الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة سعيدة – د. الطاهر مولاي –

UniversitéSaïda– Dr. Tahar Moulay –

Faculté des Mathématiques, Informatique et Télécommunications

MEMOIRE

Présenté pour l'obtention du **Diplôme** de **MASTER en** Télécommunications

Spécialité : Réseaux et Télécommunications

Par : M. BENABDERRAHMANE Mohamed Ibrahim
M. MEKKI Bencherif

Test de Kolmogorov-Smirnov: Application à la détection des cyber-attaques

Soutenu, le 18/06/ 2025, devant le jury composé de :

Melle. OUESSAI Asmaa	MCB	Président
Mr. BOUYEDDOU Benamar	MCA	Rapporteur
Mme. BOUCHENAK Sofia	MCB	Examinateur

2024 / 2025

REMERCIMENTS

Nous tenons à exprimer notre profonde gratitude à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire.

À Monsieur **Bouyeddou Benamar**, notre éminent directeur de recherche :
Nous vous sommes infiniment reconnaissants pour votre encadrement rigoureux et votre disponibilité tout au long de ce travail. Vos conseils avisés et votre expertise ont été déterminants dans l'élaboration de ce mémoire. Votre approche méthodologique et votre esprit critique nous ont permis d'approfondir notre réflexion et d'améliorer considérablement la qualité de ce travail.

À Madame **Ouessai Asmaa**, membre du jury :
Nous vous remercions chaleureusement pour votre lecture attentive et vos commentaires constructifs lors de la soutenance. Vos remarques pertinentes et vos questions éclairées ont grandement contribué à l'amélioration de ce travail. Votre expertise dans le domaine a été particulièrement précieuse.

À Madame **Bouchenak Sofia**, membre du jury :
Nous tenons à vous exprimer toute notre reconnaissance pour votre évaluation minutieuse et vos suggestions judicieuses. Vos observations ont permis d'enrichir ce mémoire et d'en affiner les conclusions. Votre regard expert a été d'une grande aide pour finaliser ce travail.

À Monsieur **Tami** :
Nous vous adressons nos remerciements les plus sincères pour votre soutien constant et vos précieux conseils tout au long de ce parcours d'études. Vos orientations ont été particulièrement utiles lors des phases cruciales de ce travail. Votre disponibilité et votre bienveillance ont été grandement appréciées.

Nous n’oublions pas de remercier l’ensemble des enseignants du département pour la qualité de leur enseignement et leurs conseils tout au long de notre parcours universitaire. Un remerciement particulier va également au personnel administratif pour son professionnalisme et son efficacité dans la gestion des divers aspects logistiques.

Ce mémoire n’aurait pu voir le jour sans l’apport et le soutien de toutes ces personnes. Qu’elles trouvent ici l’expression de notre profonde gratitude.

Nous remercions vivement nos familles Benabderahmane et mekki pour leur aide morale et matérielle durant toute la période de préparation.

DEDICASE

Dédicace spéciale du fond du cœur

À ma famille, les morceaux de mon cœur et la lumière de ma vie...

Mon cher papa... Mon plus grand soutien, toi qui as tout donné pour moi, financièrement et moralement, pour que je puisse poursuivre mes études sans crainte. Merci pour tes sacrifices innombrables et pour ta foi en moi, même quand je doutais de moi-même. Tu as été et resteras à jamais mon modèle de générosité et de patience.

Ma chère maman... Toi qui as transformé la fatigue en amour et la patience en un don inépuisable, tu es ma tendresse éternelle et ta persévérance a fait de moi ce que je suis aujourd'hui. Chaque réussite dans ma vie est le fruit de tes efforts et de tes prières incessantes.

Mehdi... Ma petite joie, toi qui remplis la maison de rires et de pureté. Malgré ton jeune âge, tu me rappelles sans cesse le sens du bonheur simple. Je te promets d'être pour toi ce que tu as été pour moi : un grand frère qui te protège et veille sur toi.

Asma et Kenza... Mes petites fleurs, vous qui semez la joie partout où vous passez. Votre présence à mes côtés rend la vie plus belle, et vos rires font oublier les soucis du monde.

Et à mes chères sœurs... Vous qui avez toujours été mon soutien et mon aide, votre présence me fait sentir comme l'homme le plus chanceux du monde.

Ces mémoires sont une partie de moi, je vous les dédie... Car c'est vous qui avez écrit mon histoire avec votre sang et vos efforts. Chaque mot ici porte un amour infini et une gratitude qui ne peut s'exprimer que par des prières pour vous, afin qu'Allah vous bénisse et vous préserve pour moi.

Par Allah, je ne vous mérite pas, mais je remercie mon Seigneur chaque jour de m'avoir accordé la grâce de vous avoir ❤.

Votre fils et frère dévoué,
Mohamed

DEDICASE

À mon père,
Merci pour ton soutien constant et ta présence tout au long de mon parcours. Ton appui m'a permis d'avancer sereinement dans mes études.

À ma mère,
Merci pour ton amour et ta patience. Ta présence m'a toujours encouragé à persévérer.

À mon frère younes ,
Merci d'avoir été là, pour ton aide et ta compagnie.

À mes sœurs, ma jumelle et ma petite sœur,
Merci pour votre présence et votre soutien au quotidien. Vous êtes importantes pour moi.

À mes grands-parents,
À mes tantes Saadia et Dalila,
À mes frères Zinedine et Abdelkader,

Merci à vous tous pour vos encouragements et votre soutien.

Ce travail vous est dédié avec toute ma reconnaissance.

Résumé

Le test de Kolmogorov-Smirnov (KS) est un puissant outil statistique pour comparer deux distributions de probabilités en mesurant la différence maximale entre leurs fonctions de distribution cumulative.

Dans ce travail, nous examinons utilité du test KS pour distinguer un comportement réseau normal d'activités malveillantes en comparant les distributions empiriques de caractéristiques du trafic (volume, segments TCP, messages ICMP…) à des références attendues. A travers des simulations, nous évaluons la sensibilité du test à différents types d'attaques (TCP SYN flood, Smurf et UDP flood).

Les résultats obtenus mettent en évidence son potentiel comme outil non-paramétrique et indépendant de tout modèle pour la détection précoce des cyber-attaques.

Mots-clés : *Test de Kolmogorov-Smirnov, détection d'anomalies, cyber-attaques, attaques DOS/DDOS, DARPA99.*

Abstract

The Kolmogorov-Smirnov (KS) test is a powerful statistical tool for comparing two probability distributions by measuring the maximum difference between their cumulative distribution functions.

In this work, we examine the usefulness of the KS test in distinguishing normal network behavior from malicious activities by comparing empirical distributions of traffic features (volume, TCP segments, ICMP messages, etc.) to expected reference distributions. Through simulations, we evaluate the test's sensitivity to different types of attacks (TCP SYN flood, Smurf, and UDP flood).

The results highlight its potential as a non-parametric and model-independent tool for the early detection of cyber-attacks.

Keywords: *Kolmogorov-Smirnov test, anomaly detection, cyber-attacks, DOS/DDOS attacks, DARPA99.*

الملخص

اختبار كولموجوروف-سميرنوف (KS) هو أداة إحصائية قوية تُستخدم لمقارنة توزيعين احتماليين عن طريق قياس أقصى فرق بين دالتي التوزيع التراكمي الخاصة بهما.

في هذا العمل، ندرس فائدة اختبار KS في التمييز بين السلوك الطبيعي للشبكة والأنشطة الضارة من خلال مقارنة التوزيعات التجريبية لخصائص حركة المرور (الحجم، مقاطعTCP ، رسائلICMP ، إلخ) مع التوزيعات المرجعية المتوقعة. ومن خلال المحاكاة، نقيم حساسية الاختبار لأنواع مختلفة من الهجمات (مثل هجومTCP SYN flood ، وهجومSmurf ، وهجوم.(UDP flood

تكشف النتائج عن إمكانات هذا الاختبار كأداة لا معلمية (غير بارامترية) ومستقلة عن أي نموذج للكشف المبكر عن الهجمات الإلكترونية.

الكلمات المفتاحية :*اختبار كولموجوروف-سميرنوف، كشف الشذوذ، الهجمات الإلكترونية، هجمات الحرمان من الخدمة/الحرمان الموزع من الخدمة(DOS/DDOS) ، مجموعة بيانات داربا 99.(DARPA99)*

Table des matières

Liste des figures

Liste des Tableaux

Liste des abréviations

ACK : Acknowledgment

ANOVA : Analysis of Variance

CDF : Cumulative Distribution Function

CL : Center Line

DDoS : Distributed Denial of Service

DoS :Denial of Service.

ECC: Elliptic Curve Cryptography

FTP : File Transfer Protocol

HTTP : HyperText Transfer Protocol

HTTPS : HyperText Transfer Protocol Secure

ICMP :Internet Control Message Protocol

IEEE :Institute of Electrical and Electronics Engineers

IDS : Intrusion Detection System

IMAP : Internet Message Access Protocol

IP : Internet Protocol

IPv4 : Internet Protocol version 4

IPv6 : Internet Protocol version 6

IPSec : Internet Protocol Security

IPS : Intrusion Prevention System

LCL : Lower Control Limit

MAC :Media Access Control).

MFA :Multi-Factor Authentication

MitM : Man-in-the-Middle

OSI : Open Systems Interconnection.

PCA :Plan de Continuité d'Activité

PIN :Personal Identification Number

PKI :Public Key Infrastructure

POP3 : Post Office Protocol v3

PRA: Plan de Reprise d'Activitaire

RSA: (Rivest-Shamir-Adleman

RST : Reset

SFTP : Secure FTP

SMTP : Simple Mail Transfer Protocol

SMS :Short Message Service

SSL : Secure Sockets Layer

SYN : Synchronize

TCP : Transmission Control Protocol

TLS : Transport Layer Security

UCL: Upper Control Limit

UDP : User Datagram Protocol

VPN :Virtual Private Network

Wi-Fi : Wireless Fidelity

WPA : Wi-Fi Protected Access

WPA2 : Wi-Fi Protected Access 2

Introduction Générale

Avec l'expansion des systèmes informatiques et des réseaux, les cyber-attaques sont devenues plus fréquentes et sophistiquées, menaçant la confidentialité, l'intégrité et la disponibilité des données. Les attaques par déni de service (DoS/DDoS), les intrusions et les malwares exploitent souvent des vulnérabilités difficiles à anticiper. Une détection précoce est cruciale pour minimiser les dommages, mais les méthodes traditionnelles basées sur des signatures peinent à identifier des attaques inédites. Dans ce sens, les approches statistiques offrent une alternative en analysant les écarts par rapport au comportement normal sans dépendre de règles prédéfinies. La cybersécurité moderne nécessite donc des outils adaptatifs capables de distinguer entre fluctuations légitimes et activités malveillantes, même dans des environnements dynamiques et bruyants.

D'autre part, Le test de Kolmogorov-Smirnov (KS) est une méthode statistique non paramétrique utilisée pour comparer une distribution de probabilité avec une distribution de référence ou deux distributions entre elles. Il mesure la distance maximale entre leurs fonctions de répartition cumulative, offrant ainsi une mesure simple mais puissante de leur similarité. Contrairement à d'autres tests, le KS ne nécessite pas d'hypothèses sur la forme de la distribution, ce qui le rend applicable à divers types de données. Sa sensibilité aux différences tant au niveau des queues de distribution qu'au centre en fait un outil polyvalent. Bien que principalement utilisé pour des distributions monovariable, il peut être adapté pour des analyses multivariées. Dans le contexte de la cybersécurité, il permet de détecter des écarts significatifs dans les modèles de trafic réseau ou les comportements système, révélant ainsi des anomalies potentielles.

Dans ce travail, nous étudions l'application du test de Kolmogorov-Smirnov pour la détection des cyber-attaques, en nous focalisant sur sa capacité à identifier des anomalies dans les données réseau. Nous évaluerons sa performance face à des attaques communes (TCP SYN flood, Smurf, UDP flood…) en comparant les distributions de trafic normal et malveillant.

Le manuscrit est organisé comme suit :

Dans le Chapitre 01, nous avons introduit les principes fondamentaux de la sécurité dans les réseaux IP avec un accent particulier sur les attaques de type DoS (Denial of Service) et DDoS (Distributed Denial of Service).

Le Chapitre 02 est dédié à l'étude du test de Kolmogorov-Smirnov. Nous exposons les fondements mathématiques de ce test, ses propriétés, son fonctionnement ainsi que ses avantages dans les contextes de détection d'anomalies.

Dans le Chapitre 03, nous explorons l'utilité de test Kolmogorov-Smirnov pour la détection des cyber-attaques en utilisant des données réelles provenant de la base DARPA99.

Chapitre I

Principes de sécurité dans les réseaux IP

Chapitre I

Principes de sécurité dans les réseaux IP

I.1. Introduction

Les réseaux IP constituent l'épine dorsale de la communication moderne. Ils facilitent la transmission de données via Internet et les réseaux privés, connectant des millions d'appareils à travers le monde. Cependant, cette interconnectivité s'accompagne de vulnérabilités qui peuvent compromettre la confidentialité, l'intégrité et la disponibilité des informations. Les cyber-attaques telles que le déni de service distribué (DDoS :Distributed Denial of Service), l'usurpation d'adresse IP (IP Spoofing) et l'écoute clandestine (Sniffing) sont des menaces fréquentes auxquelles les entreprises doivent faire face. La mise en œuvre de mesures de sécurité adéquates est donc essentielle pour garantir un fonctionnement sécurisé des réseaux IP.

Dans ce chapitre nous introduisons l'architecture TCP/IP et ses composantes clés avant d'aborder les principes fondamentaux de la sécurité dans les réseaux IP.

I.2. Fonctionnement du Modèle TCP/IP

Le modèle TCP/IP (Transmission Control Protocol / Internet Protocol) constitue l'architecture de référence des réseaux modernes, notamment Internet. Développé par le Department of Defense (DoD) dans les années 1970, il repose sur une approche pragmatique et modulaire permettant d'assurer l'interconnexion de réseaux hétérogènes à grande échelle [1].

Ce modèle repose sur une architecture en quatre couches, chacune ayant des responsabilités spécifiques en matière de transmission des données. Comparé au modèle OSI (Open Systems Interconnection), qui en compte sept, le modèle TCP/IP est plus simple et aligné sur les protocoles réellement déployés sur Internet.

Les principales caractéristiques du modèle TCP/IP sont :

- **Modularité** : Chaque couche remplit une fonction distincte et peut évoluer indépendamment des autres [1].
- **Fiabilité** : La couche Transport, avec TCP, assure une communication robuste et sans erreur.
- **Interopérabilité** : TCP/IP permet la communication entre des systèmes hétérogènes, quel que soit leur fabricant ou leur système d'exploitation.
- **Scalabilité** : Il supporte un très grand nombre d'appareils connectés, d'où le passage à IPv6 pour répondre à la pénurie d'adresses IPv4.

Nous détaillerons dans ce qui suit les rôles et les fonctions de chaque couche du modèle TCP/IP.

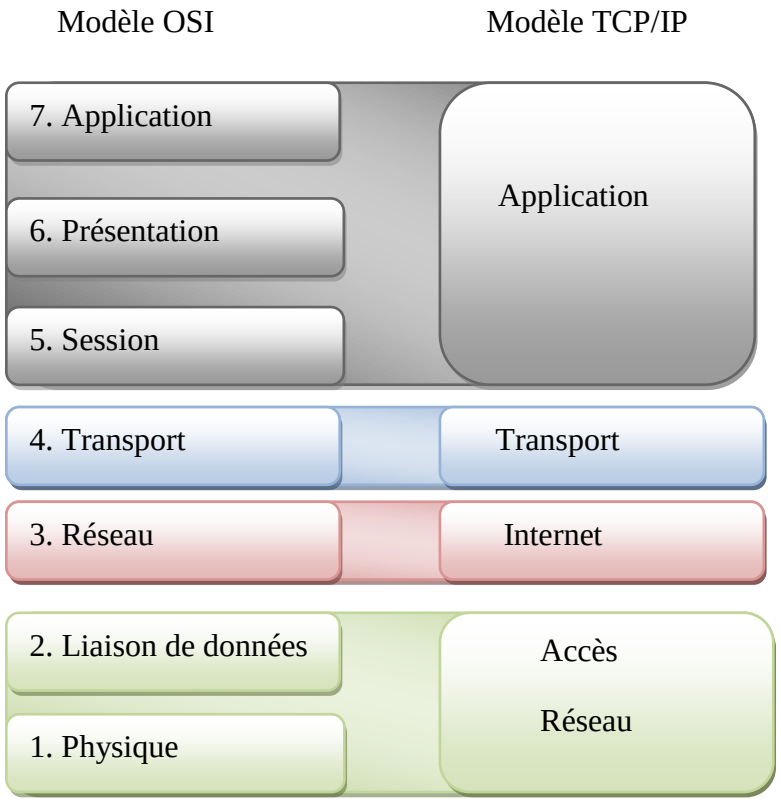


Figure I.1 : Le modèle TCP/IP vs modèle OSI

I.2.1 Couche Accès Réseau

La couche Accès Réseau assure la liaison entre un appareil et le support physique de transmission. Elle est responsable de l'encapsulation des paquets IP en trames et de la gestion de l'accès aux supports de communication [2]. Les principaux protocoles et technologies sont [1] [2] :

- **Ethernet (IEEE 802.3)** : Norme dominante dans les réseaux filaires, utilisant des adresses **MAC** pour identifier les périphériques sur un réseau local.
- **Wi-Fi (IEEE 802.11)** : Permet une connectivité sans fil et utilise des techniques comme **le chiffrement WPA/WPA2** pour sécuriser les échanges.
- **PPP (Point-to-Point Protocol)** : Utilisé pour les connexions directes entre deux nœuds, souvent dans les connexions Internet via modem.
- **Fibre optique** : Utilise des signaux lumineux pour transmettre des données à **très haute vitesse** et sur de longues distances.

I.2.2. Couche Internet

La couche Internet est chargée de l'acheminement des paquets à travers différents réseaux. Son objectif principal est de garantir que les données envoyées d'un appareil source atteignent leur destination, même si ces appareils sont connectés à des réseaux différents [3].

Elle repose principalement sur le protocole IP (Internet Protocol), qui attribue une adresse unique à chaque appareil connecté à un réseau. Grâce à ces adresses, les routeurs peuvent prendre des décisions sur le chemin optimal à emprunter pour transmettre les paquets [4]. Parmi les protocoles de la couche réseau on trouve :

- **IP (Internet Protocol)** : Protocole principal qui gère l'adressage et le routage des paquets. Il existe deux versions majeures :
  - **IPv4** : Utilise des adresses sur 32 bits (exemple : 192.168.1.1) [3].
  - **IPv6** : Utilise des adresses sur 128 bits, offrant un espace d'adressage beaucoup plus vaste [4].
- **ICMP (Internet Control Message Protocol)** : Utilisé pour envoyer des messages d'erreur et de diagnostic (exemple : commande ping) [3].

- **ARP (Address Resolution Protocol)** : Permet de traduire une adresse IP en adresse MAC dans un réseau local (**[5]**).

- **Fonctions essentielles soitles supprimer soit les fusionner avec Définition et rôle**
- **Adressage logique** : Attribution d'une adresse IP unique à chaque appareil connecté (**[3]**).
- **Routage des paquets** : Transmission des données en empruntant le chemin le plus efficace grâce aux routeurs (**[5]**).
- **Fragmentation et réassemblage** : Découpage des paquets trop volumineux pour qu'ils puissent être transmis sur certains réseaux (**[3]**).

I.2.3. Couche Transport

La couche Transport est chargée d'assurer une communication fiable et efficace entre les applications des différents hôtes. Elle gère le découpage et le réassemblage des données, la gestion des connexions et la correction des erreurs [1].

Elle repose principalement sur deux protocoles majeurs :

1. **TCP (Transmission Control Protocol)** : Protocole orienté connexion qui garantit la livraison des paquets dans le bon ordre, sans perte ni duplication [3]**.**
2. **UDP (User Datagram Protocol)** : Protocole plus léger et rapide, mais qui ne garantit pas la fiabilité des transmissions [5].

- **Comparaison entre TCP et UDP**

Fonction	TCP	UDP
Fiabilité	Oui (accusé de réception, retransmission)	Non
Vitesse	Plus lent (car contrôle d'erreur)	Très rapide
Utilisation	Web, e-mails, transferts de fichiers	Vidéo en streaming, VoIP, jeux en ligne

Tableau I.1 : Comparaison entre TCP et UDP

- **Fonctions essentielles [1-3]**

1. **Multiplexage des connexions** : Permet à plusieurs applications de partager une même connexion réseau.

2. **Contrôle de flux** : TCP ajuste la vitesse d'envoi des paquets en fonction de la capacité du récepteur.
3. **Correction d'erreurs** : TCP détecte et corrige les erreurs grâce à des accusés de réception et des retransmissions.

I.2.4. Couche Application

La couche Application est celle qui interagit directement avec les utilisateurs. Elle regroupe tous les services qui permettent l'échange de données et l'exécution d'applications sur le réseau [1]. Les principaux protocoles applicatifs sont :

- **HTTP/HTTPS (HyperText Transfer Protocol/ http Secure)** : Utilisé pour la navigation Web. HTTPS ajoute une couche de chiffrement (**[2]**).
- **FTP/SFTP** : Permet le transfert de fichiers. SFTP est sécurisé (**[5]**).
- **DNS** : Convertit les noms de domaine en adresses IP (**[3]**).
- **SMTP/IMAP/POP3** : Gèrent l'envoi et la réception des e-mails (**[1]**).

I.3. Principes Fondamentaux de la Sécurité des Réseaux IP

La sécurité des réseaux IP est un enjeu majeur pour garantir la confidentialité, l'intégrité et la disponibilité des données échangées sur Internet. Avec l'augmentation des cyber-attaques et des menaces sophistiquées, il est essentiel de mettre en place des mécanismes robustes pour protéger les communications et les infrastructures informatiques [1].

Le modèle TCP/IP, bien qu'efficace pour l'acheminement des données, ne possède pas de mécanismes de sécurité natifs. C'est pourquoi divers protocoles de sécurité et solutions techniques ont été développés pour renforcer sa protection contre les attaques potentielles [2].

Les principes fondamentaux de la sécurité des réseaux IP reposent sur cinq piliers essentiels :

I.3. Confidentialité

La confidentialité garantit que seules les entités autorisées peuvent accéder aux données échangées. Sans mesures de protection adéquates, les communications peuvent être interceptées par des attaquants qui peuvent exploiter ces informations à des fins malveillantes [3].

Pour garantir la confidentialité des données échangées sur les réseaux IP, plusieurs techniques sont couramment utilisées. L'une des méthodes les plus efficaces est le chiffrement, qui permet de rendre les informations illisibles pour toute personne non autorisée. Par exemple, les protocoles comme (SSL/TLS : Secure Socket Layer/Transport Layer Security) sont largement utilisés pour sécuriser les communications sur Internet, notamment lors des trans actions en ligne ou des échanges d'e-mails.

L'authentification joue également un rôle essentiel : elle permet de vérifier l'identité des utilisateurs ou des systèmes avant de leur accorder l'accès aux données. Cela peut passer par des mots de passe robustes, des certificats numériques, ou encore des systèmes biométriques.

Les réseaux privés virtuels (VPN : Virtual Private Network) sont une autre solution largement adoptée pour renforcer la confidentialité. Ils permettent de créer un tunnel sécurisé entre l'utilisateur et le réseau, empêchant ainsi toute interception des données en cours de transmission. Enfin, des politiques de contrôle d'accès bien définies au sein des systèmes et des applications permettent de limiter l'accès aux informations sensibles uniquement aux personnes autorisées.

I.3.2. Authentification

L'authentification permet de vérifier l'identité des utilisateurs et des appareils qui tentent d'accéder aux ressources d'un réseau. Sans authentification robuste, un attaquant pourrait se faire passer pour un utilisateur légitime et obtenir un accès non autorisé [4].

Il existe plusieurs méthodes d'authentification utilisées pour vérifier l'identité des utilisateurs et des appareils. La méthode la plus courante est l'authentification par mot de passe, bien que cette technique puisse être vulnérable si les mots de passe sont faibles ou compromis. C'est pourquoi il est fortement recommandé d'utiliser des mots de passe complexes, comprenant des lettres, des chiffres et des symboles.

Cependant, certains systèmes sont plus vulnérables que d'autres aux attaques. Par exemple, les mots de passe et les PIN (Personal Identitication Number) sont des systèmes classiques mais peuvent facilement être contournés par des attaques par force brute ou du phishing [5]. Pour pallier cette faiblesse, l'authentification à deux facteurs (2FA) ou l'authentification multi-facteurs (MFA) est de plus en plus utilisée. Cette méthode combine

plusieurs facteurs d'authentification (comme un mot de passe et un code SMS, ou un mot de passe et des données biométriques), ce qui renforce considérablement la sécurité [1].

Les certificats numériques et les infrastructures à clé publique (PKI : Public Key Infrastructure) sont également utilisés pour authentifier les serveurs Web et garantir des connexions sécurisées via HTTPS, notamment dans le cadre des transactions en ligne [2]. Ces systèmes reposent sur des clés cryptographiques pour assurer l'intégrité des communications.

Une méthode encore plus avancée et sécurisée est l'authentification biométrique, qui inclut des technologies comme la reconnaissance d'empreintes digitales, la reconnaissance faciale, ou la reconnaissance rétinienne. Ces méthodes offrent un niveau de sécurité élevé et sont de plus en plus utilisées pour l'authentification des utilisateurs [5].

I.3.3. Intégrité des Données

L'intégrité des données garantit que les informations ne sont pas altérées accidentellement ou intentionnellement pendant leur transmission. Une modification non autorisée des données peut compromettre la fiabilité des transactions et des communications [3].

Pour garantir la confidentialité des données sur les réseaux IP, plusieurs mesures peuvent être mises en place. Le chiffrement est sans doute l'une des protections les plus utilisées. Il permet de rendre les informations illisibles pour toute personne non autorisée. Par exemple, des protocoles comme SSL/TLS ou IPsec sont souvent utilisés pour sécuriser les communications. Ces technologies assurent que même si les données sont interceptées, elles ne pourront pas être exploitées.

L'utilisation d'un VPN est aussi une solution courante. Il crée une sorte de tunnel sécurisé entre deux appareils, ce qui rend les échanges invisibles aux regards extérieurs. D'un autre côté, il est important de vérifier l'identité des personnes qui accèdent aux données sensibles. C'est pourquoi l'authentification forte, avec plusieurs niveaux de vérification, est de plus en plus répandue.

Enfin, la gestion des accès joue un rôle essentiel : il faut s'assurer que seules les personnes autorisées puissent consulter certaines informations. Cela passe par des règles précises et des droits d'accès bien définis. En complément, les pare-feux et les systèmes de détection d'intrusion permettent de surveiller le trafic et de bloquer les tentatives d'accès suspectes.

Ensemble, ces mesures contribuent à protéger efficacement la confidentialité des échanges sur Internet.

I.3.4. Disponibilité

La disponibilité garantit que les services et les données restent accessibles aux utilisateurs légitimes, même en cas de tentatives de sabotage telles que les attaques DDoS (Distributed Denial of Service) [1].

Parmi les menaces majeures contre la disponibilité dans les réseaux IP, les attaques par DoS et DDoS sont parmi les plus redoutables. Elles consistent à inonder un service ou un serveur de requêtes afin de le rendre indisponible pour les utilisateurs légitimes. Ces attaques sont souvent lancées à partir de réseaux de machines infectées appelés botnets, ce qui les rend très difficiles à bloquer. En dehors de ces attaques, d'autres facteurs peuvent également compromettre la disponibilité, comme les défaillances matérielles, les erreurs de configuration, les coupures d'électricité, ou encore les actes de sabotage physique. Même une simple panne logicielle ou une mauvaise manipulation peut entraîner l'interruption d'un service essentiel.

Pour garantir une disponibilité constante des services, il est essentiel de mettre en place plusieurs niveaux de protection. L'une des premières stratégies consiste à utiliser la redondance, en dupliquant les serveurs et les équipements critiques pour qu'un élément puisse prendre le relais en cas de panne. Les systèmes de répartition de charge (load balancers) permettent également de distribuer intelligemment le trafic afin d'éviter la surcharge d'un seul point.

De plus, des solutions spécialisées dans la détection et la mitigation des attaques DDoS sont nécessaires pour filtrer le trafic malveillant avant qu'il n'impacte les services. La mise en place de plans de continuité d'activité (PCA : Plan de Continuité d'Activité) et de reprise après sinistre (PRA : Plan de Reprise d'Activitaire) est aussi primordiale pour permettre une reprise rapide en cas d'incident. Enfin, une surveillance permanente du réseau, des tests réguliers de résilience, ainsi qu'une maintenance proactive des systèmes permettent d'anticiper et de limiter les interruptions potentielles.

I.3.5. Non-répudiation

La non-répudiation empêche un utilisateur ou un système d'affirmer faussement qu'il n'a pas effectué une action donnée, comme l'envoi d'un e-mail ou une transaction en ligne [2].

Pour garantir la non-répudiation, plusieurs techniques sont mises en place afin d'assurer que les actions réalisées dans un système peuvent être vérifiées de manière fiable et authentique. L'une des méthodes les plus courantes est l'utilisation de signatures numériques, qui sont générées à l'aide de la cryptographie asymétrique, comme RSA (Rivest, Shamir, Adlman) ou ECC (Elliptic Curve Cryptography). Ces signatures permettent d'authentifier un message ou une transaction, garantissant ainsi qu'il n'y a pas de doute quant à l'origine des données échangées [6].

Une autre technique importante est l'horodatage électronique, qui associe une date et une heure vérifiables à une action, ce qui permet de prouver son existence à un instant donné. Cela est particulièrement utile pour la validation des transactions et des événements dans un contexte juridique ou commercial [4].

Enfin, les journaux et logs d'audit jouent un rôle crucial dans la non-répudiation. Ces enregistrements détaillent les événements critiques et les actions effectuées sur un réseau. Ils permettent de suivre et d'analyser les actions passées afin de vérifier toute contestation et garantir l'intégrité des processus dans un système informatique [3].

I.4. Attaques TCP/IP :

Les réseaux TCP/IP sont largement utilisés pour la communication de données, mais ils ne sont pas à l'abri des attaques. Les attaquants peuvent exploiter différentes vulnérabilités pour perturber ou accéder illégalement aux systèmes et aux informations. Cette section explore plusieurs exemples pratiques d'attaques sur la pile TCP/IP.

I.4.1. Attaque par Déni de Service Distribué (DDoS)

Une attaque DDoS consiste à surcharger un serveur, un réseau ou un service en envoyant une quantité massive de trafic. Cette surcharge empêche les utilisateurs légitimes d'accéder au service. L'attaque repose souvent sur un réseau de machines compromises, également appelé botnet, qui envoie simultanément des requêtes vers la cible. Par exemple : lors d'une attaque DDoS contre un site de commerce en ligne, l'attaquant peut générer un volume énorme de requêtes HTTP, rendant le site inaccessible pour ses clients légitimes, ce qui entraîne une perte de revenus et un dommage à la réputation de l'entreprise [1].

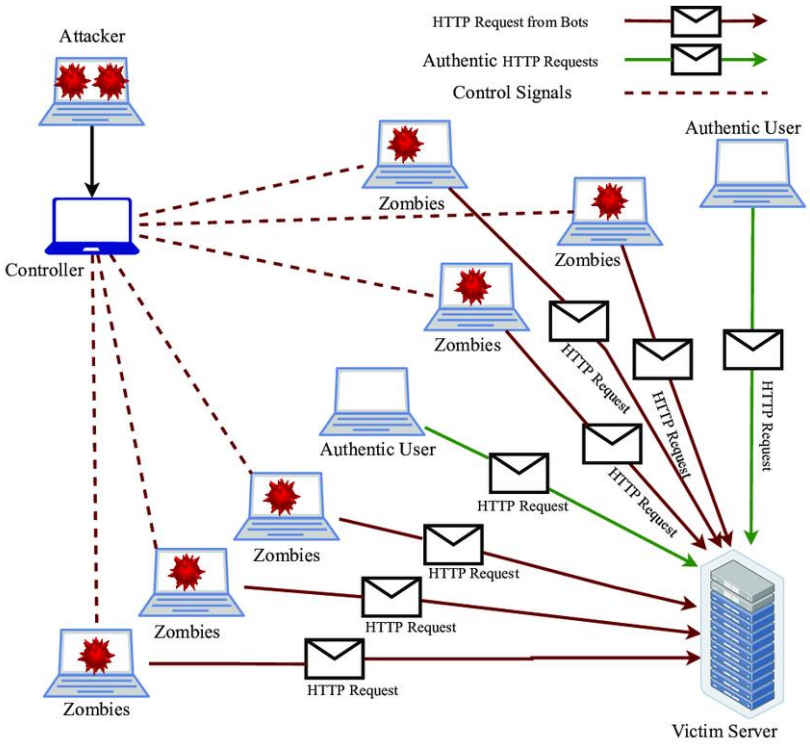


Figure I.2 : Attaque par Déni de Service Distribué

I.4.2. Attaque par Spoofing IP

Le spoofing IP est une technique où l'attaquant falsifie l'adresse IP source d'un paquet pour le faire passer pour une autre machine. Cette attaque peut être utilisée pour contourner des filtres de sécurité ou pour mener des attaques de type Man-in-the-Middle (MitM). Dans ce type d'attaque, un attaquant peut envoyer des paquets à un serveur, en se faisant passer pour un client légitime, afin de lancer une attaque sur le serveur sans être détecté. Cela peut être particulièrement dangereux dans des environnements où les identités sont vérifiées par adresse IP [3].

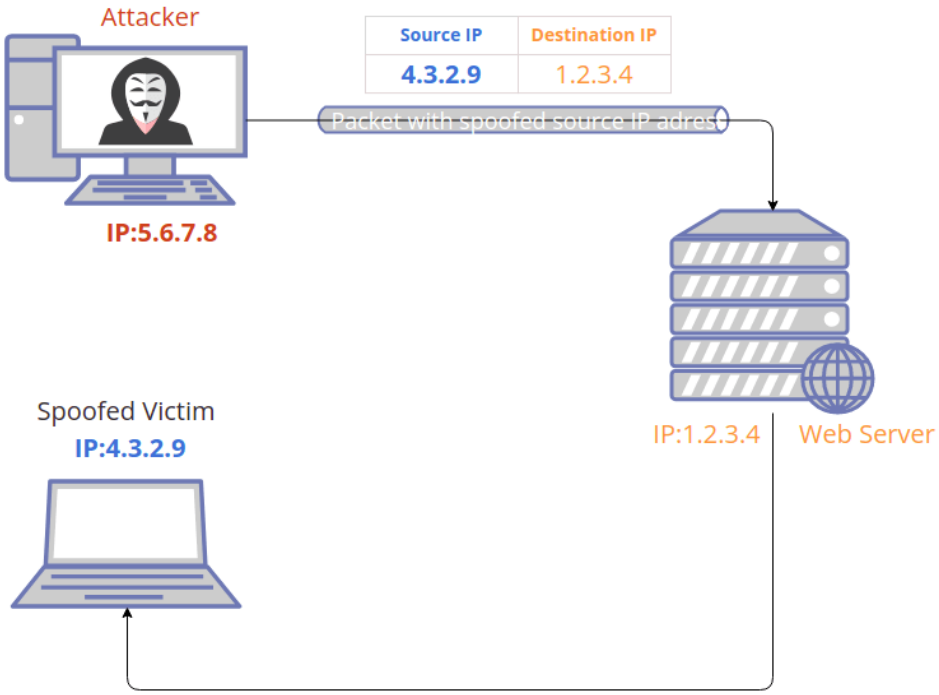


Figure I.3: Attaque par Spoofing IP

I.4.3.Attaque de Rejeu (Replay Attack)

Une attaque de Rejeu se produit lorsqu'un attaquant capture et rejoue des paquets de données précédemment transmis pour contourner les mécanismes d'authentification ou pour manipuler des transactions. Ici, dans un réseau sans chiffrement, un attaquant peut intercepter une demande de transfert de fonds entre deux parties, puis rejouer cette demande pour dupliquer la transaction. L'attaque peut être utilisée pour détourner des fonds ou pour perturber la communication sécurisée [4].

I.4.4. Attaque par Injection de Paquets Malveillants

L'injection de paquets malveillants est une attaque dans laquelle l'attaquant insère des paquets de données spécialement conçus dans une communication TCP/IP existante pour perturber ou altérer le flux de données. Dans ce cas, l'attaquant peut injecter un paquet malveillant dans une session de communication existante entre deux machines pour provoquer une défaillance du service ou accéder à des informations sensibles [2].

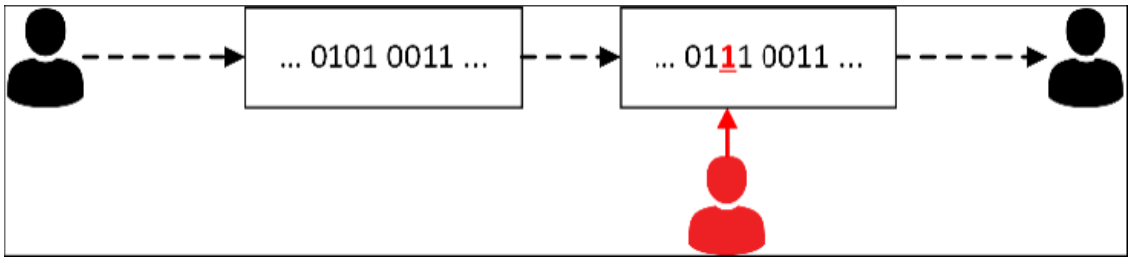


Figure I.4: Attaque par injection de paquets

I.4.5. Attaque TCP Reset (RST Attack)

Cette attaque consiste à injecter un paquet TCP avec le drapeau RST afin de forcer la terminaison immédiate d'une connexion TCP active. Elle est souvent utilisée pour interrompre une communication entre deux systèmes sans avoir à compromettre l'un d'eux.Lors d'une session de téléchargement de données, un attaquant envoie un paquet RST falsifié vers l'une des parties, forçant l'arrêt de la session. Cela peut être exploité pour saboter des connexions VPN ou HTTP [7].

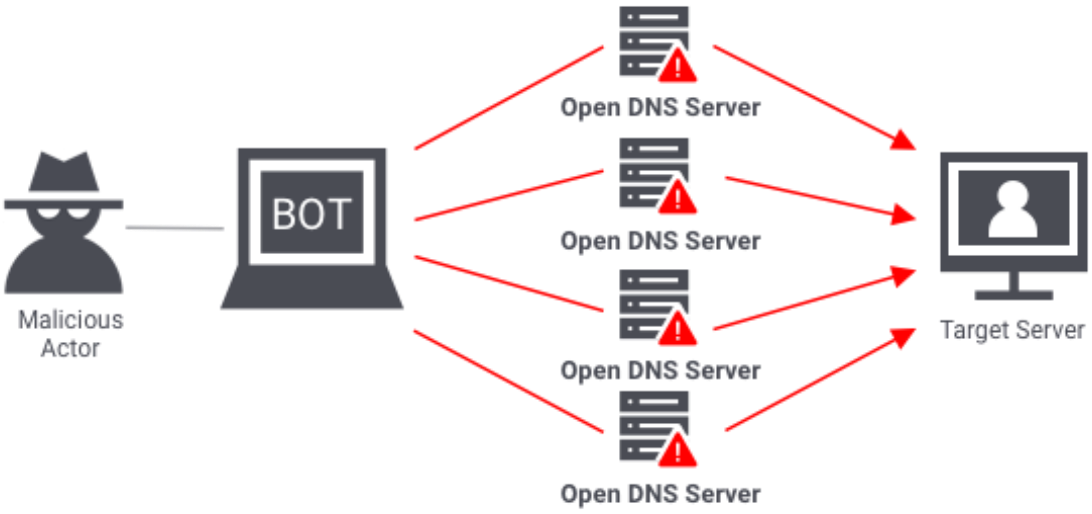


Figure I.5: Attaque RST

I.4.6. Attaque ICMP Redirect (Redirection ICMP)

Une attaque de redirection ICMP permet à un attaquant de rediriger le trafic réseau d'une machine cible en envoyant de faux paquets ICMP de type redirect. Ces paquets indiquent un nouveau chemin prétendument plus court, mais qui passe par un nœud compromis. Pour y faire, l'attaquant envoie un message ICMP modifié à une machine, lui indiquant que la passerelle réseau est une machine sous son contrôle, ce qui permet de surveiller ou modifier les communications [7].

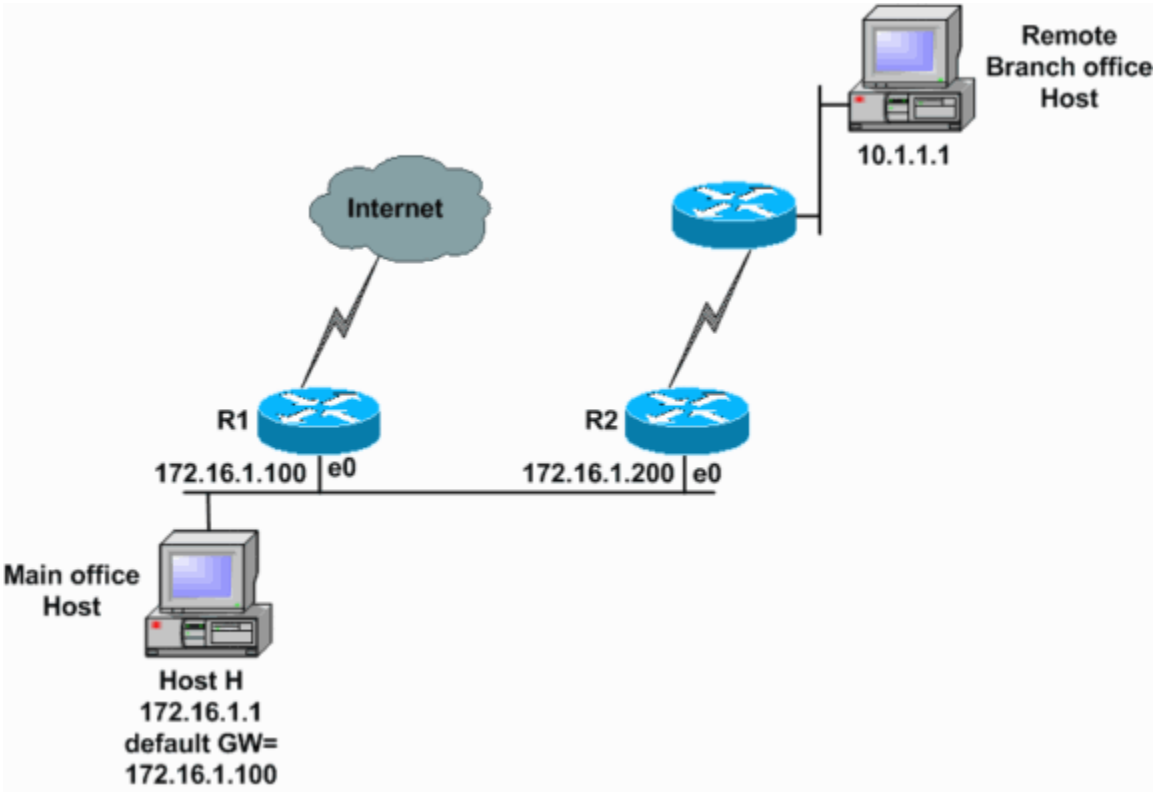


Figure I.6: Attaque ICMP Redirect

I.4.7. Tempête TCP ACK (TCP ACK Storm)

Cette attaque repose sur la création d'une boucle de paquets ACK entre deux hôtes. Chaque réponse ACK déclenche une autre réponse de l'autre hôte, provoquant ainsi une saturation du réseau ou des ressources systèmes. Par exemple, deux serveurs mal configurés peuvent être manipulés pour qu'ils échangent indéfiniment des paquets TCP ACK, ce qui entraîne une congestion du réseau local et une dégradation significative des performances [7].

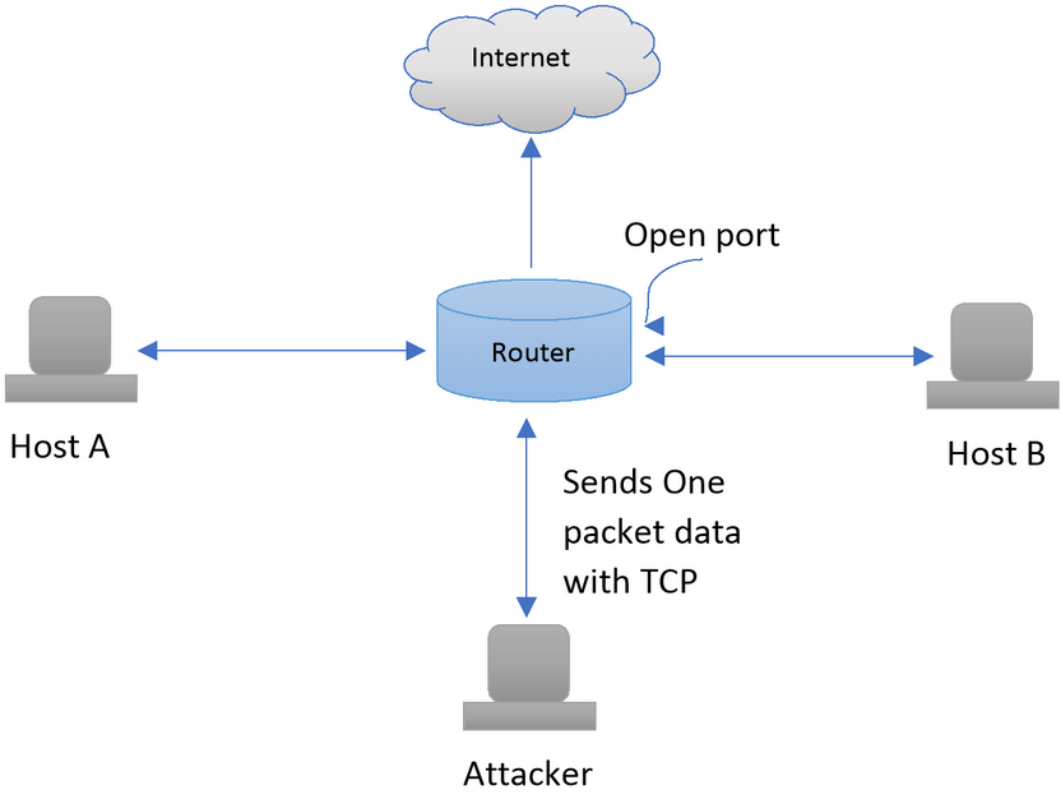


Figure I.7: Attaque TCP ACK Storm

I.4.8. Attaque par Routage IP Source (IP Source Routing)

Le routage IP source, bien que rarement utilisé aujourd'hui, permettait à l'expéditeur de spécifier la route que devait suivre le paquet. Un attaquant peut exploiter cette fonctionnalité pour faire passer les paquets par des machines sous son contrôle.En configurant un itinéraire détourné, il intercepte les paquets d'une cible sans que celle-ci ne s'en aperçoive, ce qui facilite l'espionnage ou l'altération des données échangées [7].

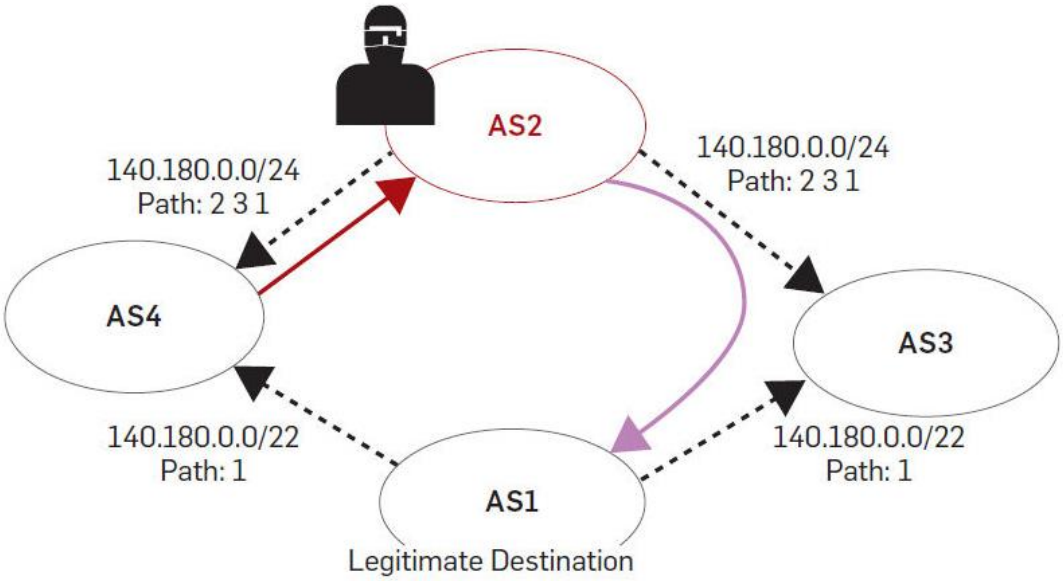


Figure I.8: Attaque par routage IP Source

I.4.9. Attaque LAND (Local Area Network Denial)

L'attaque LAND repose sur l'envoi d'un paquet TCP dont l'adresse IP source et l'adresse IP de destination sont les mêmes, provoquant une confusion au niveau du système ciblé, et parfois un plantage [7].

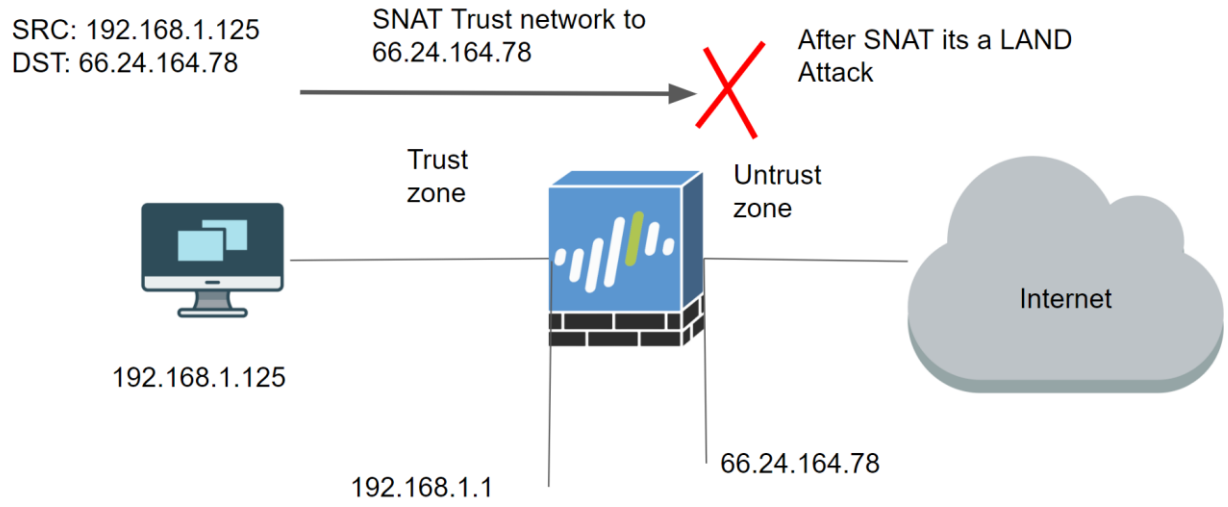


Figure I.9: Attaque LAND

I.5.Conclusion

Ce chapitre a exposé les principes fondamentaux de la cybersécurité dans les réseaux IP, en analysant les différentes cyber-attaques ainsi que les protocoles sous-jacents et leurs mécanismes de fonctionnement. Dans un environnement numérique de plus en plus menaçant, la mise en œuvre de protections efficaces pour atteindre un niveau de sécurité optimal s'impose comme une priorité absolue.

Face à ces défis, les approches statistiques émergent comme une solution prometteuse, captant l'attention croissante de la communauté scientifique et des experts en sécurité. C'est dans cette optique que le chapitre suivant présente le test de Kolmogorov-Smirnov, une méthode statistique robuste offrant des perspectives intéressantes pour la détection de diverses formes de cyber-attaques.

Chapitre II

Test de Kolmogorov-Smirnov

Chapitre II

Test de Kolmogorov-Smirnov

II.1. Introduction

Dans de nombreuses analyses de données l'hypothèse d'une distribution paramétrique spécifique (comme la loi normale) s'avère souvent irréaliste. Les tests non paramétriques émergent alors comme solution idéale, permettant d'effectuer des inférences statistiques sans suppositions restrictives sur la distribution sous-jacente. Leur robustesse et leur applicabilité étendue en font des outils privilégiés pour l'étude des données réseau complexes.

Parmi ces méthodes, le test de Kolmogorov-Smirnov (KS) se distingue par sa puissance et son élégance mathématique. Ce chapitre propose une exploration approfondie de ce test fondamental : ses fondements théoriques, son mécanisme de comparaison de distributions et son application potentielle en détection d'anomalies réseau.

II.2.Tests statistiques en math

Les tests statistiques non paramétriques représentent une classe essentielle de méthodes d'inférence qui s'affranchissent des contraintes distributionnelles rigides imposées par les approches paramétriques classiques. Leur principal avantage réside dans leur capacité à traiter des données ne satisfaisant pas les hypothèses de normalité ou de linéarité, tout en restant applicables à des échantillons de taille réduite ou à des variables qualitatives ordinales/nominales. Ces méthodes exploitent principalement les propriétés ordinales des données ou comparent directement les fonctions de distribution cumulative, ce qui les rend particulièrement robustes face aux distributions asymétriques, multimodales ou contaminées par des valeurs aberrantes. Leur développement remonte aux travaux fondateurs de Wilcoxon (1945) et Mann-Whitney

(1947), constituant depuis une alternative indispensable lorsque les conditions d'application des tests paramétriques ne sont pas vérifiées [8-10].

Les applications des tests non paramétriques couvrent un spectre analytique étendu en recherche et en analyse de données :

- **Validation distributionnelle :** permet d'évaluer l'adéquation empirique/théorique, teste la normalité ;
- **Comparaisons inter-groupes :** pour échantillons indépendants, échantillons appariés et pour extensions multi-groupes ;
- **Détection d'anomalies :**sert à l'identification des valeurs extrêmes univariées via l'écart normalisé maximal ;
- **Designs répétés :** pour mesures répétées sur données non normales

Ces tests reposent fondamentalement sur les principes suivants :

- **Indépendance distributionnelle :** ils comparent directement les fonctions de répartition ou transforment les données en rangs ;
- **Robustesse métrique :** Leur statistique de test basée sur les rangs ou percentiles minimise l'impact des outliers. Par exemple, le test de Wilcoxon utilise la médiane plutôt que la moyenne ;
- **Flexibilité opérationnelle :** ils permettent d'analyser des données censurées (tests de survie non paramétriques) ou des échelles ordinales [11] ;

II.2.1. Test de Mann-Whitney (Mann-Whitney U Test)

Le test de Mann-Whitney (ou test U) est une méthode non paramétrique permettant de comparer deux groupes indépendants lorsque les données ne suivent pas une distribution normale ou sont de nature ordinale. Ce test évalue si les deux populations proviennent d'une même distribution en se basant sur le classement des observations (rangs). Son principe repose sur le calcul d'une statistique U (eq. II.1), obtenue en comparant la somme des rangs entre les groupes, avec une hypothèse nulle (H_0) postulant l'égalité des distributions et une alternative (H_1) pouvant être bilatérale ou unilatérale. Particulièrement robuste aux valeurs extrêmes et aux petits échantillons ($n < 30$), ce test est largement employé dans divers domaines : comparaison de

l'efficacité de traitements médicaux, analyse de salaires entre secteurs, ou évaluation de méthodes pédagogiques. Il constitue ainsi une alternative fiable au test t lorsque les conditions paramétriques ne sont pas satisfaites, bien qu'il soit moins puissant si ces dernières sont remplies [12].

$$U = R_1 - \frac{n_1(n_1+1)}{2} \qquad \text{(II.1)}$$

où :

- (R_1) est la somme des rangs pour le premier échantillon.

- (n_1) est la taille du premier échantillon.

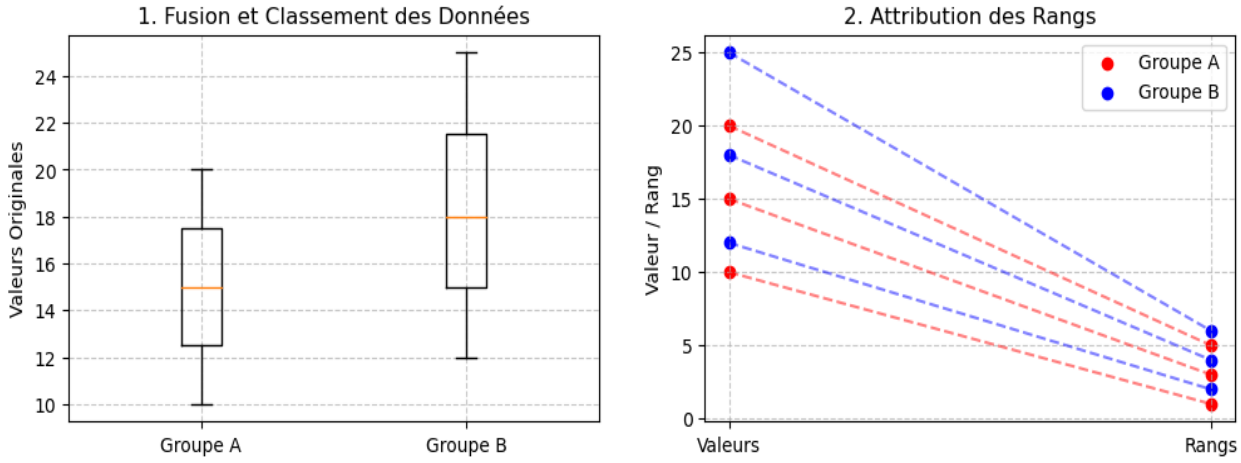


Figure II.1 : Principe du Test de Mann-Whitney (Test U)

II.2.2. Test de Wilcoxon (WilcoxonSigned-Rank Test)

Le test des rangs signés de Wilcoxon est une méthode non paramétrique utilisée pour comparer deux échantillons appariés lorsque les données ne suivent pas une distribution normale, se concentrant sur la médiane des différences plutôt que sur la moyenne. Il s'applique à des données appariées (mesures répétées, avant/après traitement) et fonctionne en calculant les différences entre paires, classant leurs valeurs absolues, attribuant des rangs signés, puis comparant la somme des rangs positifs et négatifs pour obtenir une statistique W (Eq. II.2). Ce test est robuste aux valeurs aberrantes et aux distributions asymétriques. Il évalue si la médiane des différences s'écarte significativement de zéro, avec une efficacité proche de 95 % par rapport au test t pour des distributions normales, tout en étant supérieur pour des données non normales.

Cependant, il ignore les paires sans différence et suppose une symétrie approximative des différences pour une interprétation fiable de la médiane [13].

$$W = \sum_{i=1}^{n} sgn(x_i) \cdot R_i \quad \text{(II.2)}$$

où :

- ($sgn(x_i)$) est le signe de la différence.

- (R_i) est le rang de la différence.

II.2.3. Test de Kruskal-Wallis (Kruskal-Wallis Test)

Le test de Kruskal-Wallis (test H) est une méthode non paramétrique permettant de comparer les distributions de trois groupes indépendants ou plus, servant d'alternative robuste à l'ANOVA unidirectionnelle lorsque les hypothèses de normalité ou d'homogénéité des variances ne sont pas satisfaites. En combinant toutes les observations pour les classer par ordre croissant, le test calcule une statistique H (Eq. II.3) basée sur les sommes des rangs de chaque groupe, qui suit approximativement une distribution du χ^2 avec k-1 degrés de liberté (k étant le nombre de groupes. Ce test évalue si au moins un groupe diffère significativement des autres par sa médiane ou la forme de sa distribution, sans exiger d'hypothèses sur la forme des distributions sous-jacentes. Particulièrement utile pour les données ordinales ou les échantillons de petite taille, il généralise le test de Mann-Whitney à plus de deux groupes. Bien que moins puissant que l'ANOVA pour des données normales, sa flexibilité et sa robustesse en font un outil privilégié pour l'analyse comparative de groupes indépendants dans des conditions non paramétriques [14].

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \quad \text{(II.3)}$$

où :

- (N) est le nombre total d'observations.

- (R_i) est la somme des rangs pour l'échantillon (i).

- (n_i) est la taille de l'échantillon (i).

II.2.4. Test de Shapiro-Wilk (Shapiro-Wilk Test)

Le test de Shapiro-Wilk est une méthode statistique permettant d'évaluer la normalité d'un ensemble de données, particulièrement recommandée pour les échantillons de petite taille (n

<50). En calculant une statistique W (Eq. II.4) qui mesure l'adéquation entre les données observées et une distribution normale théorique, via une analyse de la corrélation entre les valeurs triées et les scores normaux attendus, ce test permet de statuer sur l'hypothèse nulle (H0 : les données suivent une loi normale) en comparant la p-value obtenue à un seuil de significativité prédéfini (généralement $\alpha = 0.05$). Contrairement à d'autres tests de normalité, le test de Shapiro-Wilk offre une meilleure puissance statistique pour les petits échantillons, bien qu'il puisse perdre en sensibilité pour n > 2000. Son interprétation repose sur un principe simple : si $p < \alpha$, on rejette H0 et conclut à une non-normalité des données, information cruciale avant d'utiliser des tests paramétriques (t-test, ANOVA) exigeant cette condition. Cependant, comme tout test d'hypothèse, il présente des limites : une puissance réduite pour les très grands échantillons (où des écarts infimes à la normalité deviennent significatifs sans être pertinents) et une sensibilité aux valeurs aberrantes [15].

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{(II.4)}$$

où :

- (a_i) sont des coefficients dépendant de la taille de l'échantillon.

- $(x_{(i)})$ sont les valeurs ordonnées de l'échantillon.

II.2.5. Test exact de Fisher (Fisher's Exact Test)

Le test exact de Fisher est une méthode statistique précise conçue pour analyser l'association entre deux variables catégorielles organisées en tableau de contingence 2×2, particulièrement adaptée aux petits effectifs où les tests asymptotiques comme le χ^2 deviennent imprécis. Contrairement aux approximations utilisées par le χ^2, ce test calcule exactement la probabilité (p-value) d'observer la configuration des données ou une configuration plus extrême sous l'hypothèse nulle (H_0 : indépendance entre les variables), en énumérant toutes les permutations possibles des données marginales fixes via la loi hypergéométrique. Son application est cruciale lorsque les effectifs attendus sont <5, garantissant une validité même pour des échantillons très petits – par exemple, en recherche médicale pour évaluer l'association entre un traitement rare et un effet secondaire. La p-value, calculée comme la somme des probabilités des tables aussi ou plus extrêmes que celle observée, permet de conclure à une association

significative si p < α ,souvent α=0.05. Bien que computatoirement intensif pour les grands échantillons (remplacé alors par des approximations ou le χ^2), sa précision en fait la référence pour les études cas-témoins ou les essais cliniques avec données rares. Implémenté couramment dans les logiciels épidémiologiques, il s'applique aussi à des tableaux plus grands via des extensions (test de Fisher-Freeman-Halton) [16].

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \qquad \text{(II.5)}$$

où :

- (a, b, c, d) sont les valeurs des cellules dans le tableau 2x2.

- (n) est le nombre total d'observations.

II.2.6.Test de Friedman

Le test de Friedman est une alternative non paramétrique à l'ANOVA à mesures répétées, spécialement conçue pour analyser des données appariées ou des mesures répétées sur les mêmes sujets lorsque les conditions de normalité ou d'homogénéité des variances ne sont pas remplies. En classant les observations intra-sujets (c'est-à-dire par rang pour chaque individu ou bloc, indépendamment des autres), puis en comparant les rangs moyens entre les différents traitements ou temps de mesure, ce test évalue si au moins un des groupes diffère significativement des autres. La statistique de test (χ^2_F) suit approximativement une distribution du chi-carré avec k-1 degrés de liberté (k étant le nombre de groupes), et une p-value significative (p < α, généralement 0.05) indique un rejet de l'hypothèse nulle (H_0 : égalité des distributions entre groupes) [17].

Particulièrement adapté aux données ordinales ou aux petits échantillons non normaux, le test de Friedman est largement utilisé en recherche clinique (comparaison de plusieurs traitements sur les mêmes patients), en psychologie (évaluations répétées), ou en sciences sociales.

$$Q = \frac{12}{nk(k+1)}\sum_{j=1}^{k} R_j^2 - 3n(k+1) \qquad \text{(II.6)}$$

où :

- n = nombre de blocs/sujets.

- k = nombre de traitements.

- R_j= somme des rangs pour le traitement j.

II.2.7.Test de Grubbs

Test de Grubbs est un test statistique utilisé pour détecter la présence de valeurs aberrantes dans un ensemble de données univariées. Il suppose que les données, à l'exception de la valeur potentiellement aberrante, suivent une distribution normale. Ce test permet d'identifier une seule valeur aberrante à la fois. Si une valeur est détectée comme aberrante, elle peut être retirée, puis le test peut être répété pour détecter d'éventuelles autres valeurs extrêmes. C'est un outil couramment utilisé en analyse de données pour assurer la qualité et la cohérence des ensembles de données [8].

$$G = \frac{max|x_i - \bar{x}|}{s} \quad \text{(II.7)}$$

où :

- $\bar{x}$ = la moyenne de l'échantillon.

- s = l'écart-type.

- x_i= la valeur la plus éloignée de la moyenne.

On compare la statistique G à une valeur critique G_{crit}obtenue à partir de la loi de student (t) selon la formule :

$$G_{crit} = \frac{(n-1)}{\sqrt{n}} \cdot \sqrt{\frac{t^2_{\frac{\alpha}{2n},n-2}}{n-2+t^2_{\frac{\alpha}{2n},n-2}}} \quad \text{(II.8)}$$

Si G >G_{crit} , on rejette H_0 et la valeur est considérée comme aberrante.

II.3. Test de Kolmogorov-Smirnov (Kolmogorov-Smirnov Test)

II.3.1. Idée de base

Le test de Kolmogorov-Smirnov est un test non paramétrique utilisé pour comparer la distribution d'un échantillon à une distribution théorique (test KS à un échantillon) ou pour comparer deux distributions d'échantillons (test KS à deux échantillons). Le test repose sur l'idée

de mesurer la distance maximale entre la fonction de distribution cumulative (CDF) de l'échantillon et la fonction de distribution cumulative théorique (ou entre deux CDF d'échantillons) [18].

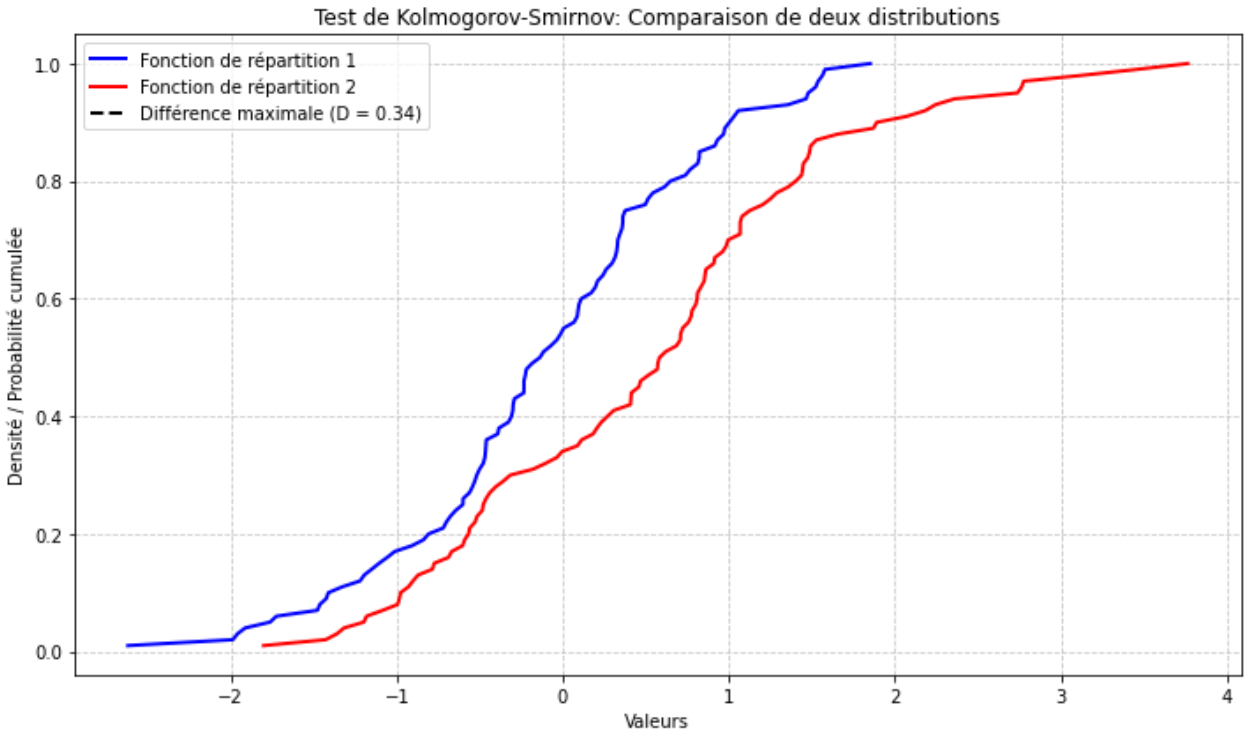


Figure II.2 : Test de Kolmogorov-Smirnov KS

II.3.2. Statistique du test

Soit la fonction de distribution cumulative (CDF) $f(x)$ définie comme la probabilité qu'une variable aléatoire soit inférieure ou égale à (x).

Pour un échantillon, la fonction de distribution cumulative empirique $f_n(x)$ est calculée comme suit :

$$f(x) = \frac{Nombre\ de\ valeurs\ dans\ l'échantillon\ \leq x}{n} \text{(II. 9)}$$

où n est la taille de l'échantillon.

La statistique du test KS, soit D_n est calculée comme la distance maximale entre $f_n(x)$ et $f(x)$:

$$D_n = sup_x|F_n(x) - F(x)| \qquad (II.10)$$

Où

sup_x désigne la valeur maximale de la distance entre les deux fonctions.

II.3.3.Etapes d'application du test

Le test KS s'applique selon deux approches principales. Pour le test à un échantillon, on commence par définir une distribution théorique de référence $F(x)$ (par exemple une loi normale , puis on calcule la fonction de distribution cumulative (CDF) empirique $F_n(x)$ de l'échantillon, qui représente la proportion des observations inférieures ou égales à chaque valeur x. La statistique du test$D(n)$ est obtenue en calculant la distance maximale verticale entre$F_n(x)$ et $F(x)$ (Eq. II.11), qui est ensuite comparée aux valeurs critiques tabulées ou à une approximation asymptotique pour déterminer si l'écart observé est statistiquement significatif (rejet de H_0 si $D(n)$dépasse le seuil critique pour un niveau α donné). Pour le test à deux échantillons, utilisé pour comparer deux ensembles de données empiriques, on calcule les CDF$F_{n1}(x)$et $F_{n2}(x)$ pour chaque échantillon, puis on détermine la distance maximale $D_{n1,n2}$ entre ces deux fonctions (Eq. II.12) ; cette statistique mesure l'écart le plus important entre les distributions cumulatives des deux échantillons, et son interprétation repose également sur des tables critiques ou des méthodes numériques pour évaluer si les différences observées sont significatives au seuil choisi. Dans les deux cas, le test KS est non paramétrique et particulièrement sensible aux différences dans les queues de distribution, mais il nécessite des distributions continues et peut perdre en puissance pour des échantillons de très petite taille.

$$D_n = sup_x|F_n(x) - F(x)|(II.11)$$

$$D_{n1,n2} = sup_x|F_{n1}(x) - F_{n2}(x)|(II.12)$$

II.3.4. Hypothèses du test de Kolmogorov-Smirnov

Le test KS formule des hypothèses distinctes selon qu'il s'agit d'une comparaison à une distribution théorique (test à un échantillon) ou d'une comparaison entre deux jeux de données (test à deux échantillons). Pour le test à un échantillon, l'hypothèse nulle H_0 postule que les données observées suivent parfaitement la distribution théorique spécifiée $F(x)$ (par exemple une loi normale ou exponentielle), tandis que l'hypothèse alternative H_1 conteste cette adéquation,

suggérant que l'échantillon provient d'une distribution différente. Dans le cas du test à deux échantillons, H_0 suppose que les deux ensembles de données sont tirés de la même distribution sous-jacente (sans préciser sa forme), alors que H_1 indique que leurs distributions diffèrent significativement, que ce soit par leur tendance centrale, leur dispersion ou leur forme globale. Ces hypothèses font du test KS un outil particulièrement utile pour détecter des écarts distributionnels sans imposer de contraintes paramétriques. La formulation non paramétrique de H_0 et H_1permet d'appliquer le test KS à des distributions continues arbitraires, mais nécessite une interprétation prudente des résultats : un rejet de H_0 n'indique pas quelle caractéristique de la distribution diffère (moyenne, variance, asymétrie...), ni dans quelle mesure, ce qui peut justifier des analyses complémentaires.

II.3.5. Avantages et inconvénients du test

Le test KS présente plusieurs avantages qui en font un outil statistique largement utilisé. Sa nature non paramétrique lui permet d'être appliqué sans exiger d'hypothèses restrictives sur la forme de la distribution sous-jacente (comme la normalité), ce qui le rend particulièrement utile pour l'analyse exploratoire de données. Contrairement à certains tests comme le χ^2, il préserve l'information en travaillant directement sur les valeurs observées sans nécessiter de regroupement arbitraire en classes. Il montre également une bonne robustesse pour les petits échantillons ($n < 30$), là où les tests paramétriques classiques perdent en fiabilité.

Néanmoins, le test KS comporte certaines limitations. Bien qu'adapté aux petits échantillons, il devient sensible avec de très grands échantillons ($n > 1000$), détectant des différences statistiquement significatives mais potentiellement négligeables en pratique. Sa puissance statistique est généralement inférieure à celle des tests paramétriques (comme le test t) lorsque les conditions de ces derniers sont remplies.

II.3.6. Domaines d'application du test de Kolmogorov-Smirnov

Le test KS trouve des applications variées dans de nombreux domaines scientifiques et techniques :

- **Economie et finance :** il sert à valider des modèles de risque (VaR, crédit) et analyser la normalité des rendements boursiers.
- **Médecine :** l'utilisent pour comparer des paramètres biologiques entre groupes de patients ou valider des hypothèses distributionnelles dans les essais cliniques.

- **Ingénierie :** il contribue au contrôle qualité en vérifiant la stabilité des procédés industriels et la durée de vie des composants.
- **Environnement :** y recourent pour étudier la distribution des polluants ou analyser des données climatiques
- **Informatique :** il est employé pour la détection d'anomalies dans les flux réseau et l'analyse de performances algorithmiques,
- **Physique :** pour valider des modèles théoriques contre des données expérimentales.

II.4.Conclusion

Dans ce chapitre nous avons vu l'utilité des tests statistiques pour identifier des écarts significatifs entre des données par comparaison de leurs CDF, avec un focus particulier sur le test de Kolmogorov-Smirnov, reconnu pour sa puissance et sa sensibilité aux variations distributionnelles. Considérant que les cyber-attaques, notamment les attaques par déni de service (DOS/DDOS), se manifestent par des anomalies marquées dans les schémas de trafic réseau, dans le chapitre III nous évaluerons les performances du test KS dans la détection de ces attaques. Nous analyserons sa capacité à discriminer différents types de menaces via une approche non paramétrique, et sous différents scénarios de trafic réseau malveillants, fournis par la base DARPA 99.

Chapitre III

Simulations et interprétations

Chapitre III

Simulations et interprétations

III.1. Introduction :

A travers la simulation sous python, nous investiguons, dans ce chapitre, l'utilité du test de Kolmogorov-Smirnov pour la détection des cyber-attaques dans un réseau IP.

Pour cela, nous avons mis en place une procédure de détection non-paramétrique pour la détection des attaques DOS de types : TCP SYN flood, UDP flood, Smurf en utilisant le trafic réseau fournit par la base de données DARPA99, comportant à la fois des traces normales et d'autres affectées différents scénarios de ces attaques.

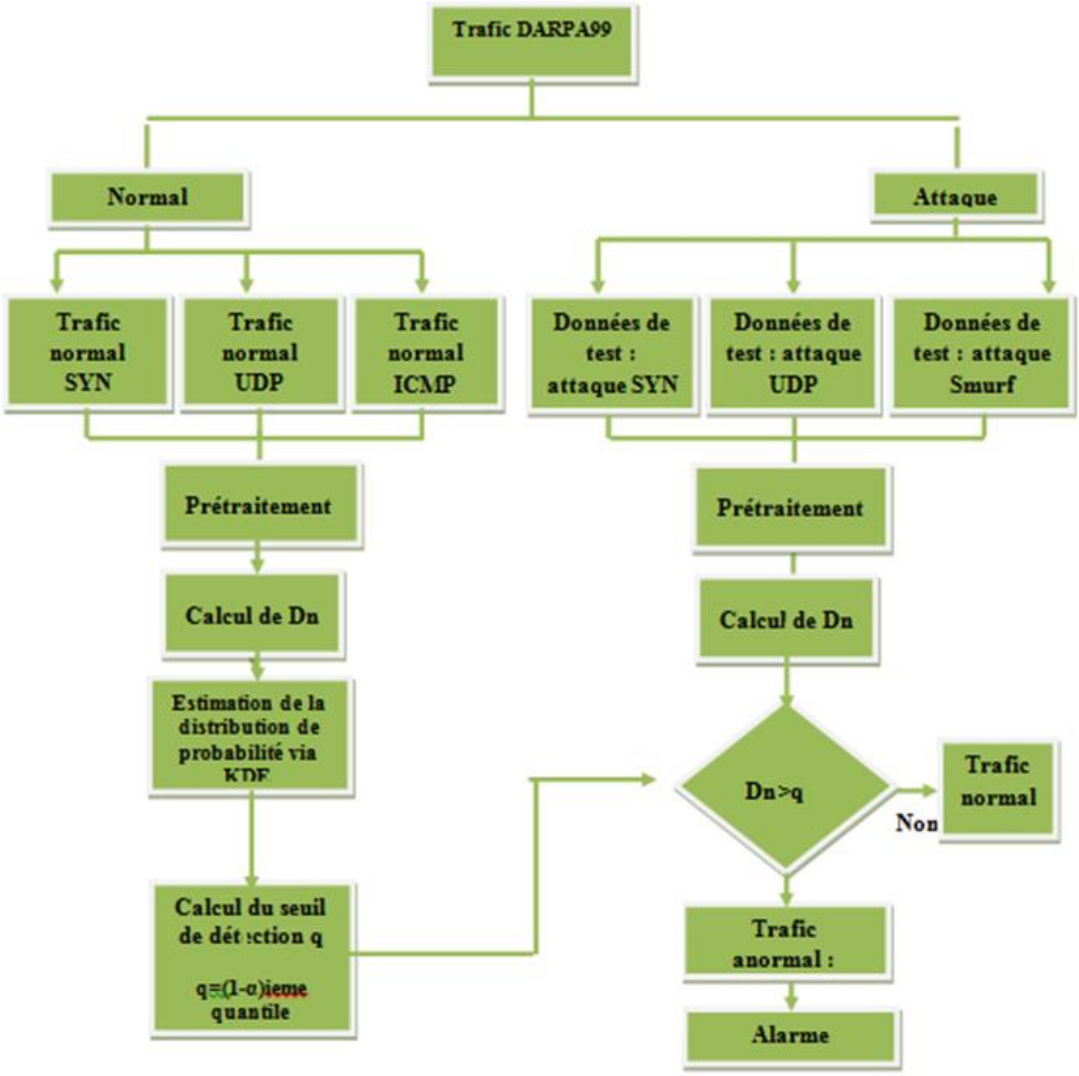
III.2. Détection des cyber-attaques via le test de Kolmogorov-Smirnov :

Pour détecter les attaques DOS comme TCP SYN flood, Smurf et UDP flood par le test de Kolmogorov-Smirnov, nous avons mis en place la procédure suivante :

- Extraction des données d'apprentissage, c à d le trafic normal en terme de structure de données ciblée (segments SYN, datagrammes UDP et messages ICMP ECHO REPLY).
- Prétraitement et normalisation des données d'apprentissage.
- Calcul la statistique *Dn* de KS des séquences de données d'apprentissages issues de l'étape précédente.
- Estimation de la distribution de probabilité reliée aux données d'apprentissage en utilisant la méthode KDE avec un kernel gaussien et la largeur de la bande de lissage optimal.
- Calcul du seuil de détection q comme le (1-α)ieme quantille de la distribution de probabilité estimée.

- Extraction de données de test c à d le trafic avec attaques en terme de structure de données ciblée
- Prétraitement et normalisation des données de test
- Calcul la statistique *Dnt* de KS des séquences de données de test issues de l'étape précédente.
- Comparaison des séquences de la statistique Dnt au seuil de détection q
- Si la statistique t est inférieure à q, alors le trafic est considéré comme normal, et ne contient pas d'attaques DOS ou DDOS.
- Sinon, si la statistique dépasse le seul q, le trafic, dans ce cas, a un comportement anormal, une alarme de détection d'attaques SYN flood est déclenchée.

La figure III.1 récapitule la méthode d'utilisation du test KS pour une détection non-paramétrique des attaques DOS/DDOS.



.**Figure III.1:**Procédure générale de détection des cyber-attaques par le test KS

III.3. La base de trafic DARPA99 [19] [20]

Pour valider notre procédure de détection non-paramétrique des cyber-attaques, basée sur le test KS, nous utiliserons la base de données DARPA99. DARPA 99 (appelé aussi IDEVAL : Intrusion Detection Evaluation) s'inscrit parmi les bases de données les plus importantes et les plus utilisées pour l'évaluation des systèmes de détection d'intrusion. Elle a été construite par Lincoln Laboratory du Massachusetts Institute of Technology (MIT) sous le support de la DARPA et de l'Air Force Research Laboratory (AFRL). Elle représente le trafic capturé dans un réseau réel similaire au réseau réel d'une base militaire de l'Air Force, connectée à Internet.

La figure III.2 montre la topologie du réseau mis en place. Les machines à gauche simulent le réseau de la base militaire (le réseau intérieur), tandis que les machines à droite simulent l'Internet (le réseau extérieur). Un routeur Cisco connecte les deux réseaux.

Les victimes sont des serveurs de la base de l'Air Force qui font les cibles des différents types d'attaques. Pascal exécutant Solaris 2.5, un shell ou un serveur de connexion fournit des services telnet, SMTP, SSH et FTP. Zeno, fonctionne avec SunOS 4.1.4, est un serveur de fichiers, sendmail et permettait le partage de fichiers entre les utilisateurs via un serveur FTP. Marx, travaille avec RedHat 5.0, est le serveur Web, il sert comme page d'accueil à la fois pour Internet et pour une utilisation interne, un serveur Web Apache a été utilisé. Hume, c'est un serveur Windows NT 4.0, était équipé avec IIS (Internet Information Server) et héberge des serveurs FTP, gopher, Web et plusieurs autres utilitaires incluent un serveur de messagerie, appelé MailSrv. Kant utilise un Windows98. Le serveur Web, Aesop, exécutant RedHat 5.0 est le serveur Web Internet et semble être des milliers de serveurs Web Internet individuels. Les hôtes virtuels internes et externes sont utilisés pour usurper différentes adresses IP.

Deux postes de travail, l'un avec Linux RedHat 5.0 et l'autre avec Windows NT4.0, sont utilisés comme attaquants internes. Trois autres postes de travail, deux avec Linux RedHat 5.2 et un avec Windows NT 4.0, sont utilisés comme attaquants extérieurs. Plusieurs types d'attaques générées dans ces données, y compris les attaques de déni de service.

Pour collecter le trafic réseau, deux renifleurs ont été installés sur le réseau. Locke, le renifleur interne, exécute Solaris 2.6 et utilisé pour capturer le trafic réseau sur le réseau intérieur. Le renifleur extérieur, Solomon, exécute également Solaris 2.6 et est utilisé pour capturer le trafic réseau (entrant/sortant) vers et depuis la base de l'Air Force. Pour les deux renifleurs, UNIX TCPDUMP a été utilisé pour collecter le trafic. Cinq semaines de données ont été collectées.

Chaque semaine comprenait cinq jours, du lundi au vendredi, avec 22 heures par jour, de 8h00 à 6h00.

Plus de 8Goctets de trafic réseau (entrant et sortant) sous forme de fichiers compressés TCPDUMP a été enregistré. Pour faciliter l'évaluation des IDS à base d'anomalies, le trafic collecté contient trois semaines de données d'apprentissage (training data) séparées de deux semaines de données de test. Les première et troisième semaines de données d'apprentissage sont totalement exemptes d'attaques, la seconde en inclue certaines. Les deux semaines de données de test (semaines quatre et cinq) contiennent différentes catégories d'attaques.

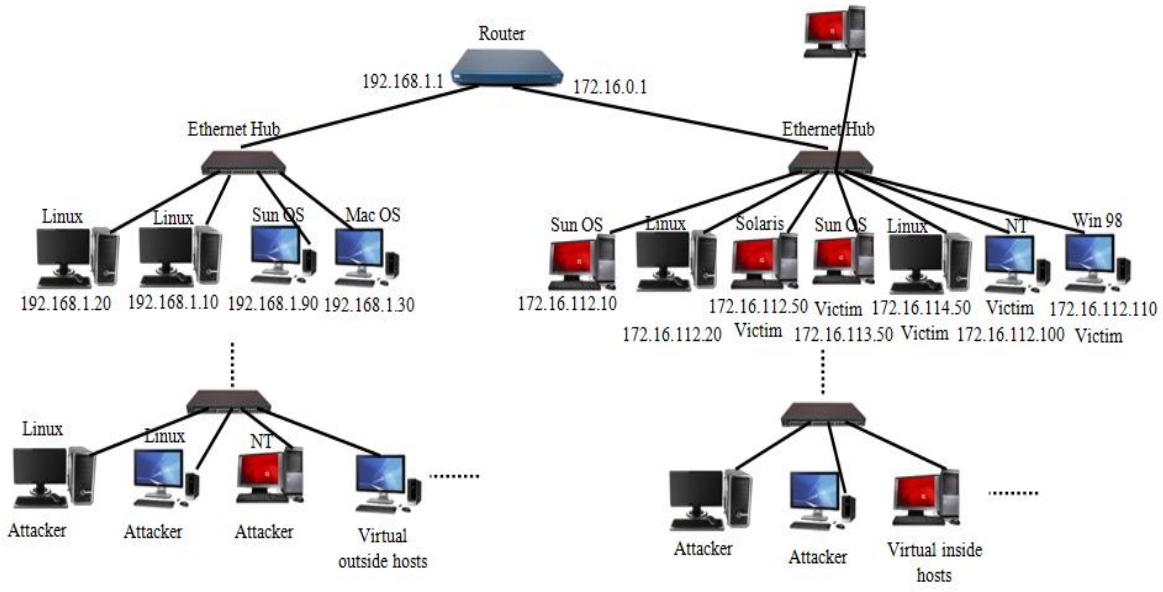


Figure III.2 : La topologie de réseau utilisé par DARPA 99 [19]

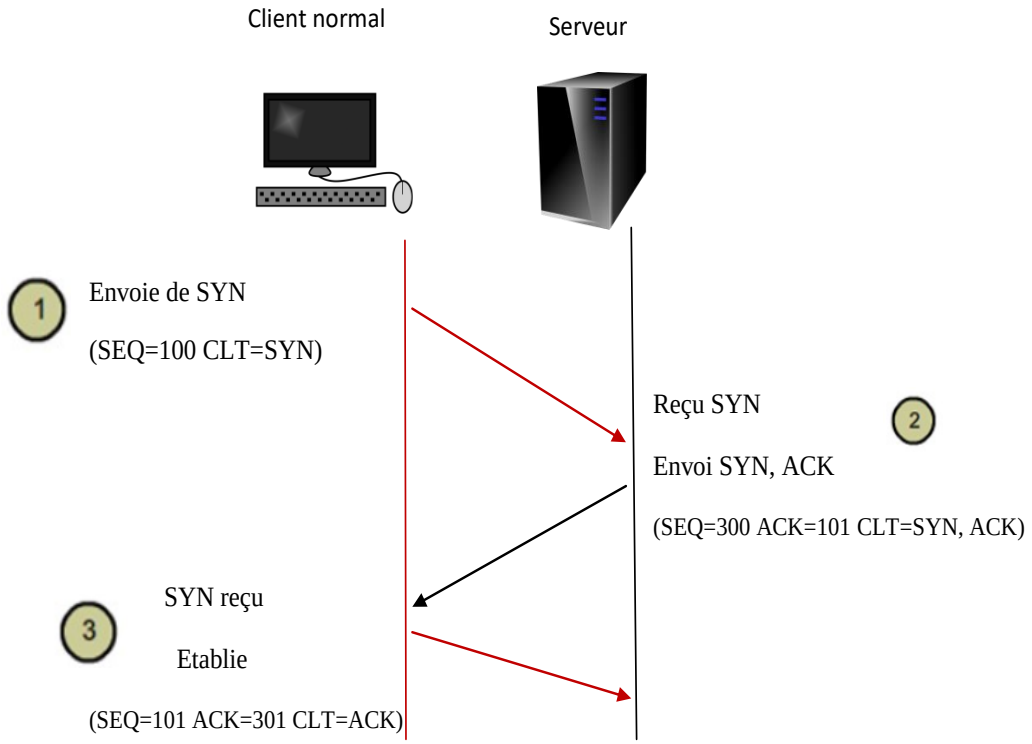
III.4. Résultats et interprétations

III.4.1. Détection des attaques SYN flood:

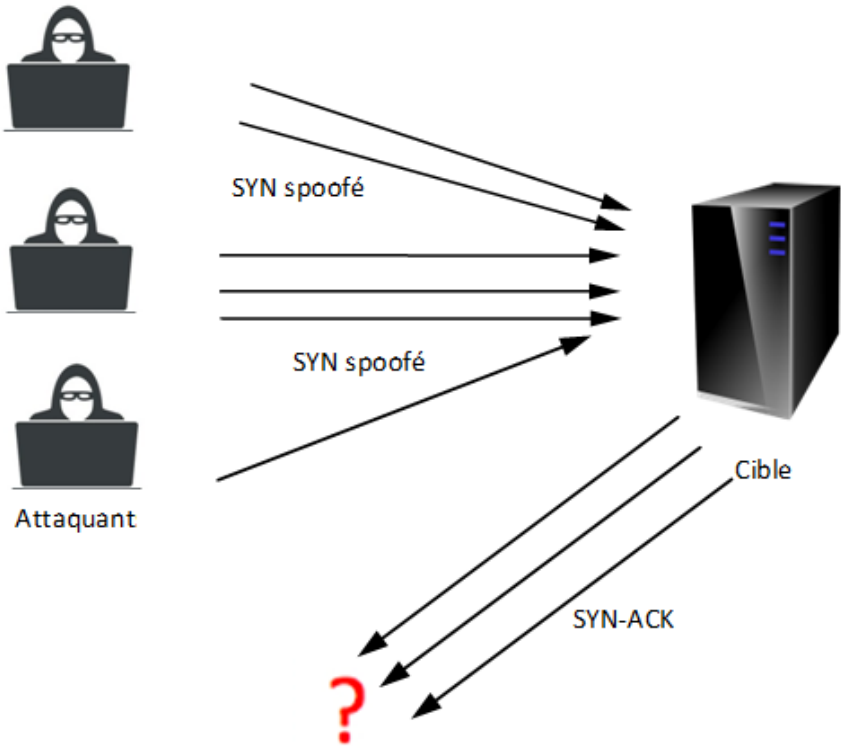
Dans cette première partie, nous avons utilisé la méthode de détection des cyber-attaques détaillée ci-dessus pour détecter les attaques SYN flood.

Dans une attaque SYN flood, l'attaquant envoie des segments SYN répétés à chaque port sur le serveur ciblé, souvent en utilisant une fausse adresse IP. Le serveur, inconscient de l'attaque, reçoit plusieurs demandes apparemment légitimes pour établir une communication. Il répond à chaque tentative avec un segment SYN-ACK de chaque port ouvert. L'utilisateur malveillant n'envoie pas le ACK attendu ou, si l'adresse IP est usurpée, ne reçoit jamais le SYN-

ACK en premier lieu. Dans les deux cas, le serveur attaqué attendra l'accusé de réception de son paquet SYN-ACK pendant un certain temps. Pendant ce temps, le serveur ne peut pas fermer la connexion en envoyant un segment RST et la connexion reste ouverte. Avant que la connexion puisse expirer, un autre segment SYN arrivera. Cela laisse un nombre de plus en plus grand de connexions semi-ouvertes, les attaques SYN flood sont également appelées attaques «semi-ouvertes». Au fur et à mesure que les tables de débordement de connexion du serveur se remplissent, le service offert aux clients légitimes sera refusé et le serveur risque même de ne pas fonctionner correctement ou de tomber en panne [21].



a) Connexion TCP normale



b) Attaque SYN flood

Figure III.3 : Principe de l'attaque SYN flood

Les données de test DARPA99 contiennent des attaques SYN flood dans les jours 1 et 2 de la semaine 5 W5D1 et W5D2.

La figure III.4 représente l'évolution du nombre de segments SYN durant le jour W5D2. On observe deux pics marqués dans l'activité du trafic, indiquant une augmentation anormale du nombre de segments TCP par rapport aux périodes normales. Ce comportement irrégulier du réseau suggère une activité suspecte, probablement liée à des tentatives d'attaque de type SYN flood.

La figure III.5 représente le résultat de détection de ces attaques. Après l'application du test de Kolmogorov-Smirnov sur les données, la statistique KS met en évidence la différence maximale entre la distribution du trafic W5D2 et la distribution de trafic normal de référence. Ainsi, toutes les valeurs qui dépassent le seuil de détection établie considérées comme des comportements anormaux ou malveillants. Dans notre cas, les deux pics identifiés dépassent nettement ce seuil, révélant ainsi la présence des attaques SYN flood.

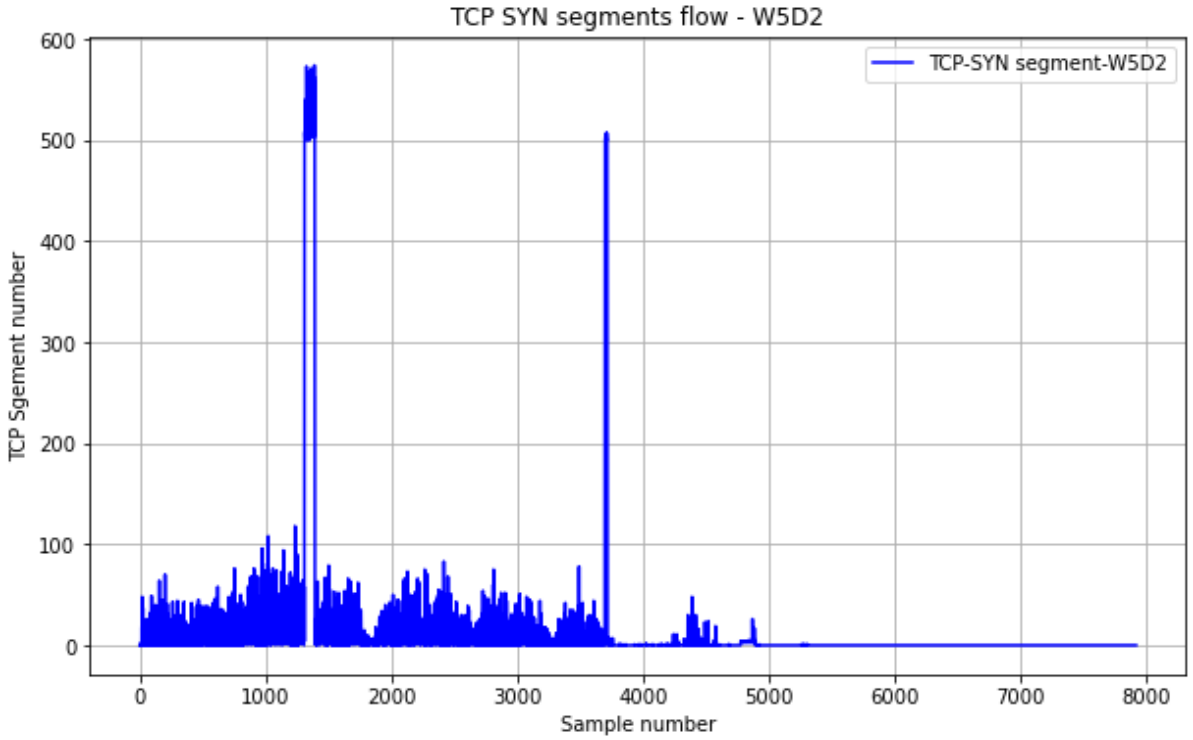


Figure III.4 : Evolution du nombre des segments SYN Durant le trafic W5D2

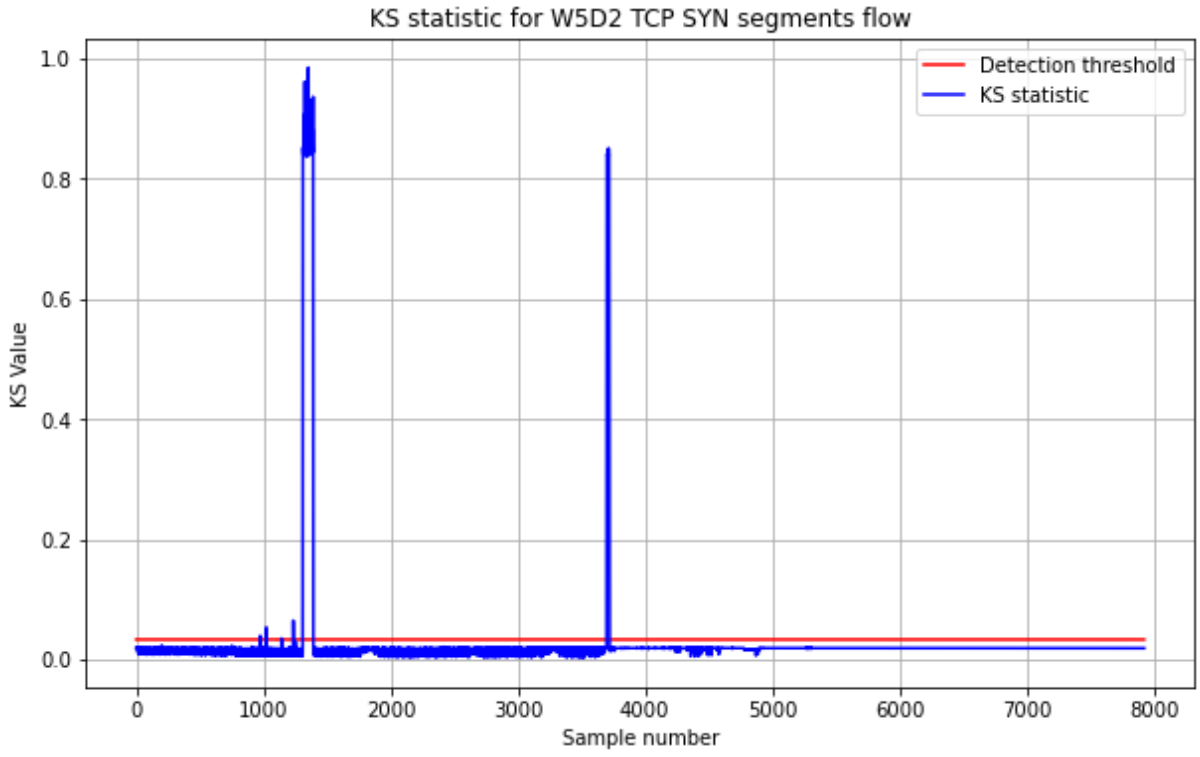


Figure III.5 : Résultat de détection en présence des attaques SYN (W5D2)

Nous avons répété la même expérience lors de la semaine 5, jour 1 (W5D1) en surveillant le flux des segments SYN de manière identique. Durant la période d'observation, trois pics distincts ont été relevés. Ces trois pics traduisent une activité anormale récurrente, caractéristique d'attaques SYN flood visant à saturer le service. La figure III.6 représente l'évolution du nombre de segments SYN durant le jour W5D1.

La figure III.7 représente le résultat de détection de ces attaques. Ces résultats confirment la présence d'attaques SYN flood répétées et de différentes intensités ciblant le réseau. Ces attaques se manifestent par une augmentation significative de la statistique KS durant les attaques.

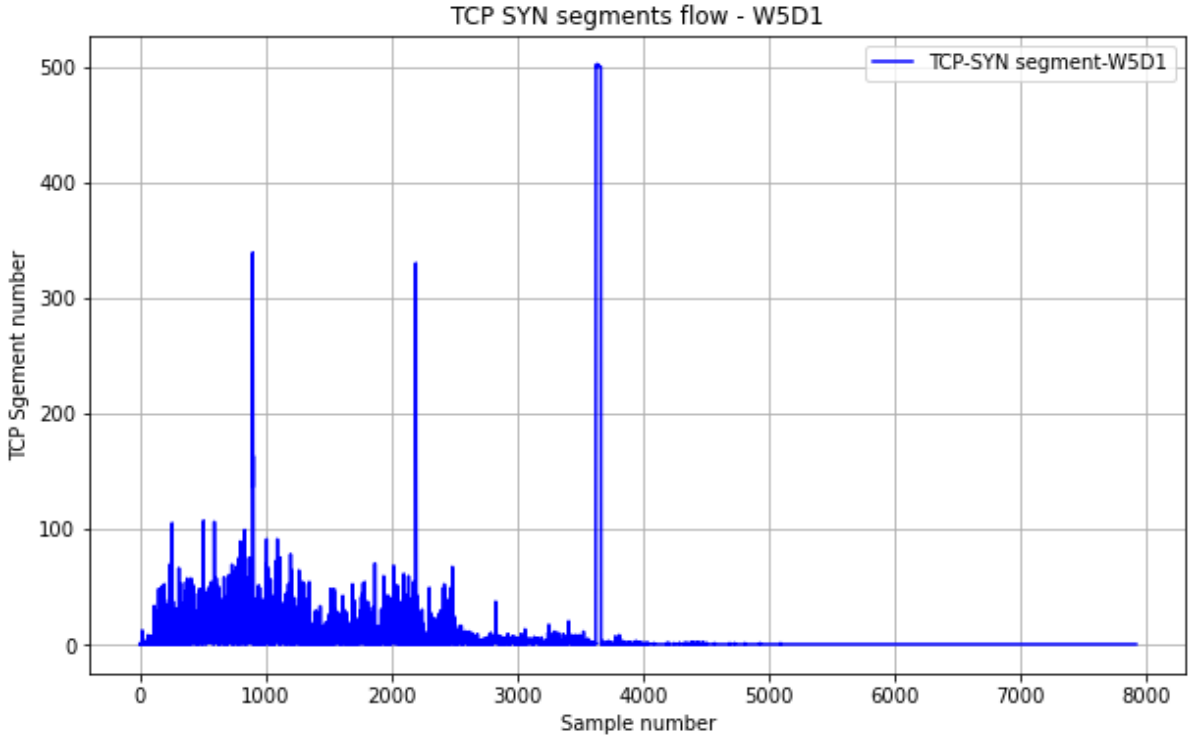


Figure III.6 : Evolution du nombre des segments SYN Durant le trafic W5D1

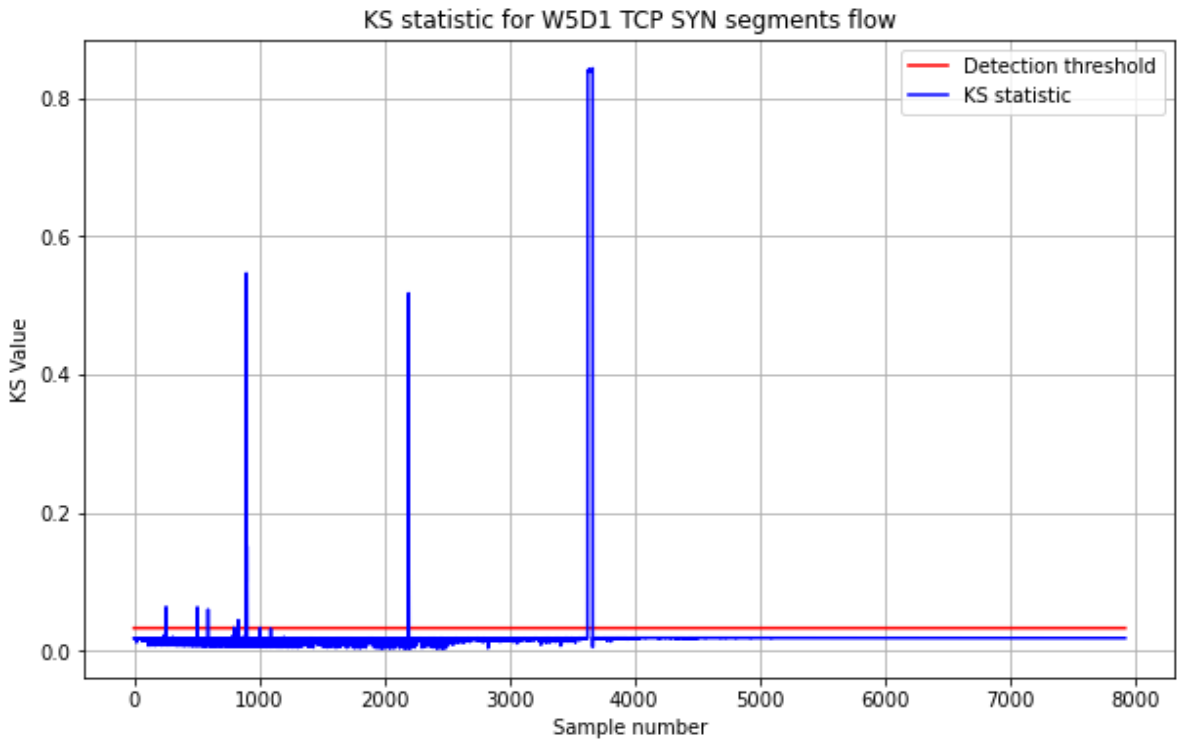


Figure III.7 : Résultat de détection en présence des attaques SYN (W5D1)

III.4.2. Détection des attaques Smurf

Dans cette partie, nous avons validons notre procédure de détection en présence des attaques Smurf.

Smurf [22] est une attaque DDoS sur la couche réseau, nommée d'après le programme malveillant DDoS smurf qui permet son exécution. Ces attaques ressemblent un peu aux inondations de ping, car les deux sont effectués en envoyant une série de paquets de requêtes ICMP ECHO. Contrairement au flot de ping classique, il s'agit d'un vecteur d'attaque par amplification qui augmente son potentiel de dommages en exploitant les caractéristiques des réseaux de diffusion.

La vérification de l'accessibilité d'un ordinateur spécifié se fait en envoyant une demande ICMP ECHO (ping) et cela déclenche une réponse ICMP automatique, en passant par un réseau de diffusion IP, la requête ping est envoyée à chaque hôte, invitant chacun des destinataires à répondre. Dans le cas d'une attaque Smurf, les logiciels malveillants Smurf permettent de générer une fausse requête Echo contenant une adresse IP source usurpée correspond à l'adresse du serveur cible.

- La demande est envoyée à un réseau de diffusion IP intermédiaire.
- La demande est transmise à tous les hôtes du réseau.
- Chaque hôte envoie une réponse ICMP à l'adresse source usurpée.
- Avec suffisamment de réponses ICMP transférées, le serveur cible est arrêté.

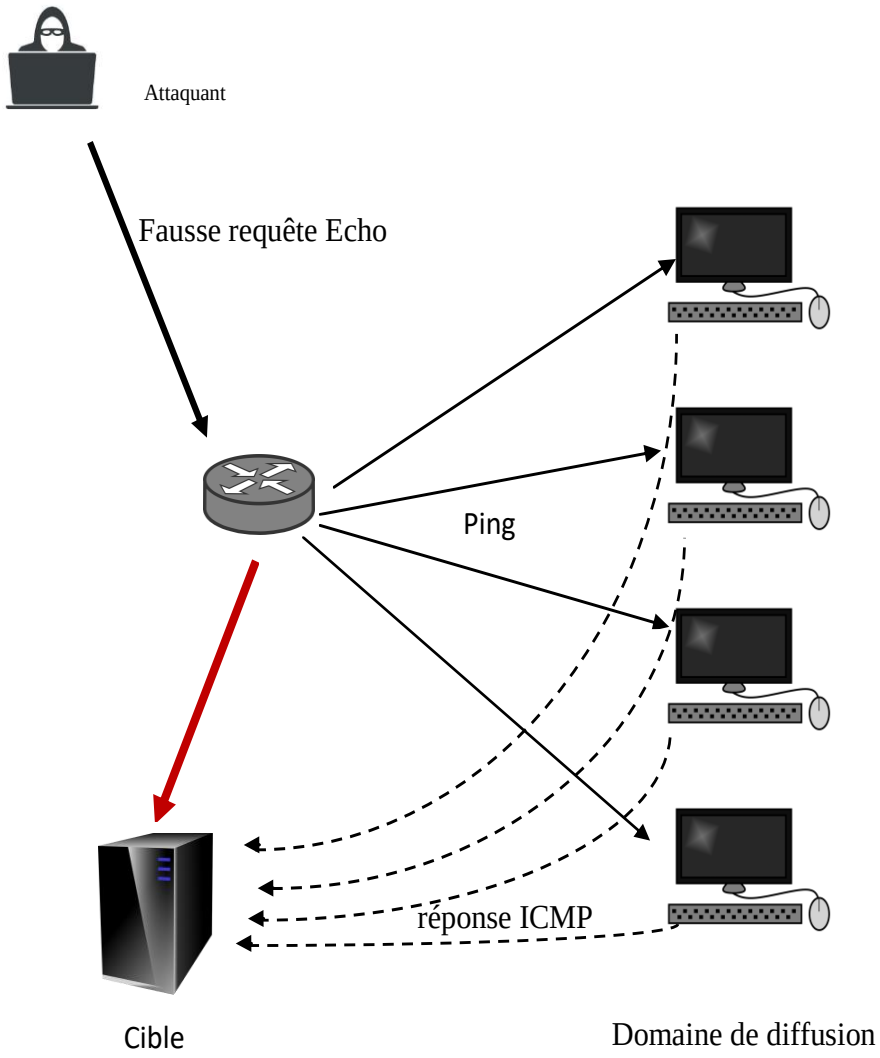


Figure III.8 : Principe de l'attaque Smurf

Les données de test DARPA99 contiennent des Smurf les jours 1 et 5 de la semaine 4 W4D1 et W4D5.

La figure III.9 représente le flux des messages ICMP ECHO REPLY durant le jour W4D1.Durant presque toute la période, le nombre de messages ICMP ECHO REPLY reste

proche de zéro, indiquant une activité réseau normale. Toutefois, un pic extrêmement élevé atteignant plus de 27 000 réponses ICMP. Ce comportement est clairement anormal et constitue un événement isolé mais critique, suggérant une attaque potentielle de type Smurf, matérialisée par ce grand nombre de requêtes envoyées dont le but de surcharger la cible.

La figure III.10 représente le résultat de détection de ces attaques. On remarque que la statistique KS est restée stable et proche de zéro, sauf autour de la période d'attaque où un pic massif a dépassé clairement le seuil de détection annonçant ainsi la présence des attaques Smurf.

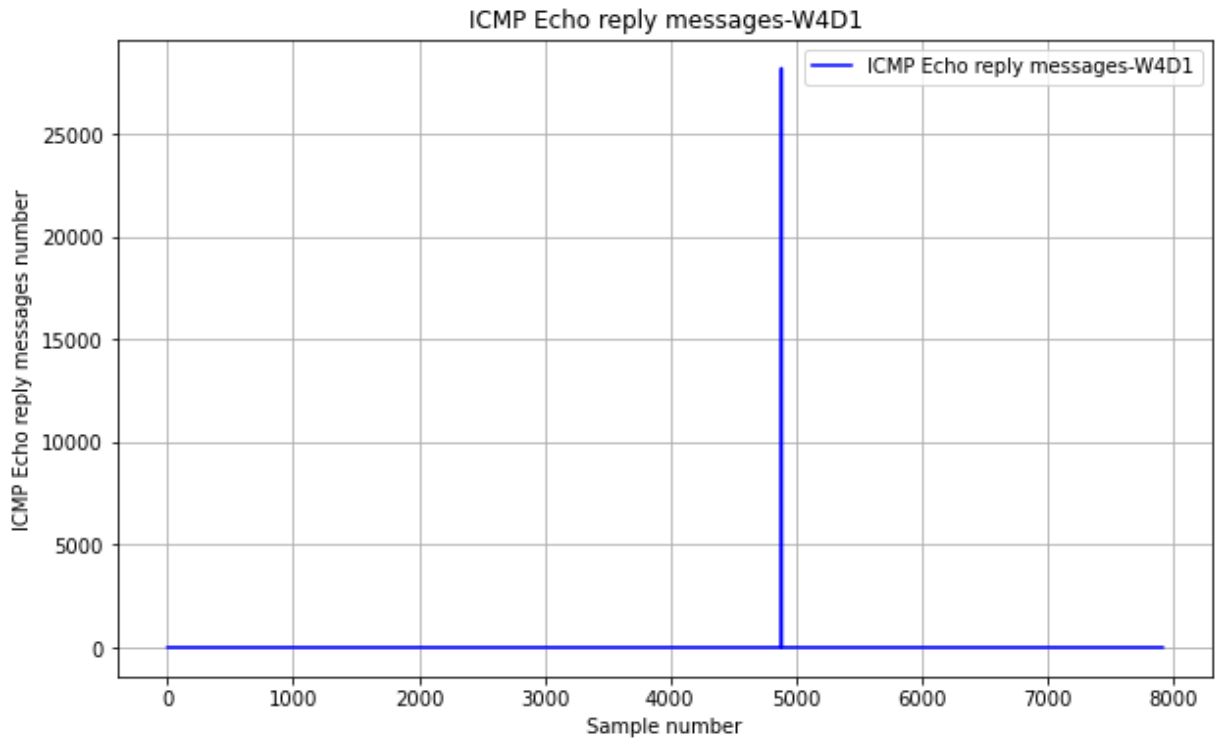


Figure III.9 : Evolution du nombre des messages ICMP ECHO REPLY Durant le trafic W4D1

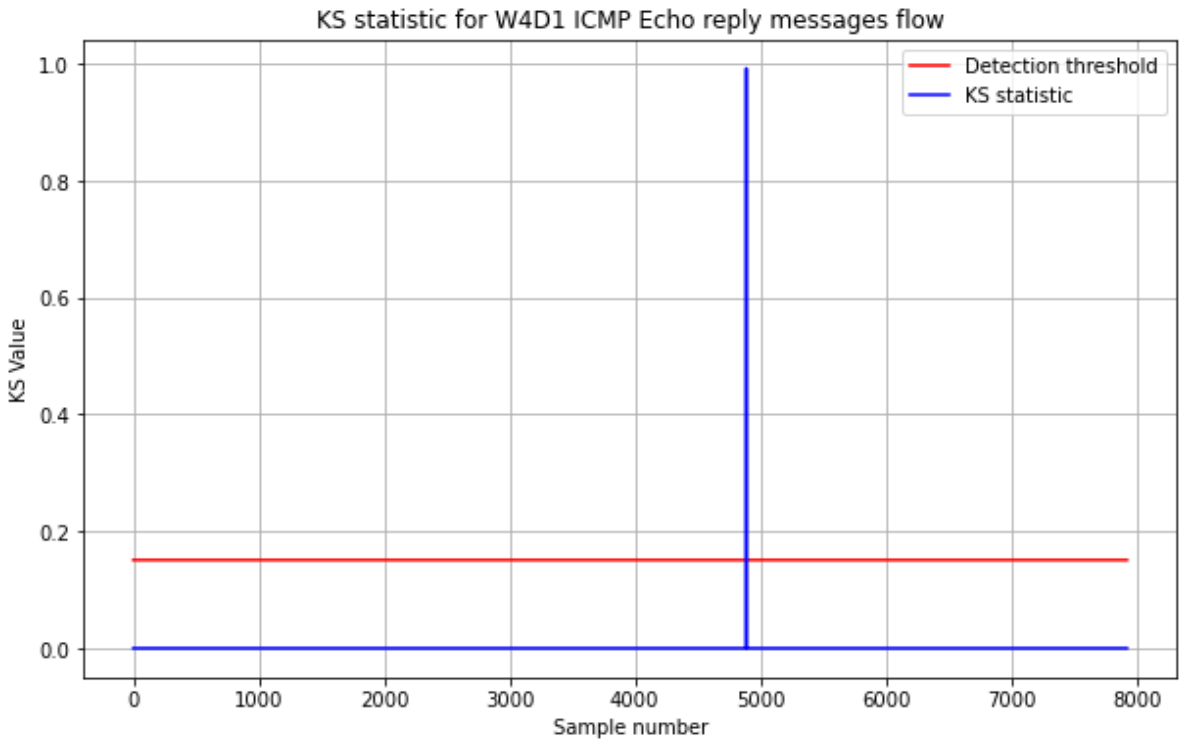


Figure III.10 : Résultat de détection en présence des attaques Smurf (W4D1)

Nous avons répété le même travail lors de la semaine 4, jour 5 (W4D5) en surveillant de le flux des messages ICMP ECHO REPLY de la même manière. Le flux correspond est représenté sur la figure III.11. On constate que l'activité est restée stable et faible pendant une longue période. Cependant, sur une courte durée d'environ 250 secondes, un pic brutal similaire a été observé, reflétant la récurrence de cet événement anormal.

Le résultat de détection est illustré sur la figure III.12. Ce résultat montre bien une attaque délibérée de type Smurf ciblant le réseau sur de courte période avec un volume très important de requêtes ICMP ECHO REPLY.

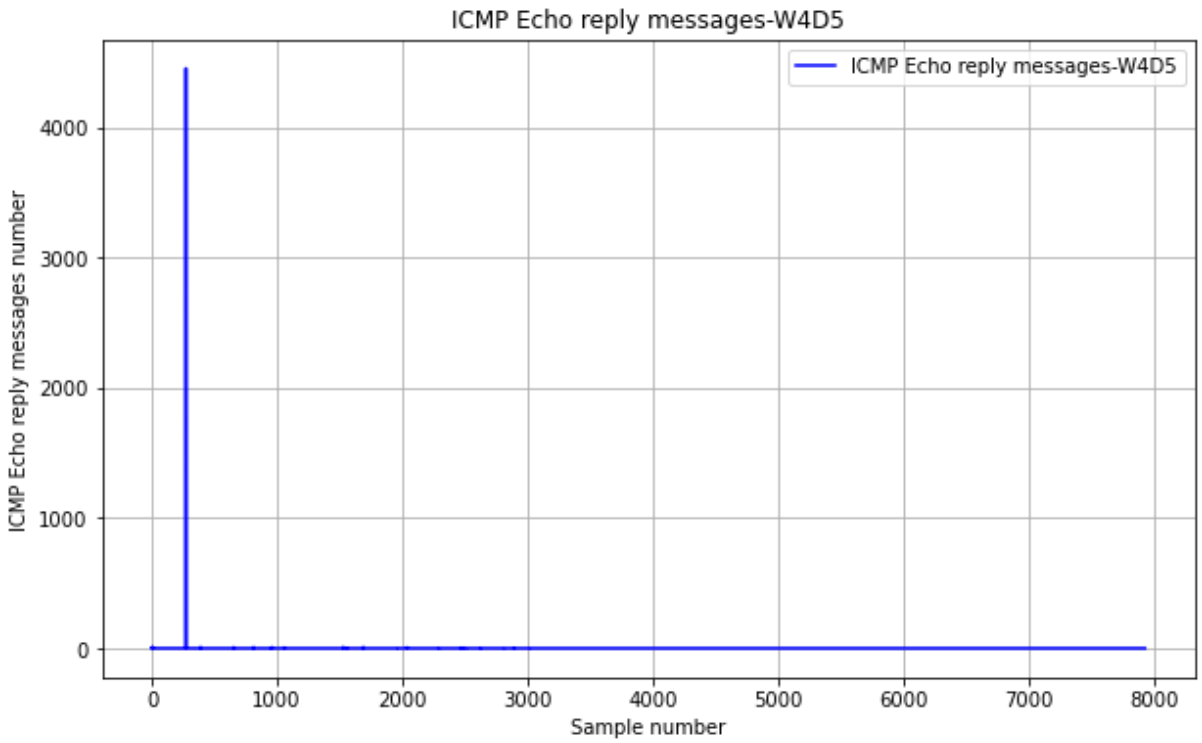


Figure III.11 : Evolution du nombre des messages ICMP ECHO REPLY Durant le trafic W4D5

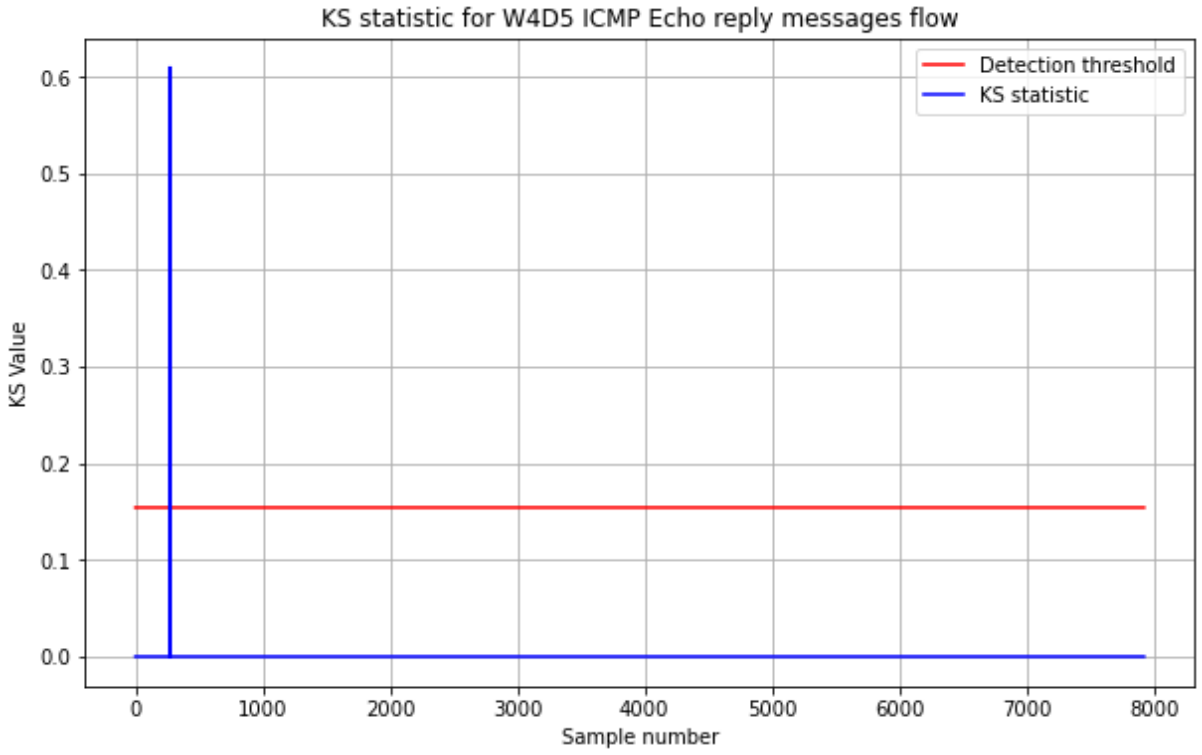


Figure III.12 : Résultat de détection en présence des attaques Smurf (W4D5)

III.4.3. Détection des attaques UDP flood

Ici, nous évaluons l'efficacité de notre procédure de détection non paramétrique face aux attaques UDP flood.

UDP flood [23] est un type d'attaque DOS dans lequel l'attaquant inonde des ports aléatoires sur l'hôte cible avec des paquets IP contenant des datagrammes UDP. L'hôte destinataire vérifie les applications associées à ces datagrammes et, sans en trouver aucune, renvoie un message « Destination inaccessible ». A mesure que de plus en plus de paquets UDP sont reçus et traites, le système devient submerge et ne répond plus aux autres clients.

Dans le cadre d'une attaque par inondation UDP, l'attaquant peut également usurper l'adresse IP des paquets, à la fois pour s'assurer que les paquets ICMP renvoyés n'atteignent pas leur hôte et pour rendre l'attaque anonyme.

L'utilisation de plusieurs machines permettra de classer cette attaque dans la catégorie de menace DDoS. Avec cette attaque, le but du délinquant est de maitriser les pare-feu et d'autres composants des infrastructures de réseau.

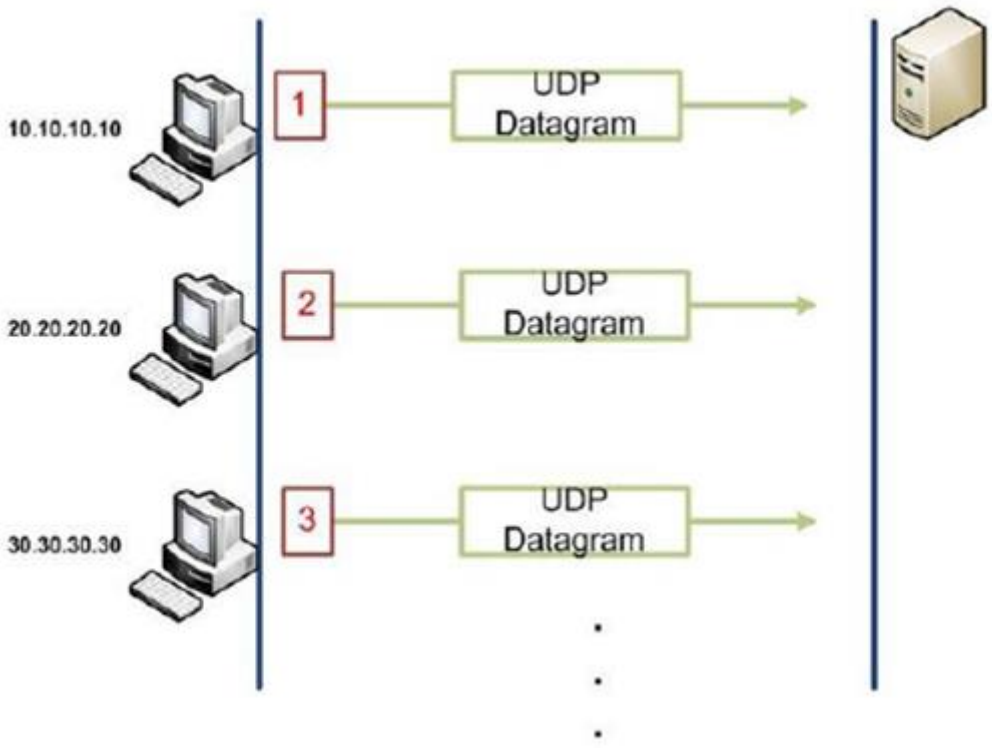


Figure III.13 : Principe de l'attaque UDP flood

Lors de cette partie portant sur le flux des datagrammes UDP pendant le jour 1 de la semaine 4 W4D5 des données de test DARPA99,deux attaques distinctes ont été observées : ces pics reflètent une activité intense du trafic UDP. La figure III.14 représente le fluxdes datagrammes UDP correspondants.

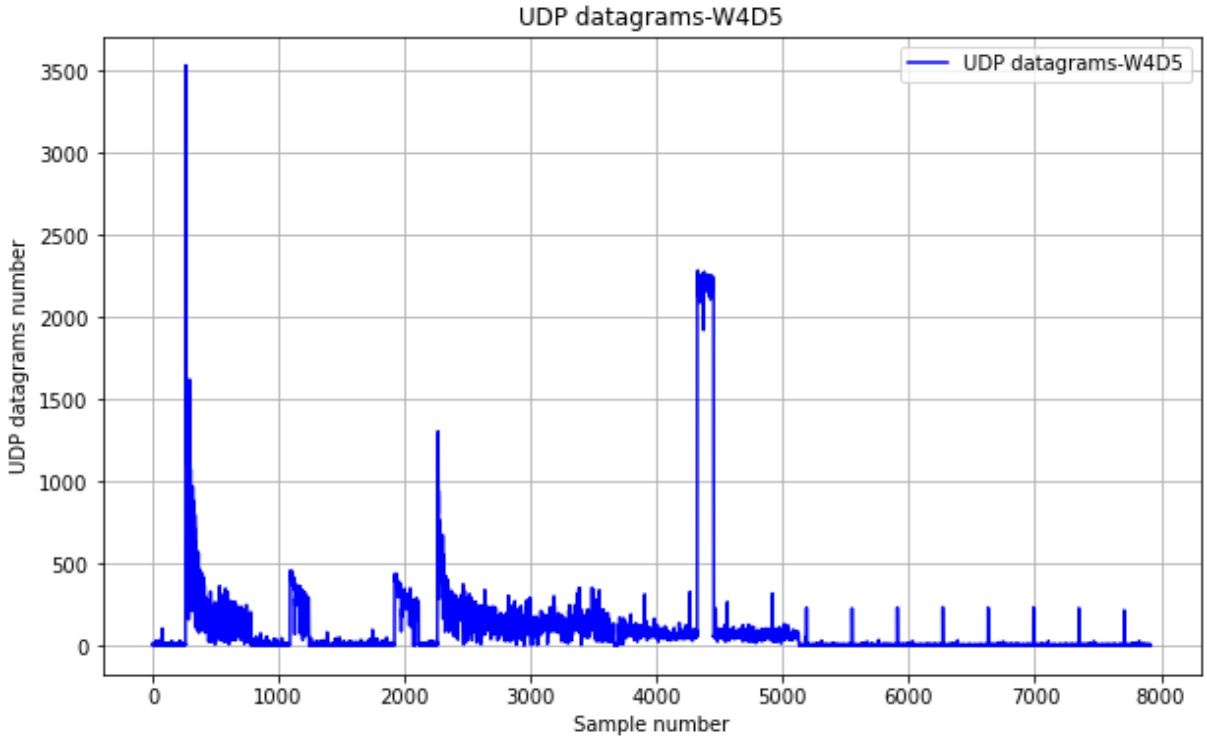


Figure III.14 : Evolution du nombre des datagrammes UDP durant le trafic W4D5

Après avoir appliqué le test de Kolmogorov-Smirnov, le résultat de détection est illustré sur la figure III.15. L'analysé des écarts (séquences de statistiques KS) entre la distribution observée et la distribution de référence, montre que ces deux événements correspondent bien à deux attaques distinctes. Cela confirme que l'activité détectée n'est pas normale, mais représente des attaques délibérées faisant appel à une charge importante de datagrammes UDP.

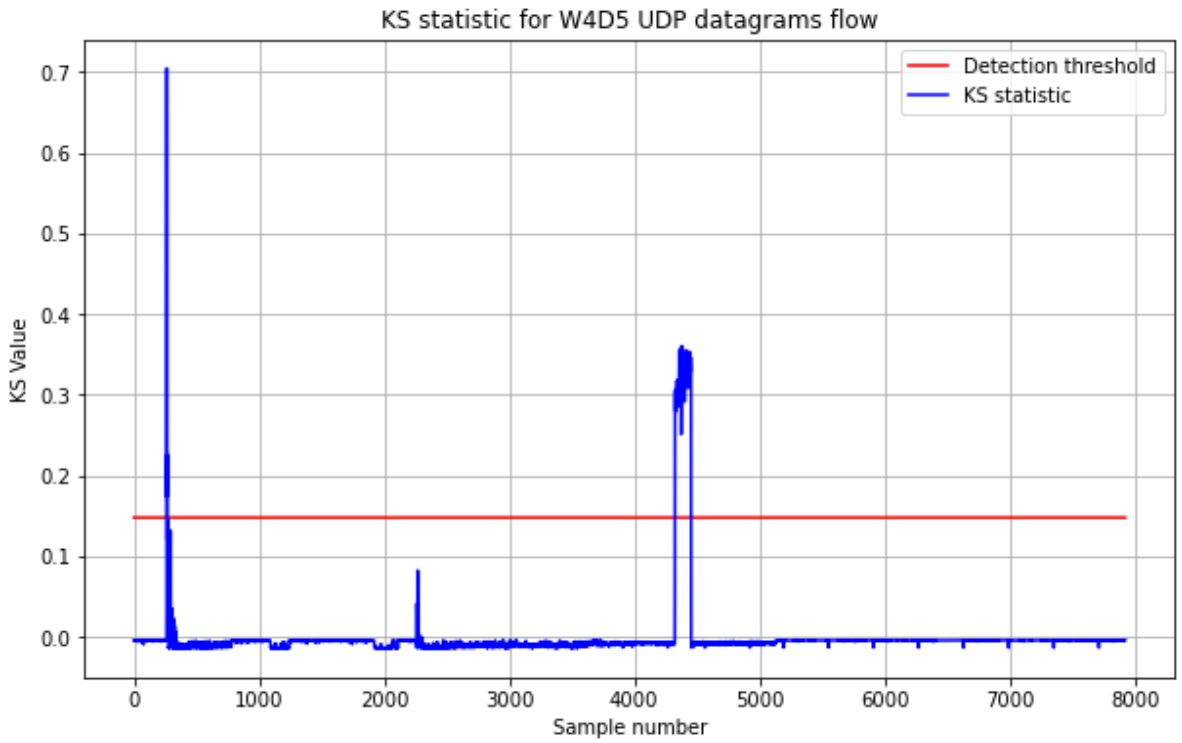


Figure III.15 : Résultat de détection en présence des attaques UDP flood (W4D5)

III.4. Conclusion :

Dans ce chapitre, nous avons mis en œuvre une approche de détection des cyber-attaques de type DoS et DDoS en nous se basant sur une analyse statistique du trafic réseau à travers le test de Kolmogorov-Smirnov. A l'aide de données représentant différents types de trafic (normal et malveillant), nous avons comparé leurs distributions statistiques et observer les déviations significatifs qui apparaissent lors des attaques.

Les résultats obtenus confirment l'efficacité du test KS dans l'identification des variations anormales du comportement réseau. Grâce à cette méthode, il est possible de détecter précocement des attaques telles que TCP SYN flood, Smurf et UDP, en se basant sur l'analyse des caractéristiques du trafic.

Conclusion Générale

Conclusion Générale

Dans un paysage numérique marqué par l'évolution rapide des cybermenaces et la complexification des attaques, la sécurisation des infrastructures IP constitue une priorité absolue. Ce travail s'inscrit dans une démarche d'amélioration des mécanismes de détection d'intrusion à travers le développement d'approches statistiques innovantes et rigoureuses.

Ce mémoire porte sur l'application des tests statistiques, et plus particulièrement du test de Kolmogorov-Smirnov (KS), pour la détection des cyber-attaques de types SYN flood, UDP flood et Smurf.

Notre procédure de détection repose sur l'utilisation du test de Kolmogorov-Smirnov (KS) comme indicateur principal des changements dans la distribution du trafic réseau IP, en comparant en les caractéristiques statistiques du trafic courant avec un profil de référence établi à partir de données normales. Pour distinguer efficacement les comportements légitimes des attaques DOS/DDOS, nous avons implémenté une approche non-paramétrique basée sur l'estimation par noyaux (KDE) des distributions de référence, avec des seuils de détection correspondant au (1-α)ieme quantille de cette distribution estimée. Concrètement, le système déclenche une alerte d'attaque lorsque les valeurs successives des statistiques KS, calculées entre les fenêtres d'observation du trafic et le modèle de référence, dépassent de manière significative les seuils prédéfinis, ce qui indique une divergence statistiquement pertinente entre les distributions comparées. Cette méthode permet une détection sensible des anomalies tout en maintenant un faible taux de faux positifs grâce à l'adaptation dynamique des seuils aux caractéristiques spécifiques du réseau surveillé.

Les performances de notre procédure de détection ont été évaluées à travers des simulations Python, en utilisant le trafic IP de la base de données DARPA99, une référence publique largement utilisée dans la recherche en cybersécurité.

L’analyse des résultats a montré sa capacité à identifier avec précision les anomalies dans le trafic, tout en maintenant un faible taux de fausses alertes, confirmant ainsi le potentiel du test KS comme outil de détection dans les systèmes de surveillance réseau.

Ces résultats ouvrent des perspectives importantes pour l'amélioration des systèmes de détection, notamment par : l'intégration multivariables de paramètres réseau, l'extension à d'autres types d'attaques et l'optimisation dynamique des seuils de détection.

Références

Références

[1] Kurose, J. F., & Ross, K. W ,Computer Networking: A Top-Down Approach. Pearson, 2017.

[2] Stallings, W, Network Security Essentials: Applications and Standards. Pearson, 2013.

[3] RFC 791 : Internet Protocol (IETF).

[4] RFC 2401 : Security Architecture for IP (IETF).

[5] Tanenbaum, A. S., & Wetherall, D, Computer Networks. Prentice Hall, 2011.

[6] Hollander, M., Wolfe, D. A., & Chicken, E. , Nonparametric statistical methods. John Wiley & Sons,2013.

[7] Behrouz A. Forouzan, Data Communications and Networking, 5th Edition, McGraw-Hill Education, 2013.

[8] Nahm, F. S. ,Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean journal of anesthesiology, 69(1), 8-14, (2016).

[9] Conover, W. J. "Practical Nonparametric Statistics. 3rd EditionWiley." New York, NY 584,1999.

[10] Grubbs, F. E. Sample Criteria for Testing Outlying Observations;. Revue : Annals of Mathematical Statistics, 21(1), 27-58.

[11] Lehmann, E. L., Romano, J. P ,Testing Statistical Hypotheses, vol3, Springer, 1986.

[12] MacFarland, Thomas W., et al. , Mann–whitney u test : Introduction to nonparametric statistics for the biological sciences using R, 103-132, 2016.

[13] Okoye, K and Hosseini, S., Wilcoxon Statistics in R: Signed-Rank Test and Rank-Sum Test. R Programming: Statistical data analysis in research. Singapore: Springer Nature Singapore, 279-303, 2024.

[14] Ostertagova, E., Oskar O., and Jozef, K., Methodology and application of the Kruskal-Wallis test, Applied mechanics and materials 611, 115-120, (2014).

[15] Dudley, R. ,The Shapiro–Wilk test for normality, 2023,

[16] Nowacki, Amy. ,Chi-square and Fisher's exact tests., Cleve Clin J Med 84.9 suppl 2 e20-5, 2017.

[17] Pereira, D. G., Afonso,A. and Melo Medeiros,F. Overview of Friedman's test and post-hoc analysis., Communications in Statistics-Simulation and Computation, Vol 44.10, 2636-2653, 2015.

[18] Berger, V.W., and Zhou,Y.Y,. Kolmogorov–smirnov test: Overview., Wiley statsref: Statistics reference online, 2014.

[19] MIT Lincoln Laboratory. (1999). 1999 DARPA Intrusion Detection Evaluation Dataset.

[20] Lippmann, R., Haines, J., Fried, D., Korba, J., & Das, K. The 1999 DARPA off-line intrusion detection evaluation. Computer Networks, 34(4), 579–595, 2000.

[21] .Almehmadi,A , Intrusion Detection System for SYN Flood Attack: Methods and Implementation , International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 1, January 2017.

[22] Azahari, M ,Yusof. M, Hani,F ,Ali,M and Yusof Darus,M, Detection and Defense Algorithms of Different Types of DDoS Attacks, International Journal of Engineering and Technology, Vol. 9, No. 5, 2017

[23] Acharya, A-A, Arpitha, K-M, Kumar S-B.J, An Intrusion Detection System Against UDP Flood Attack and Ping of Death Attack (DDOS) in MANET, International Journal of Engineering and Technology (IJET), Vol 8 No 2 Apr-May 2016.