

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή στα Πληροφοριακά Συστήματα



**Εφαρμογή και Αξιολόγηση Τεχνικής Εξόρυξης Συχνών Στοιχείων
στη Σύνοψη και Παρακολούθηση της Κίνησης Δικτύων
Υπολογιστών**

Στυλιανός Μαυροματίδης

**Επιβλέπων Καθηγητής
Ξενοφώντας Δημητρόπουλος**

Μάιος 2012

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

**Εφαρμογή και Αξιολόγηση Τεχνικής Εξόρυξης Συχνών Στοιχείων
στη Σύνοψη και Παρακολούθηση της Κίνησης Δικτύων
Υπολογιστών**

Στυλιανός Μαυροματίδης

**Επιβλέπων Καθηγητής
Ξενοφώντας Δημητρόπουλος**

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε
προς μερική εκπλήρωση των απαιτήσεων για απόκτηση

μεταπτυχιακού τίτλου σπουδών
στα Πληροφοριακά Συστήματα

από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου

Μάιος 2012

Περίληψη

Η επιτακτική ανάγκη για αρτιότερη και πληρέστερη διαχείριση δικτύων υπολογιστών με στόχο την εξασφάλιση της αδιάλειπτης λειτουργίας τους έχει οδηγήσει στην αναζήτηση νέων αποδοτικότερων μεθόδων παρακολούθησης, εντοπισμού και ανάλυσης δικτυακών προβλημάτων και γεγονότων που εμφανίζονται σε αυτά. Μία φιλόδοξη μέθοδος η οποία μπορεί να συνδράμει προς την κατεύθυνση αυτή είναι η χρήση τεχνικών εξόρυξης «κανόνων συσχέτισης» (association rules) και «συχνών στοιχείων» (frequent items). Οι τεχνικές αυτές αξιοποιούνται για τη δημιουργία μίας ολοκληρωμένης εικόνας της κίνησης που διέρχεται από το υπό παρατήρηση δίκτυο υπολογιστών καθώς και για τον εντοπισμό προβλημάτων-ανωμαλιών. Για την εξαγωγή ασφαλών συμπερασμάτων σημαντικό ρόλο διαδραματίζουν οι τεχνικές γραφικής αναπαράστασης-οπτικοποίησης (visualization) των παραγόμενων «συχνών στοιχείων».

Η παρούσα μεταπτυχιακή διατριβή παρουσιάζει μία νέα προσέγγιση στην απεικόνιση της κυκλοφορίας δικτύων υπολογιστών η οποία χρησιμοποιεί τεχνικές εξόρυξης «συχνών στοιχείων» πριν την απεικόνιση, για να συνοψίσει ένα μεγάλο αριθμό ροών δεδομένων. Στο παρελθόν αντίστοιχες προσεγγίσεις απεικονίζουν επιλεγμένες ή όλες τις συλλέχθεντες ροές δεδομένων γεγονός το οποίο δεν μπορεί να επεκταθεί σε μεγάλο αριθμό ροών. Με τη χρήση αλγορίθμων παραγωγής «συχνών στοιχείων» ο όγκος της πληροφορίας που απεικονίζεται είναι σχεδόν ανεξάρτητος από τον όγκο της εισόδου. Πραγματοποιήθηκε εφαρμογή, αξιολόγηση και λεπτομερείς σύγκριση δύο τεχνικών που παρουσιάζουν γραφικά τις συχνά εμφανιζόμενες ομάδες αντικειμένων στην παρακολούθηση-σύνοψη της κίνησης δικτύων υπολογιστών, της τεχνικής γραφημάτων παράλληλων συντεταγμένων και της τεχνικής γράφων. Τα δεδομένα που χρησιμοποιήθηκαν για τη σύγκριση προέρχονται από πραγματικό δίκτυο εταιρείας. Πραγματοποιήθηκε αξιολόγηση της ποιότητας και της χρησιμότητας των παραγόμενων συνόψεων καθώς και των τεχνικών αναπαράστασης. Οι παρατηρήσεις έδειξαν ότι η δημιουργία γράφων για την αναπαράσταση «συχνών στοιχείων» και κυρίως “maximal frequent itemsets” μπορεί να αποτελέσει χρήσιμο εργαλείο στην γενικότερη παρακολούθηση ενός δικτύου υπολογιστών. Μπορεί να παρέχει γρήγορα και εύκολα την πληροφορία ενώ παράλληλα μπορεί να χρησιμοποιηθεί για τον γρήγορο εντοπισμό κάποιων τύπων επιθέσεων, αποτελώντας έτσι τη βάση για τη δημιουργία εργαλείων ανίχνευσης ανωμαλιών.

Summary

The urgent need for optimal management of computer networks in order to ensure their uninterrupted operation requires new efficient methods for monitoring, tracking and analyzing network traffic. One ambitious method which can assist network administrators is the use of association rules and frequent-item mining. These techniques can be used to create a comprehensive picture of the observed network traffic and can identify anomalies. For reliable conclusions one can use graphical representation techniques (visualization) of produced frequent items.

This thesis presents a new approach to display computer networks traffic which uses frequent itemset mining before visualization to summarize a large number of traffic flows. In the past, similar approaches visualize each individual flow, which does not scale to a large number of flows. Using frequent item-set mining the amount of information shown is almost independent of the number of input flows. We implement, evaluate and conduct a detailed comparison of two frequent item-set visualization techniques based on parallel-coordinate and graph based plots. The traffic data used for comparison come from a real enterprise network. The quality and usefulness of the generated frequent itemsets and of their visualization were also evaluated. Observations showed that the graph-based visualization of the frequent items and especially the visualization of maximal frequent itemsets is a useful tool for monitoring computer networks. It easily and quickly provides the needed information and it can be used to detect some types of attacks, thus forming the basis for creating anomaly detection tools.

Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέποντα καθηγητή μου κ. Ξενοφώντα Δημητρόπουλο για την πολύτιμη καθοδήγησή του. Οι αξιόλογες συμβουλές, υποδείξεις και οι εύστοχες διορθώσεις του συνέβαλαν στην άρτια ολοκλήρωση της παρούσας μεταπτυχιακής διατριβής.

Ιδιαίτερα θερμές ευχαριστίες απευθύνω στην οικογένειά μου για την ατελείωτη συμπαράστασή, στήριξη και κατανόηση που μου έχουν προσφέρει απλόχερα τόσα χρόνια.

Περιεχόμενα

Περίληψη	2
Summary	3
Ευχαριστίες	4
Περιεχόμενα	5
Η Σημασία της Παρακολούθησης στα Δίκτυα Υπολογιστών	8
1.1 Εισαγωγή	9
1.2 Απειλές της Ομαλής Λειτουργίας Ενός Δικτύου Υπολογιστών	11
1.2.1 Computer Virus, Worms, Trojan Horses	11
1.2.2 Port Scan Attack	13
1.2.3 Brute Force Attacks	14
1.2.4 Denial-of-Service Attack, DDoS Attack, Buffer Overflow	14
1.2.5 Υπερβολική Χρήση Ιστοσελίδων (Spam Πληροφορίες)	15
1.2.6 Bandwidth Congestion	16
1.2.7 Δυσλειτουργία Συσκευών	16
1.2.8 Υπερβολικό Φορτίο Εργασίας (Διαμοιρασμός Υπηρεσιών)	17
1.2.9 Απότομες αλλαγές στην ποσότητα των δεδομένων	17
1.3 Τρόποι Προφύλαξης-Αντιμετώπισης Κινδύνων	18
1.4 Η Συνεισφορά της Παρακολούθησης της Κίνησης στην Ομαλή Λειτουργία Ενός Δικτύου Υπολογιστών	20
1.5 Τρόποι Παρακολούθησης σε Ένα Δίκτυο Υπολογιστών	29
1.5.1 SNMP	30
1.5.2 IDS – IPS	33
1.5.3 Snort	35
1.5.4 Συστήματα Σύλληψης και Ανάλυσης Πακέτων (sniffers)	36
1.5.5 Bandwidth Monitoring	39
1.6 Μέθοδοι Γραφικής Αναπαράστασης (Οπτικοποίησης) Καταγεγραμμένων Δεδομένων Κίνησης Δικτύου Υπολογιστών – Πλεονεκτήματα που Προσφέρουν	43

1.7 Συμπεράσματα	52
Χρήση Τεχνικών Εξόρυξης Δεδομένων στη Σύνοψη Δικτυακής κίνησης.....	54
2.1 Εισαγωγή.....	55
2.2 Διαδικασία Εξόρυξης δεδομένων	58
2.3 Ορισμοί που Σχετίζονται με τη Διαδικασία “Data Mining”	61
2.4 Αλγόριθμοι για την Παραγωγή Association Rules και Frequent Itemsets.....	67
2.4.1 Περιγραφή Αλγορίθμου Apriori.....	67
2.5 «Εξόρυξη Δεδομένων» στη Διαχείριση Δικτύων Υπολογιστών	71
2.6 Εφαρμογές Διαδικασίας Εξόρυξης Δεδομένων στα Δίκτυα Υπολογιστών	73
2.7 Συμπεράσματα	80
Πειραματική Εγκατάσταση και Παραγόμενα Δεδομένα.....	82
3.1 Εισαγωγή.....	83
3.2 Υφιστάμενη Κατάσταση Δικτύου Υπολογιστών Εταιρείας στο Οποίο θα Εφαρμοστεί Αλγόριθμος «Συχνών Στοιχείων».....	84
3.3 Σημείο Συλλογής Δεδομένων Κίνησης – Είδος Δεδομένων (Ροές Δεδομένων-Netflows).....	88
3.4 Απομόνωση Δικτύου Παρακολούθησης.....	92
3.5 Παραμετροποίηση Συσκευής Παρακολούθησης (Router)	94
3.6 Εργαλείο Συλλογής Δεδομένων	97
3.7 Υλοποίηση και Εργαλεία για την Εξαγωγή Συχνών Στοιχείων (Λογισμικό R)	99
3.8 Παράθεση Πειραματικών Δεδομένων - Αποτελεσμάτων.....	108
3.9 Support και Αριθμός Παραγόμενων «Συχνών Στοιχείων».....	129
Τεχνικές Οπτικής Αναπαράστασης «Συχνών Στοιχείων».....	154
4.1 Εισαγωγή.....	155
4.2 Τεχνικές Εμφάνισης Συχνά Εμφανιζόμενων Προτύπων	156
4.2.1 Γραφήματα Παράλληλων Συντεταγμένων (Parallel Coordinates).....	157
4.2.2 Γραφήματα Δύο Διαστάσεων	161
4.2.3 Graph-Based Visualizations.....	165
4.3 Εφαρμοσιμότητα Μεθόδων Οπτικοποίησης στην Αναπαράσταση «Συχνών Στοιχείων» που Προέρχονται από Δεδομένα Κίνησης Δικτύου Υπολογιστών.	169

Μια Πρώτη Προσέγγιση στην Οπτική Αναπαράσταση της Κίνησης Δικτύου Υπολογιστών με τη Χρήση «Συχνών Στοιχείων»	172
5.1 Εισαγωγή	173
5.2 Αποτελέσματα Γραφικής Αναπαράστασης «Συχνών Στοιχείων» με τη Χρήση Παράλληλων Συντεταγμένων	174
5.3 Αποτελέσματα Γραφικής Αναπαράστασης “Maximal Frequent Itemsets” και “Closed Frequent Itemsets” με τη Χρήση Παράλληλων Συντεταγμένων.	197
5.4 Συμπεράσματα που Αφορούν τα Γραφήματα Αναπαράστασης Παράλληλων Συντεταγμένων	213
5.4 Αποτελέσματα Γραφικής Αναπαράστασης «Συχνών Στοιχείων» με τη Χρήση της Τεχνικής “Graph Based Visualization”	214
5.5 Αποτελέσματα Γραφικής Αναπαράστασης “Maximal Frequent Itemsets” και “Closed Frequent Itemsets” με τη Χρήση της Τεχνικής “Graph Based Visualization”	229
5.6 Συμπεράσματα που Αφορούν τους γράφους	240
5.7 Συμπεράσματα	242
 Συμπεράσματα	244
6.1 Εισαγωγή	245
6.2 Πλεονεκτήματα και Μειονεκτήματα Παράλληλων Συντεταγμένων	245
6.3 Πλεονεκτήματα και Μειονεκτήματα των γράφων.	247
6.4 Προτάσεις για τη Βελτίωση (Καλύτερη Προσαρμογή σε Δεδομένα Κίνησης Δικτύου Υπολογιστών) των Τεχνικών Αναπαράστασης	249
6.4.1 Προτάσεις Τροποποιήσεων στα Γραφήματα Παράλληλων Συντεταγμένων	249
6.4.2 Προτάσεις Τροποποιήσεων των γράφων.	251
6.5 Επίλογος	252
 Βιβλιογραφία	255

Κεφάλαιο 1

Η Σημασία της Παρακολούθησης στα Δίκτυα Υπολογιστών.

Στην ενότητα αυτή θα πραγματοποιηθεί περιγραφή της σημασίας που έχει στην ομαλή λειτουργία καθώς και στην αρτιότερη διαχείριση ενός δικτύου υπολογιστών η παρακολούθηση της κίνησης του δικτύου και η καταγραφή των συμβάντων που παρατηρούνται σε αυτό. Η ενότητα αυτή καταγράφει τους συνηθέστερους τρόπους παρακολούθησης της κίνησης ενός δικτύου υπολογιστών και πραγματοποιεί μία μικρή εισαγωγή στην πρόσθετη αξία που μπορεί να προσδώσουν οι τεχνικές αναπαράστασης δεδομένων στην παρακολούθηση αυτή.

Κύριος στόχος αυτής της ενότητας είναι να τονιστεί η σημασία που μπορεί να έχει η παρακολούθηση της κίνησης σε ένα δίκτυο υπολογιστών.

1.1 Εισαγωγή

Γνωρίζουμε όλοι ότι ο όρος «Δίκτυο υπολογιστών» δηλώνει ένα σύνολο από υλικό (hardware) και λογισμικό (software) το οποίο συνεργάζεται με τέτοιο τρόπο ώστε να πετυχαίνεται ευκολότερη και γρηγορότερη ανταλλαγή πληροφοριών. Η ανάπτυξη και η χρήση των δικτύων υπολογιστών για την ανταλλαγή πληροφοριών στις μέρες μας είναι ραγδαία, με συνέπεια όλο και περισσότερα προσωπικά-ευαίσθητα δεδομένα να διακινούνται μέσα από αυτά. Στις μέρες μας δεν υφίσταται εταιρεία η οποία να μην χρησιμοποιεί και κατά συνέπεια να εκμεταλλεύεται τα πλεονεκτήματα που παρέχει η δικτύωση υπολογιστών. Ανάλογα με το μέγεθός της κάθε εταιρεία διαθέτει μικρό ή μεγάλο αριθμό υπολογιστών συνδεδεμένων σε κάποιο τοπικό δίκτυο. Home network, Local Area Networks (LAN), Personal Area Networks (PAN), Campus Network, Metropolitan Area Network (MAN), Wide Area Network (WAN), Enterprise Private Network, Virtual Private Networks, Intranet, Extranet, Internet [1] είναι κάποιοι από τους όρους που έχουν αναπτυχθεί για να περιγράψουν διάφορα είδη και μεγέθη δικτύων υπολογιστών.

Η ραγδαία και επιβαλλόμενη χρήση δικτύων υπολογιστών δημιούργησε νέες καταστάσεις και νέες ανάγκες. Δημιουργήθηκαν νέες θέσεις εργασίας αλλά ταυτόχρονα και νέου είδους απειλές-εγκληματικές ενέργειες. Η ανάγκη για αδιάκοπη ροή πληροφοριών διαμέσου δικτύων έκανε επιτακτική την ανάγκη για δημιουργία νέων εξελιγμένων τρόπων αντιμετώπισης επιθέσεων. Έγινε επιτακτική η ανάγκη χρήσης πρόσθετου υλικού, λογισμικού και ανάπτυξη νέων αλγορίθμων-μεθόδων για την προστασία των δεδομένων που διακινούνται σε αυτά. Έγινε απαραίτητη η ανάγκη παρακολούθησης των δεδομένων που διακινούνται (κίνηση δικτύου) μεταξύ των συσκευών του δικτύου για την εξαγωγή συμπερασμάτων που αφορούν την «ομαλή λειτουργία» του δικτύου υπολογιστών. Αναφέροντας «ομαλή λειτουργία», δεν εννοούμε μόνο την αδιάκοπτη λειτουργία του δικτύου υπολογιστών, αλλά τη συνεχή και χωρίς προβλήματα ανταλλαγή πληροφοριών, την εκμετάλλευση όλων των δυνατοτήτων που αυτό παρέχει (πχ χρήση της βέλτιστης ταχύτητας ανταλλαγής δεδομένων) καθώς και την απαλλαγή από κάθε είδους επίθεση που μπορεί να διαταράξει την ποιότητα των παρεχόμενων υπηρεσιών. Μερικά παραδείγματα «ομαλής λειτουργίας» ενός δικτύου υπολογιστών αποτελούν η πλήρης εκμετάλλευση των πόρων του συστήματος δικτύου (network resources), η αποφυγή προβλημάτων επικοινωνίας (πχ το φαινόμενο bandwidth congestion με αποτέλεσμα την «αργή» ανταλλαγή πληροφοριών διαμέσου του δικτύου), η μυστικότητα της διαδιδόμενης πληροφορίας και γενικότερα η αδιάλειπτη χωρίς

προβλήματα λειτουργία όλων των πόρων αλλά και γενικότερα όλου του δικτύου υπολογιστών. Σημαντικό μερίδιο στην εξασφάλιση των παραπάνω κατέχει η παρακολούθηση και ανάλυση της κίνησης δικτύου (network traffic). Η καταγραφή και ανάλυση των δεδομένων – πακέτων που ρέουν σε ένα δίκτυο υπολογιστών έχει ως στόχο την αποτελεσματικότερη – γρηγορότερη και ευκολότερη εξαγωγή συμπερασμάτων που έχουν να κάνουν με την κατάσταση και τα προβλήματα του υπό εξέταση δικτύου υπολογιστών. Εξάγοντας γρηγορότερα αποτελέσματα και συμπεράσματα, μπορεί να γίνει ευκολότερη και γρηγορότερη πρόβλεψη, εντοπισμός και αντιμετώπιση καταστάσεων που μπορούν να οδηγήσουν σε προβλήματα επικοινωνίας τα οποία θα διαταράξουν την «ομαλή λειτουργία» του δικτύου υπολογιστών.

Για την παρακολούθηση της κίνησης ενός δικτύου υπολογιστών με στόχο την ανάλυση και αντιμετώπιση πιθανών ή και πραγματικών κινδύνων-προβλημάτων έχουν αναπτυχθεί διάφορες μέθοδοι-τεχνικές και έχουν κατασκευαστεί μεγάλος αριθμός εργαλείων. Έχει δημιουργηθεί πρωτόκολλο (SNMP) για την παρακολούθηση συσκευών και αντικειμένων με στόχο την ενημέρωση για τυχόν προβλήματα συσκευών που μπορούν να οδηγήσουν σε προβλήματα στην ομαλή λειτουργία του δικτύου. Έχουν δημιουργηθεί συστήματα για την παρακολούθηση και εντοπισμό των εισβολών σε ένα δίκτυο (Intrusion detection systems) καθώς και συστήματα για την αντιμετώπιση αυτών (Intrusion Prevention Systems). Έχουν κατασκευαστεί μέχρι σήμερα πάρα πολλά εργαλεία που μπορούν να καταγράψουν και να αποτυπώσουν της ροές των δεδομένων, αποθηκεύοντας τις πληροφορίες καταγραφής σε αρχεία (log files) για μελλοντική ανάλυση. Εργαλεία όπως το nfdump, το εργαλείο snort, επώνυμα προϊόντα της GFI και πολλά άλλα κατασκευασμένα για εμπορικούς και μη λόγους έρχονται να βοηθήσουν στην καταγραφή και παρακολούθηση της κίνησης ενός δικτύου υπολογιστών προσπαθώντας να βοηθήσουν την ομάδα διαχείρισης αυτού να εξασφαλίσει την ομαλή του λειτουργία.

Για την ευκολότερη ανάλυση των καταγεγραμμένων δεδομένων και κατά συνέπεια την ευκολότερη εξαγωγή συμπερασμάτων έχουν αναπτυχθεί διάφορες μέθοδοι γραφικής απεικόνισης των καταγραφέντων δεδομένων. Είναι γνωστό ότι «μία εικόνα ίσον χίλιες λέξεις». Στόχος των μεθόδων απεικόνισης αποτελεί η διευκόλυνση του παρατηρητή – αναλυτή στην εξαγωγή συμπερασμάτων και διορθωτικών ενεργειών για την ομαλή λειτουργία του υπό παρατήρηση δικτύου υπολογιστών.

Στην ενότητα αυτή θα πραγματοποιηθεί περιγραφή της σημασίας που έχει στην ομαλή λειτουργία καθώς και στην αρτιότερη διαχείριση ενός δικτύου υπολογιστών η παρακολούθηση της κίνησης του δικτύου και η καταγραφή των συμβάντων που παρατηρούνται σε αυτό. Θα καταγραφεί το ποιοι μπορεί να είναι οι πιθανοί λόγοι οι οποίοι να καθιστούν επιτακτική την ανάγκη για αυτή την παρακολούθηση. Θα καταγραφούν οι συνηθέστεροι αλλά και οι πιο εξεζητημένοι (“state of art”) τρόποι παρακολούθησης της κίνησης ενός δικτύου υπολογιστών τονίζοντας την πρόσθετη αξία που μπορεί να προσδώσουν οι τεχνικές αναπαράστασης δεδομένων στην παρακολούθηση αυτή.

1.2 Απειλές της Ομαλής Λειτουργίας Ενός Δικτύου Υπολογιστών

Πριν προχωρήσουμε στην αναφορά τρόπων παρακολούθησης της κίνησης και την καταγραφή συμβάντων σε ένα δίκτυο υπολογιστών κρίνεται σκόπιμο να γίνει μία μικρή αναφορά στις απειλές που δέχεται σήμερα ένα δίκτυο υπολογιστών και κυρίως στις απειλές και κινδύνους που υπάρχουν και μπορούν να διαταράξουν την ομαλή λειτουργία ενός δικτύου υπολογιστών. Κάποιες από αυτές μπορούν να γίνουν αντιληπτές με την απλή παρατήρηση της κίνησης σε ένα δίκτυο υπολογιστών, οπότε σε αυτή την παράγραφο, εκτός της αναφοράς σε πιθανές απειλές θα γίνει και μία μικρή εισαγωγή στον τρόπο με τον οποίο θα μπορούσαν αυτές να εντοπιστούν με την παρατήρηση της κίνησης ενός δικτύου υπολογιστών. Φυσικά στόχος της παρούσας παραγράφου δεν είναι να καλύψει το σύνολο των επιθέσεων που μπορούν να πραγματοποιηθούν σε ένα δίκτυο υπολογιστών αλλά να τονίσει-ενδυναμώσει την πεποίθηση ότι η παρακολούθηση της κίνησης δεδομένων (traffic) σε ένα δίκτυο υπολογιστών αποτελεί ένα από τα σημαντικότερα εργαλεία που μπορεί να χρησιμοποιήσει ο διαχειριστής δικτύου για να εξασφαλίσει την ομαλή και αδιάκοπη λειτουργία αυτού.

1.2.1 Computer Virus, Worms, Trojan Horses

Μία από τις συνηθέστερες απειλές ενός δικτύου υπολογιστών είναι οι «Ιοί υπολογιστών» (Computer Viruses), τα «σκουλήκια υπολογιστών» (Computer Worms) και οι «δούρειοι ίπποι» (Trojan horses).

Computer virus είναι έναν εκτελέσιμο πρόγραμμα σε έναν υπολογιστή το οποίο έχει ως στόχο την καταστροφή του υπολογιστή στον οποίο έχει διεισδύσει (καταστροφή σκληρού δίσκου, μνήμης και γενικότερα διατάραξη της ομαλής λειτουργίας του υπολογιστή) [2]. Έχει τη δυνατότητα να αντιγράφεται και να διαδίδεται διαμέσου του δικτύου υπολογιστών στους υπόλοιπους υπολογιστές. Worm είναι ένα είδος απειλής (πρόγραμμα και αυτό) το οποίο έχει τη δυνατότητα να κλωνοποιείται και να στέλνει αντίγραφα του εαυτού του σε άλλους υπολογιστές-κόμβους του δικτύου χωρίς να χρειάζεται την διαμεσολάβηση του χρήστη (σε αντίθεση με έναν ιό) [3]. Αν οι Ιοί είναι προγράμματα που έχουν ως κύριο στόχο να βλάψουν αρχεία και προγράμματα υπολογιστών όχι μόνο κατά ανάγκη ενός υπολογιστή, τα worms έχουν ως στόχο κυρίως να βλάψουν την «ομαλή λειτουργία» του δικτύου υπολογιστών με την κατανάλωση χώρου αποθήκευσης και εύρους ζώνης δικτύου (consuming bandwidth). Ο μεγαλύτερος αριθμός worms έχει σχεδιαστεί με κύριο στόχο να διαδίδεται και όχι να προσπαθεί να αλλάξει τα συστήματα μέσω των οποίων περνά. Τα περισσότερα worms αφού προσβάλουν έναν υπολογιστή που ανήκει στο δίκτυο, πραγματοποιούν ανίχνευση (scanning) για ανοιχτές θύρες (listening-open ports) σε άλλους υπολογιστές και εφόσον εντοπίσουν τις χρησιμοποιούν για να πολλαπλασιαστούν-αντιγραφούν σε αυτούς. Με αυτό τον τρόπο αυξάνεται δραματικά ο αριθμός των μη σχετικών με τη λειτουργία του δικτύου υπολογιστών πακέτων, με συνέπεια να αυξάνεται η κίνηση (network traffic) και να μειώνεται κατ' επέκταση το διαθέσιμο-ωφέλιμο για τις απαραίτητες λειτουργίες του δικτύου εύρος ζώνης (bandwidth). Συνέπεια αυτών η διατάραξη της «ομαλής λειτουργίας» του δικτύου υπολογιστών.

Οι ενέργειες ενός worm, σε αντίθεση με τις ενέργειες ενός Ιού, θα μπορούσαν να εντοπιστούν με την παρακολούθηση της κίνησης του δικτύου υπολογιστών. Πχ η ανίχνευση μεγάλου αριθμού πακέτων (υψηλής κίνησης) από έναν υπολογιστή (μολυσμένος από ένα worm υπολογιστής – bot, zombie computer) προς διάφορους υπολογιστές και μάλιστα προς διαφορετικές θύρες δικτύου (network ports) θα μπορούσε να δώσει ενδείξεις για την ύπαρξη κάποιου worm το οποίο προσπαθεί να πολλαπλασιαστεί σε άλλους υπολογιστές αναζητώντας ανοιχτές θύρες για να εξαπλωθεί διαμέσου αυτών (Πχ Το “Slammer worm” στέλνει πακέτα των 376-Bytes στην UDP θύρα 1434 τυχαίων hosts με στόχο να διαδοθεί). Η ανίχνευση και εντοπισμός ενός τέτοιου είδους υπερβολικού φορτίου κίνησης μπορεί να γίνει παρατηρώντας απόκλιση από ένα σταθερό επίπεδο τιμών (baseline) που καταγραφόταν ήδη στο υπό παρατήρηση δίκτυο.

Αντίστοιχα, «Δούρειος Ίππος» (Trojan Horse) είναι το λογισμικό το οποίο έχει ως στόχο την υποκλοπή πληροφοριών από ένα σύνολο υπολογιστών [4]. Αντίθετα με ένα worm δεν έχει ως στόχο τον πολλαπλασιασμό του εαυτού του αλλά εφόσον «καθίσει» σε έναν υπολογιστή στέλνει δεδομένα-πληροφορίες σε άλλους υπολογιστές (είτε εντός του τοπικού δικτύου- επίθεση εκ των έσω, είτε εκτός του τοπικού δικτύου υπολογιστών-εξωτερική επίθεση) με στόχο την υποκλοπή σημαντικών ίσως και εμπιστευτικών πληροφοριών. Μπορεί να δώσει τη δυνατότητα στον εισβολέα, μετά την εγκατάστασή του, να έχει απομακρυσμένη πρόσβαση στον μολυσμένο υπολογιστή και να αντλήσει πληροφορίες από αυτόν .

Και αυτού του είδους η απειλή μπορεί να εντοπιστεί (αφού πραγματοποιηθεί) παρατηρώντας την κίνηση ενός δικτύου υπολογιστών. Εφόσον ένα Trojan Horse εγκατασταθεί σε έναν υπολογιστή, στέλνοντας πληροφορίες σε έναν συγκεκριμένο υπολογιστή είτε του εσωτερικού τοπικού δικτύου είτε όχι, δημιουργεί ασυνήθιστα μεγάλο όγκο (σε σχέση με την περίπτωση ομαλής λειτουργίας – baseline) ροής δεδομένων (αριθμός πακέτων) που προορίζονται προς τον υπολογιστή αποδέκτη. Εφόσον καταγράφεται η κίνηση από ή προς τον μολυσμένο από το Trojan υπολογιστή, θα παρατηρηθεί αύξηση του όγκου πακέτων που αποστέλλονται από αυτόν και έχουν ως παραλήπτη έναν συγκεκριμένο υπολογιστή (συγκεκριμένη IP διεύθυνση). Στην ουσία θα παρατηρηθεί απόκλιση από την ομαλή λειτουργία (baseline) που καταγραφόταν μέχρι την συγκεκριμένη χρονική στιγμή γεγονός που αποτελεί ένδειξη για πιθανή ύπαρξη απειλής.

1.2.2 Port Scan Attack

Η ανίχνευση θυρών είναι πολύ σημαντική διαδικασία για έναν διαχειριστή δικτύου. Με αυτή τη διαδικασία μπορεί να δει ποιες δικτυακές υπηρεσίες εκτελούνται σε ένα host του δικτύου του. Αντίστοιχα βέβαια αυτή η διαδικασία μπορεί να αποτελέσει τρόπο εύρεσης των αδύνατων σημείων του δικτύου από κάποιον με κακές διαθέσεις. Ο βασικός στόχος μιας επίθεσης “port scan” είναι να καθοριστεί-ανιχνευτεί αν υπάρχουν διαθέσιμες (αν ακούν – listening) σε κάποια συσκευή υπηρεσίες και που μπορούν να ανταποκριθούν σε αιτήματα διαμέσου δικτύου. Η πιο κοινή μέθοδος για την μέθοδο “port scan” είναι η “targeted scan” κατά την οποία γίνεται σάρωση των θυρών που αντιστοιχούν σε πολύ συχνά χρησιμοποιούμενες υπηρεσίες δικτύου όπως web και mail services. Οι παραλλαγές αυτής είναι αρκετές αλλά η πιο διαδεδομένη είναι η σάρωση του συνόλου των θυρών και η εύρεση με αυτό τον τρόπο των «ανοιχτών» [5].

Η επίθεση αυτή στην απλή της μορφή περιλαμβάνει την αποστολή αριθμού πακέτων από έναν host (μία IP διεύθυνση) προς έναν άλλον με διαφορετική θύρα προορισμού για κάθε πακέτο. Δημιουργείται δηλαδή μεγάλη ροή δεδομένων-πακέτων με την ίδια (IP) διεύθυνση αφετηρίας (ίσως και την ίδια θύρα αφετηρίας) την ίδια διεύθυνση προορισμού έχοντας βέβαια διαφορετική θύρα προορισμού. Με την παρατήρηση και την ανάλυση της κίνησης σε ένα δίκτυο υπολογιστών θα μπορούσε να ανιχνευτεί μίας τέτοιου είδους επίθεση, με αποτέλεσμα να είναι δυνατή η καταγραφή καθώς και ο αποκλεισμός-περιορισμός του επιτιθέμενου.

1.2.3 Brute Force Attacks

Ως Brute-force attack χαρακτηρίζεται η προσπάθεια ανίχνευσης του αλγορίθμου κρυπτογράφησης των δεδομένων και την αποκρυπτογράφηση κρυπτογραφημένων μηνυμάτων χρησιμοποιώντας τη μέθοδο δοκιμής όλων των δυνατών κλειδιών κρυπτογράφησης [6]. Ο επιτιθέμενος σε αυτή την περίπτωση δεν γνωρίζει το κλειδί κρυπτογράφησης, έχει μόνο το κρυπτογραφημένο κείμενο και προσπαθεί γνωρίζοντας τον αλγόριθμο κρυπτογράφησης να βρει το κλειδί που χρησιμοποιήθηκε δοκιμάζοντας όλους τους δυνατούς συνδυασμούς. Τέτοιου είδους επιθέσεις αποτελούν κοινή απειλή για τους web developers. Σε τέτοιες περιπτώσεις γίνεται προσπάθεια ανίχνευσης του κωδικού πρόσβασης ενός web site δοκιμάζοντας όλους τους δυνατούς συνδυασμούς των γραμμάτων, των αριθμών και των συμβόλων μέχρι να βρεθεί ο σωστός συνδυασμός [7].

Αν και αυτού του είδους οι επιθέσεις πραγματοποιούνται σε επίπεδο εφαρμογής (application layer – OSI model) θα μπορούσε η παρακολούθηση της κίνησης του δικτύου να εντοπίσει ενδείξεις ύπαρξης τέτοιας επίθεσης. Υπερβολική κίνηση και κατανάλωση εύρους ζώνης (bandwidth) από μία διεύθυνση μπορεί να αποτελεί συνθήκη ύπαρξης τέτοιου είδους επίθεσης. Εντοπίζοντας μεγάλο όγκο πακέτων από συγκεκριμένη διεύθυνση προς hosts, εφαρμογές και ιστοσελίδες που απαιτούν πιστοποίηση ταυτότητας χρηστών μπορούν να ειδοποιηθούν (alert) τον διαχειριστή ώστε αυτός να πραγματοποιήσει περεταίρω έλεγχο για την εύρεση ανωμαλιών-επίθεσης αυτού του τύπου.

1.2.4 Denial-of-Service Attack, DDoS Attack, Buffer Overflow

Πολύ γνωστή στις μέρες μας είναι η επίθεση Denial of service. Όπως αναφέρει και το όνομά της πρόκειται για προσπάθεια να καταστούν υπηρεσίες ενός υπολογιστή ή και ολόκληρου δικτύου μη

διαθέσιμες Attack [6]. Μπορεί αυτού του τύπου η επίθεση να πραγματοποιηθεί με πολλούς τρόπους. Θα μπορούσαν να σταλθούν πακέτα με συγκεκριμένο περιεχόμενο σε ευάλωτες εφαρμογές ή λειτουργικά συστήματα με αποτέλεσμα την παύση λειτουργίας της εφαρμογής ή του λειτουργικού συστήματος. Θα μπορούσε να σταλεί πολύ μεγάλος αριθμός πακέτων σε έναν υπολογιστή (bandwidth flooding) έτσι ώστε η σύνδεση πρόσβασης του να καταστεί μη λειτουργική. Θα μπορούσε να πραγματοποιηθεί στέλνοντας TCP πακέτα με στόχο την υπερχειλίση των buffers (buffer overflow) του παραλήπτη.

Στην περίπτωση που τέτοιου είδους επιθέσεις πραγματοποιηθούν, λόγω του μεγάλου αριθμού των πακέτων που στέλνονται, θα μπορούσε η παρατήρηση της ροής πακέτων του δικτύου να ανιχνεύσει αυτή την μεγάλη αλλαγή στον αριθμό των πακέτων σε σχέση με τον αριθμό που καταγραφόταν κατά την περίοδο φυσιολογικής κατάστασης. Και σε αυτή λοιπόν την περίπτωση η παρακολούθηση της κίνησης του δικτύου μπορεί να βοηθήσει τον διαχειριστή να εντοπίσει την αιτία απόκλισης από την ομαλή λειτουργία του δικτύου που διαχειρίζεται.

1.2.5 Υπερβολική Χρήση Ιστοσελίδων (Spam Πληροφορίες)

Η υπερβολική χρήση του Internet άσκοπα από κάποιους χρήστες ενός δικτύου υπολογιστών μπορεί να έχει πολλά και δυσάρεστα αποτελέσματα για την ομαλή λειτουργία του δικτύου. Η χρήση υπερβολικά του internet χωρίς να υπάρχει λόγος από κάποιους χρήστες θα μπορούσε να μειώσει αισθητά το διαθέσιμο εύρος ζώνης (bandwidth) το οποίο θα μπορούσε να χρησιμοποιηθεί από άλλες χρήσιμες για τη λειτουργία της επιχείρησης εφαρμογές. Η άσκοπη αυτή χρήση μεταφέροντας πληροφορίες στο δίκτυο οι οποίες δεν σχετίζονται με λειτουργία της εταιρείας (πληροφορίες που δεν χρειάζονται οι χρήστες) θα μπορούσε να οδηγήσει στην εισαγωγή προς το εσωτερικό του δικτύου υπερβολικά μεγάλου όγκου δεδομένων (και κατά συνέπεια πακέτων). Μία τέτοια περίπτωση θα μπορούσε να είναι η χρήση εφαρμογών όπως p2p software (π.χ. bit torrents) ή ακόμη και η αποστολή και λήψη (εισαγωγή στο εσωτερικό δίκτυο) spam emails. Με αυτό τον τρόπο επιβαρύνονται οι συνδέσεις του τοπικού δικτύου με ανεπιθύμητο traffic, επιβαρύνονται οι servers με πρόσθετο φορτίο εργασίας, ενώ οι χρήστες του δικτύου καταναλώνουν ώρες εργασίας για ενασχόληση με τον διαχωρισμό των σημαντικών από τις ασήμαντες-άχρηστες πληροφορίες που δέχονται διαμέσου email.

Η παρατήρηση της κίνησης ενός δικτύου υπολογιστών θα μπορούσε να βοηθήσει στην αρτιότερη διαχείριση και έλεγχο του τοπικού δικτύου από τον διαχειριστή ανιχνεύοντας τέτοιου είδους καταστάσεις. Θα μπορούσε η παρατήρηση και καταγραφή της κίνησης αδικαιολόγητα μεγάλου όγκου πακέτων από συγκεκριμένο ή συγκεκριμένους υπολογιστές προς διευθύνσεις στο internet να ενεργοποιήσει (trigger – alert) τον διαχειριστή του δικτύου να προβεί σε λεπτομερέστερη ανάλυση των δεδομένων.

1.2.6 Bandwidth Congestion

Τα περισσότερα από τα παραπάνω αναφερθέντα προβλήματα-καταστάσεις που μπορεί να παρατηρηθούν σε ένα δίκτυο υπολογιστών (υπερβολική χρήση ιστοσελίδων, spam πληροφορίες, Denial-of-service attack, Brute force attacks, worms κλπ) οδηγούν έμμεσα ή άμεσα στην συμφόρηση του εύρους ζώνης (bandwidth congestion) της σύνδεσης του τοπικού δικτύου με το παγκόσμιο δίκτυο (Internet). Αυτό έχει ως συνέπεια τη δημιουργία πρόσθετων καθυστερήσεων στην επικοινωνία (queuing delays in routers, packet loss, blocking of new connections κλπ). Η παρακολούθηση και καταγραφή της ροής δεδομένων από και προς το internet και η σύγκριση με προηγούμενες καταστάσεις ορθής λειτουργία του δικτύου (baseline) μπορούν να βοηθήσουν στον εντοπισμό των αιτιών του προβλήματος και να βοηθήσουν στην δημιουργία διορθωτικών ενεργειών για την αντιμετώπιση αυτών (πχ αντικατάσταση σύνδεσης με τον έξω κόσμο με υψηλότερης ταχύτητας σύνδεση ή και συνδυασμός περισσότερων της μίας συνδέσεων).

1.2.7 Δυσλειτουργία Συσκευών

Πολλά από τα καθημερινά προβλήματα που αντιμετωπίζει ένα διαχειριστής δικτύου και τα οποία δημιουργούν προβλήματα στην ομαλή λειτουργία του τοπικού δικτύου οφείλονται σε δυσλειτουργία δικτυακών συσκευών όπως κάρτες δικτύου, δρομολογητές (routers), switches, hubs, υπολογιστές, διακομιστές (servers) κλπ. Η παρακολούθηση και η αναφορά προς τον διαχειριστή γεγονότων-προβλημάτων που προέρχονται από τέτοιες συσκευές μπορεί να βοηθήσει όχι μόνο στη γρήγορη αντιμετώπιση των προβλημάτων αλλά και στην πρόληψη και αποφυγή αυτών. Πχ αν ένας διαχειριστής παρατηρήσει αύξηση στα checksum errors στα frames που στέλνονται από μία κάρτα δικτύου μπορεί να την αντικαταστήσει έγκαιρα πριν καταστραφεί, προλαμβάνοντας έτσι μελλοντικά προβλήματα [6].

1.2.8 Υπερβολικό Φορτίο Εργασίας (Διαμοιρασμός Υπηρεσιών)

Ένας ακόμη παράγοντας που προκαλεί πονοκέφαλο στους διαχειριστές επηρεάζοντας την ομαλή λειτουργία του δικτύου είναι προβλήματα αργής απόκρισης κάποιων πολύ σημαντικών για τη σωστή λειτουργία της επιχείρησης προγραμμάτων. Το υπερβολικό φορτίο εργασίας κάποιων διακομιστών που δημιουργείται λόγω έλλειψης πόρων (resources), όπως διαθέσιμη μνήμη, υπολογιστική ισχύ κλπ, μπορεί να είναι ένας από τους παράγοντες τέτοιων καταστάσεων. Μια χρήσιμη ενέργεια για την επίλυση του προβλήματος μπορεί να είναι η παρατήρηση της κίνησης από και προς τον διακομιστή (server) του τοπικού δικτύου που φέρει συγκεκριμένο και σημαντικό για τη λειτουργία της επιχείρησης λογισμικό, όπως και η καταγραφή πιθανής μη φυσιολογικής αύξησης της κίνησης αυτής που ενδεχομένως να καθιστά αργή την επικοινωνία με το συγκεκριμένο λογισμικό. Μπορεί ο διαχειριστής τοποθετώντας έναν ακόμη διακομιστή με το ίδιο λογισμικό να μοιράσει το φορτίο εργασίας πετυχαίνοντας καλύτερη απόδοση και απόκριση στα αιτήματα των χρηστών.

1.2.9 Απότομες αλλαγές στην ποσότητα των δεδομένων

Υπάρχουν ακόμη περιπτώσεις κατά τις οποίες μπορεί να υπάρξει απότομη αλλαγή στον αριθμό των παρατηρούμενων ροών (σύνολο πακέτων) είτε π.χ. λόγο της παύσης λειτουργίας κάποιας υπηρεσίας, είτε λόγο π.χ. εμφάνισης κάποιας νέας στο δίκτυο. Μπορεί κάποιος διακομιστής λόγο βλάβης να σταμάτησε να λειτουργεί στο δίκτυο με αποτέλεσμα να σταμάτησαν και τα πακέτα τα οποία έφταναν ή ξεκινούσαν από αυτόν. Αυτό το πρόβλημα θα μπορούσε να ανιχνευτεί με την παρακολούθηση της κίνησης του δικτύου και την παρατήρηση της αλλαγής στην ποσότητα των δεδομένων.

Αντίστοιχα, η ξαφνική εμφάνιση μεγάλου όγκου δεδομένων (συγκεκριμένου τύπου) μπορεί να δηλώνει την εγκατάσταση κάποιας ανεπιθύμητης υπηρεσίας στο δίκτυο. Και αυτό το γεγονός θα μπορούσε να εντοπιστεί από την παρατήρηση της κίνησης σε ένα δίκτυο υπολογιστών.

Όπως φαίνεται και από τις παραπάνω παραγράφους υπάρχουν πολλές καταστάσεις και απειλές που μπορεί να διαταράξουν την ομαλή λειτουργία ενός δικτύου υπολογιστών (έχουν επιγραμματικά αναφερθεί οι σημαντικότερες και όχι όλες από αυτές διότι κάτι τέτοιο δεν αποτελεί στόχο της συγκεκριμένης εργασίας). Οι περισσότερες από αυτές όπως αναλύθηκε και παραπάνω μπορούν να αντιμετωπιστούν με τη σωστή και μεθοδική παρατήρηση-καταγραφή συμβάντων καθώς και τις κινήσεις του υπό διαχείριση δικτύου υπολογιστών.

1.3 Τρόποι Προφύλαξης-Αντιμετώπισης Κινδύνων

Υπάρχουν γενικά στην καθημερινή μας ζωή (ξεφεύγοντας λίγο από την αναφορά μας σε δίκτυα υπολογιστών) δύο τρόποι αντιμετώπισης προβλημάτων-κινδύνων, η πρόληψη και η καταστολή (αντιμετώπιση μετά την εμφάνιση). Κάθε φορά οι δύο αυτοί όροι συνδέονται με το κόστος του σχεδιασμού και της υλοποίησης της λύσης για την πρόληψη και την αντιμετώπιση. Κάτι αντίστοιχο έχουμε και στα δίκτυα υπολογιστών. Ο υπεύθυνος του δικτύου υπολογιστών καλείται να βρει τρόπους πρόληψης και αποφυγής κινδύνων καθώς και τρόπους γρήγορης αντιμετώπισής τους εφόσον πραγματοποιηθούν. Θα πρέπει να σχεδιάσει την πολιτική ασφαλείας του δικτύου του ώστε να αποτρέψει μελλοντικούς κινδύνους και να είναι σε θέση να τους αντιμετωπίσει γρήγορα εφόσον εμφανιστούν. Θα πρέπει να είναι σε θέση να εντοπίσει και να απομονώσει γρήγορα την αιτία του προβλήματος κάνοντας ποιο εύκολη την αντιμετώπισή τους.

Υπάρχουν διάφοροι τρόποι πρόληψης και αντιμετώπισης κινδύνων-απειλών που μπορούν να εμφανιστούν σε ένα δίκτυο υπολογιστών. Μεταξύ αυτών είναι η χρήση αντιβιοτικών προγραμμάτων, η χρήση software και hardware firewalls και proxies, η χρήση λογισμικών παρακολούθησης ενημέρωσης (alerts) και αποτροπής συμβάντων (events), λογισμικά ανίχνευσης αδυναμιών (vulnerabilities scanners, π.χ. Nessus, VulScan κλπ), καθώς και η χρήση λογισμικών σύλληψης (capture), παρακολούθησης και καταγραφής πακέτων ή και ροών δεδομένων και γενικότερα κίνησης ενός δικτύου υπολογιστών. Οι περισσότερες από τις λύσεις που μπορεί να χρησιμοποιήσει κάποιος για την πρόληψη και την αποφυγή κινδύνων χωρίζονται κυρίως σε δύο κατηγορίες, στις network based και στις host based. Στις λύσεις δηλαδή σε επίπεδο δικτύου και στις λύσεις που αφορούν συγκεκριμένο κάθε φορά υπολογιστή.

Τα Firewalls και οι proxy servers αποτελούν την «πύλη του κάστρου» από την οποία θα πρέπει να περάσουν και να πιστοποιηθούν για το επίπεδο ασφάλειας που πληρούν όλα τα εισερχόμενα και εξερχόμενα αιτήματα του δικτύου [6]. Θα επιτραπεί να περάσει μόνο η κίνηση που πληρεί τα κριτήρια που καθορίζονται από την τοπική πολιτική ασφαλείας, ενώ σε όλη την υπόλοιπη κίνηση θα απαγορευτεί η πρόσβαση. Παρόλη την διάδοσή τους τα firewalls και τα proxies δεν παύει να είναι συσκευές στο δίκτυο ή λογισμικό σε κάποιο υπολογιστή, γεγονός που τα καθιστά και αυτά αντικείμενο επιθέσεων. Σε περίπτωση παραβίασής τους το εσωτερικό δίκτυο είναι εκτεθειμένο σε επιθέσεις και εισβολείς. Η καλύτερη τεχνική είναι η παρακολούθηση και ο έλεγχος της σωστής λειτουργίας τους σε τακτά χρονικά διαστήματα από την ομάδα διαχείρισης του δικτύου.

Ένας επίσης αρκετά διαδεδομένος τρόπος πρόληψης καθώς και αντιμετώπισης προβλημάτων που βοηθά στην ομαλή λειτουργία του δικτύου και διευκολύνει τον διαχειριστή στην αρτιότερη διαχείριση του είναι η χρήση και ρύθμιση παρακολούθησης συμβάντων που πραγματοποιούνται στο δίκτυο. Υπάρχουν πάρα πολλά λογισμικά στο εμπόριο που μπορούν να σαρώσουν το δίκτυο να βρουν δυσλειτουργίες και να ενημερώσουν τον διαχειριστή ή ακόμη και να διορθώσουν αυτόματα το πρόβλημα. (πχ. Το εργαλείο Network Server Monitor της εταιρείας GFI).

Για την περίπτωση εντοπισμού συμβάντων σε ανεξάρτητους υπολογιστές-συσκευές τα περισσότερα λειτουργικά συστήματα έχουν ενσωματωμένη διαδικασία και εργαλεία για την παρακολούθηση συμβάντων (πχ εργαλείο Event Viewer του λειτουργικού συστήματος Microsoft Windows).

Τέλος, μία ακόμη κατηγορία εργαλείων είναι αυτά που δίνουν τη δυνατότητα στον διαχειριστή να συλλάβει (capture) και να πραγματοποιήσει ανάλυση σε επίπεδο πακέτου της κίνησης του δικτύου. Προβλήματα όπως “network bottleneck”, μικρό εύρος ζώνης (bandwidth) στο δίκτυο ή και σε συγκεκριμένους υπολογιστές, επιθέσεις buffer overflow ή denial of service attacks μπορούν αναγνωριστούν και να αντιμετωπιστούν.

1.4 Η Συνεισφορά της Παρακολούθησης της Κίνησης στην Ομαλή Λειτουργία Ενός Δικτύου Υπολογιστών.

"If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle."

"The Art of War" by Sun Tzu

Όπως αναφέρθηκε και σε προηγούμενη παράγραφο τα δίκτυα υπολογιστών έχουν γίνει στις μέρες μας ο ακρογωνιαίος λίθος της επιχείρησης, αποτελώντας κρίσιμα συστατικά της επιτυχίας αυτής. Όταν το δίκτυο αποτύχει πελάτες και υπάλληλοι δεν μπορούν να επικοινωνήσουν, να έχουν πρόσβαση σε σημαντικές πληροφορίες και υπηρεσίες με αποτέλεσμα προβλήματα που έχουν να κάνουν με την παραγωγικότητα και την απόδοση των λειτουργιών της επιχείρησης [8]. Μία επιτυχημένη εταιρεία στηρίζει τη λειτουργία της σε ένα άρτια λειτουργικά δίκτυο υπολογιστών το οποίο εξυπηρετεί τις λειτουργικές ανάγκες της και είναι έτοιμο να ανταπεξέλθει και να αντιμετωπίσει το σύνολο των κινδύνων που υπάρχουν με στόχο την αδιάλειπτη λειτουργία και εξυπηρέτηση των χρηστών. Για την σωστή αντιμετώπιση αυτών σημαντικό παράγοντα παίζει να γνωρίζει κανείς τις απειλές και να μπορεί με κατάλληλα μέσα να τις αντιμετωπίσει.

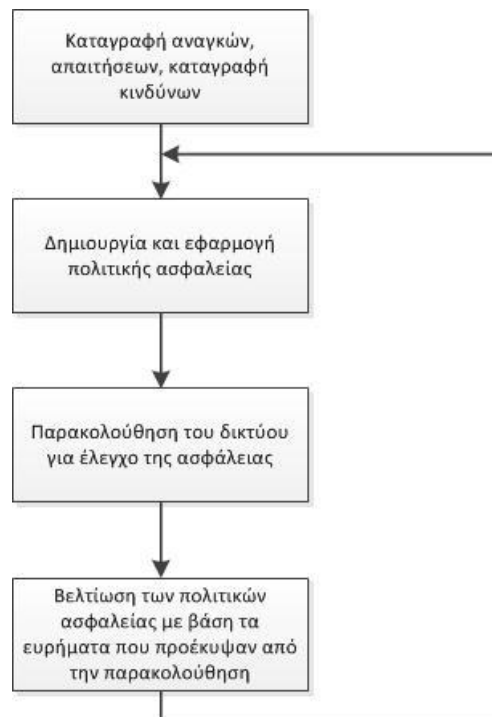
Η παρακολούθηση της κίνησης ενός δικτύου υπολογιστών μπορεί να εξασφαλίσει την παραπάνω αναφερθείσα ανάγκη αδιάλειπτης λειτουργίας και σταθερότητας ενός δικτύου και να αποτελέσει ένα μέσο αναγνώρισης και αντιμετώπισης κινδύνων-απειλών. Με την παρακολούθηση μπορεί να ανιχνευτούν επικίνδυνες καταστάσεις να βρεθούν τα αίτια των προβλημάτων και στη συνέχεια να γίνουν οι απαραίτητες διορθωτικές ενέργειες για την επιδιόρθωση – αποφυγή των σφαλμάτων - προβλημάτων. Μπορούν να ανιχνευτούν προσπάθειες εισβολής, εκμετάλλευσης πόρων, διάδοσης ιών (computer viruses) και σκουληκιών υπολογιστών (computer worms), προβλήματα υλικού (hardware), προσπάθειες πρόσβασης σε κρίσιμες και απαγορευμένες πληροφορίες, Denial of service attacks (και DDoS attacks), να ανιχνευτούν προβλήματα απόδοσης (network and server bottleneck), να αναγνωριστούν εφαρμογές, χρήστες και υπολογιστές οι οποίοι καταναλώνουν μεγάλο ποσοστό εύρους ζώνης (bandwidth), να δημιουργηθούν αναφορές (reports) και τάσεις στο δίκτυο καθώς και πάρα πολύ άλλοι κίνδυνοι, απειλές και γενικότερα καταστάσεις του δικτύου.

Παράλληλα, αν κάποιος διαχειριστής επιθυμεί να έχει μία συνολική εικόνα του δικτύου του, θα μπορούσε να χρησιμοποιήσει ένα σύστημα καταγραφής συμβάντων και γεγονότων καθώς και εργαλεία παρακολούθησης της κίνησης του δικτύου τα οποία θα του παρείχαν την πληροφορία που επιθυμεί. Ο συνδυασμός των παραπάνω μπορεί να δώσει συνολική, καθαρή εικόνα για την κίνηση που περνά μέσα από το δίκτυο υπολογιστών. Μπορεί δηλαδή η παρακολούθηση της κίνησης που περνά από το τοπικό δίκτυο υπολογιστών να βοηθήσει την ομάδα διαχείρισης του στην αρτιότερη διαχείριση ενός δικτύου υπολογιστών [9].

Ας ξεκινήσουμε την ανάλυσή μας για την υποστήριξη του γεγονότος ότι η παρακολούθηση της κίνησης ενός δικτύου συνεισφέρει στην ομαλή λειτουργία αυτού αναφέροντας πως μπορεί να χρησιμοποιηθεί για την αντιμετώπιση κινδύνων. Η σχεδίαση και επίτευξη του επιθυμητού επιπέδου ασφάλειας ενός δικτύου υπολογιστών είναι μία δυναμική διαδικασία που περιλαμβάνει τα ακόλουθα βήματα [9]:

1. Καταγραφή των αναγκών και των απαιτήσεων για ασφάλεια, αναγνώριση των πιθανών κινδύνων που μπορεί να παρουσιαστούν.
2. Δημιουργία της κατάλληλης πολιτικής ασφάλειας και εφαρμογή της στο υπό διαχείριση δίκτυο υπολογιστών.
3. Παρακολούθηση του δικτύου υπολογιστών για έλεγχο σωστής εφαρμογής της ασφάλειας, τυχόν κενών, παραλήψεων ή και αδυναμιών της πολιτικής ασφαλείας.
4. Βελτίωση των πολιτικών ασφαλείας με βάση τα ευρήματα που προέκυψαν από την παρακολούθηση του βήματος 3.
5. Επανάληψη της παρακολούθησης του δικτύου στο βήμα 3.

Θα μπορούσε αυτή η αλληλουχία βημάτων να αναπαρασταθεί σχηματικά στο ακόλουθο σχεδιάγραμμα (σχήμα 1.1):



Σχήμα 1.1: Βήματα για την επίτευξη επιθυμητού επιπέδου ασφάλειας ενός δικτύου υπολογιστών

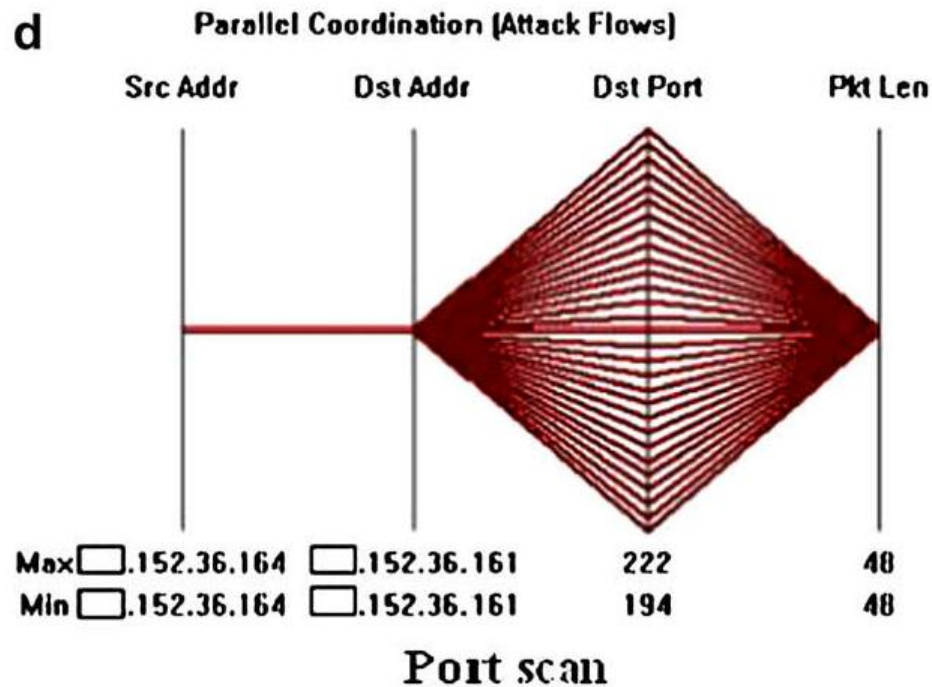
Όπως φαίνεται τόσο από την παραπάνω λίστα όσο και από το σχήμα η παρακολούθηση του δικτύου (σε αυτή συμπεριλαμβάνεται η παρακολούθηση της κίνησης του δικτύου) αποτελεί το σημαντικότερο βήμα για την επίτευξη του επιθυμητού επιπέδου ασφάλειας. Μέσα από την παρατήρηση παίρνονται αποφάσεις και πραγματοποιούνται διορθωτικές ενέργειες ώστε να εντοπιστούν και να αντιμετωπιστούν κενά ασφαλείας. Εντοπίζοντας πραγματοποιηθείσες επιθέσεις οι οποίες καταργούν την ήδη εφαρμοσμένη πολιτική ασφαλείας μπορούν να γίνουν διορθωτικές ενέργειες για την ενδυνάμωση αυτής. Παρατηρώντας την κίνηση σε ένα δίκτυο και καταγράφοντας και αποθηκεύοντας τα αποτελέσματα αυτής είναι δυνατόν να γίνουν συγκρίσεις και να βρεθούν αποκλίσεις από την ομαλή λειτουργία (baseline) του δικτύου υπολογιστών. Σε αυτή τη φιλοσοφία-μέθοδο (anomaly detection) [10] στηρίζονται πολλά συστήματα καταγραφής, ανάλυσης της κίνησης και εντοπισμού εισβολών (Intrusion Detection Systems, IDS).

Στην παράγραφο με τίτλο «**Απειλές της ομαλής λειτουργίας ενός δικτύου υπολογιστών**» αναφέρθηκαν ενδεικτικά κάποιοι από τους κινδύνους και τις απειλές οι οποίες θα μπορούσαν να εντοπιστούν χρησιμοποιώντας τεχνικές παρακολούθησης της κίνησης του δικτύου.

Καταγράφοντας την κίνηση (network traffic) σε ένα δίκτυο υπολογιστών και εντοπίζοντας σε αυτή υπερβολικά μεγάλο αριθμό πακέτων από ένα host (συσκευή με IP address, συνήθως υπολογιστής) προς έναν άλλον host και μάλιστα καταγράφοντας πακέτα τα οποία έχουν διαφορετικές θύρες προορισμού μπορεί ο παρατηρητής (τις περισσότερες φορές ο διαχειριστής του δικτύου ή σε μεγαλύτερα δίκτυα ο αναλυτής ασφαλείας) να συμπεράνει ότι πρόκειται για κάποιο είδος επίθεσης. Θα μπορούσε να είναι επίθεση σάρωσης θυρών (port scan) με στόχο να βρεθεί ανοιχτή (listening) θύρα ώστε να μπορέσει να πραγματοποιηθεί κάποια κακόβουλη ενέργεια (πχ διάδοση worm). Στο παρακάτω σχήμα (σχήμα 1.2) εμφανίζεται ένα διάγραμμα με τη χρήση parallel coordinates το οποίο αναπαριστά γραφικά τέτοιου είδους επίθεση. Πρόκειται για διάγραμμα το οποίο εξάγεται από εργαλείο καταγραφή και γραφικής αναπαράστασης κίνησης δικτύου υπολογιστών [11]. Οι τέσσερις κάθετες γραμμές στο σχήμα χρησιμοποιούνται για την απεικόνιση στο διάγραμμα των πληροφοριών των πακέτων που καταγράφονται από την παρατήρηση της κίνησης. Συγκεκριμένα στις καθέτους εμφανίζονται:

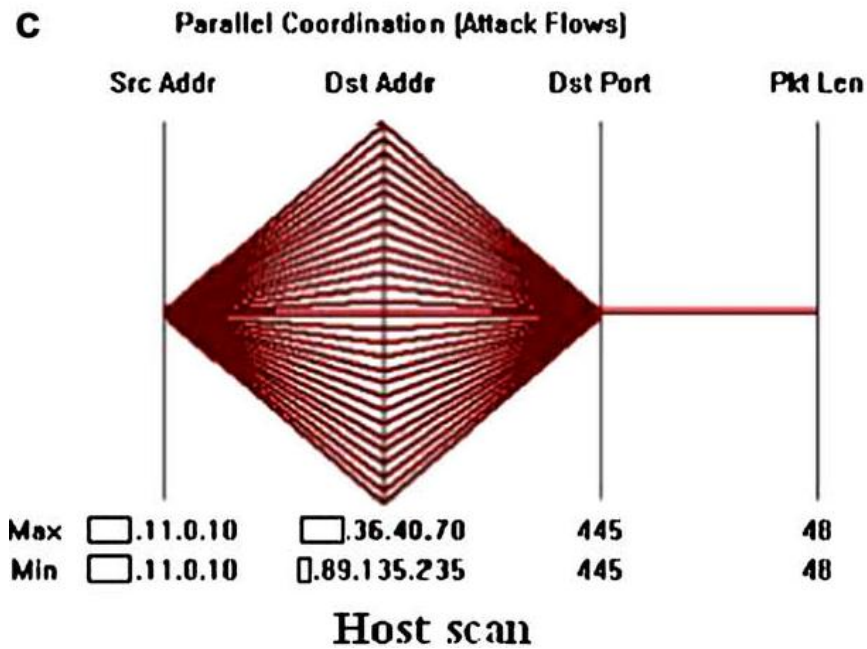
- Διεύθυνση αφετηρίας (source IP address) – πρώτη κάθετη,
- Διεύθυνση προορισμού (destination IP address) – δεύτερη κάθετος
- Θύρα προορισμού (destination port number) – τρίτη κάθετος
- Θύρα αφετηρίας (source port number) – τέταρτη κάθετος

Παρατηρώντας τέτοιου είδους σχεδιαγράμματα ανάμεσα στα αποτελέσματα παρακολούθησης κίνησης δικτύου μπορεί ο αναλυτής να συμπεράνει ύποπτη κίνηση στο υπό παρατήρηση δίκτυο και συγκεκριμένα επίθεση post scan.

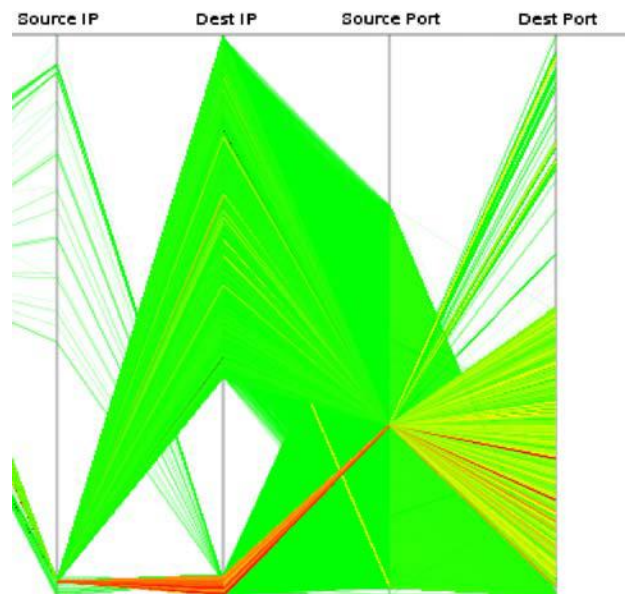


Σχήμα 1.2: Γραφική αναπαράσταση Port scan attack με τη χρήση parallel coordinates (Εικόνα από το pdf με τίτλο “Fast detection and visualization of network attacks on parallel coordinates” από τους Hyunsang Choi, Heejo Lee, Hyogon Kim [11])

Υπερβολικά μεγάλος αριθμός πακέτων (εκτός φυσιολογικών ορίων που έχουν καταγραφεί σε κατάσταση ομαλής λειτουργίας) από συγκεκριμένο host προς πολλούς άλλους hosts του ίδιου τοπικού δικτύου μπορεί να δηλώσει στον παρατηρητή ύπαρξη επίθεσης. Αυτή θα μπορούσε να είναι κάποιος ιός υπολογιστή ή σκουλήκι υπολογιστών τα οποία προσπαθούν να διαδοθούν, θα μπορούσε να είναι “host scan” attack ή γενικότερα κάποιο “botnet attack”. Στα παρακάτω σχήματα (σχήμα 1.3 και 1.4) εμφανίζονται, πάλι με τη χρήση parallel coordinates, γραφικές αναπαραστάσεις καταγραφείσας κίνησης δικτύου υπολογιστών με πιθανές τέτοιες επιθέσεις.



Σχήμα 1.3: Γραφική αναπαράσταση Host scan attack με τη χρήση parallel coordinates. (Εικόνα από το pdf με τίτλο “Fast detection and visualization of network attacks on parallel coordinates” από τους Hyunsang Choi, Heejo Lee, Hyogon Kim [11])



Σχήμα 1.4: Γραφική αναπαράσταση botnet attack attack με τη χρήση parallel coordinates. (Εικόνα από το pdf με τίτλο “Applied parallel coordinates for logs and network traffic attack analysis” από τους Sebastien Tricaud, Philippe Saade [12])

Η καταγραφή αδικαιολόγητου μεγάλου αριθμού πακέτων από έναν υπολογιστή θα μπορούσε να δηλώνει και πολλά άλλα είδη επιθέσεων για τον ακριβή εντοπισμό των οποίων θα πρέπει ο αναλυτής να διενεργήσει περαιτέρω έλεγχο. Θα μπορούσε η υπερβολική (και μη συνηθισμένη) κίνηση από και προς ένα host να δηλώνει:

- κάποια επίθεση Denial of Service,
- την ύπαρξη κάποιου Trojan Horse που υποκλέπτει πληροφορίες,
- κάποιο είδος Brute force attack προς κάποιο ιστοχώρο που απαιτεί πιστοποίηση της ταυτότητας του χρήστη,
- την δυσλειτουργία του host,

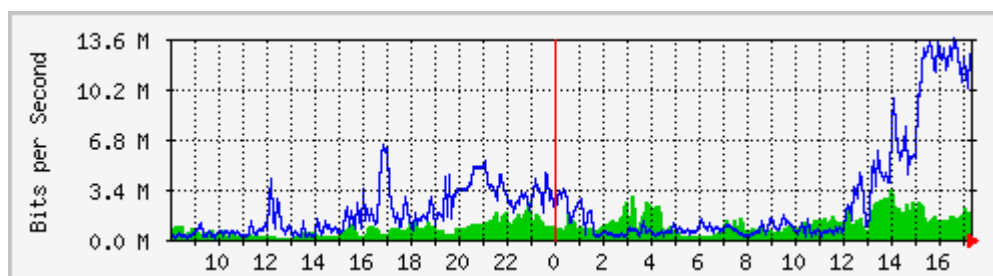
Η παρακολούθηση, σύλληψη και ανάλυση των πακέτων που διακινούνται μεταξύ των υπολογιστών του δικτύου θα μπορούσε να αποτελέσει μέθοδο εντοπισμού του προβλήματος και κατά συνέπεια αντιμετώπισης του. Για την καλύτερη κατανόηση του παραπάνω πλεονεκτήματος που προφέρει η παρακολούθηση της κίνησης του δικτύου δίνεται το ακόλουθο παράδειγμα. Ας υποθέσουμε ότι σε ένα δίκτυο υπάρχει ένας υπολογιστής μολυσμένος με ένα computer worm το οποίο να προσπαθεί να κλωνοποιηθεί και να διαδοθεί στους υπόλοιπους υπολογιστές του δικτύου. Ένα από τα πρώτα βήματα στην διαδικασία διάδοσής του είναι η εύρεση ευάλωτων υπολογιστών στο δίκτυο, οπότε και θα προσπαθήσει να βρει τέτοιους υπολογιστές (θα μπορούσε να χρησιμοποιήσει ICMP requests για την εύρεση υπολογιστών στο δίκτυο). Το γεγονός αυτό θα δημιουργήσει μεγάλο όγκο πακέτων από τον μολυσμένο υπολογιστή προς τους υπόλοιπους υπολογιστές του τοπικού δικτύου. Παρακολουθώντας και συγκρίνοντας την κίνηση του δικτύου με παλαιότερα αποθηκευμένα δεδομένα είναι δυνατόν να γίνει αμέσως αντιληπτή η αύξηση της κίνησης στο δίκτυο, γεγονός που θα αφυπνίσει (alert) τον διαχειριστή στην αναζήτηση των αιτιών. Χρησιμοποιώντας ένα εργαλείο sniffing πακέτων θα μπορούσε να εντοπίσει τον υπολογιστή που στέλνει αυτό τον μεγάλο όγκο πληροφοριών, εντοπίζοντας έτσι τον μολυσμένο υπολογιστή και λύνοντας έτσι το πρόβλημα [9].

Εκτός από τις περιπτώσεις που η καταγραφή της κίνησης ενός δικτύου υπολογιστών πραγματοποιείται με στόχο τον εντοπισμό και την αντιμετώπιση κινδύνων που μπορούν να απειλήσουν ένα δίκτυο υπολογιστών, είναι σκόπιμο η καταγραφή της κίνησης του δικτύου να αποτελεί τμήμα των καθημερινών εργασιών του διαχειριστή, ώστε να βοηθήσει στην ορθή διαχείριση του δικτύου υπολογιστών που υποστηρίζει. Όπως αναφέρεται και στο site

<http://oss.oetiker.ch/mrtg/>, για να γνωρίζει ο διαχειριστής «τι κάνουν» όλη μέρα οι διάφορες συσκευές που χρησιμοποιούνται για την προώθηση της κίνησης του δικτύου του (πχ switches και routers), θα μπορούσε να χρησιμοποιεί τα αποτελέσματα της παρακολούθησης-καταγραφής της κίνησης αυτού. Με αυτό τον τρόπο μπορεί να εξάγει συμπεράσματα για την απόδοση των συσκευών του δικτύου του, το ποσοστό εκμετάλλευσης των πόρων του δικτύου, την δημιουργία αναφορών και τάσεων της κίνησης του δικτύου. [13]

Στο παρακάτω σχήμα (σχήμα 1.5) φαίνονται τα αποτελέσματα του εργαλείου “Multi Router Traffic Grapher” το οποίο καταγράφει την κίνηση που διέρχεται από έναν router και αποτυπώνει γραφικά το ποσό του bandwidth που χρησιμοποιείται την τρέχουσα χρονική στιγμή (Ο άξονας X αντιστοιχεί στις ώρες της ημέρας ενώ ο άξονας Y εμφανίζει το εύρος ζώνης bits per second). Μάλιστα δίνει τη δυνατότητα σύγκρισης με το αντίστοιχο κατά μέσο όρο ποσοστό bandwidth τις αντίστοιχες ώρες (η μπλέ γραμμή στο γράφημα). Με αυτό τον τρόπο μπορεί ο αναλυτής να εξάγει χρήσιμες πληροφορίες όπως:

- Να καταγράψει ποιες είναι οι περίοδοι όπου υπάρχει συμφόρηση εύρους (bandwidth congestion) στη σύνδεση με το δημόσιο δίκτυο (internet).
- Να εντοπίσει εάν υπάρχουν προβλήματα bottleneck στο δίκτυό του.
- Να εντοπίσει αποκλίσεις από το baseline και να δραστηριοποιηθεί κατάλληλα (triggered) για να αναζητήσει την ύπαρξη «περίεργης» κίνησης που διέρχεται από το δίκτυό του.
- Να καταγράψει τις χρονικές στιγμές που οι πόροι του συστήματος υπερ-χρησιμοποιούνται.
- Να εντοπίσει hosts οι οποίοι διαθέτουν πολλές υπηρεσίες ο αριθμός των οποίων είναι πολύ μεγαλύτερος από αυτόν στις οποίες μπορεί να ανταπεξέλθει (υπάρχει μεγάλος αριθμός αιτημάτων λόγω του μεγάλου αριθμού υπηρεσιών προς έναν host-server ο οποίος δεν μπορεί να ανταπεξέλθει με αποτέλεσμα να δημιουργούνται καθυστερήσεις).



Σχήμα 1.5: Αποτελέσματα εργαλείου “Multi Router Traffic Grapher”. (Εικόνα από το Web site <http://oss.oetiker.ch/mrtg/> [13])

Καθημερινή καταγραφή της κίνησης ενός δικτύου υπολογιστών και μάλιστα της κίνησης που ξεκινά από το σύνολο των υπολογιστών αυτού, θα μπορούσε να βοηθήσει τον διαχειριστή να καταγράψει το ποιες μπορεί να είναι οι εφαρμογές που χρησιμοποιούνται ποιο πολύ από τους χρήστες. Ο πιο απλός τρόπος πραγματοποίησης αυτού είναι εντοπίζοντας τις θύρες προορισμού των πακέτων που εξέρχονται από τους υπολογιστές των χρηστών και προορίζονται προς τους servers της επιχείρησης είτε προς το internet. Ο παρατηρητής θα μπορούσε να αξιοποιήσει αυτή την πληροφορία για να εντοπίσει εφαρμογές που δεν θα έπρεπε να χρησιμοποιούν οι χρήστες και για λόγους ασφαλείας και για λόγους που αφορούν την ανούσια ενασχόλησή τους με εφαρμογές που οδηγούν σε απώλεια χρήσιμων παραγωγικών εργατωρών.

Στα περισσότερα δίκτυα υπολογιστών υπάρχουν δύο κατηγορίες υπαλλήλων. Υπάρχουν οι χρήστες των οποίων η εργασία σχετίζεται με την χρήση του internet, ενώ υπάρχει και η κατηγορία υπαλλήλων οι οποίοι θα χρησιμοποιήσουν περιστασιακά το internet. Με την παρακολούθηση της κίνησης ενός δικτύου θα μπορούσαν να εντοπιστούν χρήστες οι οποίοι καταναλώνουν το μεγαλύτερο ποσοστό bandwidth της σύνδεσης στο internet ή ακόμη και αυτοί για τους οποίους δεν δικαιολογείται υψηλό ποσοστό χρήσης bandwidth. Στην πρώτη περίπτωση η διαπίστωση αυτή θα μπορούσε να χρησιμοποιηθεί από τον διαχειριστή του δικτύου με στόχο να βελτιώσει την ταχύτητα σύνδεσης των χρηστών και να διαμοιράσει σωστά το διαθέσιμο εύρος ζώνης σε αυτούς, πετυχαίνοντας έτσι βελτίωση της απόδοσης. Στην δεύτερη περίπτωση η διαπίστωση αυτή είναι χρήσιμη για τον προσδιορισμό εκείνων των χρηστών που ασχολούνται με δραστηριότητες που δε σχετίζονται σε καμία περίπτωση με το εργασιακό τους αντικείμενο ή με τις δραστηριότητες της ίδιας της επιχείρησης εν ώρα εργασίας.

Τέλος, λόγοι δημιουργίας αναφορών (reporting) είτε γιατί έτσι απαιτείται από τη νομοθεσία (κάποιοι δημόσιοι οργανισμοί αναγκάζονται από το νόμο να κρατούν πληροφορίες για τις ενέργειες που έχουν πραγματοποιηθεί στο δίκτυο), είτε για λόγους καταγραφής τάσεων που επικρατούν στο δίκτυο υπολογιστών, η παρακολούθηση της κίνησης βοηθά στην δημιουργία πληρέστερης εικόνας.

Συνοψίζοντας τα παραπάνω γίνεται αντιληπτό ότι η παρακολούθηση ενός δικτύου υπολογιστών και συγκεκριμένα η παρακολούθηση της κίνησης του αποτελεί έναν από τους πιο σημαντικούς παράγοντες της σωστής διαχείρισης του. Μπορεί να χρησιμοποιηθεί όχι μόνο για την αντιμετώπιση κινδύνων και απειλών αλλά και για την κατανόηση του ίδιου του δικτύου υπολογιστών και των λειτουργιών αυτού. Είναι μία διαδικασία που θα πρέπει όλοι οι διαχειριστές να εντάξουν στις καθημερινές τους εργασίες για να εξασφαλίσουν ότι έχουν πλήρη εικόνα του δικτύου που διαχειρίζονται εξασφαλίζοντας ότι αυτό δεν θα κινδυνεύσει με κανέναν τρόπο. Χρησιμοποιώντας την παρακολούθηση εξασφαλίζονται οι τρεις βασικές αρχές της ομαλής λειτουργίας του δικτύου, η εμπιστευτικότητα, η ακεραιότητα και η διαθεσιμότητα που παρέχει.

1.5 Τρόποι Παρακολούθησης σε Ένα Δίκτυο Υπολογιστών

Όπως αναφέρει η Alisha Cecil στην δουλειά της με τίτλο “A Summary of Network Traffic Monitoring and Analysis” η παρακολούθηση του δικτύου (Network monitoring) είναι μία δύσκολη και απαιτητική εργασία η οποία είναι κρίσιμο τμήμα της εργασίας των διαχειριστών δικτύων [9]. Οι διαχειριστές δικτύων χρειάζεται να καταγράφουν και να αναλύουν την κίνηση που διέρχεται διαμέσου του δικτύου που παρακολουθούν ώστε να χρησιμοποιούν τα συμπεράσματα για να βελτιώσουν την απόδοση καθώς και την αποφυγή κενών ασφαλείας και κινδύνων. Επίσης εντοπίζουν ποιες είναι οι κύριες πηγές του συνολικού κόστους της κίνησης του δικτύου που παρακολουθούν.

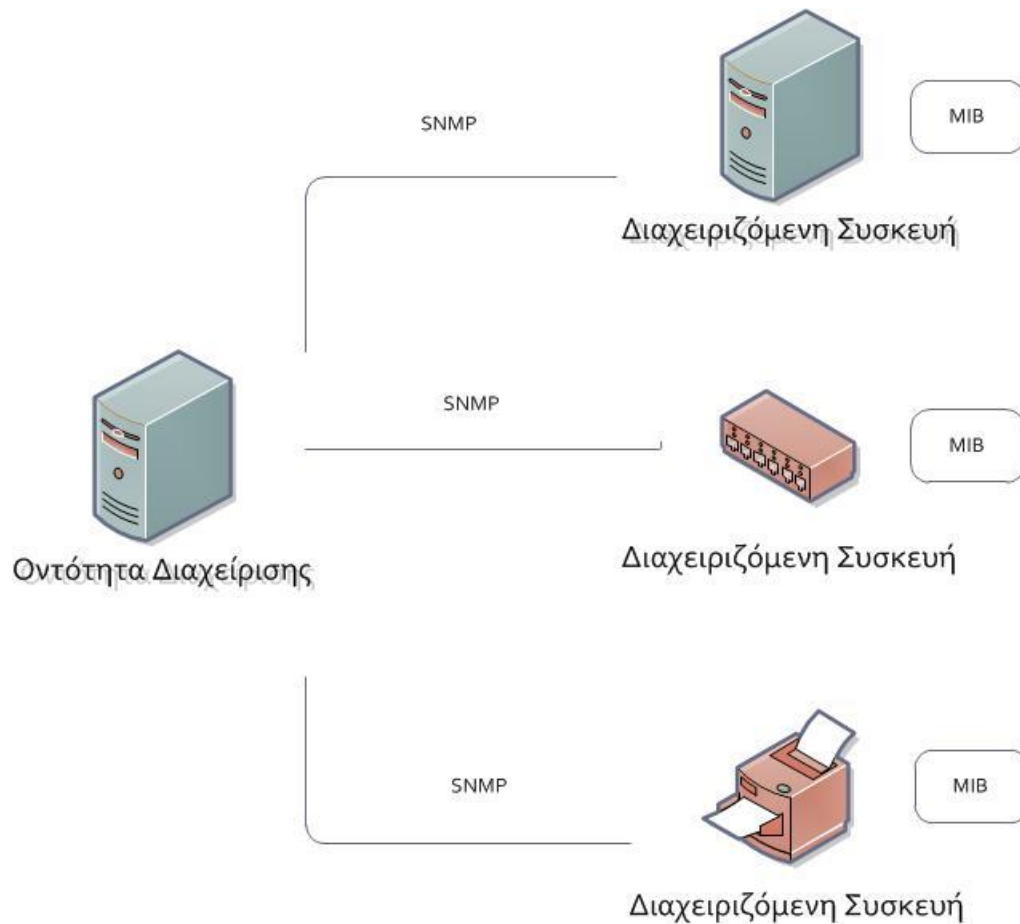
Λόγω λοιπόν της προστιθέμενης αξίας που προσδίδει η παρακολούθηση, ανάλυση και κατανόηση της κίνησης που περνά από ένα δίκτυο υπολογιστών στην ομαλή διαχείρισή του, έχουν αναπτυχθεί μέχρι σήμερα πάρα πολλές τεχνικές, μέθοδοι και εργαλεία που βοηθούν στην παρακολούθηση αυτή. Δύο είναι οι κλασικότερες μέθοδοι που βοηθούν στην παρακολούθηση δικτύων υπολογιστών, η χρήση SNMP (Simple Network Management Protocol) για την παρακολούθηση της κατάστασης δικτυακών συσκευών και η χρήση τεχνολογιών σύλληψης κίνησης δικτύου υπολογιστών όπως οι

τεχνικές RMON και NetFlow. Στην δεύτερη κατηγορία έχουν στηριχθεί και αναπτυχθεί πολλά εργαλεία παρακολούθησης. Έχουν αναπτυχθεί συστήματα ανίχνευσης και αποτροπής εισβολών (IDS Intrusion detection systems, IPS Intrusion Prevention Systems), συστήματα σύλληψης και ανάλυσης πακέτων (packet sniffers) καθώς και συστήματα παρακολούθησης του εύρους ζώνης συνδέσεων. Στην ενότητα αυτή θα κάνουμε μία μικρή αναφορά στις τεχνικές και τα εργαλεία αυτά.

1.5.1 SNMP

Μία κλασική δομή παρακολούθησης της κατάστασης δικτυακών συσκευών περιλαμβάνει την Οντότητα Διαχείρισης (Managing Entity), την Διαχειριζόμενη Συσκευή (Managed Device), το Αντικείμενο Διαχείρισης (Managed Object) και το Δικτυακό Πρωτόκολλο Διαχείρισης (Network Management Protocol) [6]. Ως Οντότητα Διαχείρισης νοείται ένα κεντρικό λογισμικό (προφανώς εγκατεστημένο σε έναν κεντρικό υπολογιστή διαχείρισης) το οποίο ελέγχει με την καθοδήγηση του διαχειριστή κεντρικά τη διαχείριση του δικτύου. Είναι αυτό το οποίο θα καθορίσει το από ποιες συσκευές θα ελέγχονται δεδομένα, πως θα γίνεται η συλλογή αυτών και πως θα εμφανίζονται οι συλλεχθέντες πληροφορίες. Διαχειριζόμενη Συσκευή είναι η υπό παρακολούθηση συσκευή η οποία μπορεί να είναι οποιαδήποτε συσκευή δικτύου (πχ router, switch, υπολογιστής, εκτυπωτής κλπ). Σε κάθε Διαχειριζόμενη Συσκευή υπάρχει τουλάχιστον ένα Διαχειριζόμενο Αντικείμενο το οποίο είναι το τμήμα υλικού (πχ κάρτα δικτύου) ή το λογισμικό της Διαχειριζόμενης Συσκευής το οποίο παρακολουθείται. Οι Διαχειριζόμενες Συσκευές και οι Οντότητες Διαχείρισης ανταλλάσσουν πληροφορίες μεταξύ τους χρησιμοποιώντας το Δικτυακό Πρωτόκολλο Διαχείρισης, το οποίο είναι το SNMP (Simple Network Management Protocol). Η έκδοση που του SNMP πρωτοκόλλου που χρησιμοποιείται σήμερα είναι η version 3. Όλες οι πληροφορίες διαχείρισης στις διάφορες Διαχειριζόμενες Συσκευές δημιουργούν μία βάση δεδομένων την ονομαζόμενη Management Information Base (MIB). Η MIB μπορεί να περιέχει διάφορες πληροφορίες από το Διαχειριζόμενο Αντικείμενο και το είδος των πληροφοριών αυτών εξαρτάται από το είδος της υπό παρακολούθηση Διαχειριζόμενης Συσκευής [14]. Πχ αν η Διαχειριζόμενη Συσκευή είναι ένας server τότε οι αποθηκευμένες πληροφορίες μπορεί να είναι το ποσοστό χρήσης της κεντρικής μονάδας επεξεργασίας ή το ποσοστό χρήσης της μνήμης, αν η διαχειριζόμενη συσκευή είναι ένας router τότε οι πληροφορίες αυτές μπορεί να είναι ο αριθμός πακέτων που φτάσανε, το πόσα απερρίφθησαν κλπ. Ο τρόπος δόμησης των πληροφοριών της MIB είναι γνωστός ως SMI (Structure of Management

Information). Στο σχήμα 1.6 φαίνονται όλα τα στοιχεία που συνθέτουν τη δομή παρακολούθησης με τη χρήση του SNMP.

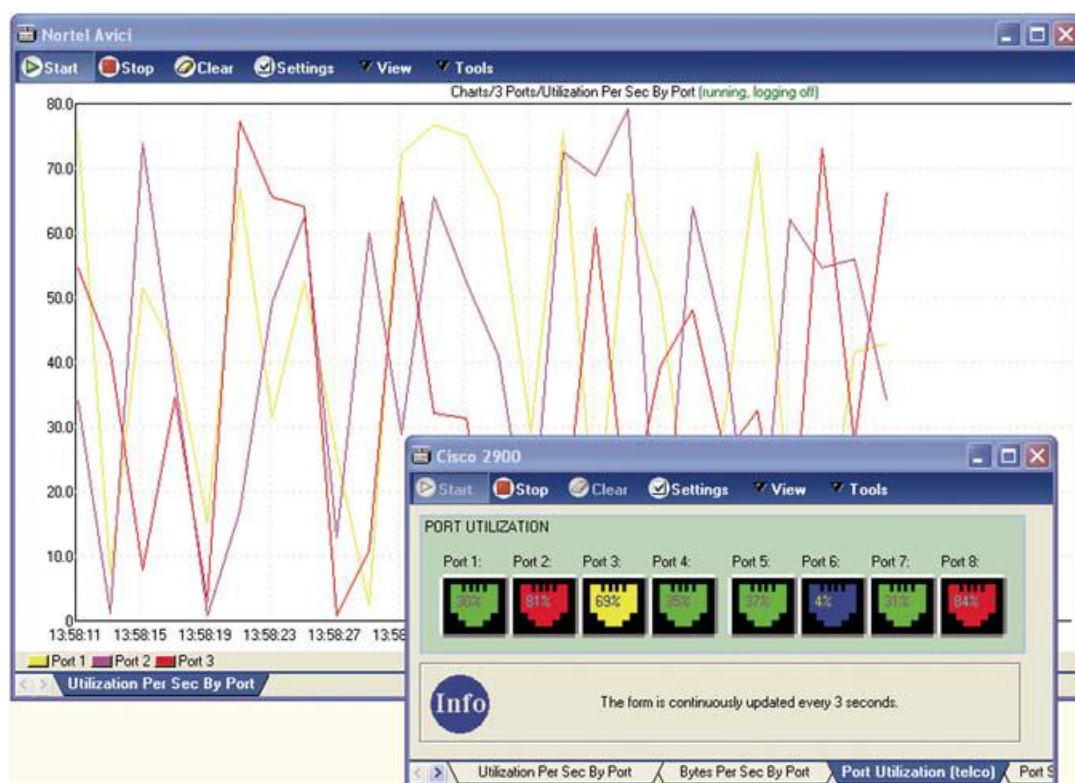


Σχήμα 1.6: Στοιχεία δομής παρακολούθησης με τη χρήση του SNMP

Ο τρόπος παρακολούθησης των συσκευών του δικτύου με τη χρήση αυτής της τεχνικής στηρίζεται στην λογική της ανταλλαγής αιτημάτων και αποκρίσεων (request-respond) μεταξύ της Διαχειριζόμενης Συσκευής και της Οντότητας Διαχείρισης. Η Οντότητα Διαχείρισης αποστέλλει αίτημα στην Διαχειριζόμενη Συσκευή (στον agent αυτής), το οποίο θα είναι κάποιο ερώτημα για άντληση πληροφοριών ή κάποιο αίτημα τροποποίησης της MIB της. Ως αποτέλεσμα, η Διαχειριζόμενη Συσκευή (ο agent αυτής) στέλνει απάντηση στο ερώτημα ή επιβεβαίωση πραγματοποίησης του αιτήματος τροποποίησης. Η εμφάνιση κάποιας μη συνηθισμένης κατάστασης στη Διαχειριζόμενη Συσκευή (απόκλιση από την ομαλή λειτουργία της) μπορεί να ενεργοποιήσει την αποστολή μηνύματος από αυτή προς την Οντότητα Διαχείρισης (trap message)

χωρίς να έχει προηγηθεί κάποιο αίτημα. Η ανταλλαγή των μηνυμάτων γίνεται με τη χρήση του UDP πρωτοκόλλου.

Αυτός ο τρόπος παρακολούθησης δίνει τη δυνατότητα στον διαχειριστή να παρακολουθεί και να καταγράφει την κατάσταση και την απόδοση των συσκευών και γενικότερα του δικτύου του, να ανιχνεύει τα προβλήματα και τις δυσλειτουργίες του δικτύου του [15]. Αυτό που δεν κάνει το SNMP είναι να δίνει ακριβώς την αιτία του προβλήματος. Όπως αναφέρεται εύστοχα και στο pdf με τίτλο “SNMP Monitoring, One Critical Component to Network Management” [14] «μπορεί το SNMP να δείχνει ότι το interface του router έχει υψηλό βαθμό χρήσης, αλλά δεν δείχνει τον τύπο της κίνησης δικτύου που χρησιμοποιεί αυτό το εύρος ζώνης καθώς και το ποια συσκευή είναι υπεύθυνη για την παραγωγή της κίνησης αυτής. Δίνεται στον διαχειριστή η πληροφορία ότι υπάρχει πρόβλημα, χωρίς όμως αυτός να γνωρίζει την αιτία με αποτέλεσμα να μην μπορεί να λύσει γρήγορα το πρόβλημα».



Σχήμα 1.7: Δείγμα αποτελεσμάτων SNMP. Παρακολούθηση χρησιμοποίησης θυρών. (Εικόνα από Network Instruments White Paper με τίτλο “SNMP Monitoring, One Critical Component to Network Management” [14])

Παράλληλα με τα συστήματα ανίχνευσης καταστάσεων συσκευών ενός δικτύου υπολογιστών έχουν αναπτυχθεί και συστήματα τα οποία στηρίζονται στη σύλληψη, καταγραφή και ανάλυση πακέτων που διέρχονται από δίκτυα υπολογιστών. Θα μπορούσε αυτά τα συστήματα να διαχωριστούν σε κατηγορίες ανάλογα με τον σκοπό για τον οποίο χρησιμοποιούνται. Υπάρχουν τα Intrusion Detection Systems ή Intrusion Prevention Systems, τα συστήματα μέτρησης χρήσης εύρους ζώνης καθώς και τα συστήματα σύλληψης και ανάλυσης πακέτων τα λεγόμενα “sniffers”.

1.5.2 IDS – IPS

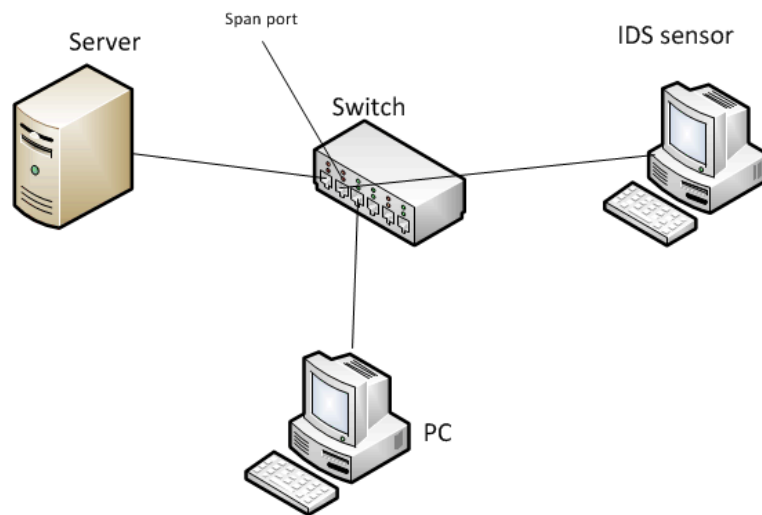
Intrusion Detection είναι η ενέργεια κατά την οποία ανιχνεύεται μη επιθυμητή κίνηση σε ένα δίκτυο ή σε μία σύνδεση. Intrusion Detection System μπορεί να είναι ένα λογισμικό ή μία συσκευή (υλικό) η οποία παρακολουθεί την κίνηση σε ένα δίκτυο υπολογιστών με στόχο να εντοπίσει μη επιθυμητή κίνηση. Αυτή η κίνηση μπορεί να είναι παράνομη ή καταστροφική κίνηση, κίνηση που παραβιάζει την πολιτική ασφαλείας ή απλά κίνηση που αποκλίνει από την καθημερινή κίνηση που διέρχεται από το υπό παρακολούθηση δίκτυο υπολογιστών [16]. Αντίστοιχα τα Intrusion Prevention Systems πραγματοποιούν ένα βήμα παραπάνω. Τα συστήματα αυτά αφού ανιχνεύσουν μη επιθυμητή κίνηση και πραγματοποιούν φιλτράρισμα αυτής με στόχο να αποτρέψουν την είσοδό της στο δίκτυο.

Μπορούμε να κατηγοριοποιήσουμε τα IDS-IPS με βάση τα ακόλουθα κριτήρια:

- Με βάση την τεχνική-μέθοδο με την οποία ανιχνεύουν την μη επιθυμητή κίνηση.
- Με βάση το που τοποθετούνται (Host IDS, Network IDS).

Υπάρχουν δύο τεχνικές-μέθοδοι οι οποίες χρησιμοποιούνται από τα IDS-IPS για την ανίχνευση μη επιθυμητής κίνησης, η “signature detection” και η “anomaly detection”. Σύμφωνα με την πρώτη τα IDS-IPS ελέγχουν την κίνηση που διέρχεται από αυτά για την ύπαρξη συγκεκριμένου είδους δεδομένων-χαρακτηριστικών τις λεγόμενες υπογραφές (signatures). Οι υπογραφές αυτές μπορεί να δημιουργηθούν είτε με τη χρήση data mining αλγορίθμων [17] είτε με τεχνικές machine learning [18]. Σύμφωνα με τη δεύτερη μέθοδο-τεχνική γίνεται προσπάθεια να εντοπιστεί κίνηση που αποκλίνει από το εύρος κίνησης που έχει κανονικά καταγραφεί στο δίκτυο αυτό [10].

Με βάση τη θέση τους τα IDS-IPS χωρίζονται σε δύο κατηγορίες, τα host based και τα network based [10]. Τα host based IDS-IPS συνήθως είναι κάποιο λογισμικό το οποίο εγκαθίσταται σε κάποιο υπολογιστή με στόχο είτε να παρακολουθεί τις διάφορες διεργασίες (processes) και τον τρόπο συμπεριφοράς τους, είτε να ελέγχει και να καταγράφει την κίνηση που φτάνει σε αυτόν. Αντίστοιχα, τα network intrusion detection systems είναι συνήθως υλικό, κάποια δηλαδή συσκευή η οποία τοποθετείται στο δίκτυο (πχ κάποια span θύρα σε ένα network switch) και έχει τη δυνατότητα να συλλαμβάνει πακέτα και να τα αναλύει.



Σχήμα 1.8: Network Intrusion Detection System example

Κάθε τύπος IDS-IPS έχει τα πλεονεκτήματα και τα μειονεκτήματά του. Τα πλεονεκτήματα και των δύο λόγω του γεγονότος ότι παρακολουθούν τη κίνηση ενός δικτύου είναι όλα τα πλεονεκτήματα που προαναφέρθηκαν και αφορούν την προστασία και εξασφάλιση της ομαλής λειτουργίας ενός δικτύου υπολογιστών. Τα signature based IDS-IPS έχουν το μεγάλο μειονέκτημα ότι χρειάζεται η επίθεση να είναι γνωστή για να είναι δυνατή η δημιουργία της υπογραφής για την ανίχνευσή της. Αυτό σημαίνει ότι τα signature based συστήματα δεν μπορούν να αναγνωρίσουν μη γνωστές-νέου είδους επιθέσεις. Το μειονέκτημα των anomaly detection systems είναι το γεγονός ότι θα πρέπει να δημιουργηθεί στην αρχή κάποιο baseline το οποίο θα πρέπει να ανανεώνεται συνεχώς με βάση τις αλλαγές που πραγματοποιούνται στο δίκτυο (πχ προσθήκη κάποιας νέας υπηρεσίας στο δίκτυο).

Και οι δύο τύποι συστημάτων είναι ευρέως διαδεδομένοι και χρησιμοποιούμενοι στις μέρες μας με αποτέλεσμα να αναπτυχθεί ένας πολύ μεγάλος αριθμός εργαλείων για τον έλεγχο της κίνησης ενός δικτύου υπολογιστών. Μία μεγάλη λίστα IDS και IPS καταγράφεται στο βιβλίο με τίτλο “Intrusion Detection Systems” [10]. Μεταξύ αυτών ευρέως γνωστό είναι και το εργαλείο “snort” για το οποίο θα γίνει συγκεκριμένη αναφορά σε ακόλουθο κεφάλαιο της παρούσας εργασίας.

1.5.3 Snort

Όπως αναφέρεται και στην ιστοσελίδα <http://www.snort.org/> [19] το εργαλείο Snort® είναι ταυτόχρονα ένα σύστημα ανίχνευσης και αποτροπής εισβολών (IDP-IPS) (και όχι μόνο) το οποίο αναπτύχθηκε από την εταιρεία Sourcefire. Η λειτουργία του συνδυάζει και τις δύο τεχνικές ανίχνευσης κινδύνων που αναλύθηκαν παραπάνω (“signature detection” και “anomaly detection”) καθιστώντας το ποιο διαδεδομένο εργαλείο ανίχνευσης και αποτροπής εισβολών (πλησιάζει τα τέσσερα εκατομμύρια downloads σύμφωνα με το <http://sourcefire.com/about-us>). Όπως αναφέρεται και στο βιβλίο με τίτλο “Solutions and Examples for Snort Administrators” «το ωραιότερο χαρακτηριστικό του Snort είναι το γεγονός ότι διαθέτει ευέλικτη τροποποιήσιμη δομή, είναι εύκολο να το κάνεις να πραγματοποιεί ενέργειες τις οποίες δεν έχει σχεδιαστεί να κάνει» [20]. Είναι ικανό να πραγματοποιεί ανάλυση κίνησης δικτύου σε πραγματικό χρόνο εξάγοντας έτσι συμπεράσματα για την απόδοση ενός δικτύου, καταγραφή πακέτων που διέρχονται από ένα δίκτυο και φυσικά ενημέρωση του διαχειριστή για τα διάφορα γεγονότα που λαμβάνουν χώρα (real time alerting system). Οι τρεις κύριες χρήσεις του είναι ως “packet sniffer”, ως “packet logger” και ως ένα πλήρες IPS. Μπορεί να ανιχνεύσει επιθέσεις όπως “Port scan”, “Buffer overflow Attacks”, “CGI attacks”, “SMB probes”, “OS fingerprinting attempts” και πολλές άλλες προσπάθειες παραβίασης της ασφάλειας. Χρησιμοποιεί το “libpcap” ως το βασικό interface σύλληψης πακέτων το οποίο χρησιμοποιείται και από άλλα γνωστά εργαλεία σύλληψης πακέτων όπως είναι το wireshark [6].

Για τη λειτουργία του χρησιμοποιεί ένα σύνολο κανόνων που καθορίζουν τον τύπο τις κίνησης που θα καταγραφεί καθώς και την ενέργεια που θα πραγματοποιήσει σε περίπτωση ανίχνευσης αυτού του είδους της κίνησης. Στο παρακάτω σχήμα (Σχήμα 1.9) εμφανίζεται ενδεικτικά ένας κανόνας που θα μπορούσε να χρησιμοποιηθεί στο εργαλείο snort για την ανίχνευση ταυτόχρονης χρήσης των σημαιών SYN και FIN στα πακέτα που συλλαμβάνονται:

alert any any -> any any (flags:SF; msg: "SYN-FIN attack");)

Σχήμα 1.9: Ενδεικτικός κανόνας εργαλείου snort

Σύμφωνα με αυτόν σε περίπτωση σύλληψης πακέτων με οποιαδήποτε διεύθυνση αφετηρίας και προορισμού, τα οποία διαθέτουν ενεργοποιημένα τα flags "SYN" και "FIN" θα ενεργοποιηθεί ένα alert με τίτλο "SYN-FIN attack". Σύνολο τέτοιων κανόνων στο snort καθορίζουν την πολιτική ασφάλειας και ελέγχων που αυτό θα πραγματοποιεί. Υπάρχουν στο internet σύνολα τέτοιων κανόνων (signature databases) που μπορούν να χρησιμοποιηθούν ελεύθερα από τους χρήστες του snort και τα οποία ενημερώνονται συνεχώς. Μέσα από τα πολλά user communities που υπάρχουν, σε ελάχιστες ώρες από την εμφάνιση κάποιας νέας επίθεσης δημιουργείται η κατάλληλη υπογραφή η οποία μπορεί να γίνει downloaded και να χρησιμοποιηθεί ελεύθερα παρέχοντας γρήγορη κάλυψη και προστασία [6].

1.5.4 Συστήματα Σύλληψης και Ανάλυσης Πακέτων (sniffers)

Σημαντική συνεισφορά στην παρακολούθηση ενός δικτύου υπολογιστών με στόχο την ομαλή λειτουργία αυτού κατέχουν και οι εφαρμογές σύλληψης και ανάλυσης πακέτων τα λεγόμενα sniffers. Μπορούν να χρησιμοποιηθούν για την διάγνωση δικτυακών προβλημάτων, για τον έλεγχο της δραστηριότητας εισβολών [21] για την ανίχνευση bottleneck στις συνδέσεις του δικτύου, καθώς και για την δημιουργία εκθέσεων (reports) και αναλύσεων κίνησης που διέρχεται από ένα δίκτυο υπολογιστών. Η κύρια λειτουργία αυτών των εφαρμογών στηρίζεται (όπως αναφέρεται και στο όνομά τους) στην σύλληψη των πακέτων της κίνησης που διέρχονται από ένα δίκτυο υπολογιστών, είτε πρόκειται για σύλληψη καθώς τα πακέτα διέρχονται από κάποια σύνδεση, από μία δικτυακή συσκευή είτε φτάνουν σε έναν υπολογιστή.

Όλα τα πακέτα ενός Ethernet δικτύου υπολογιστών περιέχουν τέσσερις πολύ σημαντικές πληροφορίες οι οποίες χρησιμοποιούνται για την σωστή δρομολόγηση και παράδοση αυτών. Οι πληροφορίες αυτές είναι οι MAC (Media Access Control) addresses αφετηρίας και προορισμού, και οι IP addresses αφετηρίας και προορισμού. Οι IP διευθύνσεις είναι αυτές που καθορίζουν σε ποιο δίκτυο/υποδίκτυο θα πρέπει να δρομολογηθεί το πακέτο, ενώ οι MAC διευθύνσεις είναι αυτές που

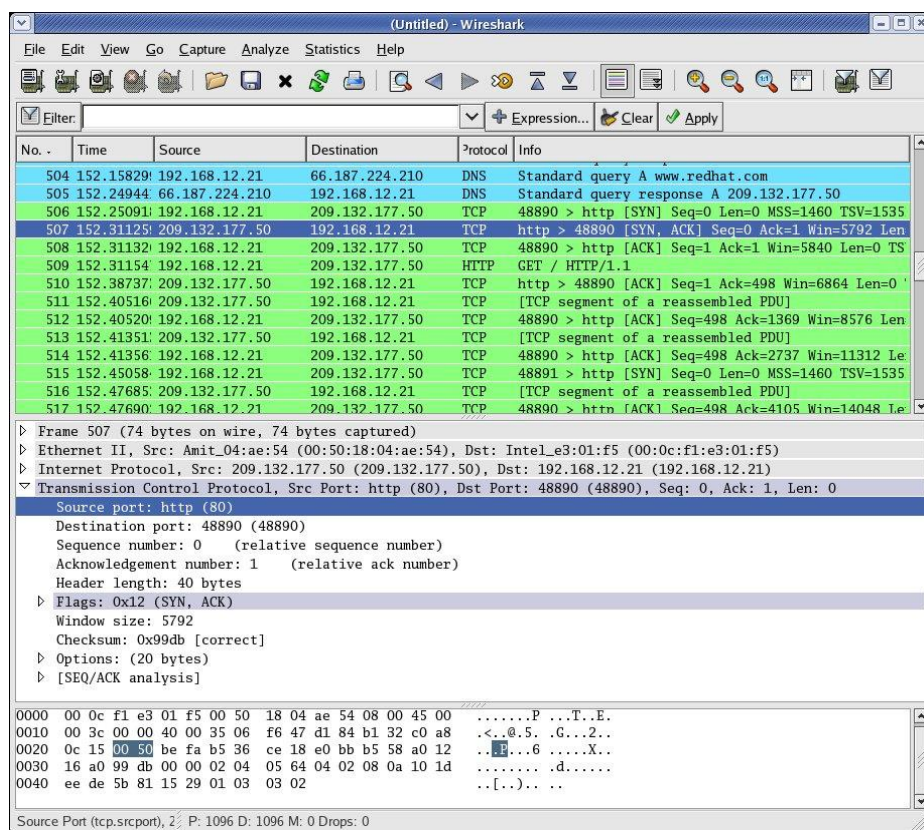
θα καθορίσουν το εάν ένα πακέτο που φτάνει στην Ethernet κάρτα ενός host θα πρέπει να προωθηθεί για επεξεργασία σε επόμενο (πιο πάνω) επίπεδο του TCP/IP stack ή θα πρέπει να απορριφθεί διότι δεν είναι για τον συγκεκριμένο Host. Όταν το πακέτο που φτάνει σε μία κάρτα ενός Host έχει MAC διεύθυνση προορισμού ίδια με την MAC διεύθυνση της κάρτας τότε το πακέτο έχει φτάσει στον σωστό παραλήπτη οπότε και χρησιμοποιείται (διαβάζεται και προωθείται σε ανώτερο επίπεδο στο TCP/IP stack) ενώ σε διαφορετική περίπτωση απορρίπτεται από την κάρτα. Οι εφαρμογές σύλληψης πακέτων στηρίζουν τη λειτουργία τους στη δυνατότητα να συλλαμβάνουν πακέτα ασχέτως με το αν έχουν MAC διεύθυνση προορισμού διαφορετική από την κάρτα δικτύου της συσκευής στην οποία έχουν φτάσει (promiscuous mode).

Ένα πακέτο θα μπορούσε να φτάσει σε διαφορετικό Host από αυτόν προς τον οποίο πραγματικά προοριζόταν, στη περίπτωση που στο δίκτυο υπάρχει ένα hub ως συσκευή σύνδεσης των υπολογιστών (Shared Ethernet), οπότε και σε αυτή την περίπτωση θα μπορούσε να γίνει captured από το sniffer. Δεν θα μπορούσε όμως να φτάσει σε διαφορετικό από τον σωστό host στην περίπτωση που στο δίκτυο χρησιμοποιείται ένα switch (Switched Ethernet). Θα μπορούσε λοιπόν κάποιος να πει ότι σε περίπτωση χρήσης switches για τη δικτύωση των υπολογιστών στο δίκτυο δεν θα μπορούσε να πραγματοποιηθεί σύλληψη πακέτων αφού τα πακέτα δεν φτάνουν παρά μόνο στον σωστό host. Αυτό όμως δεν είναι απόλυτο. Υπάρχουν τεχνικές που θα μπορούσαν να χρησιμοποιηθούν για την σύλληψη πακέτων ακόμη και σε ένα switched network. Θα μπορούσε πχ να χρησιμοποιηθούν “span ports” (Switch Port Analyzer) ενός switch ώστε τα πακέτα τα οποία φτάνουν σε μία θύρα ενός switch να αντιγράφονται και σε κάποια άλλη. Θα μπορούσε να τοποθετηθούν συσκευές tap σε κάποια σύνδεση ώστε να συλλαμβάνουν όλη τη σύνδεση που περνά από αυτή κλπ.

Αν και αρχικά οι τεχνικές και οι εφαρμογές σύλληψης πακέτων δημιουργήθηκαν για να βοηθήσουν τους διαχειριστές δικτύων να εντοπίσουν τις αιτίες προβλημάτων καθώς και αδυναμίες της δομής του TCP/IP, παρόλα αυτά αργότερα παρατηρήθηκαν περιπτώσεις στις οποίες χρησιμοποιήθηκαν λανθασμένα και μετατράπηκαν σε λογισμικά επιθέσεων. Κακόβουλοι χρησιμοποιούν την κατηγορία αυτή λογισμικού για να συλλέξουν παράνομα τα πακέτα ενός δικτύου, να τα αναλύσουν και να έχουν πρόσβαση σε εμπιστευτικές πληροφορίες όπως user names και passwords.

Γνωστά εργαλεία σύλληψης πακέτων και ανάλυσης τους είναι τα TCPdump, Ngrep, Snoop, Sniffit, Libpcap, Hunt, Ettercap, Dsniff, Lcrzoex, Wireshark/Ethereal, το ενσωματωμένο εργαλείο στα MS Windows® Network Monitor και πολλά άλλα. Ακόμη και το λογισμικό snort που αναφέρθηκε και συμπεριλήφθηκε στην κατηγορία των συστημάτων καταγραφής και αποτροπής εισβολών χρησιμοποιεί τεχνολογία σύλληψης πακέτων. Με τη χρήση αυτής είναι σε θέση να πραγματοποιεί τις απαραίτητες συγκρίσεις των περιεχομένων των πακέτων με τις υπογραφές (signatures) που έχουν καθοριστεί στους κανόνες του.

Στο παρακάτω σχήμα (Σχήμα 1.10) εμφανίζεται οθόνη από σύλληψη πακέτων χρησιμοποιώντας το εργαλείο Wireshark.



Σχήμα 1.10: Εικόνα σύλληψης πακέτων με τη χρήση του εργαλείου Wireshark

1.5.5 Bandwidth Monitoring

Κλείνοντας την αναφορά μας στους τρόπους και τα εργαλεία που χρησιμοποιούνται για την παρακολούθηση ενός δικτύου υπολογιστών θα κάνουμε και μία μικρή αναφορά και στα εργαλεία που χρησιμοποιούνται για την καταγραφή και ανάλυση της χρήσης εύρους ζώνης συνδέσεων.

Όπως αναφέρεται και στην δουλειά (διδακτορική διατριβή) με τίτλο “Network Monitoring and diagnosis Based on available Bandwidth Measurement” από τον Ningning Hu [22] η παρακολούθηση της απόδοσης ενός δικτύου γίνεται συνήθως για να βρεθούν ανωμαλίες στην απόδοση και να αναγνωριστούν οι αιτίες προβλημάτων. Αυτή η παρακολούθηση είναι σημαντική διότι έτσι καθορίζεται η ποιότητα των παρεχόμενων υπηρεσιών ενός δικτύου. Ένα δίκτυο θεωρείται αποδοτικό όταν μεταξύ άλλων παρέχει αδιάλειπτα τις υπηρεσίες του. Μετρώντας το διαθέσιμο εύρος ζώνης των συνδέσεων προς διάφορες υπηρεσίες (services) μπορούν να εντοπιστούν και να επιλυθούν προβλήματα ταχύτητας πρόσβασης και κατ’ επέκταση αξιοπιστίας των υπηρεσιών. Μεταξύ των άλλων για την μέτρηση της απόδοσης ενός δικτύου σημαντικές είναι:

- οι μετρήσεις των καθυστερήσεων των πακέτων στις συνδέσεις,
- του αριθμού των πακέτων που χάνονται στο δίκτυο,
- του διαθέσιμου εύρους ζώνης των συνδέσεων του δικτύου.

Οι μετρήσεις αυτές μπορούν να χωριστούν σε δύο κατηγορίες, στις μετρήσεις που γίνονται σε επίπεδο end-to-end (μετρήσεις σε σχέση με τον παραλήπτη) και στις μετρήσεις που γίνονται σε επίπεδο σύνδεσης (ανεξαρτήτως παραλήπτη) [22].

Οι μετρήσεις αυτές και κυρίως η μέτρηση των καθυστερήσεων των πακέτων στις συνδέσεις γίνεται σε καθημερινή βάση από τους διαχειριστές δικτύων και αποτελεί συνήθως τον πρώτο έλεγχο που κάνουν για τον εντοπισμό προβλημάτων συνδέσεων. Κλασικά εργαλεία που χρησιμοποιεί η πλειοψηφία αυτών είναι τα εργαλεία Ping και Traceroute. Παράπονα χρηστών της μορφής «δεν μπορώ να συνδεθώ στα κοινόχρηστα αρχεία», «παρατηρώ καθυστέρηση στη σύνδεσή μου με τον server και με το Internet» οδηγούν τους διαχειριστές δικτύων στον έλεγχο της καθυστέρησης της σύνδεσης των χρηστών χρησιμοποιώντας το γνωστό σε όλους εργαλείο ping (Σχήμα 1.11). Η αιτία των προβλημάτων της μορφής «δεν μπορώ να ανοίξω το site A» εντοπίζονται τις περισσότερες φορές από τους διαχειριστές με τη χρήση του traceroute εργαλείου (Σχήμα 1.12). Οι μετρήσεις που

γίνονται με τη χρήση αυτών των εργαλείων με βάση τη βιβλιογραφία θεωρούνται μετρήσεις σε επίπεδο end-to-end.

```
C:\>ping www.google.gr

Pinging www.l.google.com [209.85.148.104] with 32 bytes of data:
Reply from 209.85.148.104: bytes=32 time=71ms TTL=57
Reply from 209.85.148.104: bytes=32 time=70ms TTL=57
Reply from 209.85.148.104: bytes=32 time=70ms TTL=57
Reply from 209.85.148.104: bytes=32 time=71ms TTL=57

Ping statistics for 209.85.148.104:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
    Approximate round trip times in milli-seconds:
        Minimum = 70ms, Maximum = 71ms, Average = 70ms
```

Σχήμα 1.11: Αποτελέσματα εκτέλεσης εργαλείου ping από το οποίο φαίνονται χρόνοι απόκρισης

```
C:\Users\smavromatidis>tracert www.google.com

Tracing route to www.l.google.com [74.125.39.99]
over a maximum of 30 hops:

  1    4 ms    2 ms    2 ms  192.168.1.1 [192.168.1.1]
  2   25 ms   26 ms   25 ms  62.169.255.1
  3   27 ms   24 ms   26 ms  62.169.224.15
  4   25 ms   24 ms   24 ms  62.169.224.5
  5   35 ms   35 ms   34 ms  62.169.252.81
  6   36 ms   35 ms   46 ms  62.169.192.5
  7   82 ms   85 ms   92 ms  195.22.193.17
  8   71 ms   74 ms   68 ms  74.125.50.241
  9   80 ms   70 ms   70 ms  216.239.47.128
 10   70 ms   71 ms   70 ms  72.14.232.78
 11   70 ms   69 ms   70 ms  209.85.254.118
 12    *     70 ms   71 ms  209.85.254.134
 13   70 ms   70 ms   71 ms  fx-in-f99.1e100.net [74.125.39.99]

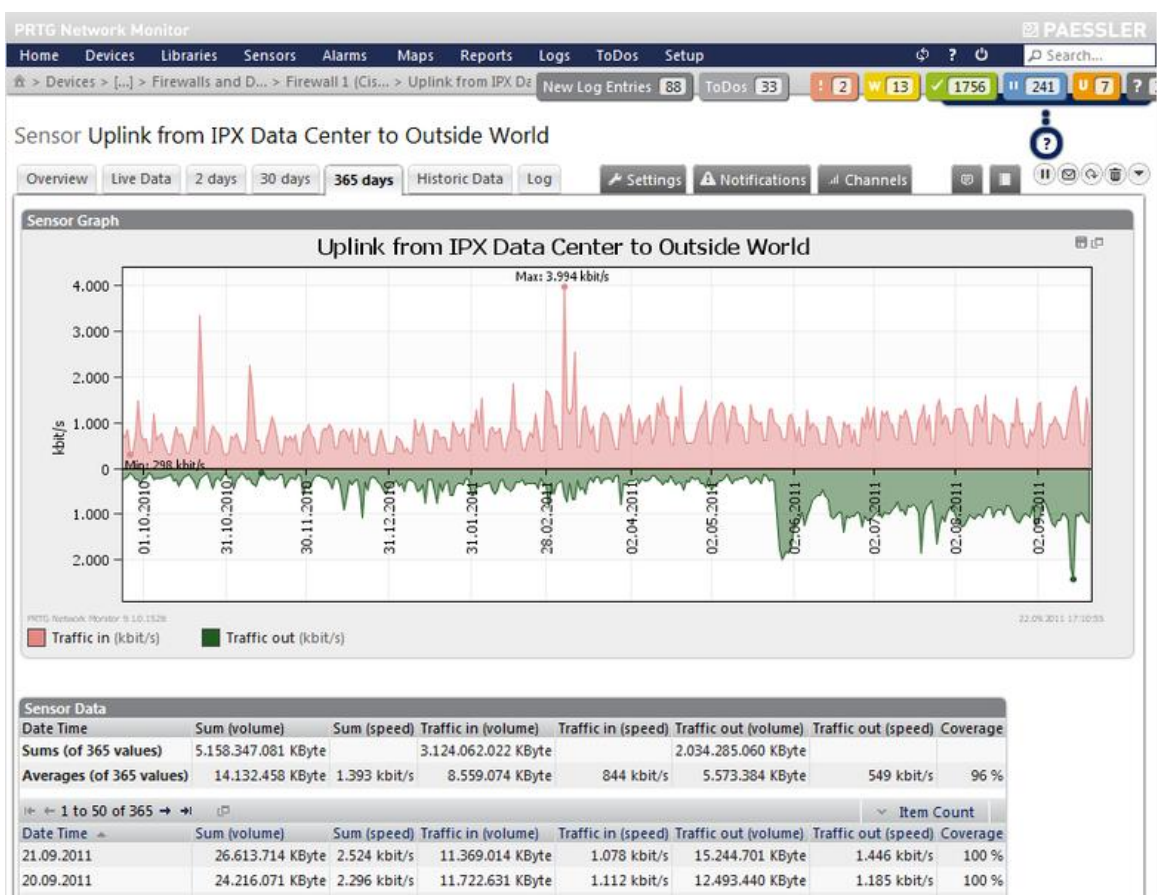
Trace complete.
```

Σχήμα 1.12: Αποτελέσματα εκτέλεσης εργαλείου traceroute

Παρακολούθηση σε επίπεδο σύνδεσης (link-level) γίνεται συνήθως με τη χρήση του SNMP (port – based monitoring), εργαλείων σύλληψης πακέτων, καθώς και με τη χρήση NetFlow (NetFlow-based monitoring) μεθόδων παρακολούθησης. Συνήθως τα εργαλεία που χρησιμοποιούνται χρησιμοποιούν κάποιο interface ενός router διαμέσου του οποίου γίνεται σύλληψη τόσο εισερχόμενων όσο και εξερχόμενων πακέτων.

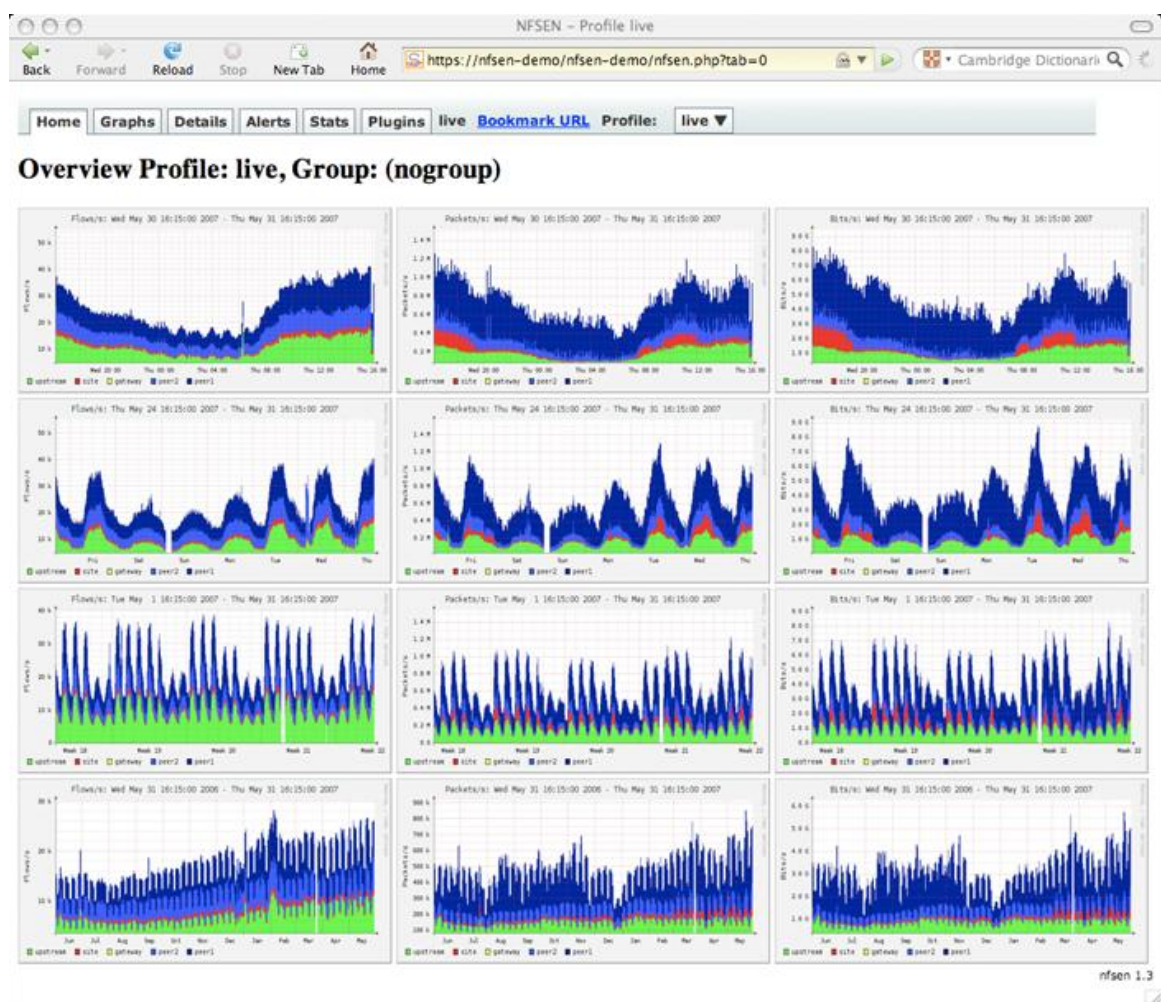
Έχουν αναπτυχθεί πάρα πολλά εργαλεία τα οποία έχουν τη δυνατότητα πραγματοποίησης μετρήσεων που σχετίζονται με το εύρος ζώνης συνδέσεων. Η πλειοψηφία αυτών δίνει τη δυνατότητα εμφάνισης των αποτελεσμάτων μέτρησης με τη μορφή σχεδιαγραμμάτων, σε πραγματικό χρόνο (real time monitoring) καθώς και εμφάνισής τους μέσα από ιστοσελίδες. Γνωστά εργαλεία είναι το εμπορικό λογισμικό PRTG καθώς και ελεύθερο λογισμικό nfdump.

Στο επόμενο σχήμα (σχήμα 1.13) εμφανίζονται αποτελέσματα του εργαλείου PRTG το οποίο είναι ένα εργαλείο παρακολούθησης και ανάλυσης της κίνησης δικτύου αποτυπώνοντας τα αποτελέσματα των μετρήσεων είτε με τη μορφή πινάκων είτε σχεδιαγραμμάτων. Συλλέγει δεδομένα διαφόρων τύπων υποστηρίζοντας πολλαπλά πρωτόκολλα για το σκοπό αυτό (όπως SNMP, Packet Sniffing, NetFlow και sFlow).



Σχήμα 1.13: Κονσόλα εμφάνισης μετρήσεων εύρους ζώνης με τη χρήση του εργαλείου PRTG. (εικόνα από το Web site http://www.paessler.com/bandwidth_monitor - Δεκέμβριος 2011 [23])

Το ελεύθερο λογισμικό nfdump (το οποίο και αποτέλεσε εργαλείο για την πραγματοποίηση των μετρήσεων της συγκεκριμένης εργασίας) δίνει τη δυνατότητα συλλογής και ανάλυσης NetFlow δεδομένων (Τεχνολογία παρακολούθησης κίνησης από την Cisco). Χρησιμοποιεί ως front end το component nfsen για τη γραφική αναπαράσταση των αναλύσεων που γίνονται στα συλλεχθέντα δεδομένα. Αναλυτικότερη περιγραφή του εργαλείου nfdump θα γίνει στην ενότητα 3 όπου και περιγράφεται η πειραματική υλοποίηση που πραγματοποιήθηκε στα πλαίσια αυτής της εργασίας. Στο σχήμα 1.14 εμφανίζονται γραφικές παραστάσεις από το εργαλείο nfsen από το Web site <http://nfsen.sourceforge.net/>.



Σχήμα 1.14: Screen shots του εργαλείου nfsen. Εικόνα από το Web site <http://nfsen.sourceforge.net/> - Δεκέμβριος 2011 [24]

1.6 Μέθοδοι Γραφικής Αναπαράστασης (Οπτικοποίησης) Καταγεγραμμένων Δεδομένων Κίνησης Δικτύου Υπολογιστών – Πλεονεκτήματα που Προσφέρουν

Όπως έχει γίνει αντιληπτό από τις αναφορές των προηγούμενων παραγράφων τα εργαλεία συλλογής καταστάσεων και πληροφοριών από ένα δίκτυο υπολογιστών είναι πάρα πολλά. Όλα στοχεύουν στο να βοηθήσουν στον εντοπισμό διαφόρων απειλών, την ανίχνευση προβλημάτων, τον εντοπισμό αιτιών και γενικότερα στην ορθή και ολοκληρωμένη διαχείριση ενός δικτύου υπολογιστών. Οι διαχειριστές έχουν να επεξεργαστούν τεράστιο αριθμό πληροφοριών που συλλέγονται καθημερινά με τη χρήση των εργαλείων αυτών. Έχουν να αναλύσουν πληροφορίες από log files, πληροφορίες που συλλέγονται από συστήματα που βρίσκονται υπό παρακολούθηση (SNMP), αποκρίσεις και alerts από συστήματα ανίχνευσης εισβολών, συστήματα παρακολούθησης διαθεσιμότητας συνδέσεων και ένα σωρό άλλες πηγές. Μεταξύ όλων αυτών, για την σωστή διαχείριση του δικτύου (όπως έχει αναφερθεί και σε προηγούμενη παράγραφο) και για να έχει ο διαχειριστής την συνολική εικόνα του δικτύου που διαχειρίζεται θα πρέπει να μπορεί να μετρήσει και να αναλύσει ώστε να κατανοήσει την κίνηση που διέρχεται από αυτό. Χρησιμοποιώντας τα εργαλεία που υπάρχουν για αυτό το σκοπό θα πρέπει ο διαχειριστής να συλλάβει πακέτα και να τα αναλύσει χρησιμοποιώντας τις περισσότερες φορές “text based” εργαλεία. Χρησιμοποιώντας όμως εργαλεία με τέτοιου είδους διεπαφές (interfaces), η χρηστικότητα των συλλεγόμενων δεδομένων και η αξιοποίηση των πληροφοριών που αυτά μπορούν να παρέχουν δεν μπορεί να κριθεί ικανοποιητική-αποδοτική. Η οπτικοποίηση των δεδομένων και αναπαράστασή τους από εικόνες είναι ο καταλληλότερος και αποδοτικότερος τρόπος κατανόησης των δεδομένων [25]. Λόγοι για τους οποίους τα εργαλεία που οπτικοποιούν τα δεδομένα (διαγραμματικός τρόπος εμφάνισης αποτελεσμάτων) είναι ποιο αποδοτικά σε σχέση με τη χρήση “text based” εργαλείων είναι οι ακόλουθοι [25]:

- δίνουν περισσότερες αντλήσιμες πληροφορίες χρησιμοποιώντας τον ίδιο χώρο εμφάνισης [26],
- χρησιμοποιούν ποιο εύκολα την αντιληπτική ικανότητα του ανθρώπου [27],
- η ιστορία έχει δείξει ότι η οπτικοποίηση (visualization) μπορεί να βοηθήσει σε μεγάλο βαθμό στη λήψη αποφάσεων [25].

Όπως λέει και το ρητό «μία εικόνα χίλιες λέξεις», η οπτικοποίηση των αποτελεσμάτων οποιουδήποτε εργαλείου καταγραφής της κίνησης δεδομένων σε ένα δίκτυο υπολογιστών θα δώσει την δυνατότητα στον διαχειριστή να αντλήσει ποιο εύκολα και ποιο γρήγορα τις πληροφορίες που θέλει. Φυσικά αυτό πάντοτε εξαρτάται από το πόσο ικανοποιητική είναι η απεικόνιση των πληροφοριών στο οπτικοποιημένο αποτέλεσμα. Την αποδοτικότητα αυτή της απεικόνισης επηρεάζουν πολλοί παράγοντες όπως:

- Το πλήθος των πληροφοριών που συμπεριλαμβάνονται και απεικονίζονται στο σχεδιάγραμμα. Όσο πιο πολλές πληροφορίες απεικονίζονται σε μία εικόνα (σχεδιάγραμμα, γράφημα) τόσο μεγαλύτερη προσπάθεια απαιτείται από τον παρατηρητή για την άντληση των απαιτούμενων πληροφοριών.
- Η καταλληλότητα του επιλεγέντος τρόπου απεικόνισης και το κατά πόσον αυτός ταιριάζει με την υπό απεικόνιση κατάσταση. Ένα γράφημα πίτας (pie chart) ταιριάζει καλύτερα και απεικονίζει πιο ικανοποιητικά την ποσοστιαία κατανομή κάποιου γεγονότος από ό,τι ένα γράφημα στηλών (bar chart or column chart) ή γραμμών (line chart).
- Το είδος της πληροφορίας που απαιτείται να αντληθεί, πχ όταν σε ένα γράφημα πίτας απεικονίζονται πολλές πληροφορίες με παρόμοια μεγέθη είναι λίγο δύσκολο να πραγματοποιηθεί σύγκριση μεταξύ των τιμών των πληροφοριών, οπότε σε αυτή την περίπτωση το γράφημα πίτας ίσως να μην είναι το καταλληλότερο για την άντληση εύκολα της πληροφορίας αυτής.

Οι κλασικές μέθοδοι απεικόνισης της κίνησης ενός δικτύου υπολογιστών χρησιμοποιούν συνήθως γραφήματα γραμμής και πίτας. Πάντοτε βέβαια η επιλογή του τρόπου απεικόνισης γίνεται με στόχο την ευκολότερη και αποδοτικότερη απεικόνιση των παραγόμενων πληροφοριών. Για την αναπαράσταση της κατανάλωσης του εύρους ζώνης μίας σύνδεσης συνηθέστερος τρόπος είναι η χρησιμοποίηση γραφημάτων γραμμών όπως αυτές που παράγουν τα εργαλεία nfsen (σχήμα 1.14) και PRTG (σχήμα 1.13) που αναφέρθηκαν παραπάνω καθώς και πάρα πολλά άλλα εργαλεία εμπορικά και μη.

Για την αναπαράσταση γραφικά των πρωτοκόλλων που χρησιμοποιούνται από τους χρήστες, των προορισμών που χρησιμοποιούν, των υπηρεσιών που παρέχονται χρησιμοποιούνται συνήθως

γραφήματα στηλών (column charts) και γραφήματα μπαρών (bar charts). Αν φυσικά χρειάζεται να δοθεί έμφαση στην αναπαράσταση ποσοστιαίας κατανομής ή να πραγματοποιηθεί σύγκριση μεταξύ των πληροφοριών που αναπαριστώνται, χρησιμοποιούνται και τα γραφήματα πίτας.



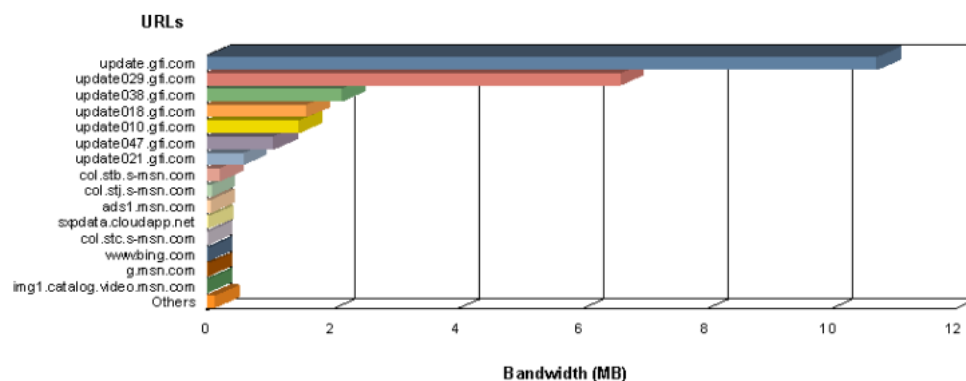
Report Title: Bandwidth Usage by Websites

Description: 'Bandwidth usage by Websites' report shows the amount of bandwidth consumed on specific websites sorted by bandwidth consumption.

Generated on: 2/23/2011

For period: 1/24/2011 - 2/23/2011

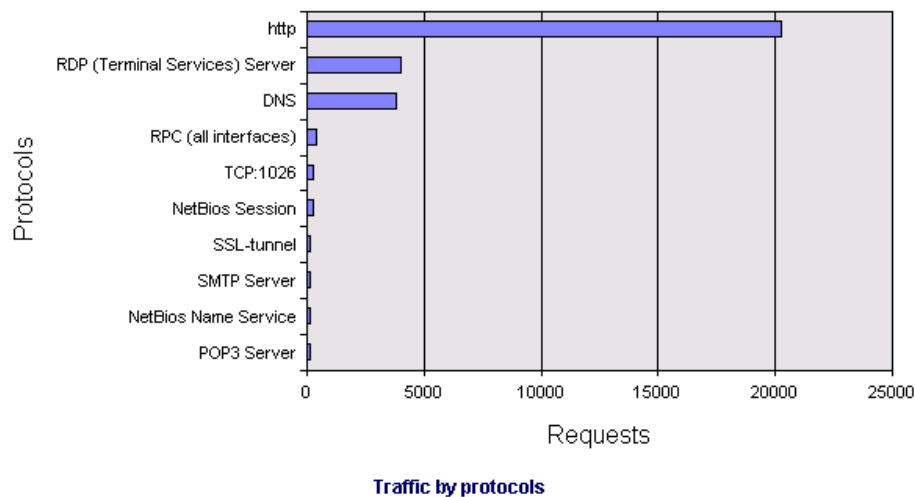
Bandwidth Usage by Websites



© 2011. All rights reserved. GFI Software Ltd

Page 1 of 4

Σχήμα 1.15: Χρήση εύρος ζώνης από web sites που επισκέπτονται οι χρήστες ενός δικτύου. (σχήμα από το Web site <http://www.gfi.com/internet-monitoring-software#screenshots> [28])

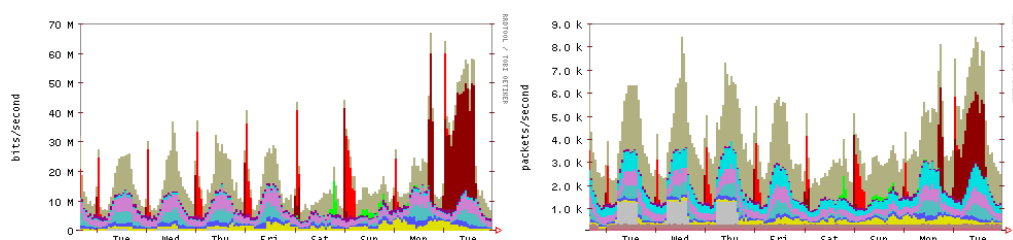


Σχήμα 1.16: Γράφημα στηλών με τα πρωτόκολλα που χρησιμοποιούνται από τους χρήστες

Μέχρι σήμερα έχουν πραγματοποιηθεί πολλές εργασίες με στόχο τη δημιουργία εργαλείων και ανάπτυξη τεχνικών οπτικοποίησης που θα βοηθήσουν στην αποδοτικότερη παρακολούθηση των δικτύων υπολογιστών. Πέραν των κλασικών μεθόδων οπτικοποίησης των αποτελεσμάτων παρακολούθησης, έχουν αναπτυχθεί και πιο προχωρημένες τεχνικές αναπαράστασης των δεδομένων. Στις επόμενες γραμμές της παραγράφου αυτής περιγράφονται δουλειές που είχαν αυτό ως στόχο.

Η παρακολούθηση της κίνησης των δικτύων υπολογιστών τις περισσότερες φορές στηρίζεται στην παρακολούθηση συγκεκριμένων χαρακτηριστικών όπως το πρωτόκολλο επικοινωνίας που χρησιμοποιείται, οι IP διευθύνσεις (αφετηρίας και προορισμού) καθώς και οι θύρες (αφετηρίας και προορισμού) που εμφανίζονται στο πακέτο. Η παρακολούθηση με αυτό τον τρόπο παρουσιάζει κάποια προβλήματα-δυσκολίες. Οι Cristian Estan, Stefan Savage και George Varghese στην δουλειά τους με τίτλο “Automatically inferring Patterns of resource Consumption in Network Traffic” [29] αναφέρονται στο πρόβλημα που απορρέει από το γεγονός ότι όσο περισσότερες λεπτομέρειες καταγράφονται τόσο αναλυτικότερη είναι η κατανόηση των δεδομένων αλλά και τόσο δυσκολότερη η εύρεση και ο εντοπισμός των σημαντικών πληροφοριών. Περιγράφουν στην εργασία τους ότι η σύνοψης της κίνησης ενός δικτύου σε μία διάσταση τη φορά (πχ ομαδοποίηση ροών με βάση μόνο ένα χαρακτηριστικό τους πχ την source IP διεύθυνσή των πακέτων) μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Από την άλλη η υπερβολική λεπτομέρεια καθιστά αδύνατη την ουσιαστική παρακολούθηση και την δυνατότητα κατανόησης της πληροφορίας. Για

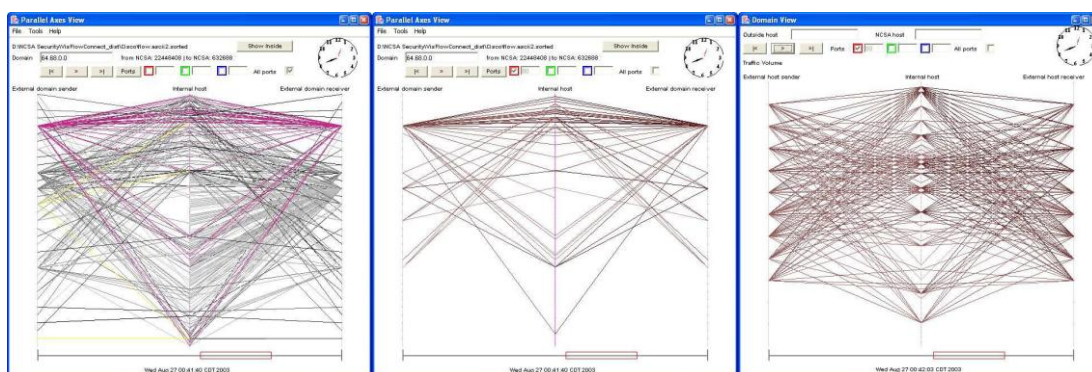
τους παραπάνω λόγους η δουλειά που εξηγείται στο άρθρο αυτό, αναφέρεται στην προσπάθεια ομαδοποίησης της κίνησης σε πολυδιάστατες ομάδες (clusters) και όχι σε ανεξάρτητες ροές. Με το σύστημα που δημιουργήθηκε από την συγκεκριμένη δουλειά (autofocus tool), δημιουργούνται αναφορές (traffic reports) οι οποίες περιέχουν ομάδες τέτοιων δεδομένων και μπορούν να χρησιμοποιηθούν για την εξαγωγή χρήσιμων συμπερασμάτων για την κίνηση του δικτύου που παρακολουθείται (πχ αναγνώριση περιπτώσεων υπερβολικής κατανάλωσης πόρων). Στο σχήμα 1.17 εμφανίζονται γραφήματα αναπαράστασης των καταγραφόμενων δεδομένων (κίνηση δικτύων υπολογιστών) από την αναφερθείσα εργασία αυτή τα οποία χρησιμοποιούν time series plots για την εμφάνιση των δεδομένων.



Σχήμα 1.17: Γραφήματα καταγραφής κίνησης δικτύων. (εικόνα από το pdf με τίτλο “Automatically inferring Patterns of resource Consumption in Network Traffic” [29])

Λόγω της αύξησης της πολυπλοκότητας των σημερινών δικτύων υπολογιστών γίνεται όλο και πιο δύσκολο για έναν διαχειριστή δικτύων να παρακολουθεί και να γνωρίζει όλες τις δραστηριότητες που πραγματοποιούνται στο δίκτυο που παρακολουθεί. Οι Xiaoxin Yin, William Yurcik, και Michael Treaster έχοντας ως στόχο να βοηθήσουν στην παρακολούθηση αυτή δημιούργησαν ένα εργαλείο (το VIsFlowConnect) [30]. Το εργαλείο αυτό χρησιμοποιεί για την αναπαράσταση ροών δεδομένων (Net Flow data) γραφήματα παράλληλων συντεταγμένων (parallel coordinates) ξεφεύγοντας έτσι από τους κλασικούς τρόπους οπτικοποίησης δεδομένων. Τα γραφήματα αυτά χρησιμοποιούνται συνήθως για την αναπαράσταση σχέσεων πολλά προς πολλά και στην συγκεκριμένη εργασία χρησιμοποιούνται για να παρουσιάσουν τις συνδέσεις μεταξύ hosts του εσωτερικού και των εξωτερικών δικτύων. Υπάρχουν τρεις κάθετοι άξονες (μία για τους εξωτερικούς αποστολείς, μία για τους εξωτερικούς παραλήπτες και μία για τους εσωτερικούς hosts) ενώ οι γραμμές σύνδεσης μεταξύ αυτών δηλώνουν τις συνδέσεις που έχουν πραγματοποιηθεί (στο χρονικό διάστημα καταγραφής των δεδομένων που αναπαριστώνται γραφικά).

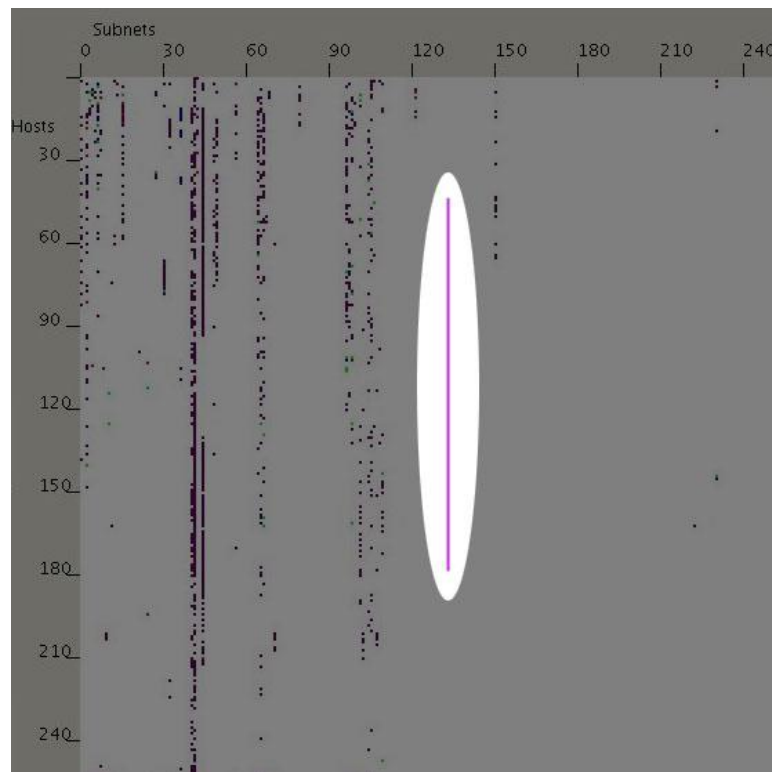
Στόχος των ερευνητών που ενεπλάκησαν στο συγκεκριμένο έργο ήταν η δημιουργία ενός εργαλείου που θα βοηθήσει τους διαχειριστές δικτύων στην ευκολότερη αναγνώριση ανωμαλιών στην κίνηση που δημιουργείται μεταξύ υπολογιστών του εσωτερικού και των εξωτερικών δικτύων. Το εργαλείο που δημιουργήθηκε στηρίζεται περισσότερο στην τεχνική «anomaly detection» την οποία χρησιμοποιεί μεγάλος αριθμός IDS σήμερα. Σημαντικό συμπέρασμα από τη συγκεκριμένη δουλειά είναι το γεγονός ότι με βάση τα γραφήματα παράλληλων συντεταγμένων που δημιουργούνται σε περιπτώσεις επιβεβαιωμένων επιθέσεων (πχ virus outbreak, denial of service attack κλπ), παράγονται πρότυπα (patterns) τα οποία μπορούν να χρησιμοποιηθούν για την αναγνώριση εύκολα επιθέσεων. Στο σχήμα 1.18 εμφανίζεται οθόνες που προκύπτουν από το εργαλείο VisFlowConnect που δημιουργήθηκε στην συγκεκριμένη εργασία και αναπαριστούν, με τη βοήθεια γραφημάτων παράλληλων συντεταγμένων, denial of service attacks.



Σχήμα 1.18: Γραφήματα παράλληλων συντεταγμένων που αναπαριστούν denial of service attacks. (εικόνα από το pdf με τίτλο “VisFlowConnect: NetFlow Visualizations of Link Relationships for Security Situational Awareness” [30])

Συνέχεια αυτής της εργασίας με στόχο να δοθεί στον διαχειριστή ενός δικτύου υπολογιστών “the big picture” του δικτύου που διαχειρίζεται, αποτελεί η δημιουργία του εργαλείου NVisionIP από τους Kiran Lakkaraju, William Yurcik και Adam J. Lee [31]. Σε αυτή την εργασία παρουσιάζεται ένας άλλος τρόπος αναπαράστασης της κίνησης σε ένα δίκτυο υπολογιστών. Χρησιμοποιείται ένα “subnet × host” σύστημα συντεταγμένων (coordinate system) στο οποίο ο ένας άξονας (κάθετος) φέρει το subnet id του υποδικτύου ενώ ο άλλος (οριζόντιος) άξονας φέρει το host id της IP διεύθυνσης του host. Κάθε host δηλώνεται με μία κουκίδα στο γράφημα διαφορετικού χρώματος με βάση τον τύπο της κίνησης που φτάνει ή φεύγει από αυτόν. Μάλιστα το εργαλείο που δημιουργήθηκε δίνει τη δυνατότητα για animation και αλλαγή των χρωμάτων με βάση την κίνηση στο χρονικό διάστημα

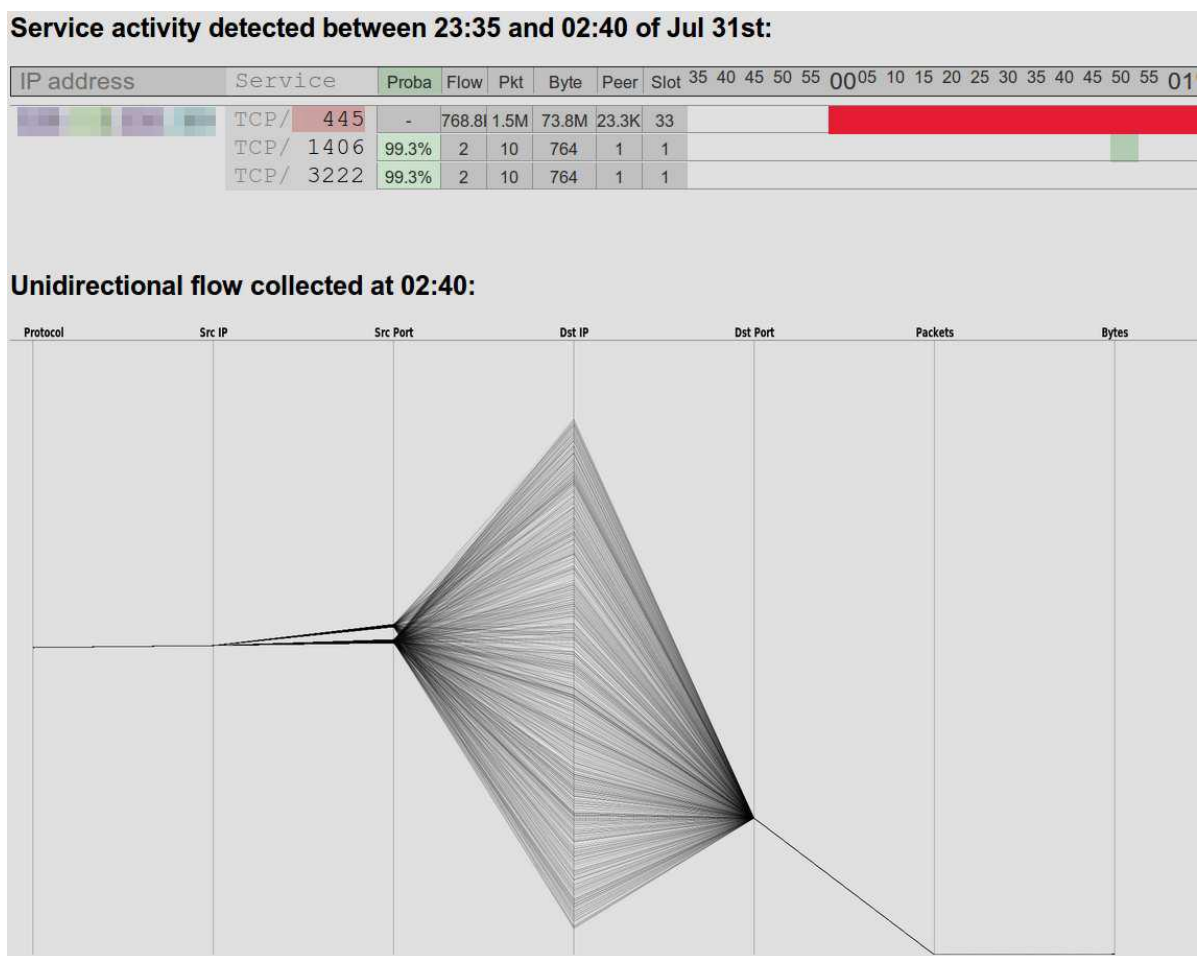
της παρακολούθησης. Στο σχήμα 1.19 εμφανίζεται οθόνη αποτελέσματος του εργαλείου NVisionIP η οποία παρουσιάζει ανίχνευση δραστηριότητας IP scan.



Σχήμα 1.19: Ανίχνευση IP scan activity με τη χρήση subnet x host coordinate system. (εικόνα από το pdf με τίτλο: “NVisionIP: NetFlow Visualizations of System State for Security Situational Awareness” [31])

Η «γνώση» του δικτύου υπολογιστών (network awareness) και η κατανόηση των λειτουργιών και των υπηρεσιών που αυτό παρέχει, όπως αναφέρθηκε και σε προηγούμενες παραγράφους, αποτελεί σημαντικό παράγοντα που βοηθά στην καλύτερη διαχείρισή του. Η παρακολούθηση της κίνησης που διέρχεται από αυτό είναι μπορεί να βοηθήσει στην λήψη αποφάσεων που σχετίζονται με την συντήρηση και την ασφάλειά του [32]. Η ομαδοποίηση των πακέτων σε ροές δεδομένων (data flows) και η ανάλυση αυτών αντί του κάθε πακέτου που καταγράφεται ξεχωριστά οδηγούν από τη λεπτομέρεια στην ευελιξία και την επεκτασιμότητα. Το άρθρο [32] από τους Robin Berthier and Michel Cukier χρησιμοποιεί ροές δεδομένων (netflow data) για την ανάλυση και εντοπισμό επιθέσεων. Το άρθρο αυτό αναφέρεται σε ένα add-on για το λογισμικό ανοιχτού κώδικα «nfsen» το οποίο δέχεται ως είσοδο ροές δεδομένων από το εργαλείο nfdump και τις αναπαριστά με γραφικό τρόπο. Ένας από τους τρόπους αναπαράστασης που παρουσιάζει μεγάλο ενδιαφέρον είναι αυτός

των παράλληλων συντεταγμένων (parallel coordinates). Τα παραγόμενα γραφήματα όπως αυτό που απεικονίζεται στην εικόνα 11 του άρθρου (παρουσιάζεται στο σχήμα 1.20) μπορούν να χρησιμοποιηθούν για την εύκολη ανίχνευση επιθέσεων. Στο σχήμα 11 του άρθρου παρουσιάζεται ένα γράφημα με παράλληλες συντεταγμένες το οποίο αναπαριστά ένα “scanning attack”, μια προσπάθεια ανίχνευσης του εάν η θύρα TCP/445 (netbios service) είναι διαθέσιμη σε ένα σύνολο υπολογιστών (IP διευθύνσεις προορισμού). Με βάση αυτό το γράφημα παρατηρείται ότι σε περίπτωση επιθέσεων η αναπαράσταση των ροών δεδομένων κατά τη διάρκεια αυτών, δημιουργεί πρότυπα-signatures. Αυτά μπορούν να χρησιμοποιηθούν για την αναγνώριση εύκολα επιθέσεων. Πχ, σε περίπτωση που κατά την οπτική αναπαράσταση ροών δεδομένων παρατηρηθεί ότι το γράφημα παράλληλων συντεταγμένων που προκύπτει μοιάζει με αυτό του σχήματος 11 μπορεί πλέον με σιγουριά ο παρατηρητής να συμπεράνει την ύπαρξη scanning attack. Παρόμοιο συμπέρασμα έχει παραχθεί και από την δουλειά που αναφέρεται στο άρθρο VisFlowConnect: NetFlow Visualizations of Link Relationships for Security Situational Awareness από τους Xiaoxin Yin, William Yurcik, και Michael Treaster [30]. Γενικότερα τονίζεται η χρησιμότητα των γραφημάτων παράλληλων συντεταγμένων στην απεικόνιση της κίνησης που διέρχεται από ένα δίκτυο υπολογιστών.



Σχήμα 1.20: Γράφημα με παράλληλες συντεταγμένες το οποίο αναπαριστά ένα “scanning attack”. Εικόνα από το άρθρο με τίτλο: “Nfsight: NetFlow-based Network Awareness Tool” [32]

Αντίστοιχο συμπέρασμα, ότι δηλαδή η χρήση των γραφημάτων παράλληλων συντεταγμένων βοηθά στον εντοπισμό επιθέσεων προκύπτει και από την δουλειά που περιγράφεται στο άρθρο με τίτλο “Applied parallel coordinates for logs and network traffic attack analysis” από τους Sebastien Tricaud και Philippe Saade [12]. Το paper αυτό κάνει μία εισαγωγή στο μαθηματικό υπόβαθρο το οποίο χρησιμοποιείται από την τεχνική απεικόνισης παράλληλων συντεταγμένων και στη συνέχεια περιγράφει το εργαλείο Picviz το οποίο χρησιμοποιεί αυτή την τεχνική για την απεικόνιση ροών δεδομένων. Στόχος της εργασίας είναι να τονίσει τη σημασία που έχει η οπτική αναπαράσταση δεδομένων στην ανίχνευση επιθέσεων. Πιο συγκεκριμένα, προσπαθεί να ενισχύσει την άποψη ότι δεδομένα κίνησης δικτύου αναπαριστώμενα με γράφημα παράλληλων συντεταγμένων, μπορούν να βοηθήσουν τους διαχειριστές δικτύων να ανιχνεύσουν «ανώμαλα» γεγονότα ανάμεσα σε τεράστιες ποσότητες πληροφοριών.

Αντίστοιχα συμπεράσματα προκύπτουν από το άρθρο με τίτλο “Fast detection and visualization of network attacks on parallel coordinates” από τους Hyunsang Choi, Heejo Lee, και Hyogon Kim [11]. Το άρθρο αυτό όπως και τα δύο προηγούμενα ενισχύει με παραδείγματα την άποψη ότι μία επίθεση μπορεί να αναγνωριστεί παρατηρώντας τη γραφική αναπαράσταση της κίνησης με τη χρήση γραφημάτων παράλληλων συντεταγμένων. Παρουσιάζει αυτό που αποκαλείται από τους συγγραφείς του ως “parallel coordinate attack visualization” και το οποίο χρησιμοποιεί parallel coordinates για την ανίχνευση επιθέσεων με βάση “visual” signatures. Παρουσιάζει και αναλύει «graphical signatures» εννέα γνωστών επιθέσεων και εξηγεί πως θα μπορούσαν να χρησιμοποιηθούν για την ανίχνευση επιθέσεων.

1.7 Συμπεράσματα

Όπως γίνεται αντιληπτό από τις προηγούμενες παραγράφους, έχουν αναπτυχθεί μέχρι σήμερα πάρα πολλές μέθοδοι-τεχνικές και εργαλεία με στόχο την πληρέστερη καταγραφή και παρακολούθηση των δικτύων υπολογιστών. Παρακολουθώντας γενικά ένα δίκτυο υπολογιστών και ειδικότερα την κίνηση που διέρχεται από αυτό μπορούν να αναγνωριστούν επιθέσεις, να εντοπιστούν αδυναμίες, να βρεθούν αιτίες προβλημάτων αναδεικνύοντας έτσι την αξία που έχει στην ομαλή λειτουργία καθώς και στην αρτιότερη διαχείριση ενός δικτύου υπολογιστών η παρακολούθηση αυτή. Πρόσθετη αξία στην προσφορά της παρακολούθησης έρχονται να προσδώσουν οι τεχνικές γραφικής αναπαράστασης δεδομένων οι οποίες και αυτές είναι πολλές σε αριθμό.

Στις ενότητες που ακολουθούν θα γίνει αναφορά στις τεχνικές εξόρυξης δεδομένων και πώς αυτές έχουν κερδίσει έδαφος τα τελευταία χρόνια, ώστε να συγκαταλέγονται στις μεθόδους παρακολούθησης δικτύου και κυρίως στην ανίχνευση επιθέσεων-ανωμαλιών που μπορεί να παρατηρούνται στα δίκτυα υπολογιστών. Θα περιγραφεί πειραματική εγκατάσταση που υλοποιήθηκε σε πραγματικό δίκτυο υπολογιστών (πραγματική εταιρεία) και θα αναλυθούν παραγόμενα δεδομένα ως αποτέλεσμα εφαρμογής αλγορίθμων εξόρυξης συχνών στοιχείων σε δεδομένα που προέρχονται από κίνηση δικτύου υπολογιστών. Θα πραγματοποιηθεί μία αναφορά στις τεχνικές που υπάρχουν διαθέσιμες για την οπτική αναπαράσταση (visualization techniques)

συχνών στοιχείων και θα πραγματοποιηθεί σύγκριση αυτών με γνώμονα την εφαρμοσιμότητά τους στην αναπαράσταση των δεδομένων που προέρχονται από κίνηση σε ένα δίκτυο υπολογιστών.

Κεφάλαιο 2

Χρήση Τεχνικών Εξόρυξης Δεδομένων στη Σύνοψη Δικτυακής κίνησης.

Η ενότητα αυτή ξεκινά με μία περιγραφή των τεχνικών data mining. Αναλύει το τι είναι data mining και πιο συγκεκριμένα frequent itemset mining. Περιγράφει το πού εφαρμόζονται ως επί το πλείστον οι τεχνικές data mining και αναφέρει αλγορίθμους εξαγωγής frequent itemsets και association rules (οι λεγόμενοι Mining Algorithms). Συνεχίζει δίνοντας μία περιγραφή των τύπων frequent itemsets όπως είναι τα maximal και closed frequent itemsets. Τέλος, περιγράφει γιατί οι αλγόριθμοι εξόρυξης δεδομένων μπορούν να αποτελέσουν χρήσιμο εργαλείο στην παρακολούθηση ενός δικτύου υπολογιστών. Παράλληλα γίνεται αναφορά σε προηγούμενες εργασίες προς αυτό το σκοπό, τη χρήση δηλαδή τεχνικών εξόρυξης στην βελτίωση της ασφάλειας ενός δικτύου και γενικότερα στη

βελτίωση της διαχείρισης αυτών (π.χ. χρήση frequent itemsets στην εύρεση-αναγνώριση ανωμαλιών).

Κύριος στόχος αυτής της ενότητας είναι να τονίσει τη σημαντικότητα και τη χρηστικότητα των τεχνικών εξόρυξης στην σύνοψη της κίνησης ενός δικτύου.

2.1 Εισαγωγή

Η εισαγωγή των υπολογιστών στη ζωή μας και η χρήση αυτών για την υλοποίηση ή υποστήριξη των περισσότερων από τις καθημερινές εργασίες του ανθρώπου είναι γεγονός εδώ και κάποιες δεκαετίες. Η εξέλιξη αυτή δε βελτιώνει μόνο τον τομέα της επιχειρηματικότητας, αλλά πλέον το σύνολο των καθημερινών εργασιών του ανθρώπου, ανεξαρτήτου οικονομικής και κοινωνικής τάξης, υλοποιούνται ή υποβοηθούνται από υπολογιστές. Ένας τεράστιος όγκος ψηφιακών πληροφοριών καταγράφονται, διακινούνται και ανταλλάσσονται συνεχώς με στόχο τη βελτίωση παρεχόμενων υπηρεσιών. Στοιχεία, όπως αγοραπωλησίες με τη χρήση πιστωτικών καρτών, αριθμός και ποσά αναλήψεων μέσω ATMs και ηλεκτρονικών ιστοσελίδων, αριθμός αγορών από συγκεκριμένους καταναλωτές σε συγκεκριμένες χρονικές περιόδους, αριθμός αυτοκινήτων που διέρχονται από παρακολουθούμενες οδούς και διασταυρώσεις, αριθμός και είδη προϊόντων που πωλούνται, αριθμός επισκέψεων (hits) σε ιστοχώρους, είδος μουσικής που προτιμάται από τους αγοραστές ανά ηλικία, αριθμός θεατών κινηματογραφικών ταινιών σε κινηματογραφικές αίθουσες, αριθμός βιβλίων που ενοικιάζονται από τοπικές βιβλιοθήκες, αριθμός πακέτων που διέρχονται από συγκεκριμένη σύνδεση, είναι λίγες από τις καθημερινές πληροφορίες οι οποίες καταγράφονται ψηφιακά με στόχο τη χρήση τους για τη βελτίωση της επιχειρηματικότητας (μεγαλύτερες πωλήσεις και κέρδη για τις εταιρείες) και την παροχή καλύτερων υπηρεσιών. Τεράστιες βάσεις δεδομένων δημιουργούνται και ενημερώνονται οι οποίες αποθηκεύουν εξαιρετικά μεγάλο όγκο τέτοιων πληροφοριών.

Για την επεξεργασία αυτών των πληροφοριών καθώς και την ευκολότερη οργάνωση των δεδομένων χρησιμοποιείται η τεχνολογία των βάσεων δεδομένων. Γλώσσες προγραμματισμού, τεχνικές και κατάλληλο λογισμικό δίνουν τη δυνατότητα να μπορούν να αντληθούν πληροφορίες από αυτές τις βάσεις δεδομένων και να εξαχθούν συμπεράσματα, που χωρίς τη χρήση των υπολογιστών και των δικτύων αυτών δεν θα ήταν εύκολη υπόθεση. Το ενδιαφέρον για τη συλλογή

και την άντληση πληροφοριών, συμπερασμάτων, τάσεων και ακολουθιών που δεν είναι ξεκάθαρες και αντλήσιμες απλά και μόνο με την απλή ανάγνωση-σάρωση των πληροφοριών, έχει αποκτήσει πολύ μεγάλο ενδιαφέρον και έχει γίνει πολύτιμο εργαλείο των περισσότερων επιχειρήσεων. Μία επιχείρηση θα χρησιμοποιήσει τα εργαλεία αυτά για να βοηθηθεί στη λήψη σημαντικών στρατηγικών αποφάσεων (decision making) με απώτερο στόχο την αύξηση του μεριδίου πωλήσεων της στην αγορά.

Η διαδικασία (επιστήμη) άντλησης πληροφοριών και εξαγωγής συμπερασμάτων μέσα από ένα συνήθως μεγάλο σύνολο δεδομένων (βάση δεδομένων) καλείται “Data Mining” ή αλλιώς “Knowledge Discovery in Data” (KDD). “Data mining” είναι η διαδικασία στην «ανακάλυψη γνώσης» κατά την οποία βρίσκονται ενδιαφέροντα πρότυπα στα δεδομένα, κανόνες συσχέτισης μεταξύ αυτών, αλληλουχίες προτύπων, ομάδες (clusters), κατηγοριοποιήσεις κ.α. [33]. Αυτά θα χρησιμοποιηθούν στη συνέχεια στην παραγωγή «γνώσης», συμπερασμάτων δηλαδή και πληροφοριών, που δεν διαφαίνονται απευθείας από τα δεδομένα. Σημαντικό ρόλο σε αυτή τη διαδικασία (data mining) παίζουν τα λεγόμενα «συχνά στοιχεία» (“frequent itemsets”). Η αναγνώριση «συχνών στοιχείων» στα δεδομένα αποτελεί ένα από τα βασικότερα βήματα στην διαδικασία του “data mining”. «Συχνά στοιχεία» είναι το σύνολο των συνδυασμών στοιχείων-αντικειμένων μίας βάσης δεδομένων που ξεπερνούν έναν συγκεκριμένο αριθμό εμφανίσεων σε αυτήν (θα ακολουθήσει πιο ακριβής ορισμός στη συνέχεια του κεφαλαίου).



Η διαδικασία εύρεσης των «συχνών στοιχείων» μέσα από ένα σύνολο δεδομένων και η ανάγκη που οδηγεί στη διαδικασία αυτή πρωτοεμφανίστηκε το 1993 στο άρθρο «Mining Association Rules between Sets of Items in Large Databases» [34] από τους Rakesh Agrawal, Tomasz Imielinski και Arun Swami. Το συγκεκριμένο άρθρο μέχρι τον Μάρτιο του 2008, σύμφωνα με το Google Scholar, είχε πάνω από 6000 βιβλιογραφικές αναφορές, κάτι που το καθιστά ένα από τα πιο δημοφιλή papers στον τομέα του “data mining”. Φυσικά, εκατοντάδες άρθρα μετά από αυτό ασχολούνται με το πρόβλημα εύρεσης των «συχνών στοιχείων» και τις χρησιμότητας – χρηστικότητα αυτών,

παρουσιάζοντας μεγάλο αριθμό αλγορίθμων και τεχνικών για την εύρεση λύσεων στο πρόβλημα του “data-frequent mining”.

Αν και η πρώτη χρήση των αλγορίθμων-τεχνικών που σχεδιάστηκαν για την εύρεση «συχνών στοιχείων» έγινε σε δεδομένα βάσεων δεδομένων που σχετίζονται με αγορές προϊόντων από super markets, τα τελευταία χρόνια πλέον η χρήση των «συχνών στοιχείων» σχετίζεται με ερευνητικά projects και τομείς διαφόρων κατηγοριών. Επιγραμματικά, χωρίς να εμβαθύνουμε, αφού δεν αποτελεί στόχο αυτής της εργασίας, παρατίθενται παρακάτω μερικές από τις εφαρμογές χρήσεις «συχνών στοιχείων». Έχουμε χρήση των «συχνών στοιχείων» [35]:

- στον επιχειρηματικό κλάδο:
 - σε εφαρμογές για τη διαχείριση σχέσεων πελατών (Customer Relationship Management Applications),
 - σε εφαρμογές που χρησιμοποιούνται από τα τμήματα προσωπικού εταιρειών (Human Resources),
 - σε εφαρμογές για τη βελτίωση των λιανικών πωλήσεων, κ.α.
- στους τομείς τεχνολογίας και επιστήμης:
 - στην επιστήμη γονιδίων (genetics) και γενικότερα στην ιατρική,
 - στην εκπαίδευση,
 - στη μηχανική,
 - στη γεωργία,
 - στη μετεωρολογία,

Μεταξύ των σημαντικών χρήσεων-εφαρμογών του “data mining” και ειδικότερα του “frequent itemsets mining” (εξόρυξης συχνών στοιχείων) συγκαταλέγεται και η χρησιμοποίησή τους στην διαχείριση δικτύων υπολογιστών. Υπάρχουν μελέτες-εργασίες στην βιβλιογραφία οι οποίες ενισχύουν τη σημαντικότητα και τονίζουν την προσφορά των «συχνών στοιχείων» στην ενδυνάμωση της παρακολούθησης και διαχείρισης των δικτύων υπολογιστών. Υπάρχουν έρευνες οι οποίες δείχνουν τη χρησιμότητα αυτών και τη βοήθεια που μπορούν να προσφέρουν στους διαχειριστές-αναλυτές της ασφάλειας των δικτύων υπολογιστών διευκολύνοντάς τους στο έργο της παρακολούθησης και ειδικότερα της αντιμετώπισης προβλημάτων ασφάλειας αυτών. Ενδεικτικές μελέτες προς αυτό τον τομέα:

- A. Lakhina, M. Crovella, and C. Diot, (2005). Mining anomalies using traffic feature distributions, In Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications, New York, USA 2005 [36].
- Brauckhoff, D., Dimitropoulos, X., Wagner, A. and Salamatian, K. (2009). Anomaly extraction in backbone networks using association rules. In Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference (Chicago, Illinois, USA, November 04 - 06, 2009). IMC '09. ACM, New York, NY, 28-34 [37],
- Paredes-Oliva, I., Dimitropoulos, X., Molina, M., Barlet-Ros, P. and Brauckhoff, D. (2010). Automating root-cause analysis of network anomalies using frequent itemset mining. In Proceedings of the ACM SIGCOMM 2010 Conference (New Delhi, India, August 30 - September 03, 2010). SIGCOMM '10. ACM, New York, NY, 467-468 [38],
- Kenichi Yoshida, Yusuke Shomura, and Yoshinori Watanabe (2007). Visualizing Network Status. Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007 [39],

και άλλες (περισσότερες θα αναπτυχθούν σε επόμενη παράγραφο αυτού του κεφαλαίου).

Στη συνέχεια στο κεφάλαιο της παρούσας μεταπτυχιακής διατριβής θα πραγματοποιηθεί περιγραφή της διαδικασίας εξόρυξης δεδομένων και των αλγορίθμων που μπορούν να χρησιμοποιηθούν για την παραγωγή «κανόνων συσχέτισης» και «συχνών στοιχείων». Συνεχίζοντας, θα γίνει αναφορά στις δουλειές-έρευνες που εφαρμόζουν αλγορίθμους παραγωγής «συχνών στοιχείων» για τον εντοπισμό αιτιών προβλημάτων και κινδύνων που εμφανίζονται σε δίκτυα υπολογιστών. Απώτερος στόχος να τονιστεί η χρησιμότητα παραγωγής «συχνών στοιχείων» στην παρακολούθηση και σύνοψη της κίνησης ενός δικτύου υπολογιστών.

2.2 Διαδικασία Εξόρυξης δεδομένων

Όπως αναφέρθηκε και στην προηγούμενη ενότητα, με τη χρήση τεχνικών εξόρυξης δεδομένων, επιτυγχάνεται η «ανακάλυψη γνώσης», η δημιουργία δηλαδή συμπερασμάτων και η παραγωγή νέων πληροφοριών οι οποίες δεν είναι ορατές χωρίς την σύγκριση και των συσχετισμό μεταξύ τους

των υπό εξέταση δεδομένων. Σύμφωνα με το βιβλίο “Fundamentals of Database Systems”, από τους R. Elmasri, S.B. Navathe [40] ο στόχος της διαδικασίας εξόρυξης δεδομένων και των πληροφοριών που μπορούν να αντληθούν με τη χρήση τέτοιων τεχνικών μπορεί μεταξύ άλλων να κατηγοριοποιηθεί ως εξής:

- Πρόβλεψη. Μπορεί τα δεδομένα που προκύπτουν να αποτελούν προβλέψεις. Μπορεί δηλαδή με τη χρήση τεχνικών εξόρυξης δεδομένων να προβλεφθούν γεγονότα τα οποία με βάση τα εμφανιζόμενα-υφιστάμενα δεδομένα έχουν μεγάλη πιθανότητα να εμφανιστούν-πραγματοποιηθούν (ως αποτέλεσμα εμφάνισης κάποιων καταχωρήσεων στα δεδομένα). Π.χ. με βάση το συνδυασμό προϊόντων που αγοράζονται μπορεί να προβλεφθεί η αγορά και κάποιου τρίτου προϊόντος. Αν ένας πελάτης αγοράσει το Α και το Β προϊόν είναι πολύ πιθανό να αγοράσει και το Γ. Αυτό μπορεί να προκύψει από το γεγονός ότι ο αριθμός των συναλλαγών που έχουν πραγματοποιηθεί μέχρι τώρα και οι οποίες περιέχουν και τα τρία παραπάνω προϊόντα είναι πολύ μεγάλος. Κάτι αντίστοιχο θα μπορούσε να χρησιμοποιηθεί και στην παρακολούθηση της κίνησης ενός δικτύου υπολογιστών. Η εμφάνιση κάποιων χαρακτηριστικών σε πακέτα που φτάνουν σε κάποιον host μπορεί να σημαίνει ότι υπάρχει μεγάλη πιθανότητα να ακολουθήσουν πακέτα με κάποια συγκεκριμένα άλλα χαρακτηριστικά [40].
- Ταυτοποίηση-Ανίχνευση. Μπορεί η εμφάνιση κάποιων δεδομένων που παράγονται (στο σύνολο των δεδομένων που παρατηρούνται) με τη χρήση τεχνικών παραγωγής «συχνών στοιχείων» να είναι ικανή, ώστε να ταυτοποιηθεί-αναγνωριστεί η ύπαρξη μίας συγκεκριμένης κατάστασης π.χ. μίας απειλής. Π.χ. η εμφάνιση κάποιων ακολουθιών συμβόλων νουκλεοτιδίων σε ακολουθία του DNA να δηλώνει την ύπαρξη ενός συγκεκριμένου γονιδίου. Αντίστοιχα στην παρακολούθηση της κίνησης σε μία σύνδεση ενός δικτύου υπολογιστών, η ύπαρξη μεγάλου σε αριθμού και όγκο με συγκεκριμένα χαρακτηριστικά πακέτων (πέραν ενός ορίου – frequent itemsets) μπορεί να δηλώνει την ύπαρξη επίθεσης τύπου π.χ. buffer overflow attack [40].

Τα βήματα για την διαδικασία άντλησης «γνώσης» σύμφωνα με το “Fundamentals of Database Systems” [40] και το paper με τίτλο “Data Mining: An Overview” του Jeffrey W. Seifert [41] μπορεί να χωριστούν στις ακόλουθες φάσεις:

- Επιλογή των κατάλληλων δεδομένων. Το πρώτο βήμα στην διαδικασία του “data mining” είναι η αναγνώριση του προβλήματος που πρέπει να λυθεί και ο καθορισμός του τύπου των συμπερασμάτων που επιθυμείται να εξαχθούν. Για αυτό το λόγο θα πρέπει να καθοριστεί και το είδος των δεδομένων που θα χρησιμοποιηθούν και θεωρείται ότι ταιριάζουν με τον επιθυμητό στόχο [33]. Π.χ. εφόσον απαιτείται η εξαγωγή συμπερασμάτων που θα σχετίζονται με τη χρήση πρωτοκόλλων και αριθμών θυρών που χρησιμοποιούνται σε ένα δίκτυο υπολογιστών θα πρέπει τα δεδομένα που θα επιλεγούν να περιέχουν πληροφορίες από τα headers πακέτων που κυκλοφορούν στο υπό παρατήρηση δίκτυο υπολογιστών. Εάν ενδιαφέρει περισσότερο ο όγκος των πακέτων και οι χρονικές στιγμές διέλευσης των πακέτων θα πρέπει να χρησιμοποιηθούν αντίστοιχα κατάλληλα δεδομένα.
- Καθαρισμός των δεδομένων. Τα δεδομένα θα πρέπει αν απαλλαγούν από λάθη και «θόρυβο». Τιμές που λείπουν μπορεί να αντικατασταθούν χειροκίνητα από σταθερές πριν προχωρήσει η διαδικασία εξαγωγής συμπερασμάτων.
- Μετασχηματισμός-κωδικοποίηση (ίσως και μείωση) των δεδομένων. Οι εγγραφές (records) μπορεί να χρειάζεται να εμπλουτιστούν με πρόσθετες πληροφορίες. Π.χ.
 - αν σε μία εγγραφή των δεδομένων υπάρχουν πληροφορίες, όπως οι κωδικοί των προϊόντων που αγοράστηκαν σε μία δοσοληψία, θα μπορούσε αυτές οι εγγραφές να εμπλουτιστούν και με τις περιγραφές των προϊόντων αυτών ή την κατηγορία στην οποία ανήκουν.
 - Σε δεδομένα τα οποία περιέχουν πληροφορίες κεφαλίδων πακέτων κίνησης δικτύου υπολογιστών θα πρέπει να είναι δυνατός ο διαχωρισμός των αριθμών θυρών προορισμού από τους αριθμούς αφετηρίας ή των IP διευθύνσεων προορισμού από τις IP διευθύνσεις αφετηρίας. Θα μπορούσε λοιπόν τα δεδομένα να εμπλουτιστούν σε αυτή την περίπτωση με κάποιο χαρακτηριστικό, ώστε να είναι δυνατός αυτός ο διαχωρισμός.
- Ανακάλυψη κρυμμένων πληροφοριών. Σε αυτή τη φάση τα δεδομένα εισέρχονται στον αλγόριθμο “data mining” και παράγονται «συχνά στοιχεία», «κανόνες συσχέτισης» και χρήσιμα πρότυπα.
- Παρουσίαση των αποτελεσμάτων και χρησιμοποίηση της παραγόμενης γνώσης. Το τελευταίο βήμα στην διαδικασία παραγωγής της γνώσης είναι η παρουσίαση των

αποτελεσμάτων. Συνήθως, λόγω του γεγονότος ότι τα παραγόμενα συμπεράσματα είναι αποτέλεσμα αναζήτησης «κρυμμένων» πληροφοριών είναι σημαντική η σωστή-κατανοητή παρουσίαση τους. Σίγουρα μία γραφική αναπαράσταση των παραγόμενων αποτελεσμάτων μπορεί να χρησιμοποιήσει καλύτερα την αντιληπτική ικανότητα του ανθρώπου και να αποδώσει πιο εύκολα τα παραχθέντα συμπεράσματα.



Σχήμα 2.1: Βήματα για τη διαδικασία εξόρυξης δεδομένων

2.3 Ορισμοί που Σχετίζονται με τη Διαδικασία “Data Mining”

Πριν προχωρήσουμε στην παράθεση των αλγορίθμων που χρησιμοποιούνται στην διαδικασία εξόρυξης δεδομένων (data mining and frequent itemset mining) καταγράφουμε κάποιους ορισμούς οι οποίοι παίζουν σημαντικό ρόλο στη διαδικασία αυτή.

Στοιχεία – Items, Εγγραφές – Records or Transactions

Η διαδικασία εξόρυξης δεδομένων χρησιμοποιεί μία βάση με δεδομένα από την οποία θα πρέπει να αντληθούν «κρυμμένες», «άξιες λόγου» πληροφορίες και να εξαχθούν συμπεράσματα τα οποία «παράγουν γνώση». Τα ανεξάρτητα δεδομένα τα οποία συνθέτουν τις εγγραφές-Records (ή transactions) της βάσης δεδομένων αποτελούν τα στοιχεία (items, i) αυτής. Το υποσύνολο των στοιχείων ($i_1, i_2, i_3, \dots, i_n$) που εμφανίζονται μαζί σε ένα στιγμιότυπο της βάσης δεδομένων αποτελούν μία εγγραφή-record αυτής. Οι εγγραφές είναι στην ουσία μία πληροφορία που αποθηκεύεται στη βάση δεδομένων. Στην περίπτωση π.χ. του “market-basket” μία εγγραφή είναι μία δοσοληψία. Στην περίπτωση παρατήρησης της κίνησης σε ένα δίκτυο υπολογιστών και της καταγραφής των πακέτων που διέρχονται από ένα σημείο, μία εγγραφή της βάσης μπορεί να είναι το σύνολο των πληροφοριών της κεφαλίδας ενός πακέτου. Μία εγγραφή (Transaction, T) μπορεί να συμβολιστεί ως:

$$T \subseteq \text{Items}$$

Ο αριθμός των στοιχείων που θα συνθέσουν μία εγγραφή της βάσης δεδομένων δεν είναι κατά ανάγκη καθορισμένος. Θα μπορούσε δηλαδή σε μία εγγραφή ο αριθμός των στοιχείων που εμφανίζονται σε αυτή να είναι π.χ. 4 ενώ σε μία άλλη να είναι π.χ. 15 ή θα μπορούσε να είναι και ίσος με τον μέγιστο αριθμό στοιχείων. Το γεγονός αυτό εξαρτάται από την μορφή των δεδομένων που καταγράφονται στη βάση. Στην περίπτωση του παραδείγματος με την καταγραφή των πακέτων ο αριθμός των στοιχείων θα μπορούσε να είναι σταθερός και ίσος με των αριθμό των πληροφοριών που καταγράφονται στην κεφαλίδα των πακέτων εφόσον αυτές οι πληροφορίες εμφανίζονται πάντοτε σε κάθε πακέτο που καταγράφεται.

Σύνολο στοιχείων – Itemset

Itemset είναι μία ομάδα στοιχείων τα οποία εμφανίζονται μαζί σε μία ή περισσότερες εγγραφές. Ένα «σύνολο στοιχείων» (Itemset, IS) είναι υποσύνολο μίας εγγραφής και φυσικά είναι και υποσύνολο του συνόλου “Items”:

$$IS \subseteq T$$

IS C Items

Παράδειγμα Itemset μπορεί να είναι το $IS = \{\text{Γάλα, Αυγά, Ζάχαρη, Σοκολάτα}\}$ σε μία βάση με αγορές από ένα super market, ή το $IS = \{\text{Source Port 2032, Destination Port 80, Source IP Address 192.168.0.23, Destination IP Address 192.168.0.14}\}$ σε μία βάση με πληροφορίες κεφαλίδας καταγραφώντων πακέτων που διέρχονται από κάποιο switch port ενός δικτύου υπολογιστών.

Συχνά στοιχεία - Frequent itemsets

Ένα «σύνολο στοιχείων» (Itemset) το οποίο εμφανίζεται συχνά σε εγγραφές της βάσης δεδομένων αποτελεί ένα «συχνό στοιχείο». Για να υπάρξει ένα μέτρο-όριο (threshold) για τον καθορισμό του ποια «σύνολα στοιχείων» θα θεωρούνται «συχνά στοιχεία», καθορίζεται (τις περισσότερες φορές αυθαίρετα) ένας ελάχιστος αριθμός εμφανίσεων. Τα «σύνολα στοιχείων» τα οποία ξεπερνούν αυτόν τον προκαθορισμένο αριθμό εμφανίσεων στην βάση δεδομένων καλούνται «συχνά στοιχεία».

Στήριξη - Support

Ο αριθμός ο οποίος καθορίζει το όριο πάνω από το οποίο τα «σύνολα στοιχείων» θεωρούνται «συχνά στοιχεία», καλείται στήριξη (support). Στήριξη ενός «συνόλου στοιχείων» ορίζεται ως ο λόγος του αριθμού των εγγραφών στις οποίες εμφανίζεται το «σύνολο στοιχείων» διά του συνολικού αριθμού εγγραφών της βάσης δεδομένων. Ο αριθμός αυτός επιλέγεται τις περισσότερες φορές με βάση τον αριθμό των «συχνών στοιχείων» που θέλει κάποιος να εξάγει από μία βάση δεδομένων. Ο αριθμός αυτός (των παραγόμενων «συχνών στοιχείων»), χρησιμοποιώντας την ίδια στήριξη, μπορεί να διαφοροποιείται από βάση σε βάση. Μπορεί δηλαδή με ίδια στήριξη να έχουμε διαφορετικό αριθμό «συχνών στοιχείων» σε μία βάση A από ότι σε μία βάση B με διαφορετικά δεδομένα από την πρώτη. Ο αριθμός αυτός καθορίζεται σε μεγάλο βαθμό από τη «διαφορετικότητα» των στοιχείων που εμφανίζονται μεταξύ των εγγραφών. Μία βάση με μικρότερη «διαφορετικότητα» στα στοιχεία μεταξύ των εγγραφών της είναι πιο λογικό να παράγει μεγαλύτερο αριθμό «συχνών στοιχείων» από μία άλλη βάση με μεγαλύτερη «διαφορετικότητα», χρησιμοποιώντας την ίδια στήριξη για την παραγωγή των «συχνών στοιχείων».

Κάθε λοιπόν «σύνολο στοιχείων» έχει και μία στήριξη που καθορίζει το πόσο συχνά αυτό εμφανίζεται στη βάση δεδομένων ή το πόσο συχνά τα στοιχεία που συνθέτουν το «σύνολο

στοιχείων» εμφανίζονται μαζί. Όσο πιο μεγάλη είναι η στήριξη τόσο πιο συχνά εμφανίζονται τα στοιχεία μαζί στις εγγραφές της βάσης δεδομένων.

Κανόνες συσχέτισης – Association rules

Ένας «κανόνας συσχέτισης» (Association rule) είναι μία έκφραση της μορφής:

$$IS_a \Rightarrow IS_b$$

όπου $IS_a = \{i_1, i_2, i_3, \dots, i_n\}$ και $IS_b = \{i_{n+1}, i_{n+2}, \dots, i_k\}$ είναι «σύνολα στοιχείων» με τα $i_1 \dots i_k$ να είναι στοιχεία μίας βάσης δεδομένων.

Παράδειγμα κανόνα συσχέτισης: $\{\text{Γάλα, Αυγά}\} \Rightarrow \{\text{Ζάχαρη, Σοκολάτα}\}$

Ένας «κανόνας συσχέτισης» καθορίζει το εξής γεγονός: όταν εμφανίζεται σε μία εγγραφή της βάσης το «σύνολο στοιχείων» IS_a είναι πολύ πιθανό (υπάρχει μαθηματικά μεγάλη πιθανότητα) στην ίδια εγγραφή να εμφανιστεί και το «σύνολο στοιχείων» IS_b . Με άλλα λόγια, αν θέλαμε να μεταφράσουμε σε ανθρώπινη λογική ένα κανόνα συσχέτισης θα μπορούσαμε να πούμε τα εξής:

«Εάν κάποιος αγοράσει τα προϊόντα IS_a τότε είναι πολύ πιθανό να αγοράσει και τα προϊόντα IS_b .»

ή

«Εάν στην κεφαλίδα του πακέτου εμφανίζεται το Destination Port Number 25, είναι πολύ πιθανό να εμφανιστεί ταυτόχρονα και το Destination IP Address 192.168.1.12».

ή

«Εάν στην κεφαλίδα του πακέτου εμφανίζονται ενεργοποιημένες οι σημαίες SYN και FIN είναι πολύ πιθανό να μιλάμε για SYN-FIN attack».

Οι «κανόνες συσχέτισης» λοιπόν δίνουν τη δυνατότητα να εξαχθούν συμπεράσματα και να δημιουργηθούν προβλέψεις εμφάνισης δεδομένων. Η εφαρμογή και χρήση τους είναι πολύ συχνή.

Π.χ. υπάρχουν δουλειές-εργασίες που εφαρμόζουν την εξαγωγή «κανόνων συσχέτισης» στην εύρεση-ανακάλυψη προτύπων (patterns) από δεδομένα που προέρχονται από log files web servers. Τα πρότυπα αυτά στο συγκεκριμένο παράδειγμα μπορεί που σχετίζονται με πρόσβαση σε συγκεκριμένες τοποθεσίες σε συγκεκριμένα χρονικά διαστήματα κλπ.

Εμπιστοσύνη - Confidence

Για την ενίσχυση της αξιοπιστίας των «κανόνων συσχέτισης» (πόσο βέβαιο είναι το γεγονός που περιγράφει ο «κανόνας συσχέτισης»), εκτός από τον παράγοντα «στήριξη» χρησιμοποιείται και ένας ακόμη παράγοντας, η «εμπιστοσύνη» (αναφέρεται στη βιβλιογραφία και ως «ισχύς κανόνα»).

Ο όρος «εμπιστοσύνη» είναι ανάλογος του όρου “conditional probability”, ο οποίος δίνει την πιθανότητα πραγματοποίησης-εμφάνισης ενός γεγονότος, όταν κάποιος άλλο έχει συμβεί. Ένας «κανόνας συσχέτισης» αναπαριστά ακριβώς αυτό το γεγονός, την εμφάνιση δηλαδή ενός «συχνού στοιχείου» (ενός γεγονότος) σε μία έγγραφη της βάσης δεδομένων, όταν κάποιος άλλο «συχνό στοιχείο» εμφανίζεται σε αυτή (ένα άλλο γεγονός συμβαίνει).

Ένας «κανόνας συσχέτισης» αποτελείται από αριστερό μέρος και το δεξί μέρος. Στον κανόνα:

$$IS_a \Rightarrow IS_b$$

το «σύνολο στοιχείων» IS_a αποτελεί το αριστερό μέρος του κανόνα, ενώ το IS_b αποτελεί το δεξί μέρος αυτού. Η εμπιστοσύνη ενός «κανόνα συσχέτισης» υπολογίζεται διαιρώντας τη στήριξη του «κανόνα συσχέτισης» με τη στήριξη του αριστερού μέρους του κανόνα.

$$\text{Confidence} = \frac{\text{Support}(IS_a \cup IS_b)}{\text{Support } IS_a}$$

Maximal Frequent Itemsets

Μέχρι σήμερα υπάρχει μεγάλος αριθμός αλγορίθμων οι οποίοι χρησιμοποιούν διαφορετικές προσεγγίσεις για τον υπολογισμό των «συχνών στοιχείων». Τις περισσότερες φορές ο αριθμός των παραγόμενων «συχνών στοιχείων» είναι μεγάλος με αποτέλεσμα η πληροφορία να είναι δύσκολα

αξιοποιήσιμη. Είναι απαραίτητο ο αριθμός αυτός να είναι ο «κατάλληλος» (ούτε πολύ λίγα «συχνά στοιχεία» αλλά ούτε και πάρα πολλά) ώστε η πληροφορία που εξάγεται να είναι ουσιαστική και εύκολα εντοπίσιμη.

Συνήθης τεχνική για τη μείωση του αριθμού των παραγόμενων «συχνών στοιχείων» είναι να τροποποιείται ο παράγοντας «στήριξη», είτε να επιλέγονται μόνο τα «συχνά στοιχεία» τα οποία περιέχουν πάνω από κάποιο αριθμό στοιχείων. Προς αυτό το στόχο, τη μείωση δηλαδή του αριθμού των παραγόμενων «συχνών στοιχείων» έχουν δημιουργηθεί αρκετοί αλγόριθμοι [42]. Το πρόβλημα μπορεί να λυθεί με την εύρεση των λεγόμενων “maximal frequent itemsets”.

“Maximal frequent itemsets” ονομάζονται τα παραγόμενα «συχνά στοιχεία» τα οποία δεν αποτελούν υποσύνολα άλλων «συχνών στοιχείων». Ένα «συχνό στοιχείο» λοιπόν είναι και “maximal frequent itemset” εφόσον το σύνολο των στοιχείων του δεν εμφανίζονται μαζί σε κάποιο άλλο «συχνό στοιχείο». Με αυτό τον τρόπο επιτυγχάνεται η μείωση των «συχνών στοιχείων». Αλγόριθμοι όπως οι Apriori, MaxEclat, Max-Miner (Bayardo, 1998), MaxClique (Zaki, 2000), Depth project (Agrawal et al., 2000), GenMax (Gouda and Zaki, 2001), Mafia (Burdick et al., 2001), Dualize and Advance (Gunopulos et al., 2003), FPMax (Grahne and Zhu, 2005) χρησιμοποιούνται για την παραγωγή “Maximal frequent itemsets” [42].

Closed Frequent Itemsets

Ένας άλλος ενδιαφέρον τρόπος αναπαράστασης «συχνών στοιχείων» είναι η εξαγωγή των λεγόμενων “closed frequent itemsets”. Ένα frequent itemset αποτελεί “closed frequent itemset” εάν κανένα από τα υπερσύνολα που το περιέχουν δεν έχει την ίδια στήριξη (support) με αυτό [43]. Με βάση τον κανόνα αυτό θα μπορούσαμε να πούμε ότι “closed frequent itemset” είναι ένα “frequent itemset” του οποίου κανένα υπερσύνολο δεν περιέχεται σε κάθε transaction στο οποίο περιέχεται το συγκεκριμένο “frequent itemset”.

Αντίστοιχα με τα “maximal frequent itemsets” δημιουργήθηκε μεγάλος αριθμός αλγορίθμων για την εξαγωγή των “closed frequent itemsets” από μία βάση δεδομένων. Αλγόριθμοι όπως οι close (Nicolas Pasquier, Yves Bastide, Rafik Taouil, Lotfi Lakhal, 1999), closet (Jian Pei, Jiawei Han, Runying Mao,

2000), charm (Mohammed Javeed Zaki, Ching-Jiu Hsiao, 2002) είναι κάποιοι από τους αλγορίθμους για την παραγωγή των closed frequent itemsets..

2.4 Αλγόριθμοι για την Παραγωγή Association Rules και Frequent Itemsets

Λόγω της μεγάλης απήχησης που είχε η διαδικασία data mining δημιουργήθηκαν και αντίστοιχα αρκετοί αλγόριθμοι για την εξαγωγή «συχνών στοιχείων» και «κανόνων συσχέτισης». Αλγόριθμοι όπως οι:

- AIS από τους Rakesh Agrawal, Tomasz Imielinski και Arun Swami (1993) [34],
- SETM από τους Maurice Houtsma και Arun Swami (1995) [44],
- Apriori από τους Rakesh Agrawal και Ramakrishnan Srikant (1994) [45],
- ECLAT - Echivalence Class Clustering and Bottom-up Lattice Traversal από τους Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara και Wei Li (1997) [46],
- FP-Growth από τους Jiawei Han, Jian Pei και Yiwen Yin (2000) [47],
- TREE-PROJECTION από τους Ramesh C. Agarwal, Charu C. Aggarwal και V.V.V. Prasad (2000) [48],
- RELIM (Recursive Elimination) από τον Christian Borgelt (2005) [49]

είναι κάποιοι από τους σημαντικότερους αλγορίθμους που χρησιμοποιούνται για αυτό το σκοπό. Λόγω του γεγονότος ότι ο αλγόριθμος Apriori αποτέλεσε τον βασικό αλγόριθμο για την επίτευξη του στόχου της παρούσας μεταπτυχιακής διατριβής (για την παραγωγή των frequent itemsets όπως θα περιγραφεί και στις επόμενες ενότητες), σε αυτή την παράγραφο θα πραγματοποιηθεί μία σύντομη περιγραφή αυτού.

2.4.1 Περιγραφή Αλγορίθμου Apriori

Το πρώτο βήμα του αλγορίθμου στη διαδικασία εύρεσης των «συχνών στοιχείων» (frequent itemsets) είναι η εύρεση της «στήριξης» (support) όλων των στοιχείων (1-itemset) της βάσης

δεδομένων [45]. Με μία πρώτη σάρωση ολόκληρης της βάσης δεδομένων μετρούνται οι εμφανίσεις των στοιχείων στη βάση (συγκεκριμένα στις εγγραφές αυτής). Ο αριθμός που προκύπτει για κάθε στοιχείο διαιρείται με το συνολικό αριθμό των εγγραφών της βάσης ώστε να προκύψει η «στήριξη» του κάθε στοιχείου (1-itemset). Εφόσον η «στήριξη» του στοιχείου (που υπολογίστηκε στο προηγούμενο βήμα) είναι μεγαλύτερη από την προκαθορισμένη ελάχιστη «στήριξη» (αυτή αποτελεί μαζί με την ελάχιστη «εμπιστοσύνη» (confidence) και τη βάση δεδομένων εισόδους του αλγορίθμου) τότε το στοιχείο αυτό αποτελεί «συχνό στοιχείο». Δημιουργείται έτσι μία ομάδα από «συχνά στοιχεία» (L1) τα οποία αποτελούνται από ένα στοιχείο το κάθε ένα.

Στο δεύτερο βήμα του αλγορίθμου πραγματοποιούνται όλοι οι δυνατοί συνδυασμοί των «συχνών στοιχείων» της προηγούμενης ομάδας L1 (γενικά L_{k-1}) δημιουργώντας με αυτό τον τρόπο υποψήφια «συχνά στοιχεία» με δύο στοιχεία το κάθε ένα (γενικά με k στοιχεία). Σαρώνεται πάλι η βάση ώστε να βρεθεί ποια από αυτά τα υποψήφια «συχνά στοιχεία» αποτελούν όντως «συχνά στοιχεία» (η «στήριξη» τους είναι μεγαλύτερη από την ελάχιστη «στήριξη» που έχει δοθεί ως είσοδος στον αλγόριθμο). Με αυτό τον τρόπο δημιουργείται μία νέα ομάδα με «συχνά στοιχεία» (L2, γενικά L_k) τα οποία αποτελούνται από δύο (γενικά k) στοιχεία.

Η διαδικασία αυτή (το δεύτερο βήμα) επαναλαμβάνεται μέχρι να φτάσουμε στην περίπτωση κατά την οποία δεν θα υπάρχουν άλλα υποψήφια «συχνά στοιχεία». Μέχρι και αυτό το βήμα έχουν βρεθεί όλα τα «συχνά στοιχεία» της βάσης δεδομένων. Τα επόμενα βήματα υλοποιούνται για την εύρεση των «κανόνων συσχέτισης».

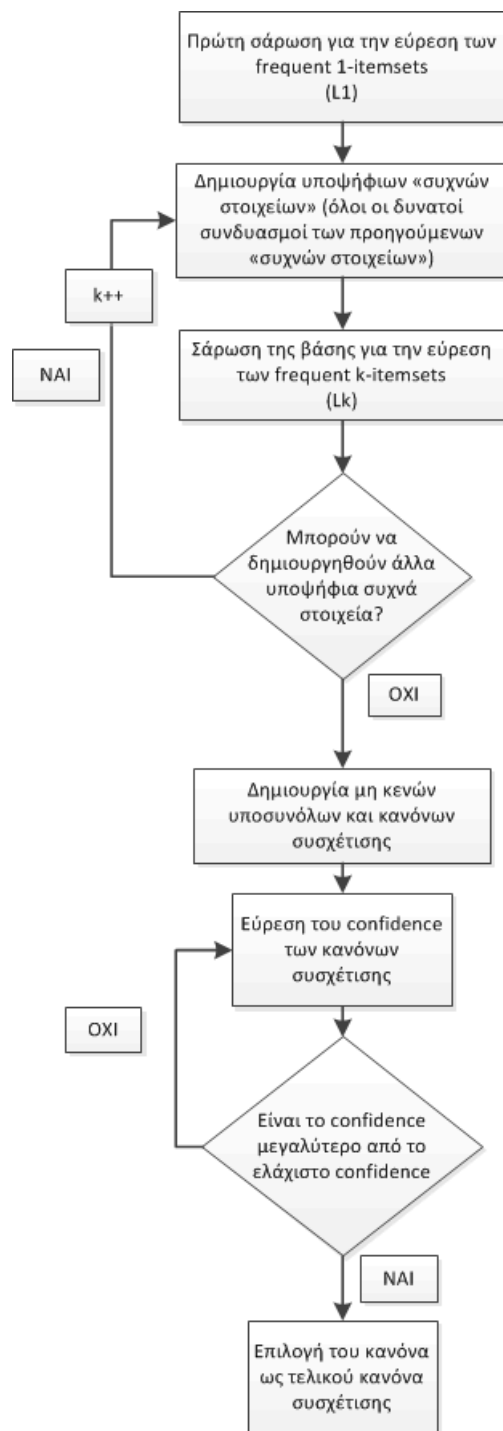
Στο βήμα αυτό δημιουργούνται όλα τα δυνατά μη κενά υποσύνολα όλων των «συχνών στοιχείων» IS. Για κάθε μη κενό υποσύνολο S ενός «συχνού στοιχείου» IS δημιουργούνται κανόνες συσχέτισης της μορφής:

$$S \Rightarrow (IS - S)$$

των οποίων η «εμπιστοσύνη» υπολογίζεται. Εφόσον αυτή είναι μεγαλύτερη από την ελάχιστη «εμπιστοσύνη» τότε ο κανόνας αποτελεί έναν από τους τελικούς «κανόνες συσχέτισης» που εξάγονται από τη βάση.

Φυσικά η επιλογή της ελάχιστης τιμής «στήριξης» και «εμπιστοσύνης» παίζουν μεγάλο ρόλο στο μέγεθος του αριθμού των παραγόμενων «συχνών στοιχείων» και «κανόνων συσχέτισης» αντίστοιχα.

Ακολουθεί flowchart με τα παραπάνω βήματα:



Σχήμα 2.2: Βήματα αλγορίθμου Apriori

2.5 «Εξόρυξη Δεδομένων» στη Διαχείριση Δικτύων Υπολογιστών

Όπως αναφέρθηκε και στην εισαγωγική παράγραφο αυτού του κεφαλαίου, η γνώση που ανακαλύπτεται από τα δοθέντα δεδομένα ύστερα από την εφαρμογή των βημάτων που αναφέρθηκαν στην παράγραφο με τίτλο «Διαδικασία Εξόρυξης δεδομένων» μπορεί να πάρει μεταξύ άλλων τη μορφή «κανόνων συσχέτισης», προτύπων ή προτύπων σε σχέση με το χρόνο. Η παραγωγή τέτοιας «γνώσης» θα μπορούσε να βοηθήσει στη βελτίωση της διαδικασίας παρακολούθησης δικτύων υπολογιστών καθώς και στην απαλλαγή τους από πιθανούς κινδύνους και επιθέσεις.

Οι «κανόνες συσχέτισης», βάσει ετυμολογίας, συσχετίζουν στοιχεία μίας βάσης δεδομένων μεταξύ τους. Η λογική τους στηρίζεται στο ότι εφόσον κάποιος συγκεκριμένος συνδυασμός στοιχείων εμφανίζεται σε μεγάλο αριθμό εγγραφών της υπό εξέτασης βάσης δεδομένων, είναι πολύ πιθανό στα δεδομένα που θα δημιουργηθούν από εδώ και πέρα να έχουμε τον ίδιο συνδυασμό στοιχείων στις εγγραφές. Π.χ. εφόσον τα στοιχεία A και B εμφανίζονται πολύ συχνά μαζί σε μεγάλο αριθμό εγγραφών της βάσης δεδομένων, υπάρχει μεγάλη πιθανότητα στις επόμενες εγγραφές, εφόσον εμφανιστεί το στοιχείο A να εμφανιστεί μαζί με το B. Αυτό μπορεί να μεταφραστεί σε πληροφορία που σχετίζεται με κάποιες από τις καθημερινές εργασίες του ανθρώπου, όπως το λεγόμενο “market-basket problem”. Σύμφωνα με αυτό όταν στις αγορές που καταγράφονται σε μία βάση δεδομένων παρατηρούνται πολλές εμφανίσεις συγκεκριμένων συνδυασμών προϊόντων μπορεί κανείς να συμπεράνει ότι υπάρχει μεγάλη πιθανότητα κάποια προϊόντα να αγοράζονται μελλοντικά μαζί. Στις μελλοντικές αγορές δηλαδή, είναι πολύ πιθανό αγοράζοντας κάποιος πελάτης ένα A προϊόν να αγοράσει και ένα συγκεκριμένο B προϊόν εφόσον αυτά τα δύο εμφανίζονται πολύ συχνά μαζί σε καταγεγραμμένες συναλλαγές. Αυτή η πληροφορία θα μπορούσε να χρησιμοποιηθεί «έξυπνα» ώστε να τοποθετηθούν σε κατάλληλες θέσεις στα ράφια κάποιου super market συγκεκριμένα προϊόντα.

Αντίστοιχη χρησιμότητα μπορεί να έχουν οι «κανόνες συσχέτισης» και στην παρακολούθηση και προστασία των δικτύων υπολογιστών. Τα δεδομένα που προέρχονται από την παρακολούθηση της κίνησης που διέρχεται από μία σύνδεση, θα μπορούσαν να χρησιμοποιηθούν ως μια βάση δεδομένων από την οποία μπορούν με την εφαρμογή αλγορίθμων εξόρυξης να δημιουργηθούν

«κανόνες συσχέτισης» με στόχο την πρόβλεψη της εμφάνισης απειλών. Πως θα μπορούσε να γίνει κάτι τέτοιο: Κάποιο είδος επιθέσεων χαρακτηρίζονται τις περισσότερες φορές από μία αλληλουχία πακέτων με συγκεκριμένα χαρακτηριστικά. Αυτά τα χαρακτηριστικά θα μπορούσαν να δημιουργήσουν κανόνες συσχέτισης οι οποίοι με τη σειρά τους θα μπορούσαν να χρησιμοποιηθούν από ένα σύστημα ανίχνευσης επιθέσεων. Το σύστημα αυτό θα μπορούσε να ανιχνεύει την εμφάνιση κάποιων χαρακτηριστικών στα πακέτα και με βάση τους «κανόνες συσχέτισης» να προβλέπει επερχόμενη πιθανή επίθεση (εφόσον παρατηρούνται κάποια χαρακτηριστικά σε κάποια πακέτα τα οποία είναι επιβεβαιωμένο ότι είναι χαρακτηριστικά κάποιων αρχικών πακέτων από μία επίθεση, είναι πολύ πιθανό να έχουμε εμφάνιση της επίθεσης). Μπορεί λοιπόν η δημιουργία «κανόνων συσχέτισης» να βοηθήσει στην εύρεση ανώμαλης δραστηριότητας προβλέποντας ή και αποκαλύπτοντας μία πραγματική επερχόμενη επίθεση.

Υπάρχουν Intrusion Detection Systems (IDS) (η λειτουργία τους περιγράφηκε στο πρώτο κεφάλαιο της παρούσας εργασίας) τα οποία χρησιμοποιούν την λογική των «κανόνων συσχέτισης» ώστε να αναγνωρίσουν «κακή χρήση» (misuse). Έρευνες παρουσιάζουν IDS που χρησιμοποιούν μεταξύ άλλων και τεχνικές εξόρυξης δεδομένων για τη δημιουργία «κανόνων συσχέτισης» με γενικότερο στόχο την προστασία των δικτύων υπολογιστών. Το άρθρο με τίτλο “Misuse Detection Using Hybrid of Association Rule Mining and Connectionist Modeling” από τους Mansour Sheikhan και Zahra Jadidi [50] παρουσιάζει ένα anomaly based (misuse based) IDS το οποίο χρησιμοποιεί «κανόνες συσχέτισης» για την αναγνώριση αποκλίσεων από την ορθή λειτουργία (baseline). Αντίστοιχη είναι και η δουλειά που αναφέρεται στο άρθρο με τίτλο “Combination of Elman Neural Network and Classification-based Predictive Association Rules to improve Computer Networks’ Security” από τους Mansour Sheikhan, Davnod Gharavian [51]. Θα παρουσιαστούν και άλλες τέτοιες εφαρμογές data mining τεχνικών στην ανίχνευση επιθέσεων στην επόμενη παράγραφο.

Αντίστοιχα με τους «κανόνες συσχέτισης» οι διαδικασίες εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν για τον εντοπισμό προτύπων ή ακόμη και χρονικών ακολουθιών. Τα πρότυπα που προκύπτουν είναι ακολουθίες καταστάσεων οι οποίες επιβεβαιώνονται από το γεγονός της συχνής εμφάνισής τους. Ένα παράδειγμα στο “market-basket problem” θα μπορούσε να αποτελέσει το γεγονός ότι κάποια προϊόντα αγοράζονται πολύ συχνότερα μαζί σε κάποιες συγκεκριμένες χρονικές εποχές σε αντίθεση με κάποιες άλλες. Κάτι αντίστοιχο θα μπορούσε να χρησιμοποιηθεί και στην ανίχνευση επιθέσεων σε δίκτυα υπολογιστών. Η παρατήρηση συγκεκριμένων χαρακτηριστικών

στα πακέτα σε συγκεκριμένες στιγμές μπορεί να υποδηλώνει την ύπαρξη κάποιου είδους επίθεσης. Π.χ. τεχνικές εξόρυξης δεδομένων μπορεί να παρατηρούν και να καταγράφουν ως πρότυπα την ύπαρξη συγκεκριμένων χαρακτηριστικών στα πακέτα μόνο κάποιες συγκεκριμένες χρονικές στιγμές, όπως εμφάνιση αιτήσεων πιστοποίησης ταυτότητας χρηστών μόνο σε ώρες λειτουργίας της επιχείρησης. Η εμφάνιση αυτών των χαρακτηριστικών-προτύπων σε άλλες στιγμές (απόκλιση από το baseline) μπορεί να δηλώνουν την ύπαρξη επίθεσης.

Στα μεγάλα και σύνθετα δίκτυα υπολογιστών με μεγάλο όγκο δεδομένων τα alarms που εγείρονται καθώς και τα γεγονότα που καταγράφονται σε log files καθημερινά είναι πάρα πολλά και κατά συνέπεια είναι επίπονη διαδικασία για τον διαχειριστή του δικτύου η απομόνωση των σημαντικότερων από αυτά. Ο όγκος πληροφοριών που συλλέγονται και πρέπει να αναλυθούν, δεδομένου και του αυξανόμενου αριθμού των επιθέσεων, είναι μεγάλος οπότε ο εντοπισμός των πιο σημαντικών από αυτές καθίσταται δύσκολη διαδικασία. Η διαδικασία εξόρυξης δεδομένων μπορεί να βοηθήσει στον περιορισμό των συμβάντων που πρέπει ένας αναλυτής να παρακολουθεί καθώς και στην εύρεση και εντοπισμό των σοβαρότερων από αυτά. Μπορεί να βοηθήσει στην αφαίρεση κανονικής δραστηριότητας (normal activity) από δεδομένα ενός συναγερμού (alarm) ώστε να επιτραπεί στον αναλυτή να εστιάσει στα δεδομένα της πραγματικής επίθεσης. Τέλος, θα μπορούσε να χρησιμοποιηθεί για την αναγνώριση ψευδών συναγερμών συνδυάζοντας και ελέγχοντας πρότυπα χαρακτηριστικών στα πακέτα [52].

Γίνεται λοιπόν αντιληπτό ότι οι διαδικασίες εξόρυξης δεδομένων μπορούν να βοηθήσουν στο έργο παρακολούθησης και σωστής λειτουργίας ενός δικτύου υπολογιστών και ειδικότερα στην παρακολούθηση της ασφάλειας αυτών. Στην επόμενη παράγραφο θα γίνει μικρή αναφορά σε μελέτες οι οποίες χρησιμοποιούν τις τεχνικές εξόρυξης δεδομένων στην επιστήμη των δικτύων υπολογιστών.

2.6 Εφαρμογές Διαδικασίας Εξόρυξης Δεδομένων στα Δίκτυα Υπολογιστών

Υπάρχουν αρκετές έρευνες προς την κατεύθυνση της χρήσης και εφαρμογής τεχνικών εξόρυξης δεδομένων στην βελτίωση της διαχείρισης ενός δικτύου υπολογιστών. Οι περισσότερες από αυτές

στοχεύουν στο να βελτιώσουν τους τρόπους ανίχνευσης και αντιμετώπισης επιθέσεων που μπορούν να παρουσιαστούν. Στην παράγραφο αυτή πραγματοποιείται μία δειγματοληπτική παρουσίαση τέτοιων εργασιών.

A Data Mining Framework for Building Intrusion Detection Models Based on IPv6 [53]

Αρκετές δουλειές έχουν χρησιμοποιήσει τις τεχνικές εξόρυξης δεδομένων για να βελτιώσουν την ακρίβεια ανίχνευσης εισβολών σε δίκτυα υπολογιστών. Πολλοί ερευνητές έχουν χρησιμοποιήσει «κανόνες συσχέτισης» για το σχεδιασμό anomaly-based Network Intrusion Detection συστημάτων. Η ανάλυση και η ομαδοποίηση των παρακολουθούμενων δεδομένων είναι μία από τις βασικότερες ενέργειες σε ένα IDS. Λόγω του γεγονότος ότι σε ένα δίκτυο υπολογιστών οι πληροφορίες των πακέτων που μπορούν να καταγραφούν και ο αριθμός αυτών σε μία σύνδεση είναι πάρα πολύ μεγάλος, χρησιμοποιούνται οι τεχνικές δημιουργίας «συχνών στοιχείων» και «κανόνων συσχέτισης» για την ομαδοποίηση και κατ' επέκταση μείωση των δεδομένων αυτών. Η δουλειά με τίτλο "A Data Mining Framework for Building Intrusion Detection Models Based on IPv6" [53] από τους Zenghui Liu και Yingxu Lai ασχολείται με τη δημιουργία και εφαρμογή «κανόνων συσχέτισης» σε IPv6 δίκτυα υπολογιστών με στόχο τη δημιουργία ενός αξιόπιστου IPv6 Network Intrusion Detection συστήματος. Σύμφωνα με τη συγκεκριμένη εργασία, λόγω του μεγάλου εύρους (μεγέθους) των subnets που μπορούν να δημιουργηθούν σε ένα δίκτυο IPv6 καθίσταται δυσκολότερη η σάρωση όλου του subnet και κατ' επέκταση και η ανίχνευση επιθέσεων. Επιθέσεις που στοχεύει η συγκεκριμένη μελέτη να εντοπίσει είναι οι ακόλουθες:

- Misuse of routing headers
- Fragmentation related attacks
- Misuse of ICMPv6 and multicast

Οι συγγραφείς χρησιμοποιούν «κανόνες συσχέτισης» που προκύπτουν από τα δεδομένα κίνησης ενός IPv6 δικτύου υπολογιστών, για τη δημιουργία προτύπων-μοντέλων αναγνώρισης επιθέσεων. Χρησιμοποιούν μία παραλλαγή (βελτίωση) του Apriori αλγορίθμου για τη δημιουργία «συχνών στοιχείων» και «κανόνων συσχέτισης» προς αυτό το σκοπό. Εφαρμόζοντας διαφορετικού τύπου επιθέσεις (DoS attacks) σε ένα IPv6 εργαστηριακό δίκτυο υπολογιστών παράγουν δεδομένα κίνησης

τα οποία και χρησιμοποιούν για τη δημιουργία αυτών των «κανόνων συσχέτισης» οι οποίοι βοηθούν στην ανίχνευση επιθέσεων.

Intrusion Detection System Research Based on Data Mining for IPv6 [54]

Πανομοιότυπη με την προηγούμενη είναι και η δουλειά με τίτλο “Intrusion Detection System Research Based on Data Mining for IPv6” από τους Chen Jun και Chen Xiaowei [54]. Η συγκεκριμένη (όπως και η προηγούμενη), εστιάζει στον εντοπισμό του ίδιου ακριβώς τύπου επιθέσεων δημιουργώντας ένα Network Intrusion Detection σύστημα με τη χρήση βελτιωμένου αλγορίθμου Apriori.

A Study of Intrusion Detection System Based on Data Mining [55]

Ανάμεσα στις έρευνες-μελέτες που έχουν πραγματοποιηθεί για τη δημιουργία Intrusion Detection Systems με την ταυτόχρονη χρήση τεχνικών εξόρυξης δεδομένων συγκαταλέγεται και η δουλειά που εμφανίζεται στο άρθρο με τίτλο “A Study of Intrusion Detection System Based on Data Mining” από τους Chunyu Miao και Wei Chen [55]. Σύμφωνα με τη μελέτη αυτή συλλέγονται δεδομένα κεφαλίδων πακέτων από ένα δίκτυο υπολογιστών με τη χρήση εργαλείου sniffer. Στα δεδομένα αυτά εφαρμόζεται ο αλγόριθμος Apriori για τη δημιουργία «κανόνων συσχέτισης» που περιέχουν source και destination IP addresses καθώς και destination port numbers. Οι κανόνες αυτοί χρησιμοποιούνται στη συνέχεια για την ανίχνευση αποκλίσεων από τη συνηθισμένη λειτουργία (baseline) του δικτύου δημιουργώντας έτσι ένα αξιόπιστο σύστημα ανίχνευσης δικτυακών εισβολών (anomaly detection NIDS).

Data Mining Approaches for Intrusion Detection [56]

Στην έρευνα που εμφανίζεται στο άρθρο με τίτλο “Data Mining Approaches for Intrusion Detection” από τους Wenke Lee και Salvatore J. Stolfo [56] χρησιμοποιούνται τεχνικές εξόρυξης δεδομένων για την εύρεση χρήσιμων προτύπων που αφορούν είτε «συμπεριφορά» χρηστών είτε προγραμμάτων με στόχο την αναγνώριση ανωμαλιών και γνωστών επιθέσεων. Για τη δημιουργία των προτύπων αυτών χρησιμοποιούνται αλγόριθμοι παραγωγής «κανόνων συσχέτισης» και «συχνών στοιχείων». Η αποτελεσματικότητα των μοντέλων που προκύπτουν στον εντοπισμό ανωμαλιών επιβεβαιώνεται από τους ερευνητές έπειτα από εφαρμογή των τεχνικών σε συστήματα sendmail και δεδομένα που προέρχονται από το εργαλείο σύλληψης πακέτων TCPdump.

Developing Custom Intrusion Detection Filters Using Data Mining [57]

Ένα μεγάλο μειονέκτημα των Intrusion Detection Systems στην προσπάθειά τους να αναγνωρίσουν «ανώμαλη» κίνηση είναι η ενεργοποίηση λανθασμένων συναγερμών (alarms). Μπορεί δηλαδή επιτρεπτή κίνηση να χαρακτηριστεί από ένα IDS ως «ύποπτη» (τμήμα κίνησης που ανήκει σε κάποιο είδος επίθεσης) ενεργοποιώντας λανθασμένα κάποιο συναγερμό. Υπάρχουν δουλειές οι οποίες χρησιμοποιούν τεχνικές εξόρυξης δεδομένων ως βοηθητικό εργαλείο για την ελαχιστοποίηση αυτών των λανθασμένων συναγερμών. Σε αυτή την κατηγορία εργασιών ανήκει και η έρευνα που εμφανίζεται στο άρθρο με τίτλο “Developing Custom Intrusion Detection Filters Using Data Mining” από τους Chris Clifton και Gary Gengo [57]. Η συγκεκριμένη εργασία προσπαθεί να δημιουργήσει προσαρμοσμένα φίλτρα για τη μείωση των λανθασμένων συναγερμών. Η ιδέα πάνω στην οποία στηρίχτηκε η συγκεκριμένη δουλειά είναι το γεγονός ότι συνήθως οι επιθέσεις δεν περιλαμβάνουν την ίδια μέθοδο κατά την επίθεση (“attackers will probably not repeatedly try the same method” όπως συγκεκριμένα αναφέρουν στο άρθρο τους οι συγγραφείς) οπότε και η επανάληψη των ίδιων συναγερμών έχει μεγάλη πιθανότητα να δηλώνει κανονική λειτουργία (normal operation) οπότε και false alarm [57]. Για τη μείωση του υπολογιστικού κόστους και την εύρεση των πιο συχνών ακολουθιών συναγερμών χρησιμοποιείται από τους ερευνητές μέθοδος εύρεσης «συχνών στοιχείων» (frequent episodes).

Anomaly Extraction in Backbone Networks using Association Rules [37]

Με στόχο τη δημιουργία πιο πρακτικών συστημάτων εντοπισμού ανωμαλιών τα οποία να χαρακτηρίζονται από όσο το δυνατόν λιγότερες λανθασμένες ενεργοποιήσεις (false positives) υλοποιήθηκε και η έρευνα που περιγράφεται στο άρθρο με τίτλο “Anomaly Extraction in Backbone Network using Association Rules” από τους Daniela Brauckhoff, Xenofontas Dimitropoulos, Arno Wagner και Kave Salamatian. Στόχος της δουλειάς αυτής ήταν η δημιουργία ενός συστήματος το οποίο θα μπορεί με πολύ μικρό ποσοστό λάθους να αναγνωρίσει και να απομονώσει τις ροές δεδομένων μέσα από ένα μεγάλο σύνολο ροών, οι οποίες είναι η αιτία για την εμφάνιση ανωμαλιών σε ένα δίκτυο υπολογιστών [37]. Η μεγάλη επιτυχία του συγκεκριμένου project είναι το γεγονός ότι επιτυγχάνει τον αρχικό του στόχο εμφανίζοντας πολύ μικρό ποσοστό λανθασμένων ενεργοποιήσεων, δηλαδή χαρακτηρισμό ροών ως υπεύθυνες για τα ανώμαλα γεγονότα που εμφανίζονται σε ένα δίκτυο. Η μέθοδος που χρησιμοποιείται στην συγκεκριμένη εργασία στηρίζεται στην γενικότερη άποψη ότι σε περίπτωση ανωμαλιών παράγονται συνήθως πολλές ροές με παρόμοια χαρακτηριστικά π.χ. IP διευθύνσεις, port numbers κλπ. Έχοντας αυτό ως κύριο κανόνα, χρησιμοποιεί τεχνικές εξόρυξης «κανόνων συσχέτισης» για τον εντοπισμό των υπεύθυνων για την εμφάνιση ανώμαλων γεγονότων, ροών δεδομένων.

Συνέχεια της δουλειάς αυτής αποτελεί η δουλειά με τίτλο “Automating root-cause analysis of network anomalies using frequent itemset mining” από τους Paredes-Oliva, Dimitropoulos, Molina, Barlet-Ros και Brauckhoff [38]. Σε αυτή πραγματοποιείται η εφαρμογή-παρουσίαση των δυνατοτήτων του συστήματος που έχει αναπτυχθεί και περιγράφεται στην εργασία [37] σε συνδυασμό με ένα τρίτο σύστημα ανίχνευσης ανωμαλιών.

Misuse Detection Using Hybrid of Association Rule Mining and Connectionist Modeling [50]

Αντίστοιχη έρευνα που στοχεύει στο συνδυασμό των τεχνικών εξόρυξης δεδομένων και της βελτίωσης της παρεχόμενης ασφάλειας σε δίκτυα υπολογιστών είναι και η έρευνα από τους Mansour Sheikhan και Zahra Jadidi με τίτλο “Misuse Detection Using Hybrid of Association Rule Mining and Connectionist Modeling” [50]. Στη μελέτη αυτή έχουμε την παρουσίαση ενός υβριδικού misuse-based Intrusion Detection System που συνδυάζει «κανόνες συσχέτισης» για την αναγνώριση προτύπων-κατηγοριών, με νευρωνικό δίκτυο για εκμάθηση (multilayer perception neural network)

παρέχοντας έτσι καλύτερα αποτελέσματα κατηγοριοποίησης. Παρόμοια είναι και η μελέτη που εμφανίζεται στο άρθρο με τίτλο “Combination of Elman Neural Network and Classification-based Predictive Association Rules to improve Computer Networks’ Security” από τους Mansour Sheikhan και Davvood Gharavian [51].

A Novel Intrusion Detection Approach Based on Data Mining [58]

Σε ένα δίκτυο υπολογιστών όπως αναφέρθηκε και στο πρώτο κεφάλαιο της παρούσας εργασίας, ένα μέσο αντιμετώπισης επιθέσεων είναι η δημιουργία αυστηρής πολιτικής ασφάλειας που μεταφράζεται σε δημιουργία κανόνων πάνω σε ένα firewall. Η αποτελεσματικότητα αυτών των κανόνων είναι ύψιστης σημασίας για την ορθή λειτουργία των υπηρεσιών που παρέχει το υπό προστασία δίκτυο υπολογιστών. Στην δουλειά με τίτλο “A Novel Intrusion Detection Approach Based on Data Mining” από τον Yan Yu [58], γίνεται αναφορά σε προσεγγίσεις ανίχνευσης εισβολών στηριζόμενες σε τεχνικές εξόρυξης δεδομένων.

Τεχνικές εξόρυξης δεδομένων χρησιμοποιούνται από τον ερευνητή για τη διαχείριση κανόνων firewalls τα οποία χρησιμοποιούνται για την ανίχνευση και αποτροπή ανωμαλιών. Σε δεδομένα δικτύου που περνούν από firewalls εφαρμόζονται τεχνικές εξόρυξης δεδομένων ώστε να ανιχνευτούν παραβιάσεις των κανόνων και γενικότερα της πολιτικής ασφάλειας. Η διαδικασία αυτή πραγματοποιείται με στόχο την αφαίρεση, αλλαγή και συγχώνευση των κανόνων αυτών. Απώτερος στόχος πάντοτε είναι βελτίωση της εφαρμοζόμενης πολιτικής ασφάλειας. Τα βήματα λοιπόν που ακολουθούνται στην αναφερόμενη εργασία είναι η καταγραφή των διερχόμενων πακέτων, η εφαρμογή αλγορίθμων εξόρυξης δεδομένων σε αυτά και τέλος η τροποποίηση της πολιτικής ασφάλειας των firewalls με βάση τα ανιχνεύσιμα κενά ασφαλείας [58].

Visualizing Network Status [39]

Η παρακολούθηση της κίνησης ενός δικτύου στις μέρες μας έχει γίνει δύσκολη υπόθεση. Ένας από τους παράγοντες που επιτείνει τη δυσκολία αυτή είναι το γεγονός ότι ο αριθμός των συσκευών που μπορούν να χρησιμοποιηθούν για την ανταλλαγή δεδομένων μέσω δικτύου έχει γίνει πλέον υπερβολικά μεγάλος [39]. Η συγκεκριμένη έρευνα από τους Kenichi Yoshida, Yusuke Shomura και Yoshinori Watanabe προσπαθεί να βοηθήσει στη βελτίωση του τρόπου εύρεσης σημαντικών

φαινομένων σε ένα δίκτυο υπολογιστών. Η προσπάθεια αυτή γίνεται χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων και συγκεκριμένα αλγόριθμο παραγωγής «συχνών στοιχείων» από μία βάση δεδομένων που αποτελείται από καταγραφέντα δεδομένα κίνησης δικτύου. Στόχος η εξαγωγή συμπερασμάτων και ο εντοπισμός σημαντικών γεγονότων-φαινομένων που μπορεί να εμφανίζονται σε ένα δίκτυο υπολογιστών. Η συγκεκριμένη μελέτη προσπαθεί να αποτυπώσει και να αναλύσει την «ποικιλία» (variety) ή αλλιώς τη διαφορετικότητα των δεδομένων που εμφανίζονται στις εγγραφές της βάσης δεδομένων στις οποίες εμφανίζονται τα «συχνά στοιχεία».

Η τεχνική που προτείνεται μπορεί να οδηγήσει-βοηθήσει στην εύρεση προβληματικών ροών δεδομένων. Σύμφωνα με αυτήν, αφού βρεθούν τα «συχνά στοιχεία», καταγράφονται οι εγγραφές της βάσης στις οποίες εμφανίζονται αυτά. Τα μη «συχνά στοιχεία» σε αυτά τα transactions και ο αριθμός των «ποικιλιών» που μετρούνται, αποτυπώνονται σε γράφημα το οποίο μπορεί να χρησιμοποιηθεί για τον εντοπισμό προβληματικών ροών.

Παράδειγμα αποτελεί η κίνηση που προκαλείται από κάποιο worm. Ένα internet worm (όπως έχει αναφερθεί και στο πρώτο κεφάλαιο) είχε, τουλάχιστον αρχικά, ως στόχο το γρήγορο πολλαπλασιασμό του και την μόλυνση άλλων υπολογιστών στο δίκτυο που έχει διεισδύσει. Δημιουργεί λοιπόν πολλά TCP/IP πακέτα που έχουν μεταξύ τους (στην επικεφαλίδα των πακέτων) κοινά στοιχεία όπως source IP address και destination port number. Ταυτόχρονα το στοιχείο destination IP address είναι διαφορετικό δημιουργώντας έτσι μεγάλη «ποικιλία» destination IP addresses. Αναπαριστώντας την πληροφορία αυτή της «ποικιλίας» μπορεί να γίνει ευκολότερος ο εντοπισμός και χαρακτηρισμός «περίεργου» φαινομένου στην κίνηση του δικτύου.

Αντίστοιχο παράδειγμα αποτελούν οι επιθέσεις Distributed Denial of Service Attacks. Οι ροές δεδομένων και τα «συχνά στοιχεία» που προκύπτουν από αυτές, έχουν μεταξύ τους κοινά πεδία όπως το destination IP address και destination port number ενώ έχουν μεγάλη «ποικιλία» (διαφορετικότητα) σε source IP address και source port number. Αντίστοιχα σε επιθέσεις port scan τα transactions στα οποία εμφανίζονται τα «συχνά στοιχεία» θα έχουν μεταξύ τους ίδιο source IP address, destination IP address και source port number και μεγάλη «ποικιλία» (διαφορετικότητα) σε port numbers.

Η μελέτη λοιπόν αυτή, αποτελεί αξιόλογη προσπάθεια και ανήκει στην κατηγορία των δημοσιεύσεων που αναφέρονται στον εντοπισμό ανωμαλιών παρακολουθώντας και αποτυπώνοντας την κίνηση σε ένα δίκτυο υπολογιστών. Με βάση αυτή γίνεται εντονότερη η πεποίθηση της σημαντικότητας που μπορεί να έχει η χρήση των αλγορίθμων εξόρυξης δεδομένων στην βελτίωση της παρακολούθησης και διαχείρισης δικτύων υπολογιστών.

2.7 Συμπεράσματα

Οι τεχνικές εξόρυξης δεδομένων μπορούν μέσα από μεγάλες βάσεις δεδομένων να παράγουν κρυμμένη, χρήσιμη γνώση η οποία μπορεί να χρησιμοποιηθεί στη διαδικασία πραγματοποίησης κρίσιμων αποφάσεων. Έχουν στην πορεία των χρόνων δημιουργηθεί διάφοροι αλγόριθμοι για την εξαγωγή «συχνών στοιχείων», «κανόνων συσχέτισης», κατηγοριοποιήσεων (classifications) καθώς και προτύπων (patterns) μέσω των τεχνικών αυτών. Έχουν προταθεί και ερευνηθεί διάφορες εφαρμογές που θα μπορούσαν να βοηθηθούν από τα παραπάνω αποτελέσματα μεταξύ των οποίων και η τεχνολογία δικτύων υπολογιστών.

Με βάση τις μελέτες που έχουν παρατεθεί στην προηγούμενη παράγραφο γίνεται αντιληπτό ότι υπάρχουν αρκετές έρευνες μέχρι σήμερα που χρησιμοποιούν τις τεχνικές εξόρυξης δεδομένων και τα αποτελέσματα αυτών στη δημιουργία εργαλείων που βοηθούν στην ασφάλεια ενός δικτύου, όπως είναι τα Intrusion Detection Systems και τα firewalls. Μια κατηγορία Intrusion Detection Systems χρησιμοποιούν «κανόνες συσχέτισης» για τον εντοπισμό επιθέσεων ή και αποκλίσεων από την ομαλή λειτουργία ενός δικτύου υπολογιστών (παράδειγμα «κανόνα συσχέτισης»: αν υπάρχουν ταυτόχρονα ενεργοποιημένα τα flags S και F σε πακέτα τότε είναι πιθανό να μιλάμε για SYN-FIN attack). Άλλη κατηγορία Intrusion Detection Systems χρησιμοποιούν παραγόμενα πρότυπα μέσω των τεχνικών εξόρυξης για να δημιουργήσουν υπογραφές επιθέσεων. Καταγράφουν δηλαδή επαναλαμβανόμενες αλληλουχίες πακέτων-δεδομένων οι οποίες εμφανίζονται πριν ή κατά τη διάρκεια κάποιας επιβεβαιωμένη επίθεσης δημιουργώντας έτσι ένα είδος υπογραφών επιθέσεων. Τέλος, έχουν υλοποιηθεί και συστήματα εντοπισμού ανωμαλιών τα οποία απομονώνουν και δείχνουν τις ροές δεδομένων οι οποίες οδήγησαν σε κάποιο συναγερμό (alert), κάνοντας ποιο εύκολο έτσι τον εντοπισμό της αιτίας του εμφανιζόμενου προβλήματος.

Από όλα όσα έχουν αναφερθεί στο παρόν κεφάλαιο γίνεται αντιληπτό ότι οι τεχνικές εξόρυξης μπορούν να βοηθήσουν στη βελτίωση της ασφάλειας των δικτύων και γενικότερα στην καλύτερη και ευκολότερη διαχείριση αυτών. Φυσικά υπάρχουν και άλλοι τομείς στο μεγάλο πεδίο διαχείρισης δικτύων στους οποίους θα μπορούσαν οι τεχνικές εξόρυξης δεδομένων να φανούν χρήσιμες. Η σύνοψη, παρακολούθηση και παρουσίαση της κίνησης που διέρχεται από ένα δίκτυο θα μπορούσε να είναι ένας από αυτούς. Μία τέτοια προσέγγιση θα μπορούσε να βοηθήσει στην ευκολότερη παραγωγή μίας γενικότερης εικόνας της κατάστασης ενός δικτύου, με στόχο τη διευκόλυνση του διαχειριστή-αναλυτή του δικτύου στη διαχείριση του εν λόγω δικτύου. Αυτός είναι και ο γενικότερος στόχος της παρούσας μεταπτυχιακής διατριβής, η εφαρμογή δηλαδή τεχνικών εξόρυξης δεδομένων και συγκεκριμένα «συχνών στοιχείων» στη σύνοψη, για λόγους παρακολούθησης, της κίνησης δικτύου.

Στις έρευνες που έχουν αναφερθεί μέχρι τώρα οι τεχνικές γραφικής απεικόνισης «συχνών στοιχείων» είναι απύσες. Δεν αναφέρεται σε κάποια από αυτές το εάν μπορεί η γραφική απεικόνιση των «συχνών στοιχείων» να χρησιμοποιηθεί στη γενικότερη διαχείριση ενός δικτύου. Ένας από τους στόχους λοιπόν της συγκεκριμένης μεταπτυχιακής διατριβής, ο οποίος και τη διαφοροποιεί σε σχέση με τις προηγούμενες έρευνες-μελέτες, είναι το γεγονός ότι θα προσπαθήσει εκτός από την αξιολόγηση της γενικότερης χρησιμότητας των «συχνών στοιχείων» στην παρακολούθηση ενός δικτύου, να αξιολογήσει τις υπάρχουσες τεχνικές οπτικοποίησης παραγόμενων «συχνών στοιχείων». Θα εκτιμηθεί το κατά πόσο η παραγωγή οπτικοποιημένων γραφημάτων αποτελούμενα από «συχνά στοιχεία» μπορεί να βοηθήσει έναν διαχειριστή δικτύων στην καλύτερη-αρτιότερη διαχείριση ενός δικτύου με τη δημιουργία ίσως νέων τρόπων-εργαλείων παρακολούθησης. Το κοινό σημείο του στόχου αυτού (της παρούσας μεταπτυχιακής διατριβής) με τις δουλειές-έρευνες που έχουν γίνει μέχρι τώρα είναι ότι ο απώτερος στόχος είναι η πρόληψη, η προστασία και ο ευκολότερος εντοπισμός ή και γρηγορότερη αντιμετώπιση των κινδύνων που μπορούν να δημιουργήσουν προβλήματα σε ένα δίκτυο υπολογιστών.

Στις ενότητες που ακολουθούν θα δοθούν πειραματικά δεδομένα και αποτελέσματα εφαρμογής της τεχνικής εξόρυξης συχνών στοιχείων σε ένα πραγματικό δίκτυο υπολογιστών.

Κεφάλαιο 3

Πειραματική Εγκατάσταση και Παραγόμενα Δεδομένα.

Στην ενότητα αυτή θα περιγραφεί η υλοποίηση που έχει πραγματοποιηθεί και αφορά την εφαρμογή αλγόριθμου εξόρυξης δεδομένων (παραγωγής «συχνών στοιχείων») σε δεδομένα που προέρχονται από πραγματικό δίκτυο υπολογιστών εταιρείας. Θα περιγραφούν η δομή της εγκατάστασης και τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση αυτής. Θα παρουσιαστούν δεδομένα που έχουν παραχθεί τα οποία και θα χρησιμοποιηθούν σε επόμενη ενότητα για την παραγωγή αποτελεσμάτων-συμπερασμάτων.

3.1 Εισαγωγή

Όπως περιγράφηκε και στα προηγούμενα κεφάλαια της παρούσας μεταπτυχιακής διατριβής η ανάγκη για αδιάκοπη ροή πληροφοριών διαμέσου δικτύων έκανε επιτακτική την ανάγκη για δημιουργία νέων εξελιγμένων τρόπων εξασφάλισης της ομαλής λειτουργίας αυτών. Η παρακολούθηση και ανάλυση της κίνησης ενός δικτύου (network traffic) αποτελεί σημαντικό τμήμα στη γενικότερη διαδικασία που έχει ως στόχο την εξασφάλιση της ομαλής λειτουργίας του. Η χρήση τεχνικών εξόρυξης δεδομένων συγκαταλέγεται μεταξύ των εξειδικευμένων τεχνικών που μπορούν να βοηθήσουν στην παρακολούθηση και κατ' επέκταση στην εξασφάλιση της ορθής λειτουργίας ενός δικτύου υπολογιστών. Συγκεκριμένα τεχνικές δημιουργίας «συχνών στοιχείων» μπορούν να χρησιμοποιηθούν για να γίνει πιο εύκολα αντιληπτή η συνολική εικόνα της κατάστασης που επικρατεί σε ένα δίκτυο υπολογιστών.

Με στόχο την ενίσχυση αυτού του επιχειρήματος, στο παρόν κεφάλαιο της μεταπτυχιακής διατριβής περιγράφεται η υποδομή σε λογισμικό και υλικό που χρησιμοποιήθηκε για την εφαρμογή τεχνικών εξόρυξης «συχνών στοιχείων» σε πραγματικά δεδομένα. Τα δεδομένα αυτά προέρχονται από καταγραφείσα κίνηση η οποία διέρχεται από έναν κόμβο πραγματικού δικτύου υπολογιστών εταιρείας της Βόρειας Ελλάδας. Η εταιρεία αυτή ασχολείται με την παροχή υπηρεσιών σε θέματα υλικού και λογισμικού υπολογιστών και το μέγεθος της είναι αντιπροσωπευτικό του μέσου μεγέθους των εταιρειών πληροφορικής της Βόρειας Ελλάδας.

Το κεφάλαιο αυτό ξεκινά με μία μικρή περιγραφή του υπό παρακολούθηση δικτύου και των υπηρεσιών που αυτό παρέχει καθώς και του τρόπου παροχής ασφάλειας στο εν λόγω δίκτυο. Στη συνέχεια αναλύεται η διαδικασία και τα εργαλεία που χρησιμοποιήθηκαν για τη συλλογή των δεδομένων και την εξαγωγή των «συχνών στοιχείων». Ακολουθεί παράθεση και ανάλυση των αποτελεσμάτων, όπως των «συχνών στοιχείων» που προκύπτουν, του αριθμού των «συχνών στοιχείων» σε σύγκριση με το επιλεχθέν support (στήριξη) του αλγορίθμου που χρησιμοποιήθηκε κλπ.

3.2 Υφιστάμενη Κατάσταση Δικτύου Υπολογιστών Εταιρείας στο Οποίο θα Εφαρμοστεί Αλγόριθμος «Συχνών Στοιχείων».

Πριν προχωρήσουμε στην περιγραφή της υποδομής που χρησιμοποιήθηκε για την παραγωγή των αποτελεσμάτων της μεταπτυχιακής αυτής διατριβής κρίνεται σκόπιμο να γίνει μία σύντομη περιγραφή της υλικοτεχνικής υποδομής δικτύου που χρησιμοποιήθηκε για το σκοπό αυτό με αναφορά στις υπηρεσίες οι οποίες χρησιμοποιούνται καθώς και τον τρόπο παροχής ασφάλειας στο εν λόγω δίκτυο. Η αναφορά αυτή θα προσδιορίσει το μέγεθος της εταιρείας και κατ' επέκταση και το μέγεθος και το είδος της κίνησης που διακινείται καθημερινά διαμέσου αυτού του δικτύου, καθώς και της ύψιστης σημασίας που παίζει το υψηλό επίπεδο ασφάλειας που θα πρέπει να πληρείται και να παρέχεται από τα μέσα που χρησιμοποιούνται για αυτό το σκοπό, είτε αυτά είναι hardware είτε software. Η περιγραφή αυτή θα χρησιμοποιηθεί στην ανάλυση που θα γίνει για τα αποτελέσματα της εργασίας, αλλά θα χρησιμοποιηθεί και στην εκτίμηση των συμπερασμάτων τα οποία θα εξαχθούν στη συνέχεια.

Το δίκτυο της εταιρείας λοιπόν, αποτελείται από δέκα διακομιστές (servers) οι οποίοι καλούνται να καλύψουν τόσο τις ανάγκες του προσωπικού της εταιρείας όσο και τις υπηρεσίες που προσφέρει η εν λόγω εταιρεία. Για την επίτευξη μιας πιο κεντρικής διαχείρισης του δικτύου και ενός πιο ενοποιημένου κεντρικού τρόπου παροχής ασφάλειας και πολιτικών, όλα τα resources του δικτύου ανήκουν σε ένα "Microsoft Windows 2003 Active Directory Domain". Για την υποστήριξη και τη διαχείριση των πόρων του Domain αυτού χρησιμοποιούνται δύο διακομιστές. Ο ένας από αυτούς φέρει τους "Operation Master Roles" του domain ("Primary Domain Controller"), ενώ ο δεύτερος χρησιμοποιείται για τον διαμοιρασμό του φορτίου εργασίας των αιτημάτων πιστοποίησης ταυτότητας των χρηστών ("Secondary Domain Controller"). Και οι δύο αποτελούν τους "Domain Name System" (DNS servers) του δικτύου δεχόμενοι όλα τα αιτήματα για "Name Resolution Process" από τους εσωτερικούς υπολογιστές του δικτύου. Υπάρχει λοιπόν μεγάλος αριθμός πακέτων προς αυτούς τους δύο διακομιστές με destination UDP port number 53. Λόγω του γεγονότος ότι και οι δύο φιλοξενούν την ίδια βάση ονομάτων ανταλλάσσουν αρκετά συχνά μεταξύ τους πακέτα με destination TCP port number 53.

Ο δεύτερος από τους παραπάνω δύο διακομιστές αποτελεί ταυτόχρονα τον “File Server” του δικτύου. Σε αυτόν αποθηκεύονται όλα τα δεδομένα των χρηστών (φυσικά αρχεία) και ο οποίος είναι συνδεδεμένος με συσκευή “tape backup” αποτελώντας παράλληλα και τον “Backup Server” του δικτύου. Το γεγονός αυτό δηλώνει ότι υπάρχουν πολλά πακέτα προς τον συγκεκριμένο διακομιστή με destination TCP port numbers 137 και 139.

Ο τρίτος διακομιστής αποτελεί τον “Mail Server” της εταιρείας. Σε αυτόν συνδέονται όλοι οι υπολογιστές των χρηστών της εταιρείας με τη χρήση των πρωτοκόλλων SMTP για την αποστολή emails (destination TCP port number 25) και με τη χρήση του πρωτοκόλλου POP3 (destination TCP port number 110) για την λήψη emails. Ο συγκεκριμένος διακομιστής (όπως όλοι οι “mail servers”) χρησιμοποιεί το πρωτόκολλο SMTP για την ανταλλαγή emails με εξωτερικούς (εκτός του εσωτερικού δικτύου) “mail servers”. Αποτελεί λοιπόν την αιτία δημιουργίας μεγάλου αριθμού πακέτων που χρησιμοποιούν τις θύρες TCP 25 τα οποία διέρχονται από την προεπιλεγμένη πύλη (Default Gateway) του δικτύου. Στα mailboxes που φιλοξενούνται σε αυτόν μπορούν οι χρήστες να συνδεθούν και εκτός δικτύου με τη χρήση ασφαλούς πρόσβασης μέσω προγράμματος περιήγησης (Web Browser). Η πρόσβαση γίνεται με τη χρήση της θύρας TCP 445.

Ο τέταρτος διακομιστής αποτελεί τον “RDBMS Server” (SQL) της εταιρείας ο οποίος φέρει διάφορες “custom” εφαρμογές πελατών. Σε αυτόν καταλήγουν πακέτα με θύρα προορισμού TCP 1443 τα οποία προέρχονται τόσο από υπολογιστές του εσωτερικού δικτύου όσο και από υπολογιστές εκτός δικτύου.

Ο πέμπτος διακομιστής αποτελεί τον “Web Server” της εταιρείας στον οποίο είναι εγκατεστημένα το “web site” της εταιρείας καθώς και κάποια ακόμη “web sites” συνεργαζόμενων εταιρειών. Σε αυτόν καταλήγουν πακέτα με “destination TCP port number” 80.

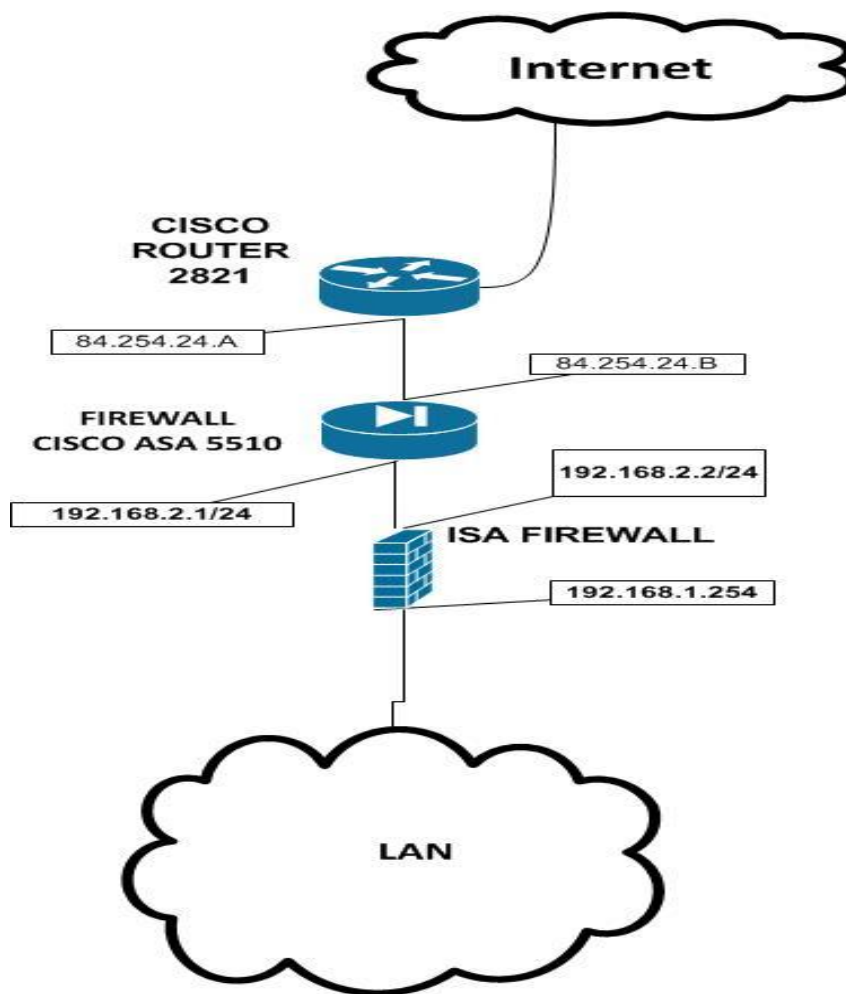
Οι επόμενοι τέσσερις διακομιστές χρησιμοποιούνται για την κάλυψη των αναγκών του ERP που χρησιμοποιεί η εταιρεία (SAP ERP). Οι θύρες που επιτρέπονται μέσω του firewall για την πρόσβαση στο λογισμικό αυτό είναι οι TCP 3200 και 3299.

Ο τελευταίος διακομιστής αποτελεί το “Software Firewall” της εταιρείας το οποίο είναι ένα “content firewall” της εταιρείας Microsoft. Συγκεκριμένα σε αυτόν το διακομιστή έχει εγκατασταθεί το

λογισμικό “Microsoft Internet Security and Acceleration Server 2006” (ISA server). Οι συνδέσεις του συγκεκριμένου διακομιστή είναι δύο. Η μία συνδέει το διακομιστή με το εσωτερικό δίκτυο, ενώ η δεύτερη με ένα “hardware firewall” και κατ’ επέκταση με το internet. Η IP διεύθυνση της εσωτερικής κάρτας δικτύου του εν λόγω διακομιστή, αποτελεί την προεπιλεγμένη πύλη όλων των υπολογιστών του εσωτερικού δικτύου. Η πολιτική ασφαλείας που είναι ρυθμισμένη διαμέσου του λογισμικού του, καθορίζει το επίπεδο ασφαλείας του εσωτερικού δικτύου.

Για την πρόσβαση στο public δίκτυο (internet) χρησιμοποιούνται ένας router εμπορικής επωνυμίας Cisco, μοντέλο 2821 k9 security bundle, καθώς και ένα firewall εμπορικής επωνυμίας Cisco, μοντέλο “ASA 5510”. Τα δύο αυτά στοιχεία είναι συνδεδεμένα σε σειρά (το ένα «πίσω από το άλλο» προς το εσωτερικό δίκτυο) ώστε να παρέχουν μεγαλύτερη ασφάλεια στο δίκτυο και στις υπηρεσίες που χρησιμοποιούνται από τους χρήστες. Στο “ASA firewall” είναι συνδεδεμένο το “application level firewall” (η εξωτερική σύνδεση του ISA server), παρέχοντας ένα ακόμη επίπεδο στην ασφάλεια των διακινούμενων πληροφοριών μεταξύ internet και εσωτερικού δικτύου.

Οι hosts του εσωτερικού δικτύου (διακομιστές, τερματικά, εκτυπωτές κλπ) χρησιμοποιούν “private” IP διευθύνσεις με network ID το 192.168.1.0/24. Η εξωτερική IP του ISA server με το εσωτερικό interface του ASA firewall συνθέτουν ένα ξεχωριστό subnet. Παρόμοια το ίδιο γίνεται με το εξωτερικό interface του ASAS firewall και το εσωτερικό Interface του router. Η συνολική δομή του δικτύου εμφανίζεται πιο παραστατικά στο επόμενο σχήμα (σχήμα 3.1). Στο σχήμα αυτό έχει αντικατασταθεί το τελευταίο octet των public IP διευθύνσεων του router και του hardware firewall με τα γράμματα Α και Β αντίστοιχα με στόχο την απόκρυψη των χρησιμοποιούμενων από την εταιρεία διευθύνσεων (λόγοι προστασίας προσωπικών δεδομένων).



Σχήμα 3.1: Δομή συνδέσεων δικτύου

Στο εν λόγω δίκτυο οι διαχειριστές δικτύου χρησιμοποιούν εργαλεία για τη διαχείριση του ακόμη και από μακριά επιτρέποντας σε πακέτα τέτοιου είδους να διέρχονται προς το εσωτερικό του δικτύου.

Από τα παραπάνω γίνεται αντιληπτό ότι λόγω των παρεχόμενων υπηρεσιών, οι περισσότερες εκ των οποίων είναι προσβάσιμες και από το internet από χρήστες ή συνεργάτες της εταιρείας, υπάρχει μεγάλος αριθμός πακέτων που διέρχονται από τη σύνδεση με το Internet του εν λόγω δικτύου. Τα πακέτα αυτά μπορεί να είναι πακέτα προς ή και από το “mail server”, πακέτα προς το “web site” της εταιρείας, προς τις εφαρμογές SQL και το λογισμικό SAP που υπάρχει στο εσωτερικό του δικτύου, καθώς και πακέτα διαχείρισης του δικτύου από μακριά. Πολιτικές ασφαλείας εφαρμόζονται τόσο στα ενδιάμεσα στοιχεία (router, hardware firewall, software firewall) όσο και στους διακομιστές και υπολογιστές των χρηστών με στόχο να αποτρέπονται εκείνες οι ενέργειες οι οποίες θεωρούνται βλαπτικές για την ορθή και αδιάκοπη λειτουργία του δικτύου. Ο αριθμός των διακινούμενων

διαμέσου της σύνδεσης στο Internet πακέτων είναι μεγάλος λόγω του μεγάλου αριθμού χρηστών και της φύσης των παρεχόμενων υπηρεσιών που είναι προσβάσιμες από συνεργαζόμενες εταιρείες και πελάτες της εταιρείας. Ο αριθμός αυτός γίνεται ακόμη μεγαλύτερος λόγω του γεγονότος ότι οι χρήστες του δικτύου συχνά δουλεύουν από διαφορετική γεωγραφική περιοχή από αυτή στην οποία βρίσκεται το δίκτυο. Ο μεγάλος αυτός αριθμός πακέτων καθιστά δύσκολη την καταγραφή και παρακολούθηση της κίνησης που διέρχεται από την σύνδεση με το παγκόσμιο δίκτυο του εν λόγω δικτύου. Γίνεται λοιπόν αισθητή η ανάγκη ενός ειδικευμένου τρόπου παρακολούθησης και εντοπισμού, είτε προβλημάτων, είτε απειλών, είτε γενικότερα της κατάστασης και της γενικής εικόνας των διακινούμενων πληροφοριών.

3.3 Σημείο Συλλογής Δεδομένων Κίνησης – Είδος Δεδομένων (Ροές Δεδομένων-Netflows).

Συνδυάζοντας την αυξημένη ανάγκη παρακολούθησης της κατάστασης του δικτύου της εν λόγω εταιρείας και τους επιθυμητούς στόχους της παρούσας μεταπτυχιακής διατριβής, αποφασίστηκε η εφαρμογή αλγορίθμων εξόρυξης «συχνών στοιχείων» σε καταγραφείσα κίνηση. Πριν την προαναφερθείσα υλοποίηση κρίθηκε αναγκαία η λήψη δύο σημαντικών αποφάσεων. Η πρώτη από αυτές αφορά την επιλογή του σημείου (σύνδεση ή host) από το οποίο θα συλλαμβάνονται δεδομένα (capture data), ενώ η δεύτερη αφορά το είδος των δεδομένων που θα πρέπει να συλλέγονται.

Για την εφαρμογή αλγορίθμου εξόρυξης «συχνών στοιχείων» χρειάζεται μία βάση με τα κατάλληλα δεδομένα. Όπως αναλύθηκε και στο δεύτερο κεφάλαιο της παρούσας μεταπτυχιακής διατριβής το πρώτο βήμα στη διαδικασία εξόρυξης δεδομένων είναι η επιλογή των κατάλληλων δεδομένων. Το είδος των δεδομένων που θα χρησιμοποιηθούν θα πρέπει να ταιριάζουν με τον επιθυμητό στόχο και το είδος συμπερασμάτων που αναμένεται από τη διαδικασία εξόρυξης. Ένας παράγοντας ο οποίος επηρεάζει τα αναμενόμενα συμπεράσματα είναι και το σημείο από το οποίο θα συλλέγονται τα δεδομένα. Το σημείο αυτό θα μπορούσε να είναι είτε ένας συγκεκριμένος host (διακομιστής) του δικτύου ή κάποια σύνδεση αυτού. Μπορεί δηλαδή να συλλέγονται δεδομένα από τα Log Files κάποιου από τους διακομιστές της εταιρείας και να εφαρμόζονται σε αυτά αλγόριθμοι παραγωγής «συχνών στοιχείων» για την εξαγωγή και ανάλυση συμπερασμάτων που μπορεί να προκύψουν από αυτά. Ή θα μπορούσε να καταγράφονται τα πακέτα που φτάνουν ή φεύγουν από ένα συγκεκριμένο

host. Αντίστοιχα αν το σημείο συλλογής δεδομένων είναι κάποια σύνδεση (link) του δικτύου θα μπορούσε να συλλέγονται και να επεξεργάζονται πακέτα που διέρχονται από αυτή.

Η επιλογή ενός συγκεκριμένου host λοιπόν θα γινόταν στην περίπτωση που ενδιέφερε να αναλυθεί η «συμπεριφορά» μίας ή κάποιων συγκεκριμένων υπηρεσιών του δικτύου. Υπηρεσίες που παρέχονται από τον host από τον οποίο θα συλλέγονται δεδομένα. Πχ. Θα μπορούσε να γίνει συλλογή πακέτων που φτάνουν ή φεύγουν από τον mail server του δικτύου. Μία τέτοια κατάσταση θα μπορούσε να αποκαλύψει χαρακτηριστικά όπως:

- το είδος της κίνησης που καλείται να διαχειριστή ο συγκεκριμένος διακομιστής,
- το είδος των επιθέσεων που αυτός δέχεται,
- τις περιόδους υψηλής κίνησης και την μεταβολή τους με το χρόνο,
- το εύρος των υπολογιστών που συνδέονται σε αυτόν, κλπ

Τα όρια που θέτει μία τέτοιου είδους καταγραφή είναι εμφανή. Η ανάλυση και τα συμπεράσματα που μπορούν να προκύψουν θα περιορίζονται σε συγκεκριμένες υπηρεσίες που παρέχονται από το δίκτυο. Τα συμπεράσματα που θα προκύπτουν σε καμία περίπτωση δεν θα αντανakλούν τη συνολική εικόνα του δικτύου και δεν θα αναλύεται η συνολική κίνηση που διέρχεται και διακινείται από το σύνολο των διακομιστών αυτού (γεγονός που αποτελεί και την κοντινότερη με τους στόχους της συγκεκριμένης μεταπτυχιακής διατριβής προσέγγιση).

Η επιλογή μίας σύνδεσης του συγκεκριμένου δικτύου και ειδικότερα μιας κεντρικής σύνδεσης αυτού ως σημείο για την άντληση και καταγραφή διερχόμενων δεδομένων, θα μπορούσε να χρησιμοποιηθεί για την εξαγωγή συμπερασμάτων που αφορούν γενικότερα το δίκτυο και την κίνηση που διέρχεται από αυτό (και όχι μία συγκεκριμένη υπηρεσία ή διακομιστή αυτού). Θα μπορούσαν δηλαδή να αναλυθούν το είδος της κίνησης, το είδος των επικίνδυνων δεδομένων, οι περίοδοι υψηλής χρήσης, για το σύνολο πλέον του δικτύου.

Με βάση τα παραπάνω και δεδομένου ότι ενδιαφέρει περισσότερο η εξαγωγή συμπερασμάτων που θα αφορούσαν τη γενικότερη εικόνα του δικτύου, αποφασίστηκε η καταγραφή των δεδομένων που διέρχονται από την κεντρική σύνδεση του δικτύου με το internet. Συγκεκριμένα αποφασίστηκε η

καταγραφή των δεδομένων-κίνησης που διέρχεται από τον router που χρησιμοποιεί η εταιρεία για την σύνδεση του τοπικού δικτύου με το παγκόσμιο δίκτυο.

Η δεύτερη απόφαση που έπρεπε να ληφθεί είναι το είδος των δεδομένων που θα πρέπει να καταγράφονται. Από τον συγκεκριμένο router διέρχεται όλη η εισερχόμενη και η εξερχόμενη κίνηση του δικτύου, τεράστιος δηλαδή αριθμός πακέτων. Θα μπορούσε να καταγράφονται με τη βοήθεια κάποιου εργαλείου sniffer κάθε ξεχωριστό πακέτο ή αντίστοιχα οι ροές δεδομένων (network flow data) που διέρχονται από τον εν λόγω router. Η καταγραφή όλων των πακέτων σημαίνει ότι ο αριθμός των καταγραφόμενων πληροφοριών θα είναι πολύ μεγάλος. Συνέπεια αυτού θα είναι η βάση δεδομένων που θα δημιουργηθεί (ή καλύτερα θα δημιουργείται κάθε φορά) και η οποία θα χρησιμοποιείται για τη δημιουργία «συχνών στοιχείων» θα είναι πολύ μεγάλη. Ο όγκος των προς επεξεργασία στοιχείων θα είναι σχετικά μεγάλος χωρίς να κερδίζουμε κάτι παραπάνω σε γνώση. Θα περιέχονται σε αυτή πληροφορίες που ίσως να μην προσφέρουν χρήσιμη πληροφορία (πχ κάποια flags των headers των πακέτων) κάνοντας πιο αργή τη διαδικασία δημιουργίας αποτελεσμάτων («συχνά στοιχεία»).

Για την αποφυγή των παραπάνω μειονεκτημάτων αποφασίστηκε η χρησιμοποίηση τεχνικής καταγραφής του συνόλου των πακέτων που συνθέτουν κάθε επικοινωνία (ροή δεδομένων) και όχι του κάθε πακέτου ξεχωριστά. Ως ροή δεδομένων ορίζεται μία ακολουθία πακέτων τα οποία μοιράζονται ένα σύνολο κοινών χαρακτηριστικών όπως:

- Source και Destination IP addresses
- Source και Destination TCP ή UDP port numbers (τοποθετείται το 0 σε περίπτωση που χρησιμοποιείται διαφορετικό πρωτόκολλο)
- IP πρωτόκολλο
- Τύπος της υπηρεσίας (type of service)
- TCP session

Μία τέτοια προσέγγιση μπορεί να μη παρέχει αναλυτικές λεπτομέρειες για κάθε πακέτο περιέχει όμως τα στοιχεία που απαιτούνται για την αναπαράσταση του είδους κίνησης. Προσδίδει όλες τις απαραίτητες πληροφορίες για την κατανόηση της κίνησης και μπορεί να χρησιμοποιηθεί για την ανίχνευση ανωμαλιών και την ενεργοποίηση συναγερμών (alerts).

Αποφασίστηκε συγκεκριμένα η χρήση της τεχνολογίας NetFlow. Η τεχνολογία αυτή αναπτύχθηκε από την εταιρεία Cisco με στόχο τη συλλογή πληροφοριών κίνησης. Με τον καιρό σύμφωνα με άρθρο της Wikipedia [59] έγινε πρότυπο παρακολούθησης κίνησης και υιοθετήθηκε από τις περισσότερες εταιρίες που ασχολούνται με την δικτύωση υπολογιστών (κατασκευάστριες εταιρείες routers, switches κλπ).

Το πρωτόκολλο αυτό δημιουργεί εγγραφές ροών δεδομένων (neflow records) οι οποίες εξάγονται από την συσκευή (πχ τον router) με τη χρήση του UDP πρωτοκόλλου. Αυτές μπορούν να συλλεχθούν από κατάλληλες συσκευές ή εφαρμογές που καλούνται συλλέκτες ροών (netflow collectors). Εφόσον αποφασιστεί από τη συσκευή (router) ότι η εγγραφή (flow record) έχει ολοκληρωθεί (αυτό μπορεί να καθοριστεί από διάφορους παράγοντες όπως το τέλος του session, η διαδικασία flow aging ή και το πέρας ενός προκαθορισμένου χρονικού διαστήματος) τότε εξάγεται η εγγραφή. Μία ροή μπορεί μεταξύ άλλων να περιέχει:

- Αύξον αριθμό και ώρα έναρξης και λήξης της ροής
- Source και Destination IP addresses
- Source και Destination port numbers
- Αριθμός περιεχόμενων πακέτων καθώς και αριθμό bytes
- Διάφορα flags, κλπ.

Στο σχήμα 3.2 ακολουθεί ένα παράδειγμα μίας καταγραφείσας ροής δεδομένων:

```
2011-03-15 22:04:33.618 14.421 TCP 84.254.24.51:11554 -> 88.218.118.221:80 .A..SF 135 7650
```

Σχήμα 3.2: Παράδειγμα ροής δεδομένων

Σε καμία περίπτωση οι ροές δεδομένων δεν περιέχουν δεδομένα χρηστών παρά μόνο πληροφορίες από τις κεφαλίδες των πακέτων.

Για τη συλλογή λοιπόν των απαιτούμενων πληροφοριών κίνησης επιλέχθηκε η χρήση της σύνδεσης της εταιρείας στο δημόσιο δίκτυο με χρήση για το σκοπό αυτό του router της εταιρείας. Το είδος των πληροφοριών που θα συλλέγονται θα είναι ροές δεδομένων, αφού παρέχουν τις απαιτούμενες

πληροφορίες για την πραγματοποίηση των στόχων της εν λόγω καταγραφής. Σε επόμενη παράγραφο θα περιγραφεί η παραμετροποίηση του router για την επίτευξη της εν λόγω διαδικασίας.

3.4 Απομόνωση Δικτύου Παρακολούθησης

Όπως αναφέρθηκε και στην προηγούμενη παράγραφο με στόχο τη δημιουργία ενός συστήματος παρακολούθησης και σύνοψης της εισερχόμενης ή εξερχόμενης κίνησης του υπό παρακολούθηση δικτύου και της εύκολης αναγνώρισης περιπτώσεων-προσπαθειών παραβίασης των κανόνων και πολιτικών ασφαλείας αυτού, αποφασίστηκε η παρακολούθηση των ροών δεδομένων που διέρχονται από αυτό. Κάθε σύστημα παρακολούθησης κίνησης μπορεί να βοηθά στην καλύτερη διαχείριση και έλεγχο της κατάστασης ενός δικτύου αλλά ταυτόχρονα έχει και κάποια αρνητικά σημεία. Σημαντικά μειονεκτήματα μπορεί να είναι:

- η επιβάρυνση με πρόσθετο φορτίο εργασίας των υπό παρακολούθηση συσκευών,
- η πρόσθετη κίνηση με την οποία επιβαρύνεται το δίκτυο,
- η ασφάλεια των υπό καταγραφή δεδομένων.

Το πρόσθετο φορτίο εργασίας προέρχεται από την παραμετροποίηση της συσκευής ώστε να μπορεί να συλλαμβάνει και να αποστέλλει πακέτα. Η πρόσθετη κίνηση με την οποία επιβαρύνεται το δίκτυο προέρχεται από τη διαδικασία αποστολής των καταγραφέντων πληροφοριών στη συσκευή η οποία θα τις συλλέξει με σκοπό την ανάλυση τους. Συνήθως οι πληροφορίες (raw data) που συλλέγονται από κάποια συσκευή ή σύνδεση του δικτύου με τη χρήση κατάλληλου εργαλείου, αποστέλλονται διαμέσου του ίδιου δικτύου σε άλλη συσκευή στην οποία θα γίνει η τελική επεξεργασία. Το γεγονός αυτό επιβαρύνει με πρόσθετη κίνηση το δίκτυο, η οποία ίσως να καταγράφεται και αυτή κατά τη μεταφορά μαζί με την υπόλοιπη κίνηση. Τέλος το θέμα της ασφάλειας των καταγραφέντων δεδομένων είναι σημαντικό διότι σε αυτές καταγράφονται σημαντικές πληροφορίες με χαρακτηριστικά του δικτύου (πχ IP addresses συσκευών ή και υπηρεσιών).

Για την αντιμετώπιση και εξάλειψη των παραπάνω μειονεκτημάτων είναι κοινή πολιτική όλων των συστημάτων παρακολούθησης να χρησιμοποιούν ανεξάρτητη δομή από το υπόλοιπο δίκτυο. Τα περισσότερα συστήματα παρακολούθησης απαιτούν την ύπαρξη ανεξάρτητου δικτύου από αυτό που παρακολουθείται ώστε να πετυχαίνεται ελάχιστη επιβάρυνση του εύρους ζώνης (bandwidth)

του υπό παρακολούθηση δικτύου με ταυτόχρονη επίτευξη ασφάλειας των καταγραφέντων πληροφοριών. Έτσι λοιπόν αποφασίστηκε κατά τη σχεδίαση του συστήματος παρακολούθησης, η ανεξαρτητοποίηση αυτού από το υπόλοιπο δίκτυο ώστε η κίνηση του συστήματος παρακολούθησης να μην ακολουθεί την κίνηση του υπόλοιπου δικτύου.

Η συσκευή πάνω στην οποία καταλήγει η σύνδεση από την οποία αποφασίστηκε να καταγραφούν δεδομένα κίνησης, όπως αναφέρθηκε και προηγουμένως, είναι ο router της εταιρείας. Η συγκεκριμένη συσκευή διαθέτει δύο Ethernet Interfaces (και ένα τρίτο το οποίο αποτελεί το Management interface του) το ένα από τα οποία χρησιμοποιείται, όπως φαίνεται και στο σχήμα 3.1, για τη σύνδεση με το εσωτερικό δίκτυο. Από αυτή τη σύνδεση καταγράφηκαν δεδομένα. Για την ανεξαρτητοποίηση του δικτύου συλλογής δεδομένων από αυτό της παρακολούθησης (με βάση τα προηγούμενα) αποφασίστηκε να δημιουργηθεί ένα ανεξάρτητο υποδίκτυο με network id 192.168.70.0 / 24 (το υπόλοιπο δίκτυο χρησιμοποιεί ως network id το 192.168.1.0 /24) χρησιμοποιώντας το δεύτερο Ethernet interface του router.

Για την αποστολή των δεδομένων παρακολούθησης σε μία συσκευή-λογισμικό συγκέντρωσης και ανάλυσης αυτών θα έπρεπε να επιλέγει μία συγκεκριμένη θύρα η οποία θα χρησιμοποιούνταν από την υπηρεσία συλλογής. Θα έπρεπε λοιπόν να ανοιχτεί η συγκεκριμένη θύρα για αυτό το σκοπό τόσο στο ASA firewall όσο και στον ISA server, ώστε να είναι δυνατή η πρόσβαση διαμέσου αυτής της θύρας στον διακομιστή-λογισμικό που θα αποθηκεύει τα δεδομένα και θα τα επεξεργάζεται. Το γεγονός αυτό θα δημιουργούσε μία πρόσθετη είσοδο για κακόβουλες επιθέσεις, καταργώντας την πολιτική ασφαλείας του εν λόγω δικτύου. Η ανάγκη παραμετροποίησης λοιπόν ενός ανεξάρτητου interface του router (διαφορετικό από αυτό με το οποίο ο router συνδέεται με το υπόλοιπο δίκτυο) ενισχύθηκε και από αυτό το γεγονός.

Το συγκεκριμένο λοιπόν interface συνδέθηκε με το κεντρικό switch της εταιρείας (παρακάμπτοντας τα δύο firewalls) σε μία συγκεκριμένη πόρτα αυτού η οποία παραμετροποιήθηκε σε διαφορετικό vlan από αυτό του υπολοίπου δικτύου για λόγους ασφαλείας και απομόνωσης των δεδομένων. Στο ίδιο vlan στη συνέχεια συνδέθηκε και ο διακομιστής στον οποίο αποθηκεύονταν οι πληροφορίες που συλλέγονταν από το router. Με αυτό τον τρόπο δημιουργήθηκε ένα ανεξάρτητο δίκτυο το οποίο δεν έχει συνδεσιμότητα με το υπόλοιπο δίκτυο, εξασφαλίζοντας έτσι την ακεραιότητα και ασφάλεια του δικτύου της εταιρείας καθώς και των δεδομένων παρατήρησης.

Τα δεδομένα λοιπόν που συλλέγονται από τον router προέρχονται από την κίνηση που διέρχεται από το interface 1 προς το WAN Interface του ενώ στέλνονται από το interface 2 στον διακομιστή με το κατάλληλο λογισμικό για την αποθήκευση και επεξεργασία αυτών.

3.5 Παραμετροποίηση Συσσκευής Παρακολούθησης (Router)

Για την καταγραφή λοιπόν των ροών δεδομένων της σύνδεσης που συνδέει το δίκτυο της εταιρείας με το παγκόσμιο δίκτυο παραμετροποιήθηκε κατάλληλα ο router εμπορικής επωνυμίας Cisco.

Λόγω του γεγονότος ότι ο Router δεν έχει προεπιλεγμένα, ενεργοποιημένη τη λειτουργία Netflow το πρώτο βήμα μεταξύ των ρυθμίσεων που έγιναν στο router είναι η ενεργοποίηση της τεχνολογίας NetFlow για το interface από το οποίο διέρχεται η κίνηση που πρέπει να καταγραφεί. Για την ενεργοποίηση αυτή στο IOS που χρησιμοποιεί ο router υπάρχει η εντολή:

IP route-cache flow

η οποία σύμφωνα με το εγχειρίδιο χρήσης του συγκεκριμένου IOS “Enables NetFlow switching for IP routing”. Οι ακόλουθες εντολές:

Ip flow ingress

Ip flow egress

ενεργοποιούν για το interface που επιλέχθηκε (Gigabit Ethernet 0/0 το οποίο είναι αυτό με το οποίο ο router επικοινωνεί με το υπόλοιπο δίκτυο της εταιρείας) την τεχνολογία NetFlow.

Πληκτρολογήθηκαν λοιπόν οι εξής εντολές με την ακόλουθη σειρά:

configure terminal	Εντολή για την εισαγωγή σε global configuration mode του router.
interface G0/0	Εντολή για την επιλογή του κατάλληλου interface στο οποίο θα ισχύσουν οι ακόλουθες εντολές. (Gigabit Ethernet 0/0)
ip route-cache flow	Ενεργοποίηση του NetFlow switching για routing
ip flow ingress	Ενεργοποίηση του NetFlow για το interface Gigabit Ethernet 0/0
ip flow egress	Ενεργοποίηση του egress NetFlow για το interface Gigabit Ethernet 0/0

Πίνακας 3.1: Ενεργοποίηση Netflow

Ο αριθμός των καταχωρήσεων NetFlows που μπορούν να διατηρηθούν στην cache του router μπορεί να παραμετροποιηθεί για τον καλύτερο έλεγχο των ρυθμών της κίνησης NetFlow. Η προεπιλεγμένη ρύθμιση είναι να κρατούνται 64K καταχωρήσεις flows στην cache. Κάθε καταχώρηση είναι περίπου 64 bytes χώρος [60]. Για να ισχύει η προεπιλεγμένη ρύθμιση στον router για τον αριθμό των NetFlow καταχωρήσεων πληκτρολογήθηκε η εντολή:

no ip flow-cache entries

Όπως αναφέρθηκε και στην προηγούμενη παράγραφο για την απαλλαγή προβλημάτων εύρους ζώνης και ταυτόχρονα ασφάλειας του δικτύου και των μεταφερόμενων δεδομένων αποφασίστηκε

η αποστολή των καταγραφέντων δεδομένων από διαφορετικό Interface από αυτό από το οποίο καταγράφεται η κίνηση. Για τον καθορισμό αυτού του interface υπάρχει η εντολή:

ip flow-export source <interface>

Για τη δήλωση του προορισμού αυτών των πακέτων καθώς και της θύρας προορισμού, στο IOS του router υπάρχει η εντολή:

Ip flow-export destination <IP address> <destination port number>

Για τον καθορισμό της έκδοσης NetFlow των εξαγόμενων δεδομένων υπάρχει η εντολή:

Ip flow-export version <version number>

Με βάση τα παραπάνω πληκτρολογήθηκαν λοιπόν οι ακόλουθες εντολές:

Ip flow-export source gigabitEthernet 0/1	Εντολή για τον καθορισμό του Interface από το οποίο θα αποστέλλονται τα δεδομένα (interface Gigabit Ethernet 0/1 του router)
Ip flow-export version 5	Εντολή για τον καθορισμό των εξαγόμενων δεδομένων
Ip flow-export destination 192.168.70.202 60001	Προορισμός θα είναι ο διακομιστής με IP address 192.168.70.202 με θύρα προορισμού της 60001

Πίνακας 3.2: Λεπτομέρειες παραμετροποίησης NetFlow

Με το πέρας πληκτρολόγησης και αυτών των εντολών ολοκληρώθηκε η παραμετροποίηση του router ο οποίος πλέον θα καταγράφει τα δεδομένα που εισέρχονται ή εξέρχονται από το GE interface

0/0 αυτού και θα αποστέλλονται από το interface GE 0/1 με προορισμό την IP 192.168.70.202 και θύρα την 60001.

3.6 Εργαλείο Συλλογής Δεδομένων

Για την συλλογή, αποθήκευση και επεξεργασία των δεδομένων που στέλνονται από τον router έπρεπε να επιλεγεί κάποιο λογισμικό. Έπειτα από έρευνα που έγινε στο internet για την εύρεση εργαλείων που να ανήκουν σε αυτή την κατηγορία μεγάλο ενδιαφέρον είχε το εργαλείο ανοιχτού κώδικα **nfdump**. Το εργαλείο αυτό αποτελεί εργαλείο NetFlow Analyzer υποστηρίζοντας την έκδοση 5 που ρυθμίστηκε στον router (πίνακας 2). Δίνει τη δυνατότητα συλλογής και αποθήκευσης ροών δεδομένων.

Ένα από τα χαρακτηριστικά του είναι η δυνατότητα που δίνει για τον καθορισμό του χρονικού διαστήματος κατά το οποίο θα ομαδοποιούνται σε αρχεία οι πληροφορίες (ροές δεδομένων). Λαμβάνει τις απεσταλμένες πληροφορίες από την «πηγή» συλλογής και δημιουργεί αρχεία με τις πληροφορίες αυτές σε αποθηκευτικό χώρο που θα καθοριστεί για αυτό το σκοπό. Τα αρχεία αυτά δημιουργούνται ανά τακτά χρονικά διαστήματα με βάση το καθορισμένο χρονικό διάστημα. Τα αρχεία αυτά αποθηκεύονται με τη μορφή δυαδικών αρχείων και στη συνέχεια μπορούν να αναλυθούν, πετυχαίνοντας έτσι το διαχωρισμό των διαδικασιών της αποθήκευσης και ανάλυσης των δεδομένων. Δίνει ακόμη τη δυνατότητα φιλτραρίσματος των εισερχόμενων δεδομένων πετυχαίνοντας απομόνωση και επιλογή συγκεκριμένων ροών.

Το συγκεκριμένο εργαλείο δε δίνει τη δυνατότητα εγκατάστασής του σε συστήματα με λειτουργικό σύστημα MS Windows. Για το λόγο αυτό αποφασίστηκε η χρησιμοποίηση ενός Linux συστήματος για την εγκατάσταση του και συγκεκριμένα της τελευταίας έκδοσης (την περίοδο της διπλωματικής διατριβής) Ubuntu server. Όλοι οι διακομιστές της εταιρείας το δίκτυο της οποίας χρησιμοποιείται για την πραγματοποίηση της μεταπτυχιακής διατριβής έχουν εγκατεστημένο λειτουργικό σύστημα (όλοι MS Windows Server) και φυσικά δεν υπήρχε η δυνατότητα επανεγκατάστασης ή αποκλειστικής χρήσης κάποιου από αυτούς. Για την αντιμετώπιση του προβλήματος αυτού αποφασίστηκε η χρησιμοποίηση της τεχνολογίας εικονικών μηχανών (virtualization). Εγκαταστάθηκε σε ένα από τους διακομιστές το λογισμικό VMWare Server το οποίο είναι freeware και μπορεί να εγκατασταθεί χωρίς να απαιτείται η προμήθεια αδειών. Χρησιμοποιώντας το

λογισμικό αυτό δημιουργήθηκε εικονικό μηχάνημα (virtual machine) και εγκαταστάθηκε σε αυτό λειτουργικό σύστημα Ubuntu Server. Σε αυτό το λειτουργικό σύστημα εγκαταστάθηκε το εργαλείο nfdump.

Για την συλλογή των πληροφοριών (ροές δεδομένων) που στέλνονται από τον cisco router και την αποθήκευση αυτών σε αρχεία δυαδικής μορφής χρησιμοποιήθηκε η εντολή nfcapd (netflow capture daemon) του εργαλείου nfdump. Συγκεκριμένα χρησιμοποιήθηκε η ακόλουθη εντολή για την παραμετροποίηση του nfcapd:

Nfcapd -w -D -l ~/flows/router1 -p 60001

Με βάση την εντολή αυτή δημιουργούνται αρχεία δυαδικών δεδομένων κάθε πέντε λεπτά στον φακέλο που δημιουργήθηκε στο εικονικό μηχάνημα για αυτό το σκοπό (~/flows/router1). Το όνομα των αρχείων που δημιουργούνται ακολουθεί τη μορφή nfcapd.YYYYMMddhhmm πχ nfcapd.201104110815. Όπως φαίνεται από την παραπάνω εντολή η θύρα στην οποία «ακούει» το εργαλείο nfdump για τις εισερχόμενες από τον router ροές δεδομένων είναι η TCP/60001 (είναι η ίδια με αυτή που καθορίστηκε στον router).

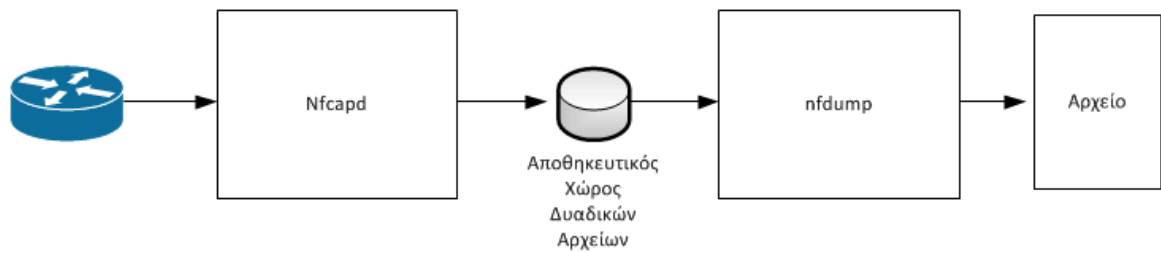
Για την ανάλυση των δεδομένων χρειάζεται να ομαδοποιηθούν τα δυαδικά αρχεία που δημιουργούνται από τη χρήση του nfcapd και να δημιουργηθούν αρχεία κειμένου. Για τη δημιουργία αυτών των αρχείων κειμένου (txt) χρησιμοποιήθηκε η ακόλουθη εντολή:

Nfdump -R ./flows/router1 >name.txt

Σύμφωνα με τη συγκεκριμένη εντολή δημιουργούνται αρχεία κειμένου με όνομα name.txt από το σύνολο (διακόπτης -R) των δυαδικών αρχείων του φακέλου «./flows/router1». Φυσικά και μπορεί να επιλεγεί συγκεκριμένος αριθμός δυαδικών αρχείων ώστε να απομονωθούν ροές δεδομένων συγκεκριμένων πχ χρονικών διαστημάτων (πχ μπορεί να χρειάζεται να αναλυθεί η κίνηση μίας μέρας ή κάποιων συγκεκριμένων ωρών κάποιας ημέρας). Σε αυτή την περίπτωση η σύνταξη της εντολής αλλάζει σε:

Nfdump -R ./flows/router1/first_file:last_file >name.txt

Στο παρακάτω σχήμα (σχήμα 3.3) δίνεται μία γραφική αναπαράσταση των παραπάνω περιγραφέντων διαδικασιών.



Σχήμα 3.3: Διαδικασίες nfdump

3.7 Υλοποίηση και Εργαλεία για την Εξαγωγή Συχνών Στοιχείων (Λογισμικό R)

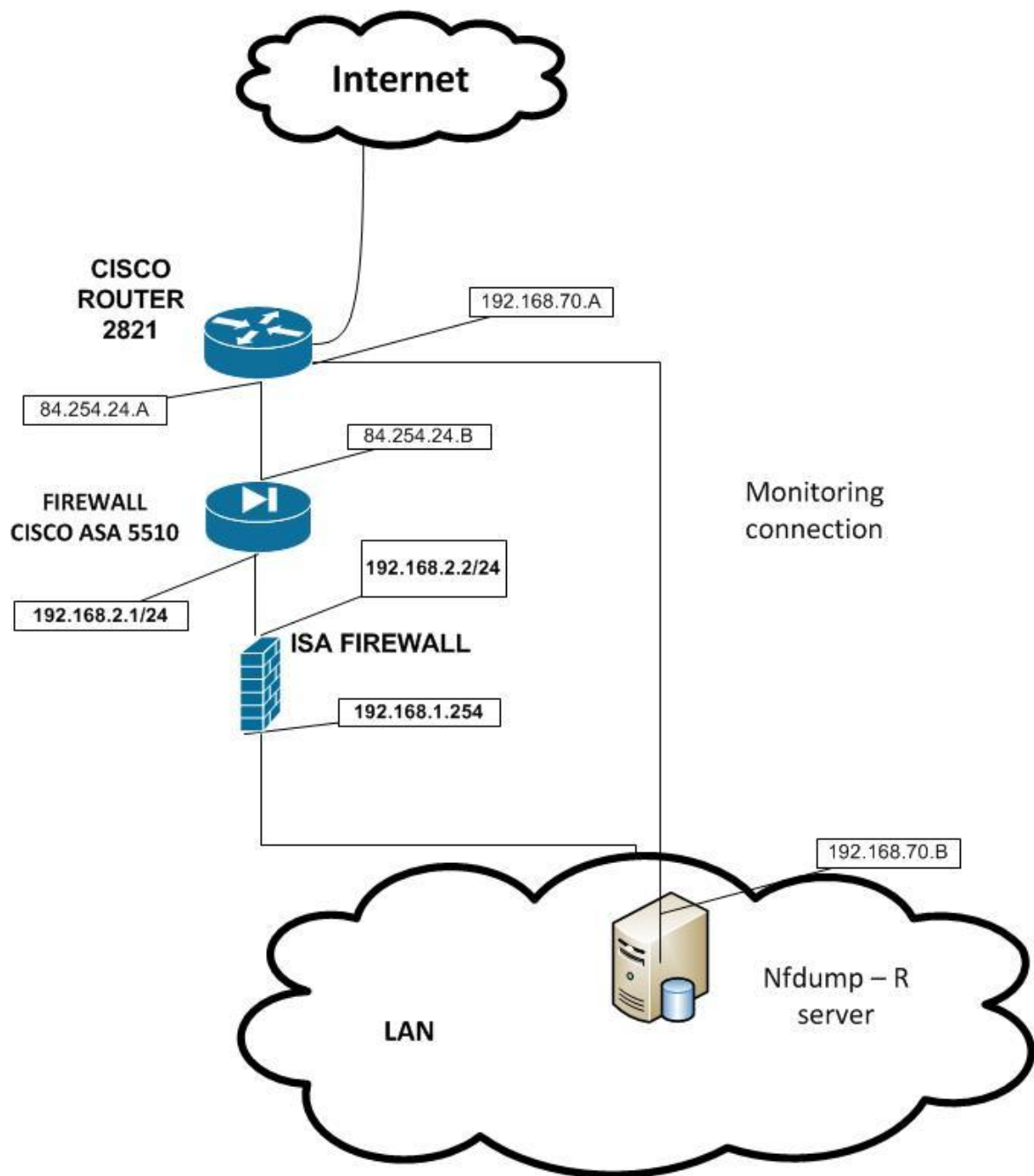
Όπως αναφέρθηκε μέχρι τώρα στόχος της παρούσης εργασίας είναι να παραχθούν από δεδομένα καταγραφείσας κίνησης δικτύου υπολογιστών «συχνά στοιχεία» να αναπαρασταθούν γραφικά ώστε να εκτιμηθεί η χρησιμότητα αυτών στην παρακολούθηση και διαχείριση των δικτύων υπολογιστών. Εκτός από τα «συχνά στοιχεία» (frequent itemsets) θα πρέπει να παραχθούν και maximal frequent itemsets καθώς και closed frequent itemsets και να αναπαρασταθούν γραφικά ώστε να εκτιμηθεί και η τυχόν χρησιμότητα αυτών. Όπως είδαμε στο προηγούμενο κεφάλαιο ο αλγόριθμος ο οποίος μπορεί να χρησιμοποιηθεί για την παραγωγή «συχνών στοιχείων», maximal frequent itemsets και closed frequent itemsets είναι ο αλγόριθμος Apriori.

Θα μπορούσε να δημιουργηθεί κατάλληλο πρόγραμμα χρησιμοποιώντας γλώσσα προγραμματισμού όπως η C για την είσοδο αρχείων και παραγωγή συχνών στοιχείων. Υπάρχουν όμως διαθέσιμες πολλές εφαρμογές στο Internet για την προσομοίωση αυτής της διαδικασίας. Μεταξύ αυτών των εργαλείων είναι και το στατιστικό πακέτο ανοιχτού κώδικα, το “R”. Το πλεονέκτημα αυτού του συγκεκριμένου λογισμικού είναι το γεγονός ότι διαθέτει έτοιμο «πακέτο» για την οπτικοποίηση (visualization) των αποτελεσμάτων, καλύπτοντας μάλιστα ένα μεγάλο αριθμό τρόπων γραφικής αναπαράστασης των δεδομένων, από αυτούς που βρέθηκαν στη βιβλιογραφία. Τα δεδομένα που δημιουργούνται από το λογισμικό nfdump (αρχεία κειμένου) μπορούν να

εισαχθούν στο «R» και με βάση τον αλγόριθμό Apriori να παραχθούν τα «συχνά στοιχεία». Αυτά μέσω του arulesViz library του πακέτου “R” μπορούν να απεικονιστούν γραφικά [61].

Αποφασίστηκε λοιπόν η χρησιμοποίηση του λογιστικού πακέτου “R” ως το λογισμικό ανάλυσης των δεδομένων τη δημιουργία δηλαδή των «συχνών στοιχείων». Το λογισμικό αυτό μπορεί να εγκατασταθεί σε λειτουργικό σύστημα MS Windows, οπότε και αποφασίστηκε η δημιουργία ενός ακόμη εικονικού μηχανήματος με Microsoft Windows XP στο οποίο και εγκαταστάθηκε το εν λόγω πακέτο.

Συνοψίζοντας λοιπόν η προκύπτουσα δομή δικτύου για την καταγραφή και ανάλυση της κίνησης του δικτύου εμφανίζεται στο ακόλουθο σχήμα (σχήμα 3.4):



Σχήμα 3.4: Τελική δομή δικτύου

Πριν την παράθεση των εντολών που ακολουθεί και που χρησιμοποιούνταν για τη διαδικασία παραγωγής «συχνών στοιχείων» μέσα από το λογισμικό “R”, θα πρέπει να τονιστεί ότι τα δεδομένα που κατέληγαν τελικά προς επεξεργασία από τον αλγόριθμο Apriori έπρεπε να υποστούν κάποια επεξεργασία. Όπως περιγράφηκε και στην παράγραφο με τίτλο «Διαδικασία Εξόρυξης δεδομένων» του δεύτερου κεφαλαίου της παρούσας μεταπτυχιακής διατριβής, ένα από τα αρχικά βήματα στη διαδικασία εξόρυξης δεδομένων είναι ο «μετασχηματισμός - κωδικοποίηση των δεδομένων». Η

ανάγκη για μετασχηματισμό των δεδομένων προέκυψε και στην περίπτωση μας διότι στα αποτελέσματα θα έπρεπε να είναι δυνατός ο διαχωρισμός των αριθμών θυρών προορισμού από αυτές της αφετηρίας και των IP διευθύνσεων προορισμού από τις IP διευθύνσεις αφετηρίας. Αυτό σημαίνει ότι κάθε εγγραφή της βάσης δεδομένων που προέκυπτε από το nfdump περιείχε τις πληροφορίες όπως φαίνονται αυτές στο σχήμα 3.5:

Date	flow start	Duration	Proto	Src IP Addr:Port		Dst IP Addr:Port	Packets	Bytes	Flows
2011-03-01	01:01:00.681	0.000	UDP	84.254.24.51:1024	->	80.76.39.10:53	1	71	1
2011-03-01	01:01:00.813	17.768	TCP	84.254.24.51:1024	->	79.140.95.130:80	2373	96549	1
2011-03-01	01:01:00.885	17.492	TCP	79.140.95.130:80	->	84.254.24.51:1024	3569	5.1 M	1
2011-03-01	01:01:08.457	10.308	TCP	84.254.24.51:1025	->	79.140.95.153:80	10	1117	1
2011-03-01	01:01:08.513	10.252	TCP	79.140.95.153:80	->	84.254.24.51:1025	12	10082	1

Σχήμα 3.5: Τμήμα περιεχομένων αρχείου που εξάγεται από το nfdump

Η IP διεύθυνση πχ 84.254.24.51 όπως φαίνεται στο παραπάνω σχήμα θα μπορούσε να αποτελεί IP διεύθυνση τόσο προορισμού όσο και αφετηρίας. Αν αυτές οι δύο διαφορετικές κατ' ουσίαν πληροφορίες δεν διαχωρίζονταν με κάποιο τρόπο, θα θεωρούνταν από τον αλγόριθμο Apriori ως ίδια πληροφορία με αποτέλεσμα τον λανθασμένο υπολογισμό των «συχνών στοιχείων». Αποφασίστηκε λοιπόν η τροποποίηση των δεδομένων με τον εξής τρόπο:

- Μπροστά από τις IP διευθύνσεις αφετηρίας προστέθηκε το πρόθεμα “SIP_”.
- Μπροστά από τις IP διευθύνσεις προορισμού προστέθηκε το πρόθεμα “DIP_”.
- Μπροστά από τις θύρες αφετηρίας προστέθηκε το πρόθεμα “SP_”.
- Μπροστά από τις θύρες προορισμού προστέθηκε το πρόθεμα “DP_”.

Μετά τις παραπάνω τροποποιήσεις τα δεδομένα μπορούσαν να εισαχθούν στον αλγόριθμο Apriori (“R” πακέτο) για την παραγωγή των «συχνών στοιχείων».

Το πρώτο βήμα με την εκκίνηση του λογισμικού “R” είναι να «φορτωθεί» η βιβλιοθήκη “arulesViz” [62]. Αυτό γίνεται πληκτρολογώντας την εντολή:

Library (“arulesViz”)

Για την παραγωγή των «συχνών στοιχείων» πρέπει να «φορτωθεί» το αρχείο κειμένου με τα δεδομένα, που έχει παραχθεί από το nfdump. Το αρχείο αυτό αποτελεί στην ουσία τη βάση

δεδομένων πάνω στην οποία θα εφαρμοστεί ο αλγόριθμος Apriori για την παραγωγή «συχνών στοιχείων». Τα περιεχόμενα αυτού του αρχείου αποθηκεύονται σε μία μεταβλητή του “R” χρησιμοποιώντας την εντολή:

```
<variable><b>←read.transactions(<file_name>,"format="basket", sep="<column separator>")
```

Όπου:

<variable> είναι το όνομα της μεταβλητής που θέλουμε να δημιουργήσουμε και η οποία θα αποτελεί τη βάση δεδομένων μας,

<file_name> είναι το όνομα του αρχείου από το nfdump

<column separator> ο διαχωριστής των στηλών στο αρχείο κειμένου (στην περίπτωσή μας είναι το «;»).

Ένα παράδειγμα τέτοιας εντολής που χρησιμοποιήθηκε είναι το ακόλουθο (σχήμα 3.6):

```
> tr5month<-read.transactions("04_2011f.txt", format="basket", sep=";")
```

Σχήμα 3.6: Παράδειγμα εντολής για την φόρτωση αρχείου δεδομένων στο “R”

Το επόμενο βήμα είναι η παραγωγή των «συχνών στοιχείων» με τη χρήση του αλγορίθμου Apriori. Η εντολή που χρησιμοποιείται συντάσσεται ως εξής:

```
<variable2> <b><-- apriori(<variable>,"parameter=list(support=<s_value>,"target=<fr_it_type>))
```

Όπου:

<variable2> είναι η μεταβλητή στην οποία θα αποθηκευτούν τα αποτελέσματα του αλγορίθμου Apriori,

<variable> είναι το όνομα της μεταβλητής η οποία δημιουργήθηκε στο προηγούμενο βήμα και περιέχει τη βάση δεδομένων,

<s_value> είναι η τιμή του support την οποία χρειάζεται ο αλγόριθμος Apriori για τον υπολογισμό των «συχνών στοιχείων»,

<frequent itemsets_it_type> ο τύπος των «συχνών στοιχείων» που θέλουμε να δημιουργήσουμε. Μπορεί να πάρει τις ακόλουθες τιμές: frequent itemsets, maximally frequent itemsets, closed frequent itemsets.

Ένα παράδειγμα τέτοιας εντολής φαίνεται στο σχήμα 3.7:

```
> mof0.003<-apriori(trimonth, parameter=list(support=0.003, target="frequent itemsets"))
```

Σχήμα 3.7: Παράδειγμα εντολής για την παραγωγή «συχνών στοιχείων» στο “R”

Για την εμφάνιση των «συχνών στοιχείων», ταξινομημένα με βάση το support (από αυτό με το μεγαλύτερο support προς αυτό με το μικρότερο support) χρησιμοποιείται η εντολή:

inspect (sort(<variable2>, by="support"))

όπου **<variable2>** το όνομα της μεταβλητής με τα αποτελέσματα του αλγορίθμου Apriori.

Παράδειγμα τέτοιας εντολής με τα αποτελέσματά της εμφανίζεται στο σχήμα 3.8:

```

> inspect(sort(f0.06, by="support"))
      items      support
1 {SIP_84.254.24.51} 0.5275881
2 {DIP_84.254.24.51} 0.4648690
3 {DP_80}           0.3143672
4 {SP_80}           0.3083091
5 {DP_80,
   SIP_84.254.24.51} 0.2940251
6 {DIP_84.254.24.51,
   SP_80}           0.2864822
7 {DP_443}          0.1253786
8 {DP_443,
   SIP_84.254.24.51} 0.1243096
9 {SP_443}           0.1228841
10 {DIP_84.254.24.51,
    SP_443}          0.1220526
> |

```

Σχήμα 3.8: Αποτελέσματα εντολής inspect

Στο παραπάνω σχήμα φαίνεται ότι το ποιο «συχνό στοιχείο» για το συγκεκριμένο set δεδομένων (f0.06) είναι αυτό με ένα στοιχείο, το SIP_84.254.24.51. το συγκεκριμένο «συχνό στοιχείο» έχει support 0,5275881. Το λιγότερο «συχνό στοιχείο» είναι αυτό με τα στοιχεία DIP_84.254.24.51 και SP_443 με support 0,1220526.

Για την παραγωγή γραφημάτων με τα αποτελέσματα του αλγορίθμου Apriori συντασσόταν η ακόλουθη εντολή [Visualizing Association Rules Introduction to the R-extension Package arulesViz, Michael Hahsler and Sudheer Chelluboina]:

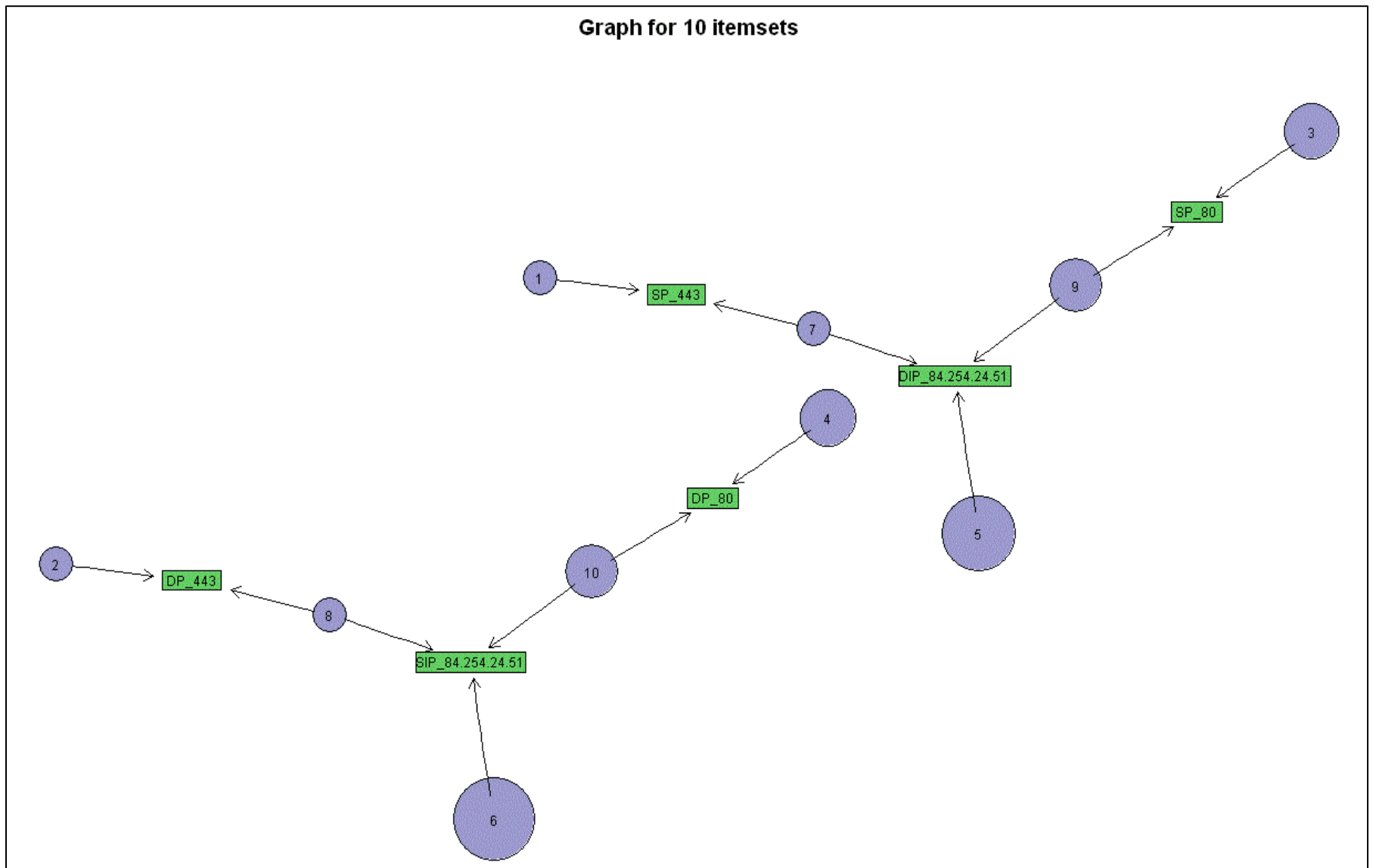
plot(<variable2>, method="<method_name>")

Όπου:

<variable2> είναι η μεταβλητή που περιέχει τα «συχνά στοιχεία»

<method_name> είναι το όνομα του τύπου γραφήματος που πρόκειται να σχεδιαστεί με την παρούσα εντολή. Αυτή η μεταβλητή μπορεί να πάρει τις ακόλουθες τιμές: matrix, matrix3D, grouped, graph, paracoord, double-decker [61]. Παράδειγμα χρήσης αυτής της εντολής βάζοντας ως μέθοδο σχεδίασης τη μέθοδο "graph" εμφανίζεται στο σχήμα 3.9:

```
> plot(cmo5f0.05, method="graph")
```



Σχήμα 3.9: Εντολή plot με method “graph”

3.8 Παράθεση Πειραματικών Δεδομένων - Αποτελεσμάτων

Χρησιμοποιώντας λοιπόν το υλοποιηθέν σύστημα παρακολούθησης, το οποίο περιγράφηκε στις προηγούμενες παραγράφους, καταγράφηκε για μεγάλο χρονικό διάστημα η εξερχόμενη και εισερχόμενη κίνηση του δικτύου. Τα δεδομένα, αφού υπέστησαν κάποια τροποποίηση, εισήχθησαν στον αλγόριθμο Apriori από τον οποίο παράχθηκαν «συχνά στοιχεία» καθώς και γραφικές αναπαραστάσεις αυτών. Ο όγκος των δεδομένων που παράχθηκαν, παρατηρήθηκαν και αναλύθηκαν είναι πολύ μεγάλος και αδύνατον να παρατεθεί στο σύνολό του. Για τον λόγο αυτό στην παρούσα παράγραφο γίνεται παράθεση ενδεικτικά τμήματος των δεδομένων και αποτελέσματα αυτών.

Στο επόμενο σχήμα (σχήμα 3.10) εμφανίζεται τμήμα των raw data τα οποία παράχθηκαν από το εργαλείο-λογισμικό nfdump. Τα δεδομένα αυτά αφορούν συλλεχθέντα δεδομένα στο εύρος ημερομηνιών από 21 έως και 25 Μαρτίου 2011. Το αρχείο που παράχθηκε συγκεντρώνοντας τα flows των ημερών αυτών περιείχε 170509 εγγραφές (όπως φαίνεται και από το παρακάτω σχήμα.

1	Date flow start	Duration	Proto	Src IP Addr:Port		Dst IP Addr:Port	Packets	Bytes	Flows
2	2011-03-21 00:56:24.914	0.000	TCP	46.182.121.166:8080	->	84.254.24.54:1024	1	48	1
3	2011-03-21 00:56:18.174	11.632	TCP	79.103.251.72:49924	->	84.254.24.54:3299	9	1712	1
4	2011-03-21 00:56:18.210	11.392	TCP	84.254.24.54:3299	->	79.103.251.72:49924	12	10361	1
5	2011-03-21 00:56:50.534	0.476	TCP	65.52.110.23:14034	->	84.254.24.51:80	5	473	1
6	2011-03-21 00:56:50.534	0.268	TCP	84.254.24.51:80	->	65.52.110.23:14034	3	1947	1
7	2011-03-21 00:56:57.990	3.452	TCP	79.103.251.72:49924	->	84.254.24.54:3299	7	1579	1
8	2011-03-21 00:56:58.166	3.016	TCP	84.254.24.54:3299	->	79.103.251.72:49924	9	7264	1
9	2011-03-21 00:57:46.814	2.892	TCP	65.52.110.23:27218	->	84.254.24.51:80	6	537	1
10	2011-03-21 00:57:46.814	2.668	TCP	84.254.24.51:80	->	65.52.110.23:27218	7	5563	1
11	2011-03-21 00:59:04.238	0.564	TCP	84.254.24.54:3299	->	79.103.251.72:49924	8	7205	1
12	2011-03-21 00:59:04.182	0.832	TCP	79.103.251.72:49924	->	84.254.24.54:3299	5	1169	1
13	2011-03-21 00:59:28.858	0.984	TCP	79.103.251.72:49924	->	84.254.24.54:3299	6	1214	1
14	2011-03-21 00:59:29.018	0.528	TCP	84.254.24.54:3299	->	79.103.251.72:49924	8	7652	1
15	2011-03-21 01:00:18.286	0.000	UDP	84.254.24.51:1025	->	80.76.39.10:53	1	71	1
16	2011-03-21 01:00:26.942	0.284	TCP	84.254.24.51:1025	->	83.171.202.8:80	7	636	1
17	2011-03-21 01:00:26.958	0.268	TCP	83.171.202.8:80	->	84.254.24.51:1025	9	9096	1
18	2011-03-21 01:00:36.806	1.008	TCP	79.103.251.72:49924	->	84.254.24.54:3299	6	1295	1
19	2011-03-21 01:00:36.934	0.604	TCP	84.254.24.54:3299	->	79.103.251.72:49924	9	7878	1
20	2011-03-21 01:00:18.414	25.052	TCP	84.254.24.51:1024	->	83.171.202.6:80	5755	231277	1
21	2011-03-21 01:00:18.434	24.836	TCP	83.171.202.6:80	->	84.254.24.51:1024	8707	12.4 M	1
22	2011-03-21 01:00:43.386	0.192	TCP	84.254.24.51:1025	->	83.171.202.8:80	2	441	1
23	2011-03-21 01:00:43.406	0.012	TCP	83.171.202.8:80	->	84.254.24.51:1025	2	517	1
24	2011-03-21 01:01:09.238	0.588	TCP	84.254.24.54:3299	->	79.103.251.72:49924	8	7361	1
25	2011-03-21 01:01:09.126	0.912	TCP	79.103.251.72:49924	->	84.254.24.54:3299	5	1171	1
26	2011-03-21 01:03:25.766	0.108	UDP	84.254.24.51:1026	->	80.76.39.10:53	2	131	1
27	2011-03-21 01:03:29.998	0.000	TCP	218.93.205.174:3310	->	84.254.24.51:44	1	52	1
28	2011-03-21 01:03:34.790	12.872	TCP	79.103.251.72:49924	->	84.254.24.54:3299	18	3945	1
29	2011-03-21 01:03:34.970	12.472	TCP	84.254.24.54:3299	->	79.103.251.72:49924	27	24453	1
30	2011-03-21 01:04:16.074	0.000	TCP	218.93.205.174:3310	->	84.254.24.53:44	1	52	1
31	2011-03-21 01:05:39.442	1.132	TCP	79.103.251.72:49924	->	84.254.24.54:3299	4	1095	1
32	2011-03-21 01:05:39.594	0.704	TCP	84.254.24.54:3299	->	79.103.251.72:49924	6	5150	1
33	2011-03-21 01:06:18.742	0.588	TCP	79.103.251.72:49924	->	84.254.24.54:3299	5	980	1
34	2011-03-21 01:06:18.746	0.376	TCP	84.254.24.54:3299	->	79.103.251.72:49924	6	4836	1
35	2011-03-21 01:07:15.878	4.904	TCP	84.254.24.54:3299	->	79.103.251.72:49924	12	10486	1
36	2011-03-21 01:07:15.862	5.216	TCP	79.103.251.72:49924	->	84.254.24.54:3299	9	1816	1
37	2011-03-21 01:08:25.950	0.196	UDP	84.254.24.51:1027	->	80.76.39.10:53	2	131	1

Σχήμα 3.10a: Raw data, 21-25/3/2011

38	2011-03-21 01:08:40.178	0.016	TCP	84.254.24.51:1025	->	83.171.202.8:80	2	80	1
39	2011-03-21 01:08:40.178	0.016	TCP	84.254.24.51:1024	->	83.171.202.6:80	2	80	1
40	2011-03-21 01:08:40.194	0.000	TCP	83.171.202.8:80	->	84.254.24.51:1025	1	40	1
41	2011-03-21 01:08:40.194	0.000	TCP	83.171.202.6:80	->	84.254.24.51:1024	1	40	1
42	2011-03-21 01:08:40.594	4.088	TCP	79.103.251.72:49924	->	84.254.24.54:3299	11	2401	1
43	2011-03-21 01:08:40.634	3.832	TCP	84.254.24.54:3299	->	79.103.251.72:49924	16	15026	1
44	2011-03-21 01:09:02.466	1.160	TCP	79.103.251.72:49924	->	84.254.24.54:3299	6	1209	1
45	2011-03-21 01:09:02.570	0.804	TCP	84.254.24.54:3299	->	79.103.251.72:49924	9	8267	1
46	2011-03-21 01:09:27.930	32.700	TCP	79.103.251.72:49924	->	84.254.24.54:3299	44	8583	1
47	2011-03-21 01:09:28.118	32.316	TCP	84.254.24.54:3299	->	79.103.251.72:49924	57	51488	1
48	2011-03-21 01:10:26.774	0.000	TCP	94.76.204.164:2621	->	84.254.24.54:1433	1	40	1
49	2011-03-21 01:10:26.774	0.000	TCP	94.76.204.164:2618	->	84.254.24.51:1433	1	40	1
50	2011-03-21 01:10:26.778	0.000	TCP	94.76.204.164:2620	->	84.254.24.53:1433	1	40	1
51	2011-03-21 01:10:26.782	0.000	TCP	94.76.204.164:2617	->	84.254.24.50:1433	1	40	1
52	2011-03-21 01:10:16.394	15.344	TCP	79.103.251.72:49924	->	84.254.24.54:3299	9	1262	1
53	2011-03-21 01:10:16.406	15.080	TCP	84.254.24.54:3299	->	79.103.251.72:49924	11	8321	1
54	2011-03-21 01:16:33.962	0.000	TCP	218.93.205.174:3310	->	84.254.24.53:44	1	52	1
55	2011-03-21 01:19:17.918	0.000	UDP	84.254.24.51:1024	->	80.76.39.10:53	1	71	1
56	2011-03-21 01:19:23.590	8.204	TCP	83.171.202.6:80	->	84.254.24.51:1025	3174	4.5 M	1
57	2011-03-21 01:19:18.002	14.192	TCP	84.254.24.51:1024	->	83.171.202.8:80	11	1163	1
58	2011-03-21 01:19:18.022	14.020	TCP	83.171.202.8:80	->	84.254.24.51:1024	14	9589	1
59	2011-03-21 01:19:23.574	8.400	TCP	84.254.24.51:1025	->	83.171.202.6:80	2093	84891	1
60	2011-03-21 01:20:51.438	1.208	TCP	79.103.251.72:49924	->	84.254.24.54:3299	6	1300	1
61	2011-03-21 01:20:51.586	0.804	TCP	84.254.24.54:3299	->	79.103.251.72:49924	9	8105	1
62	2011-03-21 01:21:11.566	0.884	TCP	79.103.251.72:49924	->	84.254.24.54:3299	6	1224	1
63	2011-03-21 01:21:11.646	0.560	TCP	84.254.24.54:3299	->	79.103.251.72:49924	8	7360	1
64	2011-03-21 01:23:14.986	0.000	TCP	85.217.190.69:23804	->	84.254.24.51:22	1	48	1
65	2011-03-21 01:23:14.990	0.000	TCP	85.217.190.69:23804	->	84.254.24.53:22	1	48	1
66	2011-03-21 01:23:14.990	0.000	TCP	85.217.190.69:23804	->	84.254.24.50:22	1	48	1
67	2011-03-21 01:23:14.990	0.000	TCP	85.217.190.69:23804	->	84.254.24.54:22	1	48	1
68	2011-03-21 01:24:19.846	0.000	TCP	218.93.205.174:3310	->	84.254.24.51:44	1	52	1
69	2011-03-21 01:24:41.646	10.676	TCP	79.103.251.72:49924	->	84.254.24.54:3299	15	2539	1
70	2011-03-21 01:24:41.822	10.260	TCP	84.254.24.54:3299	->	79.103.251.72:49924	22	21985	1

Σχήμα 3.10b: Raw data, 21-25/3/2011

170477	2011-03-25 19:32:16.633	2.516	TCP	92.240.235.156:80	->	84.254.24.51:3397	71	100632	1
170478	2011-03-25 19:32:18.161	1.000	TCP	84.254.24.51:2855	->	74.125.232.112:80	67	3692	1
170479	2011-03-25 19:32:18.241	0.920	TCP	74.125.232.112:80	->	84.254.24.51:2855	80	110919	1
170480	2011-03-25 19:32:18.645	0.496	TCP	84.254.24.51:2859	->	74.125.232.114:80	15	1301	1
170481	2011-03-25 19:32:18.741	0.400	TCP	74.125.232.114:80	->	84.254.24.51:2859	19	25083	1
170482	2011-03-25 19:32:18.945	1.112	TCP	84.254.24.51:3241	->	74.125.77.101:80	3	1738	1
170483	2011-03-25 19:32:19.917	0.000	TCP	74.125.77.101:80	->	84.254.24.51:3241	1	370	1
170484	2011-03-25 19:32:35.761	0.000	TCP	84.254.24.51:2856	->	74.125.232.115:80	1	40	1
170485	2011-03-25 19:32:35.761	0.000	TCP	84.254.24.51:3344	->	74.125.232.115:80	1	40	1
170486	2011-03-25 19:32:22.425	1.132	TCP	84.254.24.51:3009	->	74.125.77.99:80	88	3875	1
170487	2011-03-25 19:32:22.533	0.868	TCP	74.125.77.99:80	->	84.254.24.51:3009	131	184962	1
170488	2011-03-25 19:32:18.797	6.072	TCP	84.254.24.51:3406	->	74.125.232.144:80	14	1813	1
170489	2011-03-25 19:32:18.841	6.064	TCP	74.125.232.147:80	->	84.254.24.51:3065	3	1684	1
170490	2011-03-25 19:32:18.861	5.856	TCP	74.125.232.144:80	->	84.254.24.51:3406	19	21390	1
170491	2011-03-25 19:32:19.165	5.376	TCP	84.254.24.51:2858	->	74.125.232.113:80	68	4178	1
170492	2011-03-25 19:32:19.225	5.208	TCP	74.125.232.113:80	->	84.254.24.51:2858	80	109955	1
170493	2011-03-25 19:32:16.761	8.144	TCP	92.240.235.156:80	->	84.254.24.51:3407	71	97467	1
170494	2011-03-25 19:32:24.213	0.544	TCP	84.254.24.51:3222	->	74.125.77.102:80	16	1873	1
170495	2011-03-25 19:32:24.325	0.276	TCP	74.125.77.102:80	->	84.254.24.51:3222	22	28654	1
170496	2011-03-25 19:32:18.737	6.348	TCP	84.254.24.51:3065	->	74.125.232.147:80	4	981	1
170497	2011-03-25 19:32:34.249	9.648	TCP	84.254.24.51:3419	->	92.240.235.156:80	17	3618	1
170498	2011-03-25 19:32:48.837	0.000	TCP	84.254.24.51:3337	->	74.125.232.113:80	1	40	1
170499	2011-03-25 19:32:35.649	0.264	TCP	84.254.24.51:3021	->	74.125.232.146:80	2	413	1
170500	2011-03-25 19:32:35.725	0.024	TCP	74.125.232.146:80	->	84.254.24.51:3021	2	434	1
170501	2011-03-25 19:32:17.005	18.892	TCP	74.125.232.115:80	->	84.254.24.51:3369	44	54116	1
170502	2011-03-25 19:32:35.481	0.000	UDP	84.254.24.51:1500	->	80.76.39.10:53	1	70	1
170503	2011-03-25 19:32:16.897	19.128	TCP	84.254.24.51:3369	->	74.125.232.115:80	34	4045	1
170504	2011-03-25 19:32:18.157	17.800	TCP	84.254.24.51:3336	->	74.125.232.112:80	127	9085	1
170505	2011-03-25 19:32:18.217	17.740	TCP	74.125.232.112:80	->	84.254.24.51:3336	158	212052	1
170506	2011-03-25 19:32:19.289	23.408	TCP	84.254.24.51:3365	->	74.125.77.104:80	101	5694	1
170507	2011-03-25 19:32:19.389	23.176	TCP	74.125.77.104:80	->	84.254.24.51:3365	143	198973	1
170508	2011-03-25 19:32:35.985	7.396	TCP	74.125.232.114:80	->	84.254.24.51:2859	3	2732	1
170509	2011-03-25 19:32:35.501	8.180	TCP	84.254.24.51:3006	->	74.125.77.100:80	19	2865	1

Σχήμα 3.10c: Raw data, 21-25/3/2011

Τα δεδομένα πριν την εισαγωγή στον αλγόριθμο nfdump πήραν την ακόλουθη μορφή (σχήμα 3.11):

```
1 SIP_46.182.121.166;SP_8080;DIP_84.254.24.54;DP_1024
2 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
3 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
4 SIP_65.52.110.23;SP_14034;DIP_84.254.24.51;DP_80
5 SIP_84.254.24.51;SP_80;DIP_65.52.110.23;DP_14034
6 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
7 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
8 SIP_65.52.110.23;SP_27218;DIP_84.254.24.51;DP_80
9 SIP_84.254.24.51;SP_80;DIP_65.52.110.23;DP_27218
10 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
11 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
12 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
13 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
14 SIP_84.254.24.51;SP_1025;DIP_80.76.39.10;DP_53
15 SIP_84.254.24.51;SP_1025;DIP_83.171.202.8;DP_80
16 SIP_83.171.202.8;SP_80;DIP_84.254.24.51;DP_1025
17 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
18 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
19 SIP_84.254.24.51;SP_1024;DIP_83.171.202.6;DP_80
20 SIP_83.171.202.6;SP_80;DIP_84.254.24.51;DP_1024
21 SIP_84.254.24.51;SP_1025;DIP_83.171.202.8;DP_80
22 SIP_83.171.202.8;SP_80;DIP_84.254.24.51;DP_1025
23 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
24 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
25 SIP_84.254.24.51;SP_1026;DIP_80.76.39.10;DP_53
26 SIP_218.93.205.174;SP_3310;DIP_84.254.24.51;DP_44
27 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
28 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
29 SIP_218.93.205.174;SP_3310;DIP_84.254.24.53;DP_44
30 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
31 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
32 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
33 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
34 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
35 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
36 SIP_84.254.24.51;SP_1027;DIP_80.76.39.10;DP_53
37 SIP_84.254.24.51;SP_1025;DIP_83.171.202.8;DP_80
```

Σχήμα 3.11a: Τροποποιημένα δεδομένα (21-25/3/2011) για εισαγωγή στον αλγόριθμο Apriori

```

37 SIP_84.254.24.51;SP_1025;DIP_83.171.202.8;DP_80
38 SIP_84.254.24.51;SP_1024;DIP_83.171.202.6;DP_80
39 SIP_83.171.202.8;SP_80;DIP_84.254.24.51;DP_1025
40 SIP_83.171.202.6;SP_80;DIP_84.254.24.51;DP_1024
41 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
42 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
43 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
44 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
45 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
46 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
47 SIP_94.76.204.164;SP_2621;DIP_84.254.24.54;DP_1433
48 SIP_94.76.204.164;SP_2618;DIP_84.254.24.51;DP_1433
49 SIP_94.76.204.164;SP_2620;DIP_84.254.24.53;DP_1433
50 SIP_94.76.204.164;SP_2617;DIP_84.254.24.50;DP_1433
51 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
52 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
53 SIP_218.93.205.174;SP_3310;DIP_84.254.24.53;DP_44
54 SIP_84.254.24.51;SP_1024;DIP_80.76.39.10;DP_53
55 SIP_83.171.202.6;SP_80;DIP_84.254.24.51;DP_1025
56 SIP_84.254.24.51;SP_1024;DIP_83.171.202.8;DP_80
57 SIP_83.171.202.8;SP_80;DIP_84.254.24.51;DP_1024
58 SIP_84.254.24.51;SP_1025;DIP_83.171.202.6;DP_80
59 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
60 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
61 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
62 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
63 SIP_85.217.190.69;SP_23804;DIP_84.254.24.51;DP_22
64 SIP_85.217.190.69;SP_23804;DIP_84.254.24.53;DP_22
65 SIP_85.217.190.69;SP_23804;DIP_84.254.24.50;DP_22
66 SIP_85.217.190.69;SP_23804;DIP_84.254.24.54;DP_22
67 SIP_218.93.205.174;SP_3310;DIP_84.254.24.51;DP_44
68 SIP_79.103.251.72;SP_49924;DIP_84.254.24.54;DP_3299
69 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924
70 SIP_84.254.24.54;SP_3299;DIP_79.103.251.72;DP_49924

```

Σχήμα 3.11b: Τροποποιημένα δεδομένα (21-25/3/2011) για εισαγωγή στον αλγόριθμο Apriori

170476	SIP_92.240.235.156;SP_80;DIP_84.254.24.51;DP_3397
170477	SIP_84.254.24.51;SP_2855;DIP_74.125.232.112;DP_80
170478	SIP_74.125.232.112;SP_80;DIP_84.254.24.51;DP_2855
170479	SIP_84.254.24.51;SP_2859;DIP_74.125.232.114;DP_80
170480	SIP_74.125.232.114;SP_80;DIP_84.254.24.51;DP_2859
170481	SIP_84.254.24.51;SP_3241;DIP_74.125.77.101;DP_80
170482	SIP_74.125.77.101;SP_80;DIP_84.254.24.51;DP_3241
170483	SIP_84.254.24.51;SP_2856;DIP_74.125.232.115;DP_80
170484	SIP_84.254.24.51;SP_3344;DIP_74.125.232.115;DP_80
170485	SIP_84.254.24.51;SP_3009;DIP_74.125.77.99;DP_80
170486	SIP_74.125.77.99;SP_80;DIP_84.254.24.51;DP_3009
170487	SIP_84.254.24.51;SP_3406;DIP_74.125.232.144;DP_80
170488	SIP_74.125.232.147;SP_80;DIP_84.254.24.51;DP_3065
170489	SIP_74.125.232.144;SP_80;DIP_84.254.24.51;DP_3406
170490	SIP_84.254.24.51;SP_2858;DIP_74.125.232.113;DP_80
170491	SIP_74.125.232.113;SP_80;DIP_84.254.24.51;DP_2858
170492	SIP_92.240.235.156;SP_80;DIP_84.254.24.51;DP_3407
170493	SIP_84.254.24.51;SP_3222;DIP_74.125.77.102;DP_80
170494	SIP_74.125.77.102;SP_80;DIP_84.254.24.51;DP_3222
170495	SIP_84.254.24.51;SP_3065;DIP_74.125.232.147;DP_80
170496	SIP_84.254.24.51;SP_3419;DIP_92.240.235.156;DP_80
170497	SIP_84.254.24.51;SP_3337;DIP_74.125.232.113;DP_80
170498	SIP_84.254.24.51;SP_3021;DIP_74.125.232.146;DP_80
170499	SIP_74.125.232.146;SP_80;DIP_84.254.24.51;DP_3021
170500	SIP_74.125.232.115;SP_80;DIP_84.254.24.51;DP_3369
170501	SIP_84.254.24.51;SP_1500;DIP_80.76.39.10;DP_53
170502	SIP_84.254.24.51;SP_3369;DIP_74.125.232.115;DP_80
170503	SIP_84.254.24.51;SP_3336;DIP_74.125.232.112;DP_80
170504	SIP_74.125.232.112;SP_80;DIP_84.254.24.51;DP_3336
170505	SIP_84.254.24.51;SP_3365;DIP_74.125.77.104;DP_80
170506	SIP_74.125.77.104;SP_80;DIP_84.254.24.51;DP_3365
170507	SIP_74.125.232.114;SP_80;DIP_84.254.24.51;DP_2859
170508	SIP_84.254.24.51;SP_3006;DIP_74.125.77.100;DP_80

Σχήμα 3.11c: Τροποποιημένα δεδομένα (21-25/3/2011) για εισαγωγή στον αλγόριθμο Apriori

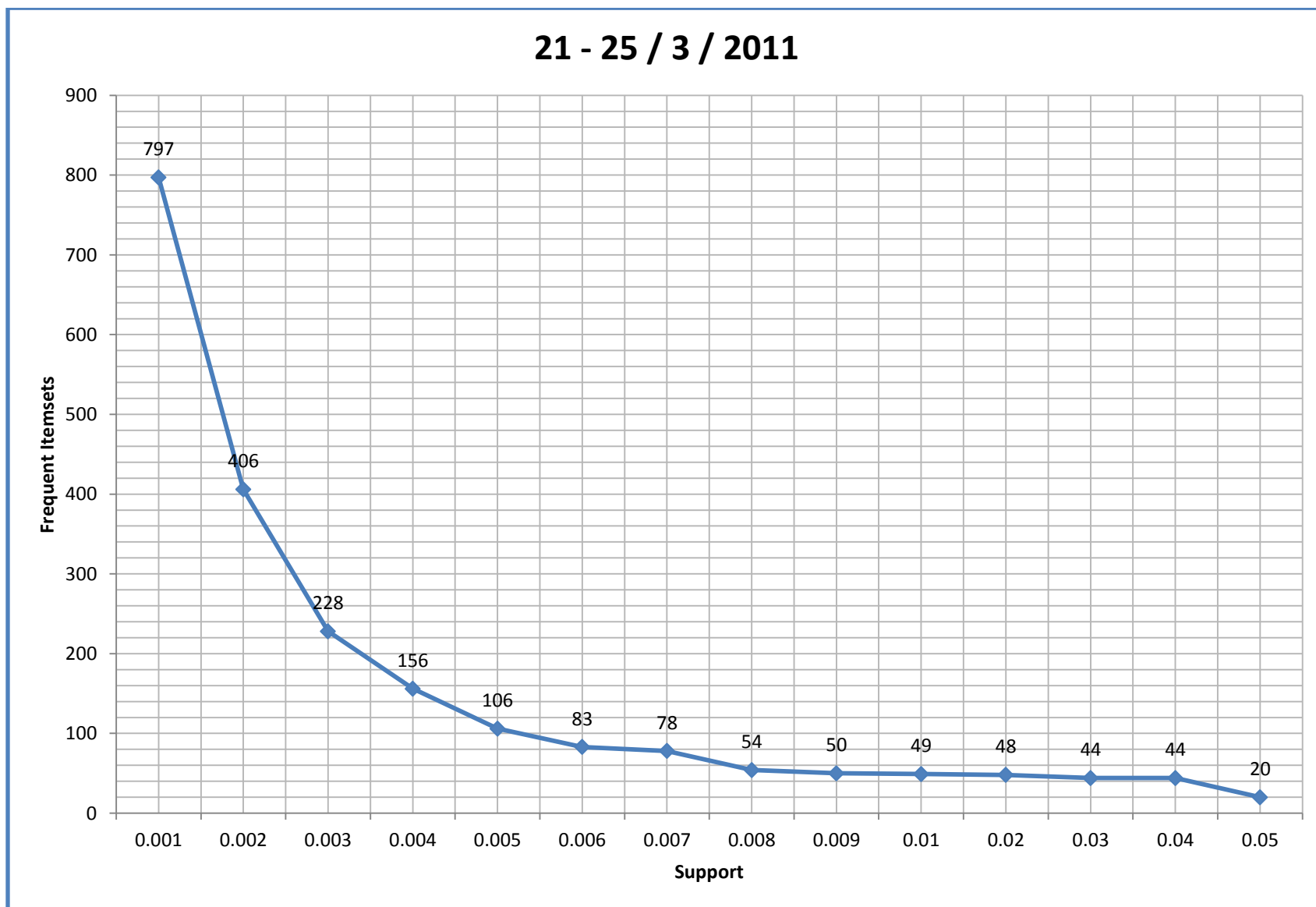
Εφαρμόζοντας τον αλγόριθμο Apriori στα παραπάνω δεδομένα και τροποποιώντας την τιμή της παραμέτρου «στήριξη» (support) προέκυψαν «συχνά στοιχεία» ο αριθμός των οποίων εμφανίζεται στον ακόλουθο πίνακα (πίνακας 3.3):

SUPPORT	NUMBER OF FREQUENT ITEMSETS
0.001	797
0.002	406
0.003	228
0.004	156
0.005	106
0.006	83
0.007	78
0.008	54
0.009	50
0.01	49
0.02	48
0.03	44
0.04	44
0.045	20
0.05	20

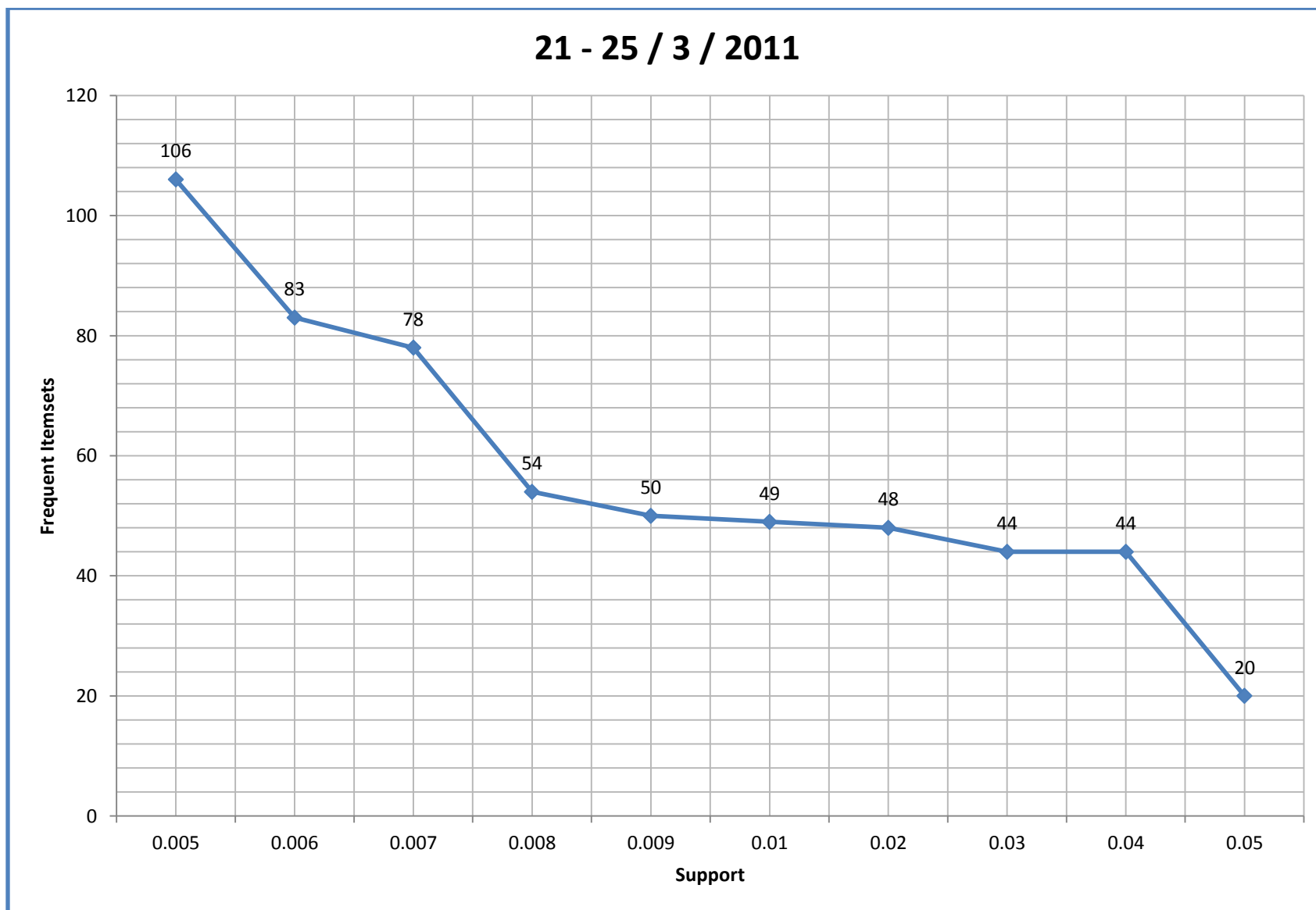
Πίνακας 3.3: Αριθμός FIs / support value – data 21-25/3/2011

Με βάση τον παραπάνω πίνακα παρατηρούμε ότι ο αριθμός των «συχνών στοιχείων» είναι υπερβολικά μεγάλος (797 «συχνά στοιχεία») όταν η «στήριξη» είναι πολύ μικρή (0.001). Φυσικά ο

αριθμός των «συχνών στοιχείων» μειώνεται με την αύξηση της τιμής της «στήριξης» φτάνοντας τα 20 «συχνά στοιχεία» για «στήριξη» 0.05. Ο πίνακας αυτός μπορεί να αναπαρασταθεί στο ακόλουθο σχεδιάγραμμα (σχήμα 3.12):



Σχήμα 3.12: Αριθμός «συχνών στοιχείων» ανά «στήριξη» για τα δεδομένα 21-25/3/2011



Σχήμα 3.13: Εστιασμένο τμήμα του σχήματος 3.12. Αριθμός «συχνών στοιχείων» ανά «στήριξη» για τα δεδομένα 21-25/3/2011

Ενδεικτικά εμφανίζονται παρακάτω τα παραγόμενα «συχνά στοιχεία» για «στήριξη» 0.04 και 0.05 αντίστοιχα (σχήμα 3.14 και σχήμα 3.15 αντίστοιχα):

	items	support
1	{SIP_84.254.24.51}	0.53216858
2	{DIP_84.254.24.51}	0.45352124
3	{DP_80}	0.30379806
4	{DP_80, SIP_84.254.24.51}	0.28976939
5	{SP_80}	0.28718301
6	{DIP_84.254.24.51, SP_80}	0.27647970
7	{DP_53}	0.07216670
8	{DP_53, SIP_84.254.24.51}	0.07216670
9	{DP_443}	0.07131630
10	{DP_443, SIP_84.254.24.51}	0.06923429
11	{DIP_80.76.39.10}	0.06876510
12	{DIP_80.76.39.10, DP_53}	0.06876510
13	{DIP_80.76.39.10, SIP_84.254.24.51}	0.06876510
14	{DIP_80.76.39.10, DP_53, SIP_84.254.24.51}	0.06876510
15	{SP_443}	0.06869472
16	{DIP_84.254.24.51, SP_443}	0.06707603
17	{SP_110}	0.05942243
18	{DP_110}	0.05925235
19	{SIP_84.254.24.51, SP_110}	0.05922889
20	{DIP_84.254.24.51, DP_110}	0.05900603
21	{DIP_207.58.204.58}	0.04108898
22	{DIP_207.58.204.58, SP_110}	0.04108898
23	{DIP_207.58.204.58, SIP_84.254.24.51}	0.04108898
24	{DIP_207.58.204.58, SIP_84.254.24.51, SP_110}	0.04108898
25	{SIP_207.58.204.58}	0.04094236
26	{DP_110, SIP_207.58.204.58}	0.04094236
27	{DIP_84.254.24.51, SIP_207.58.204.58}	0.04094236

```

28 {DIP_84.254.24.51,
    DP_110,
    SIP_207.58.204.58} 0.04094236
29 {SIP_83.171.202.6} 0.04054942
30 {DIP_84.254.24.51,
    SIP_83.171.202.6} 0.04054942
31 {SIP_83.171.202.6,
    SP_80} 0.04050250
32 {DIP_84.254.24.51,
    SIP_83.171.202.6,
    SP_80} 0.04050250
33 {SIP_83.171.202.8} 0.04049663
34 {DIP_84.254.24.51,
    SIP_83.171.202.8} 0.04049663
35 {SIP_83.171.202.8,
    SP_80} 0.04045558
36 {DIP_84.254.24.51,
    SIP_83.171.202.8,
    SP_80} 0.04045558
37 {DIP_83.171.202.8} 0.04028550
38 {DIP_83.171.202.8,
    SIP_84.254.24.51} 0.04028550
39 {DIP_83.171.202.6} 0.04027377
40 {DIP_83.171.202.6,
    SIP_84.254.24.51} 0.04027377
41 {DIP_83.171.202.8,
    DP_80} 0.04024445
42 {DIP_83.171.202.8,
    DP_80,
    SIP_84.254.24.51} 0.04024445
43 {DIP_83.171.202.6,
    DP_80} 0.04022685
44 {DIP_83.171.202.6,
    DP_80,
    SIP_84.254.24.51} 0.04022685
.

```

Σχήμα 3.14: Σύνολο «συχνών στοιχείων» για «στήριξη» 0.04 (Data 21-25/3/2011)

	items	support
1	{SIP_84.254.24.51}	0.53216858
2	{DIP_84.254.24.51}	0.45352124
3	{DP_80}	0.30379806
4	{DP_80, SIP_84.254.24.51}	0.28976939
5	{SP_80}	0.28718301
6	{DIP_84.254.24.51, SP_80}	0.27647970
7	{DP_53}	0.07216670
8	{DP_53, SIP_84.254.24.51}	0.07216670
9	{DP_443}	0.07131630
10	{DP_443, SIP_84.254.24.51}	0.06923429
11	{DIP_80.76.39.10}	0.06876510
12	{DIP_80.76.39.10, DP_53}	0.06876510
13	{DIP_80.76.39.10, SIP_84.254.24.51}	0.06876510
14	{DIP_80.76.39.10, DP_53, SIP_84.254.24.51}	0.06876510
15	{SP_443}	0.06869472
16	{DIP_84.254.24.51, SP_443}	0.06707603
17	{SP_110}	0.05942243
18	{DP_110}	0.05925235
19	{SIP_84.254.24.51, SP_110}	0.05922889
20	{DIP_84.254.24.51, DP_110}	0.05900603

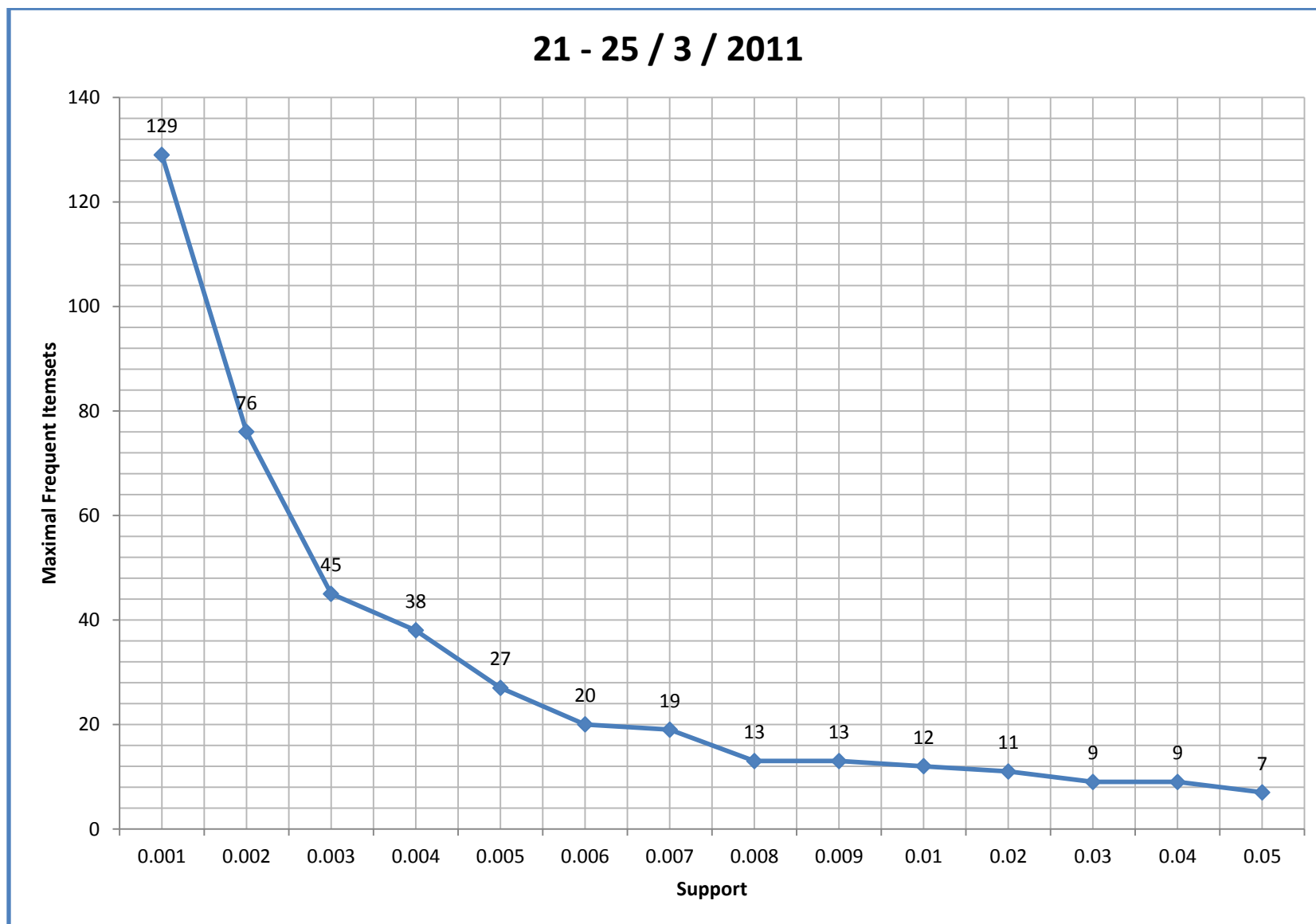
Σχήμα 3.15: Σύνολο «συχνών στοιχείων» για «στήριξη» 0.05

Για τα ίδια δεδομένα ο αριθμός των maximal frequent itemsets που παράγονται εμφανίζονται στον ακόλουθο πίνακα (Πίνακας 3.4):

SUPPORT	NUMBER OF FREQUENT ITEMSETS
0.001	129
0.002	76
0.003	45
0.004	38
0.005	27
0.006	20
0.007	19
0.008	13
0.009	13
0.01	12
0.02	11
0.03	9
0.04	9
0.05	7

Πίνακας 3.4: Αριθμός Maximal FIs / support value – data 21-25/3/2011

Ο πίνακας αυτός μπορεί να αναπαρασταθεί στο ακόλουθο σχεδιάγραμμα (σχήμα 3.16):



Σχήμα 3.16: Αριθμός maximal frequent itemsets ανά «στήριξη» για τα δεδομένα 21-25/3/2011

Ενδεικτικά τα παραγόμενα maximal frequent itemsets για «στήριξη» 0.01 εμφανίζονται στο σχήμα 3.17:

	items	support
1	{DP_443, SIP_84.254.24.51}	0.06923429
2	{DIP_80.76.39.10, DP_53, SIP_84.254.24.51}	0.06876510
3	{DIP_84.254.24.51, SP_443}	0.06707603
4	{DIP_207.58.204.58, SIP_84.254.24.51, SP_110}	0.04108898
5	{DIP_84.254.24.51, DP_110, SIP_207.58.204.58}	0.04094236
6	{DIP_84.254.24.51, SIP_83.171.202.6, SP_80}	0.04050250
7	{DIP_84.254.24.51, SIP_83.171.202.8, SP_80}	0.04045558
8	{DIP_83.171.202.8, DP_80, SIP_84.254.24.51}	0.04024445
9	{DIP_83.171.202.6, DP_80, SIP_84.254.24.51}	0.04022685
10	{DIP_84.254.24.51, DP_3389}	0.02142421
11	{SIP_84.254.24.51, SP_3389}	0.02126000
12	{DIP_84.254.24.51, DP_80}	0.01389378

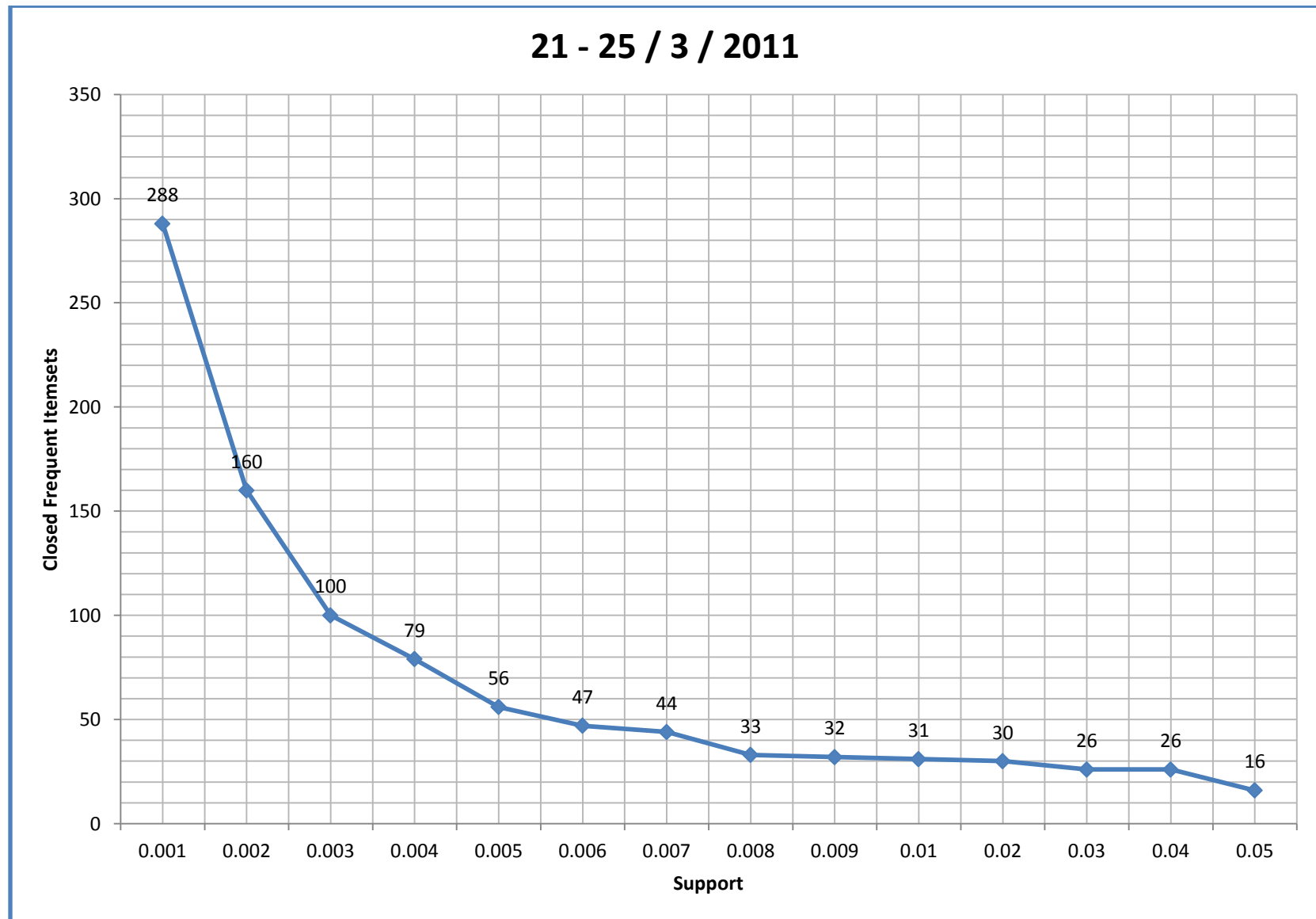
Σχήμα 3.17: Σύνολο maximal frequent itemsets για «στήριξη» 0.01

Για τα ίδια δεδομένα ο αριθμός των closed frequent itemsets που παράγονται εμφανίζονται στον ακόλουθο πίνακα (Πίνακας 3.5):

SUPPORT	NUMBER OF FREQUENT ITEMSETS
0.001	288
0.002	160
0.003	100
0.004	79
0.005	56
0.006	47
0.007	44
0.008	33
0.009	32
0.01	31
0.02	30
0.03	26
0.04	26
0.05	16

Πίνακας 3.5: Αριθμός Closed FIs / support value – data 21-25/3/2011

Ο πίνακας αυτός μπορεί να αναπαρασταθεί στο ακόλουθο σχεδιάγραμμα (σχήμα 3.18):



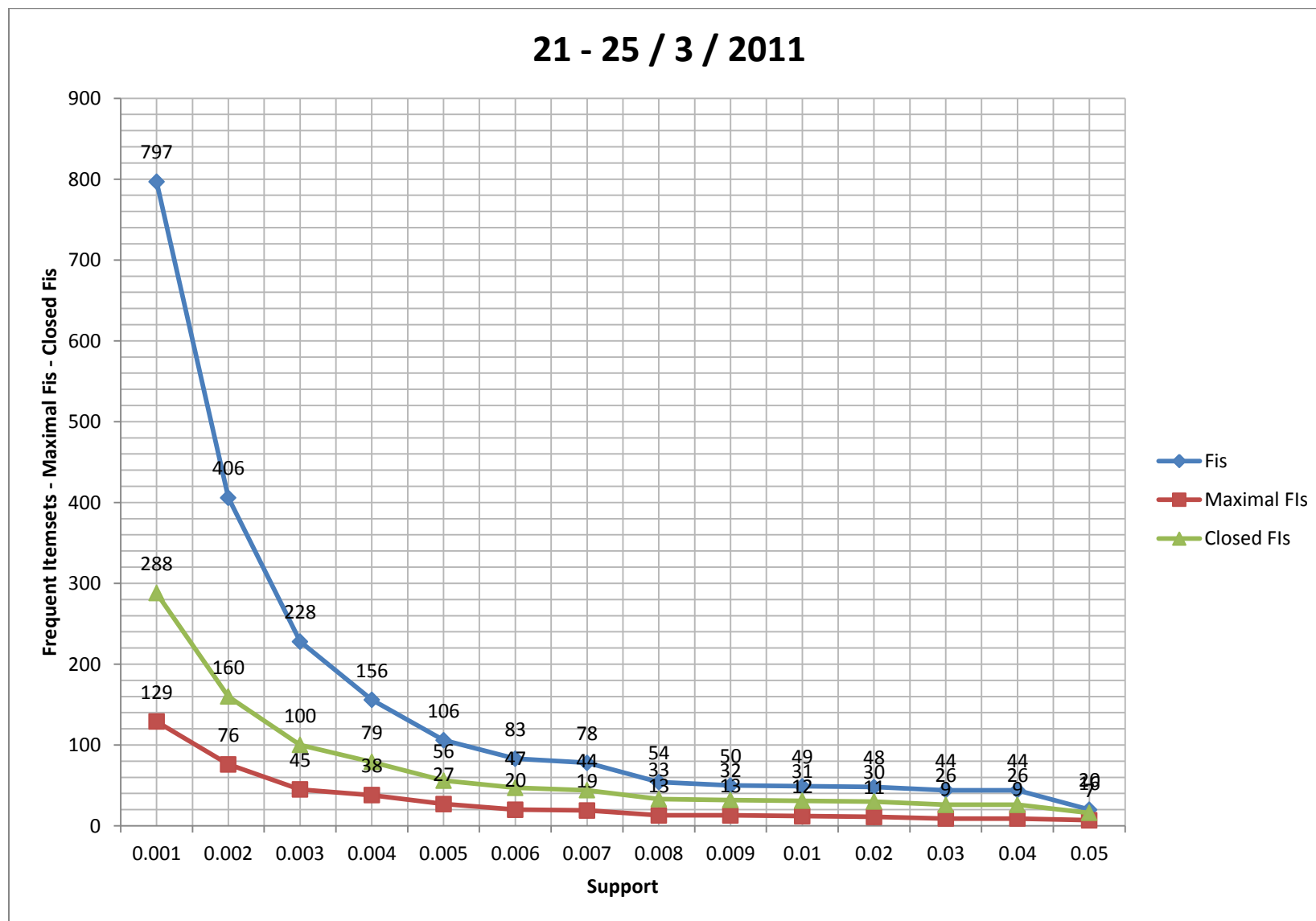
Σχήμα 3.18: Αριθμός closed frequent itemsets ανά «στήριξη» για τα δεδομένα 21-25/3/2011

Ενδεικτικά τα παραγόμενα closed frequent itemsets για «στήριξη» 0.05 εμφανίζονται στο σχήμα 3.19:

	items	support
1	{SIP_84.254.24.51}	0.53216858
2	{DIP_84.254.24.51}	0.45352124
3	{DP_80}	0.30379806
4	{DP_80, SIP_84.254.24.51}	0.28976939
5	{SP_80}	0.28718301
6	{DIP_84.254.24.51, SP_80}	0.27647970
7	{DP_53, SIP_84.254.24.51}	0.07216670
8	{DP_443}	0.07131630
9	{DP_443, SIP_84.254.24.51}	0.06923429
10	{DIP_80.76.39.10, DP_53, SIP_84.254.24.51}	0.06876510
11	{SP_443}	0.06869472
12	{DIP_84.254.24.51, SP_443}	0.06707603
13	{SP_110}	0.05942243
14	{DP_110}	0.05925235
15	{SIP_84.254.24.51, SP_110}	0.05922889
16	{DIP_84.254.24.51, DP_110}	0.05900603

Σχήμα 3.19: Σύνολο closed frequent itemsets για «στήριξη» 0.05

Για να μπορέσει να γίνει μία καλύτερη σύγκριση μεταξύ των «συχνών στοιχείων», “maximal frequent itemsets” και “closed frequent itemsets” σχετικά με τον αριθμό των παραγόμενων «συχνών στοιχείων» και την τιμή «στήριξης», δημιουργήθηκε το παρακάτω σχήμα (σχήμα 3.20) το οποίο περιέχει και τους τρεις παραπάνω τύπους «συχνών στοιχείων».



Σχήμα 3.20: Σύγκριση αριθμού frequent itemsets, Closed FIs και Maximal FIs ανά «στήριξη» για τα δεδομένα 21-25/3/2011

Επειδή ένας από τους στόχους της παρούσας εργασίας ήταν και η αξιολόγηση της γραφικής αναπαράστασης των «συχνών στοιχείων» και το κατά πόσον θα μπορούσε να βοηθήσει μια τέτοια αναπαράσταση στην παρατήρηση και εξαγωγή χρήσιμων συμπερασμάτων για την αρτιότερη παρακολούθηση ενός δικτύου υπολογιστών, δημιουργήθηκαν και γραφήματα από τα παραγόμενα συχνά στοιχεία. Τέτοια θα παρατεθούν σε επόμενο κεφάλαιο, αφού αναλυθούν οι διάφορες μελέτες-έρευνες προς αυτό το σκοπό.

3.9 Support και Αριθμός Παραγόμενων «Συχνών Στοιχείων».

Με βάση τα παραγόμενα πειραματικά δεδομένα και όπως μπορεί να φανεί και από τους πίνακες της προηγούμενης παραγράφου, ο αριθμός των παραγόμενων «συχνών στοιχείων» ποικίλει με βάση την επιλεγμένη τιμή «στήριξης». Γενικά οι παρατηρήσεις των αποτελεσμάτων έδειξαν ότι όταν ο αριθμός των «συχνών στοιχείων» είναι μεγάλος είναι δύσκολη η διαδικασία εξαγωγής χρήσιμης πληροφορίας από τον παρατηρητή. Όπως προέκυψε από την παρατήρηση αυτή (καθώς και την παρατήρηση γραφημάτων που προκύπτουν από τη γραφική αναπαράσταση των «συχνών στοιχείων» και θα αναλυθούν σε επόμενο κεφάλαιο) ένας αριθμός 20 με 30 «συχνά στοιχεία» είναι ικανοποιητικός ώστε να γίνει πολύ πιο χρηστική και ευκολότερα αντλήσιμη η πληροφορία από τον διαχειριστή-παρατηρητή. Σε ένα σύστημα παρακολούθησης ο παρατηρητής-διαχειριστής που καλείται να βγάλει συμπεράσματα για την κίνηση του δικτύου του δεν θα πρέπει να δοκιμάζει τιμές «στήριξης» μέχρι να βγάλει τον επιθυμητό αριθμό «συχνών στοιχείων». Θα πρέπει να γνωρίζει εκ των προτέρων ποια τιμή (ή τιμές) «στήριξης» θα πρέπει να χρησιμοποιήσει στον αλγόριθμο Apriori.

Προέκυψε λοιπόν η ανάγκη εύρεσης ενός “threshold” τιμής για τη «στήριξη» (η καλύτερα ενός εύρους τιμών) ώστε χρησιμοποιώντας αυτόν να παράγεται σχετικά μικρός αλλά ικανοποιητικός αριθμός «συχνών στοιχείων». Για την εύρεση αυτού του threshold πραγματοποιήθηκαν δοκιμές παίρνοντας δεδομένα διαφόρων διαστημάτων (ημερών, βδομάδων και μηνών) με βάση τα οποία και τον αλγόριθμο Apriori βρέθηκε ο αριθμός «συχνών στοιχείων» για διάφορες τιμές «στήριξης». Συγκεκριμένα δημιουργήθηκαν αρχεία για εισαγωγή στον αλγόριθμο Apriori:

- 10 ανεξάρτητων ημερών (10 διαφορετικά αρχεία από 01/03/2011 μέχρι και 10/08/2011),
- 5 ανεξάρτητων εβδομάδων (5 διαφορετικά αρχεία με δεδομένα βδομάδων, 01/04/2011 έως 08/04/2011, 09/04/2011 έως 16/04/2011, 17/04/2011 έως 24/04/2011, 25/04/2011 έως 02/05/2011 και 03/05/2011 έως 10/05/2011).
- 5 ανεξάρτητων μηνών (5 διαφορετικά αρχεία με δεδομένα μηνών, Δεκέμβριος 2010, Ιανουάριος 2011, Φεβρουάριος 2011, Μάρτιος 2011 και Απρίλιος 2011).

Για τα δεδομένα αυτά παράχθηκαν «συχνά στοιχεία», maximal frequent itemsets και closed frequent itemsets για διάφορες τιμές «στήριξης». Τα αποτελέσματα εμφανίζονται στους ακόλουθους πίνακες:

SUPPORT	NUMBER OF FREQUENT ITEMSETS 1/3/2011	NUMBER OF FREQUENT ITEMSETS 2/3/2011	NUMBER OF FREQUENT ITEMSETS 3/3/2011	NUMBER OF FREQUENT ITEMSETS 4/3/2011	NUMBER OF FREQUENT ITEMSETS 5/3/2011*
0.001	914	873	989	808	1152
0.002	584	472	465	457	413
0.003	380	325	353	338	145
0.004	336	296	290	258	105
0.005	229	276	230	226	64
0.006	183	264	204	210	51
0.007	166	204	164	172	50
0.008	166	201	146	152	48
0.009	158	200	146	116	48
0.01	158	191	126	116	48

0.015	94	110	48	56	36
0.02	44	72	48	42	36
0.025	40	20	32	32	36
0.03	40	19	30	28	36
0.035	40	18	28	24	36
0.04	40	16	28	24	36
0.045	32	16	28	24	36
0.05	16	16	28	24	36
0.055	16	16	28	16	14
0.06	10	16	28	16	14

Πίνακας 3.6: Δεδομένα ημέρας (ημέρες 1 έως 5/3/2011) – Frequent Itemsets

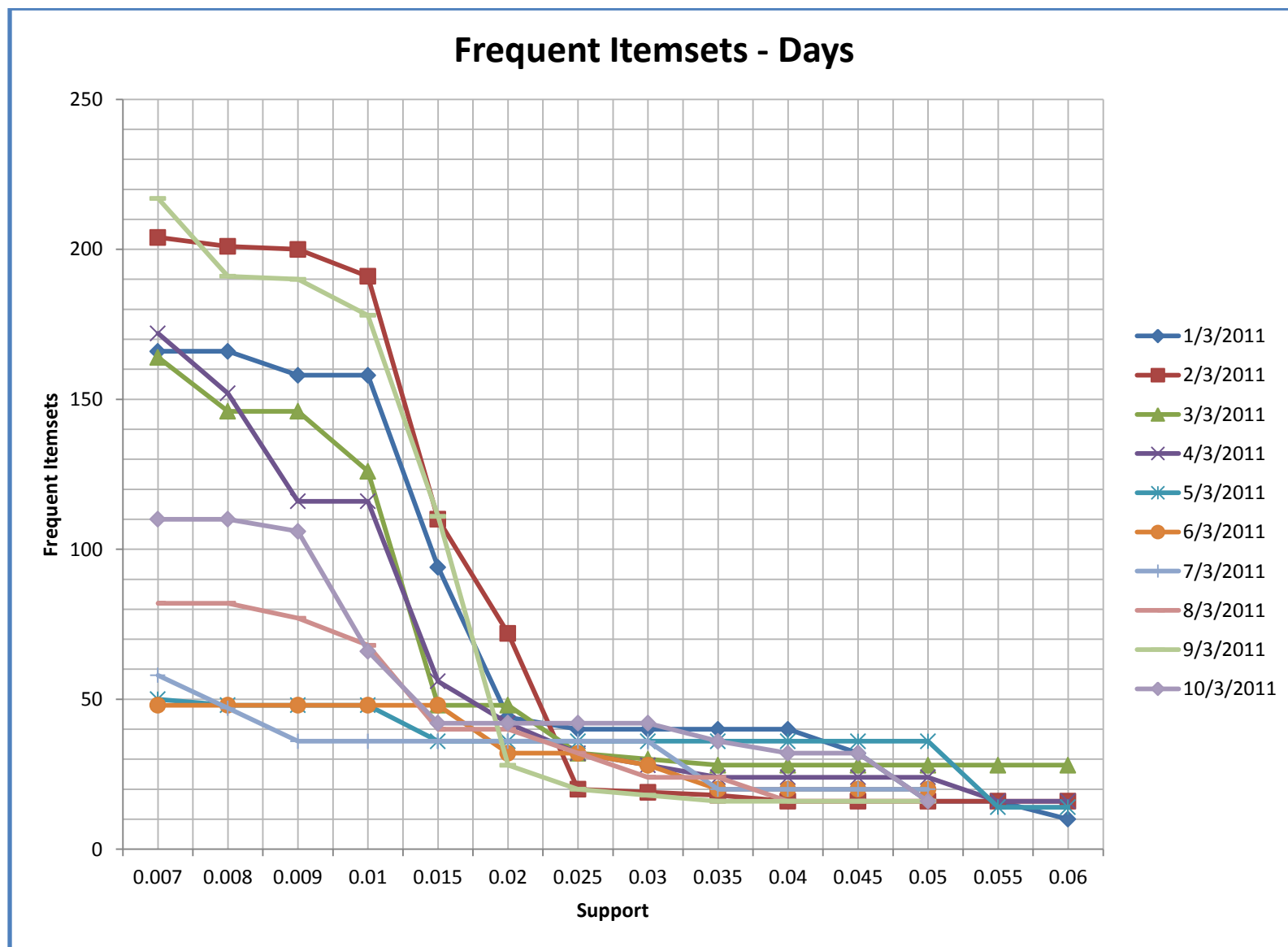
SUPPORT	NUMBER OF FREQUENT ITEMSETS 6/3/2011*	NUMBER OF FREQUENT ITEMSETS 7/3/2011	NUMBER OF FREQUENT ITEMSETS 8/3/2011	NUMBER OF FREQUENT ITEMSETS 9/3/2011	NUMBER OF FREQUENT ITEMSETS 10/3/2011
0.001	1305	1114	1190	1113	944
0.002	261	203	526	476	489
0.003	92	115	314	304	284
0.004	82	74	188	258	190
0.005	82	70	186	218	162

0.006	65	60	114	218	118
0.007	48	58	82	217	110
0.008	48	47	82	191	110
0.009	48	36	77	190	106
0.01	48	36	68	178	66
0.015	48	36	40	111	42
0.02	32	36	40	28	42
0.025	32	36	32	20	42
0.03	28	36	24	18	42
0.035	20	20	24	16	36
0.04	20	20	16	16	32
0.045	20	20	16	16	32
0.05	20	20	16	16	16

Πίνακας 3.7: Δεδομένα ημέρας (ημέρες 6 έως 10/3/2011) – Frequent Itemsets

Οι στήλες με τον * δηλώνουν ότι οι συγκεκριμένες μέρες αφορούν Σαββατοκύριακο όπου οι κίνηση του δικτύου διαφοροποιείται σχετικά με τις υπόλοιπες μέρες της εβδομάδας διότι οι χρήστες δε δουλεύουν στην εταιρεία οπότε και δε χρησιμοποιούν το υπό παρατήρηση δίκτυο υπολογιστών.

Τα δεδομένα των δύο παραπάνω πινάκων δίνονται και με τη μορφή σχεδιαγράμματος στο επόμενο σχήμα (σχήμα 3.21):

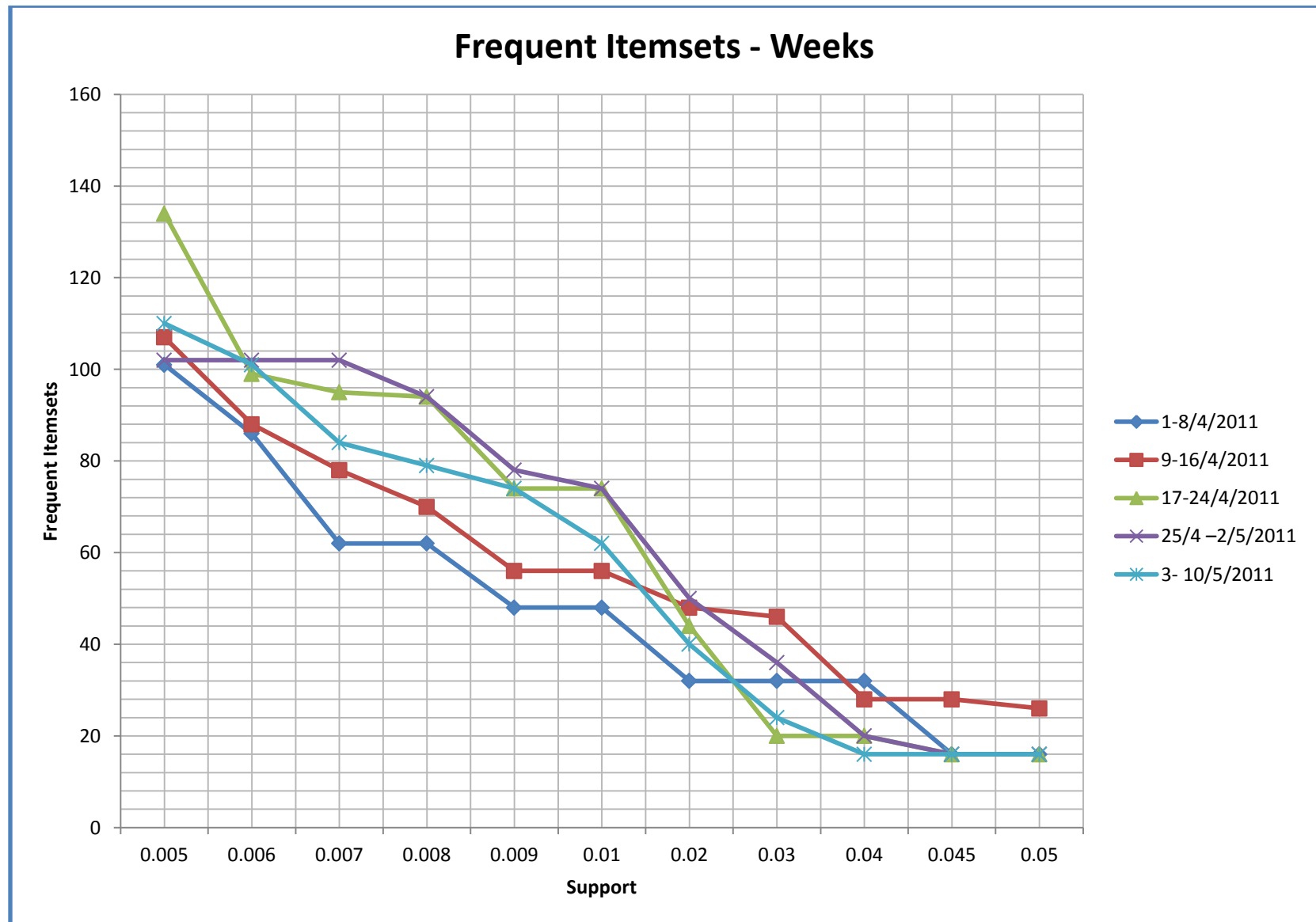


Σχήμα 3.21: Frequent Itemsets - Days

SUPPORT	NUMBER OF FREQUENT ITEMSETS 1-8/4/2011	NUMBER OF FREQUENT ITEMSETS 9-16/4/2011	NUMBER OF FREQUENT ITEMSETS 17- 24/4/2011	NUMBER OF FREQUENT ITEMSETS 25/4 – 2/5/2011	NUMBER OF FREQUENT ITEMSETS 3- 10/5/2011
0.005	101	107	134	102	110
0.006	86	88	99	102	101
0.007	62	78	95	102	84
0.008	62	70	94	94	79
0.009	48	56	74	78	74
0.01	48	56	74	74	62
0.02	32	48	44	50	40
0.03	32	46	20	36	24
0.04	32	28	20	20	16
0.045	16	28	16	16	16
0.05	16	26	16	16	16

Πίνακας 3.8: Δεδομένα εβδομάδων – Frequent Itemsets

Τα δεδομένα του παραπάνω πίνακα δίνονται και με τη μορφή σχεδιαγράμματος στο επόμενο σχήμα (σχήμα 3.22):

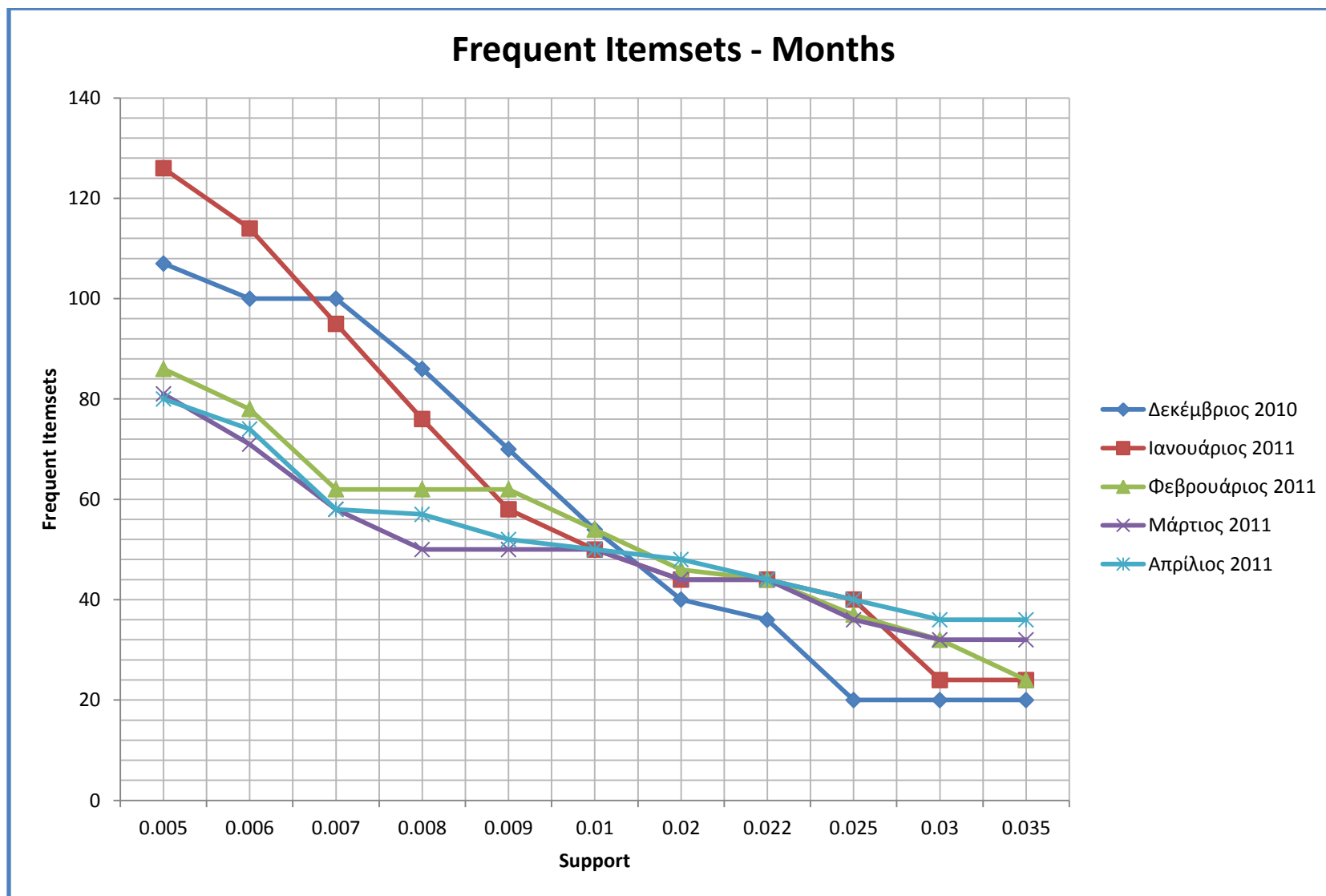


Σχήμα 3.22: Frequent Itemsets - Weeks

SUPPORT	NUMBER OF FREQUENT ITEMSETS Δεκέμβριος 2010	NUMBER OF FREQUENT ITEMSETS Ιανουάριος 2011	NUMBER OF FREQUENT ITEMSETS Φεβρουάριος 2011	NUMBER OF FREQUENT ITEMSETS Μάρτιος 2011	NUMBER OF FREQUENT ITEMSETS Απρίλιος 2011
0.005	107	126	86	81	80
0.006	100	114	78	71	74
0.007	100	95	62	58	58
0.008	86	76	62	50	57
0.009	70	58	62	50	52
0.01	54	50	54	50	50
0.02	40	44	46	44	48
0.022	36	44	44	44	44
0.025	20	40	37	36	40
0.03	20	24	32	32	36
0.035	20	24	24	32	36

Πίνακας 3.9: Δεδομένα μηνών – Frequent Itemsets

Τα δεδομένα του παραπάνω πίνακα δίνονται και με τη μορφή σχεδιαγράμματος στο επόμενο σχήμα (σχήμα3.23):



Σχήμα 3.23: Frequent Itemsets - Months

SUPPORT	NUMBER OF M. F. I. 1/3/2011	NUMBER OF M. F. I. 2/3/2011	NUMBER OF M. F. I. 3/3/2011	NUMBER OF M. F. I. 4/3/2011	NUMBER OF M. F. I. 5/3/2011*
0.001	183	157	189	155	243
0.002	100	88	80	82	83
0.003	56	50	61	53	33
0.004	47	39	44	36	25
0.005	28	37	35	32	18
0.006	24	34	32	30	14
0.007	23	26	26	26	13
0.008	23	25	21	22	13
0.009	21	24	21	18	13
0.01	21	22	21	18	13

Πίνακας 3.10: Δεδομένα ημέρας (ημέρες 1 έως 5/3/2011) – Maximal Frequent Itemsets

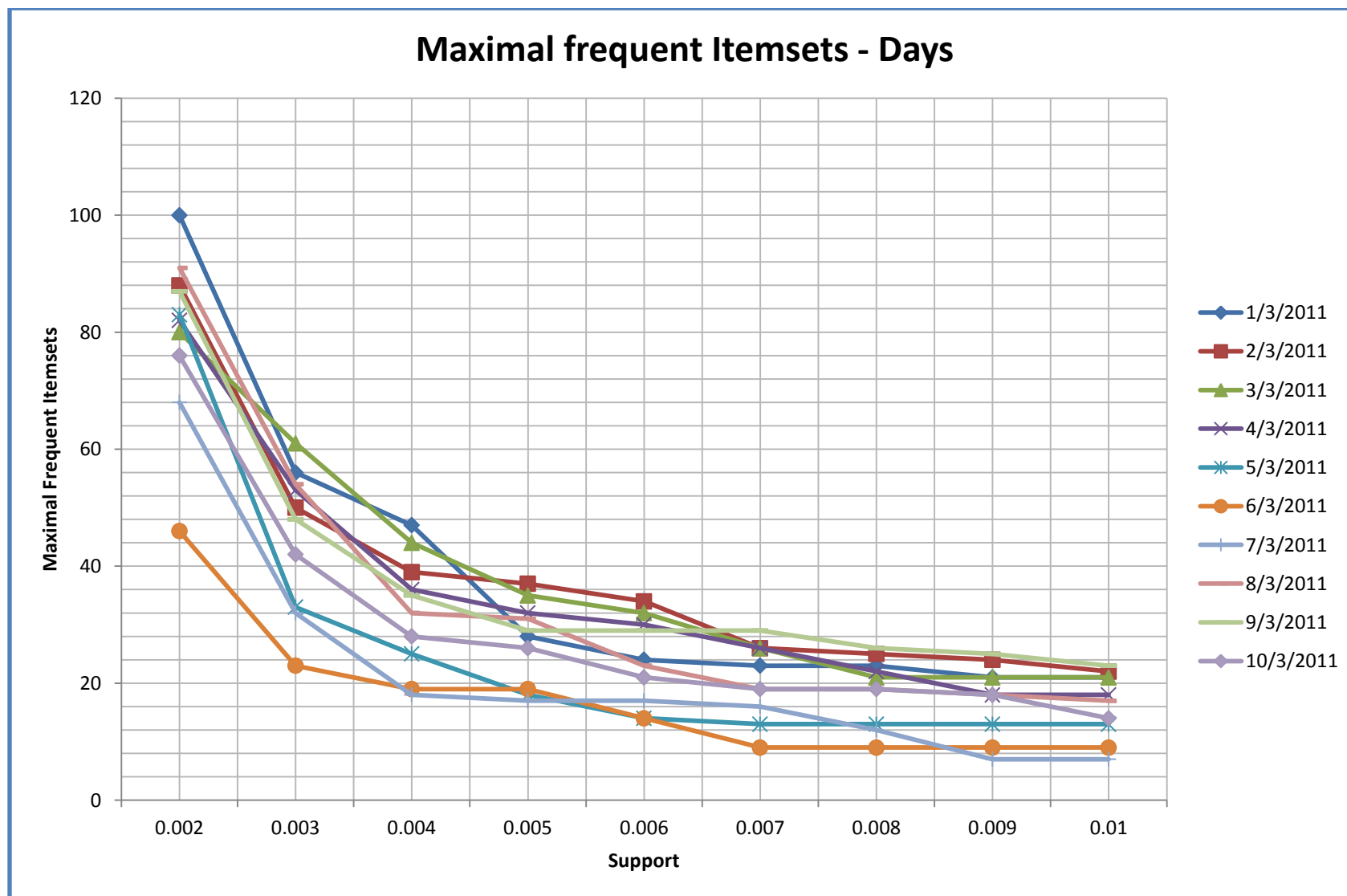
SUPPORT	NUMBER OF M. F. I. 6/3/2011*	NUMBER OF M. F. I. 7/3/2011	NUMBER OF M. F. I. 8/3/2011	NUMBER OF M. F. I. 9/3/2011	NUMBER OF M. F. I. 10/3/2011
0.001	241	214	189	212	161
0.002	46	68	91	87	76
0.003	23	32	54	48	42
0.004	19	18	32	35	28

0.005	19	17	31	29	26
0.006	14	17	23	29	21
0.007	9	16	19	29	19
0.008	9	12	19	26	19
0.009	9	7	18	25	18
0.01	9	7	17	23	14

Πίνακας 3.11: Δεδομένα ημέρας (ημέρες 6 έως 103/2011) – Maximal Frequent Itemsets

Οι στήλες με τον * δηλώνουν ότι οι συγκεκριμένες μέρες αφορούν Σαββατοκύριακο όπου οι κίνηση του δικτύου διαφοροποιείται σχετικά με τις υπόλοιπες μέρες της εβδομάδας διότι οι χρήστες δε δουλεύουν στην εταιρείας οπότε και δε χρησιμοποιούν το υπό παρατήρηση δίκτυο.

Τα δεδομένα των δύο παραπάνω πινάκων δίνονται και με τη μορφή σχεδιαγράμματος στο επόμενο σχήμα (σχήμα 3.24):

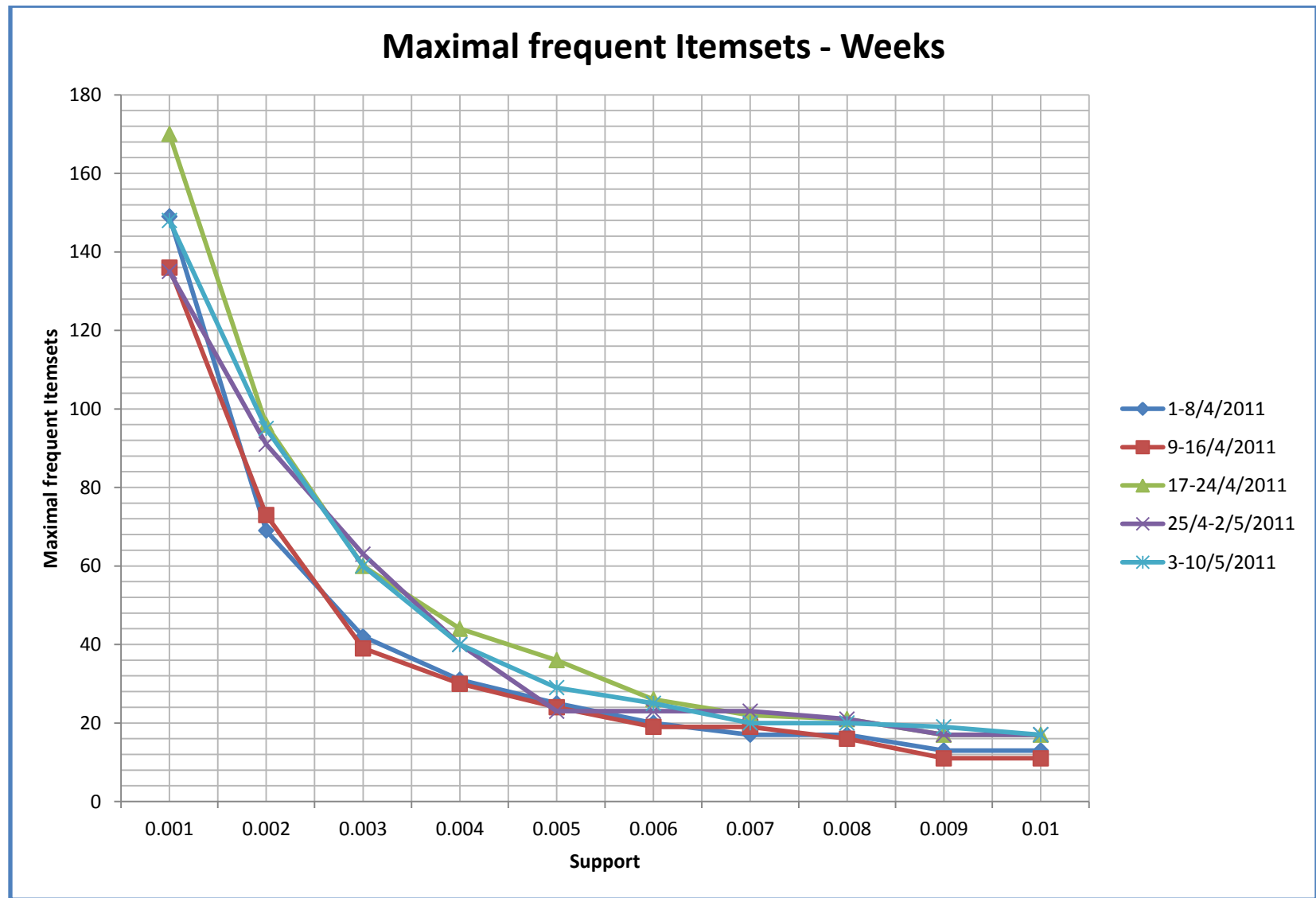


Σχήμα 3.24: Maximal Frequent Itemsets – Days

SUPPORT	NUMBER OF M. F. I. 1-8/4/2011	NUMBER OF M. F. I. 9-16/4/2011	NUMBER OF M. F. I. 17- 24/4/2011	NUMBER OF M. F. I. 25/4 – 2/5/2011	NUMBER OF M. F. I. 3- 10/5/2011
0.001	149	136	170	135	148
0.002	69	73	96	91	95
0.003	42	39	60	63	60
0.004	31	30	44	40	40
0.005	25	24	36	23	29
0.006	20	19	26	23	25
0.007	17	19	22	23	20
0.008	17	16	21	21	20
0.009	13	11	17	17	19
0.01	13	11	17	17	17

Πίνακας 3.12: Δεδομένα εβδομάδων – Maximal Frequent Itemsets

Τα δεδομένα του παραπάνω πίνακα δίνονται και με τη μορφή σχεδιαγράμματος στο επόμενο σχήμα (σχήμα3.25):

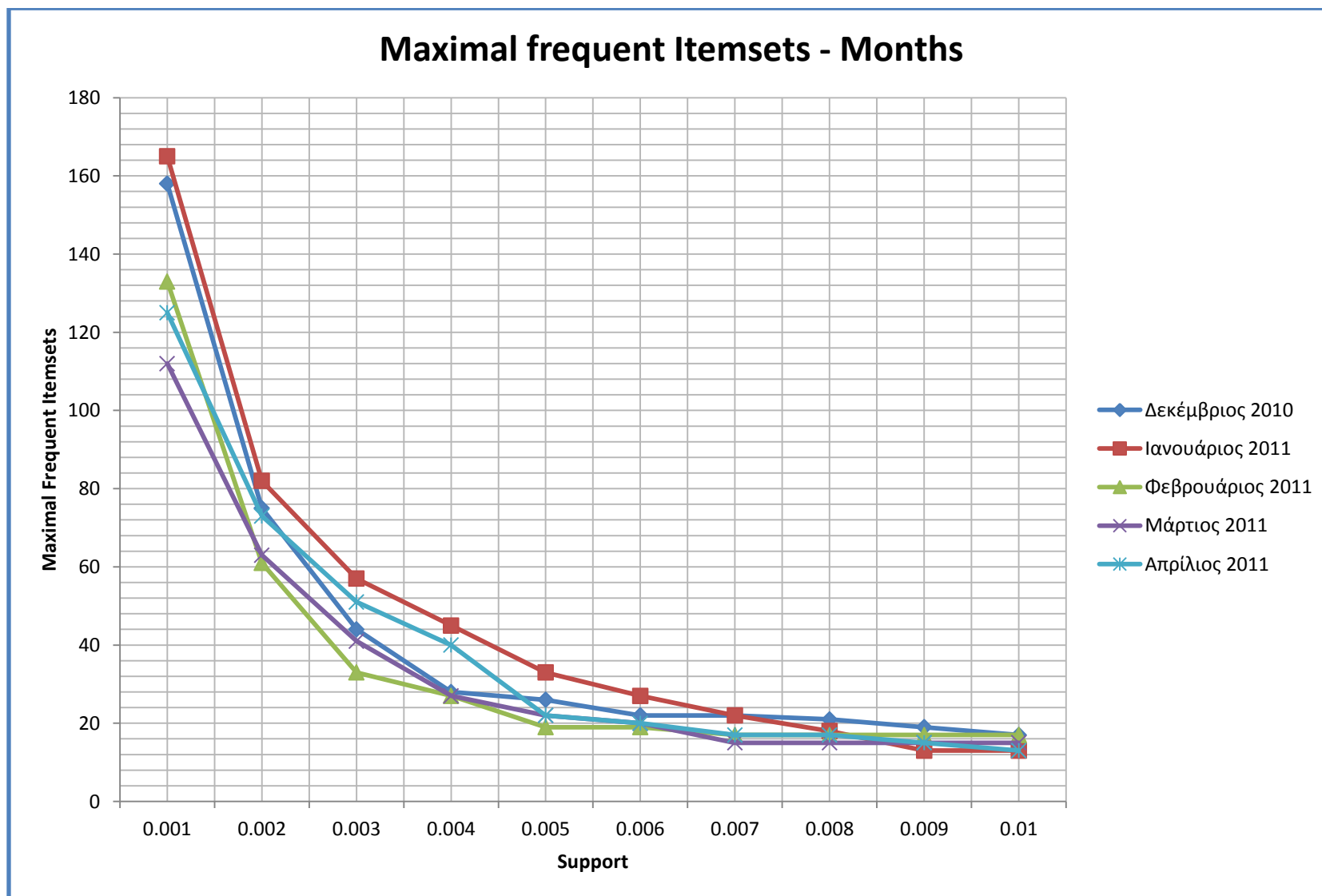


Σχήμα 3.25: Maximal Frequent Itemsets – Weeks

SUPPORT	NUMBER OF M. F. I. Δεκέμβριος 2010	NUMBER OF M. F. I. Ιανουάριος 2011	NUMBER OF M. F. I. Φεβρουάριος 2011	NUMBER OF M. F. I. Μάρτιος 2011	NUMBER OF M. F. I. Απρίλιος 2011
0.001	158	165	133	112	125
0.002	75	82	61	63	73
0.003	44	57	33	41	51
0.004	28	45	27	27	40
0.005	26	33	19	22	22
0.006	22	27	19	20	20
0.007	22	22	17	15	17
0.008	21	18	17	15	17
0.009	19	13	17	15	15
0.01	17	13	17	15	13

Πίνακας 3.13: Δεδομένα μηνών – Maximal Frequent Itemsets

Τα δεδομένα του παραπάνω πίνακα δίνονται και με τη μορφή σχεδιαγράμματος στο επόμενο σχήμα (σχήμα 3.26):



Σχήμα 3.26: Maximal Frequent Itemsets – Months

SUPPORT	NUMBER OF C. F. I. 1/3/2011	NUMBER OF C. F. I. 2/3/2011	NUMBER OF C. F. I. 3/3/2011	NUMBER OF C. F. I. 4/3/2011	NUMBER OF C. F. I. 5/3/2011*
0.001	252	247	286	251	323
0.002	151	146	150	152	129
0.003	95	98	116	103	57
0.004	86	85	93	79	40
0.005	61	79	78	69	28
0.006	51	75	71	65	24
0.007	48	61	64	56	23
0.008	48	59	58	50	21
0.009	46	58	58	42	21
0.01	46	53	53	42	21
0.02	26	26	30	22	15
0.03	22	15	20	16	15

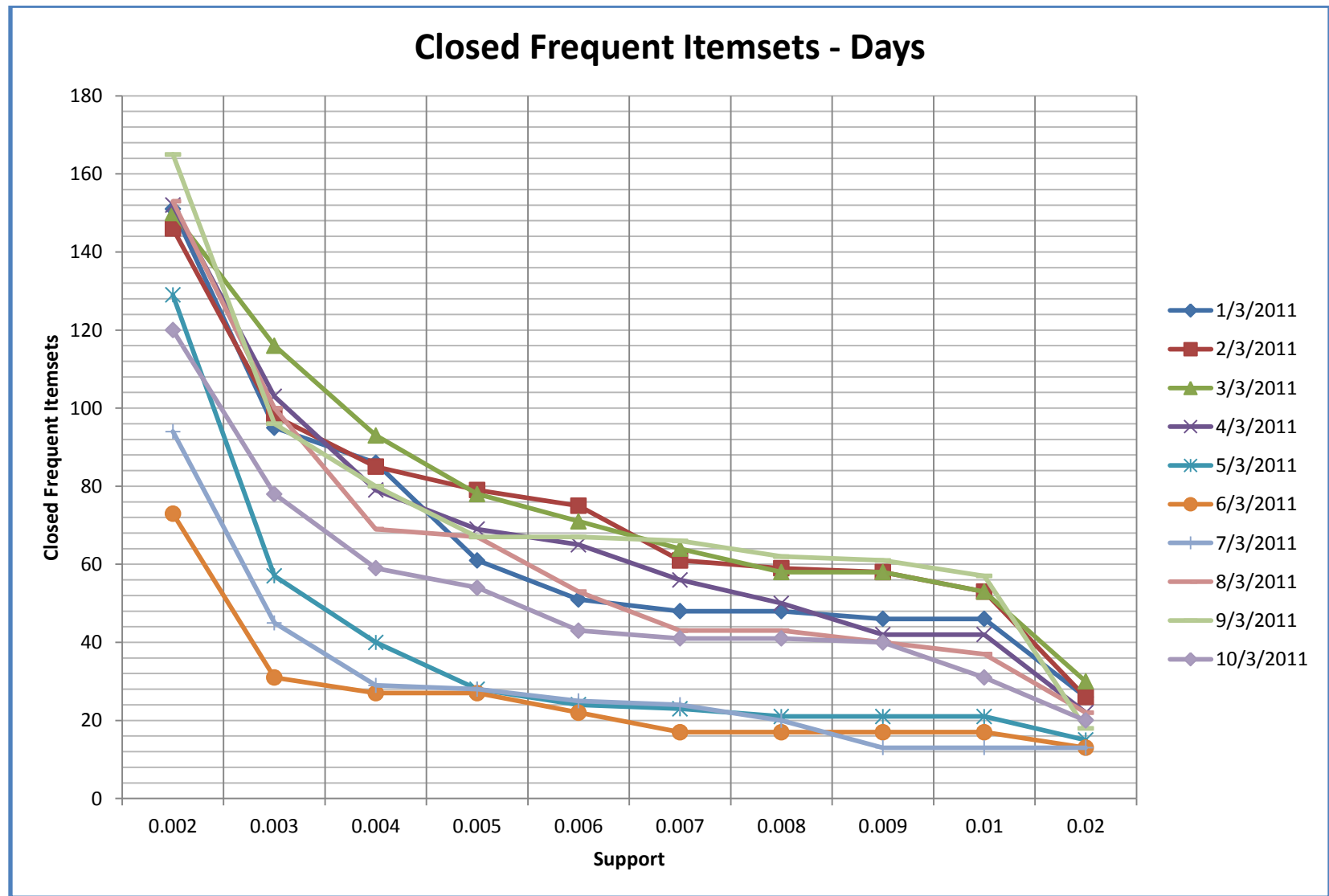
Πίνακας 3.14: Δεδομένα ημέρας (ημέρες 1 έως 5/3/2011) – Closed Frequent Itemsets

SUPPORT	NUMBER OF C. F. I. 6/3/2011*	NUMBER OF C. F. I. 7/3/2011	NUMBER OF C. F. I. 8/3/2011	NUMBER OF C. F. I. 9/3/2011	NUMBER OF C. F. I. 10/3/2011
0.001	323	259	296	330	221
0.002	73	94	153	165	120
0.003	31	45	100	96	78
0.004	27	29	69	80	59
0.005	27	28	67	67	54
0.006	22	25	53	67	43
0.007	17	24	43	66	41
0.008	17	20	43	62	41
0.009	17	13	40	61	40
0.01	17	13	37	57	31
0.02	13	13	22	18	20

Πίνακας 3.15: Δεδομένα ημέρας (ημέρες 6 έως 10/3/2011) – Closed Frequent Itemsets

Οι στήλες με τον * δηλώνουν ότι οι συγκεκριμένες μέρες αφορούν Σαββατοκύριακο όπου οι κίνηση του δικτύου διαφοροποιείται σχετικά με τις υπόλοιπες μέρες της εβδομάδας διότι οι χρήστες δε δουλεύουν στην εταιρεία οπότε και δε χρησιμοποιούν το υπό παρατήρηση δίκτυο υπολογιστών.

Τα δεδομένα των δύο παραπάνω πινάκων δίνονται και με τη μορφή σχεδιαγράμματος στο επόμενο σχήμα (σχήμα 3.27):

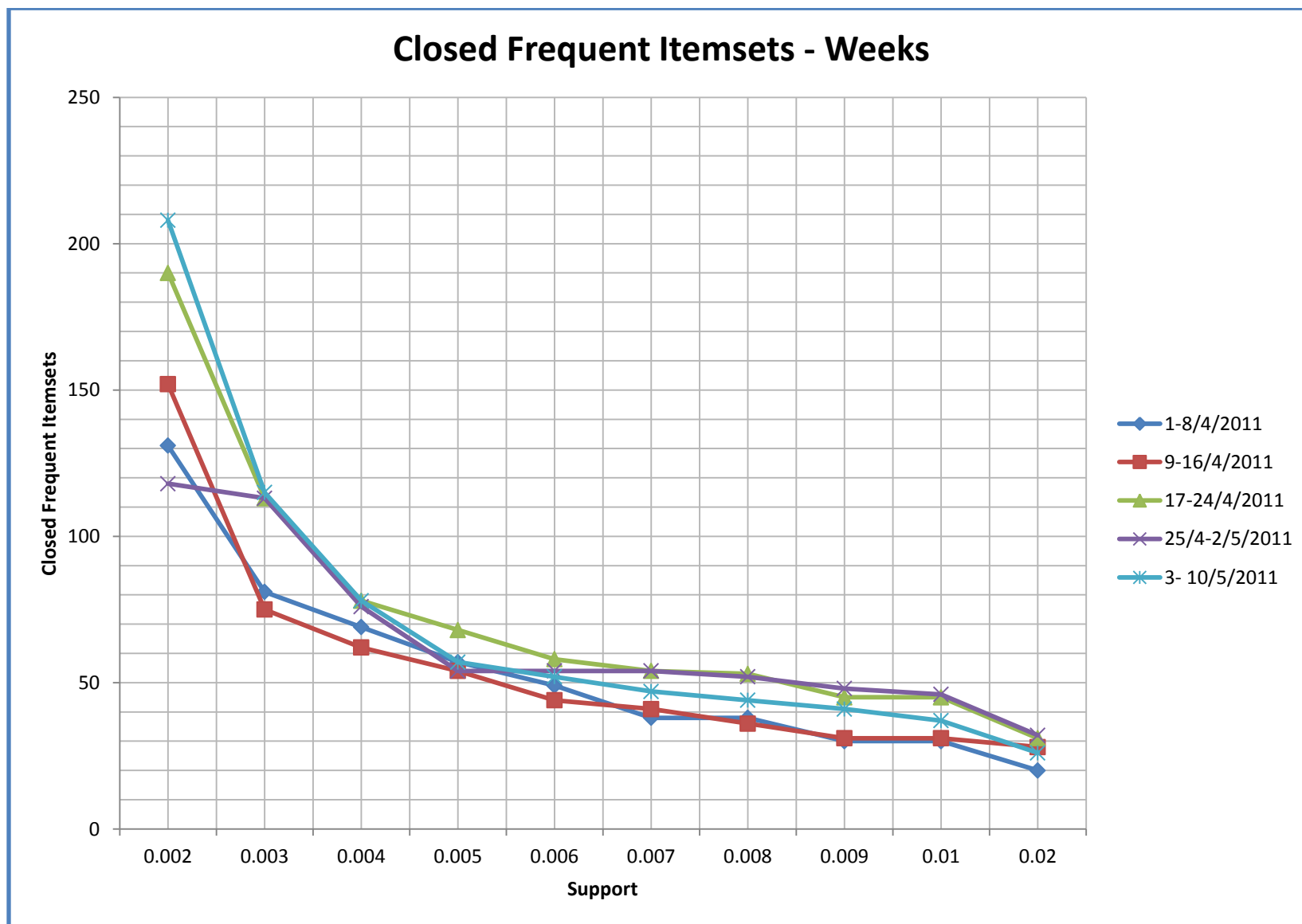


Σχήμα 3.27: Closed Frequent Itemsets – Days

SUPPORT	NUMBER OF C. F. I. 1-8/4/2011	NUMBER OF C. F. I. 9-16/4/2011	NUMBER OF C. F. I. 17- 24/4/2011	NUMBER OF C. F. I. 25/4 – 2/5/2011	NUMBER OF C. F. I. 3- 10/5/2011
0.001	285	287	319	285	308
0.002	131	152	190	118	208
0.003	81	75	113	113	115
0.004	69	62	78	76	78
0.005	57	54	68	54	57
0.006	49	44	58	54	52
0.007	38	41	54	54	47
0.008	38	36	53	52	44
0.009	30	31	45	48	41
0.01	30	31	45	46	37
0.02	20	28	31	32	26

Πίνακας 3.16: Δεδομένα εβδομάδων – Closed Frequent Itemsets

Τα δεδομένα του παραπάνω πίνακα δίνονται και με τη μορφή σχεδιαγράμματος στο επόμενο σχήμα (σχήμα 3.28):

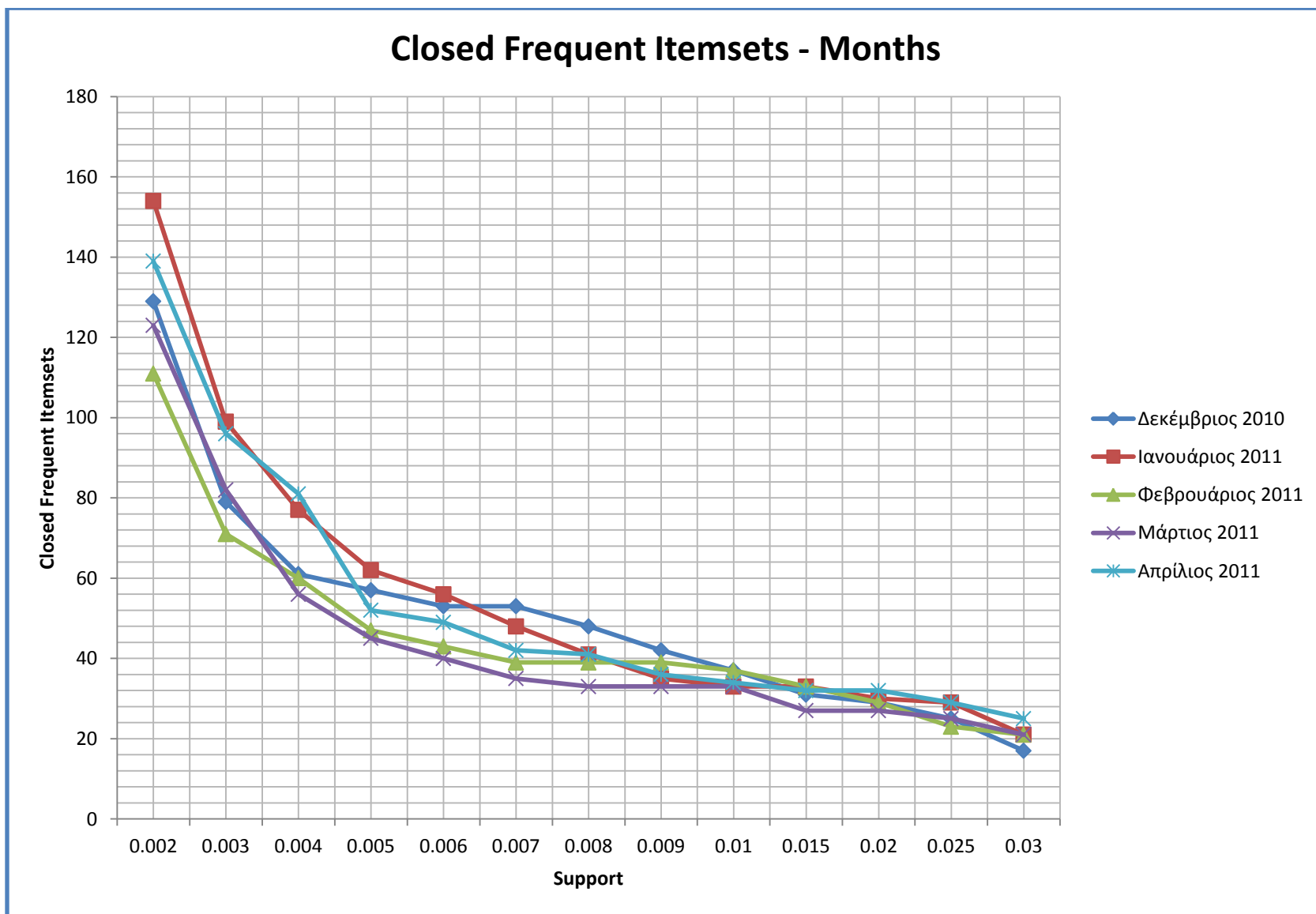


Σχήμα 3.28: Closed Frequent Itemsets – Weeks

SUPPORT	NUMBER OF C. F. I. Δεκέμβριος 2010	NUMBER OF C. F. I. Ιανουάριος 2011	NUMBER OF C. F. I. Φεβρουάριος 2011	NUMBER OF C. F. I. Μάρτιος 2011	NUMBER OF C. F. I. Απρίλιος 2011
0.001	279	353	228	209	223
0.002	129	154	111	123	139
0.003	79	99	71	82	96
0.004	61	77	60	56	81
0.005	57	62	47	45	52
0.006	53	56	43	40	49
0.007	53	48	39	35	42
0.008	48	41	39	33	41
0.009	42	35	39	33	36
0.01	37	33	37	33	34
0.015	31	33	33	27	32
0.02	29	30	29	27	32
0.025	25	29	23	25	29
0.03	17	21	21	21	25

Πίνακας 3.17: Δεδομένα μηνών – Closed Frequent Itemsets

Τα δεδομένα του παραπάνω πίνακα δίνονται και με τη μορφή σχεδιαγράμματος στο επόμενο σχήμα (σχήμα 3.29):



Σχήμα 3.29: Closed Frequent Itemsets – Months

Με βάση τα παραπάνω δεδομένα καταλήγουμε στα εξής συμπεράσματα που έχουν να κάνουν με την επιλογή της τιμής «στήριξης»:

Δεδομένα ημερών

Σε περίπτωση που χρειάζεται να δημιουργηθούν περίπου είκοσι «συχνά στοιχεία» (συγκεκριμένα 20 με 24 «συχνά στοιχεία») για δεδομένα που προέρχονται από καταγραφή κίνησης μίας μέρας καλό είναι να χρησιμοποιήσει τιμές «στήριξης» από 0.05 έως 0.04.

Σε περίπτωση που χρειάζεται να εξαχθούν περίπου είκοσι maximal frequent itemsets (συγκεκριμένα από 16 μέχρι 20 maximal frequent itemsets) για δεδομένα που προέρχονται από καταγραφή κίνησης μίας μέρας καλό είναι να χρησιμοποιήσει τιμή «στήριξης» 0.007 έως 0.01.

Σε περίπτωση που χρειάζεται να δημιουργηθούν περίπου είκοσι closed frequent itemsets για δεδομένα που προέρχονται από καταγραφή κίνησης μίας μέρας καλό είναι να χρησιμοποιήσει τιμή «στήριξης» 0.02.

Δεδομένα εβδομάδας

Σε περίπτωση που χρειάζεται να εξαχθούν περίπου είκοσι «συχνά στοιχεία» για δεδομένα που προέρχονται από καταγραφή κίνησης μίας εβδομάδας καλό είναι να χρησιμοποιήσει τιμή «στήριξης» μεταξύ 0.04 και 0.05.

Σε περίπτωση που χρειάζεται να δημιουργηθούν λιγότερα από είκοσι maximal frequent itemsets για δεδομένα που προέρχονται από καταγραφή κίνησης μίας εβδομάδας καλό είναι να χρησιμοποιήσει τιμή «στήριξης» μεταξύ 0.008 και 0.01.

Σε περίπτωση που χρειάζεται να δημιουργηθούν λιγότερα από τριάντα closed frequent itemsets για δεδομένα που προέρχονται από καταγραφή κίνησης μίας εβδομάδας καλό είναι να χρησιμοποιήσει τιμή «στήριξης» περίπου 0.02.

Δεδομένα μήνα

Σε περίπτωση που χρειάζεται να εξαχθούν μεταξύ είκοσι και τριάντα «συχνά στοιχεία» για δεδομένα που προέρχονται από καταγραφή κίνησης ενός μήνα καλό είναι να χρησιμοποιήσει τιμή «στήριξης» 0.03.

Σε περίπτωση που χρειάζεται να δημιουργηθούν λιγότερα από είκοσι maximal frequent itemsets για δεδομένα που προέρχονται από καταγραφή κίνησης ενός μήνα καλό είναι να χρησιμοποιήσει τιμή «στήριξης» μεταξύ 0.007 και 0.01.

Σε περίπτωση που χρειάζεται να δημιουργηθούν περίπου είκοσι closed frequent itemsets για δεδομένα που προέρχονται από καταγραφή κίνησης ενός μήνα καλό είναι να χρησιμοποιήσει τιμή «στήριξης» 0.03.

Θα πρέπει να τονιστεί για ακόμη μία φορά ότι τα νούμερα αυτά προέρχονται από εφαρμογή του αλγορίθμου Apriori στα δεδομένα καταγραφείσας κίνησης του συγκεκριμένου δικτύου υπολογιστών, το οποίο αποτέλεσε το δίκτυο πειραματισμού της συγκεκριμένης μεταπτυχιακής διατριβής. Ενδεχομένως αυτά τα αποτελέσματα να διαφοροποιούνται σε κάποιο άλλο δίκτυο και μάλιστα σε δεδομένα τα οποία παρουσιάζουν μεγαλύτερη διαφοροποίηση μεταξύ τους.

Κεφάλαιο 4

Τεχνικές Οπτικής

Αναπαράστασης «Συχνών

Στοιχείων».

Στην ενότητα αυτή θα πραγματοποιηθεί μία αναφορά στις τεχνικές που υπάρχουν διαθέσιμες γενικά για την οπτική αναπαράσταση (visualization techniques) «συχνών στοιχείων» και θα γίνει παράθεση ενδεικτικών ερευνών και εργασιών προς σε αυτό τον τομέα. Πρόσθετα θα πραγματοποιηθεί μικρή αναφορά στην εφαρμοσιμότητά που μπορεί να έχουν τέτοιες τεχνικές στην αναπαράσταση «συχνών στοιχείων» που προέρχονται από κίνηση σε ένα δίκτυο.

4.1 Εισαγωγή

Γενικά είναι ευρύτερα αποδεκτό το γεγονός ότι η οπτικοποίηση (διαγραμματικός τρόπος εμφάνισης) δεδομένων είναι τις περισσότερες φορές πιο αποδοτικός τρόπος εμφάνισης σε σχέση με την απλή παράθεση αυτών. Τις περισσότερες φορές η γραφική αναπαράσταση των δεδομένων δίνει τη δυνατότητα για πιο εύκολη άντληση χρήσιμων πληροφοριών χρησιμοποιώντας καλύτερα την αντιληπτική ικανότητα του ανθρώπου, με απώτερο στόχο να βοηθήσει στην ευκολότερη λήψη αποφάσεων. Όπως λέει και το ρητό «μία εικόνα είναι χίλιες λέξεις». Η γραφική αναπαράσταση δίνει τη δυνατότητα να αντληθούν πιο γρήγορα πληροφορίες και συμπεράσματα. Φυσικά αυτό πάντοτε εξαρτάται από το κατά πόσο ο τρόπος απεικόνισης που έχει επιλεγεί είναι ο καταλληλότερος και πόσο ικανοποιητική είναι η απεικόνιση των περιγραφόμενων πληροφοριών. Όπως έχει αναφερθεί και σε προηγούμενο κεφάλαιο την αποδοτικότητα αυτή της απεικόνισης επηρεάζουν πολλοί παράγοντες όπως:

- Το πλήθος των πληροφοριών που απεικονίζονται στο σχεδιάγραμμα. Όταν οι πληροφορίες που εμφανίζονται σε μία εικόνα είναι πάρα πολλές τότε απαιτείται μεγαλύτερη προσπάθεια από τον παρατηρητή για την άντληση των απαιτούμενων πληροφοριών πετυχαίνοντας το ακριβώς αντίθετο από το επιθυμητό αποτέλεσμα.
- Η καταλληλότητα του επιλεχθέντος τρόπου απεικόνισης και το κατά πόσον αυτός ταιριάζει με τα υπό απεικόνιση δεδομένα. Ένα γράφημα πίτας (pie chart) απεικονίζει ικανοποιητικότερα την ποσοστιαία κατανομή κάποιου γεγονότος από ό,τι ένα γράφημα στηλών (bar chart or column chart) ή γραμμών (line chart).
- Το είδος των πληροφοριών που απαιτείται να αντληθούν, πχ όταν σε ένα γράφημα πίτας απεικονίζονται πολλές πληροφορίες με παρόμοια μεγέθη είναι λίγο δύσκολο να πραγματοποιηθεί σύγκριση μεταξύ των τιμών των πληροφοριών, οπότε σε αυτή την περίπτωση το γράφημα πίτας ίσως να μην είναι το καταλληλότερο για την άντληση εύκολα της πληροφορίας αυτής.

Μία διαδικασία (αλγόριθμος) εξόρυξης δεδομένων έχει ως αποτέλεσμα την παραγωγή «συχνών στοιχείων» τα οποία είναι κατ' ουσίαν ομάδες στοιχείων. Για την εμφάνιση αυτών των ομάδων στοιχείων θα μπορούσε να γίνει απλή παράθεση αυτών σχηματίζοντας π.χ. έναν πίνακα ή μία λίστα τιμών, αλλά θα μπορούσε να χρησιμοποιηθεί και κάποια μέθοδος γραφικής απεικόνισης αυτών,

προσπαθώντας να βελτιωθεί ο αριθμός, η ποιότητα καθώς και η ευκολία εξαγωγής συμπερασμάτων.

Στην ενότητα αυτή θα παρατεθούν διάφορες μελέτες – εργασίες που στόχο έχουν την επίτευξη ικανοποιητικής γραφικής αναπαράστασης «συχνών στοιχείων» και «κανόνων συσχέτισης» που πηγάζουν από αλγορίθμους εξόρυξης δεδομένων. Πρόσθετα θα πραγματοποιηθεί σύγκριση αυτών με γνώμονα την εφαρμοσιμότητά τους στην αναπαράσταση των δεδομένων («συχνών στοιχείων») που προέρχονται από κίνηση σε ένα δίκτυο υπολογιστών.

4.2 Τεχνικές Εμφάνισης Συχνά Εμφανιζόμενων Προτύπων

Γενικά, όπως έχει αναφερθεί και στο δεύτερο κεφάλαιο της παρούσας μεταπτυχιακής διατριβής, μέχρι σήμερα έχουν αναπτυχθεί διάφορες τεχνικές για την παραγωγή των πιο συχνά εμφανιζόμενων (επαναλαμβανόμενων) προτύπων (patterns). Τα σύνολα αυτά ονομάζονται «συχνά στοιχεία» ή “frequent itemsets”. Οι περισσότερες από αυτές τις τεχνικές εμφανίζουν τα αποτελέσματά τους με τη μορφή απλού κειμένου (plain text). Οι υπόλοιπες προσπαθούν με γραφικό τρόπο να αναδείξουν τη γνώση που μπορεί κάποιος να λάβει από την επαναλαμβανόμενη εμφάνιση αυτών των προτύπων (“frequent itemsets” και “association rules”). Τα συστήματα οπτικής απεικόνισης (visualization systems) που σχεδιάστηκαν και υλοποιήθηκαν προς αυτό το στόχο, την ανάδειξη δηλαδή της γνώσης (Knowledge Discovery) χωρίζονται σε δύο κατηγορίες:

1. Στα συστήματα που εστιάζουν στην οπτική απεικόνιση με στόχο την ανάδειξη της συσχέτισης των μεταβλητών-παραμέτρων του αλγορίθμου παραγωγής «συχνών στοιχείων», π.χ. support σε σχέση με τον αριθμό των στοιχείων.
2. Στα συστήματα που παρουσιάζουν και αναλύουν τις συχνά εμφανιζόμενες ομάδες αντικειμένων (τα «συχνά στοιχεία») με στόχο να κάνουν πιο εύκολη και αξιόπιστη την εξαγωγή συμπερασμάτων.

Στην παρούσα μεταπτυχιακή διατριβή γίνεται αναφορά στα συστήματα της δεύτερης κατηγορίας. Η δυσκολία των μεθόδων γραφικής απεικόνισης «συχνών στοιχείων» πηγάζει από τη μορφή των δεδομένων («συχνά στοιχεία») που πρέπει να απεικονιστούν. Το σύνολο των «συχνών στοιχείων» δεν είναι τίποτε άλλο από μία σχέση αντικειμένων της μορφής «πολλά προς πολλά» γεγονός που

καθιστά δύσκολη την εποικοδομητική απεικόνισή τους. Η πρόκληση είναι η απεικόνιση ταυτόχρονα πολλών ομάδων στοιχείων ώστε ο παρατηρητής να μπορεί εύκολα να αναγνωρίσει τις πιο ενδιαφέρουσες σχέσεις μεταξύ αυτών. Ταυτόχρονα η απεικόνιση θα πρέπει να είναι τέτοια ώστε να μη χαθεί ή αποκρυφτεί καμία από τις παρεχόμενες πληροφορίες-συμπεράσματα.

Ένας από τους στόχους της συγκεκριμένης μεταπτυχιακής διατριβής είναι η εφαρμογή τεχνικών γραφικής απεικόνισης (όπως αυτές της προαναφερόμενης δεύτερης κατηγορίας) σε «συχνά στοιχεία» που προέρχονται από την εφαρμογή αλγορίθμων εξόρυξης «συχνών στοιχείων» σε δεδομένα καταγραφείσας κίνησης δικτύου υπολογιστών, όσο και η εξεύρεση της καταλληλότερης από αυτές.

Κατηγορίες τεχνικών γραφικών απεικονίσεων «συχνών στοιχείων» και «κανόνων συσχέτισης» είναι οι ακόλουθες:

4.2.1 Γραφήματα Παράλληλων Συντεταγμένων (Parallel Coordinates)

Μία κατηγορία εργαλείων για τη γραφική αναπαράσταση «συχνών στοιχείων» είναι η χρήση γραφημάτων παράλληλων συντεταγμένων (Parallel coordinates). Ενδεικτικές εργασίες-μελέτες προς αυτή την κατεύθυνση είναι οι ακόλουθες:

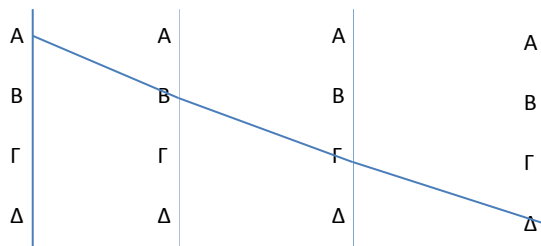
Li Yang (2003). Visualizing Frequent Itemsets, Association Rules, and Sequential Patterns in Parallel Coordinates. ICCSA'03 Proceedings of the 2003 international conference on Computational science and its applications, Berlin, Heidelberg 2003 [63].

Li Yang (2008). Visual Exploration of Frequent Itemsets and Association Rules. Published in book: Visual Data Mining, Springer-Verlag Berlin, Heidelberg ©2008 [64].

Η δουλειά που παρουσιάζεται από τα παραπάνω άρθρα χρησιμοποιεί για την απεικόνιση «συχνών στοιχείων» και «κανόνων συσχέτισης» γραφήματα παράλληλων συντεταγμένων (parallel coordinates). Ο τρόπος αυτός περιγράφεται από τον ερευνητή ως «ένας εύκολος και κομψός τρόπος να παρουσιαστούν σχετιζόμενα δεδομένα σε εύκολα αναγνωρίσιμη μορφή δύο διαστάσεων» [63]. Στη συγκεκριμένη εργασία κατασκευάστηκε και χρησιμοποιήθηκε στην πράξη ένα εργαλείο το

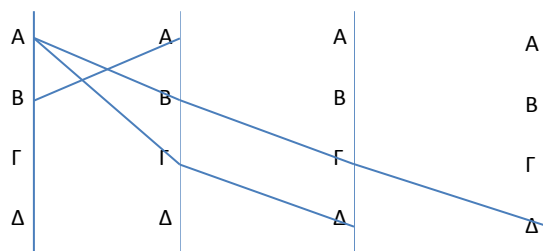
οποίο δίνει τη δυνατότητα δημιουργίας τέτοιων γραφημάτων (παράλληλων συντεταγμένων) για την απεικόνιση «συχνών στοιχείων» καθώς και «κανόνων συσχέτισης».

Ως τρόπος για τη γραφική αναπαράσταση «συχνών στοιχείων» (“frequent itemsets”), τα γραφήματα παράλληλων συντεταγμένων παρουσιάζουν μεγάλο ενδιαφέρον. Πρώτο βήμα στη διαδικασία σχεδίασης της γραφικής παράστασης είναι η εύρεση των “frequent itemsets”. Αφού βρεθούν αυτά απομονώνονται τα ανεξάρτητα στοιχεία που συνθέτουν τα “frequent itemsets”, π.χ. αν τα frequent itemsets είναι τα $\{AB\Gamma\}$, $\{AB\Delta\}$ και $\{B\Delta\}$ τότε τα ανεξάρτητα στοιχεία θα είναι τα Α, Β, Γ, Δ. Αυτά ταξινομούνται με βάση τη συχνότητα εμφάνισής τους και τοποθετούνται στις καθέτους του σχεδιαγράμματος. Στο σχεδιάγραμμα που θα δημιουργηθεί υπάρχουν τόσες κάθετοι όσες και ο αριθμός των στοιχείων του μεγαλύτερου “frequent itemset”. Αν δηλαδή το μεγαλύτερο “frequent itemset” αποτελείται από τέσσερα στοιχεία τότε θα σχεδιαστεί ένα γράφημα με τέσσερις καθέτους. Για την αναπαράσταση κάθε “frequent itemset” σχεδιάζεται γραμμή που συνδέει τις καθέτους ενώνοντας τα στοιχεία που συνθέτουν το “frequent itemset” μεταξύ τους. Παράδειγμα εμφανίζεται στο ακόλουθο σχήμα (σχήμα 4.1) όπου απεικονίζεται γράφημα παράλληλων συντεταγμένων του “frequent itemset” $\{AB\Gamma\Delta\}$.



Σχήμα 4.1: Γραφική αναπαράσταση «συχνού στοιχείου» $\{AB\Gamma\Delta\}$ με τη μέθοδο παράλληλων συντεταγμένων

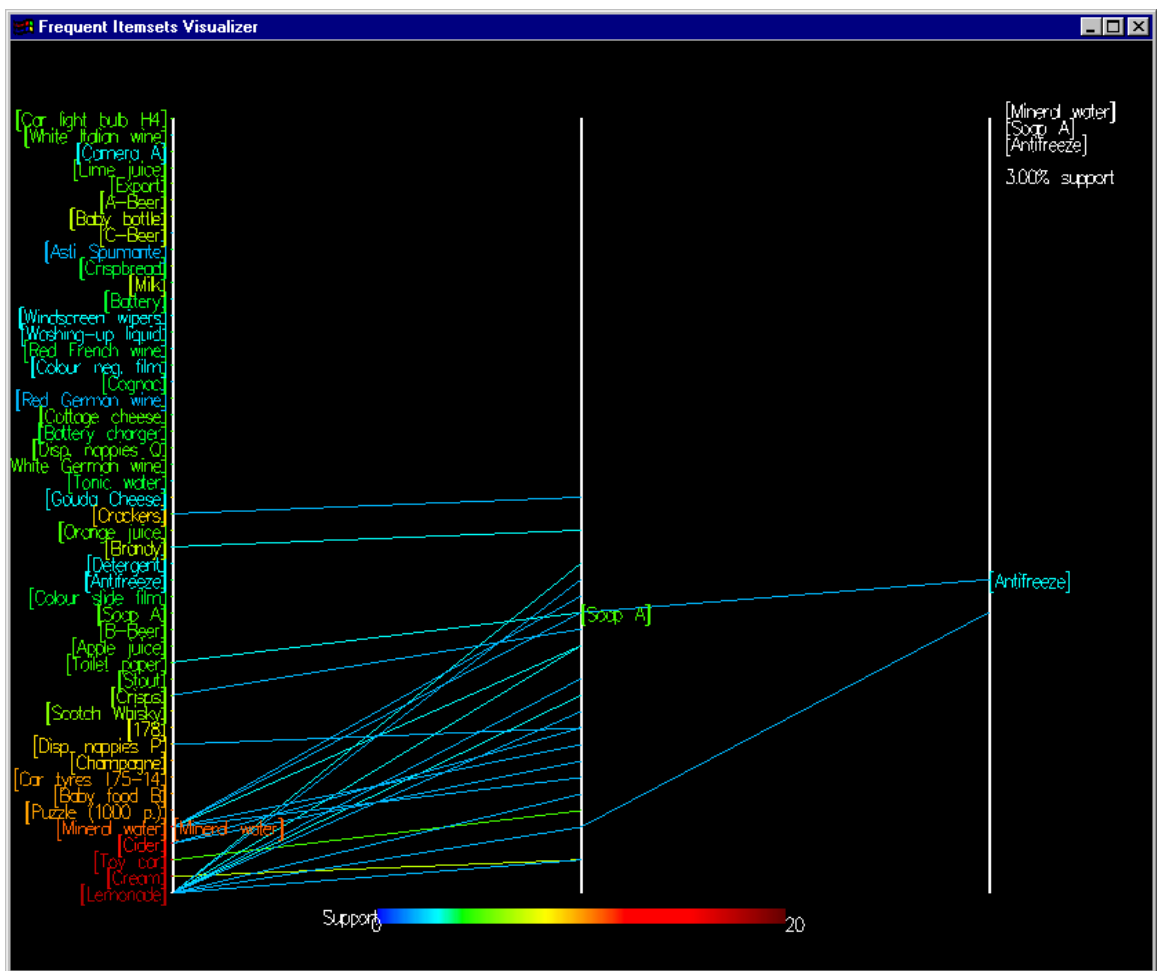
Στο γράφημα αυτό το στοιχείο Α εμφανίζεται πρώτο (από πάνω) στις καθέτους διότι έχει τη μεγαλύτερη συχνότητα εμφάνισης. Στο σχήμα 4.2 εμφανίζεται ένα αντίστοιχο σχεδιάγραμμα με περισσότερα συχνά στοιχεία. Συγκεκριμένα στο σχεδιάγραμμα αυτό εμφανίζονται τρία «συχνά στοιχεία» τα $\{AB\Gamma\Delta\}$, $\{A\Gamma\Delta\}$ και $\{AB\}$.



Σχήμα 4.2: Γραφική αναπαράσταση «συχνών στοιχείων» $\{AB\Gamma\Delta\}$, $\{A\Gamma\Delta\}$ και $\{AB\}$ με τη μέθοδο παράλληλων συντεταγμένων

Κατά την παραγωγή “frequent itemsets” σημασία στην εξαγωγή συμπερασμάτων και γενικότερα «γνώσης» παίζει και η τιμή της στήριξης του κάθε “frequent itemset”. Όσο μεγαλύτερη είναι αυτή η τιμή τόσο μεγαλύτερη αξία έχει αυτό το “frequent itemset”, διότι στην πράξη εμφανίζεται πολλές φορές μεταξύ των δεδομένων. Θα πρέπει λοιπόν και αυτή η πληροφορία (της τιμής «στήριξης» του κάθε “frequent itemset”) να αναπαριστάται στο γράφημα. Αυτό γίνεται με τη χρήση διαφορετικών χρωμάτων για κάθε διαφορετική τιμή «στήριξης», οπότε και κάθε γραμμή που δηλώνει ένα “frequent itemset” παίρνει το κατάλληλο χρώμα με βάση την τιμή «στήριξης» του αντίστοιχου “frequent itemset”. Αντίστοιχα εναλλακτικός τρόπος δήλωσης της τιμής «στήριξης» είναι το πάχος της γραμμής που δηλώνει ένα “frequent itemset”. Ανάλογα δηλαδή με το πάχος που έχει μία γραμμή δηλώνεται και η τιμή «στήριξης» του αντίστοιχου “frequent itemset” που περιγράφεται από την γραμμή αυτή.

Τελικά μία ολοκληρωμένη εικόνα η οποία αναπαριστά σύνολο “frequent itemsets” που προέρχονται από μία βάση δεδομένων με αγορές από super market είναι η ακόλουθη (σχήμα 4.3):



Σχήμα 4.3: Αποτέλεσμα εργαλείου για τη αναπαράσταση «συχνών στοιχείων» με τη χρήση γραφήματος παράλληλων συντεταγμένων. Εικόνα από το PDF με τίτλο “Visualizing Frequent Itemsets, Association Rules, and Sequential Patterns in Parallel Coordinates” - Li Yang – 2003 [63].

Μειονεκτήματα που θα μπορούσε κάποιος να εντοπίσει στο συγκεκριμένο τρόπο απεικόνισης «συχνών στοιχείων» είναι τα εξής:

- Υπάρχει δυσκολία στον ξεκάθαρο εντοπισμό των «συχνών στοιχείων». Το πρόβλημα μάλιστα επιδεινώνεται στην περίπτωση που ο αριθμός των «συχνών στοιχείων» είναι πολύ μεγάλος.
- Υπάρχει δυσκολία στον εντοπισμό με ακρίβεια της τιμής «στήριξης» του κάθε «συχνού στοιχείου» τόσο στην περίπτωση που θα χρησιμοποιηθεί χρωματικός κώδικας όσο και στην περίπτωση χρήσης πάχους γραμμών για την αναπαράσταση της τιμής αυτής.

- Μπορεί τα μεμονωμένα στοιχεία να εμφανίζονται στις καθέτους ταξινομημένα με βάση της συχνότητα εμφάνισής τους αλλά δεν προσδιορίζεται ακριβώς ποια είναι η τιμή τους. Ακόμη περισσότερο δεν αναπαρίσταται η διαφορά μεταξύ των τιμών συχνοτήτων αυτών, π.χ. μπορεί το Α εμφανίζεται 3 φορές στα «συχνά στοιχεία» που αναπαρίστανται από το σχήμα 2 και τα Β, Γ και Δ να εμφανίζονται 2 φορές, παρόλα αυτά η απόστασή μεταξύ τους στον κάθετο άξονα δεν διαφέρει.

Παρόλα αυτά πρέπει να τονιστεί το γεγονός ότι η τεχνική των παράλληλων συντεταγμένων είναι η πιο γνωστή και συχνά χρησιμοποιούμενη τεχνική στην αναπαράσταση των «συχνών στοιχείων». Ταυτόχρονα, όπως έχει αναφερθεί αρκετές φορές μέχρι τώρα στην παρούσα μεταπτυχιακή διατριβή, υπάρχουν πολλές έρευνες που αφορούν την αναπαράσταση-οπτικοποίηση δεδομένων κίνησης δικτύου (χωρίς τη χρήση «συχνών στοιχείων») οι οποίες κάνουν αναφορά στη συγκεκριμένη τεχνική.

4.2.2 Γραφήματα Δύο Διαστάσεων

Στην κατηγορία αυτή ανήκουν εργασίες που στόχο είχαν την αναπαράσταση «συχνών στοιχείων» με τη χρήση δυσδιάστατων γραφημάτων. Ενδεικτικές εργασίες-μελέτες προς αυτή την κατεύθυνση είναι οι ακόλουθες:

Carson Kai-Sang Leung, Pourang P. Irani, and Christopher L. Carmichael (2008). FIsViz: A Frequent Itemset Visualizer. PAKDD'08 Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining, Berlin, Heidelberg ©2008 [65].

Carson Kai-Sang Leung, Pourang P. Irani, and Christopher L. Carmichael (2008). WiFIsViz: Effective Visualization of Frequent Itemsets. ICDM '08 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, IEEE Computer Society Washington, DC, USA ©2008 [66].

Carson Kai-Sang Leung and Christopher L. Carmichael (2009). FpVAT: A Visual Analytic Tool for Supporting Frequent Pattern Mining. Published in ACM SIGKDD Explorations Newsletter, New York, USA 2009 [67].

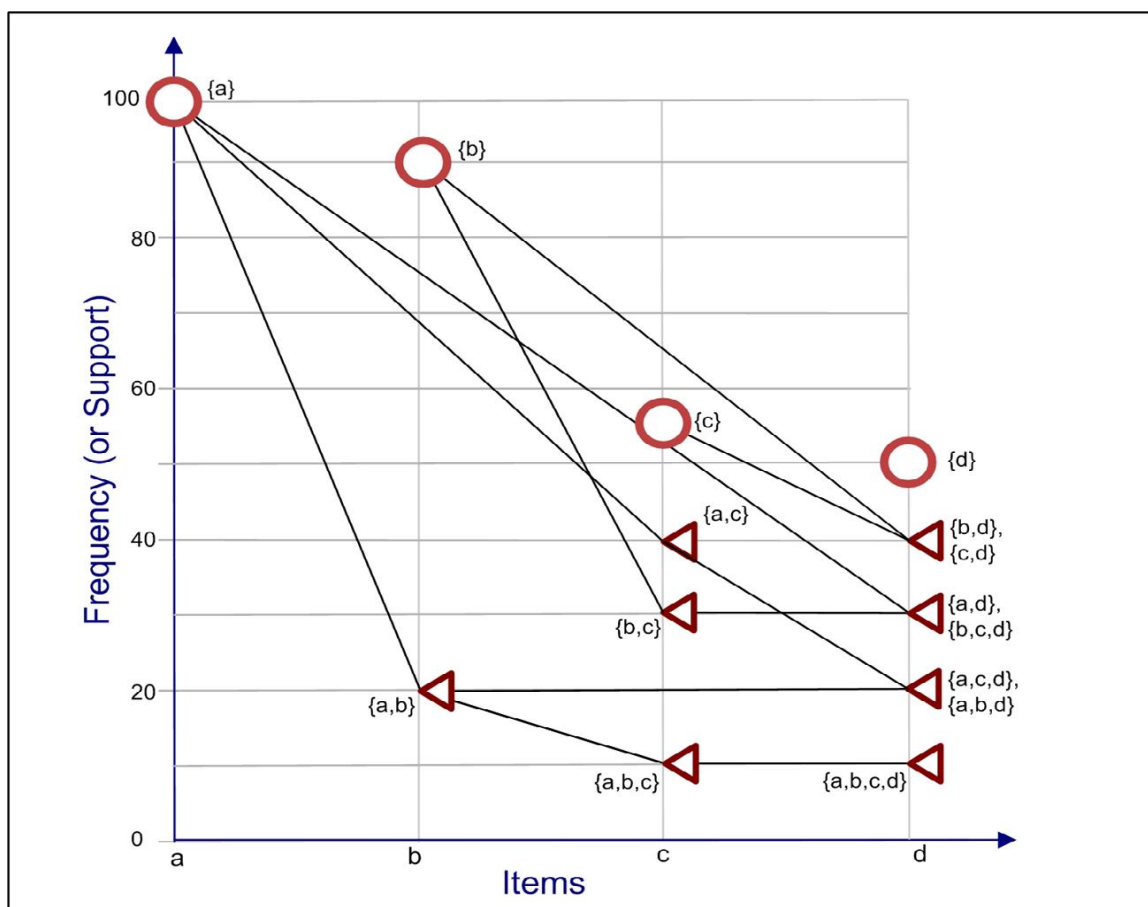
Οι δουλειές που περιγράφονται στα τρία παραπάνω άρθρα χρησιμοποιούν για την απεικόνιση «συχνών στοιχείων» και «κανόνων συσχέτισης» δυσδιάστατα γραφήματα. Στα γραφήματα αυτού του τύπου:

- Στον οριζόντιο άξονα X εμφανίζονται οι συχνότητες εμφάνισης των στοιχείων των “frequent itemsets” ως μεμονωμένα στοιχεία.
- Στον κατακόρυφο άξονα Y εμφανίζονται οι συχνότητες (τιμές «στήριξης») των “frequent itemsets”.

Για την αναπαράσταση κάθε “frequent itemset” σχεδιάζεται μία γραμμή η οποία συνδέει όλα τα στοιχεία του “frequent itemset”. Αυτή ξεκινά από το στοιχείο με τη μεγαλύτερη συχνότητα και καταλήγει στο στοιχείο με τη μικρότερη συχνότητα που ανήκει στο περιγραφέν “frequent itemset”. Για να δηλωθεί το τέλος της γραμμής αυτής (τελευταίο στοιχείο του “frequent itemset”) σχεδιάζεται ένα τρίγωνο. Η τιμή του κατακόρυφου άξονα η οποία αντιστοιχεί στη θέση του τριγώνου δηλώνει την τιμή «στήριξης» του “frequent itemset”. Εφόσον υπάρχει “frequent itemset” με ένα μόνο στοιχείο αυτό αναπαριστάτε με έναν κύκλο.

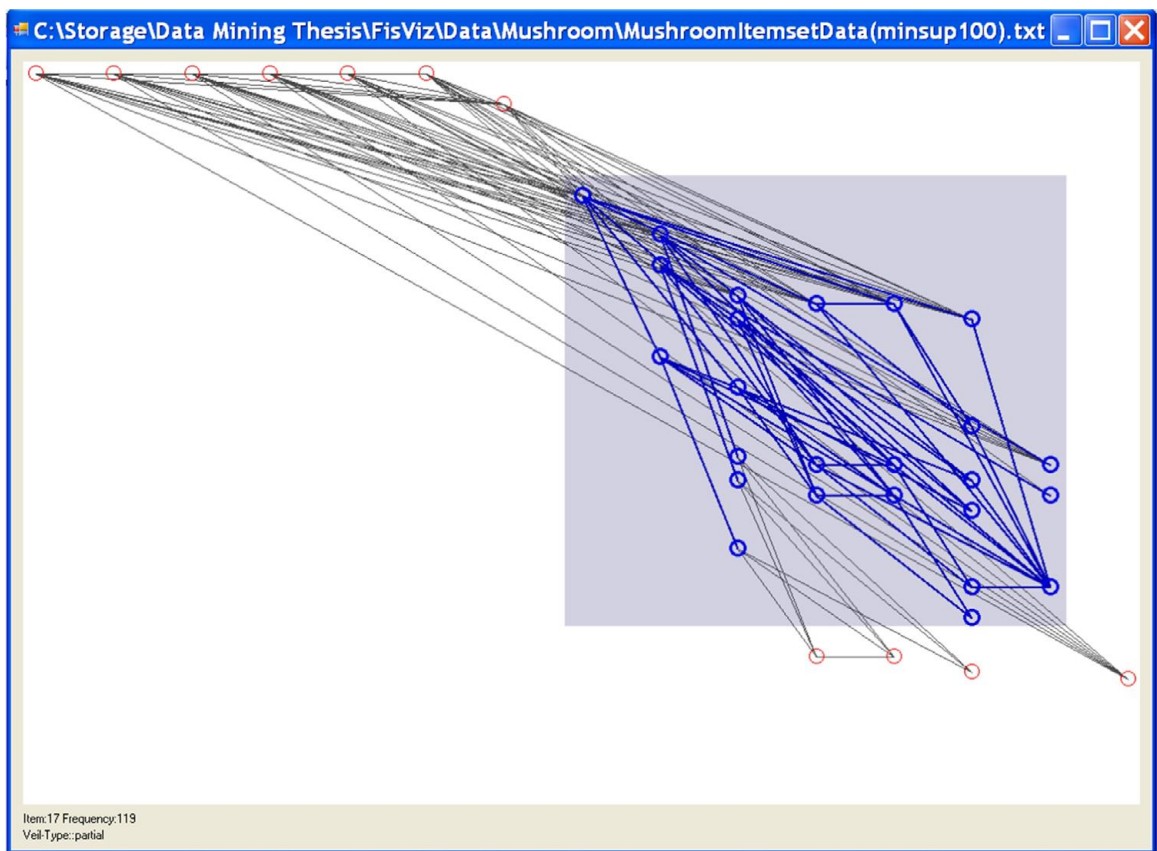
Στο παρακάτω σχήμα (σχήμα 4.4) φαίνεται ένα γράφημα τέτοιου είδους το οποίο αναπαριστά 15 “frequent itemsets” τα ακόλουθα:

{a}, {b}, {c}, {d}, {a,b}, {a,c}, {b,c}, {b,d}, {c,d}, {a,d}, {b,c,d}, {a,c,d}, {a,b,d}, {a,b,c}, {a,b,c,d}



Σχήμα 4.4: Διάγραμμα δύο διαστάσεων για την αναπαράσταση “frequent itemsets”. Εικόνα από το PDF με τίτλο “FIsViz: A Frequent Itemset Visualizer” από τους Carson Kai-Sang Leung, Pourang P. Irani, και Christopher L. Carmichael [65]

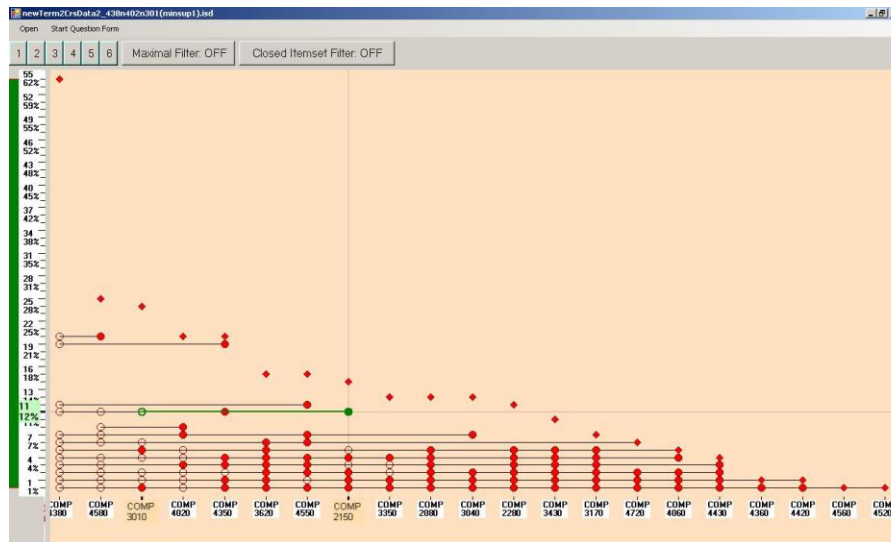
Στην δουλειά που περιγράφεται στο άρθρο με τίτλο “FIsViz: A Frequent Itemset Visualizer” από τους Carson Kai-Sang Leung, Pourang P. Irani, και Christopher L. Carmichael περιγράφεται εργαλείο που δημιουργήθηκε και το οποίο χρησιμοποιεί την τεχνική αυτή για την αναπαράσταση “frequent itemsets”. Το εργαλείο ονομάζεται FIsViz και στο παρακάτω σχήμα (σχήμα 4.5) εμφανίζεται οθόνη από αποτελέσματα αυτού.



Σχήμα 4.5: Screen shot από το εργαλείο FIsViz. Εικόνα από το PDF με τίτλο “FIsViz: A Frequent Itemset Visualizer” από τους Carson Kai-Sang Leung, Pourang P. Irani, και Christopher L. Carmichael [65]

Η δουλειά που περιγράφεται στο άρθρο με τίτλο “WiFIsViz: Effective Visualization of Frequent Itemsets” από τους Carson Kai-Sang Leung, Pourang P. Irani, και Christopher L. Carmichael παρουσιάζεται εργαλείο που δημιουργήθηκε (WiFIsViz) και το οποίο αποτελεί επέκταση-βελτίωση του συστήματος FIsViz που περιγράφεται από το προηγούμενο άρθρο. Η βελτίωση έγινε με στόχο να αντιμετωπιστεί ο μεγάλος αριθμός διασταυρώσεων μεταξύ των “polylines” στην περίπτωση ύπαρξης πολλών “frequent itemsets” και το οποίο καθιστούσε την κατανόηση του αποτελέσματος «δύσκολη υπόθεση».

Αντίστοιχη με την προηγούμενη είναι και η δουλειά που περιγράφεται στο άρθρο με τίτλο “FrVAT: A Visual Analytic Tool for Supporting Frequent Pattern Mining” από τους Carson Kai-Sang Leung και Christopher L. Carmichael [67]. Στο σχήμα 4.6 εμφανίζεται οθόνη από το εργαλείο FrVAT που δημιουργήθηκε.



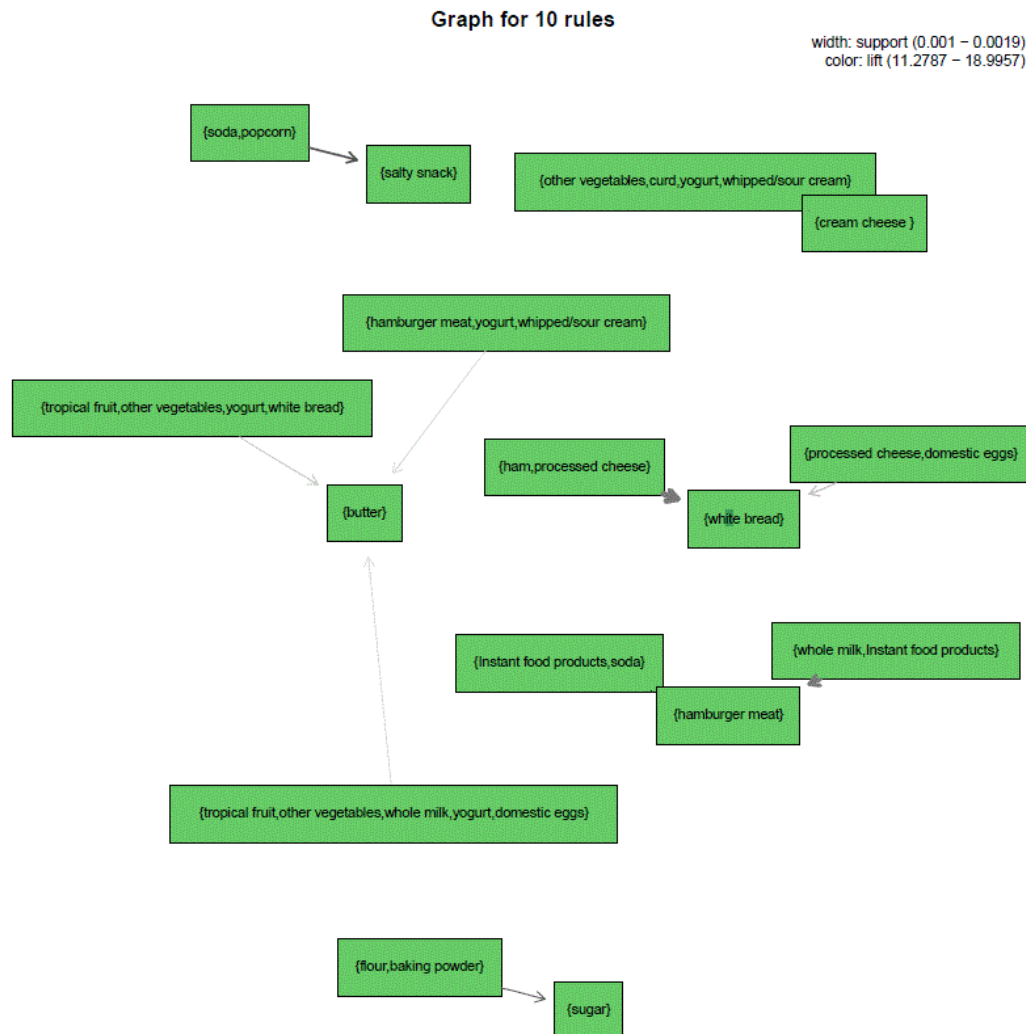
Σχήμα 4.6: Screen shot από το εργαλείο FpVAT. Εικόνα από το PDF με τίτλο “FpVAT: A Visual Analytic Tool for Supporting Frequent Pattern Mining” [67].

Στην κατηγορία αυτή ανήκει και η δουλειά που περιγράφεται στο άρθρο με τίτλο “Exploring Social Networks: A Frequent Pattern Visualization Approach” από τους Carson Kai-Sang Leung και Christopher L. Carmichael [68].

Τα πλεονεκτήματα γενικά του τρόπου απεικόνισης με τη χρήση γραφημάτων δύο διαστάσεων όπως αυτά περιγράφηκαν παραπάνω, είναι ότι ο παρατηρητής μπορεί να δει ξεκάθαρα τις τιμές της «στήριξης» τόσο των ανεξάρτητων στοιχείων όσο και των “frequent itemsets”. Αντίθετα το μεγάλο μειονέκτημα που μπορεί κάποιος να διακρίνει είναι το γεγονός ότι όσο μεγαλύτερος είναι ο αριθμός των “frequent itemsets” και όσα περισσότερα στοιχεία έχει το κάθε ένα από αυτά, τόσο δυσκολότερη έως αδύνατη γίνεται η δυνατότητα διαχωρισμού και αναγνώρισης αυτών.

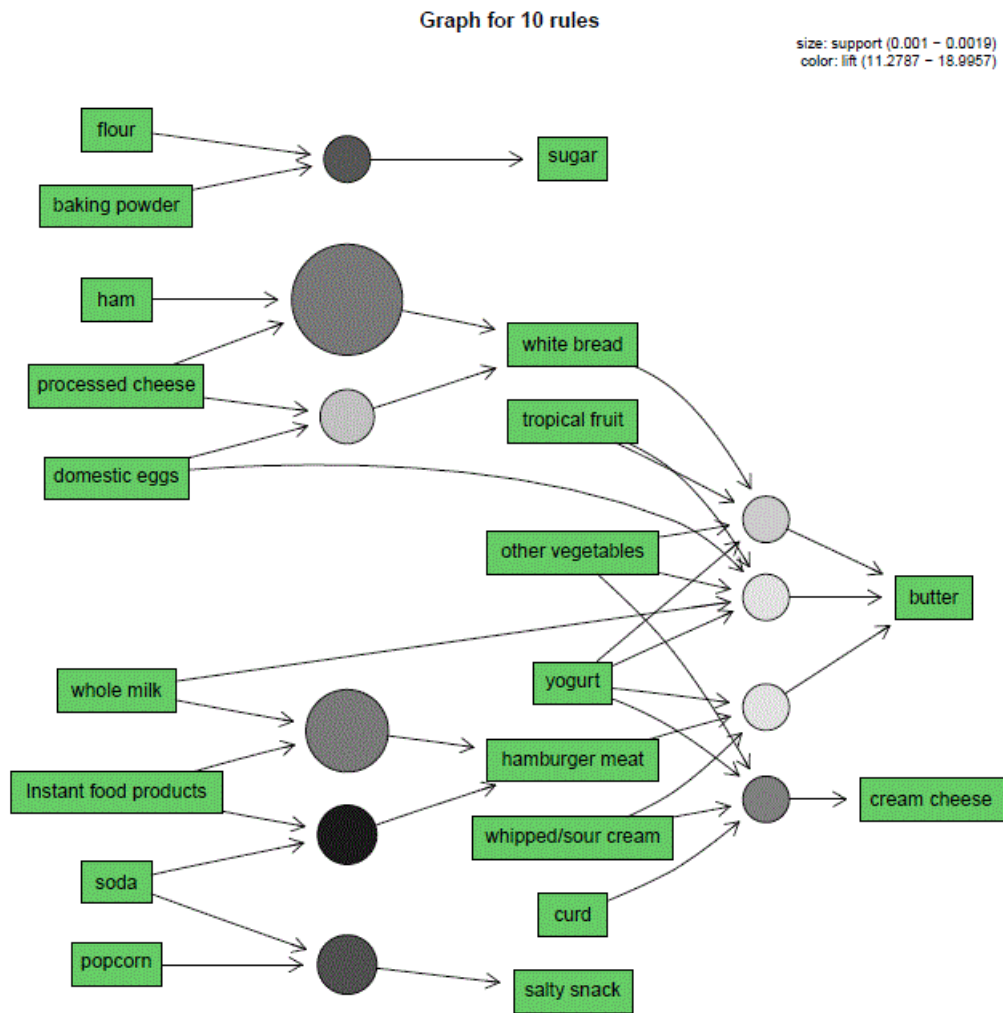
4.2.3 Graph-Based Visualizations

Ένας πολύ ενδιαφέρον τρόπος αναπαράστασης «κανόνων συσχέτισης» είναι αυτός με τη χρήση “graph based visualization” [61]. Για την αναπαράσταση των «κανόνων συσχέτισης» χρησιμοποιούνται κορυφές και ακμές. Οι κορυφές στο γράφημα αναπαριστούν «συχνά στοιχεία» ενώ οι ακμές αναπαριστούν τις σχέσεις μεταξύ των «συχνών στοιχείων». Τέτοιο γράφημα εμφανίζεται στο σχήμα 4.7 το οποίο αποτελεί αποτέλεσμα τέτοιου είδους απεικόνισης με τη χρήση του πακέτου “R”.



Σχήμα 4.7: Απεικόνιση «κανόνων συσχέτισης» με τη χρήση graph based visualization. Εικόνα από το PDF με τίτλο “Visualizing Association Rules Introduction to the R-extension Package arulesViz” [61]

Αντίστοιχος τρόπος αναπαράστασης (μία μικρή παραλλαγή της ίδια τεχνικής) είναι αυτός που εμφανίζεται στο σχήμα 4.8. Εδώ τα στοιχεία που συνθέτουν τα «συχνά στοιχεία» αποτελούν ανεξάρτητες κορυφές (τα παραλληλόγραμμα στο σχήμα). Κάθε «συχνό στοιχείο» αναπαριστάται από έναν κύκλο. Τα στοιχεία κάθε «συχνού στοιχείου» είναι τα παραλληλόγραμμα από τα οποία ξεκινούν ακμές οι οποίες καταλήγουν στον κύκλο που αναπαριστά το «συχνό στοιχείο». Τα παραλληλόγραμμα στα οποία καταλήγει μία ακμή αποτελούν το δεξί τμήμα του κανόνα συσχέτισης.



Σχήμα 4.8: Απεικόνιση «κανόνων συσχέτισης» με τη χρήση graph based visualization. Εικόνα από το PDF με τίτλο “Visualizing Association Rules Introduction to the R-extension Package arulesViz” [61].

Αυτός ο τρόπος γραφικής αναπαράστασης θα μπορούσε να χρησιμοποιηθεί και για την αναπαράσταση «συχνών στοιχείων». Η διαφορά θα είναι το γεγονός ότι στην αναπαράσταση «συχνών στοιχείων» οι ακμές θα είναι βέλη τα οποία είτε θα ξεκινούν όλα, είτε θα καταλήγουν όλα σε κύκλους. Δεν θα υπάρχει δηλαδή κύκλος στην γραφική αναπαράσταση «συχνών στοιχείων» από τον οποίο ταυτόχρονα να ξεκινούν και να καταλήγουν «βέλη» (ακμές). Φυσικά και θα μπορούσαν σε αυτού του είδους τα γραφήματα να προστεθούν και περεταίρω πληροφορίες που αφορούν τα συχνά στοιχεία όπως η τιμή της «στήριξης» του κάθε «συχνού στοιχείου».

Αυτή η μέθοδος απεικόνισης αναγνωρίστηκε ως η πιο πολλά υποσχόμενη από αυτές που αναφέρονται στην παρούσα μεταπτυχιακή διατριβή. Ένα από τα κριτήρια που οδηγούν σε αυτή την απόφαση είναι η ικανότητά του συγκεκριμένου τρόπου απεικόνισης να δίνει μία πιο ξεκάθαρη εικόνα των «συχνών στοιχείων» και των στοιχείων από τα οποία αποτελούνται αυτά. (περισσότερα επιχειρήματα-λόγοι που οδηγούν στο πόρισμα αυτό θα επεξηγηθούν με παραδείγματα στο επόμενο κεφάλαιο). Και η συγκεκριμένη βέβαια μέθοδος απεικόνισης έχει μειονεκτήματα-περιορισμούς όπως είναι ο περιορισμένος αριθμός «συχνών στοιχείων» που μπορούν να αναπαρασταθούν εφόσον απαιτείται λεπτομερείς απεικόνιση των «συχνών στοιχείων». Στην περίπτωση δηλαδή που ο αριθμός αυτών αυξηθεί αρκετά επιτυγχάνεται το αντίθετο από το επιθυμητό αποτέλεσμα, η δημιουργία δηλαδή ενός δυσνόητου γραφήματος.

Εκτός των μεθόδων οπτικής αναπαράστασης «συχνών στοιχείων» που παρατέθηκαν στο παρόν κεφάλαιο, το πακέτο επέκτασης “arulesViz” του λογιστικού πακέτου “R” περιέχει και άλλες μεθόδους «οπτικοποίησης» παραγόμενων «συχνών στοιχείων». Τέτοιες μέθοδοι είναι οι: “scatter plot”, “matrix based visualization” και “grouped matrix-based visualizations” [61]. Οι συγκεκριμένες μέθοδοι δεν αναλύονται στην παρούσα μεταπτυχιακή διατριβή, διότι θεωρήθηκε ότι δεν καλύπτουν ικανοποιητικά τους στόχους αυτής. Κρίθηκε δηλαδή ότι δεν μπορεί να προσφέρουν στην αναπαράσταση «συχνών στοιχείων» που προέρχονται από δεδομένα καταγραφείσας κίνησης δικτύου υπολογιστών.

Ας αναφερθεί στο σημείο αυτό ότι, με βάση τα παραπάνω, το εργαλείο “R” αποτέλεσε το κεντρικό εργαλείο για την εξαγωγή συμπερασμάτων στην παρούσα μεταπτυχιακή διατριβή. Χρησιμοποιήθηκε για την εξαγωγή γραφημάτων και τη σύγκριση μεταξύ τους. Σε αυτό εισήχθησαν πειραματικά δεδομένα τα οποία προέρχονται από κίνηση πραγματικού δικτύου υπολογιστών και, αφού παράχθηκαν «συχνά στοιχεία» με βάση τον αλγόριθμο Apriori, δημιουργήθηκαν κατάλληλα γραφήματα. Στα γραφήματα αυτά έγιναν αναλυτικές παρατηρήσεις με στόχο την εξαγωγή συμπερασμάτων για τη χρησιμότητα των παραγόμενων γραφημάτων και την ευκολία με την οποία μπορεί ένας διαχειριστής δικτύου να τα χρησιμοποιήσει και να εξάγει χρήσιμες πληροφορίες για την κίνηση που διέρχεται από το τοπικό δίκτυο. Τα δεδομένα που εισήχθησαν και χρησιμοποιήθηκαν για την παραγωγή συχνών στοιχείων ήταν από διαφορετικές χρονικές στιγμές και διαφορετικά χρονικά διαστήματα (μικρά και μεγάλα) στην προσπάθεια εξαγωγής συμπερασμάτων.

4.3 Εφαρμοσιμότητα Μεθόδων Οπτικοποίησης στην Αναπαράσταση «Συχνών Στοιχείων» που Προέρχονται από Δεδομένα Κίνησης Δικτύου Υπολογιστών.

Ένας από τους στόχους της παρούσας μεταπτυχιακής διατριβής είναι η αξιολόγηση των τεχνικών οπτικοποίησης (visualization) των «συχνών στοιχείων» καθώς και ο εντοπισμός της καταλληλότερης από αυτές με βασικό κριτήριο το κατά πόσο θα μπορούσε να χρησιμοποιηθεί και να συνδράμει στην ολοκληρωμένη διαχείριση-παρακολούθηση δικτύων υπολογιστών. Επιμέρους κριτήρια τα οποία θα μπορούσαν να χρησιμοποιηθούν για την επιλογή της καταλληλότερης τεχνικής είναι τα ακόλουθα:

- Η αποδοτικότητά της. Η ευκολία δηλαδή που παρέχει στην εξαγωγή «κρυμμένων» πληροφοριών.
- Η απλότητά της. Η ευκολία δηλαδή κατανόησης τόσο της ίδιας οπτικής αναπαράστασης όσο και των δεδομένων που αναπαριστά.
- Τη δυνατότητα που παρέχει ώστε παρατηρώντας την κάποιος να μπορεί να εξάγει συμπεράσματα για την κίνηση που διέρχεται από το δίκτυο υπολογιστών, να δημιουργεί δηλαδή εύκολα μία εικόνα για το είδος της κίνησης που διέρχεται από το δίκτυο.
- Η χρησιμότητά της στον εντοπισμό κάποιου είδους επιθέσεων.
- Το πόσο γρήγορα και εύκολα μπορούν να εντοπιστούν τα πιο συχνά χρησιμοποιούμενα στοιχεία όπως η πιο συχνά χρησιμοποιούμενη διεύθυνση προορισμού, η πιο συχνά χρησιμοποιούμενη θύρα προορισμού κ.λ.π.
- Το πόσο ευανάγνωστη αναπαράσταση των «συχνών στοιχείων» παρέχει. Δηλαδή το πόσο εύκολα μπορούν να διακριθούν (και να ξεχωρίσουν μεταξύ τους) πληροφορίες, όπως οι IP διευθύνσεις και οι αριθμοί θυρών που αναπαριστώνται.
- Ο αριθμός των «συχνών στοιχείων» που μπορούν να αναπαρασταθούν.

Ένα ακόμη κριτήριο για την επιλογή της καταλληλότερης τεχνικής αποτελεί και το κατά πόσο η συγκεκριμένη τεχνική μπορεί να αναπαραστήσει δεδομένα κίνησης δικτύου λόγω κάποιων ιδιαιτεροτήτων που παρουσιάζουν αυτά. Τα δεδομένα γενικά που προέρχονται από την καταγραφή

κίνησης δικτύου υπολογιστών και τα οποία περιέχουν πληροφορίες κεφαλίδων πακέτων-ροών διαφοροποιούνται από τα δεδομένα των συνηθισμένων βάσεων δεδομένων που χρησιμοποιούνται για την εξαγωγή «συχνών στοιχείων». Μία από τις σημαντικότερες διαφοροποιήσεις είναι ο διαχωρισμός των δεδομένων σε κατηγορίες. Υπάρχουν τέσσερις κύριες κατηγορίες πληροφοριών στα δεδομένα: οι διευθύνσεις προορισμού, οι διευθύνσεις αφετηρίας, οι θύρες προορισμού και οι θύρες αφετηρίας. Οι πληροφορίες αυτές είναι ανεξάρτητες μεταξύ τους και πρέπει να μπορούν να διαχωρίζονται στο αποτέλεσμα της οπτικής αναπαράστασης. Ένας τέτοιος διαχωρισμός δεν υπάρχει στην περίπτωση αναπαράστασης «συχνών στοιχείων» που προέρχονται π.χ. από δεδομένα αγορών από ένα super market.

Από τις τεχνικές που αναλύθηκαν στην προηγούμενη ενότητα οι δύο που τελικά αποφασίστηκε να χρησιμοποιηθούν στην παρούσα μεταπτυχιακή διατριβή είναι αυτές των γραφημάτων παράλληλων συντεταγμένων και της απεικόνιση με τη χρήση γράφων (“graph based visualization”). Με μία πρώτη ματιά δείχνουν να μπορούν να καλύψουν το σύνολο των προαναφερθέντων κριτηρίων. Δείχνουν ότι μπορούν να ανταποκριθούν στις προαναφερθέντες ιδιαιτερότητες των δεδομένων διαχωρίζοντας τις τέσσερις κατηγορίες των δεδομένων. Έχουν απλό τρόπο δημιουργίας και μπορούν να χρησιμοποιηθούν για να μπορέσει ο παρατηρητής να εξάγει μία γρήγορη εικόνα για την κίνηση που περιγράφεται από αυτά. Θα μπορούσαν να χρησιμοποιηθούν ακόμη και για τον εντοπισμό κάποιου είδους επιθέσεων όπως θα φανεί τελικά και στην πράξη έπειτα από την παράθεση των πειραματικών δεδομένων του επόμενου κεφαλαίου. Φαίνεται να μπορούν να απεικονίσουν με αποδοτικό τρόπο τα δημιουργημένα από δεδομένα δικτύου «συχνά στοιχεία» και γενικότερα να αποτελέσουν ένα πολλά υποσχόμενο εργαλείο βοηθώντας στη γενικότερη παρακολούθηση-διαχείριση ενός δικτύου υπολογιστών.

Η τεχνική απεικόνισης με τη χρήση γραφημάτων δύο διαστάσεων (§ 4.2.2) δε θα μπορούσε να ικανοποιήσει τα παραπάνω αναφερθέντα κριτήρια (παρατηρήσεις συνηγορούν σε αυτό), γι’ αυτό και δεν χρησιμοποιήθηκε για περαιτέρω ανάλυση. Αν και η συγκεκριμένη τεχνική θα μπορούσε να τονίζει τη συχνότητα εμφάνισης κάθε συχνού στοιχείου», δε θα μπορούσε να δώσει ευδιάκριτα τις απαιτούμενες πληροφορίες λόγω του τρόπου με τον οποίο δημιουργείται. Για να μπορέσει να διακρίνει κανείς τις πληροφορίες των «συχνών στοιχείων» με τη χρήση αυτού του τρόπου αναπαράστασης θα έπρεπε ο αριθμός των αναπαριστώμενων «συχνών στοιχείων» να είναι πάρα πολύ μικρός καθιστώντας έτσι το γράφημα μη αποδοτικό.

Τέλος, για την επιλογή των δύο αυτών τεχνικών ως βάση των παρατηρήσεων, σημαντικό ρόλο έπαιξε και ο αριθμός των μελετών-ερευνών που έχουν πραγματοποιηθεί μέχρι τώρα και οι οποίες χρησιμοποιούν τις μεθόδους αυτές. Υπάρχουν πολλές έρευνες που χρησιμοποιούν την τεχνική των γραφημάτων παράλληλων συντεταγμένων ενώ αντίθετα πολύ λίγες που χρησιμοποιούν την τεχνική δημιουργίας γράφων. Συγκεκριμένα, όπως αναφέρθηκε και στο πρώτο κεφάλαιο της παρούσας μεταπτυχιακής διατριβής, υπάρχουν διάφορες μελέτες οι οποίες χρησιμοποιούν την τεχνική των παράλληλων συντεταγμένων για τον εντοπισμό επιθέσεων σε δίκτυα υπολογιστών χρησιμοποιώντας βέβαια raw data (χωρίς τη χρήση «συχνών στοιχείων»). Αποτέλεσε κίνητρο λοιπόν το κατά πόσο το ίδιο θα μπορούσε να συμβεί με την αναπαράσταση «συχνών στοιχείων». Αντίθετα η χρήση απεικονίσεων με τη χρήση γράφων δεν χρησιμοποιήθηκε μέχρι τώρα για απεικόνιση «συχνών στοιχείων» προερχόμενα από κίνηση δικτύου υπολογιστών. Αυτό αποτέλεσε πρόσθετο κίνητρο για τη χρήση της συγκεκριμένης μεθόδου και τον εντοπισμό του κατά πόσο θα μπορούσε και σε ποιο βαθμό, να προσφέρει ικανοποιητική αναπαράσταση.

Οι δύο λοιπόν αυτές μέθοδοι οπτικής αναπαράστασης αποτέλεσαν τον κύριο άξονα για τις παρατηρήσεις που έγιναν. Δημιουργήθηκε με βάση αυτές μεγάλος αριθμός γραφημάτων προς παρατήρηση. Στο κεφάλαιο που ακολουθεί μεταξύ άλλων θα γίνει αναλυτική παρουσίαση παραδειγμάτων τέτοιων σχεδιάσεων και θα παρατεθούν προκύπτοντα πλεονεκτήματα και μειονεκτήματα της καθεμιάς από τις δύο μεθόδους απεικόνισης «συχνών στοιχείων» με στόχο την επιλογή της καταλληλότερης.

Κεφάλαιο 5

Μια Πρώτη Προσέγγιση στην Οπτική Αναπαράσταση της Κίνησης Δικτύου Υπολογιστών με τη Χρήση «Συχνών Στοιχείων».

Στην ενότητα αυτή θα πραγματοποιηθεί ανάλυση με μεγαλύτερη λεπτομέρεια των τεχνικών αναπαράστασης «συχνών στοιχείων» (τεχνική γραφημάτων παράλληλων συντεταγμένων και τεχνική “graph based visualization”) χρησιμοποιώντας παραδείγματα, με στόχο την επιλογή της καταλληλότερης τεχνικής. Θα παρατεθούν συμπεράσματα που προκύπτουν από τις παρατηρήσεις ενώ παράλληλα θα δίδονται διάφορα παραδείγματα για την υποστήριξη αυτών.

5.1 Εισαγωγή

Όπως αναλύθηκε και στην προηγούμενη ενότητα οι δύο ίσως πιο σημαντικές τεχνικές αναπαράστασης «συχνών στοιχείων» (frequent itemsets) είναι η μέθοδος αναπαράστασης με τη χρήση παράλληλων συντεταγμένων (parallel coordinates) και η μέθοδος που στηρίζεται στη χρήση γράφων (graph based visualization).

Οι περισσότερες μελέτες και εργασίες που έχουν γίνει προς αυτό το σκοπό, τη γραφική δηλαδή αναπαράσταση «συχνών στοιχείων», υποστηρίζουν το γεγονός ότι η μέθοδος αναπαράστασης με τη χρήση παράλληλων συντεταγμένων αποτελεί την πιο ενδιαφέρουσα (καλύτερη) μέθοδο αναπαράστασης «συχνών στοιχείων». Το μεγαλύτερο επιχείρημα υποστήριξης της μεθόδου αυτής αποτελεί το γεγονός ότι μεταξύ των υπό αναπαράσταση δεδομένων υφίσταται σχέση της μορφής «πολλά προς πολλά» για την αναπαράσταση της οποίας ο καταλληλότερος τύπος γραφήματος είναι αυτός των παράλληλων συντεταγμένων ([63], [67], [65], [66], [67]). Μάλιστα δημιουργήθηκαν αρκετές παραλλαγές αυτής της μεθόδου αναπαράστασης για την απεικόνιση είτε των «συχνών στοιχείων», είτε των «κανόνων συσχέτισης» που προκύπτουν από την εφαρμογή αλγορίθμων εξόρυξης «συχνών στοιχείων» σε δεδομένα.

Καμία από τις έρευνες που αναφέρονται στη βιβλιογραφία και οι οποίες κάνουν αναφορά σε τεχνικές γραφικής αναπαράστασης «συχνών στοιχείων» με τη χρήση παράλληλων συντεταγμένων δε χρησιμοποιεί και δεν αναπαριστά δεδομένα που προέρχονται από καταγραφείσα κίνηση (traffic) δικτύου υπολογιστών. Κάποιες δουλειές όπως αυτή που περιγράφεται στο άρθρο με τίτλο “Fast detection and visualization of network attacks on parallel coordinates” από τους Hyunsang Choi, Heejo Lee και Hyogon Kim (και στην οποία έγινε αναφορά και στο πρώτο κεφάλαιο της παρούσης μεταπτυχιακής διατριβής) απλά προσπαθούν να παρουσιάσουν με γραφήματα παράλληλων συντεταγμένων την κίνηση δικτύου, χωρίς όμως να εμπλέκουν στην όλη διαδικασία αλγορίθμους εξαγωγής «συχνών στοιχείων». Χρησιμοποιούν για την αναπαράσταση τα δεδομένα όπως αυτά είναι πριν εισέλθουν σε κάποιο αλγόριθμο παραγωγής «συχνών στοιχείων». Η συγκεκριμένη δουλειά παρουσιάζει αυτό που αποκαλείται από τους συγγραφείς του ως “parallel coordinate attack visualization” και το οποίο χρησιμοποιεί γραφήματα παράλληλων συντεταγμένων για την ανίχνευση επιθέσεων με βάση “visual signatures”. Προσπαθεί δηλαδή με βάση τα αποτελέσματα

αναπαράστασης διαφόρων περιπτώσεων επίθεσης να δημιουργήσει συστήματα ανίχνευσης εισβολών με βάση το αποτέλεσμα της αναπαράστασης.

Πολύ λιγότερες είναι οι αναφορές στη βιβλιογραφία για την αναπαράστασης αποτελεσμάτων αλγορίθμων εξόρυξης γνώσης με τη χρήση γράφων (graph based visualization). Παρατηρώντας κανείς όμως γραφήματα που προκύπτουν χρησιμοποιώντας αυτό τον τρόπο απεικόνισης, μπορεί να συμπεράνει (παρά την περιορισμένη συχνότητά εμφάνισής του τρόπου αυτού στη βιβλιογραφία) ότι ο τρόπος αυτός υπερτερεί και μπορεί να βοηθήσει στη γενικότερη παρακολούθηση της κίνησης ενός δικτύου υπολογιστών χρησιμοποιώντας «συχνά στοιχεία». Φαίνεται ότι μπορεί να βοηθήσει στην εξαγωγή εύκολα χρήσιμων συμπερασμάτων που θα μπορούσαν να βοηθήσουν έναν διαχειριστή δικτύων.

Στις επόμενες παραγράφους λοιπόν, γίνεται παράθεση παραδειγμάτων αυτών των δύο τεχνικών χρησιμοποιώντας δεδομένα που προέρχονται από ένα πραγματικό δίκτυο υπολογιστών (η δομή του περιγράφηκε στο κεφάλαιο 3 της παρούσας μεταπτυχιακής διατριβής). Γίνεται αναφορά σε πλεονεκτήματα και μειονεκτήματα των δύο αυτών μεθόδων και γίνονται συγκρίσεις μεταξύ τους ώστε να είναι δυνατή η εύρεση του καταλληλότερου τρόπου αναπαράστασης χρησιμοποιώντας τα συγκεκριμένα δεδομένα.

5.2 Αποτελέσματα Γραφικής Αναπαράστασης «Συχνών Στοιχείων» με τη Χρήση Παράλληλων Συντεταγμένων

Στην παρούσα παράγραφο θα γίνει παράθεση γραφημάτων παράλληλων συντεταγμένων τα οποία προέκυψαν από την απεικόνιση «συχνών στοιχείων» των δεδομένων που αφορούν καταγραφείσα κίνηση μεταξύ 21 και 25 Μαρτίου 2011 στο υπό παρατήρηση δίκτυο υπολογιστών (σε αυτά τα δεδομένα έγινε αναφορά και στο κεφάλαιο 3). Φυσικά οι παρατηρήσεις που πραγματοποιήθηκαν στην πράξη, χρησιμοποίησαν μεγαλύτερο όγκο δεδομένων, διαφορετικών χρονικών στιγμών καθώς και διαφορετικών κάθε φορά διαστημάτων, ώστε να ενισχυθεί η εγκυρότητα των συμπερασμάτων-αποτελεσμάτων. Είναι βέβαια αδύνατο να παρατεθούν όλα τα γραφήματα τα οποία δημιουργήθηκαν και χρησιμοποιήθηκαν για παρατήρηση για αυτό το λόγο στην παρούσα παράγραφο παρατίθενται κάποια από αυτά.

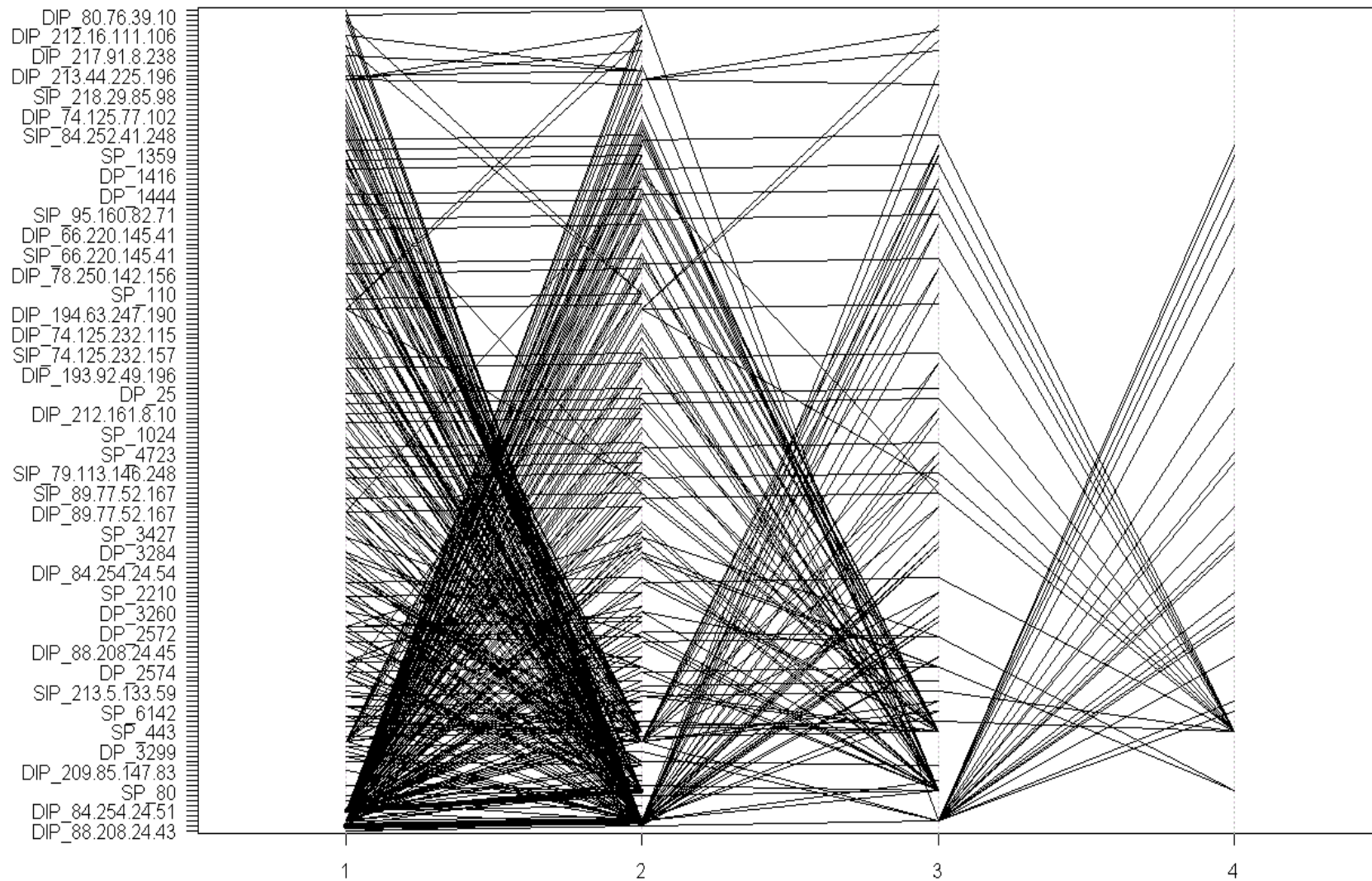
Χρησιμοποιώντας τα δεδομένα μεταξύ 21 και 25 Μαρτίου 2011 και χρησιμοποιώντας ως τιμή «στήριξης» την τιμή 0.001 ο αριθμός των «συχνών στοιχείων» που προκύπτουν είναι 797. Αν αυτά παρασταθούν με ένα γράφημα παράλληλων συντεταγμένων προκύπτει το παρακάτω γράφημα (σχήμα 5.1).

Χρησιμοποιώντας κανείς το γράφημα αυτό μπορεί να παρατηρήσει τα εξής:

- Στην πρώτη κάθετο, σύμφωνα με τον αλγόριθμο με βάση τον οποίο δημιουργείται ο συγκεκριμένος τύπος γραφήματος, τοποθετούνται μεμονωμένα τα στοιχεία τα οποία εμφανίζονται έστω και μία φορά σε κάποιο από τα υπό αναπαράσταση «συχνά στοιχεία». Τα στοιχεία αυτά δεν εμφανίζονται στις υπόλοιπες καθέτους για λόγους σαφήνειας της παρουσίασης του γραφήματος. Αν τα στοιχεία αυτά εμφανιζόταν σε κάθε κάθετο τότε θα υπήρχαν πάρα πολλές πληροφορίες στο γράφημα με αποτέλεσμα να απαιτείται πρόσθετη προσπάθεια για την ανάγνωση και κατανόησή του.
- Τα στοιχεία τα οποία εμφανίζονται στην πρώτη κάθετο δεν είναι ταξινομημένα με κάποιο τρόπο. Το συχνότερο στοιχείο για τα δεδομένα αυτά είναι το {SIP_84.254.24.51} με τιμή στήριξης 0.5232168579, ενώ το δεύτερο συχνότερο «συχνό στοιχείο» είναι το {DIP_84.254.24.51} με τιμή στήριξης 0.4533521242. Στο γράφημα, στον κάθετο άξονα ξεκινώντας από κάτω το πρώτο στοιχείο είναι το {DIP_88.208.24.43} ενώ δεύτερο εμφανίζεται το {DIP_84.254.24.51}.
- Οι ακμές που χρησιμοποιούνται για να δηλώσουν τα «συχνά στοιχεία» είναι πάρα πολλές και αντίστοιχα μεγάλος είναι ο αριθμός των διασταυρώσεων μεταξύ τους γεγονός που καθιστά το γράφημα δυσνόητο.
- Είναι φυσικά κατανοητό ότι ο αριθμός των «συχνών στοιχείων» (797) είναι υπερβολικά μεγάλος με αποτέλεσμα να είναι αδύνατος ο διαχωρισμός των πληροφοριών των «συχνών στοιχείων»:
 - Είναι αδύνατο να ξεχωρίσει κανείς από ποια στοιχεία αποτελούνται τα «συχνά στοιχεία».
 - Δεν αναγράφεται κάπου η πληροφορία της τιμής «στήριξης» των «συχνών στοιχείων».

- Ένα από τα συχνότερα στοιχεία είναι το {DIP_84.254.24.51}. Αυτό μπορεί να εντοπιστεί άμεσα (με μία γρήγορη ματιά) διότι από αυτό ή σε αυτό, καταλήγει μεγάλος αριθμός ακμών. Η πυκνότητα δηλαδή των ακμών που καταλήγουν σε ένα στοιχείο δίνει τη δυνατότητα για εύκολη εύρεση των συχνότερων στοιχείων.

Parallel coordinates plot for 797 itemsets



Σχήμα 5.1: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 797 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Αυξάνοντας την τιμή «στήριξης» για τα ίδια πάντοτε δεδομένα, μειώνεται ο αριθμός των «συχνών στοιχείων» που παράγονται. Στο σχήμα 5.2 εμφανίζεται ένα γράφημα παράλληλων συντεταγμένων το οποίο απεικονίζει τα 406 «συχνά στοιχεία» τα οποία προκύπτουν από τον αλγόριθμο Apriori χρησιμοποιώντας την τιμή στήριξης 0.002. Αντίστοιχα στο σχήμα 5.3 απεικονίζονται 228 «συχνά στοιχεία» τα οποία προκύπτουν χρησιμοποιώντας τιμή στήριξης 0.003.

Οι διαπιστώσεις που μπορεί κανείς να κάνει παρατηρώντας τα συγκεκριμένα γραφήματα είναι παρόμοιες με αυτές που αναφέρθηκαν παραπάνω.

- Ο αριθμός των «συχνών στοιχείων» που απεικονίζονται είναι πάρα πολύ μεγάλος για να μπορεί κανείς να πει ότι μπορεί να πει με βεβαιότητα ότι μπορεί να ξεχωρίσει τα στοιχεία που συνθέτουν τα «συχνά στοιχεία».
- Παρατηρώντας το γράφημα στο σχήμα 3 μπορεί κανείς να αρχίσει να διακρίνει τις διαφορές στα πάχη των ακμών που συνδέουν τα στοιχεία μεταξύ τους. Όσο πιο παχιά είναι μία ακμή τόσο μεγαλύτερη είναι η τιμή «στήριξης» του συχνού στοιχείου που αναπαριστά.

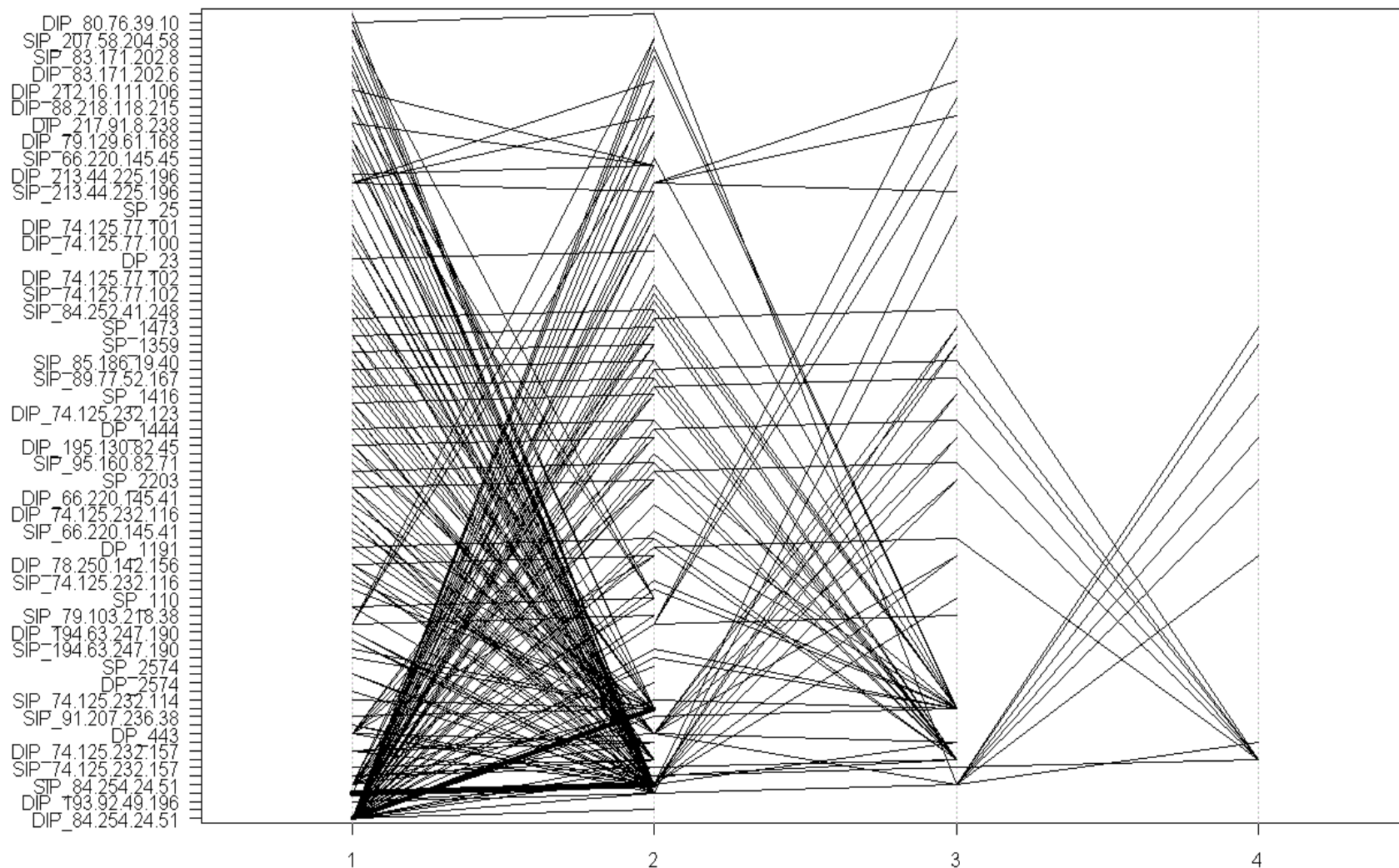
Τι ίδιες παρατηρήσεις κάνει κανείς παρατηρώντας τα γραφήματα που αναπαριστούν τα «συχνά στοιχεία» για τιμές στήριξης μέχρι την τιμή 0.005 (σχήματα 5.4 και 5.5 αντίστοιχα για τιμές στήριξης 0.004 και 0.005).

Στα γραφήματα που εμφανίζονται στα σχήματα 5.4 και 5.5 παρατηρείται ότι οι κάθετες γραμμές που εμφανίζονται στο γράφημα από τέσσερις που ήταν στις προηγούμενες αναπαραστάσεις τώρα έχουν γίνει τρεις. Αυτό οφείλεται στο γεγονός ότι τα «συχνά στοιχεία» που προκύπτουν αποτελούνται από το πολύ τρία στοιχεία.

Κάτι που γίνεται πάλι αντιληπτό από την παρατήρηση αυτών των γραφημάτων είναι το γεγονός ότι οι τέσσερις κατηγορίες πληροφοριών που υπάρχουν στα δεδομένα που πρέπει να αναπαρασταθούν (Source IP address, Destination IP Address, Source Port Number και Destination Port Number) και οι οποίες αποτελούν ιδιομορφία των δεδομένων (μιλάμε για δεδομένα κίνησης δικτύου) δεν αναπαρίστανται από συγκεκριμένη κάθε φορά κάθετο. Δεν έχουμε δηλαδή αντιστοιχία των καθέτων του γραφήματος με μία συγκεκριμένη κάθε φορά κατηγορία πληροφορίας πχ δεν

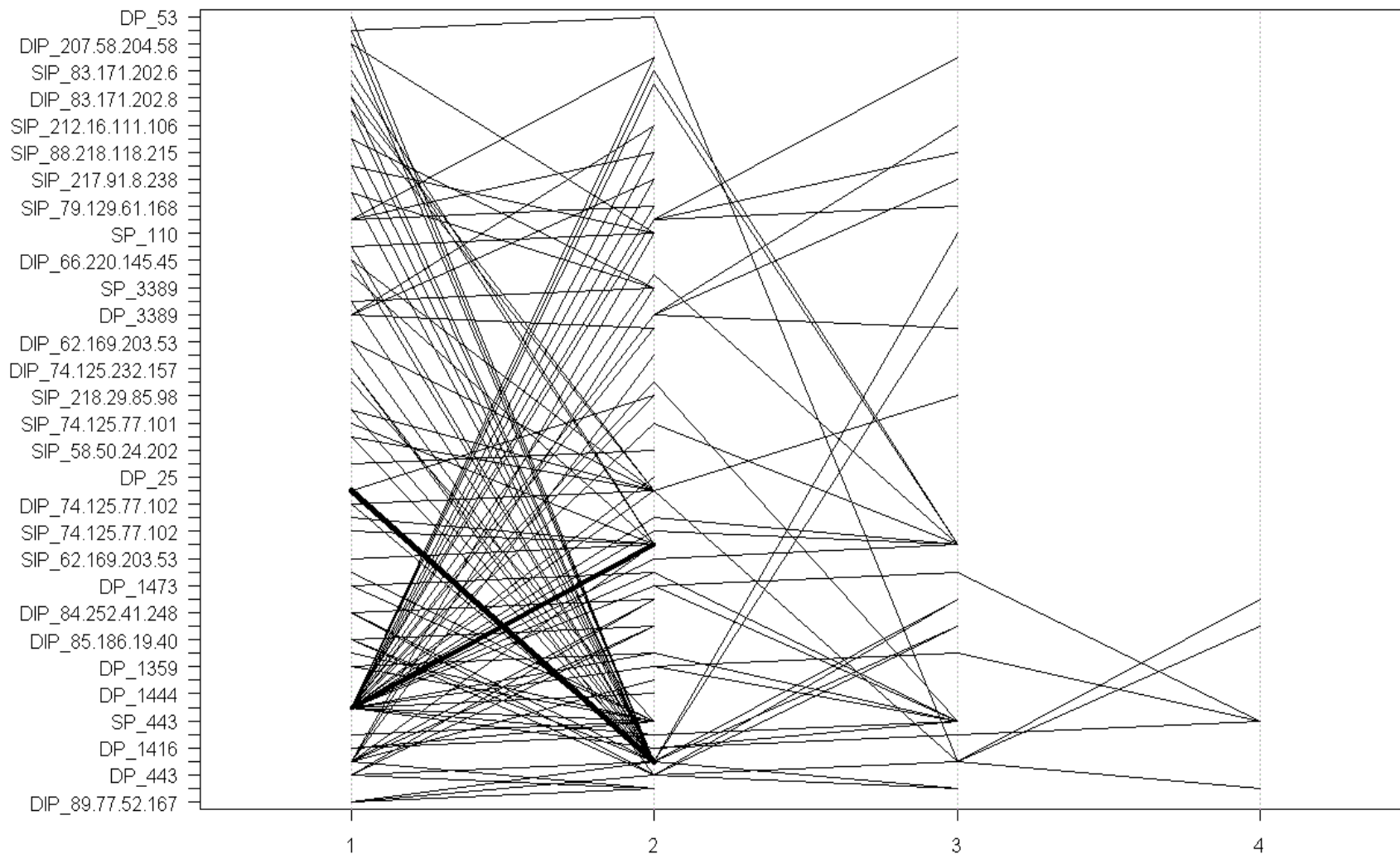
υπάρχει αντιστοιχία της πρώτης καθέτου με την πληροφορία Source IP Address, της δεύτερης καθέτου με την πληροφορία Destination IP address κλπ. Ίσως κάτι τέτοιο να μπορούσε να βοηθήσει στον εντοπισμό επιθέσεων όπως αναφέρεται στο άρθρο με τίτλο “Fast detection and visualization of network attacks on parallel coordinates” [11].

Parallel coordinates plot for 406 itemsets



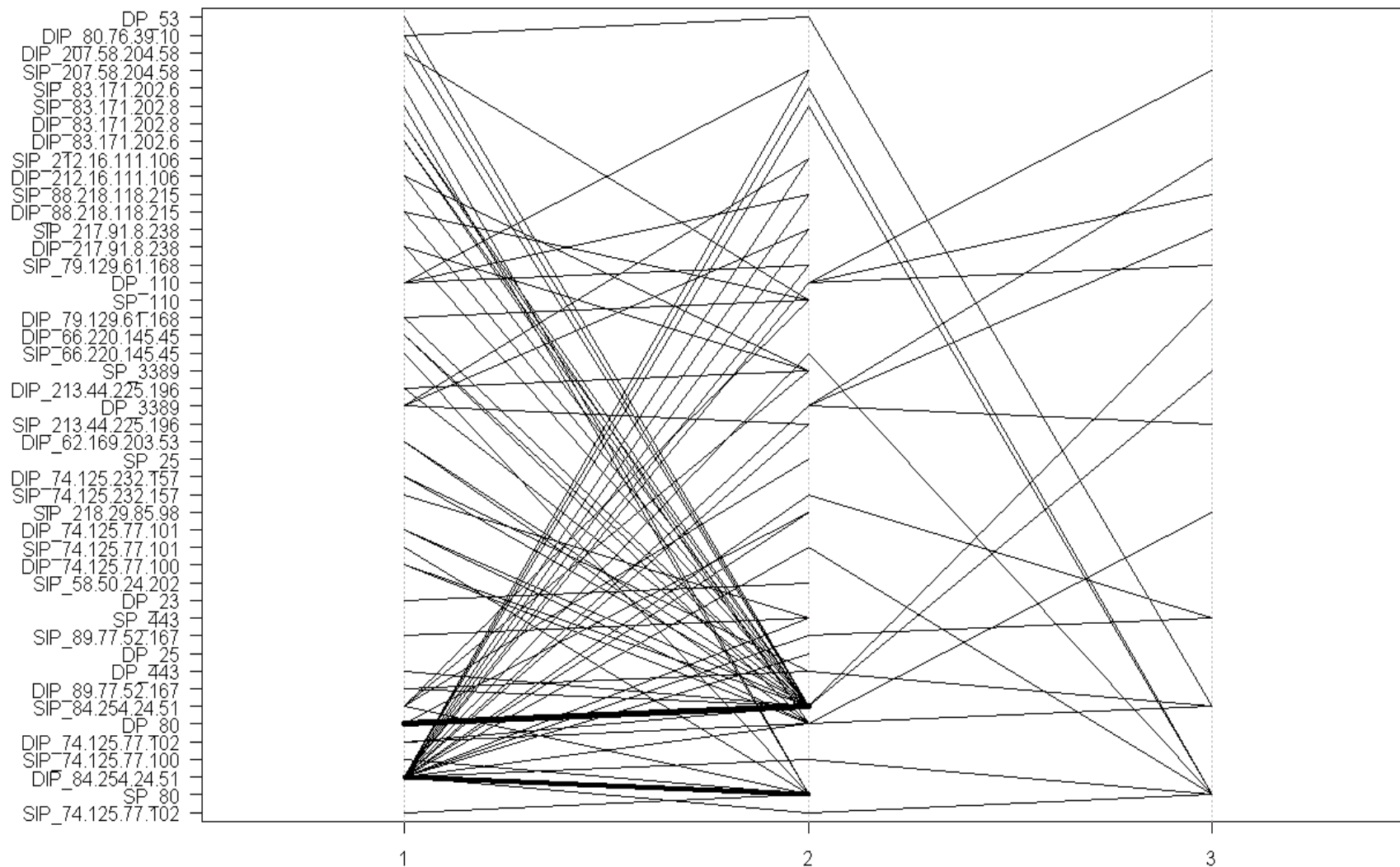
Σχήμα 5.2: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 406 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Parallel coordinates plot for 228 itemsets



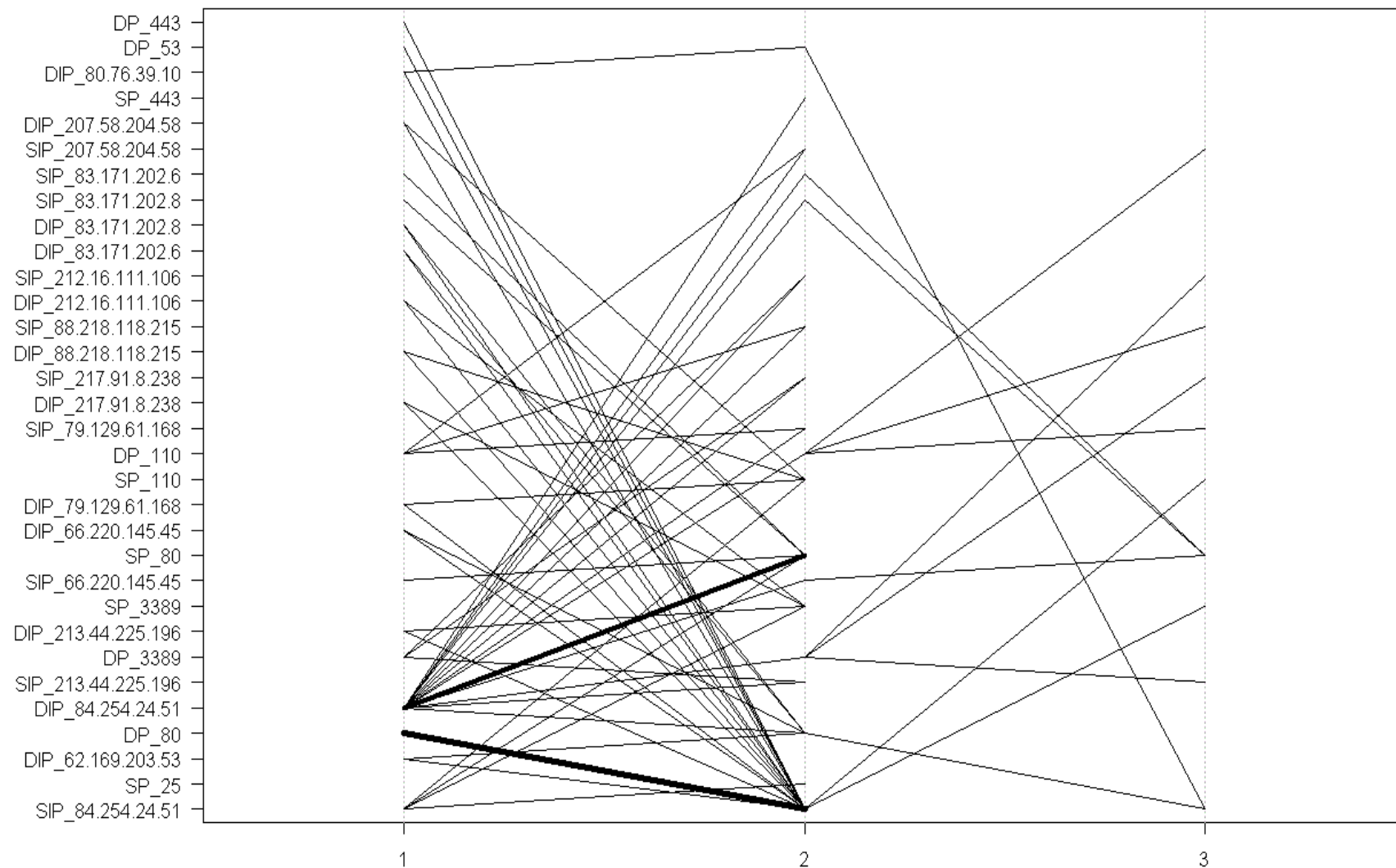
Σχήμα 5.3: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 228 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Parallel coordinates plot for 156 itemsets



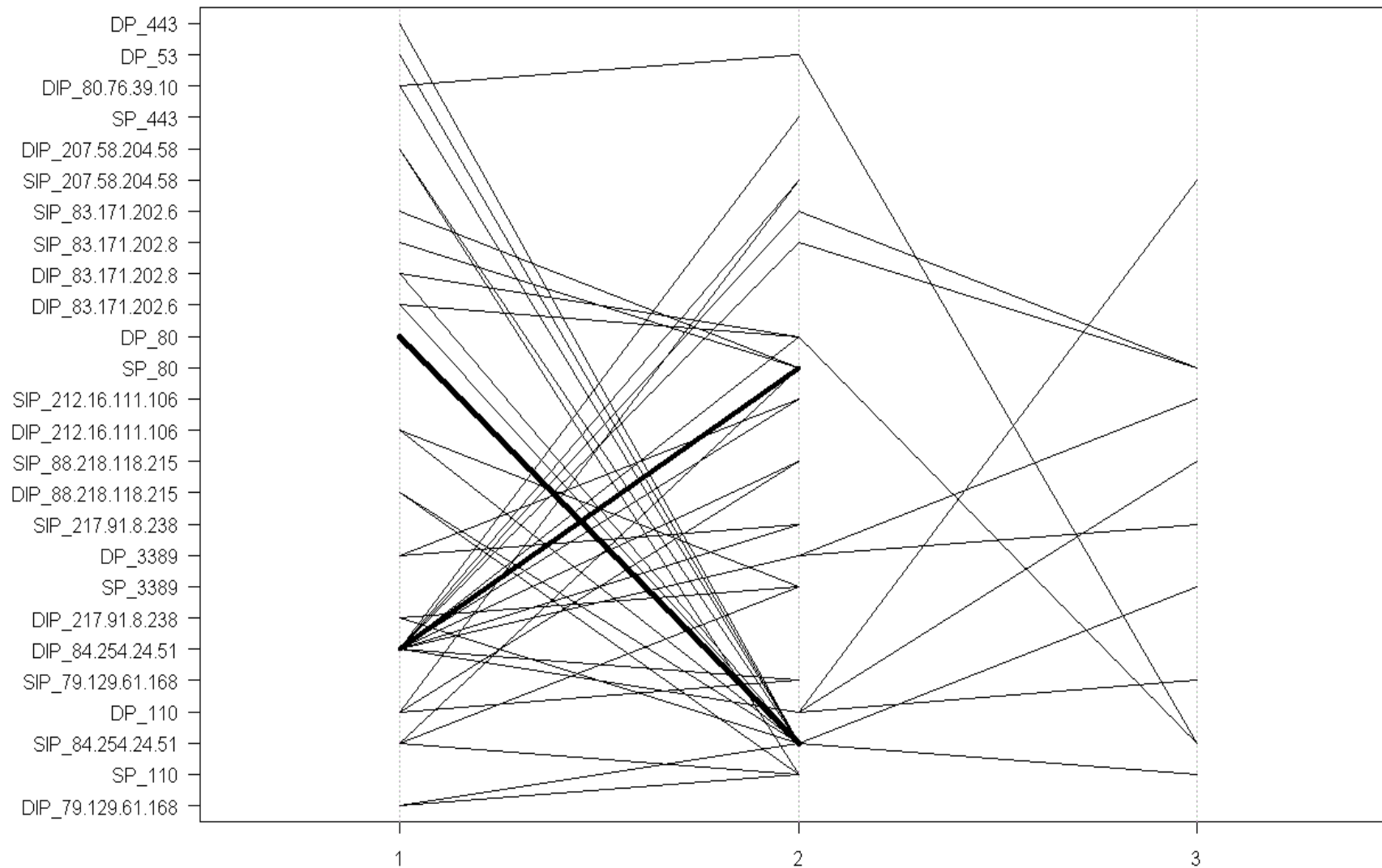
Σχήμα 5.4: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 156 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Parallel coordinates plot for 106 itemsets



Σχήμα 5.5: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 106 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Parallel coordinates plot for 83 itemsets



Σχήμα 5.6: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 83 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Με τη χρήση τιμής «στήριξης» 0.006 ο αριθμός των «συχνών στοιχείων» πέφτει κάτω 100. Συγκεκριμένα ο αριθμός των «συχνών στοιχείων» για τιμή «στήριξης» 0.006 είναι 83. Η αναπαράσταση με τη χρήση γραφήματος παράλληλων συντεταγμένων αυτών των 83 «συχνών στοιχείων» φαίνεται στο σχήμα 6. Οι διαπιστώσεις που γίνονται παρατηρώντας αυτό το γράφημα είναι οι ακόλουθες:

- Είναι πλέον πιο ευδιάκριτα τα «συχνά στοιχεία», μπορεί κάποιος παρατηρώντας το γράφημα να καταλάβει ποια στοιχεία συνθέτουν τα «συχνά στοιχεία» (τουλάχιστον με μικρότερη προσπάθεια από αυτή που θα κατέβαλε στην περίπτωση των γραφημάτων με μεγαλύτερο αριθμό «συχνών στοιχείων»).
- Υπάρχουν τρεις κάθετοι στο γράφημα γεγονός το οποίο δηλώνει ότι ο μέγιστος αριθμός στοιχείων στα υπό αναπαράσταση «συχνά στοιχεία» είναι τρία στοιχεία.
- Γίνεται αντιληπτό ότι δεν υπάρχει η δυνατότητα απεικόνισης των «συχνών στοιχείων» τα οποία αποτελούνται από ένα και μόνο στοιχείο. Αν ταξινομηθούν τα «συχνά στοιχεία» με βάση την τιμή «στήριξης» (support) τα πρώτα έξι «συχνά στοιχεία» είναι τα {SIP_84.254.24.51}, {DIP_84.254.24.51}, {DP_80}, {DP_80, SIP_84.254.24.51}, {SP_80} και {DIP_84.254.24.51, SP_80} όπως φαίνεται και από το παρακάτω σχήμα 5.7:

```
> inspect(sort(f21_25_0.006, by="support"))
  items                                support
1 {SIP_84.254.24.51}          0.532168579
2 {DIP_84.254.24.51}          0.453521242
3 {DP_80}                      0.303798062
4 {DP_80,
  SIP_84.254.24.51}          0.289769395
5 {SP_80}                      0.287183006
6 {DIP_84.254.24.51,
  SP_80}                      0.276479696
```

Σχήμα 5.7: Τα έξι πρώτα «συχνά στοιχεία» με το μεγαλύτερο support

Παρατηρώντας όμως το αντίστοιχο γράφημα (σχήμα 5.6) τα πρώτα τρία από τα παραπάνω «συχνά στοιχεία» δεν εμφανίζονται διότι δεν μπορούν να αναπαρασταθούν.

- Τα πάχη των ακμών δίνουν μία εικόνα συγκριτικά μεταξύ των στοιχείων για το ποια είναι τα συχνότερα «συχνά στοιχεία». Συγκεκριμένα στο γράφημα η πιο παχιά ακμή συνδέει τα στοιχεία DP_80 και SIP_84.254.24.51 ο συνδυασμός των οποίων δίνει το πιο συχνό «συχνό στοιχείο». Στην πραγματικότητα αυτό το «συχνό στοιχείο» είναι τα τέταρτο πιο συχνό αλλά τα τρία πριν από αυτό είναι «συχνά στοιχεία» με ένα μόνο στοιχείο οπότε και δεν μπορούν να αναπαρασταθούν στο γράφημα. Η επόμενη πιο παχιά ακμή είναι αυτή που συνδέει τα στοιχεία DIP_84.254.24.51 και SP_80 τα οποία δίνουν το έκτο πιο συχνό «συχνό στοιχείο». Τα πάχη των υπόλοιπων ακμών δεν μπορούν να ξεχωρίσουν μεταξύ τους.
- Από τα γράφημα και με βάση τις δύο αυτές πιο παχιές ακμές (άρα τα πιο συχνά «συχνά στοιχεία») μπορεί κανείς να εξάγει αμέσως μία πρώτη εικόνα για την κίνηση του δικτύου. Το γεγονός ότι το πιο συχνό «συχνό στοιχείο» (με βάση το γράφημα) είναι ο συνδυασμός των DP_80 και SIP_ 84.254.24.51 δηλώνει ότι το μεγαλύτερο ποσοστό της κίνησης του δικτύου αποτελείται από πακέτα-ροές δεδομένων που περιέχουν στην κεφαλίδα τους ως διεύθυνση αφετηρίας την IP 84.254.24.51 (μία από τις public IP διευθύνσεις του δικτύου – κάθε εξερχόμενο request έχει αυτή ως IP διεύθυνση αφετηρίας) και ως θύρα προορισμού την 80. Αυτό με απλά λόγια σημαίνει ότι τα περισσότερα από τα requests είναι requests προς web sites εκτός του τοπικού δικτύου υπολογιστών. Το συμπέρασμα αυτό ενισχύεται και από το δεύτερο πιο συχνό «συχνό στοιχείο» (DIP_84.254.24.51 και SP_80) το οποίο δηλώνει τις απαντήσεις στα requests από τα web sites.
- Λόγω του τρόπου σχεδιασμού των γραφημάτων παράλληλων συντεταγμένων δεν είναι σαφής ο διαχωρισμός και εντοπισμός όλων των «συχνών στοιχείων». Υπάρχουν δηλαδή «κρυμμένα» «συχνά στοιχεία» στα γραφήματα αυτού του τύπου. Όπως φαίνεται π.χ. από το σχήμα 5.6, στο γράφημα υπάρχουν δύο ακμές οι οποίες συνδέουν τα στοιχεία DIP_83.171.202.6 και DIP_83.171.202.8 από την πρώτη κάθετο, με το στοιχείο DP_80 από τη δεύτερη κάθετο. Από το στοιχείο αυτό (DP_80) ξεκινά μία τρίτη ακμή η οποία καταλήγει στο στοιχείο SIP_84.254.24.51 στην τρίτη κάθετο. Το σχήμα 5.8 δημιουργήθηκε από το σχήμα 5.6 για να γίνει κατανοητό σε ποιες ακμές αναφερόμαστε (ακμές με κόκκινο χρώμα). Το σύνολο αυτών των ακμών και ο τρόπος με τον οποίο αποτυπώνονται στο γράφημα δεν ξεκαθαρίζει ακριβώς το ποια είναι τα «συχνά στοιχεία» τα οποία αναπαρίστανται. Οι δυνατοί

συνδυασμοί των «συχνών στοιχείων» τα οποία θα μπορούσαν να αναπαρίστανται από το σχεδιασμό αυτό είναι οι εξής:

- {DIP_83.171.202.6, DP_80}, {DIP_83.171.202.8, DP_80} και {DP_80, SIP_84.254.24.51}
- {DIP_83.171.202.6, DP_80} και {DIP_83.171.202.8, DP_80, SIP_84.254.24.51}
- {DIP_83.171.202.8, DP_80} και {DIP_83.171.202.6, DP_80, SIP_84.254.24.51}
- {DIP_83.171.202.6, DP_80}, {DIP_83.171.202.6, DP_80, SIP_84.254.24.51} και {DIP_83.171.202.8, DP_80}
- {DIP_83.171.202.8, DP_80}, {DIP_83.171.202.8, DP_80, SIP_84.254.24.51} και {DIP_83.171.202.6, DP_80}
- {DIP_83.171.202.6, DP_80}, {DIP_83.171.202.6, DP_80, SIP_84.254.24.51}, {DIP_83.171.202.8, DP_80} και {DIP_83.171.202.8, DP_80, SIP_84.254.24.51}
- {DIP_83.171.202.6, DP_80}, {DIP_83.171.202.8, DP_80}, {DP_80, SIP_84.254.24.51}, {DIP_83.171.202.6, DP_80, SIP_84.254.24.51} και {DIP_83.171.202.8, DP_80, SIP_84.254.24.51}

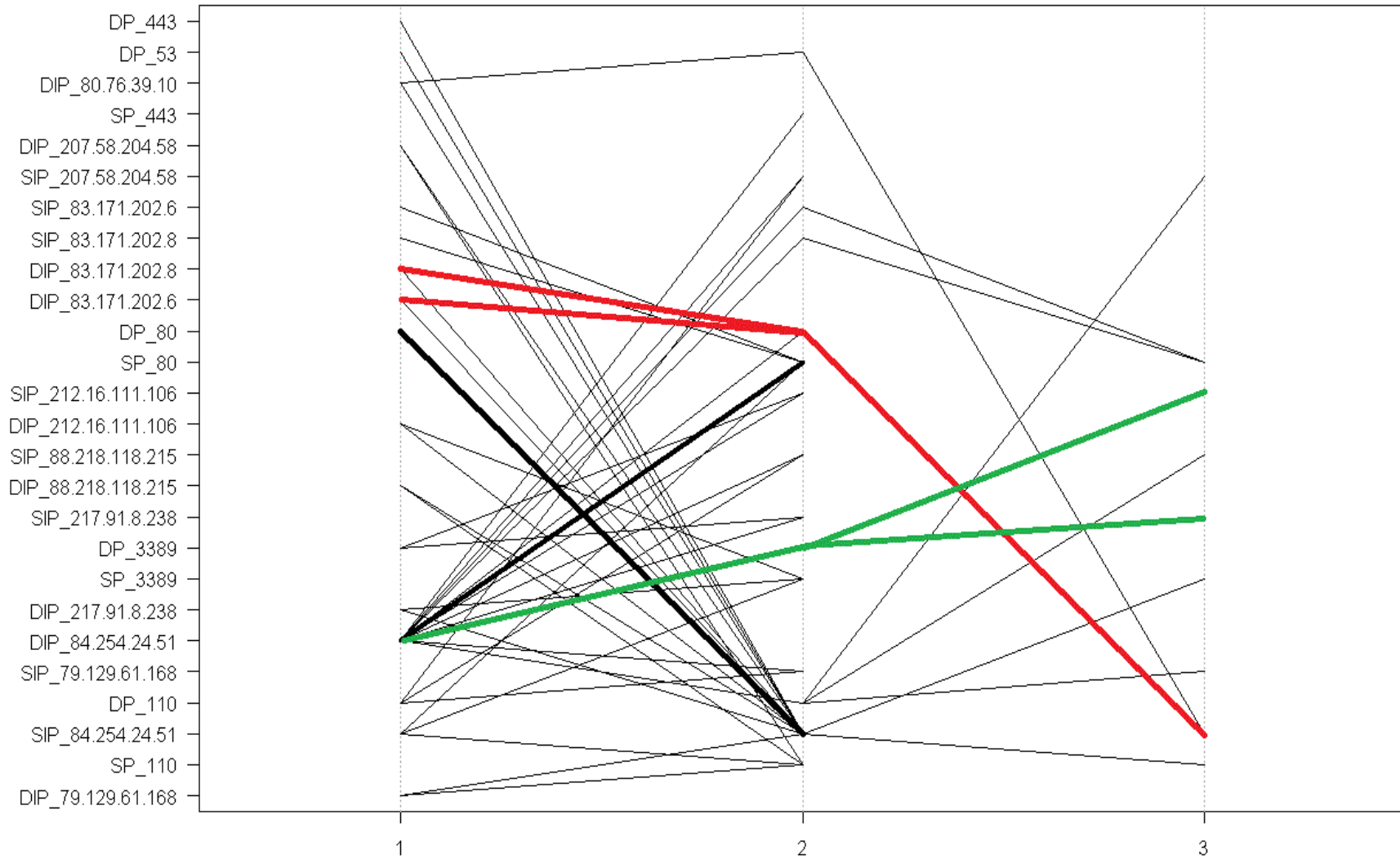
Ελέγχοντας τα αποτελέσματα του αλγορίθμου Apriori (μη γραφική μορφή), τα «συχνά στοιχεία» που τελικά αναπαρίστανται (στο συγκεκριμένο σημείο) στο γράφημα είναι τα τέσσερα ακόλουθα «συχνά στοιχεία»:

- {DIP_83.171.202.6, DP_80},
- {DIP_83.171.202.8, DP_80},
- {DIP_83.171.202.6, DP_80, SIP_84.254.24.51},
- {DIP_83.171.202.8, DP_80, SIP_84.254.24.51}.

Επιβεβαιώνεται λοιπόν η αδυναμία που υπάρχει από τα γραφήματα παράλληλων συντεταγμένων στην ακριβή αναπαράσταση των «συχνών στοιχείων». Αντίστοιχη περίπτωση (σχήμα 5.8, ακμές με πράσινο χρώμα) έχουμε και για τια ακμές που συνδέουν τα στοιχεία DIP_84.254.24.51 (1^η κάθετος), DP_3389 (2^η κάθετος), SIP_217.91.8.238 (3^η κάθετος) και SIP_212.16.111.106 (3^η κάθετος). Τα «συχνά στοιχεία» που αναπαριστούν οι ακμές αυτές είναι τα:

- {DIP_84.254.24.51, DP_3389},
- {DIP_84.254.24.51, DP_3389, SIP_212.16.111.106},
- {DIP_84.254.24.51, DP_3389},
- {DIP_84.254.24.51, DP_3389, SIP_217.91.8.238},
- {DP_3389, SIP_217.91.8.238} και
- {DP_3389, SIP_212.16.111.106}

Parallel coordinates plot for 83 itemsets



Σχήμα 5.8: Γράφημα παράλληλων συντεταγμένων – πρόβλημα αναπαράστασης

Για τιμή στήριξης 0.008 ο αριθμός των «συχνών στοιχείων» είναι 54. Το γράφημα που δημιουργείται εμφανίζεται το σχήμα 5.9. Πλέον τα «συχνά στοιχεία» είναι πολύ λιγότερα από τις προηγούμενες περιπτώσεις και ξεχωρίζουν τα στοιχεία που αποτελούν τα «συχνά στοιχεία». Ισχύουν παρόλα αυτά όλα όσα έχουν επισημανθεί μέχρι τώρα όπως το γεγονός ότι εξακολουθεί και υπάρχει το πρόβλημα που αναφέρθηκε προηγουμένως (σχήμα 5.8).

Αντίστοιχα στο σχήμα 5.10 εμφανίζεται το γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 50 συχνών στοιχείων (τιμή στήριξης 0.009) το οποίο δεν διαφέρει πολύ από το προηγούμενο.

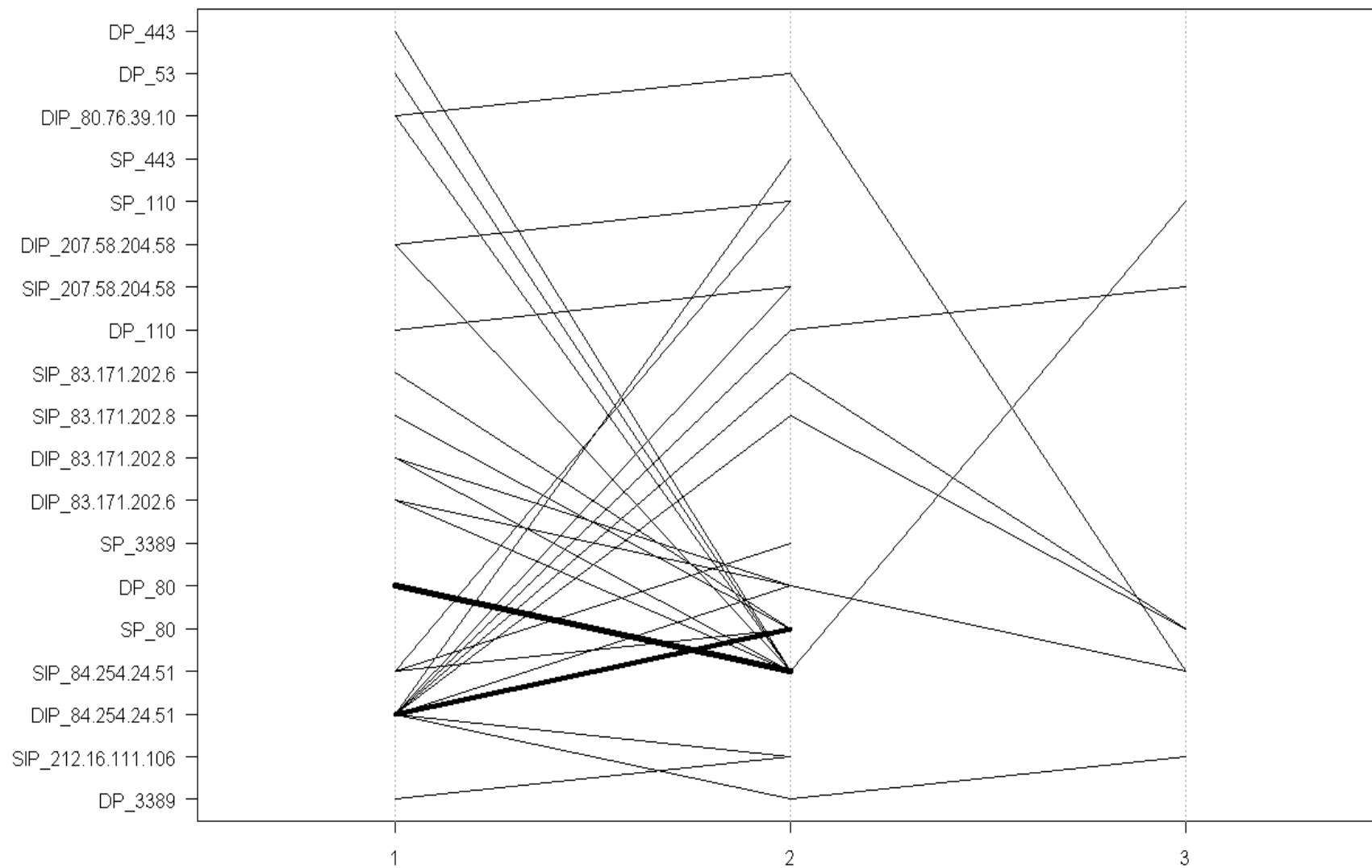
Αυξάνοντας αρκετά την τιμή στήριξης μειώνεται σημαντικά ο αριθμός των «συχνών στοιχείων». Συγκεκριμένα για τιμή στήριξης 0.05 τα «συχνά στοιχεία» γίνονται 20 οπότε και το γράφημα που τα αναπαριστά (σχήμα 5.11) είναι πολύ πιο ευδιάκριτο. Ο παρατηρητής (λογικά ο διαχειριστής δικτύου ή ο αναλυτής της κίνησης δικτύου) ο οποίος θα χρησιμοποιήσει το σχεδιάγραμμα αυτό με στόχο να βγάλει μία εικόνα της κίνησης του υπό παρατήρηση δικτύου θα μπορέσει πιο εύκολα σε σχέση με τα προηγούμενα γραφήματα να εξαγει συμπεράσματα. Μπορεί πιο εύκολα να αντλήσει πληροφορίες όπως να δει ποια είναι τα πιο «συχνά στοιχεία» της κίνησης του δικτύου, δηλαδή ποιος είναι ο μεγαλύτερος όγκος ροών δεδομένων που διακινούνται κυρίως στο υπό παρατήρηση δίκτυο υπολογιστών καθώς και να δει αναλυτικά ποια είναι τα στοιχεία που συνθέτουν τα πιο συχνά «συχνά στοιχεία».

Ρίχνοντας λοιπόν μία γρήγορη ματιά στο γράφημα του σχήματος 5.11 μπορεί να δει κανείς ότι ο κύριος όγκος δεδομένων (για το συγκεκριμένο χρονικό διάστημα για το οποίο συλλέχθηκαν τα δεδομένα) αφορά τρία είδη κινήσεων:

- κίνηση από και προς Web sites (θύρες 80 και 443),
- κίνηση ηλεκτρονικής αλληλογραφίας (θύρα 110)
- κίνηση για name resolution (προς DNS servers) (θύρες 53) (η οποία βέβαια θα μπορούσε να θεωρηθεί τμήμα της διαδικασίας πρόσβασης σε web sites οπότε και ανήκει στην πρώτη κατηγορία).

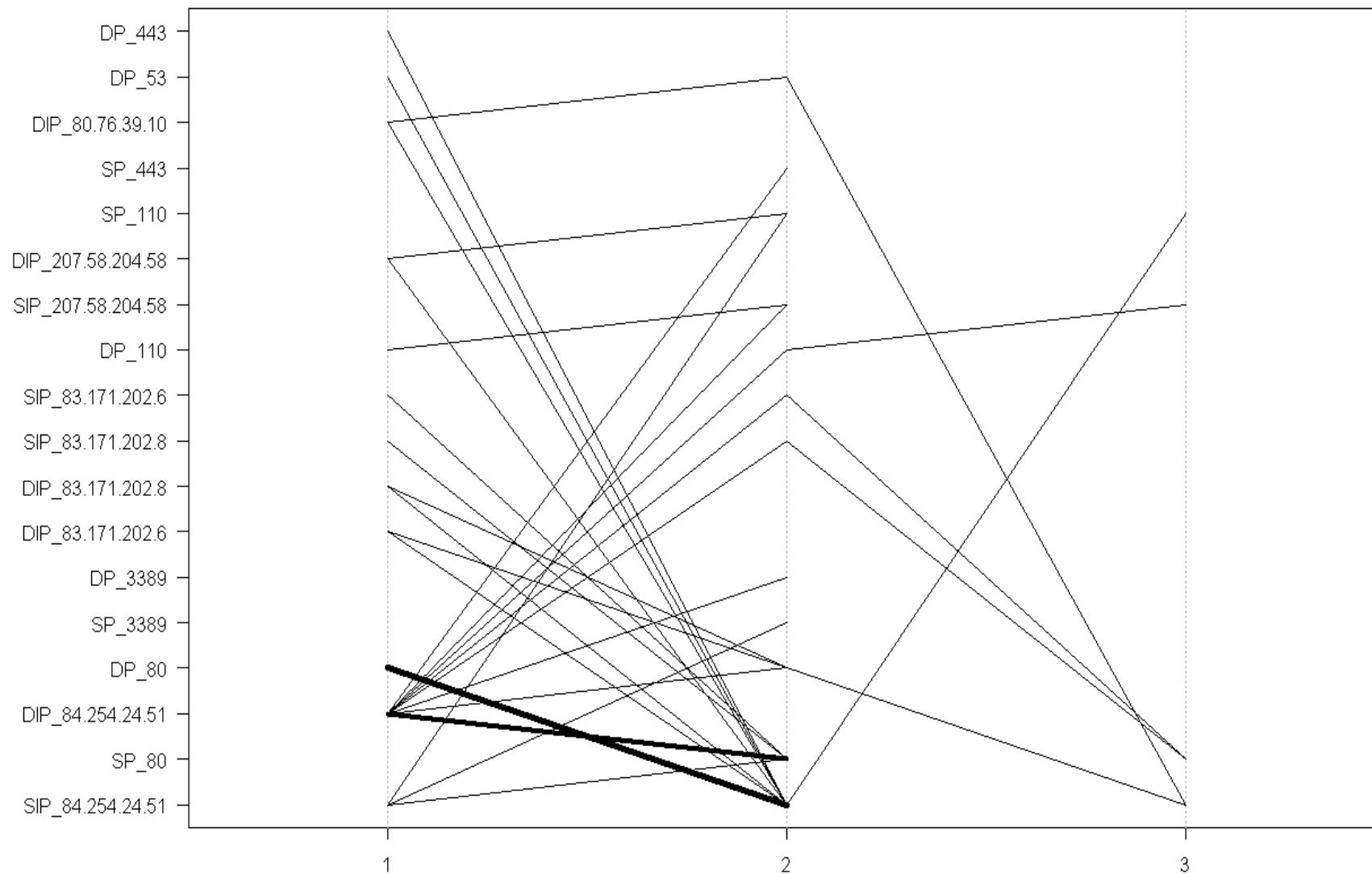
Για την κατανόηση των παραπάνω δημιουργήθηκε το σχεδιάγραμμα του σχήματος 5.12 στο οποίο έχουν χρωματιστεί με κόκκινο οι ακμές του σχεδιαγράμματος παράλληλων συντεταγμένων που αφορούν κίνηση Web sites, με πράσινο αυτές που αφορούν κίνηση ηλεκτρονικής αλληλογραφίας και με μπλε αυτές που αφορούν κίνηση προς DNS servers (για καλύτερη κατανόηση του γραφήματος και του είδους τις υπό αναπαράσταση κίνησης να σημειωθεί ότι η εμφανιζόμενη IP διεύθυνση 80.76.39.10 είναι IP διεύθυνση ενός public DNS εταιρείας ISP.

Parallel coordinates plot for 54 itemsets

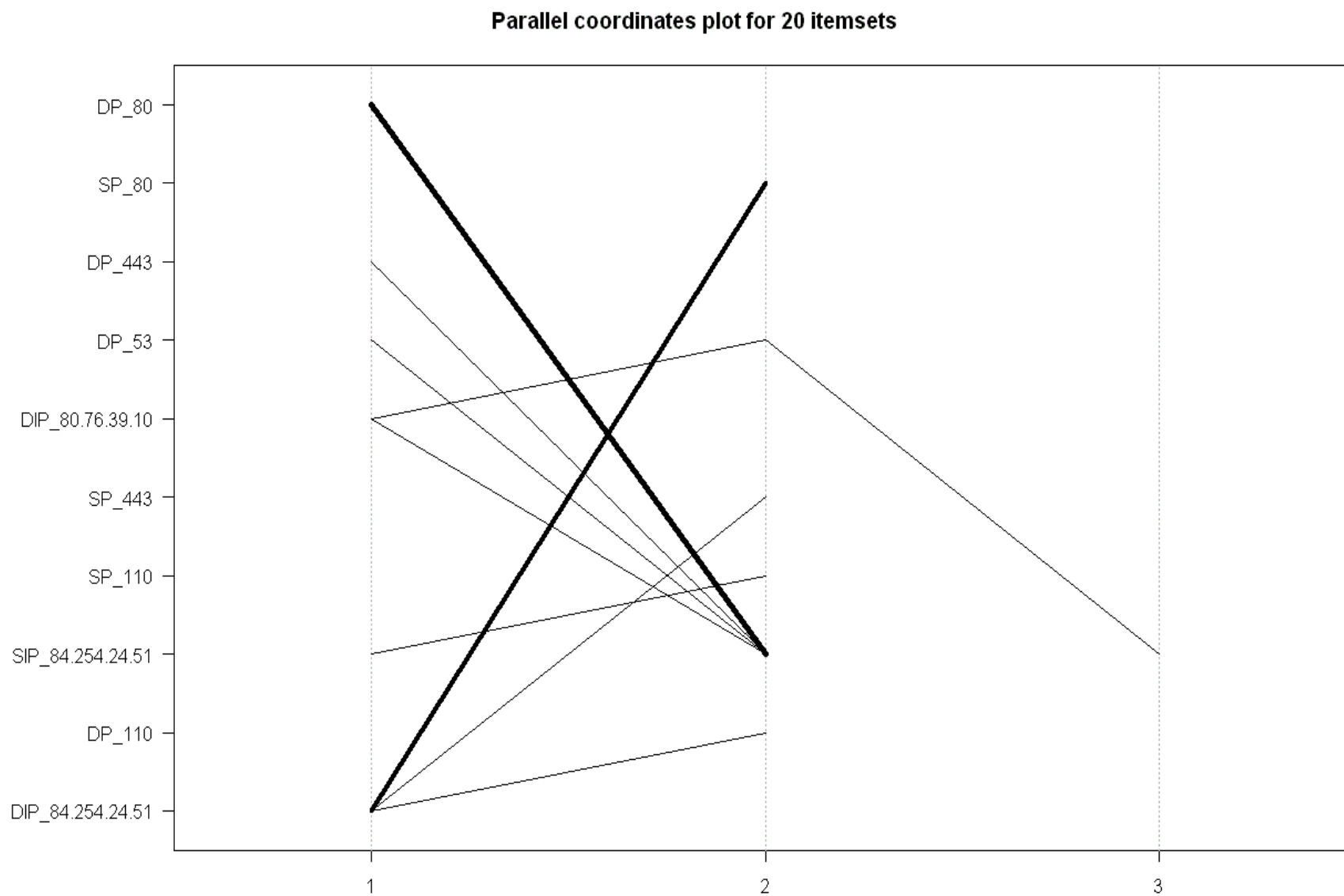


Σχήμα 5.9: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 54 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

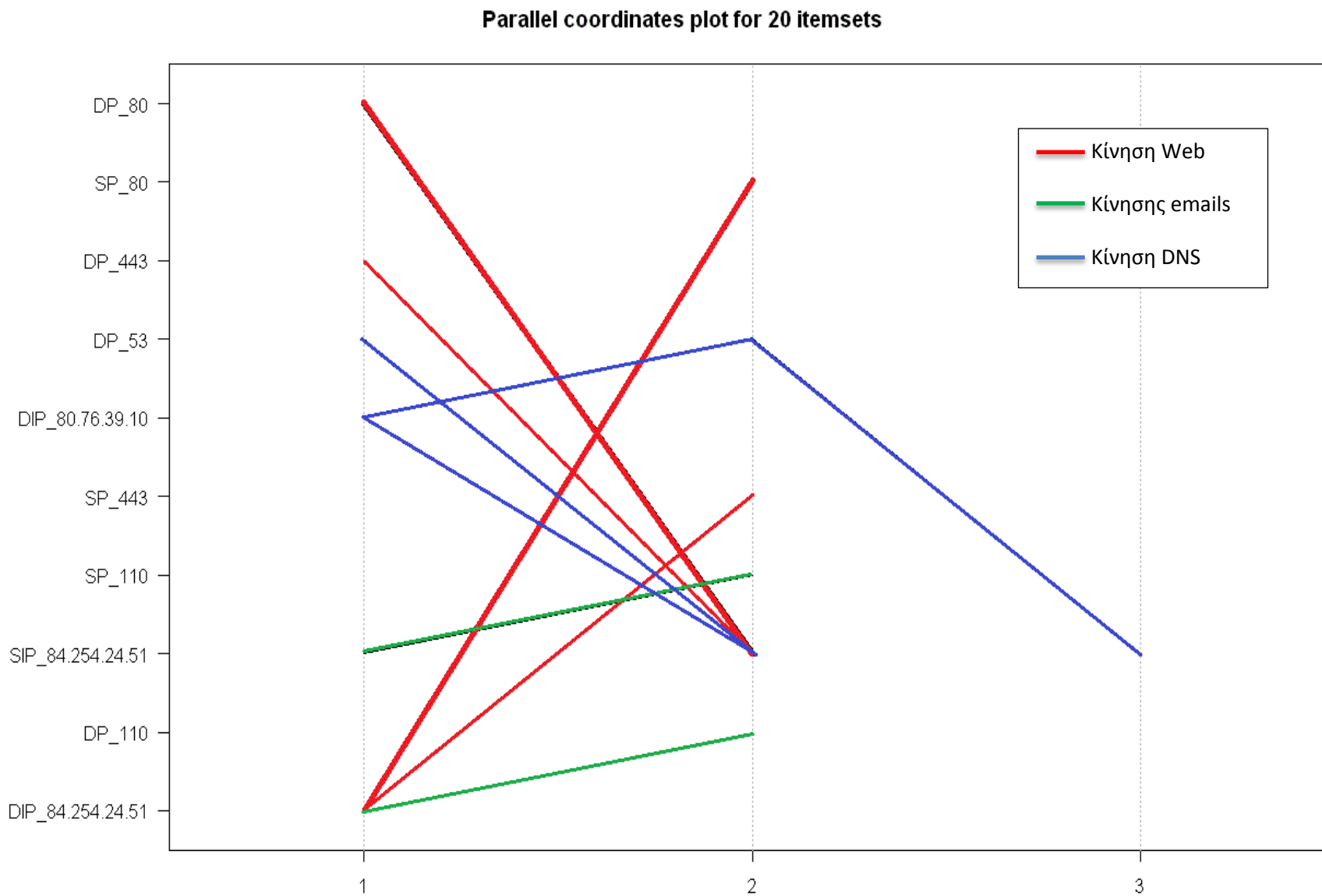
Parallel coordinates plot for 50 itemsets



Σχήμα 5.10: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 50 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)



Σχήμα 5.11: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 20 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)



Σχήμα 5.12: Γράφημα παράλληλων συντεταγμένων με χρωματισμό για την κατηγοριοποίηση είδους κίνησης (FIs)

Το πρόβλημα βέβαια της μη αναπαράστασης των «συχνών στοιχείων» τα οποία αποτελούνται από ένα στοιχείο εξακολουθεί και υπάρχει και σε αυτή την περίπτωση (σχήματα 5.11 και 5.12). Ενώ το γράφημα αναπαριστά είκοσι «συχνά στοιχεία» εμφανίζονται μόνο δέκα ακμές, στην ουσία αναπαριστώνται έντεκα μόνο «συχνά στοιχεία». Αυτό συμβαίνει διότι από τα είκοσι «συχνά στοιχεία» που προκύπτουν από τον αλγόριθμο Apriori τα εννέα είναι «συχνά στοιχεία» τα οποία αποτελούνται από ένα μόνο στοιχείο.

Μπορεί ο μικρός αριθμός στοιχείων στο γράφημα να το κάνει ποιο ευδιάκριτο σε σχέση με τα γραφήματα παράλληλων συντεταγμένων που αναπαριστούν μεγάλο αριθμό «συχνών στοιχείων» παρόλα αυτά θέλει κάποια προσπάθεια για την κατανόηση των αναπαριστώμενων πληροφοριών. Παράδειγμα αποτελεί το γεγονός ότι θα έπρεπε ο παρατηρητής να γνωρίζει ότι η IP διεύθυνση 80.76.39.10 είναι IP διεύθυνση ενός public DNS διαφορετικά δεν θα μπορούσε να βγάλει εύκολα το συμπέρασμα ότι το «συχνό στοιχείο» {DIP_80.76.39.10, DP_53, SIP_84.254.24.51} αφορά αιτήματα (requests) προς DNS server. Αυτό σημαίνει ότι για την καλύτερη και ευκολότερη άντληση των πληροφοριών θα έπρεπε πριν τη δημιουργία του γραφήματος να ενσωματωθεί ένα πρόσθετο βήμα στη διαδικασία, το οποίο θα μετέτρεπε (ίσως χρησιμοποιώντας κάποια τεχνική-αλγόριθμο WHO IS) τις προς αναπαράσταση διευθύνσεις σε φιλικά προς τον χρήστη ονόματα.

Άλλο μειονέκτημα που παρατηρείται αποτελεί το γεγονός ότι δεν υπάρχουν οριζόντιες γραμμές για την ευκολότερη αντιστοίχιση των τιμών του άξονα Υ με τις καταλήξεις των ακμών με αποτέλεσμα να χρειάζεται προσπάθεια για τη δημιουργία οπτικά των «συχνών στοιχείων».

Μία ακόμη παρατήρηση που μπορεί να γίνει αφορά τον συσχετισμό της αναπαράστασης «συχνών στοιχείων» με τη χρήση παράλληλων συντεταγμένων και την ανίχνευση επιθέσεων στο δίκτυο. Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο έχουν πραγματοποιηθεί μέχρι τώρα αρκετές μελέτες που σχετίζονται με την δημιουργία “virus signatures” με τη χρήση σχεδιαγραμμάτων παράλληλων συντεταγμένων (όπως στην δουλειά [11]). Κάτι τέτοιο δεν διαφαίνεται να είναι εφικτό με βάση τον τρέχον τρόπο δημιουργίας των γραφημάτων αυτών για την αναπαράσταση «συχνών στοιχείων». Αυτό οφείλεται κυρίως στο γεγονός ότι κάθε κάθετος στο σχεδιάγραμμα δεν αναπαριστά συγκεκριμένη πληροφορία κάθε φορά, δηλαδή π.χ. η πρώτη κάθετος δεν σχετίζεται μόνο με source IP addresses, η δεύτερη με Destination IP addresses κλπ., με αποτέλεσμα να μην είναι δυνατή η δημιουργία «γραφικών υπογραφών» επιθέσεων.

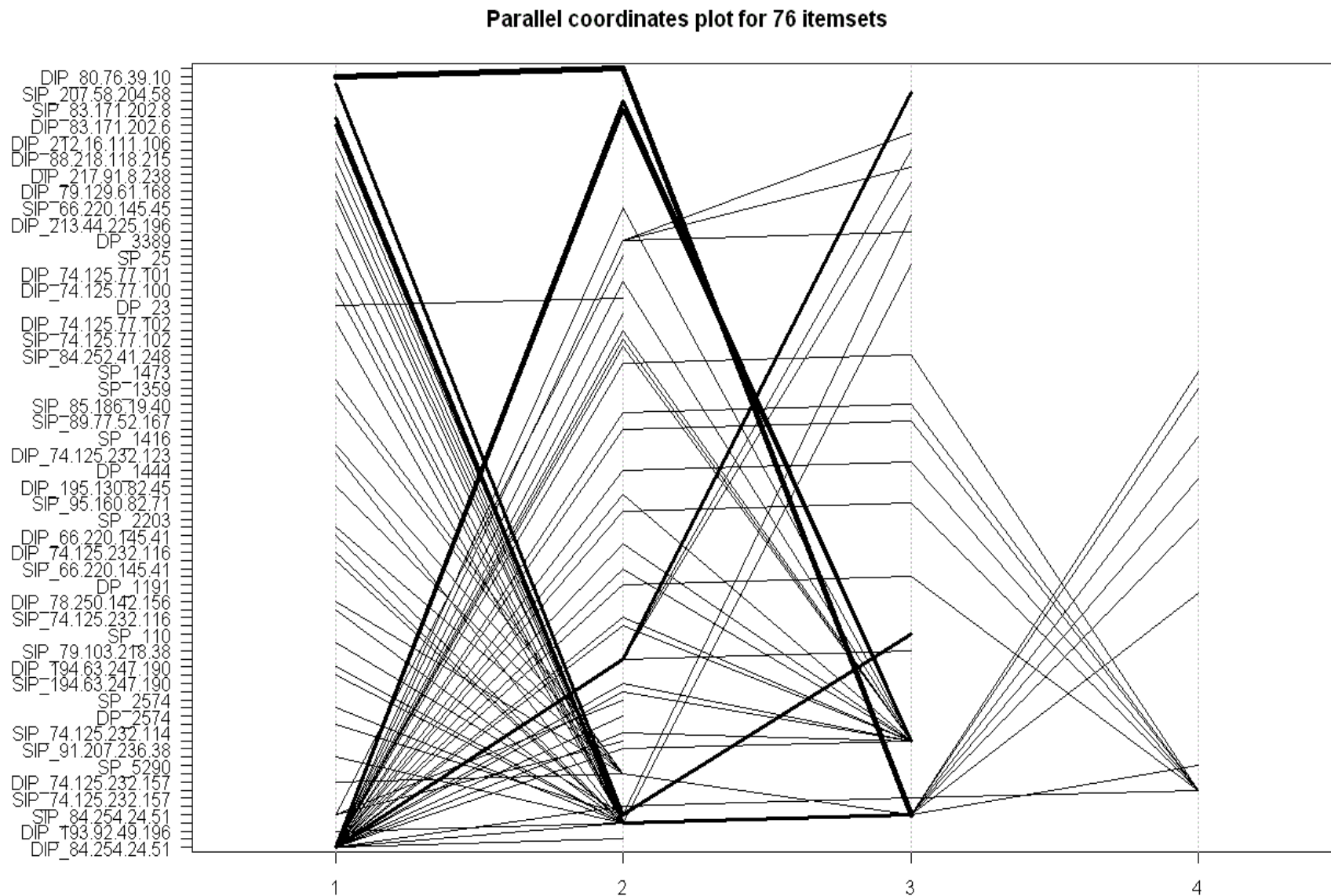
5.3 Αποτελέσματα Γραφικής Αναπαράστασης “Maximal Frequent Itemsets” και “Closed Frequent Itemsets” με τη Χρήση Παράλληλων Συντεταγμένων.

Τα «συχνά στοιχεία» τα οποία παράγονται από κάποιον αλγόριθμο εξόρυξης γνώσης όπως είναι ο αλγόριθμος Apriori μπορεί να επαναλαμβάνονται. Μπορεί δηλαδή να υπάρχουν «συχνά στοιχεία» μεταξύ των παραγόμενων, τα οποία να αποτελούν υποσύνολα άλλων. Θα μπορούσε δηλαδή «συχνά στοιχεία» να είναι τα {DIP_192.168.0.1}, {DP_3389} και {DIP_192.168.0.1, DP_3389}. Η ύπαρξη του τελευταίου «συχνού στοιχείου» δηλώνει (λόγο της διαδικασία παραγωγής με βάση τον αλγόριθμο Apriori) ότι και τα στοιχεία που το συνθέτουν είναι και αυτά «συχνά στοιχεία» οπότε μία αναπαράσταση και των τριών δεν θα είχε κάποια πρόσφορα στη γνώση. Για αυτό το σκοπό και την αποφυγή εμφάνισης επαναλαμβανόμενων «συχνών στοιχείων» υπάρχουν τα maximal frequent itemsets, η οπτική αναπαράσταση των οποίων με τη χρήση γραφημάτων παράλληλων συντεταγμένων αναφέρεται στην παρούσα παράγραφο.

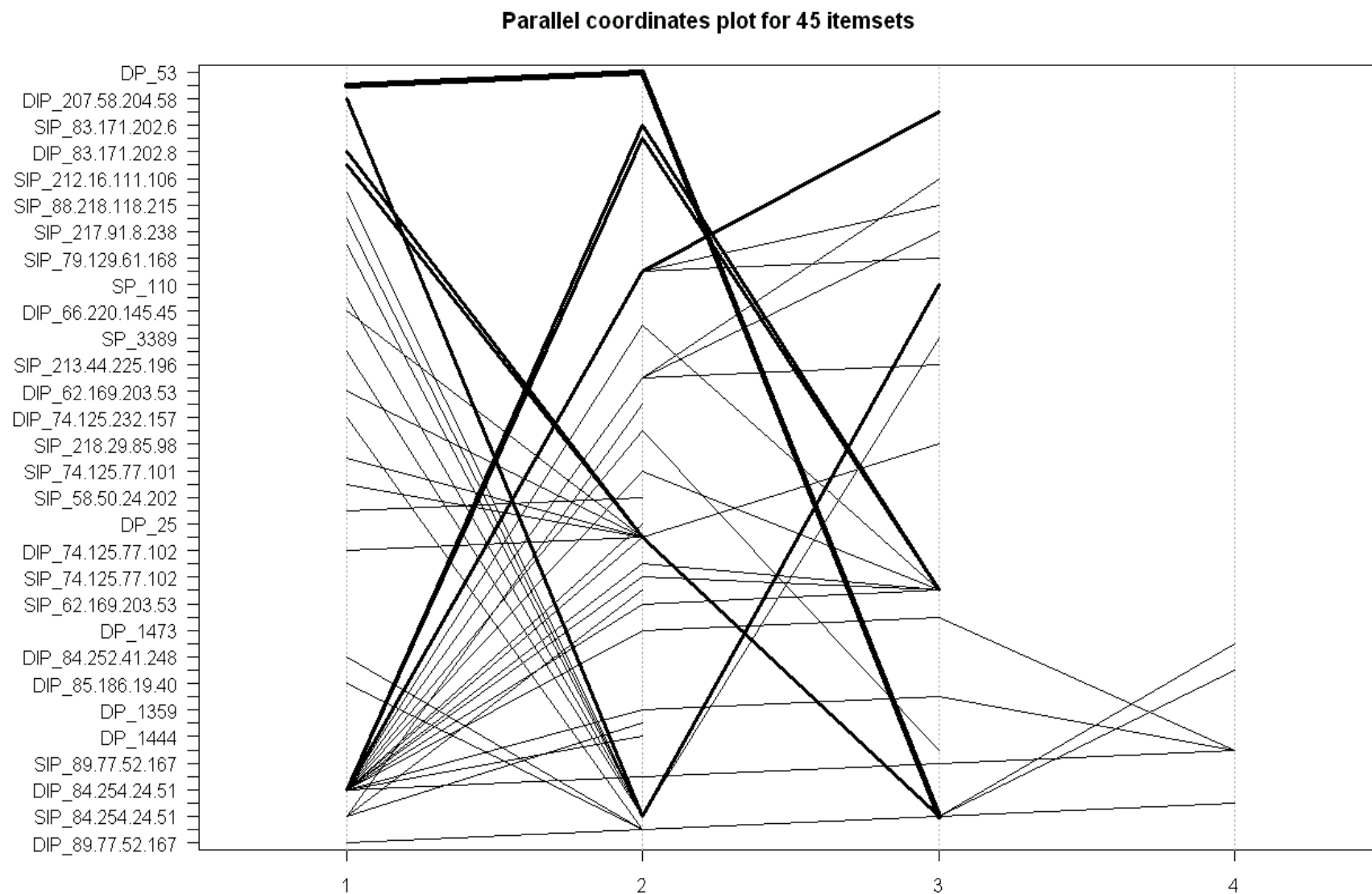
Θα παρατεθούν λοιπόν γραφήματα τα οποία χρησιμοποιούν την τεχνική των παράλληλων συντεταγμένων για την αναπαράσταση δεδομένων που προέρχεται από καταγραφείσα κίνηση μεταξύ 21 και 25 Μαρτίου 2011 στο υπό παρατήρηση δίκτυο υπολογιστών.

Στο σχήμα 5.13 φαίνεται ένα γράφημα παράλληλων συντεταγμένων το οποίο αναπαριστά 76 maximal frequent itemsets. Οι παρατηρήσεις δεν αλλάζουν σε σχέση με τις παρατηρήσεις που καταγράφηκαν στην προηγούμενη παράγραφο και αφορούσαν τη γραφική αναπαράσταση με τη χρήση παράλληλων συντεταγμένων των «συχνών στοιχείων». Ίσως η μόνη διαφοροποίηση είναι το γεγονός ότι υπάρχει εδώ μεγαλύτερος αριθμός παχιών ακμών το οποίο όμως οφείλεται στο γεγονός ότι τα υπό αναπαράσταση «συχνά στοιχεία» έχουν μεταξύ τους κοντινές τιμές «στήριξης».

Αυξάνοντας την τιμή «στήριξης» (0.03) τα παραγόμενα «συχνά στοιχεία» πέφτουν στα 45. Το σχήμα με τη γραφική αναπαράστασή τους χρησιμοποιώντας γράφημα παράλληλων συντεταγμένων φαίνεται το σχήμα 5.14. Δεν μπορεί να πει κανείς παρατηρώντας το γράφημα αυτό ότι η οπτική αναπαράσταση τόσο μεγάλου αριθμού «συχνών στοιχείων» προσφέρει κάτι περισσότερο σε σχέση με την απλή (μη γραφική) παράθεση σε κείμενο των «συχνών στοιχείων».



Σχήμα 5.13: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 76 Maximal FIs (δεδομένα 21-25 Μαρτίου 2011)



Σχήμα 5.14: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 45 Maximal FIs (δεδομένα 21-25 Μαρτίου 2011)

Μια ποιο ξεκάθαρη εικόνα των «συχνών στοιχείων» έχουμε εφόσον στον αλγόριθμο χρησιμοποιηθεί τιμή «στήριξης» 0.006 (σχήμα 5.15). Σε αυτή την περίπτωση παράγονται 20 «συχνά στοιχεία».

Όπως αναφέρθηκε και πριν, κανένα “maximal frequent itemset” από αυτά που αναπαριστώνται δεν είναι υπερσύνολο κάποιου άλλου, γεγονός το οποίο σημαίνει ότι μειώνεται το πρόβλημα απεικόνισης (εντοπισμού των ακριβή «συχνών στοιχείων»), όπως αυτό αναφέρθηκε ποιο πάνω στην περίπτωση της γραφικής απεικόνισης «συχνών στοιχείων». Για να γίνει ποιο κατανοητό αυτό δίνεται το εξής παράδειγμα: Στο σχήμα 5.15 η ακμή που συνδέει τα στοιχεία {DIP_80.76.39.10}, {DP_53} και {SIP_84.254.24.51}, είναι σίγουρο ότι αναπαριστά το «συχνό στοιχείο» {DIP_80.76.39.10, DP_53, SIP_84.254.24.51} και μόνο αυτό. Για την καλύτερη κατανόηση του παραδείγματος δημιουργήθηκε το σχήμα 5.16 το οποίο είναι αντίγραφο του σχήματος 5.15 με τη διαφορά ότι η αναφερόμενη στο παράδειγμα ακμή έχει χρωματιστεί κόκκινη. Στην περίπτωση που το γράφημα αναπαριστούσε «συχνά στοιχεία» και όχι maximal frequent itemsets δεν θα μπορούσαμε με σιγουριά να πούμε το ποια είναι ακριβώς τα «συχνά στοιχεία» που αναπαριστώνται. Θα μπορούσε να αναπαριστάται οποιοσδήποτε συνδυασμός των τριών στοιχείων:

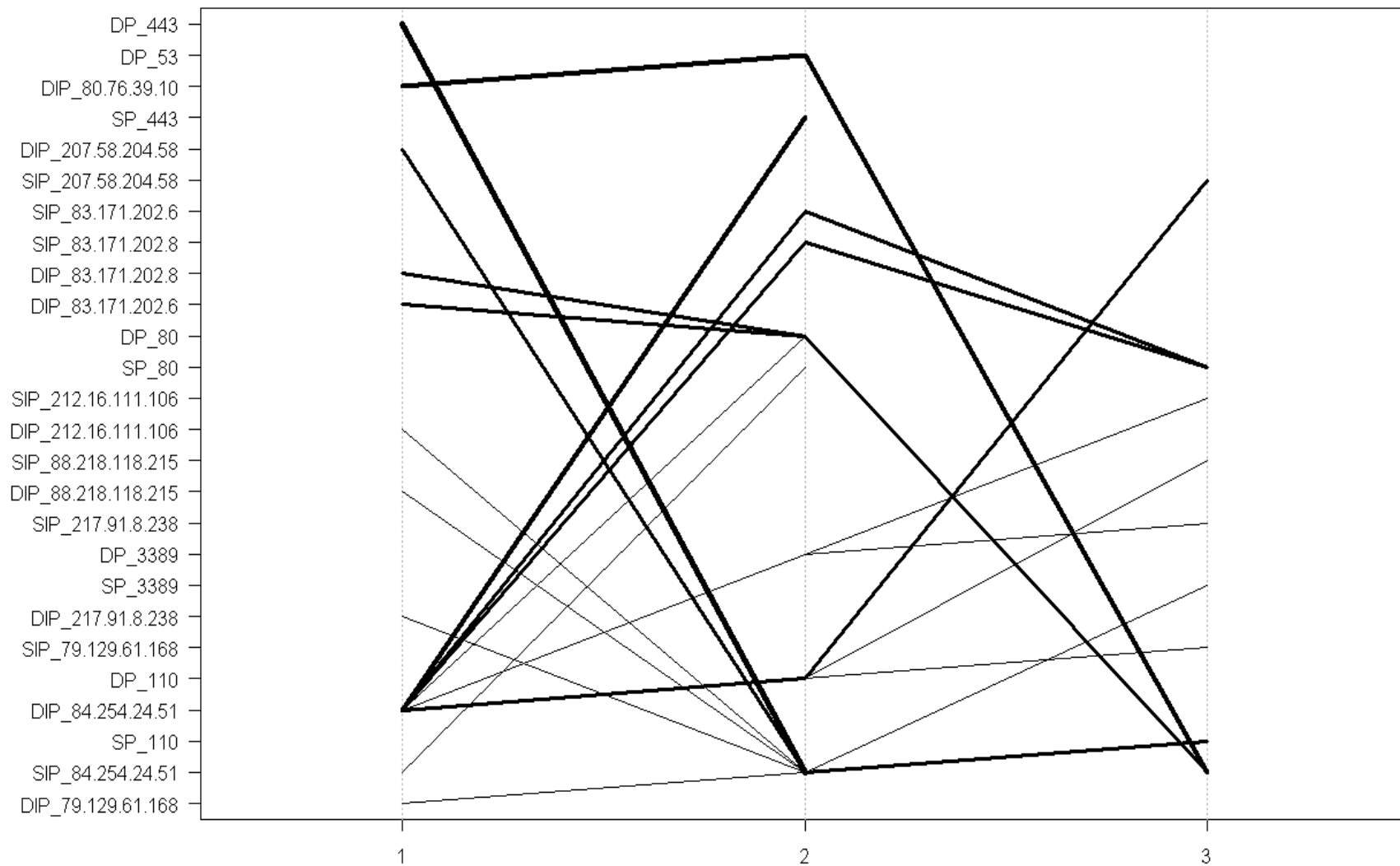
- {DIP_80.76.39.10, DP_53, SIP_84.254.24.51}
- {DIP_80.76.39.10, DP_53}, {DIP_80.76.39.10, DP_53, SIP_84.254.24.51}
- {DP_53, SIP_84.254.24.51}, {DIP_80.76.39.10, DP_53, SIP_84.254.24.51}
- {DIP_80.76.39.10, DP_53}, {DP_53, SIP_84.254.24.51}

Φυσικά το φαινόμενο δεν εξαφανίζεται τελείως διότι όπως φαίνεται με τις πράσινες ακμές στο σχήμα 16, δεν μπορούμε να πούμε με ακρίβεια ποια είναι τα «συχνά στοιχεία» που αναπαριστώνται. Στην περίπτωση αναπαράστασης των maximal frequent itemsets υπάρχουν 3 πιθανοί συνδυασμοί «συχνών στοιχείων» που μπορεί να αναπαριστώνται. Στην περίπτωση βέβαια αναπαράστασης «συχνών στοιχείων» (όχι maximal frequent itemsets) θα είχαμε πολύ περισσότερους πιθανούς συνδυασμούς.

Μία άλλη παρατήρηση που μπορεί να γίνει με βάση το σχεδιάγραμμα του σχήματος 5.15 και συγκρίνοντάς το με το σχεδιάγραμμα του σχήματος 5.11 το οποίο αναπαριστά τον ίδιο αριθμό «συχνών στοιχείων» (το σχήμα 5.11 αναπαριστά 20 συχνά στοιχεία, ενώ το σχήμα 5.15 αναπαριστά 20 maximal frequent itemsets) είναι το γεγονός ότι ο αριθμός των ακμών είναι μεγαλύτερος όπως

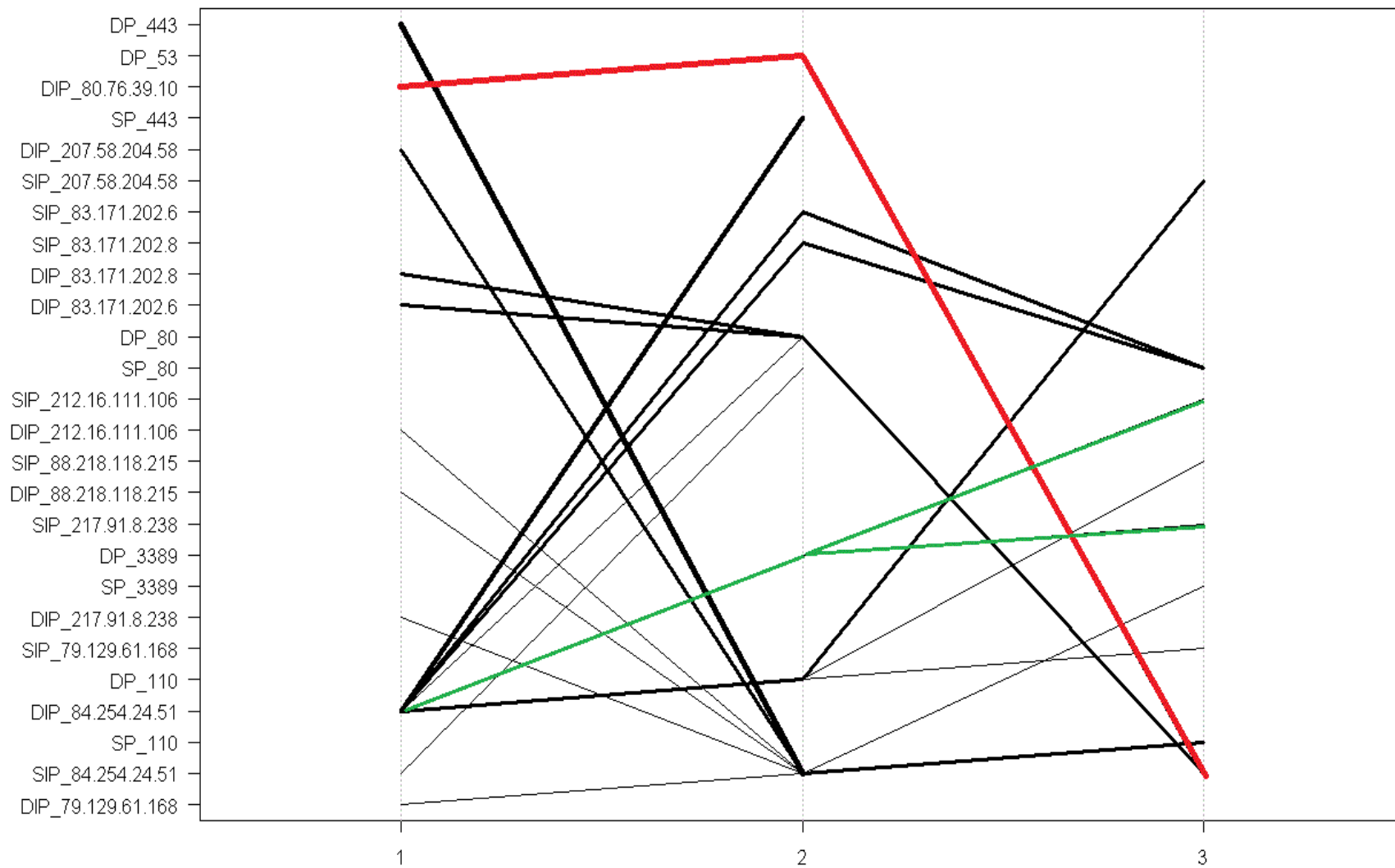
μεγαλύτερος είναι και ο αριθμός των στοιχείων στον άξονα Υ. Αυτό είναι αποτέλεσμα του γεγονότος ότι τα αναπαριστώμενα στοιχεία είναι maximal frequent itemsets (κανένα δεν είναι υποσύνολο του άλλου) και όχι απλά «συχνά στοιχεία».

Parallel coordinates plot for 20 itemsets



Σχήμα 5.15: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 20 Maximal FIs (δεδομένα 21-25 Μαρτίου 2011)

Parallel coordinates plot for 20 itemsets



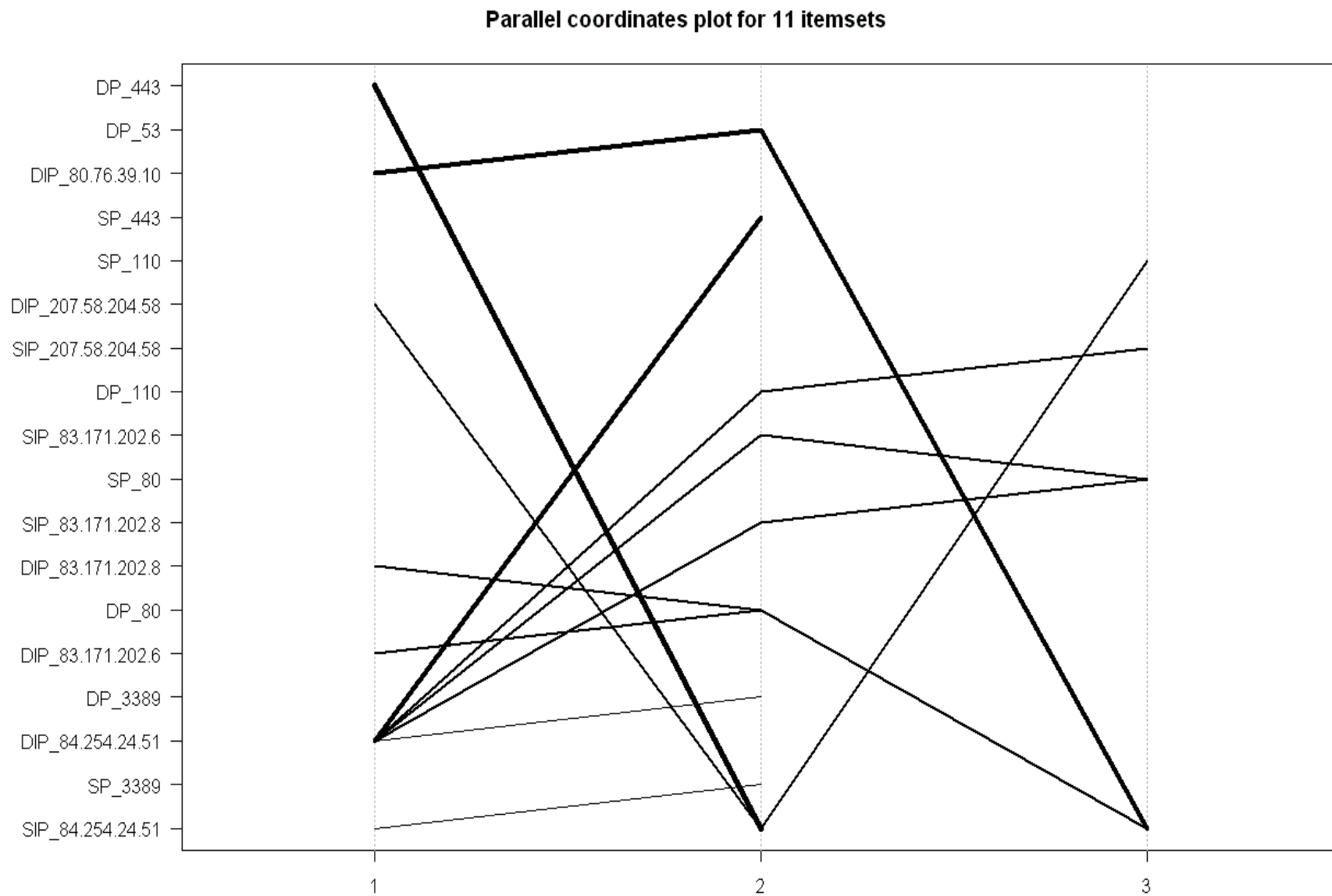
Σχήμα 5.16: Γράφημα παράλληλων συντεταγμένων – πρόβλημα αναπαράστασης

Ένα γράφημα το οποίο θα μπορούσε κανείς να χρησιμοποιήσει για να καταλάβει εύκολα και γρήγορα τα αναπαριστώμενα «συχνά στοιχεία» (maximal frequent itemsets) είναι το γράφημα του σχήματος 5.17 το οποίο αναπαριστά 11 maximal frequent itemsets (Η τιμή στήριξης που χρησιμοποιήθηκε για την εξαγωγή τους ήταν 0.02). Στο γράφημα αυτό εμφανίζονται πέντε (5) είδη κίνησης:

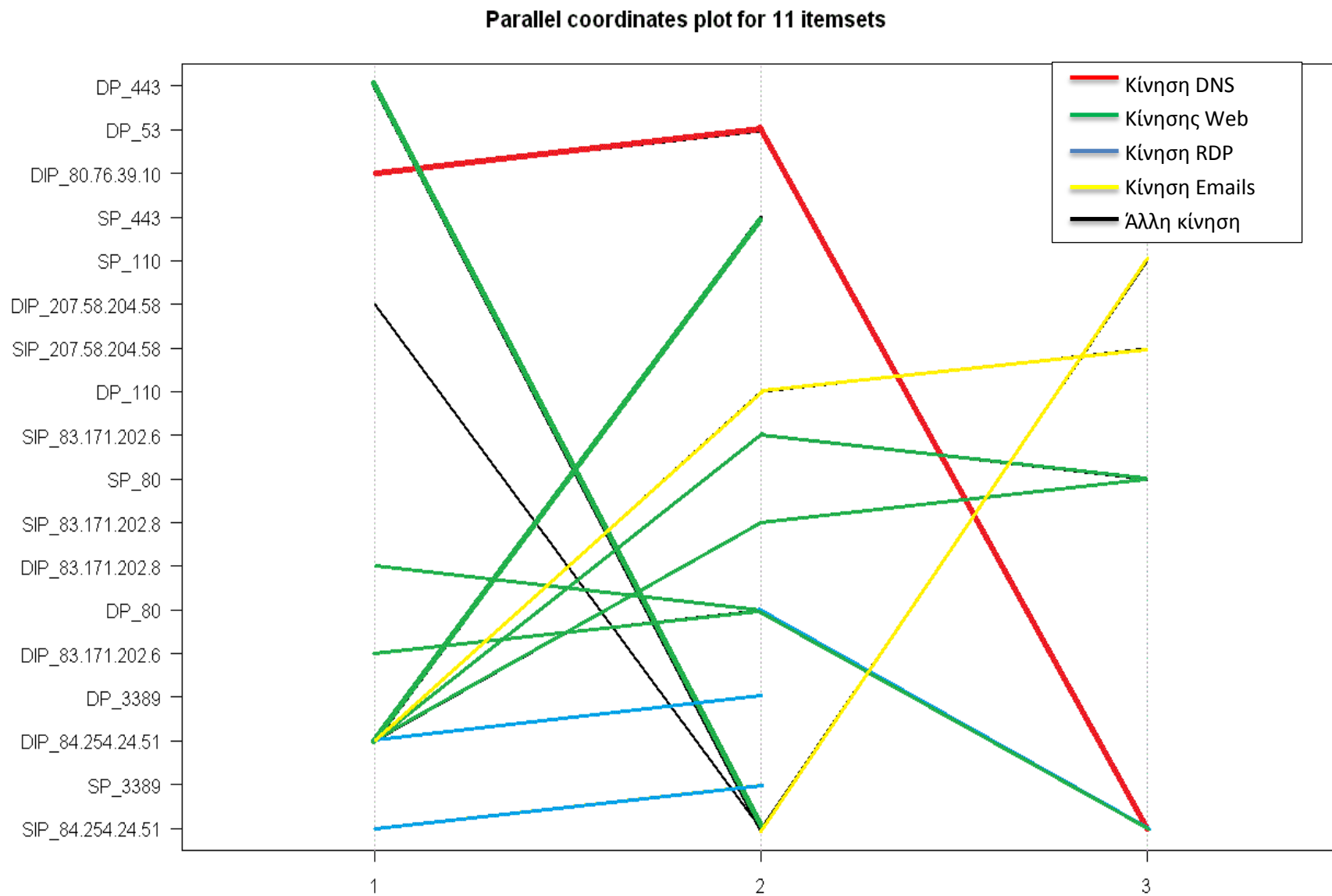
1. κίνηση από και προς Web sites (θύρες 80 και 443),
2. κίνηση για name resolution (προς DNS servers) (θύρες 53) (η οποία βέβαια θα μπορούσε να θεωρηθεί τμήμα της διαδικασίας πρόσβασης σε web sites οπότε και ανήκει στην πρώτη κατηγορία)
3. κίνηση ηλεκτρονικής αλληλογραφίας (θύρα 110)
4. κίνηση remote desktop connections (θύρα 3389)
5. άλλη κίνηση (ακμή που συνδέει τα SIP_84.254.24.51, DIP_207.58.204.58)

Για την κατανόηση των παραπάνω δημιουργήθηκε το σχεδιάγραμμα του σχήματος 5.18 στο οποίο έχουν χρωματιστεί με πράσινο οι ακμές του σχεδιαγράμματος παράλληλων συντεταγμένων που αφορούν κίνηση Web sites, με κόκκινο αυτές που αφορούν κίνηση προς DNS servers, με κίτρινο αυτές που αφορούν κίνηση ηλεκτρονικής αλληλογραφίας, με μπλε αυτές που δείχνουν κίνηση συνδέσεων απομακρυσμένης πρόσβασης και με μαύρο η άλλη κίνηση.

Η κίνηση που αναπαριστάται με την ακμή μαύρου χρώματος στο γράφημα δεν περιέχει κάποιο χαρακτηριστικό για να αναγνωριστεί εύκολα από τον παρατηρητή. Θα πρέπει αυτός να καταβάλει πρόσθετη προσπάθεια για την κατανόησή της. Σε περίπτωση όμως που πριν το σχεδιασμό του γραφήματος εφαρμοζόταν τεχνική ώστε να αντικατασταθούν οι υπό αναπαράσταση IP διευθύνσεις (τεχνική WHO IS) με τα ονόματα στα οποία ίσως να αντιστοιχούν (hosts names) θα ήταν σίγουρα πιο απλή η κατανόησή της κίνησης που αντιπροσωπεύουν.



Σχήμα 5.17: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 11 Maximal FIs (δεδομένα 21-25 Μαρτίου 2011)



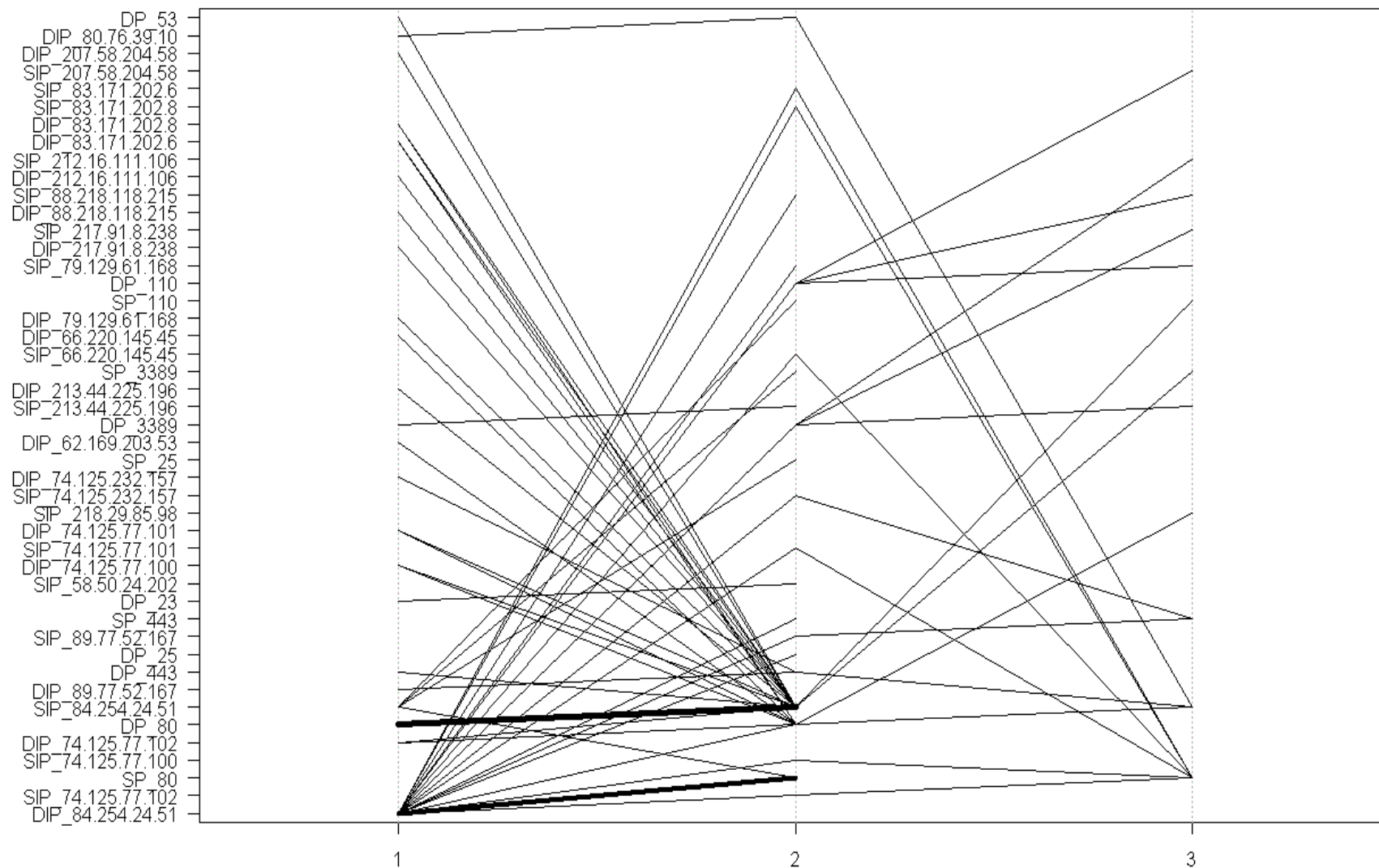
Σχήμα 5.18: Γράφημα παράλληλων συντεταγμένων με χρωματισμό για την κατηγοριοποίηση είδους κίνησης (MFIs)

Αντίστοιχες είναι οι παρατηρήσεις που μπορεί να γίνουν στα γραφήματα τα οποία αναπαριστούν “closed frequent itemsets”. Παρακάτω εμφανίζονται γραφήματα τα οποία αναπαριστούν 79, 44, 26 και 16 “closed frequent itemsets” σχήματα 5.19, 5.20, 5.21 και 5.22 αντίστοιχα.

Χρησιμοποιώντας το σχήμα 5.22 το οποίο περιέχει το μικρότερο αριθμό αναπαριστώμενων “closed frequent itemsets” και από το οποίο μπορούμε να διακρίνουμε ποιο εύκολα τα στοιχεία που τα συνθέτουν παρατηρούμε ότι υπάρχουν τριών ειδών κατηγορίες κίνησης που αναπαριστώνται:

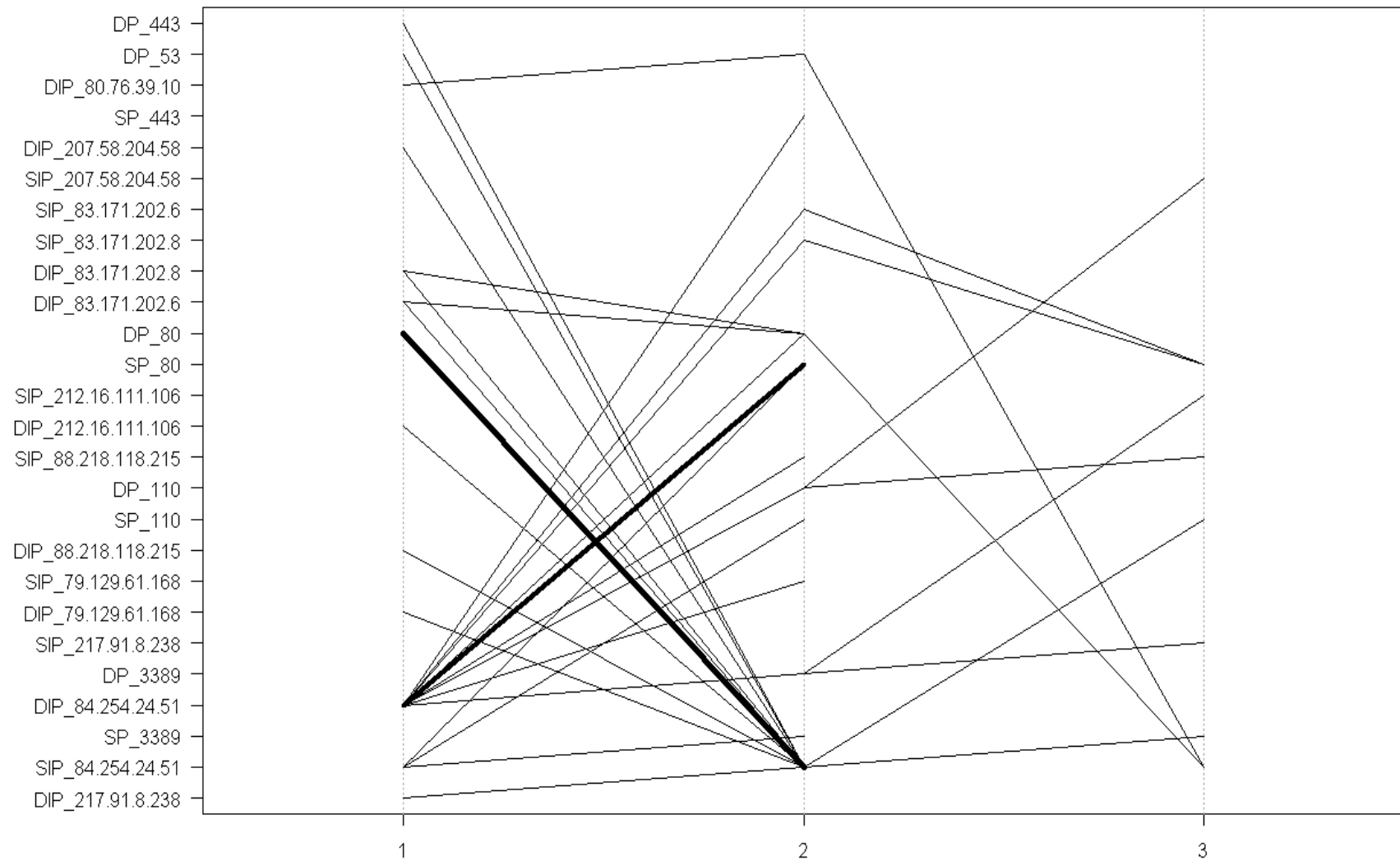
1. κίνηση από και προς Web sites (θύρες 80 και 443),
2. κίνηση για name resolution (προς DNS servers) (θύρες 53) (η οποία βέβαια θα μπορούσε να θεωρηθεί τμήμα της διαδικασίας πρόσβασης σε web sites οπότε και ανήκει στην πρώτη κατηγορία)
3. κίνηση ηλεκτρονικής αλληλογραφίας (θύρα 110)

Parallel coordinates plot for 79 itemsets



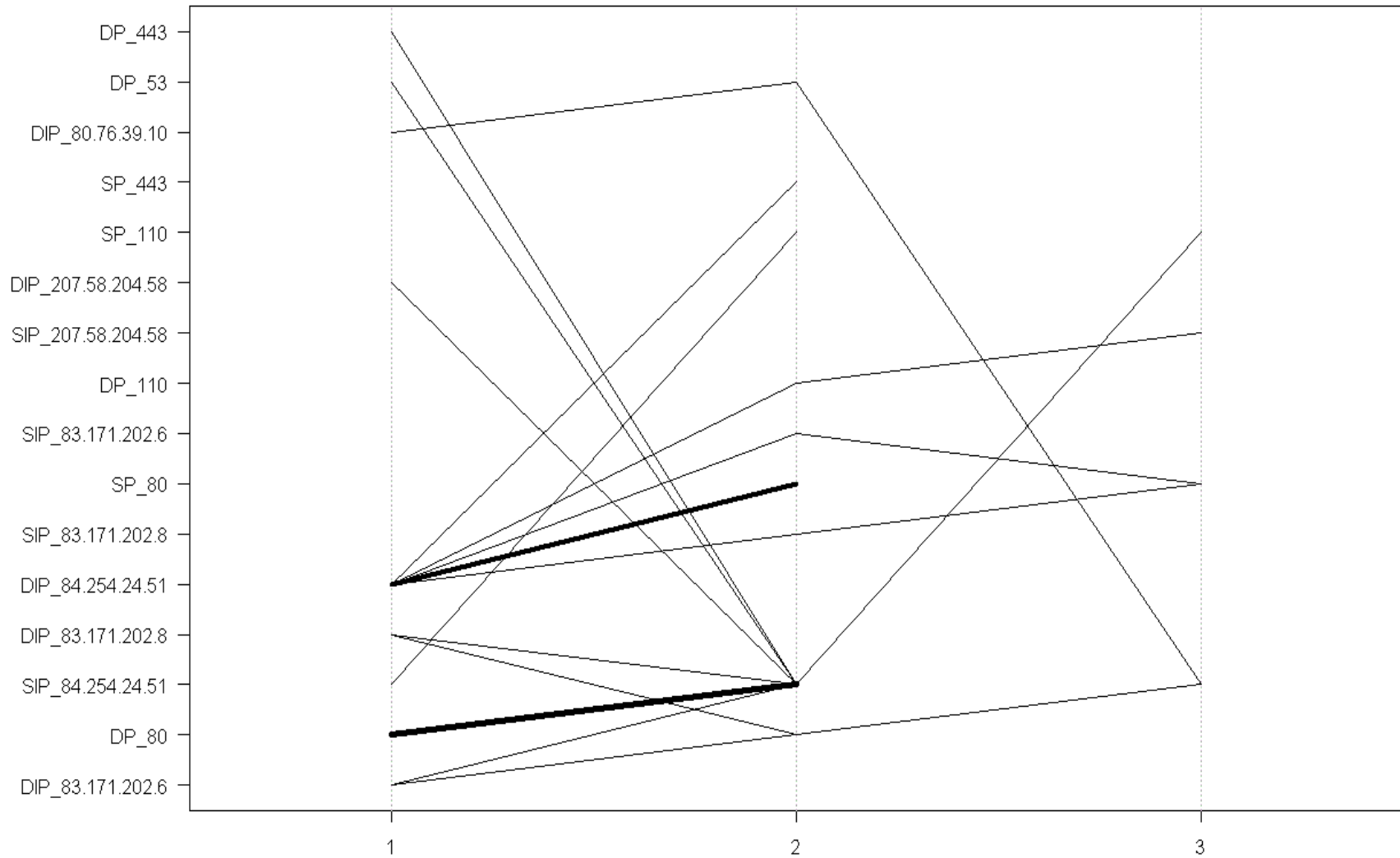
Σχήμα 5.19: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 79 Closed FIs (δεδομένα 21-25 Μαρτίου 2011)

Parallel coordinates plot for 44 itemsets



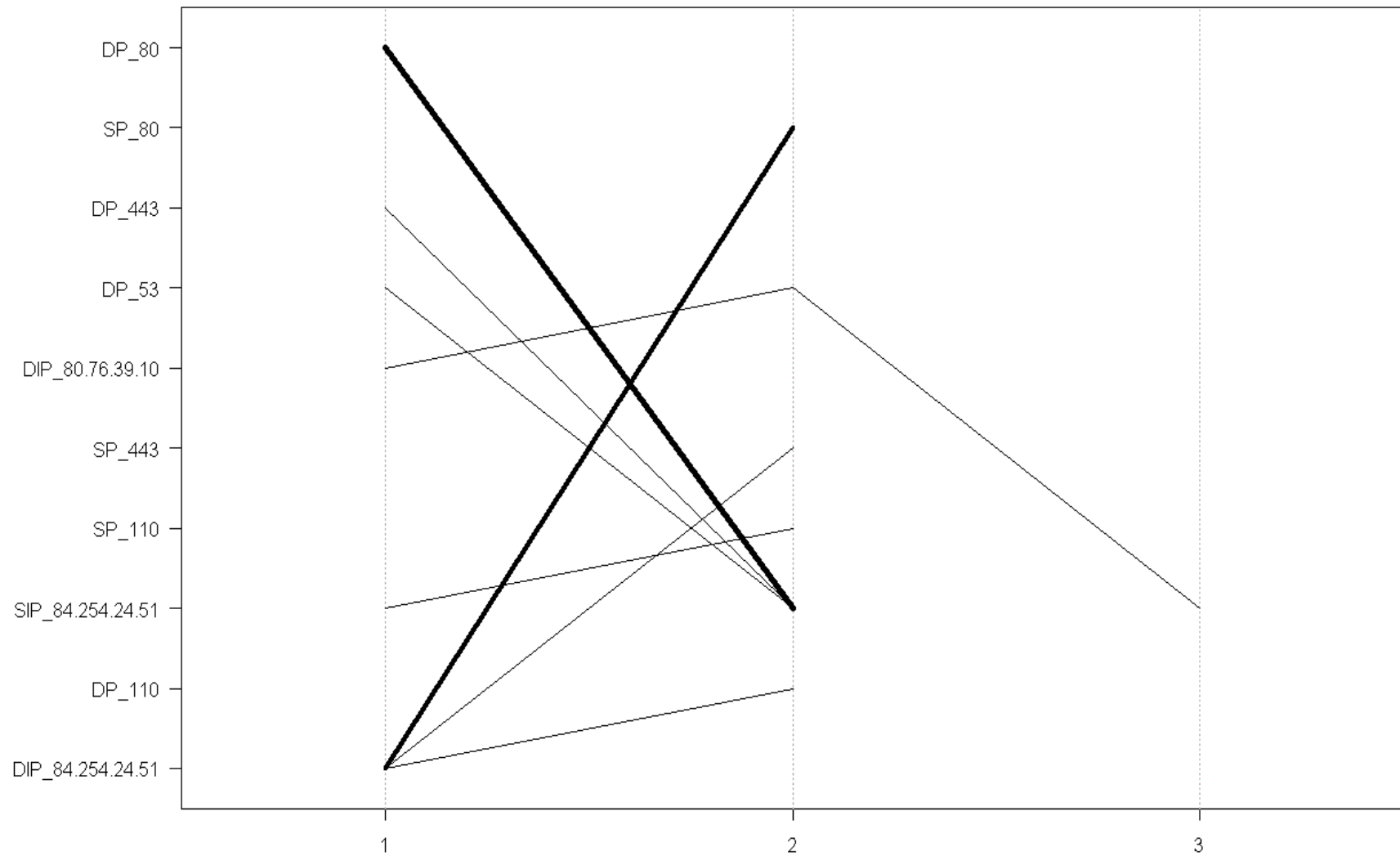
Σχήμα 5.20: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 44 Closed FIs (δεδομένα 21-25 Μαρτίου 2011)

Parallel coordinates plot for 26 itemsets

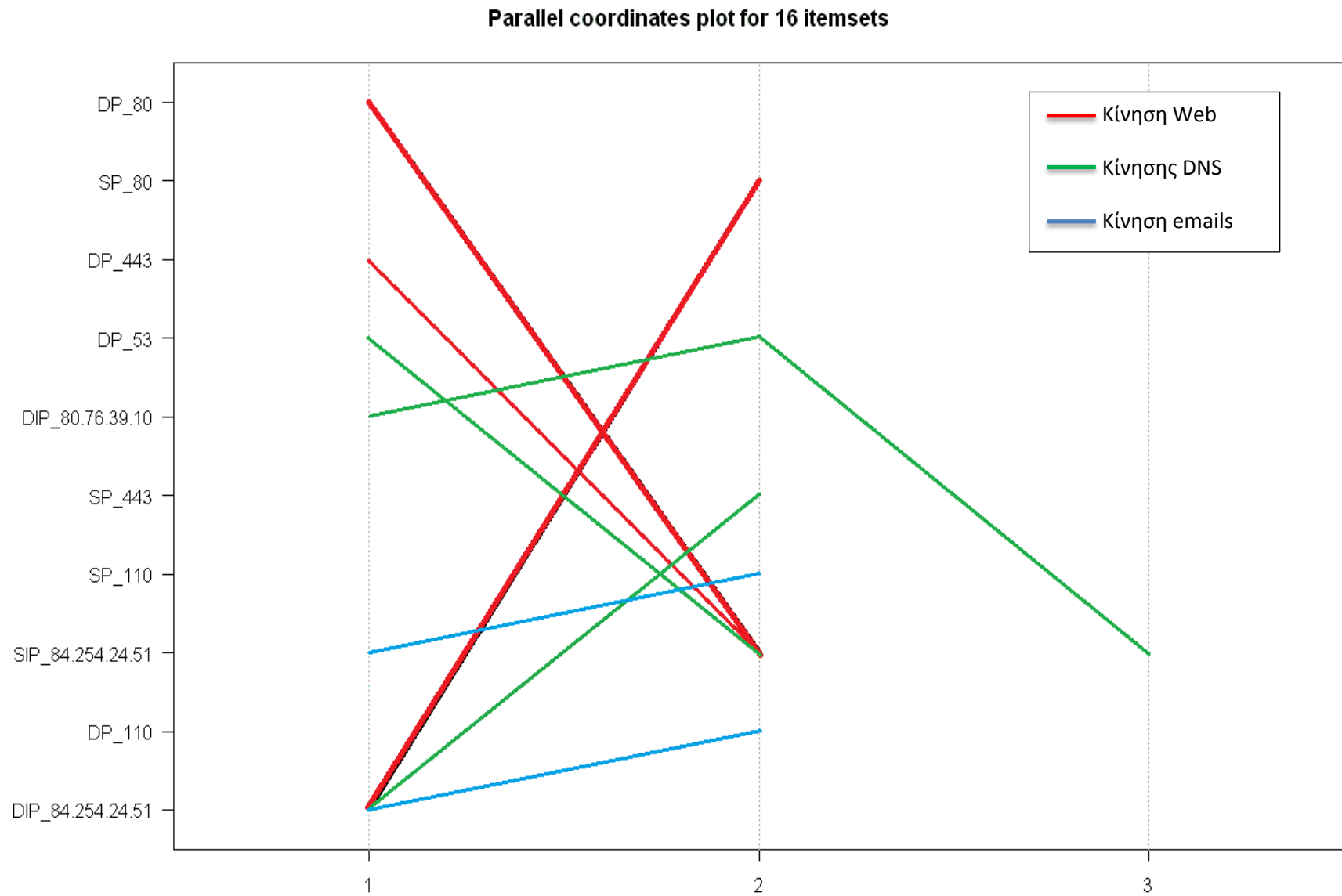


Σχήμα 5.21: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 26 Closed FIs (δεδομένα 21-25 Μαρτίου 2011)

Parallel coordinates plot for 16 itemsets



Σχήμα 5.22: Γράφημα παράλληλων συντεταγμένων για την αναπαράσταση 16 Closed FIs (δεδομένα 21-25 Μαρτίου 2011)



Σχήμα 5.23: Γράφημα παράλληλων συντεταγμένων με χρωματισμό για την κατηγοριοποίηση είδους κίνησης (CFIs)

5.4 Συμπεράσματα που Αφορούν τα Γραφήματα Αναπαράστασης Παράλληλων Συντεταγμένων

Συγκεντρώνοντας όλες τις παρατηρήσεις που έγιναν και αφορούν τις αναπαραστάσεις “frequent itemsets”, “maximal frequent itemsets” και “closed frequent itemsets” με τη χρήση γραφημάτων παράλληλων συντεταγμένων μπορούμε να καταλήξουμε στα ακόλουθα συμπεράσματα.

Τα γραφήματα παράλληλων συντεταγμένων δεν είναι ο καταλληλότερος τρόπος αναπαράστασης «συχνών στοιχείων» τα οποία προέρχονται από καταγραφείσα κίνηση δικτύου υπολογιστών όταν ο αριθμός των παραγόμενων «συχνών στοιχείων» είναι σχετικά μεγάλος. Με τον τρέχον τρόπο δημιουργίας των γραφημάτων αυτών δεν μπορούν να χρησιμοποιηθούν για το γρήγορο και εύκολο προσδιορισμό-εντοπισμό των «συχνών στοιχείων». Ένας αριθμός μεταξύ δέκα και είκοσι «συχνών στοιχείων» θα μπορούσε να δημιουργήσει ένα γράφημα, το οποίο να παρέχει αξιοποιήσιμες στον παρατηρητή πληροφορίες που αφορούν την κίνηση του δικτύου που καταγράφεται (σχήματα 5.11, 5.17 και 5.22).

Μία συνολική εικόνα της σημαντικότερης κίνησης που διέρχεται από το δίκτυο (και το χαρακτηρίζει) θα μπορούσε να δημιουργηθεί χρησιμοποιώντας την αναπαράσταση «συχνών στοιχείων» με τη χρήση γραφημάτων παράλληλων συντεταγμένων αρκεί και πάλι ο αριθμός των αναπαριστώμενων «συχνών στοιχείων» να είναι μικρός (κάτω από είκοσι). Συγκρίνοντας τα σχήματα 5.12, 5.18 και 5.23 μπορούμε να δούμε ότι στην περίπτωση αναπαράστασης “maximal frequent itemsets”, αν και ο αριθμός των αναπαριστώμενων στοιχείων είναι μικρότερος (στο σχήμα 5.18 τα «συχνά στοιχεία» είναι 11) αναπαριστώνται περισσότερα είδη-κατηγορίες κίνησης (5 είδη κίνησης). Ακολουθεί ο παρακάτω πίνακας για την καλύτερη κατανόηση αυτού:

Σχήμα	Αριθμός αναπαριστώμενων «συχνών στοιχείων»	Αριθμός κατηγοριών κίνησης που αποτυπώνονται διαμέσου του γραφήματος
5.12	20 F.I.	3
5.18	11 M. F. I.	5
5.23	16 C. F. I.	3

Πίνακας 5.1: Αριθμός κατηγοριών κίνησης ανά γράφημα

Γενικότερα η γραφική αναπαράσταση των “maximal frequent itemsets” με τη χρήση γραφημάτων παράλληλων συντεταγμένων είναι πιο αποδοτική σε σχέση με την αναπαράσταση των άλλων ειδών «συχνών στοιχείων». Η πληροφορία που αναπαριστάται είναι πιο περιεκτική με αποτέλεσμα τα γραφήματα να είναι πιο αποδοτικά. (Εκτός από τα σχεδιαγράμματα τα οποία φαίνονται στο κείμενο της παρούσας μεταπτυχιακής διατριβής παρατηρήσεις έγιναν σε μεγάλο όγκο δεδομένων διαφορετικών χρονικών στιγμών και διαστημάτων).

Κάποιες διορθώσεις (αντιμετώπιση προβλημάτων) στην τεχνική αναπαράστασης (χρήση παράλληλων συντεταγμένων) θα μπορούσε να βοηθήσουν ώστε η πληροφορία που αναπαριστάται να είναι πιο εύκολα αντλήσιμη. Τέτοιες θα περιγραφούν στο επόμενο κεφάλαιο.

5.4 Αποτελέσματα Γραφικής Αναπαράστασης «Συχνών Στοιχείων» με τη Χρήση της Τεχνικής “Graph Based Visualization”.

Αφού παρουσιάστηκαν γραφήματα παράλληλων συντεταγμένων για την απεικόνιση των παραγόμενων «συχνών στοιχείων» ακολουθεί στην παρούσα παράγραφο παράθεση γραφημάτων «συχνών στοιχείων» χρησιμοποιώντας την τεχνική “graph based visualization”. Τα δεδομένα που χρησιμοποιήθηκαν και παρουσιάζονται στην παρούσα παράγραφο είναι του ίδιου διαστήματος με αυτά της προηγούμενης παραγράφου (καταγραφείσα κίνηση μεταξύ 21 και 25 Μαρτίου 2011). Παρατίθενται τα αντίστοιχα σχεδιαγράμματα με αυτά της προηγούμενης παραγράφου ώστε να είναι δυνατή σύγκριση μεταξύ τους. Φυσικά οι παρατηρήσεις που πραγματοποιήθηκαν στην πράξη αφορούσαν περισσότερα δεδομένα, δεδομένα διαφορετικών χρονικών στιγμών καθώς και διαφορετικών κάθε φορά διαστημάτων, ώστε να ενισχυθεί η εγκυρότητα των συμπερασμάτων.

Χρησιμοποιώντας λοιπόν τα δεδομένα αυτά και εισάγοντας τιμή «στήριξης» την 0.001 στον αλγόριθμο Apriori δημιουργούνται (όπως αναφέρθηκε και σε προηγούμενη παράγραφο) 797 «συχνά στοιχεία». Δημιουργώντας το γράφο για την αναπαράσταση των στοιχείων αυτών προκύπτει το γράφημα του σχήματος 5.24.

Σχήμα 5.24: Graph Based Visualization για την αναπαράσταση 797 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Στο γράφημα αυτό, όπως αναφέρθηκε και στο κεφάλαιο 4, τα «συχνά στοιχεία» αποτελούνται από παραλληλόγραμμα, ακμές (βέλη) και κύκλους. Κάθε κύκλος ενώνεται με ακμές με ένα ή περισσότερα παραλληλόγραμμα δηλώνοντας έτσι ένα «συχνό στοιχείο (κάθε παραλληλόγραμμο αποτελεί και ένα ανεξάρτητο στοιχείο). Στον κάθε κύκλο εμφανίζεται ένας αριθμός ο οποίος είναι ο αύξον αριθμός του «συχνού στοιχείου» με βάση τη σειρά παραγωγής του από τον αλγόριθμο Apriori. Το μέγεθος του κάθε κύκλου προσδιορίζει την τιμή «στήριξης» του «συχνού στοιχείου». Όσον ποιο μεγάλος είναι ο κύκλος τόσο μεγαλύτερη είναι η τιμή «στήριξης» του αντίστοιχου «συχνού στοιχείου».

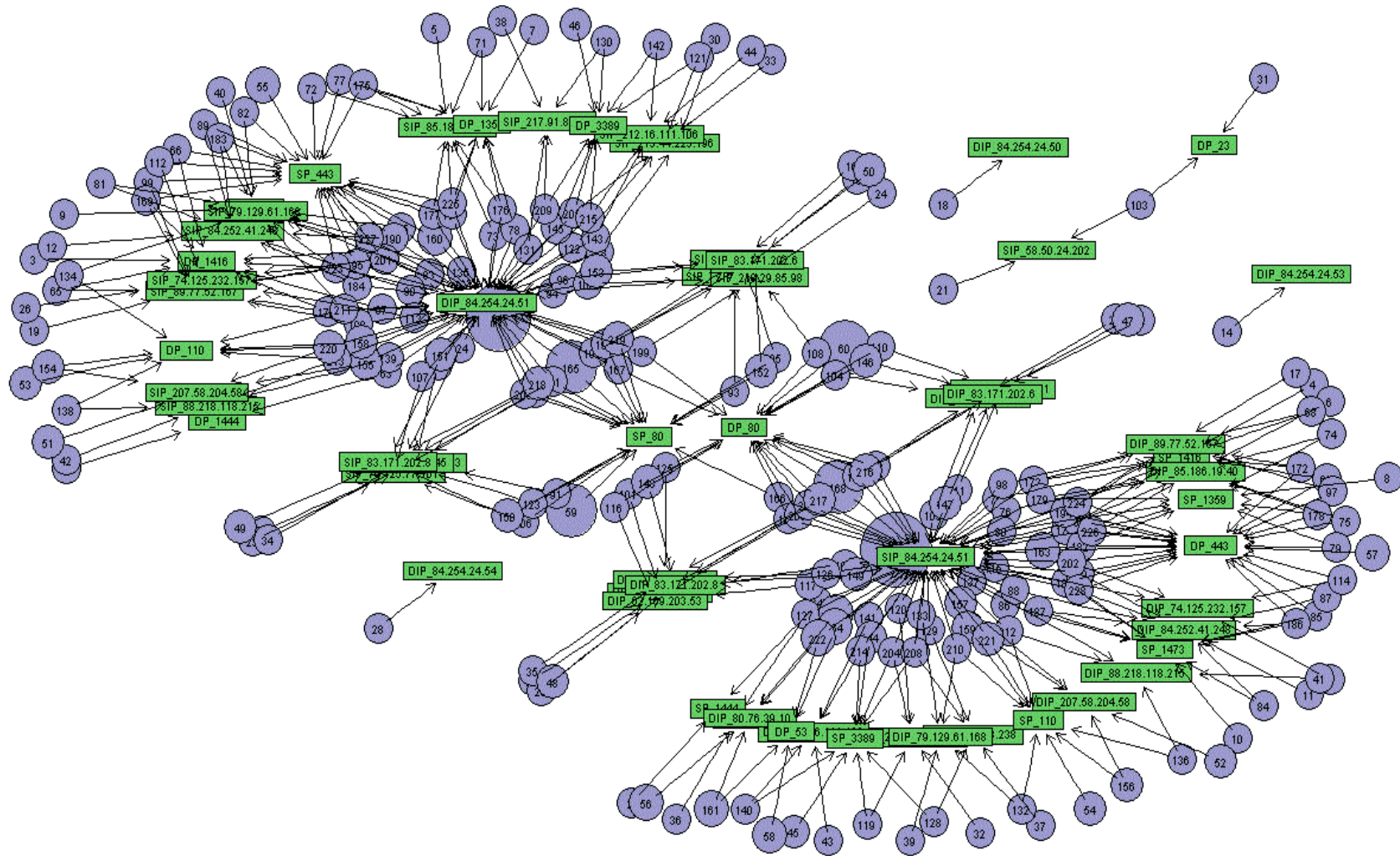
Παρατηρώντας το συγκεκριμένο γράφημα του σχήματος 5.24 μπορεί κανείς να δηλώσει τα εξής:

- Ο αριθμός των υπό αναπαράσταση «συχνών στοιχείων» είναι πάρα πολύ μεγάλος για τον χώρο που χρησιμοποιείται από το γράφημα αυτό. Τα στοιχεία τις περισσότερες φορές επικαλύπτουν το ένα το άλλο οπότε δεν είναι δυνατό να δει κανείς ξεκάθαρα τις πληροφορίες που παρέχουν αυτά και περισσότερο δεν μπορεί να χρησιμοποιήσει το γράφημα για τον εντοπισμό των στοιχείων των «συχνών στοιχείων».
- Δεν γίνεται δε λόγος για την παρατήρηση της τιμής «στήριξης» των «συχνών στοιχείων».
- Το συχνότερο στοιχείο μπορεί να εντοπιστεί από το γράφημα και είναι αυτό στο οποίο η πυκνότητα των ακμών που καταλήγουν είναι μεγάλη. Στο συγκεκριμένο σχήμα φαίνεται ότι τα συχνότερα στοιχεία είναι τα DIP_84.254.24.51 και το SIP_84.254.24.51.
- Με τη χρήση της τεχνικής αυτή είναι δυνατή η αναπαράσταση των «συχνών στοιχείων» τα οποία αποτελούνται από ένα μόνο στοιχείο (σε αντίθεση με τα γραφήματα παράλληλων συντεταγμένων).
- Το σχήμα που προκύπτει είναι συμμετρικό διότι κάθε εισερχόμενη κίνηση έχει την αντίστοιχη εξερχόμενη (τα πακέτα απαντήσεις). Η συμμετρικότητα αυτή μπορεί να χρησιμοποιηθεί για τον εντοπισμό επιθέσεων και θα αναλυθεί σε επόμενο σχήμα όπου ο αριθμός των συχνών στοιχείων θα είναι μικρότερος και θα μπορούν να εξαχθούν ευκολότερα συμπεράσματα.

Στα ίδια συμπεράσματα καταλήγει κανείς παρατηρώντας και τα γραφήματα των σχημάτων 5.25, 5.26, 5.27 και 5.28.

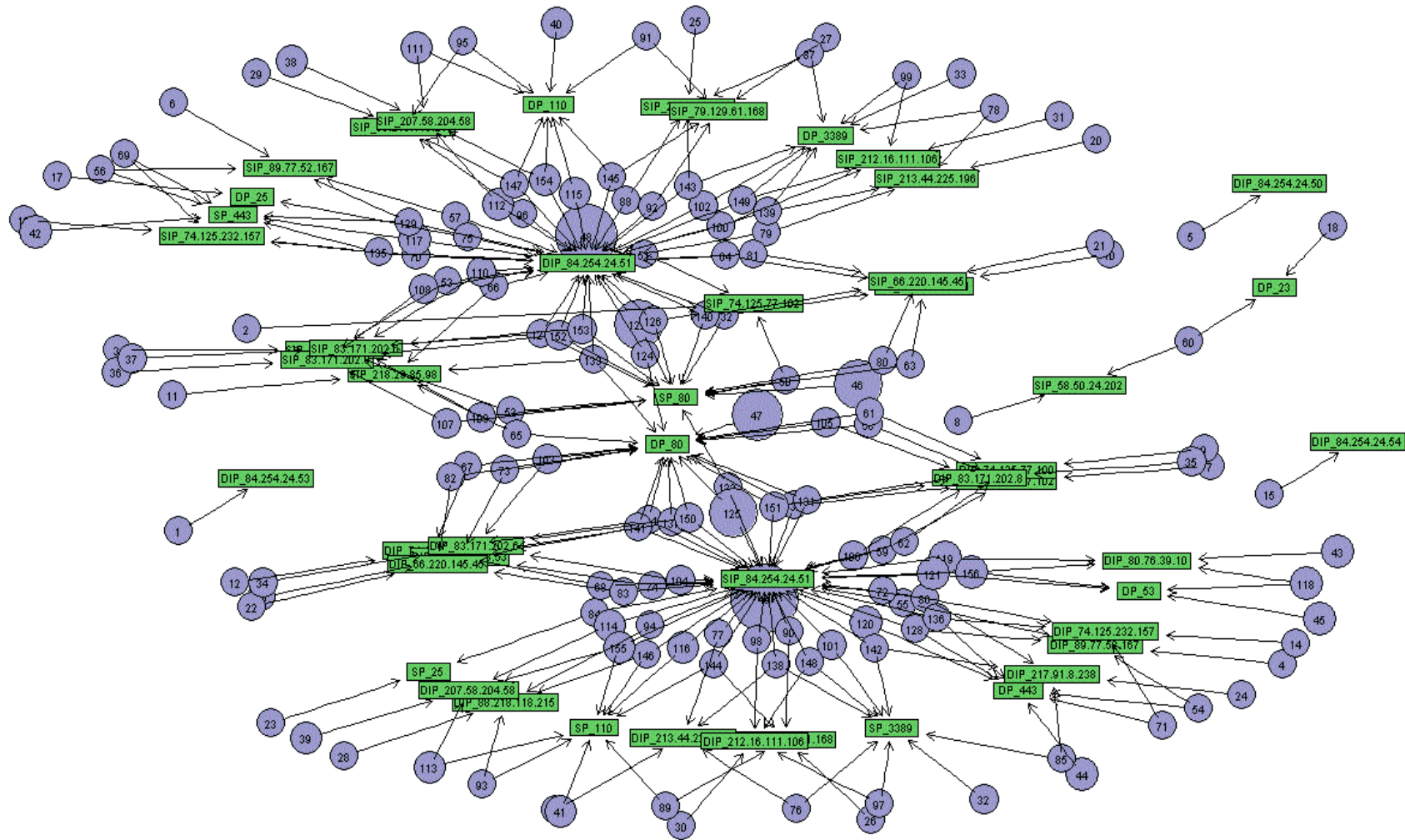
218

Graph for 228 itemsets



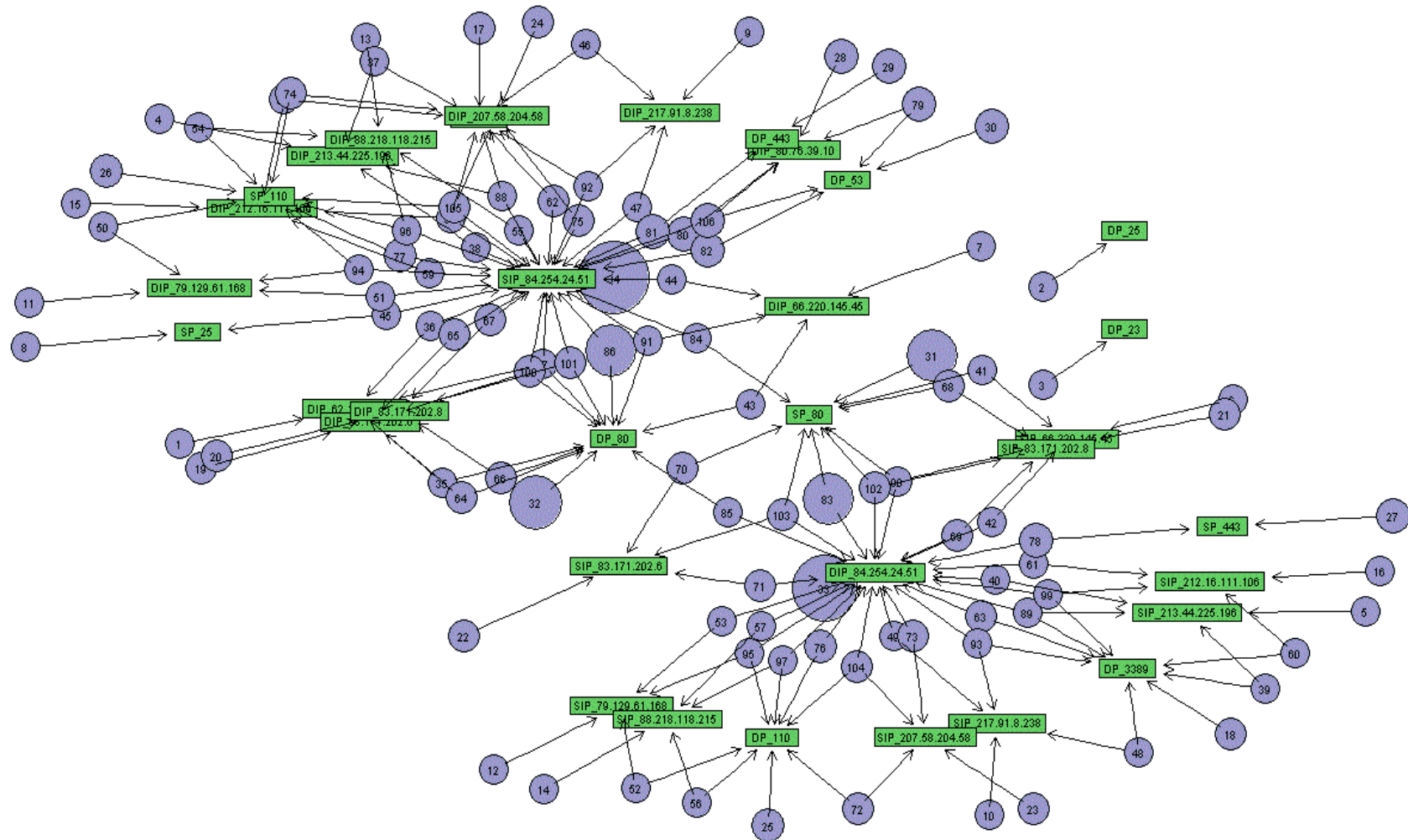
Σχήμα 5.26: Graph Based Visualization για την αναπαράσταση 228 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Graph for 156 itemsets



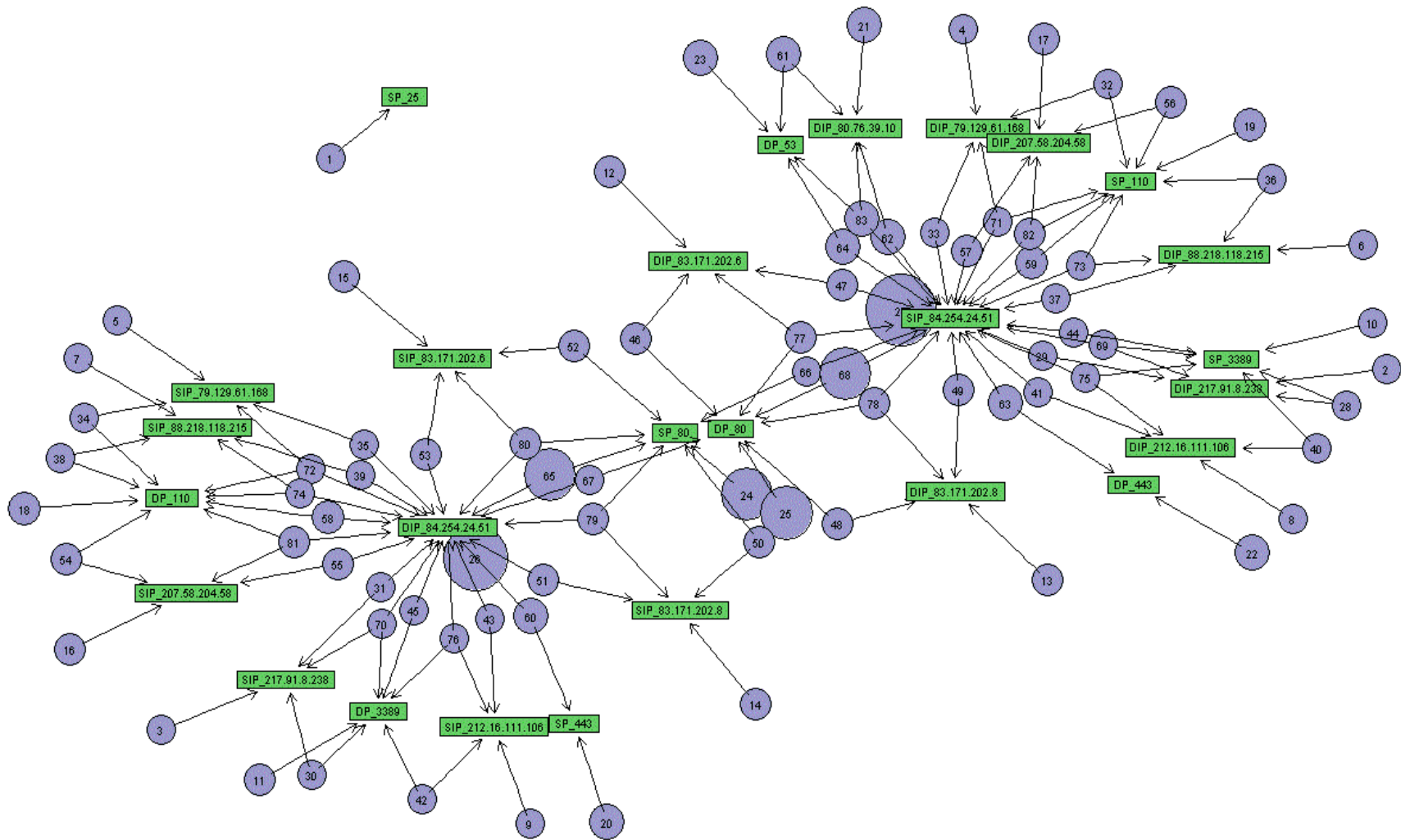
Σχήμα 5.27: Graph Based Visualization για την αναπαράσταση 156 «συχών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Graph for 106 itemsets



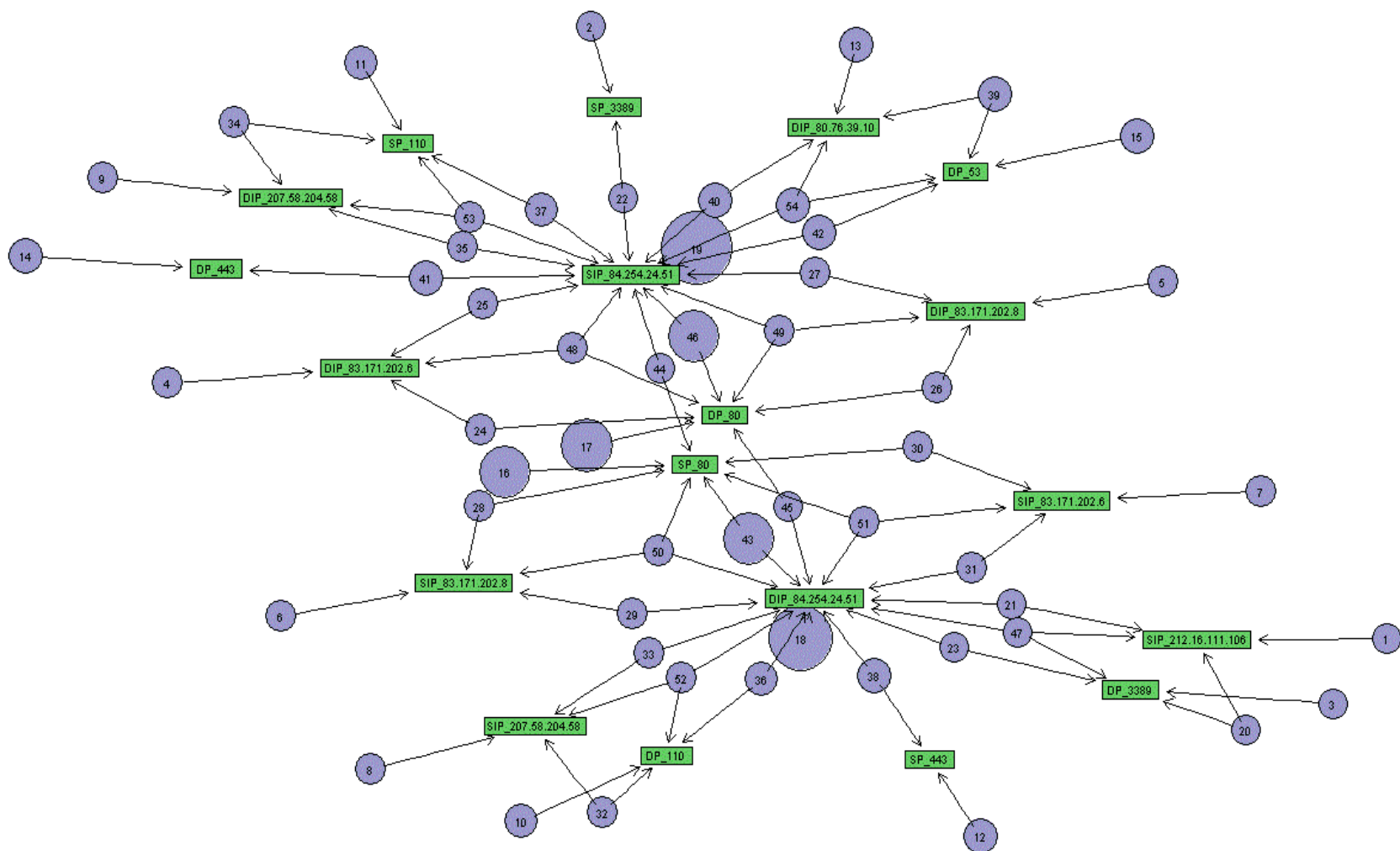
Σχήμα 5.28: Graph Based Visualization για την αναπαράσταση 106 «συχών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Graph for 83 itemsets



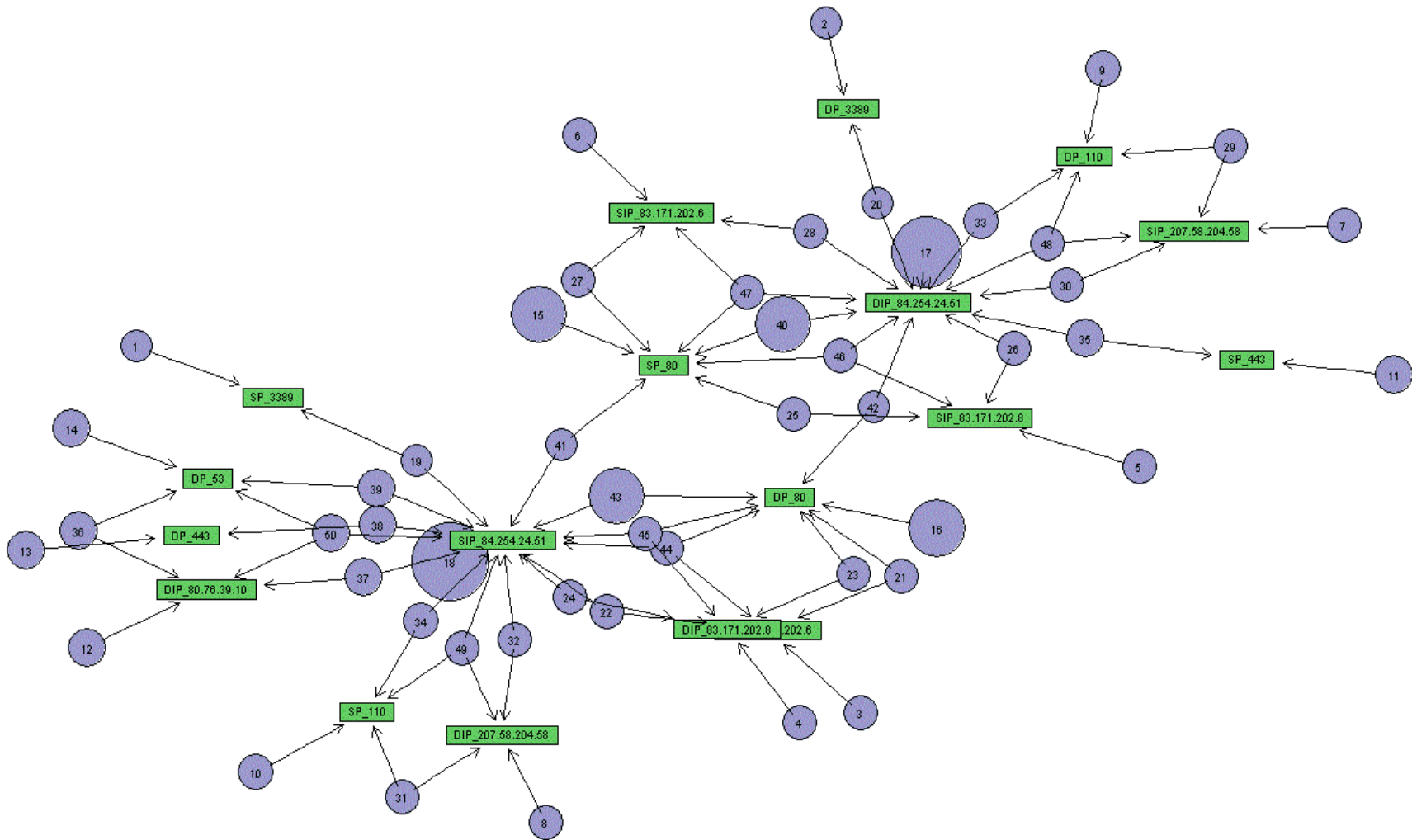
Σχήμα 5.29: Graph Based Visualization για την αναπαράσταση 83 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Graph for 54 itemsets



Σχήμα 5.30: Graph Based Visualization για την αναπαράσταση 54 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Graph for 50 itemsets



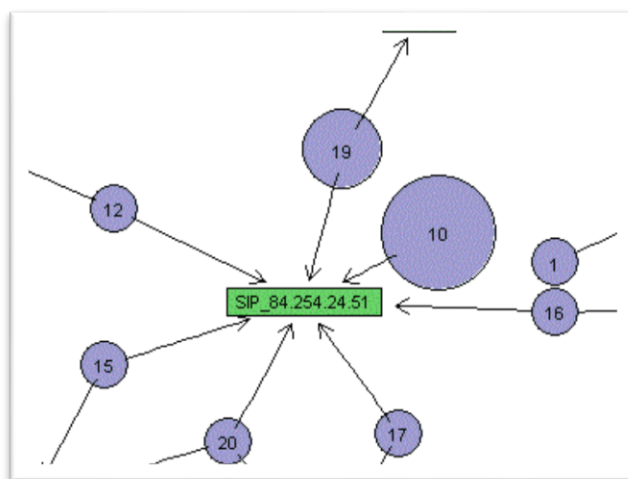
Σχήμα 5.31: Graph Based Visualization για την αναπαράσταση 50 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Σχήμα 5.32: Graph Based Visualization για την αναπαράσταση 20 «συχνών στοιχείων» (δεδομένα 21-25 Μαρτίου 2011)

Στο σχήμα 5.29 εμφανίζεται το “graph based visualization” το οποίο αναπαριστά 83 «συχνά στοιχεία». Εδώ έχει αρχίσει το σχήμα να γίνεται πιο καθαρό με αποτέλεσμα να διακρίνονται (με μία μικρή προσπάθεια) τα στοιχεία που συνθέτουν τα «συχνά στοιχεία». Και σε αυτό το σχήμα φαίνεται η συμμετρικότητα του γραφήματος (όπως αναφέρθηκε και παραπάνω). Από το σχήμα αυτό έχει αρχίσει να φαίνεται καθαρά και η διαφορά στα μεγέθη των κύκλων το οποίο δηλώνει έμμεσα το μέγεθος της τιμής «στήριξης» του κάθε «συχνού στοιχείου». Δεν δηλώνεται όμως ακριβώς η τιμή της τιμής «στήριξης». Τα νούμερα στους κύκλους δεν σχετίζονται με την τιμή «στήριξης» αλλά δείχνουν την θέση που είχε το «συχνό στοιχείο» κατά την παραγωγή αυτών από τον αλγόριθμο Apriori.

Ακολουθούν επιπλέον σχήματα (σχήματα 5.30 και 5.31) από τα οποία προκύπτουν οι ίδιες παρατηρήσεις.

Στο γράφημα του σχήματος 5.32 είναι πλέον ξεκάθαρα τα στοιχεία τα οποία συνθέτουν τα «συχνά στοιχεία». Φαίνεται άμεσα ότι το συχνότερο «συχνό στοιχείο» είναι το {SIP_84.254.24.51} αφού σε αυτό καταλήγει ακμή που ξεκινά από τον μεγαλύτερο κύκλο του γραφήματος (σχήμα 5.33).



Σχήμα 5.33: Συχνότερο «συχνό στοιχείο»

Από το σχήμα αυτό (σχήμα 5.32) θα μπορούσε κανείς να εντοπίσει τα τρία ήδη κίνησης όπως είχε παρατηρηθεί χρησιμοποιώντας το αντίστοιχο γράφημα παράλληλων συντεταγμένων:

- κίνηση από και προς Web sites (θύρες 80 και 443),
- κίνηση ηλεκτρονικής αλληλογραφίας (θύρα 110)

- κίνηση για name resolution (προς DNS servers) (θύρες 53).

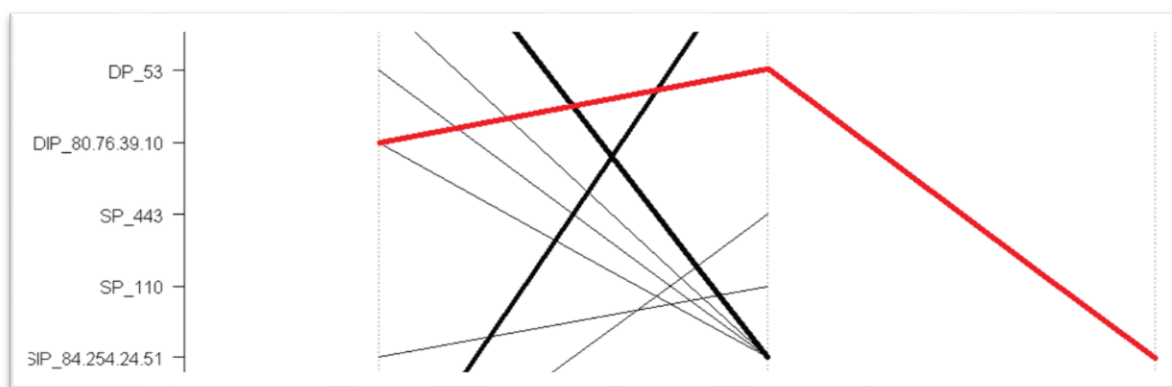
Βέβαια η παρατήρηση και εντοπισμός αυτών των κατηγοριών κίνησης δεν είναι τόσο εύκολος συγκρινόμενος με την τεχνική γραφημάτων παράλληλων συντεταγμένων. Αυτό οφείλεται στο γεγονός ότι στα γραφήματα με τη χρήση παράλληλων συντεταγμένων μπορεί κανείς να ξεχωρίσει τις ακμές που καταλήγουν ή που ξεκινούν από στοιχεία που αφορούν port numbers πιο εύκολα από ότι στα σχεδιαγράμματα με γράφους. Στην τεχνική με γράφους αυτό θα μπορούσε να γίνει ευκολότερα αν τα «συχνά στοιχεία» που αφορούν θύρες (source και destination port numbers) χρωματιζόταν ίσως με διαφορετικό χρώμα από ότι τα υπόλοιπα.

Στα γραφήματα τα οποία δημιουργούνται με τη χρήση της τεχνικής γράφων, σε αντίθεση με αυτά τα οποία δημιουργούνται με τη χρήση παράλληλων συντεταγμένων, απεικονίζονται όλα τα «συχνά στοιχεία» ακόμη και αυτά τα οποία περιέχουν ένα μόνο στοιχείο. Αντίθετα λοιπόν με το αντίστοιχο γράφημα των παράλληλων συντεταγμένων (σχήμα 5.20), στο γράφημα του σχήματος 5.32 εμφανίζονται ξεκάθαρα και τα «συχνά στοιχεία» τα οποία αποτελούνται από ένα μόνο στοιχείο. Παρατηρώντας το συγκεκριμένο γράφημα τα «συχνά στοιχεία» με ένα μόνο στοιχείο είναι:

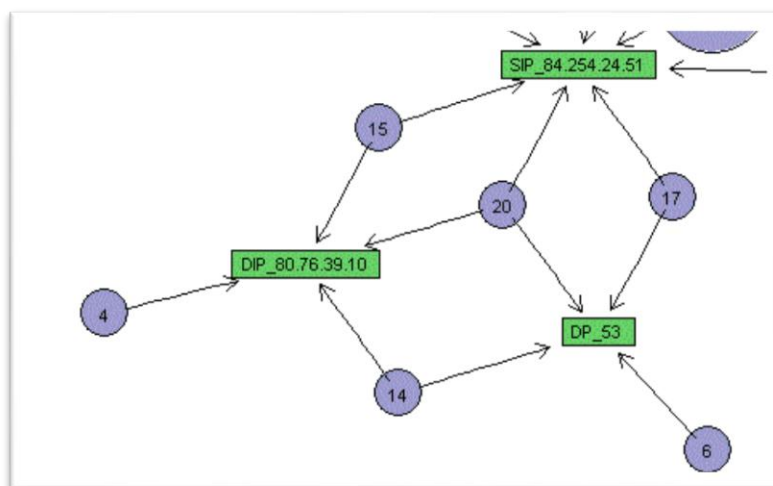
- DP_110
- SP_110
- SP_443
- DIP_80.76.39.10
- DP_443
- DP_53
- SP_80
- DP_80
- DIP_84.254.24.51
- SIP_84.254.24.51

Τα γραφήματα τα οποία δημιουργούνται με τη χρήση “graph based visualization” δεν περιέχουν «κρυμμένα» «συχνά στοιχεία» όπως παρατηρήθηκε στα γραφήματα με τη χρήση παράλληλων συντεταγμένων. Όπως αναφέρθηκε σε προηγούμενη παράγραφο υπάρχουν περιπτώσεις που ο σχεδιασμός των γραφημάτων με τη χρήση παράλληλων συντεταγμένων δεν δίνει τη δυνατότητα

για τον ξεκάθαρο προσδιορισμό των αναπαριστώμενων «συχνών στοιχείων». Κάτι τέτοιο αναφέρθηκε στο σχήμα 5.8 στο οποίο το ποια είναι τα «συχνά στοιχεία» δεν αναπαριστώνται ξεκάθαρα. Στο σχήμα 5.34 φαίνεται μία μεγέθυνση τμήματος του σχήματος 5.8 στο οποίο φαίνονται με κόκκινο χρώμα οι δύο ακμές που δηλώνουν αυτά τα πέντε «συχνά στοιχεία». Παρατηρώντας λοιπόν αυτό το γράφημα δεν φαίνονται ξεκάθαρα το ποια είναι τα στοιχεία αυτά. Αντίθετα στο σχήμα 5.32 (είναι το αντίστοιχο του σχήματος 5.8) τα πέντε αυτά «συχνά στοιχεία» απεικονίζονται όλα ξεκάθαρα. Στο σχήμα 5.35 φαίνεται μία μεγέθυνση τμήματος του σχήματος 5.32 στο οποίο φαίνονται τα στοιχεία αυτά (τα «συχνά στοιχεία» χαρακτηρίζονται από τους κύκλους με τα νούμερα 4, 14, 15, 17 και 20).



Σχήμα 5.34: Αναπαράσταση 5 συχνών στοιχείων με παράλληλες συντεταγμένες (κόκκινες ακμές)



Σχήμα 5.35: Αναπαράσταση 5 συχνών στοιχείων με τεχνική “graph based visualization”

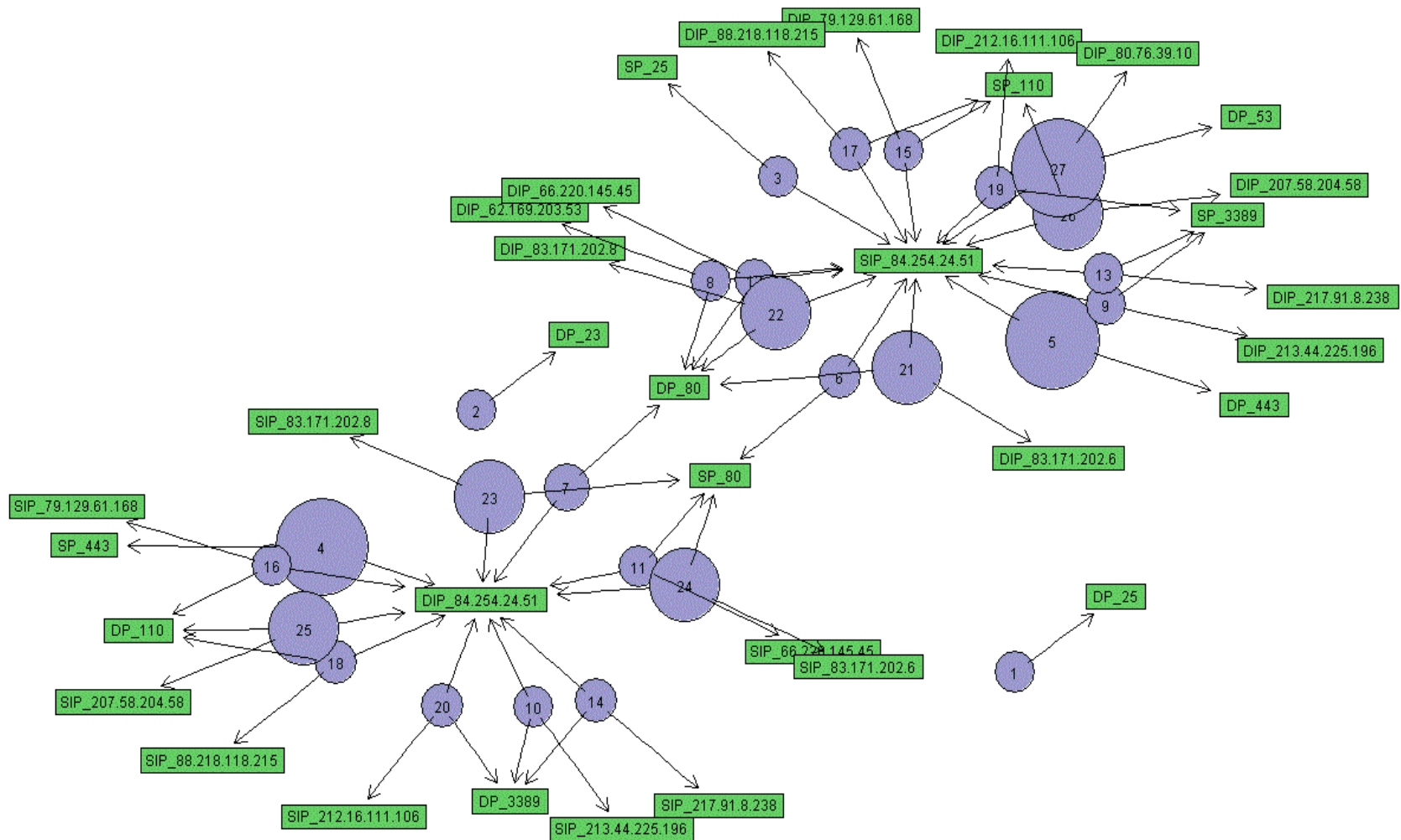
5.5 Αποτελέσματα Γραφικής Αναπαράστασης “Maximal Frequent Itemsets” και “Closed Frequent Itemsets” με τη Χρήση της Τεχνικής “Graph Based Visualization”.

Για την αποφυγή απεικόνισης επαναλαμβανόμενων «συχνών στοιχείων» και την μείωση των κύκλων που εμφανίζονται στους γράφους υπάρχουν τα “maximal frequent itemsets”. Σε αυτή την παράγραφο παρατίθενται λοιπόν γραφήματα τα οποία χρησιμοποιούν την τεχνική “graph based visualization” για την αναπαράσταση “maximal frequent itemsets” που προέρχεται από καταγραφείσα κίνηση μεταξύ 21 και 25 Μαρτίου 2011 στο υπό παρατήρηση δίκτυο υπολογιστών. Δημιουργήθηκαν λοιπόν τα σχήματα 5.36 έως και 5.40 τα οποία αναπαριστούν “maximal frequent itemsets” με τη χρήση γράφων.

Σχήμα 5.36: Graph Based Visualization για την αναπαράσταση 76 “maximal frequent Itemsets” (δεδομένα 21-25 Μαρτίου 2011)

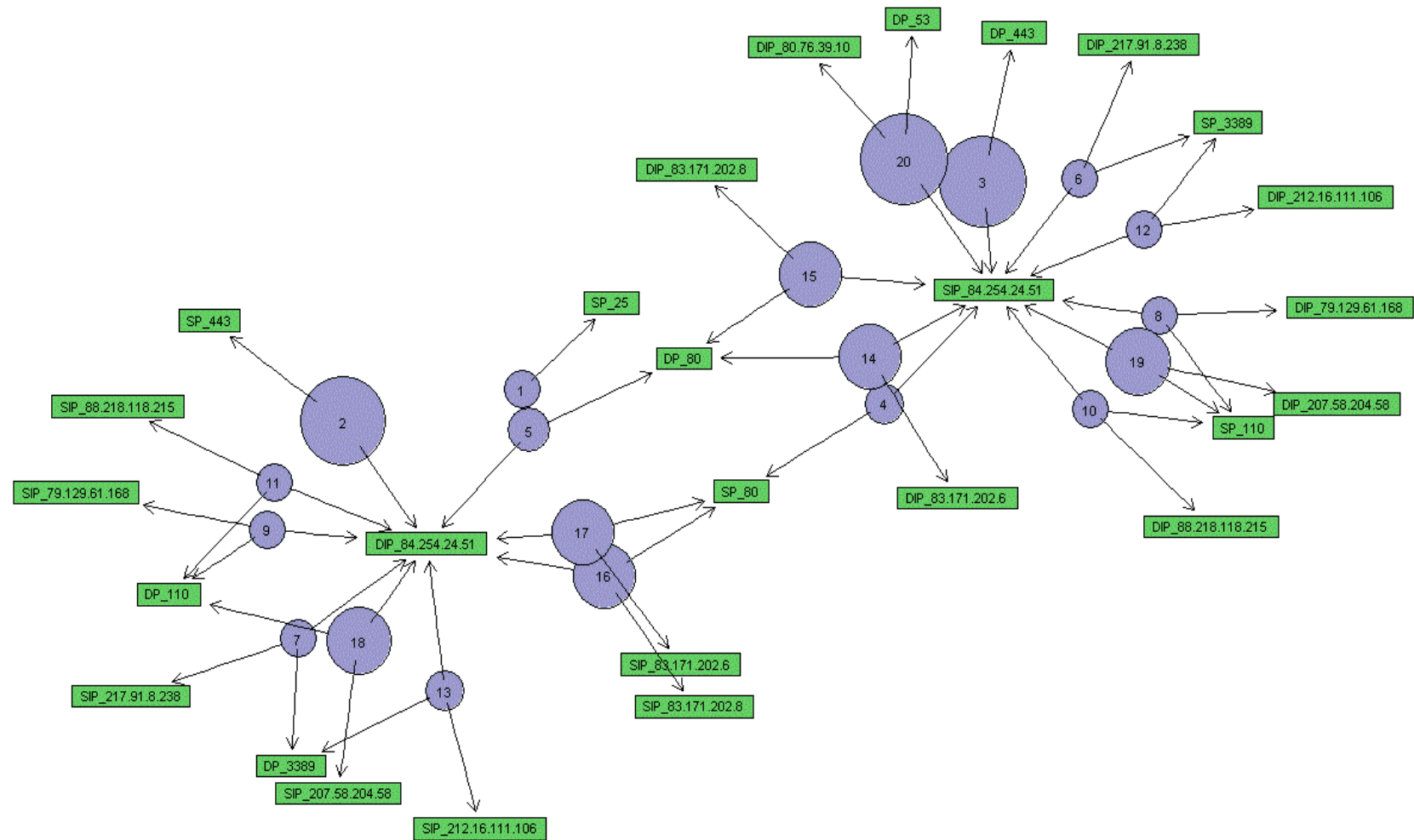
Σχήμα 5.37: Graph Based Visualization για την αναπαράσταση 45 “maximal frequent Itemsets” (δεδομένα 21-25 Μαρτίου 2011)

Graph for 27 itemsets



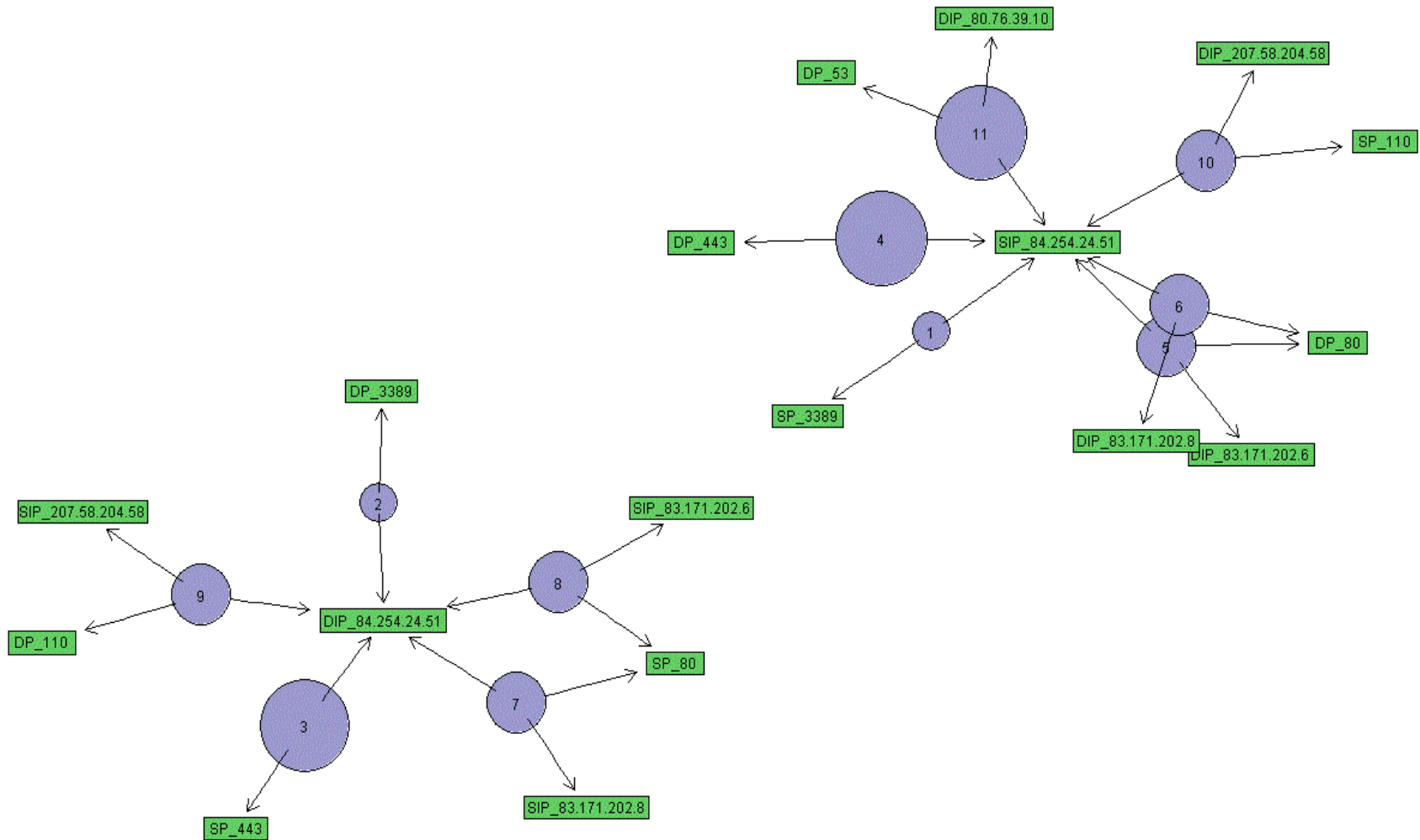
Σχήμα 5.38: Graph Based Visualization για την αναπαράσταση 27 “maximal frequent Itemsets” (δεδομένα 21-25 Μαρτίου 2011)

Graph for 20 itemsets



Σχήμα 5.39: Graph Based Visualization για την αναπαράσταση 20 “maximal frequent Itemsets” (δεδομένα 21-25 Μαρτίου 2011)

Graph for 11 itemsets



Σχήμα 5.40: Graph Based Visualization για την αναπαράσταση 11 “maximal frequent Itemsets” (δεδομένα 21-25 Μαρτίου 2011)

Στο σχήμα 5.36 τα “maximal frequent itemsets” είναι πάρα πολλά για να μπορέσει κανείς εύκολα να εντοπίσει τα στοιχεία που συνθέτουν τα «συχνά στοιχεία». Παρόλα αυτά ο αριθμός των κύκλων («συχνά στοιχεία») έχει μειωθεί σημαντικά σε σχέση με το αντίστοιχο γράφημα που αναπαριστά τα απλά «συχνά στοιχεία».

Αντίστοιχα από επόμενα σχήματα, 5.37 έως και 5.40 που ο αριθμός των «συχνών στοιχείων» έχει μειωθεί είναι πλέον εύκολη υπόθεση ο εντοπισμός των «συχνών στοιχείων». Μάλιστα ο εντοπισμός των «συχνών στοιχείων» γίνεται ακόμη και αν ο αριθμός «στήριξης» είναι ακόμη σχετικά μεγάλος. Σε σχέση λοιπόν με τα γραφήματα παράλληλων συντεταγμένων η αναπαράσταση των «συχνών στοιχείων» γίνεται πιο ξεκάθαρη από ποιο νωρίς (από μικρότερη τιμή «στήριξης»).

Μία πολύ σημαντική παρατήρηση η οποία έγινε όχι μόνο με τα δεδομένα τα οποία παρατίθενται εδώ (21-25 Μαρτίου) αλλά και με άλλα δεδομένα, σχετίζεται με τη δυνατότητα για εντοπισμό επιθέσεων που δίνει η αναπαράσταση των “maximal frequent itemsets” με τη χρήση “graph based visualization”. Παρατηρώντας το σχήμα 5.36 (76 M.F.I.) φαίνεται ένα “maximal frequent itemset” το {DP_23, SIP_58.50.24.202}. Το ίδιο “maximal frequent itemset” παρατηρείται στο σχήμα 5.37 (45 M.F.I.). Αντίστοιχα παρατηρείται από το σχήμα 5.38 (27 M.F.I.) ότι υπάρχει ένα “maximal frequent itemset” με ένα μόνο στοιχείο, το {DP_23}. Μάλιστα στα αντίστοιχα σχήματα δεν υπάρχουν οι συμμετρικές κινήσεις, π.χ. SP_23, DIP_58.50.24.202. Η παρατήρηση αυτή σε συνδυασμό με το γεγονός ότι στο υπό παρατήρηση δίκτυο υπολογιστών η θύρα TCP/UDP 23 είναι μπλοκαρισμένη για τα εισερχόμενα αιτήματα κίνησε το ενδιαφέρον για περισσότερη διερεύνηση. Γενικότερα το γεγονός ότι εμφανίζεται μεταξύ των “maximal frequent itemsets” ένα “maximal frequent itemset” με ένα μόνο στοιχείο δηλώνει ότι το συγκεκριμένο στοιχείο εμφανίζεται σε πολλά (με βάση την τιμή στήριξης) transactions στα οποία όμως τα υπόλοιπα στοιχεία μεταξύ των transactions δεν είναι κοινά. Υπάρχει δηλαδή μία επανάληψη του στοιχείου αυτού ενώ ταυτόχρονα υπάρχει μεγάλη «ποικιλία» (διαφοροποίηση) στις τιμές των υπόλοιπων στοιχείων (SIP, SP, DIP) στα transactions στα οποία εμφανίζεται αυτό. Αυτό σε συνδυασμό με το γεγονός ότι το συγκεκριμένο στοιχείο δεν είναι από τα επιτρεπόμενα στοιχεία πάνω στο firewall της εταιρείας αυξάνει την πιθανότητα το γεγονός αυτό να υποδηλώνει πιθανή επίθεση. Κάνοντας μία αναλυτικότερη έρευνα στα δεδομένα παρατηρήθηκε ότι το {DP_23} εμφανίζεται σε 2361 transactions. Από αυτά τα:

- 767 έχουν ως Source IP address την 58.50.24.202 (από το οποίο προκύπτει και το “maximal frequent itemset” {DP_23, SIP_58.50.24.202} των σχημάτων 5.36 και 5.37),
- 4 έχουν ως Source IP την 80.224.230.115,
- 4 έχουν ως Source IP την 122.57.9.97
- 4 έχουν ως Source IP την 186.0.110.220
- 4 έχουν ως Source IP την 188.34.68.101 κ.λ.π.

Πραγματοποιώντας διαδικασία «who is» για τον εντοπισμό της τοποθεσίας αυτών των IP διευθύνσεων αφετηρίας βρέθηκε ότι πρόκειται για τοποθεσίες από τις οποίες δεν υπάρχει περίπτωση να υπήρχαν «νόμιμα» request με θύρα προορισμού τη θύρα 23 του δικτύου της εταιρείας της οποίας το δίκτυο χρησιμοποιήθηκε για την παρακολούθηση. Πχ η IP 58.50.24.202 εντοπίζεται στην Κίνα, η IP 80.224.230.115 στην Ισπανία, η IP 122.57.9.97 στην Νέα Ζηλανδία κλπ. Με βάση λοιπόν τα παραπάνω η κίνηση που εντοπίζεται και αποτυπώνεται στα γραφήματα αποτελεί πιθανότατα Distributed Denial of Service Attack.

Αντίστοιχα στο σχήμα 5.38 υπάρχει ένα ακόμη “maximal frequent itemset” με ένα στοιχείο το {DP_25} το οποίο όμως δεν θα μπορούσε να παρακινήσει για τον έλεγχο επίθεσης διότι είναι στοιχείο το οποίο επιτρέπεται από το firewall της εταιρείας (ο mail server της εταιρείας βρίσκεται στο εσωτερικό δίκτυο αυτής).

Μπορούμε λοιπόν να συμπεράνουμε ότι στην περίπτωση που σε ένα γράφημα που απεικονίζει “maximal frequent itemsets” εντοπιστεί “maximal frequent itemset” με ένα στοιχείο (και μάλιστα απαγορευμένο στο firewall) τότε η αντίστοιχη κίνηση αποτελεί τμήμα κάποιας επίθεσης.

Σε όλα τα γραφήματα που δημιουργούνται με τη χρήση “graph based visualization” δημιουργούνται δύο ομάδες (υπο-γράφοι) που αντιστοιχούν στην εισερχόμενη και στην εξερχόμενη κίνηση του δικτύου. Αυτοί οι γράφοι είναι σχεδόν (και τις περισσότερες φορές όχι απόλυτα) συμμετρικοί. Το γεγονός αυτό οφείλεται στο γεγονός ότι όλες οι ροές έχουν και μία αντίστροφα συμμετρική ροή. Κάθε αίτημα προς εξυπηρέτηση έχει και την αντίστοιχη απάντηση του για αυτό και δημιουργούνται αυτές οι συμμετρίες. Από το σχήμα 5.40 στο οποίο εμφανίζονται 11 συχνά στοιχεία εκτός του γεγονότος ότι είναι πολύ εύκολο να εντοπιστούν τα στοιχεία που συνθέτουν τα «συχνά στοιχεία»

είναι ξεκάθαρη αυτή η συμμετρία, ο διαχωρισμός δηλαδή της εισερχόμενης από την εξερχόμενη κίνηση. Στο σχήμα εμφανίζονται δύο ξεχωριστές περιοχές με δεδομένα, μία περιοχή πάνω δεξιά στο γράφημα και μία κάτω αριστερά στο γράφημα. Η περιοχή με τα “maximal frequent itemsets” πάνω δεξιά αναφέρεται σε εξερχόμενη κίνηση ενώ η περιοχή κάτω αριστερά αναφέρεται σε κίνηση που εισέρχεται στο εσωτερικό δίκτυο. Συγκρίνοντας τις δύο περιοχές παρατηρούμε ότι είναι συμμετρικές εκτός από ένα είδος (κατηγορία) κίνησης η οποία εμφανίζεται μόνο ως εξερχόμενη και όχι ως εισερχόμενη. Αυτή αφορά το «συχνό στοιχείο» {DP_53, DIP_80.76.30.10} το οποίο αφορά κίνηση προς DNS server ενός ISP. Γενικά θα περίμενε κανείς κάθε κίνηση προς τα έξω να έχει και την αντίστοιχη απάντησή της, γεγονός το οποίο δεν επαληθεύεται στην συγκεκριμένη περίπτωση. Αυτό ίσως στην συγκεκριμένη περίπτωση να δηλώνει την απώλεια κάποιων πακέτων ή ροών δεδομένων. Γενικότερα απόκλιση από την απόλυτη συμμετρία οφείλεται στις λεγόμενες one-way flows που δεν έχουν αντίστροφες ροές και συνήθως οφείλονται σε ενέργειες όπως scanning είτε προς τυχαίες θύρες ή και διευθύνσεις, από τις οποίες δεν αναμένεται απάντηση. Η εμφάνιση λοιπόν μη συμμετρικών γραφημάτων μπορεί να οδηγήσει στην ανεύρεση είτε χαμένων ροών είτε στην ανίχνευση επιθέσεων.

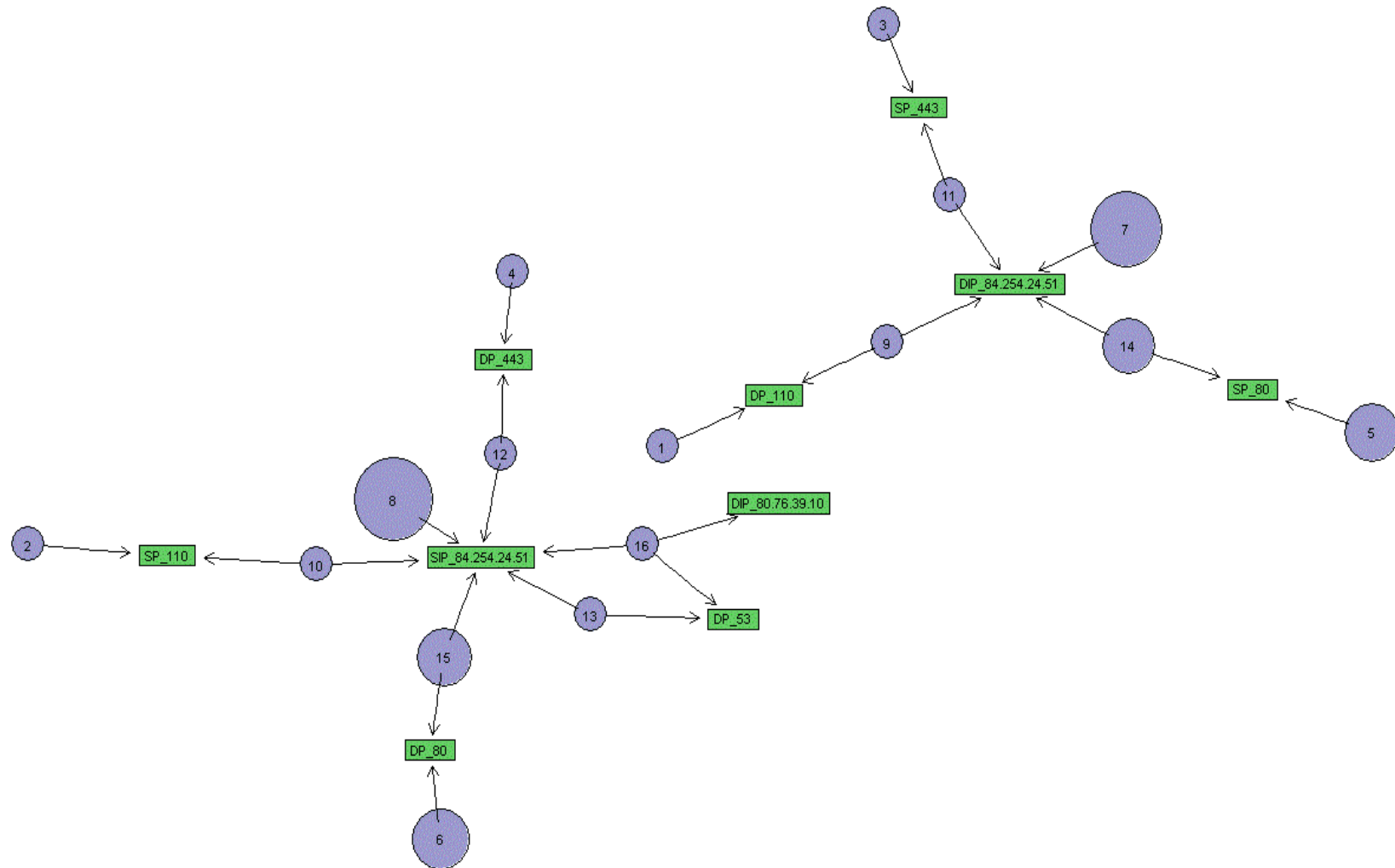
Για τη σύγκριση των εισερχόμενων με τα εξερχόμενα αιτήματα καλό θα ήταν η σχεδίαση του γραφήματος να προσπαθεί να γίνεται με τέτοιο τρόπο ώστε με μία γρήγορη ματιά να μπορεί κανείς να εντοπίσει διαφοροποιήσεις σε σχέση με τη συμμετρικότητα. Καλό θα ήταν δηλαδή εφόσον υπάρχουν δύο «συχνά στοιχεία» τα οποία είναι «συμμετρικά» αυτά να εμφανίζονται ακριβώς στις αντίστοιχες «συμμετρικές» θέσεις στο γράφημα γεγονός το οποίο δεν συμβαίνει απόλυτα στα γραφήματα που εμφανίζονται σε όλα τα γραφήματα που παρατηρήθηκαν.

Στο επόμενο σχήμα (σχήμα 5.41) εμφανίζεται γράφημα το οποίο αναπαριστά “closed frequent itemsets” με τη χρήση “graph based visualization”.

Αυτό το οποίο μπορεί να παρατηρήσει κανείς από το συγκεκριμένο γράφημα είναι η ύπαρξη πολλών “closed frequent itemsets” τα οποία αποτελούνται από ένα μόνο στοιχείο. Συγκεκριμένα στο γράφημα φαίνονται οκτώ τέτοια “closed frequent itemsets”. Η αναπαράσταση των “closed frequent itemsets” μας δίνει και αυτή σημαντική πληροφορία για τα flows του δικτύου και το συνδυασμό των στοιχείων μεταξύ τους. Με βάση αυτά μπορεί ο παρατηρητής να δει και να εντοπίσει σχέσεις μεταξύ των στοιχείων. Αν στα “closed frequent itemsets” π.χ. υπάρχουν τα {DP_80} και {DP_80,

SIP_84.254.24.51} αυτό σημαίνει ότι το DP_80 εμφανίζεται αρκετές φορές σε transactions (flows στην δικιά μας περίπτωση) τα οποία δεν περιέχουν το SIP_84.254.24.51. Αυτό προκύπτει από το γεγονός ότι "closed frequent itemsets" είναι ταυτόχρονα και το {DP_80} αλλά και το {DP_80, SIP_84.254.24.51}.

Graph for 16 itemsets



Σχήμα 5.41: Graph Based Visualization για την αναπαράσταση 16 closed frequent Itemsets” (δεδομένα 21-25 Μαρτίου 2011)

5.6 Συμπεράσματα που Αφορούν τους γράφους.

Με βάση τις παρατηρήσεις των γράφων μπορεί κανείς να εντοπίσει τη χρησιμότητα των γραφημάτων αυτών στην παρακολούθηση της κίνησης ενός δικτύου υπολογιστών. Χρησιμοποιώντας τα γραφήματα αυτά μπορεί να δημιουργηθεί μία γενική εικόνα της κίνησης που διέρχεται από ένα δίκτυο υπολογιστών και να εντοπιστούν ακόμη και επιθέσεις που πραγματοποιήθηκαν προς το υπό παρακολούθηση δίκτυο.

Συγκεντρώνοντας λοιπόν τα συμπεράσματα μετά την παρατήρηση των γράφων διαπιστώνει κανείς ότι για την καλύτερη αναπαράσταση και εντοπισμό των «συχνών στοιχείων» (ανεξαρτήτου τύπου, είτε αναπαριστούνται «συχνά στοιχεία» είτε “maximal frequent itemsets” είτε “closed frequent itemsets”) καλό είναι ο αριθμός των στοιχείων αυτών να μη ξεπερνά τα τριάντα. Πέραν του αριθμού αυτού τα στοιχεία επικαλύπτονται το ένα με το άλλο με αποτέλεσμα να γίνεται δυσκολότερος ο εντοπισμός της απαιτούμενης πληροφορίας. Από την άλλη μεριά ένας αριθμός αναπαριστώμενων «συχνών στοιχείων» μικρότερος των δέκα είναι πολύ μικρός και δεν δίνει χρήσιμη πληροφορία. Η μόνη περίπτωση που θα χρειαστεί να αναπαρασταθεί μεγάλος αριθμός «συχνών στοιχείων» (πάνω από εκατό) είναι η περίπτωση που χρειάζεται με μία γρήγορη ματιά να εντοπιστούν τα συχνότερα «συχνά στοιχεία». Αυτά είναι τα στοιχεία στα οποία καταλήγει μεγάλος αριθμός ακμών (βέλη), η πυκνότητα δηλαδή των ακμών που καταλήγουν στα πιο συχνά «συχνά στοιχεία» είναι μεγάλη.

Τα σχεδιαγράμματα με γράφους σε αντίθεση με τα γραφήματα παράλληλων συντεταγμένων δίνουν τη δυνατότητα αναπαράστασης ακόμη και των «συχνών στοιχείων» τα οποία αποτελούνται από ένα μόνο στοιχείο. Το γεγονός αυτό δηλώνει ότι δεν χάνεται πληροφορία κατά τη γραφική αναπαράσταση των «συχνών στοιχείων». Μάλιστα με τη χρήση των γραφημάτων αυτών δεν υπάρχει «κρυμμένη» πληροφορία («κρυμμένα» «συχνά στοιχεία») όπως συμβαίνει με τη χρήση των γραφημάτων παράλληλων συντεταγμένων. Μπορεί κανείς κοιτώντας το γράφημα να είναι σίγουρος ότι δεν χάνει κάποια πληροφορία.

Με τη χρήση των γράφων μπορεί να γίνει εντοπισμός κάποιων τύπων επιθέσεων προς το υπό παρατήρηση δίκτυο. Λόγω του γεγονότος ότι αναπαριστάται κίνηση (ροές δεδομένων) δικτύου όλα τα γραφήματα πρέπει να είναι συμμετρικά. Τα περισσότερα αν όχι όλα τα είδη κινήσεων είναι

διπλής κατεύθυνσης (two way flows). Υπάρχει δηλαδή πάντοτε μία αίτηση (request) και μία απάντηση σε αυτή. Εφόσον δηλαδή υπάρχει μία εξερχόμενη ροή δεδομένων θα υπάρχει και η αντίστοιχη εισερχόμενη (απάντηση στο request που έχει γίνει) με αποτέλεσμα τα προκύπτοντα γραφήματα να είναι συμμετρικά (εισερχόμενη-εξερχόμενη κίνηση). Απόκλιση από τη συμμετρικότητα αυτή μπορεί να δηλώνει είτε απώλεια κάποιων πακέτων-ροών δεδομένων, είτε την ύπαρξη one-way flows προς ή από το υπό παρατήρηση δίκτυο, γεγονός το οποίο μπορεί να δηλώνει κάποια επίθεση, όπως port scan attack είτε denial of service attack.

Μεγάλο ενδιαφέρον παρουσιάζει και η αναπαράσταση των “maximal frequent itemsets”. Αυτά μπορούν να χρησιμοποιηθούν για την αποφυγή εμφάνισης επαναλαμβανόμενων «συχνών στοιχείων», τη μείωση δηλαδή των εμφανιζόμενων κύκλων στο γράφημα, γεγονός το οποίο το κάνει ποιο «ξεκάθαρο» το γράφημα. Το γράφημα γίνεται ποιο εύκολο στην ανάγνωση χωρίς να χάνεται κάποια πληροφορία, ενώ μπορεί να είναι εύκολο στην ανάγνωση ακόμη και στην περίπτωση που ο αριθμός των υπό αναπαράσταση στοιχείων είναι σχετικά μεγάλος. Ταυτόχρονα, τα γραφήματα αναπαράστασης “maximal frequent itemsets” μπορούν χρησιμοποιηθούν ακόμη και για τον εντοπισμό-ανίχνευση κάποιου είδους επιθέσεων που έχουν πραγματοποιηθεί στο υπό παρακολούθηση δίκτυο. Κάτι τέτοιο μπορεί να εντοπιστεί από την εμφάνιση στο γράφημα «συχνών στοιχείων» τα οποία αποτελούνται από ένα μόνο στοιχείο (και πολύ περισσότερο όταν το στοιχείο αυτό είναι στοιχείο το οποίο δεν επιτρέπεται από την πολιτική ασφαλείας του δικτύου). Η εμφάνιση ενός “maximal frequent itemset” το οποίο αποτελείται από ένα μόνο στοιχείο δηλώνει τον συνδυασμό του στοιχείου αυτού με μεγάλη ποικιλία άλλων στοιχείων. Υπάρχουν δηλαδή πολλά transactions (ροές δεδομένων στην περίπτωσή μας) τα οποία περιέχουν το στοιχείο αυτό ενώ παράλληλα τα υπόλοιπα στοιχεία των συγκεκριμένων transactions είναι διαφορετικά μεταξύ αυτών (των transactions). Το γεγονός αυτό μπορεί να δηλώνει απλά τη μεγάλη χρήση του στοιχείου αυτού, αλλά ταυτόχρονα μπορεί να δηλώνει και distributed denial of service attack (π.χ. εμφάνιση της ίδιας θύρας προορισμού στις ροές ενώ η διεύθυνση αφετηρίας είναι διαφορετική σε κάθε ένα από τα transactions στα οποία εμφανίζεται η θύρα αυτή).

Και ο τρόπος σχεδιασμού των γράφων επιδέχεται τροποποιήσεις ώστε να βελτιωθεί και να γίνει ποιο χρηστική η αναπαράσταση των «συχνών στοιχείων». Τέτοιες προτάσεις θα παρατεθούν στο επόμενο κεφάλαιο.

5.7 Συμπεράσματα

Οι παρατηρήσεις των γραφημάτων (τόσο των γράφων όσο και γραφημάτων παράλληλων συντεταγμένων) έδειξαν ότι η γραφική αναπαράσταση «συχνών στοιχείων» μπορεί να βοηθήσει χρησιμοποιούμενη παράλληλα με τα υπόλοιπα εργαλεία ενός διαχειριστή δικτύου, στη βελτίωση της παρακολούθησης του δικτύου και κατ' επέκταση στη σωστότερη διαχείριση αυτού. Κάθε τύπος γραφημάτων (από αυτά που αναφέρθηκαν στην παρούσα ενότητα) έχει να προσφέρει διαφορετικά στοιχεία στην ολοκληρωμένη παρακολούθηση.

Τα γραφήματα παράλληλων συντεταγμένων μπορεί να χρησιμοποιηθούν ώστε να βοηθήσουν στον εντοπισμό των κατηγοριών κίνησης που διέρχονται από το υπό παρατήρηση δίκτυο υπολογιστών δίνοντας έτσι ευκολότερα τη γενικότερη εικόνα της κίνησης που διέρχεται από αυτό. Από την άλλη όμως παρουσιάζουν πολλά μειονεκτήματα-προβλήματα συγκρινόμενα με την τεχνική δημιουργίας γράφων. Δε δίνουν τη δυνατότητα για αναπαράσταση των «συχνών στοιχείων» που αποτελούνται από ένα μόνο στοιχείο, είναι δύσκολος ο εντοπισμός των στοιχείων στα οποία καταλήγουν οι ακμές του γραφήματος, δε δίνουν σε μερικές περιπτώσεις εύκολα την εικόνα των ακριβή στοιχείων συνθέτουν τα «συχνά στοιχεία» (κρυμμένα στοιχεία) κ.λπ. Γενικότερα χρειάζεται να γίνουν αρκετές παρεμβάσεις-διορθώσεις στον τρόπο με τον οποίο δημιουργούνται ώστε να μπορεί κανείς να πει με βεβαιότητα ότι η παρακολούθηση της κίνησης μέσα από αυτά μπορεί να επωφεληθεί και να βοηθηθεί.

Τα σχεδιαγράμματα γράφων από την άλλη είναι πιο συγκεκριμένα και πιο παραστατικά συγκρινόμενα με τα γραφήματα παράλληλων συντεταγμένων. Στην περίπτωση που ο αριθμός των προς αναπαράσταση «συχνών στοιχείων» είναι λογικός τα περιεχόμενα αυτών εμφανίζονται καθαρά χωρίς να υπάρχουν «κρυμμένες» πληροφορίες. Μπορεί κανείς να εντοπίσει ευκολότερα την πληροφορία που χρειάζεται ενώ χρησιμοποιώντας τη συμμετρικότητα αυτών μπορούν να χρησιμοποιηθούν και στον εντοπισμό κάποιων τύπων επιθέσεων που έχουν πραγματοποιηθεί στο υπό παρατήρηση δίκτυο.

Με βάση λοιπόν τις παρατηρήσεις αυτές γίνεται κατανοητό ότι οι γράφοι υπερτερούν έναντι των γραφημάτων παράλληλων συντεταγμένων στην αναπαράσταση «συχνών στοιχείων» που προέρχονται από δεδομένα κίνησης δικτύου υπολογιστών.

Στο κεφάλαιο που ακολουθεί αφού πραγματοποιηθεί μία γενικότερη σύνοψη των συμπερασμάτων που αφορούν τη βοήθεια που προσφέρουν τα «συχνά στοιχεία» στην παρακολούθηση ενός δικτύου υπολογιστών και αναφερθεί η συνεισφορά σε αυτή της γραφικής αναπαράστασης των «συχνών στοιχείων», θα δοθούν προτάσεις για την βελτίωση των γραφημάτων.

Κεφάλαιο 6

Συμπεράσματα

Η ενότητα αυτή έχει ως στόχο να τονίσει τη συνεισφορά των «συχνών στοιχείων» στην παρακολούθηση ενός δικτύου υπολογιστών. Σε αυτή πραγματοποιείται μία γενικότερη σύνοψη των συμπερασμάτων που προέκυψαν από την παρακολούθηση των γραφημάτων που δημιουργήθηκαν με βάση τις τεχνικές γραφικής αναπαράστασης «συχνών στοιχείων» και αναλύθηκαν σε προηγούμενο κεφάλαιο. Παράλληλα δίνονται προτάσεις για τη βελτίωση αυτών των τεχνικών (παράλληλων συντεταγμένων και “graph based visualization”).

6.1 Εισαγωγή

Η ασφάλεια και γενικότερα η σωστή και αδιάκοπη λειτουργία ενός δικτύου υπολογιστών στηρίζεται σε μεγάλο βαθμό στην αποφυγή απρόσμενων και δυσάρεστων γεγονότων που μπορούν να προκύψουν σε αυτό. Για την επίτευξη αυτού του στόχου πολύ σημαντικό ρόλο παίζει η ολοκληρωμένη διαχείριση του δικτύου στην οποία σημαντικό τμήμα κατέχει η γνώση της κίνησης που διέρχεται από αυτό. Έχουν μέχρι σήμερα αναπτυχθεί διάφορα συστήματα-τεχνικές για τη σωστότερη παρακολούθηση των γεγονότων που πραγματοποιούνται καθώς και την παρακολούθηση της κίνησης που διέρχεται από ένα δίκτυο υπολογιστών. Οι τεχνικές εξόρυξης «συχνών στοιχείων» έρχονται να βοηθήσουν στον τομέα αυτό. Όπως αναφέρθηκε και στα προηγούμενα κεφάλαια υπάρχουν αρκετές μελέτες-έρευνες στη βιβλιογραφία που επιβεβαιώνουν τη συνεισφορά των τεχνικών αυτών στη διαχείριση και εντοπισμό συμβάντων σε δίκτυα υπολογιστών (όπως η [37]). Η γραφική αναπαράσταση των «συχνών στοιχείων» που προκύπτουν από την καταγραφή της κίνησης ενός δικτύου υπολογιστών θα μπορούσε να συμπεριληφθεί και αυτή μεταξύ των τεχνικών για την παρακολούθηση και κατανόηση της κίνησης που διέρχεται από ένα δίκτυο υπολογιστών.

Το παρόν κεφάλαιο συνοψίζει τα συμπεράσματα που προέκυψαν από την παρακολούθηση των γραφικών αναπαραστάσεων των «συχνών στοιχείων» που δημιουργήθηκαν με βάση το σύστημα παρακολούθησης-καταγραφής που αναπτύχθηκε σε δίκτυο υπολογιστών υφιστάμενης εταιρείας. Παράλληλα συμπεριλαμβάνει προτάσεις για τη βελτίωση αυτών καθώς και τρόπους με τους οποίους θα μπορούσαν να αξιοποιηθούν.

6.2 Πλεονεκτήματα και Μειονεκτήματα Παράλληλων Συντεταγμένων

Στην παρούσα παράγραφο συνοψίζονται τα πλεονεκτήματα και μειονεκτήματα των γραφημάτων παράλληλων συντεταγμένων που αναπαριστούν «συχνά στοιχεία» τα οποία προέκυψαν ύστερα από εφαρμογή τεχνικών παραγωγής «συχνών στοιχείων» σε καταγραφείσα κίνηση δικτύου υπολογιστών.

Τα γραφήματα που προκύπτουν όταν ο αριθμός των υπό αναπαράσταση «συχνών στοιχείων» είναι πολύ μεγάλος μπορούν να βοηθήσουν ώστε ο παρατηρητής με μία γρήγορη ματιά να εντοπίσει τα συχνότερα στοιχεία. Αυτά εντοπίζονται γρήγορα λόγω του γεγονότος ότι σε αυτά καταλήγει μεγάλος αριθμός ακμών. Με βάση λοιπόν την πυκνότητα των ακμών που ξεκινούν ή καταλήγουν σε ένα στοιχείο μπορεί ο παρατηρητής να εντοπίσει άμεσα τα συχνότερα χρησιμοποιούμενα στοιχεία, όπως κάποιες διευθύνσεις ή θύρες, αφετηρίας ή προορισμού. Μπορεί λοιπόν να αποκτήσει μία πρώτη εικόνα για την βασικότερη κίνηση που διέρχεται από το υπό παρατήρηση δίκτυο. Μάλιστα το πάχος των ακμών του γραφήματος δηλώνει την τιμή «στήριξης» του αντίστοιχου «συχνού στοιχείου» δηλαδή το πόσο συχνά εμφανίζεται αυτό μεταξύ των transactions (flows) που έχουν καταγραφεί (σχήματα 5.1, 5.2, 5.3, 5.4 και 5.13).

Τα γραφήματα αυτού του τύπου μπορούν, εφόσον ο αριθμός των υπό αναπαράσταση «συχνών στοιχείων» είναι μικρότερος από έναν αριθμό (πρακτικά ο αριθμός αυτός δεν πρέπει να ξεπερνά τα τριάντα «συχνά στοιχεία»), να χρησιμοποιηθούν για τον εντοπισμό των κατηγοριών κίνησης που διέρχεται από το υπό παρατήρηση δίκτυο (σχήματα 5.12, 5.18 και 5.23). Οι παρατηρήσεις έδειξαν ότι η αναπαράσταση των “maximal frequent itemsets” είναι ποιο περιεκτική, το γράφημα είναι ποιο συνεκτικό στις πληροφορίες που αναπαριστά διότι δεν εμφανίζονται επαναλαμβανόμενα «συχνά στοιχεία» με αποτέλεσμα να περιγράφεται και μεγαλύτερος αριθμός πληροφοριών.

Ο αριθμός των καθέτων που εμφανίζονται στο γράφημα καθορίζει τον μέγιστο αριθμό στοιχείων που συνθέτουν το μεγαλύτερο (σε αριθμό στοιχείων) «συχνό στοιχείο».

Τα μειονεκτήματα βέβαια που πηγάζουν από τον τρόπο δημιουργίας αυτού του τύπου γραφημάτων είναι πολύ σημαντικότερα. Τα γραφήματα παράλληλων συντεταγμένων δεν μπορούν να αναπαραστήσουν «συχνά στοιχεία» τα οποία αποτελούνται από ένα μόνο στοιχείο. Αυτό έχει ως αποτέλεσμα να μην αναπαριστάται σημαντική πληροφορία.

Είναι δύσκολος ο εντοπισμός των «συχνών στοιχείων» διότι είναι δύσκολη η αντιστοίχιση των ακμών με τα στοιχεία στις καθέτους. Το πρόβλημα μεγαλώνει όσο μεγαλώνει ο αριθμός των στοιχείων που πρέπει να αναπαρασταθούν στο γράφημα. Αυτό βέβαια θα μπορούσε να αντιμετωπισθεί με την τοποθέτηση οριζόντιων γραμμών πλέγματος (gridlines) στο γράφημα.

Όπως αναφέρθηκε και στο κεφάλαιο 5 υπάρχουν περιπτώσεις γραφημάτων στα οποία δεν είναι ξεκάθαρο το ποια είναι τα στοιχεία τα οποία αναπαριστώνται. Υπάρχουν δηλαδή ακμές οι οποίες συνδυάζονται μεταξύ τους με τέτοιο τρόπο, με αποτέλεσμα ο παρατηρητής να μην μπορεί να καταλάβει ποια είναι ακριβώς τα «συχνά στοιχεία» τα οποία αναπαριστώνται. Αυτό οφείλεται στο γεγονός ότι η σχεδίαση του γραφήματος γίνεται με στόχο των περιορισμό όσο το δυνατό περισσότερο των εμφανιζόμενων ακμών ώστε να είναι και ποιο ευδιάκριτο το γράφημα (σχήματα 5.8 και 5.16).

Τέλος, οι κάθετοι που εμφανίζονται στο γράφημα δεν αντιστοιχούν σε κάποια συγκεκριμένη πληροφορία. Όλα τα στοιχεία τα οποία θα χρησιμοποιηθούν από το γράφημα εμφανίζονται στην ίδια θέση πάνω σε όλες τις καθέτους, π.χ. δεν χρησιμοποιείται αποκλειστικά η πρώτη κάθετος για την εμφάνιση των IP διευθύνσεων αφετηρίας, η δεύτερη για την απεικόνιση των διευθύνσεων προορισμού κ.λπ.. Μία τέτοια προέκταση εκτός του ότι θα βοηθούσε πολύ στην ευκρίνεια του παραγώμενου γραφήματος, έχει ως αποτέλεσμα να μην είναι δυνατή μια πιθανή χρήση αυτών των τύπων γραφημάτων για την αναγνώριση κάποιου είδους επιθέσεων με βάση “visual signatures” (όπως αναφέρεται στο [11]).

6.3 Πλεονεκτήματα και Μειονεκτήματα των γράφων.

Η τεχνική “graph based visualization” (γράφοι) είναι σίγουρα χρηστικότερη από αυτή των παράλληλων συντεταγμένων διότι δίνει ευκολότερα τη δυνατότητα στον παρατηρητή να εντοπίσει τα στοιχεία τα οποία συνθέτουν τα «συχνά στοιχεία»

Όπως και στην περίπτωση των γραφημάτων παράλληλων συντεταγμένων όσο περισσότερα είναι τα «συχνά στοιχεία» τα οποία αναπαριστώνται τόσο ποιο εύκολα εντοπίζονται τα συχνότερα στοιχεία. Αυτά είναι που συγκεντρώνουν πάνω τους τις περισσότερες ακμές (βέλη) (σχήματα 5.25 και 5.36). Ακόμη υπάρχουν και τα μεγέθη των κύκλων που εμφανίζονται στο γράφημα, τα οποία μπορούν να χρησιμοποιηθούν για να γίνει εντοπισμός των συχνότερων στοιχείων. Από την άλλη στην περίπτωση εμφάνισης μεγάλου αριθμού «συχνών στοιχείων» χάνεται η λεπτομέρεια διότι εφόσον ο αριθμός των υπό αναπαράσταση στοιχείων είναι μεγάλος γίνεται επικάλυψη στοιχείων με αποτέλεσμα να μην είναι δυνατός ο διαχωρισμός και η αναγνώριση των «συχνών στοιχείων».

Σε αντίθεση με τα γραφήματα παράλληλων συντεταγμένων τα σχεδιαγράμματα με γράφους έχουν τη δυνατότητα αναπαράστασης των «συχνών στοιχείων» που αποτελούνται από ένα μόνο στοιχείο. Το γράφημα δεν δημιουργεί παρανοήσεις για το ποια είναι τα «συχνά στοιχεία» τα οποία αναπαριστούνται, δεν υπάρχουν δηλαδή «κρυμμένα» στοιχεία όπως αναφέρθηκε στο Κεφάλαιο 5 για τα γραφήματα παράλληλων συντεταγμένων.

Όλοι οι γράφοι εμφανίζουν μία συμμετρικότητα η οποία οφείλεται στο γεγονός ότι τα υπό αναπαράσταση δεδομένα είναι δεδομένα κίνησης δικτύου και συγκεκριμένα ροές δεδομένων κάθε μία από τις οποίες έχει και μία αντίστροφα συμμετρική ροή. Απόκλιση από την απόλυτη συμμετρία μπορεί να οφείλεται στις λεγόμενες ροές μίας κατεύθυνσης οι οποίες δεν έχουν αντίστροφες ροές. Συνήθως οφείλονται σε ενέργειες όπως scanning είτε στην ύπαρξη χαμένων πακέτων (πακέτα τα οποία δεν έφτασαν ποτέ στον προορισμό τους οπότε δεν πήραν και απάντηση). Η εμφάνιση λοιπόν μη συμμετρικών γραφημάτων μπορεί να οδηγήσει στην ανεύρεση είτε χαμένων ροών είτε στην ανίχνευση επιθέσεων. Φυσικά ο εντοπισμός της απόκλισης από την απόλυτη συμμετρία θα μπορούσε να γίνει ευκολότερα εφόσον ο αλγόριθμος σχεδιασμού των γραφημάτων πριν τη δημιουργία αυτών έβρισκε τα συμμετρικά μεταξύ τους στοιχεία και τα τοποθετούσε σε ακριβής συμμετρικές θέσεις στο γράφημα. Μία άλλη προσέγγιση (κατά την οποία δεν ενδιαφέρει τόσο ο εντοπισμός επιθέσεων) θα μπορούσε να μην εμφανίζει τα συμμετρικά στοιχεία μειώνοντας έτσι το μέγεθος του γράφου βελτιώνοντας παράλληλα την ευκρίνειά του.

Στην ανίχνευση distributed denial of server attacks ή port ή IP address scanning μπορεί να οδηγήσει και ο εντοπισμός στην περίπτωση αναπαράστασης “maximal frequent itemsets”, «συχνών στοιχείων» τα οποία αποτελούνται από ένα μόνο στοιχείο. Όπως αναφέρθηκε και στο πέμπτο κεφάλαιο, η ύπαρξη “maximal frequent itemsets” με ένα στοιχείο δηλώνει την εμφάνιση αυτού του στοιχείου σε πολλά transactions στα οποία όμως τα υπόλοιπα στοιχεία μεταξύ των transactions δεν είναι κοινά. Υπάρχει δηλαδή μία επανάληψη του στοιχείου αυτού ενώ ταυτόχρονα υπάρχει μεγάλη «ποικιλία» (διαφοροποίηση) στις τιμές των υπόλοιπων στοιχείων, στα transactions στα οποία εμφανίζεται αυτό.

6.4 Προτάσεις για τη Βελτίωση (Καλύτερη Προσαρμογή σε Δεδομένα Κίνησης Δικτύου Υπολογιστών) των Τεχνικών Αναπαράστασης

Με βάση όλα τα παραπάνω για να μπορέσουν οι αναφερόμενες τεχνικές γραφικής αναπαράστασης να προσαρμοστούν στην αναπαράσταση «συχνών στοιχείων» που προέρχονται από δεδομένα κίνησης που καταγράφεται σε ένα δίκτυο υπολογιστών θα πρέπει να γίνουν σε αυτά κάποιες τροποποιήσεις-βελτιώσεις.

6.4.1 Προτάσεις Τροποποιήσεων στα Γραφήματα Παράλληλων Συντεταγμένων

Για τον ευκολότερο εντοπισμό των συχνότερων «συχνών στοιχείων» καλό θα ήταν τα στοιχεία που εμφανίζονται στις καθέτους να είναι ταξινομημένα με βάση την συχνότητα εμφάνισής τους στα transactions της βάσης που χρησιμοποιείται για την εύρεση των «συχνών στοιχείων». Με αυτό τον τρόπο θα μπορεί ο παρατηρητής γρηγορότερα να εντοπίσει ποιες είναι οι IP διευθύνσεις ή και θύρες που χρησιμοποιούνται περισσότερο στο δίκτυό του. Η προσθήκη και οριζόντιων γραμμών πλέγματος (gridlines) θα βοηθούσε στον ευκολότερο εντοπισμό των στοιχείων στα οποία καταλήγουν οι ακμές και κατά συνέπεια στον ευκολότερο εντοπισμό των «συχνών στοιχείων».

Η μη αναπαράσταση των «συχνών στοιχείων» τα οποία αποτελούνται από ένα μόνο στοιχείο από τα γραφήματα παράλληλων συντεταγμένων έχει ως αποτέλεσμα την απώλεια πληροφορίας. Θα μπορούσαν αυτά τα «συχνά στοιχεία» να εμφανίζονται στις καθέτους εμπλουτίζοντας τα γραφήματα με πρόσθετα δεδομένα. Θα μπορούσαν αυτά να αναπαριστώνται ξεχωριστά από τα υπόλοιπα «συχνά στοιχεία» π.χ. στην αρχή (κάτω χαμηλά) ή στο τέλος (στο άνω άκρο) κάποιας καθέτου ως σημεία, χωρίς να ξεκινά από αυτά κάποια ακμή.

Λόγω του μεγάλου αριθμού των διασταυρώσεων των ακμών μεταξύ τους είναι δύσκολος ο εντοπισμός των «συχνών στοιχείων». Ακόμη, δεν είναι ξεκάθαρο σε μερικές περιπτώσεις το ποια είναι τα «συχνά στοιχεία» που αναπαριστώνται. Καλό θα ήταν λοιπόν ο τρόπος δημιουργία των γραφημάτων αυτών να τροποποιηθεί ώστε να αποφεύγονται οι πολλές διασταυρώσεις των ακμών καθώς και οι ύπαρξη κρυμμένων «συχνών στοιχείων».

Όπως αναφέρθηκε στο πέμπτο κεφάλαιο τα γραφήματα παράλληλων συντεταγμένων μπορούν να χρησιμοποιηθούν ευκολότερα για τον εντοπισμό του είδους της κίνησης που διέρχεται από το υπό παρατήρηση δίκτυο υπολογιστών. Αυτό βοηθά τον παρατηρητή των γραφημάτων να εντοπίσει με κάποια προσπάθεια π.χ. το γεγονός ότι στο δίκτυο μεγάλο ποσοστό της καταγραφείσας κίνησης προέρχεται από την πρόσβαση των χρηστών σε δικτυακού τύπους (web sites). Για τον ευκολότερο εντοπισμό αυτού θα μπορούσε να χρωματίζονται με συγκεκριμένο χρώμα οι ακμές που αφορούν «συχνά στοιχεία» συγκεκριμένου τύπου κίνησης (σχήματα 5.12, 5.18 και 5.23). Θα μπορούσε δηλαδή ακμές που καταλήγουν σε στοιχεία που σχετίζονται με well known port numbers που μπορούν να χαρακτηρίσουν κίνηση, να χρωματίζονται με το κατάλληλο χρώμα. Παράδειγμα, ακμές που καταλήγουν σε στοιχεία όπως το SP_110, DP_110, SP_25, DP_25, SIP_143, DIP_143 να χρωματίζονται όλες με το ίδιο χρώμα αφού όλες οι θύρες αυτές χαρακτηρίζουν κίνηση που σχετίζεται με αποστολή και λήψη ηλεκτρονικών μηνυμάτων (ηλεκτρονική αλληλογραφία).

Πριν τον σχεδιασμό της γραφικής αναπαράστασης «συχνών στοιχείων» με την τεχνική των γραφημάτων παράλληλων συντεταγμένων θα μπορούσε να πραγματοποιείται ένα πρόσθετο βήμα το οποίο θα βοηθούσε στην καλύτερη κατανόηση της αναπαραστώμενης πληροφορίας. Το βήμα αυτό θα μπορούσε να είναι μία διαδικασία reverse DNS ή μία διαδικασία WHO IS ώστε στο τελικό γράφημα να μην εμφανίζονται IP διευθύνσεις αλλά φιλικά προς το χρήστη ονόματα (τα αντίστοιχα host names των IP addresses). Με αυτό τον τρόπο θα μπορούσε ο παρατηρητής να εντοπίζει ακόμη και τις γεωγραφικές τοποθεσίες με τις οποίες το δίκτυό του επικοινωνεί.

Τέλος, θα μπορούσε να δημιουργηθεί μία παραλλαγή των γραφημάτων παράλληλων συντεταγμένων με στόχο τον εντοπισμό κάποιου είδους επιθέσεων στο υπό παρατήρηση δίκτυο υπολογιστών. Θα μπορούσε κάθε κάθετος στο γράφημα να σχετίζεται με συγκεκριμένη πληροφορία, θα μπορούσε π.χ. η πρώτη κάθετος να χρησιμοποιείται μόνο από IP διευθύνσεις προορισμού, η δεύτερη από IP διευθύνσεις αφετηρίας, η τρίτη από θύρες προορισμού και η τέταρτη από θύρες αφετηρίας. Αυτή η τροποποίηση θα μπορούσε να βοηθήσει στη δημιουργία “visual signatures” και κατ’ επέκταση στον εντοπισμό επιθέσεων ενώ παράλληλα θα βελτίωνε την ευκρίνεια του γραφήματος.

6.4.2 Προτάσεις Τροποποιήσεων των γράφων.

Με τη χρήση των γράφων παίρνουμε μία πιο ευανάγνωστη αναπαράσταση των «συχνών στοιχείων». Παρόλα αυτά και σε αυτού του τύπου τα γραφήματα θα μπορούσαν να γίνουν τροποποιήσεις ώστε να βελτιωθεί η αποδοτικότητά τους.

Στα γραφήματα αυτού του τύπου κάθε κύκλος ο οποίος στην ουσία αναπαριστά ένα «συχνό στοιχείο», περιέχει έναν αριθμό ο οποίος δηλώνει την σειρά με την οποία παράχθηκε το συγκεκριμένο «συχνό στοιχείο» με βάση τον αλγόριθμο παραγωγής (Apriori). Αυτό δεν προσφέρει κάποια αξιοσημείωτη πληροφορία. Θα μπορούσε αντί αυτού μέσα στον κύκλο να εμφανίζεται ο αριθμός «στήριξης» του «συχνού στοιχείου» ή ένας αριθμός ο οποίος θα δηλώνει τη σειρά του «συχνού στοιχείου» σε μια λίστα ταξινομημένη με βάση την τιμή «στήριξης» των «συχνών στοιχείων». Αυτή η αλλαγή θα προσέφερε την πρόσθετη πληροφορία του πόσο σημαντικό είναι το υπό παρατήρηση «συχνό στοιχείο» (όσο πιο μεγάλη είναι η τιμή «στήριξης» τόσο πιο συχνά εμφανίζεται το «συχνό στοιχείο» στα transactions της βάσης).

Για τον ευκολότερο διαχωρισμό των στοιχείων που αναπαριστώνται, δηλαδή το διαχωρισμό των στοιχείων που αναπαριστούν IP διευθύνσεις από αυτά τα οποία αναπαριστούν θύρες καθώς και το διαχωρισμό των στοιχείων που αναπαριστούν στοιχεία προορισμού από αυτά που αναπαριστούν στοιχεία αφετηρίας, θα μπορούσε τα γραφήματα να εμπλουτιστούν με χρωματικό κώδικα. Θα μπορούσε λοιπόν κάθε παραλληλόγραμμο στο γράφημα να χρωματίζεται ανάλογα με το περιεχόμενό του. Θα μπορούσε δηλαδή οι IP διευθύνσεις αφετηρίας να χρωματίζονται με κάποιο συγκεκριμένο χρώμα μεταξύ τους, διαφορετικό από το χρώμα με το οποίο θα χρωματίζονται τα παραλληλόγραμμα που περιέχουν IP διευθύνσεις προορισμού, θύρες προορισμού και θύρες αφετηρίας. Με αυτό τον τρόπο θα μπορεί κανείς να δει άμεσα αν το στοιχείο που παρατηρεί είναι IP διεύθυνση ή αριθμός θύρα και να ξεχωρίσει για το αν πρόκειται για στοιχείο αφετηρίας ή προορισμού.

Όπως αναφέρθηκε οι γράφοι παρουσιάζουν ένα είδος συμμετρίας το οποίο οφείλεται στο είδος δεδομένων που αναπαριστούν (two way flows). Το γεγονός αυτό όπως αναφέρθηκε και πριν μπορεί να βοηθήσει σημαντικά στον εντοπισμό κάποιων προβλημάτων που μπορεί να έχουν να κάνουν είτε με απώλεια κάποιων πακέτων είτε με κάποιο είδος επίθεσης. Για να γίνει πιο εύκολος ο εντοπισμός των διαφοροποιήσεων από την απόλυτη συμμετρία θα μπορούσε να τροποποιηθεί ο αλγόριθμος

σχεδιασμού των γραφημάτων ώστε πριν τη δημιουργία αυτών να εντόπιζε τα συμμετρικά μεταξύ τους «συχνά στοιχεία» και να τα τοποθετούσε σε ακριβής συμμετρικές θέσεις μεταξύ τους στο γράφημα. Θα μπορούσε ακόμη, εφόσον απλά ενδιαφέρει μόνο η βελτίωση της ευκρίνειας του γράφου να μην εμφανίζονται τα συμμετρικά στοιχεία μειώνοντας έτσι το μέγεθος του γράφου.

Τονίστηκε ακόμη ότι υπάρχει περίπτωση στην αναπαράσταση των “maximal frequent itemsets” με τη χρήση γράφων η εμφάνιση «συχνών στοιχείων» με ένα μόνο στοιχείο να έχει αρκετά μεγάλη πιθανότητα να δηλώνει ένα είδος επίθεσης. Θα μπορούσε λοιπόν κατά τη δημιουργία του γραφήματος με “maximal frequent itemsets” εφόσον υπήρχαν «συχνά στοιχεία» με ένα μόνο στοιχείο αυτά είτε να χρωματίζονται με κάποιο συγκεκριμένο χρώμα είτε να τοποθετούνται σε συγκεκριμένο σημείο του προκύπτοντος γραφήματος. Έτσι θα μπορούσαν να εντοπιστούν ευκολότερα και κατ’ επέκταση να εξεταστούν για την περίπτωση που δηλώνουν κάποια επίθεση.

Χρήσιμο θα ήταν στα γραφήματα να εμφανίζονται και πληροφορίες όπως ο συνολικός αριθμός των ροών (flows/transactions) που απεικονίζονται καθώς και το σύνολο των κόμβων και των ακμών που εμφανίζονται στο γράφο.

Μια κατηγορία συστημάτων ανίχνευσης εισβολών (I.D.S.) στηρίζονται στην τεχνική ανίχνευσης ανωμαλιών (anomaly based IDS). Δημιουργούν ένα baseline και καταγράφουν αποκλίσεις από αυτό. Κάτι αντίστοιχο θα μπορούσε να γίνει και στην περίπτωση γραφημάτων αναπαράστασης συχνών στοιχείων. Θα μπορούσε να δημιουργηθεί ένα σύστημα το οποίο, χρησιμοποιώντας δεδομένα διαφορετικών αλλά κοντινών μεταξύ τους χρονικών στιγμών, θα δημιουργεί γραφήματα «συχνών στοιχείων» και θα τα αναπαριστά το ένα μετά το άλλο με στόχο τον εντοπισμό «συχνών στοιχείων» (ροών) που κάποιες φορές εμφανίζονται και κάποιες άλλες όχι. Στηριζόμενο λοιπόν στην λογική του εντοπισμού αποκλίσεων από το Baseline θα μπορούσε κάτι τέτοιο να αξιοποιηθεί από κάποια IDS.

6.5 Επίλογος

Το σύστημα παρακολούθησης που στήθηκε και οι παρατηρήσεις των δεδομένων και κυρίως των γραφημάτων «συχνών στοιχείων» που προέκυψαν από αυτό έδειξαν ότι τα «συχνά στοιχεία» και η γραφική αναπαράστασή τους μπορούν να συνεισφέρουν σε μεγάλο βαθμό στην ολοκληρωμένη παρακολούθηση ενός δικτύου υπολογιστών. Συγκεκριμένα, όπως αναφέρθηκε και στο Κεφάλαιο 5,

η εφαρμογή και παρατήρηση γράφων στο υπό παρατήρηση δίκτυο βοήθησε στον εντοπισμό επίθεσεων προς το τοπικό δίκτυο (θύρα 23 και IPs αφετηρίας από την Κίνα). Βοήθησε στο να μπορεί ο παρατηρητής να βγάζει γρήγορα συμπεράσματα για την γενικότερη εικόνα του είδους κίνησης που διέρχεται από το δίκτυό του.

Ο τρόπος αναπαράστασης με γράφους και κατ' επέκταση η πληροφορία που παρέχεται μέσα από αυτόν αντλείται ευκολότερα από τον παρατηρητή. Το παραγόμενο σχεδιάγραμμα παρέχει γρήγορα και εύκολα την πληροφορία χωρίς να χρειάζεται ο παρατηρητής να συνδυάσει πληροφορίες για την εξαγωγή του αποτελέσματος. Η μέθοδος δημιουργίας γράφων λοιπόν δίνει την πιο ευανάγνωστη για τον διαχειριστή δικτύου αναπαράσταση των «συχνών στοιχείων» (όταν αυτά δεν ξεπερνούν ένα συγκεκριμένο αριθμό), βοηθώντας τον να βγάλει συμπεράσματα για την κίνηση στο δίκτυο που παρακολουθεί.

Γενικότερα, η αναπαράσταση των «συχνών στοιχείων», σε αντίθεση με την αναπαράσταση του συνόλου των δεδομένων, δίνει γρηγορότερα την πληροφορία των συχνότερα χρησιμοποιούμενων στοιχείων στον παρατηρητή. Δε χρειάζεται να ψάξει κανείς για να βρει τις πιο συχνά χρησιμοποιούμενες ροές δεδομένων, αλλά αυτές εμφανίζονται άμεσα αφού αυτό παράγει ο αλγόριθμος «συχνών στοιχείων».

Πρόσθετα ο αριθμός των στοιχείων που αναπαριστώνται είναι πολύ μικρότερος από την περίπτωση που θα έπρεπε να αναπαρασταθεί το σύνολο των δεδομένων της κίνησης με αποτέλεσμα το γράφημα που δημιουργείται να έχει μεγαλύτερη ευκρίνεια και να αναπαριστά σημαντικές μόνο πληροφορίες. Σίγουρα χάνονται κάποιες πληροφορίες με την αναπαράσταση μόνο των «συχνών στοιχείων», θέτοντας όμως ως βάση το γεγονός ότι η σημαντικότητα της πληροφορίας εξαρτάται σε μεγάλο ποσοστό από τη συχνότητα εμφάνισης-επανάληψης αυτής, η απώλεια δεν είναι μεγάλη. Φιλτράρονται και εμφανίζονται μόνο τα πιο συχνά εμφανιζόμενα δεδομένα λογικά λοιπόν τα πιο χρήσιμα.

Με το γραφικό τρόπο δίνεται η δυνατότητα να αναπαρασταθούν περισσότερες πληροφορίες χρησιμοποιώντας μικρότερο χώρο εμφάνισης. Επιτυγχάνεται η μεγαλύτερη εκμετάλλευση της αντιληπτικής ικανότητας του ανθρώπου (μία εικόνα χίλιες λέξεις).

Η γραφική αναπαράσταση λοιπόν των «συχνών στοιχείων» μπορεί να συνεισφέρει στην καλύτερη και πιο ολοκληρωμένη παρακολούθηση και κατ' επέκταση γνώση ενός δικτύου υπολογιστών.

Βιβλιογραφία

- [01] «Computer Network - Wikipedia, the free encyclopedia» [Ηλεκτρονική σελίδα], http://en.wikipedia.org/wiki/Computer_network. [Πρόσβαση 1/11/2011].
- [02] «Computer Virus - Wikipedia, the free encyclopedia» [Ηλεκτρονική σελίδα], http://en.wikipedia.org/wiki/Computer_virus. [Πρόσβαση 1/11/2011].
- [03] «Computer Worm - Wikipedia, the free encyclopedia» [Ηλεκτρονική σελίδα], http://en.wikipedia.org/wiki/Computer_worm. [Πρόσβαση 1/11/2012].
- [04] «Trojan horse (computing) - Wikipedia, the free encyclopedia» [Ηλεκτρονική σελίδα], [http://en.wikipedia.org/wiki/Trojan_horse_\(computing\)](http://en.wikipedia.org/wiki/Trojan_horse_(computing)). [Πρόσβαση 2/11/2011].
- [05] J. Harrison, «ISA Server Port Scan Alerts» [Ηλεκτρονική σελίδα], <http://technet.microsoft.com/en-us/library/cc512655.aspx>. [Πρόσβαση 2/11/2011].
- [06] J. F. Kurose και K. W. Ross, «Computer Networking, A Top Down Approach», Addison-Wesley, 2010.
- [07] M. Burnett , «Blocking Brute Force Attacks - System Administration Database» [Ηλεκτρονική σελίδα], http://www.cs.virginia.edu/~csadmin/gen_support/brute_force.php. [Πρόσβαση 5/11/2011].
- [08] «Free Network Monitoring Software for Small Networks» [Ηλεκτρονική σελίδα], <http://www.techrepublic.com/whitepapers/free-network-monitoring-software-for-small-networks/345366>. [Πρόσβαση 2/12/2011].
- [09] S. M. GadAllah, «The Importance of Logging and Traffic Monitoring for Information Security» SANS Institute, 2003.
- [10] S. Northcutt, L. Zeltser, S. Winters, K. Kent και R. W. Ritchey, «*Inside Network Perimeter Security*», Sams Publishing, 2005, pp. 201-221.
- [11] C. Hyunsang, L. Heejo και K. Hyogon, «Fast Detection and Visualization of Network Attacks on Parallel Coordinates», *Computers & Security*, τόμ. 28, αρ. 5, pp. 276-288, 2009.
- [12] S. Tricaud και P. Saadé, «Applied parallel coordinates for logs and network traffic attack analysis» *Journal in Computer Virology*, τόμ. 6, αρ. 1, pp. 1-29, 2010.
- [13] T. Oetiker, «MRTG - Tobi Oetiker's MRTG - The Multi Router Traffic Grapher» [Ηλεκτρονική σελίδα], <http://oss.oetiker.ch/mrtg/>. [Πρόσβαση 7/12/2011].

- [14] One Network Complete Control, «SNMP Monitoring One Critical Component to network Management», 2005.
- [15] A. Cecil, «A Summary of Network Traffic Monitoring and Analysis», www.cse.wustl.edu/~jain/cse567-06/ftp/net_monitoring.pdf.
- [16] W. M. Tzeyoung, «Information Assurance Tools Report – Intrusion Detection Systems», Information Assurance Technology Analysis Center (IATAC), Herndon, 2009.
- [17] D. Barbara, J. Couto, S. Jajodia, L. Popyack και N. Wu, «ADAM: Detecting intrusion by data mining», Proceedings of IEEE Workshop on Information Assurance and Security, 2001.
- [18] K. Fox, R. Henning, J. Reed και R. Simonian, «A Neutral Network Approach Towards Intrusion Detection», Technical report, Harris Corporation, 1990.
- [19] «Snort :: Home Page», SourceFire [Ηλεκτρονική σελίδα], <http://www.snort.org/>. [Πρόσβαση 2/11/2011].
- [20] A. Orebaugh, S. Biles και J. Babbin, «Miscellaneous Other Uses» σε *Snort Cookbook*, O'Reilly Media, 2005, pp. 225-264.
- [21] S. Niphadkar, «Analysis of Packet Sniffers – TCPdump VS Ngrep VS Snoop», Dept of Computer Science, George Mason University.
- [22] H. Ningning, «Network monitoring and diagnosis based on available bandwidth measurement» Ph.D. dissertation, School of Computer Science, Computer Science Department, Carnegie Mellon University, Pittsburgh, 2006.
- [23] «Network bandwidth monitor - test PRTG for free!» [Ηλεκτρονική σελίδα], http://www.paessler.com/bandwidth_monitor. [Πρόσβαση 10/12/2011].
- [24] P. Haag, [Ηλεκτρονική σελίδα], <http://nfsen.sourceforge.net/>. [Πρόσβαση 15/12/2011].
- [25] J. Zhihua, «Visualization of Network Traffic to Detect Malicious Network Activity», Master's thesis, Helsinki University of Technology, Faculty of Information and Natural Sciences, Department of Computer Science and Engineering, Norway, 2008.
- [26] E. R. Tufte, «The Visual Display of Quantitative Information», Connecticut: Graphics Press, Connecticut, 2001.
- [27] C. Ware, «Information Visualization: Perception for Design», Academic Press, San Diego , 2000.
- [28] «Internet Monitoring and Web Filtering Software with Web Security and Internet Access Control - GFI WebMonitor» [Ηλεκτρονική σελίδα], <http://www.gfi.com/internet-monitoring->

software#screenshots. [Πρόσβαση 18/12/2011].

- [29] C. Estan, S. Savage και G. Varghese, «Automatically inferring patterns of resource consumption in network traffic», σε *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, New York, 2003.
- [30] X. Yin, W. Yurcik, M. Treaster, Y. Li και K. Lakkaraju, «VisFlowConnect: Netflow Visualizations of Link Relationships for Security Situational Awareness», σε *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, New York, 2004.
- [31] K. Lakkaraju, W. Yurcik και A. J. Lee, «NVisionIP: Netflow Visualizations of System State for Security Situational Awareness», σε *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, New York, 2004.
- [32] R. Berthier, M. Cukier, M. Hiltunen, D. Kormann, G. Vesonder και D. Sheleheda, «Nfsight: NetFlow-based Network Awareness Tool», σε *Proceedings of the 24th international conference on Large installation system administration*, CA, 2010.
- [33] L. B. Ayre, «Data Mining for Information Professionals», Διαθέσιμο στο <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.105.2789&rep=rep1&type=pdf>, 2006.
- [34] R. Agrawal, T. Imieliński και A. Swami, «Mining Association Rules Between Sets of Items in Large Databases», σε *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, New York, 1993.
- [35] «Data Mining» [Ηλεκτρονική σελίδα], http://en.wikipedia.org/wiki/Data_mining#Application_examples. [Πρόσβαση 1/2/2012].
- [36] A. Lakhina, M. Crovella και C. Diot, «Mining Anomalies Using Traffic Feature Distributions», σε *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, New York, 2005.
- [37] D. Brauckhoff, X. Dimitropoulos, A. Wagner και K. Salamatian, «Anomaly Extraction in Backbone Networks Using Association Rules», σε *Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference*, Chicago, Illinois, USA, 2009.
- [38] I. Paredes-Oliva, X. Dimitropoulos, M. Molina, P. Barlet-Ros και D. Brauckhoff, «Automating Root-Cause Analysis of Network Anomalies Using Frequent Itemset Mining», *Proceedings of the ACM SIGCOMM 2010 Conference*, New Delhi, India, 2010.
- [39] K. Yoshida, Y. Shomura και Y. Watanabe, «Visualizing Network Status», σε *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*, Hong Kong, 2007.
- [40] R. A. Elmasri και S. B. Navathe, «Fundamentals of Database Systems», Addison-Wesley Longman

Publishing Co,Boston, 1998.

- [41] J. W. Seifert, «Data Mining: An Overview», *International Journal of Computer Science And Technology* , τόμ. 2, αρ. 2, pp. 336-339, 2010.
- [42] M. Rajalakshmi, T. Purusothaman και R. Nedunchezian, «Maximal Frequent Itemset Generation Using Segmentation Approach», *International Journal of Database Management Systems*, τόμ. 3, αρ. 3, 2011.
- [43] C. Lucchese, S. Orlando και R. Perego, «Mining frequent closed itemsets out of core», *Performance Computing*, pp. 417-427, 2006.
- [44] M. Houtsma και A. Swami, «Set-Oriented Mining for Association Rules in Relational Databases», σε *ICDE '95 Proceedings of the Eleventh International Conference on Data Engineering* , Washington, 1995.
- [45] R. Agrawal και R. Srikant, «Fast Algorithms for Mining Association Rules», *Citeseer*, τόμ. 1215, αρ. 1558601538, pp. 487-499, 1994.
- [46] M. J. Zaki, S. Parthasarathy, M. Ogihara και W. Li, «New Algorithms for Fast Discovery of Association Rules», *AAAI Press*, τόμ. 20, αρ. 651, pp. 283-286, 1997.
- [47] J. Han, J. Pei, Y. Yin και R. Mao, «Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach», *Data Mining and Knowledge Discovery* , τόμ. 8, αρ. 1, 2004.
- [48] R. C. Agarwal, C. C. Aggarwal και V. Prasad, «A Tree Projection Algorithm for Generation of Frequent Item Sets», *Journal of Parallel and Distributed Computing* , τόμ. 61, αρ. 3, pp. 350-371, 2001.
- [49] C. Borgelt, «Keeping things simple: finding frequent item sets by recursive elimination», σε *OSDM '05 Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, New York, 2005.
- [50] M. Sheikhan και Z. Jadidi, «Misuse Detection Using Hybrid of Association Rule Mining and Connectionist Modeling», *World Applied Sciences Journal*, τόμ. 7, pp. 31-37, 2009.
- [51] M. Sheikhan και D. Gharavian, «Combination o Elman Neural Network and Classification-Bsed Predictive Association Rules to Imrov Computer Networks' Security», *World Applied Sciences Journal*, τόμ. 7, pp. 80-86, 2009.
- [52] E. Bloedorn, A. D. Christiansen, W. Hill, C. Skorupka, L. M. Talbot και J. Tivel, «Data Mining for Network Intrusion Detection: How to Get Started», διαθέσιμο στο <http://citeseerx.ist.psu.edu>.
- [53] Z. Liu και Y. Lai, «A Data Mining Framework for Building Intrusion Detection Models Based on IPv6», σε *ISA '09 Proceedings of the 3rd International Conference and Workshops on Advances in*

Information Security and Assurance, Heidelberg, 2009.

- [54] C. Jun και C. Xiaowei, «Intrusion Detection System Research Based on Data Mining for IPv6», σε *IFITA '10 Proceedings of the 2010 International Forum on Information Technology and Applications - Volume 01*, Washington, 2010.
- [55] C. Miao και W. Chen, «A Study of Intrusion Detection System Based on Data Mining», σε *Information Theory and Information Security (ICITIS), 2010 IEEE International Conference on*, Beijing, 2010.
- [56] W. Lee και S. J. Stolfo, «Data Mining Approaches for Intrusion Detection», σε *Proceedings of the 7th USENIX Security Symposium*, San Antonio, Texas, 1998.
- [57] C. Clifton και G. Gengo, «Developing Custom Intrusion Detection Filters Using Data Mining», σε *MILCOM 2000 Proceedings 21st Century Military Communications Architectures and Technologies for Information Superiority Cat No00CH37155*, 2000.
- [58] Y. Yu, «A novel intrusion detection approaches based on data mining», σε *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, 2010.
- [59] «NetFlow - Wikipedia, the free encyclopedia» [Ηλεκτρονική σελίδα], <http://en.wikipedia.org/wiki/Netflow>. [Πρόσβαση 15/2/2012].
- [60] «Configuring NetFlow», σε «*Cisco IOS Switching Services Configuration Guide*», διαθέσιμο στο site <http://www.cisco.com>, 2001, pp. xc_47 - xc_54.
- [61] M. Hahsler και S. Chelluboina, «Visualizing Association Rules: Introduction to the R-extension Package arulesViz», διαθέσιμο στο site <http://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf>.
- [62] M. Hahsler, B. Grün και K. Hornik, «arules – A Computational Environment for Mining Association Rules and Frequent Item Sets», *Journal Of Statistical Software*, τόμ. 14, αρ. 15, pp. 1-25, 2005.
- [63] L. Yang, «Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates», σε *ICCSA'03 Proceedings of the 2003 international conference on Computational science and its applications: Part I*, Heidelberg, 2003.
- [64] L. Yang, «*Visual Data Mining*», Heidelberg, Springer-Verlag, 2008, pp. 60-75.
- [65] C. K.-S. Leung, P. P. Irani και C. L. Carmichael, «FlsViz: A Frequent Itemset Visualizer», σε *PAKDD'08 Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining*, Heidelberg, 2008.
- [66] C. K.-S. Leung, P. P. Irani και C. L. Carmichael, «WiFlsViz: Effective Visualization of Frequent Itemsets», σε *ICDM '08 Proceedings of the 2008 Eighth IEEE International Conference on Data*

Mining , Washington, 2008.

- [67] C. K.-S. Leung και C. L. Carm, «FpVAT: a visual analytic tool for supporting frequent pattern mining», *ACM SIGKDD Explorations Newsletter*, τόμ. 11, αρ. 2, pp. 39-48, 2009.
- [68] C. K.-S. Leung και C. L. Carmichael, «Exploring Social Networks: A Frequent Pattern Visualization Approach», σε *SOCIALCOM '10 Proceedings of the 2010 IEEE Second International Conference on Social Computing* , Washington, 2010.