ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Πρόγραμμα Μεταπτυχιακών Σπουδών

Επιστήμη και Τεχνολογία της Πληροφορικής και των Υπολογιστών

Ειδίκευση Λογισμικού και Πληροφοριακών Συστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σύνολα Δεδομένων και Δοκιμές για Συστήματα Ανίχνευσης Εισβολών (Datasets and Testing for Intrusion Detection Systems - IDSs)

Ιωάννης Θ. Δεληγιαννίδης
Α.Μ. 19051

Εισηγητής: Δρ Ιωάννα Κανζάβελου, Επίκ. Καθηγήτρια

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σύνολα Δεδομένων και Δοκιμές για Συστήματα Ανίχνευσης Εισβολών

(Datasets and Testing for Intrusion Detection Systems - IDSs)

Ιωάννης Θ. Δεληγιαννίδης
Α.Μ. 19051

Εισηγητής:

Ιωάννα Καντζάβελου, Επίκ. Καθηγήτρια

Εξεταστική Επιτροπή:

Καντζάβελου Ιωάννα, Επίκ. Καθηγήτρια

Ioanna Kantzavelou
Digitally signed by Ioanna Kantzavelou
Date: 2021.07.27 18:35:49 +03'00'

Μάμαλης Βασίλειος, Καθηγητής

Vasileios Mamalis
Digitally signed by Vasileios Mamalis
Date: 2021.07.27 23:19:38 +03'00'

Μπόγρης Αντώνιος, Καθηγητής

Antonios Bogris
Digitally signed by Antonios Bogris
Date: 2021.07.28 09:13:03 +03'00'

Ημερομηνία εξέτασης 13/7/2021

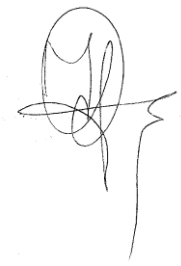
ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος ΙΩΑΝΝΗΣ ΔΕΛΗΓΙΑΝΝΙΔΗΣ του ΘΕΟΔΩΡΟΥ, με αριθμό μητρώου 19051 φοιτητής του Προγράμματος Μεταπτυχιακών Σπουδών ΕΠΙΣΤΗΜΗ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑ ΤΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

«Είμαι συγγραφέας αυτής της μεταπτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο Δηλών



Ιωάννης Δεληγιαννίδης

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία ολοκληρώθηκε μετά από επίμονες προσπάθειες, σε ένα ενδιαφέρον γνωστικό αντικείμενο, όπως αυτό των συνόλων δεδομένων και δοκιμές για Συστήματα Ανίχνευσης Εισβολών

Την προσπάθειά μου αυτή υποστήριξε η επιβλέπουσα καθηγήτριά μου, την οποία θα ήθελα να ευχαριστήσω.

Ακόμα θα ήθελα να ευχαριστήσω την οικογένειά μου που συμπαραστάθηκαν για την ολοκλήρωση των σπουδών στο μεταπτυχιακό πρόγραμμα.

ΠΕΡΙΛΗΨΗ

Ένα από τα πλέον σημαντικά και ιδιαίτερα θέματα που έχουν ανακύψει με την ραγδαία ανάπτυξη των ηλεκτρονικών υπολογιστών, εφαρμογών και δικτύων, ασύρματων και ενσύρματων, είναι η παράλληλη ανάπτυξη αποτελεσματικών συστημάτων ανίχνευσης εισβολών (IDS) με κατάλληλα και αποτελεσματικά σύνολα δεδομένων, για την διασφάλιση της ασφάλειας σε εμπιστευτικότητα, διαθεσιμότητα και ακεραιότητα των δεδομένων και διαδικασιών τόσο από εξωτερικές όσο και από εσωτερικές επιθέσεις. Οι προσπάθειες ανάπτυξης τέτοιων συστημάτων ξεκίνησαν το 1987 ως μία δεύτερη γραμμή άμυνας.

Για την επίτευξη του στόχου αυτού συνεχώς γίνεται προσπάθεια να αναπτυχθούν διάφοροι μηχανισμοί ανίχνευσης εισβολών – επιθέσεων, χρησιμοποιώντας μεθόδους που εντάσσονται στο ευρύτερο πεδίο της εξόρυξης γνώσης από δεδομένα μέσω μηχανικής μάθησης. Στο πλαίσιο αυτό πραγματοποιείται πλήθος διαρκών και επίπονων δοκιμών σε σύνολα δεδομένων κατασκευασμένα για το σκοπό αυτό, τα οποία θα πρέπει να είναι ενημερωμένα και επικαιροποιημένα με τους τελευταίους τύπους επιθέσεων. Τέτοια σύνολα δεδομένων είναι τα DARPA98, KDD99, NSL-KDD, ISC2012, ADFA13, CAIDA (2011), ICSI κ.α.

Σκοπός της παρούσας εργασίας είναι η ανάδειξη των προβλημάτων που δημιουργούνται από τα σύνολα δεδομένων κατά τη δοκιμή IDSs και των μειονεκτημάτων που συνεπάγεται η ύπαρξη και η εκδήλωση αυτών των προβλημάτων, μετά από ενδελεχή έρευνα και μελέτη των περισσότερο δημοφιλών και επεξεργασμένων δημόσιων συνόλων δεδομένων.

Προσδιορίζονται τα προβλήματα που παρατηρούνται στα σύνολα δεδομένων και πως αυτά περιορίζουν την αποτελεσματικότητα των συστημάτων ανίχνευσης εισβολών.

Στη συνέχεια προτείνονται λύσεις των προαναφερόμενων προβλημάτων και επισημαίνονται πεδία που χρήζουν περαιτέρω διερεύνησης και τέλος παρουσιάζονται τα συμπεράσματα της παρούσας εργασίας.

ABSTRACT

One of the most important and particular issues that have arisen with the rapid development of computers, applications and networks, wireless and Wired, is the parallel development of effective intrusion detection systems (IDS) with appropriate and effective data sets, to ensure security in confidentiality, availability and integrity of data and processes from both external and internal attacks. Efforts to develop such systems began in 1987 as a second line of Defense.

In order to achieve this goal, various intrusion – attack detection mechanisms are constantly being developed, using methods that are part of the wider field of knowledge extraction from data through machine learning. In this context, a number of constant and painstaking tests are carried out on datasets built for this purpose, which should be up to date and up to date with the latest types of attacks. Such data sets are DARPA98, KDD99, NSL-KDD, ISC2012, ADFA13, CAIDA (2011), ICSI, etc.

The purpose of this paper is to highlight the problems created by datasets during the IDSs test and the disadvantages involved in the existence and manifestation of these problems, after thorough research and study in the most popular and processed public datasets.

The problems observed in data sets and how these limit the effectiveness of intrusion detection systems were identified.

Then proposed solutions of the aforementioned problems and highlighted fields that need further investigation and finally presented the conclusions of this paper.

ΕΠΙΣΤΗΜΟΝΙΚΗ ΠΕΡΙΟΧΗ:

Ασφάλεια πληροφοριακών και επικοινωνιακών συστημάτων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ:

Intrusion Detection System (IDS), Dataset

Περιεχόμενα

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

- ABC : Association Based Classification
- ANN: Artificial Neural Network
- APD : Anomaly Pattern Detection
- ART : Adaptive Resonance Theory
- AUC : area under the ROC curve
- BFS-CFS: Best First Search with Correlation Features Selection
- BON : Back-Propagation Network
- BSPNN : Boosted Subspace Probabilistic Neural Network
- CAMNEP : The Cooperative Adaptive Mechanism for Network Protection
- CSOACN : Clustering based on Self-Organized Ant Colony Network
- CUSUM : CUmulative SUM
- CAMNEP The Cooperative Adaptive Mechanism for Network Protection
- DL : Deep Learning
- DoS : Denial of Service
- DT : Decision Tree
- ELM : Extreme Learning Machine
- ENN : Elman Neural Network
- FCM : Fuzzy C-Mean
- FFNN : Feed Forward Neural Network
- GMDH : Group Method for Data Handling
- GR : Gain Ratio * U2R: User to Root
- GRNN : Generalized Regression Neural Network
- GS-CFS : Greedy Stepwise with Correlation Features Selection
- GT : Ground Truth
- FLN : Fast Learning Network
- IRL : Iterative Rule Learning
- k-NN : k-Nearest Neighbors
- MLP : Multi-Layer Perceptron
- NB : Naive Bayes
- NDAE : Non-Symmetric Deep Auto-Encoder
- NN : Neural Network
- N/A : Not Available
- OCSVM: One Class Support Vector Machine
- PCA : Principal Component Analysis
- PNN : Probabilistic Neural Network
- PSO : Particle Swarm Optimization
- R2L : Remote to Local
- RBF : Radial Basis Function
- RBNN: Radial Basis Neural Network
- RNN : Recurrent Neural Networks
- SA : Simulated Annealing
- SOM : Self-Organizing Map
- SVDD : Support Vector Data Description
- SVM : Support Vector Machine
- VSM : vector space model
- WSARE: WhatβĂŹs Strange About Recent Events
- XSS : Cross Site Scripting

1 Κεφάλαιο : Εισαγωγή.

Με την ραγδαία ανάπτυξη των ηλεκτρονικών υπολογιστών, εφαρμογών, δικτύων υπολογιστών, κινητών συσκευών, με υψηλή συνδεσιμότητα, καθώς και την αύξηση των εφαρμογών που εκτελούνται σε αυτά όπως ηλεκτρονικό εμπόριο, ηλεκτρονική επιχειρηματικότητα, ηλεκτρονική εκπαίδευση κ.α. , η συνεχή εξέλιξη των στρατηγικών επίθεσης, θέματα ασφαλείας και τρωτά σημεία των συστημάτων, καθιστούν την ανάπτυξη συστημάτων ανίχνευσης εισβολών (IDS - NIDS) κρίσιμο ζήτημα το οποίο έχει απασχολήσει τους ειδικούς στην ασφάλεια συστημάτων. Οι επιθέσεις σε υπολογιστές ή και συστήματα δικτύων, ασύρματων ή ενσύρματων, στοχεύουν στην αποσταθεροποίηση τους, θέτοντας σε κίνδυνο την ασφάλεια σε εμπιστευτικότητα, διαθεσιμότητα ή ακεραιότητα τους.

Οι ειδικοί στην ασφάλεια έχουν προτείνει πολλές λύσεις, ωστόσο, τα τρέχοντα εργαλεία συχνά αποτυγχάνουν να προσαρμοστούν στις συνεχώς μεταβαλλόμενες αρχιτεκτονικές των συστημάτων και των στρατηγικών των επιθέσεων.

Τα υπάρχοντα μέχρι σήμερα συστήματα ανίχνευσης εισβολών (IDS) υπολογίζεται ότι καλύπτουν περίπου το 25% των απειλών, ενώ τα τρέχοντα σύνολα δεδομένων δείχνουν την σαφή έλλειψη απειλών πραγματικού δικτύου και αποτύπωσης επιθέσεων, αφού συχνά είναι ξεπερασμένα , δεν αντικατοπτρίζουν την πραγματικότητα των επιθέσεων, καθιστώντας αυτά, αναποτελεσματικά απέναντι σε νέες απειλές και απειλές μηδενικών ημερών.

1.1 Περιγραφή του αντικειμένου της διπλωματικής εργασίας

Σκοπός της παρούσας εργασίας είναι η ανάδειξη των προβλημάτων που δημιουργούνται από τα σύνολα δεδομένων κατά τη δοκιμή IDSs και των μειονεκτημάτων που συνεπάγεται η ύπαρξη και η εκδήλωση αυτών των προβλημάτων, μετά από ενδελεχή έρευνα και μελέτη στα περισσότερο δημοφιλών και επεξεργασμένων δημόσιων συνόλων δεδομένων.

Προσδιορίσθηκαν τα προβλήματα που παρατηρούνται στα σύνολα δεδομένων και πως αυτά περιορίζουν την αποτελεσματικότητα των συστημάτων ανίχνευσης εισβολών.

Στη συνέχεια προτείνονται λύσεις των προαναφερόμενων προβλημάτων και επισημαίνονται πεδία που χρήζουν περαιτέρω διερεύνησης και τέλος παρουσιάζονται τα συμπεράσματα της παρούσας εργασίας.

1.2 Η λειτουργία των συστημάτων ανίχνευσης εισβολών.

Ένα σύστημα ανίχνευσης εισβολών παρακολουθεί συμβάντα σε ένα σύστημα υπολογιστή (IDS - HIDS) ή σε δίκτυο υπολογιστών (NIDS) και αναλύει πιθανά σημάδια εισβολών [1] την στιγμή που συμβαίνουν ή έχουν συμβεί καθώς και μπορεί να διακόψει συνδέσεις δικτύου, να καταγράψει συμβάντα και να ειδοποιήσει διαχειριστές συστημάτων για να λάβουν τα κατάλληλα μέτρα. [2], [3].

Τα συστήματα ανίχνευσης εισβολής δικτύου (NIDS) αφορούν τη διασφάλιση του δικτύου τόσο από εξωτερικές όσο και από εσωτερικές επιθέσεις. Τα NIDS ταξινομούνται κυρίως σε Online και Offline σύμφωνα με την λειτουργία τους. Σε απευθείας σύνδεση ή σε πραγματικό χρόνο NIDS, όπως Snort, Bro, κλπ., εξετάζουν τη δομή του πακέτου για να εντοπίσουν πιθανές εισβολές και αν συμβεί αυτό παράγουν ειδοποιήσεις προς το διαχειριστή για να λάβει τα απαραίτητα μέτρα. Από την άλλη πλευρά, τα NIDS εκτός σύνδεσης καταγράφουν τα πακέτα που ρέουν προς και από το δίκτυο, κατασκευάζουν χαρακτηριστικά που βασίζονται σε συνδέσεις και δημιουργούν ένα σύνολο δεδομένων. Τέτοια σύνολα δεδομένων NIDS χρησιμοποιούνται σε ερευνητικούς σκοπούς για την εφαρμογή εξόρυξης δεδομένων, μηχανικής μάθησης, εξελικτικών αλγορίθμων κλπ., για την ανίχνευση επιθέσεων.

1.3 Κατηγορίες συστημάτων ανίχνευσης εισβολών

Τα συστήματα ανίχνευσης εισβολών (IDS) μπορούν να ταξινομηθούν βάση της πηγής που προέρχονται, σε Host-based και Network-based. Τα συστήματα host-based βασίζουν τις αποφάσεις τους σε πληροφορίες που λαμβάνονται από έναν μόνο κεντρικό υπολογιστή ή ακόμη και μια ενιαία εφαρμογή και εξαρτάται από δεδομένα που μπορούν να εντοπιστούν από το λειτουργικό σύστημα του ελεγχόμενου, συνήθως αρχεία καταγραφής, χρήση πόρων, κίνηση δικτύου προς και από τον κεντρικό υπολογιστή ή πληροφορίες σχετικά με τις διαδικασίες που εκτελούνται στον κεντρικό υπολογιστή, [19], [20], ενώ τα συστήματα Network-based λαμβάνουν δεδομένα παρακολουθώντας την κίνηση του δικτύου μεταξύ κεντρικών υπολογιστών με την βοήθεια ενός εργαλείου sniffer και συνήθως

εκτελούνται σε ξεχωριστό μηχάνημα [21], [22], προσπαθώντας να ανιχνεύσουν πακέτα που θα μπορούσαν να είναι μέρος μιας επίθεσης.

Μία άλλη κατηγοριοποίηση των συστημάτων ανίχνευσης εισβολών είναι με βάση τρεις διαφορετικές τεχνικές ανίχνευσης που είναι πολύ σημαντικές :

- ανίχνευση βάσει ανωμαλιών,
- ανίχνευση βάσει υπογραφής - κακής χρήσης και
- ανίχνευση βάσει προδιαγραφών.

Τα IDS με βάση την ανίχνευση ανωμαλιών προσπαθούν να δημιουργήσουν ένα μοντέλο κανονικής συμπεριφοράς για τα παρακολουθούμενα συστήματα ή για τους χρήστες τους, προκειμένου να επισημαίνουν ως ύποπτη οποιαδήποτε απόκλιση από αυτή την "κανονική" συμπεριφορά που υπερβαίνει τα όρια. Εάν παρατηρηθεί μεγάλη απόκλιση, οι ενέργειες ταξινομούνται ως ύποπτες και καταγράφονται. Τα συστήματα που βασίζονται σε ανωμαλίες έχουν γενικά καλή ικανότητα ανίχνευσης επιθέσεων μηδενικής ημέρας, αλλά πάσχουν από υψηλά ποσοστά ψευδών συναγερμών. Αυτό συμβαίνει λόγω της δυσκολίας δημιουργίας ακριβών βασικών γραμμών σε ένα εξαιρετικά δυναμικό σύγχρονο περιβάλλον του λειτουργικού συστήματος [26].

Μπορούν να ταξινομηθούν σε υποκατηγορίες με βάση τη χρησιμοποιούμενη τεχνική κατάρτισης. Αυτές οι κατηγορίες αναγνωρίζονται αντίστοιχα ως στατιστικές, βασισμένες στη γνώση και βασισμένες στη μηχανική μάθηση.

Οι στατιστικές τεχνικές ανίχνευσης περιλαμβάνουν χρονοσειρές.

Τεχνικές κατάρτισης με βάση τη γνώση χρησιμοποιούν μηχανές πεπερασμένων καταστάσεων και κανόνων όπως συστήματα βασισμένα σε περιπτώσεις, βασισμένα σε εξειδικευμένα συστήματα και γλώσσες περιγραφής.

Τεχνικές μηχανικής μάθησης – κατάρτισης, περιλαμβάνουν τεχνητά νευρωνικά δίκτυα, ομαδοποίηση, γενετικούς αλγόριθμους, βαθιά μάθηση κ.α.

Τα IDS βάσει υπογραφής (με βάση την κακή χρήση), χρησιμοποιούν μια βάση γνώσεων - σύνολο υπογραφών από προηγούμενη ανίχνευση, για να αναγνωρίσουν άμεσα τις προσπάθειες εισβολής, το οποίο σημαίνει ότι αντί να προσπαθεί να περιγράψει την κανονική συμπεριφορά ενός συστήματος προσπαθεί να περιγράψει τις ανώμαλες συμπεριφορές. Το κύριο πλεονέκτημα αυτής της

τεχνικής είναι η υψηλή ακρίβεια για γνωστές επιθέσεις, ωστόσο, το IDS δεν είναι σε θέση να ανιχνεύσει απειλές μηδενικής ημέρας ή πολυμορφικές απειλές.

Τα IDS με βάση τις προδιαγραφές συνδυάζουν τις ιδιότητες τόσο της τεχνικής βάσει ανωμαλίας όσο και της τεχνικής βάσει υπογραφής για να σχηματίσει ένα υβριδικό μοντέλο.

Οι διαφορές μεταξύ των δύο προσεγγίσεων αναφέρεται στον ακόλουθο πίνακα (πίνακας 1)

Πίνακας 1
Διαφορές των δύο διαφορετικών προσεγγίσεων των IDS

Βάσει Κατάχρησης	**Βάσει Ανωμαλίας**
Απαιτούν συνεχείς ενημερώσεις	Δεν απαιτούν ενημερώσεις
Δεν υπάρχει αρχική εκπαίδευση	Μακρά και πολύπλοκη εκπαίδευση
Χρειάζεστε ρύθμιση συντονισμού	Η ρύθμιση περιλαμβάνεται στην εκπαίδευση
Δεν είναι δυνατή η ανίχνευση νέων επιθέσεων	Μπορεί να ανιχνεύσει νέες επιθέσεις
Ακριβείς ειδοποιήσεις	Ασαφείς ειδοποιήσεις
Σχεδόν καθόλου ψευδώς θετικά	Τεράστιοι αριθμοί ψευδών θετικών
Ευκολότερο να σχεδιάσει	Πιο δύσκολο να σχεδιάσει

1.4 Τεχνικές ανίχνευσης εισβολών

Για την ανίχνευση επιθέσεων έχουν προταθεί διάφορες τεχνικές χρησιμοποιώντας συμβατικές στατιστικές τεχνικές καθώς και τεχνικές εξόρυξης δεδομένων που συνήθως περιλαμβάνουν εποπτευόμενες, μη εποπτευόμενες τεχνικές και ακραίες τιμές.

Μέχρι σήμερα τα περισσότερα από τα προτεινόμενα IDS συνήθως αντιμετωπίζουν προβλήματα όπως χαμηλή ακρίβεια ανίχνευσης, υψηλό ποσοστό ψευδών συναγερμών, χαμηλή απόδοση σε πραγματικό χρόνο, περιορισμένη επεκτασιμότητα, λόγω των προβλημάτων που παρουσιάζουν τα σύνολα δεδομένων που χρησιμοποιούνται για επεξεργασία, τα οποία μεταφέρονται σε αυτά, όπως ο τεράστιος όγκος πληροφορίας που περιλαμβάνουν, μεγάλο πλήθος χαρακτηριστικών κ.α. Επομένως ο σχεδιασμός αποτελεσματικών συνόλων

δεδομένων καθώς και η συνεχή ενημέρωσή τους, είναι σημαντικά θέματα για τη βελτίωση της απόδοσης τους των συστημάτων ανίχνευσης εισβολών.

Παρόλο που η ανίχνευση εισβολών την στιγμή που συμβαίνουν (σε πραγματικό χρόνο) είναι ένα σημαντικό χαρακτηριστικό των συστημάτων ανίχνευσης εισβολών, τα περισσότερα από αυτά λειτουργούν σε εκτός σύνδεση κατάσταση λόγω της ανάγκης ανάλυσης του τεράστιου όγκου δεδομένων δραστηριότητας, της συνεχούς εξέλιξης των στρατηγικών επιθέσεων, της πολυπλοκότητας των δικτύων και της ταχύτητας τους, την απαίτηση σε εξαντλητικές δοκιμές, της αξιολόγησης και του συντονισμού αυτών. Το γεγονός αυτό δίνει την δυνατότητα της σε βάθος ανάλυση των προτύπων και των συμπεριφορών επιθέσεων – εισβολών που προκύπτουν από τα συλλεγόμενα ή προϋπάρχοντα σύνολα δεδομένα και κατά προέκταση, την ευκαιρία για την δημιουργία και τον έλεγχο αλγορίθμων ανίχνευσης εισβολών για μελλοντικά συμβάντα επίθεσης.

Τα IDS που λειτουργούν σε δίκτυο (NIDS) χρησιμοποιούν συστήματα καταγραφής κίνησης τα οποία τοποθετούνται σε υπολογιστικές μονάδες, switches, routers, προκειμένου να παρακολουθούν και να καταγράφουν την κυκλοφορία στο δίκτυο, με στόχο την ανάλυση και δημιουργία αναφορών για εντοπισμό πιθανών επιθέσεων - παραβιάσεων.

Τα περισσότερα από αυτά τα συστήματα, εκτελούν ανίχνευση με βάση την υπογραφή που συνδυάζονται με υπογραφές γνωστών επιθέσεων. Αυτό παρουσιάζει συνήθως πρόβλημα, διότι δεν μπορούν να ανταποκριθούν στην συνδεσιμότητα δικτύων υψηλών ταχυτήτων. Έτσι, υπάρχει η πιθανότητα ορισμένα κρίσιμα πακέτα να παραλειφθούν από το λογισμικό και ως συνέπεια και η ανίχνευση εισβολής. Επίσης, τέτοια IDS δεν μπορούν να αντιμετωπίσουν τις μηδενικής ημέρας επιθέσεις που παρατηρούνται την στιγμή που συμβαίνουν.

Η βασική λειτουργία ενός IDS είναι να μπορεί να ειδοποιεί για τις απόπειρες εισβολής. Τα ποσοστά εμφάνισης αυτών των ειδοποιήσεων τα οποία αναφέρονται στον πίνακα 2, καθορίζουν την αποτελεσματικότητα του συστήματος.

Οι ψευδώς θετικές ειδοποιήσεις σε ένα IDS με βάση την ανίχνευση ανωμαλίας είναι ένα γνωστό και σύνηθες πρόβλημα, έχοντας αποτέλεσμα να το καθιστά αναξιόπιστο και να μην χρησιμοποιείται, ενώ τα συστήματα που βασίζονται στην υπογραφή συνήθως δεν έχουν σημαντικό ποσοστό ψευδών θετικών.

Τα ψευδώς θετικά μπορούν επίσης να δημιουργήσουν συνθήκες άρνησης εξυπηρέτησης νόμιμων χρηστών (DoS, DDoS).

Οι ψευδώς αρνητικές ειδοποιήσεις αποτελούν κι αυτές πρόβλημα. Στα συστήματα με βάση την κακή χρήση, οι περισσότερες νέες επιθέσεις δημιουργούν τέτοιου είδους ειδοποιήσεις, εκτός αν είναι πολύ παρόμοιες με μια υπάρχουσα επίθεση.

Πίνακας 2
Ειδοποιήσεις συστημάτων ανίχνευσης εισβολών

Τύπος	Περιγραφή
True Positives	ειδοποιήσεις που προκαλούνται όταν παρουσιάζονται προσπάθειες εισβολής
False Positives	ειδοποιήσεις που προκαλούνται ενώ δεν παρουσιάζονται προσπάθειες εισβολής
True Negative	δεν υπάρχουν ειδοποιήσεις όταν δεν υπάρχουν προσπάθειες εισβολής
False Negative	δεν υπάρχουν ειδοποιήσεις ενώ υπάρχουν προσπάθειες εισβολής

1.5 Δομή της διπλωματικής εργασίας.

Η συνέχεια της εργασίας οργανώνεται ως εξής :

Στο κεφάλαιο 2 παρουσιάζεται μία σύντομη περιγραφή για τα θέματα που δυσχεραίνουν το έργο των συστημάτων ανίχνευσης εισβολών (IDS)

Στο κεφάλαιο 3 παρουσιάζεται μία ανάλυση των περισσότερο δημοφιλών και επεξεργασμένων συνόλων δεδομένων που διατίθενται δημόσια, και των θεμάτων που παρατηρήθηκαν σε αυτά.

Στο κεφάλαιο 4, συγκεντρώνονται και παρουσιάζονται τα προβλήματα που παρατηρήθηκαν από την μελέτη των συνόλων δεδομένων που αναφέρονται στο κεφάλαιο 3.

Στο κεφάλαιο 5 παρουσιάζονται οι προτεινόμενες λύσεις για την αντιμετώπιση των προβλημάτων που παρατηρήθηκαν και παρουσιάστηκαν και τέλος,

Στο κεφάλαιο 6 αναφέρονται τα συμπεράσματα της παρούσας εργασίας.

2 Κεφάλαιο : Θέματα που δυσχεραίνουν το έργο των συστημάτων ανίχνευσης εισβολών

Οι συνεχώς εξελισσόμενες στρατηγικές εισβολών, νέα σφάλματα και τρωτά σημεία στις εφαρμογές και στα δίκτυα, θέματα ασφάλειας και εμπιστευτικότητας, η προστασία των προσωπικών δεδομένων, απασχολούν την ερευνητική κοινότητα στους τομείς των συστημάτων ανίχνευσης εισβολής, σχεδιάζοντας συνεχώς νέες τεχνικές ανίχνευσης για να μειώσουν τις συνέπειες στην ασφάλεια. Η δημιουργία ή η πρόσβαση σε κατάλληλα συνόλων δεδομένων για την αξιολόγηση διαφόρων ερευνητικών σχεδίων σε αυτούς τους τομείς αποτελεί σημαντική πρόκληση.

Η δημιουργία συνόλων δεδομένων από τους ερευνητές έχει συνήθως τον χαρακτήρα και την κατεύθυνση του στόχου της έρευνας που πραγματοποιείται. Το περιβάλλον δημιουργίας αυτών των συνόλων είναι περιορισμένο και ελεγχόμενο, οι κινήσεις που παράγονται είναι κυρίως προσομοίωση επιθέσεων με διάφορα εργαλεία, στην προσπάθειά τους να επιτύχουν την όσο το δυνατόν πιο ρεαλιστικό το περιβάλλον συλλογής του και να αντικατοπτρίζουν την κίνηση στην πραγματικότητα, αλλά αυτό τις περισσότερες από τις περιπτώσεις δεν είναι επιτυχές.

Ως αποτέλεσμα αυτών οι περισσότερες τεχνικές ανίχνευσης εισβολής που χρησιμοποιούν υπάρχοντα δημόσια σύνολα δεδομένων, αξιολογούνται χρησιμοποιώντας λανθασμένες κατηγορίες από αυτά επειδή οι περιορισμοί κάθε κατηγορίας είναι άγνωστοι.

Με βάση τη μελέτη των πιο δημοφιλών και όχι μόνο, συνόλων δεδομένων που περιγράφονται στο επόμενο κεφάλαιο της παρούσας εργασίας, διαπιστώνουμε ότι πολλά από αυτά είναι ξεπερασμένα ή αναξιόπιστα για να χρησιμοποιηθούν. Τα περισσότερα από αυτά πάσχουν από έλλειψη ποικιλομορφίας και όγκο κυκλοφορίας, ενώ ορισμένα από αυτά δεν καλύπτουν ποικιλία επιθέσεων, περιέχουν ανώνυμα δεδομένα πακέτων και ωφέλιμου φορτίου τα οποία δεν αντανακλούν τις τρέχουσες τάσεις, ή στερούνται από πλήθος ή καταλληλόλητα χαρακτηριστικών και μεταδεδομένων.

3 Κεφάλαιο : Σύνολα δεδομένων.

Όπως αναφέρθηκε και παραπάνω, τα σύνολα δεδομένων διαδραματίζουν σημαντικό ρόλο στη δοκιμή και την επικύρωση των τεχνικών ή των συστημάτων ανίχνευσης εισβολών. Ένα σύνολο δεδομένων καλής ποιότητας όχι μόνο επιτρέπει τον προσδιορισμό της ικανότητας μιας τεχνικής ή ενός συστήματος να ανιχνεύει ανώμαλες συμπεριφορές, αλλά παρέχει δυνητική αποτελεσματικότητα όταν αναπτύσσεται σε πραγματικά λειτουργικά περιβάλλοντα.

Τα σύνολα δεδομένων μπορούν να ταξινομηθούν σε συνθετικά και πραγματικής ζωής.

Τα συνθετικά σύνολα δεδομένων δημιουργούνται για να καλύψουν συγκεκριμένες ανάγκες ή συγκεκριμένες συνθήκες ή δοκιμές που ικανοποιούν τα πραγματικά δεδομένα. Τέτοια σύνολα δεδομένων είναι χρήσιμα κατά το σχεδιασμό οποιουδήποτε πρωτότυπου συστήματος για θεωρητική ανάλυση με στόχο την βελτιστοποίηση του σχεδιασμού. Ένα συνθετικό σύνολο δεδομένων μπορεί να χρησιμοποιηθεί για τη δοκιμή και τη δημιουργία πολλών διαφορετικών τύπων σεναρίων δοκιμών, το οποίο επιτρέπει στους σχεδιαστές να δημιουργήσουν όσο το δυνατόν ρεαλιστικότερα προφίλ συμπεριφοράς για τους κανονικούς χρήστες και τους επιτιθέμενους με βάση το σύνολο δεδομένων για να δοκιμάσουν ένα προτεινόμενο σύστημα. Αυτό παρέχει την αρχική επικύρωση μιας συγκεκριμένης τεχνικής ή ενός συστήματος και αν τα αποτελέσματα αποδειχθούν ικανοποιητικά, οι ειδικοί στην ασφάλεια συνεχίζουν να σχεδιάζουν και να αξιολογούν τεχνικές ή συστήματα για ένα συγκεκριμένο τομέα δεδομένων πραγματικής κυκλοφορίας.

Τα σύνολα δεδομένων πραγματικής ζωής ουσιαστικά δεν είναι δημοσίως διαθέσιμα λόγω ζητημάτων ασφάλειας για τους δημιουργούς τους. Τα λίγα, υπάρχοντα σύνολα δεδομένων, προέρχονται από καταγραφή κίνησης σε πραγματικά δίκτυα για κάποιο χρονικό διάστημα, που όμως έχουν υποστεί επεξεργασία για την εκκαθάριση των ευαίσθητων δεδομένων που θα επηρέαζαν την ασφάλεια του δικτύου και των συστημάτων από όπου προήλθαν. Τα πλεονεκτήματα αυτών των συνόλων δεδομένων είναι πολύ περισσότερα από αυτά των συνθετικών, διότι περιέχουν πραγματική κίνηση κανονική ή μεγάλου εύρους τύπων επίθεσης. Συνήθως τα υπάρχοντα σύνολα δεδομένων πραγματικής κίνησης

δημιουργήθηκαν με τη συλλογή της κυκλοφορίας δικτύου σε αρκετές διαδοχικές ημέρες από δίκτυα πανεπιστημιουπόλεων.

3.1 Σύνολο δεδομένων DARPA (Lincoln Laboratory 1998,1999, 2000)

Τα σύνολα δεδομένων DARPA (1998, 1999, 2000) [74], [75], [76], δημιουργήθηκαν στα εργαστήρια MIT Lincon, με την εισαγωγή μη αυτόματων επιθέσεων που βασίζονται στο δίκτυο. Στόχος του είναι η ανάλυση της ασφάλειας του δικτύου.

Περιλαμβάνουν δραστηριότητες ηλεκτρονικού ταχυδρομείου, περιήγησης, FTP, Telnet, IRC και SNMP. Περιέχει επιθέσεις όπως DoS, Guess password, Buffer υπερχείλιση, απομακρυσμένη πρόσβαση FTP, Syn πλημμύρα, Nmap και Rootkit.

Ειδικότερα το DARPA 99 αποτελείται από περίπου τέσσερα GB συμπιεσμένων ακατέργαστων (δυαδικών) εγγραφών tcpdump. Προέρχεται από δοκιμές που διενεργήθηκαν για διάστημα μηνών βάση δειγματοληψίας της Αμερικάνικης Πολεμικής Αεροπορίας, ενώ ελήφθησαν υπόψη μέρος αυτών των βάσεων δεδομένων λόγω του απόρρητου των πληροφοριών. Η δημιουργία των χαρακτηριστικών περιορίσθηκε στην δημιουργία προφίλ τόσο για το υπόβαθρο όσο και για τα δεδομένα των εισβολών. Θεωρήθηκε ότι τα συλλεγμένα στοιχεία είναι ικανά να αντικατοπτρίσουν την πραγματική κίνηση σε πραγματικό χρόνο. Για την συλλογή των δεδομένων χρησιμοποιήθηκε εργαλεία όπως το TCPflap.

Σύμφωνα με τα [8], [9] το σύνολο δεδομένων μελετήθηκε τόσο για την διαδικασία δημιουργίας του, όσο και για την συλλογή κυκλοφορίας που αποτυπώνει.

Εντοπίσθηκαν στοιχεία από αντικείμενα προσομοίωσης που θα μπορούσαν να οδηγήσουν σε υπερεκτίμηση των επιδόσεων ορισμένων τεχνικών ανίχνευσης ανωμαλιών.

Σύμφωνα με τους παραπάνω το σύνολο δεδομένων δεν υποβλήθηκε σε αναλυτική ή πειραματική επικύρωση των στοιχείων του για ψευδή χαρακτηριστικά συναγερμού.

Τα συνθετικά δεδομένα δεν φαίνεται να είναι παρόμοια με την κυκλοφορία σε πραγματικά δίκτυα.

Κατά την διαδικασία συλλογής κυκλοφορίας με το εργαλείο TCPflap, το οποίο χρησιμοποιήθηκε στο DARPA 98, υπάρχει ενδεχόμενο να μην ελήφθησαν πακέτα

με μεγάλο φορτίο κυκλοφορίας, για το οποίο δεν πραγματοποιήθηκε κάποιος έλεγχος ως προς την απώλεια ή τις συνέπειας της απώλειας τέτοιων πακέτων.

Δεν υπάρχει ακριβής ορισμός ή ειδικοί ορισμοί των επιθέσεων. Περιπτώσεις πακέτων που προκαλούν υπερχείλιση θεωρήθηκαν επίθεση, κάτι το οποίο δεν ισχύει πάντα.

Υπάρχουν ορισμένες κριτικές για τις ταξινομήσεις επίθεσης και για το μέτρο απόδοσης.

Το μέτρο επιδόσεων που εφαρμόστηκε για αξιολόγηση στο DARPA 98 καθώς και οι καμπύλες ROC, έχουν αμφισβητηθεί σε μεγάλο βαθμό, ενώ έχουν προταθεί από τότε νέα μέτρα για την αντιμετώπιση των ελλείψεων [10], [11], [12], [13], [14].

Η ανάλυση των επιθέσεων έδειξε ότι αρκετές δεν ταιριάζουν σε καμία από αυτές τις κατηγορίες, οι οποίες είναι πιθανό να προκαλούνται από αντικείμενα προσομοίωσης, καθώς και ορισμένες επιθέσεις μπορούν να ταυτοποιηθούν με ανώμαλες διευθύνσεις IP πηγής ή ανωμαλίες στο πεδίο μεγέθους παραθύρου TCP.

3.2 Σύνολο δεδομένων KDD CUP 99

Το σύνολο δεδομένων KDD'99 [4] το οποίο δημιουργήθηκε το 1999 από την Lincoln Labs στο MIT ως συνέχεια του προηγούμενου (1998), αποτελεί το βασικό σύνολο δεδομένων για την αξιολόγηση των μεθόδων ανίχνευσης ανωμαλιών το οποίο βασίστηκε [5] στα δεδομένα του DARPA'98. Το DARPA'98 [7] αποτελείται από 4 GB συμπιεσμένων δεδομένων tcp dump τα οποία συγκεντρωθήκαν σε διάστημα 7 εβδομάδων κυκλοφορίας δικτύου. Το σύνολο δεδομένων εκπαίδευσης KDD αποτελείται από περίπου 4,9 εκατομμύρια διανύσματα μεμονωμένης σύνδεσης, καθένα από τα οποία περιέχει 41 χαρακτηριστικά κανονικής ή χαρακτηριστικά επίθεσης, με συγκεκριμένο τύπο. Οι προσομοιωμένες εισβολές ανήκουν σε μία από τις παρακάτω κατηγορίες [17] :

1. Denial of Service Attack (DoS) : Επιθέσεις κατά τις οποίες ο εισβολέας προκαλεί υπερφόρτωση σε πόρους υπολογιστών ή μνήμης προκαλώντας αδυναμία εξυπηρέτησης νόμιμων αιτημάτων ή άρνησης πρόσβασης νόμιμων χρηστών.
2. User to Root Attack (U2R) : Επιθέσεις κατά τις οποίες ο εισβολέας ξεκινάει με την πρόσβαση με χρήση κανονικού λογαριασμού εσωτερικού χρήστη στο σύστημα την οποία υπέκλεψε με κάποια τεχνική και εντοπίζει, εκμεταλλεύεται πιθανές ευπάθειες ή αποκτάει πρόσβαση στο root του συστήματος.

3. Remote to Local Attack (R2L) : Επιθέσεις κατά τις οποίες ο εισβολέας αποκτά την δυνατότητα να στείλει πακέτα σε ένα μηχάνημα μέσω δικτύου από μακριά, αλλά δεν έχει λογαριασμό σε αυτό το μηχάνημα, εκμεταλλευόμενος κάποια ευπάθεια για να αποκτήσει τοπική πρόσβαση ως χρήστης αυτού του συστήματος.
4. Probing Attack : Επιθέσεις παρακολούθησης κατά τις οποίες γίνεται προσπάθεια συγκέντρωσης πληροφοριών δικτύου υπολογιστών με σκοπό την παράκαμψη των ελέγχων ασφαλείας του.

Τα δεδομένα δοκιμών αποτελούνται από 2 εκατομμύρια εγγραφές σύνδεσης που καταγράφηκαν για διάστημα δύο εβδομάδων, δεν προέρχονται από την ίδια κατανομή πιθανότητας με τα δεδομένα εκπαίδευσης και περιλαμβάνουν συγκεκριμένους τύπους επίθεσης. Θεωρείται ότι οι περισσότερες νέες επιθέσεις είναι παραλλαγές γνωστών επιθέσεων και η υπογραφή γνωστών επιθέσεων μπορεί να εντοπίσει νέες παραλλαγές τους.

Το σύνολο δεδομένων περιέχει συνολικά 24 τύπους επίθεσης κατάρτισης, με επιπλέον 14 τύπους μόνο στα δεδομένα δοκιμών [6].

Τα χαρακτηριστικά του συνόλου δεδομένων KDD 99 μπορούν να ταξινομηθούν σε τρεις ομάδες:

1. Βασικά χαρακτηριστικά : Ενσωματώνουν όλα τα χαρακτηριστικά που μπορούν να εξαχθούν από συνδέσεις TCP/IP, τα οποία προκαλούν έμμεσα καθυστέρηση στην ανίχνευση.
2. Χαρακτηριστικά κυκλοφορίας : Δημιουργούνται σε σχέση με το διάστημα παράθυρου (χρονολογικά) και ομαδοποιούνται σε «same host» τα οποία εξετάζουν μόνο τις συνδέσεις κατά τα τελευταία δύο δευτερόλεπτα που έχουν τον ίδιο στόχο προορισμού με την τρέχουσα σύνδεση, και συγκεντρώνουν στατιστικά στοιχεία σχετικά με τη συμπεριφορά του πρωτοκόλλου, της υπηρεσίας κ.λ.π. και σε «same service» χαρακτηριστικά τα οποία εξετάζουν μόνο τις συνδέσεις των δύο τελευταίων δευτερολέπτων που έχουν την ίδια υπηρεσία με την τρέχουσα σύνδεση. Για επιθέσεις που σαρώνουν τους στόχους (ή πόρτες) με αργή ανίχνευση με χρονικό διάστημα πάνω από δύο δευτερόλεπτα, οι ομάδες αυτών των χαρακτηριστικών έχουν υπολογιστεί με βάση παράθυρου σύνδεσης 100 συνδέσεων και όχι με το χρονικό όριο των δύο δευτερολέπτων, μετατρέποντάς τα ως χαρακτηριστικά κυκλοφορίας με βάση τη σύνδεση.

3. Χαρακτηριστικά περιεχομένου : Σε αντίθεση με τις εισβολές τύπου DoS, Proving, οι εισβολές τύπου R2L και U2R δεν προκαλούν συχνές και συνεχόμενες επιθέσεις. Αυτό οφείλεται στο γεγονός ότι οι εισβολές τύπου DoS και Proving προκαλούν πολλές συνδέσεις – αιτήματα σε συγκεκριμένους στόχους σε πολύ μικρό χρονικό διάστημα· ενώ, οι επιθέσεις R2L και U2R είναι ενσωματωμένες στα τμήματα των περιεχομένων των πακέτων, και συνήθως περιλαμβάνουν μόνο μία σύνδεση.

Στην προσπάθεια αντιμετώπισης του μεγάλου όγκου πληροφορίας που περιέχει το σύνολο δεδομένων με συνέπεια την επιβάρυνση πόρων, καθώς και για την βελτίωση της απόδοσης του συστήματος για τον εντοπισμό εισβολών, επιχειρήθηκε ο διαχωρισμός του σε 10 υποσύνολα που το κάθε ένα από αυτά περιείχε περίπου το 10% από το αρχικό σύνολο [15]. Όπως παρατηρήθηκε και στο [16], η κατανομή των επιθέσεων π.χ. smurf και neptune, ήταν δυσανάλογη γεγονός το οποίο δυσκόλευε αρκετά προσπάθειες διασταύρωσης των αποτελεσμάτων δοκιμών επειδή πολλά από τα υποσύνολα που δημιουργήθηκαν από τον διαχωρισμό περιείχαν περιπτώσεις μόνο ενός τύπου επιθέσεων.

Η κυριαρχία αυτών των δύο τύπων εισβολής οι οποίοι αφορούν άρνηση υπηρεσίας (DoS), περί του 70% του συνόλου δεδομένων δοκιμών, επηρεάζει – δημιουργεί τάση στην αξιολόγηση, ενώ ο τεράστιος αυτός όγκος κυκλοφορίας προκαλεί μεγάλο υπολογιστικό κόστος στο σύστημα χωρίς ουσιαστικό λόγο αφού θα μπορούσαν να εντοπιστούν με άλλες μεθόδους εκτός των συστημάτων ανίχνευσης εισβολών.

Αρκετά μεγάλο μέρος του συνόλου δεδομένων αποτελείται από περιττές εγγραφές οι οποίες προκαλούν προκατάληψη στους αλγόριθμους εκμάθησης προς αυτές τις συχνές εγγραφές, εμποδίζοντας έτσι την αποτελεσματική εκμάθηση σε μη συχνές εγγραφές που είναι συνήθως πιο επιβλαβείς για δίκτυα όπως οι επιθέσεις U2R και R2L. Επιπλέον, η ύπαρξη αυτών των επαναλαμβανόμενων εγγραφών στο σύνολο δοκιμών προκαλεί την προκατάληψη των αποτελεσμάτων της αξιολόγησης από τις μεθόδους που έχουν καλύτερα ποσοστά ανίχνευσης στις συχνές εγγραφές. Στο [17] επιχειρείται η κατάργηση όλων των εγγραφών που επαναλαμβάνονται σε ολόκληρο το σύνολο εκμάθησης και δοκιμών διατηρώντας μόνο ένα αντίγραφο κάθε εγγραφής. Ακολούθως δημιουργείται το σύνολο δεδομένων NLS-KDD [18] το οποίο αποτελείται από επιλεγμένα αρχεία του συνόλου δεδομένων KDD το οποίο

περιγράφεται στην συνέχεια της παρούσας εργασίας, με στόχο την αντιμετώπιση ελλείψεων [8].

Τέλος, η πραγματική κίνηση παρασκηνίου που είναι απαραίτητη για οποιαδήποτε IDS δεν έχει συμπεριληφθεί.

Με την πάροδο του χρόνου, τα KDD σύνολα δεδομένων έχουν χάσει το μεγαλύτερο μέρος της αποτελεσματικότητας τους αλλά εξακολουθούν να χρησιμοποιούνται ως βάση για την επικύρωση IDS από τους περισσότερους ερευνητές,. Η συνεχιζόμενη χρήση τους οφείλεται κυρίως στην απουσία εναλλακτικών δημοσίων συνόλων δεδομένων.

3.3 Σύνολο δεδομένων GureKDDcup

Το σύνολο δεδομένων GureKDDcup [24] περιέχει συνδέσεις του KDDCUP99 το οποίο περιέχει το ωφέλιμο φορτίο δηλαδή το περιεχόμενο των πακέτων δικτύου, σε κάθε μία από τις συνδέσεις, επιτρέποντας έτσι την εξαγωγή πληροφοριών απευθείας από το ωφέλιμο φορτίο κάθε σύνδεσης που χρησιμοποιείται σε διαδικασίες μηχανικής μάθησης.

Για την δημιουργία του συνόλου δεδομένων χρησιμοποιήθηκε η ίδια μεθοδολογία με αυτή του συνόλου δεδομένων KDDCUP99. Το σύνολο δεδομένων επισημαίνεται σε κάθε σύνδεση με βάση τα αρχεία κλάσης συνδέσεων (tcpdump.list) που παρέχει το MIT. Περιέχει 41 χαρακτηριστικά τα οποία χωρίζονται σε τρεις ομάδες. Τα εγγενή χαρακτηριστικά λαμβάνονται από την κεφαλίδα των πακέτων, ενώ τα χαρακτηριστικά περιεχομένου λαμβάνονται από το περιεχόμενο του πακέτου με βάση τη γνώση ειδικού. Τα χαρακτηριστικά κυκλοφορίας υπολογίζονται λαμβάνοντας υπόψη την προηγούμενη σύνδεση.

Το μέγεθος του συνόλου δεδομένων GureKDDcup 99 είναι 9,3 GB

3.4 Σύνολο δεδομένων DEFCON-8-11 CTF (Όμιλος Shmoo, 2000-2002)

Τα DEFCON-8-11 είναι διαφορετικές εκδόσεις συνόλων δεδομένων που δημιουργήθηκαν κατά τη διάρκεια διαφόρων εκδόσεων του UCSB International Capture The Flag (iCTF). [43], [34]. Τα σύνολα δεδομένων παρουσιάζουν τις

προσπάθειες της ομάδας Shmoo για την προώθηση μεθόδων για την επαρκή προστασία των πόρων υπολογιστών σε όλο τον κόσμο [70].

Περιέχουν επιθέσεις υπερχείλισης και port scanning, ενώ το σύνολο δεδομένων DEFCON-10 (2002) περιέχει port scanning και sweeps, bad packages, δικαιώματα διαχείρισης και FTP από επιθέσεις πρωτοκόλλου telnet. Η κίνηση που παράγεται κατά τη διάρκεια του διαγωνισμού "capture the Flag (CTF)" είναι διαφορετική από την κυκλοφορία δικτύου πραγματικού κόσμου, δεδομένου ότι αποτελείται κυρίως από παρεμβατική κίνηση σε αντίθεση με την κανονική κυκλοφορία παρασκηνίου. Αυτό το θέμα κάνει το σύνολο δεδομένων να χρησιμοποιείται για την αξιολόγηση των τεχνικών συσχέτισης προειδοποίησης [33], [37], [39].

3.5 Σύνολα δεδομένων CAIDA 2002/2016

Το CAIDA [49] περιέχει τρία διαφορετικά σύνολα δεδομένων, το CAIDA OC48, το οποίο περιλαμβάνει διαφορετικούς τύπους δεδομένων που παρατηρούνται σε έναν σύνδεσμο OC48 στο Σαν Χοσέ, το CAIDA DDoS (2007), το οποίο περιλαμβάνει μία ώρα κίνησης DDoS επιθέσεων διάρκειας 5 λεπτών αρχείων pcap, περιέχει μόνο επίθεση κυκλοφορίας προς το θύμα και ανταπόκριση από αυτό και το CAIDA Internet traces 2016, το οποίο περιέχει παθητικά ίχνη κυκλοφορίας από την CAIDA Equinix-Chicago monitor παρακολούθησης του διαδικτύου υψηλής ταχύτητας. Τα περισσότερα από τα σύνολα δεδομένων CAIDA είναι πολύ ειδικά για συγκεκριμένα γεγονότα ή επιθέσεις και είναι ανώνυμα με το ωφέλιμο φορτίο, τις πληροφορίες πρωτοκόλλου και τον προορισμό τους, ενώ περιέχουν κυκλοφορία χωρίς επίθεση. Χρησιμοποιούνται για τον εντοπισμό επιθέσεων DDoS χαμηλού και υψηλού ρυθμού [37], [39], [46].

Ορισμένα από τα σύνολα δεδομένων CAIDA είναι ελεύθερα διαθέσιμα για λήψη, ενώ άλλα είναι διαθέσιμα μόνο σε εγγεγραμμένους χρήστες όπως μέλη της CAIDA και ερευνητές από ακαδημαϊκούς, κυβερνητικούς λειτουργούς και μη κερδοσκοπικούς οργανισμούς.

3.6 Σύνολο δεδομένων LBNL (Εθνικό Εργαστήριο Lawrence Berkeley)

Το σύνολο δεδομένων LBNL [47], περιέχει πλήρη κίνηση δικτύου που καταγράφεται από μεσαίου μεγέθους εγκατάσταση. Τα εσωτερικά ίχνη κίνησης της

LBNL είναι ίχνη δικτύου πλήρους κεφαλίδας χωρίς ωφέλιμο φορτίο και ανωνυμία για να αφαιρέσει οποιαδήποτε πληροφορία που θα μπορούσε να προσδιορίσει ένα μεμονωμένο IP [37], [45].

Η κυκλοφορία παρασκηνίου LBNL αποτελείται από εισερχόμενες, εξερχόμενες και εσωτερικά δρομολογημένες ροές κυκλοφορίας στους δρομολογητές LBNL edge υπηρεσιών ιστού, ηλεκτρονικού ταχυδρομείου και ονομάτων, ενώ η κυκλοφορία επίθεσης προσδιορίζεται απομονώνοντας σαρώσεις σε συνολικά ίχνη κυκλοφορίας. Η κακόβουλη κίνηση αποτελείται κυρίως από αποτυχημένα εισερχόμενα αιτήματα TCP SYN. Το ποσοστό επίθεσης είναι σημαντικά χαμηλότερο από το ποσοστό κυκλοφορίας παρασκηνίου. Η πολυπλοκότητα και η ιδιωτικότητα ήταν δύο κύριες επιφυλάξεις στην συλλογή δεδομένων τελικού σημείου. Για την αντιμετώπιση αυτών των επιφυλάξεων, οι δημιουργοί του συνόλου δεδομένων ανέπτυξαν ένα προσαρμοσμένο εργαλείο MS Windows πολλαπλών επιπέδων χρησιμοποιώντας το W INPCAP API [48] για τη συλλογή δεδομένων. Για να μειώσουν την πολυπλοκότητα καταγραφής πακέτων στα τελικά σημεία, κατέγραψαν μόνο πολύ στοιχειώδεις πληροφορίες επιπέδου συνεδρίας (αμφίδρομη επικοινωνία μεταξύ δύο διευθύνσεων IP σε διαφορετικές θύρες) για τα πακέτα TCP και UDP. Για να εξασφαλιστεί το απόρρητο των χρηστών, χρησιμοποιήθηκε μια πολιτική ανωνυμοποίησης για την ανωνυμοποίηση όλων των παρουσιών κυκλοφορίας [46].

Το σύνολο δεδομένων LBNL μπορούν να χρησιμοποιηθεί για τη διερεύνηση χαρακτηριστικών της κυκλοφορίας του Διαδικτύου.

3.7 Σύνολο δεδομένων ECML-PKDD 2007

Το σύνολο δεδομένων ECML-PKDD 2007 Discovery Challenge [82], δημιουργήθηκε για τις ανάγκες διαγωνισμού εξόρυξης δεδομένων που διεξήχθη σε συνδυασμό με το 18ο Ευρωπαϊκό Συνέδριο για τη μηχανική μάθηση (ECML) και το 11ο Ευρωπαϊκό Συνέδριο για τις αρχές και την πρακτική της ανακάλυψης γνώσης σε βάσεις δεδομένων (PKDD).

Το σύνολο δεδομένων βασίζεται σε κυκλοφορία ιστού πραγματικού κόσμου. Τα δεδομένα είναι διαχωρισμένα σε ένα σετ εκπαίδευσης και ένα σετ δοκιμών, με το σετ εκπαίδευσης να αποτελείται από 50.000 δείγματα και των δοκιμών από 70.000

δείγματα. Τα αιτήματα επισημαίνονται με προδιαγραφές κατηγοριών επίθεσης ή κανονικής κυκλοφορίας.

Κάθε παράδειγμα εκπαίδευσης και δοκιμής στο σύνολο δεδομένων περιέχει το πλήρες κείμενο της αίτησης http, χωρισμένο στα ακόλουθα στοιχεία : τεχνική, πρωτόκολλο, uri, ερώτημα, κεφαλίδες και σώμα. Επιπλέον, κάθε αίτημα HTTP περιλαμβάνει τα ακόλουθα χαρακτηριστικά :

1. Λειτουργικό σύστημα που εκτελείται στον διακομιστή ιστού (UNIX, WINDOWS, άγνωστο)
2. HTTP Server που στοχεύει το αίτημα (APACHE, MIIS, άγνωστο)
3. Η τεχνολογία XPATH είναι κατανοητή από το διακομιστή; (ΑΛΗΘΗΣ, ΨΕΥΔΗΣ, άγνωστο)
4. Βάση δεδομένων LDAP στο διακομιστή Web; (ΑΛΗΘΗΣ, ΨΕΥΔΗΣ, άγνωστο)
5. Βάση δεδομένων SQL στο διακομιστή Web; (ΑΛΗΘΗΣ, ΨΕΥΔΗΣ, άγνωστο)

Τέλος, σε κάθε αίτημα ανατέθηκε μία ή περισσότερες από τις 8 πιθανές ετικέτες κλάσης. Ο πίνακας 4 δείχνει τα μεγέθη και τις κατανομές τάξεων τόσο των σετ εκπαίδευσης όσο και των σετ δοκιμών.

Ωστόσο, τα αιτήματα επίθεσης αυτού του συνόλου δεδομένων κατασκευάστηκαν γενικά και δεν στόχευσαν καμία πραγματική εφαρμογή Ιστού [68], [79], [80], [81].

Πίνακας 4
Κατανομές τάξεων συνόλου δεδομένων ECML-PKDD

Τάξεις	Σύνολο εκπαίδευσης	Σύνολο δοκιμών
Συνολικά αιτήματα	50.116	70.143
έγκυρα αιτήματα	35.006	42.006
Επιθέσεις	15.110	28.137
Cross Site Scripting	12%	11%
SQL Injection	17%	18%
LDAP Injection	15%	16%
XPath Injection	15%	16%
Path Traversal	20%	18%
Command Execution	23%	23%
SSI	13%	12%

3.8 Σύνολο δεδομένων CDX (Στρατιωτική Ακαδημία των Ηνωμένων Πολιτειών 2009)

Το σύνολο δεδομένων CDX [52], δημιουργήθηκε με την συλλογή δεδομένων που διεξήχθη κατά τη διάρκεια του 4ήμερου διαγωνισμού άσκησης Cyber Defense 2009 μεταξύ της αμερικανικής στρατιωτικής ακαδημίας και της Εθνικής Υπηρεσίας Ασφαλείας (NSA) προκειμένου να αντιμετωπίσει προβλήματα του συνόλου δεδομενων DARPA 1998, 1999.

Αντιπροσωπεύει τους διαγωνισμούς δικτύου επίθεσης - άμυνας, με στόχο να χρησιμοποιηθούν για τη δημιουργία σύγχρονου συνόλου δεδομένων με επισήμανση και να δημιουργηθεί κυκλοφορία δικτύου που μιμείται με ακρίβεια την κυκλοφορία σε πραγματικό κόσμο, αλλά παράγεται από κεντρικούς υπολογιστές. Περιλαμβάνει κυκλοφορία όπως Web, e-mail, αναζητήσεις DNS και άλλες απαιτούμενες υπηρεσίες. Με χρήση εργαλείων επίθεσης όπως το Nikto, Nessus, και WebScarab και αισθητήρες δικτύου πραγματοποιήθηκε η αυτόματη αναγνώριση και επιθέσεις.

Το εν λόγω σύνολο δεδομένων μπορεί να χρησιμοποιηθεί για τη δοκιμή των κανόνων προειδοποίησης IDS, αλλά πάσχει από την έλλειψη ποικιλομορφίας και όγκου κυκλοφορίας δεδομένου ότι τα παιχνίδια αυτά διεξάγονται συνήθως σε απομονωμένα δίκτυα. Επομένως, αυτές οι ασκήσεις στερούνται τυπικού θορύβου παρασκηνίου στο Διαδίκτυο [37], [52], [53].

3.9 Σύνολο δεδομένων Kyoto (Πανεπιστήμιο του Κιότο-2009)

Το σύνολο δεδομένων του Πανεπιστημίου του Κιότο [50], είναι μια συλλογή δεδομένων πραγματικής κίνησης δικτύου, που λαμβάνονται από honeypots [51].

Το σύστημα αποτελείται από 348 honeypot συμπεριλαμβανομένων δύο αισθητήρων μαύρης τρύπας με 318 αχρησιμοποίητες διευθύνσεις IP, ενώ τα περισσότερα από αυτά επανεκκινούνται αμέσως μετά την παρατήρηση ενός κακόβουλου εξερχόμενου πακέτου.

Το ακατέργαστο σύνολο δεδομένων που λήφθηκε σε διάστημα 3 ετών, αποτελείται από 14 στατιστικά χαρακτηριστικά που προέρχονται από το σύνολο

δεδομένων KDD Cup 99 καθώς και 10 πρόσθετα χαρακτηριστικά που μπορούν να χρησιμοποιηθούν για περαιτέρω ανάλυση και αξιολόγηση των NIDSs.

Παρουσιάζει περιορισμένη εικόνα κυκλοφορίας δικτύου, δεδομένου ότι καταγράφει μόνο επιθέσεις που στοχεύουν στα honeypots. Μόνο 14 συμβατικά χαρακτηριστικά χρησιμοποιήθηκαν κατά τη διάρκεια της εκπαίδευσης και των δοκιμών. Δεδομένου ότι τα περισσότερα από τα δεδομένα επισκεψιμότητας honeypot αποτελούνται από δεδομένα επίθεσης, δημιουργήθηκαν δεδομένα κανονικής κίνησης για τις υπηρεσίες αλληλογραφίας και διακομιστή DNS, τα οποία δεν αντικατοπτρίζονται στην κανονική κυκλοφορία του δικτύου στον πραγματικό κόσμο, επομένως δεν υπάρχουν ψευδή θετικά στοιχεία, τα οποία είναι σημαντικά για την ελαχιστοποίηση του αριθμού των καταχωρίσεων [37], [46].

Σύμφωνα με τους δημιουργούς του συνόλου δεδομένων [50], τα αποτελέσματα μελετών 3 ετών που βασίζονται σε διάφορους τύπους honeypots είναι εξαιρετικά αποτελεσματικές για την καλύτερη κατανόηση των κυριότερων τάσεων και χαρακτηριστικών των πρόσφατων απειλών στον κυβερνοχώρο και για την εκπόνηση αντίμετρων εναντίον τους.

3.10 Σύνολο δεδομένων Twente (Πανεπιστήμιο του Twente - 2009)

Το σύνολο δεδομένων [54], είναι ένα επισημασμένο σύνολο για ανίχνευση εισβολής με βάση τη ροή (ένα σύνολο πακέτων IP που διέρχονται ένα σημείο παρατήρησης στο δίκτυο κατά τη διάρκεια ενός συγκεκριμένου χρονικού διαστήματος και έχουν ένα σύνολο κοινών ιδιοτήτων [55].

Βασίζεται σε ένα honeypot από το Netflow, που εκτελεί αναπτυγμένες υπηρεσίες και συνδέεται άμεσα με το διαδίκτυο, εγκλωβίζοντας επιθέσεις- εκθέσεις. Το τελικό σύνολο δεδομένων αποτελείται από ροές 14,2 M και ειδοποιήσεις 7,6 M το οποίο συλλέχθηκαν σε διάστημα 6 ημερών, με πάνω από το 98% αυτών να έχει επισημανθεί, ενώ το περιεχόμενο πλήρους πακέτου χρησιμοποιήθηκε μόνο ως πρόσθετη πηγή πληροφοριών κατά τη διάρκεια της διαδικασίας επισήμανσης. Περιλαμβάνει τρεις υπηρεσίες, το OpenSSH, τον εξυπηρετητή ιστού Apache και το Protp, χρησιμοποιώντας το auth/ident στην θύρα 113.

Δεν χρησιμοποιήθηκε έγχυση επίθεσης προκειμένου το σύνολο δεδομένων να αντικατοπτρίζει την κατάσταση σε πραγματικά δίκτυα.

Υπάρχει κάποια ταυτόχρονη κυκλοφορία δικτύου όπως auth/dent, ICMP και IRC, οι οποίες δεν είναι καθαρά καλοήθεις ή κακόβουλες κινήσεις. Επιπλέον, περιέχει ορισμένες άγνωστες και μη συνδεδεμένες καταχωρίσεις κυκλοφορίας καθώς και υπάρχει έλλειψη όγκου και ποικιλομορφία επιθέσεων [37], [38], [54].

3.11 Σύνολο δεδομένων UNIBS (University of Brescia Dataset, 2009)

Στα δεδομένα εκπαίδευσης περιλαμβάνονται τα πρωτόκολλα POP3, SMTP, HTTP, MSN, FTP και BitTorrent.

Το σύνολο δεδομένων στο [73], περιέχει ίχνη κυκλοφορίας και τις σχετικές πληροφορίες που συλλέχθηκαν χρησιμοποιώντας το εργαλείο Ground Truth (GT) [12]. Τα ίχνη κίνησης καταγράφηκαν από δίκτυο πανεπιστημιουπόλεων του πανεπιστημίου για τρεις συνεχόμενες ημέρες. Είκοσι σταθμοί εργασίας που χρησιμοποιούν το εργαλείο GT χρησιμοποιήθηκαν για τη συλλογή των ιχνών. Το σύνολο δεδομένων δημιουργήθηκε σε ένα μη ελεγχόμενο περιβάλλον, όπου όλοι οι τύποι κυκλοφορίας μπορούσαν να ρέουν στο σύστημα χωρίς σχεδόν κανέναν περιορισμό, και στη συνέχεια το GT χρησιμοποιήθηκε για την ανάλυση των πακέτων που συλλήφθηκαν. Έχει ίχνη κυκλοφορίας περίπου 27 GB δεδομένων, ενώ οι κλάσεις επισκεψιμότητας που χρησιμοποιούνται είναι Browser, Mail, RSS feed, BitTorrent και Skype. Ο αριθμός των πακέτων που χρησιμοποιούνται από το σύνολο δεδομένων UNIBS είναι 45541 πακέτα, εκ των οποίων 22249 ανήκουν στο πρόγραμμα περιήγησης, 1291 ανήκουν στο Mail, 279 ανήκουν στο RSS feed, 946 ανήκουν στο BitTorrent.

3.12 Σύνολο δεδομένων NSL-KDD

Το σύνολο δεδομένων NSL-KDD [18] αφορά δεδομένα δικτύου εκτός σύνδεσης με βάση το σύνολο δεδομένων KDD 99, δημιουργήθηκε σε μία προσπάθεια αντιμετώπισης γνωστών θεμάτων [17] που επηρεάζουν την αποτελεσματικότητα των IDS. Αποτελείται από επιλεγμένες εγγραφές του πλήρους συνόλου δεδομένων KDD [32].

Σύμφωνα με το [17], το σύνολο εκπαίδευσης του NSL-KDD δεν περιλαμβάνει περιττές - εγγραφές και επομένως μειώνει το επίπεδο πολυπλοκότητας και κυρίως

αντιμετωπίζει το πρόβλημα δημιουργίας προκατάληψης στους αλγόριθμους εκμάθησης προς τις πιο συχνά εμφανιζόμενες εγγραφές.

Το πλήθος των επιλεγμένων εγγραφών που κατατάσσονται σε ομάδα είναι αντιστρόφως ανάλογο με το ποσοστό εγγραφών στο αρχικό σύνολο δεδομένων KDD, επιτυγχάνοντας έτσι τα ποσοστά κατάταξης των διαφορετικών μεθόδων μηχανικής εκμάθησης να ποικίλλουν σε ένα μεγαλύτερο εύρος, γεγονός το οποίο καθιστά αποτελεσματικότερη την ακριβή αξιολόγηση των διαφορετικών τεχνικών μάθησης.

Το πλήθος των εγγραφών στα σύνολα εκπαίδευσης και στα σύνολα δοκιμών είναι περιορισμένο σε σχέση με το αρχικό, γεγονός που περιορίζει το υπολογιστικό κόστος για την διεξαγωγή των πειραμάτων στο σύνολο χωρίς να χρειάζεται να επιλεγεί τυχαία ένα μικρό τμήμα, με στόχο, τα αποτελέσματα της αξιολόγησης διαφόρων ερευνητικών έργων να είναι συνεκτικά και συγκρίσιμα.

Στο [32] πραγματοποιείται μία ανάλυση απόδοσης του συνόλου δεδομένων χρησιμοποιώντας τεχνικά νευρωνικά δίκτυα (ANN) που χρησιμοποιούνται για εποπτευόμενη εκπαίδευση με σκοπό την ταξινόμηση [35]. Χρησιμοποιούνται τα σύνολα δεδομένων το KDDTrain+.ARFF για εκπαίδευση και το KDDTest+.ARFF για δοκιμή. Με την προτεινόμενη τεχνική για την ταξινόμηση δυαδικών τάξεων επιτυγχάνεται μεγαλύτερη ακρίβεια ανίχνευσης επίθεσης από άλλων αναφερόμενων τεχνικών, ενώ για την ταξινόμηση των πέντε τάξεων διαπιστώθηκε ότι το σύστημα έχει καλή ικανότητα να βρει την επίθεση για συγκεκριμένη κλάση στο σύνολο δεδομένων.

Σύμφωνα με το [17], το σύνολο δεδομένων NLS-KDD συνεχίζει να περιέχει ορισμένα από τα προβλήματα του αρχικού KDD, αλλά λόγω της έλλειψης δημόσιων συνόλων δεδομένων για δικτυακούς IDS, μπορεί να χρησιμοποιηθεί ως αποτελεσματικό σύνολο δεδομένων αναφοράς για έρευνα με διαφορετικές μεθόδους ανίχνευσης εισβολών.

3.13 Σύνολα δεδομένων LLDOS 1.0 και LLDOS 2.0.2 (DARPA 2011)

Τα σύνολα δεδομένων LLDOS 1.0 και LLDOS 2.0.2, είναι το πρώτο και δεύτερο σενάριο του συνόλου δεδομένων DARPA 2000 [76]. Περιλαμβάνουν μια επίθεση Distributed Denial of Service (DDoS), αλλά με διαφορετικά επίπεδα μυστικότητας.

Και στα δύο σενάρια, πραγματοποιείται εισβολή σε έναν κεντρικό υπολογιστή εκμεταλλευόμενος μία ευπάθεια της υπηρεσίας Solaris sadmind RPC. Στη συνέχεια, εγκαθίσταται το λογισμικό trojan mstream DDoS και ξεκινάει μια επίθεση DDoS σε έναν διακομιστή εκτός του χώρου από τον προσβεβλημένο κεντρικό υπολογιστή. Η κύρια διαφορά μεταξύ LLDoS 1.0 και LLDoS 2.0.2 είναι ότι στο LLDoS 2.0.2 ο εισβολέας ανιχνεύει τον κεντρικό υπολογιστή, την πλατφόρμα, το λειτουργικό σύστημα κάνοντας ερωτήματα DNS HINFO, αντί να σαρώνει τις θύρες IP και RPC και να σπάει ένα host_rst, ή να επιτίθεται σε κάθε κεντρικό υπολογιστή ξεχωριστά. Το LLDDoS 1.0 περιλαμβάνει πέντε βήματα και διαρκεί περίπου τρεις ώρες. Στον πίνακα 3 εμφανίζονται οι ειδοποιήσεις που δημιουργούνται με το εργαλείο RealSecure για κάθε φάση των LLDoS 1.0 όσο και των LLDoS 2.0.2.

Πίνακας 3
Ποσοστό απόδοσης LLDoS 1.0, LLDoS 2.0.2

	LLDoS 1.0	**LLDoS 2.0.2.**
Top 1 Rate (%)	82,75%	88,46%
Top 2 Rate (%)	78,42%	85,22%
Top 3 Rate (%)	76,03%	83,60%

3.14 Σύνολο δεδομένων UMASS (Πανεπιστήμιο της Μασαχουσέτης - 2011)

Το σύνολο δεδομένων UMASS [56], περιλαμβάνει στοιχεία από ίχνη, πακέτων δικτύου, και ορισμένα ίχνη από ασύρματες εφαρμογές [43]. Δημιουργήθηκε με τη χρήση ενός σεναρίου επίθεσης αιτήματος λήψης από το TCP. Το σύνολο δεδομένων δεν είναι χρήσιμο για τη δοκιμή συστημάτων IDS και IPS λόγω της έλλειψης ποικιλομορφίας κυκλοφορίας και επιθέσεων [57], [37],[38].

3.15 Σύνολο δεδομένων ISCX – 2012 UNB (Πανεπιστήμιο του New Brunswick)

Το ISCXIDS2012 σύνολο δεδομένων [39], περιέχει δυναμικά παραγόμενα δεδομένα που αποτυπώνουν πραγματικά ίχνη δικτύου και εισβολών, με περίοδο σύλληψης 7 ημερών.

Εισάγεται και γίνεται χρήση της έννοιας προφίλ που περιέχουν λεπτομερείς περιγραφές παρεμβολών και αφηρημένων μοντέλων διανομής για εφαρμογές, πρωτόκολλα ή οντότητες δικτύου χαμηλότερου επιπέδου. Τα πραγματικά ίχνη αναλύονται για τη δημιουργία προφίλ που δημιουργούν πραγματική κίνηση για HTTP, SMTP, SSH, IMAP, POP3 και FTP με πλήρες ωφέλιμο φορτίο πακέτου.

Το προφίλ-A το πραγματοποιεί διάφορα σενάρια επίθεσης πολλαπλών σταδίων, και το προφίλ-B, ως καλοήθης γεννήτρια κυκλοφορίας, παράγει ρεαλιστική κυκλοφορία δικτύου με θόρυβο παρασκηνίου.

Μειονεκτήματα του συνόλου δεδομένων είναι η πολυπλοκότητα που προκύπτει λόγω της ανάγκης αρχικής δημιουργίας και στη συνέχεια εκτέλεσης α- και β- προφίλ. Τα α- προφίλ απαιτούν ειδικές γνώσεις σχετικά με τον τρόπο κατασκευής μιας επίθεσης. Η δημιουργία β- προφίλ απαιτεί φιλτραρισμένα ίχνη δικτύου που ενδέχεται να μην είναι εύκολα προσβάσιμα. Επίσης, ανάλογα με την πολυπλοκότητα του α - προφίλ, μπορεί να απαιτούν ανθρώπινη βοήθεια για την εκτέλεση, η οποία μπορεί να εισάγει διαφορές στον τρόπο εκτέλεσης μιας επίθεσης και να μειώσει την ποιότητα επαναληψιμότητας ενός παραγόμενου συνόλου δεδομένων [39].

Δεν αντιπροσωπεύει νέα πρωτόκολλα δικτύου δεδομένου ότι σχεδόν 70% των σημερινών κυκλοφοριακών γραμμών δικτύου είναι HTTPS και δεν υπάρχουν ίχνη HTTPS σε αυτό το σύνολο δεδομένων. Επιπλέον, η διανομή των προσομοιωμένων επιθέσεων δεν βασίζεται σε στατιστικές του πραγματικού κόσμου [37].

Στερείται ρεαλιστικού θορύβου παρασκηνίου στο διαδίκτυο και η συνολική κανονική κίνηση δεν είναι συγκρίσιμη με ένα πραγματικό δίκτυο [53].

3.16 Σύνολο δεδομένων CTU-13 (Τεχνικό Πανεπιστήμιο Τσεχίας - 2011)

Το σύνολο δεδομένων CTU-13 [83], [84], είναι ο συνδυασμός σύλληψης 13 διαφορετικών κακόβουλων προγραμμάτων σε ένα πραγματικό περιβάλλον δικτύου. Ο στόχος αυτού του συνόλου δεδομένων είναι η καταγραφή πραγματικής μικτής κίνησης botnet. Οι μολυσμένοι κεντρικοί υπολογιστές δημιουργούν κίνηση botnet και οι επαληθευμένοι κανονικοί κεντρικοί υπολογιστές δημιούργησαν κανονική κίνηση, ενώ οι υπόλοιπες κινήσεις είναι άγνωστη κίνηση παρασκηνίου. Το σύνολο δεδομένων CTU-13 περιλαμβάνει 13 λήψεις διαφορετικών δειγμάτων

botnet, ως σενάρια (πίνακας 5). Καθένα από όλα τα σενάρια εκτελέστηκε με ένα συγκεκριμένο κακόβουλο λογισμικό που χρησιμοποίησε διάφορα πρωτόκολλα και πραγματοποίησε διάφορες ενέργειες. Αυτό το σύνολο δεδομένων είναι ένα από τα μεγαλύτερα και πιο επισημασμένα σε υπάρχοντα σύνολα δεδομένων και δημιουργήθηκε από το Πανεπιστήμιο CTU της.

Για την αξιολόγηση της απόδοσης εντοπισμού botnet, χρησιμοποιήθηκαν οι αλγόριθμοι BClus, CAMNEP και BotHunter.

Πλεονέκτημα του συνόλου δεδομένων είναι ότι είναι προσεκτικά επισημασμένο και η διαδικασία λήψης πραγματοποιείται σε ελεγχόμενο περιβάλλον [85], [86].

Πίνακας 5
Πλήθος δεδομένων για κάθε σενάριο Botnet

DataSet	Διάρκεια (h)	NetFlows	Μέγεθος (GB)	Bot Name	Numbers of Bots	Botnet Flow
1	6.15	2.824,637	52,00	Neris	1	39.933
2	4.20	1.808,123	60,00	Neris	1	18.839
3	67.00	4.710,639	121,00	Rbot	1	26.759
4	4.0	1.121.077	53,00	Rbot	1	1.719
5	12.00	129.833	37,60	Virut	1	695
6	2.20	558.920	30,00	Menti	1	4.431
7	0.30	114.078	5,80	Sogou	1	37
8	20.00	2.954.231	123,00	Murlo	1	5.052
9	5.20	2.753.885	94,00	Neris	10	179.880
10	5.00	1.309.792	73,00	Rbot	10	106.315
11	0.30	107.252	5,20	Rbot	3	8.161
12	1.20	325.472	8,30	NSIS.ay	3	2.143
13	16.30	**1.925.150**	**34,00**	**Virut**	**1**	**38.791**

3.17 Σύνολο δεδομένων UNSW-NB15 (Australian Centre for Cyber Security (ACCS))

Το βασικό χαρακτηριστικό του συνόλου δεδομένων UNSW-NB15 [25], είναι ένα πλήθος πραγματικών σύγχρονων φυσιολογικών συμπεριφορών και συνθετικών

δραστηριοτήτων επίθεσης. Για την δημιουργία κανονικής και ανώμαλης κυκλοφορίας δικτύου χρησιμοποιήθηκε το IXIA PerfectStorm tool στο εργαστήριο Cyber Range του ACCS. Προσομοιώνει εννέα κατηγορίες επίθεσης, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms, από ανώμαλη κυκλοφορία μέσω του εργαλείου IXIA, το οποίο περιέχει πληροφορίες σχετικά με νέες επιθέσεις. Για την καταγραφή πακέτων από την κίνηση δικτύου χρησιμοποιήθηκε το εργαλείο tcpdump. Η περίοδος προσομοίωσης διήρκησε 31 ώρες σε διάστημα 2 ημερών με λήψη 100 GBs. Επιπλέον, κάθε αρχείο pcap είναι χωρισμένο σε 1000 MB χρησιμοποιώντας το εργαλείο tcpdump. Δημιουργήθηκαν 49 χαρακτηριστικά από τα αρχεία pcap, με την χρήση των εργαλείων Argus και Bro-IDS. Για την ανάλυση των ροών των πακέτων σύνδεσης, χρησιμοποιήθηκαν 12 αλγόριθμοι σε γλώσσα C++. Το σύνολο δεδομένων επισημαίνεται από έναν πίνακα αλήθειας ως βάση, που περιέχει όλους τους προσομοιωμένους τύπους επίθεσης. Η γεννήτρια κυκλοφορίας IXIA χρησιμοποιεί τρεις εικονικούς κεντρικούς υπολογιστές, τους 2 για την κανονική κυκλοφορία, ενώ ο άλλος ένας χρησιμοποιείται για τις μη φυσιολογικές/κακόβουλες δραστηριότητες στην κυκλοφορία δικτύου. Ο πίνακας 7 αποτυπώνει τους τύπους επιθέσεων του συνόλου δεδομένων.

Πίνακας 7
Τύποι επιθέσεων συνόλου δεδομένων UNSW-NB15

Κατηγορία	Σύνολο εκπαίδευσης	Σύνολο δοκιμών
Normal	56.000	37.000
Analysis	2.000	677
BackDoor	1.746	583
DoS	12.264	4.089
Exploits	33.393	11.132
Fuzzers	18.184	6.062
Generic	40.000	18.871
Reconnaissance	10.491	3.496
ShellCode	1.133	378
Worms	130	44
Σύνολα εγγραφών	**175.341**	**82.332**

3.18 Σύνολο δεδομένων SLINGbot

Το σύστημα ανίχνευσης bots επόμενης γενιάς σε πραγματικό χρόνο SLINGbot, αναπτύχθηκε από την BBN Technologies [72] με στόχο να δώσει την δυνατότητα στους ερευνητές να μπορούν να δημιουργούν καλοήθη botnets και να χαρακτηρίζουν τις τρέχουσες και μελλοντικές πιθανές δομές εντολών και ελέγχου botnet και να σχεδιάζουν αποτελεσματικές τεχνικές άμυνας. Οι επιθέσεις στο σύνολο δεδομένων SLINGbot περιλαμβάνουν κλοπή ταυτότητας, επιθέσεις άρνησης εξυπηρέτησης (DoS), phishing, κατασκοπεία, σάρωση, spam ηλεκτρονικού ταχυδρομείου και απάτη.

Επικεντρώνεται μόνο στη δημιουργία της επικοινωνίας C2 μέσα στο botnet. Αυτό περιλαμβάνει τη λήψη λογισμικού bot, τη σύνδεση με διακομιστές bot C2 και τη λήψη εντολών botnet, ενώ, αποκλείει την κίνηση που σχετίζεται με τον εντοπισμό μελλοντικών bots, την πρόσβαση σε αυτά τα bots και την πραγματική εκτέλεση κακόβουλων δραστηριοτήτων.

Εκτελεί πραγματικό λογισμικό botnet, καθώς επιτρέπει την προσομοίωση ενός botnet με επιλεγμένα C2 χαρακτηριστικά, τα οποία δημιουργούνται με την χρήση του εργαλείου TrenMicro [73], είτε για τις υπάρχουσες ή μελλοντικές πιθανές botnets με ελεγχόμενο και ασφαλή τρόπο.

Χρησιμοποιήθηκε η Python για την ταχύτητα εφαρμογής της, την ανεξαρτησία της πλατφόρμας, παρέχοντας υποστήριξη για πρωτόκολλα δικτύου τόσο σε χαμηλό επίπεδο όσο και σε επίπεδο πρωτοκόλλου όπως XMLRPC, HTTP, SSH, SMTP, FTP κ.λ.π.

Το SLINGbot είναι επίσης επεκτάσιμο και επιτρέπει τη χρήση κοινόχρηστων βιβλιοθηκών μονάδων botnet και χρησιμοποιείται επί του παρόντος για να χαρακτηρίσει διάφορες αρχιτεκτονικές εντολών και ελέγχου botnet.

3.19 Σύνολο δεδομένων ADFA-LD

Στο [27] εισάγεται το σύνολο δεδομένων ADFA-LD12 το οποίο προσπαθεί να αντιμετωπίσει επιθέσεις με σύγχρονη δομή μεθοδολογία και βάση των αδυναμιών και προβλημάτων που παρουσιάζουν τα σύνολα δεδομένων KDD [17].

Είναι σχεδιασμένο για συστήματα βασισμένα σε ανωμαλίες και όχι για εντοπισμό υπογραφών, επιτυγχάνοντας μία διαφορά απόδοσης με αυτή των συνήθως χρησιμοποιούμενων αλγορίθμων ανίχνευσης εισβολών. Αποτελείται από τρεις

διαφορετικές ομάδες δεδομένων, με κάθε ομάδα να περιέχει ακατέργαστα ίχνη κλήσης συστήματος για εκπαίδευση ή επικύρωση, τα οποία συλλέχθηκαν κατά τη διάρκεια της κανονικής λειτουργίας κεντρικού υπολογιστή, με δραστηριότητες στο web και την προετοιμασία εγγράφων LATEX.

Για την δημιουργία αυτού του συνόλου δεδομένων [27] επιλέχθηκε μία υλοποίηση με λειτουργικό σύστημα Ubuntu Linux 11.04 και μεθοδολογίες επιθέσεων αιχμής [28], [29]. Για να είναι δυνατή μία web-based επίθεση, εγκαταστάθηκε και ενεργοποιήθηκε η έκδοση Apache 2.2.17 που εκτελεί την έκδοση PHP 5.3.5. Το λειτουργικό σύστημα ήταν πλήρως ενημερωμένο και οι FTP, SSH και MySQL Ver. 14.14 ξεκίνησαν ως υπηρεσίες με τις προεπιλεγμένες θύρες τους. Επίσης εγκαταστάθηκε το TikiWiki Ver. 8.1 [30] ως διαδικτυακό εργαλείο το οποίο επιλέχθηκε έχοντας μία γνωστή τρωτότητα [31], βάση της οποία μπορεί να επιτευχθούν web-based επιθέσεις. Αυτή η ρύθμιση έγινε για να αντιπροσωπεύει μια γενική λογική ενός σύγχρονου τοπικού εξυπηρετητή Linux, για την ανταλλαγή αρχείων, υπηρεσίες βάσης δεδομένων, απομακρυσμένη πρόσβαση και λειτουργικότητα του εξυπηρετητή ιστού, με ορισμένες μικρές ευπάθειες.

Για τις επιθέσεις στο λειτουργικό σύστημα χρησιμοποιήθηκαν σύγχρονοι μηχανισμοί επίθεσης, όπως, Hydra-FTP, Hydra-SSH, Adduser, Java-Meterpreter, και Web shell, διατηρώντας μία ισορροπία μεταξύ τρωτότητας του συστήματος και ρεαλισμού. Αν και η υλοποίηση ήταν πολύ αποτελεσματική για τον στόχο, είναι σπάνια σε σύγχρονα συστήματα. Για την επίτευξη ενός πιο ρεαλιστικού σεναρίου ως μια καλή προσομοίωση του πραγματικού κόσμου, προσαρμόσθηκαν ο στόχος και ο κώδικας TikiWiki. Δοκιμάστηκαν επιθέσεις "ωμής βίας", επιθέσεις στις υπηρεσίες FTP και SSH, επιθέσεις δικαιωμάτων πρόσβασης, εγκατάσταση εργαλείων BackDoor, που αντιπροσωπεύουν μεθοδολογίες επίθεσης που χρησιμοποιούνται από επιτιθέμενους μεσαίου επιπέδου. Εκμεταλλεύονται κυμαινόμενους τρόπους επίθεσης, από χαμηλού βαθμού κωδικούς πρόσβασης εντοπίζοντας τους μέσω κοινωνικής μηχανικής και διαδικτυακών επιθέσεων [59].

Κατά την διαδικασία συλλογής του συνόλου δεδομένων οι συντάκτες [27], εντόπισαν ένα χαρακτηριστικό των σύγχρονων επιθέσεων, οι οποίες παρουσιάζουν μια κατανεμημένη προσέγγιση στο συμβιβασμό του συστήματος, διαδίδοντας τη δραστηριότητά τους μέσω πολλαπλών διαδικασιών, σε αντίθεση των παλαιότερων τεχνικών οι οποίες περιορίζονταν σε μια ενιαία διαδικασία, δημιουργώντας έτσι ένα μεγαλύτερο και πιο ανιχνεύσιμο αποτύπωμα κλήσης συστήματος.

Για την αντιμετώπιση αυτού του τρόπου επίθεσης, προτείνουν να χρησιμοποιηθεί ένα νέο πλούσιο σε δεδομένα χαρακτηριστικό για την παροχή των εισροών στους υπάρχοντες αλγόριθμους, ή να αναπτυχθούν νέοι μηχανισμοί λήψης αποφάσεων για τον συνδυασμό των πληροφοριών που περιέχονται σε αυτά τα διαφορετικά πολλαπλά ίχνη, ενέργειες οι οποίες θα βοηθήσουν στην βελτίωση των επιδόσεων των IDS [27].

Λόγω της ποικίλης και δυναμικής φύσης των μοτίβων κλήσεων συστήματος, υπάρχει δυσκολία διαχωρισμού φυσιολογικών και μη φυσιολογικών συμπεριφορών [58], [59].

Στον πίνακα 6 εμφανίζεται η σύνθεση του συνόλου δεδομένων ADFA-LD.

Πίνακας 6
Σύνθεση συνόλου δεδομένων ADFA-LD

Ίχνη	Πλήθος	Τύπος
Training	833	Normal
Validation	4.373	Normal
Hydra-FTP	162	Attack
Hydra-SSH	148	Attack
AddUser	91	Attack
Java- Meterpreter	125	Attack
Meterpreter	75	Attack
WebShell	118	Attack

3.20 Σύνολο δεδομένων TUIDS (Tezpur University Intrusion Detection System)

Στο [68] δημιουργείται το σύνολο δεδομένων TUIDS (Tezpur University Intrusion Detection System). Συμπεριλαμβάνει τρία σύνολα δεδομένων εισβολής δικτύου πραγματικής κίνησης :

- TUIDS (Tezpur University Intrusion Detection System) σύνολο δεδομένων εισβολής,
- TUIDS συντονισμένο σύνολο δεδομένων σάρωσης, και

- TUIDS Σύνολο δεδομένων DDoS τόσο σε επίπεδα πακέτων όσο και σε επίπεδα ροής

Το δίκτυο testbed TUIDS αποτελείται από 250 κεντρικούς υπολογιστές, 15 διακόπτες L2, 8 διακόπτες L3, 3 ασύρματους ελεγκτές και 4 δρομολογητές που συνθέτουν 5 διαφορετικά δίκτυα μέσα στην πανεπιστημιούπολη του Πανεπιστημίου Tezpur.

Οι Hosts χωρίζονται σε πολλά VLAN, ενώ κάθε VLAN ανήκει σε ένα switch L3 ή L2 μέσα στο δίκτυο. Το δίκτυο αποτελείται από 6 διασυνδεδεμένους σταθμούς εργασίας Ubuntu 10.10, διακομιστή αρχείων δικτύου, e-mail server, διακομιστή telnet, διακομιστή FTP, web server, και μια SQL server με συμβατότητα PHP. Επίσης, διαμορφώθηκαν 4 διακομιστές Windows 2003 για την εκμετάλλευση ένα διαφορετικό σύνολο γνωστών τρωτών σημείων.

Η κανονική κίνηση δικτύου δημιουργείται με βάση τις καθημερινές δραστηριότητες των χρηστών και ειδικά την κυκλοφορία που δημιουργείται από τους διαμορφωμένους διακομιστές. Η κίνηση επίθεσης δημιουργείται με την εκτόξευση επιθέσεων εντός του δικτύου testbed σε τρία διαφορετικά υποσύνολα, δηλαδή., ένα σύνολο δεδομένων εισβολής TUIDS, ένα συντονισμένο σύνολο δεδομένων σάρωσης και ένα σύνολο δεδομένων DDoS.

Χρησιμοποιήθηκαν 22 τύποι επίθεσης για τη δημιουργία της κυκλοφορίας επίθεσης για το σύνολο δεδομένων εισβολής TUIDS, 6 επιθέσεις για τη δημιουργία της κυκλοφορίας επίθεσης για το συντονισμένο σύνολο δεδομένων σάρωσης και 6 επιθέσεις για τη δημιουργία της κυκλοφορίας επίθεσης για ένα σύνολο δεδομένων DDoS με συνδυασμό πρωτοκόλλων TCP, UDP και ICMP.

Η περίοδος λήψης διήρκησε επτά ημέρες για την εισβολή TUIDS και τα συντονισμένα σύνολα δεδομένων σάρωσης, ενώ η κίνηση DDoS συλλέχθηκε επίσης για το ίδιο χρονικό διάστημα, αλλά σε διαφορετική χρονική περίοδο με αρκετές παραλλαγές επιθέσεων DDoS σε πραγματικό χρόνο με χρήση του εργαλείου NetFlow.

3.21 Σύνολο δεδομένων CIC DoS (2017)

Το σύνολο δεδομένων CIC DoS [60], έχει στόχο την ανίχνευση επιθέσεων DoS με αργό ρυθμό επιπέδου εφαρμογής που βασίζονται σε HTTP, συμπεριλαμβανομένων επιθέσεων πλημμύρας, με βάση τον μη παραμετρικό

αλγόριθμο CUSUM στο πλαίσιο σύγχρονων διακομιστών ιστού. Εξετάζονται χαρακτηριστικά όπως περιορισμένη κατανάλωση πόρων από την πλευρά ενός εισβολέα, στοχευμένη ζημιά και επίθεση μυστικότητας. Χρησιμοποιούνται δύο σύνολα δεδομένων, με χρήση δειγματοληψίας που επιτρέπει την αντιμετώπιση των προβλημάτων κλιμάκωσης.

Ο αλγόριθμος ανίχνευσης εφαρμόστηκε σε μια ροή παρατηρούμενης εισερχόμενης κίνησης δικτύου που εστιάζει σε δύο χαρακτηριστικά : τον αριθμό των αιτήσεων επιπέδου εφαρμογής και τον αριθμό των πακέτων με μέγεθος ωφέλιμου φορτίου ίσο με μηδέν.

Το δοκιμαστικό περιβάλλον περιλαμβάνει έναν εξυπηρετητή θυμάτων Apache Linux v.2.2.22, PHP5 και Drupal v.7 ως σύστημα διαχείρισης περιεχομένου.

Η δημιουργία επιθέσεων του επιπέδου εφαρμογής DoS αναμείχθηκε με τα ίχνη χωρίς επίθεση από το σύνολο ISCX [39]. Το σύνολο που προέκυψε περιλαμβάνει 24h κυκλοφορίας δικτύου με συνολικό μέγεθος 4.6 GB. Το συνολικό ποσοστό για καθαρή κυκλοφορία στο τροποποιημένο σύνολο δεδομένων ISCX ήταν στο μέσο όρο 69kt/s.

Πραγματοποιήθηκαν επιθέσεις στον εξυπηρετητή που περιελάμβανε επίθεση τύπου πλημμύρας με το Goldeneye και αργή επίθεση με κεφαλίδες αποστολής με το Slowloris, παράλληλα με συλλογή πραγματικής κίνησης διαδικτυακών εξυπηρετητών σε ακαδημαϊκή καθαρή εργασία για δύο εβδομάδες μεγέθους 3.5 GB δεδομένων από κυκλοφορία για την πόρτα 80 και με ποσοστό κατά μέσο όρο 2 pkt/s.

Για να εξασφαλιστεί ίση ποσότητα δεδομένων που λαμβάνονται με τις μεθόδους δειγματοληψίας χρησιμοποιήθηκε η τεχνική ποσόστωσης ροών δειγματοληψίας. που οδηγούν σε ίσες ποσότητες ροών. Διαπιστώνεται [60] ότι μείωση της ποσότητας των δεδομένων που διατηρούνται επηρεάζει αρνητικά την ακρίβεια ανίχνευσης.

3.22 Σύνολο δεδομένων (CIC-IDS 2017, 2018)

Με το [37] δημιουργείται το νέο σύνολο δεδομένων IDS, CIC-IDS2017, με βάση έντεκα κριτήρια που θέτουν ως ένα πλαίσιο αξιολόγησης για σύνολα δεδομένων ανίχνευσης εισβολής οι [38], το οποίο καλύπτει κοινές επικαιροποιημένες επιθέσεις, όπως DoS, DDoS, Brute Force, XSS, SQL έγχυση, Infinltration, Port Scan και

Botnet. Το σύνολο δεδομένων είναι πλήρως επισημασμένο με πάνω από 80 χαρακτηριστικά κυκλοφορίας δικτύου που εξάγονται και υπολογίζονται για όλες τις καλοήθεις και παρεμβατικές ροές που μοιάζουν με τα πραγματικά δεδομένα του πραγματικού κόσμου (PCAPs), με την χρήση του λογισμικού CICFlowMeter [40].

Η δοκιμαστική υποδομή για την δημιουργία του συνόλου δεδομένων αποτελείται από δύο διαφορετικά δίκτυα, το δίκτυο θυμάτων και το δίκτυο επίθεσης, με πλήρη εξοπλισμό όπως δρομολογητές, τείχος προστασίας, switches, και διαφορετικές εκδόσεις των κοινών τριών λειτουργικών συστημάτων όπως Windows, Linux και Macintosh.

Για την δημιουργία του συνόλου δεδομένων CSE-CIC-IDS2018 ως βελτίωση του προηγούμενου [42], χρησιμοποιήθηκε το B-Προφίλ το οποίο προτείνεται στο [39] για την εξαγωγή της αφηρημένης συμπεριφοράς των 25 χρηστών με βάση τα πρωτόκολλα HTTP, HTTPS, FTP, SSH και email. Δημιουργούνται 6 προφίλ επίθεσης : Επίθεση ωμής βίας, Heartbleed Attack, Botnet, Επίθεση DoS, Επίθεση DDoS, Web Attack και Infiltration Attack.

Σύμφωνα με το [41], το σύνολο δεδομένων αποτελείται από περισσότερα από οκτώ αρχεία για την κυκλοφορία 5 ημερών που το καθιστά αρκετά μεγάλο γεγονός το οποίο προκαλεί μεγάλο υπολογιστικό κόστος, περιέχει πολλές περιττές εγγραφές που δυσκολεύει την δημιουργία χαρακτηριστικών και παρουσιάζει μεγάλη ανισορροπία στην ομαδοποίηση γεγονός το οποίο μπορεί να προκαλέσει προκατάληψη του ταξινομητή προς την τάξη πλειοψηφίας.

Στο [61] επιχειρείται μία προσέγγιση για την βελτίωση της απόδοσης εφαρμόζοντας το AdaBoost με SMOTE και PCA.

3.23 Σύνολο δεδομένων αξιολόγησης DDoS (CIC-DDoS 2019)

Το σύνολο δεδομένων CIC-DDoS 2019 [62], προτείνει μια προσέγγιση ανίχνευσης και ομαδικής ταξινόμησης με βάση ένα σύνολο χαρακτηριστικών ροής δικτύου για την αντιμετώπιση των επιθέσεων DDos σε πρωτόκολλα, όπως TCP, UDP, ICMP και HTTP.

Είναι επισημασμένο με 80 χαρακτηριστικά κυκλοφορίας δικτύου που έχουν εξαχθεί και υπολογιστεί για όλες τις καλοήθεις και αρνητικές ροές υπηρεσιών με τη χρήση του λογισμικού CICFlowMeter [40], και προτείνει τα καλύτερα σύνολα

χαρακτηριστικών για την ανίχνευση διαφορετικών τύπων επιθέσεων DDoS, συμπεριλαμβανομένων αντανακλαστικών DDoS (όπως DNS, LDAP, MSSQL και TFTP), UDP, UDP-Lag και SYN.

Για την δημιουργία του συνόλου δεδομένων δημιουργήθηκαν δύο δίκτυα, το δίκτυο Attack-Network και το Network. Το Network είναι μια υψηλής ασφαλείας υποδομή με firewall, router, διακόπτες, και αρκετά κοινά λειτουργικά συστήματα μαζί με έναν πράκτορα που παρέχει τις καλοήθεις συμπεριφορές σε κάθε PC. Το δίκτυο Attack-Network είναι μια εντελώς διαχωρισμένη υποδομή που εκτελεί διαφορετικά είδη επιθέσεων DDoS.

Για την δημιουργία ρεαλιστικής κυκλοφορίας στο παρασκήνιου γίνεται χρήση της προσέγγισης B-προφίλ [63], για τη σκιαγράφηση της αφηρημένης συμπεριφοράς των ανθρώπινων αλληλεπιδράσεων και να δημιουργήσει μια φυσική καλοήθης κυκλοφορία στο παρασκήνιο. Το B-προφίλ εξάγει την αφηρημένη συμπεριφορά των 25 χρηστών με βάση τα πρωτόκολλα HTTP, HTTPS, FTP, SSH και email.

Πραγματοποιήθηκαν 12 επιθέσεις DDoS που περιελάμβαναν NTP, DNS, LDAP, MSSQL, NetBIOS, SVMP, SSDP, UDP, UDP-Lag, WebDDoS, SYN και TFTP για κατάρτιση και 7 επιθέσεις, συμπεριλαμβανομένων των επιθέσεων Port Scan, NetBIOS, LDAP, MSSQL, UDP, UDPPLAG και SYN για δοκιμές.

Σύμφωνα με το [64], το σύνολα δεδομένων εκπαίδευσης και δοκιμών διαφέρουν ως προς τη διανομή των δεδομένων, εμφανίζοντας κλάσεις οι οποίες είναι χαμηλής δυναμικής στο σύνολο δεδομένων εκπαίδευσης, ενώ είναι σημαντικές στο σύνολο δεδομένων δοκιμών.

Υπάρχει μία ανισορροπία στις τάξεις με τις χαμηλές τάξεις να βρίσκονται πολύ χαμηλά.

Μεγάλο μέρος του συνόλου δεδομένων εκπαίδευσης ανήκει σε κατηγορίες που απουσιάζουν από το σύνολο δοκιμών, ενώ η επίθεση Portmap στο σύνολο δοκιμών δεν υπάρχει στα δεδομένα εκπαίδευσης για την αξιολόγηση του συστήματος ανίχνευσης.

3.24 Συνοπτική παρουσίαση Συνόλων Δεδομένων.

Στον πίνακα 8, παρουσιάζονται τα σύνολα δεδομένων που παρουσιάσθηκαν με τις αναφορές σε αυτά, τις τεχνικές ανίχνευσης που εφαρμόσθηκαν σε κάθε ένα από αυτά και οι τύποι επιθέσεων που εντοπίσθηκαν.

Πίνακας 8
Αλγόριθμοι ανίχνευσης και τύποι επιθέσεων που εφαρμόσθηκαν στα σύνολα δεδομένων

DataSet Αναφορές	**Αλγόριθμοι που εφαρμόστηκαν**	**Είδη επιθέσεων που αντιμετώπισαν**

3.1
DARPA 1998
 	
[8], [9], [10], [11], [12], [13], [14], [74]
	• SVM
	• ELMs:
	  ○ Basic
	  ○ Kernel-Based
	• SVDD
		• Probing
		• DoS
		• R2L
		• U2R

3.1 DARPA 1999 [8], [9], [10], [11], [12], [13], [14], [75]	• RBF • Elman NN • SNORT • Non-Parametric CUSUM • EM based Clustering	• Probing • DoS • R2L • U2R • 13 Attack Types

3.1
DARPA 2000
 	
[8], [9], [10], [11], [12], [13], [14], [76]
	• Προηγμένη Πιθανοτική προσέγγιση για αναγνωριστικά με βάση το δίκτυο (APAN) με χρήση :
	  ○ Markov Chain
	  ○ - Kmeans Clustering
		• DDoS

3.2
KDD-99
1998 - 1999
 	
[4], [5], [6], [8], [15], [16], [17], [18]
	• Tree Classifiers
	• Bayesian Clustering
	• Parzen Classifier
	• v-SVC
	• k-means clustering
	• Weighted k-NN
	• AdaBoost
	• ABC
	• Fuzzy Association Rules
	• Genetic-based Algorithm
	• C4.5
	• BSPNN using:
	  ○ Adaptive Boosting
	  ○ Semi-parametric NN
	• FC-ANN based on:
	  ○ ANN
	  ○ Fuzzy Clustering
		• Probing
		• DoS
		• R2L
		• U2R
		• DDoS
		• neptune-dos
		• pod-dos
		• smurf-dos
		• buffer-overflow
		• rootkit
		• satan
		• teardrop,

	• Logistic Regression
	• Genetic Fuzzy Systems based on:
	  ○ Michigan
	  ○ Pittsburgh
	  ○ IRL
	• DT
	• Ripper Rule
	• Back-Propagation NN
	• RBF NN
	• Bayesian Network
	• SOM
	• Modified SOM
	• SVM
	• Rule-Based
	• BON
	• ART Network
	• Hidden NB
	• SA
	• Decision Stump
	• NB Tree
	• Random Forest
	• Random Tree
	• Representative Tree Model
	• Ant Colony
	• Fuzzy C means
	• Fuzzy NN / Neurofuzzy
	• RBF SVM
	• Δύο παραλλαγές του GMDH:
	  ○ Monolithic
	  ○ Ensemble-based
	• K-medoids
	• Weighted ELM
	• PCA and Fuzzy PCA
	• k-NN
	• Binary PSO
	• R-tree
	• Polynomial Feature Correlation
	• Softmax Regression
	• Kernel Clustering
	• FLN
	• DL
	• NDAE
	• Stacked NDAEs	

3.3 Gure KDD cup Gure KDD cup 6 percent 1999 [24]	• SVM	• R2L

DataSet Αναφορές	Αλγόριθμοι που εφαρμόστηκαν	Είδη επιθέσεων που αντιμετώπισαν
3.4 DEFCON-8 DEFCON-11 2000 - 2002 [33], [34], [37], [39], [43], [70]	N/A	• port scanning • sweeps • bad packages
3.5 CAIDA 2002 / 2016 [49], [37], [39], [46].	N/A	• DoS
3.6 LBNL 2004 / 2005 [37], [45], [46], [47]	N/A	• TCP SYN
3.7 ECML / PKDD 2007 [68], [79], [80], [81], [82]	• AUC • VSM	• Cross-Site Scripting, • SQL Injection, • LDAP Injection, • XPath Injection, • Path traversal • Command execution SSI attacks
3.8 CDX 2009 [37], [52], [53].	• Nikto • Nessus • WebScarab	• Probing • DoS • R2L • U2R
3.9 Kyoto 2009 [37], [46], [50]	• honeypots	• Probing • DoS • R2L • U2R
3.10 Twente – 2009 [37], [38], [54], [55]	• honeypots	• ICMP • IRC

DataSet Αναφορές	Αλγόριθμοι που εφαρμόστηκαν	Είδη επιθέσεων που αντιμετώπισαν
3.11 UNIBS – 2009 [73]	• GT	• Κανονική κίνηση και επιθέσεις
3.12 NSL-KDD [17], [18], [32], [35]	• Fuzzy Clustering NN • ANN • Bayesian Net with GR feature selection • C4.5 DT • One-class SVM • PCA • SVM • MLP • NB • DL RNN • K-means • Deep Auto-Encoder • Information Gain • DL • NDAE • Stacked NDAEs	• Probing • DoS • R2L • U2R
3.13 LLDOS 1.0 LLDOS 2.0.2 2011 [76], [77], [78]	• ANN • SNORT • Variable Order Markov • Probabilistic Suffix Tree	• DDoS
3.14 UMASS 2011 [37], [38], [43], [56], [57].	N/A	• TCP Base Attack
3.15 ISCX 2012 [37], [39], [53]	• K-Means Clustering και NB Classifier KMC+NBC	• Κανονική κίνηση και επιθέσεις
3.16 CTU-13 2011 [83], [84], [85], [86]	• BClus • CAMNEP • BotHunter • SVM	• botnet

DataSet Αναφορές	Αλγόριθμοι που εφαρμόστηκαν	Είδη επιθέσεων που αντιμετώπισαν
3.17 UNSW- NB15 [25]	• Deep Auto-Encoder • ANN • IXIA PerfectStorm tool	• Fuzzers, Analysis • Backdoors • DoS • Exploits • Generic • Reconnaissance, Shellcode • Worms
3.18 SLINGbot [72], [73]	• TrenMicro	• Botnet • Probing • DoS • Phishing • port scanning • email spam
3.19 ADFA-LD [27], [28], [29], [58], [59].	• Metasploit Penetration Testing Software • Offensive Security Training and Services	• Hydra-FTP • HydraSSH • Adduser • Java-Meterpreter • Meter-preter • Webshell
3.20 TUIDS [68]	• NetFlow	• Probing • DDoS • Phishing • port scanning
3.21 CIC DoS 2017 [60]	• CUSUM • Goldeneye • Slowloris	• DoS • Flood attacks
3.22 CIC-IDS 2017, 2018 [37], [38], [39], [41], [42], [61]	• CICFlowMeter	• DoS • DDoS • Brute Force • XSS • SQL injection • Infinltration • Port Scan • Botnet

DataSet Αναφορές	Αλγόριθμοι που εφαρμόστηκαν	Είδη επιθέσεων που αντιμετώπισαν
3.23 CIC-DDoS 2019 [62], [64]	• CICFlowMeter	• DDoS

4 Κεφάλαιο : Προβλήματα που παρατηρήθηκαν στα σύνολα δεδομένων

Σύμφωνα με τη παραπάνω μελέτη - καταγραφή επί των πιο δημοφιλή δημόσια σύνολα δεδομένων παρατηρούνται διάφορα θέματα τα οποία είναι καθοριστικά στην μέχρι σήμερα επίτευξη του στόχου της ερευνητικής κοινότητας για την δημιουργία ενός συνόλου δεδομένων ικανού να αντιμετωπίζει δυναμικά ένα μεγάλο εύρος επιθέσεων γνωστών και όχι μόνο.

Τα παραπάνω σύνολα δεδομένων τα οποία παρουσιάστηκαν στο προηγούμενο κεφάλαιο αποτελούν αναμφισβήτητα πολύτιμα στοιχεία για την ερευνητική κοινότητα για την ανίχνευση εισβολών. Ωστόσο, υποφέρουν από το γεγονός ότι δεν αποτυπώνουν ικανοποιητικά την κυκλοφορία σε πραγματικό συνθήκες.

Πολλά σύνολα δεδομένων δεν επισημαίνονται σωστά λόγω μη διαθεσιμότητας πραγματικών πληροφοριών επίθεσης.

Η αναλογία των κανονικών και των αναλογιών επίθεσης είναι διαφορετική σε διαφορετικά σύνολα δεδομένων.

Αρκετά υπάρχοντα σύνολα δεδομένων δεν είναι ενημερωμένα ώστε να αντικατοπτρίζουν τις πρόσφατες τάσεις στην κυκλοφορία δικτύου με την ενσωμάτωση εξελιγμένων επιθέσεων δικτύου.

Στα περισσότερα υπάρχοντα σύνολα δεδομένων έχει γίνει επεξεργασία για την ανωνυμοποίηση τους για την αντιμετώπιση θεμάτων ασφαλείας του Οργανισμού από τον οποίο παρήχθησαν. Τα ακατέργαστα δεδομένα δεν είναι δημόσια στην ερευνητική κοινότητα.

Στον ακόλουθο πίνακα, (πίνακας 9), παρουσιάζονται τα σύνολα δεδομένων που παρουσιάσθηκαν στο προηγούμενο κεφάλαιο με τις αναφορές σε αυτά, τους στόχους δημιουργίας τους, τον τύπο δεδομένων που περιλαμβάνουν και θα θέματα που εντοπίσθηκαν.

Πίνακας 9
Στόχοι δημιουργίας, τύπος δεδομένων που περιλαμβάνουν και θέματα που εντοπίσθηκαν

DataSet Αναφορές	Στόχος	Δεδομένα	Θέματα που εντοπίστηκαν
3.1 DARPA 1998, 1999, 2000 [8], [9], [10], [11], [12], [13], [14], [74], [75] , [76]	• Ανάλυση ασφάλειας δικτύων • Ανίχνευση ανωμαλιών	• Προσομείωση πραγματικής κίνησης	• Προκατάληψη τεχνικών ανίχνευσης • Μη επιτυχιμένη προσομείωση πραγματικής κίνησης • Απώλεια μεγάλων πακέτων • Μη ακριβή επισήμανση
3.2 KDD-99 1998 - 1999 [4], [5], [6], [8], [15], [16], [17], [18]	• Κυκλοφορία δικτύου • Ανίχνευση ανωμαλιών	• Πραγματική κίνηση	• Μεγάλος όγκος συνόλου δεδομένων • Επιβάρυνση υπολογιστικών πόρων • Περιττές εγγραφές • Επαναλαμβανόμενες εγγραφές • Προκατάληψη τεχνικών ανίχνευσης • Μη επικαιοοποιημένο
3.3 Gure KDD cup Gure KDD cup 6 percent 1999 [24]	• Κυκλοφορία δικτύου • Ανίχνευση ανωμαλιών	• Δεδομένα KDD 99 • Πακέτα με πλήρη ωφέλιμο φορτίο	N/A
3.4 DEFCON-8 DEFCON-11 2000 - 2002 [33], [34], [37], [39], [43], [70]	• Προστασία πόρων υπολογιστών	• Προσομείωση πραγματικής κίνησης • Παρεμβατικά δεδομένα	• Δεδομένα μη πραγματικής κίνησης

DataSet Αναφορές	Στόχος	Δεδομένα	Θέματα που εντοπίστηκαν
3.5 CAIDA 2002 / 2016 [49], [37], [39], [46].	• Κυκλοφορία δικτύου	N/A	• Πολύ ειδικά δεδομένα για συγκεκριμένους τύπους επίθεσης • Ανώνυμα δεδομένα χωρίς φορτίο πληροφορίας πρωτοκόλλου και προορισμού • Μη πλήρως δημόσιο σύνολο δεδομένων
3.6 LBNL 2004 / 2005 [37], [45], [46], [47]	• Κυκλοφορία δικτύου • ροές κυκλοφορίας	• Πραγματική κίνηση	• Τα ποσοστά επιθέσεων στις εγγραφές δυσανάλογα με τις εγγραφές παρασκηνίου • Πολυπλοκότητα • Ανωνυμοποίηση
3.7 ECML / PKDD 2007 [68], [79], [80], [81], [82]	• Κυκλοφορία δικτύου	• Πραγματική κίνηση	• Κατασκευή των αιτημάτων επίθεσης
3.8 CDX 2009 [37], [52], [53].	• Αντιμετώπιση προβλημάτων συνόλου δεδομένων DARPA 1998, 1999	• Προσομοίωση πραγματικής κίνησης	• Έλλειψη ποικιλομορφίας και όγκου κυκλοφορίας • Στερείται τυπικού θορύβου παρασκηνίου στο διαδίκτυο
3.9 Kyoto 2009 [37], [46], [50]	• Κυκλοφορία δικτύου	• Πραγματική κίνηση δικτύου • Στατιστικά χαρακτηριστικά του KDD Cup 99	• Περιορισμένη εικόνα κυκλοφορίας δικτύου • Καταγραφή μόνο επιθέσεις με στόχο στα honeypots • Κατασκευή αιτημάτων επίθεσης • Δεν υπάρχουν ψευδή θετικά στοιχεία

DataSet Αναφορές	Στόχος	Δεδομένα	Θέματα που εντοπίστηκαν
3.10 Twente – 2009 [37], [38], [54], [55]	• Ροές κυκλοφορίας δικτύου	• Ροές από συγκεκριμένο σημείο παρατήρησης στο δίκτυο	• Μη ακριβή επισήμανση • Περιέχει ορισμένες άγνωστες και μη συνδεδεμένες καταχωρίσεις κυκλοφορίας • Έλλειψη όγκου και ποικιλομορφίας επιθέσεων
3.11 UNIBS – 2009 [73]	• Κυκλοφορία δικτύου	• Πραγματική κίνηση δικτύου	N/A
3.12 NSL-KDD [17], [18], [32], [35]	• Αντιμετώπιση θεμάτων του ΚΔΔ 99	• Δεδομένα δικτύου εκτός σύνδεσης με βάση το KDD 99 • Επιλεγμένες εγγραφές του KDd 99	• Προβλήματα που κληρονομούνται από το KDD 99 σε μικρότερο βαθμό
3.13 LLDOS 1.0 LLDOS 2.0.2 2011 [76], [77], [78]	• Βελτίωση του DARPA 2000	• Προσομοίωση πραγματικής κίνησης	N/A
3.14 UMASS 2011 [37], [38], [43], [56], [57].	• Παρακολούθηση κίνησης δικτύου και ίχνη ασύρματων εφαρμογών	• Πραγματική κίνηση δικτύου	• Έλλειψη ποικιλομορφίας κυκλοφορίας και επιθέσεων

DataSet Αναφορές	Στόχος	Δεδομένα	Θέματα που εντοπίστηκαν
3.15 ISCX 2012 [37], [39], [53]	• Παρακολούθηση κίνησης δικτύου	• Πραγματική κίνηση δικτύου	• Δεν αντιπροσωπεύει νέα πρωτόκολλα δικτύου • Πολυπλοκότητα συνόλου • Απαίτηση ειδικών γνώσεων για την επισήμανση • Μείωση στην ποιότητα επαναληψιμότητας • Στερείται ρεαλιστικού θορύβου παρασκηνίου στο διαδίκτυο • Δεν υπάρχουν ίχνη HTTPS • Η συνολική κανονική κίνηση δεν είναι συγκρίσιμη με ένα πραγματικό δίκτυο
3.16 CTU-13 2011 [83], [84], [85], [86]	• Κίνηση κακόβουλων προγραμμάτων • Καταγραφή πραγματικής μικτής κίνησης botnet	• Πραγματική κίνηση δικτύου	N/A
3.17 UNSW- NB15 [25]	• Ανίχνευση ανωμαλιών	• Προσομοίωση πραγματικής κίνησης	• Επισήμανση μόνο για συγκεκριμένους τύπους επίθεσης
3.18 SLINGbot [72], [73]	• Δυνατότητα δημιουργίας καλοηθών botnets • Να μπορούν να χαρακτηρίζονται οι τρέχουσες και μελλοντικές πιθανές δομές εντολών και ελέγχου botnet • Να σχεδιάζονται αποτελεσματικές τεχνικές άμυνας.	• Πραγματική κίνηση δικτύου	• Αποκλείει κίνηση σχετική με τον εντοπισμό μελλοντικών bots, την πρόσβαση σε αυτά τα bots και την πραγματική εκτέλεση κακόβουλων δραστηριοτήτων

DataSet Αναφορές	Στόχος	Δεδομένα	Θέματα που εντοπίστηκαν
3.19 ADFA-LD [27], [28], [29], [58], [59].	• Αντιμετώπιση επιθέσεων με σύγχρονες τεχνικές • Αντιμετώπιση αδυναμιών και προβλημάτων KDD • Ανίχνευση ανωμαλιών	• Πραγματική κίνηση δικτύου	• Δυσκολία επισημάνσεων
3.20 TUIDS [68]	• Παρακολούθηση κίνησης δικτύου	• Πραγματική κίνηση δικτύου	N/A
3.21 CIC DoS 2017 [60]	• Ανίχνευση επιθέσεων DoS με αργό ρυθμό επιπέδου εφαρμογής που βασίζονται σε HTTP	• Προσομοίωση πραγματικής κίνησης	• Χρήση δειγματοληψίας για μείωση όγκου με αρνητική επίδραση στην ακρίβεια ανίχνευσης
3.22 CIC-IDS 2017, 2018 [37], [38], [39], [41], [42], [61]	• Παρακολούθηση κίνησης δικτύου	• Προσομοίωση πραγματικής κίνησης	• Αρκετά μεγάλο σύνολο δεδομένων • Μεγάλο υπολογιστικό κόστος • Περιττές εγγραφές • Δυσκολεύει την δημιουργία χαρακτηριστικών • Πιθανή προκατάληψη ταξινομητή προς την τάξη πλειοψηφίας
3.23 CIC-DDoS 2019 [62], [64]	• Ανίχνευσης και ομαδοποίηση με βάση ένα σύνολο χαρακτηριστικών ροής δικτύου	• Προσομοίωση πραγματικής κίνησης	• Ανισορροπία στις τάξεις • Προκατάληψη τεχνικών ανίχνευσης • Ασυμφωνία μεταξύ συνόλου δοκιμών και εκπαίδευσης

Στην συνέχεια παρουσιάζουμε μία αποτύπωση των προβλημάτων που υπάρχουν και εμποδίζουν την δημιουργία αποτελεσματικών συνόλων δεδομένων προς χρήση από τα συστήματα ανίχνευσης εισβολών.

4.1 Η έλλειψη κατάλληλων, δυναμικών, δημόσιων συνόλων δεδομένων

Η έλλειψη κατάλληλων δημόσιων συνόλων δεδομένων αποτελεί ένα σημαντικό πρόβλημα στην προσπάθεια της ερευνητικής κοινότητας για την δημιουργία ενός αποτελεσματικού IDS. Συνήθως, αυτή η έλλειψη οφείλεται σε ανησυχίες σχετικά με το απόρρητο των πληροφοριών που περιέχουν οι οποίες πληροφορίες μπορεί να εκθέσουν στοιχεία που προσβάλουν την ασφάλεια του συστήματος από το οποίο προήλθαν ή και στοιχεία που αφορούν προσωπικά δεδομένα.

Ειδικότερα η έλλειψη κατάλληλων δυναμικών δημόσιων συνόλων δεδομένων ιδίως όσον αφορά την ανίχνευση ανωμαλιών, αποτελεί σοβαρό θέμα το οποίο απασχολεί την ερευνητική κοινότητα [65], [66]. Αυτό συμβαίνει επειδή τα δεδομένα αυτά είναι προϊόντα Οργανισμών που για λόγους ασφάλειας δεν τα δημοσιοποιούν προς χρήση λόγω των ζητημάτων απορρήτου ενώ, αυτά που είναι διαθέσιμα είναι σε μεγάλο βαθμό ανώνυμα, δεν περιέχουν σημαντικές πληροφορίες όπως περιεχόμενο επικεφαλίδων πακέτων, στοιχεία διευθύνσεων κλπ, που θα μπορούσαν θα προκαλέσουν κίνδυνο στην ασφάλεια, στατιστικών χαρακτηριστικών καθώς και δεν αντικατοπτρίζουν τις τρέχουσες τάσεις.

Για τη χρήση των ιχνών στο διαδίκτυο ένα σημαντικό θέμα είναι ότι οι περισσότερες εγγραφές από αυτά, δεν είναι άμεσα διαθέσιμα χωρίς να έχουν προεπεξεργαστεί [43], το οποίο οφείλεται στο γεγονός ότι τα περισσότερα από τα διαθέσιμα ίχνη διαδικτύου είναι εγγραφές tcpdump που καταγράφηκαν και συμπιέστηκαν σε μορφές λήψης πακέτων (PCAP).

4.2 Κατάλληλα σύνολα δεδομένων

Η συμπεριφορά του δικτύου αλλάζει από δίκτυο σε δίκτυο. Έτσι, ένα σύνολο δεδομένων που προέρχεται από ένα συγκεκριμένο δίκτυο, το πιο πιθανό είναι να μην είναι αντιπροσωπευτικό σε διαφορετικά δίκτυα. Δεδομένου ότι η συμπεριφορά του διαδικτύου περιέχει μεγάλες ποσότητες ανώμαλης κυκλοφορίας, το

παραγόμενο σύνολο δεδομένων θα πρέπει να είναι σε θέση να αντιπροσωπεύει μια τέτοια συμπεριφορά.

4.3 Παλαιότητα συνόλων δεδομένων.

Πολλά από τα διαθέσιμα δημόσια σύνολα δεδομένων είναι πολύ παλαιά με αποτέλεσμα να καθίστανται παρωχημένα και ακατάλληλα για ισχυρούς επιστημονικούς ισχυρισμούς [23].

Δεδομένου ότι τόσο τα συστήματα όσο και οι στρατηγικές επιθέσεων εξελίσσονται συνεχώς και με γρήγορους ρυθμούς είναι αναμενόμενο ότι τα υπάρχοντα σύνολα δεδομένων που κατά την δημιουργία τους κάλυπταν τις τότε προδιαγραφές και προκλήσεις, σήμερα, να έχουν ξεπεραστεί και να μην είναι σε θέση να ανταποκριθούν με επάρκεια ή και καθόλου στις νέες σύγχρονες τεχνολογίες και ανάγκες.

4.4 Πεδίο εφαρμογής των συνόλων δεδομένων

Διάφοροι ερευνητές χρησιμοποιούν συγκεκριμένες μεθόδους για να εργαστούν σύμφωνα με τους στόχους τους. Αυτό έχει ως αποτέλεσμα η καταγραφή των κινήσεων στα σύνολα δεδομένων που χρησιμοποιούν να είναι ελεγχόμενη στις περισσότερες περιπτώσεις προκειμένου να ταιριάζει με τους στόχους της μελέτης τους. Χρησιμοποιώντας μεθόδους εξόρυξης δεδομένων, εκτελώντας προεπεξεργασία δεδομένων ή εκκαθάριση τους, προσπαθούν να μειώσουν πιθανά προβλήματα στην αντιστοίχιση των δεδομένων με τους στόχους των μελετών τους [67].

4.5 Στατικά σύνολα δεδομένων

Καθώς εξελίσσονται οι συμπεριφορές και τα πρότυπα του δικτύου όπως και οι στρατηγικές και τεχνικες των σύγχρονων επιθέσεων, η χρήση στατικών συνόλων δεδομένων από τα συστήματα ανίχνευσης εισβολών, περιορίζει τις δυνατότητες και την αποτελεσματικότητα τους. Τα στατικά σύνολα δεδομένων αφορούν συγκεκριμένες στρατηγικές και τύπους επιθέσεων και μπορούν να ανταποκριθούν με μεγάλη επιτυχία στον εντοπισμό τους, αλλά πάσχουν ολοκληρωτικά στην δυνατότητα εντοπισμού νέων - άγνωστων επιθέσεων.

Αντίθετα η δημιουργία και χρήση δυναμικά παραγόμενων συνόλων δεδομένων δίνουν την δυνατότητα εντοπισμού νέων τύπων επιθέσεων καθώς και είναι εύκολο να τροποποιηθούν, να επεκταθούν και να αναπαραχθούν.

4.6 Ζητήματα απορρήτου δεδομένων

Πολλά υπάρχοντα σύνολα δεδομένων είναι εσωτερικά δηλαδή έχουν δημιουργηθεί από κάποιο Οργανισμό και δεν μπορούν να είναι δημόσια προς χρήση από την ερευνητική κοινότητα λόγω ζητημάτων ιδιωτικότητας, ενώ άλλα είναι σε μεγάλο βαθμό ανώνυμα και δεν αντικατοπτρίζουν τις τρέχουσες τάσεις, ή στερούνται ορισμένων στατιστικών χαρακτηριστικών.

Το πρόβλημα της προστασίας της ιδιωτικής ζωής των δεδομένων που υποκαθιστά τις πολιτικές ασφαλείας, η ευαισθησία των ρεαλιστικών δεδομένων, οι κίνδυνοι αποκάλυψης ψηφιακών πληροφοριών και η έλλειψη εμπιστοσύνης [23], είναι οι βασικές αιτίες του προβλήματος.

Σύνολα δεδομένων τα οποία περιέχουν καταγραφές δεδομένων χρηστών είναι πιο ευαίσθητα. Περιλαμβάνουν δεδομένα που παρατηρούνται τοπικά, όπως πλήρες περιεχόμενο πακέτων, ωφέλιμων φορτίων δεδομένων κλπ. Ενώ είναι πολύ χρήσιμα τέτοιου είδους καταγραφές που θα ήταν χρήσιμες για την ανάλυση των πακέτων σε βάθος για την ανίχνευση εφαρμογών ή κακόβουλου λογισμικού, αυτά τα δεδομένα δεν είναι διαθέσιμα. Τα δεδομένα των χρηστών θέτουν σημαντικά ζητήματα απορρήτου και νομιμότητας, επομένως τέτοια δεδομένα σπάνια μπορούν να δημοσιοποιηθούν στην ερευνητική κοινότητα.

Δεδομένα που περιέχουν διευθύνσεις IP ή MAC, κεφαλίδες πακέτων, εγγραφές ροής και αρχεία καταγραφής συστήματος, δημιουργούν θέματα ασφαλείας, επειδή είναι δυνατόν να συσχετισθούν με την ταυτότητα του χρήστη, καθώς και στοιχεία που αφορούν την δομή του δικτύου.

4.7 Επισήμανση δεδομένων

Ορισμένα διαθέσιμα σύνολα δεδομένων πάσχουν από επισήμανση λόγω κενών ή ελλείπων στοιχείων. Κάποια από αυτά είναι τεχνητά επισημασμένα από ειδικούς στην ασφάλεια, διαδικασία η οποία πέρα του γεγονότος ότι απαιτεί χρόνο και κόστος, εγκυμονεί κινδύνους λαθών, προβλήματα ρεαλισμού, ή και ελλιπούς εύρους τύπων επιθέσεων. Τα επισημασμένα ίχνη είναι σημαντικά για τη σύγκριση της απόδοσης διαφορετικών τεχνικών ανίχνευσης, για τη μέτρηση της αποτελεσματικότητας των παραμέτρων και για την τελειοποίηση των συστημάτων.

Η επισήμανση αυτή απαιτείται να είναι ρεαλιστική, πλήρη και όσο το δυνατό να καλύπτει μεγάλο εύρος τύπου επιθέσεων.

4.8 Μέγεθος συνόλου δεδομένων

Αρκετά υπάρχοντα σύνολα δεδομένων αντιμετωπίζουν θέματα χαμηλής ακρίβειας ανίχνευσης, χαμηλής απόδοσης σε πραγματικό χρόνο, περιορισμένη επεκτασιμότητα λόγω του μεγάλου όγκου δεδομένων που αντιμετωπίζουν.

Το μέγεθος του συνόλου δεδομένων τόσο για εκπαίδευση όσο και για δοκιμές πρέπει να είναι λογικό. Αυτό καθιστά το σύνολο δεδομένων οικονομικό ως προς την κατανάλωση πόρων για την επεξεργασία του στη διεξαγωγή των πειραμάτων χωρίς να χρειάζεται τυχαία επιλογή ενός μικρού μέρους ή δειγματοληψία που εγκυμονεί αρκετά προβλήματα ως προς την μεθοδολογία της.

4.9 Ασυμφωνία μεταξύ συνόλου δοκιμών και συνόλου εκπαίδευσης.

Σε κάποια σύνολα δεδομένων, το σύνολο δοκιμών περιέχει ορισμένους τύπους επίθεσης που δεν περιλαμβάνονται στο σύνολο εκπαίδευσης. Αυτό αποτελεί πρόβλημα διότι το σύνολο εκπαίδευσης και δοκιμών έχει διαφορετική κατανομή. Η διαφορά αυτή στην κατανομή προκαλεί προκατάληψη στα συστήματα ταξινομητών με αποτέλεσμα να μειώνεται η αποτελεσματικότητα και η απόδοσή τους.

4.10 Περιττές ή επαναλαμβανόμενες εγγραφές στα σύνολα δεδομένων

Πολλά σύνολα δεδομένων αντιμετωπίζουν το πρόβλημα της παραγωγής μεγάλου πλήθους περιττών εγγραφών, το οποίο προκαλεί μεγάλο όγκο συνολικά, μεγάλο υπολογιστικό κόστος στην επεξεργασία του συνόλου, επηρεάζοντας έτσι την αποτελεσματικότητά και αμεσότητα του. Εκτός από το πρόβλημα των επαναλαμβανόμενων ίδιων εγγραφών το οποίο προκαλεί μεγάλο υπολογιστικό κόστους, κατανάλωση πόρων, προκαλεί προκατάληψη στους αλγόριθμους εκμάθησης προς τις συχνές εγγραφές και έτσι τους εμποδίζει να εκπαιδευτούν σε μη συχνές εγγραφές που είναι συνήθως πιο επιβλαβείς για δίκτυα όπως οι επιθέσεις U2R και R2L. Επιπλέον, η ύπαρξη επαναλαμβανόμενων εγγραφών στο

σύνολο δοκιμών προκαλεί την προκατάληψη των αποτελεσμάτων της αξιολόγησης από τις μεθόδους που έχουν καλύτερα ποσοστά ανίχνευσης στις συχνές εγγραφές.

4.11 Υπολογιστικό κόστος

Η συνεχώς αυξανόμενη πολυπλοκότητα των υποδομών δικτύων, η αυξανόμενη χωρητικότητα τους κ.α. προκαλεί την δημιουργία τεράστιου όγκου στα σύνολα δεδομένων. Η ανάγκη ανάλυσης του τεράστιου όγκου αυτού από τα συστήματα ανίχνευσης εισβολών, είναι από τα σημαντικά θέματα που επηρεάζουν την αποτελεσματικότητα τους, όγκος ο οποίος καταναλώνει πόρους και αυξάνει τον υπολογιστικό χρόνο σε εκπαίδευση και δοκιμές.

Για την αντιμετώπιση του προβλήματος έχουν προταθεί και δοκιμαστεί τεχνικές ανίχνευσης βασισμένες στην δειγματοληψία, που επιτρέπει την αντιμετώπιση των προβλημάτων κλιμάκωσης και την εφαρμογή παραδοσιακών προσεγγίσεων ανίχνευσης επιθέσεων. Η τεχνική της δειγματοληψίας χρησιμοποιείται ήδη σε σύγχρονες υποδομές δικτύου (π.χ., οι Cisco, Juniper Fortinet κ.α. (τυχαία δειγματοληψία n-out-of-N) όμως οι τεχνικές δειγματοληψίας εγκυμονούν κινδύνους όπως η παράλειψη σημαντικών εγγραφών κίνησης που επηρεάζουν την αποτελεσματικότητα των συστημάτων ανίχνευσης εισβολών.

4.12 Μεγάλο πλήθος χαρακτηριστικών

Συνέπεια του θέματος μεγάλου όγκου των συνόλων δεδομένων, είναι το μεγάλο πλήθος χαρακτηριστικών που προκύπτουν, με την επεξεργασία τους να αποτελεί κι αυτό με την σειρά του ένα σοβαρό θέμα αφού μπορεί να αυξήσει τον υπολογιστικό χρόνο στα συστήματα ανίχνευσης εισβολών.

Για την αντιμετώπισή του θέματος αυτού, είναι σημαντικό να επιλέγονται τα ποιο σημαντικά χαρακτηριστικά που αντιπροσωπεύουν τις συνολικές ιδιότητες του δικτύου ώστε να συμβάλλουν αποτελεσματικά στην διαδικασία ανίχνευσης, θέμα το οποίο απαιτεί εξειδικευμένη γνώση, χρόνο, κόπο και ακρίβεια από τους ειδικούς στον τομέα της ασφάλειας.

4.13 Επιλογή χαρακτηριστικών

Συνέπεια του μεγάλου όγκου δεδομένων πολλών συνόλων δεδομένων είναι η δημιουργία πλήθους άσχετων και περιττών χαρακτηριστικών που προκαλούν αργή

εκπαίδευση και διαδικασία δοκιμών, υψηλότερη κατανάλωση πόρων καθώς και χαμηλό ποσοστό ανίχνευσης. Η χρήση μηχανισμών επιλογής χαρακτηριστικών μπορεί να βελτιώσει την αποδοτικότητα του συστήματος. Η δημιουργία χαρακτηριστικών του συνόλου δεδομένων θα πρέπει να αντιπροσωπεύει τις συνολικές ιδιότητες του δικτύου επομένως είναι σημαντικό να επιλέγονται τα ποιο σημαντικά χαρακτηριστικά που αντιπροσωπεύουν τις συνολικές ιδιότητες του δικτύου ώστε να συμβάλλουν αποτελεσματικά στην διαδικασία ανίχνευσης όπως την μείωση στα ποσοστά ψευδούς συναγερμού, θέμα το οποίο απαιτεί εξειδικευμένη γνώση, χρόνο, κόπο και ακρίβεια από τους ειδικούς στον τομέα της ασφάλειας.

4.14 Περιορισμένη ικανότητα IDSs

Εκτός από τις απαιτήσεις υλικού, τα IDSs έχουν περιορισμό στο πλήθος πακέτων δικτύου που μπορούν να επεξεργαστούν. Παρατηρείται ότι τείνουν να ρίχνουν πολλά πακέτα κάθε φορά που είναι υπερφορτωμένα [69]. Για παράδειγμα, η εμπειρία δείχνει ότι οι περισσότεροι ανιχνευτές εισβολής δεν είναι σε θέση να επεξεργαστούν όλα τα πακέτα σε ένα σύνολο δεδομένων εκτός σύνδεσης που είναι μεγαλύτερο από 5 Gigabytes.

5 Συζήτηση και προτάσεις για λύσεις

Η ανάπτυξη αποτελεσματικών συστημάτων ανίχνευσης εισβολών με χρήση κατάλληλων και αποτελεσματικών συνόλων δεδομένων, για την διασφάλιση της ασφάλειας σε εμπιστευτικότητα, διαθεσιμότητα και ακεραιότητα των δεδομένων και διαδικασιών τόσο από εξωτερικές όσο και από εσωτερικές επιθέσεις είναι ένα θέμα το οποίο έχει απασχολήσει και εξακολουθεί να απασχολεί, τους ειδικούς στην ασφάλεια.

Στην συνέχεια παραθέτουμε προτάσεις οι οποίες συμβάλουν στην αντιμετώπιση προβλημάτων που καθιστούν δύσκολη την δημιουργία συνόλων δεδομένων ικανών για την ανάπτυξη αποτελεσματικών συστημάτων ανίχνευσης εισβολών.

Οι ερευνητές βασίστηκαν σε υπάρχοντα δημόσια σύνολα δεδομένων ή δημιούργησαν νέα σύμφωνα με τις ανάγκες της έρευνάς τους. Ωστόσο, τα διαθέσιμα σύνολα δεδομένων δεν διαθέτουν στοιχεία της πραγματικής κίνησης, γεγονός το οποίο ευθύνεται στο ότι τα περισσότερα συστήματα ανίχνευσης εισβολών να μην είναι εφαρμόσιμα για περιβάλλοντα παραγωγής [36]. Επιπλέον, τέτοια σύνολα δεδομένων δεν μπορούν να προσαρμοστούν στις συνεχείς αλλαγές στην δομή των δικτύων όπως νέοι κόμβοι, μεταβαλλόμενα φορτία κυκλοφορίας, μεταβαλλόμενη τοπολογία κλπ.

Στο [36] αναφέρεται ότι για να ληφθεί υπόψη ένα σύνολο δεδομένων, πρέπει να καλύπτει τις ακόλουθες ιδιότητες:

α. Να περιέχει πραγματική κυκλοφορία δικτύου (παρόμοια με την παραγωγή). Ένα σύνολο δεδομένων θα πρέπει να δημιουργείται με την παρακολούθηση της καθημερινής κίνησης με ρεαλιστικό τρόπο, όπως η καθημερινή, πραγματική κίνηση του δικτύου του Οργανισμού από το οποίο συλλέχθηκαν τα στοιχεία και να περιλαμβάνει τόσο την κανονική όσο και την ανώμαλη κυκλοφορία. Όπως έχει διαπιστωθεί από υπάρχουσες εργασίες μέχρι σήμερα, κάθε τεχνητή εισαγωγή ιχνών μετά τη σύλληψη επηρεάσει αρνητικά τα πρωτογενή δεδομένα και πιθανό είναι να δημιουργηθούν ασυνέπειες στο τελικό σύνολο δεδομένων. Ωστόσο, η λήψη ρεαλιστικών συνόλων δεδομένων μπορεί συνήθως να απαιτεί επίσημη έγκριση από τους διαχειριστές των δικτύων. Επίσης, ένα άλλο πολύ σημαντικό θέμα είναι ότι πρέπει να

λαμβάνονται υπόψη οι ηθικές αξίες, η προστασία προσωπικών δεδομένων και οι βέλτιστες πρακτικές.

β. Να είναι έγκυρο, με πλήρη σενάρια.

γ. Να είναι επισημασμένο, προσδιορίζοντας την τάξη κάθε εγγραφής ως κανονική ή επίθεση.

Η επισήμανση της κυκλοφορίας ως καλοήθους ή κακόβουλης πρέπει να στηρίζεται από κατάλληλα στοιχεία για κάθε περίπτωση. Στόχος είναι η παροχή επισημασμένων συνόλων δεδομένων τόσο σε επίπεδα πακέτων όσο και σε επίπεδα ροής για κάθε περίπτωση καλοήθους και κακόβουλης κυκλοφορίας.

δ. Να είναι παραμετρικό.

Για την δημιουργία ενός ορθού μοντέλου ταξινόμησης κανονικής από κακόβουλης κίνησης, είναι απαραίτητη η δυνατότητα παραμετροποίησης του μοντέλου. Η ανίχνευση ανωμαλίας δικτύου αναλαμβάνει το μοντέλο κανονικής κίνησης για τον εντοπισμό κακόβουλης κυκλοφορίας.

ε. Να είναι ακριβής.

Η επισήμανση κάθε στιγμιότυπου κυκλοφορίας σε ένα σύνολο δεδομένων πρέπει να είναι ακριβής, διότι έχει μεγάλη σημασία στην αξιολόγηση διαφόρων μηχανισμών ανίχνευσης. Η δημιουργία ενός συνόλου δεδομένων σε ένα ελεγχόμενο και ντετερμινιστικό περιβάλλον επιτρέπει τη διάκριση της ανωμαλίας από την κανονική κυκλοφορία συνεπώς, εξαλείφει την μη πρακτική διαδικασία της χειροκίνητης επισήμανσης η οποία απαιτεί χρόνο, κόστος, εξειδικευμένο άτομο στην ασφάλεια και εγκυμονεί και κινδύνους λαθών.

στ. Να μπορεί να ενημερωθεί εύκολα,

Οι νέες τεχνικές ανίχνευσης και αλγόριθμοι αναπτύσσονται συνεχώς για την ανίχνευση ανωμαλιών δικτύου. Είναι απαραίτητο να είναι δυνατή και εύκολη η ενημέρωση του σε κάθε νέα προσέγγιση, καθώς και να παρουσιάζουν βελτιώσεις σε σχέση με τις παλαιότερες μεθόδους στην απόδοση με βελτίωση της ποσοτικοποίησης τους.

ζ. Να είναι αναπαραγώγιμο.

Στο παραγόμενο σύνολο δεδομένων, η ερευνητική κοινότητα θα πρέπει να είναι σε θέση να επαναλάβει πειράματα και να έχει παρόμοια αποτελέσματα, όταν χρησιμοποιούνται ίδιες προσεγγίσεις. Αυτό είναι σημαντικό επειδή η

προτεινόμενη τεχνική πρέπει να αντιμετωπίσει τις συνεχώς εξελισσόμενες στρατηγικές και τύπους επιθέσεων και της δομής δικτύου.

η. Να είναι δημόσιο, επομένως δεν πρέπει να περιέχει εμπιστευτικά δεδομένα.

Επίσης στο [63] αναφέρεται ότι :

α. Η ύπαρξη παραλλαγής πρωτοκόλλων είναι μια σημαντική πτυχή του συνόλου δεδομένων IDS και,
β. Να υπάρχει κατάλληλη τεκμηρίωση για το περιβάλλον συλλογής χαρακτηριστικών και συνόλων δεδομένων.

Ένα βέλτιστο σύνολο χαρακτηριστικών θα πρέπει να χρησιμοποιείται για να αντιπροσωπεύει κανονικές και όλες τις πιθανές περιπτώσεις επίθεσης. Ένα σύνολο χαρακτηριστικών κατά τη δημιουργία ενός συνόλου δεδομένων είναι σημαντικό επειδή τα χαρακτηριστικά αυτά διαδραματίζουν σημαντικό ρόλο κατά την επικύρωση των μηχανισμών ανίχνευσης.

Η μελέτη της κυκλοφορίας δικτύου αποτελεί σημαντικό παράγοντα για την αποτελεσματικότητα ενός συνόλου δεδομένων. Η μελέτη της κίνησης δικτύου είναι σημαντική για τον χαρακτηρισμό τόσο της τυπικής όσο και της άτυπης κίνησης, συχνά κακόβουλης επισκεψιμότητας. Στόχος είναι η κατανόηση ποια κίνηση κυριαρχεί στο διαδίκτυο και πώς επηρεάζει τη μηχανική κυκλοφορίας, τις αρχιτεκτονικές δικτύου και το σχεδιασμό πρωτοκόλλων, δρομολογητών, τείχους προστασίας και άλλων συσκευών δικτύου.

Σημαντικό ρόλο στα συστήματα ανίχνευσης εισβολών διαδραματίζει το πλήθος των διαθέσιμων πληροφοριών, καθώς αποτελεί προϋπόθεση για την ανίχνευση ανώμαλης συμπεριφοράς, την αξιολόγηση και την ορθή ερμηνεία των αποτελεσμάτων. Επομένως, είναι σημαντικό για ένα σύνολο δεδομένων να περιλαμβάνει όλες τις αλληλεπιδράσεις δικτύου, εντός και μεταξύ εσωτερικών δικτύων.

Θέματα τα οποία αφορούν την προστασία της ιδιωτικής ζωής και σχετίζονται με την ανταλλαγή πραγματικών ιχνών δικτύου αποτελούν ένα από τα σημαντικότερα εμπόδια για τους ερευνητές ασφάλειας δικτύου, καθώς οι πάροχοι δεδομένων είναι συχνά απρόθυμοι να κοινοποιήσουν αυτού του είδους ευαίσθητες πληροφορίες. Για τον λόγο αυτό τα περισσότερα ευαίσθητα δεδομένα χρησιμοποιούνται εσωτερικά,

ή είναι ανώνυμα με το ωφέλιμο φορτίο να έχει αφαιρεθεί εντελώς, με αποτέλεσμα να περιορίζει την ερευνητική κοινότητα από την ακριβή αξιολόγηση και τη σύγκριση των συστημάτων τους και να μειώνεται η χρησιμότητα τους. Οι πληροφορίες αυτές δεν είναι χρήσιμες μόνο για την αξιολόγηση των συστημάτων που βασίζονται στην ανάλυση ωφέλιμων φορτίων με την ανάλυση σε βάθος πακέτου, αλλά και των συστημάτων που βασίζονται στην ακριβή ανάλυση εφαρμογών για την ορθή αξιολόγηση των συστημάτων και των μεθόδων τους.

Η δημιουργία συνόλων δεδομένων από πειραματικά δίκτυα για την προσομοίωση ορισμένων επιθέσεων σύμφωνα με τους στόχους για τους οποίους προορίζεται δίνει στους ερευνητές τον απόλυτο έλεγχο των παραγόμενων συνόλων δεδομένων. Επομένως η έρευνα είναι στοχευμένη για τα πρότυπα των επιθέσεων στο σύνολο δεδομένων. Οι ερμηνεία των αποτελεσμάτων είναι πιο εύκολη και αποτελεσματική για τον ερευνητή σε σύγκριση με σύνολα δεδομένων που λαμβάνονται από άλλες πηγές, όμως, οι ηθικές αξίες και οι βέλτιστες πρακτικές πρέπει επίσης να λαμβάνονται αυστηρά υπόψη για την προσομοίωση επιθέσεων με υπολογιστές και για τη δημοσίευση αποτελεσμάτων που προκύπτουν από πειραματικά δίκτυα.

6 Συμπεράσματα

Οι επιθέσεις παρουσιάζουν μία συνεχόμενη αύξηση σε συχνότητα, μέγεθος, ποικιλία και πολυπλοκότητα τα τελευταία χρόνια με στόχο τις υπηρεσίες και τις εφαρμογές.

Η πιθανότητα απόκτησης ή δημιουργίας ενός τέλειου συνόλου δεδομένων δεν είναι μεγάλη, δεδομένου ότι πάντα θα υπάρχουν προβλήματα τα οποία σχετίζονται με την εκρηκτική ανάπτυξη – εξέλιξη των υποδομών δικτύων και εφαρμογών καθώς και των στρατηγικών που εφαρμόζουν οι επιτιθέμενοι.

Η αποτύπωση μίας συστηματικής προσέγγισης για την δημιουργία συνόλων δεδομένων θα μπορούσε ενδεχομένως να δημιουργήσει τις προϋποθέσεις για το σκοπό αυτό.

Σε κάθε περίπτωση είναι επιτακτικός ο προσανατολισμός από στατικά σύνολα δεδομένων προς πιο δυναμικά παραγόμενα σύνολα δεδομένων τα οποία αντικατοπτρίζουν τις τρέχουσες συνθήκες κυκλοφορίας και τις σύγχρονες μεθόδους εισβολών, καθώς και είναι τροποποιήσιμα, επεκτάσιμα και αναπαραγώγιμα.

Η προσπάθεια αντιμετώπισης εισβολών δεν πρέπει να παραβλέπει την μελέτη του ίδιου του διαδικτύου ως ένα απαραίτητο συμπλήρωμα των παραπάνω εργαλείων, θέμα το οποίο είναι αρκετά δύσκολο. Το διαδίκτυο έχει ενσωματωθεί στη ζωή των ανθρώπων και έχουν προκύψει σοβαρά ζητήματα για προστασία της ιδιωτικής.

Επίσης, η ερευνητική κοινότητα μέχρι το πρόσφατο παρελθόν είχε επικεντρωθεί σε σύνολα δεδομένων για το παραδοσιακό, ενσύρματο διαδίκτυο. Δεδομένα ειδικά για άλλους τύπους πρόσβασης όπως τα ασύρματα δίκτυα, τηλεφωνικά δίκτυα κ.α. τα οποία βασίζονται όλο και περισσότερο σε IP, αλλά και ο διαφορετικός συνδυασμός εφαρμογών και μοτίβων χρήσης, μπορεί να επηρεάσουν τις έρευνες. Ορισμένοι πάροχοι συνόλων δεδομένων έχουν ήδη στρέψει την προσοχή τους σε ασύρματα συγκεκριμένα σύνολα δεδομένων.

Με αυτή η εργασία παρουσιάσαμε μερικές κρίσιμες προκλήσεις που αντιμετωπίζουν οι ερευνητές στις προσπάθειες αξιολόγησης των ερευνών στους τομείς των IDSs καθώς και των μειονεκτημάτων που συνεπάγεται η ύπαρξη και η εκδήλωση αυτών των προβλημάτων.

Θεωρούμε ότι σύνολα δεδομένων που δημιουργούνται από μια πηγή είναι ανεπαρκές για τη διερεύνηση επιθέσεων λόγω της ποικιλομορφίας των τύπων, στρατηγικών και πολυπλοκότητας των επιθέσεων παγκοσμίως [71].

Τα εμπόδια στις αξιολογήσεις των ερευνών σε αυτούς τους τομείς, όπως ζητήματα προστασίας της ιδιωτικής ζωής, περιορισμένο πεδίο εφαρμογής των συνόλων δεδομένων ακατάλληλες τεκμηριώσεις, ανωνυμία και πλαστά σύνολα δεδομένων, μειώνουν την αποτελεσματικότητα τους, σε αντίθεση με τη χρήση συνόλων δεδομένων από πολλαπλές πηγές. Στο κεφάλαιο 3 πραγματοποιήσαμε μια επισκόπηση των δημοφιλέστερων και ευρέως χρησιμοποιούμενων δημόσιων συνόλων δεδομένων. Πιστεύουμε ότι οι πληροφορίες που παρέχουμε σε αυτή την εργασία συμβάλει στις προσπάθειες των ερευνητών στον τομέα της ασφάλειας.

Στη συνέχεια προτείναμε λύσεις των προαναφερόμενων προβλημάτων και επισημάναμε πεδία που χρήζουν περαιτέρω διερεύνησης.

Δεδομένου ότι η ασφάλεια στον κυβερνοχώρο είναι μία διαρκή προσπάθεια ενάντια στις κακόβουλες πράξεις πιστεύουμε ότι συνεργατικές προσπάθειες από ερευνητές και πηγές δεδομένων προς δημόσια διάθεση, μπορούν να εξασφαλίσουν την δημιουργία ενημερωμένων αξιολογικών συνόλων δεδομένων. Τέλος, δεν πρέπει να παραβλέπεται από όλους τους συμμετέχοντες στην ασφάλεια η τήρηση και η διασφάλιση των ηθικών αξιών, των προσωπικών δεδομένων, η χρήση βέλτιστων πρακτικών για τη λήψη αξιολογικών συνόλων δεδομένων και τη δημοσίευση των αποτελεσμάτων των ευρημάτων τους.

7 Βιβλιογραφικές αναφορές

[1] Bace, R., & Mell, P. (2001). NIST special publication on intrusion detection systems. BOOZ-ALLEN AND HAMILTON INC MCLEAN VA.
[2] Mukherjee, B., Heberlein, L. T., & Levitt, K. N. (1994). Network intrusion detection. IEEE network, 8(3), 26-41.
[3] Heady, R., Luger, G., Maccabe, A., & Servilla, M. (1990). The architecture of a network level intrusion detection system (No. LA-SUB-93-219). Los Alamos National Lab., NM (United States); New Mexico Univ., Albuquerque, NM (United States). Dept. of Computer Science.
[4] Cup, K. D. D. (1999). http://kdd. ics. uci. edu/databases/kddcup99/kddcup99. html. The UCI KDD Archive.
[5] Stolfo, S. J., Fan, W., Prodromidis, A., Chan, P. K., & Lee, W. (2000). Cost-sensitive modeling for fraud and intrusion detection: Results from the JAM project. In In Proceedings of the 2000 DARPA Information Survivability Conference and Exposition.
[6] MIT Lincoln Labs, 1998 DARPA IntrusionDetection Evaluation. Available on : https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset
[7] Lippmann, R. P., Fried, D. J., Graf, I., Haines, J. W., Kendall, K. R., McClung, D., ... & Zissman, M. A. (2000, January). Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00 (Vol. 2, pp. 12-26). IEEE.
[8] McHugh, J. (2000). Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Transactions on Information and System Security (TISSEC), 3(4), 262-294.
[9] Mahoney, M. V., & Chan, P. K. (2003, September). An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection. In International Workshop on Recent Advances in Intrusion Detection (pp. 220-237). Springer, Berlin, Heidelberg.
[10] Axelsson, S. (2000). The base-rate fallacy and the difficulty of intrusion detection. ACM Transactions on Information and System Security (TISSEC), 3(3), 186-205.
[11] Gaffney, J. E., & Ulvila, J. W. (2000, May). Evaluation of intrusion detectors: A decision theory approach. In Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001 (pp. 50-61). IEEE.
[12] Di Crescenzo, G., Ghosh, A., & Talpade, R. (2005, September). Towards a theory of intrusion detection. In European Symposium on Research in Computer Security (pp. 267-286). Springer, Berlin, Heidelberg.
[13] Cárdenas, A. A., Baras, J. S., & Seamon, K. (2006, May). A framework for the evaluation of intrusion detection systems. In 2006 IEEE Symposium on Security and Privacy (S&P'06) (pp. 15-pp). IEEE.
[14] Gu, G., Fogla, P., Dagon, D., Lee, W., & Skorić, B. (2006, March). Measuring intrusion detection capability: An information-theoretic approach. In Proceedings of the 2006 ACM Symposium on Information, computer and communications security (pp. 90-101).
[15] Portnoy, L., Esking, L., & Stolfo, S. J. (2001). Intrusion detection with unlabeled data using clustering [CJ//Proc of ACM CSS Workshop on Data mining Applied to Security (DMSA-2001).
[16] Leung, K., & Leckie, C. (2005, January). Unsupervised anomaly detection in network intrusion detection using clusters. In Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38 (pp. 333-342).
[17] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE symposium on computational intelligence for security and defense applications (pp. 1-6). IEEE.
[18] “Nsl-kdd data set for network-based intrusion detection systems.” Available on: http://nsl.cs.unb.ca/NSL-KDD/, March 2009.

[19] Yasin, M. M., & Awan, A. A. (2004, June). A study of host-based IDS using system calls. In 2004 International Networking and Communication Conference (pp. 36-41). IEEE.
[20] Lichodzijewski, P., Zincir-Heywood, A. N., & Heywood, M. I. (2002, May). Host-based intrusion detection using self-organizing maps. In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290) (Vol. 2, pp. 1714-1719). IEEE.
[21] Lichodzijewski, P., Zincir-Heywood, A. N., & Heywood, M. I. (2002, May). Dynamic intrusion detection using self-organizing maps. In The 14th Annual Canadian Information Technology Security Symposium (CITSS).
[22] Lei, J. Z., & Ghorbani, A. (2004, May). Network intrusion detection using an improved competitive learning neural network. In Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004. (pp. 190-197). IEEE.
[23] Ojo, N. J. (2011). Methods for reducing workload during investigations of Intrusion Logs (Doctoral dissertation, The University of Essex).
[24] GureKDDCup: http://www.sc.ehu.es/acwaldap/
[25] Moustafa, N., & Slay, J. (2015, November). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In 2015 military communications and information systems conference (MilCIS) (pp. 1-6). IEEE.
[26] Xi, K., & Hu, J. (2010). Bio-cryptography. In Handbook of Information and Communication Security (pp. 235-255). Springer, Berlin, Heidelberg.
[27] Creech, G. (2013). Hu J Generation of a new IDS test dataset: time to retire the KDD collection. IEEE Wireless Communications and Networking Conference (WCNC.2013)
[28] “Metasploit Penetration Testing Software,” http://www.metasploit.com, Accessed 24 Apr. 2012.
[29] “Offensive Security Training and Services,” http://www.offensivesecurity.com, Accessed 22 May 2012.
[30] “TikiWiki: CMS groupware,” http://info.tiki.org/Tiki+Wiki+CMS+ Groupware, Accessed 19 May 2012.
[31] “Tiki Wiki CMS Groupware Remote PHP Code Injection.”
[32] Ingre, B., & Yadav, A. (2015, January). Performance analysis of NSL-KDD dataset using ANN. In 2015 international conference on signal processing and communication engineering systems (pp. 92-96). IEEE.
[33] Nehinbe, J. O. (2009, September). A simple method for improving intrusion detections in corporate networks. In International Conference on Information Security and Digital Forensics (pp. 111-122). Springer, Berlin, Heidelberg.
[34] T. S. Group, “Defcon 8, 10 and 11,” 2000. [Online]. Available: http://cctf.shmoo.com/
[35] Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In Neural networks for perception (pp. 65-93). Academic Press.
[36] Eduardo K. Viegas, Altair O. Santin, and Luiz S. Oliveira. 2017. Toward a reliable anomaly-based intrusion detection in real-world environments. Computer Networks 127 (2017), 200–216. https://doi.org///doi.org/10.1016/j.comnet.
[37] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp. 2018.
[38] Gharib, A., Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2016, December). An evaluation framework for intrusion detection dataset. In 2016 International Conference on Information Science and Security (ICISS) (pp. 1-6). IEEE.
[39] Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A. A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. computers & security, 31(3), 357-374.
[40] Lashkari, A. H., Draper-Gil, G., Mamun, M. S. I., & Ghorbani, A. A. (2017, February). Characterization of tor traffic using time based features. In ICISSp (pp. 253-262).
[41] Panigrahi, R., & Borah, S. (2018). A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems. International Journal of Engineering & Technology, 7(3.24), 479-482.

[42] Karatas, G., Demir, O., & Sahingoz, OK (2020). Αύξηση της απόδοσης των IDS που βασίζονται σε μηχανική μάθηση σε ένα μη ισορροπημένο και ενημερωμένο σύνολο δεδομένων. Πρόσβαση IEEE , 8 , 32150-32162.
[43] Nehinbe, J. O. (2011, September). A critical evaluation of datasets for investigating IDSs and IPSs researches. In 2011 IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS) (pp. 92-97). IEEE.
[44] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2015). An empirical evaluation of information metrics for low-rate and high-rate DDoS attack detection. Pattern Recognition Letters, 51, 1-7.
[45] Nechaev, B., Allman, M., Paxson, V., & Gurtov, AV (2010). Μια προκαταρκτική ανάλυση της απόδοσης TCP σε ένα εταιρικό δίκτυο. INM / WREN , 10.
[46] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2015). Towards Generating Real-life Datasets for Network Intrusion Detection. IJ Network Security, 17(6), 683-701.
[47] Lawrence Berkeley National Laboratory (LBNL), ICSI, LBNL/ICSI Enterprise Tracing Project, 2005. (http://www.icir.org/enterprise-tracing/)
[48] CACE Technologies, WinPcap, June 2015. (http://www.winpcap.org)
[49] CAIDA, The Cooperative Analysis for Internet Data Analysis, 2011. (http://www.caida.org)
[50] Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D., & Nakao, K. (2011, April). Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. In Proceedings of the first workshop on building analysis datasets and gathering experience returns for security (pp. 29-36).
[51] Mokube, I., & Adams, M. (2007, March). Honeypots: concepts, approaches, and challenges. In Proceedings of the 45th annual southeast regional conference (pp. 321-326).
[52] Sangster, B., O'Connor, T. J., Cook, T., Fanelli, R., Dean, E., Morrell, C., & Conti, G. J. (2009, August). Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets. In CSET.
[53] Zuech, R., Khoshgoftaar, T. M., Seliya, N., Najafabadi, M. M., & Kemp, C. (2015, April). A new intrusion detection benchmarking system. In The Twenty-Eighth International Flairs Conference.
[54] Sperotto, A., Sadre, R., Van Vliet, F., & Pras, A. (2009, October). A labeled data set for flow-based intrusion detection. In International Workshop on IP Operations and Management (pp. 39-50). Springer, Berlin, Heidelberg.
[55] Quittek, J., Zseby, T., Claise, B., Zander, S.: Requirements for IP Flow Information Export (IPFIX). RFC 3917 (Informational).
[56] U. of Massachusetts Amherst, “Optimistic tcp acking,” 2011. [Online]. Available: http://traces.cs.umass.edu/
[57] Prusty, S., Levine, B. N., & Liberatore, M. (2011, October). Forensic investigation of the OneSwarm anonymous filesharing system. In Proceedings of the 18th ACM conference on Computer and communications security (pp. 201-214).
[58] Chawla, A., Lee, B., Fallon, S., & Jacob, P. (2018, September). Host based intrusion detection system with combined CNN/RNN model. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 149-158). Springer, Cham.
[59] Xie, M., & Hu, J. (2013, December). Evaluating host-based anomaly detection systems: A preliminary analysis of adfa-ld. In 2013 6th International Congress on Image and Signal Processing (CISP) (Vol. 3, pp. 1711-1716). IEEE.
[60] Jazi, H. H., Gonzalez, H., Stakhanova, N., & Ghorbani, A. A. (2017). Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling. Computer Networks, 121, 25-36.
[61] Yulianto, A., Sukarno, P., & Suwastika, N. A. (2019, March). Improving adaboost-based intrusion detection system (IDS) performance on CIC IDS 2017 dataset. In Journal of Physics: Conference Series (Vol. 1192, No. 1, p. 012018). IOP Publishing.

[62] Sharafaldin, I., Lashkari, A. H., Hakak, S., & Ghorbani, A. A. (2019, October). Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In 2019 International Carnahan Conference on Security Technology (ICCST) (pp. 1-8). IEEE.
[63] Sharafaldin, I., Gharib, A., Lashkari, A. H., & Ghorbani, A. A. (2018). Towards a reliable intrusion detection benchmark dataset. Software Networking, 2018(1), 177-200.
[64] Can, D. C., Le, H. Q., & Ha, Q. T. (2021). Detection of Distributed Denial of Service Attacks using Automatic Feature Selection with Enhancement for Imbalance Dataset. ACIIDS 2021.
[65] Sommer R, Paxson V. Outside the closed world: on using machine learning for network intrusion detection. In: Security and privacy, IEEE Symposium on; 2010. p. 305e16.
[66] Tavallaee M, Stakhanova N, Ghorbani AA. Toward credible evaluation of anomaly-based intrusion detection methods. Trans Sys Man Cyber Part C 2010;40:516e24.
[67] Heidemann, J., & Papdopoulos, C. (2009, March). Uses and challenges for network datasets. In 2009 Cybersecurity Applications & Technology Conference for Homeland Security (pp. 73-82). IEEE.
[68] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2015). Towards Generating Real-life Datasets for Network Intrusion Detection. IJ Network Security, 17(6), 683-701.
[69] Buchanan, B., Flandrin, F., Macfarlane, R., & Graves, J. (2011). A methodology to evaluate rate-based intrusion prevention system against distributed denial-of-service (DDoS). Cyberforensics 2011.
[70] CTFC (Capture the flag contest) defcon datasets, http://cctf.shmoo.com/data/, Accessed 09.
[71] Ghorbani, A. A., Lu, W., & Tavallaee, M. (2009). Network intrusion detection and prevention: concepts and techniques (Vol. 47). Springer Science & Business Media.
[72] Jackson, A. W., Lapsley, D., Jones, C., Zatko, M., Golubitsky, C., & Strayer, W. T. (2009, March). SLINGbot: A system for live investigation of next generation botnets. In 2009 Cybersecurity Applications & Technology Conference for Homeland Security (pp. 313-318). IEEE.
[73] UNIBS, University of Brescia Dataset, 2009. (http://www.ing.unibs.it/ntw/tools/traces/)
[74] MIT Lincon, 1998 DARPA Intrusion Detection Evaluation Dataset (https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset)
[75] MIT Lincon, 1999 DARPA Intrusion Detection Evaluation Dataset (https://www.ll.mit.edu/r-d/datasets/1999-darpa-intrusion-detection-evaluation-dataset)
[76] MIT Lincon, 2000 DARPA Intrusion Detection Scenario Specific Datasets (https://www.ll.mit.edu/r-d/datasets/2000-darpa-intrusion-detection-scenario-specific-datasets)
[77] Li, Z. T., Lei, J., Wang, L., & Li, D. (2007, August). A data mining approach to generating network attack graph for intrusion prediction. In Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007) (Vol. 4, pp. 307-311). IEEE.
[78] Kholidy, H. A., Yousof, A. M., Erradi, A., Abdelwahed, S., & Ali, H. A. (2014, October). A finite context intrusion prediction model for cloud systems with a probabilistic suffix tree. In 2014 European Modelling Symposium (pp. 526-531). IEEE.
[79] Wang, J., & Paschalidis, I. C. (2016). Botnet detection based on anomaly and community detection. IEEE Transactions on Control of Network Systems, 4(2), 392-404.
[80] U. Wijesinghe, T. Udaya and V. Vijay, "An enhanced model for network flow based botnet detection", Proceedings of the 38th Australasian Computer Science Conference (ACSC 2015), vol. 27, 2015.
[81] F. Haddadi, L. Le, D. Porter and L. Zincir-Heywood, "On the Effectiveness of Different Botnet Detection Approaches", ISPEC, pp. 121-135, 2015.
[82] Gallagher, B., & Eliassi-Rad, T. (2009). Classification of http attacks: a study on the ECML/PKDD 2007 discovery challenge (No. LLNL-TR-414570). Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).
[83] CTU-13 botnet traffic dataset, https://mcfp.weebly.com/, 2011.

[84] Garcia, S., Grill, M., Stiborek, J., & Zunino, A. (2014). An empirical comparison of botnet detection methods. computers & security, 45, 100-123.
[85] Kato, K., & Klyuev, V. (2014). An intelligent ddos attack detection system using packet analysis and support vector machine. IJICR, 14(5), 478-485.
[86] Bhamare, D., Salman, T., Samaka, M., Erbad, A., & Jain, R. (2016, December). Feasibility of supervised machine learning for cloud security. In 2016 International Conference on Information Science and Security (ICISS) (pp. 1-5). IEEE.