de domaine de l'organisation, qui sont des éléments pouvant être utilisés pour identifier une organisation. Dans cette partie, nous présentons deux méthodes pour identifier le nom de l'organisation qui possède l'adresse IP, à partir de sa **description RDAP**. La première analyse est basée sur l'observation de la fréquence d'apparition des mots dans certains champs spécifiés, première méthode que nous avons proposée dans la première version **IPSeen Fast** [161]. La deuxième méthode emploie un outil de traitement automatique du langage naturel pour identifier les noms propres et noms d'organisation, seconde méthode que nous avons proposée dans la seconde version **IPSeen Accurate**.

3.2.4.1 IPSeen Fast : Analyse de la fréquence d'apparition des mots dans RDAP

Dans le but d'extraire le vecteur d'identification, nous mettons en œuvre une analyse de texte qui nettoie toutes les entrées contenues dans le JSON renvoyé par le RDAP : elle permet d'identifier les sous-chaînes les plus pertinentes. Nous sélectionnons par la suite le top 4 des mots qui ont les plus hautes fréquences d'apparition. Nous avons décidé d'utiliser le top 4, puisque nous avons mesuré expérimentalement qu'il s'agit d'un nombre de résultats suffisamment discriminant. Nous considérons les sous-chaînes les plus fréquentes contenant ces mots les plus fréquents. La chaîne sélectionnée est la chaîne composée des différents mots qui a la fréquence d'apparition la plus haute. En cas d'égalité, nous prenons la chaîne la plus longue, car dans l'étape suivante, nous pourrons toujours décomposer à nouveau cette chaîne et cela nous permet de ne pas trop réduire le potentiel des informations.

1. https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

Exemple 3 Réponse RDAP pour l'IP 134.214.58.24

```
{[...]
  "name": "FR-ROCAD",
  "type": "LEGACY",
  "country": "FR",
  [...]
  "entities": [{[...]
    {
      "handle": "RENATER-MNT",
      "roles": [
        "registrant"
      ],
      "objectClassName": "entity"
    },
    [...]
    }],
  "remarks": [{
      "description": [
        "Reseau Optique du Campus de la Doua",
        "Centre Inter-etablissement pour les Services Reseaux",
        "Universite Claude Bernard Lyon 1, bat Braconnier",
        "21 avenue Claude Bernard",
        "69622 Villeurbanne CEDEX, France"]
    }],
  [...]
  "notices": [
    {"title": "Filtered",
        "description": [
        "This output has been filtered."]},
    [...]
}
```

[...] signifie que des champs ont été retirés pour une lecture plus fluide. Le résultat complet est présent en Annexe A

Dans cette version d'IPSeen, nous incluons également dans le vecteur des facteurs discriminants potentiels tels que le code pays, le nom de domaine, un numéro d'AS. Toutes ces informations proviennent également de RDAP. Elles nous serviront dans un second temps afin de filtrer et d'ordonner les résultats obtenus via Wikidata.

Exemple 4 Réponse reçue pour l'IP 134.214.58.24

Parmi le tableau des fréquences, voici le top 4 des chaînes identifiées : ['claude', 'bernard', 'univ-lyon1', 'centre'].
Après recherche de la chaîne la plus longue contenant ces mots, nous obtenons : "claude bernard".
Nous identifions aussi les domaines suivants : ['renater.fr', 'univ-lyon1.fr'].

L'Exemple 4 illustre les résultats obtenus à la suite de cette étape pour la version **IPSeen Fast**. On obtient ici le nom de l'université Claude Bernard, université qui partage le réseau avec l'INSA. Nous identifions également deux noms de domaine : celui de l'université Lyon 1 et celui de Renater, fournisseur d'accès des réseaux universitaires en France.

Cette méthode n'était pas parfaite, mais nous a permis d'identifier précisément les problèmes auxquels nous étions confrontés en utilisant RDAP : le RGPD qui remplace le nom de l'entité par un identifiant, l'apparence aléatoire du nom de l'entité que nous souhaitions identifier. Mais nous avons pu aussi identifier des solutions face à ces difficultés : des éléments discriminants que nous pouvons utiliser par la suite, l'apparition dans quasiment tous les cas du nom de l'entité qui correspond à celui de Wikidata (évaluation présentée dans le chap. 4).

Par la suite, nous avons décidé d'implémenter un algorithme d'analyse de langage de type NLP (Natural Language Processing) pour voir si les résultats étaient meilleurs.

3.2.4.2 IPSeen Accurate : Utilisation d'un outil de traitement automatique du langage naturel

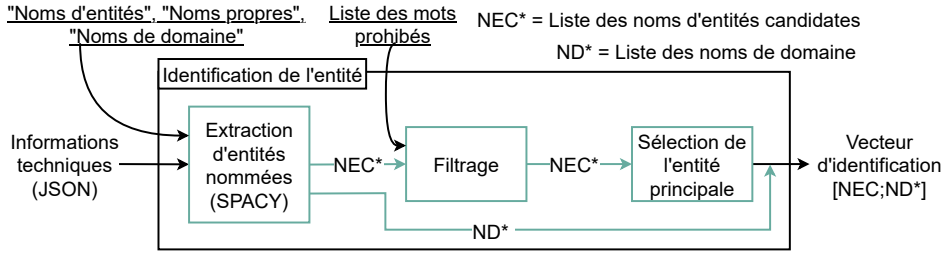


FIGURE 3.3 – Identification de l'entité propriétaire d'**IPSeen Accurate**

Afin d'améliorer nos résultats, nous avons décidé de mettre en œuvre le traitement du langage naturel, dans le but de réduire les faux positifs et d'affiner le processus. Une vue générale de cette nouvelle méthode est disponible sur la Figure 3.3. Nous expliquons dans la suite chaque bloc composant cette étape.

3.2.4.2.1 Extraction d'entités nommées

L'objectif de ce bloc est d'extraire une liste de mots susceptibles d'être le nom de l'entité, ainsi qu'une liste de noms de domaine. Dans notre exemple, visible dans l'annexe A, nous aimerions identifier *Renater* et, pour l'exemple en annexe B nous aimerions récupérer *Wikimedia* dans la liste des noms de l'entité et *wikimedia.org* dans la liste des noms de domaine.

L'entrée de ce bloc est le JSON contenant les informations sur le propriétaire. La sortie est constituée de deux listes : l'une de noms de domaine associés au propriétaire, nommée ND*, et

l'autre de noms potentiels pour l'entité, nommée NEC*. Comme nous ne voudrons qu'un nom, cette liste NEC* devra ensuite être nettoyée et filtrée.

Les algorithmes NLP sont une bonne option pour automatiser l'analyse de texte. Conformément à notre objectif, nous avons besoin d'une solution capable d'identifier les noms d'organisations et les noms de domaines. Nous avons opté pour une bibliothèque appelée Spacy [2] qui est adaptée à ce cas d'utilisation. Spacy est une bibliothèque Python open-source qui met en œuvre le traitement du langage naturel. Nous avons choisi Spacy parce qu'elle est entraînée pour reconnaître le nom de l'organisation et les noms de domaine. De plus, lorsque les noms d'organisations sont plus difficiles à identifier en raison de la structure du texte, nous pouvons également identifier les noms propres. En effet, le texte n'étant pas un texte classique, mais plutôt une suite d'informations ne suivant pas la structure d'une langue (pronom-verbe-complément), élargir le champ d'analyse est fructueux.

Afin d'extraire le vecteur d'identification, Spacy est implémenté et entraîné en anglais uniquement, étant donné que la plupart du texte dans le JSON est en anglais et qu'il n'y a pas beaucoup de phrases complètes. L'entraînement se fait via un ensemble de données fourni par Spacy.

En utilisant Spacy, nous extrayons les informations suivantes : les chaînes de caractères étant identifiées comme des noms d'organisation et les chaînes de caractères étant identifiées comme des noms propres si celles-ci ne font pas partie des noms d'organisation précédemment identifiés. Enfin, Spacy permet aussi d'identifier les noms de domaines. Nous extrayons donc une liste de noms de domaine uniques.

Finalement, nous obtenons 2 listes : une liste de noms d'entités candidates (composée de noms d'organisations et de noms propres) et une liste de noms de domaines. La liste des noms d'entités candidates est ensuite envoyée à l'étape de filtrage.

3.2.4.2.2 Filtrage

L'objectif de ce bloc est de filtrer la liste des noms d'entités des candidats. Dans l'étape précédente, nous avons prévu un large panel de résultats afin d'obtenir les meilleures chances d'obtenir les noms de l'organisation. Par conséquent, nous devons filtrer la liste pour éliminer le bruit ajouté. Un exemple de bruit peut être le nom des RIR ou les noms propres contenus dans les adresses postales, ou sur la figure 3, 'Claude Bernard', 'Villeurbanne Cedex', etc.

Afin de filtrer nos résultats, nous avons créé des listes de mots interdits. Différents types de listes ont été construits, en fonction du type de mot qui peut être filtré à partir de la liste identifiée des noms propres. Tout d'abord, il existe une liste de noms d'entités interdits. Cette liste contient les noms des différents RIR, car, dans les résultats de RDAP, ils peuvent apparaître plus souvent que le nom de l'entité. Par exemple, sur les deux exemples en annexe A et B, on voit de nombreuses fois respectivement *RIPE* et *ARIN*. Il s'agit bien de noms propres et il est tout à fait normal que ceux-ci soient identifiés par Spacy. Néanmoins, dans le cas de notre analyse, ces résultats ne sont pas pertinents pour nous. Cependant, faire le choix de les retirer, génère un faux négatif dans le cas où les adresses ne seraient pas attribuées et appartiennent bien à ces organisations (RIR). Néanmoins, dans ce cas, il est très facile d'identifier ces adresses par un humain en regardant les résultats RDAP.

Il existe également une liste permettant de filtrer les formes de sociétés telles que *s.a., LMT, MNT...* afin de donner des noms d'entités au texte plus simple. Dans ce cas, on va simplement

2. https://spacy.io/

venir nettoyer les chaînes de caractères afin de simplifier les recherches via Wikidata par la suite.

Enfin, une liste de vocabulaire lié aux télécommunications qui apparaissait régulièrement dans nos résultats a été créée pour les supprimer des noms des entités. En voici quelques exemples : *réseau, ADSL, ISP. . ..*

Nous supprimons également certains noms qui peuvent apparaître en raison du format de l'entrée que nous étudions, comme par exemple : *role, abuse. . .*, comme on peut le voir dans l'annexe B, 'Wikimedia abuse' peut être identifié comme un candidat pour le nom de l'entité. Il faut donc retirer le mot abuse.

En sortie de cette étape, nous avons donc une liste de noms propres filtrés, inférieure ou égale à la liste donnée en entrée.

3.2.4.2.3 Sélection de l'entité principale

L'objectif de la dernière étape est d'identifier 1 entité qui est renvoyée en sortie de ce bloc. L'entrée de ce bloc est la liste préfiltrée du NEC et la sortie est une chaîne de caractères, qui peut être composée de plusieurs mots.

Exemple 5 Réponse reçue pour l'IP 134.214.58.24 avec la v2.0

Les résultats obtenus avec Spacy sont les suivants :

```
{'fr': 4, 'renater': 5, 'eu': 1, 'reseau': 1, 'optique': 1, 'campus':
    1, 'la': 1, 'douancentre': 1, 'inter': 4, '-': 4, 'etablissement':
    3, 'les': 3, 'services': 1, 'reseaux universite': 4, 'claude': 8, '
    bernard': 4, 'lyon': 5, 'bat': 1, 'villeurbanne': 4, 'cedex': 4, '
    inaccuracy': 1, 'reporting': 1, 'rocad': 1, 'cisr': 1, 'centre': 3,
     'avenue': 2, 'france': 2, 'remi': 1, 'sader': 1, 'thomas': 1, '
    petit': 1, '': 9}
```

Par la suite, nous filtrons les résultats pour obtenir :

```
{'villeurbanne': 4, 'petit': 1, 'cisr': 1, 'inter': 4, 'renater': 5, '
    optique': 1, 'cedex': 4, 'sader': 1, 'les' : 3, 'reporting': 1, '
    avenue': 2, 'inaccuracy': 1, 'bernard': 4, 'douancentre': 1, '
    reseaux universite': 4, 'reseau': 1, 'etablissement': 3, 'lyon': 5,
     'campus': 1, 'centre': 3, 'rocad': 1, 'remi': 1, 'claude': 8, '
    france': 2, 'services': 1, 'bat': 1, 'thomas': 1}
```

Avec cette proposition qui utilise Spacy, la chaîne extraite est la suivante : claude.
Nous obtenons également les noms de domaines suivants : ['renater.fr', 'univ-lyon1.fr'].

Pour ce faire, nous calculons l'occurrence d'apparition de chaque chaîne de caractères. Comme nous l'avions vu auparavant, la plupart du temps, la chaîne de texte correspondant à l'occurrence la plus élevée est le nom de l'organisation ou quelque chose d'assez proche. Nous avons donc décidé d'exploiter le nombre d'occurrences de chaque chaîne de caractères. En outre, l'occurrence de chaque élément d'une chaîne de caractères composée de plusieurs mots a été ajoutée si un mot a été observé séparément de la chaîne complète. Cela permet d'ajouter

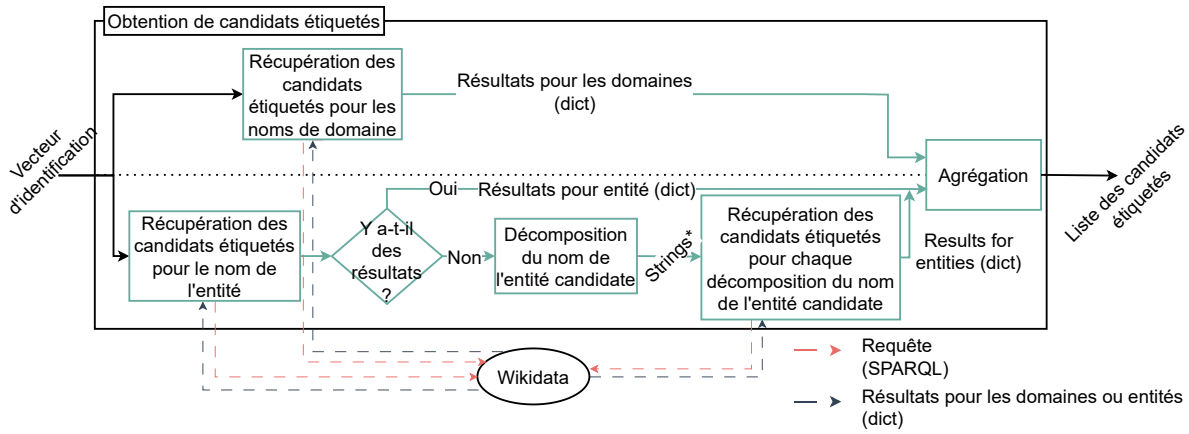


FIGURE 3.4 – Obtention des candidats étiquetés, via Wikidata

FIGURE 3.5 – Extrait de la page Wikidata de l'INSA Lyon - Exemple pour illustrer l'architecture de Wikidata

de l'importance aux chaînes plus précises, car plus longues. De plus, dans l'étape suivante, nous décomposons potentiellement les chaînes trop longues. Nous préférons favoriser trop d'informations que pas assez.

En dernier lieu, la chaîne de texte ayant l'occurrence la plus élevée est sélectionnée.

L'Exemple 5 illustre le vecteur d'identification obtenu pour notre exemple conducteur. Avec cette version, nous obtenons une chaîne plus courte. Les noms de domaine, quant à eux, restent identiques.

3.2.5 Obtention de candidats étiquetés

Grâce à l'étape précédente, nous avons obtenu le nom de l'entité ainsi qu'une liste de noms de domaines qui correspondent à cette entité, rassemblés sous la notion de **vecteur d'identification**. On peut donc à présent effectuer des requêtes sur Wikidata pour nous permettre d'obtenir les informations concernant l'entité trouvée.

Dans cette section, nous allons dans un premier temps expliquer le fonctionnement de Wikidata, puis nous verrons comment nous l'utilisons. Suite à l'introduction de Wikidata et du vocabulaire associé, nous allons à présent nous intéresser aux différentes requêtes utilisées dans notre processus.

Le processus suivi pour l'obtention des candidats étiquetés est visible sur la Figure 3.4. Ce processus passe par l'exploitation de Wikidata. Nous allons ici faire une recherche par chaîne de caractères pour obtenir nos candidats. Nous allons aussi, dans un but d'optimisation, venir récupérer les différentes étiquettes recherchées. Nous obtiendrons donc, à la suite de cette étape, un sous-ensemble de Wikidata, comprenant les **candidats** pour une entité nommée, ainsi que les étiquettes associées à ces candidats. Nous allons à présent décrire les différentes étapes pour la réalisation de cette étape.

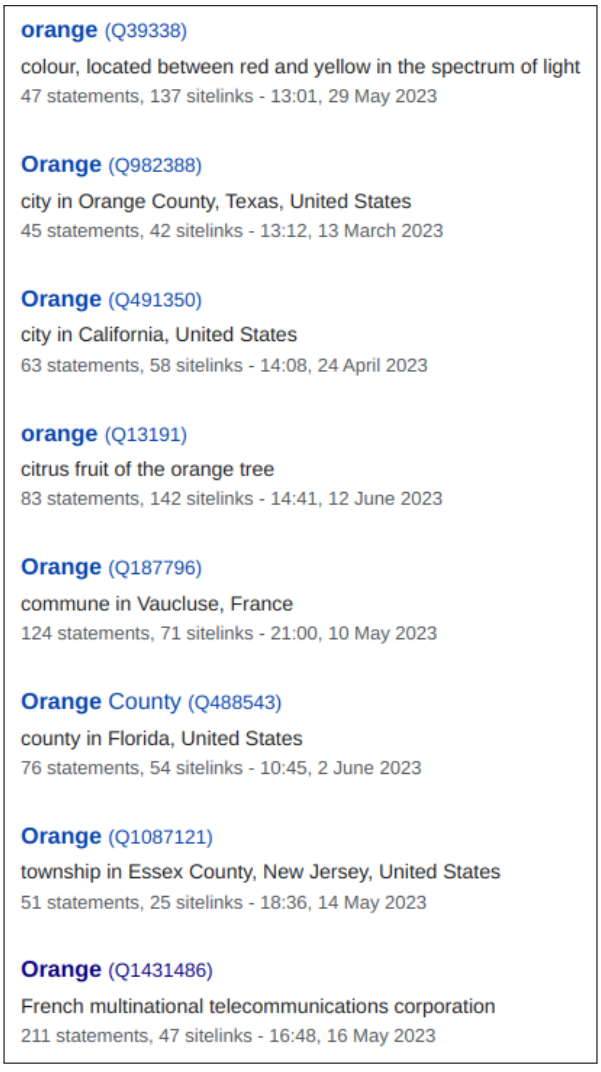


FIGURE 3.6 – Résultat possible pour la chaîne de caractères *Orange*

3.2.5.1 Introduction au fonctionnement de Wikidata

Afin de mieux comprendre la suite du processus, nous allons d'abord expliquer le fonctionnement de Wikidata. Comme il a été discuté précédemment, Wikidata fonctionne comme une encyclopédie et rassemble des connaissances générales sur un grand nombre d'éléments, d'idées, d'entités. Wikidata présente un grand avantage dans notre utilisation : son API. Nous pouvons facilement faire des requêtes en SPARQL pour obtenir des informations sur l'ensemble des éléments.

Pour exemplifier et mieux comprendre, prenons un exemple avec un extrait de la page Wikidata de l'INSA Lyon présenté sur la Figure 3.5. Sur cette Figure, on peut voir que l'INSA possède un identifiant, *Q1633859*, appelé le Qid. C'est un identifiant unique qui correspond à cette page et à cet ensemble de

champs. Chaque Qid possède une liste de champs appelés *Statments* qui décrivent l'élément. Chaque champ correspond à une propriété que l'on peut utiliser pour construire des requêtes SPARQL. Chacun des champs peut avoir plusieurs valeurs.

Wikidata présente beaucoup d'avantages, mais il faut aussi naviguer autour de ses inconvénients. Une subtilité qui nous oblige à devoir analyser les résultats à la suite d'une requête est que Wikidata contient des homographes. Par conséquent, même dans le cas où nous aurions identifié une chaîne de caractères qui correspond parfaitement à celle de Wikidata, nous pouvons avoir plusieurs pages qui ont le même nom. La Figure 3.6 est un exemple pour une recherche de *Orange*. Il nous faudra donc trouver un moyen de trier et de filtrer les résultats. Cela sera présenté dans la section 1.2.6.

3.2.5.2 Récupération de candidats étiquetés pour le nom de l'entité

Les objectifs des blocs **Récupération des candidats étiquetés pour le nom de l'entité** et **Récupération des candidats étiquetés pour chaque décomposition du nom de l'entité candidate** sont de récupérer les candidats étiquetés en utilisant une requête SPARQL. Ce paragraphe décrit les deux blocs, car le dernier est le même que le premier répété plusieurs fois. En entrée, ce bloc a besoin d'une chaîne de caractères qui sera incluse dans la requête SPARQL pour rechercher du texte. Avec SPARQL, un service appelé *MediaWikiAPI* permet d'effectuer une recherche de texte. Tous les éléments qui contiennent une certaine chaîne de caractères dans leur étiquette d'élément ou leur étiquette *also known as* peuvent être récupérés à l'aide de ce service. Cela permet de rechercher et d'extraire les propriétés sélectionnées dans la même recherche et d'accélérer le temps de calcul.

Notre requête par chaîne de caractères est celle présentée en Listing 3.1. Dans cette requête, on peut voir les différents champs que nous allons récupérer :

1. item : le Qid identifiant l'item
2. itemlabel : le nom de l'item, celui qui doit correspondre au nom que nous avons utilisé pour faire notre requête.
3. type et typelabel : le nom et identifiant qui correspond au type de l'item
4. legal et legallabel : la forme légale de l'entité ainsi que l'identifiant qui correspond à cette forme légale. Par exemple : corporation.
5. employees : le nombre d'employés de l'entité
6. subscriber : nombre de souscripteurs à un FAI
7. sub : entreprise subsidiaire. Par exemple : Orange a pour subsidiaire Dailymotion ou encore Orange Espagne.

Pour chacun des champs textuels, nous récupérons l'identifiant Wikidata de la propriété ainsi que le *Label* (représenté par le nom de la variable suivi de Label).

Une fois tous les résultats collectés, un dictionnaire est créé avec tous les résultats trouvés et leurs étiquettes. La sortie de ce bloc est un dictionnaire formaté contenant les candidats. Chaque candidat est identifié par son *Qid* et un dictionnaire d'étiquettes lui est attribué.

```
SELECT ?item ?type ?itemLabel ?typeLabel ?legal ?legalLabel ?employees
?itemAltLabel ?industryLabel ?suivi_parLabel ?suivi_par ?ASN ?siteweb
?countryLabel ?student ?subscribers ?sub WHERE {
  SERVICE wikibase:mwapi {
     bd:serviceParam wikibase:endpoint "www.wikidata.org";
                 wikibase:api "EntitySearch" ;
                 mwapi:search "{Entity_name}" ;
                 mwapi:language "en" .
     ?item wikibase:apiOutputItem mwapi:item .}
  OPTIONAL {?item (wdt:P279|wdt:P31) ?type;}
  OPTIONAL {?item wdt:P1454 ?legal;}
  OPTIONAL {?item wdt:P1128 ?employees;}
  OPTIONAL {?item wdt:P452 ?industry;}
  OPTIONAL {?item wdt:P156|wdt:P1366 ?suivi_par;}
  OPTIONAL {?item wdt:P856 ?siteweb;}
  OPTIONAL {?item wdt:P355 ?sub;}
  OPTIONAL {?item wdt:P17 ?country;}
  OPTIONAL {?item p:P3797 ?statement.
    ?statement ps:P3797 ?ASN.
    ?statement wikibase:rank ?rank.}
  OPTIONAL {
     ?article schema:about ?item.
     ?article schema:inLanguage "en" .
     FILTER (SUBSTR(str(?article), 1, 25) = "https://en.wikipedia.org/")
   }
  OPTIONAL {?item wdt:P3744 ?subscribers.}
  OPTIONAL {?item wdt:P2196 ?students.}
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
LIMIT 50
```

LISTING 3.1 – Requête SPARQL : récupération des labels correspondant au nom de l'entité *{Entity_name}*

3.2.5.3 Récupération de candidats étiquetés en utilisant le nom de domaine

Premièrement, nous essayons d'obtenir des candidats en partant des noms de domaines que nous avons identifiés précédemment. Pour cela, nous allons comparer les noms de domaines identifiés aux noms de domaine du champ *official website* de Wikidata. Ce champ donne les URLs des sites officiels des différentes entités. Nous aurions aussi pu comparer les noms de domaines avec le champ *email address*, mais avons choisi de laisser ce champ de côté, car les entités liées à ce champ étaient principalement des personnes. Afin de gagner en efficacité, nous avons extrait tous les noms de domaines disponibles dans Wikidata en utilisant la requête présentée en Listing 3.2.

Cette requête permet d'obtenir une correspondance entre les sites Web et les Qids. Lorsque nous avons une correspondance entre les noms de domaine, nous faisons une seconde requête afin d'obtenir les étiquettes liées au candidat. La sortie de ce bloc est un dictionnaire formaté

contenant les candidats. Chaque candidat est identifié par son *Qid* et un dictionnaire d'étiquettes qui lui est attribué.

```
  SELECT ?plage_IPv4 ?item ?itemLabel ?itemAltLabel ?rank WHERE {
   SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
   OPTIONAL {
      ?item p:P3761 ?statement.
      ?statement ps:P3761 ?plage_IPv4;
      wikibase:rank ?rank.
  }
}
```

LISTING 3.2 – Rêquete SPARQL : récupération de l'ensemble des plage IPv4 contenues dans Wikidata

3.2.5.4 Décomposition du nom de l'entité candidate

Ce bloc a été ajouté dans le cas où il n'y aurait pas de résultats après la requête initiale. Nous avons remarqué que ce cas se produit lorsque le texte utilisé pour la requête est trop long. Cela peut être dû à ce qui est extrait de RDAP, ou simplement à ce qui est identifié comme un nom propre ou un nom d'entité par Spacy. Il est donc nécessaire d'explorer d'autres possibilités.

En entrée de ce bloc, on retrouve le nom d'entité identifié comme étant propriétaire de l'adresse IP. En listant toutes les combinaisons possibles, ainsi que chaque mot individuellement, on décompose le nom de l'entité en plusieurs sous-noms. La sortie de ce bloc contient donc une liste des différentes combinaisons des mots de la chaîne de caractères, que nous allons pouvoir par la suite tester.

3.2.5.5 Agrégation

Ce bloc regroupe tous les résultats obtenus à partir des requêtes portant sur les noms de domaine et d'entité, ou sur les deux à la fois, et renvoie la liste agrégée des candidats étiquetés qu'il faudra analyser plus avant pour sélectionner les résultats. Il supprime également les entrées en double. Une fois que nous avons obtenu toutes les informations nécessaires pour mener l'analyse, ainsi que toutes les étiquettes finales que nous souhaitons obtenir, nous agrégeons toutes ces données dans un dictionnaire où la clé est le Qid associé aux informations. À présent, nous avons une liste d'entités candidates. Nous allons par la suite filtrer et ordonner cette liste afin de trouver le meilleur candidat. Nous présentons cela dans la section suivante.

Exemple 6 Liste des candidats obtenus suite aux requêtes

En ce qui concerne notre IP, voici la liste des résultats que l'on obtient à ce stade :
Pour **IPSeen Fast** :

label	industry Label	aka	type	Site web
Claude Bernard University Lyon 1	higher education	Université Lyon 1, UCBL, univ-lyon1.fr, Université Claude Bernard Lyon 1	university in France	https://www.univ-lyon1.fr/
Lyon research team in Information and Communication Sciences			facility	http://www.elico-recherche.eu/
Fédération Informatique de Lyon			facility	http://fil.cnrs.fr/
Management et économie Lyon Saint Etienne			facility, organization,	https://maelyse.universite-lyon.fr/
Faculté des Sciences de Lyon			faculty	
Centre for Astronomical Research of Lyon			Joint Research Unit	http://cral.univ-lyon1.fr/?lang=en, https://cral.univ-lyon1.fr/?lang=fr,
Center for Research in Image Acquisition and Treatment for Health			Joint Research Unit, laboratory,	http://www.creatis.insa-lyon.fr

Exemple 7 Liste des candidats obtenus suite aux requêtes

Pour **IPSeen Accurate** : Résultats similaires au tableau obtenu pour **IPSeen Fast**. Il contient néanmoins plus de résultat à analyser puisque la chaîne de caractères utilisée pour la requête Wikidata était moins précise.

Les Exemples 6 et 7 illustrent les résultats obtenus pour les versions respectives d'**IPSeen Fast** et **IPSeen Accurate**. Les résultats sont différents puisque nous n'avions pas obtenu le même vecteur d'identification à l'étape précédente de notre processus.

3.2.6 Sélection du meilleur candidat

Nous allons à présent devoir analyser ce résultat afin d'identifier le **résultat** correct, s'il existe, parmi la **liste de candidats** que nous avons en entrée.

À ce stade de notre processus, nous avons une liste de résultats potentiels de laquelle nous devons extraire l'entité correcte si celle-ci est présente. La sortie de cet ultime bloc doit être les étiquettes organisationnelles du résultat identifié, si celui-ci figure parmi la liste en entrée. Si cela n'est pas le cas, alors nous souhaitons que notre solution ne retourne aucun résultat.

Pour cela, nous avons besoin de métriques pour pouvoir évaluer ce qui fait d'un résultat le résultat correct. Dans cette partie, nous allons présenter deux solutions différentes qui sont deux possibilités de réponse à cette problématique. La première solution sera basée sur des métriques que nous avons observées et sur un seuillage manuel de ces métriques. La seconde solution sera basée sur l'entraînement d'un algorithme de Machine Learning qui a pour objectif d'identifier le résultat correct.

3.2.6.1 IPSeen Fast : Empirique et résultat rapide

Dans ce paragraphe, nous présentons une méthode qui permet d'identifier un grand nombre d'entités correctement, mais a aussi un plus grand nombre de faux positifs que la solution que nous allons présenter par la suite. Néanmoins, cette première approche nous a permis d'obtenir des résultats corrects et d'identifier des métriques qualifiant les résultats.

Tout d'abord, nous identifions l'élément de Wikidata ayant le plus grand nombre de mots en commun avec le nom de l'entité que nous avions précédemment identifié. À cette fin, nous inspectons les champs suivants : *itemlabel*, *also known as*, et *subsidiaries*. Tous ces champs donnent des noms possibles de notre entité. Nous cherchons à identifier celui dont nous nous rapprochons le plus.

Une autre métrique nous permet de filtrer nos résultats en fonction des champs qu'ils ont sur Wikidata. Instinctivement, pour filtrer les résultats que nous avions obtenus pour la chaîne de caractères Orange présentée sur la figure 3.6, nous comprenons que notre résultat doit avoir certaines propriétés. Par exemple, comme nous cherchons une entité, notre résultat devrait posséder une forme légale, un domaine d'activité, un nombre d'étudiants s'il s'agit d'une école, ou un nombre de clients s'il s'agit d'un FAI. En partant de cette observation, nous avons conclu que de compter le nombre d'étiquettes pour lesquelles nous avons obtenu un résultat non nul nous permet d'exclure les résultats les plus faibles. Avec cela, nous excluons les résultats ayant une faible probabilité de correspondre à ce que nous cherchons.

Nous avons aussi mis en place d'autres métriques calculées pour chaque entité restante, ces métriques sont présentées dans ce paragraphe. Premièrement, dans Wikidata, il existe un champ ASN qui peut être spécifié pour certaines entités. Nous pouvons vérifier si le champ correspond à celui que nous pouvons voir dans les résultats obtenus via le protocole RDAP. L'origine géographique est aussi un facteur qui peut nous aider à discriminer les résultats. En effet, il existe un champ pays dans Wikidata que nous pouvons aussi comparer au champ *asn_country_code* des résultats de RDAP. Enfin, pour chaque résultat, nous allons aussi mesurer la similitude entre le nom de domaine que l'on peut trouver dans le champ *official website* de Wikidata, et la liste des noms de domaines que nous avons dans notre vecteur d'identification de notre entité. Une forte similarité nous met sur la piste qu'il s'agit de la bonne entité.

Toutes ces métriques et informations sont calculées et donnent pour chaque résultat de notre liste de résultats un score. Les résultats sont classés en fonction du score, et nous considérons le résultat comme étant celui avec le meilleur score. Néanmoins, nous voulons aussi nous assurer de la qualité du résultat et écarter les résultats qui seraient trop mauvais : soit de par leur manque d'information, soit parce qu'il n'y avait aucun résultat correct dans la liste.

Pour mettre en place le filtrage sur les résultats finaux, nous avons mis en place un seuillage. Le seuillage est fait sur le score global que nous calculons en nous basant sur l'ensemble des métriques, mais aussi sur chacun des critères. Un résultat, pour être bon, a donc besoin d'être au-dessus du seuil de chaque métrique, mais aussi au-dessus du seuil global.

Si le résultat candidat ne valide pas ces conditions, nous considérons qu'il n'y a pas de résultat. Ces seuils ont été définis expérimentalement à l'aide d'un ensemble de données composées des adresses IP présentes dans Wikidata. Nous avons fait varier les seuils de chaque valeur ainsi que les coefficients d'importance de chaque résultat pour le calcul du score global. Nous avons fixé les seuils avec les valeurs qui nous offraient le meilleur compromis entre vrai positif et faux positif.

Cette étape nous a également permis d'identifier un problème avec les données contenues dans RDAP. Parfois, probablement en raison d'un manque de mises à jour de la part des entités, nous avons remarqué que l'entité que nous avions identifiée a été acquise par une autre société ou a simplement changé de nom. Parfois, il existe encore un article dans Wikidata désignant l'ancienne société. Pour remédier à cela, nous utilisons le champ *followed by* de Wikidata, qui nous redirige vers l'élément qui désigne les informations les plus récentes concernant l'entité propriétaire de l'adresse IP.

Le résultat obtenu avec **IPSeen Fast** pour notre IP exemple est présenté en Exemple 8. On peut voir que nous n'avons pas trouvé l'INSA précisément. Néanmoins, nous avons trouvé l'Université qui partage le campus et le sous-réseau avec l'INSA et nous avons obtenu le bon type ainsi que la bonne industrie.

Exemple 8 Résultats obtenus avec **IPSeen Fast**

Pour cet exemple, notre meilleur résultat correspond au label : *Claude Bernard University Lyon 1.*
Le résultat obtenu pour notre exemple à l'aide de la solution présentée dans cette partie est la suivante :

```
industryLabel : ['higher education']
typeLabel : ['university in France']
```

Pour cette entité, nous n'avons pas d'information quant à la taille de celle-ci (nombre d'étudiants).

3.2.6.2 IPSeen Accurate : Utilisation d'un algorithme de Machine learning

La sélection du meilleur candidat telle que décrite en 3.2.6.1 est intéressante, car rapide à répondre avec une précision raisonnable. Dans ce paragraphe, nous allons présenter une solution alternative permettant d'améliorer la qualité des résultats. L'objectif reste le même, à savoir identifier, parmi la liste de candidats étiquetés, le résultat correct s'il existe. Une vue globale de notre processus est proposée sur le schéma 3.7.

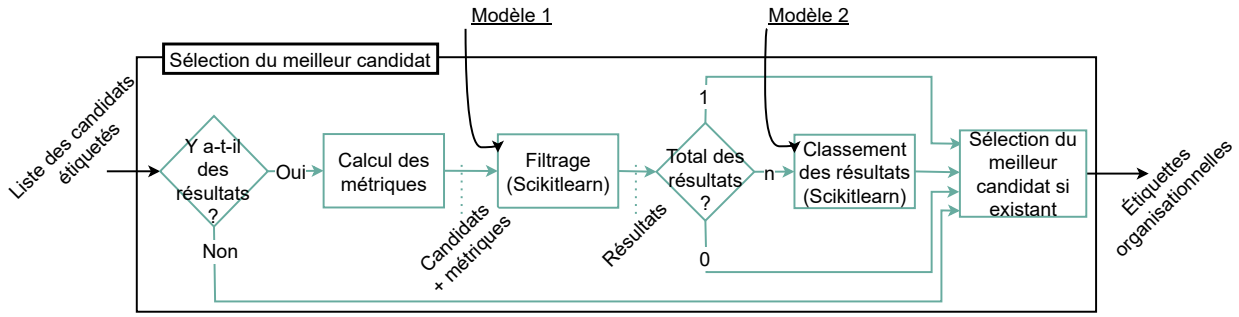


FIGURE 3.7 – Sélection du meilleur candidat et des étiquettes associées **IPSeen Accurate**

En entrée de l'étape de sélection du meilleur candidat est la liste des candidats étiquetés qui peut être vide. La sortie, elle, est constituée des étiquettes organisationnelles liées au candidat qui a été déterminé comme étant le bon résultat. Dans le cas contraire, si l'algorithme détermine qu'aucun des candidats n'était approprié ou si la liste des candidats était vide au départ, il n'y a pas de résultats.

Deux blocs contenant des algorithmes d'apprentissage sont utilisés dans cette nouvelle version afin de déterminer le meilleur résultat : le bloc de filtrage et le bloc de classement des résultats. Il est donc nécessaire d'avoir deux modèles pré-entrainés.

Sur la Figure 3.7, nous commençons par calculer les différentes métriques pour chaque candidat. Ces métriques nous servent pour l'étape de filtrage par la suite. En fonction du nombre de résultats disponibles après l'étape de filtrage, nous avons potentiellement besoin de l'étape de classement des résultats. Cette étape n'est appliquée que lorsque nous avons plusieurs résultats. Nous voulons donc les ordonner pour n'en garder qu'un seul : le meilleur. Enfin, nous extrayons les étiquettes de ce meilleur résultat. L'ensemble des étapes de ce processus est décrit dans les sous-sections suivantes.

Cette solution est plus coûteuse en temps et nécessite un ensemble de données d'apprentissage, néanmoins, elle est plus précise et surtout limite le nombre de faux positifs. Pour la mettre en place, nous avons dû créer un ensemble de données d'entraînement, construit via les IPs disponibles dans Wikidata, et un ensemble d'évaluation qui sera plus longuement présenté dans le chapitre 4.

3.2.6.2.1 Calcul des métriques

En entrée de notre processus, nous avons à disposition des candidats étiquetés. Nous souhaitons utiliser des algorithmes de Machine Learning pour évaluer ces données. Pour cela, nous devons prétraiter les données en entrée pour pouvoir les rendre exploitables par un algorithme de Machine Learning. Pour le moment, nous avons seulement des étiquettes avec des valeurs textuelles. Nous devons transformer ces données en métriques évaluables par notre algorithme d'apprentissage.

Pour cela, nous devons définir les métriques générales. Dans la version précédente, **IPSeen Fast**, nous avions trouvé des métriques performantes, comme on peut le voir dans le chapitre 4. Nous allons les réutiliser. Nous ajoutons aussi d'autres métriques dans cette nouvelle version.

Ainsi, pour chaque candidat, on extrait les éléments suivants :

— Nombre d'entrées non nulles : nombre d'étiquettes qui devraient être attribuées et qui ne sont pas nulles. Comme nous l'avons vu précédemment, cette métrique nous permet

de spécifier une forme des résultats que nous souhaitons obtenir. Nous allons valoriser les résultats qui ont certaines étiquettes, plus particulièrement les étiquettes que nous cherchons, puisqu'elles décrivent bien une entité.

— Similitude du nom de l'entité : nombre de mots communs entre le texte interrogé et le nom (ainsi que le *also kown as*). Ici, le résultat est un chiffre égal au nombre de mots en commun.
— Entity name Jaro distance [162] : une évaluation de la distance Jaro entre le texte interrogé et le nom (ainsi que le *also known as*). La distance de Jaro a été privilégiée puisqu'elle accentue la similarité entre les premiers caractères des deux chaînes comparées.
— Domain name similarity : un test de correspondance entre le domaine de la liste de domaines et le domaine trouvé dans la propriété *official website* de Wikidata. Ici, le résultat vaut 1 s'il s'agit du même nom de domaine, 0 sinon.
— Domain name Jaro distance : évaluation de la distance Jaro entre le domaine de la liste de domaines et le domaine trouvé dans la propriété *official website* de Wikidata.
— Item name in RDAP (nom de l'élément dans le résultat RDAP) : test pour savoir si l'étiquette de l'entité candidate se trouve dans le RDAP. Permets de vérifier une éventuelle erreur d'identification lors de l'étape d'extraction du nom de l'entité.
— Type in RDAP : ici, nous venons tester si le type de l'entité trouvé dans Wikidata est présent dans les résultats RDAP. Par exemple, on peut essayer de retrouver la chaîne de caractères *ISP* pour une entreprise ISP. Le résultat renvoyé est 1 si nous trouvons le type, 0 sinon.
— Country validation : tester si le nom du pays se trouve dans le RDAP.
— ASN validation : tester si le numéro de système autonome provenant de RDAP est identique à la propriété *ASN* de Wikidata.

Toutes ces métriques ont été choisies suite à des expérimentations. Elles vont nous permettre de faire par la suite un filtrage. Nous avons sélectionné une fois de plus la distance de Jaro qui avantage les similarités en début de chaines de caractères comparées.

La sortie de ce bloc est constituée des candidats associés aux métriques précédemment itérées.

3.2.6.2.2 Filtrage

Maintenant, il est nécessaire de filtrer les résultats afin d'en obtenir un seul : le résultat exact. Cependant, nous allons constater que lors de cette étape, il est possible de repérer plusieurs candidats appropriés ou même aucun. Ainsi, la liste de candidats filtrés qui sort de notre bloc peut être de longueur 0, 1 ou n. L'étape qui suit ce bloc diffère selon le nombre de résultat.

En entrée de cette étape, nous avons donc une liste de candidats avec leur métrique. Pour pouvoir évaluer nos candidats, nous utilisons un algorithme d'apprentissage automatique. Plusieurs algorithmes (Random Forrest, Perceptron, Perceptron multicouche) ont été testés et les résultats de chacun sont présentés dans le chapitre d'évaluation.

Nous décrivons par la suite la création de notre modèle 1 ainsi que la mise en oeuvre de l'étape de filtrage.

Création d'un ensemble de données d'entraînement Afin d'entraîner nos solutions, nous avons créé un ensemble de données d'entraînement basé sur les adresses IP déjà disponibles dans Wikidata.

Nous avons extrait toutes les adresses IPv4 présentes dans Wikidata, ce qui constitue un ensemble de 2228 blocs d'adresse IPv4. Cela nous a constitué un ensemble de données qui associe adresse IP et Qid. Par la suite, nous avons passé notre ensemble de données à travers toutes les étapes d'**IPSeen Accurate**. Cela nous a permis de prendre en compte le biais créé par notre solution pour l'obtention du vecteur d'identification. Ce biais est lié au fait que l'identification de l'entité n'est pas parfaite. Nous obtenons aussi tous les candidats après l'étape d'obtention des candidats. Pour chaque candidat, nous affectons une étiquette résultat qui vaut 1 lorsque le Qid du candidat est celui qui correspond à l'adresse IP recherchée et 0 sinon. Nous avons donc pour chaque adresse IP de notre ensemble de données la liste de candidats correspondante avec au plus 1 résultat correct.

Par la suite, nous avons également préfiltré les résultats de l'ensemble de données afin de créer un ensemble de données dans lequel seuls les résultats corrects figurent dans l'ensemble de données initial. Par conséquent, si la liste ne contient pas de résultats corrects, les lignes correspondant à l'adresse IP interrogée sont supprimées.

Mise en œuvre du filtrage Divers algorithmes ont été entraînés avec cet ensemble de données : un Random Forest, un perceptron et un perceptron multicouche de la bibliothèque Scikit-learn [163]. Les différents paramètres des solutions ont été optimisés manuellement. Nous nous sommes arrêtés sur l'algorithme de Random Forest, car il offrait le meilleur compromis entre vrai positif et faux positifs. Les évaluations de chacune des propositions sont présentées dans le Chapitre 4 qui détaille l'évaluation de la solution .

Pour chaque entrée de notre liste de candidats, l'algorithme va déterminer s'il s'agit d'une solution correcte ou non, en renvoyant un entier qui vaut 1 s'il appartient à la première catégorie ou 0 s'il appartient à la seconde catégorie. Notre liste de candidats est alors entièrement évaluée. Nous gardons seulement le ou les candidats qui appartiennent à la catégorie des résultats corrects.

À présent, nous avons filtré et sélectionné 0 ou 1 ou plusieurs candidats grâce à notre algorithme Random Forest et à notre ensemble d'entraînement.

3.2.6.2.3 Classement des résultats

Dans ce bloc, nous traitons le cas particulier où nous avons identifié plusieurs candidats suite au filtrage des résultats. Il nous faut départager les candidats dans le cas particulier où nous en avons plus d'un.

Nous avons favorisé l'ajout d'une étape supplémentaire puisque nous souhaitons que notre solution soit entièrement autonome et nous ne souhaitons pas d'intervention humaine pour départager les résultats. Nous ne souhaitons pas avoir une intervention humaine pour départager les derniers résultats.

Pour réaliser cet objectif, nous avons exploré les 2 solutions ci-dessous. La seconde est celle que nous avons finalement retenue dans la version de référence d'**IPSeen Accurate**, comme nous le justifions par la suite en Section 4.6.3.2.

1ère solution : regarder les probabilités obtenues La première solution consiste à renvoyer la probabilité la plus élevée obtenue par l'algorithme RF du bloc précédent. En effet, lors de la classification au niveau du bloc de filtrage, nous pouvons choisir de calculer les probabilités d'appartenance aux deux catégories : correct et incorrect. Au lieu d'avoir un résultat qui vaut 0 ou 1 pour l'appartenance à la catégorie, nous aurions une valeur comprise entre 0 et 1. Nous pouvons alors ordonner les candidats restants et sélectionner celui avec la plus grande probabilité.

2ème solution : introduire un second algorithme de Random Forrest avec de nouvelles métriques La seconde option que nous avons décidé d'explorer, consiste à employer une seconde fois l'algorithme de random forrest, mais cette fois-ci en intégrant de nouvelles métriques.

Comme nous l'avons expliqué auparavant, une des intuitions qui nous a fait nous tourner vers les algorithmes de Machine Learning est que chaque candidat est en réalité une liste d'étiquettes qui lui sont affectées par Wikidata. Nous pouvons transformer nos candidats en une liste d'étiquettes, qui prendra la valeur 1 si le candidat possède l'étiquette et 0 si celui-ci ne la possède point.

Pour entraîner notre second algorithme de RF, notre jeu de données d'entraînement doit également être mis à jour. Nous allons utiliser le précédent ensemble de données, mais cette fois-ci, nous allons garder seulement les entrées qui ont été filtrées par le premier algorithme de RF. Nous attribuons alors pour chacun de ces résultats les étiquettes complémentaires. Enfin, nous utilisons ce second ensemble de données pour entraîner notre seconde implémentation de RF. Nous pouvons par la suite utiliser le modèle créé pour classer nos derniers résultats.

La deuxième solution ayant donné de meilleurs résultats, nous avons opté pour celle-ci. Les résultats de la comparaison des différents algorithmes sont présentés dans le Chapitre d'évaluation (Section 4.6.3.2) ainsi qu'une évaluation de l'ensemble de notre processus.

En définitive, la sortie est le résultat identifié par cette seconde implémentation de l'algorithme de Random Forest comme ayant la meilleure probabilité d'appartenir à la catégorie des résultats corrects.

3.2.6.2.4 Sélection du meilleur candidat

Enfin, il nous reste une ultime étape, celle de retourner les étiquettes organisationnelles qui correspondent à notre résultat. Cela se fait seulement dans le cas où nous avons réussi à identifier un résultat. Si nous n'avons pas identifié de résultat, nous ne retournons aucune étiquette.

La non-présence de résultat peut indiquer plusieurs causes. Tout d'abord, il se peut qu'il n'y ait aucun résultat existant dans Wikidata. Il se peut aussi que le résultat existant n'était pas assez précis, et que, par conséquent, il n'a pas passé l'étape de filtrage. Enfin, il se peut aussi que le vecteur d'identité qui a servi à faire les requêtes vers Wikidata n'était pas assez précis ou correct.

Avec **IPSeen Accurate**, nous obtenons les résultats présentés en Exemple 9 pour notre IP exemple. Nous obtenons des résultats similaires à la version **IPSeen Fast**.

Exemple 9 Résultats obtenus avec **IPSeen Accurate**

Après le calcul des métriques et le premier filtrage, nous obtenons 1 seul résultat avec les métriques suivantes :

```
{
  "number_of_not_None_entries": 5,
  "eval_sim_name": 1,
  "eval_sim_jaro_name": 0.6631944444444445,
  "eval_sim_domain": 1,
  "eval_sim_jaro_domain": 1,
  "name_in_rdap": "False",
  "type_in_rdap": "False",
  "country_validation": "True",
  "asn_validation": "False",
}
```

Nous n'avons donc pas besoin de passer par l'étape de classement des résultats. Notre candidat devient le candidat résultat.
Le résultat obtenu pour notre exemple à l'aide de la solution présentée dans cette partie est le suivant :

```
industryLabel : ['higher education']
typeLabel : ['university in France']
```

3.3 Conclusions du Chapitre

A travers ce chapitre, nous avons tout d'abord discuté les objectifs de ce que nous souhaitons atteindre avec nos travaux. Nous les avons par la suite comparés avec la réalité du terrain pour voir ce qui est techniquement atteignable. À l'aide de cela, nous avons pu définir une série d'étiquettes qui nous permettent de différencier les différents acteurs finaux qui sont présents sur Internet. Cette proposition de classification s'appuie sur des critères aussi bien techniques qu'organisationnels. Au niveau des critères techniques, nous avons décidé de nous concentrer sur les services associés à une adresse IP, sur sa géolocalisation ainsi que sur la présumée taille de l'infrastructure. De l'autre côté, sur le côté organisationnel, nous avons décidé de qualifier l'organisation propriétaire de la machine utilisant l'adresse IP en fonction de son type d'organisation, de son domaine d'activité et de sa taille humaine. Ces informations nous semblent pertinentes et utiles pour différents contextes d'analyses. Tout d'abord, dans un contexte de sécurité. Ces informations peuvent nous aider à comprendre les politiques de sélection d'appareils participant à des activités malveillantes. Ces étiquettes peuvent aussi profiter aux analystes de log de sites Web. Ils permettent d'avoir une meilleure compréhension des acteurs qui se connectent à un site.

Nous avons ensuite proposé un outil qui nous permet, en partant d'une adresse IP, d'obtenir ces étiquettes. Notre solution est décomposable en deux sous-parties. La première sous-partie utilise des outils bien connus de la littérature pour obtenir les étiquettes techniques qui correspondent aux adresses IP : Onyphe et GeoIP. La seconde sous-partie se concentre sur l'obtention

des étiquettes organisationnelles en s'appuyant sur deux sources de données : Wikidata et les bases RIRs accessibles via le protocole RDAP. Nous avons proposé deux versions d'implémentation pour notre solution IPSeen. La première est plus rapide à exécuter et nous a permis d'expérimenter avec les différentes sources. La seconde solution nécessite des ensembles de données d'apprentissage puisqu'elle utilise du Machine Learning, mais propose des résultats plus précis, résultats qui seront présentés dans le chapitre 4, notamment aussi grâce à l'utilisation d'algorithmes de traitement automatique du langage naturel.

Nous avons présenté ici une solution pour répondre à notre problématique initiale. Nous allons, dans le chapitre suivant, évaluer cette proposition de solution. Nous allons aussi énoncer les limites de notre solution ainsi que des pistes pour l'amélioration de celle-ci.

Chapitre 4

Évaluation des sources de données et de notre solution

Dans ce chapitre, nous proposons une évaluation de **IPSeen Fast** et **IPSeen Accurate**, dont les algorithmes ont été proposés dans le chapitre 3. Nous devons également questionner la qualité des données externes utilisées par IPSeen : données d'enregistrement (RIR), informations techniques (crawlers), données collaboratives publiques (Wikidata).

En ce qui concerne les informations techniques récupérées, nous avons vu dans l'état de l'art que les outils utilisés sont bien connus de la communauté scientifique. C'est le cas notamment pour notre base de géolocalisation étudiée [99]. Nous avons aussi vu que les solutions de géolocalisation sont très performantes au niveau de la localisation par pays [99], ce qui, dans notre cas, est suffisant. En ce qui concerne les crawlers Onyphe, Shodan et Cencys, nous pouvons aussi avoir confiance et utiliser l'un ou l'autre de manière indifférente. En effet, Cencys est une solution proposée par l'université du Michigan[103] qui a, par la suite, été commercialisée. Nous avons aussi étudié une comparaison proposée par Onyphe entre Shodan et Onyphe [1]. De manière générale, Shodan et Cencys sont des outils utilisés, et validés par la communauté scientifique [111, 164, 165, 166].

En ce qui concerne les bases RIRs, elles sont construites par les RIRs au moment de leur enregistrement. Elles sont donc forcément correctes. Nous avons pu remarquer qu'elles n'étaient pas toujours à jour. Mais nous pensons avoir pu gommer ce défaut grâce à la richesse des données de Wikidata. Du fait de la nature de Wikidata, qui est une base de donnée collaborative, nous pouvons légitimement nous poser la question de la qualité des données qui y sont contenues. Pour cela, nous avons réalisé une évaluation de la qualité des données qui sera présentée en section 4.1 de ce chapitre.

Enfin, nous nous devons d'évaluer notre solution **IPSeen Fast** et **IPSeen Accurate**. Pour cela, nous avons utilisé un ensemble de données entièrement qualifié, qui sert de contrôle et qui est présenté en section 4.2. L'évaluation de la qualité des résultats d'IPSeen est réalisée dans les sections 4.3 à 4.6. Pour structurer cette présentation, nous nous sommes basés sur l'algorithme général d'IPSeen (**IPSeen Fast** et **IPSeen Accurate**), rappelé en Figure 4.1. Sur cette Figure, nous avons 5 points d'évaluation notés A à E, qui correspondent à des étapes clés de notre solution. Chaque point est abordé dans l'une des sections suivantes. Ils correspondent chacun à l'évaluation des résultats d'un bloc de notre solution.

Pour finir ce chapitre, nous proposons une lecture des problèmes identifiés ainsi que différentes pistes de résolution. Ces pistes ne sont pas implémentées et ne sont pas forcément

1. https://www.onyphe.io/docs/write-ups/onyphe-vs-shodan

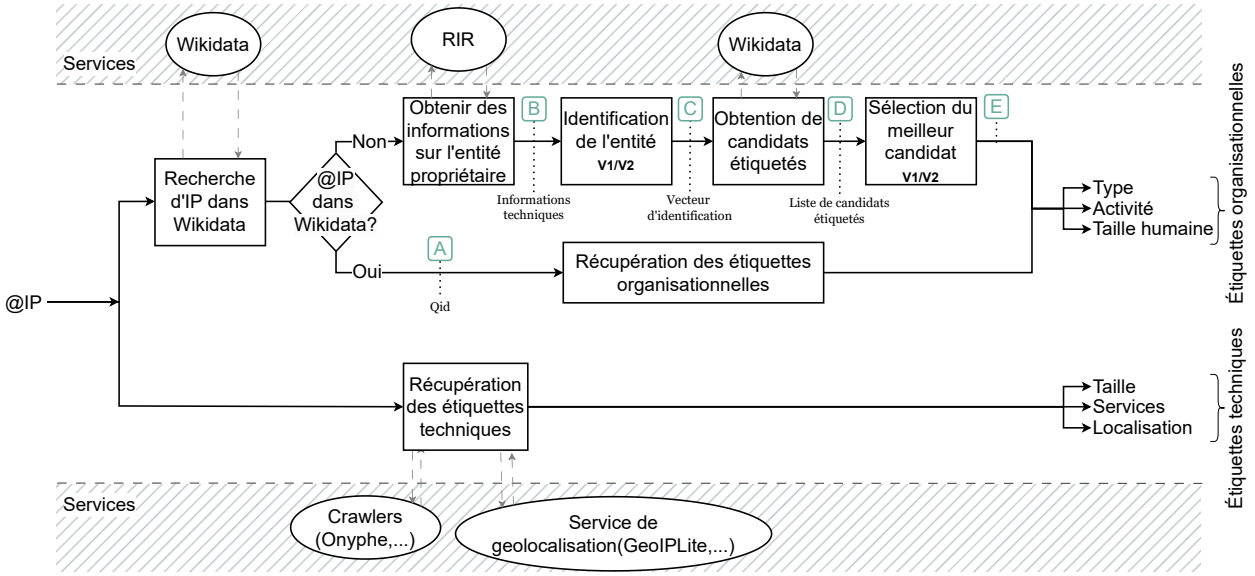


FIGURE 4.1 – Vue d'ensemble de l'algorithme et des points d'évaluation

liées à notre implémentation d'IPSeen. En revanche, elles nous permettent de prendre du recul sur notre outil et d'établir des politiques de bonnes pratiques en ce qui concerne la qualification d'adresse IP.

4.1 Point A : Évaluation de la qualité des données de Wikidata

Pour rappel, Wikidata est une base de données collaborative, c'est-à-dire que tout utilisateur peut y contribuer en modifiant/ajoutant des informations. Wikimedia, la fondation à l'origine de Wikidata, propose de nombreuses méthodes pour lutter contre les mauvaises contributions dans sa base. Cela passe, par exemple, par la détection de vandalisme des données [167].

Néanmoins, la construction de notre solution suppose que les informations contenues dans Wikidata sont correctes puisque nous renvoyons directement le résultat lorsque l'adresse IP est contenue dans Wikidata. Il est donc primordial que les adresses IP renseignées soient correctes. Nous proposons donc une évaluation qualitative des données concernant les adresses IPs contenues dans Wikidata.

Nous commençons par décrire l'ensemble de données utilisées, puis nous détaillons le protocole d'évaluation, puis nous présentons les résultats.

4.1.1 Ensemble de données des IPs contenues dans Wikidata

Afin de vérifier cela, nous avons extrait toutes les adresses IP contenues dans Wikidata (2228 blocs d'adresses) à l'aide de la requête SPARQL visible sur le Listing 4.1. Parmi ces adresses IP, nous avons retiré les plages qui correspondent aux RIRs puisqu'il ne s'agit pas d'entités finales. Nous avons aussi retiré toutes les plages qui correspondent à des adresses privées, qui sont aussi déclarées dans Wikidata, ou encore les adresses réservées comme l'adresse de loopback[2].

2. https://www.wikidata.org/wiki/Q1758736

LISTING 4.1 – Requête SPARQL de récupération de l'ensemble des IPs de Wikidata

```
SELECT ?plage_IPv4 ?item ?itemLabel ?itemAltLabel ?rank WHERE {
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
  OPTIONAL {
    ?item p:P3761 ?statement.
    ?statement ps:P3761 ?plage_IPv4;
      wikibase:rank ?rank.
  }
}
```

Une fois le filtrage réalisé, nous avons un ensemble de données exploitable contenant plus de 2000 plages d'adresses IPs. Cet ensemble de données est nommé par la suite ensemble de données Wikidata.

4.1.2 Protocole d'évaluation des données de Wikidata

Pour évaluer la qualité des informations contenues dans Wikidata, nous avons besoin d'une source à laquelle nous pouvons comparer les informations contenues dans Wikidata. Pour nous, aucune source ne représente mieux la vérité du terrain que les informations contenues dans les bases de données des RIRs puisque ce sont ces entités qui sont responsables de l'attribution des adresses IPs. Nous comparons donc les données de notre ensemble de données de Wikidata aux informations obtenues pour les plages d'adresses IP via le protocole RDAP qui interroge les bases de données des RIRs.

Il s'agit en réalité du processus inverse que celui que nous avons implémenté dans IPSeen. Nous partons des entrées de Wikidata, dont les champs sont parfaitement détaillés, et nous cherchons dans les entrées du JSON RDAP le nom de notre entrée Wikidata. Le problème est donc inversé et beaucoup plus facile. Et la méthode de résolution n'est pas redondante avec l'usage que nous faisons d'IPSeen. Le Protocole 1 illustre notre protocole d'évaluation.

Protocole 1 Protocole de correspondance des adresses IPs contenues dans Wikidata

Entrée : Ensemble de données de Wikidata
Protocole :
Pour chaque bloc d'adresses IP, on prend la première adresse machine du bloc.

1. On récupère les informations auprès des RIRs avec le protocole RDAP.
2. Si on trouve le nom de l'entité dans le résultat :
   (a) Correspondance
3. Sinon Si on trouve une équivalence partielle :
   (a) Correspondance
4. Sinon Si validation manuelle d'une correspondance :
   (a) Correspondance
5. Sinon :
   (a) Pas de correspondance

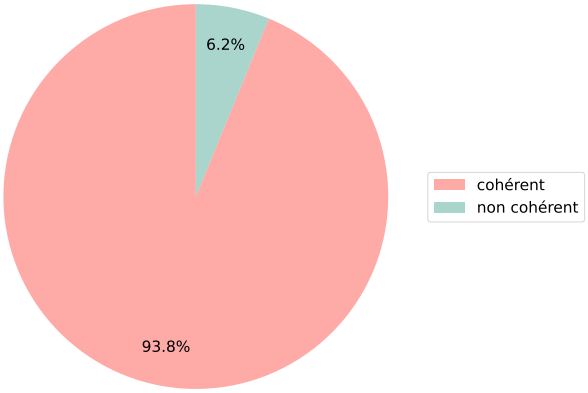


FIGURE 4.2 – Comparaison des résultats obtenus via Wikidata et RDAP pour une même IP

En suivant ce protocole, nous commençons par récupérer le JSON d'information sur le propriétaire via le protocole RDAP. Par la suite, nous croisons les informations de noms disponibles dans Wikidata avec le texte structuré du JSON de résultat. L'objectif ici est de confirmer que nous arrivons bien à retrouver le nom de l'entité dans le JSON. Si nous trouvons le nom de l'entité, nous pouvons affirmer que la plage d'adresse IP est correcte. Nous essayons aussi avec les informations des champs *also known as* de Wikidata si le premier essai n'a pas été concluant.

Nous utilisons la première adresse de chaque plage. Cela nous permet d'obtenir une intuition rapide. Nous avons aussi fait ce choix, car nous sommes limités en nombre de requêtes journalières par certains RIR (notamment RIPE NCC).

4.1.3 Résultats de l'application du protocole d'évaluation de Wikidata

En suivant le protocole 1, nous obtenons les résultats de la Figure 4.2. Nous pouvons voir que dans 93% des cas, nous trouvons une correspondance parmi les résultats. Dans les cas restants, nous n'avons pu identifier de cohérence. Il y a plusieurs raisons à cela. Tout d'abord, parmi l'ensemble de données, nous avons de nombreuses universités qui font régulièrement appel à des prestataires réseaux externes. La déclaration au niveau des RIRs peut être faite par ces partenaires, et le nom de l'université n'est alors pas mentionné dans les résultats obtenus.

La méthodologie d'évaluation ainsi que le code sont disponibles sur GitLab[3].

En conclusion, nous pouvons utiliser les données de Wikidata avec assurance. Les données contenues dans la base possèdent un haut niveau de fiabilité. Nous pouvons donc conclure que la source est pertinente et que nous pouvons valider notre approche qui considère Wikidata comme une source fiable.

4.2 Ensemble de données pour l'évaluation des points B à E

Afin d'évaluer notre solution, nous avons besoin d'un ensemble de données entièrement qualifié, c'est-à-dire d'un ensemble de données qui contient l'association entre des plages

3. https://gitlab.inria.fr/cmoriot/lcn_2022

d'adresse IP et des Qid de Wikidata. Nous avons choisi d'utiliser les Qid de Wikidata plutôt qu'un ensemble d'étiquettes, puisque ainsi, nous pouvons confirmer que notre solution identifie bien la bonne organisation, et non seulement les bonnes étiquettes. En effet, si plusieurs entités peuvent avoir les mêmes étiquettes (par exemple ISP), aucune ne peut avoir les mêmes Qid.

4.2.1 Choix d'un ensemble de données représentatif de la réalité

Pour pouvoir valider nos résultats, nous avons besoin d'un Oracle, c'est-à-dire d'un ensemble de données qualifié de manière sûre. Nous pourrions faire un tirage aléatoire d'IP. Cependant, afin d'obtenir un ensemble de données non biaisé, nous aurions eu beaucoup d'informations à prendre en considération : origine géographique, nature du réseau, taille, etc. Mais nous avons mieux. En effet, nous avons à notre disposition un ensemble de données fourni par Sekoia [4], entreprise de cybersécurité. Pour les besoins de notre travail, ils ont accepté de le rendre public. Nous avons décidé d'utiliser un ensemble de données provenant d'observation d'activité malveillante, car nous pensons que c'est sur ce type d'adresse IP que notre outil doit être performant. Prendre un ensemble de données réelles nous permet aussi d'évaluer notre algorithme dans un contexte d'attaque.

Cet ensemble de données contient des adresses IPs participant à des actions malveillantes. En plus des IPs, il y a un label désignant le type d'activité malveillante et un label contenant la date d'observation de l'activité. L'ensemble de données fourni par Sekoia comporte plus de 73 000 IPs et sont classées en 133 types d'attaque. Nous n'utilisons pas l'ensemble de données dans son ensemble, car la caractérisation manuelle de l'ensemble de cet ensemble serait trop longue. Nous avons choisi de qualifier un sous-ensemble qui s'approche des 1000 adresses IPs. Ce nombre nous parait un bon compromis entre un ensemble de données étant qualifiable manuellement et un ensemble de données suffisamment varié pour évaluer notre algorithme. Nous avons sélectionné les adresses IPs qui sont labellisées comme ayant participé au botnet Quack (label de type d'activité malveillante) puisqu'il s'agit de l'ensemble de données qui s'approche le plus de 1000 adresses IP uniques. Nous reviendrons plus en détail dans le chapitre suivant sur l'ensemble de données de Sekoia ainsi que sur ce label (Quack) en particulier.

L'ensemble de données contient 8 122 lignes, regroupées en 934 adresses IPs uniques. Dans la suite de ce chapitre, l'ensemble de données sera nommé ensemble de données de test.

4.2.2 Qualification de l'ensemble de données de test

Parmi les IPs de notre ensemble de données, certaines sont qualifiables très rapidement, car elles sont déclarées dans Wikidata. Nous avons vu dans la Section 2.1 que les IPs de Wikidata sont fiables. Nous ne les re-contrôlons donc pas.

En ce qui concerne l'attribution du reste des Qid aux adresses IPs, nous avons suivi le Protocole 2.

4. https://github.com/SEKOIA-IO/Sekoia.io_ipseen_dataensemble

Protocole 2 Protocole de qualification de l'ensemble de données de test

Pour toute adresse IP non contenue dans Wikidata.

1. Exécution de l'algorithme **IPSeen Fast** pour l'adresse IPs.
2. Si résultat :
   (a) Vérification manuelle du résultat.
   (b) Si correct :
      i. attribution du Qid à l'adresse IP.
   (c) Sinon :
      i. on passe à l'étape 3.
3. Sinon : recherche manuelle d'un résultat potentiel.
   (a) Si on trouve un résultat :
      i. attribution du Qid à l'adresse IP.
   (b) Sinon :
      i. Résultat null pour cette adresse IP.

Dans ce Protocole 2, par la vérification des résultats (étape 2.a), on entend la cohérence entre les résultats de RDAP et de Wikidata données par **IPSeen Fast**. Afin de mener cette évaluation, nous extrayons en plus d'IPSeen les informations de l'organisation obtenues via RDAP. Nous restons toujours sur la même hypothèse raisonnable qui est que les résultats donnés par RDAP sont corrects. Nous cherchons donc le nom de l'entité retourné dans les résultats RDAP obtenu pour l'IP. Mais aussi d'autres détails, comme par exemple une correspondance au niveau du pays. Par exemple, dans Wikidata, nous avons les entités suivantes : Vodafone (Q122141), Vodafone Spain (Q7939295), Vodafone Italy (Q4015982), Vodafone Netherlands (Q17031826), Vodafone India (Q2719715), Vodafone Germany (Q2529830). Il existe encore d'autres subsidiaires de l'entreprise Vodafone dans d'autres pays/régions du monde. Dans notre étape de recherche manuelle, nous vérifions ce type de cas.

En ce qui concerne l'étape de recherche manuelle (étape 3), nous avons utilisé différents outils de recherche. Tout d'abord, après observation des résultats obtenus via RDAP où nous identifions le nom de l'entreprise, nous avons effectué une recherche manuelle du résultat dans Wikidata. Si nous ne trouvons rien, nous faisons une recherche sur Wikipédia du nom de l'entreprise. Il est possible qu'un lien existe entre une entrée Wikipédia et une autre entrée Wikidata. Nous avons aussi utilisé les sites web des entreprises directement pour identifier un potentiel rachat d'une entité par une autre qui pourrait être contenue dans Wikidata. Enfin, la dernière étape de recherche manuelle utilise Google Maps. Nous venons reporter les adresses postales données dans RDAP sur Google Maps pour tenter d'identifier un nom d'entreprise, d'association ou d'université. Si aucune des méthodes précédentes ne s'avère concluante, alors nous considérons qu'il n'y a pas de résultat dans Wikidata pour l'entité identifiée.

En suivant ce protocole, nous obtenons les résultats présentés sur la Figure 4.3. Nous avons réussi à qualifier 805 adresses IPs de notre ensemble de données. 124 adresses IPs étaient contenues dans Wikidata, 448 adresses IPs ont été correctement qualifiées par **IPSeen Fast**.

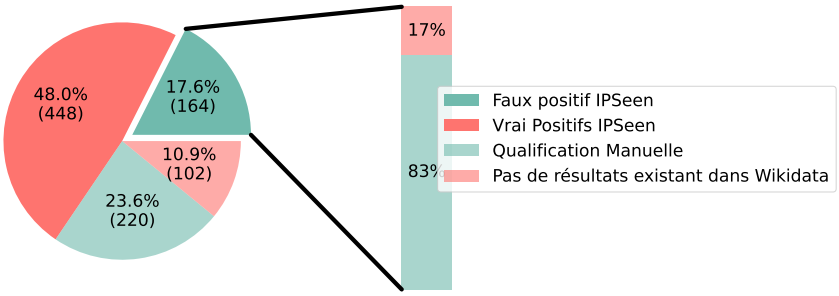


FIGURE 4.3 – Résultat de la caractérisation de l'ensemble de test

Parmi le reste des adresses, 17.6% ont été mal qualifiées par **IPSeen Fast**. Nous avons pu requalifier manuellement 135 adresses IPs et qualifier 220 adresses IPs qui n'avaient pas obtenu de résultats. Il reste donc 129 adresses IPs qui n'ont pas obtenu de Qid, car aucune entité correspondante n'existe dans Wikidata. Cette qualification manuelle a pris une semaine.

4.3 Point B : Évaluation du potentiel des informations obtenues via RDAP

En partant d'une IP, le premier bloc de l'algorithme d'IPSeen ne calcule rien et renvoie un JSON (Point B de la figure 4.1). La question ici est d'estimer s'il est possible de faire le lien entre ce JSON et Wikidata. Nous évaluons ici l'usage de RDAP en tant que source d'information. Nous souhaitons connaître, dès la source, le pourcentage d'IP qu'il sera possible de caractériser et donc de valider la correspondance entre nos deux sources d'information.

Ici, nous évaluons la qualité des informations que nous obtenons à la suite de l'étape de récupération des informations sur l'entité propriétaire. Nous cherchons à quantifier le nombre d'entrées de notre ensemble de données qui pourrait aboutir au bon résultat. Pour cela, nous cherchons à retrouver le nom de l'entité dans les données fournies par RDAP. En effet, si nous sommes en capacité de trouver le nom de l'entité, alors notre processus devrait pouvoir aboutir. Ce n'est pas le cas lorsque nous ne trouvons pas le nom de l'entité.

4.3.1 Protocole d'évaluation

Afin d'estimer le pourcentage d'IP qu'il sera possible de caractériser, nous mesurons pour les résultats la présence du nom de l'entreprise (ou des noms contenus dans le champ *also known as*) dans les résultats de RDAP. Car si le nom est présent, alors nous avons une potentielle chance de l'identifier par la suite pour l'utiliser dans une requête vers Wikidata. Afin de mesurer ce potentiel de caractérisation, nous suivons le protocole 3.

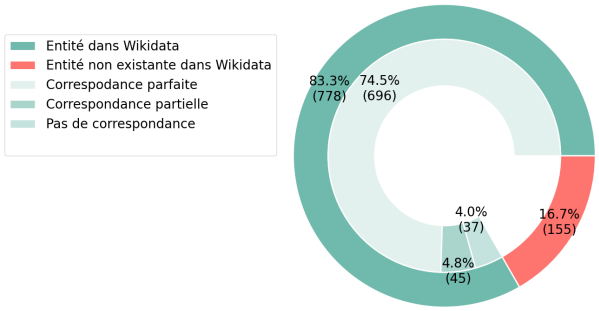


FIGURE 4.4 – Évaluation du potentiel des données de RDAP

Protocole 3 Protocole d'évaluation du potentiel des informations obtenues via RDAP

Entrée : Ensemble de données de test *Protocole :*
Pour chaque IP de l'ensemble de données totalement qualifié :

1. Obtention des informations sur l'entité propriétaire via IPSeen.
2. Recherche du nom de l'organisation, de ses alias correspondants à l'IP dans l'ensemble de données qualifié dans les résultats obtenus via RDAP.
3. si on trouve :
   (a) Nous avons une correspondance parfaite pour le résultat.
4. Sinon, On cherche à identifier une correspondance partielle
5. Si on en trouve une :
   (a) Nous avons une correspondance partielle.
6. Sinon :
   (a) nous n'arrivons pas à retrouver une chaîne de caractères nécessaire pour identifier correctement l'entité.

Nous définissons *une correspondance partielle* comme une chaîne de texte qui renverrait toujours le résultat correct si elle était utilisée dans une requête Wikidata. La définition est la suivante : si cette chaîne a été utilisée comme nom d'entité dans le contexte du bloc *Identification de l'entité,* la requête SPARQL renverrait le bon *Qid* parmi ceux figurant dans les résultats de Wikidata.

Ce protocole se différencie du protocole d'évaluation des données de Wikidata puisque ce ne sont pas les mêmes ensembles de données qui sont évalués, et de plus, ici, nous ne cherchons pas seulement une correspondance entre les données de notre ensemble de données ainsi que les informations dans RDAP. Nous cherchons à établir si une chaine de caractères présente dans les entrées RDAP pourrait nous permettre d'effectuer une recherche textuelle via une requête SparQL qui nous permettrait d'obtenir le Qid correct parmi la liste de candidats potentiels.

4.3.2 Résultat de l'application du protocole d'évaluation

Les résultats sont présentés dans la Figure 4.4. Ce bloc peut identifier parfaitement 696 sur 778. Le total n'est pas 934 (pour rappel : taille de l'ensemble de donnée), car dans cette partie, seuls 778 résultats de notre ensemble de donnée de test ont pu être qualifiés, donc nous pouvons évaluer seulement cette proportion de notre ensemble de donnée de test. En ce qui concerne *une correspondance partielle*, 45 des 82 restants ont été trouvés.

Pour les autres cas, nous n'avons pas pu établir une corrélation entre ce qui se trouve dans les résultats RDAP et dans Wikidata. Notre ensemble de données, qualifié manuellement, pourrait contenir des liens qui n'ont pas pu être récupérés automatiquement. Par exemple, il peut s'agir peut-être d'un lien entre une filiale et son organisation mère qui n'apparaît pas dans Wikidata. Il peut également s'agir de sociétés ayant changé de nom. Enfin, il peut s'agir d'informations contenues dans le RDAP qui ne sont pas suffisamment proches des noms déclarés dans Wikidata et qui ne peuvent donc pas être utilisées dans le contexte susmentionné.

Cette évaluation est valable pour les deux versions d'IPSeen proposées puisqu'il s'agit de la même méthodologie appliquée pour l'exécution de cette partie.

En conclusion de cette évaluation, à ce stade (point B) du parcours, nous pouvons donc considérer que 79% des IPs peuvent être qualifiables par IPSeen (pour les deux versions) en utilisant une recherche via le nom de l'entité. Dans les sections suivantes, nous déterminons le pourcentage d'entre elles qui arrivent réellement à être caractérisées par les deux versions d'IPSeen.

4.4 Point C : Qualités des informations d'identification extraites de RDAP

Dans cette section, nous évaluons le bloc d'obtention des informations sur l'entité propriétaire. Pour rappel, en entrée de ce bloc, nous avons les informations techniques obtenues via RDAP et en sortie, nous avons un vecteur d'identification de l'entité. Pour valider la qualité du vecteur d'identification, nous devons évaluer la validité du nom d'entité candidat et de la liste des noms de domaine extraits. Il s'agit ici de déterminer si nous avons trouvé un résultat qui nous permet, en se projetant dans le bloc suivant (obtention des candidats), d'obtenir le Qid parmi la liste des résultats.

Nous avons deux propositions pour l'étape d'identification de l'entité (associée à **IPSeen Fast** et **IPSeen Accurate**). Nous allons donc évaluer les deux propositions et comparer les résultats obtenus.

4.4.1 Protocole d'évaluation

Afin d'évaluer nos algorithmes, nous définissons les termes suivants :

— une correspondance parfaite : l'égalité entre le nom de l'entité candidate et le nom de l'élément (ou de son appellation) dans Wikidata. Avec cette égalité, nous retrouvons le résultat parmi la liste de candidats lors de l'étape *Obtention de candidats étiquetés* du processus.

— une correspondance partielle : une chaîne de texte qui, lorsqu'elle est utilisée dans une requête Wikidata, donnerait néanmoins le Qid en retour, même si celle-ci n'est pas le nom de l'entité donnée dans Wikidata.

En plus des noms, nous comparons également les noms de domaines que nous trouvons. Nous évaluons la pertinence de ces noms de domaine, c'est-à-dire, si nous les utilisons pour faire une recherche sur la base de données de Wikidata, nous pouvons retrouver l'entité voulue.

Nous décrivons à présent le protocole 4 qui nous permet de réaliser cette évaluation. Nous évaluons le nombre de correspondances parfaites et partielles pour l'ensemble des adresses IP de l'ensemble de données de test. Nous commençons par établir le nombre de correspondances parfaites. Pour cela, nous recherchons les noms des organisations ainsi que leurs alias (*also known as*). Par la suite, nous allons chercher à établir le nombre de correspondances partielles. Pour cela, nous allons utiliser la chaîne de caractères identifiée comme étant le nom de l'entité, et nous allons effectuer la requête Wikidata. Nous allons voir dans les résultats si nous sommes en capacité de trouver le bon Qid parmi les résultats. Enfin, parmi les résultats restants qui ne correspondent ni à une correspondance parfaite, ni à une correspondance partielle, nous allons comparer les noms de domaine obtenus avec les noms de domaine que nous avons à disposition dans Wikidata. Cela nous permettra d'évaluer le supplément de résultat apporté par l'analyse des noms de domaines.

Protocole 4

Entrée : ensemble de donnée de test
Protocole : Protocole d'évaluation de la qualité du vecteur d'identification
Pour chaque IP de l'ensemble de données totalement qualifié :

1. Obtention du vecteur d'identification via IPSeen (pour la version évaluée).
2. Nous comparons le nom d'entité obtenue avec le nom et les alias de notre entrée Wikidata.
3. Si l'une d'entre elle est identique avec le nom compris dans notre vecteur d'identification :
   (a) Nous avons une correspondance parfaite pour le résultat.
4. Sinon, on cherche à identifier une correspondance partielle
5. Si on en trouve une :
   (a) Nous avons une correspondance partielle.
6. Sinon Si : on a une correspondance au niveau des noms de domaines
   (a) On a une correspondance des noms de domaines.
7. Sinon :
   (a) Pas de correspondance.

4.4.2 Résultats de l'application du protocole :

À présent, nous appliquons notre méthodologie pour les deux versions proposées d'IPSeen. Nous allons pouvoir par la suite comparer les deux versions.

4.4.2.1 Résultat pour IPSeen Fast

La méthode que nous évaluons avec **IPSeen Fast** correspond à l'occurrence la plus grande de la même chaîne de caractères que nous supposons être le nom de l'entité, comme vu en

Correspondance parfaite	112
Correspondance partielle	280
Correspondance des noms de domaines	149
Total	778

TABLE 4.1 – Résultat de l'évaluation pour **IPSeen Fast** au niveau du point C

Correspondance parfaite	250
Correspondance partielle	255
Correspondance des noms de domaines	138
Total	778

TABLE 4.2 – Résultat de l'évaluation pour **IPSeen Accurate** au niveau du point C

section 3.2.4.1.

Pour cette version, les résultats sont présentés dans la table 4.1. Nous identifions 112 correspondances parfaites, 280 correspondances partielles et 149 correspondances entre les noms de domaine. Le nombre total d'IP analysé étant de 778 adresses IPs puisqu'il s'agit de la portion de l'ensemble de données pour laquelle nous avons pu attribuer un Qid, et donc avoir les informations pour mener cette évaluation.

Nous pouvons donc, avec cette version, obtenir des résultats d'une qualité suffisante pour 541 adresses IP sur 778. Cela correspond à presque 70%. Nous pouvons donc constater qu'il existe une marge de progression pour cette étape. Néanmoins, cette version présente une rapidité d'exécution comme avantage principal.

4.4.2.2 Résultat pour IPSeen Accurate

Pour cette version d'IPSeen, nous devons évaluer l'utilisation de Spacy, ainsi que le filtrage que nous réalisons par la suite, ainsi que le choix de l'occurrence maximale. Si dans l'exemple présenté dans la section 3.2.4.2 la chaîne de caractères identifiée par cette version d'IPSeen s'avérait moins précise, nous allons voir dans cette section qu'en majorité, cela n'est pas le cas.

Pour cette version, les résultats sont présentés dans la table 4.2. Avec cette nouvelle version, nous obtenons 250 correspondances parfaites, 255 correspondances partielles et 138 correspondances entre les noms de domaine. Sur les 778 adresses IPs, nous obtenons donc des résultats satisfaisants pour 82% des adresses IPs.

4.4.2.3 Comparaison des résultats pour les deux versions d'IPSeen

La Figure 4.5 propose une comparaison entre les deux versions d'IPSeen. Comme on peut l'observer, la version de l'algorithme utilisant Spacy est plus précise. Cette version permet d'obtenir plus de deux fois plus de correspondances parfaites. Le nombre de correspondances partielles est juste inférieur, car certaines correspondances partielles sont devenues des correspondances parfaites. La quantité de domaines pour **IPSeen Accurate** est légèrement inférieure à la quantité pour **IPSeen Fast** car nous n'examinons que les adresses IP manquantes, c'est-à-dire les adresses IP qui n'ont pas été incluses dans les parties précédentes.

Nous pouvons donc constater une progression entre les deux versions. Néanmoins, cette progression a un coup : le temps d'analyse nécessaire. Une comparaison des temps d'exécution

IPSeen Fast	IPSeen Accurate
3s	1h25min

TABLE 4.3 – Comparaison des temps d'exécution pour l'ensemble de l'ensemble de données de test

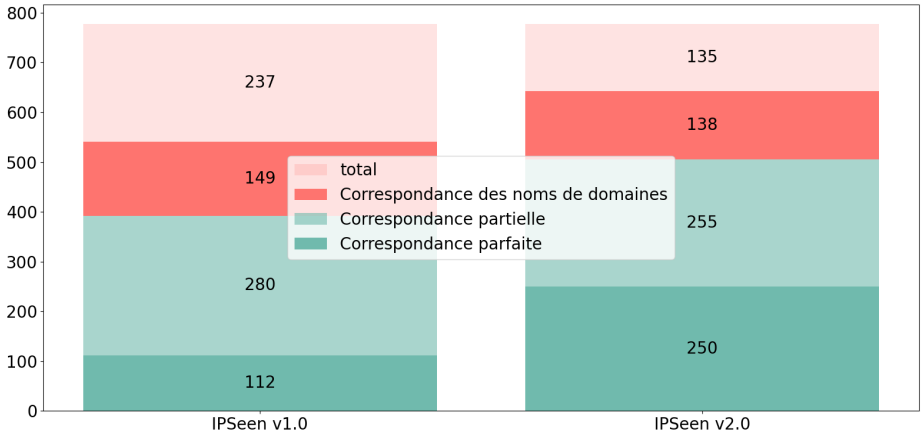


FIGURE 4.5 – Organisation pertinente identifiée à partir de RDAP
Le total représente les 778 IPs caractérisable dans notre ensemble de données

pour l'ensemble de l'ensemble de données est présentée dans la table 4.3. En effet, la première version utilise une méthode d'analyse assez rapide : le temps d'exécution est totalement négligeable par rapport au temps d'exécution des requêtes RDAP. Ce n'est pas le cas pour la deuxième solution. Il est nécessaire d'entraîner notre modèle à la reconnaissance des chaînes de caractères et le résultat prend également plus de temps à être généré. C'est pourquoi cette solution est beaucoup plus performante à partir du moment où l'on souhaite analyser un grand nombre d'adresses IPs, puisque l'entraînement n'aura lieu qu'une seule fois pour l'ensemble des IPs.

4.5 Point D : Évaluation de l'exhaustivité après les requêtes vers Wikidata

À présent, nous allons chercher à évaluer l'étape *Obtention de candidat étiqueté*. Pour rappel, cette étape a en entrée un vecteur d'identification de l'entité propriétaire de l'adresse IP et retourne en sortie une liste de candidats étiquetés obtenue via l'interrogation de Wikidata. Nous avons choisi d'évaluer l'exhaustivité de cette étape, car il est nécessaire dans cette étape que l'on obtienne un sous-ensemble des résultats de Wikidata, mais ce sous-ensemble doit contenir notre résultat. Nous évaluons pour l'ensemble des entrées de notre ensemble de données de test la proportion pour laquelle nous arrivons à avoir le bon candidat parmi la liste de candidats étiquetés.

Nous avons observé tout au long de notre expérimentation avec Wikidata certaines formes de pratique du moteur de recherche. Tout d'abord, les chaînes de caractères trop longues ne retournent pas toujours de résultat. Mais aussi, un grand nombre d'homonymes est présent

dans Wikidata. Il est donc possible pour certaines chaînes de caractères recherchées d'avoir une grande quantité de résultats, tous n'étant bien sûr pas des entités. L'objectif de cette étape était donc d'utiliser des requêtes suffisamment larges pour être certain d'avoir le bon résultat parmi les résultats tout en respectant une limite de résultats sur les requêtes de Wikidata.

Afin d'évaluer cette étape, nous proposons tout d'abord un protocole d'évaluation, puis nous l'appliquerons aux deux versions d'IPSeen. Ici, si l'étape est la même dans les deux cas, il est nécessaire de noter que, puisque les chaînes de caractères correspondant à l'entité identifiée ne sont pas les mêmes, les résultats varient d'une version à l'autre.

4.5.1 Protocole d'évaluation pour l'exhaustivité de notre solution IPSeen

Le protocole suivi pour cette évaluation est assez simple. Pour chaque chaîne de caractères et nom de domaine que nous avons récupéré suite à l'étape d'analyse des résultats de RDAP, nous effectuons les requêtes de notre bloc de requête vers Wikidata. Les résultats des différentes requêtes sont agrégés, et nous venons vérifier si oui ou non le Qid associé à l'adresse IP est présent parmi la liste des résultats. Si nous savons qu'il est présent, alors nous avons une chance d'obtenir le bon résultat lorsque nous passerons à l'étape suivante. Si ce n'est pas le cas, alors nous avons un risque de faux positifs. Si nous n'avons pas de résultats suite à l'ensemble des requêtes, alors il ne sera pas possible d'identifier le résultat correct et nous obtiendrons seulement un faux négatif ou un vrai négatif.

Le protocole 5 reprend les éléments précédents.

Protocole 5 Protocole d'évaluation de l'exhaustivité d'IPSeen

---

Entrée : ensemble de donnée de test

Protocole :

Pour chaque IP de l'ensemble de données totalement qualifié :

1. Obtention de la liste des candidats en utilisant IPSeen et le vecteur d'identification.
2. si le Qid associé à l'IP dans notre ensemble de données est dans la liste des candidats :
   (a) Qid dans la liste des candidats
3. Sinon :
   (a) Non présent

4.5.2 Résultats pour les deux versions d'IPSeen

Les deux sections suivantes présentent les résultats pour les deux versions d'IPSeen.

4.5.2.1 Résultat pour IPSeen Fast

Les résultats pour la première version d'IPSeen sont présentés dans la figure 4.6. Sur les 934 adresses IP, 778 correspondent à un identifiant Wikidata. Sur 778 Qid récupérables, dans 519 cas, nous avons trouvé le Qid parmi la liste des résultats à la suite des requêtes vers Wikidata. Nous avons donc 67% des adresses susceptibles d'être caractérisées à la suite de cette étape. Comme illustré dans la Figure 4.4, il est important de se rappeler que nous avons identifié 741

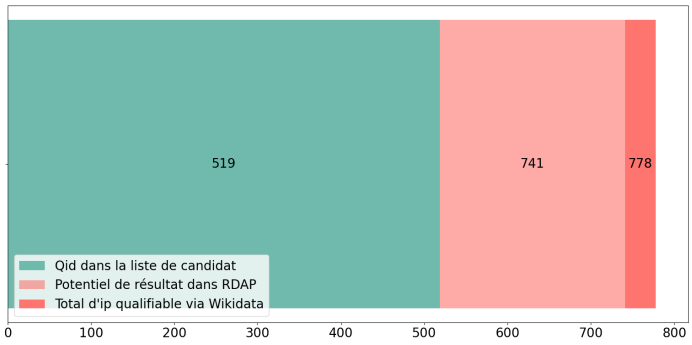


FIGURE 4.6 – Évaluation cumulée de l'exhaustivité de **IPSeen Fast**

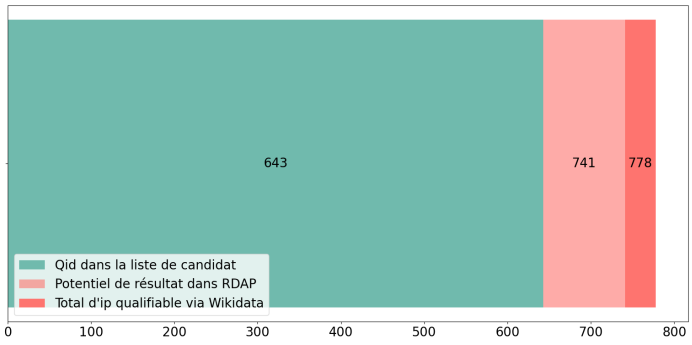


FIGURE 4.7 – Évaluation cumulée de l'exhaustivité de **IPSeen Accurate**

IPs qui présentaient une corrélation adéquate au niveau du point B afin d'obtenir le résultat correct. La portion rose clair représente donc le potentiel que nous pouvons améliorer.

Par ailleurs, nous obtenons également 44,6 Qid en moyenne par entrée de notre ensemble de données de test. Il faudra donc identifier la valeur correcte parmi cet ensemble pour chaque entrée. Même si cette valeur reste élevée, il s'agit tout de même d'un ensemble beaucoup plus petit que l'ensemble des entrées Wikidata.

4.5.2.2 Résultat pour IPSeen Accurate

Les résultats pour la seconde version d'IPSeen sont présentés dans la figure 4.7. Sur les 778 *Qid* récupérables, 643 avaient le bon *Qid* parmi la liste des *Qid* récupérés après la campagne d'interrogation de Wikidata. Cela signifie que 88 % des adresses IP susceptibles d'être caractérisées ont été trouvées après cette étape. Comme illustré dans la Figure, il est important de se rappeler que nous avons identifié 741 IPs qui présentaient une corrélation adéquate au niveau du point B afin d'obtenir le résultat correct. La portion rose clair représente donc le potentiel que nous pouvons améliorer.

Ce pourcentage est plus grand que celui que nous avions observé lors de l'évaluation de l'étape précédente. En effet, lors de l'évaluation de l'étape précédente, nous ne divisons pas les chaînes de caractères identifiées lorsque nous n'avions pas de résultat disponible. Et lors de

cette étape, nous évaluions les choix techniques que nous avions faits pour pouvoir être sûr d'avoir une chance d'obtenir le bon résultat parmi la liste de résultats à analyser.

Avec cette version, nous obtenons en moyenne 45 résultats par entrée de notre ensemble de données.

4.5.3 Comparaison des deux versions d'IPSeen

Si la brique utilisée pour récupérer les Qid est similaire dans les deux cas, du fait de la variation lors de l'identification de la chaîne de caractères représentant l'entité, nous obtenons différents résultats. Nous souhaitons comparer 2 résultats de cette évaluation.

Tout d'abord, le nombre de résultats potentiels obtenus à ce stade est pertinent à observer. Nous souhaitons voir dans quel cas nous avons un plus grand potentiel d'identification. Nous pouvons voir que c'est avec la deuxième version d'IPSeen que cette valeur est plus élevée.

Néanmoins, il est aussi pertinent de comparer pour chacune des versions le nombre moyen de résultats obtenus pour une entrée. En effet, par la suite, nous devons filtrer les résultats. Plus la liste de résultats est longue, plus nous devons redoubler d'efficacité dans notre étape de sélection du résultat. Si l'on observe cela, nous voyons qu'avec la version d'**IPSeen Fast**, nous avons en moyenne moins de résultats obtenus à la suite de cette étape. Mais la différence est négligeable.

Ici, il s'agit de faire un compromis entre ces deux valeurs. Si, effectivement, nous favorisons un ratissage large pour obtenir un maximum de chances d'avoir le bon résultat, nous devons néanmoins réduire au maximum le nombre de résultats à analyser afin d'identifier plus rapidement notre résultat correct par la suite. Les valeurs étant très proches, mais la version 2 ayant un plus grand nombre de résultats identifiables par la suite, nous préférons favoriser celle-ci, même si l'étape précédente, *Identification de l'entité*, est plus longue.

4.6 Point E : Évaluation de la sélection du meilleur résultat

Dernière étape de notre processus : la sélection du meilleur candidat. Pour rappel, en entrée du bloc, nous avons une liste de candidats étiquetés et nous devons sélectionner parmi cette liste au plus un candidat comme étant le candidat correct.

Intéressons-nous à présent aux différentes propositions techniques que nous avons détaillées en section 3.2.6 afin de sélectionner le meilleur résultat.

4.6.1 Protocole d'évaluation

Dans cette évaluation, nous souhaitons mesurer la qualité de notre choix. Pour cela, nous allons nous intéresser aux différentes métriques suivantes :

— Nombre de Vrai Positif : le Qid retourné correspond bien à celui présent dans notre ensemble de données de test,
— Nombre de Faux Positif : un Qid a été identifié comme étant le résultat, néanmoins il n'est pas correct (ce n'est pas le bon, ou il n'y en avait pas en premier lieu),
— Nombre de Faux Négatif : aucun Qid n'a été identifié par notre solution, néanmoins il existait bien un Qid qui correspondait à l'entité,
— Nombre de Vrai Négatif : Aucun Qid n'a été identifié comme étant correct et aucun résultat n'était disponible dans Wikidata.

Vrais positifs	448
Vrais négatifs	102
Faux positifs	164
Faux négatifs	220

TABLE 4.4 – Matrice de confusion pour **IPSeen Fast**

Dans notre cas d'application, les faux positifs représentent une plus grande menace que les faux négatifs. Nous avons fait le choix de favoriser et d'optimiser notre algorithme pour qu'il soit précis lors de la sélection et de la comparaison des différentes solutions de chaque implémentation. Nous préférons obtenir moins de résultats, plutôt que d'avoir des résultats qui sont faux.

Le protocole 6 décrit l'évaluation de la dernière étape de notre solution.

Protocole 6 Evaluation de la justesse de notre solution

---

ensemble d'entrée : ensemble de donnée de test, (+ensemble d'entraînement pour la v2 d'IPSeen).

Protocole :

Pour chaque IP de notre ensemble de données :

1. Obtention du résultat via le dernier bloc d'IPSeen.
2. Pour chaque résultat, on le compare au résultat de notre ensemble de données de test.
3. Si le résultat est identique :
   (a) on le classe comme vrai positif
4. Sinon Si le résultat est nul et qu'il n'y avait pas de résultat dans Wikidata :
   (a) on le classe comme vrai négatif.
5. Sinon Si IPSeen identifie un Qid, alors que le résultat de l'ensemble de données est nulle ou un autre Qid :
   (a) on le classe comme faux positif.
6. Sinnon Si IPSeen ne retourne pas de résultat, alors qu'un résultat était identifiable :
   (a) on le classe comme faux négatif.

On affiche la matrice de confusion.

4.6.2 Résultat pour IPSeen Fast

Les résultats obtenus par ce protocole via **IPSeen Fast** pour l'ensemble des adresses IPs de notre ensemble de données de test sont consultables dans le tableau 4.4. Nous pouvons y voir que nous obtenons 448 résultats corrects, 164 résultats faux, et que nous aurions pu obtenir 220 résultats supplémentaires. Nous avons donc un taux de 57% de réussite et un taux de 17% de faux positifs.

Avec l'utilisation de cette première version, nous avons vu que nous arrivons à obtenir un aperçu assez correct avec une majorité d'adresses qualifiées correctement. De plus, cette

version est beaucoup plus rapide que la seconde, car elle s'affranchit de l'étape d'entraînement des différents modèles ainsi que de la construction des ensembles d'entraînements. Car, pour rappel, les ensembles d'entraînement sont créés grâce à l'extraction des informations présentes dans Wikidata concernant les adresses IPs.

Le défaut que nous avons trouvé à cette version est le nombre de faux positifs qui est assez élevé. En effet, les faux positifs peuvent venir polluer l'analyse. Car si, dans un cas, il s'agit de résultats non trouvés (faux négatifs), alors l'analyse n'est pas complète, mais dans le cas de faux résultats retournés (faux positifs), nos résultats peuvent prédire une fausse origine du trafic. Bien sûr, parmi les faux positifs, nous en avons qui sont plus "acceptables" que d'autres. C'est le cas par exemple si nous identifions un groupe à la place de sa filière nationale. Nous avons des cas aussi différents où nous venons identifier le fournisseur de réseau d'une université plutôt que celle-ci. Dans ce cas, il nous faudrait pouvoir identifier les fournisseurs de réseau des universités. Nous pouvons bien sûr penser à la liste des réseaux universitaires NREN[5], mais elle ne comprend que les réseaux des grandes écoles ou des universités. Enfin, nous pouvons aussi établir des règles sur les résultats retournés, comme de refuser certaines catégories de résultats : objets, personnes, etc. Néanmoins, dans un but d'analyse entièrement automatisée, il serait préférable de limiter le nombre de faux positifs dès le rendu du résultat.

4.6.3 Résultat pour IPSeen Accurate

Nous évaluons la seconde version d'IPSeen ici. Pour rappel, cette version s'appuie sur l'utilisation du Machine Learning pour la sélection du meilleur candidat et nécessite un ensemble de données d'apprentissage. Ce fut présenté en section 3.2.6.2.

Nous évaluons deux choix techniques dans cette section : le choix de l'algorithme de Machine Learning, puis le choix technique d'implémenter un second algorithme de Machine Learning dans le cas où nous avons une égalité.

4.6.3.1 Choix de l'algorithme de Machine Learning

Pour sélectionner le meilleur bloc candidat, le processus utilise un algorithme ML. Pour choisir celui que nous utilisons, nous avons comparé 3 algorithmes possibles. Les résultats varient légèrement entre nos trois propositions : Random Forest (RF), Perceptron, et Multi-layer Perceptron. Dans le tableau 4.5 sont comparés les résultats des métriques standard.

À la matrice de confusion, nous ajoutons les métriques suivantes qui sont des métriques courantes du Machine Learning : Où VP désigne Vrai Positif, VN désigne Vrai Négatif, FP désigne Faux Positif et VN désigne Vrai Négatif.

1. Rappel : définie le pourcentage de résultats corrects. Un rappel élevé signifie qu'il ne ratera aucun positif. Cependant, cela ne révèle aucune information sur la qualité de sa prédiction des négatifs définie par la formule mathématique suivante :

$$Rappel = \frac{VP}{VP + FN}$$

2. Précision : définie le nombre de prédictions positives bien effectuées. Plus le nombre de faux positifs est faible dans le modèle de machine learning, plus cette métrique est

5. https://en.wikipedia.org/wiki/National_research_and_education_network

Métrique	RF	Perceptron	Multi-LayerPerceptron
Rappel	0.72	0.32	0.49
Précision	0.76	0.58	0.70
F1 score	0.74	0.41	0.58
Vrais positifs	446	196	304
Faux positifs	140	140	133
Vrais négatifs	36730	36730	36737
Faux négatifs	170	420	312

TABLE 4.5 – Métriques pour les différents algorithmes de ML proposés

élevée. La majorité des prédictions positives du modèle sont bien prédites lorsque la précision est élevée. Définie par la formule suivante :

$$Precision = \frac{VP}{VP + FP}$$

3. F1 Score : Combinaison d'évaluation du Rappel et de la Précision. Plus le F1 Score est élevé, plus le modèle est performant. Définie par la formule suivante :

$$F1Score = \frac{2 * Precision * Rappel}{Precision + Rappel}$$

Le tableau 4.5 illustre le résultat obtenu sur l'ensemble de données d'entrainement lors de l'entraînement avec un ensemble de données différent. On peut y voir le potentiel d'identification du résultat correct. En regardant le tableau des résultats, nous avons décidé de continuer avec l'algorithme de Random Forest (RF) car il y a une bonne combinaison entre le rappel et la précision pour celui-ci.

4.6.3.2 Choix d'implémentation d'un second algorithme de Machine Learning

Dans le chapitre précédent, nous avons vu que nous avions envisagé deux solutions pour traiter le cas particulier où plusieurs solutions potentielles seraient identifiées par la solution de Machine Learning.

Ici, nous comparons les deux solutions envisagées. Pour rappel, la première solution envisagée était l'utilisation des résultats du premier algorithme et la sélection du résultat avec la plus haute probabilité. Par la suite, nous désignons cette solution par *RF*. La seconde proposition consiste à implémenter un second algorithme de Random Forest. Nous désignons cette solution par *Double RF*. L'ensemble d'entraînement a été décrit en section 3.2.6.2.2.

Sur les 934 adresses IP de notre ensemble de données de test, nous évaluons le résultat obtenu à l'aide de l'algorithme RF. Les résultats sont présentés dans le tableau 4.6. Comme on peut le voir, notre méthode de sélection de la probabilité la plus élevée a permis de sélectionner 407 résultats corrects tout en réduisant le nombre de faux positifs à 88.

En ce qui concerne la solution de double RF, le nombre de faux positifs est inférieur à celui de la version précédente, de même que le nombre de vrais positifs. On peut également constater qu'il reste un potentiel de qualification, car 168 adresses IP ne se sont pas vu attribuer de *Qid*, alors que cela était possible.

Métriques	RF	double RF
Vrais positifs	407	419
Vrais négatifs	273	275
Faux positifs	88	69
Faux négatifs	166	171

TABLE 4.6 – Matrice de confusion pour RF et double RF

Métriques	IPSeen Fast	IPSeen Accurate
Vrais positifs	448	419
Vrais négatifs	102	275
Faux positifs	164	69
Faux négatifs	220	171

TABLE 4.7 – Comparaison des résultats des deux versions

4.6.4 Comparaison des deux versions d'IPSeen

Nous souhaitons à présent comparer les deux versions d'IPSeen afin d'identifier leurs singularités et leurs points forts. La Table 4.7 présente les matrices de confusions obtenues pour chacune des versions d'IPSeen ainsi que les valeurs de sensibilités et spécificités.

Les vrais positifs correspondent aux résultats qui ont été correctement affectés : le Qid identifié correspond bel et bien à celui que nous souhaitions retrouver. Les vrais négatifs correspondent aux cas où aucun résultat n'a été identifié lorsqu'il n'existait pas de résultat disponible dans Wikidata. Ces deux mesures, si ajoutées, permettent d'évaluer le nombre de résultats correctement retournés par notre solution.

De l'autre côté, les faux positifs quantifient les cas où IPSeen identifie un résultat alors que soit celui-ci n'est pas correct (c'est-à-dire qu'il existe un autre résultat), soit il n'y aurait pas à l'origine de résultat qui devrait être identifié. Enfin, les faux négatifs quantifient les cas où un résultat aurait pu exister, mais celui-ci n'a pas été trouvé.

En prenant en compte toutes ces mesures, on peut voir que la seconde version d'IPSeen est meilleure, car si celle-ci a moins de vrais positifs, le nombre de faux positifs a lui été divisé par 2. Nous avons donc deux solutions différentes. Si la première est rapide à l'exécution et à la mise en place, la seconde, elle est beaucoup plus précise.

4.7 Comparaison des résultats à l'Oracle

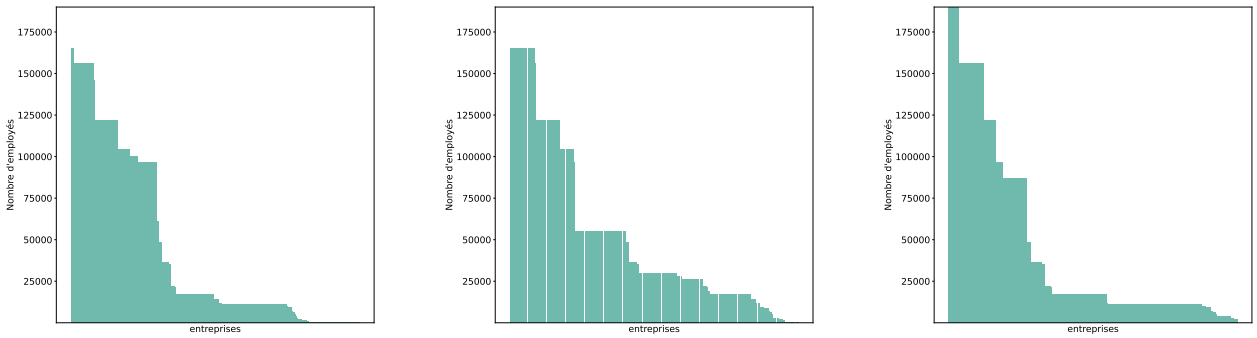
Cette section fournit les résultats de ce qui est retourné par notre solution et les compare à ce qui est la vérité de terrain (l'Oracle). La vérité de terrain est caractérisée par ce que nous avons caractérisé manuellement pour créer le jeu de données. L'objectif est de voir, ici, si nous nous rapprochons de la réalité en termes d'analyse. Nous souhaitons voir si, à cause des faux positifs et des faux négatifs, nous ne trouvons pas des résultats complètement à l'opposé de ce que nous devrions observer.

Nous allons donc dans cette section comparer trois ensembles d'étiquettes. Les trois ensembles découlent de l'ensemble de données de test. Le premier ensemble d'étiquettes correspond aux étiquettes que nous avons obtenues après la qualification manuelle de l'ensemble de données de test. Nous nommons cet ensemble l'ensemble de vérité. Le second ensemble

d'étiquettes correspond aux étiquettes obtenues après l'application de la solution **IPSeen Fast**. Nous nommons cet ensemble ensemble **IPSeen Fast**. Enfin, le dernier ensemble d'étiquettes correspond aux étiquettes obtenues après l'application de la seconde version d'IPSeen, **IPSeen Accurate**. Cet ensemble sera donc nommé ensemble **IPSeen Accurate**.

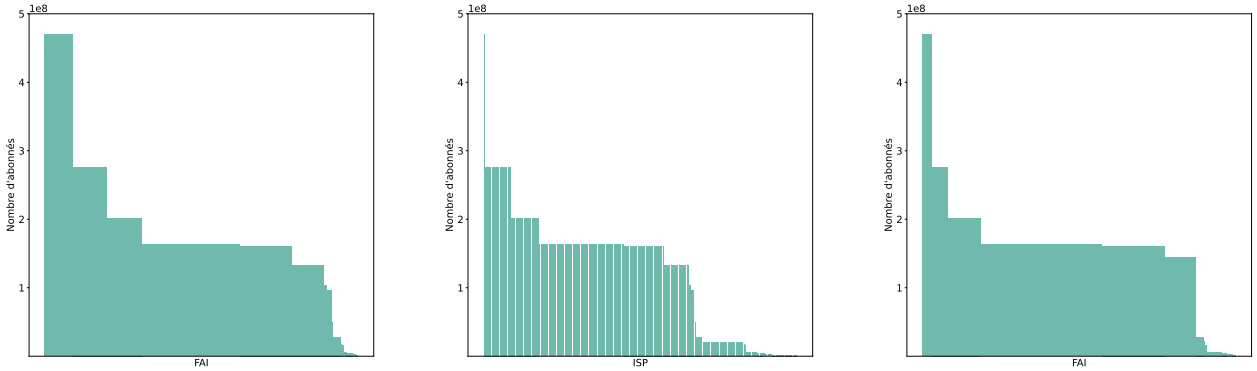
Les résultats de l'analyse obtenus par notre algorithme sont présentés dans 4.8,4.9,4.10 et 4.11. Pour chaque figure, nous affichons l'ensemble **IPSeen Fast** à gauche, l'ensemble de vérité au centre et l'ensemble **IPSeen Accurate** à droite.

Comme on peut le constater, un peu de nuance a été perdue, mais les grandes catégories sont toujours bien représentées. Nous allons à présent examiner chaque graphique.



(A) **IPSeen Fast** (B) Oracle (C) **IPSeen Accurate**

FIGURE 4.8 – Nombre d'employés par entreprise
Une barre = une entreprise, trié par ordre décroissant



(A) **IPSeen Fast** (B) Oracle (C) **IPSeen Accurate**

FIGURE 4.9 – Nombre d'abonnés par FAI

En ce qui concerne la taille humaine, caractérisée à la fois par le nombre d'employés pour les entreprises dans la figure 4.8 et le nombre d'abonnés pour les FAI dans la figure 4.9, on peut constater que pour la première, les valeurs extrêmes sont conservées, ce qui permet de garder la forme générale de la distribution.

En ce qui concerne le nombre d'abonnés, en revanche, nous obtenons principalement la valeur moyenne. En fonction de l'analyse prévue, les deux peuvent être utiles, car, en ce qui concerne le FAI, l'obtention d'une moyenne pourrait aider à calculer en moyenne combien de réseaux résidentiels pourraient être utilisés comme sources de trafic malveillant par l'intermédiaire de dispositifs corrompus. Le nombre d'employés, quant à lui, permet de mieux

comprendre la taille des entreprises confrontées à ce type d'activité malveillante. Dans le cas présent, le nombre d'employés varie entre 25 000 et 175 000. Cela correspond probablement à des multinationales ou à de grandes entreprises nationales.

(A) **IPSeen Fast**

(B) **IPSeen Accurate**

(C) Oracle

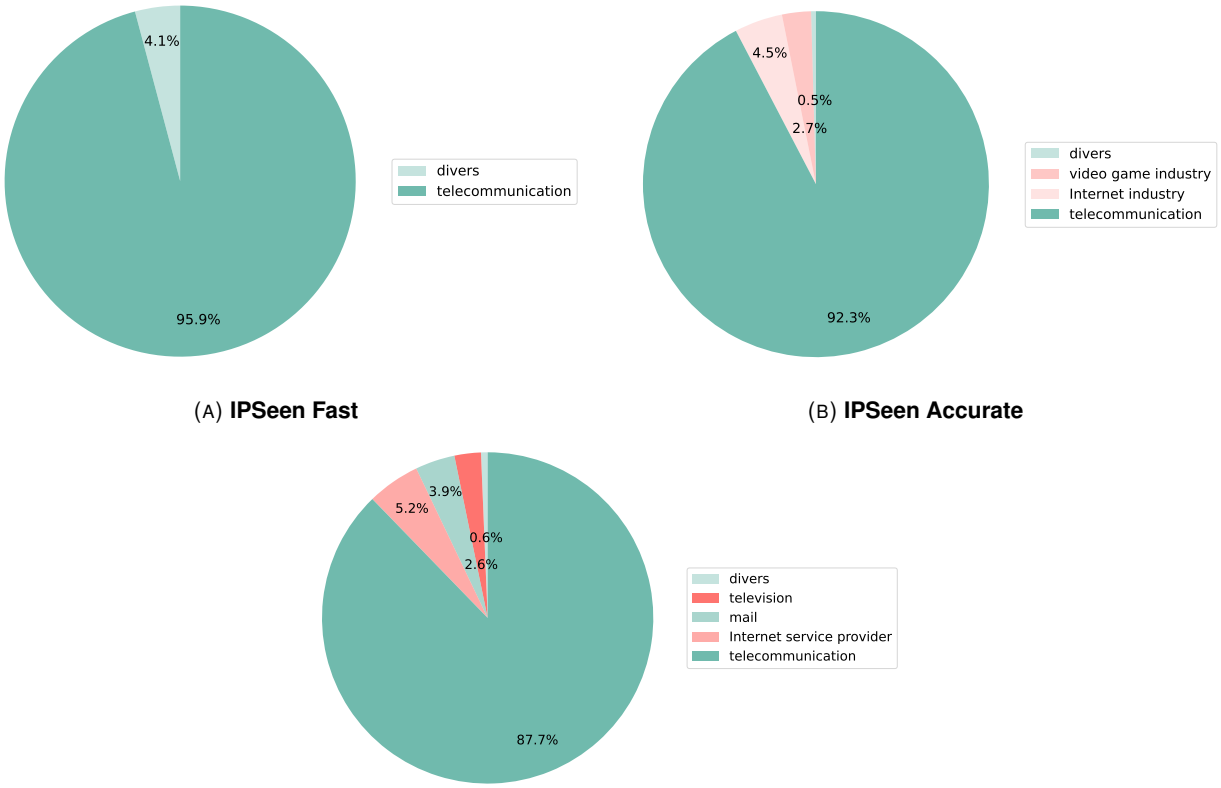
FIGURE 4.10 – Répartition par type d'organisation

Si l'on passe au type d'organisation de la figure 4.10, les catégories sont essentiellement les mêmes, avec moins de nuances sur le résultat renvoyé par IPSeen en ce qui concerne les entreprises. Avec moins de résultats attribués, les secteurs du diagramme à secteurs sont différents. Cela est dû aux résultats faussement positifs ainsi qu'au nombre réduit de résultats caractérisés. C'est également pour cette raison que l'on peut observer des catégories de type pour les deux versions d'IPSeen qu'on ne retrouve pas au niveau de l'Oracle.

Enfin, dans la figure de l'industrie 4.11, les résultats obtenus par IPSeen sont moins précis que la vérité de terrain, car il y a moins de catégories de résultats d'industries. Cela est dû au fait que les différentes catégories présentées via l'Oracle n'ont pas été obtenues. Toutefois, les résultats tendent à montrer la même chose que les résultats de base : les entreprises à l'origine de ce trafic malveillant sont très probablement des entreprises de télécommunications.

4.8 Proposition de réponse aux problèmes identifiés

Tout au long de ce chapitre, nous avons évoqué différentes limitations. Nous les rappelons ici en proposant des pistes d'amélioration.



(A) **IPSeen Fast**

(B) **IPSeen Accurate**

(C) Oracle

FIGURE 4.11 – Répartition des domaines d'activité parmi les entreprises

4.8.1 Limites identifiées via les différentes versions d'IPSeen

L'évaluation a permis d'identifier plusieurs causes à l'origine des erreurs d'identification de la chaîne de texte dans le RDAP. Nous fournirons un exemple pour chacune des limites que nous avons pu identifier.

4.8.1.1 Informations incorrectes dans les bases RIRs

Tout d'abord, comme nous l'avons vu dans la section 4.4, dans 4% des cas, les informations dans le RDAP ne sont pas suffisamment précises pour nous permettre de remonter une organisation jusqu'à Wikidata.

Plusieurs causes peuvent être à l'origine de celles-ci. Tout d'abord, cela peut être dû au fait que les informations ne sont pas mises à jour après un changement de propriétaire ou de nom d'une entreprise. Dans ce cas, si l'ancien nom de l'entité n'apparaît pas dans Wikidata, nous ne pouvons remonter jusqu'à la bonne entité. Cela vient donc de deux causes : le manque de mise à jour côté RIR et le manque d'information sur les entités qui n'existent plus dans Wikidata. Par exemple, nous avons trouvé des blocs d'adresses IPs dont l'entité propriétaire déclarée est un groupe qui a changé de nom il y a plus de 15 ans. Comme l'ancien nom était mentionné dans Wikidata comme étant un alias, nous avons pu remonter jusqu'au bon résultat, néanmoins ce n'est pas forcément le cas pour toutes les entités.

Parfois, la relation est complexe à établir, car une entité peut faire appel à des prestataires pour la gestion de son réseau. Dans ces cas particuliers, on pourra observer une dissonance entre ce qui est disponible dans Wikidata, et ce qui est déclaré auprès des RIR. C'est le cas par exemple pour certaines universités ou écoles.

Dans certains cas extrêmes, une des conclusions des observations que nous avons pu mener est, dans certains cas, le manque de rigueur dans la collecte des informations des détenteurs des blocs d'adresses IPs. Au long de notre expérimentation avec RDAP, nous avons pu rencontrer différents cas. Tout d'abord, lors de la qualification manuelle de notre ensemble de données d'évaluation, nous avons pu voir qu'il nous est parfois impossible d'identifier l'entité qui possède les adresses IPs. Nous sommes tombées sur des adresses postales qui ne tombent nulle part, sur des entités avec des adresses e-mails spécifiées, pour les signalisations d'abus en "abuse@gmail.com".

4.8.1.2 Nombre élevé de faux négatifs

Une autre limite de notre solution est le nombre élevé de faux négatifs, c'est-à-dire de résultats potentiels qui n'ont pas été identifiés. Dans 18 % des cas, aucun résultat n'a été renvoyé alors qu'un résultat aurait dû être obtenu. Cela peut s'expliquer par plusieurs raisons, et nous présentons ici les principales.

Tout d'abord, les informations provenant de RDAP identifiées appartiennent à une filiale et Wikidata ne dispose d'aucune information sur cette filiale. C'est le cas par exemple entre Numéricable et SFR en France. Nous avons identifié SFR comme résultat, alors que nous aurions pu identifier Numéricable.

Nous avons aussi des faux négatifs où une filiale régionale n'a pas pu être identifiée, mais à la place, nous avons le résultat pour le groupe dans son ensemble. C'est le cas pour Vodafone (Q122141) et Vodafone Egypt (Q1540525).

Deuxièmement, le vecteur d'identification extrait de RDAP n'est pas correct et aucun nom de domaine ne peut être utilisé (ce qui signifie qu'il faut soit extraire un élément qui n'a pas de déclaration de site web, soit que le site web a un autre domaine légèrement différent).

Troisièmement, le nom identifié est trop long et ne renvoie aucun résultat, et perd son sens lorsqu'il est décomposé en plus petites chaînes. Nous avons eu le cas par exemple lorsque nous avions identifié Cable and Wireless Jamaica, mais n'avons pu remonter à l'identifiant qui correspond à "Cable & Wireless plc" (Q1024869).

Enfin, notre solution a été entraînée sur un ensemble de données où les éléments ont beaucoup d'étiquettes. Cependant, certains des résultats que nous recherchons dans Wikidata ont moins d'informations, et donc moins d'étiquettes pour les caractériser. Pour cette raison, le résultat est filtré dans la phase de ML. Cela peut aussi être la cause des faux négatifs.

Nous avons également fait un choix d'implémentation au moment de la sélection de l'algorithme de Machine Learning qui priorise un nombre de faux positifs bas, ce qui a aussi entraîné un fort taux de faux négatifs.

4.8.1.3 Informations manquantes dans Wikidata

Sur les 934 adresses IP, seules 778 adresses IP ont reçu un résultat de Wikidata. Ce résultat a été obtenu en effectuant des recherches manuelles. Il s'agit du potentiel optimal à atteindre si toutes les informations nécessaires peuvent être extraites du RDAP ou de Wikidata. Pour les 156 adresses IP restantes, aucun *Qid* n'a pu être trouvé dans Wikidata, même avec une recherche manuelle. La base de donnée ne contenant pas des informations sur toutes les entités, nous avons forcément des adresses qui seront non qualifiées. Néanmoins, la base que nous avons choisie contient quand même un grand nombre d'informations à propos des entités, en plus de ces avantages d'utilisation.

4.8.2 Piste potentielle de résolution des limites identifiées

Suite à l'identification des limites de notre solution, nous proposons également des pistes d'amélioration qui nous semblent accessibles, mais que nous n'avons pas implémentées dans le temps de cette thèse.

4.8.2.1 Automatiser la contribution à Wikidata afin de compléter la base de données

Lorsque nous avons choisi d'utiliser Wikidata comme source d'information, nous l'avons aussi fait de par sa spécificité : la possibilité d'y contribuer. La base de données n'est certes pas complète, mais nous pouvons contribuer à l'ajout des informations manquantes. Dans le cas où notre solution n'aurait pas identifié de résultat, mais que nous prenions le temps d'analyser manuellement un résultat, nous pouvons alors compléter les informations manquantes dans la base de Wikidata. Il peut s'agir directement des adresses IPs dans le cas où l'entité existerait dans la base, mais il peut aussi s'agir d'alias pour venir compléter des informations historiques à propos des entreprises.

Une fois que les données ont été trouvées et examinées par un utilisateur de notre outil, elles peuvent être ajoutées à Wikidata et servir pour les prochaines personnes qui viendraient étudier la même adresse. Seuls points négatifs, cela peut prendre beaucoup de temps et être parfois difficile à réaliser si l'organisation n'est pas connue ou si l'information n'est disponible que dans une seule langue, car rappelons-le, nous ne gérons que la langue anglaise.

4.8.2.2 Obtenir l'étiquette de type même si l'organisation n'a pas de page Wikidata

Wikidata ne décrit pas toutes les organisations à travers le monde. Pour résoudre ce problème, il existe différentes solutions. On peut ajouter soi-même des informations ou attendre que quelqu'un de la communauté les ajoute, comme suggéré ci-dessus.

Cependant, nous proposons une autre solution, qui ne s'appuie pas sur Wikidata, pour obtenir l'étiquette de type de réseau. Comme nous l'avons vu précédemment, un des autres outils que nous avons utilisés pour faire de la qualification, Onyphe, propose de nombreuses informations à propos des adresses IPs. Cela va des services ouverts à la version des systèmes d'exploitation, en passant aussi par les informations DNS. Nous avons le sentiment que nous pouvons déterminer le type de réseaux en fonction de la typologie de celui-ci. Nous pouvons utiliser Onyphe pour établir les typologies des différents types de réseaux.

Par exemple, un réseau d'entreprise peut avoir des dispositifs particuliers qui sont habituellement utilisés par une entreprise, alors qu'un réseau résidentiel devrait avoir le même type de routeur selon le FAI. Un autre exemple peut être celui d'une entreprise dont le domaine d'activité est l'hébergement, qui peut avoir un grand nombre de noms de domaine associés aux services qu'elle héberge. Nous pensons qu'en examinant tous les champs renvoyés par Onyphe, nous pouvons trouver des particularités associées à chaque type/nature de réseau.

Nous pensons qu'en croisant le type d'entité des données disponibles dans Wikidata et les informations d'Onyphe pour le bloc d'adresses IP, une typologie de réseau associée à chaque étiquette de type peut être déterminée. Nous pensons que nous pouvons extraire des typologies de réseaux spécifiques à partir de toutes les étiquettes disponibles dans Onyphe. Nous pensons que certaines informations doivent être pré-traitées et non analysées en tant que telles. Par exemple, pour le nom de domaine associé à une série d'adresses IP, il peut être nécessaire d'examiner le nombre et les différences entre tous les noms de domaine. Par exemple, certaines entreprises peuvent avoir quelques noms de domaine, mais ils peuvent être assez similaires d'une certaine manière, ainsi que le nom d'un domaine subsidiaire. Mais leur nombre est inférieur à celui des sociétés d'hébergement.

4.9 Conclusions du chapitre

Afin de conclure ce chapitre, nous pouvons revenir sur les évaluations proposées. Nous avons évalué étape par étape, brique par brique, notre solution afin d'identifier en profondeur les points forts ainsi que les limitations de notre solution. Nous avons démontré une certaine efficacité de notre solution. Même si nous l'avons vu, celle-ci reste améliorable. Nous avons comparé les deux versions et présenté des avantages à chacune d'entre elles : la rapidité pour la version 1 et la précision pour la version 2. L'évolution entre les deux propositions d'implémentation est positive, puisque nous avons réussi à réduire grandement le nombre de faux positifs. Rappelons aussi que les résultats sont très encourageants pour la seconde version, puisque les algorithmes de ML ont été entraînés avec un ensemble de données contenues dans Wikidata, puis testés sur un ensemble de données différent qui contient des adresses qui ont participé à des attaques.

De plus, nous avons pu constater que chaque bloc, s'il est isolé, présente de très bons résultats. Néanmoins, il nous reste une marge de progression que nous avons pu définir pour chacun des blocs. Il serait aussi possible de venir combiner nos deux versions d'IPSeen afin de voir si nous pouvons obtenir de meilleures performances et de résultats.

Nous avons aussi identifié des causes d'échec qui ne sont malheureusement pas toutes sous notre contrôle. Néanmoins, pour certaines d'entre elles, nous avons proposé des solutions en exploitant par exemple la collaboration rendue possible grâce aux sources de données que nous avons choisies d'utiliser.

Nous avons aussi proposé deux pistes d'amélioration pour contourner les limites d'une des sources de données : Wikidata. Une piste s'appuie sur la collaboration, propriété de la base de Wikidata, l'autre sur une autre source que nous avons utilisée pour l'analyse technique du réseau. C'est à cela, que nous pouvons voir qu'une fois de plus, les analyses techniques et organisationnelles sont complémentaires.

Chapitre 5

Application et étude de cas

Dans ce chapitre, nous traitons de l'application de notre outil, IPSeen, dans un contexte d'analyse de sécurité. Comme nous l'avons évoqué précédemment, un des axes des analyses de sécurité se concentre sur la connaissance des infrastructures des attaquants. Dans ce chapitre, nous abordons deux thèmes. Premièrement, nous décrivons notre méthodologie pour l'obtention des captures de trafic d'attaques à analyser. Nous expliquons et justifions nos choix de critères d'acceptation pour les captures d'attaques. Puis, nous présentons les captures que nous avons obtenues et proposons un retour d'expérience sur l'obtention de ces captures. Dans un second temps, nous présentons dans ce chapitre les analyses qui ont été produites sur les ensembles de données obtenus via notre outil présenté et évalué dans les chapitres précédents. Nous proposons aussi des conclusions basées sur nos observations.

5.1 Méthodologie de collecte de traces dans un contexte d'analyse de sécurité

La motivation principale de cette thèse est l'analyse d'attaques et plus particulièrement des infrastructures utilisées pour générer des attaques. Notre outil, présenté dans le chapitre 3, a pour vocation d'être appliqué sur des traces d'attaques afin d'en caractériser plus précisément les machines participantes, et surtout d'identifier un/des points d'actions dans le réseau. Ces points pourront, par la suite, faire l'objet de déploiement de processus de sécurité afin de limiter le trafic d'attaque, et cela, dès la source. Pour cela, nous rappelons qu'il est nécessaire de relever un maximum d'informations possibles sur ces sources de trafic.

Dans le but d'atteindre cet objectif, nous avons donc besoin de trafic d'attaques à analyser. Dans la première partie de cette section, nous présentons les critères que nous avons établis pour la collecte de traces. Par la suite, nous présentons les traces de trafics d'attaques que nous avons obtenues. Enfin, nous partageons un retour d'expérience de la collecte de traces d'attaques.

5.1.1 Quelles traces collecter ?

Pour réaliser cette analyse, nous avons besoin de traces d'attaques qui répondent à un certain cahier des charges.

Tout d'abord, les traces doivent être récentes. En effet, les blocs d'adresses IPs n'étant pas toujours affectés aux mêmes entités, il est nécessaire de se rapprocher au plus près de l'instant présent. De plus, nous souhaitons faire un état des lieux du système Internet actuel et pour cela, nous nous limitons à des traces de moins de quatre ans.

Dans un second point, nous cherchons également à caractériser un maximum de sources d'attaques. Pour cela, il est nécessaire d'identifier un point de collecte stratégique dans le réseau. Nous favorisons donc les traces collectées auprès des victimes d'attaques, puisque c'est à ce point précis que nous allons pouvoir observer la plus grande pluralité de sources d'attaques. Observer auprès de la victime nous permet également d'avoir une vue sur l'ensemble des ressources exploitées pour une seule attaque et donc de pouvoir faire une classification des résultats en fonction des types d'attaques. Cela nous permet également de voir si les stratégies de sélection des machines changent en fonction du type d'attaque lancé ainsi que du type de victime visée.

Pour finir, nous avons besoin de traces réelles. Il existe quelques ensembles de données d'attaques disponibles [1] qui permettent de tester des mécanismes de défense. Néanmoins, les sources des attaques ont bien souvent été anonymisées afin de répondre aux règles du RGPD [127]. Dans notre cas d'analyse, obtenir des traces d'attaques générées n'a aucun intérêt, puisque nous essayons de rendre compte de l'état réel des attaques sur Internet. Il est donc impératif que nous qualifiions des traces réelles et non générées.

Pour répondre aux problématiques de conformité RGPD que nous avons pu rencontrer avec la transmission des traces, nous proposons une alternative qui consiste à analyser des blocs de sous-réseaux (/24 ou plus grand) afin d'apporter une certaine anonymisation qui n'impacte que très peu notre étude. En effet, nous obtenons alors des informations plus larges sur les sous-réseaux au niveau des spécificités techniques des machines, mais l'organisation propriétaire reste la même sur l'ensemble du sous-réseau.

Afin de proposer une analyse approfondie, nous voulons varier aussi les points de collecte des attaques. En effet, en fonction des points de collecte des attaques, nous pouvons potentiellement constater des différences. Nous avons émis l'hypothèse que les réseaux d'attaques peuvent varier en fonction du type de victime : entreprise, université, site Web à caractère politique, etc. Ce qui nous pousse à avoir émis cette hypothèse est que les attaquants qui s'en prennent à ces victimes ne sont pas forcément les mêmes. Il peut s'agir de grands groupes d'attaques, d'acteurs proposant des attaques comme un service, ou encore d'attaquants isolés. La variation que nous pourrions observer permet de voir si les infrastructures d'attaques varient d'un type de victime à un autre, ou d'un type d'attaquant à un autre.

Pour résumer, nous avons trois critères principaux pour la collecte des traces : la véracité, la pluralité des sources observées ainsi que l'actualité des traces.

5.1.2 Traces collectées

Afin d'obtenir des traces d'attaques, nous avons pu échanger avec de nombreux acteurs des réseaux. Parmi eux, des acteurs gérant des réseaux universitaires, des acteurs de la sécurité, des fournisseurs d'accès à Internet grand public et des hébergeurs. Il s'agit principalement d'organisations françaises.

Si parmi ces acteurs, nombreux étaient très intéressés par notre projet, nous n'avons pu approfondir notre collaboration en raison de problèmes juridiques. Nous reviendrons plus longuement sur ce détail en section 5.1.3.

1. https://github.com/shramos/Awesome-Cybersecurity-Dataensembles, https://www.unb.ca/cic/dataensembles/index.html

Nombre de labels	174
Nombre moyen d'IP par label	15270
Nombre moyen d'IP unique par label	575.5
Nombre d'IP unique observé divisé par le nombre total d'IP en moyenne	0.11
Période d'observation	03/12/2019-29/08/2022
Nombre d'adresses IP uniques	89533
Nombre de labels moyen par IP unique	1.12
Nombre d'observations moyen journalier par IP	9.9

TABLE 5.1 – Résumé des principales caractéristiques de l'ensemble de donnée de Sekoia

Afin de réaliser nos travaux d'analyses, nous avons collaboré avec Sekoia. Sekoia est une entreprise française de cybersécurité. Parmi ses activités, Sekoia a un projet de suivi des infrastructures de *commande et contrôle*[2] des plus grosses menaces.

Sekoia a mis à notre disposition un ensemble de données contenant des adresses IPs, la date de collecte de celles-ci ainsi que le type d'activité malveillante qui a été observée. Parmi cet ensemble de données, on retrouve notamment des malwares, des outils de test d'intrusion pouvant être utilisés à des fins malveillantes, des adresses participant à des campagnes de phishing, mais aussi des IPs utilisées par différents groupes d'attaques.

Sekoia a accepté de mettre à disposition cet ensemble de données pour les besoins de nos travaux de recherche. Ils sont accessibles ici [168]. Il est nécessaire de faire une demande d'accès à Sekoia.

L'ensemble de données de Sekoia contient les traces de différentes infrastructures auxquelles plus de 89 000 adresses IPs uniques ont participé. Ces traces sont regroupées en 174 labels. Chaque label correspond à une infrastructure de *command and control* d'attaque, ainsi qu'à son suivi sur le temps. De notre côté, nous avons réuni les différents labels en catégories pour faciliter une analyse plus macroscopique de nos résultats. L'ensemble des descriptions des labels est présent en Annexe D. Les données ont été collectées entre décembre 2019 et août 2022. Cet ensemble contient différentes adresses IPs, dont certaines ont participé à plusieurs attaques. L'ensemble des caractéristiques des ensembles de données est agrégé dans la table 5.1.

Nous avons aussi eu accès à d'autres captures d'attaques, néanmoins celles-ci n'étaient pas suffisamment complètes pour réaliser notre travail de caractérisation. Nous avons donc décidé de nous concentrer sur la caractérisation de cet ensemble de données. Les résultats sont présentés dans la section 5.2.

5.1.3 Limites rencontrées dans l'obtention des données

Enfin, nous proposons un retour d'expérience de la collecte des traces. Ce fut un travail long et difficile avec de nombreuses discussions pour arriver à identifier des sources pertinentes. Il était difficile d'identifier les bonnes personnes et les bons services. À chaque échange, nous avons présenté notre méthodologie puis discuté des possibilités, avant d'entamer une discussion avec un aspect juridique pour la transmission des informations.

Nous avons échangé avec 9 différents acteurs qui ont accès à des traces d'attaques. Parmi ces 9 acteurs à qui nous avons présenté notre projet, nous avons poursuivi les discussions avec 6

2. https://blog.sekoia.io/command-control-infrastructures-tracked-by-sekoia-io-in-2022/

d'entre eux. Les acteurs avec qui nous avons échangé étaient tous liés au département technique de leurs entités respectives. Les 6 acteurs avec qui nous avons plus longuement discuté nous ont fait part de leur engouement et intérêt pour le projet. En effet, la caractérisation des sources d'attaques pouvait présenter des intérêts de développement et d'implémentation de solutions de leur côté. D'une manière générale, cela pouvant apporter une meilleure connaissance de l'écosystème des attaques.

Néanmoins, malgré cet engouement prononcé, nous avons vite rencontré diverses limites.

Tout d'abord, les adresses IPs étant considérées comme des données personnelles par le RGPD, nos interlocuteurs nous ont vite communiqué leurs inquiétudes quant à la transmission des adresses IPs sources. À cela, nous avons proposé 2 alternatives : la possibilité de déployer notre solution dans leurs locaux, ou l'anonymisation au niveau des sous-réseaux avec comme taille la plus petite un /24.

À l'exception de 3 acteurs pour qui ces conditions étaient acceptables, il en a été autrement pour le reste des acteurs. Parmi les acteurs les plus intéressés par le projet, un acteur nous a confié pouvoir uniquement nous fournir les informations de géographie de sources, cela dû aux limites imposées par le service juridique. Nous avons dû refuser puisque cela ne nous permettait pas de produire des analyses suffisamment pertinentes. Sans adresses IPs, nous ne pouvons pas appliquer notre méthodologie.

Si des traces sont bien disponibles pour la communauté scientifique, celles-ci sont toujours anonymisées. Des initiatives comme DDoSDB[3] permettent de mettre à disposition des traces d'attaques pour analyse et rejeux, néanmoins, elles sont d'abord anonymisées pour une grande partie. Si d'autres traces sont disponibles, elles sont soit trop anciennes pour répondre à notre objectif, soit de même anonymisées, soit sont trop peu qualifiées pour proposer une analyse approfondie.

Dans le cadre de cette thèse, il nous était donc nécessaire de partir à la recherche de nos propres traces et de trouver des partenaires en capacité à détecter et à identifier les attaques. La détection et l'identification des attaques étant un domaine de la recherche à part entière sur lequel nous avons décidé de ne pas nous concentrer, puisque déjà très étudié.

À la suite de ces recherches, nous avons choisi de concentrer notre étude sur les traces mises à disposition par Sekoia.

5.2 Étude des traces de Sekoia

Dans cette section, nous présentons une analyse des sources de trafic identifiées par Sekoia comme participants à des activités malveillantes. Comme nous l'avons dit précédemment, l'ensemble de données de Sekoia est très complet et possède de nombreux labels d'attaques. Dans cette section, nous proposons un regroupement des labels par catégories d'attaques afin d'étudier les possibles motifs que nous pouvons observer en fonction des catégories. Par la suite, nous présentons les résultats obtenus pour les différentes catégories de données.

Cet ensemble de données correspond à des machines participant à des activités de *Command and Control*. Selon le référentiel Mitre [169], la *Command and Control* correspond à la définition suivante : "*des techniques que les adversaires peuvent utiliser pour communiquer avec les systèmes qu'ils contrôlent au sein d'un réseau victime. Les adversaires tentent généralement d'imiter le trafic normal et attendu afin d'éviter d'être détectés. Il existe de nombreuses façons*

3. https://github.com/ddos-clearing-house

Catégories	Malwares	Outils	Groupes d'attaques	Virus
Sous catégories	RAT Portes dérobées (backdoor) Trojan Chevaux de troies (trojan) Logiciels espion (Spyware) Logiciels malveillants bancaires Downladers Vulnérabilités d'exploitation CryptoJacking autre	C2 botnet complete red team web pen test remote administration and post exploitation pen test phishing network sniffers remote access tool (RAT) post exploitation reverse proxy/socket interaction detection tool activity ensemble	APT Crypto actor autres groupes	non applicable

TABLE 5.2 – Catégories et sous-catégories de l'ensemble de données

pour un adversaire d'établir un commandement et un contrôle avec différents niveaux de furtivité en fonction de la structure et des défenses du réseau de la victime." Il s'agit donc d'un point central de l'architecture des attaques.

L'ensemble de données fourni par Sekoia contient donc les adresses IPs des machines communiquant avec les machines infectées. Pour rappel, dans cet ensemble de données, il y a 174 labels différents qui caractérisent les IPs et qui font référence aux activités qui ont été caractérisées par Sekoia (exemple : Cryptocore). L'analyse que nous présentons ne contient pas l'ensemble des labels. Nous avons écarté certains labels pour deux raisons majoritaires : trop peu d'adresses avaient été qualifiées, ou Sekoia ne suivait plus certains labels.

Dans cette section, nous commençons par décrire ces différents labels ainsi que les regrouper en sous-catégories. Par la suite, nous présentons les analyses qui ont été produites par nos soins sur l'ensemble de données. Enfin, nous présentons les grandes conclusions que cette analyse nous a permis d'émettre.

5.2.1 Description des labels et regroupement des labels par catégories

L'ensemble de données contient 174 labels différents. L'ensemble des labels est décrit dans l'annexe D. Comme on peut le voir sur cette annexe, les sous-ensembles de données varient très largement en taille en fonction des labels. Pour extraire des conclusions pertinentes et identifier potentiellement des infrastructures en fonction des catégories, nous proposons ici un

regroupement des labels en fonction des types d'attaques auxquelles correspondent les labels. Ce découpage a été réalisé en fonction des natures des infrastructures observées par Sekoia. En effet, certaines se rassemblent grâce à leur type d'activité malveillante, d'autres par les groupes de personnes qui sont regroupés derrière ces activités. Nous faisons donc ce découpage, car il ne s'agit pas des mêmes composantes socio-organisationnelles qui sont traquées par Sekoia.

Nous proposons de décomposer l'ensemble des données en 4 sous-catégories d'attaques : les malwares, les outils, les groupes d'attaques et le virus. Les 4 catégories ainsi que leurs sous-catégories respectives sont présentées dans la Table 5.2 et discutées dans les 4 sous-sections suivantes.

5.2.1.1 Les catégories de la famille Malware

La première catégorie correspond à la famille des **Malwares**. Les malwares sont des logiciels qui ont pour objectif d'infecter les victimes sans se faire découvrir. On peut décomposer la famille des *malwares* en de nombreuses sous-catégories. Celles-ci peuvent se concentrer sur l'objectif du malware, le type d'action mené par celui-ci ou encore la façon par laquelle il a été distribué. Du fait de la variété de notre ensemble de données, nous proposons une classification qui utilise l'ensemble de ces dénominateurs. Les catégories que nous avons identifiées sont les suivantes :

— Les RAT (Remote Access Trojan) : ils ont pour objectifs de permettre un accès distant à la machine infectée.
— Les portes dérobées (ou Backdoor) : regroupe les méthodes par lesquelles les attaquants ont la possibilité d'accéder à un système via l'exploitation d'une vulnérabilité de celui-ci.
— Les chevaux de Troie (Trojan) : Malware qui se fait passer en apparence pour un logiciel légitime.
— Les logiciels espions (Spyware) : ils ont pour but de collecter des informations que les systèmes sur lesquels ils sont installés dans le but de les transmettre aux attaquants par la suite.
— Les logiciels malveillants bancaires : ces logiciels se concentrent particulièrement sur la récupération d'identifiant bancaire ainsi que sur la réalisation de transaction bancaire.
— Les downladers : ils ont une seule tâche, la récupération des exécutables malveillants à partir d'un serveur contrôlé par un attaquant.
— Les vulnérabilités d'exploitation (exploit) : ils exploitent une vulnérabilité d'une application sur la machine victime.
— Les bots : un malware qui a pour but de transformer une machine en bot, pour l'intégrer par la suite dans un réseau de bot (botnet).
— CryptoJacking : le but est d'utiliser l'appareil infecté dans le but de miner de la crypto-monnaie.

La table 5.3 présente les différents labels de la famille Malware attribué à chacune des sous-catégories.

5.2.1.2 Les catégories de la famille Outils

Dans une seconde famille, nous regroupons les différents **Outils** qui ont pu être utilisés à des fins malveillantes. Il s'agit d'outils qui sont généralement disponibles en ligne et qui permettent d'être utilisés dans des campagnes de test d'intrusions sur des infrastructures. Si ces outils

TABLE 5.3 – Classification des malwares

Sous-catégorie des Malwares	Liste des labels de la sous-catégorie, identifiés par Sekoia
RAT	plugX ; evilosx ; brata ; Badrat ; BitRat ; DcRat ; nanocore ; Orcus ; Vajra ; NetDooka ; Rafel ; ArrowRat ; JSSLoader
Portes Dérobées (ou Backdoor)	poshC2 ; ShadowPad ; felixRoot ; bazarbackdoor ; tunna ; bazarloader ; Tomiris ; SolarMaker ; GoMet ; phantomLance
Chevaux de Troie (Trojan)	octopus ; nexus ; xhelper ;
Logiciels espions (Spyware)	predatorTheThief ; chaes ; vulturi ; SolarMaker ; FFDroider ; CreepySnail ; furball ; FinFisher ; Cytrox ; trickbot ; Xloader ; XploitSPY
Logiciels malveillants bancaires	hydra ; dridex ; Grandoreiro ; Ramnit ; SharkBot ; SeaFlower
downlader	BUMBLEBEE ; PowGoop
Vulnérabilités d'exploitation (exploit)	tagBarnakle
CryptoJacking autre	LemonDuck Subzero ; panda ; GrandaMisha ; Teardroid ; SVCRAT

sont développés pour des utilités à des fins bienveillantes, du fait de leurs disponibilités, ils peuvent être aussi utilisés à des fins malveillantes.

Comme expliqué en introduction de cette section, la famille **Outils** agrège les labels qui correspondent à des outils connus qui sont développés dans un but de test des infrastructures, mais qui, du fait de leurs disponibilités, peuvent aussi être utilisés par des acteurs à but malveillant. Nous avons décidé de décomposer cette famille en diverses sous-catégories, car les outils identifiés ne réalisent pas tous les mêmes actions. Nous avons choisi de regrouper les labels de la famille *outils* en utilisant les sous-catégories suivantes :

— C2 : outils qui permettent de gérer une infrastructure de command and control.
— botnet : outils qui permettent de gérer un réseau de bot.
— complete red team : Outils de test d'intrusions très complets avec de nombreuses fonctionnalités.
— web pen test : outils de test d'intrusions visant la couche applicative web.
— remote administration and exploitation : Outils qui permettent l'administration à distance des machines infectées et l'exploitation des ressources de celles-ci.
— pen test : Outils de test d'intrusions.
— phishing : Outils qui permettent de générer des campagnes de phishing.
— network sniff : Outils qui servent à faire des analyses réseaux au niveau des machines infectées.
— remote access tool (RAT) : Outils d'accès à distance des machines. Les RAT ici sont différents que ceux de la famille *malwares* car ici l'accès est désiré.
— reverse proxy/socket : Outils qui permettent d'exploiter les sockets et les proxys pour la communication entre les machines.
— interraction detection tools : Outils qui surveillent les réseaux.
— activity ensemble : Ensemble d'activités menées par un groupe.

TABLE 5.4 – Classification des outils

Sous-catégorie des Outils	Liste des labels de la sous-catégorie, identifiés par Sekoia
C2	apfell ; factionC2 ; covenant ; merlin ; sliver ; shad0w ; cs2modrewrite ; BruteRatel ; TrevorC2 ; NorthStar ; DeimosC2 ; RedGuard
botnet	armitage
complete red team	cobaltstrike ; Caldera
web pen test	beef
remote administration and post exploitation	empirev1 ; empirev2 ; pupy ; silenttrinity ; empirev3 ;empire ;
pen test	metasploit ; koadic
phishing	gophish ; phishery ; Phishmonger
network sniff	responder ;
remote access tool	quasarRAT ; asyncRat ; meterpreter
post exploitation	fruityC2 ; AlanFramework ; PickleC2
reverse proxy/socket (communication)	modlishka revsocks ; HTTP-Revshell
interaction detection tool	interactsh
activity ensemble	GhostwriterTool
autre	throwback ; satellite

Les labels qui appartiennent à la famille des outils ont donc été repartis dans ces différentes catégories. Un résumé de cette répartition se trouve dans la table 5.4.

5.2.1.3 Les catégories de la famille Groupe d'attaque

La troisième catégorie regroupe, elle, des labels qui correspondent à des **Groupes d'attaques**. Ici, ce n'est pas un type d'activité malveillante qui est relevé par Sekoia, mais plutôt un ensemble d'activités qui sont liées à un groupe de personnes rassemblées sous la forme d'une organisation. Cela présente un grand intérêt pour notre analyse puisque cela nous permet de regarder des groupes d'attaques indifféremment de leurs activités.

Une des spécificités de notre ensemble de données est qu'il contient aussi des IPs qui sont attribuées non pas à des activités malveillantes, mais à des groupes d'attaques. Les groupes d'attaques sont des groupes, une association, d'humains qui œuvrent dans un but commun via des attaques. Il existe différents groupes d'attaque et nous proposons une classification en 4 sous-catégories :

— APT (Advanced Persitent Threats) : menace cataloguée persistante. Le terme désigne également le groupe de personnes à l'origine de l'attaque.
— crypto actor : acteur qui s'intéresse uniquement au vol de cryptomonnaies.
— autres groupes : groupe de personnes à l'origine d'attaque communes.

L'ensemble des labels correspondants à des groupes d'attaques ont été réparties parmi les catégories précédentes. La table 5.5 présente la répartition.

TABLE 5.5 – Classification des groupes d'attaques

Sous-catégorie des groupes d'attaques	Liste des labels de la sous-catégorie, identifiés par Sekoia
APT	apt28-xagent ; sidewinder ; thalium ; APT35 ; APT29 ; UNC2452 ; APT27 ; Deathstalker
autres groupes	turla ; CostaRicto ; cerium ; tontoteam ; wizardspider ; muddyWater ; RoamingMantis ; TAG38 ; TEMPHERETIC ; Sandworm ; Karkadann
crypto actor	cryptocore

5.2.1.4 Le virus

Enfin, la dernière famille que nous avons identifiée correspond au **virus**. Un virus est un logiciel qui agit à des fins malveillantes et qui, de plus, cherche à s'autorépliquer sur d'autres machines. Ici, nous avons un unique label qui correspond à un virus. Nous n'avons donc pas de sous-catégorie au sein de cette famille.

5.2.2 Analyses produites par IPSeen

Nous allons à présent présenter les résultats d'analyse de notre ensemble de données d'attaques. Les résultats présentés ont été générés en suivant la procédure suivante : Utilisation de **IPSeen Accurate** sur l'ensemble des ensembles de données, utilisation de **IPSeen Fast** sur les ensembles de données qui n'aurait été que très peu qualifiée. L'analyse est lancée sur l'ensemble des IP, car cela nous permet également de confirmer si les résultats obtenus sont similaires pour les deux versions. Ensuite, une analyse des résultats les plus étonnants a été réalisée, ainsi qu'une correction manuelle des ensembles de données où il y avait peu d'informations manquantes. Cette vérification manuelle sur un sous-ensemble des plus petits ensembles de données a permis d'établir une liste d'entités qui n'ont pas été qualifiées par IPSeen, car :

1. Les données de Wikidata ne sont pas assez complètes.
2. Aucunes données n'étaient existantes dans Wikidata, mais il fût très simple d'identifier le type de l'entité qui possède l'IP.

Via cette procédure manuelle, nous avons pu rapidement identifier environ 50 entités supplémentaires qui revenaient très régulièrement. Il s'agit principalement de petites entreprises qui proposent des services d'hébergement. Pour les ensembles de données plus conséquents, nous avons automatisé l'attribution de la caractérisation manuelle en s'aidant du nom de l'entité. Cela nous a permis de rapidement attribuer le label de type à des résultats qui n'avaient pas été qualifiés.

Au niveau des analyses, comme l'entreprise Sekoia a déjà produit une analyse de la géographie des adresses IP avec des résultats disponibles [4], nous avons décidé de ne pas retracer l'analyse géographique des adresses IP et d'utiliser les résultats déjà produits.

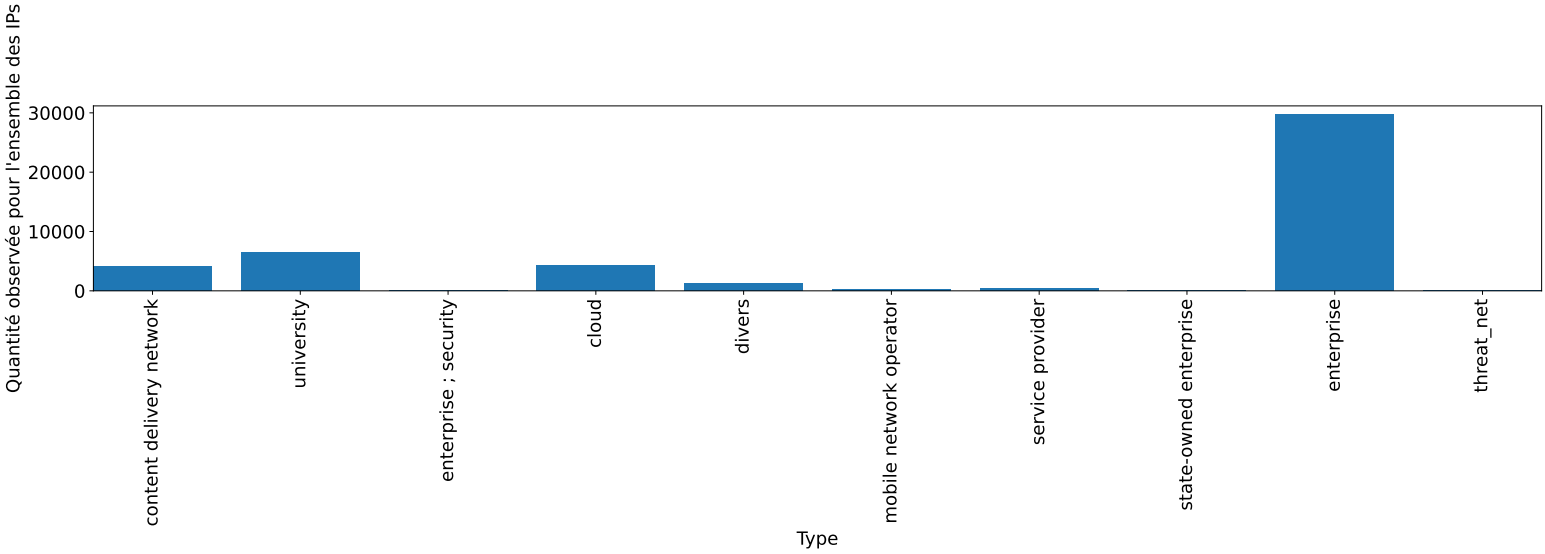


FIGURE 5.1 – Répartition de l'ensemble des adresses IP qualifiées

5.2.2.1 Vue de l'ensemble des adresses IP

La figure 5.1 présente l'ensemble des types de réseaux identifiés pour l'entièreté de l'ensemble de données. Sur ce graphique, nous pouvons constater qu'il y a quatre pics majeurs : entreprise, Cloud, université et CDN. Autrement dit, la majorité des infrastructures d'attaque se trouvent dans les réseaux de ces acteurs. En ce qui concerne les entreprises, nous avons également regardé les domaines d'activités des différentes entreprises. La figure 5.2 présente les différents domaines d'activité des entreprises. Sur ce graphique, nous pouvons voir qu'une majorité des entreprises ont des activités liées au domaine de l'informatique. On voit à nouveau apparaitre de nombreuses entreprises Cloud. Il n'est pas surprenant de retrouver des entreprises de ce domaine puisque les infrastructures réseaux sont au cœur de leurs métiers.

Discutons à présent de la taille des entités impliquées pour l'ensemble des données et visible en figure 5.3a et 5.3b. Sur la figure 5.3a, nous pouvons constater que nous observons en grande majorité des très grandes entreprises. Cela s'explique par deux faits. Premièrement, les grandes entreprises possèdent beaucoup d'adresses IP en comparaison avec des structures plus petites. Également, les grandes entreprises sont généralement plus représentées dans Wikidata, ce qui crée un biais. Au niveau des fournisseurs d'accès, on peut voir qu'il s'agit majoritairement de grands fournisseurs d'accès, avec un potentiel de machines infectées ou utilisées à des fins malveillantes qui peut être très grand.

Nous observons aussi quelques réseaux universitaires. En revanche, il s'agit, pour une grande majorité de ce pic, de deux faux positifs. Tout d'abord, une grande partie des résultats sont affectés au MIT, en cause, un bloc d'adresses déclaré dans Wikidata qui n'est pas correct. En effet, l'entièreté du /8 est déclarée comme appartenant au MIT, alors que le MIT ne possède que le premier /11 de ce bloc. Le reste appartient à une grande entreprise fournisseur d'hébergement. Dans cette catégorie, le deuxième faux positif correspond à une erreur de résultat d'IPSeen, et il s'agit en fait d'une entreprise d'hébergement.

4. https://blog.sekoia.io/command-control-infrastructures-tracked-by-sekoia-io-in-2022/

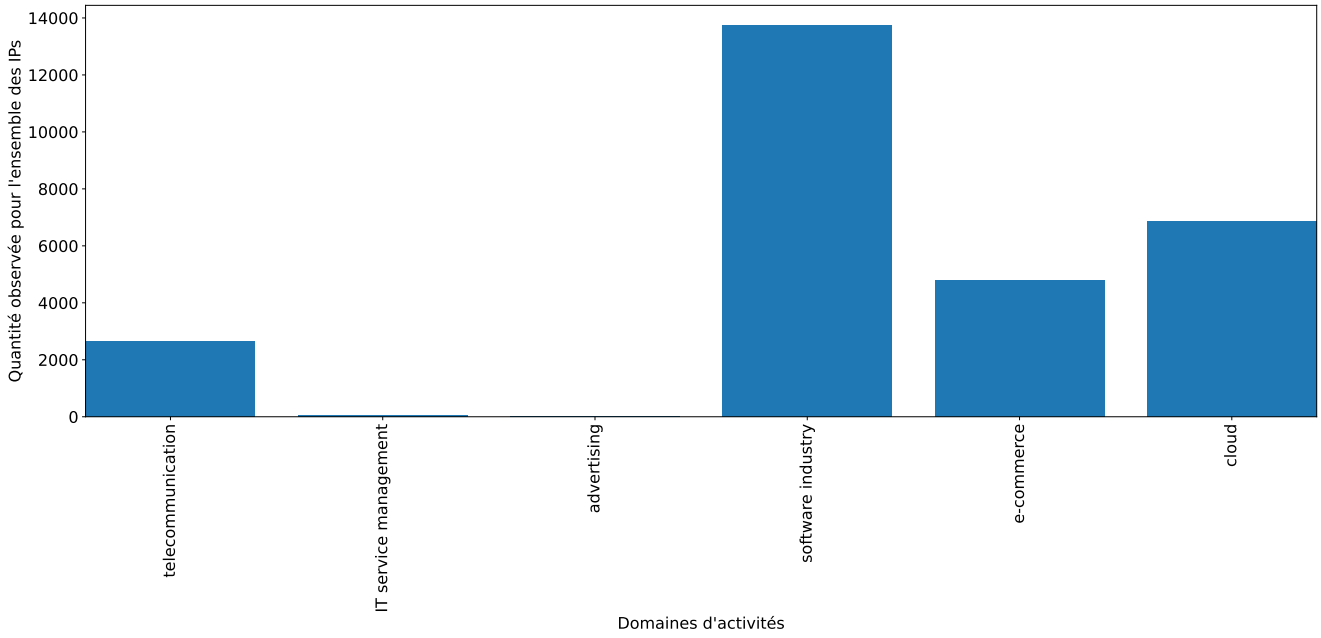


FIGURE 5.2 – Répartition des domaines d'activités parmi les entreprises sur l'ensemble du jeu de données qualifié

En ce qui concerne la quantité importante d'entreprises au domaine d'activité lié au développement de logiciel, nous avons derrière ce pic de très grosses entreprises qui possèdent des grands blocs d'adresses IPv4. Ces grandes entités ont généralement plusieurs domaines d'activités et ne se concentrent pas seulement sur le développement de logiciel. Néanmoins, l'ensemble des domaines d'activités sont liés à l'informatique et aux réseaux de manière générale.

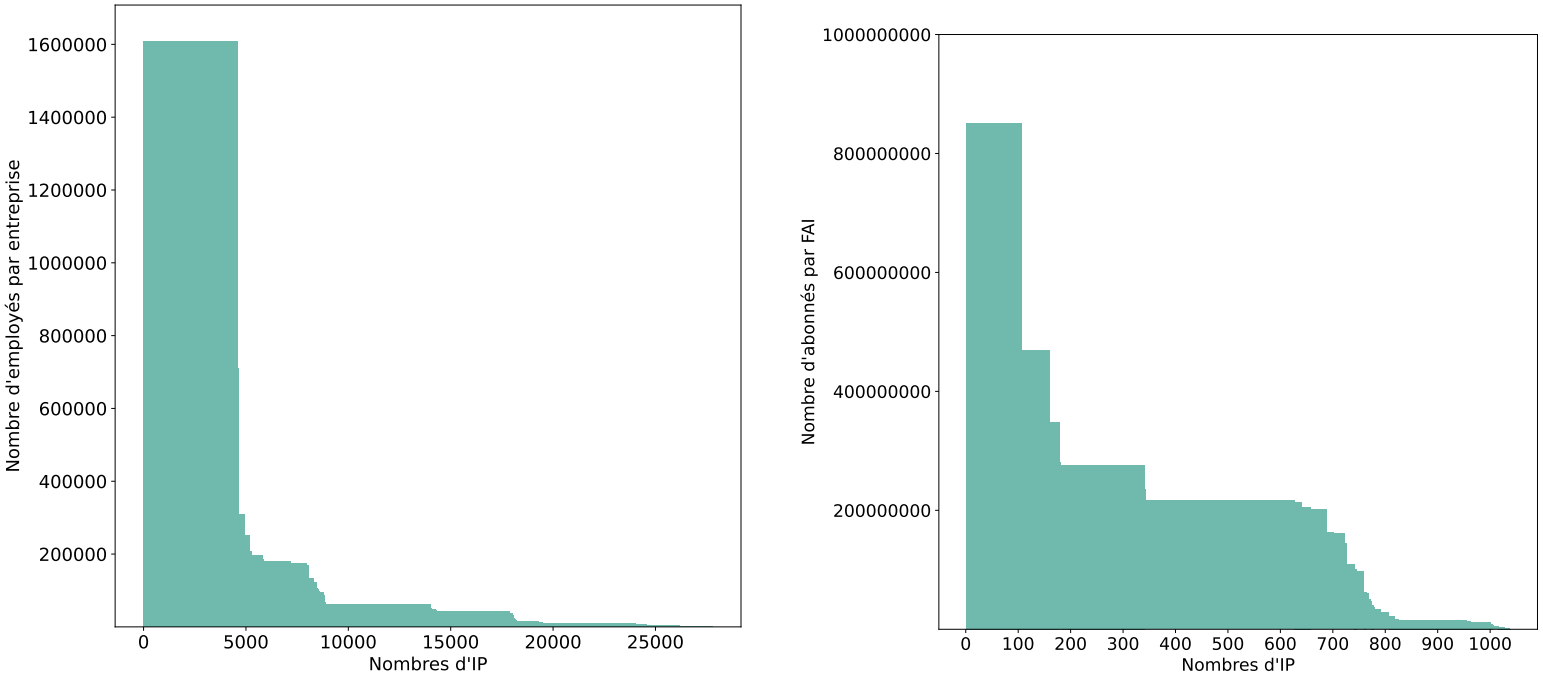
Parmi les labels, nous avons un label threat_net. Ce label a été affecté manuellement et désigne des réseaux que nous n'avons pas pu qualifier manuellement, car nous n'étions pas en capacité d'identifier une entité propriétaire. Néanmoins, après des recherches, nous avons vu que ces réseaux remontent très régulièrement parmi les listes d'alerte à la fraude. Nous avons donc créé une catégorie qui représente ces réseaux qui sont difficiles à qualifier, mais qui en revanche sont très présents parmi les listes noires.

Nous pouvons donc conclure que dans l'ensemble des cas, nous avons une forte dominance des entreprises de télécommunications (donc des fournisseurs d'accès) et donc des réseaux domestiques. Mais également, une forte dominance des entreprises d'hébergement et des hébergeurs. Nous pouvons donc envisager qu'il soit nécessaire de regarder plus particulièrement dans ces deux directions pour la sécurisation des réseaux.

5.2.2.2 Graphiques par catégorie

Dans un second temps, nous séparons l'ensemble des résultats en les classant en quatre catégories précédemment évoquées : Malwares, Outils, Groupes d'attaques et Virus.

Les sections suivantes détaillent les résultats obtenus pour chacune des catégories. L'ensemble des répartitions parmi les différentes catégories présentées par la suite n'inclut pas les adresses IP non qualifiées.



(A) Taille des entités

(B) Taille des fournisseurs de services

FIGURE 5.3 – Taille humaine des entités
Une barre représente une IP

5.2.2.2.1 Malware

Intéressons-nous à présent à la catégorie des malwares. Nous cherchons à voir s'il y existe un motif, ou une répartition commune dans la famille des malwares, ou dans les différentes sous familles. Les résultats pour la catégorie des malwares sont présentés sur la Figure 5.4. Sur cette figure, nous pouvons voir sur l'axe des ordonnées les différentes attaques classifiées en catégorie. Les catégories correspondent à celles présentées dans la table 5.3. Sur cette figure, nous pouvons observer une répartition des différents types de réseaux ainsi que les différents domaines d'activité pour le type "entreprise". Les domaines d'activités sont représentés par les portions encadrées par la couleur qui correspond à la catégorie "entreprise", ici vert clair.

Sur l'ensemble des attaques, nous pouvons observer une majorité de réseaux de type cloud, entreprise et université (sachant qu'il s'agit de faux positifs correspondant à des IP Cloud). Nous pouvons constater que pour environ une dizaine des attaques de type malware, nous avons une infrastructure avec plus de 95% des IPs qui appartiennent à un unique type de réseaux, et que le plus souvent, il s'agit de réseaux de type cloud. Néanmoins, il ne semble pas y avoir de tendance en fonction de la catégorie de malware.

5.2.2.2.2 Outils

Nous cherchons à présent à voir s'il existe des motifs communs pour les différentes catégories d'outils.

Les résultats pour cette catégorie sont présentés sur la Figure 5.5. Comme la figure précédente, nous pouvons voir sur l'axe des ordonnées les différentes infrastructures d'attaques classifiées en sous-catégorie (5.4). De même, nous avons séparé la catégorie entreprise en sous

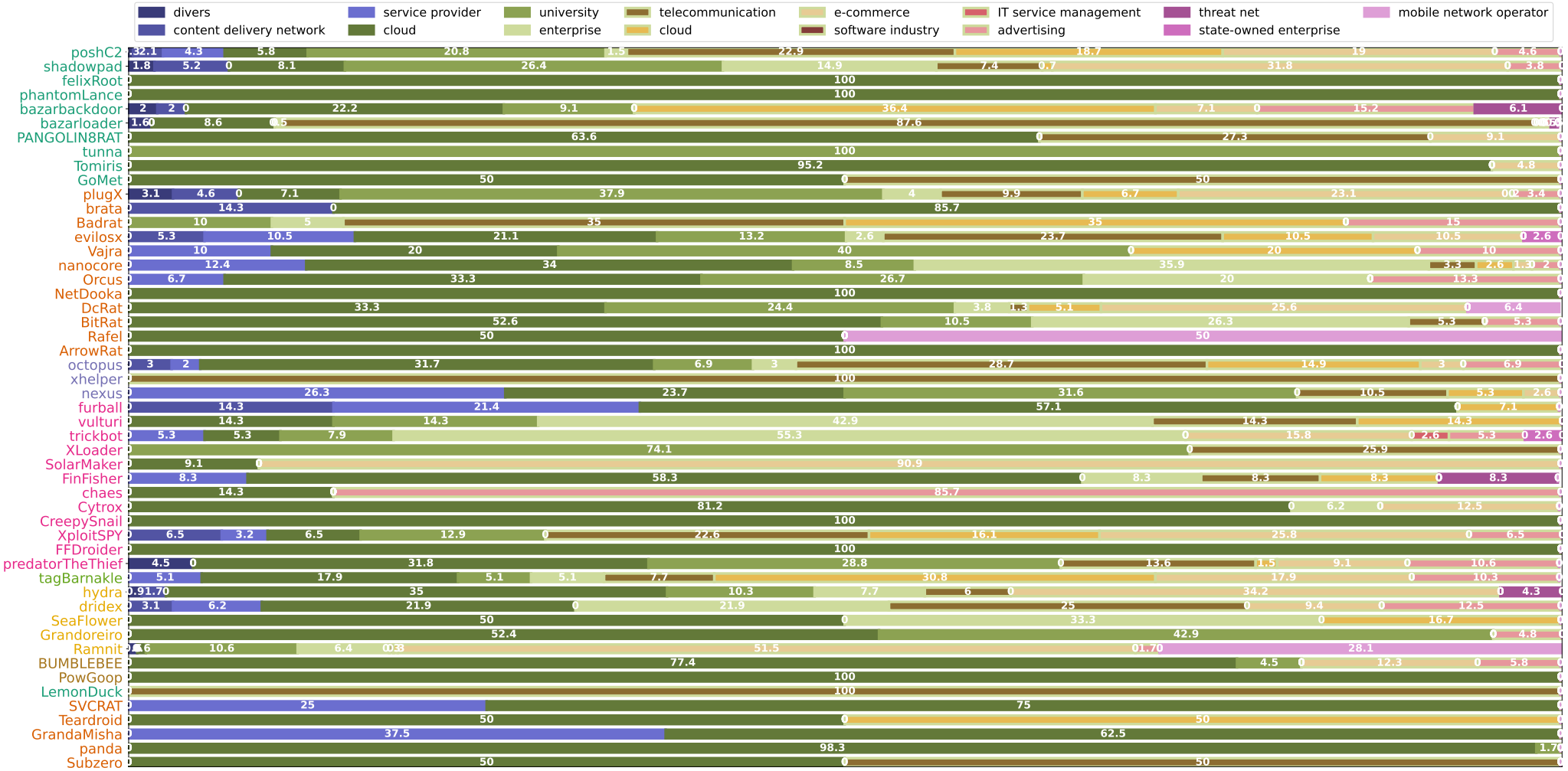


FIGURE 5.4 – Type de réseau pour la catégorie des malwares

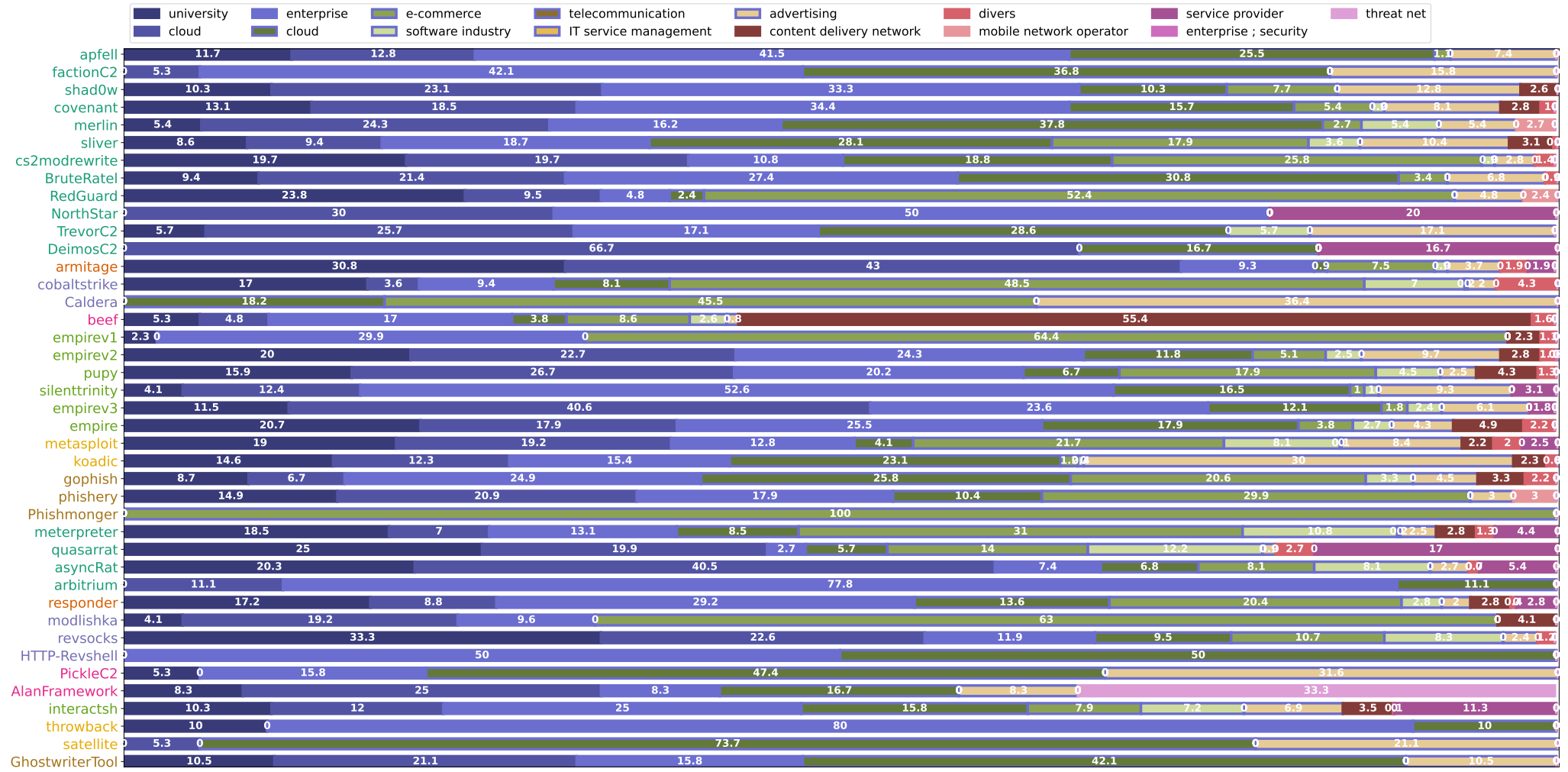


FIGURE 5.5 – Type de réseaux pour la catégorie des outils

domaines d'activités. Les barres encadrées par du bleu représentent les différents domaines d'activités.

En ce qui concerne les outils, nous avons une écrasante proportion pour la catégorie des entreprises. L'interprétation que nous faisons de ces résultats est plus particulière. En effet, du fait de la nature de la catégorie, nous pouvons nous demander s'il s'agit réellement d'activités malveillantes qui ont été capturées, ou alors s'il s'agit d'activités de test d'intrusions menées vers des infrastructures qui se savaient sous attaque. En effet, pour un certain label, nous avons pu retracer l'ensemble des adresses IPs présentes dans la capture vers une entreprise dont le site Internet met en avant les audits de sécurité comme activité. Nous avons choisi de retirer ce label de l'ensemble de données que nous avons analysées. La plupart des outils traqués sont en effet disponibles en ligne et sont présentés comme des outils servant à faire des tests de sécurité. Une question reste donc en suspens pour cet ensemble d'attaque : s'agit-il d'activités qui se sont déroulées dans un cadre légal ou s'agit-il d'une exploitation d'outils disponibles pour faire de la gestion d'infrastructure ? Enfin, nous voyons dans la légende le label "enterprise ; security". Via ce label, nous désignons l'entité précédemment évoquée que nous avons identifiée comme étant une entreprise de sécurité spécialisée dans les activités de test d'intrusions. Nous pouvons par ce biais voir que celle-ci est présente parmi d'autres activités.

On observe également une tendance pour quasiment l'ensemble des outils, où nous avons une partie de l'infrastructure dans le cloud, et une autre dans des réseaux d'université, autrement dit des faux positifs pour du cloud. On observe aussi une plus grande variété des domaines d'activités au niveau des entreprises.

Parmi l'ensemble des résultats, nous devons noter quelques résultats particuliers. Tout d'abord, pour *Phishmonger*, nous avons une unique entité, de type entreprise de logiciel, derrière l'infrastructure d'attaque. Au niveau du label *beef*, nous observons un nombre étonnant de CDN par rapport au reste des outils. Pour rappel, *beef* est le seul label de la catégorie des outils de test d'intrusions Web.

5.2.2.2.3 Groupes d'attaques

Les résultats pour la catégorie des groupes d'attaques sont présentés sur la Figure 5.6. De la même façon que les graphiques des autres catégories, nous faisons apparaître la classification des groupes en sous-catégories et nous faisons apparaitre les différents domaines d'activité.

Une fois de plus, on constate des résultats similaires aux précédents avec une majorité d'équipement dans le Cloud, avec deux groupes ayant une infrastructure à 100% dans le Cloud. Néanmoins, nous ne pouvons pas conclure sur des tendances qui seraient liées au type de groupe. Mais il serait intéressant de continuer à tracer les profils de ces groupes pour étudier leurs stabilités et donc définir des contre-mesures.

5.2.2.2.4 Virus

Les résultats pour la catégorie des virus sont présentés sur la Figure 5.7. Comme nous n'avions qu'une seule attaque identifiée comme un virus, nous ne pouvons pas conclure sur une tendance générale. Néanmoins, l'entité qui est identifiée comme une entreprise d'e-commerce a également des activités d'hébergement. Cependant, nous ne pouvons pas savoir avec certitude s'il s'agit de machines Cloud, ou liées aux activités d'e-commerce.

FIGURE 5.6 – Type de réseau pour la catégorie groupe d'attaque

FIGURE 5.7 – Classification des types de réseaux pour la catégorie des Virus

5.2.2.3 Observations dans le temps

En observant les graphiques de la section précédente, nous avons pu voir deux pics se dégager : entreprise et hébergement. Du fait de la nature particulière des activités d'hébergement, nous avons voulu vérifier si les machines d'hébergement utilisées dans le cadre des infrastructures d'attaques sont utilisées temporairement, peut-être pour remplacer une ressource manquante, ou à plus long terme. Grâce à cette analyse temporelle, nous sommes en capacité de savoir quelle est la nature des réseaux qui hébergent des adresses IPs persistantes.

La Figure 5.8 présente les résultats de notre analyse.

Les différentes couleurs représentent sur le premier graphique les différents types de réseaux et sur le second les domaines d'activités. En abscisse, nous avons le nombre d'observations pour une IP. En ordonnée, nous avons le cumul d'IP observées.

Afin d'obtenir ce graphique, nous avons agrégé chaque adresse IP unique de chaque ensemble de données d'attaques que nous avons associé avec son type. Nous pouvons voir sur ce graphique que les adresses IPs appartenant à des réseaux de type hébergement sont seulement présentes sur un court intervalle et ont été de manière générale moins observées que le reste des adresses IPs. Les adresses IPs d'entreprises ont néanmoins été observées beaucoup plus de fois. Nous constatons donc une stabilité au niveau des entreprises infectées alors que les hébergeurs sont plus volatils.

Comme nous l'avons vu en section 5.2.2.1, parmi les entreprises, un grand nombre d'entre elles ont des activités d'hébergement. Il est donc pertinent de regarder les résultats précédents

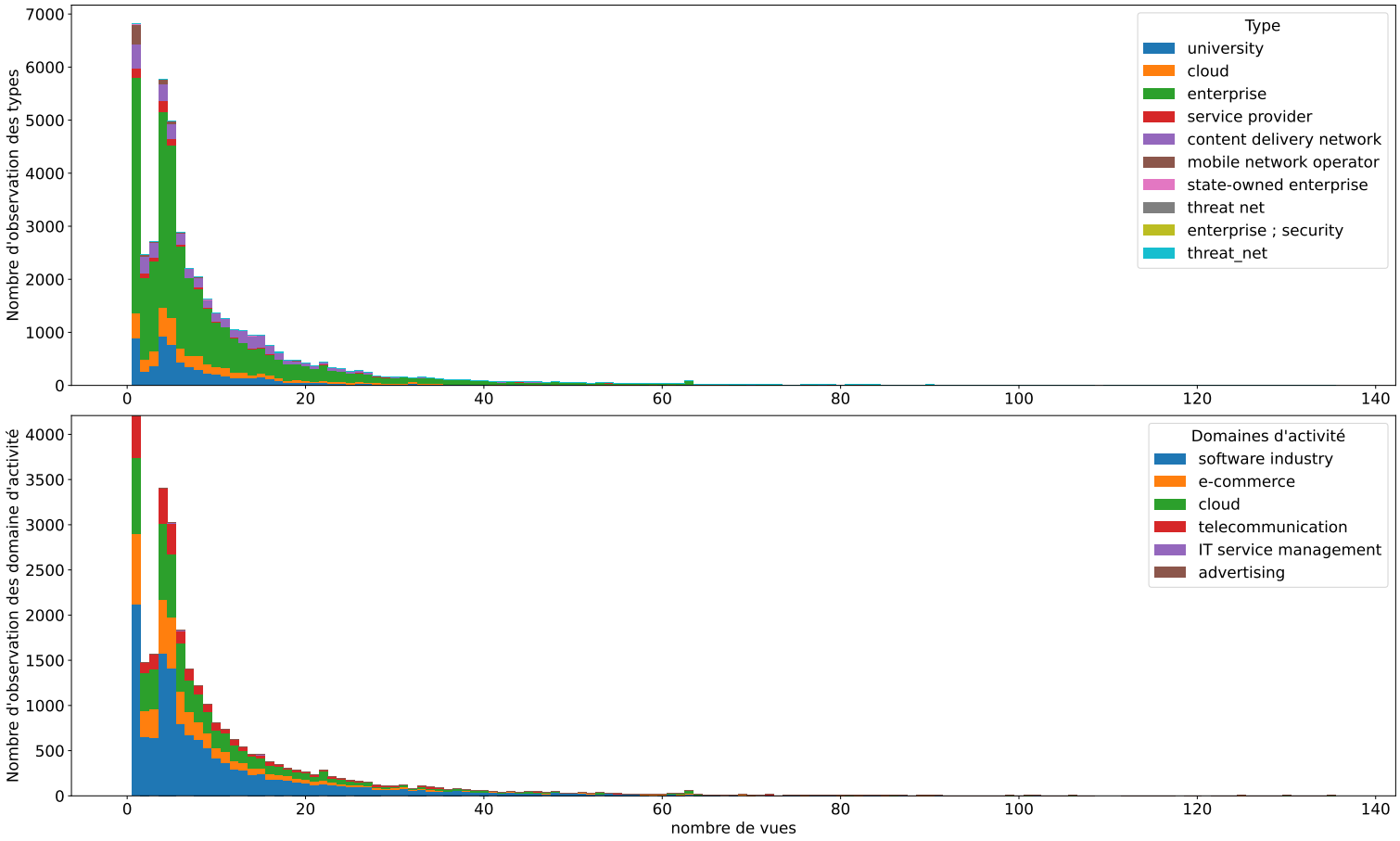


FIGURE 5.8 – Nombre d'observations des IP pour chaque type de réseaux auxquels elles appartiennent

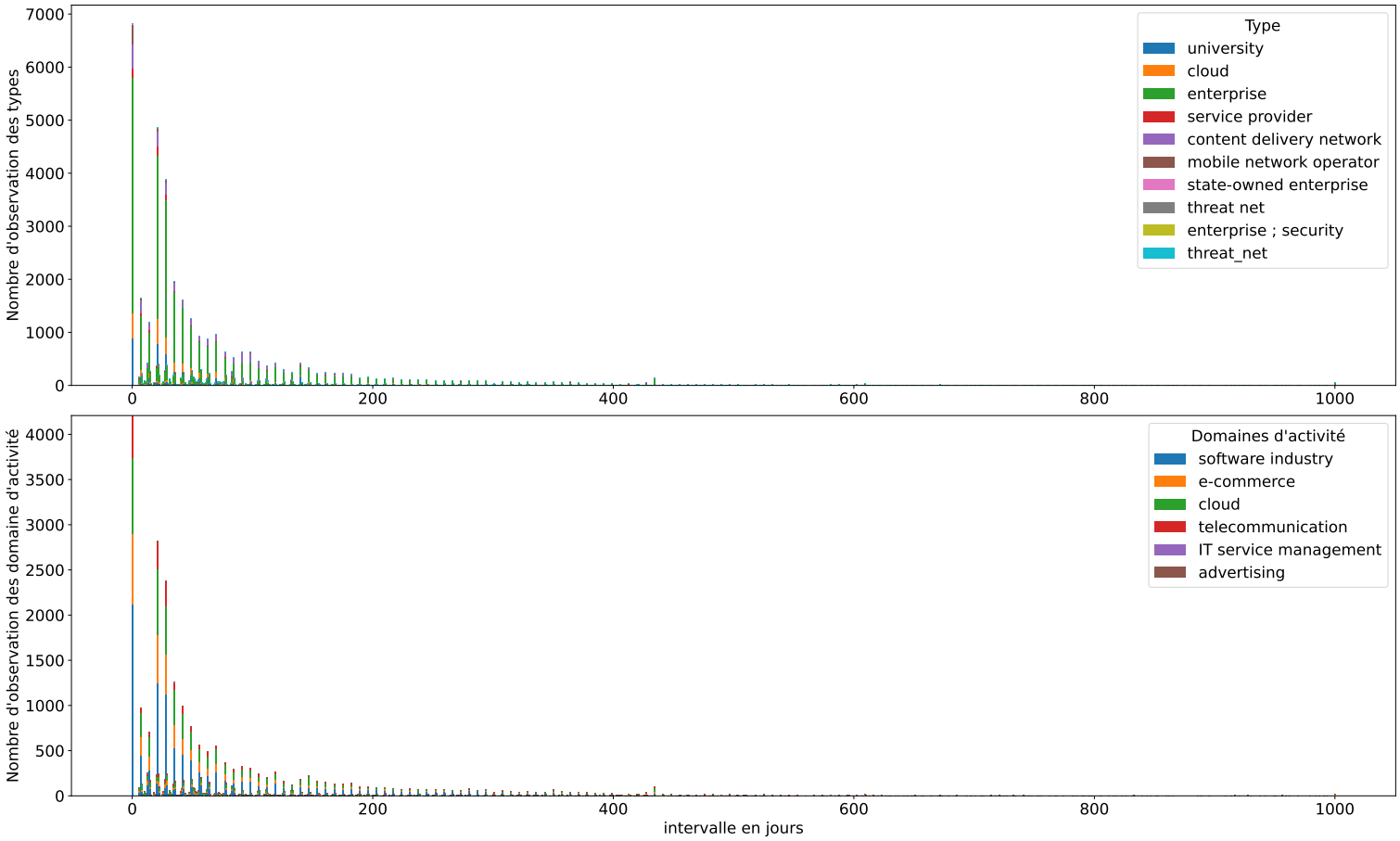


FIGURE 5.9 – Nombre de jours entre la première et la dernière observation des IP pour chaque type de réseaux auxquels elles appartiennent

avec un angle d'analyse des domaines d'activités. On peut constater que les ressources cloud au niveau des domaines d'activités ont également tendance à être présentes seulement sur les nombres d'observations plus bas, c'est-à-dire inférieurs à 20 fois.

Nous nous sommes également interrogés sur la durée de rétention des ressources. Pour cela, nous avons regardé pour chaque adresse IP le nombre de jours entre la première observation dans une infrastructure d'attaque et la dernière. Les résultats sont présentés sur la figure 5.9. Comme précédemment, les différentes couleurs représentent sur le premier graphique les différents types de réseaux et sur le second les domaines d'activités. En abscisse, nous avons le nombre d'observations pour une IP. En ordonnée, nous avons le cumul d'IP observé.

Premièrement, nous pouvons constater que des IPs ont été observées sur des périodes équivalentes à 1000 jours. Néanmoins, la grande majorité des IPs a une période d'observation inférieure à 50 jours. De même que précédemment, on peut constater que les machines appartenant à des réseaux de types Cloud ont tendance à moins persister dans le temps.

Ces observations sont en cohérence avec le type de service proposé par ces infrastructures. Néanmoins, la période de location des ressources est tout de même assez conséquente. On peut supposer deux conclusions à cela, soit une très grande rentabilité, soit un très gros usage parmi les différentes entités.

5.2.3 Analyse des entités présentes

Enfin, afin de finir cette section, nous nous intéressons aux entités présentes. Celles-ci sont anonymisées. La Figure 5.10 présente les entités qui apparaissent à plus de 50 reprises sur le total de l'ensemble catégorisé. Ce graphique représente leur nombre d'apparitions dans chacun des ensembles de données d'attaques. Nous pouvons voir qu'il y a 69 entités très récurrentes. Parmi celles-ci, sept entités reviennent sur plus de la moitié des ensembles de données d'attaques caractérisés.

Le nombre maximum d'observations d'une entité est de plus de 3 500 fois. Le nombre d'observations est compté par IP unique pour chaque sous-ensemble de données. Nous ne prenons donc pas en compte le nombre de fois qu'une adresse IP a été ajoutée à l'ensemble de données.

Nous avons identifié en totalité 1029 entités distinctes. Parmi celles-ci, nous avons des faux positifs néanmoins, mais ceux-ci apparaissent moins.

De plus, comme nous l'avons discuté très largement précédemment, nous pouvons observer une grosse présence des entreprises de Cloud. Nous nous sommes interrogés sur la taille de ces entreprises de Cloud. En effet, nous voulions savoir si des grosses infrastructures avaient plus de moyens pour potentiellement détecter les machines participantes aux activités malveillantes. La Figure 5.11 représente les différentes tailles de cloud observées pour les entités de type cloud non qualifiées manuellement. Comme nous pouvons le voir, il y a des infrastructures de toutes tailles. Il faut aussi savoir, pour avoir une vision globale de l'analyse, que l'ensemble des entités qualifiées manuellement qui n'existaient pas dans Wikidata sont des petites et moyennes entreprises. Nous n'avons pas pris en compte les entités Cloud qualifiées manuellement, car nous n'avons pas recherché les informations sur leurs tailles.

FIGURE 5.10 – Heatmap des apparences des entités présentes plus de 50 fois sur l'ensemble des attaques (échelle logarithmique)

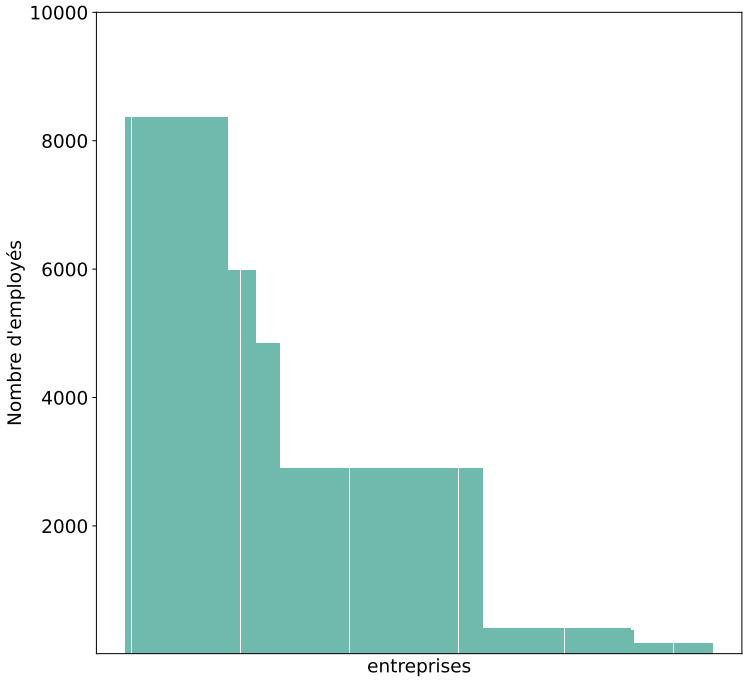


FIGURE 5.11 – Taille des entités cloud et entreprise de domaine d'activité cloud

5.2.4 Conclusions du chapitre

Les analyses produites en grande partie par IPSeen, ainsi que par analyse manuelle des extraits produite à chaque étape, permettent d'établir plusieurs conclusions :

— Les entreprises qui offrent des services d'hébergement sont très largement présentes dans l'ensemble des infrastructures d'attaques observées par Sekoia. Parmi ces infrastructures, il y a tous types de tailles. Il ne s'agit donc pas seulement d'infrastructures avec de petits moyens. Afin de limiter les attaques, il est donc impératif d'arriver à limiter le trafic malveillant dans les infrastructures Cloud. De plus, nous avons pu noter que parmi ces entreprises, un nombre important commercialisent également des solutions de protection contre les attaques.

— Il existe des faux positifs rapidement corrigibles dans Wikidata. Nous avons vu le cas avec l'université MIT, mais aussi, nous avons pu voir que des blocs d'adresses IPs ont pu être attribués à des résultats Wikidata portant le même nom. Ce phénomène n'a pas été mis en lumière lors de notre analyse de la qualité des données de Wikidata puisque nous évaluions seulement la première adresse du bloc d'IP. En constatant cela, nous pouvons affirmer qu'en améliorant les contributions et la collaboration autour de Wikidata, nous pourrons à l'avenir proposer de meilleures analyses. Nous n'avons pas pu corriger certaines erreurs, car certaines pages sont protégées.

— Au niveau des infrastructures de C2 des attaques, nous pouvons constater que les fournisseurs d'accès et, par conséquent, les réseaux domestiques ne représentent pas la plus grande proportion.

— Nous n'avons pas observé trop de chevauchement sur les adresses IPs, néanmoins, sur l'ensemble de toutes les attaques, nous avons identifié 69 entités qui sont présentes plus de 50 fois.

— Les observations dans le temps montrent que les infrastructures d'attaque semblent s'appuyer sur des ressources disponibles à la location pour des périodes plus courtes. Néanmoins, la période d'utilisation de ces ressources semble approcher jusqu'à 1000 jours dans certains cas avant que les ressources ne soient potentiellement plus utilisées.

Grâce à l'ensemble de ces résultats, nous avons pu identifier un point d'action (les réseaux d'hébergement), qu'il s'agisse de grande ou petite infrastructure. Nous avons pu constater qu'il n'y a pas forcément de corrélation entre la sélection des machines composantes de l'infrastructure d'attaques et le type d'attaque menée par l'infrastructure. Nous savons aussi que les périodes de rétention des ressources sont assez longues, mais que le nombre de machines décroit considérablement en fonction de la durée d'observation. Enfin, nous avons vu qu'un ensemble d'environ 70 entités est très largement présent parmi l'ensemble des attaques. Il serait donc possible, en déployant des solutions au niveau de ces entités, d'atteindre efficacement les différentes infrastructures d'attaques observées par Sekoia. Néanmoins, l'analyse produite peut présenter un biais puisqu'il faut prendre en compte le fait qu'il s'agit d'observations menées par Sekoia, une entreprise spécialisée en sécurité qui cherche à produire de la connaissance autour de menaces dont leurs clients peuvent être les victimes.

Chapitre 6

Conclusions et perspectives

La sécurité des réseaux reste un enjeu majeur de notre société actuelle. Il est nécessaire de continuer à mitiger les attaques, et cela passe entre autres par une meilleure connaissance de celle-ci. Au travers de cette thèse, nous avons souhaité établir une méthodologie permettant d'approfondir les connaissances sur les infrastructures d'attaque ainsi qu'une proposition d'implémentation de celle-ci. Pour cela, nous nous sommes concentrés sur l'analyse des machines participantes aux attaques, en proposant d'investiguer les infrastructures aussi bien sous un angle technique qu'un angle organisationnel. Nous avons vu que ces infrastructures sont un enjeu majeur de la caractérisation des attaques et que le suivi de celles-ci est un moyen efficace de protection.

De nombreux travaux traitent aujourd'hui si bien de la caractérisation des machines présentes sur Internet, mais également de la caractérisation de celles-ci dans un contexte de sécurité. Or, ces analyses s'appuient principalement sur des composantes techniques des machines ou des infrastructures. Nous avons questionné la nécessité d'intégrer des composantes socio-organisationnelles dans les analyses d'infrastructure. C'est une dimension qui doit nécessairement être prise en compte, puisque les attaques ne résultent pas seulement de failles techniques, mais s'inscrivent dans des contextes sociaux, politiques, économiques et organisationnels.En outre, les commanditaires des attaques demeurent des individus, il est donc envisageable d'observer des comportements, tant au niveau de l'attaque que de la sélection des ressources pour la mener. Nous affirmons donc qu'on ne peut laisser la dimension d'analyse socio-organisationnelle de côté dans la caractérisation des infrastructures d'attaques.

Nous avons présenté une analyse socio-organisationnelle des réseaux dans cette thèse, en identifiant des catégories organisationnelles que nous souhaitons distinguer. Ces classifications sont le fruit d'une réflexion sur les divers acteurs présents sur Internet.

Pour départager ces acteurs, nous avons établi une liste de labels, qui peut être décomposée en deux axes : un axe d'analyse organisationnelle et un axe d'analyse technique. Ces labels s'appuient sur des sources de données externes : des crawlers réseaux, une base de géolocalisation, les bases de données des RIRs accessibles via RDAP et enfin Wikidata. Pour obtenir toutes ces informations, nous avons développé un outil nommé IPSeen. Il existe deux implémentations de cet algorithme : une version rapide qui permet d'avoir un premier aperçu rapide et une version qui nécessite plus de temps d'analyse, mais qui apporte des résultats plus précis. Notre outil IPSeen est accessible à tous [160].

Cet outil et ces deux versions ont été évalués en profondeur. Pour chacun, nous avons

analysé les points forts et identifié les différentes limites. Les deux versions ont été comparées et chacune présente des avantages : la rapidité pour la première et la précision pour la seconde. Les deux versions proposent des résultats satisfaisants, même si perfectibles. IPSeen présente des perspectives nombreuses et enthousiasmantes pour la suite. Nous avons également évalué nos différentes sources de données. Nous avons montré qu'il s'agit de sources de confiance, mais que le maintien à jour des informations est un processus complexe qui n'est pas toujours réalisé dans les meilleurs délais.

Enfin, nous avons appliqué notre méthodologie à un ensemble de données d'attaque. Cet ensemble de données nous a été fourni par une entreprise française de sécurité, Sekoia. Cet ensemble de données est composé d'adresses IPs qui ont été observées par Sekoia dans leurs activités de tracking des infrastructures de command and control. Parmi cet ensemble, nous avons analysé des malwares, des outils, des groupes d'attaques ainsi qu'un virus. Nous avons classifié le set de données en 4 catégories et jusqu'à 13 sous catégories. Ce travail résulte d'une investigation des différents labels de l'ensemble de données. Ces analyses nous ont permis d'établir les conclusions suivantes. Tout d'abord, nous avons pu noter une importante participation aux infrastructures d'attaques de la part des machines de type hébergement. Nous avons pu identifier plus de 1000 entités distinctes parmi lesquelles 69 entités reviennent plus de 50 fois. Enfin, nous avons pu voir que des entités de toute taille étaient concernées par la présence de machines d'infrastructure d'attaque. Nous avons également proposé une analyse dans le temps des adresses IPs et avons vu que certaines des IPs ont pu être observées sur des périodes très longues.

À présent, si nous revenons à notre objectif initial, c'est-à-dire la caractérisation des machines participant aux attaques, nous pouvons voir que nous avons apporté de nouvelles analyses. Ces analyses s'appuient sur une méthodologie et un outil que nous avons développés. Par ailleurs, nous affirmons qu'il est aujourd'hui nécessaire d'inclure un aspect, social, organisationnel, économique, lorsque l'on s'intéresse à la caractérisation des attaques. Cela n'est pas seulement valable pour la caractérisation des motivations des attaquants, mais également pour l'analyse des infrastructures utilisées. En effet, les politiques de sélection des machines membres de l'infrastructure peuvent varier en fonction des moyens, des besoins, des objectifs de l'attaque. L'apport d'information qualitative est donc pertinent et permet également d'envisager des solutions à déployer au plus près des machines participant à l'infrastructure.

L'originalité de notre travail s'appuie sur l'analyse organisationnelle des entités qui possèdent les machines membres d'infrastructures d'attaque. Nous proposons d'ajouter aux caractérisations techniques des adresses IPs, des labels qui servent à décrire les entités possédant les machines. L'ensemble de ces labels permet d'avoir une meilleure observation du système et, par conséquent, d'affiner les contre-mesures possibles et de lutter contre les attaques à des points dans le réseau différents que celui de la victime. Cela permet aussi de lutter contre la centralisation d'Internet derrière les acteurs de la protection.

Néanmoins, notre outil possède des limites que nous avons identifiées grâce à notre évaluation et notre application. Tout d'abord, nous avons vu que notre outil est dépendant de la qualité des bases d'informations externes que nous avons sélectionnées. Nous les avons évaluées et

même si nous pouvons avoir confiance dans ces sources de données, elles contiennent parfois des erreurs pour Wikidata et/ou des informations pas assez complètes ou précises pour les bases de données des RIRs ou Wikidata. Il serait donc pertinent d'envisager l'ajout de sources d'informations supplémentaires afin d'améliorer la qualité des informations en croisant les données. Par exemple, nous avons vu que d'autres sources de données permettent d'obtenir des informations similaires, qu'il s'agisse au niveau des informations techniques que des informations organisationnelles. Il serait donc pertinent d'envisager le croisement des sources pour augmenter la confiance de nos résultats ainsi que de combler des résultats potentiellement manquants. De plus, nous avons mis en évidence lors de l'évaluation de notre solution qu'il existe une marge de progression pour notre outil. Chaque étape peut être améliorée pour atteindre des résultats plus précis. Cela passe par l'optimisation des outils choisis en grande partie, ainsi que par l'étude d'autres ensembles d'apprentissages.

Les perspectives de notre travail sont encourageantes. Nous avons proposé plusieurs pistes d'amélioration et d'exploration que nous n'avons pas eu le temps de finir d'explorer, ou simplement seulement commencé à envisager.

Premièrement, nous avons fait le choix de choisir une base collaborative, Wikidata, afin d'obtenir des informations sur les différentes entités. Ce choix est motivé par la possibilité de complétion de la base une fois de nouvelles entités identifiées ou l'ajout des informations de possession de bloc IPv4 directement dans la base. Une piste de perspective serait donc de pouvoir automatiser la complétion des informations afin de mettre en avant l'aspect communautaire de notre solution. Le langage SPARQL permet d'automatiser des contributions vers Wikidata. La mise en place d'un outil permettant de faciliter la contribution pourrait s'avérer bénéfique pour accentuer l'aspect collaboratif de notre solution.

Une seconde piste, qui a été longuement présentée dans le chapitre 4, concerne l'analyse des typologies des réseaux. Nous avons commencé à explorer la possibilité d'identifier des typologies de réseaux en nous basant sur les qualifications des blocs d'adresses IP produites par des outils comme Onyphe. En effet, Onyphe propose également d'autres informations que nous n'avons pas exploitées, comme des analyses géographiques ou encore des analyses de certificat. Ce travail aurait pour but de pouvoir affecter les labels de type d'entité sur des entités qui n'existeraient pas dans Wikidata en analysant leurs topologies réseaux et en les comparant à des clusters préétablis. Nous pensons qu'il est pertinent de poursuivre plusieurs pistes pour l'analyse des topologies, comme le nombre de noms de domaines ainsi que leur variance sur un réseau, l'étude des services et les types de machines et de systèmes d'exploitation. Toutes ces composantes pourraient être pertinentes à mesurer et à intégrer dans un système de classification automatisé.

Une troisième piste d'amélioration concerne la gestion des adresses IPv6. Dans notre ensemble de données, nous n'avions que très peu d'adresses IPv6. Notre outil a été capable de caractériser ces adresses, mais il a principalement été développé pour caractériser des adresses de type IPv4. Il serait donc pertinent pour l'avenir d'enquêter sur les changements nécessaires pour des meilleures caractérisations des adresses IPv6, comme par exemple la recherche d'adresse IPv6 dans Wikidata. C'est une composante qui pourrait être rapidement intégrée à notre outil et qui pourrait permettre de faire des analyses sur le domaine d'IPv6.

Une quatrième piste d'amélioration concerne le potentiel de ressources offertes par Wikidata. En effet, nous avons établi 3 labels organisationnels pertinents, néanmoins Wikidata offre de

nombreux qualifiants sous la forme de labels. Nous pouvons donc envisager d'aller chercher d'autres labels pertinents qui pourraient approfondir la caractérisation des adresses IPs de manière générale. Par exemple, les marques associées, les types de condamnations, les cartes de localisation de l'entité ou encore des identifiants liant l'entité vers différentes encyclopédies.

En lien avec ceci, nous pourrions envisager d'autres domaines d'analyse, hors sécurité. Par exemple, afin de faire de la métrologie et d'adapter des services, notre solution pourrait aider des entreprises proposant des services afin d'affiner leurs connaissances sur les personnes qui joignent le service.

Enfin, une ultime piste d'analyse concerne un type d'attaque en particulier. Les attaques DDoS volumétriques semblent être un cas d'application très pertinent pour notre outil. En effet, limiter les attaques un cran plus haut dans la chaine est parfois nécessaire dans ce type d'attaque, car les contre-mesures existantes ne sont pas forcément efficaces face à l'afflux soudain de trafic toujours grandissant. Il en va de même pour les Botnets. Identifier les réseaux de bots et surtout leur localisation, non géographique, dans l'infrastructure peut être pertinente pour limiter les infrastructures de ce type. Il est plus facile de mettre en place les outils lorsque l'on sait où chercher les machines qui participent aux activités malveillantes.

Bibliographie

[1] Compatible time-sharing system (1961-1973) fiftieth anniversary commemorative overview. https ://www.multicians.org/thvv/compatible-time-sharing-system.pdf, dernière consultation le : 18/07/2024.

[2] Abhishta Abhishta, Marianne Junger, Reinoud Joosten, and Lambert JM Nieuwenhuis. A note on analysing the attacker aims behind ddos attacks. In Intelligent Distributed Computing XIII, pages 255–265. Springer, 2020.

[3] Nouvelle caledonie : une cyberattaque pas vraiment inedite et beaucoup de confusion. https ://www.lemonde.fr/article-offert/mlqakpvecpbw-6234855/nouvelle-caledonie-une-cyberattaque-pas-vraiment-inedite-et-beaucoup-de-confusion, dernière consultation le : 18/07/2024.

[4] Cyberattacks since the murder of george floyd. https ://blog.cloudflare.com/cyberattacks-since-the-murder-of-george-floyd/,dernière consultation le : 18/07/2024.

[5] The role of cyber in the russian war against ukraine : Its impact and the consequences for the future of armed conflict. https ://www.europarl.europa.eu/RegData/etudes/BRIE/2023/702594/EXPO_BRI(2023)702594_EN.pdf, dernière consultation le : 19/07/2024.

[6] Cyber attacks in the israel-hamas war. https ://blog.cloudflare.com/cyber-attacks-in-the-israel-hamas-war,dernière consultation le : 19/07/2024.

[7] Un symbole, une cause anonymous. https ://www.arte.tv/fr/videos/115512-005-A/un-symbole-une-cause/,dernière consultation le : 19/07/2024.

[8] Communiqué officiel. https ://x.com/LyoneSport/status/1099744167557808128 ?, dernière consultation le : 19/07/2024.

[9] Lck hit by severe ddos attacks. https ://ggboost.com/blog/post/lol-lck-hit-by-ddos, dernière consultation le : 19/07/2024.

[10] 12 most common types of cyberattacks. https ://www.crowdstrike.com/cybersecurity-101/cyberattacks/most-common-types-of-cyberattacks/, dernière consultation le : 19/07/2024.

[11] What is a cyberattack ? https ://www.cisco.com/c/en/us/products/security/common-cyberattacks.html, dernière consultation le : 19/07/2024.

[12] Rapport attaques ddos observées par ovh en 2017. https ://www.ovh.com/fr/blog/rapport-attaques-ddos-observees-par-ovh-en-2017/, dernière consultation le : 03/01/2023.

[13] 911 s5 botnet dismantled and its administrator arrested in coordinated international operation. https ://www.justice.gov/opa/pr/911-s5-botnet-dismantled-and-its-administrator-arrested-coordinated-international-operation, dernière consultation le : 19/07/2024.

[14] Generative ai is increasing cyber attacks. https ://www.deepinstinct.com/pdf/infographic-voice-of-secops-4th-edition-generative-ai-in-cyber, dernière consultation le : 19/07/2024.

[15] Wormgpt – the generative ai tool cybercriminals are using to launch business email compromise attacks. https ://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/, dernière consultation le : 19/07/2024.

[16] Digital attack map. https ://www.digitalattackmap.com/.

[17] Malpedia. https ://malpedia.caad.fkie.fraunhofer.de/.

[18] No more ransom. https ://www.nomoreransom.org/fr/decryption-tools.html.

[19] Eric Osterweil, Angelos Stavrou, and Lixia Zhang. 20 years of ddos : a call to action. arXiv preprint arXiv :1904.02739, 2019.

[20] Ddos attacks can threaten the independent internet. https ://www.macchaffee.com/blog/2024/ddos-attacks/, dernière consultation le : 19/07/2024.

[21] The democratization of censorship. https ://krebsonsecurity.com/2016/09/the-democratization-of-censorship/, dernière consultation le : 19/07/2024.

[22] Why we terminated daily stormer. https ://blog.cloudflare.com/why-we-terminated-daily-stormer/, dernière consultation le : 19/07/2024.

[23] Criticism received by cloudflare for content censorship. https ://www.cloudflare.com/cloudflare-criticism/, dernière consultation le : 19/07/2024.

[24] Cloudflare’s ceo is right : We can’t count on him to police the internet,will oremus. https ://slate.com/technology/2017/08/cloudflare-ceo-matthew-prince-is-right-we-can-t-count-on-him-to-police-online-speech.html, dernière consultation le : 19/07/2024.

[25] Ddos mitigation firm founder admits to ddos. https ://krebsonsecurity.com/2020/01/ddos-mitigation-firm-founder-admits-to-ddos/, dernière consultation le : 19/07/2024.

[26] Derrière le cauchemar iot mirai, le business lucratif des serveurs minecraft. https ://www.usine-digitale.fr/article/cybersecurite-derriere-le-cauchemar-iot-mirai-le-business-lucratif-des-serveurs-minecraft.N489934, dernière consultation le : 19/07/2024.

[27] Blake E Strom, Andy Applebaum, Doug P Miller, Kathryn C Nickels, Adam G Pennington, and Cody B Thomas. Mitre att&ck® : Design and philosophy. MITRE, 2020.

[28] Hacker releases 500,000 iot credentials. https ://www.iottechtrends.com/hacker-releases-500000-iot-credentials/, dernière consultation le : 19/07/2024.

[29] Kimberly M Christopherson. The positive and negative implications of anonymity in internet social interactions :“on the internet, nobody knows you’re a dog”. Computers in Human Behavior, 23(6) :3038–3056, 2007.

[30] Cristoffer Leite, Jerry Den Hartog, Daniel R Dos Santos, and Elisa Costante. Automated cyber threat intelligence generation on multi-host network incidents. In 2023 IEEE International Conference on Big Data (BigData), pages 2999–3008. IEEE, 2023.

[31] Bongsik Shin and Paul Benjamin Lowry. A review and theoretical explanation of the ‘cyberthreat-intelligence (cti) capability’that needs to be fostered in information security practitioners and how this can be accomplished. Computers & Security, 92 :101761, 2020.

[32] Ghaith Husari, Ehab Al-Shaer, Mohiuddin Ahmed, Bill Chu, and Xi Niu. Ttpdrill : Automatic and accurate extraction of threat actions from unstructured text of cti sources. In Proceedings of the 33rd Annual Computer Security Applications Conference, ACSAC '17, page 103–115, New York, NY, USA, 2017. Association for Computing Machinery.

[33] Sergio Caltagirone, Andrew Pendergast, and Christopher Betz. The diamond model of intrusion analysis. Threat Connect, 298(0704) :1–61, 2013.

[34] Eric M Hutchins, Michael J Cloppert, Rohan M Amin, et al. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. Leading Issues in Information Warfare & Security Research, 1(1) :80, 2011.

[35] Yussuf Ahmed, A. Taufiq Asyhari, and Md Arafatur Rahman. A cyber kill chain approach for detecting advanced persistent threats. Computers, Materials and Continua, 67(2) :2497–2513, 2021. Funding Information : Funding Statement : The work of Y. Ahmed and A. T. Asyhari was supported in part by the School of Computing and Digital Technology at Birmingham City University. The work of M. A. Rahman was supported in part by the Flagship Grant RDU190374. Publisher Copyright : © 2021 Tech Science Press. All rights reserved.

[36] Shanto Roy, Emmanouil Panaousis, Cameron Noakes, Aron Laszka, Sakshyam Panda, and George Loukas. Sok : The mitre att&ck framework in research and practice. arXiv preprint arXiv :2304.07411, 2023.

[37] Nitin Naik, Paul Jenkins, Paul Grace, and Jingping Song. Comparing attack models for it systems : Lockheed martin's cyber kill chain, mitre attck framework and diamond model. In 2022 IEEE International Symposium on Systems Engineering (ISSE), pages 1–7, 2022.

[38] Adrien Hemmer, Mohamed Abderrahim, Remi Badonnel, and Isabelle Chrisment. An ensemble learning-based architecture for security detection in iot infrastructures. In 2021 17th International Conference on Network and Service Management (CNSM), pages 180–186, 2021.

[39] Tanwir Ahmad, Dragos Truscan, Jüri Vain, and Ivan Porres. Early detection of network attacks using deep learning. In 2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), pages 30–39, 2022.

[40] Lyman Chapin, Dr. David D. Clark, Robert T. Braden, Russ Hobby, and Dr. Vinton G. Cerf. Towards the Future Internet Architecture. RFC 1287, December 1991.

[41] Internet Protocol. RFC 791, September 1981.

[42] Bob Hinden and Dr. Steve E. Deering. Internet Protocol, Version 6 (IPv6) Specification. RFC 2460, December 1998.

[43] Wesley Eddy. Transmission Control Protocol (TCP). RFC 9293, August 2022.

[44] User Datagram Protocol. RFC 768, August 1980.

[45] Baromètre annuel de la transition vers ipv6 en france. https ://www.arcep.fr/cartes-et-donnees/nos-publications-chiffrees/transition-ipv6/barometre-annuel-de-la-transition-vers-ipv6-en-france.html, dernière consultation le : 19/07/2024.

[46] Quaizar Vohra and Enke Chen. BGP Support for Four-Octet Autonomous System (AS) Number Space. RFC 6793, December 2012.

[47] Caida as rank. http ://as-rank.caida.org/.

[48] Peeringdb. https ://www.peeringdb.com/.

[49] Annika Baumann and Benjamin Fabian. Who runs the internet ? classifying autonomous systems into industries. volume 1, 04 2014.

[50] Maya Ziv, Liz Izhikevich, Kimberly Ruth, Katherine Izhikevich, and Zakir Durumeric. Asdb : a system for classifying owners of autonomous systems. In Proceedings of the 21st ACM Internet Measurement Conference, IMC '21, page 703–719, New York, NY, USA, 2021. Association for Computing Machinery.

[51] Leslie Daigle. Whois protocol specification. Technical report, 2004.

[52] Quel est le rôle de l'icann ? https ://www.icann.org/resources/pages/what-2012-02-25-fr.

[53] Michael M. Roberts, Fred Baker, and Brian E. Carpenter. Memorandum of Understanding Concerning the Technical Work of the Internet Assigned Numbers Authority. RFC 2860, June 2000.

[54] Ripe ncc. https ://www.ripe.net/.

[55] Arin. https ://www.arin.net/.

[56] Apnic. https ://www.apnic.net/.

[57] Afrinic. https ://www.afrinic.net/fr/.

[58] Lacnic. https ://www.lacnic.net/.

[59] Ivan Cvitić, Dragan Peraković, Marko Periša, and Mate Botica. Novel approach for detection of iot generated ddos traffic. Wireless Networks, 2021.

[60] Kai Yang, Qiang Li, and Limin Sun. Towards automatic fingerprinting of iot devices in the cyberspace. Computer Networks, 148 :318–327, 2019.

[61] Bruhadeshwar Bezawada, Maalvika Bachani, Jordan Peterson, Hossein Shirazi, Indrakshi Ray, and Indrajit Ray. Behavioral fingerprinting of iot devices. In Proceedings of the 2018 Workshop on Attacks and Solutions in Hardware Security, ASHES '18, page 41–50, New York, NY, USA, 2018. Association for Computing Machinery.

[62] Jon Postel. Internet protocol. Technical report, 1981.

[63] Leigh Metcalf and Jonathan M. Spring. Blacklist ecosystem analysis : Spanning jan 2012 to jun 2014. In Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security, WISCS '15, page 13–22, New York, NY, USA, 2015. Association for Computing Machinery.

[64] Arya Renjan, Karuna Pande Joshi, Sandeep Nair Narayanan, and Anupam Joshi. Dabr : Dynamic attribute-based reputation scoring for malicious ip address detection. In 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), pages 64–69, 2018.

[65] Henanksha Sainani, Josephine M. Namayanja, Guneeti Sharma, Vasundhara Misal, and Vandana P. Janeja. Ip reputation scoring with geo-contextual feature augmentation. ACM Trans. Manage. Inf. Syst., 11(4), oct 2020.

[66] George Tsirtsis and Pyda Srisuresh. Network address translation-protocol translation (nat-pt). Technical report, 2000.

[67] Sivaramakrishnan Ramanathan, Anushah Hossain, Jelena Mirkovic, Minlan Yu, and Sadia Afroz. Quantifying the impact of blocklisting in the age of address reuse. In IMC '20 : ACM Internet Measurement Conference, 2020.

[68] Vikas Mishra, Pierre Laperdrix, Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Martin Lopatka. Don't count me out : On the relevance of IP address in the tracking ecosystem. In WWW '20 : The Web Conference 2020.

[69] T. Kohno, A. Broido, and K.C. Claffy. Remote physical device fingerprinting. IEEE Transactions on Dependable and Secure Computing, 2(2) :93–108, 2005.

[70] Steven M Bellovin. A technique for counting natted hosts. In Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment, pages 267–272, 2002.

[71] Ali Safari Khatouni, Lan Zhang, Khurram Aziz, Ibrahim Zincir, and Nur Zincir-Heywood. Exploring nat detection and host identification using machine learning. In 2019 15th International Conference on Network and Service Management (CNSM), pages 1–8. IEEE, 2019.

[72] Sanjeev Shukla and Himanshu Gupta. Identification and counting of hosts behind nat using machine learning. SN Computer Science, 3(2) :126, 2022.

[73] Hanbyeol Park, Seung-hun Shin, Byeong-hee Roh, and Cheolho Lee. Identification of hosts behind a nat device utilizing multiple fields of ip and tcp. In 2016 International Conference on Information and Communication Technology Convergence (ICTC), pages 484–486, 2016.

[74] Nino Vincenzo Verde, Giuseppe Ateniese, Emanuele Gabrielli, Luigi Vincenzo Mancini, and Angelo Spognardi. No nat'd user left behind : Fingerprinting users behind nat from netflow records alone. In 2014 IEEE 34th International Conference on Distributed Computing Systems, pages 218–227. IEEE, 2014.

[75] Sarthak Grover, Mi Seon Park, Srikanth Sundaresan, Sam Burnett, Hyojoon Kim, Bharath Ravi, and Nick Feamster. Peeking behind the nat : an empirical study of home networks. In Proceedings of the 2013 conference on Internet measurement conference, pages 377–390, 2013.

[76] Yair Meidan, Vinay Sachidananda, Hongyi Peng, Racheli Sagron, Yuval Elovici, and Asaf Shabtai. A novel approach for detecting vulnerable iot devices connected behind a home nat. Computers Security, 97 :101968, 2020.

[77] Daniel Senie. Network ingress filtering : Defeating denial of service attacks which employ ip source address spoofing. RFC 2827, 2000.

[78] Zhenhai Duan, Xin Yuan, and Jaideep Chandrashekar. Controlling ip spoofing through interdomain packet filters. IEEE transactions on Dependable and Secure computing, 5(1) :22–36, 2008.

[79] M Luckie, R Beverly, R Koga, K Keys, J Kroll, and k claffy. Network Hygiene, Incentives, and Regulation : Deployment of Source Address Validation in the Internet. In ACM Computer and Communications Security (CCS), Nov 2019.

[80] Marcin Nawrocki, Mattijs Jonker, Thomas C. Schmidt, and Matthias Wählisch. The far side of dns amplification : tracing the ddos attack ecosystem from the internet core. In Proceedings of the 21st ACM Internet Measurement Conference, IMC '21, page 419–434, New York, NY, USA, 2021. Association for Computing Machinery.

[81] Jakub Czyz, Michael Kallitsis, Manaf Gharaibeh, Christos Papadopoulos, Michael Bailey, and Manish Karir. Taming the 800 pound gorilla : The rise and decline of ntp ddos attacks. In Proceedings of the 2014 Conference on Internet Measurement Conference, pages 435–448, 2014.

[82] Mizuki Kondo, Rui Tanabe, Natsuo Shintani, Daisuke Makita, Katsunari Yoshioka, and Tsutomu Matsumoto. Amplification chamber : Dissecting the attack infrastructure of memcached drdos attacks. In Lorenzo Cavallaro, Daniel Gruss, Giancarlo Pellegrino, and Giorgio Giacinto, editors, Detection of Intrusions and Malware, and Vulnerability Assessment, pages 178–196, Cham, 2022. Springer International Publishing.

[83] Qiang Xu, Rong Zheng, Walid Saad, and Zhu Han. Device fingerprinting in wireless networks : Challenges and opportunities. IEEE Communications Surveys Tutorials, 18(1) :94–104, 2016.

[84] Patricia Callejo, Marco Gramaglia, Rubén Cuevas, and Ángel Cuevas. A deep dive into the accuracy of ip geolocation databases and its impact on online advertising. IEEE Transactions on Mobile Computing, 22(8) :4359–4373, 2023.

[85] Ethan Katz-Bassett, John P. John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. Towards ip geolocation using delay and topology measurements. In Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, IMC '06, page 71–84, New York, NY, USA, 2006. Association for Computing Machinery.

[86] Qian Zhao, Fei Wang, Can Huang, and Chuan Yu. Improving ip geolocation databases based on multi-method classification. In 2020 IEEE 14th International Conference on Anti-counterfeiting, Security, and Identification (ASID), pages 44–48, 2020.

[87] Ethan Katz-Bassett, John P John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. Towards ip geolocation using delay and topology measurements. In Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, 2006.

[88] Ovidiu Dan, Vaibhav Parikh, and Brian D Davison. Ip geolocation through reverse dns. ACM Transactions on Internet Technology (TOIT), 22(1) :1–29, 2021.

[89] Brian Eriksson, Paul Barford, Joel Sommers, and Robert Nowak. A learning-based approach for ip geolocation. In Arvind Krishnamurthy and Bernhard Plattner, editors, Passive and Active Measurement, pages 171–180, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[90] Hellen Maziku, Sachin Shetty, Keesook Han, and Tamara Rogers. Enhancing the classification accuracy of ip geolocation. In MILCOM 2012 - 2012 IEEE Military Communications Conference, pages 1–6, 2012.

[91] Phillipa Gill, Yashar Ganjali, and Bernard Wong. Dude, where's that {IP}? circumventing measurement-based {IP} geolocation. In 19th USENIX Security Symposium (USENIX Security 10), 2010.

[92] Geolite database. https ://dev.maxmind.com/geoip/geolite2-free-geolocation-data.

[93] Geoip database. https ://www.maxmind.com/en/geoip-databases.

[94] Ip2location lite database. https ://lite.ip2location.com/.

[95] Geonetmap database. https ://geobytes.com/iplocator/.

[96] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. Ip geolocation databases : unreliable ? SIGCOMM Comput. Commun. Rev., 41(2) :53–56, apr 2011.

[97] Yuval Shavitt and Noa Zilberman. A geolocation databases study. IEEE Journal on Selected Areas in Communications, 29(10) :2044–2056, 2011.

[98] James Saxon and Nick Feamster. Gps-based geolocation of consumer ip addresses. In Oliver Hohlfeld, Giovane Moura, and Cristel Pelsser, editors, Passive and Active Measurement, pages 122–151, Cham, 2022. Springer International Publishing.

[99] Ioana Livadariu, Thomas Dreibholz, Anas Saeed Al-Selwi, Haakon Bryhni, Olav Lysne, Steinar Bjørnstad, and Ahmed Elmokashfi. On the accuracy of country-level IP geolocation. In ANRW '20 : Applied Networking Research Workshop, 2020, pages 67–73, 2020.

[100] Irom Lalit Meitei, Khundrakpam Johnson Singh, and Tanmay De. Detection of ddos dns amplification attack using classification algorithm. In Proceedings of the International Conference on Informatics and Analytics, ICIA-16, New York, NY, USA, 2016. Association for Computing Machinery.

[101] Nmap. https ://nmap.org/.

[102] Zakir Durumeric, Eric Wustrow, and J Alex Halderman. {ZMap} : fast internet-wide scanning and its security applications. In 22nd USENIX Security Symposium (USENIX Security 13), pages 605–620, 2013.

[103] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J. Alex Halderman. A search engine backed by internet-wide scanning. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, pages 542–553, 2015.

[104] Censys. https ://search.censys.io/.

[105] Shodan. https ://www.shodan.io/.

[106] Onyphe. https ://www.onyphe.io/.

[107] Natalija Vlajic and Daiwei Zhou. Iot as a land of opportunity for ddos hackers. Computer, 51(7) :26–34, 2018.

[108] Roland Bodenheim, Jonathan Butts, Stephen Dunlap, and Barry E. Mullins. Evaluation of the ability of the shodan search engine to identify internet-facing industrial control devices. Int. J. Crit. Infrastructure Prot., 7 :114–123, 2014.

[109] Béla Genge and Călin Enăchescu. Shovat : Shodan-based vulnerability assessment tool for internet-facing services. Security and Communication Networks, 9(15) :2696–2714, 2016.

[110] Haneen Al-Alami, Ali Hadi, and Hussein Al-Bahadili. Vulnerability scanning of iot devices in jordan using shodan. In 2017 2nd International Conference on the Applications of Information Technology in Developing Renewable Energy Processes Systems (IT-DREPS), pages 1–6, 2017.

[111] Marc Arnaert, Yoann Bertrand, and Karima Boudaoud. Modeling vulnerable internet of things on shodan and censys : An ontology for cyber security. In Proceedings of the Tenth International Conference on Emerging Security Information, Systems and Technologies (SECUREWARE 2016), pages 299–302, 2016.

[112] Andrea Tundis, Wojciech Mazurczyk, and Max Mühlhäuser. A review of network vulnerabilities scanning tools : types, capabilities and functioning. In Proceedings of the 13th International Conference on Availability, Reliability and Security, ARES '18, New York, NY, USA, 2018. Association for Computing Machinery.

[113] Seungwoon Lee, Seung-Hun Shin, and Byeong-hee Roh. Abnormal behavior-based detection of shodan and censys-like scanning. In 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), pages 1048–1052, 2017.

[114] Kim Hubbard, Mark Kosters, David Conrad, Daniel Karrenberg, and Jon Postel. Internet registry ip allocation guidelines. Technical report, 1996.

[115] Russ Housley, John Curran, Geoff Huston, and D Conrad. The internet numbers registry system. Technical report, 2013.

[116] Andrew Newton, Byron Ellacott, and Ning Kong. Http usage in the registration data access protocol (rdap). Technical report, 2015.

[117] A Newton and S Hollenbeck. Registration data access protocol (rdap) query format. Technical report, 2015.

[118] S Hollenbeck and A Newton. Rfc 9083 : Json responses for the registration data access protocol (rdap), 2021.

[119] Ipinfo.io. https ://ipinfo.io/.

[120] Ipinfodb. https ://www.ipinfodb.com/.

[121] Yoshitaka Nakamura, Shihori Kanazawa, Hiroshi Inamura, and Osamu Takahashi. Classification of unknown web sites based on yearly changes of distribution information of malicious ip addresses. In 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), pages 1–4, 2018.

[122] Sophie Gastellier-Prevost and Maryline Laurent. Defeating pharming attacks at the client-side. In 2011 5th International Conference on Network and System Security, pages 33–40. IEEE, 2011.

[123] Maryam Feily, Alireza Shahrestani, and Sureswaran Ramadass. A survey of botnet and botnet detection. In 2009 Third International Conference on Emerging Security Information, Systems and Technologies, pages 268–273. IEEE, 2009.

[124] Maryam Feily, Alireza Shahrestani, and Sureswaran Ramadass. A survey of botnet and botnet detection. In 2009 Third International Conference on Emerging Security Information, Systems and Technologies, pages 268–273, 2009.

[125] Jehyun Lee, Jonghun Kwon, Hyo-Jeong Shin, and Heejo Lee. Tracking multiple cc botnets by analyzing dns traffic. In 2010 6th IEEE Workshop on Secure Network Protocols, pages 67–72, 2010.

[126] Markus Ring, Alexander Dallmann, Dieter Landes, and Andreas Hotho. Ip2vec : Learning similarities between ip addresses. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pages 657–666. IEEE, 2017.

[127] Leon Böck, Martin Fejrskov, Katerina Demetzou, Shankar Karuppayah, Max Mühlhäuser, and Emmanouil Vasilomanolakis. Processing of botnet tracking data under the gdpr. Computer Law Security Review, 45 :105652, 2022.

[128] Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J Alex Halderman, Luca Invernizzi, Michalis Kallitsis, et al. Understanding the mirai botnet. In 26th {USENIX} security symposium ({USENIX} Security 17), pages 1093–1110, 2017.

[129] Sérgio SC Silva, Rodrigo MP Silva, Raquel CG Pinto, and Ronaldo M Salles. Botnets : A survey. Computer Networks, 57(2) :378–403, 2013.

[130] Georgios Kambourakis, Constantinos Kolias, and Angelos Stavrou. The mirai botnet and the iot zombie armies. In MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM), pages 267–272, 2017.

[131] Teng-Fei Tu, Jia-Wei Qin, Hua Zhang, Miao Chen, Tong Xu, and Yue Huang. A comprehensive study of mozi botnet. International Journal of Intelligent Systems, 37(10) :6877–6908, 2022.

[132] Ihsan Ali, Abdelmuttlib Ibrahim Abdalla Ahmed, Ahmad Almogren, Muhammad Ahsan Raza, Syed Attique Shah, Anwar Khan, and Abdullah Gani. Systematic literature review on iot-based botnet attack. IEEE Access, 8 :212220–212232, 2020.

[133] Bryson Lingenfelter, Iman Vakilinia, and Shamik Sengupta. Analyzing variation among iot botnets using medium interaction honeypots. In 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), pages 0761–0767, 2020.

[134] Saman Taghavi Zargar, James Joshi, and David Tipper. A survey of defense mechanisms against distributed denial of service (ddos) flooding attacks. IEEE communications surveys & tutorials, 15(4) :2046–2069, 2013.

[135] Abhishta Abhishta, Wouter van Heeswijk, Marianne Junger, Lambert JM Nieuwenhuis, and Reinoud Joosten. Why would we get attacked ? an analysis of attacker's aims behind ddos attacks. J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl., 11(2) :3–22, 2020.

[136] Derek Manky. Cybercrime as a service : a very modern business. Computer Fraud & Security, 2013(6) :9–13, 2013.

[137] Kraesten L Arnold, Paul AL Ducheine, and Peter BMJ Pijpers. Comprehensive cyberattack chain. Amsterdam Law School Research Paper, (2023-02), 2023.

[138] Alex Kigerl. Cyber crime nation typologies : K-means clustering of countries based on cyber crime rates. International Journal of Cyber Criminology, 10(2), 2016.

[139] Jonathan Lusthaus, Miranda Bruce, and Nigel Phair. Mapping the geography of cybercrime : A review of indices of digital offending by country. In 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pages 448–453. IEEE, 2020.

[140] Christian Konradt, Andreas Schilling, and Brigitte Werners. Phishing : An economic analysis of cybercrime perpetrators. Computers & Security, 58 :39–46, 2016.

[141] Xianghang Mi, Xuan Feng, Xiaojing Liao, Baojun Liu, XiaoFeng Wang, Feng Qian, Zhou Li, Sumayah A. Alrwais, Limin Sun, and Ying Liu. Resident evil : Understanding residential IP proxy as a dark service. In 2019 IEEE Symposium on Security and Privacy, SP 2019, pages 1185–1201, 2019.

[142] Kassidy Clark, Martijn Warnier, and Frances MT Brazer. The future of cloud-based botnets ? In CLOSER 2011–International Conference on Cloud Computing and Services Science, pages 597–603. SciTePress–Science and Technology Publications Noordwijkerhout, 2011.

[143] Johannes Krupp, Michael Backes, and Christian Rossow. Identifying the scan and attack infrastructures behind amplification ddos attacks. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16, page 1426–1437, New York, NY, USA, 2016. Association for Computing Machinery.

[144] Oleg Boyarchuk, Sebastiano Mariani, Stefano Ortolani, and Giovanni Vigna. Keeping up with the emotets : Tracking a multi-infrastructure botnet. Digital Threats : Research and Practice, 4(3) :1–29, 2023.

[145] Hatem A. Almazarqi, Mathew Woodyard, Troy Mursch, Dimitrios Pezaros, and Angelos K. Marnerides. Macroscopic analysis of iot botnets. In GLOBECOM 2022 - 2022 IEEE Global Communications Conference, pages 2674–2679, 2022.

[146] Hatem A. Almazarqi, Angelos K. Marnerides, Troy Mursch, Mathew Woodyard, and Dimitrios Pezaros. Profiling iot botnet activity in the wild. In 2021 IEEE Global Communications Conference (GLOBECOM), pages 1–6, 2021.

[147] Who makes it work ? https ://www.internetsociety.org/internet/who-makes-it-work/.

[148] Internet governance – why the multistakeholder approach works. https ://www.internetsociety.org/resources/doc/2016/internet-governance-why-the-multistakeholder-approach-works/.

[149] Internet ecosystem. https ://www.centr.org/library/library/download/8452/4572/41.html.

[150] Amogh Dhamdhere and Constantine Dovrolis. Ten years in the evolution of the internet ecosystem. In Proceedings of the 8th ACM SIGCOMM conference on Internet measurement, pages 183–196, 2008.

[151] David Murray, Terry Koziniec, Sebastian Zander, Michael Dixon, and Polychronis Koutsakis. An analysis of changing enterprise network traffic characteristics. In 2017 23rd Asia-Pacific Conference on Communications (APCC), pages 1–6. IEEE, 2017.

[152] David Murray and Terry Koziniec. The state of enterprise network traffic in 2012. 10 2012.

[153] Jie Li, Andreas Aurelius, Viktor Nordell, Manxing Du, Åke Arvidsson, and Maria Kihl. A five year perspective of traffic pattern evolution in a residential broadband access network. In 2012 Future Network & Mobile Summit (FutureNetw), pages 1–9. IEEE, 2012.

[154] Gregor Maier, Anja Feldmann, Vern Paxson, and Mark Allman. On dominant characteristics of residential broadband internet traffic. In Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, pages 90–102, 2009.

[155] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, et al. The lockdown effect : Implications of the covid-19 pandemic on internet traffic. In Proceedings of the ACM internet measurement conference, pages 1–18, 2020.

[156] Baromètre du numérique. https ://www.arcep.fr/uploads/tx_gspublication/barometre-du-numerique_2023_infographie_mai2024.pdf.

[157] Kompass. https ://fr.kompass.com/.

[158] Linkedin api. https ://developer.linkedin.com/.

[159] Wikidata. https ://www.wikidata.org/wiki/Wikidata :Main_Page.

[160] Ipseen. https ://gitlab.inria.fr/cmoriot/ipseen.

[161] Camille Moriot, François Lesueur, Nicolas Stouls, and Fabrice Valois. How to build socio-organizational information from remote ip addresses to enrich security analysis ? In 2022 IEEE 47th Conference on Local Computer Networks (LCN), pages 287–290. IEEE, 2022.

[162] Matthew A Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. Journal of the American Statistical Association, 84 :414–420, 1989.

[163] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn : Machine learning in python. the Journal of machine Learning research, 12 :2825–2830, 2011.

[164] Seungwoon Lee, Sun-young Im, Seung-Hun Shin, Byeong-hee Roh, and Cheolho Lee. Implementation and vulnerability test of stealth port scanning attacks using zmap of censys engine. In 2016 International Conference on Information and Communication Technology Convergence (ICTC), pages 681–683, 2016.

[165] Béla Genge and Călin Enăchescu. Shovat : Shodan-based vulnerability assessment tool for internet-facing services. Security and communication networks, 9(15) :2696–2714, 2016.

[166] Haneen Al-Alami, Ali Hadi, and Hussein Al-Bahadili. Vulnerability scanning of iot devices in jordan using shodan. In 2017 2nd International Conference on the Applications of Information Technology in Developing Renewable Energy Processes & Systems (IT-DREPS), pages 1–6. IEEE, 2017.

[167] Stefan Heindorf, Martin Potthast, Benno Stein, and Gregor Engels. Vandalism detection in wikidata. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, page 327–336, New York, NY, USA, 2016. Association for Computing Machinery.

[168] Set de données de sekoia. https ://github.com/SEKOIA-IO/Sekoia.io_ipseen_dataset/.

[169] Command and control, matrice attck, mitre. https ://attack.mitre.org/tactics/TA0011/.

Annexe A

Exemple de résultat entier obtenu via le protocole RDAP pour l'adresse IP 134.214.58.24

Résultats au 06/11/2023.


```
{
  {"nir": null,
  "asn_registry": "ripencc",
  "asn": "2060",
  "asn_cidr": "134.214.0.0/16",
  "asn_country_code": "FR",
  "asn_date": "1989-07-25",
  "asn_description": "FR-RENATER RENATER_ASNBLOCK1, EU",
  "query": "134.214.58.24",
  "network": {
    "handle": "134.214.0.0 - 134.214.255.255",
    "status": ["active"],
    "remarks":
        [{"title": null,
        "description": "Reseau Optique du Campus de la Doua\nCentre Inter-
           etablissement pour les Services Reseaux\nUniversite Claude Bernard
           Lyon 1, bat Braconnier\n21 avenue Claude Bernard\n69622 Villeurbanne
            CEDEX, France\ntrouble reports : reseau@univ-lyon1.fr\nchanged:
           rensvp@renater.fr 19990908\nchanged: rensvp@renater.fr 20110111", "
           links": null}],
    "notices":
        [{"title": "Filtered",
            "description": "This output has been filtered.",
            "links": null},
        {"title": "Whois Inaccuracy Reporting",
            "description": "If you see inaccuracies in the results, please
               visit:",
            "links": ["https://www.ripe.net/contact-form?topic=ripe_dbm&
               show_form=true"]},
```

```
        {"title": "Source",
            "description": "Objects returned came from source\nRIPE",
            "links": null},
        {"title": "Terms and Conditions",
            "description": "This is the RIPE Database query service. The
               objects are in RDAP format.",
            "links": ["http://www.ripe.net/db/support/db-terms-conditions.pdf"]
               }],

    "links": [
        "https://rdap.db.ripe.net/ip/134.214.58.24",
        "http://www.ripe.net/data-tools/support/documentation/terms"],
    "events": [
        {"action": "registration",
            "timestamp": "2001-11-02T14:06:59Z",
            "actor": null},
        {"action": "last changed",
            "timestamp": "2019-12-04T13:06:16Z",
            "actor": null}],
    "raw": null,
    "start_address": "134.214.0.0",
    "end_address": "134.214.255.255",
    "cidr": "134.214.0.0/16",
    "ip_version": "v4",
    "type": "LEGACY",
    "name": "FR-ROCAD",
    "country": "FR",
    "parent_handle": "0.0.0.0 - 255.255.255.255"
    },
    "entities": ["CU291-RIPE", "RENATER-MNT", "RS7087-RIPE", "TP2569-RIPE"],
    "objects": {"CU291-RIPE": {
        "handle": "CU291-RIPE",
        "status": null,
        "remarks": null,
        "notices": null,
        "links": ["https://rdap.db.ripe.net/entity/CU291-RIPE", "http://www.
           ripe.net/data-tools/support/documentation/terms"],
        "events": null,
        "raw": null,
        "roles": ["administrative"],
        "contact": {
            "name": "CISR U-1-LYON",
            "kind": "group",
            "address": [
                {"type": null,
```

```
                "value": "Centre Inter-etablissement pour les Services Reseaux\
                   nUniversite Claude Bernard Lyon 1\n21 Avenue Claude Bernard\
                   n69622 Villeurbanne CEDEX, France"}],
            "phone": [{"type": "voice", "value": "+33 4 72 44 79 99"},
                {"type": "fax", "value": "+33 4 72 44 84 10"}],
            "email": [{"type": "email", "value": "reseau@univ-lyon1.fr"}, {"
               type": "abuse", "value": "reseau.securite@univ-lyon1.fr"}],
            "role": null,
            "title": null},
            "events_actor": null,
            "entities": null},
        "RENATER-MNT": {
            "handle": "RENATER-MNT",
            "status": null,
            "remarks": null,
            "notices": null,
            "links": ["https://rdap.db.ripe.net/entity/RENATER-MNT", "http://
               www.ripe.net/data-tools/support/documentation/terms"],
            "events": null,
            "raw": null,
            "roles": ["registrant"],
            "contact": {
                "name": "RENATER-MNT",
                "kind": "individual",
                "address": null,
                "phone": null,
                "email": null,
                "role": null,
                "title": null},
            "events_actor": null,
            "entities": null},
        "RS7087-RIPE": {
            "handle": "RS7087-RIPE",
            "status": null,
            "remarks": null,
            "notices": null,
            "links": ["https://rdap.db.ripe.net/entity/RS7087-RIPE", "http://
               www.ripe.net/data-tools/support/documentation/terms"],
            "events": null,
            "raw": null,
            "roles": ["technical"],
            "contact": {
                "name": "Remi SADER",
                "kind": "individual",
```

```
                "address": [{"type": null, "value": "Centre Inter-etablissement
                    pour les Services Reseaux\nUniversite Claude Bernard Lyon 1
                   , bat Braconnier\n21 avenue Claude Bernard\n69622
                   Villeurbanne CEDEX,"}],
                "phone": [{"type": "voice", "value": "+33 4 72 44 79 99"}, {"
                   type": "fax", "value": "+33 4 72 44 84 10"}],
                "email": [{"type": "email", "value": "Remi.Sader@univ-lyon1.fr"
                   }],
                "role": null, "title": null},
            "events_actor": null,
            "entities": null},
        "TP2569-RIPE": {
            "handle": "TP2569-RIPE",
            "status": null,
            "remarks": null,
            "notices": null,
            "links": ["https://rdap.db.ripe.net/entity/TP2569-RIPE", "http://
               www.ripe.net/data-tools/support/documentation/terms"],
            "events": null,
            "raw": null,
            "roles": ["technical"],
            "contact": {
                "name": "Thomas PETIT",
                "kind": "individual",
                "address": [{"type": null, "value": "Centre Inter-etablissement
                    pour les Services Reseaux\nUniversite Claude Bernard Lyon 1
                   \n21 Avenue Claude Bernard\n69622 Villeurbanne CEDEX, France
                   "}],
                "phone": [{"type": "voice", "value": "+33 4 72 43 18 99"}, {"
                   type": "fax", "value": "+33 4 72 44 84 10"}],
                "email": [{"type": "email", "value": "thomas.petit@univ-lyon1.
                   fr"}],
                "role": null, "title": null},
                "events_actor": null,
                "entities": null}},
        "raw": null}
```

Annexe B

Résultat entier obtenu via le protocole RDAP pour l'adresse IP de Wikidata.org

Résultat au 06/11/2023


```
{"nir": null,
"asn_registry": "arin",
"asn": "14907",
"asn_cidr": "208.80.154.0/23",
"asn_country_code": "US",
"asn_date": "2007-07-23",
"asn_description": "WIKIMEDIA, US",
"query": "208.80.154.224",
"network": {
    "handle": "NET-208-80-152-0-1",
    "status": ["active"],
    "remarks": [{
        "title": "Registration Comments",
        "description": "http://www.wikimediafoundation.org",
        "links": null}],
    "notices": [{
        "title": "Terms of Service",
        "description": "By using the ARIN RDAP/Whois service, you are
           agreeing to the RDAP/Whois Terms of Use",
        "links": ["https://www.arin.net/resources/registry/whois/tou/"]}, {
        "title": "Whois Inaccuracy Reporting",
        "description": "If you see inaccuracies in the results, please
           visit: ", "links":["https://www.arin.net/resources/registry/
           whois/inaccuracy_reporting/"]},{
        "title": "Copyright Notice",
        "description": "Copyright 1997-2023, American Registry for Internet
            Numbers, Ltd.", "links": null}],
        "links": ["https://rdap.arin.net/registry/ip/208.80.152.0", "https:
           //whois.arin.net/rest/net/NET-208-80-152-0-1"],
        "events": [{
            "action": "last changed",
            "timestamp": "2021-12-14T20:28:45-05:00",
```

```
                "actor": null}, {
                "action": "registration",
                "timestamp": "2007-07-23T17:32:30-04:00",
                "actor": null}],
            "raw": null,
            "start_address": "208.80.152.0",
            "end_address": "208.80.155.255",
            "cidr": "208.80.152.0/22",
            "ip_version": "v4",
            "type": "DIRECT ALLOCATION",
            "name": "WIKIMEDIA",
            "country": null,
            "parent_handle": "NET-208-0-0-0-0"},
    "entities": ["WIKIM"],
    "objects": {
        "WIKIM": {
            "handle": "WIKIM",
            "status": null,
            "remarks": null,
            "notices": null,
            "links": ["https://rdap.arin.net/registry/entity/WIKIM", "https://
               whois.arin.net/rest/org/WIKIM"],
            "events": [{
                "action": "last changed",
                "timestamp": "2022-11-07T04:52:46-05:00",
                "actor": null}, {
                "action": "registration",
                "timestamp": "2006-05-30T14:12:34-04:00",
                "actor": null}],
            "raw": null,
            "roles": ["registrant"],
            "contact": {"name": "Wikimedia Foundation Inc.", "kind": "org", "
               address": [{"type": null, "value": "1 Montgomery Street\nSuite 1
               600\nSan Francisco\nCA\n94104\nUnited States"}],
            "phone": null,
            "email": null,
            "role": null,
            "title": null},
            "events_actor": null,
            "entities": ["WIKIM-ARIN", "YOUNS-ARIN", "MBE96-ARIN", "WNA11-ARIN"
               , "MOONE85-ARIN"]}},
    "raw": null}
```

Annexe C

Liste de candidats étiquetés obtenus avec IPSeen Fast

label	industry Label	aka	type	Site web
Claude Bernard University Lyon 1	higher education	Université Lyon 1, UCBL, univ-lyon1.fr, Université Claude Bernard Lyon 1	university in France	https://www.univ-lyon1.fr/
Lyon research team in Information and Communication Sciences			facility	http://www.elico-recherche.eu/
Fédération Informatique de Lyon			facility	http://fil.cnrs.fr/
Management et économie Lyon Saint Etienne			facility, organization,	https://maelyse.universite-lyon.fr/
Field Observatory in Urban Hydrology			facility	http://www.graie.org/othu/index.htm
Automation and Process Engineering Laboratory			Joint Research Unit	http://www.lagep.cpe.fr/wwwlagep7/?lang=en
Faculté des Sciences de Lyon			faculty	
Centre for Astronomical Research of Lyon			Joint Research Unit	http://cral.univ-lyon1.fr/?lang=en, https://cral.univ-lyon1.fr/?lang=fr
Center for Research in Image Acquisition and Treatment for Health			Joint Research Unit, laboratory,	http://www.creatis.insa-lyon.fr

Institut de biologie marine Michel Pacha			Centre for maritime biology	
ICBMS			Joint Research Unit, institute,	http://www.icbms.fr
Institute of Physics of 2 Infinites Lyon			Joint Research Unit	http://www.ip2i.in2p3.fr/, https://www.ip2i.in2p3.fr/?lang=en,
Institute of Researches on Catalysis and Environment in Lyon			Joint Research Unit	http://www.ircelyon.univ-lyon1.fr/en/, https://www.ircelyon.univ-lyon1.fr/,
Lyon Institute of Nanotechnology			Joint Research Unit, laboratory, research institute	https://inl.cnrs.fr/
LIRIS			Joint Research Unit, research institute,	https://liris.cnrs.fr/
Biometry and Evolutionary Biology Laboratory			Joint Research Unit, laboratory,	http://lbbe.univ-lyon1.fr/
Laboratoire de Géologie de Lyon : Terre, Planètes et Environnement			Joint Research Unit, laboratory,	http://lgltpe.ens-lyon.fr/accueil?set_language=en&cl=en, http://lgltpe.ens-lyon.fr/accueil?set_language=fr&cl=fr,
LMI			Joint Research Unit, research institute,	http://lmi.cnrs.fr
Microbiology, Adaptation and Pathogenesis Lab			Joint Research Unit	https://map.insa-lyon.fr/en/, https://map.insa-lyon.fr/fr/,
Microbial Ecology			Joint Research Unit, laboratory,	http://www.ecologiemicrobiennelyon.fr/
Institut Lumière Matière			Joint Research Unit, research institute,	http://ilm.univ-lyon1.fr/, https://ilm.univ-lyon1.fr/?&lang=2,
Institute of Analytical Sciences			research institute	http://isa-lyon.fr/
Lyon Neuroscience Research Center			Joint Research Unit	https://crnl.univ-lyon1.fr
Stem-Cell and Brain Research Institute			facility	http://www.rhone-alpes-auvergne.inserm.fr/rubriques/l-inserm-en-region/laboratoires/lyon/annexes/u846, http://www.sbri.fr/,

Institut de Génomique Fonctionnelle de Lyon			Joint Research Unit	http://igfl.ens-lyon.fr/
Cancer Center of Lyon			Joint Research Unit	http://www.crcl.fr/
International Center for Infectiology Research			Joint Research Unit	http://ciri.inserm.fr/en/
Institut de Biologie et de Chimie des Protéines			facility	http://www.ibcp.fr/?lang=en
Laboratoire de Physique de l'ENS de Lyon			Joint Research Unit	http://www.ens-lyon.fr/PHYSIQUE
Matériaux Ingénierie et Science			Joint Research Unit	http://mateis.insa-lyon.fr/
Laboratoire de Reproduction et Développement des Plantes			Joint Research Unit	http://www.ens-lyon.fr/RDP/?lang=fr
Groupe d'Analyse et de Théorie Economique Lyon St Etienne			Joint Research Unit	https://www.gate.cnrs.fr/?lang=en
Institut NeuroMyoGène			facility	http://inmg.fr/
Centre for Cognitive Neuroscience			Joint Research Unit	http://cnc.isc.cnrs.fr/
Laboratoire de Biologie Moléculaire de la Cellule			Joint Research Unit	http://www.ens-lyon.fr/LBMC/spip/
Laboratoire de l'Informatique du Parallélisme			Joint Research Unit	http://www.ens-lyon.fr/LIP/web-n/
Laboratory for Language, Brain and Cognition			Joint Research Unit	http://l2c2.isc.cnrs.fr/drupal7/en/
Archéologie et Archéométrie			Joint Research Unit	http://www.arar.mom.fr/, https://lefildarar.hypotheses.org,
Applications des Ultrasons à la Thérapie			facility, laboratory,	http://labtau.univ-lyon1.fr/, http://www.rhone-alpes-auvergne.inserm.fr/rubriques/l-inserm-en-region/laboratoires/lyon/annexes/u1032, http://www.univ-lyon1.fr/recherche/entites-de-recherche/laboratoire-therapie-et-applications-ultrasonores-labtau--618088.kjsp

Structural and Molecular Basis of Infectious Systems			Joint Research Unit	http://www.ibcp.fr/mmsb/?lang=en
Laboratoire de Chimie			Joint Research Unit	http://www.ens-lyon.fr/CHIMIE
Bioingénierie et Dynamique Microbienne aux Interfaces Alimentaires			facility, laboratory,	http://biodymia.univ-lyon1.fr/
Laboratory of Polymer Materials Engineering			Joint Research Unit	http://www.imp.cnrs.fr/
Biologie Tissulaire et Ingénierie Thérapeutique			Joint Research Unit	https://www.ibcp.fr/lbti/?lang=fr
Laboratory of Chemistry, Catalysis, Polymers and Process			Joint Research Unit	http://c2p2-cpe.com/index.php
Laboratoire de Mécanique des Fluides et d'Acoustique			Joint Research Unit	http://lmfa.ec-lyon.fr/
Biologie Fonctionnelle Insectes et Interactions			facility	http://bf2i.insa-lyon.fr/
Virology and Human Pathology			Q43371093	http://www.virpath.com/
Institute for Cognitive Science			Joint Research Unit	http://www.isc.cnrs.fr
Decision & Information Sciences for Production Systems			facility	http://disp-lab.fr/en/node/22
Renater	research and development in other physical and natural sciences		network service provider, national research and education network,	https://www.renater.fr
La Doua			campus, neighborhood,	https://www.univ-lyon1.fr/campus/plan-des-campus/campus-lyontech-la-doua-731724.kjsp

Pathophysiology , Diagnosis and Treatments of Bone Diseases			laboratory	http://www.univ-lyon1.fr/recherche/entites-de-recherche/physiopathologie-diagnostic-et-traitements-des-maladies-osseuses-618131.kjsp

TABLE C.1 – Résultats pour **IPSeen Fast** post requêtes à Wikidata

Annexe D

Description de l'ensemble des labels d'attaques

nom	categorie	souscat	extrait de documentation des sources	source
apfell	tool	C2	Mythic is a multiplayer, command and control platform for red teaming operations. It is designed to facilitate a plug-n-play architecture where new agents, communication channels, and modifications can happen on the fly.	https://docs.mythic-c2.net/
factionC2	tool	C2	C2 framework	https://www.foregenix.com/blog/a-first-look-at-todays-command-and-control-frameworks
plugX	malware	RAT	PlugX is a remote access tool (RAT) with modular plugins that has been used by multiple threat groups.	https://attack.mitre.org/software/S0013/
apt28-xagent	group	APT	APT28 is a threat group that has been attributed to Russia's General Staff Main Intelligence Directorate (GRU). This group has been active since at least 2004. APT28 reportedly compromised the Hillary Clinton campaign, the Democratic National Committee, and the Democratic Congressional Campaign Committee in 2016 in an attempt to interfere with the U.S. presidential election. In 2018, the US indicted five GRU Unit 26165 officers associated with APT28 for cyber operations (including close-access operations) conducted between 2014 and 2018 against the World Anti-Doping Agency (WADA), the US Anti-Doping Agency, a US nuclear facility, the Organization for the Prohibition of Chemical Weapons (OPCW), the Spiez Swiss Chemicals Laboratory, and other organizations.	https://attack.mitre.org/groups/G0007/

armitage	tool	bot	Armitage est un logiciel d'exploit inclus dans Kali Linux et Parrot OS servant à rechercher les failles de sécurités dans un système informatique. Il permet de transformer l'ordinateur victime en PC zombie.	https://fr.wikipedia.org/wiki/Armitage_(logiciel)
cobalt strike	tool	redteam	Cobalt Strike is a commercial, full-featured, remote access tool that bills itself as "adversary simulation software designed to execute targeted attacks and emulate the post-exploitation actions of advanced threat actors". Cobalt Strike's interactive post-exploit capabilities cover the full range of ATT&CK tactics, all executed within a single, integrated system.	https://attack.mitre.org/software/S0154/
beef	tool	web pentest	BeEF is short for The Browser Exploitation Framework. It is a penetration testing tool that focuses on the web browser.	https://github.com/beefproject/beef
empirev1	tool	remote administration	Empire is an open source, cross-platform remote administration and post-exploitation framework that is publicly available on GitHub. Empire was one of five tools singled out by a joint report on public hacking tools being widely used by adversaries.	https://attack.mitre.org/software/S0363/
empirev2	tool	remote administration	Empire is an open source, cross-platform remote administration and post-exploitation framework that is publicly available on GitHub. Empire was one of five tools singled out by a joint report on public hacking tools being widely used by adversaries.	https://attack.mitre.org/software/S0363/
pupy	tool	remote administration	Pupy is an open source, cross-platform (Windows, Linux, OSX, Android) remote administration and post-exploitation tool. Pupy is publicly available on GitHub.	https://attack.mitre.org/software/S0192/
metasploit	tool	pentest	Metasploit, Metasploit Pen Testing Tool, est un projet (open source, sous Licence BSD modifiée2) en relation avec la sécurité des systèmes informatiques. Son but est de fournir des informations sur les vulnérabilités de systèmes informatiques, d'aider à la pénétration et au développement de signatures pour les systèmes de détection d'intrusion (IDS, Intrusion Detection System).	https://fr.wikipedia.org/wiki/Metasploit
covenant	tool	C2	Covenant is a .NET command and control framework that aims to highlight the attack surface of .NET, make the use of offensive .NET tradecraft easier, and serve as a collaborative command and control platform for red teamers.	https://www.foregenix.com/blog/a-first-look-at-todays-command-and-control-frameworks
gophish	tool	phishing tool	Gophish is a powerful, open-source phishing framework that makes it easy to test your organization's exposure to phishing.	https://getgophish.com/

koadic	tool	pentest	Koadic is a Windows post-exploitation framework and penetration testing tool that is publicly available on GitHub. Koadic has several options for staging payloads and creating implants, and performs most of its operations using Windows Script Host.	https://attack.mitre.org/software/S0250/
merlin	tool	C2	Merlin is a cross-platform post-exploitation Command & Control server and agent written in Go.	https://www.foregenix.com/blog/a-first-look-at-todays-command-and-control-frameworks
meterpreter	tool	rat tool	Meterpreter (l'interpréteur de Metasploit) permet aux utilisateurs de contrôler l'écran d'un appareil à l'aide de VNC et de parcourir, charger et télécharger des fichiers.	https://fr.wikipedia.org/wiki/Metasploit
responder	tool	net_sniff	Responder is an open source tool used for LLMNR, NBT-NS and MDNS poisoning, with built-in HTTP/SMB/MSSQL/FTP/LDAP rogue authentication server supporting NTLMv1/NTLMv2/LMv2, Extended Security NTLMSSP and Basic HTTP authentication.	https://attack.mitre.org/software/S0174/
octopus	malware	trojan	Octopus is a Windows Trojan written in the Delphi programming language that has been used by Nomadic Octopus to target government organizations in Central Asia since at least 2014.	https://attack.mitre.org/software/S0340/
sliver	tool	C2	Sliver is an open source, cross-platform, red team command and control framework written in Golang.	https://attack.mitre.org/software/S0633/
evilosx	malware	RAT	EvilOSX is a pure python, post-exploitation, RAT (Remote Administration Tool) for macOS / OSX.	https://medium.com/@lucideus/evilosx-a-remote-administration-tool-rat-for-macos-os-x-lucideus-research-da0551ed3969

panda	malware			https://i.blackhat.com/asia-19/Fri-March-29/bh-asia-Jang-When-Voice-Phishing-Met-Malicious-Android-App-updated.pdf
predator The Thief	malware	spyware	Predator is a feature-rich information stealer. It is sold on hacking forums as a bundle which includes : Payload builder and Command and Control web panel. It is able to grab passwords from browsers, replace cryptocurrency wallets, and take photos from the web-camera.	https://malpedia.caad.fkie.fraunhofer.de/details/win.predator
nexus	malware	trojan	Android Trojan	https://www.liansecurity.com/#/main/news/RWt_ZocBrFZDfCElFqw_/detail
turla	group	other	Turla is a cyber espionage threat group that has been attributed to Russia's Federal Security Service (FSB). They have compromised victims in over 50 countries since at least 2004, spanning a range of industries including government, embassies, military, education, research and pharmaceutical companies. Turla is known for conducting watering hole and spearphishing campaigns, and leveraging in-house tools and malware, such as Uroburos.	https://attack.mitre.org/groups/G0010/
phishery	tool	phishing tool	Phishery is a Simple SSL Enabled HTTP server with the primary purpose of phishing credentials via Basic Authentication. Phishery also provides the ability easily to inject the URL into a .docx Word document.	https://github.com/ryhanson/phishery
shadowpad	malware	backdoor	ShadowPad is a modular backdoor that was first identified in a supply chain compromise of the NetSarang software in mid-July 2017. The malware was originally thought to be exclusively used by APT41, but has since been observed to be used by various Chinese threat activity groups.	https://attack.mitre.org/software/S0596/
sidewinder	group	APT	XAgentOSX is a trojan that has been used by APT28 on OS X and appears to be a port of their standard CHOPSTICK or XAgent trojan.	https://attack.mitre.org/software/S0161/
silenttrinity	tool	remote administration	Sidewinder is a suspected Indian threat actor group that has been active since at least 2012. They have been observed targeting government, military, and business entities throughout Asia, primarily focusing on Pakistan, China, Nepal, and Afghanistan	https://attack.mitre.org/groups/G0121/

thalium	group	APT	SILENTTRINITY is an open source remote administration and post-exploitation framework. SILENTTRINITY was used in a 2019 campaign against Croatian government agencies by unidentified cyber actors.	https://attack.mitre.org/software/S0692/
throwback	tool		HTTP/S Beaconing Implant	https://github.com/silentbreaksec/Throwback
APT35	group	APT	Magic Hound is an Iranian-sponsored threat group that conducts long term, resource-intensive cyber espionage operations, likely on behalf of the Islamic Revolutionary Guard Corps. They have targeted European, U.S., and Middle Eastern government and military personnel, academics, journalists, and organizations such as the World Health Organization (WHO), via complex social engineering campaigns since at least 2014.	https://attack.mitre.org/groups/G0059/
CostaRicto	group	other	CostaRicto was a suspected hacker-for-hire cyber espionage campaign that targeted multiple industries worldwide, with a large number being financial institutions. CostaRicto actors targeted organizations in Europe, the Americas, Asia, Australia, and Africa, with a large concentration in South Asia (especially India, Bangladesh, and Singapore), using custom malware, open source tools, and a complex network of proxies and SSH tunnels.	https://attack.mitre.org/campaigns/C0004/
cerium	group	other	Ruby Sleet is a threat actor linked to North Korea's Ministry of State Security. Cerium has been involved in spearphishing campaigns, compromising devices, and conducting cyberattacks alongside other North Korean threat actors. They have also targeted companies involved in COVID-19 research and vaccine development.	https://malpedia.caad.fkie.fraunhofer.de/actor/ruby_sleet
APT29	group	APT	APT29 is threat group that has been attributed to Russia's Foreign Intelligence Service (SVR). They have operated since at least 2008, often targeting government networks in Europe and NATO member countries, research institutes, and think tanks. APT29 reportedly compromised the Democratic National Committee starting in the summer of 2015.	https://attack.mitre.org/groups/G0016/

modlishka	tool	reverse proxy sock	A new reverse proxy tool called Modlishka can easily automate phishing attacks and bypass two-factor authentication (2FA) — and it's available for download on GitHub.	https://securityintelligence.com/news/new-reverse-proxy-tool-can-bypass-two-factor-authentication-and-automate-phishing-attacks/
tontoteam	group	other	Tonto Team is a suspected Chinese state-sponsored cyber espionage threat group that has primarily targeted South Korea, Japan, Taiwan, and the United States since at least 2009 ; by 2020 they expanded operations to include other Asian as well as Eastern European countries. Tonto Team has targeted government, military, energy, mining, financial, education, healthcare, and technology organizations, including through the Heartbeat Campaign (2009-2012) and Operation Bitter Biscuit (2017).	https://attack.mitre.org/groups/G0131/
empirev3	tool	remote administration	Empire is an open source, cross-platform remote administration and post-exploitation framework that is publicly available on GitHub. Empire was one of five tools singled out by a joint report on public hacking tools being widely used by adversaries.	https://attack.mitre.org/software/S0363/
wizardspider	group	other	Wizard Spider is reportedly associated with Grim Spider and Lunar Spider. The WIZARD SPIDER threat group is the Russia-based operator of the TrickBot banking malware. The WIZARD SPIDER threat group, known as the Russia-based operator of the TrickBot banking malware, had focused primarily on wire fraud in the past.	https://malpedia.caad.fkie.fraunhofer.de/actor/wizard_spider
xhelper	malware	trojan	was the middle of last year that we detected the start of mass attacks by the xHelper Trojan on Android smartphones, but even now the malware remains as active as ever. The main feature of xHelper is entrenchment — once it gets into the phone, it somehow remains there even after the user deletes it and restores the factory settings.	https://securelist.com/unkillable-xhelper-and-a-trojan-matryoshka/96487/
felixRoot	malware	backdoor	FELIXROOT is a backdoor that has been used to target Ukrainian victims.	https://attack.mitre.org/software/S0267/

furball	malware	spyware	an operation dubbed "Domestic Kitten", which uses malicious Android applications to steal sensitive personal information from its victims : screenshots, messages, call logs, surrounding voice recordings, and more. This operation managed to remain under the radar for a long time, as the associated files were not attributed to a known malware family and were only detected by a handful of security vendors.	https://malpedia.caad.fkie.fraunhofer.de/details/apk.furball
UNC2452	group	APT	APT29 is threat group that has been attributed to Russia's Foreign Intelligence Service (SVR). They have operated since at least 2008, often targeting government networks in Europe and NATO member countries, research institutes, and think tanks. APT29 reportedly compromised the Democratic National Committee starting in the summer of 2015. In April 2021, the US and UK governments attributed the SolarWinds Compromise to the SVR; public statements included citations to APT29, Cozy Bear, and The Dukes.Industry reporting also referred to the actors involved in this campaign as UNC2452, NOBELIUM, StellarParticle, Dark Halo, and SolarStorm.	https://attack.mitre.org/groups/G0016/
muddyWater	group	other	MuddyWater is a cyber espionage group assessed to be a subordinate element within Iran's Ministry of Intelligence and Security (MOIS).Since at least 2017, MuddyWater has targeted a range of government and private organizations across sectors, including telecommunications, local government, defense, and oil and natural gas organizations, in the Middle East, Asia, Africa, Europe, and North America.	https://attack.mitre.org/groups/G0069/
tagBarnakle	malware	exploit	we will disclose the details behind one such ongoing malvertising campaign that is perpetrated by an attacker via mass compromise of Revive Adserver instances. They then append their malicious payload to existing ad slots, all of which results in free access to publisher inventory. We have named the attacker Tag Barnakle.	https://blog.confiant.com/tag-barnakle-one-year-later-120-more-revive-adserver-hacks-f3e5b3bc8e70
fruityC2	tool	post_exp	FruityC2 is a post-exploitation (and open source) framework based on the deployment of agents on compromised machines. Agents are managed from a web interface under the control of an operator. It works as a command-and-control model and is language and system agnostic.	https://github.com/xtr4nge/FruityC2
shad0w	tool	C2	SHAD0W is a modular C2 framework designed to successfully operate on mature environments.	https://github.com/bats3c/shad0w

phantom Lance	malware	backdoor	backdoor trojan in Google Play,	https://securelist.com/apt-phantomlance/96772/
hydra	malware	banking	Hydra is an Android BankBot variant, a type of malware designed to steal banking credentials. The way it does this is by requesting the user enables dangerous permissions such as accessibility and every time the banking app is opened, the malware is hijacking the user by overwriting the legit banking application login page with a malicious one. The goal is the same, to trick the user to enter his login credentials so that it will go straight to the malware authors.	https://malpedia.caad.fkie.fraunhofer.de/details/apk.hydra
SVCRAT	malware			
cryptocore	group	crypto actor	CryptoCore is an attack campaign against crypto-exchange companies that has been ongoing for three years and was discovered by ClearSky researchers. This cybercrime campaign is focused mainly on the theft of cryptocurrency wallets.	https://otx.alienvault.com/pulse/5ef36f8f63a7d8a11972ca54
bazar backdoor	malware	backdoor	A Domain Generation Algorithm (DGA) is a technique used by malware authors to generate new domain names for malware command and control. Typically malware will contain a configuration which will house any number of things, including the Command and Control (C2) domains/IPs. While these configurations are typically encrypted within the binary, malware analysts and reverse engineers can often extract these C2s through sandboxes or configuration extractors. This makes it fairly easy, if not trivial, to extract these C2s and put in network blocks. To combat this, malware authors use DGAs to generate domains over time, allowing for a sometimes infinite stream of C2s. This allows for increased persistence if C2 infrastructure is taken down and makes it more difficult to block network traffic.	https://malpedia.caad.fkie.fraunhofer.de/details/win.bazarbackdoor
ninja	virus		Ce virus infecte les feuilles de calcul Excel (fichiers XLS).	https://threats.kaspersky.com/fr/threat/Virus.MSExcel.Ninja/
tunna	malware	backdoor	webshell :Un code encoquillé (en Anglais web shell) est une interface de type shell Web malveillante qui permet un accès et un contrôle à distance à un serveur Web en permettant l'exécution de commandes arbitraires.	https://malpedia.caad.fkie.fraunhofer.de/details/asp.tunna

dridex	malware	banking	Dridex is a prolific banking Trojan that first appeared in 2014. By December 2019, the US Treasury estimated Dridex had infected computers in hundreds of banks and financial institutions in over 40 countries, leading to more than $100 million in theft. Dridex was created from the source code of the Bugat banking Trojan (also known as Cridex)	https://attack.mitre.org/software/S0384/
APT27	group	APT	Threat Group-3390 is a Chinese threat group that has extensively used strategic Web compromises to target victims. The group has been active since at least 2010 and has targeted organizations in the aerospace, government, defense, technology, energy, manufacturing and gambling/betting sectors.	https://attack.mitre.org/groups/G0027/
SharkBot	malware	banking	SharkBot is a banking malware, first discovered in October 2021, that tries to initiate money transfers directly from compromised devices by abusing Accessibility Services.	https://attack.mitre.org/software/S1055/
brata	malware	RAT	“BRATA” is a new Android remote access tool malware family. We used this code name based on its description – “Brazilian RAT Android”. It exclusively targets victims in Brazil : however, theoretically it could also be used to attack any other Android user if the cybercriminals behind it want to. It has been widespread since January 2019, primarily hosted in the Google Play store, but also found in alternative unofficial Android app stores.	https://malpedia.caad.fkie.fraunhofer.de/details/apk.brata
empire	tool	remote administration	Empire is an open source, cross-platform remote administration and post-exploitation framework that is publicly available on GitHub. Empire was one of five tools singled out by a joint report on public hacking tools being widely used by adversaries.	https://attack.mitre.org/software/S0363/
bazarloader	malware	backdoor	avec bazarbackdoor	https://malpedia.caad.fkie.fraunhofer.de/details/win.bazarbackdoor

Badrat	malware	RAT	This threat can give a malicious hacker unauthorized access and control of your PC.	https://www.microsoft.com/en-us/wdsi/threats/malware-encyclopedia-description?Name=Backdoor:Win32/Badrat&threatId=-2147389654
cs2 modrewrite	tool	C2	This project converts a Cobalt Strike profile to a functional mod_rewrite to support HTTP reverse proxy redirection to a Cobalt Strike teamserver. The use of reverse proxies provides protection to backend C2 servers from profiling, investigation, and general internet background radiation.	https://github.com/threatexpress/cs2modrewrite
interactsh	tool	interaction tool	Interactsh is an open-source tool for detecting out-of-band interactions. It is a tool designed to detect vulnerabilities that cause external interactions.	https://github.com/projectdiscovery/interactsh
BitRat	malware	RAT	BitRAT is a notorious remote access trojan (RAT) marketed on underground cybercriminal web markets and forums. Its price tag of $20 for lifetime access makes it irresistible to cybercriminals and helps the malicious payload spread. Furthermore, each buyer's modus operandi makes BitRAT even harder to stop, considering it can be employed in various operations, such as trojanized software, phishing and watering hole attacks. The malicious tool can perform a wide range of operations, including data exfiltration, UAC bypass, DDoS attacks, clipboard monitoring, gaining unauthorized webcam access, credential theft, audio recording, XMRig coin mining and generic keylogging.	https://malpedia.caad.fkie.fraunhofer.de/details/win.bit_rat
Caldera	tool	redteam	CALDERA™ offers an intelligent, automated red team system that can reduce resources needed by security teams for routine testing, freeing them to address other critical problems. CALDERA can be used to test endpoint security solutions and assess a network's security posture against the common post-compromise adversarial techniques contained in the MITRE ATT&CK® model.	https://www.mitre.org/ourimpact/intellectual-property/caldera
BruteRatel	tool	C2	Brute Ratel is a a Customized Command and Control Center for Red Team and Adversary SimulationSMB and TCP payloads provide functionality to write custom external C2 channels over legitimate websites such as Slack, Discord, Microsoft Teams and more.	https://malpedia.caad.fkie.fraunhofer.de/details/win.brute_ratel_c4

PickleC2	tool	post_exp	PickleC2 is a post-exploitation and lateral movements framework	https://github.com/xRET2pwn/PickleC2
chaes	malware	spyware	Chaes is a multistage information stealer written in several programming languages that collects login credentials, credit card numbers, and other financial information. Chaes was first observed in 2020, and appears to primarily target victims in Brazil as well as other e-commerce customers in Latin America.	https://attack.mitre.org/software/S0631/
DcRat	malware	RAT	DCRat is a typical RAT that has been around since at least June 2019.	https://malpedia.caad.fkie.fraunhofer.de/details/win.dcrat
TrevorC2	tool	C2	TrevorC2 is a client/server model for masking command and control through a normally browsable website. Detection becomes much harder as time intervals are different and does not use POST requests for data exfil.	https://github.com/trustedsec/trevorc2?source=post_page-----2a9ce6f1f425----------
BUMBLEBEE	malware	downlader	Bumblebee is a custom loader written in C++ that has been used by multiple threat actors, including possible initial access brokers, to download and execute additional payloads since at least March 2022.	https://attack.mitre.org/software/S1039/
FinFisher	malware	spyware	FinFisher is a government-grade commercial surveillance spyware reportedly sold exclusively to government agencies for use in targeted and lawful criminal investigations. It is heavily obfuscated and uses multiple anti-analysis techniques.	https://attack.mitre.org/software/S0182/
nanocore	malware	RAT	NanoCore is a modular remote access tool developed in .NET that can be used to spy on victims and steal information. It has been used by threat actors since 2013.	https://attack.mitre.org/software/S0336/
NorthStar	tool	C2	An open source C2 framework intended for pentest and red teaming activities.	https://malpedia.caad.fkie.fraunhofer.de/details/win.northstar
satellite	tool	spyware	Satellite is an web payload hosting service which filters requests to ensure the correct target is getting a payload. This can also be a useful service for hosting files that should be only accessed in very specific circumstances.	https://github.com/t94j0/satellite

Roaming Mantis	group	other	Roaming Mantis Group have continued to evolve since they were first discovered in April 2018. As part of their activities, this group hacks into exploitable routers and changes their DNS configuration. This allows the attackers to redirect the router user's traffic to malicious Android apps disguised as Facebook and Chrome or to Apple phishing pages that were used to steal Apple ID credentials. Recently, Kaspersky has discovered that this group is testing a new monetization scheme by redirecting iOS users to pages that contain the Coinhive in-browser mining script rather than the normal Apple phishing page. When users are redirected to these pages, they will be shown a blank page in the browser, but their CPU utilization will jump to 90% or higher.	https://malpedia.caad.fkie.fraunhofer.de/actor/roaming_mantis
Cytrox	malware	spyware	Predator est un logiciel espion développé par Cytrox qui permet de surveiller et de suivre les appareils de ses cibles, en collectant leurs données de manière discrète. Plus précisément, il fournit à ses utilisateurs et clients des capacités de : 1. Surveillance à distance : le logiciel espion Predator permet la surveillance à distance des appareils ciblés, quel que soit leur emplacement. Ses utilisateurs et clients peuvent accéder aux données collectées via un portail, assurant ainsi des capacités de surveillance en temps réel. 2. Collecte de données : le logiciel recueille des données détaillées à partir d'appareils ciblés, y compris les journaux d'appels, les messages, les fichiers multimédias, les informations de localisation, l'historique de navigation, etc. Cette collecte complète de données permet une compréhension complète des activités numériques de la cible. 3. Mode furtif : le logiciel espion Predator fonctionne en mode furtif, garantissant une présence quasi-indétectable sur l'appareil cible. 4. Compatibilité : le logiciel espion Predator est compatible avec les appareils iOS et Android, ce qui le rend très polyvalent pour cibler un large éventail d'appareils. 5. Personnalisation et contrôle : les utilisateurs ont la possibilité d'adapter les paramètres de surveillance et de contrôler les données qu'ils souhaitent collecter et surveiller.	https://www.sekoia.io/fr/glossaire/logiciel-espion-predator/

revsocks	tool	reverse proxy sock	Reverse socks5 tunneler with SSL/TLS and proxy support	https://www.trendmicro.com/vinfo/ie/threat-encyclopedia/malware/hacktool.linux.revsocks.aa
Tomiris	malware	backdoor	Tomiris is a backdoor written in Go that continuously queries its C2 server for executables to download and execute on a victim system. It was first reported in September 2021 during an investigation of a successful DNS hijacking campaign against a Commonwealth of Independent States (CIS) member.	https://attack.mitre.org/software/S0671/
Phishmonger	tool	phishing tool	Phishing platform designed for pentesters. This tool allows us to craft phishing emails in Outlook, clone them quickly, automatically template them for mass distribution, test email templates, schedule phishing campaigns, and track phishing results.	https://github.com/fkasler/phishmonger
Orcus	malware	RAT	Orcus has been advertised as a Remote Administration Tool (RAT) since early 2016.	https://malpedia.caad.fkie.fraunhofer.de/details/win.orcus_rat
LemonDuck	malware	crypto	Lemon Duck is a monerocrypto-mining malware with capabilitiy to spread rapidly across the entire network. The malware runs its payload mainly in memory. Internal network spreading is performed by SMB RCE Vulnerability (CVE-2017-0144), or brute-force attacks.	https://malpedia.caad.fkie.fraunhofer.de/details/win.lemonduck
TAG38	group	other	Threat Activity Group 38 (TAG-38) is a hacking group which used ShadowPad Command and Control (C2) and compromised IOT devices (devices used to launch the intrusions were based in South Korea and Taiwan) to hack Ladakh's Power-Grid, which affected Industrial Control System (ICS) which is used for electricity routing and load balancing.	https://arpitbaj21.medium.com/a-full-breakdown-of-ladakhs-power-grid-hack-1f873596bbe5
TEMP HERETIC	group	other	TEMP_Heretic is a threat actor that has been observed engaging in targeted spear-phishing campaigns. They exploit vulnerabilities in email platforms, such as Zimbra, to exfiltrate emails from government, military, and media organizations.	https://malpedia.caad.fkie.fraunhofer.de/actor/temp_heretic

vulturi	malware	spyware	Vulturi is a straightforward stealer family that appeared in early 2021 and was reported by our friends over at SEKOIA.IO earlier this year. There is currently very limited public analysis of the family available but it seems to be a fairly typical infostealer capable of extracting data from a comprehensive list of software. It is sold through forums for anyone to use (or at least it was until the cracked version became available in June 2021).	https://hatching.io/blog/tt-2022-07-21/
Vajra	malware	RAT	android rat	https://malpedia.caad.fkie.fraunhofer.de/details/apk.vajraspy
trickbot	malware	spyware	TrickBot is a Trojan spyware program written in C++ that first emerged in September 2016 as a possible successor to Dyre. TrickBot was developed and initially used by Wizard Spider for targeting banking sites in North America, Australia, and throughout Europe; it has since been used against all sectors worldwide as part of "big game hunting" ransomware campaigns.	https://attack.mitre.org/software/S0266/
XLoader	malware	spyware	XLoader for iOS is a malicious iOS application that is capable of gathering system information.It is tracked separately from the XLoader for Android.	https://attack.mitre.org/software/S0490/
SolarMaker	malware	spyware	SolarMarker, a malware family known for its infostealing and backdoor capabilities, mainly delivered through search engine optimization (SEO) manipulation to convince users to download malicious documents. Some of SolarMarker's capabilities include the exfiltration of auto-fill data, saved passwords and saved credit card information from victims' web browsers. Besides capabilities typical for infostealers, SolarMarker has additional capabilities such as file transfer and execution of commands received from a C2 server. The malware invests significant effort into defense evasion, which consists of techniques like signed files, huge files, impersonation of legitimate software installations and obfuscated PowerShell scripts.	https://malpedia.caad.fkie.fraunhofer.de/details/win.solarmarker
PowGoop	malware	downlader	PowGoop is a loader that consists of a DLL loader and a PowerShell-based downloader; it has been used by MuddyWater as their main loader.	https://attack.mitre.org/software/S1046/
GrandaMisha	malware			

FFDroider	malware	spyware	FDroider is a malicious program classified as a stealer. It is designed to extract and exfiltrate sensitive data from infected devices. FFDroider targets popular social media and e-commerce platforms in particular.	https://malpedia.caad.fkie.fraunhofer.de/details/win.ffdroider
HTTP-Revshell	tool	reverse proxy sock	HTTP-revshell is a tool focused on redteam exercises and pentesters. This tool provides a reverse connection through the http/s protocol. It uses a covert channel to gain control over the victim machine through web requests and thus evade solutions such as IDS, IPS and AV.	https://github.com/3v4Si0N/HTTP-revshell
Ghostwriter Tool	tool	activity set	Ghostwriter is referred as an 'activity set', with various incidents tied together by overlapping behavioral characteristics and personas, rather than as an actor or group in itself.	https://malpedia.caad.fkie.fraunhofer.de/actor/ghostwriter
Deathstalker	group	APT	the APT group behind the Evilnum malware previously seen in attacks against financial technology companies. While said malware has been seen in the wild since at least 2018 and documented previously, little has been published about the group behind it and how it operates. The group's targets remain fintech companies, but its toolset and infrastructure have evolved and now consist of a mix of custom, homemade malware combined with tools purchased from Golden Chickens, a Malware-as-a-Service (MaaS) provider whose infamous customers include FIN6 and Cobalt Group.	https://malpedia.caad.fkie.fraunhofer.de/actor/evilnum
CreepySnail	malware	spyware	CreepySnail is a custom PowerShell implant that has been used by POLONIUM since at least 2022	https://attack.mitre.org/software/S1024/
NetDooka	malware	RAT	A RAT written in .NET, delivered with a driver to protect it from deletion. Observed being dropped by PrivateLoader.	https://malpedia.caad.fkie.fraunhofer.de/details/win.netdooka

SeaFlower	malware	banking	As of today, the main current objective of SeaFlower is to modify web3 wallets with backdoor code that ultimately exfiltrates the seed phrase.	https://blog.confiant.com/how-seaflower-%E8%97%8F%E6%B5%B7%E8%8A%B1-installs-backdoors-in-ios-android-web3-wallets-to-steal-your-seed-phrase-d25f0ccdffce
Teardroid	malware		This threat can perform a number of actions of a malicious hacker's choice on your PC.	https://www.microsoft.com/en-us/wdsi/threats/malware-encyclopedia-description?Name=Trojan:AndroidOS/Teardroid.A&ThreatID=2147899821
Sandworm	group	other	Sandworm Team is a destructive threat group that has been attributed to Russia's General Staff Main Intelligence Directorate (GRU) Main Center for Special Technologies (GTsST) military unit 74455.[1][2] This group has been active since at least 2009.	https://attack.mitre.org/groups/G0034/
DeimosC2	tool	C2	Trend Micro describes DeimosC2 as an open-source C&C framework that was released in June 2020. It is a fully-functional framework that allows for multiple attackers to access, create payloads for, and interact with victim computers. As a post-exploitation C&C framework, DeimosC2 will generate the payloads that need to be manually executed on computer servers that have been compromised through other means such as social engineering, exploitation, or brute-force attacks. Once it is deployed, the threat actors will gain the same access to the systems as the user account that the payload was executed as, either as an administrator or a regular user. Note that DeimosC2 does not perform active or privilege escalation of any kind.	https://malpedia.caad.fkie.fraunhofer.de/details/win.deimos_c2

GoMet	malware	backdoor	The story of this backdoor is rather curious — there are two documented cases of its usage by sophisticated threat actors. First, in 2020, attackers were deploying this malware after the successful exploitation of CVE-2020-5902, a vulnerability in F5 BIG-IP so severe that USCYBERCOM posted a tweet urging all users to patch the application. The second is more recent and involved the successful exploitation of CVE-2022-1040, a remote code execution vulnerability in Sophos Firewall.	https://blog.talosintelligence.com/attackers-target-ukraine-using-gomet/
Alan Framework	tool	post_exp	post exploitation framework	https://reconshell.com/alanframework-a-post-exploitation-framework/
RedGuard	tool	C2	RedGuard, a derivative tool based on command and control (C2) front flow control technology, has a lighter design, efficient traffic interaction, and reliable compatibility with development in the go programming language.	https://github.com/wikiZ/RedGuard
Karkadann	group	other	Karkadann is a threat actor that has been active since at least October 2020, targeting government bodies and news outlets in the Middle East. They have been involved in watering hole attacks, compromising high-profile websites to inject malicious JavaScript code. The group has been linked to another commercial spyware company called Candiru, suggesting they may utilize multiple spyware technologies. There are similarities in the infrastructure and tactics used by Karkadann in their campaigns.	https://malpedia.caad.fkie.fraunhofer.de/actor/karkadann
Subzero	malware		MSTIC found the Subzero malware being deployed through a variety of methods, including 0-day exploits in Windows and Adobe Reader, in 2021 and 2022. As part of our investigation into the utility of this malware, Microsoft's communications with a Subzero victim revealed that they had not commissioned any red teaming or penetration testing, and confirmed that it was unauthorized, malicious activity.	https://malpedia.caad.fkie.fraunhofer.de/details/win.subzero
Rafel	malware	RAT	RAT ;MALWARE	
ArrowRat	malware	RAT	It is available as a service, purchasable by anyone to use in their own campaigns. It's features are generally fairly typical of a RAT, with its most notable aspect being the hVNC module which basically gives an attacker full remote access with minimal need for technical knowledge to use it.	https://malpedia.caad.fkie.fraunhofer.de/details/win.arrowrat

Grandoreiro	malware	banking	Grandoreiro is a banking trojan written in Delphi that was first observed in 2016 and uses a Malware-as-a-Service (MaaS) business model. Grandoreiro has confirmed victims in Brazil, Mexico, Portugal, and Spain.	https://attack.mitre.org/software/S0531/
Ramnit	malware	banking	According to Check Point, Ramnit is primarily a banking trojan, meaning that its purpose is to steal login credentials for online banking, which cybercriminals can sell or use in future attacks. For this reason, Ramnit primarily targets individuals rather than focusing on particular industries. Ramnit campaigns have been observed to target organizations in particular industries. For example, a 2019 campaign targeted financial organizations in the United Kingdom, Italy, and Canada.	https://malpedia.caad.fkie.fraunhofer.de/details/win.ramnit
XploitSPY	malware	spyware	In several cases, this group used a modified version of commodity Android malware known as XploitSPY available on Github. While XploitSPY appears to have been originally developed by a group of self-reported ethical hackers in India, APT36 made modifications to it to produce a new malware variant we call LazaSpy A cloud based Android Monitoring Tool, powered by NodeJS	https://about.fb.com/wp-content/uploads/2022/08/Quarterly-Adversarial-Threat-Report-Q2-2022.pdf
JSSLoader	malware	RAT	JSS Loader is Remote Access Trojan (RAT) with .NET and C++ variants that has been used by FIN7 since at least 2020.	https://attack.mitre.org/software/S0648/

FOLIO ADMINISTRATIF

THESE DE L'INSA LYON, MEMBRE DE L'UNIVERSITE DE LYON

NOM : **MORIOT** DATE de SOUTENANCE : **24/09/2024**

Prénoms : **Camille**

TITRE : **Méthodologie de caractérisation socio-organisationnelle des adresses IPs appliquée à la sécurité**

NATURE : **Doctorat** Numéro d'ordre : **2024ISAL0077**

École Doctorale : **Informatique et Mathématique**

Spécialité : **Informatique**

RÉSUMÉ :
Internet est un système clé dans la société contemporaine. Il s'agit d'un système complexe réparti entre de nombreuses organisations ayant une variété de rôles et d'intérêts. Depuis leur création, les cyberattaques sont devenues des actifs précieux, car elles donnent aux rivaux des avantages, par exemple dans les domaines politique ou économique. Il est nécessaire d'analyser ces attaques, d'identifier leurs singularités et les mécanismes sur lesquels elles s'appuient afin de les contrer. Cela permettra d'établir des signatures plus précises et plus pertinentes et aidera la conception des contre-mesures. Un des aspects d'analyse des attaques sont les infrastructures utilisées par les attaquants pour générer les attaques. De nombreux outils aujourd'hui permettent de caractériser l'aspect technique des machines qui composent ces infrastructures. Mais comme les attaques ont lieu dans un environnement social, politique, économique et organisationnel, nous revendiquons qu'il est nécessaire d'évaluer ces machines d'un point de vue organisationnel. Cette thèse propose une méthodologie originale de catégorisation des adresses IP, à l'aide de 6 étiquettes décrivant deux axes : un axe technologique et un axe organisationnel. Nous proposons également un outil d'investigation, IPSeen, qui implémente cette méthodologie, en affectant les étiquettes aux adresses IP. Il s'appuie sur différentes sources de données : Wikidata, RDAP, Onyphe, GeoIPLite. Deux versions d'IPSeen sont proposées et évaluées dans ce manuscrit. Ces deux versions se différencient par leur rapidité et leur niveau de précision. Enfin, nous appliquons notre méthodologie à un ensemble de données réelles de suivi d'infrastructure de type "command and control". L'analyse produite propose une description des infrastructures des organisations qui maintiennent les machines participant aux infrastructures d'attaques. Nous montrons que notre approche apporte un éclairage essentiel sur la compréhension des attaques, en complément des nombreuses caractérisations techniques par ailleurs disponibles.

MOTS-CLÉS : **Qualification des IPs, Sécurité, Réseaux, Métrologie**
Laboratoire(s) de recherche : **CITI**

Directeur de thèse : **VALOIS Fabrice, Professeur des Universités, INSA-Lyon**

Président du Jury : **BOUABDALLAH Abdelmadjid**

Composition du Jury :

LAURENT Maryline
CHRISMENT Isabelle
OWEZARSKI Philippe
BOSSERT Georges
LESUEUR François
STOULS Nicolas
VALOIS Fabrice