Classification automatique d'anomalies du trafic

Philippe Owezarski, Guilherme Fernandes

Classification automatique d'anomalies du trafic

Philippe Owezarski, Guilherme Fernandes

LAAS-CNRS, université de Toulouse, 7 Avenue du Colonel Roche, 31077 Toulouse cedex 4, France

La détection d'anomalies dans le trafic réseau et leur caractérisation est un sujet d'importance depuis de nombreuses années. Une gestion efficace de grands réseaux dépend clairement de la capacité à identifier et limiter l'effet de ces anomalies. En particulier, les anomalies causées par une attaque de déni de service par inondation ont un très fort impact sur la qualité de service des réseaux, même lorsque les liens sont largement surdimensionnés et pas saturés par ces attaques. Bien que le domaine de la recherche sur la détection d'anomalies soit bien avancé, une classification automatique précise et efficace de ces anomalies reste toujours un problème d'actualité non résolu. En effet, avec les propositions actuelles, le nombre d'anomalies détectées submerge rapidement les opérateurs réseaux qui ne savent pas comment faire face à un tel afflux, surtout si des informations supplémentaires adaptées ne sont pas fournies. Certaines approches parmi les plus récentes identifient les flux anormaux, mais elles ne donnent pas suffisamment d'informations aux administrateurs réseaux pour prioritiser le temps à passer pour analyser manuellement ces anomalies. Cela fait de la classification automatique le besoin le plus important aujourd'hui et donc la prochaine étape logique pour les recherches à venir dans ce domaine.

Dans cet article, un nouvel algorithme pour la classification automatique des anomalies du trafic est proposé. L'algorithme agit en 3 étapes : (i) après qu'une anomalie ait été détectée, il identifie tous (ou la plupart) des paquets ou flux en cause ; (ii) utilise ces informations sur les paquets et flux pour produire plusieurs métriques en rapport direct avec l'anomalie ; et (iii) classifie l'anomalie à partir de ces métriques en suivant une approche orientée signature. Nous montrons dans l'article (i) le caractère expressif nécessaire pour distinguer de façon fiable plusieurs types d'anomalies, (ii) un riche ensemble d'informations sur les anomalies détectées, et (iii) la flexibilité requise pour les administrateurs réseau afin de comprendre et dominer le processus de classification. Nous montrons également comment la phase de classification agit comme un filtre pour réduire le taux de faux positifs des algorithmes de détection. Cet algorithme a été validé sur deux bases de traces de trafic : la base produite dans le cadre du projet METROSEC, et la base MAWI.

Mots-clés: Détection d'anomalies du trafic réseau, classification d'anomalies du trafic réseau

I INTRODUCTION

L'Internet n'a cessé de se complexifier ces dernières années, passant d'un réseau offrant un service best effort unique à un réseau multi-services. Il est donc de plus en plus nécessaire d'offrir des services à qualité de service (QoS) garantie. Les anomalies du trafic sont dans ce contexte de nature a sérieusement dégrader le fonctionnement normal de ces réseaux. Il est donc vital que les administrateurs réseaux puissent rapidement les identifier et les contenir. En particulier, les anomalies en volume sont responsables de variations inhabituelles sur les caractéristiques du trafic (en général identifiées par les # de paquets[†], # d'octets ou # de nouveaux flux). Ces anomalies peuvent être causées par de très nombreux événements : de problèmes physiques ou techniques au niveau du réseau (e.g. pannes, mauvaises configurations des routeurs), à des comportements intentionnellement malicieux (e.g. attaques de déni de service , trafic lié à la propagation de vers), jusqu'à des modifications abruptes dues à du trafic légitime (e.g. foules subites, alpha flux). Cette diversité, couplée au fait que le trafic Internet normal est intrinsèquement caractérisé par des propriétés d'auto-similarité [30], de (multi-)fractalité [16] et de dépendance à long terme (LRD) [12], qui expliquent son importante variabilité naturelle, font de l'identification et de la lutte contre ces anomalies un challenge important et difficile. Les anomalies causées par les attaques de déni de service sont particulièrement importantes car ces attaques sont extrêmement fréquentes [28, 17] et provoquent des ruptures drastiques sur

[†] '#' signifie 'nombre de'.

les caractéristiques du trafic. Même les attaques DoS de faible intensité ont un impact sur la LRD du trafic, et par conséquent, engendrent des dégradation sur la QoS du réseau [29].

Malgré ces difficultés, des progrès constants ont été réalisés dans la détection des anomalies du trafic réseau. Des méthodes ont été conçues pour détecter les anomalies aussi bien sur un lien unique qu'à l'échelle d'un réseau dans son ensemble. Ces algorithmes de détection d'anomalies du trafic ont évolué, et ce, de systèmes juste capables de signaler une anomalies dans le temps à des systèmes fournissant des informations sur les flux qui créent l'anomalie. Ces informations sont particulièrement instructives pour les administrateurs réseaux qui doivent manuellement vérifier et gérer ces anomalies potentielles... Mais ce n'est pas encore assez. A cause des caractéristiques du trafic réseau et de la fréquence des anomalies, il n'est pas humainement possible pour les administrateurs réseaux de manuellement analyser toutes les anomalies détectées et signalées par les algorithmes de détection actuels. Les opérateurs réseaux ont besoin de plus d'informations que la simple indication du flux anormal pour pouvoir adapter efficacement les priorités à donner aux différentes alarmes à traiter. La classification automatique des anomalies est donc la prochaine étape que la recherche doit aborder pour pouvoir donner cette information, aujourd'hui manquante, aux administrateurs réseaux.

Bien qu'il y ait eu beaucoup d'efforts de recherche sur la caractérisation des anomalies du trafic réseau, la classification automatique n'a que très peu été abordée (une exception notable est [22]). La classification automatique a pour objectif d'ajouter des informations utiles et riches de sens à l'alerte générée lors de la détection d'une anomalie. En plus des informations basiques sur le volume et les attributs simples d'une anomalie, un algorithme de classification automatique devrait pouvoir effectuer des dérivations complexes sur les caractéristiques de l'anomalie. Idéalement, l'information produite peut ainsi être utilisée pour définir le type d'anomalie, ou au moins aider à caractériser ses causes sous-jacentes. Dans cet article, nous proposons un nouvel algorithme de classification automatique des anomalies du trafic réseau. Nous montrons comment les informations obtenues en analysant en profondeur les flux identifiés comme anormaux peuvent servir à abstraire les attributs de l'anomalie. Ces attributs peuvent ainsi être utilisés dans un module de classification orientée signature pour caractériser de façon fiable les différents types d'anomalies. En utilisant des attributs significatifs et clairs, les opérateurs réseaux peuvent adapter les règles de classification selon leurs besoins. Nous définissons dans cet article plusieurs de ces attributs et montrons que différents types d'anomalies peuvent vraiment être caractérisés en fonction d'eux. Nous montrons également les propriétés d'expressivité de cette approche qui permet de classifier de façon fiable différents types d'anomalies (e.g. DDoS, scans réseaux, réponses à des attaques) et comment elle offre aux opérateurs réseau la flexibilité requise pour l'utiliser de manière efficace selon les situations. Nous avons ensuite réalisé une validation statistique de notre méthode de classification automatique pour les attaques DDoS, et analysé les résultats obtenus avec d'autres types d'anomalies contenues dans deux bases de traces : celle issue du projet METROSEC [26] et la base de traces MAWI [7]. Ces bases de traces se complètent bien pour permettre de proposer une validation appropriée de notre algorithme.

La suite de l'article s'articule comme suit : dans la seconde partie, nous donnons une vue d'ensemble des travaux ayant des objectifs similaires aux notres. Dans la partie 3, nous présentons notre algorithme et notamment comment nous caractérisons les différents types d'anomalies. Dans la partie 4, nous décrivons les données que nous utilisons ainsi que la méthodologie de validation. Dans la partie 5, nous présentons et analysons les résultats d'évaluation. Enfin, la partie 6 conclut cet article.

II ETAT DE L'ART

Il existe aujourd'hui une littérature riche sur la question de la détection d'anomalies du trafic réseau. La plupart des approches proposées analysent les variation statistiques des volumes de trafic (i.e. nombres de paquets, d'octets ou de nouveaux flux) et/ou des distributions de certains attributs du trafic (i.e. adresses IP et numéros de ports) dans le temps ou l'espace. Les anomalies peuvent être découvertes à partir de données venant d'un seul lien ou de tout un réseau. Les références les plus fréquemment citées sont [4, 20, 21, 23], avec en plus quelques travaux très récents comme [24, 10, 5, 31]. La réduction de la dimentionalité du trafic agrégé, tel qu'il est capturé et observé aujourd'hui est aussi un domaine qui commence à être étudié, et des techniques basées sur des projections statistiques [20, 24, 10] ou les décomposition par composantes princi-

pales (PCA) [21] sont extrêmement prometteuses pour la détection d'anomalies en ligne. Il a également été montré que les projections statistiques permettaient de détecter les anomalies d'intensités faibles et peuvent les identifier (ce qui n'est pas possible avec des techniques qui travaillent sur du trafic agrégé ou sur des paires origines-destinations). La partie détection et identification de notre algorithme est issue des travaux présentés dans [14, 15]. La détection des anomalies est faite simultanément en étudiant différents niveaux d'agrégation du trafic par une méthode d'analyse statistique simple - la méthode dite des *deltoïdes* [8] appliquée aux métriques volumiques classiques du trafic - qui permet également la détection des anomalies de faibles intensités et l'identification des flux qui en sont responsables.

La caractérisation des anomalies du trafic réseau est un domaine qui a également été très largement exploré. [2] a utilisé une méthode d'analyse à base de décomposition par ondelettes sur des données du trafic issues d'un unique lien et pouvait ainsi caractériser 4 classes d'anomalies : les pannes, les foules subites, les attaques DoS et celles liées à des erreurs de mesures. Lakhina et al. ont utilisé la méthode des sous-espaces pour caractériser différents types d'anomalies à l'échelle d'un réseau à partir de métriques volumiques [23] et d'attributs du trafic [22]. Des travaux plus anciens s'étaient focalisés sur un type spécifique d'anomalie. Par exemple, les attaques DoS et DDoS ont été étudiées en détail dans [28, 17, 27]. Jung et al. [18] ont étudié les différences de comportement entre les DDoS et les foules subites au niveau d'un serveur web. Nous avons évidemment utilisé au maximum ces résultats pour sélectionner les différents attributs des anomalies de trafic qui sont utilisées aujourd'hui dans notre module de classification.

D'autres travaux ont proposé d'introduire d'autres informations sur le trafic réseau (e.g. en les regroupant en clusters [22, 13]) et de donner des priorités (e.g. en sélectionnant les contributeurs principaux - "heavy hitters" [33], ou en utilisant des heuristiques comme l'inespérance [13]). Spécifiquement aux anomalies du trafic, l'approche non supervisée de [22] crée des clusters en fonction de la représentation de l'anomalie dans l'espace entropique de leurs attributs (i.e. adresses IP et numéros de ports). Même si toutes les anomalies qui se retrouvent dans un même cluster partage une caractéristique commune, cette approche ne suffit pas pour caractériser de façon unique une anomalie. Les opérateurs réseaux auraient encore besoin de vérifier manuellement chaque anomalie signalée, même si, avec suffisamment d'anomalies connues comme faisant partie d'un cluster, ils pourraient mieux fixer leurs priorités d'actions entre les différents clusters représentant les familles d'anomalies. D'autre part, même si le travail de [13] n'a pas de rapport direct avec la détection d'anomalies, il montre comment la caractérisation automatique du trafic et la présentation de ses informations significatives peut grandement aider les administrateurs réseaux. Toutefois, dans ces travaux, catégoriser les clusters est toujours une tâche qui doit être faite manuellement. Choisir les "heavy-hitters" est une technique classique en gestion des réseaux (supervision du trafic, facturation, etc.). Cependant, pour la détection d'anomalies, ce n'est pas nécessairement une méthode appropriée, car les anomalies de faibles intensités peuvent être dominées par le trafic normal, et donc ne pas apparaître dans les listes *top N*. Caractériser les anomalies avec des attributs significatifs et donner le contrôle aux opérateurs réseaux sur le processus de classification automatique est à notre avis l'approche qui semble la plus constructive.

Plus proche de nos travaux, les auteurs de [19] étudient comment différents types d'attaques DoS et de scans de ports se comportent, créant ainsi des règles pour les détecter et les classifier, soit à partir des informations contenues dans les entêtes des flux, soit à partir d'une analyse statistique du trafic lié au flux en question. Malheureusement, [19] ne propose pas de résultats de validation sur des données réelles. Nous avons adopté le même concept consistant à exploiter le maximum d'informations possibles pour créer des signatures avec un taux de faux positifs le plus bas possible (mauvaises classifications dans notre cas). Notre algorithme a ainsi pour objectif de fournir un ensemble d'informations le plus riche possible sur les anomalies de trafic *déjà détectées* en dérivant différents attributs des caractéristiques du trafic, des métriques volumiques du trafic et des algorithmes de détection, et cela en utilisant une approche orientée signature pour les classifier. La validation de la méthode a été effectuée sur deux bases de traces de trafic complètement différentes, et les résultats se montrent très prometteurs.

III CLASSIFICATION D'ANOMALIES

Notre algorithme comporte trois étapes pour classifier les anomalies : (i) une fois qu'une anomalie a été détectée, il identifie tous (ou la plupart) des paquets ou des flux la composant ; (ii) il utilise ces informations

sur les paquets et flux impliqués dans l'anomalie pour dériver plusieurs métriques distinctes directement en rapport avec l'anomalie ; et (iii) classifie l'anomalie en utilisant ces métriques selon une approche orientée signature. Ces étapes reposent sur le besoin de disposer de beaucoup d'informations pour classifier de façon fiable les différents types d'anomalies, et même de pouvoir faire la distinction entre les différents sous-types, comme la multitude d'attaques DoS. Comme les algorithmes actuels de détection d'anomalies reposent sur un petit nombre de paramètres (i.e. métriques volumiques ou attributs du trafic comme les adresses IP et les numéros de ports), une étape est nécessaire pour obtenir plus d'informations sur l'anomalie. De façon évidente, la meilleure source d'informations sont les détails sur les paquets et les flux qui causent l'anomalie. Nous allons à partir de maintenant nous concentrer sur des traces paquets, mais des résultats similaires pourraient être obtenus à partir de traces de flux.

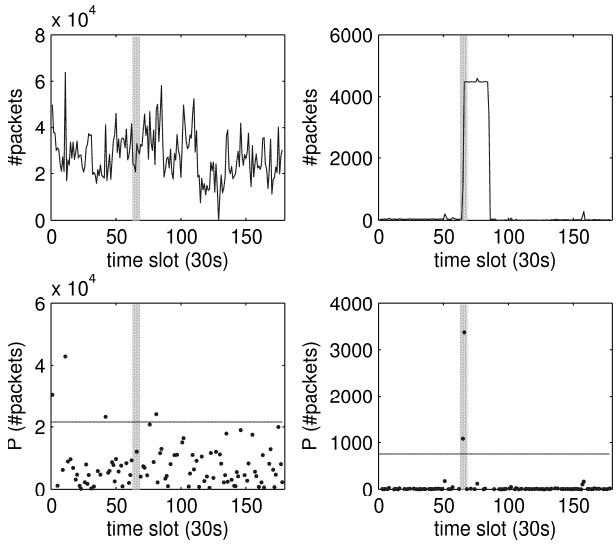

FIG. 1: Les courbes montrent comment les anomalies de faibles intensités peuvent être masquées en considérant du trafic agrégé (gauche), mais peuvent apparaître clairement sur le réseau de la victime (de préfixe /24, à droite).

Pour tester notre algorithme de classification, nous utilisons un algorithme très simple de détection d'anomalies basé sur l'utilisation de *deltoïdes absolus* [8] sur des séries temporelles associées au volume du trafic acheminé (i.e. #paquets, #octets, #syn) et une formule simple pour le seuil fixé à la moyenne + K fois l'écart type sur ces deltoïdes. La formule pour la détection (voir équation 1) se justifie de la manière suivante : Etant donnée une trace de durée T et une granularité d'observation Δ (i.e. 30s dans cet article), elle est découpée en N slots où $N \in [1, T/\Delta]$. Pour chaque slot i on obtient la série temporelle X pour chaque métrique volumique du trafic. On obtient ensuite les deltoïdes absolus P pour X et on calcule leur écart type. Pour tout p_i au dessus du seuil $K * \sigma_p$, on marque le slot comme anormal. L'utilisation des deltoïdes sur ces séries temporelles est primordiale pour considérer la variation par rapport à l'amplitude au lieu de la simple variation du trafic, cette dernière n'étant pas significative à cause de la forte variabilité naturelle du trafic. Notre choix de métriques s'appuie sur celui de [23] (avec #syn au lieu du # de nouveaux flux), mais la formule permet d'utiliser des séries temporelles associées à l'évolution de n'importe quel autre type de données.

$$\begin{aligned} X = \{x_1, x_2, \ldots, x_n\},\ x_i = \{\#packets | \#bytes | \#syn\}/\Delta \\ P = \{p_1, p_2, \ldots, p_{n-1}\},\ p_i = x_{i+1} - x_i \\ \begin{cases} p_i \geq K\sigma_p, & anomalous \\ p_i < K\sigma_p, & normal \end{cases} \end{aligned} \tag{1}$$

La détection des anomalies de faibles intensités est essentielle notamment par rapport aux attaques DDoS. Si la détection d'une attaque par inondation se fait près de la cible, la dégradation des performances du réseau et de sa QoS est déjà importante, et l'attaque par conséquent un succès. Il faut donc absolument

détecter l'attaque au plus tôt, soit près de ses sources (i.e. où l'intensité est faible). Pour rendre l'algorithme de détection sensible aux anomalies de faibles intensités, nous appliquons la formule de détection avec différents niveaux d'agrégation. Les différents niveaux d'agrégation reposent sur l'utilisation d'un masque de taille variable sur les adresses IP destination. Dans cet article, nous utilisons les niveaux d'agrégation correspondant à des masques /0 (i.e. tout le trafic), /8, /16 et /24. La figure 1 montre une attaque DDoS aux composantes de faible intensité qui ne peut être détectée qu'avec un préfixe de taille /24. La courbe du haut montre la série temporelle associée au nombre de paquets (i.e. le trafic total à gauche, et celui à destination du seul réseau /24 ciblé à droite) ; la courbe du bas représente les *deltoïdes positifs* P correspondants ; les slots correspondant au début de l'attaque sont grisés. Ce genre de situation où une *anomalie locale* est masquée par la grande variabilité du trafic agrégé normal est quasiment systématique aujourd'hui et pose des problèmes de gestion des réseaux très importants.

Comme pour les autres algorithmes de détection, l'augmentation de la sensibilité implique un plus grand nombre de faux positifs (i.e. beaucoup des variations normales du trafic sont alors considérées comme anormales). A cause du principe d'analyse multi-niveaux, la formule présentée ci-dessus est particulièrement sensible aux communications peu fréquentes pour lesquelles peu de paquets sont généralement vus pour un réseau ou un masque d'agrégation particulier. Le taux de faux positifs est largement supérieur au nombre d'anomalies de faibles intensités effectivement détectées. Même si cette lacune aurait tendance à rendre ce type d'algorithme de détection inutilisable, nous montrons comment le processus de classification peut être utilisé par les administrateurs réseaux pour sélectionner a priori les anomalies qu'ils jugent importantes, réduisant ainsi fortement le taux de faux positifs. La simplicité de l'algorithme rend l'étape suivante (i.e. l'identification des paquets correspondants et la dérivation des métriques) évidente, et nous a permis de nous concentrer sur la caractérisation des anomalies. Nous envisageons d'étudier dans de futurs travaux comment de meilleurs algorithmes de détection pourraient être utilisés avec cette approche de classification.

III.1 Collecte des Informations

Grâce à la caractérisation des anomalies de trafic faite dans des travaux précédents [2, 23, 22], nous avons pu constater que différents types d'anomalies peuvent affecter les métriques volumiques et les attributs du trafic, comme les adresses IP et les numéros de ports, de la même manière. Cela démontre clairement que l'on ne peut pas proposer une classification fiable reposant sur ces métriques, et que l'on doit identifier des informations additionnelles pour y parvenir. Nous allons donc maintenant introduire la notion d'*attributs* d'anomalies pour exprimer les informations que nous pouvons récupérer en analysant les paquets directement associés à une anomalie détectée. Un attribut est un élément qui aide à caractériser une anomalie spécifique. Les attributs peuvent être simples et obtenus directement (e.g. si l'anomalie a un impact sur le nombre de paquets) ou provenir de dérivations plus complexes (e.g. le ratio entre le nombre de source et le nombre de destinations impliquées dans l'anomalie). Le module de classification utilise des signatures basées sur les attributs ainsi obtenus. Un processus de classification fiable a besoin d'attributs significatifs qui contiennent suffisamment d'informations pour pouvoir faire la distinction entre différents types d'anomalies). Par exemple, une variante de l'algorithme de clustering de [18] pourrait être implémentée et représentée par un attribut pour permettre de distinguer DDoS et foules subites. Mais, pour pouvoir dériver ce genre d'informations, il faut d'abord pouvoir identifier les paquets de l'anomalie.

L'algorithme de détection que nous utilisons rend évidente l'identification de ces paquets. Une anomalie, une fois détectée, est identifiée par son slot, son adresse réseau et son masque. Nous savons également exactement pourquoi elle est considérée comme anormale (i.e. quel deltoïde pour l'une ou l'autre des métrique était au dessus du seuil). A partir de ces informations, nous commençons donc à lire tous les paquets du slot en question qui vont vers le réseau ciblé, de façon à pouvoir en identifier la machine source (i.e. adresse IP /32). Nous faisons de même pour identifier les ports et protocoles responsables. Cela peut également être appliqué à tout autre caractéristique du trafic.

Pendant la phase de détection et d'identification des flux responsables de l'anomalie, nous produisons l'ensemble des attributs indiqués dans le Tableau I. Les attributs *found*, *impactlevel*, *duration* and *decrease* sont spécifiques à l'algorithme de détection que nous avons sélectionné pour présenter cette article, mais des attributs différents pourrait être considérés si d'autres algorithmes de détection étaient utilisés. *Found* and *impactlevel* sont les attributs les plus importants dans notre cas. Alors que l'attribut *found* est simple-

Attribute	Type	Description
found{p,b,s}	integer	If the corresponding metric was anomalous, value of P, zero otherwise
impactlevel{p,b,s}	integer	The impactlevel of the anomaly (see Section 3.1)
duration{p,b,s}	integer	For how many slots the metric stayed above the threshold
decrease{p,b,s}	float	The biggest negative deltoid during the anomaly as a fraction of the threshold
#respdest	integer	Number of responsible destinations
#rsrc/#rdst	integer	Ratio of responsible sources to responsible destinations
avg#rdstports	integer	Average number of responsible destination ports
avg#rsrcports	integer	Average number of responsible source ports
#rpkt/#rdstport	integer	Ratio of number of packets to responsible destination ports
#rpkt/#rsrc	integer	Average number of packets of responsible sources
bpprop	integer	Average packet size (only packets of the anomaly)
spprop	float	Ratio of number of syn to number of packets of the anomaly
samesrcpred	boolean	If a specific responsible source appears for the majority of destinations
samesrcportpred	boolean	If the majority of responsible sources use the same source ports
oneportpred	boolean	If only one destination port dominated
invalidpred	boolean	If the anomaly was mainly consisted of invalid packets (e.g. malformed headers)
invprotopred	boolean	If the anomaly was dominated by packets using invalid protocol numbers or types
landpred	boolean	If most packets had the same source and destination IPs
echopred	boolean	If most packets were of type ICMP Echo Request/Reply
icmppred	boolean	If most packets were ICMP of any other type
rstpred	boolean	If most packets were TCP with RST flag set

TAB. I: Attributs dérivés d'une anomalie donnée. *p*, *b* and *s* sont utilisés pour paquets, octets (bytes) et syn respectivement.

ment l'indication du deltoïde qui a détecté l'anomalie, l'attribut *impactlevel* représente la corrélation entre les différents niveaux d'agrégation. Plus simplement, l'*impactlevel* d'une anomalie est le nombre de niveaux d'agrégation parents pour lesquels l'anomalie a aussi été décelée. Les autres attributs sont dérivés de l'identification des flux responsables de l'anomalie. La liste n'est pas absolue et pourrait être étendue. Ces attributs sont ceux qui ont été identifiés comme utiles au cours de ce travail, et leur justification est apportée dans la partie 3.2.

III.2 Classification

III.2.1 Idée Générale

L'objectif principal de notre algorithme est d'étiqueter automatiquement des anomalies de trafic qui viennent d'être détectées. Cette classification automatique doit donner aux administrateurs réseaux assez d'informations pour qu'ils puissent utiliser efficacement leur temps disponible à l'inspection manuelle des anomalies. Cela pose un problème important : Les opérateurs réseaux doivent être capables de facilement et complètement comprendre et manipuler le processus de classification. Eventuellement, un processus de classification fiable (i.e. signatures avec un taux de faux positif très bas) pourrait être utilisé dans un système complètement autonome qui traiterait les cas anormaux sans aucune intervention humaine (e.g. en changeant la politique de sécurité des routeurs). Malheureusement, cet objectif d'un processus de classification automatique des anomalies du trafic avec un très faible taux de mauvaises classifications est un problème fortement complexe et encore ouvert. Le grand nombre d'anomalies de types différents [23] et les grandes variations à l'intérieur d'un même type (e.g. pour les DDoS [27]) oblige à créer des signatures très spécifiques pour atteindre de faibles taux de faux positifs. Nous avons utilisés les travaux précédents sur la caractérisation d'anomalies (e.g. sur des anomalies de trafic générales [2, 23, 22], et plus spécifiquement sur les attaques [28, 17, 27]) pour faire un pas supplémentaire dans cette direction. Pour cela, nous avons défini 3 types de signatures : (i) universelle, (ii) forte et (iii) locale. Les signatures universelles sont des règles qui ne devraient jamais mal classifier une anomalie, indépendamment des caractéristiques du réseau. Malheureusement, ce type de signatures est rare, ayant surtout trait à des spécifications de protocoles. Les signatures fortes sont supposées présenter des taux de faux positifs bas, mais en général reposent sur une forme ou une autre de seuillage (et les seuils sont très difficiles à fixer). Les signatures locales devraient être définies par les administrateurs réseaux pour signaler les anomalies qui ont un intérêt pour eux. A noter qu'ils peuvent choisir comment étiqueter au mieux ces anomalies et changer les seuils de façon à ce qu'ils

s'adaptent au mieux à leurs besoins.

Nous allons maintenant décrire les anomalies que nous avons étudiées et montrer sur quelques exemples comment les attributs que nous avons identifiés peuvent être utilisés pour créer des signatures fortes, ou même universelles. Même si nous nous sommes concentrés sur l'étude et la validation de méthodes de classification des anomalies dues aux attaques de déni de service (potentiellement distribuées), nous avons également étudié les scans réseaux, les scans de ports, les foules subites, les alpha-flux et les anomalies dues aux réponses aux attaques. Nous allons également montrer quelques exemples de signatures locales qui pourraient intéresser les administrateurs réseaux. L'idée est de permettre au lecteur de mieux appréhender comment la classification automatique peut se faire à partir de ces attributs, et de montrer l'expressivité de notre algorithme. De nouveaux attributs et de nouvelles règles pourront certainement être identifiés grâce à l'expertise des administrateurs réseaux.

III.2.2 Caractérisation des attaques DoS

Les attaques de déni de service (DoS) sont des tentatives malicieuses pour interdire l'accès à des ressources réseaux pour des utilisateurs légitimes. Ces attaques DoS peuvent se classer grossièrement en attaques logiques et par inondation [28]. Les attaques logiques utilisent des failles logicielles pour dégrader les performances ou complètement stopper un service. Les attaques par inondation demandent plus de ressources que la victime ne peut en offrir. Même si ces deux types sont importants, nous allons seulement étudier les attaques par inondation car ce sont celles qui sont de nature à créer des anomalies dans le trafic. Les attaques DDoS sont des attaques par inondation qui utilisent de multiples sources pour maximiser les dégats tout en restant difficilement détectables. D'ailleurs elles sont communément utilisées pour des activités malicieuses (e.g. blackmailing [32], cyber-guerres politiques [9]), et peuvent grandement réduire la QoS d'un Réseau même lorsque ce dernier a suffisamment de ressources pour fonctionner [29]. La nature (des-)agrégée de cette attaque la rend difficile à combattre une fois qu'elle a atteint sa cible. Il est donc important de les détecter et les stopper aussi près que possible de leurs sources. Pour atteindre cet objectif, il faut absolument pouvoir détecter les anomalies de faibles intensités. Mais les attaques DoS avec de faibles intensités et un petit nombre de sources sont comportementalement très proches du trafic normal, ce qui rend indispensable de bien les caractériser et de trouver une signature associée forte.

Les anomalies du trafic causées par des attaques DDoS ont été caractérisées comme créant des pics dans les séries temporelles dénombrant le nombre de paquets, le nombre de flux, ou les deux [23, 2] et affectant les distributions des attributs de base du trafic comme les adresses et ports sources et destinations [22]. Même si cela donne une bonne idée de l'impact des attaques DDoS, ces attributs ne suffisent pas pour créer des signatures d'anomalies robustes et permettant leur classification automatique (i.e. pour les distinguer d'avec d'autres types d'anomalies). C'est pour cela que nous devons fournir plus d'informations sur les différents types d'attaques DDoS [17, 27] et même sur le fonctionnement des différents outils d'attaque utilisés par les pirates (e.g. [11]).

Des signatures universelles pour les DDoS peuvent être définies en analysant les types d'attaques DDoS qui utilisent des paquets qui ne respectent pas la spécification du protocole. Par exemple, ont été observées de nombreuses attaques qui utilisaient soit la taille minimale des paquets IP (i.e. 40 octets) [17], un protocole invalide (e.g. le champs protocole de l'entête IP à 0 ou 255 [28, 17], ou utilisant des paquets ayant la même adresse source et destination [10]. Cela peut s'étendre aux couches transport, avec des attaques qui utilisent des combinaisons invalides de flags TCP ou des types ICMP invalides [17]. Comme notre phase d'identification des flux responsables peut travailler sur ce genre de détails, il est facile de dériver des attributs qui peuvent être utilisés pour classifier ces attaques spécifiques (i.e. dans le Tableau I, les attributs *invalidpred*, *invprotopred* et *landpred* respectivement par rapport aux exemples précédents). Une règle simple et directe serait *if invalidpred or invprotopred or landpred then label as DoS*. A noter que toutes les informations d'identification (e.g. adresses et ports source(s) et destination, protocole, etc.) sont données dans l'alerte.

Créer des signatures universelles pour des attaques DDoS qui utilisent des paquets conformes est très difficile (si ce n'est impossible). Pour ce type d'attaques, nous avons essayé de développer des signatures fortes en utilisant une grande variété d'attributs. Le Tableau II montre certaines des signatures utilisées pour cet article. La signature de l'attaque ICMP Echo prise en exemple repose sur le fait que l'anomalie n'a

Id	Anomaly Type	Signature
1	ICMP Echo DDoS	#respdest == 1 and echopred and (#rpkt/#rdstport > 30*gr or #rsrc/#rdest > 15)
2	TCP SYN DDoS	#respdest == 1 and founds and spprop > 0.9 and oneportpred and #rpkt/#rdstport > 10*gr
3	Network Scan	#respdest > 200 and samesrcpred
4	SYN Port Scan	#respdest == 1 and #rsrc/#rdest == 1 and spprop > 0.8 and avg#rdstports > 5
5	Attack Response	#respdest == 1 and (rstpred or icmppred) and foundp > 20*gr and (not (impactlevelp == 3)) and (#rsrc/#rdest == 1 or samesrcportspred)

TAB. II: Exemples de signatures fortes utilisées dans ce travail. *gr* est la granularité de la série temporelle.

qu'une destination responsable (*#respdest*), se compose essentiellement de paquets ICMP Echo (*echopred*), et présente en moyenne plus de 30 paquets par seconde (*#rpkt/#rdstport*) ou plus de 15 sources responsables (*#rsrc/#rdest*). Ces seuils ont été fixés de manière à détecter les anomalies de faibles intensités, mais devraient conduire à un taux très faible de mauvaises classifications. La seconde signature du Tableau II classifie les attaques TCP SYN à destination d'un service particulier (*oneportpred*) avec une moyenne de 10 ou plus paquets par seconde. Elle utilise *founds* et *spprop* pour vérifier que la plupart des paquets qui créent l'anomalie ont (seulement) le flag TCP SYN positionné. Les autres signatures pour les anomalies de faibles intensités repose sur le fait que les attaques vont utiliser soit de nombreuses adresses sources aléatoires *#rsrc/#rdst*), plusieurs ports sources (*avg#rsrcports*), plusieurs ports destinations (*avg#rdstports*) ou une combinaison de ces 3 caractéristiques. Les signatures d'anomalies de fortes intensités sont plus précises car on peut utiliser la valeur forte de l' impactlevel (i.e. 3) du #paquets ou du #syn. La taille moyenne des paquets (*bpprop*) peut aussi être utilisée car il est rare qu'une attaque DDoS utilise de gros paquets [28]. Il est important de noter que nous utilisons ces attributs pour classifier les anomalies de trafic en volume, et pas pour détecter des comportements anormaux. Cela rend les seuils plus robustes face aux changements dans les caractéristiques du réseau. Par exemple, la signature de l'attaque ICMP Echo peut être levée par la détection d'anomalies sur les réponses des systèmes de gestion du réseau, alors que leur activité planifiée et régulière (et courte) ne devrait pas générer d'anomalie.

III.2.3 Autres Anomalies

Nous allons maintenant nous intéresser rapidement aux autres types d'anomalies et leurs attributs les plus intéressants que nous avons identifiés pour chacune d'elles. Les *scans réseaux* (souvent appelés aussi scans horizontaux) [27] sondent de nombreuses machines pour savoir si un service spécifique est disponible. Ce type d'anomalie est très important car il est souvent un signe précurseur à une tentative de subversion de machines vulnérables par un pirate — qui va potentiellement les utiliser comme sources pour des attaques de DDoS — ou par des vers. Un scan réseau peut être caractérisé de façon fiable par une source unique communicant avec de multiples destinations. Les attributs les plus utiles pour classifier cette anomalie sont donc *#respdest* et *samesrcpred*. Nous avons empiriquement remarqué qu'un seuil de 20 pour *#respdest* est efficace, et agit presque comme une signature universelle (le seul cas pour lequel une mauvaise classification apparait est dans les mises-à-jour massives pour des anti-virus ou des systèmes opératoires avec de nombreux clients sur un même réseau). Des signatures plus fortes peuvent également utiliser *bpprop*, *foundsyn*, *spprop*, *oneportpred* et *#rpkt/#rdstport* pour améliorer la précision et peut être aussi baisser le seuil pour *#respdest*. Nous avons choisi d'augmenter le seuil pour *#respdest* à 200 dans ces travaux à cause du grand nombre de scan réseaux qui ont été décelés (voir partie 5).

Les *Port scans* sont similaires mais se concentrent sur une seule destination afin de découvrir tous les services qu'elle offre. En général, ils ne crée qu'un tout petit trafic mais peuvent avoir un impact sur le nombres de paquets syn. Ils sont caractérisés par une source, une destination et de multiples ports et un petit nombre de paquets échangés. Aussi, les attributs suivants sont importants : *#respdest*, *#rsrc/#rdst*, *avg#rdstports*, *#rpkt/#rdstport*, *foundsyn*, *spprop*. La signature 4 du Tableau II décrit un exemple pour classifier les ports scans TCP SYN. Les *Foules subites* (ou Flash Crowds - FC) peuvent être définies comme une arrivée massive et soudaine de requêtes de clients légitimes pour une ressource. Elles peuvent être planifiées (e.g. webcasts programmés) ou inattendues (e.g. un site web populaire établissant un lien vers un petit site). La nature distribuée des FCs rend délicate la distinction d'avec les attaques DDoS [18]. Les attributs caractéristiques comprennent *#rsrc/#rdst*, *oneportpred*, *foundsyn*, *foundpkts*, *#rpkt/#rsrc* et *decrease*, tout en prenant aussi en compte qu'elles seront seulement détectables sur des granularités d'observation élevées

(i.e. $>$ 5min). Les *Alpha flux* sont de gros transferts à hauts débits inhabituels d'une source vers une destination unique et qui ont un fort impact sur #bytes et #packets [23]. Ils utilisent généralement des paquets bien plus gros que les attaques DoS. Normalement, les informations sur les ports sont utilisées pour identifier des opérations connues qui crée des alpha flux (e.g. backups programmés ou estimations de bandes passantes). Les attributs concernés sont *impactlevelbytes*, *impactlevelpkts*, *#respdest*, *#rsrc/#rdst*, *bpprop* et *foundsyn*, et les ports réels restent à définir.

Finalement, les anomalies liées à des *réponses à des attaques* sont générées par des victimes d'attaques (e.g. DDoS ou scans). Ces paquets de réponse sont normalement soit des paquets TCP avec les flags RST ACK, RST ou SYN ACK levés, soit des paquets de contrôle ICMP (e.g. Destination Unreachable, TTL Exceeded) [28]. Les réponses à des attaques devraient avoir des tailles de paquets moyennes très faibles. Pour des réponses à des attaques DoS, il n'existe qu'une source responsable, alors que pour les scans réseaux (en particulier ceux dus à des vers) il en existe plusieurs. La frontière entre des réponses à des attaques et des DDoS de faibles intensités est très étroite, notamment parce qu'il est connu que ces paquets sont utilisés dans les attaques de déni de service par réflexion [17]. Par conséquent, des signatures fortes sont difficiles à définir et il faut être prudent sur la priorité à donner aux anomalies ainsi labelisées. La liste des attributs pour caractériser les réponses à des attaques comprend *#respdest*, *impactlevelpkts*, *bpprop*, *icmppred*, *rstpred*, *#rsrc/#rdst*, *samesrcportspred*. La signature 5 du Tableau II détaille une signature unifiée pour détecter des réponses à des attaques par inondation et à des tentatives de scan. Pour cette signature, nous avons choisi de considérer des anomalies fortes (i.e. des anomalies qui ont un impact visible au niveau 0).

III.2.4 Signatures locales

Cette liste d'attributs clairs permets maintenant aux administrateurs réseaux de construire leurs propres règles à partir de leur précieuse expérience. Même si ces règles peuvent ne pas être transposables sur d'autres réseaux, la flexibilité qu'apporte la possibilité de comprendre, ajouter et modifier des règles de classification d'anomalies est un avantage certain pour l'applicabilité de la classification automatique d'anomalies sur des réseaux réels. Les signatures fortes peuvent être modifiées (i.e. en changeant les seuils) ou désactivées, alors que des signatures locales peuvent être implémentées. Les opérateurs réseaux peuvent préférer définir des signatures pour des intensités d'anomalies différentes, ou utiliser leurs propres conventions de labelisation. Par exemple, au lieu d'essayer de séparer les réponses aux attaques des attaques DDoS qui utilisent des paquets TCP RST, la signature suivante pourrait être définie : *if #respdest == 1 and rstpred and impactlevelp $>$ 2 then label as StrongRSTAnomaly*. Cette signature pourrait ensuite être spécialisée en versions simple ou multi-sources. La flexibilité offerte par cette approche peut aussi être utilisée pour réduire le nombre de faux positifs générés par les algorithmes de détection.

IV VALIDATION

Dans cette partie, nous évaluons les performances de notre algorithme de classification automatique des anomalies du trafic réseau. Nous validons statistiquement les performances pour les attaques de DDoS et commentons les résultats obtenus pour d'autres types d'anomalies. En outre, les multiples types d'attaques de DDoS nous ont permis de vérifier l'expressivité de notre approche. Si nous pouvons efficacement différencier différentes attaques DDoS du trafic normal et d'autres types d'anomalies, cela prouve que la classification automatique est possible. Nous présentons les données utilisées pour cette validation, ainsi que la méthodologie. Les résultats sont présentés dans la partie 5.

IV.1 Données expérimentales

Une bonne méthode de validation statistique de procédures de détection d'anomalies (et de classification) nécessite d'utiliser des traces de trafic contenant des anomalies identifiées. Ces traces doivent être collectées sur une réseau réel opérationnel et être labelisées ensuite grâce à l'expertise d'administrateurs réseaux. Cela génère des bases de traces avec des anomalies réelles et connues, mais peut aussi être sujettes à des erreurs humaines (i.e. les opérateurs réseaux peuvent mal classifier une anomalie) et ne permet pas de contrôler les caractéristiques de l'anomalie (e.g. leur intensité). Générer de telles bases de traces est coûteux et aujourd'hui, très peu sont publiquement disponibles. L'autre façon de générer de telles bases de traces

id	Tt	At	Type	#bots	Intensité
tN	7200	600	UDP	3	22.90%
tT	7200	600	UDP	4	86.80%
1,tM	7200	600	UDP	2	7%
2	3600	600	UDP	4	4%
3	5400	600	UDP	4	7%
4	7200	600	TCP SYN	2	12%
5	5400	1800	ICMP Echo	4	8%
6	3600	600	ICMP Echo	4	10%
7	3600	600	Mixed	4	27%
8	3600	600	Smurf	4	4%
9	3600	600	TCP SYN	3	33%
10	7200	660	UDP	4	92%
11	10800	160	TCP SYN	1	90.50%
		420		1	70.80%
		500		1	45.62%
12	1800	300	TCP RST	1	91.70%

TAB. III: Caractéristiques des traces sélectionnées dans la base METROSEC. *Tt* est la durée des traces en secondes, *At* est la durée de l'attaque en secondes, *#bots* est le nombre de sources attaquantes et *Intensité* est le ratio entre de débit de l'attaque et le débit total.

documentées est de produire les anomalies dans des réseaux réels ou simulés. Avec cette approche, les anomalies peuvent être complètement documentées et ne sont pas sujettes à de mauvaises interprétations. Les caractéristiques des anomalies peuvent aussi être contrôlées (i.e. en faisant varier leurs intensités, leurs durées, etc.) pour permettre des évaluations avec des paramétrages différents. L'inconvénient potentiel est que les anomalies ne soient pas très représentatives des occurrences courantes. Pour valider notre algorithme de façon complète nous utilisons ces deux types de bases de traces : les traces du projet METROSEC avec des anomalies artificiellement créées et la base de trace MAWI avec des anomalies observées dans le monde Internet réel.

La base de traces METROSEC consiste en des traces de trafic réel collectées sur RENATER avec des attaques qui ont été perpétrées en utilisant des outils d'attaques DDoS. Cette base de trace a été créée dans le contexte du projet de recherche METROSEC [26] pour, entre autres objectifs, étudier la nature et l'impact des anomalies sur la QoS des réseaux. Elles ont été utilisées pour valider différents outils de détections d'anomalies produits dans le cadre du projet (e.g. [31, 1, 15]). Les traces ont été collectées par un système DAG de fin 2004 à fin 2006, et contiennent des anomalies dont les intensités vont de très faible (i.e. moins de 4% en volume) à très forte (i.e. plus de 80%). Pour les attaques DDoS, de une à quatre sources ont été utilisées (laboratoires à Mont-de-Marsan, Lyon, Nice et Paris) pour recréer des attaques complexes et réalistes vers le LAAS. Ces traces sont complètement documentées. Le Tableau III liste les traces utilisées pour la validation de notre algorithme. Nous nous sommes concentrés sur des anomalies de faibles intensités, et plus spécifiquement pour différents types d'attaques DDoS (comme indiqué dans le Tableau III). Les 3 premières attaques ont été réalisées avec Trin00 [11] et les autres avec TFN2K [3].

D'une autre côté, la base de traces MAWI contient des anomalies réelles non documentées. Elle se compose de traces de paquets de 15 minutes collectées tous les jours à 2 heures (PM) depuis 1999 sur un lien entre le Japon et les Etats-Unis. Ces traces sont publiques après avoir été anonymisées et dépourvues de leur charge utile (voir http ://mawi.wide.ad.jp/). Même si ces traces ne sont pas documentées, les auteurs de [10] ont commencé à documenter les différentes anomalies trouvées dans cette base de traces. Nous avons aléatoirement sélectionné 30 de ces traces parmi lesquelles certaines contiennent des attaques DDoS (identifiées par [10]). Utiliser cette seconde base de traces est important pour vérifier que notre algorithme ne fonctionne pas seulement sur un seul réseau et face à des anomalies artificielles. Nous allons appliquer les mêmes signatures aux deux bases de traces.

IV.2 Méthodologie

Pour représenter nos résultats avec les traces METROSEC, nous utilisons la technique ROC (Receiver Operating Characteristic). La courbe ROC que nous utilisons peut se définir comme la probabilité de bonne

détection (i.e. true positive rate ou TPR) en fonction de la probabilité de fausses alarmes (i.e. false positive rate ou FPR). Cette technique proposée pour l'évaluation du système de détection d'intrusion du DARPA en 1998 a été utilisée de façon quasi générale depuis pour évaluer ce type de systèmes. Malgré son adoption massive, plusieurs lacunes ont été mises en évidence [25]. Une explication prudente de la méthodologie doit donc être donnée de sorte qu'elle puisse être utilisable sans imper pour d'autres analyses. Nous allons donc maintenant expliquer et justifier notre méthodologie pour l'évaluation de notre algorithme.

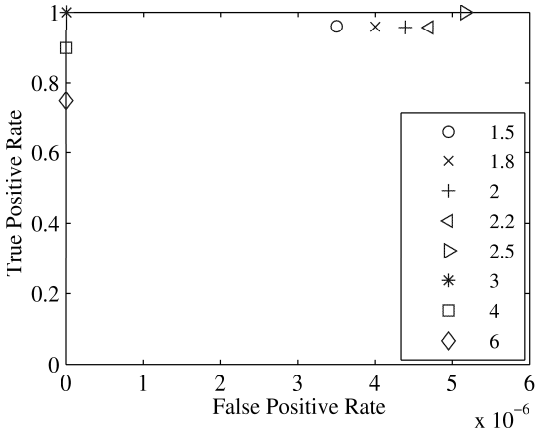

FIG. 2: Courbe ROC pour la classification des attaques DDoS de la base METROSEC (à noter l'échelle de l'axe des abscisses). Chaque point représente une valeur différente du paramètre K de l'algorithme de détection.

Comme l'objectif de notre algorithme est de classifier automatiquement des anomalies du trafic *déjà détectées*, il remplira son objectif s'il peut en dériver les attributs et les signatures utilisées. Nous avons choisi de valider l'algorithme dans son ensemble (i.e. avec l'algorithme de détection basé sur les deltoïdes choisi pour l'occasion). Nous nous concentrons aussi sur les attaques DDoS qui sont parmi les plus problématiques à transporter et celles pour lesquelles le traitement correctif est le plus évident à implémenter. En s'éloignant du contexte classique de la *seule détection*, nous introduisons quelques subtilités notamment pour déterminer la bonne *unité d'analyse* [25]. L'unité d'analyse de base de notre algorithme est toute anomalie détectée qui entre dans le module de classification (i.e. celles au niveau /24). La première subtilité est que seules les anomalies détectées par la formule de détection (voir équation 1) vont être traitées par le module de classification. Le second problème aussi lié à la formule de détection est dû au fait qu'elle utilise des deltoïdes pour déterminer les comportements anormaux (i.e. les slots postérieurs à celui du début d'une attaque peuvent ne pas présenter de variations significatives). Cela signifie que seul le début de l'anomalie sera certainement détecté, mais c'est certainement le plus important. De plus, la nature multi-types de notre processus de classification doit être réduite ponctuellement à un cas binaire de décision par rapport à des DDoS seulement.

Nous proposons une solution à ces problèmes de la façon suivante : nous définissons une *anomalie d'attaque* comme une anomalie détectée par notre formule de détection mais qui est due principalement à des paquets d'attaque. Par conséquent, l'algorithme doit correctement identifier et classifier l'anomalie détectée, pour qu'elle soit comptabilisée comme une vrai positif. L'identification est considérée comme correcte si l'adresse IP du destinataire est celle de la victime, et si la liste des sources responsables ne contient que des adresses IP sources des paquets d'attaque (mais pas nécessairement de tous). Nous considérons un décalage de deux slots (i.e. une minute pour une granularité de 30 secondes) pour la détection du début d'une attaque DDoS, principalement à cause d'un effet potentiellement montant de l'attaque à son démarrage. Cela signifie que si la formule de détection ne génère pas une alerte durant la première minute par rapport au début de l'anomalie telle que documentée, un faux négatif sera comptabilisé. Le problème par rapport à la nature multi-types de notre algorithme a été résolu en ignorant simplement toutes les signatures qui ne sont pas directement associées à des DDoS (les signatures locales aussi). Si une anomalie détectée est liée à n'importe quel autre type d'événement (e.g. scan réseau, foule subite, alpha flux), il ne sera que comptabilisé pour les faux positifs (s'il est classifié comme un DDoS) et les faux négatifs (s'il ne correspond à aucune signature

K	TP	FP	FN	TN	TPR	FPR
1.5	25	2	1	571715	0.961	3.50E-06
1.8	24	2	1	500448	0.96	4.00E-06
2	23	2	1	455731	0.958	4.39E-06
2.2	23	2	1	425752	0.958	4.70E-06
2.5	23	2	0	387719	1	5.16E-06
3	22	0	0	346656	1	0
4	19	0	2	288446	0.9	0
6	12	0	4	192247	0.75	0

TAB. IV: Résultats de la classification des attaques DDoS des traces METROSEC.

de DDoS).

Comme les traces MAWI ne sont pas complètement documentées, nous ne pouvons pas utiliser la même méthode. En utilisant les résultats des travaux de [10], nous avons une première indication sur les anomalies contenues dans ces traces. Mais de nombreuses autres peuvent être encore cachées et nous ne pouvons donc pas être sûrs des taux de faux positifs et faux négatifs. Pour ces traces, nous utilisons toutefois ces premiers résultats et analysons avec soins les alertes générées par l'algorithme pour faire une analyse descriptive des faux et vrais positifs ainsi que des faux négatifs connus. Pour les deux bases de traces, nous avons réalisé une analyse complète des résultats montrant comment les différents composants (i.e. algorithme de détection, processus de dérivation des attributs et des signatures) affectent les performances. Nous avons aussi testé des signatures pour d'autres types d'anomalies et commenté les résultats obtenus. Comme mentionné dans la partie 3, nous avons utilisé les paramètres suivants : granularité = 30s, niveaux = 0, 8, 16, 24. Nous illustrons les lacunes de notre problème d'unité d'analyse en choisissant différentes valeurs du paramètre K de l'algorithme de détection sur les traces METROSEC, et avons représenté les résultats dans l'espace ROC. Nous avons utilisé un paramètre K fixé à 2 pour les traces MAWI et pour l'analyse des signatures pour d'autres types d'anomalies que les DDoS.

V RESULTATS ET DISCUSSIONS

La Figure 2 montre l'espace ROC avec les résultats obtenus pour les traces METROSEC. Chaque point représente les résultats (i.e. probabilité de bonne alarme vs. probabilité de fausse alarme) obtenus avec un paramètre K différent. On peut voir à partir des points dessinés que, dans ce cas, baisser ce paramètre fait augmenter le taux de fausses alarmes. Bien que ce soit intuitif, cela peut s'expliquer par le fait que les nouvelles anomalies détectées ont tendance à ne pas correspondre aux signatures fortes, mais augmentent le nombre total d'anomalies classifiées (i.e. augmentent le nombre de vrais négatifs). Ces résultats sont mieux illustrés dans le Tableau IV. On peut voir que même s'il y a seulement 16 attaques dans la base de traces, le nombre d'anomalies d'attaques pour le K considéré peut monter jusqu'à 26. Ces "extra" anomalies signifient que certaines attaques présentent des augmentations de leur débits après leurs débuts. Ces résultats démontrent une incroyable précision des signatures pour classifier les anomalies de type DDoS. Si on regarde le plus mauvais résultat, seulement 2 fausses alarmes ont été relevées et 1 anomalie d'attaque a été considérée comme normale. Une analyse approfondie a révélé qu'une des fausses alarmes était une très forte réponse aux attaques SYN Floods de la trace 11, et l'autre était en fait une attaque ICMP par réflexion, réelle et inattendue. Le faux négatif est dû à l'attaque de la trace 8 démarrant à 7 secondes du slot et ne créant ainsi qu'un faible début d'anomalie, suivi par un beaucoup plus fort détecté dans le slot suivant. Le début de l'anomalie évite ainsi nos signatures, mais la partie la plus forte, qui se trouve au sein de nos deux slots considérés, est correctement identifiée. Aucune des ces anomalies fautives n'est détectée en utilisant un paramètre K égal à 3, alors qu'il détecte toutes les anomalies de type DDoS. Naturellement, la performance parfaite de ces signatures avec un K ègal à 3 est une artéfact de notre base de traces, et l'utiliser sur un réseau réel générerait éventuellement quelques (peu nombreux) faux positifs (comme cela sera montré plus loin sur l'analyse des traces japonaises).

Nous avons ensuite évalué l'expressivité de notre approche en utilisant les signatures autres que celles relatives aux DDoS du Tableau II sur les traces METROSEC. Grâce à ces signatures nous avons trouvé au total 16 port scans, 13 réponses à des attaques et 2471 scans réseaux (i.e. les scans réseaux sont relatifs à

des réseaux /24). Leur analyse manuelle a montré que tous les port scans sont de vrais positifs. Parmi les 13 réponses à des attaques, 9 sont des réponses à des attaques DDoS et 1 est une réponse à un scan réseau. Nous n'avons pas été en mesure d'identifier la nature des 3 autres réponses. Ces anomalies *indéfinies* sont composées de trafics ICMP multi-sources soudains, normalement venant du même réseau (même préfixe), et qui seraient logiquement des réponses à des tentatives de scan réseaux, mais qui sont dirigées vers des machines inexistantes. Il est assez invraisemblable que ces anomalies proviennent de tentatives d'*idlescan* [6], car on en trouve tout au long de l'année 2006, et il est très improbable qu'un pirate ait pu sniffer le réseau commuté du LAAS pour voir les réponses. Les scans réseaux n'ont pas tous été analysés manuellement (vu leur très grand nombre), mais la signature est sans aucun doute très précise et ne peut conduire à beaucoup d'erreurs de classification (a priori le taux d'erreur devrait être nul). Evidemment, en utilisant plus de signatures ou en abaissant les seuils, on trouverait certainement de nouvelles (vraies) anomalies, mais pourrait conduire à augmenter le taux de mauvaise classification de l'algorithme. Par exemple, si on change le seuil de *#rpkt/#rdstport* pour la signature des DDoS ICMP echos (voir Tableau II) à une valeur $> 15 * gr$, le faux négatif rencontré pour la trace 8 est alors correctement classifié, mais 3 réponses à des attaques sont classifiées comme des DDoS (i.e. cela ajoute donc 3 faux positifs). De façon similaire, pour la signature associée à la détection des réponses à des attaques du Tableau II, si on fixe le seuil de *findp* à une valeur $> 10 * gr$, on trouve une nouvelle réponse à un scan réseau et 4 nouvelles anomalies *indéfinies*. Même si ce ne sont certainement pas les vraies anomalies contenues dans cette trace, cette analyse préliminaire des signatures non relatives à des DDoS démontre qu'une classification fiable des différents types d'anomalies est toutefois possible. Ces résultats montrent aussi que différents types d'anomalies DDoS peuvent être précisément classifiées par notre approche, même en présence d'autres types d'anomalies. Un des points clés de ce travail est que le choix des signatures et seuils utilisés peuvent être définis par les administrateurs réseaux pour coller au mieux à leurs besoins.

Les résultats obtenus à partir des traces METROSEC sont très prometteuse, mais nous voulions être sûrs que les signatures n'étaient pas seulement adaptées aux caractéristiques de ce réseau (RENATER), en particulier pour les faux positifs. Nous avons utilisé exactement les mêmes signatures sur les traces MAWI. Par rapport aux anomalies de type DDoS, les tests sur les traces de 15 minutes ont donné un total de 19 vrais positifs, 3 faux positifs et 9 faux négatifs (connus), sur un total de plus de 2,5 millions d'anomalies détectées. Parmi ces 19 alarmes correctes, 6 n'avaient pas été identifiées précédemment par les travaux de [10] et consistaient principalement en de tous petits paquets IP (i.e. 34 octets, pas d'entête de niveau transport) et/ou des paquets avec les mêmes adresses source et destination. Les 3 faux positifs trouvés sont issus de la même adresse IP source et sont tous classifiés par par la signature ICMP Echo. Comme les machines destinatrices (chaque faux positif est pour une destination responsable différente) émettent des paquets de demande d'echo et que l'anomalie est détectée sur les réponses echo, il n'est pas clair que ce soit du trafic normal (i.e. elles peuvent être des attaques par réflexion). Le nombre de faux négatifs montre une lacune connue de notre algorithme de détection. Tous ces faux négatifs ont trait à des attaques qui avaient déjà démarré lorsque le processus de capture a été lancé, et ne présentaient pas de variation tout au long se son existence. Notre algorithme de détection manque complètement ces anomalies car il travaille sur des deltoïdes et aucune variation de trafic n'est créée par l'attaque (i.e. l'attaque a déjà stabilisé son débit au début de la capture, et varie peu ensuite). Ce n'est pas vraiment une lacune de notre algorithme de classification ou des signatures, car une analyse manuelle montre que ces anomalies auraient été correctement classifiées si elles avaient été détectées (i.e. elles sont essentiellement composées de paquets invalides et d'attaques TCP SYN). En considérant les autres types d'anomalies, l'algorithme a trouvé 4429 scans réseaux, 5233 port scans et 72 réponses à des attaques. Une analyse manuelle préliminaire des résultats montre que plusieurs de ces anomalies sont générées par des scan de vers (et les réponses associées), avec des variantes de Sasser et Dabber qui sont particulièrement fréquentes.

VI CONCLUSIONS

Dans cet article, nous avons présenté une nouvelle approche pour la classification automatique des anomalies du trafic réseau. Nous avons mis en évidence comment les capacités des algorithmes de détection existants pour l'identification des flux anormaux peuvent être utilisés pour collecter un riche ensemble d'in-

formations sur les anomalies, de sorte à rendre possible une classification précise. Ces informations sont représentées par un large ensemble d'attributs d'anomalies (i.e. métriques qui comportent les informations sur les anomalies) qui sont utilisés dans un module de classification orienté signature. Notre approche de classification est la première à offrir aux opérateurs réseaux la possibilité de pouvoir facilement comprendre et contrôler le processus complet d'étiquetage.

Nous avons défini un ensemble initial d'attributs d'anomalies et caractérisé différents types d'anomalies (e.g. DDoS, scans réseaux, etc.) en les utilisant. Nous avons aussi montré comment la classification automatique peut être utilisée comme un filtre pour réduire considérablement le nombre de faux négatifs des algorithmes de détection. Nous avons évalué notre travail en utilisant deux bases de traces différentes contenant du trafic normal et plusieurs anomalies. Les résultats obtenus illustrent l'expressivité de notre approche pour différencier différents types d'anomalies (DDoS et autres). Dans des travaux à venir, nous envisageons de voir comment d'autres algorithmes de détection et d'identifications peuvent être intégrés à notre approche de classification.

Références

[1] P. Abry, P. Borgnat, and G. Dewaele. Invited talk : Sketch based anomaly detection, identification and performance evaluation. *International Symposium on Applications and the Internet Workshops (SAINT)*, January 2007.

[2] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *Internet Measurment Workshop*, Marseille, November 2002.

[3] J. Barlow and W. Thrower. Tfn2k - an analysis, February 2000. `http ://packetstormsecurity.org/distributed/TFN2k%5FAnalysis-1.3.txt`.

[4] J. D. Brutlag. Aberrant behavior detection in time series for network monitoring. In *USENIX LISA*, New Orleans, December 2000.

[5] P. Chhabra, C. Scott, E. Kolaczyk, and M. Crovella. Distributed spatial anomaly detection. In *IEEE INFOCOM*, Phoenix, AZ, April 2008.

[6] T. Chmielarski. Intrusion detection faq : Reconnaissance techniques using spoofed ip addresses, April 2001. `http ://www.sans.org/resources/idfaq/spoofed%5Fip.php`.

[7] K. Cho, K. Mitsuya, and A. Kato. Traffic data repository at the wide project. In *USENIX ATEC*, San Diego, California, 2000.

[8] G. Cormode and S. Muthukrishnan. What's new : finding significant differences in network data streams. *IEEE/ACM Trans. Netw.*, 13(6) :1219–1232, 2005.

[9] D. Danchev. Coordinated russia vs georgia cyber attack in progress, August 2008. `http ://blogs.zdnet.com/security/ ?p=1670`.

[10] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, and K. Cho. Extracting hidden anomalies using sketch and non gaussian multiresolution statistical detection procedures. In *Workshop on Large-Scale Attack Defense (LSAD)*, Kyoto, Japan, 2007.

[11] D. Dittrich. The dos project's "trinoo" distributed denial of service attack tool, October 1999. `http ://staff.washington.edu/dittrich/misc/trinoo.analysis`.

[12] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. Netw.*, 4(2) :209–223, 1996.

[13] C. Estan, S. Savage, and G. Varghese. Automatically inferring patterns of resource consumption in network traffic. In *ACM SIGCOMM*, Karlsruhe, 2003.

[14] S. Farraposo, P. Owezarski, and E. Monteiro. A multi-scale tomographic algorithm for detecting and classifying traffic anomalies. In *IEEE ICC*, Glasgow, June 2007.

[15] S. Farraposo, P. Owezarski, and E. Monteiro. Detection, classification et identification d'anomalies de trafic. In *Colloque Francophone d'Ingenierie des Protocoles (CFIP)*, Les Arcs, France, March 2007.

[16] A. Feldmann, A. C. Gilbert, and W. Willinger. Data networks as cascades : investigating the multifractal nature of internet wan traffic. In *ACM SIGCOMM*, Vancouver, 1998.

[17] A. Hussain, J. Heidemann, and C. Papadopoulos. A framework for classifying denial of service attacks. In *ACM SIGCOMM*, Karlsruhe, 2003.

[18] J. Jung, B. Krishnamurthy, and M. Rabinovich. Flash crowds and denial of service attacks : Characterization and implications for cdns and web sites. In *WWW*, Honolulu, Hawaii, May 2002.

[19] M.-S. Kim, H.-J. Kong, S.-C. Hong, S.-H. Chung, and J. Hong. A flow-based method for abnormal network traffic detection. In *IEEE/IFIP Network Operations and Management Symposium*, Seoul, April 2004.

[20] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen. Sketch-based change detection : methods, evaluation, and applications. In *Internet Measurement Conference*, Miami Beach, FL, 2003.

[21] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM*, Portland, August 2004.

[22] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *ACM SIGCOMM*, Philadelphia, August 2005.

[23] A. Lakhina, M. Crovella, and C. Diot. Characterization of network-wide anomalies in traffic flows. In *Internet Measurement Conference*, Taormina, Italy, October 2004.

[24] X. Li, F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone, and A. Lakhina. Detection and identification of network anomalies using sketch subspaces. In *Internet Measurement Conference*, Rio de Janeiro, Brazil, 2006.

[25] J. McHugh. Testing intrusion detection systems : a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Trans. Inf. Syst. Secur.*, 3(4) :262–294, 2000.

[26] METROSEC. At `http ://www.laas.fr/METROSEC`.

[27] J. Mirkovic and P. Reiher. A taxonomy of ddos attack and ddos defense mechanisms. *SIGCOMM Comput. Commun. Rev.*, 34(2) :39–53, 2004.

[28] D. Moore, G. M. Voelker, and S. Savage. Inferring internet denial-of-service activity. In *USENIX SSYM*, Washington, D.C., 2001.

[29] P. Owezarski. On the impact of dos attacks on internet traffic characteristics and qos. *ICCCN*, October 2005.

[30] K. Park and W. Willinger. *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons, Inc., New York, NY, USA, 2000.

[31] A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat, and P. Abry. Non-gaussian and long memory statistical characterizations for internet traffic with anomalies. *IEEE Trans. Dependable Secur. Comput.*, 4(1) :56–70, 2007.

[32] J. Swartz and B. Acohido. Botnets can be used to blackmail targeted sites, March 2008. `http ://www.usatoday.com/tech/news/computersecurity/2008-03-16-botside%5FN.htm`.

[33] Y. Zhang, S. Singh, S. Sen, N. Duffield, and C. Lund. Online identification of hierarchical heavy hitters : algorithms, evaluation, and applications. In *Internet Measurement Conference*, Taormina, Italy, October 2004.