ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μελέτη της Επίδρασης Δειγματοληψίας στη Διαδικασία Ανίχνευσης Ανωμαλιών στο Διαδίκτυο

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Γεώργιος Ο. Ανδρουλιδάκης

Αθήνα, Ιούλιος 2009

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μελέτη της Επίδρασης Δειγματοληψίας στη Διαδικασία Ανίχνευσης Ανωμαλιών στο Διαδίκτυο

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Γεώργιος Ο. Ανδρουλιδάκης

Συμβουλευτική Επιτροπή : Συμεών Παπαβασιλείου
Βασίλειος Μάγκλαρης
Ανδρέας Γεώργιος Σταφυλοπάτης

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την:

.....	.....	.....
Συμεών Παπαβασιλείου	Βασίλειος Μάγκλαρης	Ανδρέας Γεώργιος Σταφυλοπάτης
Επ. Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ
.....	.....	.....
Ευστάθιος Συκάς	Μιλτιάδης Αναγνώστου	Μιχαήλ Θεολόγου
Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ
	.....	
	Χρήστος Δουληγέρης	
	Καθηγητής Παν. Πειραιώς	

Αθήνα, Ιούλιος 2009

.....

Γεώργιος Ο. Ανδρουλιδάκης

Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στα πλαίσια της παρούσας διδακτορικής διατριβής γίνεται μελέτη της επίδρασης της δειγματοληψίας πάνω στο πεδίο της ανίχνευσης ανωμαλιών στο περιβάλλον του Διαδικτύου. Η ερευνητική προσπάθεια σχετικά με την δειγματοληψία που έχει γίνει μέχρι σήμερα αφορά γενικές διαχειριστικές εφαρμογές σε ένα δίκτυο και εστιάζει κυρίως στη μελέτη διάφορων στατιστικών στοιχείων κίνησης όπως είναι η κατανομή του μεγέθους των ροών πακέτων, του μέσου μήκους ροής, του συνολικού αριθμού ροών, κτλ. σε ένα δείγμα δεδομένων κίνησης δικτύου. Στην παρούσα διατριβή προτείνεται και αναλύεται μια μεθοδολογία δειγματοληψίας κατάλληλη για εφαρμογή στο πεδίο της ανίχνευσης ανωμαλιών στο περιβάλλον του Διαδικτύου. Η ιδέα της συγκεκριμένης δειγματοληπτικής μεθόδου βασίζεται στην παρατήρηση ότι οι περισσότερες από τις ανωμαλίες του δικτύου προκαλούνται από μικρές ροές (με μικρό αριθμό πακέτων). Η προτεινόμενη μέθοδος ονομάζεται «Επιλεκτική Δειγματοληψία» και εστιάζει στην δυναμική και προνομιακή επιλογή των μικρών ροών (σε πακέτα) στις οποίες και παρουσιάζονται ανωμαλίες όπως είναι οι επιθέσεις άρνησης υπηρεσίας (DDoS attacks) και οι αυτοδιαδιδόμενοι ιοί (worm propagation). Επιπρόσθετα, προτείνουμε και αναλύουμε μια μέθοδο δειγματοληψίας δύο σταδίων, συνδυάζοντας κατάλληλα τη δειγματοληψία ροών και πακέτων, και κάνοντας χρήση της «Επιλεκτικής Δειγματοληψίας». Τα αποτελέσματα καταδεικνύουν ότι η προτεινόμενη προσέγγιση βελτιώνει σε μεγάλο βαθμό την αποτελεσματικότητα της ανίχνευσης ανωμαλιών, ενώ συγχρόνως μειώνει τον αριθμό των επιλεγέντων δεδομένων. Στη συνέχεια, γίνεται μελέτη της επίδρασης «Ευφυών Δειγματοληπτικών Μεθόδων» στην ανίχνευση και κατηγοριοποίηση διαφόρων ανωμαλιών δικτύου. Συγκεκριμένα, δείχνουμε πως μέσα από συγκεκριμένες δειγματοληπτικές μεθόδους επιτυγχάνουμε την «μεγέθυνση» των ανωμαλιών, και επομένως την αποτελεσματικότερη ανίχνευσή τους. Με αυτόν τον τρόπο, η δειγματοληψία από μία διαδικασία με απώλειες πληροφοριών από το σύνολο των δεδομένων, μετατρέπεται σε ένα ευεργετικό χαρακτηριστικό γνώρισμα στην περίπτωση της ανίχνευσης ανωμαλιών δικτύου, επιτρέποντας την ανίχνευση ανωμαλιών οι οποίες δεν θα ήταν ανιχνεύσιμες σε άλλη περίπτωση, προσδίδοντας μεγαλύτερη αποτελεσματικότητα στην ανίχνευση και κατηγοριοποίηση των ανωμαλιών δικτύου. Η αξιολόγηση των προτεινόμενων δειγματοληπτικών μεθόδων πραγματοποιείται με την εφαρμογή τους σε διαφορετικές τεχνικές ανίχνευσης ανωμαλιών σε πραγματικά δεδομένα δικτύου που έχουν συλλεχθεί από τη σύνδεση του δικτύου του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ) με το Εθνικό Δίκτυο Έρευνας και Τεχνολογίας (ΕΔΕΤ).

Λέξεις Κλειδιά

Ανίχνευση Ανωμαλιών Δικτύου, Δειγματοληψία, Δικτυακές Μετρήσεις, Ασφάλεια Δικτύων

Abstract

In this work we study and evaluate the effect of sampling on the field of network anomaly detection in the Internet. Most of the existing work on sampling is related to general network management functions and focuses on the study of some general statistical traffic properties such as flow size distribution, average flow size, or total number of flows in a sample of network data. In the present thesis a novel sampling methodology suitable for application on the field of network anomaly detection is proposed and analyzed. The idea of the proposed sampling method is based on the observation that most of the network anomalies are caused by small flows (with a small number of packets). The proposed method is named "Selective Sampling" and it focuses on the dynamic and preferential selection of small flows (in packets) which are usually the source of many network attacks such as Distributed Denial of Service (DDoS) Attacks and worm propagation. In addition, we propose and analyze a method of two-stage sampling, combining suitably the sampling of flows and packets, while making use of the proposed "Selective Sampling" method. The results show that the proposed approach improves significantly the anomaly detection effectiveness, while at the same time decreases the number of selected data. Moreover, we study the effect of "Intelligent Sampling Methods" in the detection and classification of various anomalies on the network. We show that through sampling techniques that opportunistically and preferentially sample traffic data we achieve to "magnify" the appearance of anomalies within the sampled data set and therefore improve their detection. Therefore, the inherently "lossy" sampling process is transformed to an advantageous feature in the anomaly detection case, allowing the revealing of anomalies that would be otherwise untraceable and thus becoming the vehicle for efficient anomaly detection and classification. The evaluation of the proposed sampling methods is achieved through the use of different anomaly detection techniques on real network data that have been collected from the link that connects the National Technical University of Athens (NTUA) with the Greek Research and Technology Network (GRNET).

Keywords

Network Anomaly Detection, Sampling, Network Measurements, Network Security

Ευχαριστίες

Κατά τη διάρκεια της εκπόνησης της διδακτορικής διατριβής μου ήταν πολλοί οι άνθρωποι που με στήριξαν και με βοήθησαν και προέρχονται τόσο από το ακαδημαϊκό όσο και από το φιλικό και συγγενικό μου περιβάλλον.

Θα ήθελα να ευχαριστήσω μέσα από την καρδιά μου τον επιβλέποντα καθηγητή μου κ. Συμεών Παπαβασιλείου για την ουσιαστική υποστήριξη που μου προσέφερε. Η συνεργασία μας όλα αυτά τα χρόνια ήταν άψογη και η επίβλεψη του διδακτορικού έγινε με μεράκι. Επίσης, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου κ. Βασίλη Μάγκλαρη για την βοήθεια του αλλά και την εμπιστοσύνη που μου έδειξε όταν με δέχθηκε ως υποψήφιο διδάκτορα στο εργαστήριο Διαχείρισης και Βέλτιστου Σχεδιασμού Δικτύων (NETMODE), καθώς και για τη διπλωματική μου εργασία την οποία επέβλεψε, και η οποία αποτέλεσε προπομπό της διδακτορικής διατριβής. Η πορεία αυτή θα ήταν πολύ πιο δύσβατη χωρίς την άψογη συνεργασία και τις γόνιμες συζητήσεις με τα υπόλοιπα μέλη του εργαστηρίου NETMODE και ιδιαίτερα τον Βασίλη Χατζηγιαννάκη και τη Μαίρη Γραμματικού.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου και τα αδέρφια μου για την ανεκτίμητη υποστήριξη και φροντίδα που μου έχουν προσφέρει. Τους αγαπώ και τους ευχαριστώ μέσα από την καρδιά μου.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1. Εισαγωγή

Καθώς το Διαδίκτυο συνεχίζει να μεγαλώνει γρήγορα τόσο σε μέγεθος όσο και σε πολυπλοκότητα, έχει γίνει πλέον φανερό ότι η εξέλιξη του είναι στενά δεμένη με την κατανόηση και μελέτη της δικτυακής κίνησής του. Οι μετρήσεις της κίνησης στο δίκτυο [McGr00][Deri00] είναι σημαντικές για μια πληθώρα εφαρμογών, όπως είναι ο σχεδιασμός της χωρητικότητας των γραμμών μεταφοράς, η κοστολόγηση των πελατών με βάση την κίνηση, οι διαγνώσεις σφαλμάτων σε στοιχεία του δικτύου, οι παραβιάσεις ασφαλείας και ειδικότερα η ανίχνευση ανωμαλιών στο δίκτυο.

Πιο συγκεκριμένα, ο εντοπισμός δικτυακών οντοτήτων που παράγουν μεγάλη δικτυακή κίνηση, όπως π.χ. οι δημοφιλέστεροι ιστοχώροι (websites), είναι μία πολύ χρήσιμη πληροφορία για τους παρόχους δικτυακών υπηρεσιών. Επιπρόσθετα, ένας φορέας παροχής δικτυακών υπηρεσιών στηριζόμενος στις δικτυακές μετρήσεις μπορεί να καθορίζει την κίνηση μεταξύ των συνόλων διευθύνσεων πηγής και προορισμού που μεταφέρονται πάνω από μια κορεσμένη δικτυακή σύνδεση. Αυτές οι μετρήσεις θα μπορούσαν να χρησιμοποιηθούν για να εξετάσουν τη δυνατότητα πραγματοποίησης αναδρομολόγησης (rerouting) τμήματος της κίνησης μακριά από την κορεσμένη σύνδεση [Feld00][Feld01].

Επιπρόσθετα, οι δικτυακές μετρήσεις συχνά χρησιμοποιούνται για την χρέωση των πελατών από τους φορείς παροχής δικτυακών υπηρεσιών, με βάση τον όγκο των δεδομένων που ανταλλάχθηκε. Το ποσό της χρέωσης μπορεί να εξαρτηθεί από τον τύπο της εφαρμογής (όπως προσδιορίζεται από τις θύρες TCP/UDP), ή τις απομακρυσμένες διευθύνσεις [Duff01b]. Επίσης, η μελέτη της δικτυακής κίνησης αποτελεί σημαντικό παράγοντα για την αποτελεσματική ανίχνευση εισβολών στα δίκτυα, συμπεριλαμβανομένων των αλλαγών στα προφίλ της χρήσης συγκεκριμένων πρωτοκόλλων και θυρών TCP/UDP, και τον εντοπισμό των περισσότερο ενεργών δικτυακών συσκευών [Tayl01][Reev02].

Λόγω της μεγάλης κίνησης των σημερινών δικτύων υπάρχει μία δυσκολία στην αποθήκευση και επεξεργασία όλων αυτών των πληροφοριών που αφορούν τα δεδομένα του δικτύου. Σε αυτό το πρόβλημα, λύση έρχεται να δώσει η δειγματοληψία (sampling), στην οποία έχει επικεντρωθεί το ενδιαφέρον της επιστημονικής κοινότητας ως ένας τρόπος συλλογής στατιστικών στοιχείων για τον τεράστιο όγκο δεδομένων του δικτύου [Duff05][Hohn06]. Στόχος της δειγματοληψίας αποτελεί η

απόκτηση όσο το δυνατόν μεγαλύτερης πληροφορίας/γνώσης για το σύστημα που μελετάμε, ενώ ταυτόχρονα γίνεται προσπάθεια να μειωθεί ο όγκος των δεδομένων που συλλέγονται, ώστε η διαδικασία αυτή να είναι εφικτή και κλιμακούμενη.

Το σημαντικότερο σημείο στη διαδικασία της δειγματοληψίας είναι η ακρίβεια των αποτελεσμάτων που παρέχει. Στο περιβάλλον του Διαδικτύου, όπου η κίνηση αλλάζει δυναμικά και ακαθόριστα, η ανακριβής δειγματοληψία μπορεί να οδηγήσει σε λανθασμένες αποφάσεις τους υπεύθυνους ενός δικτύου. Ένα άλλο σημαντικό θέμα που σχετίζεται με την ακρίβεια στη δειγματοληψία σε ένα δίκτυο είναι η αποτελεσματικότητα και η αποδοτικότητα που αυτή παρέχει, δηλαδή με απλά λόγια, πόσα πακέτα χρειάζεται να συλλέξουμε για να έχουμε αξιόπιστα αποτελέσματα.

Η ερευνητική προσπάθεια σχετικά με την δειγματοληψία που έχει γίνει μέχρι σήμερα αφορά κυρίως τη μελέτη διάφορων στατιστικών στοιχείων κίνησης όπως είναι η κατανομή του μεγέθους των ροών πακέτων, του μέσου μήκους ροής, του συνολικού αριθμού ροών, κτλ. σε ένα δείγμα δεδομένων κίνησης δικτύου. Παρόλο που αυτές οι προσεγγίσεις είναι ενδιαφέρουσες για γενικές διαχειριστικές εφαρμογές σε ένα δίκτυο, δεν επαρκούν για το πεδίο της ανίχνευσης ανωμαλιών σε ένα δίκτυο. Οι τεχνικές ανίχνευσης ανωμαλιών δικτύου [Barf02][Ye02][Lee01] βασίζονται στην ανάλυση της δικτυακής κίνησης και στον προσδιορισμό των στατιστικών στοιχείων της κίνησης του δικτύου. Η κεντρική ιδέα της ανίχνευσης ανωμαλιών στηρίζεται στο γεγονός ότι σε περίπτωση εμφάνισης κάποιας ανωμαλίας στο δίκτυο, π.χ. κάποια επίθεση, τότε παρατηρείται απόκλιση στα στατιστικά στοιχεία του δικτύου από τις συνήθεις τιμές τους κατά την ομαλή λειτουργία.

Πρέπει να σημειωθεί εδώ ότι το συγκεκριμένο πρόβλημα της επίδρασης της δειγματοληψίας στη διαδικασία ανίχνευσης ανωμαλιών δικτύου είναι αρκετά διαφορετικό και πιο περίπλοκο από τα αντίστοιχα προβλήματα επίδρασης της δειγματοληψίας σε άλλες διαδικασίες διαχείρισης του δικτύου. Αυτό οφείλεται κυρίως στο γεγονός ότι η ανίχνευση ανωμαλιών είναι μία διαδικασία που εφαρμόζεται κάτω από ανώμαλες συνθήκες δικτύου (π.χ. επιθέσεις), ενώ από τη φύση της περιλαμβάνει ταυτόχρονα διάφορους παράγοντες, όπως η κανονική κίνηση του δικτύου, η ανώμαλη κίνηση, τα διάφορα μετρικά ανίχνευσης, των οποίων οι στατιστικές ιδιότητες και η συμπεριφορά μπορεί να επηρεαστεί σε μεγάλο βαθμό και με αρκετά διαφορετικούς τρόπους με την εφαρμογή της δειγματοληψίας.

Από τα παραπάνω είναι προφανές ότι το πρόβλημα της ανίχνευσης ανωμαλιών δικτύου γίνεται σαφώς πιο πολύπλοκο όταν εφαρμόζεται δειγματοληψία στα

δεδομένα του δικτύου. Η υπάρχουσα ερευνητική προσπάθεια γύρω από το πεδίο αυτό εστιάζει κυρίως στον τρόπο που οι υπάρχουσες μέθοδοι δειγματοληψίας επηρεάζουν συγκεκριμένους αλγόριθμους ανίχνευσης ανωμαλιών. Στόχος της παρούσας διδακτορικής διατριβής είναι να μελετήσει την επίδραση της δειγματοληψίας πάνω στο πεδίο της ανίχνευσης ανωμαλιών δικτύου και να προτείνει εναλλακτικές αποτελεσματικές μεθόδους δειγματοληψίας κατάλληλες για την εφαρμογή τους στο συγκεκριμένο πεδίο.

Ένας από τους βασικούς στόχους αυτής της διατριβής είναι να αποκτηθεί γνώση για τη δυνατότητα πραγματοποίησης αποτελεσματικής ανίχνευσης ανωμαλιών δικτύου, με την ανάλυση και κατανόηση της ισορροπίας ανάμεσα στη μείωση του όγκου των συλλεχθέντων δεδομένων και την διατήρηση της ακρίβειας και της αποτελεσματικότητας στην ανίχνευση ανωμαλιών.

Συγκεκριμένα, στην παρούσα διατριβή προτείνεται και αναλύεται μια μεθοδολογία δειγματοληψίας κατάλληλη για εφαρμογή στο πεδίο της ανίχνευσης ανωμαλιών στο περιβάλλον του Διαδικτύου. Η ιδέα της συγκεκριμένης δειγματοληπτικής μεθόδου βασίζεται στην παρατήρηση ότι οι περισσότερες από τις ανωμαλίες του δικτύου προκαλούνται από μικρές ροές (με μικρό αριθμό πακέτων). Η προτεινόμενη μέθοδος ονομάζεται «Επιλεκτική Δειγματοληψία» (Selective Sampling) και εστιάζει στην επιλογή των μικρών ροών (σε πακέτα) στις οποίες και παρουσιάζονται ανωμαλίες όπως είναι οι επιθέσεις άρνησης υπηρεσίας (DoS attacks) και οι αυτοδιαδιδόμενοι ιοί (worm propagation). Ένα από τα κύρια χαρακτηριστικά της προτεινόμενης μεθόδου δειγματοληψίας είναι η προσαρμοστικότητα της σε σχέση με άλλες προσεγγίσεις καθώς μπορεί να ελέγξει και να μειώσει τον αριθμό των επιλεγμένων ροών. Με την κατάλληλη παραμετροποίηση μπορούμε αφενός να επιλέξουμε ένα σημαντικό ποσοστό των μικρών ροών χωρίς να μειωθεί η αποτελεσματικότητα στην ανίχνευση ανωμαλιών, και αφ' ετέρου, να μειωθεί περαιτέρω η συμμετοχή στη διαδικασία δειγματοληψίας ροών με μεγάλο αριθμό πακέτων. Αυτό το γεγονός είναι ιδιαίτερα σημαντικό επειδή μειώνεται δραστικά η ποσότητα των επιλεγέντων πακέτων που πρέπει να επεξεργαστούν από τον αλγόριθμο ανίχνευσης ανωμαλιών. Η αξιολόγηση της μεθόδου πραγματοποιείται με την εφαρμογή της σε διαφορετικές τεχνικές ανίχνευσης ανωμαλιών σε πραγματικά δεδομένα δικτύου που έχουν συλλεχθεί από τη σύνδεση του δικτύου του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ) με το Εθνικό Δίκτυο Έρευνας και Τεχνολογίας (ΕΔΕΤ).

Επίσης προτείνουμε και αναλύουμε μια μέθοδο δειγματοληψίας δύο σταδίων, συνδυάζοντας κατάλληλα τη δειγματοληψία ροών και πακέτων, και κάνοντας χρήση της «Επιλεκτικής Δειγματοληψίας». Τα αποτελέσματα καταδεικνύουν ότι η προτεινόμενη προσέγγιση βελτιώνει σε μεγάλο βαθμό την αποτελεσματικότητα της ανίχνευσης ανωμαλιών, ενώ συγχρόνως μειώνει τον αριθμό των επιλεγέντων δεδομένων, επιτυγχάνοντας στις περισσότερες περιπτώσεις να ξεπεράσει ακόμη και τα αντίστοιχα αποτελέσματα στην περίπτωση που δεν εφαρμόζεται κάποια μορφή δειγματοληψίας. Πρέπει να υπογραμμιστεί εδώ ότι ο στόχος της προτεινόμενης μεθόδου είναι να επιλεγούν τα κατάλληλα δεδομένα προκειμένου να «μεγεθυνθούν» οι ανωμαλίες, βελτιώνοντας κατά συνέπεια περαιτέρω την αποτελεσματικότητα της ανίχνευσης, και όχι η διατήρηση των στατιστικών ιδιοτήτων του αρχικού δείγματος της δικτυακής κίνησης.

Με βάση την ανάλυση και την παρατήρηση ότι για συγκεκριμένες εφαρμογές όπως η ανίχνευση ανωμαλιών, ένα μεγάλο μέρος των πληροφοριών που μας είναι χρήσιμες περιλαμβάνεται σε ένα μικρό μέρος των ροών, καταδεικνύουμε ότι με τη χρησιμοποίηση «Ευφυών Τεχνικών Δειγματοληψίας», επιτυγχάνουμε τη μεγέθυνση της εμφάνισης των ανωμαλιών. Η «μεγέθυνση» μίας ανωμαλίας πραγματοποιείται με την επιλογή των κατάλληλων δεδομένων στο επιλεγέν δείγμα από το σύνολο των δικτυακών δεδομένων. Με τον τρόπο αυτό επιτυγχάνουμε τη βελτίωση της αποτελεσματικότητας της ανίχνευσης και σε μερικές περιπτώσεις πετυχαίνουμε την ανίχνευση ανωμαλιών που θα ήταν αόρατες σε άλλη περίπτωση.

Επομένως, υποστηρίζουμε πως μπορεί να προκύψει μία ολόκληρη νέα κατηγορία μεθόδων δειγματοληψίας, αποκαλούμενη «Ευκαιριακή Δειγματοληψία» (Opportunistic Sampling), η οποία στοχεύει να αντιστρέψει το μειονέκτημα της απώλειας πληροφοριών κατά τη δειγματοληψία, σε ένα σημαντικό ευεργετικό χαρακτηριστικό γνώρισμα στην περίπτωση της ανίχνευσης ανωμαλιών δικτύου.

Πιο αναλυτικά η παρούσα διατριβή διαρθρώνεται ως εξής.

- Στο κεφάλαιο 2 γίνεται ταξινόμηση των μεθόδων δειγματοληψίας και αναλυτική παρουσίαση των κυριοτέρων τεχνικών σε κάθε κατηγορία. Ταυτόχρονα, παρουσιάζονται οι σημαντικότερες ερευνητικές εργασίες που έχουν εφαρμόσει τις συγκεκριμένες τεχνικές δειγματοληψίας στο χώρο των δικτύων.

- Στη συνέχεια στο κεφάλαιο 3 γίνεται μία εισαγωγή στα συστήματα ανίχνευσης εισβολών. Περιγράφονται οι διάφορες κατηγορίες και εστιάζουμε κυρίως στα συστήματα ανίχνευσης ανωμαλιών δικτύου, όπου παραθέτουμε αναλυτικά τους μηχανισμούς λειτουργίας της κάθε μεθόδου, περιγράφοντας τις αντίστοιχες ερευνητικές εργασίες που πραγματοποιούν αξιολόγηση αυτών των μεθόδων.

- Στο κεφάλαιο 4 συνοψίζονται τα μειονεκτήματα των υπάρχουσων ερευνητικών εργασιών πάνω στο πρόβλημα που μελετά αυτή η διατριβή και αναλύουμε την επίδραση γνωστών τεχνικών δειγματοληψίας πακέτων που έχουν οριστεί στο PSAMP IETF draft [Psamp] σε διάφορους αλγόριθμους ανίχνευσης ανωμαλιών μέσα από πειραματικά αποτελέσματα που έχουν γίνει με πραγματικά δικτυακά δεδομένα που έχουν συλλεχθεί από την σύνδεση του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ) με το Εθνικό Δίκτυο Έρευνας και Τεχνολογίας (ΕΔΕΤ).

- Στο κεφάλαιο 5 προτείνουμε μία νέα προσέγγιση πάνω στο πρόβλημα της ανίχνευσης ανωμαλιών κάτω από συνθήκες δειγματοληψίας. Συγκεκριμένα προτείνουμε μία νέα μέθοδο δειγματοληψίας που τείνει να επιλέγει τις μικρές ροές στο δίκτυο, στις οποίες συγκεντρώνονται συνήθως τα περιστατικά των ανωμαλιών. Αναλυτικότερα, γίνεται μελέτη της προτεινόμενης δειγματοληπτικής μεθόδου (την αποκαλούμε «Επιλεκτική Δειγματοληψία»), σε δύο τεχνικές ανίχνευσης ανωμαλιών στο δίκτυο και πραγματοποιείται σύγκριση με προϋπάρχουσες τεχνικές δειγματοληψίας.

- Στο κεφάλαιο 6, περιγράφεται μια σύνθετη μέθοδος δειγματοληψίας δύο σταδίων και γίνεται αξιολόγηση της επίδοσης της μεθόδου στην ανίχνευση ανωμαλιών δικτύου. Επίσης προτείνεται η ενσωμάτωση και υιοθέτηση της «Επιλεκτικής Δειγματοληψίας» στην μέθοδο των δύο σταδίων για την αποτελεσματικότερη ανίχνευση των ανωμαλιών. Η αξιολόγηση της προτεινόμενης σύνθετης μεθόδου γίνεται με εφαρμογή της σε μία μέθοδο ανίχνευσης ανωμαλιών που βασίζεται στην έννοια της εντροπίας.

- Το κεφάλαιο 7 αποτελεί μία συνολική μελέτη της επίδρασης «Ευφυών Δειγματοληπτικών Μεθόδων» στην ανίχνευση και κατηγοριοποίηση διαφόρων ανωμαλιών δικτύου. Συγκεκριμένα, δείχνουμε πως μέσα από συγκεκριμένες δειγματοληπτικές μεθόδους επιτυγχάνουμε την «μεγέθυνση» των ανωμαλιών, και επομένως αποτελεσματικότερη ανίχνευσή τους. Στην μελέτη μας εστιάζουμε σε τρεις γνωστές ανωμαλίες (προερχόμενες από κακόβουλη χρήση του δικτύου), οι οποίες θα μπορούσαν να χαρακτηριστούν ως επιθέσεις δικτύων (DDoS, worm propagation, portscan) και δύο άλλες ανωμαλίες που προκαλούνται από νόμιμη χρήση του δικτύου (flash crowd, alpha flows)

- Τέλος στο κεφάλαιο 8, ολοκληρώνουμε την διατριβή παρουσιάζοντας τα συμπεράσματα αλλά και τις κατευθύνσεις μελλοντικής έρευνας που προέκυψαν κατά την εκπόνηση της παρούσας διατριβής.

2. Δειγματοληψία (Sampling)

Ως δειγματοληψία ορίζουμε την διαδικασία με την οποία επιλέγουμε ένα μέρος από τα δεδομένα του συστήματος το οποίο μελετούμε, εξάγοντας στη συνέχεια συμπεράσματα για την ολική συμπεριφορά του συστήματος από τα συλλεχθέντα δεδομένα. Το βασικό κίνητρο για τη δειγματοληψία στο περιβάλλον των δικτύων είναι να επιλεγεί ένα αντιπροσωπευτικό υποσύνολο των πακέτων που να επιτρέπει τις ακριβείς εκτιμήσεις των ιδιοτήτων του συνολικού δείγματος της κίνησης του δικτύου. Στη συνέχεια αυτού του κεφαλαίου περιγράφουμε τις σημαντικότερες τεχνικές δειγματοληψίας, χωρισμένες στις εξής κατηγορίες: Δειγματοληψία Πακέτων, Δειγματοληψία Ροών, Προσαρμοστικές μέθοδοι δειγματοληψίας, Σύνθετες μέθοδοι δειγματοληψίας.

2.1. Δειγματοληψία πακέτων (Packet Sampling)

Διάφορες τεχνικές δειγματοληψίας πακέτων χρησιμοποιούνται προκειμένου να μειωθεί το πλήθος των επεξεργαζόμενων δεδομένων. Στη συνέχεια περιγράφουμε τις κυριότερες από αυτές που έχουν οριστεί στο PSAMP IETF-draft [Psamp]. Το Σχήμα 2.1 απεικονίζει τις τρεις βασικές τεχνικές δειγματοληψίας που καθορίζονται στο προαναφερθέν κείμενο.

- ***Συστηματική δειγματοληψία (Systematic Sampling)***

Η συστηματική δειγματοληψία περιγράφει τη διαδικασία επιλογής των σημείων έναρξης και τη διάρκεια των διαστημάτων επιλογής σύμφωνα με μια ντετερμινιστική (deterministic) συνάρτηση. Πρακτικά, στο πεδίο των δικτύων αυτό εφαρμόζεται με την περιοδική επιλογή πακέτων με βάση την χωρική τους θέση στην ακολουθία του συνόλου, και συχνά απλουστεύεται στην επιλογή του κ-πακέτου κάθε φορά.

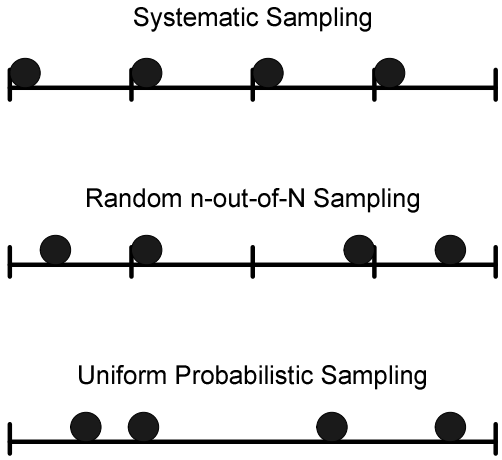
- ***Τυχαία δειγματοληψία ν-από-N (Random n-out-of-N sampling)***

Στην τυχαία δειγματοληψία ν-από-N, ο αρχικός πληθυσμός των στοιχείων διαιρείται σε τμήματα των N πακέτων κάθε ένα, και ν στο πλήθος πακέτα επιλέγονται τυχαία από κάθε τμήμα. Ένα παράδειγμα θα ήταν να παραχθούν οι διαφορετικοί τυχαίοι αριθμοί ν από το διάστημα [1,N] και να επιλεγούν όλα τα πακέτα που έχουν

μια θέση πακέτων ίση με έναν από τους τυχαίους αριθμούς. Συνήθως αυτή η μεθοδολογία χρησιμοποιείται με ν=1.

- ***Ομοιόμορφη πιθανολογική τυχαία δειγματοληψία (Uniform Probabilistic Random Sampling)***

Στην πιθανολογική δειγματοληψία η απόφαση για το εάν ένα πακέτο επιλέγεται ή όχι γίνεται σύμφωνα με μια προκαθορισμένη πιθανότητα επιλογής. Για την ομοιόμορφη πιθανολογική τυχαία δειγματοληψία κάθε πακέτο επιλέγεται ανεξάρτητα με την ίδια πιθανότητα *p*.



Σχήμα 2.1. Οι τρεις βασικές μέθοδοι δειγματοληψίας πακέτων

Το πρότυπο PSAMP του IETF καθορίζει επίσης μεθόδους δειγματοληψίας με βάση τον χρόνο (time-driven). Πιο συγκεκριμένα, καθορίζει τη συστηματική μέθοδο με βάση το χρόνο δειγματοληψίας στην οποία τα πακέτα επιλέγονται σε ένα συγκεκριμένο διάστημα έναρξης και λήξης, καθώς επίσης και το ανάλογο προς το χρόνο της ομοιόμορφης πιθανολογικής δειγματοληψίας που έχει το χρόνο μεταξύ των σημείων δειγματοληψίας έναρξης και λήξης κατανεμημένο εκθετικά. Στην παρούσα εργασία δεν πρόκειται να εξετάσουμε τις μεθόδους δειγματοληψίας που είναι βασισμένες στο χρόνο, επειδή τείνουν να χάνουν τις περιόδους έκρηξης (bursty periods) με πολλά πακέτα σε μικρό χρονικό διάστημα, με συνέπεια να αποτελούν ακατάλληλες επιλογές για τη χρήση στον τομέα της ανίχνευσης ανωμαλιών δικτύου.

Η εφαρμογή της δειγματοληψίας πακέτων στις μετρήσεις κίνησης των δικτύων μελετήθηκε αρχικά χρησιμοποιώντας πραγματικά δικτυακά δεδομένα από το δίκτυο NSFNET [Nsfnt]. Ο Claffy [Claf93] αξιολόγησε τις κλασσικές μεθόδους

δειγματοληψίας που περιγράψαμε παραπάνω (τόσο αυτές που βασίζονται στην αρίθμηση όσο και αυτές που βασίζονται στον χρόνο) και υπολόγισε τα στατιστικά των κατανομών του μεγέθους των πακέτων και του ενδιάμεσου χρόνου μεταξύ της άφιξης δύο διαδοχικών πακέτων (interarrival time). Τα αποτελέσματα έδειξαν ότι οι τεχνικές που βασίζονταν στο χρόνο (time-driven) δεν είχαν τα αναμενόμενα αποτελέσματα σε σχέση με τις μεθόδους που είχαν σαν βάση την αρίθμηση (count-driven), ενώ οι διαφορές στην απόδοση των μεθόδων της ίδιας κατηγορίας (με βάση την αρίθμηση ή το χρόνο) ήταν μικρές. Το πρόβλημα του υπολογισμού των κατανομών ροών (flow distributions) κάτω από συνθήκες δειγματοληψίας πακέτων έχει μελετηθεί στις εργασίες [Duff05] και [Hohn06]. Συγκεκριμένα στην εργασία [Duff05] μία στατιστική τεχνική καθώς και οι πληροφορίες των επιπέδων πρωτοκόλλων χρησιμοποιούνται για να υπολογίσουν τη μορφή των κατανομών ροών, εξάγοντας συμπεράσματα για τις ιδιότητες των ροών της αρχικής κίνησης του δικτύου. Επιπλέον, οι συγγραφείς της εργασίας [Hohn06] συγκρίνουν τη δειγματοληψία πακέτων με τη δειγματοληψία ροών και δείχνουν ότι η δειγματοληψία ροών βελτιώνει την ακρίβεια εκτίμησης διαφόρων στατιστικών μεγεθών όπως είναι η κατανομή μεγέθους των ροών (flow size distribution).

2.2. Δειγματοληψία Ροών (Flow Sampling)

Στη «δειγματοληψία ροών» τα πακέτα ταξινομούνται πρώτα σε ροές (flows). Μια ροή ορίζεται ως ένα σύνολο πακέτων που έχουν κοινά τα ακόλουθα χαρακτηριστικά: IP διεύθυνση πηγής (source IP address), θύρα πηγής (source port), IP διεύθυνση προορισμού (destination IP address), θύρα προορισμού (destination port) και πρωτόκολλο. Σε αυτήν την περίπτωση, η δειγματοληψία εφαρμόζεται στις ροές, διαδικασία η οποία οδηγεί στην επιλογή όλων των πακέτων που αποτελούν μια ροή. Συνήθης τρόπος δειγματοληψίας σε ροές είναι η **«Τυχαία Δειγματοληψία Ροών» (Random Flow Sampling)** κατά την οποία κάθε μεμονωμένη ροή επιλέγεται ανεξάρτητα με πιθανότητα *p*.

Επιπρόσθετα, στην εργασία [Duff03] προτείνεται μια νέα μέθοδος δειγματοληψίας ροών που ονομάζεται **«Έξυπνη Δειγματοληψία» (Smart Sampling)**. Στην «Έξυπνη δειγματοληψία» επιλέγονται όλες οι ροές που έχουν μέγεθος πάνω από κάποιο κατώφλι, ενώ οι υπόλοιπες ροές επιλέγονται με πιθανότητα

ανάλογη του μεγέθους τους. Αυτό το σχήμα δειγματοληψίας που έχει ως στόχο τις μεγάλες ροές είναι κατάλληλο για την χρησιμοποίηση του από παρόχους δικτύου που θέλουν να χρεώνουν τους πελάτες τους με βάση το μέγεθος των δεδομένων που διακινούν στο δίκτυο [Duff01b]. Άλλες δικτυακές εφαρμογές στις οποίες είναι χρήσιμη η συλλογή των μεγάλων ροών είναι οι μηχανισμοί ελέγχου της ποιότητας υπηρεσίας (Quality of Service) [Fang99] και η σχεδίαση και επέκταση του παρόντος δικτύου [Feld00].

Μία άλλη μέθοδος δειγματοληψίας που βασίζεται στην έννοια των ροών είναι η μέθοδος **«sample-and-hold»** που αποτελεί πρόταση των Estan και Varghese [Esta02]. Στη μέθοδο αυτή, για κάθε πακέτο γίνεται ένας έλεγχος ύπαρξης της ροής, στην οποία ανήκει το συγκεκριμένο πακέτο, στον πίνακα με τις επιλεγμένες ροές. Αν η ροή αυτή υπάρχει (από προηγούμενο επιλεγμένο πακέτο), τα στατιστικά της ροής ενημερώνονται, ενώ εάν το πακέτο αυτό ανήκει σε ροή που δεν είναι γνωστή προηγουμένως τότε δημιουργείται η νέα αυτή ροή με πιθανότητα $1 – (1 – p)^s$ όπου s είναι το μέγεθος του πακέτου σε bytes. Κατά συνέπεια η πιθανότητα μιας ροής που περιλαμβάνει b bytes να μην επιλεγεί καθόλου είναι $(1 - p)^b$, ανεξάρτητα του τρόπου με τον οποίο τα bytes της ροής είναι κατανεμημένα μεταξύ των πακέτων της.

2.3. Προσαρμοστικές μέθοδοι δειγματοληψίας (Adaptive Sampling)

Ο ρυθμός δειγματοληψίας (sampling rate) καθορίζει σε μεγάλο βαθμό την ακρίβεια στις εκτιμήσεις των πραγματικών μεγεθών [Jedw92][Choi02][Choi07]. Υπάρχουν συμπεριφορές δικτύων που είναι δύσκολο να ανιχνευθούν επακριβώς με χαμηλούς ρυθμούς δειγματοληψίας, όπως για παράδειγμα η ανίχνευση κάποιας ανωμαλίας στο δίκτυο. Από την άλλη πλευρά, η εφαρμογή δειγματοληψίας με υψηλό ρυθμό παράγει τεράστιο όγκο δεδομένων που πρέπει να σταλούν σε κάποιο σύστημα-συλλέκτη για να υποβληθούν σε επεξεργασία. Επιπρόσθετα, σε περιόδους υψηλής κίνησης στο δίκτυο, ο εξοπλισμός μπορεί να μην καταφέρει να ανταποκριθεί στον απαραίτητο ρυθμό δειγματοληψίας με αποτέλεσμα να απορρίψει ένα ποσοστό της κίνησης. Κατά συνέπεια είναι προφανές ότι υπάρχει μια λεπτή σχέση ανάμεσα στην ακρίβεια των μετρήσεων και στην απόδοση του συστήματος που πραγματοποιεί τη

δειγματοληψία. Αυτό που δεν είναι προφανές είναι ο τρόπος να επιλεγεί ο κατάλληλος ρυθμός δειγματοληψίας.

Τα προαναφερθέντα ζητήματα αντιμετωπίζονται από μια ομάδα προσαρμοστικών τεχνικών δειγματοληψίας. Αυτές οι μέθοδοι υιοθετούν είτε κάποιον ειδικό ευριστικό (heuristic) αλγόριθμο για την πραγματοποίηση της δειγματοληπτικής διαδικασίας, είτε κάποιους μηχανισμούς πρόβλεψης για τη μελλοντική κίνηση στο δίκτυο, προσαρμόζοντας κατάλληλα τον ρυθμό δειγματοληψίας. Η προσαρμογή του ρυθμού δειγματοληψίας έχει βέβαια μερικά μειονεκτήματα. Το κυριότερο από αυτά είναι η αδυναμία στην ταχύτητα της προσαρμογής του ρυθμού δειγματοληψίας ανάλογα με την κίνηση του δικτύου. Ώσπου να πραγματοποιηθεί η μεταβολή στο ρυθμό δειγματοληψίας, υπάρχει περίπτωση σε μεγάλη έκρηξη κίνησης (traffic burst) η συσκευή που πραγματοποιηθεί την δειγματοληψία να μην μπορεί να ανταποκριθεί. Παρόλα αυτά, αρκετές τεχνικές έχουν αναπτυχθεί με σκοπό την προσαρμογή του ρυθμού δειγματοληψίας.

Στην εργασία [Drob98] περιγράφονται δύο μέθοδοι προσαρμοστικής δειγματοληψίας για τη διαχείριση της χρήσης του επεξεργαστή σε συσκευές δικτύου. Η μία μέθοδος χρησιμοποιεί τις πληροφορίες για την τρέχουσα χρησιμοποίηση του επεξεργαστή προκειμένου να ρυθμιστεί το ποσοστό επιλογής πακέτων. Η άλλη μέθοδος χρησιμοποιεί τους interarrival χρόνους πακέτων (που μπορούν να χρησιμοποιηθούν για να προσδιορίσουν την εισερχόμενη έκρηξη της κίνησης) από κοινού με τη γνώση για το χρόνο επεξεργασίας που απαιτείται για την ανάλυση ενός δείγματος. Η μελέτη διαπίστωσε ότι οι προτεινόμενες προσαρμοστικές μέθοδοι παρήγαγαν ακριβέστερες εκτιμήσεις των παραμέτρων κίνησης του δικτύου (παράμετρος Hurst που χρησιμοποιείται για την περιγραφή των self-similar φαινομένων) κάτω από τον προκαθορισμένο περιορισμό των πόρων σε σχέση με την απλή συστηματική δειγματοληψία. Στην εργασία [Hern01] μελετάται η εφαρμογή ενός βρόχου ελέγχου (control loop) για τη ρύθμιση του ρυθμού δειγματοληψίας. Αναλύονται μέθοδοι γραμμικής πρόβλεψης (linear prediction) και ασαφούς λογικής (fuzzy logic) και εφαρμόζονται σε διαφορετικούς τύπους κίνησης (κανονική κίνηση του δικτύου και εκρηκτική κίνηση βίντεο (bursty video traffic)).

Ο Choi στην εργασία [Choi04] πρότεινε μια προσαρμοστική τεχνική δειγματοληψίας πακέτων για την μέτρηση της κίνησης δικτύου σε επίπεδο ροών, η οποία παρέχει μία μη πολωμένη εκτίμηση του μεγέθους ροής (από την άποψη του αριθμού πακέτων και χαρακτήρων) για τις μεγάλες ροές. Στην εργασία [Xu05a]

προτείνεται μια προσαρμοστική δειγματοληπτική μέθοδος για διαφορετικές διακυμάνσεις της κίνησης στο δίκτυο. Η μέθοδος μπορεί δυναμικά να ρυθμίσει την πιθανότητα δειγματοληψίας των πακέτων με βάση το μέγεθος της διακύμανσης της κίνησης και μπορεί να επιτύχει μεγαλύτερη ακρίβεια δειγματοληψίας σε σχέση με την στατική τυχαία δειγματοληψία πακέτων. Επίσης, οι συγγραφείς της εργασίας [Esta04] προτείνουν την προσαρμοστική έκδοση του NetFlow (Adaptive NetFlow) που επιλύει πολλές από τις ανεπάρκειες του κλασσικού Cisco NetFlow [NetFl] με την εισαγωγή δυναμικής προσαρμογής του ποσοστού δειγματοληψίας στο μίγμα της κίνησης, για να επιτύχουμε την ευρωστία χωρίς να θυσιάσουμε την ακρίβεια στις μετρήσεις.

2.4. Σύνθετες μέθοδοι δειγματοληψίας

Άλλη μία πιο σύνθετη κατηγορία δειγματοληψίας είναι οι μέθοδοι βασιζόμενες σε κατακερματισμό (hash-based sampling) οι οποίες εφαρμόζονται κατανεμημένα σε όλο το υπό μελέτη δίκτυο. Μία γνωστή μέθοδος αυτού του είδους είναι η **«Δειγματοληψία Τροχιάς» (Trajectory sampling)** [Duff01] κατά την οποία όλοι οι δρομολογητές του υπό μελέτη δικτύου εφαρμόζουν στα πακέτα του δικτύου την ίδια συνάρτηση κατακερματισμού (hash function) και το ίδιο εύρος τιμών επιλογής. Το τμήμα του πακέτου στο οποίο εφαρμόζουμε την συνάρτηση κατακερματισμού περιορίζεται στα πεδία της επικεφαλίδας του πακέτου που παραμένουν αμετάβλητα καθώς κινούνται στο δίκτυο (π.χ. το πεδίο Time-To-Live της επικεφαλίδας IP δεν μπορεί να επιλεγεί αφού αλλάζει σε κάθε βήμα). Κατά συνέπεια ένα συγκεκριμένο πακέτο επιλέγεται είτε σε όλα τα σημεία κατά την πορεία του μέσα στο δίκτυο, είτε σε κανένα. Η εφαρμογή αυτής της μεθόδου μπορεί να βοηθήσει στην ανίχνευση λαθών δρομολόγησης στο δίκτυο, καθώς και στον εντοπισμό της πραγματικής διαδρομής που ακολούθησαν πακέτα επίθεσης με παραποιημένη διεύθυνση πηγής (source IP spoofing).

Μία άλλη εργασία [Seka07] που κάνει εφαρμογή της δειγματοληψίας κατακερματισμού και ονομάζεται **«Συντονισμένη δειγματοληψία» (Coordinated sampling)**, ακολουθεί το ακριβώς αντίθετο σκεπτικό από την προηγούμενη. Η δειγματοληψία πακέτων στους δρομολογητές γίνεται με τρόπο τέτοιο ώστε ο κάθε δρομολογητής να διαλέγει πακέτα που ανήκουν σε ροές που δεν θα επιλεγούν σε

άλλον δρομολογητή. Αυτό έχει ως αποτέλεσμα στο τέλος να έχουμε πληροφορία για όσο το δυνατόν περισσότερες ροές που υπάρχουν στο υπό μελέτη δίκτυο.

Ένα άλλο παράδειγμα σύνθετης δειγματοληπτικής μεθόδου είναι η **«Στρωματοποιημένη Τυχαία Δειγματοληψία» (Stratified Random Sampling)**. Σε αυτήν την προσέγγιση τα πακέτα ομαδοποιούνται σε υποσύνολα σύμφωνα με κάποιο δεδομένο χαρακτηριστικό. Κατόπιν, ο αριθμός των δειγμάτων επιλέγεται τυχαία από κάθε ομάδα. Η Στρωματοποιημένη Τυχαία Δειγματοληψία παρουσιάζει μικρότερη διακύμανση των στατιστικών από την απλή τυχαία δειγματοληψία εάν η διακύμανση μέσα στην ομάδα είναι μικρή σε σχέση με την διακύμανση μεταξύ των ομάδων [Zseb03]. Στην εργασία [Bosc06] αποδεικνύεται πως η μέθοδος αυτή υπερτερεί σε σχέση με την συστηματική και την τυχαία δειγματοληψία, όταν χρησιμοποιείται για την παρακολούθηση SLA (Service Level Agreements).

Τέλος, έχουν προταθεί και μέθοδοι δειγματοληψίας οι οποίες συνδυάζουν κάποιες από τις τεχνικές που περιγράψαμε παραπάνω. Μία από αυτές είναι η **«Δειγματοληψία σε Δύο Στάδια» (Two-Stage Sampling)** [Yang07] κατά την οποία το πρώτο στάδιο περιλαμβάνει δειγματοληψία ροών, ενώ στο δεύτερο στάδιο εκτελείται δειγματοληψία πακέτων. Πιο συγκεκριμένα, στο πρώτο στάδιο οι ροές επιλέγονται τυχαία με πιθανότητα p_f ανεξάρτητα από το μέγεθός τους, ενώ στο δεύτερο στάδιο, τα πακέτα επιλέγονται τυχαία με την πιθανότητα p_p από τις επιλεγείσες ροές του πρώτου σταδίου. Επειδή κάθε ροή έχει ίση πιθανότητα επιλογής, το επιλεγέν σύνολο διατηρεί το μέσο μήκος ροής (σε πακέτα) του αρχικού συνόλου ροών. Κατά συνέπεια, το τελικό ποσοστό των επιλεγμένων πακέτων από αυτό το σχήμα των δύο σταδίων δειγματοληψίας είναι ίσο με $p = p_f \cdot p_p$.

3. Ανίχνευση Εισβολών (Intrusion Detection)

3.1. Εισαγωγή στα Συστήματα Ανίχνευσης Εισβολών

Η Ανίχνευση Εισβολών (Intrusion Detection) [Mukh94] αποτελεί μια προσέγγιση για την παροχή μιας αίσθησης ασφάλειας στους υπάρχοντες υπολογιστές και δίκτυα, επιτρέποντας παράλληλα σε αυτά να λειτουργούν με μη περιορισμένο τρόπο. Ο στόχος της ανίχνευσης επιθέσεων είναι να προσδιοριστεί, κατά προτίμηση σε πραγματικό χρόνο, η κακή χρήση και η κατάχρηση των συστημάτων ηλεκτρονικών υπολογιστών τόσο από τα ίδια τα εσωτερικά μέλη των συστημάτων όσο και από εξωτερικούς χρήστες. Το πρόβλημα αυτό γίνεται μια πρόκληση καθώς η αυξανόμενη δικτύωση των ηλεκτρονικών υπολογιστών δίνει μεγαλύτερη πρόσβαση στους εξωτερικούς χρήστες και διευκολύνει τους εισβολείς να αποφεύγουν την αναγνώρισή τους. Τα Συστήματα Ανίχνευσης Εισβολών (Intrusion Detection Systems) βασίζονται στο γεγονός ότι η συμπεριφορά του εισβολέα θα είναι διαφορετική από αυτήν κάποιου κανονικού χρήστη και συνεπώς οι ανάρμοστες πράξεις μπορούν να γίνουν άμεσα ανιχνεύσιμες.

Η συμβατική προσέγγιση για την ασφάλεια ενός υπολογιστικού συστήματος ή δικτύου περιλαμβάνει τη δημιουργία μιας προστατευτικής ασπίδας γύρω από αυτό. Η ασπίδα αυτή θα πρέπει να αποτρέπει την ροή πληροφοριών από προστατευμένες περιοχές του δικτύου προς τον εξωτερικό κόσμο. Τεχνικές ελέγχου πρόσβασης μπορεί να χρησιμοποιηθούν για το σχεδιασμό τέτοιων ασφαλών συστημάτων. Ωστόσο υπάρχουν κάποιοι περιορισμοί στο σχεδιασμό ασφαλών υπολογιστικών συστημάτων και δικτύων. Ο βασικότερος από αυτούς αναφέρεται στο ότι είναι πολύ δύσκολο αν όχι ακατόρθωτο να σχεδιάσει κανείς ένα σύστημα που θα είναι εύχρηστο και συγχρόνως ασφαλές. Επίσης, δεν μπορεί κανείς να αποκλείσει από ένα θεωρητικά ασφαλές σύστημα κάποιο λάθος στην παραμετροποίηση από τον διαχειριστή το οποίο θα οδηγήσει σε πρόβλημα ασφάλειας.

Για το λόγο αυτό, στα μέσα της δεκαετίας του ‘80 μία εναλλακτική προσέγγιση που ονομαζόταν ανίχνευση εισβολής (intrusion detection) έκανε την εμφάνισή της. Η νέα αυτή προσέγγιση της ασφάλειας, δεν είχε σκοπό να αλλάξει την υπάρχουσα υποδομή των πιθανά ανασφαλών συστημάτων με καινούρια συστήματα που θα ήταν

ασφαλή, αλλά επιδιώκει να δράσει συμπληρωματικά προς αυτά. Τα συστήματα ανίχνευσης επιθέσεων θα ήταν βασισμένα σε μια τεχνολογία που θα επέτρεπε να ανιχνεύει επιθέσεις σε υπολογιστές και σε δίκτυα, κατά προτίμηση σε πραγματικό χρόνο, και να ειδοποιούν για αυτές το διαχειριστή ασφαλείας. Η προσέγγιση αυτή με το πέρασμα του χρόνου κέρδισε όλο και περισσότερο έδαφος στο χώρο της ασφάλειας με αποτέλεσμα ένας μεγάλος αριθμός από πρωτότυπα τέτοια συστήματα να έχουν δημιουργηθεί σήμερα σε πολλά ερευνητικά κέντρα και μερικά από αυτά να έχουν εγκατασταθεί σε παραγωγικά συστήματα.

3.2. Ταξινόμηση Συστημάτων Ανίχνευσης Εισβολών

Υπάρχουν δύο κύριες ταξινομήσεις των συστημάτων ανίχνευσης εισβολών. Η πρώτη διαιρεί τις τεχνικές ανίχνευσης επιθέσεων σε δύο κύριους τύπους: ανίχνευση ανωμαλίας (anomaly detection) και κακής χρήσης (misuse detection). Το πρότυπο ανίχνευσης ανωμαλίας χρησιμοποιεί ένα σύνολο στατιστικών στοιχείων που διαμορφώνουν τη συμπεριφορά μιας οντότητας. Οντότητα μπορεί να είναι ένας χρήστης, μια ομάδα χρηστών ή ένας υπολογιστής. Το προφίλ μιας οντότητας χρηστών, μπορεί να περιλάβει πληροφορίες όπως η μέση διάρκεια των συνόδων του Telnet και FTP, το ποσό των bytes που μεταδίδονται και προς τις δύο κατευθύνσεις, τις ώρες της ημέρας ή τα τερματικά από τα οποία συνδέεται ο χρήστης. Το προφίλ ενός υπολογιστή μπορεί να περιλαμβάνει τη μέση χρησιμοποίηση της CPU, το μέσο αριθμό συνδεδεμένων χρηστών, κ.α. Ένα IDS (Intrusion Detection System) ελέγχει τη λειτουργία ενός υπολογιστικού συστήματος και συγκρίνει συνεχώς το τρέχον προφίλ ενός συστήματος, με το προφίλ που είναι αποθηκευμένο στη βάση δεδομένων του. Σε περίπτωση που ανιχνεύσει μια μεγάλη απόκλιση από την κανονική συμπεριφορά στέλνει μία ειδοποίηση στο διαχειριστή ασφάλειας των υπολογιστικών συστημάτων. Το μέγεθος μιας μεγάλης απόκλισης ορίζεται ως ένα κατώτατο όριο που τίθεται από το IDS ή τον διαχειριστή ασφάλειας των συστημάτων. Συνήθως τα αποθηκευμένα προφίλ ενημερώνονται συνεχώς προκειμένου να απεικονιστούν οι αλλαγές στη συμπεριφορά των χρηστών ή του συστήματος. Δεδομένου ότι αυτό το πρότυπο λειτουργεί με βάση την ανίχνευση συνόδων που διαφέρουν σημαντικά από τις συνηθισμένες συνόδους ενός χρήστη, καλείται πρότυπο ανίχνευσης ανωμαλίας. Ανάμεσα στα βασικά πλεονεκτήματα της προσέγγισης αυτής, είναι η ικανότητά της

να ανιχνεύει καινούριες και άγνωστες ανωμαλίες που επηρεάζουν το προφίλ της κίνησης του χρήστη ή/και του δικτύου.

Το πρότυπο ανίχνευσης κακής χρήσης (misuse detection), λειτουργεί με βάση την ανίχνευση ενός συνόλου γνωστών επιθέσεων που έχουν αποθηκευτεί στη βάση δεδομένων του συστήματος. Η γνώση των επιθέσεων κωδικοποιείται ως ένα σύνολο από "υπογραφές επιθέσεων", οι οποίες είναι ουσιαστικά ακολουθίες από χαρακτήρες, που εμφανίζονται κάθε φορά που πραγματοποιείται μια επίθεση. Ο τρόπος που μια γνωστή επίθεση αντιπροσωπεύεται στο σύστημα είναι ένα σημαντικό χαρακτηριστικό της λειτουργίας του. Η εφαρμογή ενός τέτοιου IDS περιλαμβάνει συνήθως ένα έμπειρο σύστημα που εκτελεί τη σύγκριση με κανόνες αποθηκευμένους σε μια βάση δεδομένων. Μια προφανής δυσκολία σε αυτή την αρχιτεκτονική είναι η ανάγκη για τη σταθερή ενημέρωση της βάσης με καινούριες υπογραφές επιθέσεων, καθώς νέες μέθοδοι επιθέσεων γίνονται γνωστές καθημερινά. Δεδομένου ότι το πρότυπο αυτό λειτουργεί με την έρευνα για δείγματα που είναι αντιπροσωπευτικά διαφόρων επιθέσεων, αναφέρεται στη βιβλιογραφία ως πρότυπο ανίχνευσης κακής χρήσης.

Η δεύτερη ταξινόμηση είναι βασισμένη στο εάν το IDS ελέγχει τη δραστηριότητα σε ένα συγκεκριμένο υπολογιστή ή σε ένα δίκτυο υπολογιστών. Τα πρώτα συστήματα ανίχνευσης εισβολών συνήθιζαν να εξετάζουν στοιχεία σε ένα μεμονωμένο υπολογιστή και να παράγουν τα συμπεράσματά τους βασισμένα στα τοπικά αρχεία του συγκεκριμένου υπολογιστή. Συνεπώς, δεν θα μπορούσαν να ανιχνεύσουν τις επιθέσεις που στοχεύουν σε πολλούς υπολογιστές σε ένα δίκτυο. Επιπλέον, τα IDS αυτά στηρίζονται σε μεγάλο ποσοστό στα αρχεία καταγραφής (log files) που παρέχονται από το λειτουργικό σύστημα του υπολογιστή, το οποίο τα καθιστά αρχιτεκτονικά εξαρτώμενα και πιο ευάλωτα σε επιθέσεις DoS (Denial of Service) ενάντια σε αυτά, δεδομένου ότι ένας εισβολέας μπορεί να κατορθώσει να καθυστερήσει το μηχανισμό καταγραφής ή ακόμα και να τον σταματήσει τελείως. Τα συστήματα αυτά είναι γνωστά ως host-based IDS.

Μια πιο αποδοτική λύση για ανίχνευση των επιθέσεων παρέχεται από τα IDS που ελέγχουν παθητικά το δίκτυο, εξετάζοντας τα πακέτα που ρέουν σε αυτό, για ύποπτη δραστηριότητα. Δεδομένου ότι στηρίζονται στα πρωτόκολλα TCP/IP, είναι ανεξάρτητα από την αρχιτεκτονική του λειτουργικού συστήματος και μπορούν να ελέγξουν τα δίκτυα υπολογιστών σε μεγάλο βαθμό. Αν συνυπολογίσουμε και τη σύγχρονη τάση προς τη σύνδεση των υπολογιστών μέσω δικτύων, σχεδόν κάθε επίθεση περιλαμβάνει χρήση του δικτύου. Επομένως αυτή η κατηγορία των IDS που

είναι γνωστή ως network-based αποτελεί σήμερα μία αναγκαιότητα για την ασφάλεια του δικτύου. Στην παρούσα διατριβή θα εστιάσουμε το ενδιαφέρον μας στα συστήματα ανίχνευσης εισβολής που αφορούν ανίχνευση ανωμαλιών στο δίκτυο (network anomaly detection).

3.3. Ανίχνευση Ανωμαλιών Δικτύου

3.3.1. Κατηγορίες Ανωμαλιών

Οι ανωμαλίες δικτύων μπορούν να ταξινομηθούν σε δύο κατηγορίες. Η πρώτη κατηγορία σχετίζεται με τις αστοχίες και τα προβλήματα απόδοσης των δικτύων. Χαρακτηριστικά παραδείγματα τέτοιων ανωμαλιών είναι αποτυχίες εξυπηρετητών, βλάβες σε ζεύξεις δικτύων, συμφόρηση δικτύων, κ.τ.λ. [Thot01][Maxi90]. Παραδείγματος χάριν, οι αποτυχίες εξυπηρετητών, όπως μια αποτυχία ενός εξυπηρετητή ιστού (web server), θα μπορούσαν να εμφανιστούν όταν υπάρχει μια μεγάλη αύξηση στον αριθμό αιτημάτων προς τον συγκεκριμένο εξυπηρετητή. Μεγάλη χρησιμοποίηση πακέτων τύπου broadcast μπορεί να οδηγήσουν στο σημείο να θέσουν εκτός λειτουργίας το δίκτυο, δημιουργώντας συνθήκες συμφόρησης. Επίσης, συμφόρηση σε σύντομα χρονικά διαστήματα μπορεί να εμφανιστεί λόγω κάποιας αστοχίας συνδέσεων. Σε μερικές περιπτώσεις, διάφορα προβλήματα λογισμικού μπορούν επίσης να προκαλέσουν ανωμαλίες δικτύων, όπως ένα λάθος σε κάποια εφαρμογή υλοποίησης ενός πρωτοκόλλου μπορεί να προκαλέσει αυξανόμενη κίνηση δικτύου.

Η δεύτερη σημαντική κατηγορία ανωμαλιών δικτύων αφορά προβλήματα σχετικά με την ασφάλεια του δικτύου. Χαρακτηριστικότερη ανωμαλία αυτής της κατηγορίας αποτελούν οι επιθέσεις άρνησης υπηρεσίας (Denial of Service Attacks) [Doul04]. Οι επιθέσεις αυτές εμφανίζονται όταν κακόβουλες οντότητες έχουν ως στόχο να αποτραπεί η νόμιμη χρήση μιας υπηρεσίας του δικτύου. Ο επιτιθέμενος θα μπορούσε να θέσει εκτός λειτουργίας μια υπηρεσία ζωτικής σημασίας όπως είναι η υπηρεσία ονοματολογίας (Domain Name Service – DNS) [Mock87] και να προκαλέσει ένα εικονικό αποκλεισμό του δικτύου [Vign98][Yang00]. Σε αυτό το παράδειγμα, η ανωμαλία μπορεί να χαρακτηριστεί από πολύ μικρή κίνηση. Αντίθετα

σε άλλες περιπτώσεις επιθέσεων σε δίκτυα, η κακόβουλη οντότητα θα μπορούσε να σπαταλήσει το εύρος ζώνης (bandwidth) του δικτύου, πλημμυρίζοντας το δίκτυο με μία περιττή και ταυτόχρονα ανεπιθύμητη κίνηση, δημιουργώντας πρόβλημα στους νόμιμους χρήστες του δικτύου [Sava00]. Τα τελευταία χρόνια επίσης έχουν κάνει την εμφάνισή τους οι αποκαλούμενοι «αυτοδιαδιδόμενοι ιοί» (Worms). Με τον όρο «worm» [Weav03] ορίζουμε ένα κακόβουλο πρόγραμμα το οποίο αυτοδιαδίδεται μέσω δικτύου και προσπαθεί να μολύνει άλλους υπολογιστές εκμεταλλευόμενο μια συγκεκριμένη ευπάθεια στον υπολογιστή-θύμα. Κατά τη διάρκεια της φάσης διάδοσης, ο μολυσμένος υπολογιστής στέλνει πακέτα σε έναν μεγάλο πλήθος υπολογιστών, δημιουργώντας έτσι συχνά συνθήκες συμφόρησης στο δίκτυο.

3.3.2. Κατηγορίες Τεχνικών Ανίχνευσης Ανωμαλιών

Στην ενότητα αυτή περιγράφονται οι κυριότερες τεχνικές ανίχνευσης ανωμαλιών δικτύου. Αυτές μπορούν να ταξινομηθούν σε τρεις κατηγορίες: Στατιστικές Μέθοδοι (Statistical Anomaly Detection), Μέθοδοι βασισμένες στην εκμάθηση συστήματος (Machine Learning based Anomaly Detection) και Μέθοδοι Κατηγοριοποίησης (Classification-based Anomaly Detection).

- **Στατιστικές Μέθοδοι (Statistical Anomaly Detection)**

Στις στατιστικές μεθόδους για την ανίχνευση ανωμαλιών, το σύστημα παρατηρεί τη δραστηριότητα των διαφόρων οντοτήτων και παράγει τα κατάλληλα προφίλ για να περιγράψει τη συμπεριφορά τους. Τυπικά, διατηρούνται δύο προφίλ για κάθε οντότητα: το τρέχον προφίλ και το αποθηκευμένο προφίλ. Καθώς τα γεγονότα του δικτύου (δηλαδή αρχεία καταγραφής, εισερχόμενα πακέτα, κ.λπ.) υποβάλλονται σε επεξεργασία, το σύστημα ανίχνευσης εισβολής ενημερώνει το τρέχον προφίλ και υπολογίζεται περιοδικά ένας βαθμός ανωμαλίας (anomaly score), που δείχνει το βαθμό παρατυπίας για το συγκεκριμένο γεγονός, συγκρίνοντας το τρέχον προφίλ με το αποθηκευμένο. Εάν ο βαθμός ανωμαλίας είναι υψηλότερος από ένα ορισμένο κατώτατο όριο (threshold), το σύστημα ανίχνευσης εισβολών παράγει ένα προειδοποιητικό μήνυμα (alert).

Οι στατιστικές προσεγγίσεις στο πεδίο της ανίχνευσης ανωμαλιών έχουν αρκετά πλεονεκτήματα. Αρχικά, αυτά τα συστήματα, όπως τα περισσότερα συστήματα ανίχνευσης ανωμαλιών, δεν απαιτούν προγενέστερη γνώση των προβλημάτων ασφαλείας ή/και των επιθέσεων που έχουν προηγηθεί. Κατά συνέπεια, τέτοια συστήματα έχουν την ικανότητα της ανίχνευσης επιθέσεων τύπου «μηδενικής μέρας» (zero day) ή επιθέσεων που δεν είναι ευρέως γνωστές. Επιπλέον, οι στατιστικές προσεγγίσεις μπορούν να παρέχουν μία ακριβή προειδοποίηση των κακόβουλων δραστηριοτήτων όπου αυτές εμφανίζονται σε παρατεταμένες χρονικά περιόδους. Ένα πολύ κοινό παράδειγμα μιας τέτοιας δραστηριότητας είναι μια δραστηριότητα σάρωσης θυρών (portscan). Χαρακτηριστικά, η κατανομή των πακέτων από δραστηριότητα portscan είναι ιδιαίτερα ανώμαλη σε σύγκριση με τη συνηθισμένη κατανομή κίνησης. Έχοντας αυτό υπόψη, τα portscans ακόμα και όταν κατανέμονται σε ένα μεγάλο χρονικό διάστημα θα γίνουν αντιληπτά από τις στατιστικές μεθόδους επειδή θα προκαλέσουν ανώμαλη δραστηριότητα.

Εντούτοις, οι στατιστικές τεχνικές ανίχνευσης ανωμαλιών έχουν και κάποια μειονεκτήματα. Οι επιτιθέμενοι μπορούν να εκπαιδεύσουν ένα σύστημα ανίχνευσης ανωμαλιών που χρησιμοποιεί στατιστική ανάλυση να δεχτεί μία ανώμαλη συμπεριφορά σαν κανονική. Μπορεί επίσης να είναι δύσκολο να καθοριστούν τα κατώτατα όρια που ελαχιστοποιούν τις πιθανότητες ψευδών θετικών και ψευδών αρνητικών περιστατικών. Επιπλέον, οι στατιστικές μέθοδοι χρειάζονται τις ακριβείς στατιστικές κατανομές των διαφόρων μετρικών του δικτύου, αλλά αυτές πάντοτε δεν μπορούν να διαμορφωθούν κατάλληλα με βάση τις συμπεριφορές των διαφόρων οντοτήτων μέσα στο δίκτυο.

Μια γνωστή στατιστική μηχανή ανίχνευσης ανωμαλιών που αποκαλείται SPADE (Statistical Packet Anomaly Detection Engine) [Stan02] είναι ενσωματωμένη στο ευρέως γνωστό IDS SNORT [Roes99], και μπορεί να χρησιμοποιηθεί για την αυτόματη ανίχνευση δραστηριότητας portscan. Το SPADE ήταν μία από τις πρώτες τεχνικές που πρότειναν την έννοια του βαθμού ανωμαλίας (anomaly score) για να ανιχνεύσουν την σάρωση των θυρών σε ένα δίκτυο, αντί της χρησιμοποίησης της παραδοσιακής προσέγγισης της ανίχνευσης με βάση τις p προσπάθειες σε ένα χρονικό διάστημα q δευτερολέπτων. Στην εργασία [Stan02], οι συγγραφείς χρησιμοποίησαν έναν απλό τρόπο βασισμένο στη συχνότητα εμφάνισης ενός πακέτου, για να υπολογίσουν το βαθμό ανωμαλίας του. Όσο λιγότερο εμφανίζεται ένα πακέτο στο δίκτυο, τόσο υψηλότερος ήταν ο βαθμός ανωμαλίας του. Με άλλα

λόγια, οι συγγραφείς καθορίζουν το βαθμό ανωμαλίας σε ένα πακέτο σαν «μέτρο περιέργειας», βασισμένο στη δραστηριότητα του πρόσφατου παρελθόντος. Μόλις ξεπεράσει ο βαθμός ανωμαλίας ένα κατώτατο όριο, τα πακέτα διαβιβάζονται σε ένα μηχανισμό συσχέτισης που σχεδιάστηκε για να ανιχνεύει τις δραστηριότητες portscan. Εντούτοις, ένα σημαντικό μειονέκτημα για το SPADE είναι ότι έχει ένα πολύ υψηλό ποσοστό από ψευδή θετικά περιστατικά. Αυτό οφείλεται στο γεγονός ότι το SPADE ταξινομεί τα περίεργα πακέτα ως επιθέσεις, ανεξάρτητα από το εάν είναι πραγματικές επιθέσεις ή όχι.

Μία άλλη κατηγορία αλγορίθμων ανίχνευσης ανωμαλιών δικτύου που βασίζονται στην στατιστική ανάλυση αποτελούν οι αλγόριθμοι Ανίχνευσης Αλλαγής Σημείου (Change Point Detection) [Blaz01][Wang04][Peng04] οι οποίοι μπορούν και απομονώνουν την αλλαγή ενός στατιστικού στοιχείου του δικτύου που προκαλείται συνήθως από επιθέσεις. Αυτοί οι αλγόριθμοι αποθηκεύουν τα στοιχεία κίνησης του δικτύου ως μία χρονική ακολουθία. Εάν μια επίθεση αρχίσει στη χρονική στιγμή *t*, η χρονική ακολουθία θα παρουσιάσει κάποια στατιστική αλλαγή γύρω από το χρονική στιγμή *t* και μετέπειτα. Ένα παράδειγμα τέτοιου αλγορίθμου είναι ο αλγόριθμος CUSUM (Cummulative Sum) [Siri04]. Για να εντοπίσει μια επίθεση, ο CUSUM προσδιορίζει τις αποκλίσεις ανάμεσα στις πραγματικές τιμές και στις αναμενόμενες μέσες τιμές της χρονικής ακολουθίας. Εάν η διαφορά υπερβαίνει κάποιο ανώτατο όριο, οι επαναλαμβανόμενες αυξήσεις στα στατιστικά του CUSUM θα σηματοδοτήσουν την ανίχνευση μιας επίθεσης. Κατά τη διάρκεια των χρονικών διαστημάτων που περιέχουν μόνο κανονική κίνηση στο δίκτυο, η διαφορά είναι κάτω από αυτό το όριο. Μέσω του καθορισμού του ορίου, ο αλγόριθμος CUSUM μπορεί να επηρεαστεί σε σχέση με την καθυστέρηση της ανίχνευσης και τα λανθασμένα ποσοστά ανίχνευσης.

Άλλη μια γνωστή στατιστική μέθοδος ανίχνευσης ανωμαλιών είναι αυτή που βασίζεται στην έννοια της εντροπίας [Cove06] που χαρακτηρίζει τις κατανομές χαρακτηριστικών γνωρισμάτων της κίνησης του δικτύου. Η εντροπία μετρά την τυχαιότητα ενός συνόλου στοιχείων. Υψηλές τιμές εντροπίας δηλώνουν μια διασκορπισμένη κατανομή πιθανότητας των στοιχείων, ενώ χαμηλές τιμές εντροπίας δηλώνουν τη συγκέντρωση της κατανομής γύρω από συγκεκριμένα στοιχεία. Η εντροπία έχει χρησιμοποιηθεί εκτενώς στη βιβλιογραφία για την ανίχνευση αυτοδιαδιδόμενων ιών [Ranj07][Yu06][Lakh05]. Μερικές χρησιμοποιούμενες κατανομές χαρακτηριστικών γνωρισμάτων της κίνησης του δικτύου που είναι

πολύτιμες στην ανίχνευση ανωμαλιών δικτύων είναι οι κατανομές IP διευθύνσεων πηγής (source IP address), IP διευθύνσεων προορισμού (destination IP address), θύρας πηγής (source port) και θύρας προορισμού (destination port). Παραδείγματος χάριν, μια ανωμαλία που προέρχεται από ένα μολυσμένο υπολογιστή που προσπαθεί να μολύνει άλλους υπολογιστές στο Διαδίκτυο (αυτοδιαδιδόμεοι ιός) οδηγεί στη μείωση της εντροπίας των IP διευθύνσεων πηγής. Η μολυσμένη μηχανή παράγει έναν μεγάλο αριθμό ροών αναγκάζοντας την ίδια IP διεύθυνση πηγής να κυριαρχεί στη κατανομή ροών των IP διευθύνσεων πηγής.

- **Μέθοδοι βασισμένες στην Εκμάθηση Συστήματος (Machine Learning based Anomaly Detection)**

Η εκμάθηση συστήματος μπορεί να οριστεί ως η δυνατότητα ενός προγράμματος/συστήματος να μαθαίνει και να βελτιώνει την απόδοσή του σε ένα συγκεκριμένο στόχο ή μια ομάδα στόχων με την πάροδο του χρόνου. Η εκμάθηση συστήματος έχει ως στόχο να απαντήσει σε πολλές από τις ίδιες ερωτήσεις με αυτές των στατιστικών μεθόδων. Εντούτοις, αντίθετα από τις στατιστικές προσεγγίσεις που τείνουν να εστιάσουν στην κατανόηση της διαδικασίας που παρήγαγε τα δεδομένα, οι τεχνικές εκμάθησης συστημάτων εστιάζουν στην οικοδόμηση ενός συστήματος το οποίο βελτιώνει την απόδοσή του βασισμένο σε προηγούμενα χρονικά αποτελέσματα. Με άλλα λόγια, συστήματα που είναι βασισμένα στην εκμάθηση έχουν τη δυνατότητα να αλλάξουν τη στρατηγική εκτέλεσής τους βάσει των πρόσφατα ληφθέντων πληροφοριών.

Μία από τις ευρέως χρησιμοποιημένες τεχνικές εκμάθησης συστημάτων για την ανίχνευση ανωμαλιών περιλαμβάνει την εκμάθηση της συμπεριφοράς του συστήματος από ένα πρόγραμμα και την αναγνώριση των σημαντικών αποκλίσεων από την κανονική λειτουργία. Στην εργασία [Forr96] ο Forrest διαμόρφωσε μια αναλογία μεταξύ του ανθρώπινου ανοσοποιητικού συστήματος και της ανίχνευσης ανωμαλιών. Συγκεκριμένα προτάθηκε μια μεθοδολογία που περιελάμβανε την ανάλυση των ακολουθιών των κλήσεων του λειτουργικού συστήματος (system calls) για ένα πρόγραμμα προκειμένου να δημιουργηθεί ένα προφίλ για την κανονική λειτουργία του. Στην εργασία αυτή, οι συγγραφείς ανέλυσαν διάφορα UNIX προγράμματα, όπως το sendmail, lpr, κ.λπ., και έδειξαν ότι οι συσχετισμοί στις καθορισμένου μήκους ακολουθίες των κλήσεων του λειτουργικού συστήματος θα

μπορούσαν να χρησιμοποιηθούν για να δημιουργήσουν το κανονικό προφίλ ενός προγράμματος. Επομένως, προγράμματα που θα παρουσίαζαν ακολουθίες που παρέκκλιναν από την κανονική ακολουθία του προφίλ, θα μπορούσαν να θεωρηθούν θύματα μιας επίθεσης.

Μία άλλη κατηγορία μεθόδων ανίχνευσης ανωμαλιών που είναι βασισμένη στην Εκμάθηση Συστήματος αποτελούν οι μέθοδοι που είναι βασισμένες σε Μπεϋζιανά Δίκτυα. Ένα Μπεϋζιανό Δίκτυο (Bayesian Network) [Krue03] είναι ένα γραφικό μοντέλο το οποίο κωδικοποιεί τις πιθανολογικές σχέσεις μεταξύ των μεταβλητών ενός συστήματος. Όταν χρησιμοποιείται σε συνδυασμό με στατιστικές τεχνικές, τα Μπεϋζιανά Δίκτυα παρουσιάζουν διάφορα πλεονεκτήματα κατά την ανάλυση των στοιχείων τους. Αρχικά, επειδή τα Μπεϋζιανά Δίκτυα κωδικοποιούν τις αλληλεξαρτήσεις μεταξύ των μεταβλητών, μπορούν να ανταποκριθούν σε καταστάσεις στις οποίες λείπουν κάποια δεδομένα. Επιπρόσθετα, τα Μπεϋζιανά Δίκτυα έχουν τη δυνατότητα να αντιπροσωπεύσουν τις "σχέσεις αιτίας" (causal relationships). Επομένως, μπορούν να χρησιμοποιηθούν για να προβλέψουν τις συνέπειες μιας ενέργειας. Τέλος, επειδή τα Μπεϋζιανά Δίκτυα μοντελοποιούν και τις πιθανολογικές σχέσεις (probabilistic relationships) και τις "σχέσεις αιτίας", μπορούν να χρησιμοποιηθούν για να περιγράψουν προβλήματα όπου υπάρχει μια ανάγκη να συνδυαστεί η προγενέστερη γνώση με τα τρέχοντα δεδομένα.

Διάφοροι ερευνητές έχουν προσαρμόσει ιδέες από την θεωρία της Μπεϋζιανής στατιστικής δημιουργώντας κατάλληλα μοντέλα για την ανίχνευση ανωμαλιών στο δίκτυο [Krue03][Vald00][Ye00]. Συγκεκριμένα, ο Valdes [Vald00] ανέπτυξε ένα σύστημα ανίχνευσης ανωμαλιών βασισμένο σε ένα Μπεϋζιανό Δίκτυο για να εφαρμόσει την ανίχνευση εισβολών σε καταστάσεις που έχουμε απότομη αύξηση κίνησης. Το μοντέλο του, που είναι μέρος της πλατφόρμας EMERALD [Porr97], έχει την ικανότητα να ανιχνεύει τις κατανεμημένες επιθέσεις στις οποίες κάθε μεμονωμένη συνιστώσα επίθεσης δεν είναι αρκετά ύποπτη ώστε να παραγάγει κάποιο συναγερμό.

Τα δεδομένα για την ανίχνευση ανωμαλιών είναι συνήθως πολύ μεγάλου όγκου και ταυτόχρονα πολυδιάστατα. Με την αύξηση των δικτύων υψηλής ταχύτητας και τα κατανεμημένα δεδομένα από διαφορετικές πηγές, η αποθήκευσή τους, η επεξεργασία και η ανάλυση των στοιχείων γίνεται ολοένα και πιο σύνθετη. Για να αντιμετωπίσουν το πρόβλημα των συνόλων δεδομένων πολλών διαστάσεων, οι ερευνητές ανέπτυξαν μια τεχνική μείωσης της διαστατικότητας γνωστή ως Ανάλυση Κύριων Συνιστωσών

(Principal Component Analysis - PCA) [Calv98][Joll02][Wang06]. Με μαθηματικούς όρους, η μέθοδος PCA αποτελεί μία τεχνική όπου n συσχετιζόμενες τυχαίες μεταβλητές μετασχηματίζονται σε d μη συσχετιζόμενες μεταβλητές, όπου $d < n$, διατηρώντας όμως όσο το δυνατόν περισσότερο την υπάρχουσα διακύμανση στο σύνολο των στοιχείων. Οι μη συνδεδεμένες μεταβλητές είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και μπορούν να χρησιμοποιηθούν για να περιγράψουν τα δεδομένα σε μια πιο μειωμένη (από άποψη μεγέθους) μορφή.

Στις εργασίες [Lakh04a] και [Lakh04b] προτείνεται η χρήση της μεθόδου PCA συνδυάζοντας μετρήσεις ροών, πακέτων και χαρακτήρων (bytes) σε δικτυακά μονοπάτια που συνθέτουν το δίκτυο υπό μελέτη. Τα αποτελέσματα δείχνουν ότι κάθε διαφορετικός τύπος μετρήσεων (ροές, πακέτα, bytes) φέρνουν στην επιφάνεια και διαφορετικό τύπο ανωμαλιών δικτύου, όπως επιθέσεις άρνησης υπηρεσίας, portscans, διάδοση ιών και διακοπή λειτουργίας του δικτύου. Η μέθοδος αυτή εκτός από την ανίχνευση ανωμαλιών σε ενσύρματα δίκτυα ευρείας ζώνης [Chat08] έχει χρησιμοποιηθεί τόσο για ανίχνευση σε δίκτυα αισθητήρων (sensor networks) [Chat07a] όσο και σε δίκτυα αυτοκίνησης (vehicular networks) [Chat07b].

- **Μέθοδοι Κατηγοριοποίησης (Classification-based Anomaly Detection)**

Το σύστημα ανίχνευσης εισβολών που ταξινομεί τα δεδομένα ως φυσιολογικά ή ανώμαλα βασισμένο σε ένα σύνολο κανόνων, σε κάποιο πρότυπο ή σε άλλες παρόμοιες τεχνικές μπορεί να οριστεί ως ένα σύστημα ανίχνευσης εισβολών βασισμένο στην κατηγοριοποίηση. Η διαδικασία κατηγοριοποίησης περιλαμβάνει τα ακόλουθα βήματα:

1. Προσδιορισμός των ιδιοτήτων και των κατηγοριών από τα δεδομένα εκμάθησης (training data).
2. Προσδιορισμός των ιδιοτήτων για την κατηγοριοποίηση.
3. Κατασκευή ενός μοντέλου χρησιμοποιώντας τα δεδομένα εκμάθησης.
4. Χρησιμοποίηση του μοντέλου εκμάθησης για την κατηγοριοποίηση των άγνωστων στοιχείων.

Ποικίλες τεχνικές κατηγοριοποίησης έχουν προταθεί στη βιβλιογραφία. Αυτές περιλαμβάνουν κυρίως τεχνικές επαγωγικής παραγωγής κανόνων, ασαφούς λογικής (fuzzy logic) και τεχνικές βασισμένες σε νευρωνικά δίκτυα.

Οι Επαγωγικοί αλγόριθμοι παραγωγής κανόνων περιλαμβάνουν χαρακτηριστικά την εφαρμογή ενός συνόλου κανόνων συσχέτισης και συχνά εμφανιζόμενων προτύπων για να κατηγοριοποιηθούν τα δεδομένα. Σε αυτό το πλαίσιο, εάν ένας κανόνας δηλώνει ότι «εάν το γεγονός *X* εμφανίζεται, τότε το γεγονός *Y* είναι πιθανό να εμφανιστεί», τότε τα γεγονότα *X* και *Y* μπορούν να περιγραφούν σαν σύνολα ζευγών (μεταβλητή, τιμή), όπου στόχος είναι να βρεθούν τα σύνολα *X* και *Y* ώστε το *X* να συνεπάγεται το *Y*. Στην περιοχή της κατηγοριοποίησης, καθορίζουμε το σύνολο *Y* και προσπαθούμε να βρούμε τα σύνολα *X* που είναι καλοί προάγγελοι (predictors) για τη σωστή κατηγοριοποίηση. Ενώ η εποπτευμένη κατηγοριοποίηση παράγει μόνο κανόνες σχετικά με μια ιδιότητα, γενικές τεχνικές επαγωγικών κανόνων, που είναι συνήθως ανεπίβλεπτης (unsupervised) φύσης, παράγουν κανόνες σχετικά με όλες τις ιδιότητες. Το πλεονέκτημα της χρησιμοποίησης των κανόνων είναι ότι τείνουν να είναι απλοί, διαισθητικοί και μη δομημένοι. Αντίθετα, τα μειονεκτήματα περιλαμβάνουν δυσκολία στην διατήρηση και συντήρηση των κανόνων, ενώ σε μερικές περιπτώσεις, οι κανόνες αυτοί είναι ανεπαρκείς για να αντιπροσωπεύσουν πολλούς τύπους πληροφοριών.

Αρκετοί επαγωγικοί αλγόριθμοι παραγωγής κανόνων έχουν προταθεί στη βιβλιογραφία. Μερικοί απ' αυτούς αρχικά κατασκευάζουν ένα δέντρο απόφασης και έπειτα παράγουν ένα σύνολο κανόνων κατηγοριοποίησης από το δέντρο απόφασης. Άλλοι αλγόριθμοι (π.χ., RIPPER [Cohe95], C4.5 [Quin93]) παράγουν άμεσα κανόνες από τα δεδομένα με την υιοθέτηση μεθόδων «διαίρει και βασίλευε» (divide-and-conquer). Ένα δεύτερο στάδιο που περιλαμβάνει την απόρριψη (C4.5) ή την περικοπή (RIPPER) μερικών από τους μαθημένους κανόνες εφαρμόζεται προκειμένου να αυξήσει την ακρίβεια της κατηγοριοποίησης. Ο αλγόριθμος RIPPER έχει χρησιμοποιηθεί επιτυχώς σε διάφορες μεθόδους ανίχνευσης ανωμαλιών που ταξινομούν τα εισερχόμενα δικτυακά δεδομένα και ανιχνεύουν τις εισβολές. Ένα από τα αρχικά πλεονεκτήματα του RIPPER είναι ότι οι παραγόμενοι κανόνες είναι εύκολο να χρησιμοποιηθούν και να επιβεβαιωθούν. Ο Lee [Lee00] χρησιμοποίησε τον αλγόριθμο RIPPER για να χαρακτηρίσει ακολουθίες που εμφανίζονται στα φυσιολογικά δεδομένα, με βάση ένα μικρότερο σύνολο κανόνων που συλλαμβάνουν τα κοινά στοιχεία μέσα σε αυτές τις ακολουθίες. Κατά τη διάρκεια της

παρακολούθησης των δεδομένων, οι ακολουθίες που παραβιάζουν εκείνους τους κανόνες χαρακτηρίζονται ως ανωμαλίες.

Οι τεχνικές ασαφούς λογικής (fuzzy logic) χρησιμοποιούνται στον τομέα της ασφάλειας υπολογιστών και δικτύων από την δεκαετία του '90 [Hosm93]. Η ασαφής λογική έχει χρησιμοποιηθεί για την ανίχνευση εισβολών για δύο πρωταρχικούς λόγους. Αρχικά, διάφορες ποσοτικές παράμετροι που χρησιμοποιούνται στο πλαίσιο της ανίχνευσης εισβολών, π.χ., του χρόνου χρήσης της CPU, του χρονικού διαστήματος σύνδεσης, κ.λπ., μπορούν ενδεχομένως να αντιμετωπισθούν ως ασαφείς μεταβλητές. Αφετέρου, όπως παρουσιάζεται στην εργασία [Brid00], η έννοια της ασφάλειας, είναι από μόνη της ασαφής. Με άλλα λόγια, η έννοια της ασάφειας βοηθά στην ομαλοποίηση της απότομης διαφοροποίησης της κανονικής συμπεριφοράς από την ανώμαλη συμπεριφορά.

Τέλος, τα συστήματα ανίχνευσης εισβολών βασισμένα σε νευρωνικά δίκτυα ήταν παραδοσιακά βασισμένα σε host-based συστήματα τα οποία εστίαζαν στην ανίχνευση αποκλίσεων στη συμπεριφορά ενός προγράμματος σαν σημάδι ανωμαλίας [Ghos99][Rama03]. Το κύριο πλεονέκτημα των νευρικών δικτύων είναι η ανοχή τους σε ανακριβή στοιχεία και αβέβαιες πληροφορίες καθώς και η δυνατότητά τους να συνθέτουν λύσεις από τα στοιχεία χωρίς την κατοχή προγενέστερης γνώσης της συστηματικότητας στα στοιχεία. Το γεγονός αυτό σε συνδυασμό με τη δυνατότητά τους να γενικεύουν από τα στοιχεία εκμάθησης τα έχει καταστήσει μία κατάλληλη προσέγγιση στην ανίχνευση ανωμαλιών. Παρόλα αυτά, οι λύσεις βασισμένες σε νευρωνικά δίκτυα έχουν και διάφορα μειονεκτήματα. Αρχικά, μπορούν να αποτύχουν να βρουν μια ικανοποιητική λύση είτε λόγω έλλειψης ικανοποιητικών στοιχείων είτε επειδή δεν υπάρχει καμία λειτουργία εκμάθησης. Αφετέρου, η διαδικασία εκπαίδευσης των νευρωνικών δικτύων μπορεί να είναι αργή και υπολογιστικά ακριβή. Η έλλειψη ταχύτητας στην όλη διαδικασία οφείλεται εν μέρει στην ανάγκη να συλλεχθούν και να αναλυθούν τα δεδομένα εκπαίδευσης και εν μέρει επειδή το νευρωνικό δίκτυο πρέπει να χειριστεί τα βάρη των μεμονωμένων νευρώνων για να φθάσει στη σωστή λύση.

4. Επίδραση Δειγματοληψίας στην Ανίχνευση Ανωμαλιών Δικτύου

4.1. Εισαγωγή

Πρόσφατα, οι ερευνητές άρχισαν να εστιάζουν τις μελέτες τους στην επίδραση της δειγματοληψίας στην ανίχνευση ανωμαλιών δικτύου. Στην εργασία [Mai06a] οι συγγραφείς αξιολόγησαν την επίδραση της δειγματοληψίας πακέτων σε τρεις αλγορίθμους ανίχνευσης δραστηριότητας portscan. Τα αποτελέσματά τους έδειξαν ότι η δειγματοληψία πακέτων μειώνει την αποτελεσματικότητα ανίχνευσης αυτών των αλγορίθμων και αυξάνει εντυπωσιακά τα ψευδή θετικά περιστατικά (false positives). Αυτή η μελέτη διευρύνθηκε στην εργασία [Mai06b] στην οποία γίνεται σύγκριση της επίδρασης της τυχαίας δειγματοληψίας πακέτων (random packet sampling), της τυχαία δειγματοληψία ροών (random flow sampling), της έξυπνης δειγματοληψίας (smart sampling) [Duff03] και της δειγματοληψίας τύπου sample-and-hold [Esta02] σε συγκεκριμένες τεχνικές ανίχνευσης ανωμαλιών δικτύου. Τα αντίστοιχα αποτελέσματα έδειξαν ότι η τυχαία δειγματοληψία ροών συμπεριφέρεται καλύτερα ενώ η έξυπνη δειγματοληψία και η δειγματοληψία τύπου sample-and-hold δεν είναι κατάλληλη για την ανίχνευση ανωμαλιών.

Επιπλέον, οι συγγραφείς της εργασίας [Brau06] μελέτησαν την επίδραση της τυχαίας δειγματοληψίας πακέτων σε μια ανωμαλία αυτοδιαδιδόμενου ιού (συγκεκριμένα στο blaster worm), και έδειξαν ότι τα μετρικά που είναι βασισμένα στην εντροπία επηρεάζονται λιγότερο από τη δειγματοληψία από ότι τα μετρικά που σχετίζονται με το πλήθος των πακέτων, των ροών και των χαρακτήρων (bytes) που διακινούνται.

Στην εργασία [Kawa07] επίσης αποδεικνύεται ότι η δειγματοληψία πακέτων περιορίζει την ανίχνευση ανωμαλιών δικτύου. Για παράδειγμα, μία δραστηριότητα portscan ή μια DoS επίθεση με SYN πακέτα παράγει μία πληθώρα από ροές που έχουν ένα μόνο πακέτο. Τέτοιες ροές έχουν λιγότερη πιθανότητα να επιλεγούν σε σχέση με τις κανονικές ροές. Σαν λύση σε αυτό το πρόβλημα, στην εργασία προτείνεται να αυξήσουμε την ανίχνευση τέτοιων ανωμαλιών χωρίζοντας το χώρο

της ελεγχόμενη κίνησης σε ομάδες (π.χ. με βάση την IP διεύθυνση πηγής) και να αναλύσουμε την κίνηση των μεμονωμένων αυτών ομάδων ξεχωριστά.

Στη συνέχεια αυτού του κεφαλαίου εξετάζουμε την επίδραση της δειγματοληψίας πακέτων σε διαφορετικές μεθοδολογίες ανίχνευσης ανωμαλιών. Ένας από τους βασικούς στόχους της μελέτης μας είναι να αποκτηθεί γνώση για τη δυνατότητα πραγματοποίησης αποτελεσματικής ανίχνευσης ανωμαλιών δικτύου, με την ανάλυση και κατανόηση της ισορροπίας ανάμεσα στη μείωση του όγκου των συλλεχθέντων δεδομένων και την διατήρηση της ακρίβειας και της αποτελεσματικότητας στην ανίχνευση ανωμαλιών.

4.2. Επίδραση κυριοτέρων τεχνικών δειγματοληψίας πακέτων σε ανίχνευση ανωμαλιών

Σε αυτή την ενότητα, δίνεται έμφαση στην αξιολόγηση της επίδρασης των κυριοτέρων τεχνικών δειγματοληψίας πακέτων που έχουν προταθεί στο PSAMP IETF draft [Psamp], σε δύο ευρέως χρησιμοποιημένες μεθοδολογίες ανίχνευσης ανωμαλιών στο δίκτυο. Συγκεκριμένα, αξιολογούμε τη συμπεριφορά μιας διαδοχικής μη παραμετρικής (non-parametric) μεθόδου Ανίχνευσης Αλλαγής Σημείου (Change Point Detection – CPD) και ενός αλγορίθμου βασισμένου στην μέθοδο της Ανάλυσης Κύριων Συνιστωσών (Principal Components Analysis - PCA) με τη χρήση διαφορετικών μετρικών, κάτω από διαφορετικές μεθοδολογίες δειγματοληψίας.

Πρέπει να σημειωθεί εδώ ότι το συγκεκριμένο πρόβλημα της επίδρασης της δειγματοληψίας στη διαδικασία ανίχνευσης ανωμαλιών δικτύου είναι αρκετά διαφορετικό και πιο περίπλοκο από τα αντίστοιχα προβλήματα επίδρασης της δειγματοληψίας σε άλλες διαδικασίες διαχείρισης του δικτύου. Όπως προαναφέρθηκε αυτό οφείλεται κυρίως στο γεγονός ότι η ανίχνευση ανωμαλιών είναι μία διαδικασία που εφαρμόζεται κάτω από ανώμαλες συνθήκες δικτύου (π.χ. επιθέσεις), ενώ από τη φύση της περιλαμβάνει ταυτόχρονα διάφορους παράγοντες, όπως η κανονική κίνηση του δικτύου, η ανώμαλη κίνηση, τα διάφορα μετρικά ανίχνευσης, των οποίων οι στατιστικές ιδιότητες και η συμπεριφορά μπορεί να επηρεαστεί σε μεγάλο βαθμό και με αρκετά διαφορετικούς τρόπους με την εφαρμογή της δειγματοληψίας.

Επομένως, όπως θα δούμε και στη συνέχεια από τα πειραματικά αποτελέσματα, οι παρατηρήσεις και οι κοινές πρακτικές που έχουν χρησιμοποιηθεί μέχρι τώρα σχετικά με την εφαρμογή των τεχνικών δειγματοληψίας πακέτων, δεν είναι

απαραιτήτως κατάλληλες επιλογές για την αποδοτική και αποτελεσματική λειτουργία των τεχνικών ανίχνευσης ανωμαλιών, δεδομένου ότι έχουν σημαντικές επιπτώσεις στην ικανότητα ανίχνευσής τους. Ως εκ τούτου, η κατανόηση και η παροχή μιας ποιοτικής και ποσοτικής αξιολόγησης της επίδρασης των διάφορων τεχνικών δειγματοληψίας στη διαδικασία ανίχνευσης ανωμαλιών είναι υψηλής ερευνητικής και πρακτικής σπουδαιότητας, και αποτελεί ένα σημαντικό μέρος της μελέτης μας, καθώς επίσης παρέχει και τις κύριες κατευθύνσεις για την ανάπτυξη έξυπνων και αποτελεσματικών καινοτόμων μεθόδων δειγματοληψίας.

4.2.1. Μέθοδος Ανίχνευσης Αλλαγής Σημείου (Change Point Detection)

Σε αυτή την ενότητα παρουσιάζουμε μια διαδοχική μη παραμετρική μέθοδο Ανίχνευσης Αλλαγής Σημείου (Change Point Detection – CPD) που αντιπροσωπεύει μια ευρεία κατηγορία χρησιμοποιούμενων στρατηγικών ανίχνευσης ανωμαλιών. Αυτή η μέθοδος είναι ανεξάρτητη από την τοπολογία και τα χαρακτηριστικά κίνησης του δικτύου και μπορεί να εφαρμοστεί για να ελέγξει οποιοδήποτε δίκτυο. Ο στόχος της μεθόδου CPD είναι να καθοριστεί εάν η παρατηρούμενη χρονική ακολουθία είναι στατιστικά ομοιογενής και, εάν όχι, να βρεθεί το σημείο εκείνο στο οποίο συνέβη η αλλαγή [Wang04][Peng04].

Ο αλγόριθμος ανίχνευσης ανωμαλίας που περιγράφεται παρακάτω ανήκει στη κατηγορία ανίχνευσης αλλαγής σημείου στην οποία οι δοκιμές γίνονται on-line με τα στοιχεία που εμφανίζονται διαδοχικά και οι αποφάσεις λαμβάνονται σε πραγματικό χρόνο. Στη δική μας περίπτωση ο μη παραμετρικός αλγόριθμος CUSUM (Cumulative Sum) εφαρμόζεται για την ανίχνευση των επιθέσεων. Η κύρια ιδέα του αλγορίθμου CUSUM είναι ότι η μέση τιμή μιας τυχαίας ακολουθίας $\{X_n\}$ είναι αρνητική κατά τη διάρκεια της κανονικής λειτουργίας του δικτύου και γίνεται θετική όταν εμφανίζεται μια αλλαγή. Κατά συνέπεια, μπορούμε να θεωρήσουμε την $\{X_n\}$ ως στάσιμη τυχαία διαδικασία η οποία κάτω από φυσιολογικές συνθήκες, έχει μέσο όρο X_n ίσο με k ($E(X_n)=k$). Μια παράμετρος α επιλέγεται έτσι ώστε να αποτελεί ένα άνω όριο της τιμής του k (δηλαδή $\alpha>k$) και μια άλλη τυχαία διαδικασία $\{Z_n\}$ ορίζεται έτσι ώστε $Z_n = X_n - \alpha$, η οποία έχει έναν αρνητικό μέσο όρο κατά τη διάρκεια της κανονικής

λειτουργίας του δικτύου. Ο σκοπός της παραμέτρου α είναι να αντισταθμίσει το πιθανό θετικό μέσο όρο της τυχαίας διαδικασίας $\{X_n\}$ που μπορεί να προκληθεί από μικρές ανωμαλίες στο δίκτυο, έτσι ώστε η στατιστική δοκιμή y_n, που θα περιγραφεί παρακάτω, να μηδενίζεται συχνά και να μην συσσωρεύεται με το χρόνο.

Όταν πραγματοποιείται μια επίθεση, η τιμή του Z_n θα αυξηθεί ξαφνικά και θα γίνει ένας μεγάλος θετικός αριθμός. Αν υποθέσουμε ότι κατά τη διάρκεια μιας επίθεσης η αύξηση στο μέσο όρο του Z_n μπορεί να έχει ως κάτω όριο την τιμή h τότε η ανίχνευση αλλαγής είναι βασισμένη στην παρατήρηση του $h>>k$.

Ειδικότερα, έχουμε: $y_n = (y_{n-1} + Z_n)^+$,

$$y_0 = 0,$$

όπου x^+ είναι ίσο με x εάν $x > 0$ και 0 σε κάθε άλλη περίπτωση. Η συνάρτηση απόφασης περιγράφεται παρακάτω ως:

$$d_N(y_n) = 0, \text{ αν } y_n \leq N,$$

$$d_N(y_n) = 1, \text{ αν } y_n > N$$

όπου $d_N(y_n)$ είναι η απόφαση τη χρονική στιγμή n: το ‘0’ υποδηλώνει κανονική λειτουργία και το ‘1’ υποδηλώνει κάποια ανωμαλία, ενώ το N καθορίζει το κατώφλι για την ανίχνευση της ανωμαλίας. Αυτή η μέθοδος ανίχνευσης ανωμαλιών έχει χρησιμοποιηθεί με διάφορα μετρικά, όπως η αναλογία πακέτων TCP SYN/FIN [Wang04] ή το ποσοστό νέων IP διευθύνσεων πηγής που παρατηρούνται σε ένα χρονικό παράθυρο [Peng04], για την ανίχνευση επιθέσεων άρνησης υπηρεσίας (Denial of Service attacks).

4.2.2. Μέθοδος Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis - PCA)

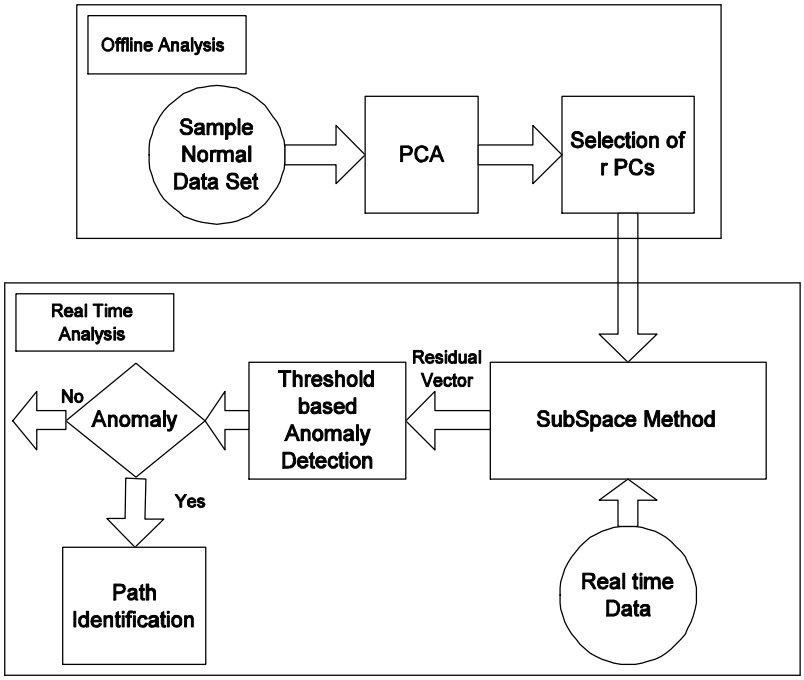
Ο στόχος της μεθόδου της Ανάλυσης Κύριων Συνιστωσών (PCA) [Chat06][Chat08] είναι να εφαρμοστεί μια μεθοδολογία σύνθεσης και συνδυασμού δεδομένων ετερογενών οργάνων μέτρησης του δικτύου, προκειμένου να παρασχεθεί ένα γενικευμένο πλαίσιο, ικανό να ανιχνεύσει ένα ευρύ φάσμα ανωμαλιών, όπως αυτές που μπορούν να οδηγήσουν σε αλλαγές στη σύνθεση της κίνησης του δικτύου

ή τις κατευθύνσεις της κίνησης μέσα στο δίκτυο. Αυτό επιτυγχάνεται με την εφαρμογή μιας βασισμένης στη μέθοδο PCA προσέγγισης σε διάφορα μετρικά ταυτόχρονα μιας ή περισσότερων συνδέσεων (links). Γενικά, για κάθε σύνδεση δικτύων υπάρχουν διάφορα μετρικά που περιγράφουν την κίνηση που περνά μέσω αυτών των συνδέσεων. Προκειμένου να διαμορφωθεί καλύτερα και να αντιπροσωπευθεί αυτό, δημιουργείται ένα σύνολο εικονικών συνδέσεων (virtual links) για κάθε πραγματική σύνδεση, με κάθε εικονική σύνδεση να αντιστοιχεί σε ένα διαφορετικό μετρικό.

Η βασική ιδέα πίσω από την Ανάλυση Κύριων Συνιστωσών είναι ο εντοπισμός των γραμμικών συνδυασμών των αρχικών μεταβλητών οι οποίοι είναι γραμμικά ανεξάρτητοι μεταξύ τους. Ο στόχος είναι να μειωθεί ο αριθμός των αρχικών μεταβλητών, διατηρώντας τόσες ώστε το σύστημα να εμπεριέχει όσο το δυνατόν μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών. Οι μεταβλητές αυτές ονομάζονται Κύριες Συνιστώσες (ΚΣ) (Principal Components - PC) και υπολογίζονται από τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης ή του πίνακα συσχετισμού των αρχικών μεταβλητών.

Στη συνέχεια περιγράφεται συνοπτικά και σχηματικά η προτεινόμενη μεθοδολογία [Chat08]. Η συνολική διαδικασία μπορεί να χωριστεί σε δυο βασικά μέρη, όπως φαίνεται στο Σχήμα 4.1: α) τη διαδικασία εκμάθησης (offline analysis), όπου δημιουργείται το μοντέλο της κίνησης από δεδομένα που θεωρούνται κανονικά (δεν περιέχουν ανωμαλίες), και β) τη διαδικασία ανίχνευσης που γίνεται σε πραγματικό χρόνο (real time analysis) και η οποία συγκρίνει την τρέχουσα κίνηση με τη μοντελοποιημένη με στόχο την ανίχνευση ανωμαλιών.

Κατά τη διαδικασία εκμάθησης, η μέθοδος εφαρμόζεται σε ένα δείγμα της κίνησης που θεωρείται ότι δεν περιέχει ανωμαλίες με σκοπό να εξαχθούν οι Κύριες Συνιστώσες (ΚΣ) που επαρκούν για τη περιγραφή των σημαντικών συσχετίσεων στα δεδομένα. Ο αριθμός τους εξαρτάται από το δίκτυο και τον αριθμό εικονικών συνδέσεων, και αντιπροσωπεύει τον αριθμό των ΚΣ που απαιτείται για τη σύλληψη του ποσοστού της διακύμανσης που χρειάζεται το σύστημα για να μοντελοποιήσει την κανονική κίνηση του δικτύου. Οι επιλεγείσες ΚΣ (αποτέλεσμα της διαδικασίας εκμάθησης) θα χρησιμοποιηθούν στη Μέθοδο Υποχώρων (Subspace Method).



Σχήμα 4.1. Παρουσίαση της μεθοδολογίας PCA

Ο στόχος της Μεθόδου Υποχώρων είναι να διαχωριστούν τα τρέχοντα δεδομένα κίνησης σε δύο διαφορετικά τμήματα: ένα που να περιέχει τα δεδομένα της κίνησης που θεωρείται κανονική ($\mathbf{y}_{\mathbf{norm}}$) και μοιάζει με την μοντελοποιημένη κίνηση και ένα τμήμα το οποίο περιγράφει το υπόλοιπο τμήμα της κίνησης ($\mathbf{y}_{\mathbf{res}}$). Γενικά, οι ανωμαλίες τείνουν να οδηγήσουν σε μεγάλες διακυμάνσεις στο υπόλοιπο τμήμα $\mathbf{y}_{\mathbf{res}}$, δεδομένου ότι παρουσιάζουν διαφορετικά χαρακτηριστικά από αυτά της μοντελοποιημένης κίνησης.

Κατά τη διάρκεια της διαδικασίας ανίχνευσης που γίνεται σε πραγματικό χρόνο (real time analysis), το τρέχον διάνυσμα της κίνησης προβάλλεται σε δύο διαφορετικούς υποχώρους, με τη χρήση των ΚΣ που υπολογίστηκαν στη διαδικασία εκμάθησης. Όταν εμφανιστεί μια ανωμαλία, το διάνυσμα $\mathbf{y}_{\mathbf{res}}$ παρουσιάζει μία μεγάλη διακύμανση σε μερικές από τις μεταβλητές του. Ένα συνηθισμένο στατιστικό μέτρο που χρησιμοποιείται για την ανίχνευση ανωμαλίας είναι το τετραγωνικό σφάλμα πρόβλεψης – Squared Prediction Error (SPE) [Choi05] – το οποίο ορίζεται ως εξής:

$$SPE \equiv \left\| y_{res} \right\|^2$$

Αν λοιπόν το SPE ξεπεράσει κάποιο προκαθορισμένο κατώφλι τότε μπορούμε να θεωρήσουμε ότι έχουμε ανωμαλία στο δίκτυο. Στην περίπτωση που οι μεταβλητές που έχουμε χρησιμοποιήσει αναφέρονται σε πολλαπλές φυσικές ζεύξεις μπορούμε να

εντοπίσουμε και το μονοπάτι της ανωμαλίας, επιλέγοντας τις συνιστώσες του $\mathbf{y}_{res}$ που συνεισφέρουν στην αύξηση της νόρμας $\| \mathbf{y}_{res} \|$.

4.2.3. Αξιολόγηση Κυριοτέρων Δειγματοληπτικών Μεθόδων Πακέτων

Στην ενότητα αυτή θα πραγματοποιηθεί αξιολόγηση στην επίδραση τριών βασικών τεχνικών δειγματοληψίας πακέτων που έχουμε περιγράψει σε προηγούμενο κεφάλαιο πάνω στις δύο μεθόδους ανίχνευσης ανωμαλιών δικτύου που περιγράψαμε προηγουμένως (μέθοδος Ανίχνευσης Αλλαγής Σημείου (CPD) και Ανάλυσης Κύριων Συνιστωσών (PCA)). Συγκεκριμένα θα μελετήσουμε τις εξής δειγματοληπτικές μεθόδους:

- Συστηματική δειγματοληψία (Systematic Sampling),
- Τυχαία ν-από-Ν δειγματοληψία (n-out-of-N Random Sampling)
- Ομοιόμορφη πιθανολογική τυχαία δειγματοληψία (Uniform Probabilistic Random Sampling)

οι οποίες έχουν προταθεί στο PSAMP IETF draft [Psamp].

Τα αποτελέσματα και οι αντίστοιχες παρατηρήσεις που παρουσιάζονται σε αυτή την ενότητα είναι βασισμένα σε πραγματικά δεδομένα δικτύου που έχουν συλλεχθεί από ένα ακαδημαϊκό δίκτυο. Συγκεκριμένα μελετήσαμε τη σύνδεση μεταξύ του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ) και του Εθνικού Δικτύου Έρευνας και Τεχνολογίας (ΕΔΕΤ) που συνδέει το ΕΜΠ με το Διαδίκτυο. Στο διάστημα των πειραμάτων μας, αυτή η σύνδεση είχε μια μέση κίνηση της τάξης των 70-80Mbit/sec και περίπου 20000 πακέτα/sec, και περιείχε ένα πλούσιο μίγμα κίνησης αποτελούμενο από κίνηση web, ηλεκτρονικού ταχυδρομείου, FTP καθώς επίσης και p2p κίνηση.

Στην ακόλουθη αξιολόγηση, μελετάται λεπτομερώς μια κατανεμημένη επίθεση άρνησης υπηρεσίας (Distributed Denial of Service Attack – DDoS) και συγκεκριμένα επίθεση TCP SYN πακέτων ενάντια σε έναν υπολογιστή μέσα στο ΕΜΠ (NTUA). Ο Πίνακας 4.1 παρουσιάζει τα διάφορα ποσοστά επίθεσης (σε σχέση με την κανονική κίνηση) που χρησιμοποιούνται σε αυτήν την μελέτη.

Πίνακας 4.1. Ποσοστά επίθεσης DDoS σε packets/sec και bytes/sec

Attack Ratio in packets/sec and bytes/sec	
4000 packets/sec (20%)	1.72Mbit/sec (2.5%)
2000 packets/sec (10%)	864kbit/sec (1.25%)
1000 packets/sec (5%)	432kbit/sec (0.6%)
400 packets/sec (2%)	172kbit/sec (0.25%)
200 packets/sec (1%)	86.4kbit/sec (0.12%)

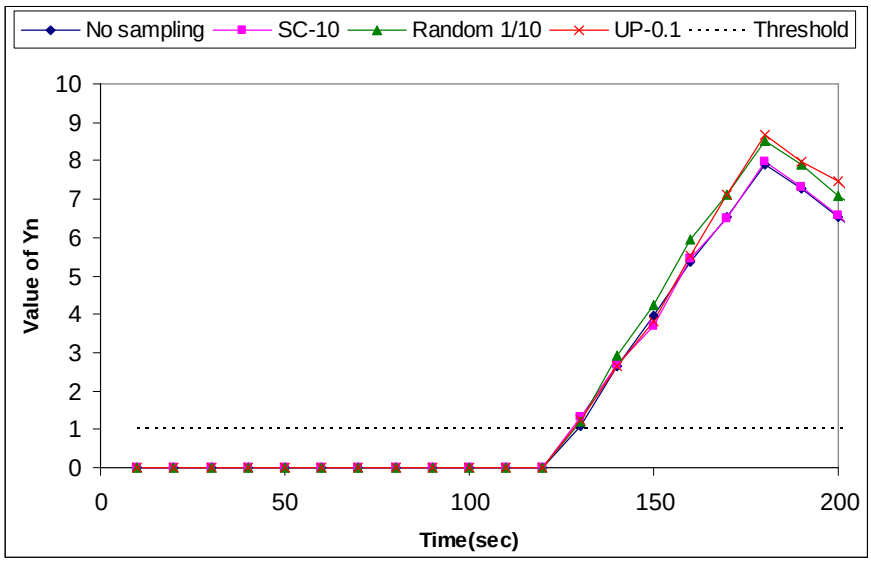
Για να καταδείξουμε καλύτερα τα αποτελέσματα και τις αντίστοιχες παρατηρήσεις, μελετάμε δύο σενάρια για τη μέθοδο Ανίχνευσης Αλλαγής Σημείου (CPD) και ένα σενάριο για τη μέθοδο της Ανάλυσης Κύριων Συνιστωσών (PCA). Σε όλα τα σενάρια εφαρμόζουμε τους προαναφερθέντες τύπους δειγματοληψίας πακέτων που περιγράψαμε προηγουμένως σε διαφορετικά ποσοστά δειγματοληψίας. Στα πειράματά μας, χρησιμοποιήσαμε ποσοστά δειγματοληψίας πακέτων 1/10, 1/50 και 1/100 για τη συστηματική και τυχαία ν-από-Ν δειγματοληψία, και πιθανότητες ίσες με 0.1, 0.02 και 0.01 αντίστοιχα, για την ομοιόμορφη πιθανολογική τυχαία δειγματοληψία.

4.2.3.1 Μέθοδος CPD με μετρικό την αναλογία SYN/FIN πακέτων

Η μέθοδος Ανίχνευσης Αλλαγής Σημείου (CPD) σε αυτή την προσέγγιση χρησιμοποιεί σαν μετρικό (μεταβλητή X_n στον αλγόριθμο) την διαφορά των πακέτων SYN και FIN διαιρούμενη με τον αριθμό των FIN πακέτων [Wang04]. Τα πακέτα SYN είναι αυτά που έχουν ενεργή τη σημαία (flag) TCP SYN που υποδηλώνει την έναρξη μιας TCP σύνδεσης, ενώ ο αριθμός πακέτων FIN που χρησιμοποιούμε εδώ είναι στην πραγματικότητα η ποσότητα των πακέτων που έχουν ενεργή την σημαία FIN (ομαλός τερματισμός TCP σύνδεσης) ή RST (μή ομαλός τερματισμός TCP σύνδεσης). Υπό κανονικές συνθήκες η μεταβλητή X_n έχει μια θετική μέση τιμή κοντά στο μηδέν ενώ η μεταβλητή Z_n έχει αρνητική τιμή (θυμίζουμε ότι $Z_n = X_n - \alpha$). Όταν πραγματοποιείται μια επίθεση με SYN πακέτα, η τιμή της μεταβλητής X_n γίνεται ένας μεγάλος θετικός αριθμός που έχει ως συνέπεια η τιμή Z_n να γίνει θετική. Αυτό οδηγεί στην αύξηση της τιμής της μεταβλητής y_n που υποδηλώνει την παρουσία κάποιας επίθεσης εάν η τιμή της υπερβαίνει ένα ορισμένο όριο (threshold). Δεδομένου ότι οι τρεις πρώτες αναλογίες επίθεσης είναι πολύ μεγάλες όσον αφορά την αναλογία

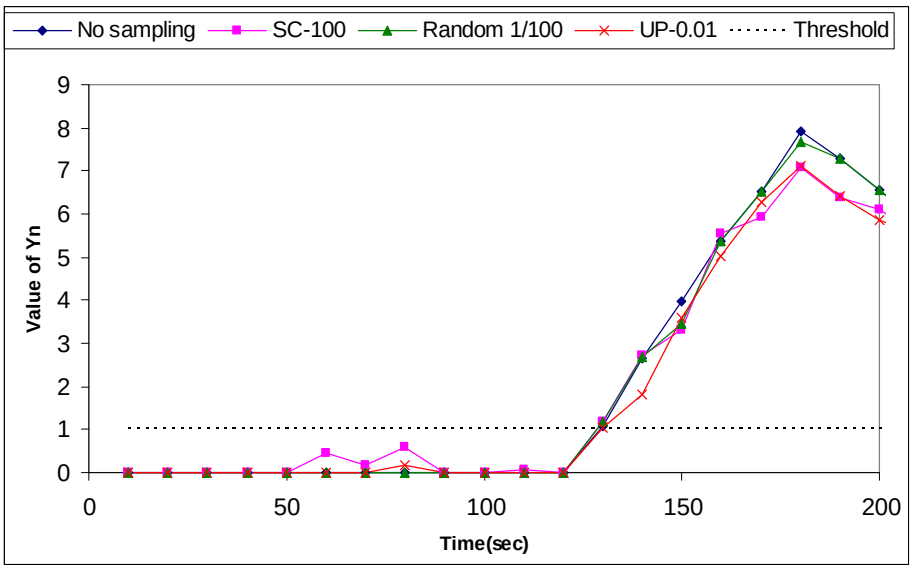
SYN/FIN παρουσιάζουμε εδώ μόνο τα αποτελέσματα για τις τελευταίες δύο αναλογίες επίθεσης του Πίνακα 4.1 (αναλογία πακέτων 2% και 1%). Στα παρακάτω γραφήματα η επίθεση SYN εμφανίζεται κατά τη διάρκεια του διαστήματος 130-180sec.

Τα αποτελέσματα που απεικονίζονται στο Σχήμα 4.2 αντιστοιχούν σε ποσοστό επίθεσης 2% (σε πακέτα) και σε ποσοστό δειγματοληψίας 1/10 (σε πακέτα). Όπως μπορούμε να παρατηρήσουμε και οι τρεις τύποι δειγματοληψίας (Συστηματική δειγματοληψία – SC-10, Τυχαία ν-από-Ν δειγματοληψία – Random 1/10, Ομοιόμορφη πιθανολογική τυχαία δειγματοληψία – UP-0.1) παρουσιάζουν παρόμοια συμπεριφορά και οι καμπύλες τους μοιάζουν με την αντίστοιχη καμπύλη για την περίπτωση που δεν εφαρμόζεται καθόλου δειγματοληψία. Παρόμοια αποτελέσματα εμφανίζονται για το ποσοστό δειγματοληψίας 1/50.



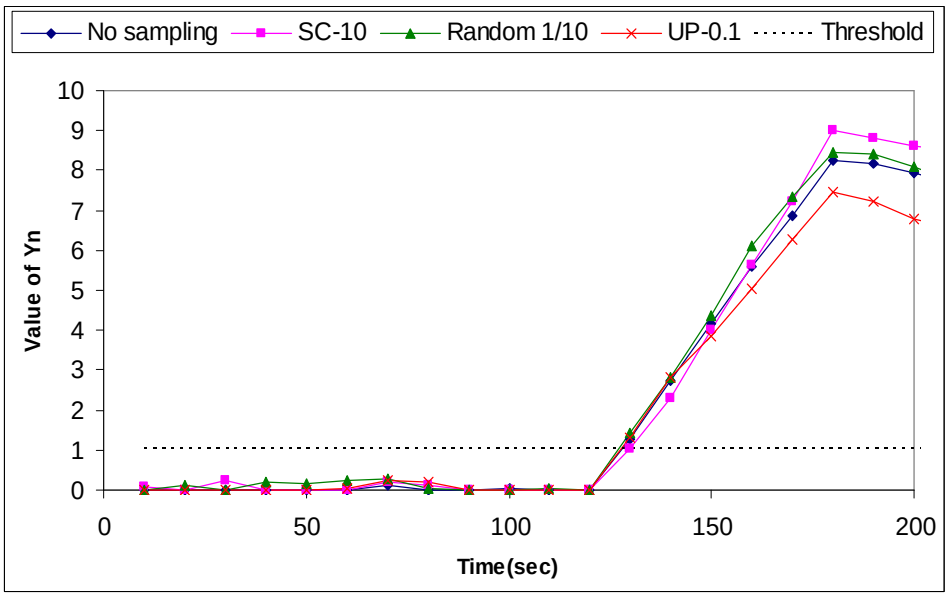
Σχήμα 4.2. Ποσοστό επίθεσης 2% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/10 (σε πακέτα)

Το Σχήμα 4.3 παρουσιάζει τα αντίστοιχα αποτελέσματα για ποσοστό επίθεσης 2% (σε πακέτα) σε ποσοστό δειγματοληψίας 1/100. Εδώ παρατηρούμε ότι η συστηματική δειγματοληψία δίνει μερικά ψευδή θετικά περιστατικά, τα οποία εντούτοις δεν υπερβαίνουν το κατώφλι επίθεσης που έχουμε ορίσει.

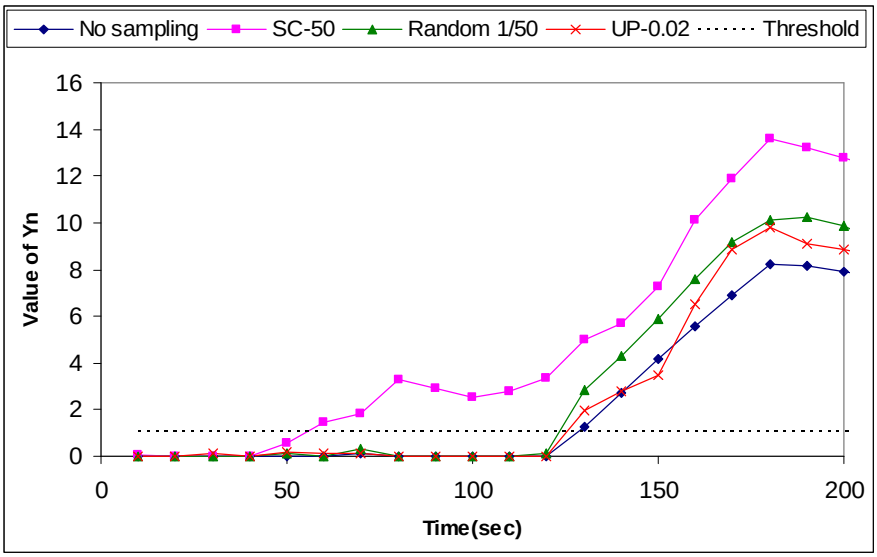


Σχήμα 4.3. Ποσοστό επίθεσης 2% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/100 (σε πακέτα)

Όταν μειώσουμε το ποσοστό επίθεσης στο 1% (σε πακέτα) και σε ποσοστό δειγματοληψίας 1/10 (σε πακέτα) δεν παρατηρούμε ψευδή θετικά περιστατικά, όπως φαίνεται στο Σχήμα 4.4, εντούτοις εάν αλλάξουμε το ποσοστό δειγματοληψίας σε 1/50, η καμπύλη της συστηματικής δειγματοληψίας παρεκκλίνει από την αρχική περίπτωση (Σχήμα 4.5), ενώ οι άλλοι δύο τύποι δειγματοληπτικών μεθόδων έχουν σχεδόν την ίδια συμπεριφορά όπως στην περίπτωση που δεν εφαρμόζεται δειγματοληψία.



Σχήμα 4.4. Ποσοστό επίθεσης 1% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/10 (σε πακέτα)



Σχήμα 4.5. Ποσοστό επίθεσης 1% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/50 (σε πακέτα)

Αυτό οφείλεται στο γεγονός ότι αυτή η μέθοδος ανίχνευσης επιθέσεων είναι βασισμένη στην ανίχνευση των πακέτων που έχουν ενεργή τη σημαία SYN ή FIN (ή RST). Αυτά τα πακέτα δεν κατανέμονται ομοιόμορφα μέσα στην ακολουθία πακέτων του δικτύου, με συνέπεια την ανεπαρκή τους δειγματοληψία όταν χρησιμοποιείται η μέθοδος της συστηματικής δειγματοληψίας. Σε αντίθεση, η τυχαία 1-από-N και η ομοιόμορφη πιθανολογική δειγματοληψία κάνουν μια «πιο τυχαία» επιλογή των πακέτων η οποία φαίνεται να είναι αποτελεσματικότερη. Κατά συνέπεια, όπως καταδεικνύεται εδώ, η συστηματική δειγματοληψία αποτελεί την χειρότερη επιλογή στη δειγματοληψία πακέτων για την ανίχνευση επιθέσεων που εξαρτώνται από συγκεκριμένα χαρακτηριστικά πακέτων (όπως οι σημαίες TCP). Πρέπει ωστόσο να σημειωθεί εδώ ότι οι περισσότεροι από τους δρομολογητές στο διαδίκτυο εφαρμόζουν τη συστηματική δειγματοληψία πακέτων λόγω της απλότητας και του μικρού υπολογιστικού φόρτου της [Choi05].

4.2.3.2 Μέθοδος CPD με μετρικό τις IP διευθύνσεις πηγής

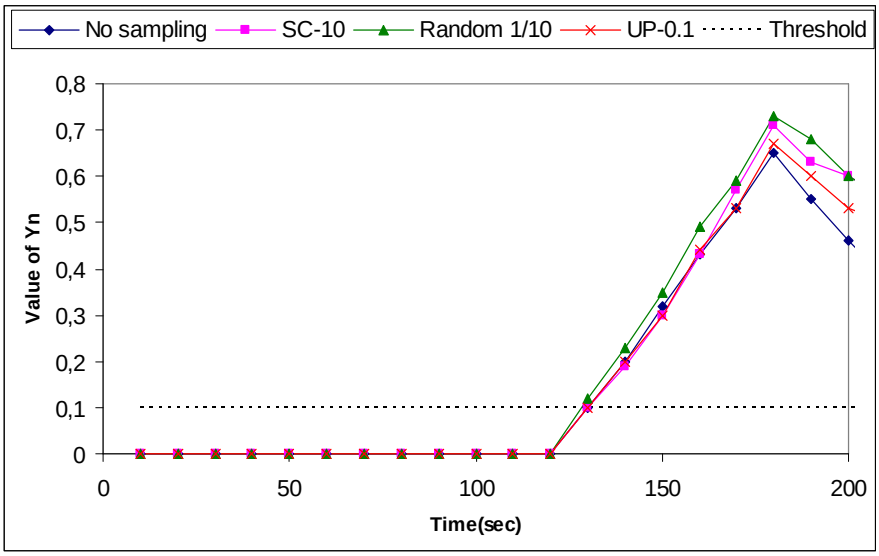
Το δεύτερο σενάριο αξιολόγησης της επίδρασης της δειγματοληψίας στην ανίχνευση ανωμαλιών χρησιμοποιεί την ίδια μέθοδο (CPD) αλλά με διαφορετικό μετρικό. Σε αυτήν την περίπτωση, η μεταβλητή X_n είναι το ποσοστό των «νέων» IP διευθύνσεων πηγής που παρατηρούνται στο σύνολο των «συχνών» IP διευθύνσεων

πηγής σε ένα παράθυρο χρόνου [Peng04]. Με τον όρο «συχνές IP διευθύνσεις» ορίζουμε τις διευθύνσεις πηγής που εμφανίζουν τους μεγαλύτερους αριθμούς σε πακέτα σε ένα χρονικό παράθυρο. Με τον όρο «νέες IP διευθύνσεις» ορίζουμε το σύνολο των IP διευθύνσεων πηγής που δεν ανήκουν στο σύνολο των IP διευθύνσεων που περιέχει τις «συχνές» διευθύνσεις IP κατά τη διάρκεια της κανονικής λειτουργίας του δικτύου.

Για αυτήν την μέθοδο που εστιάζει στις IP διευθύνσεις πηγής, πειραματιστήκαμε με ποσοστά επίθεσης 20% και 10% (σε πακέτα), διότι για μικρότερα ποσοστά επίθεσης ακόμη και χωρίς δειγματοληψία η μέθοδος αυτή αποτυγχάνει να ανιχνεύσει την επίθεση. Στην περίπτωση του ποσοστού επίθεσης 20%, οι πηγές επίθεσης κατανέμονται μεταξύ 254 διαφορετικών IP διευθύνσεων πηγής, ενώ στην περίπτωση 10%, οι πηγές επίθεσης περιλαμβάνουν μόνο 100 διαφορετικές διευθύνσεις IP.

Το Σχήμα 4.6 παρουσιάζει τα αποτελέσματα για το ποσοστό επίθεσης 20% (σε πακέτα) σε ποσοστό δειγματοληψίας 1/10 (σε πακέτα). Όπως απεικονίζεται σε αυτό το σχήμα, η δειγματοληψία δεν έχει σημαντικές επιπτώσεις στην αποτελεσματικότητα της ανίχνευσης. Παρόμοια αποτελέσματα εμφανίζονται και για το ποσοστό δειγματοληψίας 1/50 (σε πακέτα).

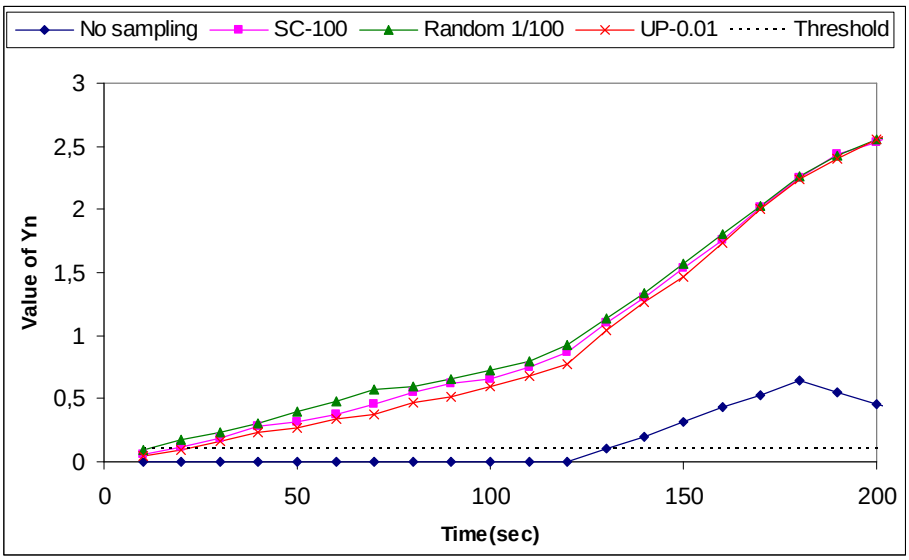
.



Σχήμα 4.6. Ποσοστό επίθεσης 20% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/10 (σε πακέτα)

Το Σχήμα 4.7 παρουσιάζει τα αντίστοιχα αποτελέσματα για το ποσοστό επίθεσης 20% σε ποσοστό δειγματοληψίας 1/100 (σε πακέτα). Είναι προφανές ότι και

οι τρεις τύποι δειγματοληψίας έχουν την ίδια συμπεριφορά, η οποία οδηγεί στη σημαντική αύξηση των ψευδών θετικών περιστατικών.



Σχήμα 4.7. Ποσοστό επίθεσης 20% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/100 (σε πακέτα)

Αυτό το γεγονός μπορεί να εξηγηθεί από τη παραμόρφωση της κατανομής πακέτων των IP διευθύνσεων πηγής όταν εφαρμόζεται η δειγματοληψία πακέτων. Πιο συγκεκριμένα, βασισμένοι στο στατιστικό προσδιορισμό του κατάλληλου μεγέθους δείγματος [Coch87], υπολογίζουμε το κατάλληλο μέγεθος δείγματος για την περίπτωσης μας (πακέτα που πρέπει να επιλεχτούν σε ένα χρονικό παράθυρο) για ένα δεδομένο επίπεδο εμπιστοσύνης στην κατανομή πακέτων των IP διευθύνσεων πηγής. Ορίζουμε μια ακρίβεια $r = 5\%$ και ένα επίπεδο εμπιστοσύνης 95% που οδηγεί σε μία τιμή $z = 1.96$ στην ακόλουθη έκφραση για το κατάλληλο μέγεθος δείγματος n:

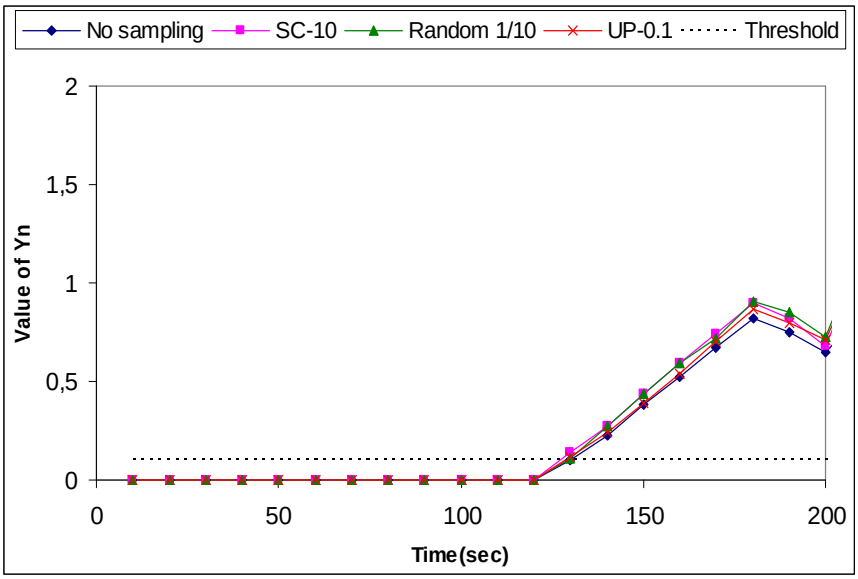
$$n = \left(\frac{100 \cdot zs}{rm} \right)^2 \quad (4.1)$$

όπου μ είναι ο μέσος όρος και σ είναι η τυπική απόκλιση της κατανομής.

Στο πείραμά μας, η κατανομή πακέτων των IP διευθύνσεων πηγής είχε μέσο όρο $\mu=16.5$ και τυπική απόκλιση $\sigma=71.6$. Επομένως, σύμφωνα με την σχέση (4.1), το κατάλληλο μέγεθος του δείγματος σε ένα χρονικό παράθυρο πρέπει να είναι 28687 πακέτα. Δεδομένου ότι ο αριθμός πακέτων σε ένα χρονικό παράθυρο είναι περίπου 195000, το κατάλληλο ποσοστό δειγματοληψίας πρέπει να είναι περίπου 1/7. Θεωρώντας εφαρμογή ποσοστού δειγματοληψίας ίσο με 1/100, κάποιος μπορεί να

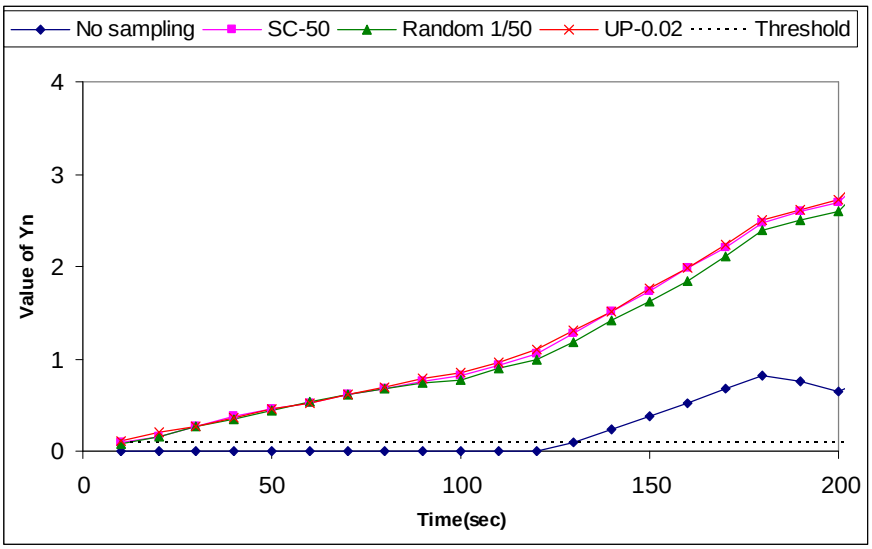
πραγματοποιήσει πολύ σημαντική παραμόρφωση της κατανομής πακέτων. Συνεπώς, η μείωση του ποσοστού δειγματοληψίας, έχει ως αποτέλεσμα την αύξηση του ποσοστού των νέων IP διευθύνσεων πηγής στις συχνές IP διευθύνσεις πηγής, ακόμη και σε ομαλή λειτουργία του δικτύου, όπως απεικονίζεται στο Σχήμα 4.7.

Το Σχήμα 4.8 παρουσιάζει τα αντίστοιχα αποτελέσματα για ποσοστό επίθεσης 10% (σε πακέτα) σε ποσοστό δειγματοληψίας 1/10. Όπως μπορούμε να παρατηρήσουμε, και οι τρεις τύποι δειγματοληψίας συμπεριφέρονται όμοια και οι γραφικές παραστάσεις τους είναι παρόμοιες με τη γραφική παράσταση στην οποία δεν εφαρμόζεται δειγματοληψία παρόλο που το ποσοστό επίθεσης έχει μειωθεί και η επίθεση έχει γίνει λιγότερο κατανεμημένη (από λιγότερες IP διευθύνσεις).



Σχήμα 4.8. Ποσοστό επίθεσης 10% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/10 (σε πακέτα)

Το Σχήμα 4.9 παρουσιάζει την περίπτωση του ποσοστού επίθεσης 10% σε ποσοστό δειγματοληψίας 1/50 (σε πακέτα), στο οποίο και οι τρεις τύποι δειγματοληψίας παρουσιάζουν μια μεγάλη απόκλιση από το αρχική καμπύλη (χωρίς δειγματοληψία). Σε αυτήν την περίπτωση (ποσοστό επίθεσης 10%) μειώσαμε το κατώφλι ανίχνευσης της επίθεσης προκειμένου να επιτευχθεί η ανίχνευση της επίθεσης ακόμη και στην περίπτωση χωρίς δειγματοληψία. Κατά συνέπεια, με το ποσοστό δειγματοληψίας 1/50, η αναλογία των «νέων» IP διευθύνσεων πηγής σε σχέση με τις «συχνές» IP διευθύνσεις παράγει ένα μεγάλο αριθμό ψευδών θετικών περιστατικών.



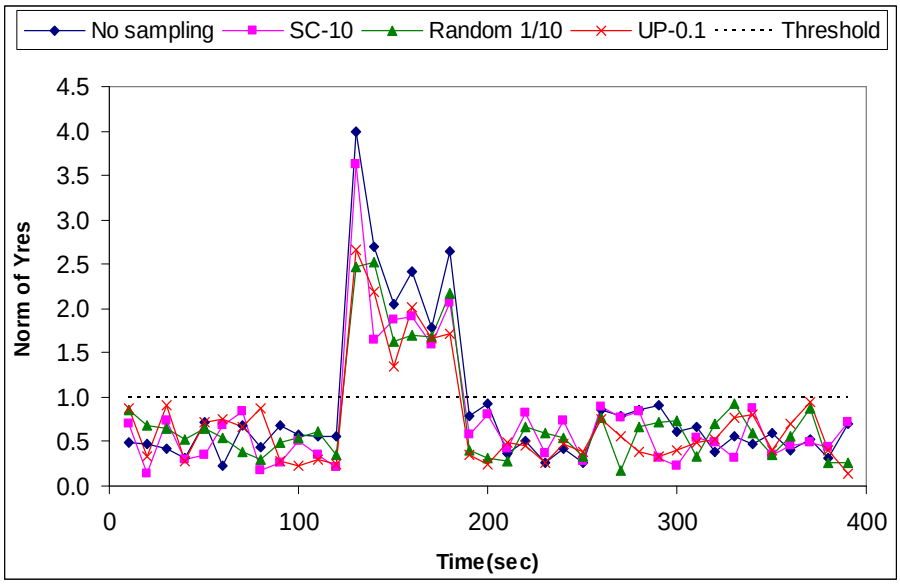
Σχήμα 4.9. Ποσοστό επίθεσης 10% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/50 (σε πακέτα)

4.2.3.3 Μέθοδος Ανάλυσης Κύριων Συνιστωσών (PCA) με τη χρήση πολλαπλών μετρικών

Όπως έχει αναφερθεί σε προηγούμενη ενότητα, η μέθοδος PCA αποτελεί μία τεχνική ανίχνευσης ανωμαλιών που μπορεί να λάβει υπόψη ταυτόχρονα πολλαπλά μετρικά που προέρχονται από μία δικτυακή σύνδεση ή και πολλαπλές συνδέσεις. Στο πείραμά μας, εφαρμόσαμε τη μέθοδο σε μία δικτυακή σύνδεση χρησιμοποιώντας πολλαπλά μετρικά. Συγκεκριμένα, χρησιμοποιήσαμε τα ακόλουθα επτά μετρικά: αριθμός ροών, αριθμός πακέτων, αριθμός ροών TCP, αριθμός πακέτων TCP, αριθμός μικρών ροών (με μικρό αριθμό πακέτων), αριθμός πακέτων SYN, και αριθμός πακέτων FIN.

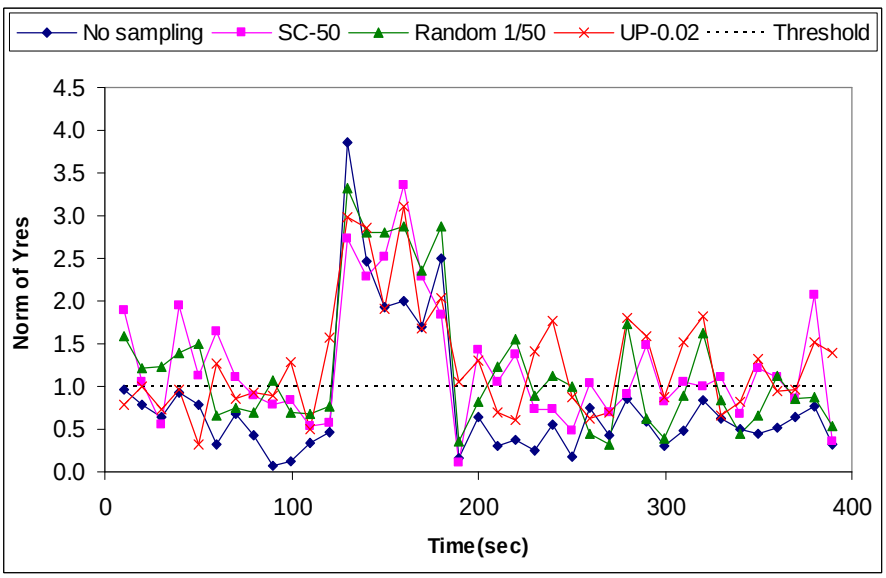
Όσον αφορά αυτήν την μέθοδο ανίχνευσης, στα πειράματά μας εξετάσαμε ποσοστά επίθεσης 20%, 10% και 5% (σε πακέτα). Τα αποτελέσματα καταδεικνύουν σχεδόν την ίδια συμπεριφορά σε όλα τα ανωτέρω ποσοστά επίθεσης. Επομένως, στη συνέχεια παρουσιάζουμε ενδεικτικά μόνο τα αντίστοιχα αποτελέσματα για το ποσοστό επίθεσης 5%.

Το Σχήμα 4.10 παρουσιάζει τα αποτελέσματα για το ποσοστό επίθεσης 5% (σε πακέτα) σε ποσοστό δειγματοληψίας 1/10. Όπως μπορούμε να παρατηρήσουμε, και οι τρεις τύποι δειγματοληψίας επιτυγχάνουν για να ανιχνεύσουν την επίθεση, ενώ ταυτόχρονα παρατηρείται μια μικρή διακύμανση στις τιμές τους.



Σχήμα 4.10. Ποσοστό επίθεσης 5% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/10 (σε πακέτα)

Το Σχήμα 4.11 παρουσιάζει την περίπτωση για το ποσοστό επίθεσης 5% (σε πακέτα) σε ποσοστό δειγματοληψίας 1/50, στο οποίο και οι τρεις τύποι δειγματοληψίας παρουσιάζουν έναν μεγάλο αριθμό ψευδών θετικών περιστατικών. Η ίδια συμπεριφορά εμφανίζεται για το ποσοστό δειγματοληψίας 1/100.



Σχήμα 4.11. Ποσοστό επίθεσης 5% (σε πακέτα) - Ποσοστό δειγματοληψίας 1/50 (σε πακέτα)

Το διάνυσμα y_{res} περιέχει τους συντελεστές συσχετισμού (correlation coefficients) που δεν ήταν παρόντες στα δεδομένα της διαδικασίας εκμάθησης (κανονική λειτουργία δικτύου). Κατά τη διάρκεια της ομαλής λειτουργίας του δικτύου το διάνυσμα y_{res} τείνει στο μηδέν. Η αύξηση του y_{res} αντιστοιχεί σε μεγάλη διακύμανση των συντελεστών συσχετισμού των μετρικών που αποδίδεται συνήθως στην παρουσία μιας επίθεσης. Όπως απεικονίζεται στο Σχήμα 4.11, η PCA μέθοδος ανίχνευσης ανωμαλιών ανιχνεύει την επίθεση που πραγματοποιήθηκε κατά τη διάρκεια του χρονικού διαστήματος 130-180sec, για όλες τις δειγματοληπτικές μεθόδους. Ο μεγάλος αριθμός ψευδών θετικών περιστατικών που παρατηρείται στο Σχήμα 4.11, πριν από την έναρξη και μετά από το τέλος της επίθεσης, οφείλεται στις σημαντικές διακυμάνσεις των συντελεστών συσχετισμού των μετρικών για το ποσοστό δειγματοληψίας 1/50 (όταν συγκρίνεται με την περίπτωση που δεν εφαρμόζεται δειγματοληψία).

Ενδεικτικά στον Πίνακα 4.2 παρουσιάζουμε την τυπική απόκλιση δύο μετρικών (συνολικός αριθμός πακέτων και συνολικός αριθμός ροών) του διανύσματος y_{res} για τα διαφορετικά ποσοστά δειγματοληψίας κατά τη διάρκεια της ομαλής λειτουργίας του δικτύου. Όπως μπορούμε να παρατηρήσουμε από τον πίνακα, η διαφορά στην τυπική απόκλιση μεταξύ των ποσοστών δειγματοληψίας 1/10 και 1/50 είναι σημαντική. Και ο αριθμός των πακέτων και των ροών στο ποσοστό δειγματοληψίας 1/50 παρουσιάζουν μια μεγάλη διαφορά από την αρχική περίπτωση όπου δεν εφαρμόζεται δειγματοληψία.

Πίνακας 4.2. Τυπική απόκλιση για το y_{res} στην ομαλή λειτουργία του δικτύου

	Sampling Rate		
Metric	**No Sampling**	**1/10**	**1/50**
Packets	0.10	0.12	0.31
Flows	0.05	0.13	0.21

Τα πειραματικά αποτελέσματα αυτής της PCA μεθόδου, που χρησιμοποιεί και μετρικά βασισμένα σε πακέτα, αλλά και σε ροές, δείχνουν ότι η αποδοτικότητα της μεθόδου αυτής είναι ανεξάρτητη από τη δειγματοληπτική μέθοδο που χρησιμοποιείται και εξαρτάται μόνο από το ποσοστό δειγματοληψίας.

Από την παραπάνω μελέτη διαπιστώνουμε ότι η δειγματοληψία πακέτων (τις περισσότερες φορές ανεξάρτητα της συγκεκριμένης μεθόδου) μειώνει σε μεγάλο βαθμό την αποτελεσματικότητα των διαφόρων μεθόδων ανίχνευσης ανωμαλιών. Αυτό εμφανίζεται άλλοτε απλά με την χρονική καθυστέρηση στην ανίχνευση της ανωμαλίας και σε άλλες περιπτώσεις με την δημιουργία πολλών ψευδών θετικών περιστατικών. Επομένως, διαπιστώνουμε ότι η δειγματοληψία πακέτων εισάγει αρκετά σφάλματα στη διαδικασία ανίχνευσης ανωμαλιών στο δίκτυο με αποτέλεσμα να την καθιστά αναποτελεσματική.

Αν συνυπολογίσουμε και το γεγονός ότι κατά τη εφαρμογή της δειγματοληψίας πακέτων, οι μικρές ροές τείνουν να μην επιλεγούν, τότε ανωμαλίες, όπως οι επιθέσεις DDoS, οι αυτοδιαδιδόμενοι ιοί (worms), κ.λπ, είναι δυσκολότερο να ανιχνευθούν. Επομένως, προκύπτει η ανάγκη για πιο εξελιγμένες μεθόδους δειγματοληψίας, ώστε να παρέχουν στη διαδικασία ανίχνευσης ορθότερα δεδομένα δικτύου. Αυτές οι εξελιγμένες μέθοδοι δειγματοληψίας, παρέχουν συνήθως ανάλυση των ροών του δικτύου. Στη συνέχεια εισάγουμε μία τέτοια μέθοδο δειγματοληψίας που βασίζεται στις ροές και μελετούμε την αποτελεσματικότητα της σε διαφορετικές τεχνικές ανίχνευσης ανωμαλιών.

5. Επιλεκτική Δειγματοληψία (Selective Sampling)

5.1. Προτεινόμενη μέθοδος

Κύριος στόχος της παρούσας διατριβής είναι να προτείνει μεθοδολογίες δειγματοληψίας που είναι κατάλληλες για την εφαρμογή τους στο πεδίο της ανίχνευσης ανωμαλιών δικτύου καθώς και την βελτίωση των υπάρχουσων μεθόδων ώστε να ανταποκρίνονται αποτελεσματικότερα στην ανίχνευση ανωμαλιών. Στο κεφάλαιο αυτό περιγράφουμε την προτεινόμενη μέθοδο δειγματοληψίας και μελετούμε την αποτελεσματικότητα της σε δύο ευρέως διαδεδομένους αλγόριθμους ανίχνευσης ανωμαλιών.

Βασισμένοι στην ιδέα της «Έξυπνης Δειγματοληψίας» (Smart Sampling) [Duff03] στην οποία οι ροές επιλέγονται με πιθανότητα ανάλογη προς το μέγεθός τους, προτείνουμε μια νέα μέθοδο δειγματοληψίας ροής – στο εξής θα αναφερόμαστε σε αυτήν ως «Επιλεκτική Δειγματοληψία» [Andr07][Andr08] – η οποία εστιάζει στην επιλογή των μικρών ροών. Πρέπει να υπογραμμιστεί εδώ ότι οι μικρές ροές είναι συνήθως η πηγή πολλών επιθέσεων σε δίκτυα όπως είναι οι επιθέσεις άρνησης υπηρεσίας (DoS attacks) και η διάδοση ιών (worm propagation) [Barf01][Srid06] και πρέπει να επιλεγούν προκειμένου να επιτευχθεί μια αποδοτική διαδικασία ανίχνευσης ανωμαλιών. Σύμφωνα με τη προτεινόμενη μεθοδολογία η επιλογή μιας μεμονωμένης ροής είναι βασισμένη στην ακόλουθη σχέση:

$$p(x) = \begin{cases} c & x \le z \\ \dfrac{z}{n \cdot x} & x > z \end{cases} \qquad (5.1)$$

όπου με x συμβολίζουμε το μέγεθος της ροής σε πακέτα, $0 < c \le 1$, $n \ge 1$ και z είναι ένα κατώφλι (μετρούμενο σε πακέτα).

Όπως μπορούμε να παρατηρήσουμε από τη σχέση (5.1), οι ροές που είναι μικρότερες από το κατώφλι z επιλέγονται με μια σταθερή πιθανότητα c, ενώ οι ροές που είναι μεγαλύτερες στο μέγεθος από το z επιλέγονται με πιθανότητα αντιστρόφως

ανάλογη προς το μέγεθός τους. Ένα από τα κύρια χαρακτηριστικά της νέας μεθόδου δειγματοληψίας είναι η προσαρμοστικότητα της σε σχέση με άλλες προσεγγίσεις (π.χ. Έξυπνη Δειγματοληψία) επειδή μπορεί να ελέγξει περαιτέρω και να μειώσει τον αριθμό των επιλεγμένων ροών. Αφενός, με την κατάλληλη επιλογή της τιμής για την παράμετρο *c* μπορούμε να επιλέξουμε ένα σημαντικό ποσοστό των μικρών ροών χωρίς να μειωθεί η αποτελεσματικότητα στην ανίχνευση ανωμαλιών. Αφ' ετέρου, η επιλογή των μεγάλων ροών μπορεί να μειωθεί περαιτέρω με την αύξηση της τιμής της παραμέτρου *n*. Για μεγάλες τιμές της παραμέτρου *n*, οι ροές με έναν τεράστιο αριθμό πακέτων αναμένονται να μη ληφθούν στη διαδικασία δειγματοληψίας. Αυτό το γεγονός είναι πολύ σημαντικό επειδή μειώνεται σημαντικά η ποσότητα των επιλεγέντων πακέτων που πρέπει να επεξεργαστούν από τον αλγόριθμο ανίχνευσης ανωμαλιών. Πιο συγκεκριμένα, ας θεωρήσουμε την τυχαία μεταβλητή *x* να αντιπροσωπεύει το μέγεθος ροής σε πακέτα και $\mathbf{l}(x)$ την αντίστοιχη συνάρτηση μάζας πιθανότητας (probability mass function). Ο αριθμός των επιλεγμένων ροών N_f δίνεται από την ακόλουθη σχέση:

$$N_f = \sum_{x=1}^{N} p(x) \cdot \mathbf{l}(x) \cdot S_f$$

όπου S_f είναι ο συνολικός αριθμός των ροών και *N* είναι το μέγιστο μέγεθος ροής.

Με τη βοήθεια της σχέσης (5.1) η παραπάνω σχέση μετασχηματίζεται στην ακόλουθη:

$$N_f = \sum_{x=1}^{z} c \cdot \mathbf{l}(x) \cdot S_f + \sum_{x=z+1}^{N} \frac{z}{n \cdot x} \cdot \mathbf{l}(x) \cdot S_f$$

Ο αριθμός των επιλεγμένων πακέτων N_p δίνεται από την παρακάτω σχέση:

$$N_p = \sum_{x=1}^{N} p(x) \cdot \mathbf{l}(x) \cdot x \cdot S_f$$

Χρησιμοποιώντας την σχέση (5.1) έχουμε τελικά:

$$N_p = \sum_{x=1}^{z} c \cdot \mathbf{l}(x) \cdot x \cdot S_f + \sum_{x=z+1}^{N} \frac{z}{n} \cdot \mathbf{l}(x) \cdot S_f \qquad (5.2)$$

5.2. Εφαρμογή Επιλεκτικής Δειγματοληψίας στη μέθοδο Ανίχνευσης Αλλαγής Σημείου

Σε αυτό το τμήμα περιγράφουμε την εφαρμογή και την επίδραση της προτεινόμενης επιλεκτικής δειγματοληψίας πάνω σε έναν αλγόριθμο που χρησιμοποιεί την μέθοδο CPD για την ανίχνευση ανωμαλιών. Ο αλγόριθμος αυτός [Wang04] χρησιμοποιεί σαν μετρικό (μεταβλητή X_n) την διαφορά μεταξύ του πλήθους των TCP SYN και FIN πακέτων σε ένα ορισμένο χρονικό διάστημα. Όπως έχουμε περιγράψει και σε προηγούμενη ενότητα, η μεταβλητή X_n έχει μια θετική μέση τιμή κοντά στο μηδέν στις κανονικές συνθήκες, ενώ η μεταβλητή Z_n έχει αρνητική τιμή. Όταν μια επίθεση SYN κάνει την εμφάνισή της, το X_n γίνεται ένας μεγάλος θετικός αριθμός που έχει σαν αποτέλεσμα το Z_n να γίνει θετικό. Αυτό οδηγεί στην αύξηση του y_n που υποδηλώνει την εμφάνιση μιας επίθεσης εάν η τιμή της υπερβαίνει ένα προκαθορισμένο κατώφλι.

Η αποδοτικότητα ανίχνευσης επιθέσεων της μεθόδου CPD απεικονίζεται από την κλίση της καμπύλης της μεταβλητής y_n και καθορίζεται από τον αντίστοιχο βαθμό ανίχνευσης (*Degree of detection*), ο οποίος ορίζεται από την σχέση (5.3). Προκειμένου να επιτευχθεί ο καλύτερος βαθμός ανίχνευσης, η τιμή του X_n, που στην περίπτωσή μας είναι η διαφορά μεταξύ των πακέτων SYN και FIN διαιρεμένη με τον αριθμό πακέτων FIN, πρέπει να μεγιστοποιηθεί. Επομένως, η ακόλουθη έκφραση που δίνει το βαθμό ανίχνευσης πρέπει να μεγιστοποιηθεί:

$$\textit{Degree of detection} = \frac{N_{SYN} - N_{FIN}}{N_{FIN}} \qquad (5.3)$$

όπου N_{SYN} είναι ο συνολικός αριθμός των επιλεγμένων πακέτων SYN και N_{FIN} ο συνολικός αριθμός επιλεγμένων πακέτων FIN σε ένα ορισμένο χρονικό διάστημα.

Ο συνολικός αριθμός των επιλεγμένων πακέτων SYN, N_{SYN}, δίνεται από την παρακάτω σχέση:

$$N_{SYN} = \sum_{x=1}^{N} \left(p(x) \cdot \mathbf{l}(x) \cdot S_f \cdot \sum_{y=1}^{x} \left(y \cdot f(y \mid x) \right) \right)$$

όπου S_f είναι ο συνολικός αριθμός των ροών,

N είναι το μέγιστο μέγεθος ροής σε πακέτα,

$\mathbf{l}(x)$ είναι η συνάρτηση μάζας πιθανότητας της τυχαίας μεταβλητής x που παριστάνει το μέγεθος της ροής σε πάκετα,

$p(x)$ είναι η πιθανότητα επιλογής μίας ροής με μέγεθος x και δίνεται από την σχέση (1) και

$f(y|x)$ είναι η δεσμευμένη πιθανότητα μίας ροής να έχει y πακέτα SYN δεδομένου ότι έχει μέγεθος ίσο με x πακέτα.

Με τη βοήθεια της σχέσης (4.2) παίρνουμε:

$$N_{SYN} = \sum_{x=1}^{z} \left(c \cdot \mathbf{l}(x) \cdot S_f \cdot \sum_{y=1}^{x} \left(y \cdot f(y \mid x) \right) \right) + \sum_{x=z+1}^{N} \left(\frac{z}{n \cdot x} \cdot \mathbf{l}(x) \cdot S_f \cdot \sum_{y=1}^{x} \left(y \cdot f(y \mid x) \right) \right) \quad (5.4)$$

Ο συνολικός αριθμός των πακέτων FIN, N_{FIN}, δίνεται από την παρακάτω σχέση :

$$N_{FIN} = \sum_{x=1}^{N} \left(p(x) \cdot \mathbf{l}(x) \cdot S_f \cdot \sum_{y=1}^{x} \left(y \cdot g(y \mid x) \right) \right)$$

όπου $g(y|x)$ είναι η δεσμευμένη πιθανότητα μιας ροής να έχει y πακέτα FIN δεδομένου ότι το μέγεθος της είναι ίσο με x. Με την βοήθεια της σχέσης (5.1) έχουμε:

$$N_{FIN} = \sum_{x=1}^{z} \left(c \cdot \mathbf{l}(x) \cdot S_f \cdot \sum_{y=1}^{x} \left(y \cdot g(y \mid x) \right) \right) + \sum_{x=z+1}^{N} \left(\frac{z}{n \cdot x} \cdot \mathbf{l}(x) \cdot S_f \cdot \sum_{y=1}^{x} \left(y \cdot g(y \mid x) \right) \right) \quad (5.5)$$

Αν αντικαταστήσουμε τα N_{SYN} και N_{FIN} στην σχέση (5.3) από τις αντίστοιχες σχέσεις (5.4) και (5.5), παίρνουμε μία έκφραση που περιγράφει τον βαθμό ανίχνευσης αναφορικά με τις παραμέτρους *z, c* και *n*. Διαλέγοντας επομένως τις κατάλληλες τιμές για τις παραμέτρους *z, c,* and *n* μπορούμε να μεγιστοποιήσουμε την τιμή της σχέσης (5.3), η οποία θα οδηγήσει σε έναν καλύτερο βαθμό ανίχνευσης των επιθέσεων. Η επίδραση των παραμέτρων αυτών στην ακρίβεια της ανίχνευσης της επίθεσης μελετάται στην επόμενη ενότητα.

5.2.1. Αξιολόγηση επίδοσης της προτεινόμενης μεθόδου

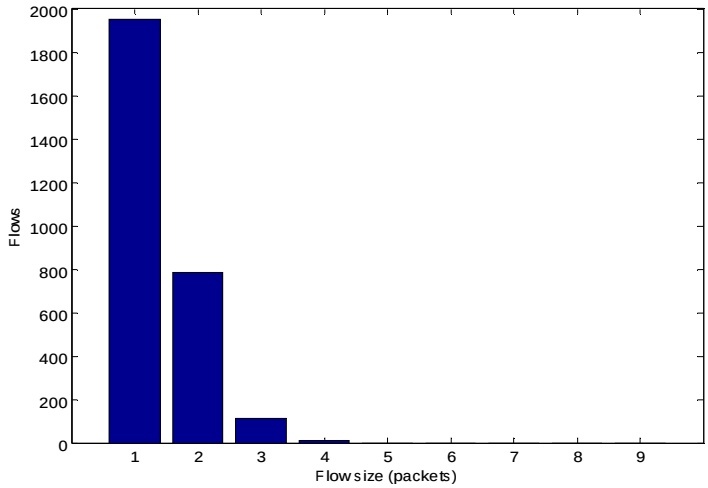
Σε αυτή την ενότητα μελετάται και αξιολογείται η αποτελεσματικότητα της προτεινόμενης επιλεκτικής δειγματοληψίας σε συνδυασμό με τη μέθοδο ανίχνευσης ανωμαλιών που περιγράφηκε πιο πάνω κάτω από διαφορετικά σενάρια επίθεσης. Τα αποτελέσματα και οι αντίστοιχες παρατηρήσεις που παρουσιάζονται σε αυτή την ενότητα είναι βασισμένα σε πραγματικά δεδομένα δικτύου που έχουν συλλεχθεί από ένα ακαδημαϊκό δίκτυο. Όπως αναφέραμε και σε προηγούμενη ανάλυση, μελετήσαμε τη σύνδεση μεταξύ του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ) και του Εθνικού Δικτύου Έρευνας και Τεχνολογίας (ΕΔΕΤ) [GRNet] που συνδέει το ΕΜΠ με το Διαδίκτυο. Στο διάστημα των πειραμάτων μας, αυτή η σύνδεση είχε μια μέση κίνηση της τάξης των 70-80Mbit/sec και περίπου 20000 πακέτα/sec. Στην ακόλουθη αξιολόγηση, μελετάται λεπτομερώς μια κατανεμημένη επίθεση άρνησης υπηρεσίας (συγκεκριμένα επίθεση TCP SYN) ενάντια σε έναν υπολογιστή μέσα στο ΕΜΠ, η οποία παράγεται από ένα πραγματικό εργαλείο επίθεσης.

- **Συγκριτική Μελέτη**

Στα πειράματά μας, προκειμένου να αποκτηθεί κάποια γνώση σχετικά με την αποτελεσματικότητα της επιλεκτικής δειγματοληψίας συγκρίνουμε αρχικά τη μέθοδο αυτή ενάντια στην τυχαία δειγματοληψία ροών (random flow sampling). Όπως έχουμε περιγράψει σε προηγούμενο κεφάλαιο, στην τυχαία δειγματοληψία ροών κάθε ροή επιλέγεται ανεξάρτητα με την ίδια πιθανότητα *p*. Το ποσοστό των επιλεγμένων ροών ορίστηκε ως κοινό κριτήριο για τη σύγκριση των δύο μεθόδων. Επιπλέον, προκειμένου να αξιολογηθούν καλύτερα και να γίνουν κατανοητά τα χαρακτηριστικά της επιλεκτικής δειγματοληψίας μελετάμε την σχέση μεταξύ του επιτεύξιμου βαθμού

ανίχνευσης και του αριθμού των επιλεγμένων πακέτων κατά τη διάρκεια της διαδικασίας της επιλεκτικής δειγματοληψίας.

Για να καταδείξουμε καλύτερα τα αποτελέσματα και τις αντίστοιχες παρατηρήσεις, μελετάμε δύο σενάρια για τη μέθοδο ανίχνευσης αλλαγής σημείου (CPD). Στο πρώτο σενάριο συγκρίνουμε την Επιλεκτική Δειγματοληψία με την Τυχαία Δειγματοληψία Ροών και πειραματιζόμαστε με επιθέσεις που αντιστοιχούν στο 2% της κανονικής κίνησης (σε πακέτα), ενώ στο δεύτερο σενάριο μειώνουμε περαιτέρω το ποσοστό επίθεσης ώστε να αποτελεί το 1% της κανονικής κίνησης του δικτύου. Το Σχήμα 5.1 παρουσιάζει τη κατανομή μεγέθους των ροών επίθεσης (σε πακέτα) στην περίπτωση του ποσοστού επίθεσης 2%.



Σχήμα 5.1. Κατανομή του μεγέθους ροών επίθεσης για ποσοστό επίθεσης 2%.

Όπως μπορούμε να παρατηρήσουμε, οι περισσότερες από τις ροές επίθεσης έχουν 1 ή 2 πακέτα. Βασιζόμενοι στο Σχήμα 5.1, επιλέγουμε μερικές χαρακτηριστικές τιμές για τις παραμέτρους της επιλεκτικής δειγματοληψίας. Συγκεκριμένα, επιλέγουμε τις τιμές του *z* προκειμένου να κάνουμε διάκριση μεταξύ των μικρών και μεγάλων ροών. Στην περίπτωσή μας, όπου η επίθεση κατανέμεται κυρίως μεταξύ ροών που έχουν 1 ή 2 πακέτα επιλέγουμε *z=1* και *z=2*. Στη συνέχεια, με την κατάλληλη τιμή για την παράμετρο *c* καθορίζουμε το ποσοστό των μικρών ροών που επιλέγονται (στα πειράματά μας χρησιμοποιούμε *c=1.0* και *c=0.2*) και με την παράμετρο *n* ρυθμίζουμε την επιλογή των μεγάλων ροών. Μεγαλύτερες τιμές του *n*

οδηγούν στην επιλογή μικρότερου αριθμού μεγάλων ροών και συνεπώς λιγότερων πακέτων. Για τον αλγόριθμο της τυχαίας δειγματοληψίας ροών επιλέγουμε την κατάλληλη πιθανότητα *p* που οδηγεί στο ίδιο ποσοστό των επιλεγμένων ροών. Ο Πίνακας 5.1 παρουσιάζει τις παραμέτρους για τις δύο τεχνικές δειγματοληψίας καθώς και το αντίστοιχο ποσοστό των επιλεγέντων πακέτων.

Πίνακας 5.1. Παράμετροι για τις τεχνικές δειγματοληψίας (Μέθοδος Ανίχνευσης CPD)

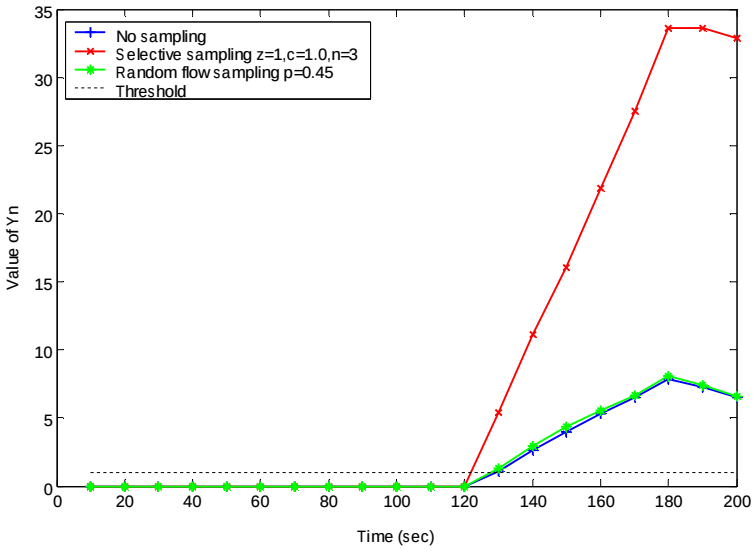
Flows (%)	Random Flow Sampling		Selective Sampling			
	p	*Packets (%)*	*z*	*c*	*n*	*Packets (%)*
45.12%	0.45	45.67%	1	1.0	3	4.25%
12.05%	0.12	12.53%	2	0.2	6	2.05%

Πρέπει να σημειωθεί εδώ ότι το ποσοστό των επιλεγμένων πακέτων στην περίπτωση της επιλεκτικής δειγματοληψίας είναι σημαντικά μικρότερο από την αντίστοιχη περίπτωση της τυχαίας δειγματοληψίας ροών. Ο αριθμός επιλεγμένων πακέτων είναι σημαντικής σπουδαιότητας, καθώς ο αλγόριθμος ανίχνευσης ανωμαλιών επιθεωρεί κάθε επιλεγμένο πακέτο προκειμένου να εξεταστεί το πεδίο TCP-flags της επικεφαλίδας.

Γενικά, λαμβάνοντας υπόψη το γεγονός ότι οι μικρές ροές είναι συνήθως η πηγή πολλών επιθέσεων δικτύων, ο διαχειριστής του δικτύου μπορεί να επιλέξει τις κατάλληλες τιμές για τις παραμέτρους *z*, *c* και *n* για να ανιχνεύσει ένα πλήθος επιθέσεων που αποτελούνται από μικρές ροές. Λόγω του γεγονότος ότι η επιλεκτική δειγματοληψία στοχεύει τις μικρές ροές, ο αριθμός των επιλεγέντων πακέτων είναι πολύ μικρός σε σχέση με την περίπτωση της τυχαίας δειγματοληψίας ροών στην οποία οι ροές επιλέγονται με την ίδια πιθανότητα, ανεξάρτητα από το μέγεθός τους. Αυτό καταδεικνύεται σαφώς από τα αποτελέσματα που παρουσιάζονται στον Πίνακα 5.1, όπου στην επιλεκτική δειγματοληψία επιλέχτηκε μόνο το 4% του συνολικού αριθμού πακέτων, ενώ στην αντίστοιχη περίπτωση της τυχαίας δειγματοληψίας ροών το ποσοστό των επιλεγέντων πακέτων ήταν περίπου 45%.

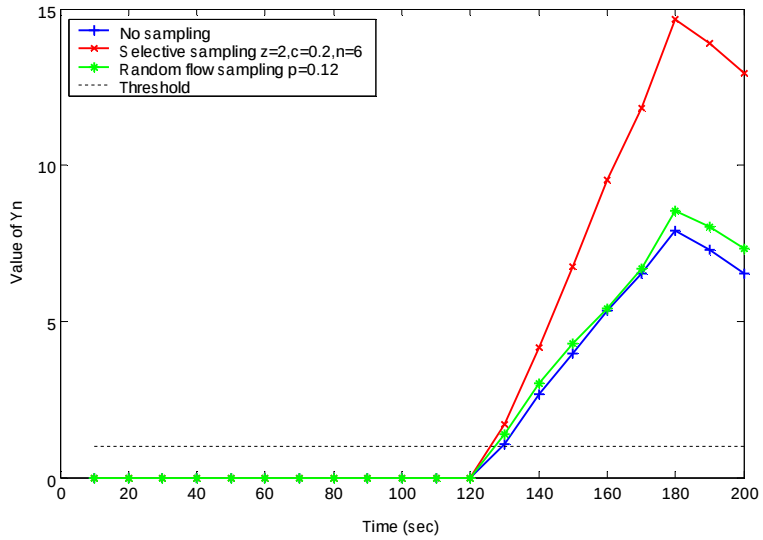
Τα αποτελέσματα που απεικονίζονται στο Σχήμα 5.2 αντιστοιχούν σε ποσοστό επίθεσης 2% και σε ποσοστό δειγματοληψίας 45% όσον αφορά τον αριθμό ροών. Όπως μπορούμε να παρατηρήσουμε, η καμπύλη που αντιστοιχεί στην τυχαία δειγματοληψία ροών μοιάζει αρκετά με την καμπύλη της αρχικής περίπτωσης που δεν εφαρμόζεται καθόλου δειγματοληψία. Αντίθετα, η επιλεκτική δειγματοληψία ξεπερνά και τις δύο αυτές περιπτώσεις δεδομένου ότι η κλίση της αντίστοιχης καμπύλης για

την επιλεκτική δειγματοληψία έχει αυξηθεί σημαντικά. Όπως αναφέραμε νωρίτερα, η κλίση αντιπροσωπεύει το βαθμό ανίχνευσης. Σε αυτό το σενάριο, είναι προφανές ότι η αποτελεσματικότητα ανίχνευσης βελτιώνεται σημαντικά στην περίπτωση της επιλεκτικής δειγματοληψίας. Αυτό αποδίδεται στο γεγονός ότι λιγότερα πακέτα FIN επιλέχτηκαν (τα οποία εμφανίζονται κανονικά στις μεγάλες ροές), και συγχρόνως όλα τα πακέτα επίθεσης SYN έχουν επιλεγεί.



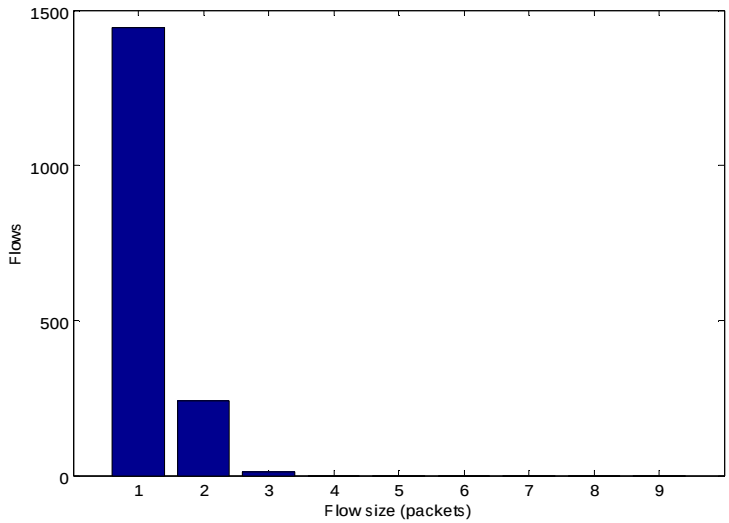
Σχήμα 5.2. Ανίχνευση επίθεσης για δειγματοληψία ροών 45% σε ποσοστό επίθεσης 2%

Στο Σχήμα 5.3 παρουσιάζουμε τα αντίστοιχα αποτελέσματα για ποσοστό δειγματοληψίας 12% όσον αφορά τον αριθμό επιλεγμένων ροών. Σε αυτήν την περίπτωση, η κλίση της καμπύλης έχει μειωθεί αισθητά, αλλά εξακολουθούμε να επιτυγχάνουμε ένα καλύτερο βαθμό ανίχνευσης επίθεσης στην περίπτωση της επιλεκτικής δειγματοληψίας έναντι της περίπτωσης που δεν έχουμε καθόλου δειγματοληψία, παρόλη την μικρή τιμή της παραμέτρου *c* που χρησιμοποιήθηκε (*c=0.2*) για την πιθανότητα επιλογής των μικρών ροών.



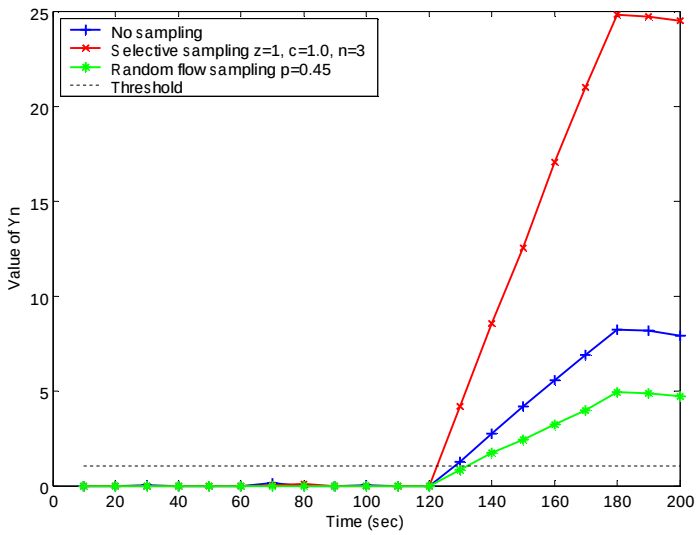
Σχήμα 5.3. Ανίχνευση επίθεσης για δειγματοληψία ροών 12% σε ποσοστό επίθεσης 2%

Στη συνέχεια, μειώσαμε το ποσοστό επίθεσης για να αντιστοιχεί σε ποσοστό 1% της κανονικής κίνησης του δικτύου. Το Σχήμα 5.4 παρουσιάζει την κατανομή μεγέθους των ροών επίθεσης (σε πακέτα). Όπως μπορούμε να παρατηρήσουμε, οι περισσότερες από τις ροές επίθεσης έχουν 1 πακέτο. Για αυτό το ποσοστό επίθεσης, εφαρμόσαμε την τυχαία δειγματοληψία ροών και την επιλεκτική δειγματοληψία με τις ίδιες παραμέτρους δειγματοληψίας όπως στην προηγούμενη περίπτωση του ποσοστού επίθεσης 2%. Τα αντίστοιχα αποτελέσματα απεικονίζονται στα Σχήματα 5.5 και 5.6.



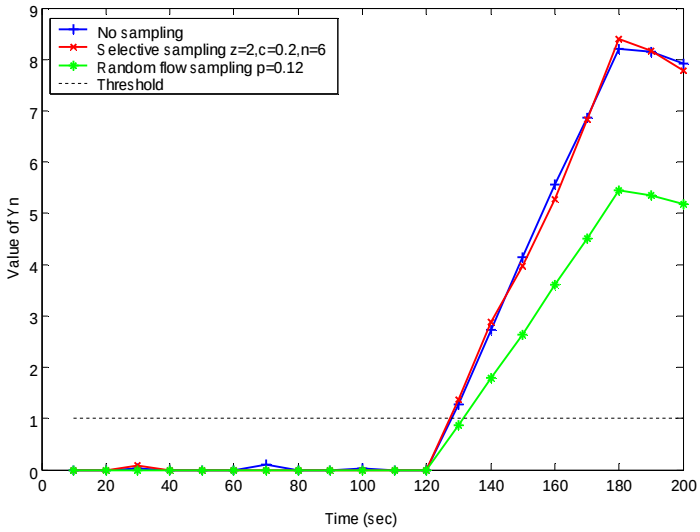
Σχήμα 5.4. Κατανομή του μεγέθους ροών επίθεσης για ποσοστό επίθεσης 1%

Συγκεκριμένα, στο Σχήμα 5.5 παρουσιάζουμε τα αποτελέσματα για το ποσοστό δειγματοληψίας 45% όσον αφορά τις επιλεγμένες ροές. Είναι προφανές ότι στην περίπτωση της επιλεκτικής δειγματοληψίας ο βαθμός ανίχνευσης επίθεσης είναι ακόμα μεγαλύτερος από την περίπτωση κατά την οποία δεν έχουμε καθόλου δειγματοληψία. Αυτό οφείλεται στο γεγονός ότι ένα σημαντικό μέρος των ροών επίθεσης έχει επιλεγεί. Αντίθετα, έχουμε έναν μικρότερο βαθμό ανίχνευσης για την περίπτωση της τυχαίας δειγματοληψίας ροών. Συγκεκριμένα, όπως μπορούμε να παρατηρήσουμε, η επίθεση δεν είναι ανιχνεύσιμη έγκαιρα στη χρονική στιγμή των 130sec για την περίπτωση της τυχαίας δειγματοληψίας ροών.



Σχήμα 5.5. Ανίχνευση επίθεσης για δειγματοληψία ροών 45% σε ποσοστό επίθεσης 1%

Το Σχήμα 5.6 παρουσιάζει τα αντίστοιχα αποτελέσματα για τη δειγματοληψία ροών σε ποσοστό 12%. Η κλίση της καμπύλης για την περίπτωση της επιλεκτικής δειγματοληψίας έχει μειωθεί και μοιάζει με την περίπτωση που δεν εφαρμόζεται καθόλου δειγματοληψία. Αυτό αποδίδεται στο γεγονός ότι ένας μικρότερος αριθμός πακέτων SYN παράγεται κατά τη διάρκεια της επίθεσης, αναγκάζοντας κατά συνέπεια τη διαφορά μεταξύ του πλήθους των πακέτων SYN και FIN να μειωθεί. Αντίθετα, ο αλγόριθμος αποτυγχάνει να ανιχνεύσει την επίθεση κατά την αρχική της φάση στην περίπτωση της τυχαίας δειγματοληψίας ροών.



Σχήμα 5.6. Ανίχνευση επίθεσης για δειγματοληψία ροών 12% σε ποσοστό επίθεσης 1%

- **Επίδραση Παραμέτρων στην Επιλεκτική Δειγματοληψία**

Στη συνέχεια μελετάμε την σχέση ανάμεσα στο βαθμό ανίχνευσης της ανωμαλίας (όπως καθορίστηκε στην σχέση (5.3)) και στον αριθμό των επιλεγμένων πακέτων κατά τη διαδικασία της επιλεκτικής δειγματοληψίας. Αυτή η μελέτη είναι βασισμένη στο σενάριο του ποσοστού επίθεσης SYN που αντιστοιχεί σε 1% της ομαλής κίνησης του δικτύου. Με βάση το γεγονός ότι το μεγαλύτερο ποσοστό των ροών επίθεσης έχει ένα μόνο πακέτο (Σχήμα 5.4) επιλέξαμε την παράμετρο *z* να έχει τιμή $z = 1$. Με τη χρήση των σχέσεων (5.3), (5.4) και (5.5) υπολογίζουμε το βαθμό ανίχνευσης για διάφορες τιμές των παραμέτρων *c* και *n*.

Ο Πίνακας 5.2 παρουσιάζει το βαθμό ανίχνευσης για τις διάφορες τιμές της παραμέτρου *c* στο διάστημα $0.1 \leq c \leq 1.0$ με βήμα 0.1, και για διάφορες τιμές της παραμέτρου *n* που κυμαίνονται από 1 έως 1000. Όπως αναφέραμε προηγουμένως, η παράμετρος *c* καθορίζει την πιθανότητα επιλογής των μικρών ροών (ροές που έχουν ίσο ή μικρότερο αριθμό πακέτων από *z*), ενώ η παράμετρος *n* χαρακτηρίζει την επιλογή των μεγάλων ροών. Η επιλογή των μεγάλων ροών μειώνεται με την αύξηση της τιμής της παραμέτρου *n*. Συγκεκριμένα, στον Πίνακα 5.3 και στον Πίνακα 5.4 παρουσιάζουμε τον απόλυτο αριθμό και το αντίστοιχο ποσοστό των επιλεγμένων πακέτων σε ένα χρονικό διάστημα 10sec αναφορικά με τις παραμέτρους *c* και *n*.

Πίνακας 5.2. Βαθμός ανίχνευσης αναφορικά με τις παραμέτρους c και n της επιλεκτικής δειγματοληψίας

c	n=1	n=5	n=10	n=20	n=50	n=100	n=1000
0.1	0.72	1.57	2.36	3.41	4.89	5.77	6.94
0.2	0.96	2.36	3.41	4.55	5.77	6.36	7.02
0.3	1.18	2.95	4.08	5.15	6.15	6.59	7.05
0.4	1.38	3.41	4.55	5.52	6.36	6.71	7.06
0.5	1.57	3.78	4.89	5.77	6.50	6.78	7.07
0.6	1.75	4.08	5.15	5.95	6.59	6.83	7.07
0.7	1.92	4.33	5.35	6.09	6.66	6.87	7.08
0.8	2.07	4.55	5.52	6.20	6.71	6.90	7.08
0.9	2.22	4.73	5.66	6.29	6.75	6.92	7.08
1.0	2.36	4.89	5.77	6.36	6.78	6.94	7.08

Πίνακας 5.3. Αριθμός επιλεγμένων πακέτων αναφορικά με τις παραμέτρους c και n της επιλεκτικής δειγματοληψίας

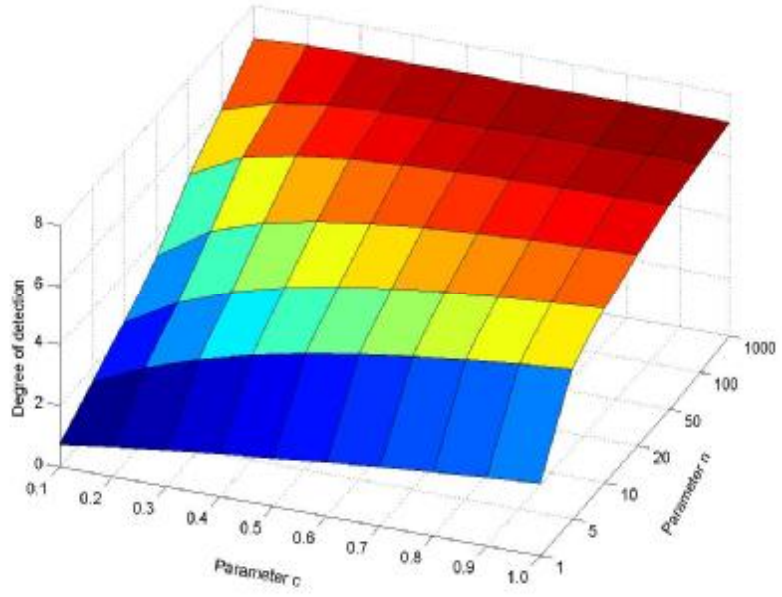
c	n=1	n=5	n=10	n=20	n=50	n=100	n=1000
0.1	9536	2471	1588	1147	882	793	714
0.2	10241	3176	2293	1852	1587	1499	1419
0.3	10946	3881	2998	2557	2292	2204	2124
0.4	11651	4587	3703	3262	2997	2909	2829
0.5	12356	5292	4409	3967	3702	3614	3534
0.6	13062	5997	5114	4672	4407	4319	4239
0.7	13767	6702	5819	5377	5112	5024	4945
0.8	14472	7407	6524	6082	5817	5729	5650
0.9	15177	8112	7229	6787	6523	6434	6355
1.0	15882	8817	7934	7493	7228	7139	7060

Πίνακας 5.4. Ποσοστό επιλεγμένων πακέτων αναφορικά με τις παραμέτρους c και n της επιλεκτικής δειγματοληψίας

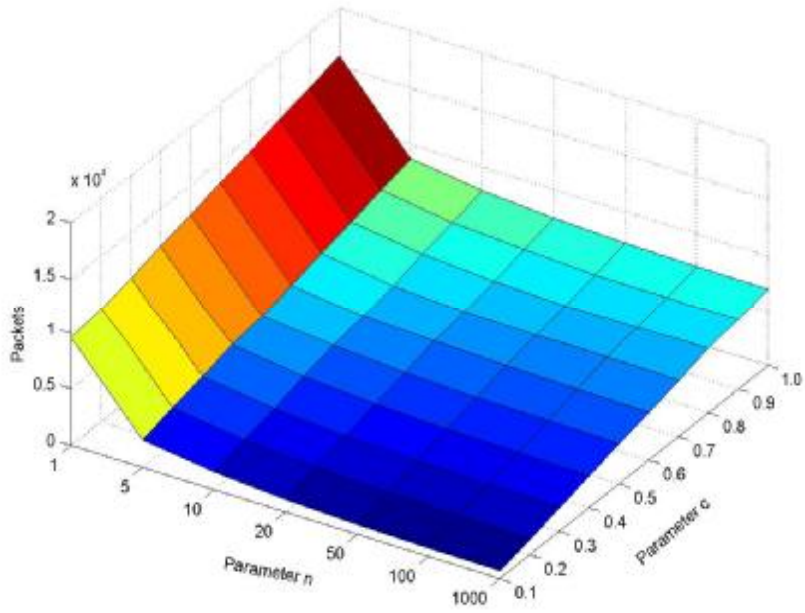
c	n=1	n=5	n=10	n=20	n=50	n=100	n=1000
0.1	4.89%	1.27%	0.81%	0.59%	0.45%	0.41%	0.37%
0.2	5.25%	1.63%	1.18%	0.95%	0.81%	0.77%	0.73%
0.3	5.61%	1.99%	1.54%	1.31%	1.18%	1.13%	1.09%
0.4	5.97%	2.35%	1.90%	1.67%	1.54%	1.49%	1.45%
0.5	6.34%	2.71%	2.26%	2.03%	1.90%	1.85%	1.81%
0.6	6.70%	3.08%	2.62%	2.40%	2.26%	2.21%	2.17%
0.7	7.06%	3.44%	2.98%	2.76%	2.62%	2.58%	2.54%
0.8	7.42%	3.80%	3.35%	3.12%	2.98%	2.94%	2.90%
0.9	7.78%	4.16%	3.71%	3.48%	3.35%	3.30%	3.26%
1.0	8.14%	4.52%	4.07%	3.84%	3.71%	3.66%	3.62%

Αναφερόμενοι στους Πίνακες 5.3 και 5.4, στην κάτω αριστερή γωνία των πινάκων απεικονίζεται ο αριθμός και το ποσοστό των επιλεγμένων πακέτων αντίστοιχα, για τις τιμές $c=1$ και $n=1$ της επιλεκτικής δειγματοληψίας. Αυτές οι τιμές αντιστοιχούν στην απλή περίπτωση της επιλεκτικής δειγματοληψίας που μοιάζει με την αντιστροφή της «έξυπνης δειγματοληψίας». Στην πάνω αριστερή γωνία, παρατηρούμε την περίπτωση όπου μόνο το 10% των μικρών ροών επιλέγονται, το οποίο οδηγεί σε μείωση των επιλεγμένων πακέτων. Στην κάτω δεξιά γωνία του πίνακα απεικονίζεται η περίπτωση στην οποία η πλειοψηφία των μεγάλων ροών δεν έχει επιλεγεί, ενώ την ίδια στιγμή όλες οι μικρές ροές έχουν επιλεγεί. Τέλος, στην πάνω δεξιά γωνία, μπορούμε να παρατηρήσουμε την περίπτωση όπου $c=0.1$ και $n=1000$. Αυτές οι τιμές αντιστοιχούν στην περίπτωση όπου έχουμε επιλέξει τον ελάχιστο αριθμό πακέτων. Όπως φαίνεται από τον Πίνακα 5.4, ο προτεινόμενος αλγόριθμος δειγματοληψίας παρέχει την ευελιξία της μείωσης του ποσοστού των επιλεγμένων πακέτων από 8,14% σε 0,37% με κατάλληλα καθορισμένες τιμές για τις παραμέτρους c και n.

Στο Σχήμα 5.7 παρουσιάζουμε γραφικά το βαθμό ανίχνευσης για τις ανωτέρω τιμές των παραμέτρων c και n. Όπως μπορούμε να παρατηρήσουμε από το γράφημα, ο βαθμός ανίχνευσης αυξάνεται καθώς η παράμετρος c μεγαλώνει, ενώ συγχρόνως διατηρούμε την παράμετρο n σταθερή, λόγω του γεγονότος ότι η πλειοψηφία των ροών της επίθεσης στο σενάριό μας έχει μόνο ένα πακέτο. Κατά συνέπεια, για μεγάλες τιμές του c, ο απόλυτος αριθμός πακέτων SYN είναι μεγαλύτερος, το οποίο οδηγεί σε μια μεγαλύτερη διαφορά μεταξύ του πλήθους των πακέτων SYN και FIN προκαλώντας την αύξηση του βαθμού ανίχνευσης. Αντιθέτως, ο βαθμός ανίχνευσης μειώνεται με τη μείωση της παραμέτρου n για σταθερές τιμές της παραμέτρου c. Αυτό αποδίδεται στο γεγονός ότι για τις μικρές τιμές του n τείνουμε να επιλέξουμε περισσότερες μεγάλες ροές. Οι μεγάλες ροές είναι πιθανό να περιέχουν ένα πακέτο SYN και ένα πακέτο FIN. Η επιλογή αυτών των ροών προκαλεί τη μείωση του πηλίκου της διαφοράς μεταξύ των πακέτων SYN και FIN με τον αριθμό πακέτων FIN, καθώς ο παρονομαστής αυξάνεται.



Σχήμα 5.7. Βαθμός ανίχνευσης ανωμαλίας σε σχέση με τις παραμέτρους c και n της επιλεκτικής δειγματοληψίας



Σχήμα 5.8. Αριθμός επιλεγέντων πακέτων σε σχέση με τις παραμέτρους c και n της επιλεκτικής δειγματοληψίας

Όπως έχει αναφερθεί προηγουμένως, στόχος μας είναι να επιτύχουμε έναν υψηλό βαθμό ανίχνευσης διατηρώντας παράλληλα έναν μικρό αριθμό πακέτων προς επεξεργασία. Ο αριθμός των επιλεγμένων πακέτων δίνεται από την σχέση (5.2). Στο

Σχήμα 5.8 παρουσιάζουμε τον αριθμό των επιλεγμένων πακέτων αναφορικά με τις παραμέτρους *c* και *n*. Όπως μπορούμε να παρατηρήσουμε, ο αριθμός επιλεγμένων πακέτων μειώνεται σημαντικά για μεγάλες τιμές της παραμέτρου *n* και μικρές τιμές της παραμέτρου *c*. Η μείωση των επιλεγμένων πακέτων είναι σημαντική από την περίπτωση *n=1* ως στην περίπτωση *n=5*, καθώς ένα αρκετά μεγάλο πλήθος μεγάλων ροών δεν επιλέγεται. Συνδυάζοντας τα αποτελέσματα των Σχημάτων 5.7 και 5.8, παρατηρούμε ότι για την περίπτωση που μελετάμε, οι βέλτιστες τιμές για τις παραμέτρους *n* και το *c* θα ήταν *n=1000* και *c=0.7*. Συγκεκριμένα, από το Σχήμα 5.7 (που αντιστοιχεί στον Πίνακα 5.2), παρατηρούμε ότι ο καλύτερος βαθμός ανίχνευσης δίνεται για τιμές *n=1000* και *c=0.7, 0.8, 0.9* και *1.0*. Το Σχήμα 5.8 (που αντιστοιχεί στους Πίνακες 5.3 και 5.4) απεικονίζει τον αριθμό των επιλεγμένων πακέτων για κάθε ζευγάρι των παραμέτρων *c* και *n*. Επομένως, οι τιμές *c=0.7* και *n=1000* παρέχουν τον καλύτερο βαθμό ανίχνευσης με τον ελάχιστο αριθμό επιλεγμένων πακέτων.

5.3. Εφαρμογή Επιλεκτικής Δειγματοληψίας στη μέθοδο Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis)

Σε αυτή την ενότητα εξετάζεται η αποτελεσματικότητα της επιλεκτικής δειγματοληψίας σε σχέση με την μέθοδο ανίχνευσης που είναι βασισμένη στην τεχνική PCA που περιγράψαμε σε προηγούμενη ενότητα, κάτω από διαφορετικά σενάρια επίθεσης/ανωμαλιών. Προκειμένου να αποκτηθεί κάποια γνώση σχετικά με τη λειτουργία και την επίδραση της επιλεκτικής δειγματοληψίας και να καταδειχθούν καλύτερα τα πλεονεκτήματά της, τη συγκρίνουμε με την τυχαία δειγματοληψία ροών. Στην τυχαία δειγματοληψία ροών, κάθε ροή επιλέγεται ανεξάρτητα με την ίδια πιθανότητα *p*. Το ποσοστό των επιλεγμένων ροών αποτελεί το κριτήριο για τη δίκαιη σύγκριση των δύο τεχνικών δειγματοληψίας.

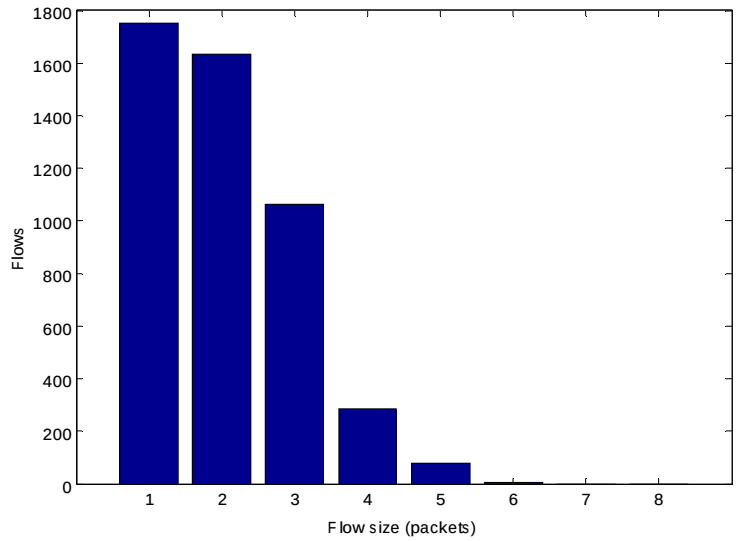
Όπως και προηγουμένως, τα αποτελέσματα και οι αντίστοιχες παρατηρήσεις που παρουσιάζονται σε αυτή την ενότητα είναι βασισμένα σε πραγματικά δεδομένα δικτύου που έχουν συλλεχθεί από τη σύνδεση μεταξύ του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ) και του Εθνικού Δικτύου Έρευνας και Τεχνολογίας (ΕΔΕΤ)

που συνδέει το ΕΜΠ με το Διαδίκτυο. Στην ακόλουθη αξιολόγηση, μελετάται λεπτομερώς μια κατανεμημένη επίθεση άρνησης υπηρεσίας (συγκεκριμένα επίθεση TCP SYN) ενάντια σε έναν υπολογιστή μέσα στο ΕΜΠ, η οποία παράγεται από ένα πραγματικό εργαλείο επίθεσης.

Όπως έχει προηγουμένως αναφερθεί, η μέθοδος Ανάλυσης Κύριων Συνιστωσών (PCA) μπορεί να λάβει υπόψη της ταυτόχρονα πολλαπλά μετρικά που προέρχονται από την ίδια σύνδεση δικτύου ή ακόμα και από πολλαπλές συνδέσεις (links). Στα πειράματά μας χρησιμοποιήσαμε εννέα μετρικά, ως εξής: *M1*: αριθμός ροών, *M2*: αριθμός πακέτων, *M3*: αριθμός ροών TCP, *M4*: αριθμός πακέτων TCP, *M5*: αριθμός ροών UDP, *M6*: αριθμός πακέτων UDP, *M7*: μικρές ροές (με έναν μικρό αριθμό πακέτων), *M8*: Αριθμός πακέτων SYN, και *M9*: Αριθμός πακέτων FIN.

Για να καταδείξουμε καλύτερα τα αποτελέσματα και τις αντίστοιχες παρατηρήσεις, μελετάμε δύο σενάρια. Στο πρώτο σενάριο, συγκρίνουμε την επιλεκτική δειγματοληψία με την τυχαία δειγματοληψία ροών και πειραματιζόμαστε με επιθέσεις που αντιστοιχούν στο 5% της ομαλής κίνησης του δικτύου (σε πακέτα), ενώ στο δεύτερο σενάριο το ποσοστό επίθεσης μειώνεται περαιτέρω για να αντιστοιχεί στο 2% της κίνησης του δικτύου.

Το Σχήμα 5.9 παρουσιάζει την κατανομή του μεγέθους των ροών επίθεσης στην περίπτωση του ποσοστού επίθεσης 5%. Από το σχήμα μπορούμε να παρατηρήσουμε ότι οι περισσότερες από τις ροές επίθεσης έχουν 1, 2 ή 3 πακέτα.



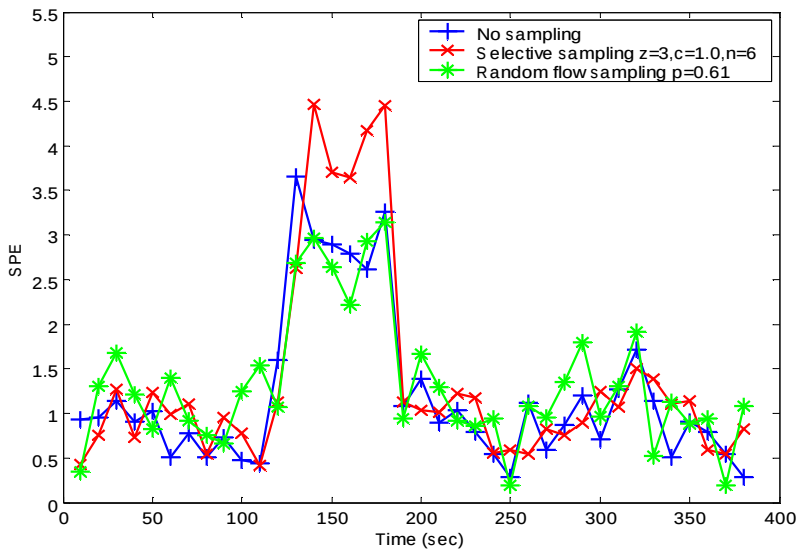
Σχήμα 5.9. Κατανομή του μεγέθους ροών επίθεσης για ποσοστό επίθεσης 5%

Με βάση το Σχήμα 5.9, επιλέγουμε μερικές χαρακτηριστικές τιμές για τις παραμέτρους της επιλεκτικής δειγματοληψίας, ενώ για την μέθοδο της τυχαίας δειγματοληψίας ροών επιλέγεται η κατάλληλη πιθανότητα *p* έτσι ώστε να οδηγεί στο ίδιο ποσοστό επιλεγμένων ροών. Ο Πίνακας 5.5 παρουσιάζει τις παραμέτρους και το αντίστοιχο ποσοστό των επιλεγέντων πακέτων και για τις δύο τεχνικές δειγματοληψίας. Οι δύο σειρές αυτού του πίνακα αντιστοιχούν σε δύο διαφορετικά ποσοστά δειγματοληψίας ροών (61% και 19%).

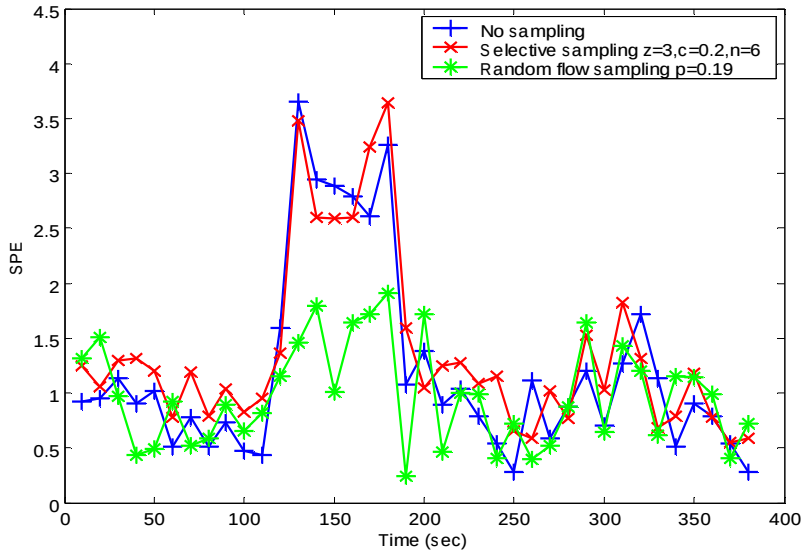
Πίνακας 5.5. Παράμετροι για τις τεχνικές δειγματοληψίας (Μέθοδος Ανίχνευσης PCA - ποσοστό επίθεσης 5%)

Flows (%)	Random Flow Sampling		Selective Sampling			
	p	*Packets (%)*	*z*	*c*	*n*	*Packets (%)*
61.22%	0.61	61.84%	3	1.0	6	7.18%
19.31%	0.19	19.57%	3	0.2	6	3.24%

Πρέπει να υπογραμμιστεί εδώ ότι το ποσοστό των επιλεγμένων πακέτων στην περίπτωση της επιλεκτικής δειγματοληψίας είναι σημαντικά μικρότερο από το αντίστοιχο για την περίπτωση της τυχαίας δειγματοληψίας ροών. Ο αριθμός επιλεγμένων πακέτων είναι μεγάλης σπουδαιότητας για τα μεγάλα δίκτυα, επειδή ο αλγόριθμος ανίχνευσης ανωμαλιών εξετάζει κάθε επιλεγμένο πακέτο προκειμένου να ελεγχθεί το πεδίο των σημαιών TCP (TCP flags).



Σχήμα 5.10. Ανίχνευση επίθεσης για δειγματοληψία ροών 61% σε ποσοστό επίθεσης 5%



Σχήμα 5.11. Ανίχνευση επίθεσης για δειγματοληψία ροών 19% σε ποσοστό επίθεσης 5%

Τα αποτελέσματα που απεικονίζονται στο Σχήμα 5.10 αντιστοιχούν στο ποσοστό επίθεσης 5% και σε ένα ποσοστό δειγματοληψίας 61% όσον αφορά τον αριθμό των ροών. Όπως μπορούμε να παρατηρήσουμε, η καμπύλη που αντιστοιχεί στην τυχαία δειγματοληψία ροών μοιάζει με την καμπύλη της περίπτωσης που δεν εφαρμόζεται δειγματοληψία, ενώ η καμπύλη στην περίπτωση της επιλεκτικής δειγματοληψίας ξεπερνά και τις άλλες δύο καμπύλες. Ομοίως, στο Σχήμα 5.11 παρουσιάζουμε τα αντίστοιχα αποτελέσματα για ποσοστό δειγματοληψίας 19% όσον αφορά τον αριθμό των επιλεγμένων ροών. Σε αυτήν την περίπτωση, η αποτελεσματικότητα ανίχνευσης έχει μειωθεί σημαντικά για την περίπτωση της τυχαίας δειγματοληψίας ροών, ενώ η περίπτωση της επιλεκτικής δειγματοληψίας μοιάζει με την αρχική περίπτωση (καθόλου δειγματοληψία), παρόλο που χρησιμοποιήθηκε μία μικρή τιμή c=0.2 για την επιλογή των μικρών ροών.

Όπως καταδεικνύεται από τα αποτελέσματα που παρουσιάζονται και στα δύο σχήματα, είναι προφανές ότι για το υπό εξέταση σενάριο η αποτελεσματικότητα ανίχνευσης είναι σημαντικά ανώτερη στην περίπτωση της επιλεκτικής δειγματοληψίας από ότι στην περίπτωση της τυχαίας δειγματοληψίας ροών. Η εξήγηση της διαφοράς στην επίδοση των δύο τεχνικών δειγματοληψίας βρίσκεται στο συσχετισμό των επιλεγμένων μετρικών. Συγκεκριμένα, η μέθοδος PCA απαιτεί τα μετρικά που θα τροφοδοτήσουν τον αλγόριθμο να παρουσιάζουν υψηλό βαθμό

συσχετισμού μεταξύ τους. Εάν δεν υπάρχει ισχυρός συσχετισμός μεταξύ των χρησιμοποιημένων μετρικών, η ακρίβεια του μοντέλου και επομένως η αποτελεσματικότητα του αλγορίθμου μειώνεται δραματικά. Η δοκιμή για τον καθορισμό εάν δύο μεταβλητές συσχετίζονται ή όχι είναι βασισμένη στον υπολογισμό του συντελεστή συσχετισμού τους, $R_{X,Y,}$ που δίνεται από τη σχέση (5.6). Μεταβλητές με συντελεστή συσχετισμού κοντά στην τιμή 1 διαφέρουν μεταξύ τους προς την ίδια κατεύθυνση, ενώ μεταβλητές με συντελεστή συσχετισμού κοντά στην τιμή -1 διαφέρουν μεταξύ τους προς αντίθετες κατευθύνσεις.

$$R_{X,Y} = \frac{Cov(X,Y)}{s_X \cdot s_Y} \qquad (5.6)$$

Ο Πίνακας 5.6 παρουσιάζει τον πίνακα συσχετισμού μεταξύ των χρησιμοποιημένων μετρικών για την περίπτωση που δεν εφαρμόζεται καθόλου δειγματοληψία. Όπως μπορούμε να παρατηρήσουμε από τον Πίνακα 5.6, τα μετρικά συσχετίζονται σημαντικά μεταξύ τους, καθώς ο πίνακας συσχετισμού τους περιλαμβάνει πολλά στοιχεία που έχουν τιμές κοντά στο 1. Στους πίνακες 5.7 και 5.8 παρουσιάζονται οι αντίστοιχες τιμές συσχετισμού των μετρικών για την επιλεκτική δειγματοληψία και την τυχαία δειγματοληψία ροών αντίστοιχα. Όπως παρατηρούμε από τις αντίστοιχες τιμές στον Πίνακα 5.7 ο συσχετισμός μεταξύ των μετρικών έχει αυξηθεί στις περισσότερες περιπτώσεις, με την εφαρμογή της επιλεκτικής δειγματοληψίας. Αντίθετα, ο συσχετισμός μεταξύ των μετρικών στην περίπτωση της τυχαίας δειγματοληψίας ροών έχει μειωθεί σημαντικά, όπως απεικονίζεται στον Πίνακα 5.8. Λαμβάνοντας υπόψη ότι η επιλεκτική δειγματοληψία εστιάζει στην επιλογή των μικρών ροών, μπορεί κανείς εύκολα να συνειδητοποιήσει ότι τα μετρικά που χρησιμοποιούνται στα πειράματα μας έχουν ένα μεγαλύτερο βαθμό συσχετισμού στις μικρές ροές. Το γεγονός αυτό οδηγεί στη βελτίωση της αποτελεσματικότητας ανίχνευσης ανωμαλιών όταν χρησιμοποιούνται αλγόριθμοι βασισμένοι στη μέθοδο PCA σε συνδυασμό με την επιλεκτική δειγματοληψία.

Στο δεύτερο σενάριο μειώσαμε το ποσοστό επίθεσης για να αντιστοιχεί στο 2% της κίνησης του δικτύου. Το Σχήμα 5.12 παρουσιάζει την κατανομή μεγέθους (σε πακέτα) των ροών επίθεσης. Όπως μπορούμε να παρατηρήσουμε οι περισσότερες από τις ροές επίθεσης έχουν 1 ή 2 πακέτα σε αυτήν την περίπτωση. Ομοίως, όπως και στο προηγούμενο σενάριο, βασιζόμενοι στο Σχήμα 5.12, επιλέγουμε μερικές

χαρακτηριστικές τιμές για τις παραμέτρους της επιλεκτικής δειγματοληψίας, ενώ για την τυχαία δειγματοληψία ροών, επιλέγεται η κατάλληλη πιθανότητα p έτσι ώστε να οδηγεί στο ίδιο ποσοστό των επιλεγμένων ροών.

Πίνακας 5.6. Πίνακας Συσχετισμού μεταξύ των μετρικών για την περίπτωση που δεν εφαρμόζεται δειγματοληψία

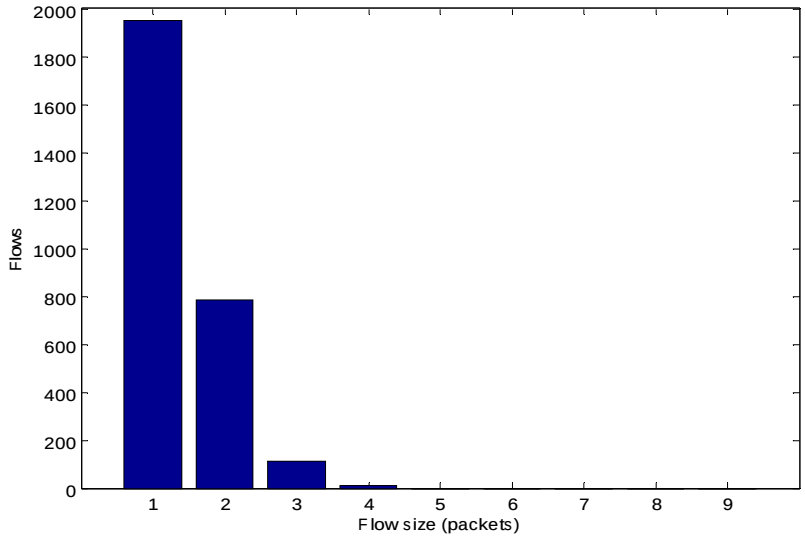
	M1	***M2***	***M3***	***M4***	***M5***	***M6***	***M7***	***M8***	***M9***
M1	1.00	0.75	0.77	0.21	0.68	0.46	0.88	0.29	0.48
M2	0.75	1.00	0.50	0.92	0.03	0.20	0.69	0.14	0.30
M3	0.77	0.50	1.00	0.31	-0.12	-0.03	0.31	0.53	0.36
M4	0.21	0.92	0.31	1.00	0.11	0.04	0.34	-0.08	0.47
M5	0.68	0.03	-0.12	0.11	1.00	0.59	0.69	0.19	0.02
M6	0.46	0.20	-0.03	0.04	0.59	1.00	0.64	0.27	-0.03
M7	0.88	0.69	0.31	0.34	0.69	0.64	1.00	0.34	0.14
M8	0.29	0.14	0.53	-0.08	0.19	0.27	0.34	1.00	0.30
M9	0.48	0.30	0.36	0.47	0.02	-0.03	0.14	0.30	1.00

Πίνακας 5.7. Πίνακας Συσχετισμού μεταξύ των μετρικών για την περίπτωση της επιλεκτικής δειγματοληψίας

	M1	***M2***	***M3***	***M4***	***M5***	***M6***	***M7***	***M8***	***M9***
M1	1.00	0.85	0.90	0.86	0.57	0.44	0.97	0.89	0.61
M2	0.85	1.00	0.84	0.97	0.18	0.13	0.81	0.13	0.26
M3	0.90	0.84	1.00	0.87	0.06	-0.10	0.48	0.79	0.30
M4	0.86	0.97	0.87	1.00	0.10	0.01	0.82	0.16	0.24
M5	0.57	0.18	0.06	0.10	1.00	0.61	0.49	0.07	0.12
M6	0.44	0.13	-0.10	0.01	0.61	1.00	-0.06	-0.10	0.14
M7	0.97	0.81	0.48	0.82	0.49	-0.06	1.00	0.48	0.27
M8	0.89	0.13	0.79	0.16	0.07	-0.10	0.48	1.00	0.30
M9	0.61	0.26	0.30	0.24	0.12	0.14	0.27	0.30	1.00

Πίνακας 5.8. Πίνακας Συσχετισμού μεταξύ των μετρικών για την περίπτωση της τυχαίας δειγματοληψίας ροών

	M1	***M2***	***M3***	***M4***	***M5***	***M6***	***M7***	***M8***	***M9***
M1	1.00	0.14	0.62	0.07	0.59	0.49	0.85	0.44	0.35
M2	0.14	1.00	0.38	0.90	-0.01	0.20	-0.19	0.09	0.20
M3	0.62	0.38	1.00	0.34	0.01	0.37	0.24	0.32	0.43
M4	0.07	0.90	0.34	1.00	-0.06	0.09	-0.25	0.08	0.16
M5	0.59	-0.01	0.01	-0.06	1.00	0.55	0.67	0.18	0.07
M6	0.49	0.20	0.37	0.09	0.55	1.00	0.32	0.05	0.29
M7	0.85	-0.19	0.24	-0.25	0.67	0.32	1.00	0.29	0.04
M8	0.44	0.09	0.32	0.08	0.18	0.05	0.29	1.00	0.28
M9	0.35	0.20	0.43	0.16	0.07	0.29	0.04	0.28	1.00

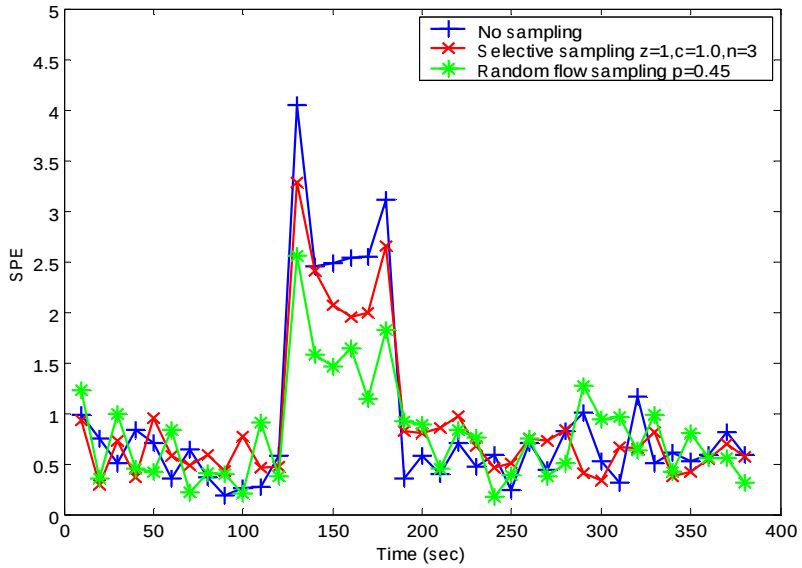


Σχήμα 5.12. Κατανομή του μεγέθους ροών επίθεσης για ποσοστό επίθεσης 2%

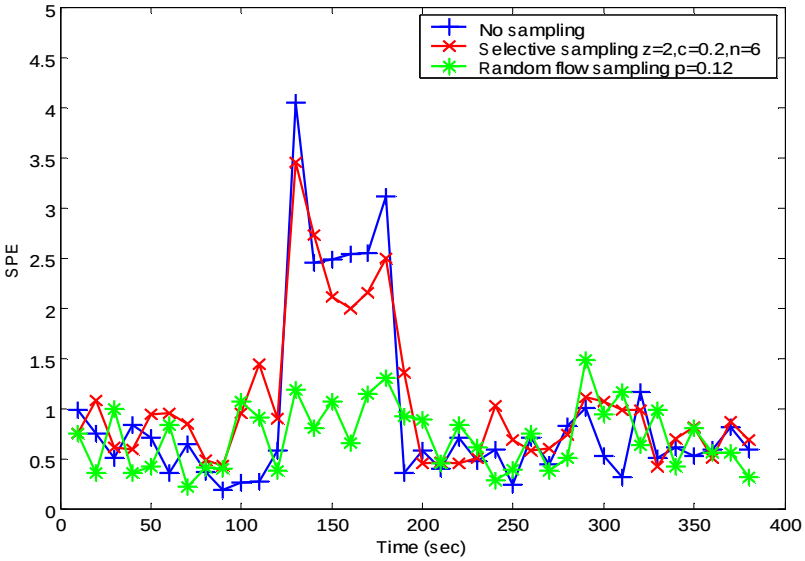
Ο Πίνακας 5.9 παρουσιάζει τις παραμέτρους και το αντίστοιχο ποσοστό των επιλεγμένων πακέτων για τις δύο τεχνικές δειγματοληψίας, για δύο διαφορετικά ποσοστά δειγματοληψίας ροών (δηλ. 45% και 12%).

Πίνακας 5.9. Παράμετροι για τις τεχνικές δειγματοληψίας (Μέθοδος Ανίχνευσης PCA - ποσοστό επίθεσης 2%)

Flows (%)	**Random Flow Sampling**		**Selective Sampling**			
	p	***Packets (%)***	***z***	***c***	***n***	***Packets (%)***
45.12%	0.45	45.67%	1	1.0	3	4.25%
12.05%	0.12	12.53%	2	0.2	6	2.05%



Σχήμα 5.13. Ανίχνευση επίθεσης για δειγματοληψία ροών 45% σε ποσοστό επίθεσης 2%



Σχήμα 5.14. Ανίχνευση επίθεσης για δειγματοληψία ροών 12% σε ποσοστό επίθεσης 2%

Στα σχήματα 5.13 και 5.14 παρουσιάζουμε τα αντίστοιχα αποτελέσματα για δύο διαφορετικά ποσοστά δειγματοληψίας ροών (45% και 12%) αντίστοιχα. Στο Σχήμα 5.13, και για τις δύο τεχνικές δειγματοληψίας ο αλγόριθμος ανίχνευσης ανωμαλιών επιτυγχάνει στην ανίχνευση της επίθεσης, με την επιλεκτική δειγματοληψία να συμπεριφέρεται ελαφρώς καλύτερα από την τυχαία δειγματοληψία ροών. Αντίθετα, για το ποσοστό δειγματοληψίας 12%, όπως μπορούμε να παρατηρήσουμε από το Σχήμα 5.14, ο αλγόριθμος ανίχνευσης ανωμαλιών αποτυγχάνει να ανιχνεύσει την

επίθεση στην περίπτωση της τυχαίας δειγματοληψίας ροών. Αυτό οφείλεται κυρίως στο γεγονός ότι ο συσχετισμός μεταξύ των μετρικών έχει μειωθεί σημαντικά όταν εφαρμόζεται η τυχαία δειγματοληψία ροών με μικρή πιθανότητα *p*. Από την άλλη πλευρά, η επιλεκτική δειγματοληψία επιτυγχάνει παρόμοια συμπεριφορά με την περίπτωση που δεν εφαρμόζεται δειγματοληψία, όσον αφορά την αποτελεσματικότητα ανίχνευσης, ενώ συγχρόνως ο αριθμός πακέτων που πρέπει να υποβληθούν σε επεξεργασία από τον αλγόριθμο ανίχνευσης έχει μειωθεί σημαντικά, βελτιώνοντας επομένως τη γενική απόδοση της διαδικασίας ανίχνευσης ανωμαλιών.

6. Η έννοια της δειγματοληψίας σε δύο στάδια (Two-Stage Sampling)

6.1. Εισαγωγή

Το κεφάλαιο αυτό επικεντρώνεται στην ανάλυση και την αξιολόγηση της επίδρασης τεχνικών δειγματοληψίας σε δύο στάδια στην ανίχνευση ανωμαλιών σε δίκτυα. Αρχικά, μελετάμε την αποτελεσματικότητα της τεχνικής δειγματοληψίας δύο σταδίων στη διαδικασία ανίχνευσης ανωμαλιών συγκρίνοντάς την με την τυχαία δειγματοληψία πακέτων. Στη συνέχεια, εκμεταλλευόμενοι το γεγονός ότι η επιλεγείσα κίνηση δικτύου είναι μια ελλιπής και ταυτόχρονα πολωμένη προσέγγιση του συνολικού δείγματος της κίνησης, προτείνουμε και αναλύουμε μια βελτιωμένη μέθοδο δειγματοληψίας δύο σταδίων, στην οποία γίνεται χρήση έξυπνης δειγματοληπτικής μεθόδου που εστιάζει στην επιλογή των μικρών ροών, οι οποίες όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο είναι συνήθως η πηγή κακόβουλης κίνησης στο δίκτυο [Barf01][Srid06].

Η αξιολόγηση της επίδρασης δειγματοληψίας δύο σταδίων στη διαδικασία ανίχνευσης ανωμαλιών πραγματοποιείται με την χρήση και την εφαρμογή μιας μεθόδου ανίχνευσης ανωμαλιών βασισμένη στην εντροπία σε πραγματικά δεδομένα δικτύου που έχουν συλλεχθεί από το περιβάλλον του ακαδημαϊκού δικτύου του ΕΜΠ. Τα αποτελέσματα καταδεικνύουν ότι η προτεινόμενη προσέγγιση βελτιώνει σε μεγάλο βαθμό την αποτελεσματικότητα της ανίχνευσης ανωμαλιών, ενώ συγχρόνως μειώνει τον αριθμό των επιλεγέντων δεδομένων, επιτυγχάνοντας στις περισσότερες περιπτώσεις να ξεπεράσει ακόμη και τα αντίστοιχα αποτελέσματα στην περίπτωση που δεν εφαρμόζεται κάποια μορφή δειγματοληψίας. Πρέπει να υπογραμμιστεί εδώ ότι ο στόχος της προτεινόμενης μεθόδου είναι να επιλεγούν τα κατάλληλα δεδομένα προκειμένου να μεγεθυνθούν οι ανωμαλίες, βελτιώνοντας κατά συνέπεια περαιτέρω την αποτελεσματικότητα της ανίχνευσης, και όχι η διατήρηση των στατιστικών ιδιοτήτων του αρχικού δείγματος της δικτυακής κίνησης.

6.2. Δειγματοληψία σε Δύο Στάδια (Two-Stage Sampling)

Στην ενότητα αυτή, περιγράφουμε αρχικά την τυχαία δειγματοληψία πακέτων και έπειτα τη δειγματοληψία σε δύο στάδια που χρησιμοποιεί τόσο μεθόδους δειγματοληψίας βασισμένη στις ροές όσο και μεθόδους δειγματοληψίας βασισμένες σε πακέτα. Να σημειώσουμε εδώ ότι η περιγραφή και ανάλυση της «Τυχαίας Δειγματοληψίας Πακέτων» (Random Packet Sampling) γίνεται με σκοπό να αποτελέσει τη βάση σύγκρισης για τις μεθόδους «Δειγματοληψίας Δύο Σταδίων» (Two-Stage Sampling).

6.2.1. Τυχαία Δειγματοληψία Πακέτων

Στην «Τυχαία Δειγματοληψία Πακέτων» (Random Packet Sampling - RPS) κάθε πακέτο επιλέγεται ανεξάρτητα με την ίδια πιθανότητα p. Πιο συγκεκριμένα, θεωρούμε την τυχαία μεταβλητή x να αντιπροσωπεύει το μέγεθος μιας ροής σε πακέτα. Στην περίπτωση της τυχαίας δειγματοληψίας πακέτων με την πιθανότητα p, μια ροή μεγέθους x επιλέγεται με πιθανότητα ίση με $1 - (1 - p)^x$. Κατά συνέπεια, ο αριθμός επιλεγμένων ροών N_f^{RPS} κατά τη διάρκεια της τυχαίας δειγματοληψίας πακέτων (RPS) δίνεται από την ακόλουθη σχέση:

$$N_f^{RPS} = \sum_{x=1}^{N} [1-(1-p)^x] \cdot \mathbf{1}(x) \cdot S_f \qquad (6.1)$$

Όπου $\mathbf{1}(x)$ είναι η συνάρτηση μάζας πιθανότητας (probability mass function - pmf) της τυχαίας μεταβλητής x,
S_f είναι ο συνολικός αριθμός των ροών και
N είναι το μέγιστο μέγεθος ροής σε πακέτα.

6.2.2. Δειγματοληψία σε Δύο Στάδια

Σε αυτή την ενότητα μελετάμε και αναλύουμε την έννοια της «Δειγματοληψίας Δύο Σταδίων» [Yang07], στην οποία η πρώτη φάση περιλαμβάνει δειγματοληψία ροών, ενώ στο δεύτερο στάδιο εφαρμόζεται δειγματοληψία πακέτων. Συγκεκριμένα,

στη πρώτη φάση οι ροές επιλέγονται τυχαία με πιθανότητα p_f ανεξάρτητα από το μέγεθός τους (σε πακέτα), ενώ στο δεύτερο στάδιο, τα πακέτα επιλέγονται τυχαία με πιθανότητα p_p από τις επιλεγείσες ροές της πρώτης φάσης. Εξαιτίας του γεγονότος ότι κάθε ροή έχει την ίδια πιθανότητα επιλογής, το επιλεγμένο σύνολο ροών διατηρεί το μέσο μήκος ροής (σε πακέτα) του αρχικού συνόλου. Κατά συνέπεια, το τελικό ποσοστό των επιλεγμένων πακέτων από αυτή τη μέθοδο δειγματοληψίας δύο σταδίων είναι ίσο με $p = p_f \cdot p_p$. Για την τεχνική «Δειγματοληψίας Δύο Σταδίων» (Two-Stage Sampling - TSS) ο αριθμός επιλεγμένων ροών N_f^{TSS} υπολογίζεται ως εξής. Μετά από τη πρώτη φάση της τυχαίας δειγματοληψίας ροών με πιθανότητα p_f ο αριθμός επιλεγμένων ροών δίνεται από την ακόλουθη σχέση:

$$N_{f1}^{TSS} = p_f \cdot S_f \qquad (6.2)$$

Ο τελικός αριθμός των επιλεγμένων ροών μετά το δεύτερο στάδιο της τυχαίας δειγματοληψίας πακέτων με πιθανότητα p_p είναι:

$$N_f^{TSS} = \sum_{x=1}^{N} [1-(1-p_p)^x] \cdot \mathbf{l}_1(x) \cdot N_{f1}^{TSS}$$

Όπου x παριστάνει το μέγεθος ροής σε πακέτα,

$\mathbf{l}_1(x)$ είναι η συνάρτηση μάζας πιθανότητας της τυχαίας μεταβλητής x για το δείγμα που έχει προκύψει μετά το πρώτο στάδιο δειγματοληψίας.

Με την χρήση της σχέσης (6.2), έχουμε:

$$N_f^{TSS} = \sum_{x=1}^{N} [1-(1-p_p)^x] \cdot \mathbf{l}_1(x) \cdot p_f \cdot S_f \qquad (6.3)$$

6.3. Επέκταση της μεθόδου Δειγματοληψίας Δύο Σταδίων για αποτελεσματικότερη ανίχνευση ανωμαλιών

Στη συνέχεια περιγράφουμε μια βελτιωμένη τεχνική δειγματοληψίας δύο σταδίων που έχει ως στόχο τη βελτίωση της αποτελεσματικότητας ανίχνευσης ανωμαλιών δικτύου. Συγκεκριμένα, κατά τη διάρκεια της πρώτης φάσης της δειγματοληψίας δύο σταδίων υιοθετούμε την «Επιλεκτική Δειγματοληψία» [Andr07][Andr08] η οποία επιλέγει με μεγαλύτερη προτίμηση τις μικρές ροές, αντί της τυχαίας δειγματοληψίας ροών. Το δεύτερο στάδιο της τυχαίας δειγματοληψίας πακέτων παραμένει το ίδιο.

Από εδώ και πέρα, θα αναφερόμαστε στην προτεινόμενη μέθοδο δειγματοληψίας ως «Επιλεκτική Δειγματοληψία Δύο Σταδίων» (Two-Stage Selective Sampling - TSSS). Πρέπει να υπογραμμιστεί εδώ ότι οι μικρές ροές είναι συνήθως η πηγή πολλών ανωμαλιών του διαδικτύου (DDoS, portscans, διάδοση worms) [Barf01][Srid06] και πρέπει να επιλεγεί προκειμένου να διατηρηθεί μια αποδοτική διαδικασία ανίχνευσης ανωμαλιών. Σύμφωνα με την «Επιλεκτική Δειγματοληψία», την οποία περιγράψαμε αναλυτικά στο προηγούμενο κεφάλαιο, η επιλογή μιας μεμονωμένης ροής είναι βασισμένη στην ακόλουθη έκφραση:

$$p(x) = \begin{cases} c & x \le z \\ \dfrac{z}{n \cdot x} & x > z \end{cases} \qquad (6.4)$$

όπου με x συμβολίζουμε το μέγεθος της ροής σε πακέτα, $0 < c \le 1$, $n \ge 1$ και z είναι ένα κατώφλι (μετρούμενο σε πακέτα).

Ας θεωρήσουμε τώρα την τυχαία μεταβλητή x να αντιπροσωπεύει το μέγεθος ροής σε πακέτα και $\mathbf{l}(x)$ την αντίστοιχη συνάρτηση μάζας πιθανότητας (probability mass function). Ο αριθμός των επιλεγμένων ροών για την περίπτωση της επιλεκτικής δειγματοληψίας N_f^{SS} δίνεται από την ακόλουθη σχέση:

$$N_f^{SS} = \sum_{x=1}^{N} p(x) \cdot \mathbf{l}(x) \cdot S_f$$

όπου S_f είναι ο συνολικός αριθμός των ροών και N είναι το μέγιστο μέγεθος ροής.

Με τη χρήση της σχέσης (6.4), ο αριθμός των επιλεγμένων ροών στο τέλος του πρώτου σταδίου δίνεται από την παρακάτω σχέση:

$$N_{f1}^{TSSS} = \sum_{x=1}^{z} c \cdot \mathbf{l}(x) \cdot S_f + \sum_{x=z+1}^{N} \frac{z}{n \cdot x} \cdot \mathbf{l}(x) \cdot S_f \qquad (6.5)$$

Ο τελικός αριθμός των επιλεγμένων ροών μετά το δεύτερο στάδιο της τυχαίας δειγματοληψίας πακέτων με πιθανότητα p_p είναι:

$$N_f^{TSSS} = \sum_{x=1}^{N} [1-(1-p_p)^x] \cdot \mathbf{l}_1(x) \cdot N_{f1}^{TSSS}$$

Όπου x παριστάνει το μέγεθος ροής σε πακέτα,

$\mathbf{l}_1(x)$ είναι η συνάρτηση μάζας πιθανότητας της τυχαίας μεταβλητής x για το δείγμα που έχει προκύψει μετά το πρώτο στάδιο δειγματοληψίας.

Με τη βοήθεια της σχέσης (6.5) παίρνουμε:

$$N_f^{TSSS} = \left[\sum_{x=1}^{N}[1-(1-p_p)^x]\cdot \mathbf{l}_1(x)\right]\cdot\left[\sum_{x=1}^{z} c\cdot \mathbf{l}(x) + \sum_{x=z+1}^{N}\frac{z}{n\cdot x}\cdot \mathbf{l}(x)\right]\cdot S_f \qquad (6.6)$$

6.4. Εφαρμογή Επιλεκτικής Δειγματοληψίας Δύο Σταδίων σε μέθοδο ανίχνευσης ανωμαλιών βασισμένη στην Εντροπία

Στη συνέχεια αυτού του κεφαλαίου αξιολογείται η αποτελεσματικότητα της τεχνικής Δειγματοληψίας Δύο Σταδίων, καθώς επίσης και της βελτιωμένης μεθόδου Επιλεκτικής Δειγματοληψίας Δύο Σταδίων, η οποία περιγράφηκε προηγουμένως, στην ανίχνευση ανωμαλιών δικτύου. Για λόγους επίδειξης, σε αυτή την μελέτη θα χρησιμοποιήσουμε μία μέθοδο ανίχνευσης ανωμαλιών βασισμένη στην εντροπία. Εδώ να σημειώσουμε ότι η μέθοδος είναι ενδεικτική και θα μπορούσε να χρησιμοποιηθεί οποιαδήποτε άλλη μέθοδος ανίχνευσης ανωμαλιών.

6.4.1. Ανίχνευση Ανωμαλιών με βάση την Εντροπία

Σε αυτή την ενότητα παρουσιάζουμε μια μέθοδο ανίχνευσης ανωμαλιών βασισμένη στην εντροπία, η οποία εντοπίζει τις ανωμαλίες δικτύου με την εξέταση διαφόρων κατανομών χαρακτηριστικών γνωρισμάτων κίνησης δικτύου, και αντιπροσωπεύει μια ευρεία κατηγορία συνήθως χρησιμοποιούμενων στρατηγικών ανίχνευσης ανωμαλιών.

Η Εντροπία $H(X)$ [Cove06] ενός συνόλου $X = \{x_1, x_2, \ldots x_n\}$ ορίζεται ως:

$$H(X) = -\sum_{i=1}^{N} p_i \log_2(p_i) \qquad (6.7)$$

όπου N είναι ο αριθμός στοιχείων που περιλαμβάνονται στο σύνολο στοιχείων X και p_i είναι η πιθανότητα $P[X = x_i]$. Η εντροπία μετράει πόσο τυχαία είναι κατανεμημένο ένα σύνολο στοιχείων. Οι υψηλές τιμές εντροπίας δηλώνουν μια

διασκορπισμένη κατανομή πιθανότητας, ενώ οι χαμηλές τιμές εντροπίας δείχνουν τη συγκέντρωση μιας κατανομής στοιχείων. Οι τιμές της εντροπίας, όπως αυτή ορίζεται στη σχέση (6.7), κυμαίνονται μεταξύ 0 και $\log_2 N$. Προκειμένου να υπάρξει ένα μετρικό ανεξάρτητο του αριθμού των διαφορετικών τιμών του συνόλου, ομαλοποιούμε την εντροπία διαιρώντας την τιμή $H(X)$ με τη μέγιστη τιμή της, η οποία είναι $\log_2 N$. Η τιμής της ομαλοποιημένης εντροπίας δίνεται από την σχέση (6.8):

$$H_n(X) = -\frac{\sum_{i=1}^{N} p_i \log_2(p_i)}{\log_2 N} \qquad (6.8)$$

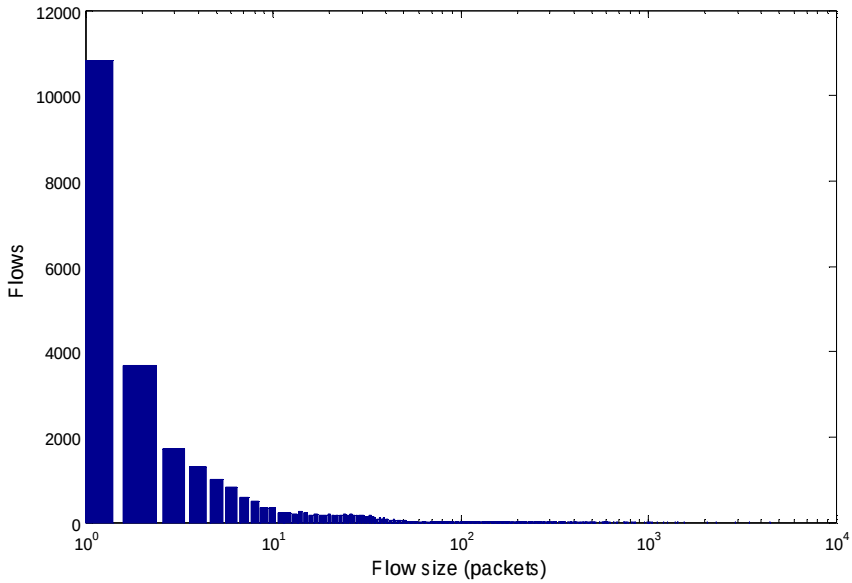
και οι τιμές της κυμαίνονται μεταξύ (0,1).

Η εντροπία έχει χρησιμοποιηθεί εκτενώς σε μεθόδους ανίχνευσης ανωμαλιών [Ranj07][Yu06][Lakh05][Xu05b]. Μερικές από τις συνήθεις χρησιμοποιούμενες κατανομές χαρακτηριστικών γνωρισμάτων δικτυακής κίνησης, οι οποίες είναι πολύτιμες στην ανίχνευση ανωμαλιών δικτύων είναι οι κατανομές IP διευθύνσεων πηγής (srcIP), IP διευθύνσεων προορισμού (dstIP), θύρα πηγής (srcPort) και θύρα προορισμού (dstPort). Παραδείγματος χάριν, μια ανωμαλία όπως ένας μολυσμένος υπολογιστής που προσπαθεί να μολύνει και άλλους υπολογιστές στο Διαδίκτυο (διάδοση worm) οδηγεί στη μείωση της εντροπίας των IP διευθύνσεων πηγής. Ο μολυσμένος υπολογιστής παράγει ένα μεγάλο αριθμό ροών αναγκάζοντας την ίδια IP διεύθυνση πηγής να κυριαρχεί στην κατανομή των IP διευθύνσεων πηγής. Αντίθετα, σε ένα παράδειγμα ανωμαλίας portscan, η εντροπία της θύρας προορισμού αυξάνεται λόγω της σάρωσης (scanning) τυχαίων θυρών προορισμού, ενώ συγχρόνως η εντροπία των IP διευθύνσεων πηγής και προορισμού μειώνεται, επειδή οι διευθύνσεις IP του επιτιθεμένου και του θύματος κυριαρχούν στις αντίστοιχες κατανομές. Στην περίπτωση μιας κατανεμημένες επίθεσης DoS, όπου οι πολλαπλές πηγές επιτιθέμενων υπολογιστών στέλνουν τα πακέτα σε ένα μόνο θύμα (συνήθως προς μία συγκεκριμένη θύρα), έχει ως αποτέλεσμα την παραγωγή ενός τεράστιου αριθμού πακέτων με τη συγκεκριμένη IP διεύθυνση προορισμού (διεύθυνση θύματος) και θύρα προορισμού, προκαλώντας στις αντίστοιχες τιμές εντροπίας σημαντική μείωση. Στη συγκεκριμένη μελέτη, εστιάζουμε στις κατανομές πιθανότητας της IP διεύθυνσης πηγής και θύρας προορισμού και στις αντίστοιχες ομαλοποιημένες τιμές εντροπίας

που υπολογίζονται με την εξέταση των ροών κατά τη διάρκεια ενός χρονικού παραθύρου (10 sec).

6.4.2. Πλαίσιο Αξιολόγησης Επιλεκτικής Δειγματοληψίας Δύο Σταδίων με βάση την Εντροπία

Τα αποτελέσματα και οι αντίστοιχες παρατηρήσεις που παρουσιάζονται σε αυτή την ενότητα είναι βασισμένα σε ένα αρχείο κίνησης δικτύου που έχει καταγραφεί σε ένα ακαδημαϊκό δίκτυο. Πιο συγκεκριμένα, μελετήσαμε τη σύνδεση μεταξύ του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ) και του Εθνικού Δικτύου Έρευνας και Τεχνολογίας (ΕΔΕΤ) [GRNet] που συνδέει το ΕΜΠ με το Διαδίκτυο. Στο διάστημα των πειραμάτων μας, αυτή η σύνδεση είχε μια μέση κίνηση της τάξης των 200Mbit/sec, και περιείχε ένα πλούσιο μίγμα κίνησης αποτελούμενο από κίνηση web, ηλεκτρονικού ταχυδρομείου, FTP καθώς επίσης και p2p κίνηση. Στο Σχήμα 6.1 παρουσιάζεται η κατανομή του μεγέθους (σε πακέτα) των ροών του δείγματος.

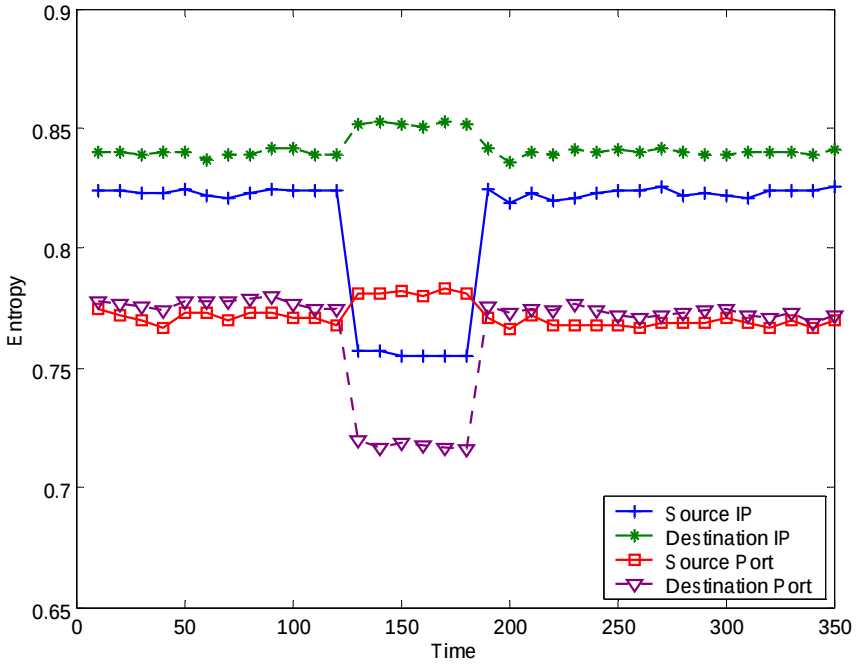


Σχήμα 6.1. Κατανομή μεγέθους ροών του δείγματος

Στην ακόλουθη αξιολόγηση, μελετάμε ως περίπτωση ανίχνευσης ανωμαλίας, ένα σενάριο διάδοσης ενός αυτοδιαδιδόμενου ιού (worm) και συγκεκριμένα το Slammer worm [Moor03]. Το Slammer worm είναι ο γρηγορότερος αυτοδιαδιδόμενος ιός στην ιστορία των δικτύων υπολογιστών, μολύνοντας περισσότερους από το 90% των

τρωτών υπολογιστών μέσα σε διάστημα 10 λεπτών. Η μεγάλη ταχύτητα στην διάδοσή του οφείλεται στο γεγονός ότι η φάση διάδοσής του περιλαμβάνει μόνο ένα πακέτο UDP μεγέθους 404 bytes που στοχεύει στην UDP θύρα 1434. Το worm αυτό έκανε την εμφάνισή του τον Ιανουαρίου του 2003 και εκμεταλλευόταν μια ευπάθεια στο λογισμικό Microsoft SQL Server [MSSQL].

Στα παρακάτω πειράματα, εισάγουμε ένα ποσοστό κίνησης του παραπάνω αυτοδιαδιδόμενου ιού στο καταγεγραμμένο αρχείο κίνησης, κατά τη διάρκεια του διαστήματος 120-180 sec. Πιο συγκεκριμένα, θεωρούμε έναν υπολογιστή μέσα στο ΕΜΠ να έχει μολυνθεί από το Slammer worm. Αυτός ο μολυσμένος υπολογιστής παράγει ροές διάδοσης του ιού που αντιστοιχούν στο 15% του συνολικού αριθμού των ροών της φυσιολογικής κίνησης και 1% του συνολικού αριθμού των πακέτων ανά χρονικό παράθυρο (στην περίπτωση που μελετάμε το παράθυρο είναι 10sec). Ο μολυσμένος υπολογιστής παράγει τις ροές που διαδίδουν τον ιό χρησιμοποιώντας μόνο ένα πακέτο UDP ανά IP διεύθυνση προορισμού, όπου η IP διεύθυνση προορισμού επιλέγεται τυχαία. Η θύρα πηγής (source port) κάθε πακέτου UDP επιλέγεται επίσης τυχαία, και κυμαίνεται από 1 έως 65535.



Σχήμα 6.2. Τιμές Εντροπίας για την ανωμαλία που εισάγαμε στο αρχικό δείγμα

Το Σχήμα 6.2 παρουσιάζει την ανωμαλία που εισάγαμε όπως φαίνεται μέσα από το φακό της εντροπίας. Όπως μπορούμε να παρατηρήσουμε, οι τιμές της εντροπίας για την IP διεύθυνση πηγής (source IP address) και την θύρα προορισμού (destination port) μειώνονται σημαντικά κατά τη διάρκεια του διαστήματος ανωμαλίας, ενώ η εντροπία για την IP διεύθυνση προορισμού (destination IP address) και την θύρα πηγής (source port) ελαφρώς αυξάνεται. Η μείωση στις τιμές εντροπίας της IP διεύθυνσης πηγής και την θύρα προορισμού οφείλεται στο γεγονός ότι μια συγκεκριμένη IP διεύθυνση (ο μολυσμένος υπολογιστής) και μία συγκεκριμένη θύρα (στην περίπτωσή μας η UDP θύρα 1434), εμφανίζονται πολλαπλές φορές μέσα σε ένα χρονικό παράθυρο, προκαλώντας έτσι τη συγκέντρωση των αντίστοιχων κατανομών γύρω από ένα συγκεκριμένο στοιχείο. Αντίθετα, οι τιμές εντροπίας για την IP διεύθυνση πηγής και την θύρα προορισμού δεν παρουσιάζουν σημαντική αλλαγή επειδή η δραστηριότητα της διάδοσης του ιού δεν προκαλεί σημαντική αλλαγή στις αντίστοιχες κατανομές. Αυτό οφείλεται στο γεγονός ότι αυτές οι κατανομές περιλαμβάνουν αρκετή «τυχαιότητα» στις φυσιολογικές συνθήκες του δικτύου, οπότε η πρόσθετη τυχαία σάρωση των IP διευθύνσεων χρησιμοποιώντας τυχαίες θύρες πηγής δεν αλλάζει σημαντικά αυτές τις κατανομές. Για αυτό τον λόγο, στα πειράματά μας εστιάζουμε στις τιμές εντροπίας της IP διεύθυνσης πηγής και της θύρας προορισμού.

Προκειμένου να αποκτηθεί σημαντική γνώση σχετικά με την αποτελεσματικότητα της τεχνικής Δειγματοληψίας Δύο Σταδίων στην ανίχνευση αυτής της ανωμαλίας, την συγκρίνουμε αρχικά ενάντια στην Τυχαία Δειγματοληψία Πακέτων (Random Packet Sampling). Κατόπιν, καταδεικνύουμε την πρόσθετη αποτελεσματικότητα της προτεινόμενης Επιλεκτικής Δειγματοληψίας Δύο Σταδίων. Το ποσοστό των επιλεγέντων πακέτων επιλέχτηκε ως κοινό κριτήριο για τη δίκαιη σύγκριση των τεχνικών δειγματοληψίας που προαναφέραμε.

6.4.3. Πειραματικά Αποτελέσματα

Για να καταδείξουμε καλύτερα τα αποτελέσματα και τις αντίστοιχες παρατηρήσεις, μελετάμε δύο σενάρια. Στο πρώτο σενάριο συγκρίνουμε την Επιλεκτική Δειγματοληψία Δύο Σταδίων (TSSS), τη Δειγματοληψία Δύο Σταδίων (TSS) και την Τυχαία Δειγματοληψία Πακέτων (RPS) πειραματιζόμενοι με μια

ανωμαλία (αυτοδιαδιδόμενος ιός) που αντιστοιχεί στο 15% της κανονικής κίνησης του δικτύου (μετρούμενη σε αριθμό ροών) με ένα ποσοστό δειγματοληψίας 10% (σε αριθμό πακέτων), ενώ στο δεύτερο σενάριο μειώνουμε περαιτέρω το ποσοστό δειγματοληψίας για να αντιστοιχεί στο 1% του συνολικού αριθμού των πακέτων.

Προκειμένου να γίνει κατανοητή η επίδραση της Δειγματοληψίας Δύο Σταδίων στην ανίχνευση ανωμαλιών δικτύου επιλέγουμε μερικές χαρακτηριστικές τιμές για τις πιθανότητες p_f και p_p της Δειγματοληψίας Δύο Σταδίων. Για την τεχνική της Τυχαίας Δειγματοληψίας Πακέτων επιλέγουμε την κατάλληλη πιθανότητα p που οδηγεί στο ίδιο ποσοστό των επιλεγμένων πακέτων. Ομοίως, για την τεχνική της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων επιλέγουμε μερικές χαρακτηριστικές τιμές για τις παραμέτρους z, c και n της Επιλεκτικής Δειγματοληψίας, και την κατάλληλη τιμή για την παράμετρο p_p στην οποία ο συνολικός αριθμός των επιλεγμένων πακέτων είναι ίσος με τον αντίστοιχο αριθμό των άλλων δύο τεχνικών δειγματοληψίας. Πιο συγκεκριμένα, υπολογίζουμε το ποσοστό των επιλεγμένων πακέτων στο τέλος της πρώτης φάσης της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων. Το ποσοστό των επιλεγμένων πακέτων μετά από τη πρώτη φάση δίνεται από την ακόλουθη σχέση:

$$p_1 = \frac{N_p}{S_p} \qquad (6.9)$$

όπου N_p είναι ο αριθμός επιλεγμένων πακέτων κατά τη διάρκεια της πρώτης φάσης δειγματοληψίας και S_p είναι ο συνολικός αριθμός των πακέτων στην αρχική περίπτωση (χωρίς δειγματοληψία). Ο αριθμός επιλεγμένων πακέτων στην πρώτη φάση (μέθοδος Επιλεκτικής Δειγματοληψίας) N_p δίνεται από την ακόλουθη σχέση:

$$N_p = \sum_{x=1}^{N} p(x) \cdot \mathbf{l}(x) \cdot x \cdot S_f$$

όπου x αναπαριστά το μέγεθος της ροής σε πακέτα,

$\mathbf{l}(x)$ είναι η συνάρτηση μάζας πιθανότητας της τυχαίας μεταβλητής x,

S_f είναι ο συνολικός αριθμός των ροών και

N είναι το μέγιστο μέγεθος ροής.

Χρησιμοποιώντας τη σχέση (6.4) έχουμε:

$$N_p = \sum_{x=1}^{z} c \cdot \mathbf{l}(x) \cdot x \cdot S_f + \sum_{x=z+1}^{N} \frac{z}{n} \cdot \mathbf{l}(x) \cdot S_f$$

Οπότε η σχέση (5.9) γίνεται:

$$p_1 = \left[\sum_{x=1}^{z} c \cdot \mathbf{l}(x) \cdot x + \sum_{x=z+1}^{N} \frac{z}{n} \cdot \mathbf{l}(x) \cdot x \right] \cdot \frac{S_f}{S_p}$$

Προκειμένου να επιτευχθεί το ίδιο ποσοστό επιλεγμένων πακέτων με τις άλλες δύο τεχνικές δειγματοληψίας, στο δεύτερο στάδιο επιλέγουμε τα πακέτα με πιθανότητα p_p, η οποία δίνεται από την ακόλουθη σχέση:

$$p_p = \frac{p}{p_1}, \quad p \le p_1$$

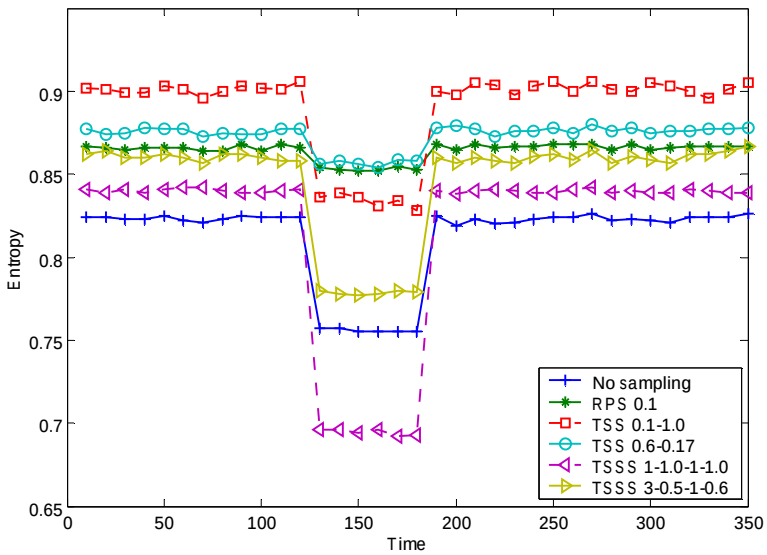
Ο Πίνακας 6.1 παρουσιάζει τις διάφορες παραμέτρους όλων των τεχνικών δειγματοληψίας που μελετάμε, οι οποίες οδηγούν σε ποσοστό δειγματοληψίας περίπου 10% (σε πακέτα). Συγκεκριμένα, για την περίπτωση της Δειγματοληψίας Δύο Σταδίων πειραματιζόμαστε με τις τιμές 0.1 και 0.6 για την παράμετρο p_f προκειμένου να μελετήσουμε πως επηρεάζει η δειγματοληψία ροών τη διαδικασία ανίχνευσης ανωμαλιών στο δίκτυο. Επιπλέον, στην πρώτη περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων όπου *z=1, c=1.0* και *n=1* επιλέγουμε τις κατάλληλες τιμές για να επιτύχουμε την επιλογή όλων των ροών της ανωμαλίας (του αυτοδιαδιδόμενου ιού στο σενάριο που μελετάμε), ενώ στη δεύτερη περίπτωση, όπου *z=3, c=0.5* και *n=1,* επιλέγουμε μια γενικότερη προσέγγιση για τη ανίχνευση των ανωμαλιών που αποτελούνται από μικρές ροές (συγκεκριμένα ροές με 3 ή λιγότερα πακέτα).

Πίνακας 6.1. Παράμετροι για την κάθε τεχνική δειγματοληψίας (ποσοστό δειγματοληψίας σε πακέτα 10%)

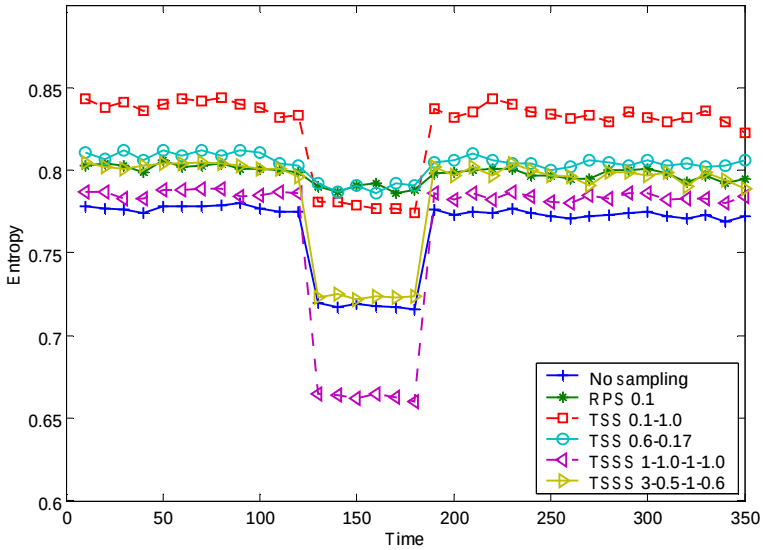
Sampling Technique	**Parameters**			
Random packet sampling	***p***		0.1	
Two-stage sampling	p_f		p_p	
	0.1		1.00	
	0.6		0.17	
Two-stage Selective sampling	***z***	***c***	***n***	p_p
	1	1.0	1	1.0
	3	0.5	1	0.6

Τα αποτελέσματα που απεικονίζονται στο Σχήμα 6.3 αντιστοιχούν στις ομαλοποιημένες τιμές εντροπίας της IP διεύθυνσης πηγής για ποσοστό δειγματοληψίας 10%, όσον αφορά τον αριθμό πακέτων. Όπως μπορούμε να παρατηρήσουμε, στην περίπτωση της Δειγματοληψίας Δύο Σταδίων (TSS), όπου $p_f=0.1$ και $p_p=1.0$ η ανωμαλία ανιχνεύεται εύκολα, αφού η εντροπία ελαττώνεται από 0.90 σε 0.83. Αν αυξήσουμε την πιθανότητα p_f της δειγματοληψίας ροών παρατηρούμε ότι η εντροπία δεν παρουσιάζει σημαντική αλλαγή κατά τη διάρκεια της ανωμαλίας. Όπως παρατηρούμε στην περίπτωση όπου $p_f=0.6$ και $p_p=0.17$ η εντροπία IP διευθύνσεων πηγής μειώνεται ελαφρώς από 0.88 σε 0.86. Η ίδια συμπεριφορά εμφανίζεται και για την Τυχαία Δειγματοληψία Πακέτων (RPS).

Στην Επιλεκτική Δειγματοληψία Δύο Σταδίων (TSSS) όπου $z=1$, $c=1.0$ και $n=1$ η ανωμαλία ανιχνεύεται πολύ εύκολα από την ελάττωση της εντροπίας από 0.84 σε 0.69, ξεπερνώντας ακόμα και την περίπτωση όπου δεν εφαρμόζεται καθόλου δειγματοληψία. Αυτή η συμπεριφορά αποδίδεται στο γεγονός ότι κατά τη διάρκεια της πρώτης φάσης δειγματοληψίας έχουν επιλεγεί όλες οι ροές της ανωμαλίας, ενώ ταυτόχρονα ένα μεγάλο ποσοστό των κανονικών ροών (μεγάλου αριθμού πακέτων) έχει απορριφθεί. Κατά αυτόν τον τρόπο, η διεύθυνση IP του μολυσμένου υπολογιστή κυριαρχεί στην κατανομή των IP διευθύνσεων πηγής. Στη δεύτερη περίπτωση, όπου $z=3$, $c=0.5$ και $n=1$ η ανωμαλία ανιχνεύεται πάλι, αλλά σε έναν χαμηλότερο βαθμό, αφού έχουν επιλεγεί περισσότερες κανονικές ροές και λιγότερες ροές ανωμαλίας.



Σχήμα 6.3. Εντροπία για τις IP διευθύνσεις πηγής (ποσοστό δειγματοληψίας 10%)



Σχήμα 6.4. Εντροπία για τις θύρες προορισμού (ποσοστό δειγματοληψίας 10%)

Στο Σχήμα 6.4 παρουσιάζουμε τα αντίστοιχα αποτελέσματα για τις τιμές εντροπίας της θύρας προορισμού. Στην περίπτωση της Δειγματοληψίας Δύο Σταδίων (TSS) με τη μικρότερη πιθανότητα p_f παρατηρούμε ότι η ομαλοποιημένη εντροπία μειώνεται από 0.84 σε 0.78, ανιχνεύοντας έτσι αποτελεσματικά την ανωμαλία του αυτοδιαδιδόμενου ιού. Όπως επισημάνθηκε και προηγουμένως, για μεγαλύτερες τιμές της πιθανότητας p_f παρατηρείται μια μικρότερη αλλαγή στην καμπύλη της εντροπίας κατά τη διάρκεια της ανωμαλίας. Η ανωμαλία είναι σχεδόν αξιοπρόσεχτη όταν αλλάζουμε σε πιθανότητα p_f=*0.6,* στην οποία η τιμή της εντροπίας θύρας προορισμού διαφέρει κατά 0.02, πέφτοντας από 0.81 έως 0.79. Παρόμοια με την περίπτωση εντροπίας για τις IP διευθύνσεις πηγής, η ομαλοποιημένη εντροπία για την Τυχαία Δειγματοληψία Πακέτων δεν παρουσιάζει σημαντική αλλαγή κατά τη διάρκεια της περιόδου δραστηριότητας του αυτοδιαδιδόμενου ιού.

Στην πρώτη περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων (TSSS) παρατηρούμε ότι η τιμή της εντροπίας μειώνεται από 0.79 σε 0.66, ανιχνεύοντας έτσι αποτελεσματικά την ανωμαλία. Ακολουθώντας την ίδια τάση, όπως στην προηγούμενη περίπτωση της εντροπίας των IP διευθύνσεων πηγής, η εντροπία για τις θύρες προορισμού, ξεπερνά σε απόδοση (ως προς την ανίχνευση της ανωμαλίας) την

περίπτωση που δεν εφαρμόζεται δειγματοληψία. Να σημειώσουμε εδώ ότι όλες οι καμπύλες στις περιπτώσεις δειγματοληψίας παρουσιάζουν μια αυξανόμενη τιμή εντροπίας κάτω από φυσιολογικές συνθήκες δικτύου σε σχέση με την περίπτωση που δεν εφαρμόζεται δειγματοληψία. Αυτό οφείλεται στο γεγονός ότι ένας μειωμένος αριθμός ροών παράγει πιο ομοιόμορφες κατανομές IP διευθύνσεων πηγής και θυρών προορισμού, καθώς επιλέγονται διαφορετικά στοιχεία.

Οι ανωτέρω παρατηρήσεις συνοψίζονται στον Πίνακα 6.2, ο οποίος παρουσιάζει την μεταβολή (σε % ποσοστό) στις τιμές της εντροπίας για τις IP διευθύνσεις πηγής (SrcIP) και τις θύρες προορισμού (DstPort) για κάθε τεχνική δειγματοληψίας. Η αποτελεσματικότητα της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων στη διαδικασία ανίχνευσης ανωμαλιών καταδεικνύεται με σαφήνεια από τα αποτελέσματα του πίνακα. Συγκεκριμένα, παρατηρείται αύξηση της μεταβολής της εντροπίας IP διευθύνσεων πηγής από 8.54% στην αρχική περίπτωση σε 17.86% στην περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων, ενώ η εντροπία των θυρών προορισμού αυξάνεται από 7.79% σε 16.45%. Μια σημαντική μεταβολή στην εντροπία IP διευθύνσεων πηγής και την εντροπία θυρών προορισμού παρατηρείται για τις μικρές τιμές της πιθανότητας p_f για την τεχνική της Δειγματοληψίας Δύο Σταδίων (TSS), αλλά η μεταβολή αυτή είναι σχετικά μικρότερη από την αρχική περίπτωση που δεν εφαρμόζεται καθόλου δειγματοληψία. Τέλος, η εντροπία για την περίπτωση της Τυχαίας Δειγματοληψίας Πακέτων (RPS) δεν φαίνεται να επηρεάζεται σχεδόν καθόλου (παρατηρείται μόνο μια αλλαγή περίπου 1%).

Πίνακας 6.2. Μεταβολή της Εντροπίας για τις IP διευθύνσεις πηγής (SrcIP) και τις θύρες προορισμού (DstPort) για ποσοστό δειγματοληψίας 10%

Sampling Technique	**Parameters**				**Change in Src IP entropy**	**Change in Dst Port entropy**
No sampling					8.54%	7.79%
Random packet sampling	***p***		0.1		1.16%	1.25%
Two-stage sampling	p_f		p_p			
	0.1		1.00		7.78%	7.14%
	0.6		0.17		2.30%	2.36%
Two-stage Selective sampling	***z***	***c***	***n***	p_p		
	1	1.0	1	1.0	17.86%	16.45%
	3	0.5	1	0.6	9.30%	9.87%

6.4.4. Ανάλυση

Σε αυτή την ενότητα εξηγούμε και τεκμηριώνουμε λεπτομερώς τα ανωτέρω αποτελέσματα βασισμένα στην ανάλυση που παρέχεται στις ενότητες 6.2 και 6.3. Συγκεκριμένα, υπολογίζουμε το ποσοστό των ροών της ανωμαλίας στις συνολικές επιλεγείσες ροές για κάθε μια από τις προαναφερθείσες τεχνικές δειγματοληψίας.

Όπως αναφέραμε σε προηγούμενη ενότητα, η ανωμαλία του αυτοδιαδιδόμενου ιού στην περίπτωσή μας, αποτελείται από ροές με ένα μόνο πακέτο UDP. Κατά συνέπεια, ο αριθμός επιλεγμένων ροών N_{wf}^{RPS} της ανωμαλίας για την περίπτωη της Τυχαία Δειγματοληψίας Πακέτων (RPS) δίνεται από την ακόλουθη σχέση:

$$N_{wf}^{RPS} = p \cdot S_{wf} \quad (6.10)$$

όπου S_{wf} είναι ο συνολικός αριθμός των ροών της ανωμαλίας.

Για την τεχνική της Δειγματοληψίας Δύο Σταδίων (TSS) ο αντίστοιχος αριθμός επιλεγμένων ροών ανωμαλίας υπολογίζεται ως εξής: Στο τέλος της πρώτης φάσης της τυχαίας δειγματοληψίας ροών με πιθανότητα p_f λαμβάνουμε το πλήθος των ροών της ανωμαλίας N_{wf1}^{TSS}, που υπολογίζεται ως εξής:

$$N_{wf1}^{TSS} = p_f \cdot S_{wf} \quad (6.11)$$

Μετά από το τέλος του δεύτερου σταδίου της τυχαίας δειγματοληψίας πακέτων με πιθανότητα p_p, ο τελικός αριθμός επιλεγμένων ροών ανωμαλίας για την περίπτωση της Δειγματοληψίας Δύο Σταδίων (TSS) δίνεται από τη σχέση (6.12):

$$N_{wf}^{TSS} = p_p \cdot N_{wf1}^{TSS} \quad (6.12)$$

Με την χρήση της σχέσης (6.11) παίρνουμε:

$$N_{wf}^{TSS} = p_f \cdot p_p \cdot S_{wf} \quad (6.13)$$

Για την τεχνική της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων (TSSS) ο αντίστοιχος αριθμός επιλεγμένων ροών ανωμαλίας υπολογίζεται ως εξής: Στο τέλος

της πρώτης φάσης της επιλεκτικής δειγματοληψίας ροών λαμβάνουμε N_{wf}^{TSSS} ροές ανωμαλίας, που υπολογίζονται ως εξής:

$$N_{wf1}^{TSSS} = c \cdot S_{wf} \qquad (6.14)$$

όπου S_{wf} είναι ο συνολικός αριθμός των ροών της ανωμαλίας. Μετά από το τέλος του δεύτερου σταδίου της τυχαίας δειγματοληψίας πακέτων με πιθανότητα p_p, ο τελικός αριθμός επιλεγμένων ροών ανωμαλίας για την περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων δίνεται από τη σχέση (6.15):

$$N_{wf}^{TSSS} = p_p \cdot N_{wf1}^{TSSS} \qquad (6.15)$$

Με την βοήθεια της σχέσης (6.14), έχουμε:

$$N_{wf}^{TSSS} = c \cdot p_p \cdot S_{wf} \qquad (6.16)$$

Χρησιμοποιώντας τις σχέσεις (6.1) και (6.10) λαμβάνουμε το ποσοστό των ροών ανωμαλίας στις συνολικές επιλεγείσες ροές για την περίπτωση της Τυχαίας Δειγματοληψίας Πακέτων (RPS), η οποία δίνεται από τη σχέση (6.17):

$$\frac{N_{wf}^{RPS}}{N_f^{RPS}} = \frac{p \cdot S_{wf}}{\sum_{x=1}^{N} [1-(1-p)^x] \cdot \mathbf{l}(x) \cdot S_f} \qquad (6.17)$$

Ομοίως, το αντίστοιχο ποσοστό για την περίπτωση της Δειγματοληψίας Δύο Σταδίων (TSS) χρησιμοποιώντας τις σχέσεις (6.3) και (6.13), δίνεται από την παρακάτω σχέση:

$$\frac{N_{wf}^{TSS}}{N_f^{TSS}} = \frac{p_f \cdot p_p \cdot S_{wf}}{\sum_{x=1}^{N} [1-(1-p_p)^x] \cdot \mathbf{l}_1(x) \cdot p_f \cdot S_f} \qquad (6.18)$$

Τέλος, το ποσοστό των ροών ανωμαλίας στις συνολικές επιλεγείσες ροές για την περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων (TSSS) λαμβάνεται με το συνδυασμό των σχέσεων (6.7) και (6.16):

$$\frac{N_{wf}^{TSSS}}{N_{f}^{TSSS}} = \frac{c \cdot p_p \cdot S_{wf}}{\left[\sum_{x=1}^{N}[1-(1-p_p)^x]\cdot \mathbf{l}_1(x)\right]\cdot\left[\sum_{x=1}^{z} c\cdot \mathbf{l}(x) + \sum_{x=z+1}^{N}\frac{z}{n\cdot x}\cdot \mathbf{l}(x)\right]\cdot S_f} \quad (6.19)$$

Πίνακας 6.3. Ποσοστά επιλεγμένων ροών για κάθε μία από τις τεχνικές δειγματοληψίας για ποσοστό δειγματοληψίας 10%

Sampling Technique	**Parameters**		**Percentage of sampled worm flows**	**Percentage of total sampled flows**	**Percentage of worm flows in total sampled flows**
No sampling			100.00%	100.00%	15.06%
Random packet sampling	***P***	0.1	10.00%	35.31%	4.26%
Two-stage sampling	p_f	p_p			
	0.1	1.00	10.00%	10.22%	14.74%
	0.6	0.17	10.00%	27.18%	5.54%
Two-stage Selective sampling	***z / c / n / p_p***				
	1/ 1.0 /1 / 1.0		100.00%	53.57%	28.11%
	3/ 0.5 /1 / 0.6		30.00%	32.86%	13.75%

Ο Πίνακας 6.3 παρουσιάζει το ποσοστό των ροών ανωμαλίας, το ποσοστό των συνολικών επιλεγμένων ροών και το ποσοστό των ροών ανωμαλίας στις συνολικές επιλεγείσες ροές για κάθε τεχνική δειγματοληψίας, όπως υπολογίζεται από τις παραπάνω σχέσεις. Όπως μπορούμε να παρατηρήσουμε από τον Πίνακα 6.3 το ποσοστό των επιλεγμένων ροών ανωμαλίας είναι το ίδιο για όλες τις περιπτώσεις της Δειγματοληψίας Δύο Σταδίων και της Τυχαίας Δειγματοληψίας Πακέτων. Με την παρατήρηση των σχέσεων (6.10) και (6.13) καταλήγουμε στη διαπίστωση ότι κάτι τέτοιο είναι αναμενόμενο, αφού $p = p_f \cdot p_p$. Αντίθετα, το ποσοστό των συνολικών επιλεγμένων ροών αυξάνεται με την αύξηση της πιθανότητας p_f για την περίπτωση της Δειγματοληψίας Δύο Σταδίων. Κατά συνέπεια, το ποσοστό των ροών ανωμαλίας στις συνολικές επιλεγείσες ροές είναι μεγαλύτερο από τις αντίστοιχες για τις μικρές τιμές της πιθανότητας p_f. Το υψηλότερο ποσοστό των ροών ανωμαλίας στις συνολικές ροές οδηγεί σε μια συγκεντρωμένη κατανομή πιθανότητας, αναγκάζοντας την εντροπία να μειωθεί περαιτέρω κατά τη διάρκεια της ανωμαλίας. Επιπλέον, το ποσοστό των επιλεγμένων ροών ανωμαλίας στην περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων είναι μεγαλύτερο από τις άλλες δύο τεχνικές, με

συνέπεια να έχουμε ένα σημαντικά αυξημένο ποσοστό των ροών ανωμαλίας στις συνολικές επιλεγείσες ροές. Όπως μπορούμε να παρατηρήσουμε από τον ανωτέρω πίνακα, στην πρώτη περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων το ποσοστό των επιλεγμένων ροών ανωμαλίας είναι όμοιο με την περίπτωση που δεν εφαρμόζεται δειγματοληψία. Το γεγονός ότι η διαδικασία δειγματοληψίας μειώνει το ποσοστό των επιλεγμένων ροών, αναγκάζει το ποσοστό των ροών ανωμαλίας στις συνολικές επιλεγείσες ροές να αυξηθεί σε 28.11% έναντι 15.06% στην περίπτωση χωρίς δειγματοληψία. Τα ανωτέρω θεωρητικά αποτελέσματα που προκύπτουν από τους μαθηματικούς τύπους που περιγράψαμε σε αυτή την ενότητα επιβεβαιώνονται επίσης και από τα πειραματικά αποτελέσματα.

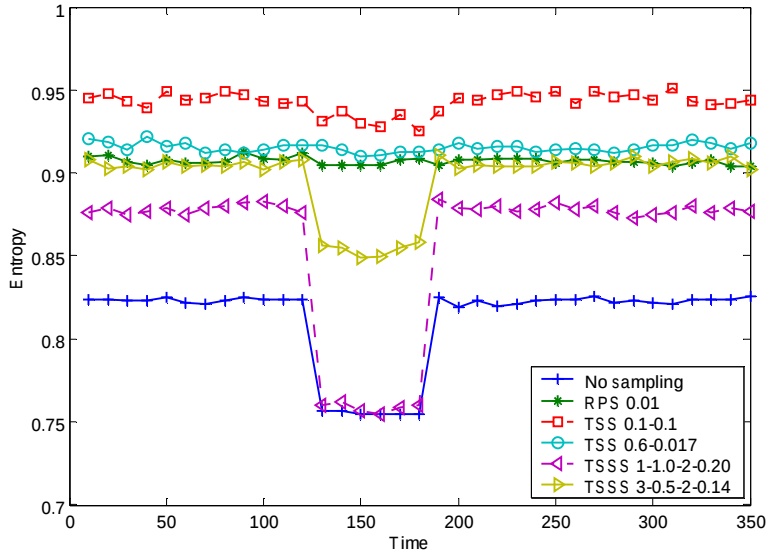
6.4.5. Μείωση του ρυθμού δειγματοληψίας

Στη συνέχεια, μειώσαμε το ποσοστό δειγματοληψίας ώστε να αντιστοιχεί στο 1% των συνολικών πακέτων, διατηρώντας το ίδιο ποσοστό της ανωμαλίας όπως στα προηγούμενα πειράματα. Για αυτό το ποσοστό δειγματοληψίας, εφαρμόσαμε τη Δειγματοληψία Δύο Σταδίων με τις ίδιες τιμές για την παράμετρο p_f όπως στην προηγούμενη περίπτωση του ποσοστού δειγματοληψίας 10%, και τις κατάλληλες τιμές για την παράμετρο p_p ώστε να προκύψει δείγμα πακέτων σε ποσοστό 1% από τα συνολικά πακέτα. Για την περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων, διατηρούμε τις ίδιες τιμές για τις παραμέτρους z και c, και αυξάνουμε την τιμή της παραμέτρου n από $n=1$ σε $n=2$ προκειμένου να μειωθεί σημαντικά ο αριθμός των επιλεγμένων πακέτων. Όπως αναφέραμε σε προηγούμενο κεφάλαιο, η τιμή της παραμέτρου n καθορίζει την επιλογή των μεγάλων ροών. Μεγαλύτερες τιμές της παραμέτρου n οδηγούν στην επιλογή μικρότερου αριθμού μεγάλων ροών και συνεπώς λιγότερων πακέτων. Τέλος, με την κατάλληλη τιμή για την παράμετρο p_p της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων μειώνουμε περαιτέρω το ποσοστό των επιλεγέντων πακέτων για να φτάσουμε σε ποσοστό 1% των συνολικών πακέτων. Οι αντίστοιχες παράμετροι για κάθε τεχνική δειγματοληψίας που οδηγούν στο ποσοστό δειγματοληψίας 1% (σε πακέτα) συνοψίζονται στον Πίνακα 6.4.

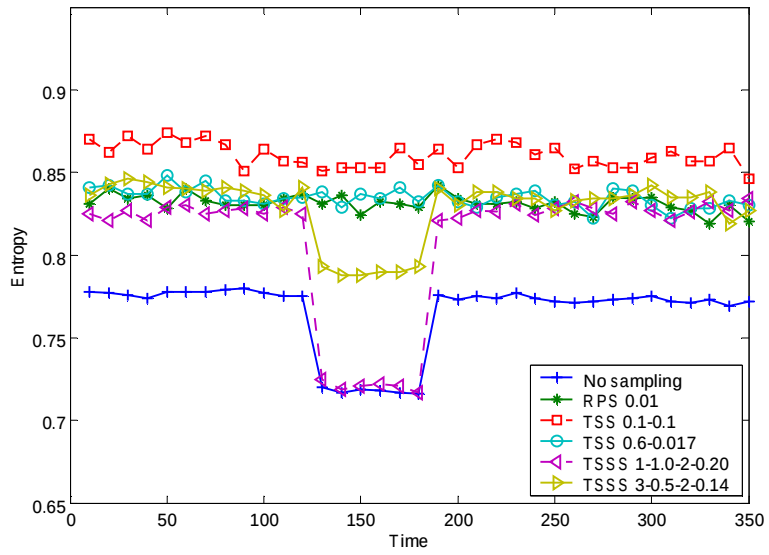
Πίνακας 6.4. Παράμετροι για την κάθε τεχνική δειγματοληψίας (ποσοστό δειγματοληψίας σε πακέτα 1%)

Sampling Technique	**Parameters**			
Random packet sampling	p		0.01	
Two-stage sampling	p_f		p_p	
	0.1		0.100	
	0.6		0.017	
Two-stage Selective sampling	z	c	n	p_p
	1	1.0	2	0.20
	3	0.5	2	0.14

Τα αντίστοιχα αποτελέσματα για το ποσοστό δειγματοληψίας 1% απεικονίζονται στα Σχήματα 6.5 και 6.6. Πιο συγκεκριμένα, στο Σχήμα 6.5 παρουσιάζουμε τις ομαλοποιημένες τιμές εντροπίας για τις IP διευθύνσεις πηγής σε ποσοστό δειγματοληψίας 1%. Μελετώντας το Σχήμα 6.5, παρατηρούμε ότι η Τυχαία Δειγματοληψία Πακέτων και η Δειγματοληψία Δύο Σταδίων, για όλες τις τιμές των παραμέτρων, παρουσιάζουν σχεδόν την ίδια συμπεριφορά. Καμία από αυτές τις δύο δειγματοληπτικές μεθόδους δεν επιτυγχάνει την ανίχνευση της ανωμαλίας του αυτοδιαδιδόμενου ιού. Αντίθετα, στη μέθοδο της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων όπου $z=1$, $c=1.0$ και $n=2$ η ανωμαλία ανιχνεύεται επιτυχώς ως αποτέλεσμα της μείωσης της εντροπίας από 0.88 σε 0.75, ξεπερνώντας ακόμη και την αρχική περίπτωση που δεν εφαρμόζεται δειγματοληψία. Αυτή η συμπεριφορά αποδίδεται στο γεγονός ότι κατά τη διάρκεια της πρώτης φάσης δειγματοληψίας έχουν επιλεγεί όλες οι ροές της ανωμαλίας και ταυτόχρονα ένα μεγάλο ποσοστό των κανονικών ροών (μεγάλου μεγέθους) έχει απορριφθεί. Στη δεύτερη περίπτωση όπου $z=3$, $c=0.5$ και $n=2$ η ανωμαλία ανιχνεύεται πάλι, αλλά σε έναν χαμηλότερο βαθμό, αφού έχουν επιλεγεί περισσότερες κανονικές ροές και λιγότερες ροές ανωμαλίας.



Σχήμα 6.5. Εντροπία για τις IP διευθύνσεις πηγής (ποσοστό δειγματοληψίας 1%)



Σχήμα 6.6. Εντροπία για τις θύρες προορισμού (ποσοστό δειγματοληψίας 1%)

Στο Σχήμα 6.6 παρουσιάζουμε τις ομαλοποιημένες τιμές εντροπίας για τις θύρες προορισμού για το ίδιο ποσοστό δειγματοληψίας. Όπως μπορούμε να παρατηρήσουμε, όπως και με την εντροπία των IP διευθύνσεων πηγής, η εντροπία για τις θύρες προορισμού (ως μετρικό) αποτυγχάνει να ανιχνεύσει την ανωμαλία για τις δύο πρώτες τεχνικές δειγματοληψίας (Τυχαία Δειγματοληψία Πακέτων και

Δειγματοληψία Δύο Σταδίων). Αντίθετα, στην πρώτη περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων παρατηρούμε ότι η ομαλοποιημένη εντροπία μειώνεται από 0.83 σε 0.71, γεγονός που έχει ως συνέπεια την αποτελεσματική ανίχνευση της ανωμαλίας.

Οι ανωτέρω παρατηρήσεις συνοψίζονται στον Πίνακα 6.5, ο οποίος παρουσιάζει τη μεταβολή (σε ποσοστό %) στις τιμές εντροπίας των IP διευθύνσεων πηγής και των θυρών προορισμού για κάθε τεχνική δειγματοληψίας. Στις περισσότερες από τις περιπτώσεις για τις μεθόδους της Τυχαίας Δειγματοληψίας Πακέτων και της Δειγματοληψίας Δύο Σταδίων, η μεταβολή στις τιμές της εντροπίας είναι λιγότερο από 1%, καθιστώντας έτσι την ανωμαλία σχεδόν μη ανιχνεύσιμη. Αντίθετα, η Επιλεκτική Δειγματοληψία Δύο Σταδίων πετυχαίνει αποτελεσματικά την ανίχνευση της ανωμαλίας, ξεπερνώντας ακόμη και την αρχική περίπτωση (καθόλου δειγματοληψία). Χαρακτηριστικά, η μεταβολή για την τιμή της εντροπίας των IP διευθύνσεων πηγής είναι από 8.54% στην αρχική περίπτωση σε 14.77% στην πρώτη περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων, ενώ η εντροπία για τις θύρες προορισμού αυξάνεται από 7.79% σε 14.46%.

Πίνακας 6.5. Μεταβολή της Εντροπίας για τις IP διευθύνσεις πηγής (SrcIP) και τις θύρες προορισμού (DstPort) για ποσοστό δειγματοληψίας 1%

Sampling Technique	**Parameters**				**Change in Src IP entropy**	**Change in Dst Port entropy**
No sampling					8.54%	7.79%
Random packet sampling	***p***		0.01		0.33%	0.47%
Two-stage sampling	p_f		p_p			
	0.1		0.100		1.06%	1.16%
	0.6		0.017		0.44%	0.56%
Two-stage Selective sampling	***z***	***c***	***n***	p_p		
	1	1.0	2	0.20	14.77%	14.46%
	3	0.5	2	0.14	5.56%	5.95%

Στον Πίνακα 6.6, παρουσιάζουμε το ποσοστό των ροών ανωμαλίας, το ποσοστό των συνολικών επιλεγμένων ροών και το ποσοστό των ροών ανωμαλίας στις συνολικές επιλεγείσες ροές και για τις τρεις τεχνικές δειγματοληψίας που έχουν υπολογιστεί από τις αναλυτικές εκφράσεις της προηγούμενης ενότητας, για ποσοστό δειγματοληψίας 1% (σε πακέτα). Παρατηρώντας τον Πίνακα 6.6, καταδεικνύεται η

ανικανότητα της μεθόδου της Τυχαίας Δειγματοληψίας Πακέτων και της Δειγματοληψίας Δύο Σταδίων για την ανίχνευση της ανωμαλίας, δεδομένου ότι το ποσοστό των ροών ανωμαλίας στις συνολικές επιλεγείσες ροές δεν είναι σημαντικό (στις περισσότερες περιπτώσεις κάτω από 3%). Αντιθέτως, το ποσοστό των επιλεγμένων ροών ανωμαλίας είναι σημαντικά μεγαλύτερο και για τις δύο περιπτώσεις της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων. Αυτό οφείλεται στο γεγονός ότι κατά τη διάρκεια της πρώτης φάσης της συγκεκριμένης μεθόδου οι ροές ανωμαλίας επιλέγονται με μεγαλύτερη προτίμηση. Κατά συνέπεια, το ποσοστό των ροών ανωμαλίας στις συνολικές επιλεγείσες ροές γίνεται μεγαλύτερο σε αυτήν την περίπτωση, με συνέπεια να οδηγούμαστε σε μια αποτελεσματικότερη ανίχνευση της ανωμαλίας. Πρέπει να σημειωθεί εδώ ότι στην πρώτη περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων το ποσοστό των ροών της ανωμαλίας στις συνολικές επιλεγείσες ροές έχει αυξηθεί από 15% σε 26.77%.

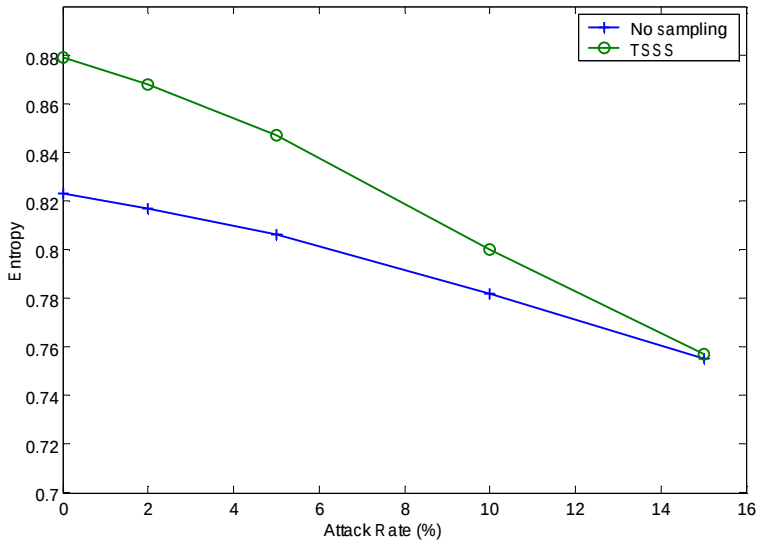
Πίνακας 6.6. Ποσοστά επιλεγμένων ροών για κάθε μία από τις τεχνικές δειγματοληψίας για ποσοστό δειγματοληψίας 1%

Sampling Technique	**Parameters**		**Percentage of sampled worm flows**	**Percentage of total sampled flows**	**Percentage of worm flows in total sampled flows**
No sampling			100.00%	100.00%	15.06%
Random packet sampling	p	0.01	1.00%	9.92%	1.52%
Two-stage sampling	p_f	p_p			
	0.1	1.000	1.00%	3.93%	3.83%
	0.6	0.017	1.00%	8.39%	1.79%
Two-stage Selective sampling	z / c / n / p_p				
	1/ 1.0 /2 / 0.20		20.00%	11.25%	26.77%
	3/ 0.5 /2 / 0.14		7.00%	9.11%	11.57%

6.4.6. Μείωση του ρυθμού επίθεσης

Τέλος, για να καταδείξουμε την αποτελεσματικότητα της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων, πειραματιστήκαμε με μικρότερα ποσοστά ανωμαλίας. Πιο συγκεκριμένα, μειώσαμε το ποσοστό επίθεσης της ανωμαλίας του

αυτοδιαδιδόμενου ιού από 15% όσον αφορά το συνολικό αριθμό των ροών, σε 10%, 5% και 2% του συνολικού αριθμού των ροών, ενώ διατηρήσαμε ένα ποσοστό δειγματοληψίας 1% (σε πακέτα). Στο Σχήμα 6.7 συγκρίνουμε την εντροπία IP διευθύνσεων πηγής της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων (που χρησιμοποιεί τις παραμέτρους $z=1$, $c=1.0$, $n=2$ και $p_p=0.20$) με την αντίστοιχη απόδοση της αρχικής περίπτωσης που δεν εφαρμόζεται δειγματοληψία. Μελετώντας τα αποτελέσματα αυτού του σχήματος, παρατηρούμε ότι η κλίση της καμπύλης στην περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων είναι πιο απότομη από την αντίστοιχη καμπύλη της αρχικής περίπτωσης, επιβεβαιώνοντας ουσιαστικά ότι η Επιλεκτική Δειγματοληψία Δύο Σταδίων ξεπερνά στην ανίχνευση την αρχική περίπτωση, ακόμη και για μικρότερα ποσοστά ανωμαλίας.



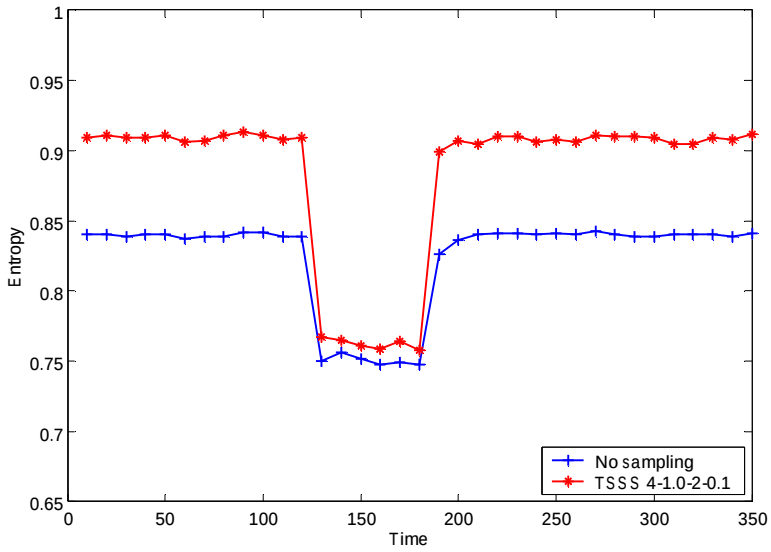
Σχήμα 6.7. Εντροπία IP διευθύνσεων πηγής για διαφορετικά ποσοστά ανωμαλίας

6.4.7. Πειράματα με διαφορετικά σενάρια ανωμαλιών

Η προτεινόμενη τεχνική της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων μπορεί να είναι αποτελεσματική με ποικίλες ανωμαλίες που προκαλούνται από μικρές ροές. Παρακάτω, αξιολογούμε την αποτελεσματικότητα ανίχνευσης της προτεινόμενης τεχνικής δειγματοληψίας σε ένα σενάριο μιας κατανεμημένης επίθεσης άρνησης

υπηρεσίας (DDoS attack). Συγκεκριμένα, θεωρούμε 100 επιτιθεμένους υπολογιστές που στέλνουν πακέτα TCP SYN σε έναν συγκεκριμένο υπολογιστή (θύμα) που βρίσκεται μέσα στο δίκτυο του ΕΜΠ στοχεύοντας την θύρα TCP 80. Οι ροές επίθεσης αποτελούνται από 1 έως 4 πακέτα ανά χρονικό παράθυρο (στην περίπτωσή μας είναι 10sec). Η δικτυακή κίνηση της συνολικής επίθεσης αντιστοιχεί σε 10% της κανονικής κίνησης του δικτύου (η οποία μετριέται σε αριθμό ροών ανά χρονικό παράθυρο).

Για τη μέθοδο της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων, επιλέγουμε τις ακόλουθες τιμές για τις παραμέτρους: $z=4$, $c=1.0$ και $n=2$ προκειμένου να επιτευχθεί η επιλογή όλων των ροών που είναι μέρος της DDoS επίθεσης. Τέλος, επιλέγουμε $p_p=0.1$ για να μειώσουμε τον αριθμό επιλεγμένων πακέτων ώστε να αντιστοιχηθεί σε ποσοστό 1% του συνολικού αριθμού των πακέτων.



Σχήμα 6.8. Εντροπία για τις IP διευθύνσεις προορισμού για σενάριο DDoS επίθεσης (ποσοστό δειγματοληψίας 1%)

Όπως περιγράψαμε νωρίτερα στην ενότητα 5.4.1, σε ένα σενάριο DDoS επίθεσης όπου ένας υπολογιστής-θύμα λαμβάνει πακέτα από έναν μεγάλο αριθμό επιτιθεμένων, η εντροπία των IP διευθύνσεων προορισμού αλλάζει σημαντικά. Στο Σχήμα 6.8 συγκρίνουμε τις τιμές εντροπίας των IP διευθύνσεων προορισμού για τη μέθοδο της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων και τη συγκρίνουμε με την

αρχική περίπτωση στην οποία δεν εφαρμόζεται δειγματοληψία. Όπως μπορούμε να παρατηρήσουμε, στην αρχική περίπτωση η εντροπία μειώνεται από 0.84 σε 0.75, ενώ στην περίπτωση της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων παρατηρούμε ότι η εντροπία μειώνεται από 0.91 σε 0.76, μεγεθύνοντας την ανωμαλία DDoS και οδηγώντας επομένως σε αποτελεσματική ανίχνευση.

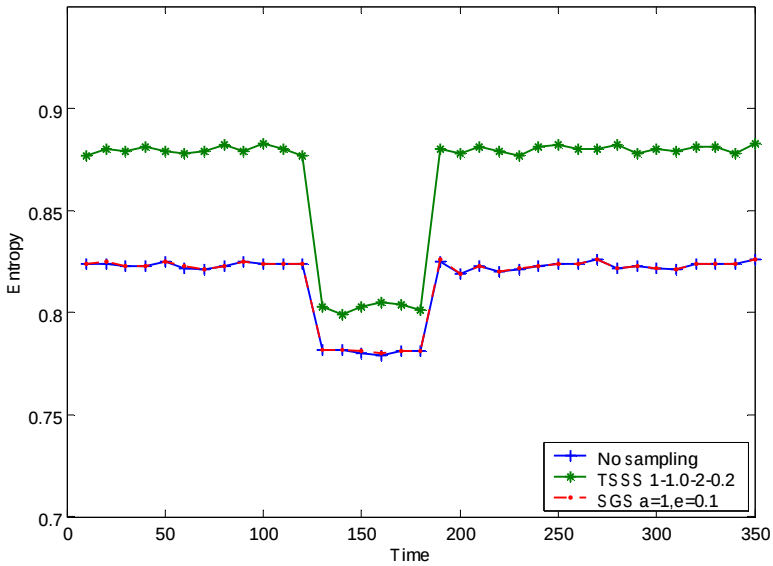
6.5. Σύγκριση με άλλες τεχνικές δειγματοληψίας

Μια άλλη προσέγγιση που ακολουθεί μια παρόμοια αρχή για την εφαρμογή δειγματοληψίας πακέτων βασισμένη στο τρέχον μέγεθος κάθε ροής που ανήκει ένα πακέτο, είναι η μέθοδος Δειγματοληψίας Βασισμένη σε Σκίτσο (Sketch Guided Sampling – SGS) [Kuma06]. Συγκεκριμένα, η μέθοδος SGS εφαρμόζει δειγματοληψία πακέτων σύμφωνα με μια συνάρτηση *f*, η οποία επιλέγει ένα πακέτο με μικρότερη πιθανότητα αν πρόκειται για πακέτο που ανήκει σε μεγάλη ροή, ενώ με μεγαλύτερη πιθανότητα επιλέγει πακέτα που ανήκουν σε μικρές ροές. Κατά αυτόν τον τρόπο, η μέθοδος SGS είναι σε θέση να επιλέγει πακέτα από τις μικρές και μεσαίου μεγέθους ροές, αντίθετα προς την Τυχαία Δειγματοληψία Πακέτων που όπως είναι γνωστό χάνει αυτά τα πακέτα, και επομένως να παράγει έτσι ακριβέστερα στατιστικά στοιχεία για την κίνηση του δικτύου. Ολόκληρη η μέθοδος αυτή είναι βασισμένη σε μία μικρή δομή περίληψης, αποκαλούμενη "Σκίτσο" η οποία κρατά και ενημερώνει τα τρέχοντα μεγέθη όλων των ροών.

Αν και αυτή η μέθοδος είναι πολύ χρήσιμη για τις περισσότερες από τις διαδικασίες διαχείρισης δικτύων, όπως η ακριβής εκτίμηση των στατιστικών του δικτύου (π.χ. κατανομή μεγέθους ροών) και η ανίχνευση των μεγάλων ροών, δεν είναι κατάλληλη για την ανίχνευση των ανωμαλιών που προκαλούνται από μικρές ροές. Στο Σχήμα 6.9, παρουσιάζουμε μια σύγκριση μεταξύ της μεθόδου SGS και της προτεινόμενης μεθόδου Επιλεκτικής Δειγματοληψίας Δύο Σταδίων (TSSS) για την ανωμαλία του αυτοδιαδιδόμενου ιού που περιγράψαμε στις προηγούμενες ενότητες σε ένα ποσοστό ανώμαλων ροών 10% σε σχέση με την κανονική κίνηση του δικτύου (μετρημένη σε ροές). Για την περίπτωση της μεθόδου SGS, εφαρμόζουμε τη συνάρτηση δειγματοληψίας $f(s) = 1/(1+\varepsilon^2 s^{(2\alpha-1)})$ [Kuma06], όπου s είναι το μέγεθος ροής, ε είναι η τιμή του σχετικού λάθους και α είναι μια σταθερά ($1/2 \leq \alpha \leq 1$), χρησιμοποιώντας τις τιμές $a=1$ και $\varepsilon=0.1$. Για την περίπτωση της Επιλεκτικής

Δειγματοληψίας Δύο Σταδίων (TSSS) επιλέγουμε τις τιμές παραμέτρων *z=1, c=1.0, n=2* και p_p*=0.20*.

Όπως μπορούμε να παρατηρήσουμε από το Σχήμα 6.9, η μέθοδος TSSS ξεπερνά στην ανίχνευση τη μέθοδο SGS, μεγεθύνοντας την ανωμαλία του αυτοδιαδιδόμενου ιού, αφού η τιμή της εντροπίας για τις IP διευθύνσεις πηγής μειώνεται από 0.88 σε 0.80, ενώ η καμπύλη της μεθόδου SGS είναι σχεδόν παρόμοια με την καμπύλη της αρχικής περίπτωσης όπου δεν εφαρμόζεται δειγματοληψία. Αυτό οφείλεται στο γεγονός ότι η μέθοδος SGS επιλέγει τα πακέτα από όλες τις ροές, με συνέπεια τη δημιουργία ενός συνόλου δειγμάτων με παρόμοια κατανομή μεγέθους ροών με το αρχικό δείγμα. Αντίθετα, ο στόχος της προτεινόμενου μεθόδου Επιλεκτικής Δειγματοληψίας Δύο Σταδίων (TSSS) είναι να επιλεγούν τα κατάλληλα στοιχεία προκειμένου να μεγεθυνθούν οι ανωμαλίες, βελτιώνοντας περαιτέρω την αποτελεσματικότητα ανίχνευσης ανωμαλιών και όχι να διατηρήσει τις στατιστικές ιδιότητες του αρχικού δείγματος.



Σχήμα 6.9. Εντροπία για τις IP διευθύνσεις πηγής για ανωμαλία σε ποσοστό 10%

6.6. Θέματα Εφαρμογής και Υλοποίησης

Στο σημείο αυτό, πρέπει να σημειωθεί ότι η μέθοδος Επιλεκτικής Δειγματοληψίας Δύο Σταδίων περιλαμβάνει μια ιδιαίτερα βαριά διαδικασία που απαιτεί πολλούς

πόρους, με αποτέλεσμα να μην επιτρέπεται να χρησιμοποιηθεί το προτεινόμενο σχέδιο δειγματοληψίας μέσα σε έναν δρομολογητή. Παρόλα αυτά, η μέθοδος της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων μπορεί να εφαρμοστεί σε κάποιο σημείο μέτρησης μέσα στο δίκτυο χρησιμοποιώντας παραδείγματος χάριν μια κάρτα επεξεργαστή δικτύου (network processor card). Οι σύγχρονες κάρτες επεξεργαστών δικτύων είναι σε θέση να πραγματοποιούν παθητικό έλεγχο (passive monitoring) σε ταχύτητες που αρχίζουν από 1 Gbps και φτάνουν μέχρι και 10 Gbps. Για ένα χρονικό παράθυρο των 10sec, τα πακέτα καταγράφονται και ταξινομούνται σε ροές, ενώ η μέθοδος αυτή δειγματοληψίας μπορεί να εφαρμοστεί στο τέλος του χρονικού παραθύρου. Επιπλέον, η Επιλεκτική Δειγματοληψία Δύο Σταδίων μπορεί να εφαρμοστεί μόνο για ένα συγκεκριμένο τμήμα της κίνησης του δικτύου (ύποπτη κίνηση), σε μια αρχιτεκτονική ανίχνευσης πολλών σταδίων όπως είναι η αρχιτεκτονική LADS – Large-scale Automated DDoS Detection System – (μεγάλης κλίμακας αυτοματοποιημένο σύστημα ανίχνευσης DDoS επιθέσεων) [Seka06]. Στην αρχιτεκτονική αυτή, τα αρχικά στάδια αποτελούνται από μηχανισμούς ανίχνευσης ανωμαλιών χαμηλού υπολογιστικού κόστους (συνήθως βασισμένο σε δεδομένα SNMP), ενώ τα επόμενα στάδια, τα οποία ενεργοποιούνται κατόπιν κάποιας διαταγής και επομένως λιγότερο συχνά, εφαρμόζονται στα δεδομένα κίνησης του δικτύου που έχουν χαρακτηριστεί ως ύποπτα. Οι τελευταίοι μηχανισμοί ανίχνευσης ανωμαλιών εκτελούν ανάλυση υψηλής υπολογιστικής πολυπλοκότητας όπως είναι η ανάλυση των ροών ή των επικεφαλίδων των πακέτων. Η προτεινόμενη μέθοδος της Επιλεκτικής Δειγματοληψίας Δύο Σταδίων θα μπορούσε επομένως να εφαρμοστεί σε ένα από τα προχωρημένα στάδια αυτής της αρχιτεκτονικής, μειώνοντας έτσι το υπολογιστικό κόστος της διαδικασίας ανίχνευσης ανωμαλιών.

7. Ευφυείς Μέθοδοι Δειγματοληψίας για Ανίχνευση Ανωμαλιών Δικτύου

7.1. Εισαγωγή

Σε αυτό το κεφάλαιο γίνεται μελέτη της επίδρασης «Ευφυών Δειγματοληπτικών Μεθόδων» στην ανίχνευση και κατηγοριοποίηση διαφόρων ανωμαλιών δικτύου. Η ανίχνευση ανωμαλιών [Barf01], όπως έχουμε περιγράψει και σε προηγούμενο κεφάλαιο, είναι βασισμένη στην έννοια των παρεκκλίσεων από την κανονική συμπεριφορά ορισμένων χαρακτηριστικών γνωρισμάτων του δικτύου.

Με βάση την παρατήρηση ότι για συγκεκριμένες εφαρμογές όπως η ανίχνευση ανωμαλιών, ένα μεγάλο μέρος των πληροφοριών που μας είναι χρήσιμες περιλαμβάνεται σε ένα μικρό μέρος των ροών, καταδεικνύουμε ότι με τη χρησιμοποίηση των «Ευφυών Τεχνικών Δειγματοληψίας», επιτυγχάνουμε τη μεγέθυνση της εμφάνισης των ανωμαλιών μέσα στο επιλεγέν δείγμα στοιχείων με την επιλογή των κατάλληλων στοιχείων. Με τον τρόπο αυτό επιτυγχάνουμε τη βελτίωση της αποτελεσματικότητας της ανίχνευσης και σε μερικές περιπτώσεις πετυχαίνουμε την ανίχνευση ανωμαλιών που θα ήταν αόρατες σε άλλη περίπτωση.

Αντίθετα της κοινής πρακτικής, όπου η δειγματοληψία θεωρείται ως μία διαδικασία με απώλειες που επηρεάζει αρνητικά την ποιότητα του επιλεγμένου υποσυνόλου των δεδομένων δικτύου σε σχέση με τις περισσότερες από τις διαδικασίες διαχείρισης και ελέγχου δικτύων, σε αυτή τη μελέτη υποστηρίζουμε και καταδεικνύουμε ότι η δειγματοληψία όχι μόνο δεν βλάπτει τη διαδικασία ανίχνευσης ανωμαλιών, αλλά σε μερικές περιπτώσεις διευκολύνει και βελτιώνει την αποτελεσματικότητά της. Επομένως, υποστηρίζουμε πως μπορεί να προκύψει μία ολόκληρη νέα κατηγορία μεθόδων δειγματοληψίας, αποκαλούμενη «Ευκαιριακή Δειγματοληψία» (Opportunistic Sampling), η οποία στοχεύει να αντιστρέψει το μειονέκτημα της απώλειας πληροφοριών κατά τη δειγματοληψία, σε ένα σημαντικό ευεργετικό χαρακτηριστικό γνώρισμα στην περίπτωση της ανίχνευσης ανωμαλιών δικτύου.

Στο υπόλοιπο αυτού του κειμένου, χρησιμοποιούμε τον όρο «Ευφυείς Μέθοδοι Δειγματοληψίας» και τον όρο «Ευκαιριακές Μέθοδοι Δειγματοληψίας» με την ίδια

σημασία. Προκειμένου, να αξιολογήσουμε ποιοτικά και ποσοτικά την επίδραση των «Ευφυών Δειγματοληπτικών Μεθόδων» στην αποτελεσματικότητα της ανίχνευσης και κατηγοριοποίησης των ανωμαλιών δικτύου, χρησιμοποιούμε μία μέθοδο βασισμένη στην εντροπία. Πρέπει να σημειωθεί εδώ ότι οι Ευφυείς Τεχνικές Δειγματοληψίας που εξετάζονται στη διατριβή αυτή μπορούν να χρησιμοποιηθούν είτε για την ανίχνευση και την ταξινόμηση των ανωμαλιών σε πραγματικό χρόνο, είτε για τη μεταγενέστερη εξέταση ενός καταγεγραμμένου αρχείου δεδομένων δικτύου για ύπαρξη ανωμαλιών (network forensics).

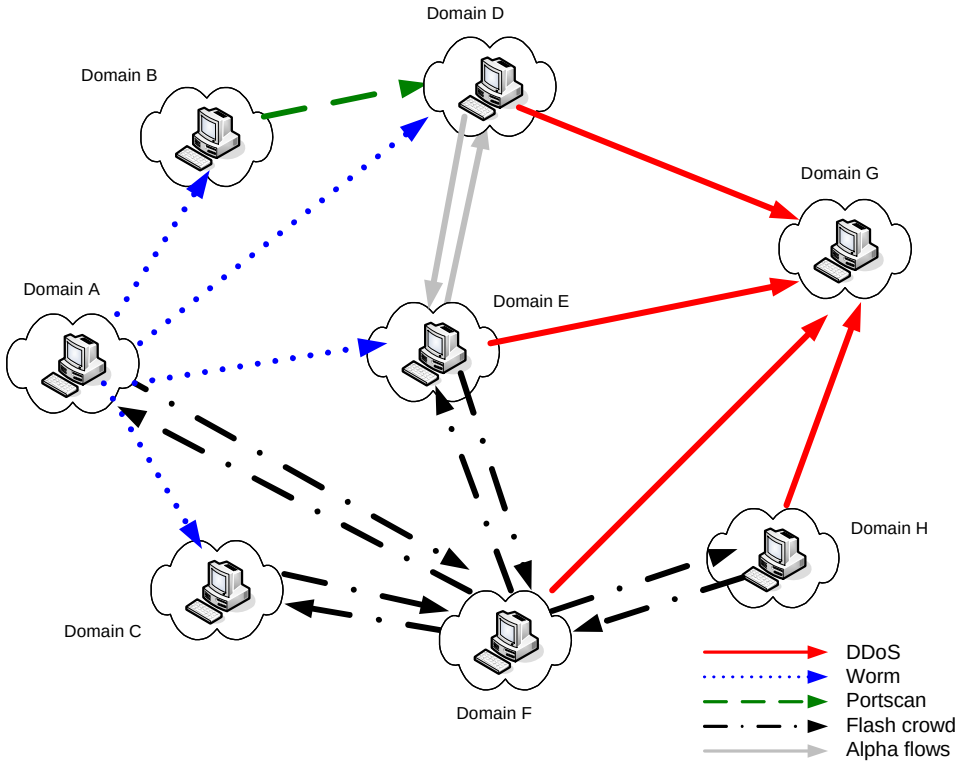
7.2. Παρουσίαση των χαρακτηριστικότερων ανωμαλιών δικτύου

Σε αυτή την ενότητα περιγράφουμε ένα ευρύ φάσμα ανωμαλιών δικτύου που προκαλούνται είτε από κακόβουλους χρήστες είτε από νόμιμα γεγονότα κίνησης δικτύου. Η ανίχνευση και η ταξινόμηση των ανωμαλιών δικτύου είναι πολύ σημαντική για τους διαχειριστές δικτύων, επειδή η κατανόηση της φύσης της ανωμαλίας θα τους βοηθήσει να πάρουν τα κατάλληλα μέτρα προκειμένου να επαναφέρουν το δίκτυο στην κανονική λειτουργία του. Σε αυτή την ενότητα, εστιάζουμε σε τρεις γνωστές ανωμαλίες (προερχόμενες από κακόβουλη χρήση του δικτύου), οι οποίες θα μπορούσαν να χαρακτηριστούν ως επιθέσεις δικτύων (DDoS, worm propagation, portscan) και δύο άλλες ανωμαλίες που προκαλούνται από νόμιμη χρήση του δικτύου (flash crowd, alpha flows). Το Σχήμα 7.1 απεικονίζει πώς αυτές οι πέντε ανωμαλίες μπορούν να εμφανιστούν σε μια τοπολογία δικτύου.

- **Επίθεση DDoS**: Μια επίθεση DDoS (Distributed Denial of Service) [Doul04] χαρακτηρίζεται από μια προσπάθεια να αποτραπεί η νόμιμη χρήση μιας υπηρεσίας. Γενικά, οι επιθέσεις DDoS εκμεταλλεύονται γνωστές ευπάθειες ενός πρωτοκόλλου επικοινωνίας προκειμένου το θύμα να μην μπορεί να εξυπηρετήσει τα νόμιμα αιτήματα υπηρεσιών που προσφέρει. Μια κοινή πρακτική για τους επιτιθεμένους, στα μεγάλης κλίμακας δίκτυα, είναι να στέλνουν ένα μεγάλο αριθμό πακέτων στο θύμα. Ένας συχνός τύπος επίθεσης DDoS είναι το SYN flooding, όπου οι κακόβουλοι υπολογιστές στέλνουν έναν

μεγάλο αριθμό πακέτων TCP SYN στο θύμα, καθιστώντας κατά συνέπεια τον υπολογιστή-στόχο ανίκανο να επεξεργαστεί όλα αυτά τα αιτήματα. Άλλοι τύποι επιθέσεων DDoS είναι οι επιθέσεις UDP flooding και ICMP flooding όπου ένας μεγάλος αριθμός πακέτων UDP ή ICMP αντίστοιχα, στέλνεται προς το δίκτυο του θύματος από πολλαπλές πηγές προκειμένου να καταναλωθεί το διαθέσιμο εύρος ζώνης (bandwidth) της σύνδεσης δικτύου του θύματος.

- **Worm propagation**: Με τον όρο «worm» [Chen05] ορίζουμε ένα κακόβουλο πρόγραμμα το οποίο αυτοδιαδίδεται μέσω δικτύου και προσπαθεί να μολύνει άλλους υπολογιστές εκμεταλλευόμενο μια συγκεκριμένη ευπάθεια στον υπολογιστή-θύμα. Κατά τη διάρκεια της φάσης διάδοσης, ο μολυσμένος υπολογιστής στέλνει ένα μικρό αριθμό πακέτων ανά στόχο, σε έναν μεγάλο πλήθος υπολογιστών στο διαδίκτυο. Σε αυτή την ενότητα, μελετάμε ένα κοινό τύπο worm που στέλνει τα κακόβουλα πακέτα σε μία συγκεκριμένη θύρα προορισμού σε μια προσπάθεια να εκμεταλλευτεί μία συγκεκριμένη ευπαθή υπηρεσία.
- **Δραστηριότητα Portscan**: Η δραστηριότητα portscan [Srid06] περιλαμβάνει την κυκλοφορία που προκαλείται από έναν υπολογιστή που στέλνει πακέτα ελέγχου σε ένα ευρύ φάσμα θυρών ενός συγκεκριμένου υπολογιστή προκειμένου να ελέγξει ποιες υπηρεσίες είναι διαθέσιμες.
- **Flash crowd**: Το Flash Crowd [Ari03] αποτελεί μια μεγάλη νόμιμη ζήτηση για μια συγκεκριμένη υπηρεσία (π.χ. πολλοί πελάτες που «κατεβάζουν» ταυτόχρονα μία νέα διανομή Linux ή κάποιο αρχείο αναβάθμισης ασφάλειας από έναν κεντρικό υπολογιστή μέσω HTTP/FTP). Αυτό το γεγονός οδηγεί στην αύξηση και της εισερχόμενης (αιτήματα) και εξερχόμενης κίνησης (απαντήσεις) από τον κεντρικό υπολογιστή.
- **Alpha flows**: Οι ροές “Alpha” [Sarv01] συνθέτουν μια ανωμαλία δικτύων στην οποία η κίνηση του δικτύου αυξάνει σε μεγάλο βαθμό από μερικές μόνο συνδέσεις (μεγάλης όμως κίνησης) μεταξύ δύο υπολογιστών. Η κίνηση αυτή προκαλείται συνήθως από μεγάλες μεταφορές αρχείων πάνω από υψηλού εύρους ζώνης συνδέσεις ή από πειράματα δικτύων μεταξύ διαφορετικών περιοχών (domains).



Σχήμα 7.1. Χαρακτηριστικότερες Ανωμαλίες Δικτύου

7.3. Εφαρμογή Ευφυών Μεθόδων Δειγματοληψίας στην ανίχνευση και ταξινόμηση ανωμαλιών δικτύου

Όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο, πρόσφατα ερευνητικά αποτελέσματα [Hohn06] κατέδειξαν ότι η δειγματοληψία ροών (σε σχέση με τη δειγματοληψία πακέτων) βελτιώνει την ακρίβεια των στατιστικών του δικτύου. Το γεγονός αυτό καθιστά τη δειγματοληψία ροών καταλληλότερη για την εφαρμογή της στο πεδίο της ανίχνευσης ανωμαλιών δικτύου. Στη συνέχεια, παρουσιάζουμε συνοπτικά δύο γνωστές τεχνικές δειγματοληψίας ροών. Η πρώτη είναι, η «Επιλεκτική Δειγματοληψία» (Selective Sampling) [Andr08] η οποία όπως έχουμε αναφέρει έχει ως στόχο τις μικρές ροές, ενώ ο δεύτερη αναφέρεται ως «Έξυπνη δειγματοληψία» (Smart Sampling) [Duff03], και επιλέγει τις μεγάλες ροές. Και οι δύο αυτές τεχνικές παρουσιάζουν έναν «ευκαιριακό» (opportunistic) χαρακτήρα στη λειτουργία τους, δεδομένου ότι στοχεύουν στην εκμετάλλευση του γεγονότος ότι ένα μεγάλο μέρος των πληροφοριών (που αφορούν κυρίως την ανωμαλία) περιλαμβάνεται μέσα σε ένα μικρό μέρος των ροών. Επομένως, ανωμαλίες που γίνονται αντιληπτές συνήθως από

«ακραία» στοιχεία (σε σχέση με τα υπόλοιπα) μπορούν να αποκαλυφθούν ευκολότερα μέσα σε ένα κατάλληλα επιλεγμένο σύνολο στοιχείων, όπως αυτό που μπορεί να προκύψει από τεχνικές «ευφυούς δειγματοληψίας».

Έχει καταδειχθεί ότι οι μικρές ροές είναι συνήθως η πηγή πολλών επιθέσεων δικτύων (π.χ. επιθέσεις DDoS, portscans, αυτοδιαδιδόμενοι ιοί) [Barf01], και επομένως πρέπει να επιλεχτούν κατά προτίμηση προκειμένου να επιτευχθεί υψηλή αποτελεσματικότητα στην ανίχνευση ανωμαλιών δικτύου. Η «Επιλεκτική Δειγματοληψία» ακολουθεί αυτό το παράδειγμα και η επιλογή μιας μεμονωμένης ροής είναι βασισμένη στην ακόλουθη έκφραση:

$$p(x) = \begin{cases} c & x \le z \\ \dfrac{z}{n \cdot x} & x > z \end{cases} \qquad (7.1)$$

όπου με x συμβολίζουμε το μέγεθος της ροής σε πακέτα, $0 < c \le 1$, $n \ge 1$ και z είναι ένα κατώφλι (μετρούμενο σε πακέτα).

Αντίθετα, η «Έξυπνη δειγματοληψία» είναι ένας τύπος δειγματοληψίας βασισμένος σε ροές που εστιάζει στην επιλογή των μεγάλων ροών. Πιο συγκεκριμένα, στην έξυπνη δειγματοληψία μια ροή του μεγέθους x επιλέγεται με πιθανότητα $p(x)$ σύμφωνα με την ακόλουθη έκφραση:

$$p(x) = \begin{cases} x/z & x < z \\ 1 & x \ge z \end{cases} \qquad (7.2)$$

όπου το x είναι το μέγεθος ροής σε bytes και z είναι ένα κατώφλι. Στη μελέτη μας, θεωρούμε το x ως μέγεθος ροής σε πακέτα. Όπως μπορούμε να παρατηρήσουμε από την σχέση (7.2), οι ροές που είναι μεγαλύτερες στο μέγεθος από το κατώφλι z επιλέγονται με πιθανότητα ίση με 1, ενώ οι ροές που είναι μικρότερες από z επιλέγονται με πιθανότητα ανάλογη προς το μέγεθός τους. Αυτή η τεχνική δειγματοληψίας είναι κατάλληλη για τον εντοπισμό ανωμαλιών που προκαλούνται από μεγάλες ροές όπως τα flash crowds και οι ροές “Alpha”.

Όπως έχουμε περιγράψει στο προηγούμενο κεφάλαιο η εντροπία έχει χρησιμοποιηθεί εκτενώς στο πεδίο της ανίχνευσης ανωμαλιών δικτύου [Ranj07]. Μερικές από τις κλασικές κατανομές χαρακτηριστικών γνωρισμάτων κίνησης δικτύου που είναι πολύτιμες στην ανίχνευση ανωμαλιών είναι η κατανομή: α) των IP

διευθύνσεων πηγής (srcIP), β) των IP διευθύνσεων προορισμού (dstIP), γ) των θυρών πηγής (srcPort), δ) των θυρών προορισμού (dstPort), και ε) του μεγέθους των ροών σε πακέτα (flow-size). Παραδείγματος χάριν, μια ανωμαλία όπως ένας μολυσμένος υπολογιστής που προσπαθεί να μολύνει και άλλους υπολογιστές στο Διαδίκτυο (διάδοση worm) οδηγεί στη μείωση της εντροπίας των IP διευθύνσεων πηγής. Ο μολυσμένος υπολογιστής παράγει ένα μεγάλο αριθμό ροών αναγκάζοντας την ίδια IP διεύθυνση πηγής να κυριαρχεί στην κατανομή των IP διευθύνσεων πηγής. Με βάση αυτές τις μεταβολές, ο διαχειριστής του δικτύου μπορεί να ανιχνεύσει την παρουσία μιας ανωμαλίας χρησιμοποιώντας προκαθορισμένα κατώφλια στις μεταβολές των αντίστοιχων τιμών εντροπίας.

Σε αυτή τη μελέτη, εστιάζουμε στις κατανομές πιθανότητας των πέντε προαναφερθέντων χαρακτηριστικών γνωρισμάτων κίνησης δικτύου και των αντίστοιχων τιμών εντροπίας που υπολογίζονται κατά τη διάρκεια ενός κυλιόμενου χρονικού παραθύρου (της τάξης των 10 sec). Ο Πίνακας 7.1 συνοψίζει τις ανωμαλίες που εξετάζονται σε αυτή τη μελέτη και τις αντίστοιχες μεταβολές στις τιμές εντροπίας των χαρακτηριστικών γνωρισμάτων κίνησης του δικτύου.

Πίνακας 7.1. Ταξινόμηση των ανωμαλιών δικτύου με βάση την μεταβολή της εντροπίας

Είδος Ανωμαλίας	**Περιγραφή**	**Μεταβολή Εντροπίας**
Distributed Denial of Service Attack (DDoS)	Επίθεση εναντίον μιας συγκεκριμένης υπηρεσίας με σκοπό αυτή να γίνει μη διαθέσιμη στους χρήστες	Σημαντική μείωση στην εντροπία dstIP και dstPort Σχεδόν καθόλου αλλαγή στο srcIP, srcPort και flow-size
Worm Propagation	Ένα αυτοδιαδιδόμενο πρόγραμμα το οποίο προσπαθεί να μολύνει και άλλους υπολογιστές εκμεταλλευόμενο μία συγκεκριμένη ευπάθεια του συστήματος	Σημαντική μείωση στην εντροπία srcIP και dstPort Μικρή αύξηση στο dstIP και srcPort Μικρή μείωση στο flow-size
Portscan	Αποστολή πακέτων σε ένα μεγάλο φάσμα θυρών ενός συστήματος με σκοπό την ανίχνευση των ενεργών υπηρεσιών που αυτό προσφέρει	Σημαντική μείωση στην εντροπία srcIP, dstIP και srcPort Μικρή αύξηση στο dstPort Μικρή μείωση στο flow-size
Flash Crowd	Μεγάλη ζήτηση για μία συγκεκριμένη υπηρεσία (π.χ. πολλοί πελάτες «κατεβάζουν» ένα συγκεκριμένο αρχείο από εναν HTTP/FTP server)	Μικρή μείωση στο srcIP, dstIP, srcPort, dstPort και flow-size
Alpha Flows	Μικρός αριθμός ροών που περιλαμβάνουν μία εξαιρετικά μεγάλη ποσότητα πακέτων (μεταφορά δεδομένων μεταξύ δύο υπολογιστών)	Μικρή μείωση στο srcIP και dstIP Σχεδόν καθόλου αλλαγή στο srcPort, dstPort και flow-size

7.4. Μεθοδολογία - Αποτελέσματα

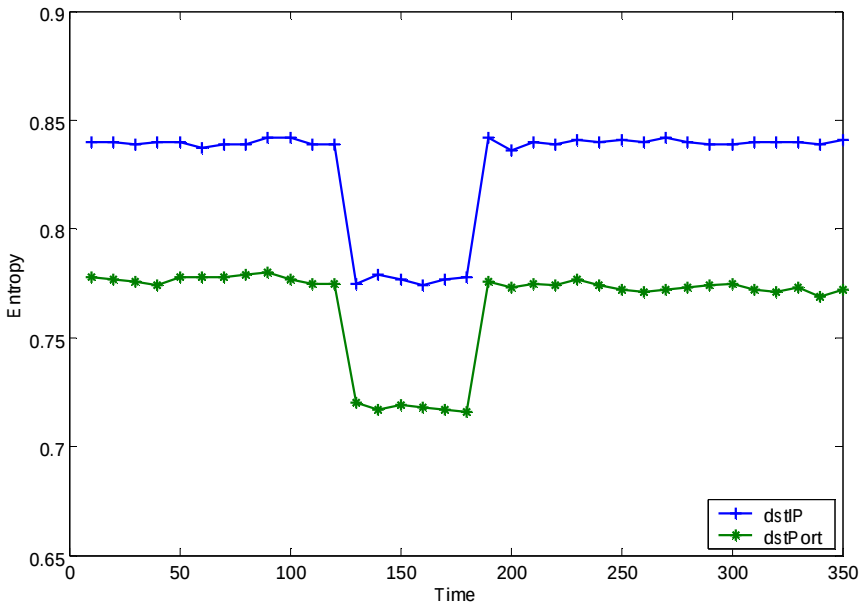
Τα αποτελέσματα και οι αντίστοιχες παρατηρήσεις που παρουσιάζονται και σε αυτή την ενότητα είναι βασισμένα σε ένα αρχείο καταγραφής δεδομένων δικτύου που συλλέχτηκε από την σύνδεση του δικτύου του ΕΜΠ με το ΕΔΕΤ. Στην ακόλουθη αξιολόγηση, προσθέσαμε ένα ποσοστό καθεμίας από τις πέντε ανωμαλίες που περιγράφηκαν νωρίτερα στο συλλεχθέν αρχείο δεδομένων και στη συνέχεια μελετάμε κάθε περίπτωση ανωμαλίας ξεχωριστά. Προκειμένου να προκύψει κάποιο συμπέρασμα σχετικά με την αποτελεσματικότητα των ευφυών τεχνικών δειγματοληψίας, τις συγκρίνουμε ενάντια στην τυχαία δειγματοληψία ροών. Στην «Τυχαία Δειγματοληψία Ροών» θυμίζουμε ότι κάθε ροή επιλέγεται ανεξάρτητα με την ίδια πιθανότητα p. Το ποσοστό των επιλεγεισών ροών επιλέχτηκε ως κοινό κριτήριο για τη σύγκριση.

Στα πειράματά μας, εστιάζουμε αρχικά στις ανωμαλίες που χαρακτηρίζονται ως επιθέσεις δικτύων (δηλ. DDoS, διάδοση worm, portscan) και έπειτα μελετάμε δύο άλλες ανωμαλίες δικτύων που προκαλούνται από νόμιμη χρήση του δικτύου (δηλ. Flash crowd, alpha flows). Αυτές οι πέντε διαφορετικές ανωμαλίες εισάγονται στο δείγμα μας σε διαφορετικές αναλογίες σε σχέση με την κανονική κίνηση δικτύου (που μετριέται σε αριθμό ροών), κυμαινόμενες από 0.2% έως 15%, προκειμένου να καταδειχθεί καλύτερα η συμπεριφορά της μεταβολής της εντροπίας όσον αφορά τη διαδικασία ανίχνευσης ανωμαλιών.

7.4.1. Περίπτωση DDoS attack

Για να προσομοιώσουμε την επίθεση DDoS εισάγουμε ένα ποσοστό κίνησης TCP SYN πακέτων μέσα στην κανονική κίνηση του δικτύου που μελετάμε. Συγκεκριμένα, θεωρούμε 200 επιτιθεμένους υπολογιστές που στέλνουν πακέτα TCP SYN σε έναν υπολογιστή μέσα στο ΕΜΠ, στοχεύοντας την θύρα TCP 80 (http). Κάθε ένας από τους επιτιθεμένους υπολογιστές θεωρούμε ότι στέλνει 20 ροές που αποτελούνται από 1-4 πακέτα, ανά χρονικό παράθυρο. Η κίνηση που παράγεται από τη συνολική επίθεση αντιστοιχεί στο 15% της κανονικής κίνησης του δικτύου (μετρούμενη σε αριθμό ροών ανά χρονικό παράθυρο).

Αυτός ο τύπος ανωμαλίας, όπως φαίνεται στον Πίνακα 7.1, προκαλεί σημαντική μείωση στη εντροπία των IP διευθύνσεων προορισμού και των θυρών προορισμού. Αυτό οφείλεται στο γεγονός ότι η διεύθυνση IP του θύματος και της συγκεκριμένης θύρας (θύρα TCP 80 στην περίπτωσή μας), εμφανίζεται πολυάριθμες φορές σε ένα χρονικό παράθυρο, προκαλώντας κατά συνέπεια τη συγκέντρωση των αντίστοιχων κατανομών γύρω από ένα συγκεκριμένο στοιχείο. Αντίθετα, οι τιμές εντροπίας για τις IP διευθύνσεις πηγής και τις θύρες πηγής δεν παρουσιάζουν σημαντική αλλαγή επειδή η επίθεση DDoS δεν προκαλεί σημαντική αλλαγή στις αντίστοιχες κατανομές. Αυτό οφείλεται στο γεγονός ότι αυτές οι κατανομές περιέχουν ήδη αρκετή «τυχαιότητα» στις κανονικές συνθήκες λειτουργίας του δικτύου και η πρόσθετη κίνηση που προκύπτει από τις 200 IP διευθύνσεις πηγής των επιτιθέμενων, οι οποίοι χρησιμοποιούν τυχαίες θύρες πηγής δεν αλλάζει σημαντικά αυτές τις κατανομές. Η εντροπία στη κατανομή του μεγέθους ροής (flow-size) δεν επηρεάζεται, αφού οι ροές επίθεσης που αποτελούνται από 1-4 πακέτα δεν αλλάζουν τη heavy-tailed κατανομή του μεγέθους των ροών.



Σχήμα 7.2. Εντροπία IP διευθύνσεων και θυρών προορισμού για την περίπτωση της επίθεσης DDoS

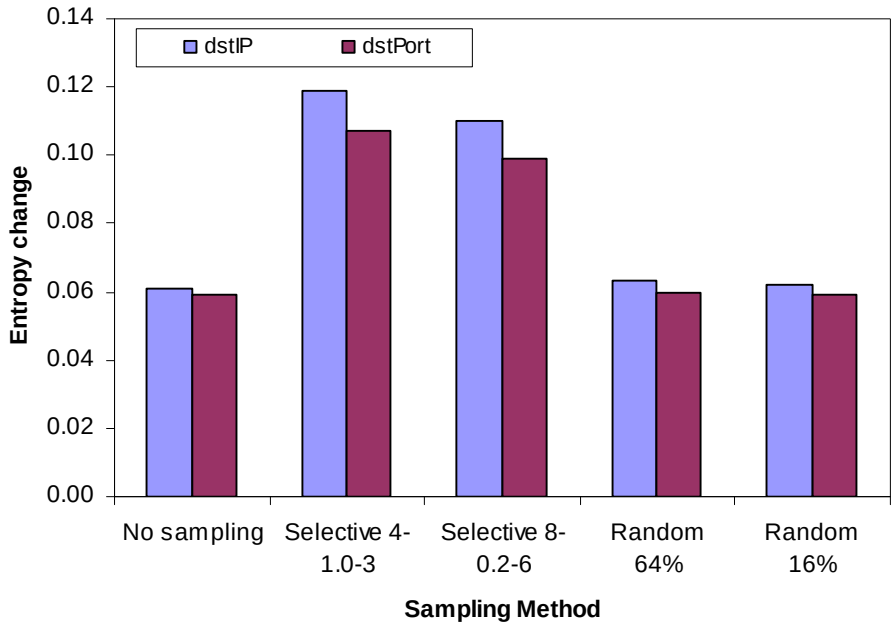
Στο Σχήμα 7.2 παρουσιάζουμε τις τιμές των δύο μετρικών εντροπίας (IP διευθύνσεις και θύρες προορισμού) που επηρεάζονται από την εισαγωγή της

ανωμαλίας, καθώς ο χρόνος εξελίσσεται. Όπως μπορούμε να παρατηρήσουμε, η εντροπία για τις IP διευθύνσεις προορισμού μειώνεται από 0.84 σε 0.78, ενώ η εντροπία θυρών προορισμού ελαττώνεται από 0.78 σε 0.72 κατά τη διάρκεια του διαστήματος 120-180 sec, στο οποίο εμφανίζεται η ανωμαλία.

Σε αυτό το σενάριο επίθεσης DDoS, προκειμένου να ενισχυθεί η αποτελεσματικότητα ανίχνευσης, γίνεται επιλογή ροών χρησιμοποιώντας τη μέθοδο της «Επιλεκτικής Δειγματοληψίας». Οι τιμές των παραμέτρων της επιλεκτικής δειγματοληψίας είναι υπολογισμένες με βάση τη φύση της επίθεσης που πρέπει να ανιχνευθεί. Για αυτό το συγκεκριμένο σενάριο, στην πρώτη περίπτωση επιλέγουμε τιμές παραμέτρων *z=4, c=1.0 και n=3* για να επιτύχουμε την επιλογή όλων των ροών επίθεσης, ενώ στη δεύτερη περίπτωση, όπου *z=8, c=0.2 και n=6*, ακολουθούμε μια γενικότερη προσέγγιση για τη ανίχνευση των επιθέσεων DDoS που αποτελούνται και από μεγαλύτερες ροές.

Στο Σχήμα 7.3 παρουσιάζουμε την μεταβολή στις τιμές της εντροπίας στις IP διευθύνσεις και τις θύρες προορισμού μεταξύ της κανονικής λειτουργίας του δικτύου και της περιόδου κατά την οποία εμφανίζεται η ανωμαλία. Όπως μπορούμε να παρατηρήσουμε η μεταβολή στην εντροπία των IP διευθύνσεων και θυρών προορισμού είναι σχεδόν διπλάσια στην πρώτη περίπτωση της επιλεκτικής δειγματοληψίας σε σχέση με την αρχική περίπτωση (καθόλου δειγματοληψία). Η μεγαλύτερη αύξηση στη μεταβολή της εντροπίας υποδηλώνει και την αποτελεσματικότερη ανίχνευση της ανωμαλίας. Αυτή η συμπεριφορά αποδίδεται στο γεγονός ότι στην πρώτη περίπτωση της επιλεκτικής δειγματοληψίας έχουν επιλεγεί όλες οι ροές επίθεσης, ενώ ταυτόχρονα ένα μεγάλο ποσοστό των κανονικών ροών έχει απορριφθεί. Κατά αυτόν τον τρόπο, η διεύθυνση IP του θύματος κυριαρχεί στη κατανομή των IP διευθύνσεων προορισμού. Στη δεύτερη περίπτωση όπου *z=8, c=0.2 και n=6* η ανωμαλία DDoS ανιχνεύεται πάλι, αλλά σε ένα μικρότερο βαθμό, αφού λιγότερες ροές επίθεσης έχουν επιλεγεί.

Αντίθετα, στις περιπτώσεις της «Τυχαίας Δειγματοληψίας Ροών» (η περίπτωση 64% αντιστοιχεί στην πρώτη περίπτωση της επιλεκτικής δειγματοληψίας, ενώ η 16% στη δεύτερη περίπτωση της επιλεκτικής δειγματοληψίας) παρουσιάζεται παρόμοια συμπεριφορά με την περίπτωση που δεν εφαρμόζεται δειγματοληψία. Αυτό οφείλεται στο γεγονός ότι η τυχαία δειγματοληψία ροών επιλέγει τις ροές με την ίδια πιθανότητα (ανεξάρτητα από το μέγεθος ροής), καταλήγοντας έτσι στην επιλογή του ίδιου ποσοστού ροών επίθεσης σε σχέση με τις κανονικές ροές.



Σχήμα 7.3. Μεταβολή εντροπίας για την περίπτωση της επίθεσης DDoS

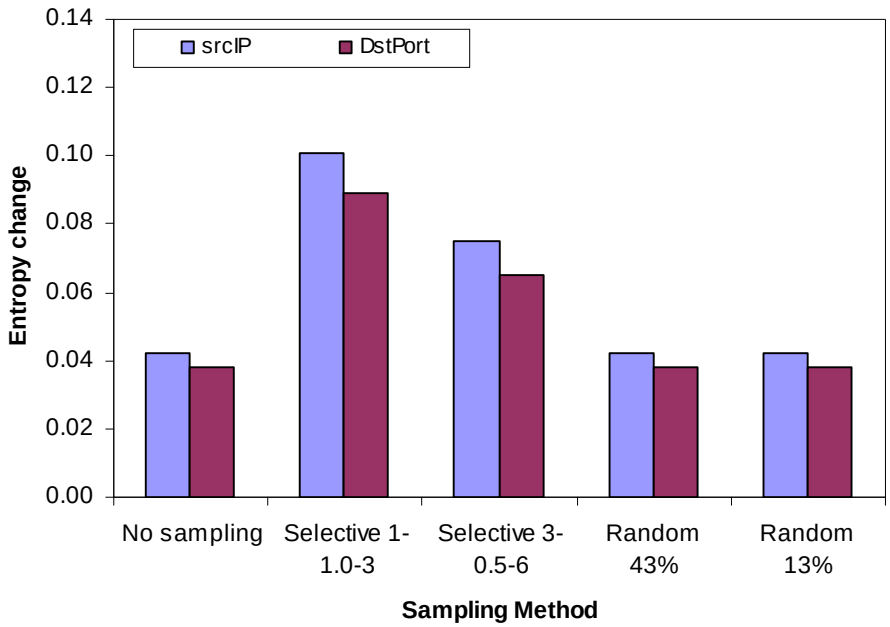
7.4.2. Περίπτωση Worm propagation

Στη συνέχεια, εξετάζουμε ένα σενάριο αυτοδιαδιδόμενου ιού (worm) βασισμένο στο Slammer worm [Moor03]. Η φάση διάδοσης αυτού του worm περιλαμβάνει μόνο ένα πακέτο UDP που στοχεύει στη θύρα 1434. Συγκεκριμένα, θεωρούμε έναν υπολογιστή μέσα στο ΕΜΠ ο οποίος έχει μολυνθεί και παράγει ροές ανωμαλίας που αντιστοιχούν στο 10% του συνολικού αριθμού των ροών σε ένα χρονικό παράθυρο. Οι ροές ανωμαλίας αποτελούνται από ένα πακέτο UDP ανά IP διεύθυνση προορισμού, όπου η διεύθυνση προορισμού IP επιλέγεται τυχαία. Η θύρα πηγής κάθε πακέτου UDP επιλέγεται επίσης τυχαία και κυμαίνεται από 1 έως 65535.

Η διάδοση του ιού οδηγεί στη σημαντική μείωση της εντροπίας στις IP διευθύνσεις πηγής και την εντροπία θυρών προορισμού, ενώ η εντροπία θυρών πηγής και των IP διευθύνσεων προορισμού ελαφρώς αυξάνεται. Η διεύθυνση IP του μολυσμένου υπολογιστή και η συγκεκριμένη θύρα προορισμού (θύρα 1434 στην περίπτωσή μας) εμφανίζονται σε κάθε κακόβουλη ροή, προκαλώντας έτσι τη

συγκέντρωση των αντίστοιχων κατανομών γύρω από τη συγκεκριμένη διεύθυνση IP και τη θύρα προορισμού. Οι τιμές εντροπίας για τις IP διευθύνσεις προορισμού και τις θύρες πηγής δεν επηρεάζονται δεδομένου ότι η πρόσθετη τυχαία σάρωση των διευθύνσεων IP χρησιμοποιώντας τυχαίες θύρες πηγής δεν αλλάζει σημαντικά τις αντίστοιχες κατανομές.

Σε αυτό το σενάριο, προκειμένου να ενισχυθεί η αποτελεσματικότητα ανίχνευσης της ανωμαλίας, επιλέγουμε κατά προτίμηση τις μικρές ροές, όπως στην προηγούμενη περίπτωση, χρησιμοποιώντας τη μέθοδο της επιλεκτικής δειγματοληψίας, και χρησιμοποιώντας συγκεκριμένα τις ακόλουθες τιμές για τις παραμέτρους: στην πρώτη περίπτωση επιλέγουμε *z=1, c=1.0 και n=3* για να επιτύχουμε την επιλογή όλων των ροών της ανωμαλίας (γνωρίζοντας ότι το Slammer worm χρησιμοποιεί ροές με ένα μόνο πακέτο), ενώ στη δεύτερη περίπτωση, όπου *z=3, c=0.5 και n=6,* ακολουθούμε μια γενικότερη προσέγγιση για την ανίχνευση των αυτοδιαδιδόμενων ιών, που χρησιμοποιούν πιο μεγάλες ροές κατά τη διάρκεια της φάσης διάδοσης τους.



Σχήμα 7.4. Μεταβολή εντροπίας για την περίπτωση του worm propagation

Το Σχήμα 7.4 παρουσιάζει τη μεταβολή στις τιμές εντροπίας των IP διευθύνσεων πηγής και των θυρών προορισμού μεταξύ της κανονικής λειτουργίας του δικτύου και της περιόδου κατά τη διάρκεια της ανωμαλίας. Η μεταβολή στην εντροπία των IP διευθύνσεων πηγής και των θυρών προορισμού έχει αυξηθεί σημαντικά στην πρώτη

περίπτωση της επιλεκτικής δειγματοληψίας, και φθάνει σε τιμή 0.1 και 0.09 αντίστοιχα, επειδή όλες οι ροές της ανωμαλίας έχουν επιλεγεί, ενώ ταυτόχρονα ένα μεγάλο ποσοστό των κανονικών ροών έχει απορριφθεί. Στη δεύτερη περίπτωση όπου *z=3, c=0.5 και n=6* η ανωμαλία ανιχνεύεται πάλι, αλλά σε έναν μικρότερο βαθμό, αφού λιγότερες ροές επίθεσης έχουν επιλεγεί. Αντίθετα, και οι δύο περιπτώσεις της μεθόδου της τυχαίας δειγματοληψίας ροών είναι παρόμοιες με την αρχική περίπτωση.

7.4.3. Περίπτωση Portscan Activity

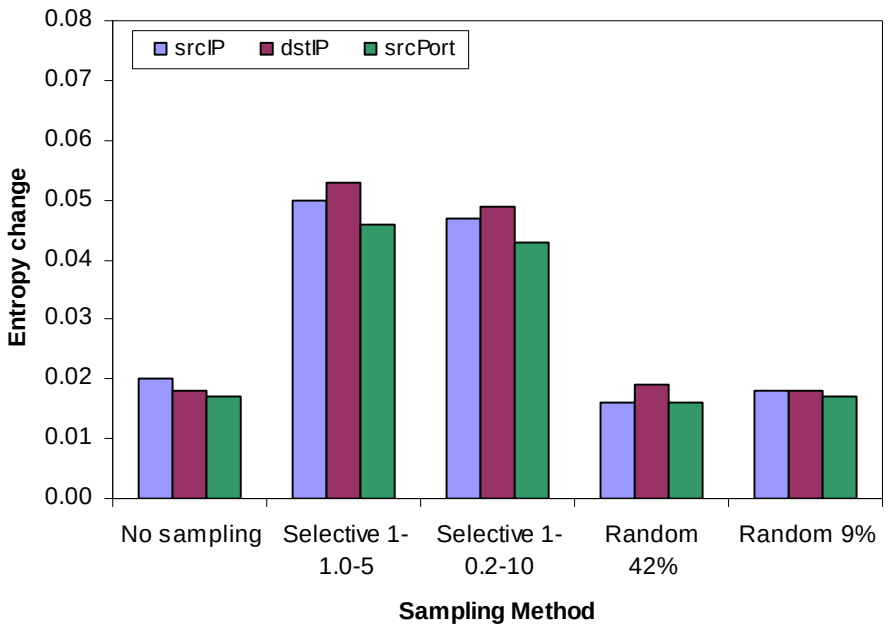
Στην περίπτωση δραστηριότητας portscan εισάγουμε κίνηση από έναν υπολογιστή (scanner) έξω από το ΕΜΠ προς έναν υπολογιστή μέσα στο ΕΜΠ. Ο υπολογιστής-scanner στέλνει πακέτα ελέγχου (probe packets) προς 1400 θύρες του υπολογιστή-στόχου χρησιμοποιώντας την ίδια θύρα πηγής. Η συνολική portscan κίνηση αντιστοιχεί στο 5% της κανονικής κίνησης του δικτύου (που μετριέται σε αριθμό ροών ανά χρονικό παράθυρο).

Σύμφωνα με τον Πίνακα 7.1, αυτή η ανωμαλία προκαλεί σημαντική μείωση στην εντροπία IP διευθύνσεων πηγής, IP διευθύνσεων προορισμού και στην εντροπία θυρών πηγής. Αυτό οφείλεται στο γεγονός ότι και η IP διεύθυνση πηγής (scanner) και η IP διεύθυνση του στόχου (θύμα), εμφανίζονται πολλές φορές μέσα σε ένα χρονικό παράθυρο, προκαλώντας κατά συνέπεια τη συγκέντρωση των αντίστοιχων κατανομών. Με τον ίδιο τρόπο, η συγκεκριμένη θύρα πηγής του υπολογιστή-scanner, από την οποία στέλνονται τα πακέτα, κυριαρχεί στην κατανομή των θυρών πηγής. Αντίθετα, η τιμή της εντροπίας για τις θύρες προορισμού παρουσιάζει μικρή αύξηση, επειδή η κατανομή θυρών προορισμού γίνεται περισσότερο διασκορπισμένη με την τυχαία ανίχνευση των θυρών. Η εντροπία στην κατανομή του μεγέθους ροών μειώνεται ελαφρώς καθώς οι ροές που αποτελούν το portscan αποτελούνται από ένα μόνο πακέτο, προκαλώντας τη συγκέντρωση της αντίστοιχης κατανομής.

Σε αυτό το σενάριο portscan, προκειμένου να βελτιωθεί η αποτελεσματικότητα της ανίχνευσης της ανωμαλίας, επιλέγουμε τις ροές χρησιμοποιώντας την επιλεκτική δειγματοληπτική μέθοδο, και χρησιμοποιώντας συγκεκριμένα τις ακόλουθες τιμές για τις παραμέτρους: στην πρώτη περίπτωση επιλέγουμε *z=1, c=1.0 και n=5*, δεδομένου ότι η δραστηριότητα portscan συμπεριλαμβάνεται στις ροές του ενός μόνο πακέτου,

ενώ στη δεύτερη περίπτωση, όπου *z=1, c=0.2 και n=10,* επιλέγουμε μόνο ένα μικρό τμήμα (συγκεκριμένα 0.2) των ροών με μέγεθος ένα μόνο πακέτο.

Στο Σχήμα 7.5 απεικονίζουμε τις τιμές της εντροπίας των χαρακτηριστικών γνωρισμάτων της κίνησης δικτύου που παρουσιάζουν σημαντική μεταβολή. Όπως μπορούμε να παρατηρήσουμε, στην περίπτωση που δεν εφαρμόζεται δειγματοληψία, οι τιμές εντροπίας για τις IP διευθύνσεις πηγής, τις IP διευθύνσεις προορισμού και των θυρών πηγής μειώνονται μόνο κατά 0.02, το οποίο αποτελεί μία μικρή διαφορά σε σχέση με την κανονική τιμή της αντίστοιχης εντροπίας. Αντίθετα, στις δύο περιπτώσεις της επιλεκτικής δειγματοληψίας, οι αντίστοιχες μεταβολές της εντροπίας είναι σχεδόν ίσες με 0.05, υποδηλώνοντας κατά συνέπεια αποτελεσματικότερη ανίχνευση της ανωμαλίας. Όπως στις προηγούμενες περιπτώσεις ανωμαλιών, και οι δύο περιπτώσεις της τυχαίας δειγματοληψίας ροών (η περίπτωση 42% αντιστοιχεί στην πρώτη περίπτωση της επιλεκτικής δειγματοληψίας, ενώ η 9% αντιστοιχεί στη δεύτερη περίπτωση της επιλεκτικής δειγματοληψίας) παρουσιάζουν τα ίδια αποτελέσματα με την περίπτωση που δεν εφαρμόζεται καθόλου δειγματοληψία.



Σχήμα 7.5. Μεταβολή εντροπίας για την περίπτωση της δραστηριότητας portscan

7.4.4. Περίπτωση Flash Crowd

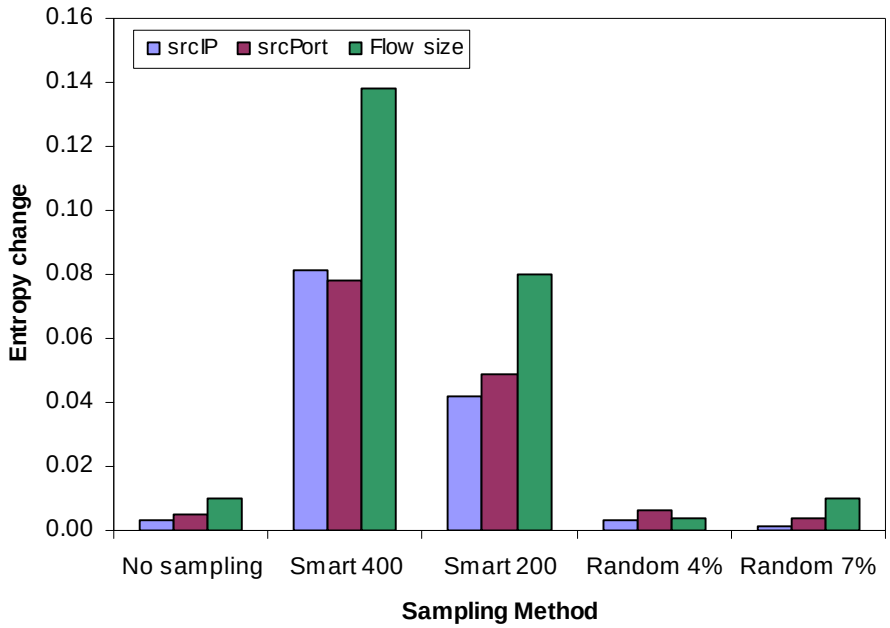
Παρακάτω, μελετάμε ένα σενάριο flash crowd στο οποίο 500 διαφορετικοί υπολογιστές-πελάτες από το Διαδίκτυο ζητούν ένα συγκεκριμένο αρχείο από έναν εξυπηρετητή (server) μέσα στο ΕΜΠ. Αυτό το γεγονός δημιουργεί μία σημαντική αύξηση των ροών και προς τις δύο κατευθύνσεις της μεταφοράς δεδομένων. Συγκεκριμένα, θεωρούμε ροές περίπου 600 πακέτων από τον εξυπηρετητή (server) προς κάθε πελάτη και τις αντίστοιχες ροές με τα πακέτα TCP ACK με κατεύθυνση από τον κάθε πελάτη προς τον εξυπηρετητή. Η συνολική κίνηση του δικτύου (ως προς τις ροές) που παράγεται από αυτήν την ανωμαλία αντιστοιχεί στο 3% της συνολικής κίνησης του δικτύου σε ένα χρονικό παράθυρο.

Όπως απεικονίζεται στον Πίνακα 7.1, η ανωμαλία flash crowd προκαλεί τη μείωση όλων των μετρικών εντροπίας. Η μείωση στις τιμές εντροπίας των IP διευθύνσεων και θυρών πηγής οφείλεται στο γεγονός ότι οι ροές από τη IP διεύθυνση και την θύρα του κεντρικού υπολογιστή κυριαρχούν στις αντίστοιχες κατανομές, ενώ οι μειωμένες τιμές εντροπίας για τις IP διευθύνσεις προορισμού και τις θύρες προορισμού προκαλούνται από τις ροές πελατών που έχουν ως προορισμό τον εξυπηρετητή.

Σε αυτό το σενάριο, όπου η ανωμαλία αποτελείται από ροές με έναν μεγάλο αριθμό πακέτων, χρησιμοποιούμε τη μέθοδο της «Έξυπνης δειγματοληψίας» για να επιλέξουμε τα κατάλληλα στοιχεία, με τις ακόλουθες τιμές για την παράμετρο z. Στην πρώτη περίπτωση επιλέγουμε *z=400*, θεωρώντας ότι τα στοιχεία που μεταφέρονται από τον εξυπηρετητή χρησιμοποιούν ροές που έχουν μέγεθος 600 πακέτα. Στη δεύτερη περίπτωση, όπου *z=200*, ακολουθούμε μια γενικότερη προσέγγιση για την ανίχνευση των γεγονότων flash crowd που χρησιμοποιούν μεγάλες ροές.

Το Σχήμα 7.6 παρουσιάζει τη μεταβολή στις τιμές της εντροπίας για τις IP διευθύνσεις πηγής, τις θύρες πηγής και του μεγέθους ροής μεταξύ της κανονικής λειτουργίας του δικτύου και της περιόδου κατά τη διάρκεια της ανωμαλίας flash crowd. Η μεταβολή στις τιμές της εντροπίας στην περίπτωση που δεν εφαρμόζεται δειγματοληψία είναι λιγότερη από 0.01, καθιστώντας την ανωμαλία σχεδόν μη ανιχνεύσιμη. Χρησιμοποιώντας την «Έξυπνη δειγματοληψία», οι αντίστοιχες τιμές εντροπίας μεταβάλλονται από 0.04 έως και 0.14 στην καλύτερη περίπτωση. Η μεγάλη αυτή μεταβολή στις τιμές εντροπίας για το μέγεθος ροής οφείλεται στο γεγονός ότι έχουμε έναν μικρό αριθμό μεγάλων ροών με διαφορετικά μεγέθη κατά τη διάρκεια

της κανονικής περιόδου, ενώ κατά τη διάρκεια της ανωμαλίας η ποσότητα ροών με μέγεθος 600 πακέτα αυξάνεται εντυπωσιακά, προκαλώντας τη συγκέντρωση της κατανομής του μεγέθους ροής γύρω από αυτή την τιμή (600). Αντίθετα, και οι δύο περιπτώσεις της μεθόδου «Τυχαία Δειγματοληψία Ροών», όπως σε όλα τα προηγούμενα σενάρια, έτσι και σε αυτό, μοιάζουν με την αρχική περίπτωση (καθόλου δειγματοληψία) και αποτυγχάνουν να ανιχνεύσουν την ανωμαλία.



Σχήμα 7.6. Μεταβολή εντροπίας για την περίπτωση του flash crowd

7.4.5. Περίπτωση Alpha Flows

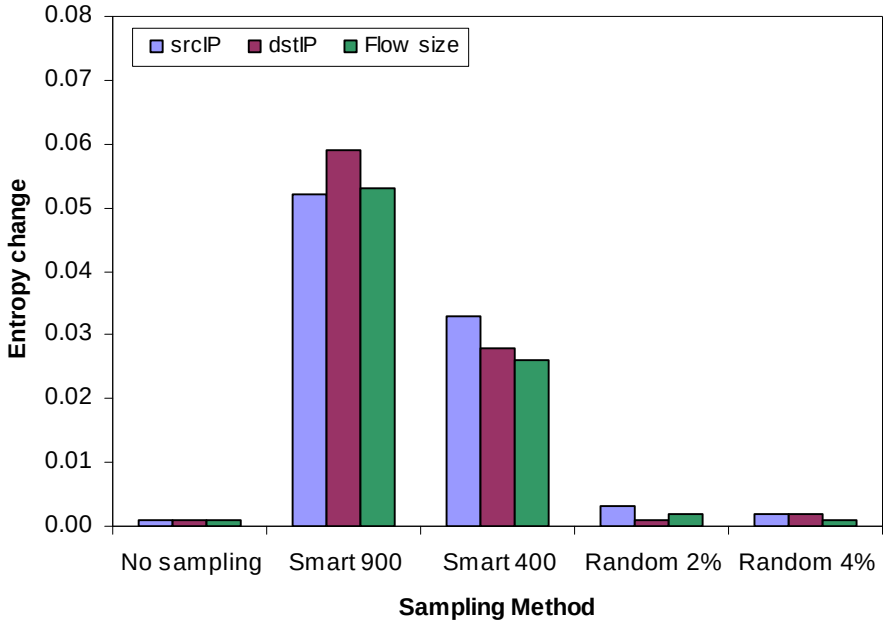
Για αυτόν τον τύπο ανωμαλίας, μελετάμε ένα σενάριο στο οποίο δύο διαφορετικοί υπολογιστές (ο ένας βρίσκεται μέσα στο δίκτυο του ΕΜΠ και ο άλλος απ' έξω) ανταλλάσσουν μεγάλες ποσότητες δεδομένων που αποτελούνται από 50 ροές με 1000 πακέτα κάθε μία από αυτές. Η συνολική κίνηση (σε ροές) που παράγεται από τις "Alpha" ροές αντιστοιχεί στο 0.2% της συνολικής κίνησης του δικτύου σε ένα χρονικό παράθυρο.

Όπως συνοψίζεται στον Πίνακα 7.1, η κίνηση των "Alpha flows" προκαλεί μια μικρή μείωση στις τιμές εντροπίας στις IP διευθύνσεις πηγής και προορισμού. Οι IP διευθύνσεις πηγής και προορισμού των δύο υπολογιστών που πραγματοποιούν την

ανταλλαγή δεδομένων εμφανίζονται πολλαπλές φορές στο δίκτυο, προκαλώντας κατά συνέπεια τη συγκέντρωση των αντίστοιχων κατανομών γύρω από αυτά τα συγκεκριμένα στοιχεία (IP διευθύνσεις).

Σε αυτήν την περίπτωση, όπου η ανωμαλία αποτελείται από μεγάλες ροές, επιλέγουμε κατά προτίμηση τις ροές με έναν μεγάλο αριθμό πακέτων χρησιμοποιώντας την μέθοδο της «Έξυπνης δειγματοληψίας», όπως και προηγουμένως, με τις ακόλουθες τιμές για την παράμετρο z: στην πρώτη περίπτωση επιλέγουμε *z=900*, θεωρώντας ότι η μετάδοση δεδομένων χρησιμοποιεί ροές μεγέθους 1000 πακέτων, ενώ στη δεύτερη περίπτωση, όπου *z=400*, ακολουθούμε μια γενικότερη προσέγγιση για να επιλέξουμε και τις μεσαίου μεγέθους ροές.

Στο Σχήμα 7.7 παρουσιάζουμε τη μεταβολή στην εντροπία των IP διευθύνσεων πηγής και προορισμού καθώς και του μεγέθους ροής μεταξύ της κανονικής λειτουργίας του δικτύου και της περιόδου κατά τη διάρκεια της ανωμαλίας. Όπως μπορούμε να παρατηρήσουμε η μεταβολή στην περίπτωση που δεν εφαρμόζεται δειγματοληψία είναι σχεδόν μηδενική, καθιστώντας κατά συνέπεια την ανίχνευση της ανωμαλίας αδύνατη. Χρησιμοποιώντας την «Έξυπνη δειγματοληψία» οι αντίστοιχες τιμές εντροπίας μειώνονται κατά 0.05 στην πρώτη περίπτωση (*z=900*) και κατά 0.03 στη δεύτερη περίπτωση (*z=400*). Σε αυτήν την περίπτωση η έξυπνη δειγματοληψία απορρίπτει τη μεγάλη πλειοψηφία των μικρών ροών, αφήνοντας έναν μικρό αριθμό ομοιόμορφα κατανεμημένων ροών στις κατανομές των IP διευθύνσεων πηγής, των IP διευθύνσεων προορισμού και του μεγέθους ροών. Έτσι, όταν εμφανίζεται η ανωμαλία, ένα ή δύο συγκεκριμένα στοιχεία κυριαρχούν στις προαναφερθείσες κατανομές. Παραδείγματος χάριν, οι διευθύνσεις IP των δύο υπολογιστών που συμμετέχουν στη μεταφορά δεδομένων κυριαρχούν στην κατανομή της IP διεύθυνσης πηγής και προορισμού, μειώνοντας κατά συνέπεια την εντροπία αυτών των μετρικών.



Σχήμα 7.7. Μεταβολή εντροπίας για την περίπτωση των ροών Alpha

7.5. Θέματα υλοποίησης της μεθόδου και προκλήσεις

Παραδοσιακά, η δειγματοληψία έχει χρησιμοποιηθεί για να παρθεί μία απόφαση σε πραγματικό χρόνο για το εάν ένα πακέτο θα επιλεγεί ή όχι. Ενώ η δειγματοληψία πακέτων είναι αρκετά απλή διαδικασία για να εφαρμοστεί, καταφέρνει να συγκεντρώσει πολύ λιγότερες στατιστικές πληροφορίες από ότι μπορούν να ληφθούν με τη δειγματοληψία ροών. Οι περισσότερες από τις τεχνικές δειγματοληψίας πακέτων που έχουν προταθεί καταλήγουν σε ένα αποτέλεσμα σημαντικά διαστρεβλωμένων κατανομών μεγέθους ροής, συνήθως με έναν απρόβλεπτο και τις περισσότερες φορές μη χρήσιμο τρόπο. Για παράδειγμα, κατά την εφαρμογή της δειγματοληψίας πακέτων, οι μεγάλες ροές έχουν μεγαλύτερη πιθανότητα να επιλεγούν σε σχέση με τις μικρές ροές. Αντίθετα, με τη χρήση της κατάλληλης δειγματοληψίας ροών μπορούμε να επηρεάσουμε τη φύση και τη κατανομή των προκυπτόντων επιλεγμένων στοιχείων, όπου μπορεί να προκύπτει είτε ένα σύνολο από «αμερόληπτα» (unbiased) επιλεγμένα στοιχεία (π.χ. με τυχαία δειγματοληψία ροής, όπου το επιλεγμένο σύνολο στοιχείων διατηρεί τις αρχικές στατιστικές ιδιότητες) είτε να γίνει επιλογή από «προκατειλημμένα» (biased) επιλεγμένα στοιχεία

(π.χ. ευκαιριακή δειγματοληψία – opportunistic sampling, κατά την οποία αποκαλύπτονται ανωμαλίες κατά προτίμηση).

Παρόλα αυτά, η εφαρμογή δειγματοληψίας βασισμένη στις ροές θέτει διάφορες προκλήσεις κατά την εφαρμογή. Συγκεκριμένα, η δειγματοληψία ροών πρέπει να λάβει μια απόφαση σχετικά με το εάν πρέπει να επιλεγεί μία ροή που έχει συλλεχθεί ήδη και έχει αποθηκευτεί στη μνήμη στο τέλος ενός χρονικού παραθύρου. Η δειγματοληψία ροών μπορεί να υλοποιηθεί χρησιμοποιώντας πίνακες κατακερματισμού (Hash Tables) με κλειδί ένα αντικείμενο πέντε πεδίων που προσδιορίζει μοναδικά μια ροή (δηλ. IP διευθύνσεις πηγής και προορισμού, θύρες πηγής και προορισμού, πρωτόκολλο), και τιμή, έναν μετρητή των πακέτων που ανήκουν στη συγκεκριμένη ροή. Για κάθε νέο πακέτο που φτάνει, ο πίνακας κατακερματισμού σαρώνεται και αυξάνει ο μετρητής της αντίστοιχης ροής (στην οποία ανήκει το πακέτο) εάν υπάρχει ήδη, ή δημιουργείται μια νέα καταχώρηση σε περίπτωση νέας ροής. Ο πίνακας κατακερματισμού δημιουργείται περιοδικά, και οι ροές επιλέγονται στο τέλος του χρονικού παραθύρου.

Μια τέτοια διαδικασία μπορεί να εφαρμοστεί εύκολα με τη χρήση μιας κάρτας επεξεργαστή δικτύου (network processor card). Οι σύγχρονες κάρτες επεξεργαστών δικτύου είναι σε θέση να επεξεργάζονται δεδομένα με ταχύτητες που αρχίζουν από 1 Gbps και φτάνουν μέχρι και 10 Gbps. Κάθε καταχώρηση στον πίνακα κατακερματισμού απαιτεί 8 bytes για τις IP διευθύνσεις πηγής και προορισμού, 4 bytes για τις θύρες πηγής και προορισμού, 1 byte για το πρωτόκολλο και 3 bytes για το μετρητή, δηλαδή 16 bytes ανά καταχώρηση. Σε σύνδεση 1 Gbps, υπάρχουν κατά μέσο όρο 100.000 πακέτα ανά δευτερόλεπτο, κατά συνέπεια στη χειρότερη περίπτωση όπου κάθε πακέτο ανήκει σε μια διαφορετική ροή θα χρειαζόμασταν 16 x 100.000 = 1.600.000 bytes (λιγότερο από 1.6 MB μνήμης). Στην περίπτωση όπου έχουμε ένα χρονικό παράθυρο 10 sec θα απαιτούνταν δηλαδή περίπου 16MB μνήμης.

8. Συμπεράσματα – Μελλοντική Έρευνα

Στο κεφάλαιο αυτό συνοψίζουμε τα αποτελέσματα και την προσφορά της διατριβής στον τεχνολογικό τομέα της ανίχνευσης ανωμαλιών δικτύου. Επίσης γίνεται αναφορά σε μερικά θέματα μελλοντικής εργασίας και επέκτασης των ερευνητικών αποτελεσμάτων της διατριβής που παρουσιάζουν ιδιαίτερο πρακτικό και ερευνητικό ενδιαφέρον.

8.1. Συμπεράσματα

Στην παρούσα διδακτορική διατριβή μελετήθηκε η επίδραση της δειγματοληψίας πάνω στο πεδίο της ανίχνευσης ανωμαλιών στο περιβάλλον του Διαδικτύου. Η ερευνητική προσπάθεια σχετικά με την δειγματοληψία που έχει γίνει μέχρι σήμερα αφορά κυρίως τη μελέτη διάφορων στατιστικών στοιχείων κίνησης όπως είναι η κατανομή του μεγέθους των ροών πακέτων, του μέσου μήκους ροής, του συνολικού αριθμού ροών, κτλ. σε ένα δείγμα δεδομένων κίνησης δικτύου. Παρόλο που αυτές οι προσεγγίσεις είναι ενδιαφέρουσες για γενικές διαχειριστικές εφαρμογές σε ένα δίκτυο, δεν επαρκούν για το πεδίο της ανίχνευσης ανωμαλιών.

Το συγκεκριμένο πρόβλημα της επίδρασης της δειγματοληψίας στη διαδικασία ανίχνευσης ανωμαλιών δικτύου είναι αρκετά διαφορετικό και πιο περίπλοκο από τα αντίστοιχα προβλήματα επίδρασης της δειγματοληψίας σε άλλες διαδικασίες διαχείρισης του δικτύου. Αυτό οφείλεται κυρίως στο γεγονός ότι η ανίχνευση ανωμαλιών είναι μία διαδικασία που εφαρμόζεται κάτω από ανώμαλες συνθήκες δικτύου (π.χ. επιθέσεις), ενώ από τη φύση της περιλαμβάνει ταυτόχρονα διάφορους παράγοντες, όπως η κανονική κίνηση του δικτύου, η ανώμαλη κίνηση, τα διάφορα μετρικά ανίχνευσης, των οποίων οι στατιστικές ιδιότητες και η συμπεριφορά μπορεί να επηρεαστεί σε μεγάλο βαθμό και με αρκετά διαφορετικούς τρόπους με την εφαρμογή της δειγματοληψίας.

Σε αυτή τη διατριβή αρχικά έγινε αξιολόγηση στην επίδραση τριών βασικών τεχνικών δειγματοληψίας πακέτων (συστηματική, τυχαία ν-από-Ν, και ομοιόμορφη πιθανολογική τυχαία δειγματοληψία) που έχουν προταθεί στο PSAMP IETF draft σε τρεις ευρέως χρησιμοποιούμενους αλγορίθμους ανίχνευσης ανωμαλιών. Τα

αποτελέσματά έδειξαν ότι οι μέθοδοι δειγματοληψίας πακέτων δεν επαρκούν για αποτελεσματική ανίχνευση ανωμαλιών. Ειδικότερα, η συστηματική δειγματοληψία δεν έχει καλή απόδοση σε χαμηλά ποσοστά δειγματοληψίας όταν εξαρτάται η ανίχνευση της ανωμαλίας από ορισμένα χαρακτηριστικά (π.χ. TCP-flags). Επιπλέον, παρατηρήθηκε ότι όταν χρησιμοποιούνται μετρικά βασισμένα σε ροές, όπως ο αριθμός IP διευθύνσεων πηγής ή ο αριθμός των ροών, η απόδοση του αλγορίθμου ανίχνευσης ανωμαλίας στηρίζεται κυρίως στο ποσοστό δειγματοληψίας που εφαρμόζεται και είναι λιγότερο εξαρτώμενη από την χρησιμοποιούμενη τεχνική δειγματοληψίας.

Στην συνέχεια, προτάθηκε και αναλύθηκε μια μεθοδολογία δειγματοληψίας κατάλληλη για εφαρμογή στο πεδίο της ανίχνευσης ανωμαλιών στο περιβάλλον του Διαδικτύου. Η ιδέα της συγκεκριμένης δειγματοληπτικής μεθόδου βασίζεται στην παρατήρηση ότι οι περισσότερες από τις ανωμαλίες του δικτύου προκαλούνται από μικρές ροές (με μικρό αριθμό πακέτων). Η προτεινόμενη μέθοδος ονομάζεται «Επιλεκτική Δειγματοληψία» (Selective Sampling) και εστιάζει στην επιλογή των μικρών ροών (σε πακέτα) στις οποίες και παρουσιάζονται ανωμαλίες όπως είναι οι επιθέσεις άρνησης υπηρεσίας (DoS attacks) και οι αυτοδιαδιδόμενοι ιοί (worm propagation).

Ένα από τα κύρια χαρακτηριστικά της νέας μεθόδου δειγματοληψίας είναι η προσαρμοστικότητα της σε σχέση με άλλες προσεγγίσεις (π.χ. «Εξυπνη Δειγματοληψία») επειδή μπορεί να ελέγξει περαιτέρω και να μειώσει τον αριθμό των επιλεγμένων ροών. Αφενός, με την κατάλληλη επιλογή της τιμής για την παράμετρο *c* μπορούμε να επιλέξουμε ένα σημαντικό ποσοστό των μικρών ροών χωρίς να μειωθεί η αποτελεσματικότητα στην ανίχνευση ανωμαλιών. Αφ' ετέρου, η επιλογή των μεγάλων ροών μπορεί να μειωθεί περαιτέρω με την αύξηση της τιμής της παραμέτρου *n*.

Η αξιολόγηση της μεθόδου πραγματοποιήθηκε με την εφαρμογή της σε δύο διαφορετικές τεχνικές ανίχνευσης ανωμαλιών σε πραγματικά δεδομένα δικτύου που έχουν συλλεχθεί από τη σύνδεση του δικτύου του ΕΜΠ με το ΕΔΕΤ. Συγκεκριμένα, αξιολογήσαμε την επίδραση αυτής της τεχνικής δειγματοληψίας κάτω από διαφορετικά σενάρια σε μια κατανεμημένη επίθεση DoS (SYN flooding attack) χρησιμοποιώντας μια μέθοδο Ανίχνευσης Αλλαγής Σημείου, καθώς και μία μέθοδο βασισμένη στην Ανάλυση Κύριων Συνιστωσών. Τα πειράματα και τα αποτελέσματά από τις δύο μεθόδους κατέδειξαν ότι ακόμη και με μικρά ποσοστά επίθεσης και

ρυθμούς δειγματοληψίας η αποτελεσματικότητα ανίχνευσης βελτιώνεται σημαντικά, ενώ συγχρόνως ο αριθμός πακέτων που πρέπει να υποβληθούν σε επεξεργασία μειώνεται.

Επίσης, στην παρούσα διατριβή προτάθηκε η ενσωμάτωση της «Επιλεκτικής Δειγματοληψίας» σε μία σύνθετη μέθοδο δειγματοληψίας που ονομάζεται «Δειγματοληψία Δύο Σταδίων» (Two-Stage Sampling) με στόχο την αποτελεσματικότερη ανίχνευση των ανωμαλιών. Η αξιολόγηση της προτεινόμενης σύνθετης μεθόδου έγινε με εφαρμογή της σε μία μέθοδο ανίχνευσης ανωμαλιών που βασίζεται στην έννοια της εντροπίας. Τα πειραματικά αποτελέσματα αποδεικνύουν ότι η προτεινόμενη προσέγγιση βελτιώνει σε μεγάλο βαθμό την αποτελεσματικότητα της ανίχνευσης ανωμαλιών, ενώ συγχρόνως μειώνει τον αριθμό των επιλεγέντων δεδομένων, επιτυγχάνοντας στις περισσότερες περιπτώσεις να ξεπεράσει ακόμη και τα αντίστοιχα αποτελέσματα της περίπτωσης που δεν εφαρμόζεται κάποια μορφή δειγματοληψίας.

Τέλος, η διατριβή ολοκληρώνεται με μία μελέτη της επίδρασης «Ευφυών Δειγματοληπτικών Μεθόδων» (μία από αυτές είναι και η προτεινόμενη «Επιλεκτική Δειγματοληψία») στην ανίχνευση και κατηγοριοποίηση διαφόρων ανωμαλιών δικτύου. Με βάση την παρατήρηση ότι για συγκεκριμένες εφαρμογές όπως η ανίχνευση ανωμαλιών, ένα μεγάλο μέρος των πληροφοριών που μας είναι χρήσιμες περιλαμβάνεται σε ένα μικρό μέρος των ροών, καταδεικνύουμε ότι με τη χρησιμοποίηση των «Ευφυών Τεχνικών Δειγματοληψίας», επιτυγχάνουμε τη μεγέθυνση της εμφάνισης των ανωμαλιών μέσα στο επιλεγέν δείγμα στοιχείων με την επιλογή των κατάλληλων στοιχείων. Με την εφαρμογή αυτών των ευφυών μεθόδων δειγματοληψίας σε πέντε διαφορετικές ανωμαλίες δικτύου (DDoS attack, worm propagation, portscan, flash crowd, alpha flows) αποδεικνύεται ότι επιτυγχάνεται βελτίωση στην αποτελεσματικότητα της ανίχνευσης και σε μερικές περιπτώσεις πετυχαίνουμε την ανίχνευση ανωμαλιών που θα ήταν αόρατες σε άλλη περίπτωση.

8.2. Θέματα μελλοντικής εργασίας

Ενδιαφέρον στοιχείο στην προτεινόμενη μεθοδολογία της «Επιλεκτικής Δειγματοληψίας» το οποίο χρήζει περαιτέρω έρευνας και πειραματισμού είναι η δυναμική αλλαγή των παραμέτρων της μεθόδου, ανάλογα με τη σύνθεση της κίνησης

του δικτύου στην τρέχουσα κατάσταση. Για παράδειγμα, σε ένα περιστατικό κατανεμημένης επίθεσης με μικρές και μεσαίες ροές ίσως το σύστημα δειγματοληψίας να επέλεγε να αυξήσει την τιμή της παραμέτρου *z* (κατώφλι διαχωρισμού ροών σε μικρές και μεγάλες) ώστε να επιλέγονται (εκτός από τις μικρές) και οι μεσαίες ροές με μεγαλύτερη πιθανότητα. Αντίθετα, σε μία περίπτωση που αυξάνουν υπερβολικά οι μικρές ροές στο δίκτυο ίσως το σύστημα θα έπρεπε να μειώνει την παράμετρο *c* με την οποία επιλέγονται οι μικρές ροές, δεδομένου ότι και με την μικρότερη τιμή θα ήταν ικανό να ανιχνεύσει την ανωμαλία. Συνολικά, μία τέτοια προσέγγιση θα οδηγούσε στην ανάπτυξη μιας δυναμικά προσαρμοστικής μεθόδου δειγματοληψίας ανάλογα με την τρέχουσα κατάσταση του δικτύου.

Επίσης, μία άλλη κατεύθυνση μελλοντικής έρευνας πάνω στο πρόβλημα που μελετά η παρούσα διατριβή θα ήταν μία μελέτη σχετικά με τις μεθοδολογίες δειγματοληψίας κατακερματισμού (hash-based sampling) και ποια από αυτές είναι καταλληλότερη για εφαρμογή της στο πεδίο της ανίχνευσης ανωμαλιών σε δίκτυα ευρείας περιοχής (wide area networks). Όπως αναφέραμε σε προηγούμενο κεφάλαιο μία γνωστή μέθοδος αυτού του είδους είναι η «Δειγματοληψία Τροχιάς» (Trajectory sampling) κατά την οποία όλοι οι δρομολογητές του υπό μελέτη δικτύου εφαρμόζουν στα πακέτα του δικτύου την ίδια συνάρτηση κατακερματισμού (hash function) και το ίδιο εύρος τιμών επιλογής. Κατά συνέπεια ένα συγκεκριμένο πακέτο επιλέγεται είτε σε όλα τα σημεία κατά την πορεία του μέσα στο δίκτυο, είτε σε κανένα. Μία άλλη μέθοδος που κάνει εφαρμογή της δειγματοληψίας κατακερματισμού ονομάζεται «Συντονισμένη δειγματοληψία» (Coordinated sampling), και εφαρμόζει δειγματοληψία στις ροές και όχι στα πακέτα όπως η προηγούμενη μέθοδος. Στη μέθοδο αυτή σε διαφορετικά σημεία μέτρησης επιλέγονται και διαφορετικές ροές, οι οποίες δεν θα επιλεγούν σε άλλο σημείο μέσα στο δίκτυο. Αυτό έχει ως αποτέλεσμα τελικά να έχουμε πληροφορία για όσο το δυνατόν περισσότερες ροές που υπάρχουν στο υπό μελέτη δίκτυο. Σε συνδυασμό με το γεγονός ότι κάθε πακέτο κατά την εισαγωγή του στο υπό μελέτη δίκτυο σημαδεύεται με ένα μοναδικό αναγνωριστικό με βάση την IP διεύθυνση πηγής και προορισμού, αποκτούμε για κάθε ροή το πλήρες μονοπάτι που ακολούθησε στο δίκτυο. Η τεχνική αυτή, συνδυαζόμενη με έναν αλγόριθμο ανίχνευσης ανωμαλιών σε δίκτυα ευρείας περιοχής μπορεί να αποτελέσει έναν αποτελεσματικό τρόπο ανίχνευσης του μονοπατιού που ακολούθησε η ανωμαλία μέσα στο υπό μελέτη δίκτυο.

9. Δημοσιεύσεις

9.1. Διεθνή Επιστημονικά Περιοδικά με Κρίση

1. G. Androulidakis, V. Chatzigiannakis and S. Papavassiliou, “Network Anomaly Detection and Classification via Opportunistic Sampling”, IEEE Network, Vol. 23, Issue 1, pp. 6-12, 2009.

2. G. Androulidakis and S. Papavassiliou, “Improving Network Anomaly Detection via Selective Flow-based Sampling”, IET Communications Journal, Vol. 2, Issue 3, pp. 399-409, March 2008.

3. G. Androulidakis and S. Papavassiliou, “Two-Stage Selective Sampling for Anomaly Detection: Analysis and Evaluation”, submitted (June 2009) to Security and Communication Networks (Wiley).

4. V. Chatzigiannakis, S. Papavassiliou and G. Androulidakis, “Improving Network Anomaly Detection Effectiveness via an Integrated Multi-Metric-Multi-Link (M3L) PCA-based Approach”, Security and Communication Networks (Wiley), Vol. 2, Issue 3, pp. 289-304, 2009.

9.2. Πρακτικά Διεθνών Επιστημονικών Συνεδρίων με Κρίση

1. G. Androulidakis and S. Papavassiliou, “Intelligent Flow-based Sampling for Effective Network Anomaly Detection”, in Proc. IEEE Global Communications Conference (GLOBECOM 2007), pp. 1948-1953, Washington D.C., USA, November 2007.

2. G. Androulidakis, V. Chatzigiannakis and S. Papavassiliou, “Using Selective Sampling for the Support of Scalable and Efficient Network Anomaly Detection”, in Proc. IEEE Distributed Autonomous Network Management

Systems Workshop (IEEE DANMS 2007), Washington D.C., USA, November 2007.

3. V. Chatzigiannakis, G. Androulidakis, K. Pelechrinis, S. Papavassiliou and V. Maglaris, “Data fusion algorithms for network anomaly detection: classification and evaluation”, in Proc. Third International Conference on Networking and Services (ICNS’07), pp. 50-57, Athens, Greece, June 2007.

4. Georgios Androulidakis, Vasilis Chatzigiannakis, Symeon Papavassiliou, Mary Grammatikou and Vasilis Maglaris, “Understanding and Evaluating the Impact of Sampling on Anomaly Detection Techniques”, in Proc. IEEE Military Communications Conference (MILCOM’06), Washington D.C., USA, October 2006.

5. V. Chatzigiannakis, S. Papavassiliou, G. Androulidakis, B. Maglaris, “On the realization of a generalized data fusion and network anomaly detection framework”, in Proc. Fifth International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP’06), pp. 251-255, Patras, Greece, July 2006.

6. Georgios Androulidakis, Vasilis Chatzigiannakis, Mary Grammatikou, Fotis Stamatelopoulos, “Network Flow-Based Anomaly Detection of DDoS Attacks”, TERENA Networking Conference 2004, Rhodes, Greece, June 2004.

7. V.Chatzigiannakis, G.Androulidakis, M.Grammatikou, B.Maglaris, “An Architectural Framework for Distributed Intrusion Detection using Smart Agents”, International Conference on Security and Management (SAM ’04), Las Vegas, Nevada, USA, 2004.

10. Βιβλιογραφία

[Andr06] G. Androulidakis, V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, V. Maglaris, “Understanding and Evaluating the Impact of Sampling on Anomaly Detection Techniques”, IEEE Military Communications Conference (MILCOM 2006), Washington D.C., USA, October 2006.

[Andr07] G. Androulidakis and S. Papavassiliou, “Intelligent Flow-based Sampling for Effective Network Anomaly Detection”, IEEE Global Communications Conference (GLOBECOM 2007), Washington D.C., USA, November 2007.

[Andr08] G. Androulidakis and S. Papavassiliou, “Improving Network Anomaly Detection via Selective Flow-based Sampling”, IET Communications Journal, Vol. 2, No. 3, March 2008.

[Andr09] G. Androulidakis, V. Chatzigiannakis and S. Papavassiliou, “Network Anomaly Detection and Classification via Opportunistic Sampling”, IEEE Network (Special Issue on Recent Developments in Network Intrusion Detection), Vol. 23, No. 1, 2009.

[Ari03] I. Ari, B. Hong, E.L. Miller, S.A. Brandt, D.D.E Long, “Managing flash crowds on the Internet”, in Proc. of the 11th IEEE International Symposium on Modeling, Analysis and Simulation of Computer Telecommunications Systems (MASCOTS’03), Orlando, Florida, USA, October 2003, pp. 246-249.

[Barf01] P. Barford and D. Plonka, ”Characteristics of network traffic flow anomalies”, In Proc. of the First ACM SIGCOMM Internet Measurement Workshop, pp. 69–74, 2001.

[Barf02] P. Barford, J. Kline, D. Plonka, and A. Ron, “A Signal Analysis of Network Traffic Anomalies”, in Proc. of the 2nd ACM SIGCOMM Workshop on Internet measurement, pp. 71-82, 2002.

[Blaz01] R. B. Blazek, H. Kim, B. Rozovskii, and A. Tartakovsky, “A novel approach to detection of Denial of Service attacks via adaptive sequential and batch sequential change point detection methods”, in Proc. of IEEE Workshop Information Assurance and Security, New York, USA, June 2001, pp. 220–226.

[Bosc06] E. Boschi, S. Denazis, T. Zseby, “A measurement framework for inter-domain SLA validation”, Computer Communications, Vol. 29, No. 6, March 2006, pp. 703-716.

[Brau06] D. Brauckhoff, B. Tellenbach, A. Wagner, M. May and A. Lakhina, “Impact of Packet Sampling on Anomaly Detection Metrics”, Internet Measurement Conference 2006, Rio de Janeiro, Brazil, October 2006.

[Brid00] S.M. Bridges, R.B. Vaughn, “Fuzzy data mining and genetic algorithms applied to intrusion detection”, in Proc. of the National Information Systems Security Conference, Baltimore, MD, 2000.

[Calv98] R.A. Calvo, M. Partridge, M.A. Jabri, “A comparative study of principal component analysis techniques”, in Proc. of the Ninth Australian Conference on Neural Networks, Brisbane, Qld, Australia, 1998.

[Chat06] V. Chatzigiannakis, S. Papavassiliou, G. Androulidakis, B. Maglaris, “On the realization of a generalized data fusion and network anomaly detection framework”, Fifth International Symposium on Communication Systems, Networks and Digital Signal Processing

(CSNDSP'06), Patras, Greece, July 2006.

[Chat07a] V. Chatzigiannakis, S. Papavassiliou, "Diagnosing Anomalies and Identifying Faulty Nodes in Sensor Networks", IEEE Sensors Journal , Vol. 7, Issue 5, pp. 637-645, 2007

[Chat07b] V. Chatzigiannakis, M. Grammatikou, S. Papavassiliou, "Extending Driver's Horizon Through Comprehensive Incident Detection in Vehicular Networks", IEEE Transactions on Vehicular Technology, Vol. 56, Issue 6, pp.3256-3265, 2007

[Chat08] V. Chatzigiannakis, G. Androulidakis, S. Papavassiliou, "Improving Network Anomaly Detection Effectiveness via an Integrated Multi-Metric-Multi-Link (M3L) PCA-based Approach", Security and Communication Networks (Wiley), 2008.

[Chen05] S. Chen and S. Ranka, "Detecting Internet worms at early stage", IEEE Journal on Selected Areas in Communications, Vol. 23, No. 10, October 2005, pp. 2003-2012.

[Choi02] Baek-Young Choi, Jaesung Park, and Zhi-Li Zhang, "Adaptive random sampling for load change detection", in Proc of ACM SIGMETRICS, July 2002.

[Choi04] B.Y. Choi, J. Park, Z.L. Zhang, "Adaptive Packet Sampling for Accurate and Scalable Flow Measurement", IEEE Global Telecommunications Conference (GLOBECOM'04), 2004.

[Choi05] Baek-Young Choi and Supratik Bhattacharyya, "On the Accuracy and Overhead of Cisco Sampled NetFlow", ACM SIGMETRICS Workshop on Large Scale Network Inference (LSNI'05), Banff, Canada, June 2005.

[Choi07] Baek-Young Choi and Zhi-Li Zhang, “Adaptive random sampling for trac volume measurement”, Telecommunication Systems Journal, Vol. 34, No. 1, pp. 71-80, Feb 2007.

[CisNet] Cisco NetFlow
http://www.cisco.com/warp/public/732/netflow/index.html

[Claf93] K.C. Claffy, G.C. Polyzos, and H.-W. Braun. “Application of Sampling Methodologies to Network Traffic Characterization”, In Proceedings of ACM SIGCOMM’93, San Francisco, CA, pp. 13–17, September 1993.

[Coch87] W. Cochran, “Sampling Techniques”, John Wiley & Sons, 1987.

[Cohe95] W. Cohen, “Fast effective rule induction”, in Proc. of the 12th International Conference on Machine Learning, Tahoe City, CA, 1995, pp. 115–123.

[Cove06] T. Cover and J. Thomas, “Elements of Information Theory”, Wiley & Sons, Second Edition, June 2006.

[Deri00] L. Deri and S. Suin, “Effective Traffic Measurement Using ntop”, IEEE Com. Mag., vol.38, no.5, pp. 138-143, May 2000.

[Doul04] C. Douligeris and A. Mitrokotsa, “DDoS attacks and defense mechanisms: classification and state-of-the-art”, Computer Networks, Vol. 44, No. 5, pp. 643-666, 2004.

[Drob98] J. Drobisz and Kenneth J. Christensen, “Adaptive sampling methods to determine network traffic statistics including the hurst parameter”, in Proce. of the 23rd Conference on Local Computer Networks, Boston, Massachusetts, USA, October 1998.

[Duff01] N. Duffield and M. Grossglauser, “Trajectory sampling for direct traffic observation”, IEEE/ACM Transactions on Networking, Vol. 9, Issue 3, 2001, pp. 280-292.

[Duff01b] N. Duffield, C. Lund and M.Thorup, “Charging from sampled network usage”, in Proc. of the 1st ACM SIGCOMM Workshop on Internet Measurement, San Francisco, California, USA, 2001.

[Duff03] N.G. Duffield and C. Lund, “Predicting Resource Usage and Estimation Accuracy in an IP Flow Measurement Collection Infrastructure”, ACM SIGCOMM Internet Measurement Conference 2003, Miami Beach, Fl, October 27-29, 2003.

[Duff05] N. Duffield, C. Lund, and M. Thorup, “Estimating Flow Distributions From Sampled Flow Statistics”, IEEE/ACM Transactions on Networking, 2005, vol. 13, no. 5, pp. 933-946.

[Esta02] C. Estan and G. Varghese, “New Directions in Traffic Measurement and Accounting”, In Proc. of ACM SIGCOMM’02, Pittsburgh, Pennsylvania, USA, Aug. 2002.

[Esta04] C. Estan, K. Keys, D. Moore, G. Varghese, “Building a Better Netflow”, Proceedings of the ACM SIGCOMM’04, Portland, Oregon, USA, August 2004

[Fang99] Wenjia Fang and Larry Peterson, “Inter-AS traffic patterns and their implications”, Technical Report TR-598-99, Princeton University, Computer Science Department, March 1999

[Feld00] Anja Feldmann, Albert G. Greenberg, Carsten Lund, Nick Reingold, Jennifer Rexford, and Fred True, “Deriving traffic demands for

operational IP networks: methodology and experience", in Proc. of ACM SIGCOMM, Stockholm, Sweden, pp. 257-270, 2000.

[Feld00] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, "NetScope: Traffic engineering for IP networks", IEEE Network, Vol. 14 No. 2, 2000, pp. 11–19.

[Feld01] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, F. True, "Deriving traffic demands for operational IP networks: Methodology and experience", IEEE/ACM Transactions on Networking, Vol. 9, No.1, pp. 265–279.

[Forr96] S. Forrest, S.A. Hofmeyr, A. Somayaji, T.A. Longstaff, "A sense of self for unix processes", in:Proc. of the IEEE Symposium on Research in Security and Privacy, Oakland, CA, USA, 1996, pp. 120–128.

[Ghos99] A.K. Ghosh, A. Schwartzbard, "A study in using neural networks for anomaly and misuse detection", in Proc. of the 8th USENIX Security Symposium, Washington, DC, 1999, pp. 141–151.

[GRNet] Greek Research and Technology Network – http://www.grnet.gr.

[Hern01] Edwin A. Hernandez, Matthew C. Chidester, and Alan D. George, "Adaptive sampling for network management", Journal of Network and Systems Management, Vol. 9, No. 4, pp. 409-434, 2001.

[Hohn06] N. Hohn, and D. Veitch, "Inverting sampled traffic", IEEE/ACM Transactions on Networking, 2006, vol. 14, no. 1, pp. 68-80.

[Hosm93] H.H. Hosmer, "Security is fuzzy: applying the fuzzy logic paradigm to the multipolicy paradigm", in: Proc. of the Workshop on New Security

Paradigms, Little Compton, RI, United States, 1993.

[Jack79] J. E. Jackson and G. S. Mudholkar, “Control Procedures for Residuals Associated with Principal Component Analysis”, Technometrics, pp. 341–349, 1979.

[Jedw92] Jonathan Jedwab and Peter Phaal, “Traffic estimation for the largest sources on a network, using packet sampling with limited storage”, Technical Report HPL-92-35, Hewlett Packard Laboratories, March, 1992.

[Joll02] I.T. Jollife. “Principal Component Analysis”, Second Edition, Springer, 2002.

[Kawa07] Ryoichi Kawahara, Tatsuya Mori, Noriaki Kamiyama, Shigeaki Harada, and Shoichiro Asano, “A study on detecting network anomalies using sampled flow statistics”, in Proc of International Symposium on Applications and the Internet Workshops, Washington D.C., USA, 2007.

[Krue03] C. Kruegel, D. Mutz, W. Robertson, F. Valeur, “Bayesian event classification for intrusion detection”, in Proc. of the 19th Annual Computer Security Applications Conference, Las Vegas, NV, 2003.

[Kuma06] A. Kumar and J. Xu. “Sketch Guided Sampling: Using On-Line Estimates of Flow Size for Adaptive Data Collection”, In Proc. of IEEE Infocom 2006, Barcelona, Spain, April 2006.

[Lakh04a] A. Lakhina, M. Crovella, C. Diot, “Diagnosing network-wide traffic anomalies”, in Proc. of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (ACM SIGCOMM), pp. 219 - 230, Portland, OR, USA, August 2004.

[Lakh04b] A. Lakhina, M. Crovella and C. Diot, “Characterization of Network-Wide Anomalies in Traffic Flows”, in Proc. of ACM/SIGCOMM Internet Measurement Conference (IMC) '04, Taormina, Sicily, Italy, October 2004.

[Lakh05] A. Lakhina, M. Crovella, and C. Diot “Mining Anomalies Using Traffic Feature Distributions”, In Proc. of ACM SIGCOMM 2005, pp. 217–228, Philadelphia, PA, USA, August 2005.

[Lee00] W. Lee, R.A. Nimbalkar, K.K. Yee, S.B. Patil, P.H. Desai, T.T. Tran, S.J. Stolfo, “A data mining and CIDF based approach for detecting novel and distributed intrusions”, in Proc of the 3rd International Workshop on Recent Advances in Intrusion Detection (RAID 2000), Toulouse, France, 2000, pp. 49–65.

[Lee01] W. Lee and D. Xiang, Information-Theoretic Measures for Anomaly Detection, In Proc. of the IEEE Symposium on Security and Privacy (S&P 2001), pp. 130 -143, 2001.

[Lee99] W. Lee, S.J. Stolfo, K.W. Mok, “A data mining framework for building intrusion detection models”, in Proc. of the IEEE Symposium on Security and Privacy, Oakland, CA, 1999, pp. 120–132.

[Mai06a] J. Mai, A. Sridharan, C.N. Chuah, H. Zang and T. Ye, “Impact of Packet Sampling on Portscan Detection”, IEEE Journal on Selected Areas in Communication, vol. 24, no 12, pp. 2285-2298, 2006.

[Mai06b] J. Mai, A. Sridharan, C.N. Chuah, H. Zang and T. Ye, “Is sampled data sufficient for anomaly detection?”, Internet Measurement Conference 2006, Rio de Janeiro, Brazil, October 2006.

[Maxi90] R. Maxion and F. E. Feather, “A case study of ethernet anomalies in a distributed computing environment,” IEEE Transactions on Reliability, Vol. 39, No. 4, pp. 433–443, 1990.

[McGr00] T. McGregor, H.W. Braun, and J. Brown, “The NLANR Network Analysis Infrastructure”, IEEE Communications Magazine, vol.38, no.5, pp. 122-128, May 2000.

[Mirk04] J. Mirkovic and P. Reiher, “A Taxonomy of DDoS Attack and DDoS Defense Mechanisms”, ACM SIGCOMM Computer Communication Review, vol.34 no.2, April 2004, pp. 39-53.

[Mock87] P. Mockapetris, “DOMAIN NAMES - CONCEPTS AND FACILITIES”, RFC 1034.

[Moor03] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, N. Weaver, “Inside the Slammer worm”, IEEE Security & Privacy Magazine, Volume 1, Issue 4, pp. 33 – 39, July-Aug 2003.

[MSSQL] Microsoft SQL Server - http://www.microsoft.com/sql/default.mspx.

[Mukh94] B. Mukherjee, L.T. Heberlein, and K.N. Levitt, “Network intrusion detection”, IEEE Network, Vol. 8, No. 3, pp. 26-41, 1994.

[NetFl] Cisco NetFlow, http://www.cisco.com/warp/public/732/netflow/index.html

[Nsfnt] NSFNET http://www.nsf.gov/about/history/nsf0050/internet/launch.htm

[Peng04] T. Peng, C. Leckie, and K. Ramamohanarao, “Proactively Detecting Distributed Denial of Service Attacks Using Source IP Address

Monitoring", In Proc. of the Third International IFIP-TC6 Net-working Conference, Athens, Greece, May 2004.

[Porr97] P.A. Porras, P.G. Neumann, "EMERALD: event monitoring enabling responses to anomalous live disturbances", in Proc. of the 20th NIST-NCSC National Information Systems Security Conference, Baltimore, MD, USA, 1997, pp. 353–365.

[Psamp] Packet Sampling (PSAMP) IETF Working Group Charter. http://www.ietf.org/html.charters/psamp-charter.html

[Quin93] J.R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufman, Los Altos, CA, 1993.

[Rama03] M. Ramadas, S.O.B. Tjaden, "Detecting anomalous network traffic with self-organizing maps", in Proc. of the 6th International Symposium on Recent Advances in Intrusion Detection, Pittsburgh, PA, USA, 2003, pp. 36–54.

[Ranj07] S. Ranjan, S. Shah, A. Nucci, M. Munafo, R. Cruz, S. Muthukrishnan, "DoWitcher: Effective Worm Detection and Containment in the Internet Core", 26th IEEE International Conference on Computer Communications (INFOCOM 2007), pp. 2541-2545, Anchorage, Alaska, USA, May 2007.

[Reev02] J. Reeves, S. Panchein, "Traffic monitoring with packet-based sampling for defence against security theats", in Proc. of Passive and Active Measurement Workshop, Fort Collins, Colorado, USA, March 2002.

[Roes99] M. Roesch, "Snort – lightweight intrusion detection for networks", in Proc. of the 13th USENIX Conference on System Administration Seattle, Washington, 1999, pp. 229–238

[Sarv01] S. Sarvotham, R. Riedi and R. Baraniuk, “Connection-level analysis and modeling of network traffic”, in Proc. of the 1st ACM SIGCOMM Workshop on Internet Measurement, San Diego, CA, USA, November 2001.

[Sava00] S. Savage, D. Wetherall, A. R. Karlin, and T. Anderson, “Practical network support for ip traceback,” in Proc. ACM SIGCOMM, 2000, pp. 295–306.

[Seka06] V. Sekar, N.G. Duffield, O. Spatscheck, J.E. van der Merwe, H. Zhang, “LADS: Large-scale Automated DDoS Detection System”, In Proc. of 2006 USENIX Annual Technical Conference, Boston, USA, pp. 171-184, June 2006.

[Seka07] V. Sekar, M. Reiter, W. Willinger, H. Zhang, “Coordinated Sampling: An Efficient Network-Wide Approach for Flow Monitoring”, Technical Report, CMU-CS-07-139, July 2007.

[Siri04] V. Siris and F. Papagalou, “Application of anomaly detection algorithms for detecting SYN flooding attacks”, in Proc. of IEEE GLOBECOM '04, Dallas, Texas, USA, 2004. pp. 2050–2054

[Snmp] Simple Network Management Protocol (SNMP) - RFC 1157.

[Srid06] A. Sridharan, T. Ye, and S. Bhattacharyya, “Connectionless Port Scan Detection on the Backbone,” in Malware Workshop (in conjunction with IEEE IPCCC 2006), Phoenix, Arizona, USA, April 2006.

[Stan02] S. Staniford, J.A. Hoagland, J.M. McAlerney, “Practical automated detection of stealthy portscans”, Journal of Computer Security, Vol. 10, 2002, pp. 105–136.

[Tayl01] C. Taylor, J. Alves-Foss, “NATE: Network Analysis of anomalous Traffic Events”, in Proc. of the ACM Workshop on New Security Paradigms, 2001, pp. 89-96.

[Thot01] M. Thottan and C. Ji, “Using network fault predictions to enable IP traffic management”, Journal of Network and Systems Management, Vol. 9, No. 3, 2001.

[Vald00] A. Valdes, K. Skinner, “Adaptive model-based monitoring for cyber attack detection”, in Proc. of Recent Advances in Intrusion Detection (RAID), Toulouse, France, 2000, pp. 80–92.

[Vign98] G. Vigna and R. A. Kemmerer, “Netstat: A network based intrusuion detection approach,” in Proc. of 14th Annual Computer Security Applications Conference, Scottsdale, Arizona, USA, 1998

[Wang04] H. Wang, D. Zhang, and K.G. Shin, “Change-Point Monitoring for the Detection of DoS Attacks”, IEEE Transactions on Dependable and Secure Computing, 2004, Vol. 1, No. 4, pp. 193-208.

[Wang06] W. Wang, R. Battiti, “Identifying intrusions in computer networks with principal component analysis”, in The First International Conference on Availability, Reliability and Security, Vienna, Austria, 2006, pp. 270–279.

[Weav03] N. Weaver, V. Paxson, S. Staniford, and R. Cunningham, “A Taxonomy of Computer Worms”, In Proc. of the First ACM Workshop on Rapid Malcode (WORM), Washington DC, USA, October 2003.

[Xu05a] Li-Bo Xu, Guo-Xin Wu, Jian-Fei Li, “Packet-Level Adaptive Sampling on Multi-Fluctuation Scale Traffic”, Proceedings of International

Conference on Communications, Circuits and Systems, pp. 604 - 608 Vol. 1, May 2005.

[Xu05b] K. Xu, Z.L. Zhang, and S. Bhattacharyya, “Profiling Internet Backbone Traffic: Behavior Models and Applications”, in Proc. of the ACM SIGCOMM’05, Philadelphia, Pennsylvania, USA, August 2005.

[Yang00] J. Yang, P. Ning, X. S. Wang, and S. Jajodia, “Cards: A distributed system for detecting coordinated attacks,” in Proc. 16th Annual Working Conference on Information Security, 2000, pp. 171–180.

[Yang07] L. Yang and G. Michailidis, “Sampled Based Estimation of Network Traffic Flow Characteristics”, 26th IEEE International Conference on Computer Communications (INFOCOM 2007), pp. 1775-1783, Anchorage, Alaska, USA, May 2007.

[Ye00] N. Ye, M. Xu, S.M. Emran, “Probabilistic networks with undirected links for anomaly detection”, in Proc. of IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop, West Point, NY, 2000.

[Ye02] N. Ye, S. Emran, Q. Chen, S. Vilbert, “Multivariate Statistical Analysis of Audit Trails for Host-Based Intrusion Detection”, IEEE Transactions on Computers, Vol. 51, No. 7, July 2002.

[Yu06] W. Yu, X. Wang, D. Xuan, D. Lee, “Effective Detection of Active Worms with Varying Scan Rate”, Second International Conference on Security and Privacy in Communication Networks, (IEEE SecureComm 2006), Baltimore, MD, USA, August 2006.

[Zseb03] T. Zseby, “Stratification strategies for sampling-based non-intrusive

measurements of one-way delay", in Proc. of Passive and Active Measurement Workshop, La Jolla, CA, USA, April 2003.