Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή
στα Πληροφοριακά Συστήματα

Χωροχρονική Επεξεργασία και Απεικόνιση Διευθύνσεων IP που Αποστέλλουν SPAM

Φώτιος Κερασιώτης

Επιβλέπων
Δρ. Ξενοφών Δημητρόπουλος

Φεβρουάριος 2012

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Χωροχρονική Επεξεργασία και Απεικόνιση Διευθύνσεων IP που Αποστέλλουν SPAM

Φώτιος Κερασιώτης

Επιβλέπων
Δρ. Ξενοφών 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

Φεβρουάριος 2012

Περίληψη

Στόχος της μεταπτυχιακής διατριβής είναι η ανάπτυξη ενός πιλοτικού ιστότοπου - διαδικτυακού συστήματος, ο οποίος θα απεικονίζει την χωρο-χρονική κατανομή των διευθύνσεων IP που αποστέλλουν spam μηνύματα μέσω του ηλεκτρονικού ταχυδρομείου ανά τον κόσμο, έτσι ώστε ο εν δυνάμει χρήστης να μπορέσει να κατανοήσει το φαινόμενο spamming ως προς την προέλευσή του, την ένταση με την οποία αυτό εμφανίζεται, καθώς και τον τρόπο με τον οποίο αυτό εξελίσσεται στο χρόνο. Στα πλαίσια αυτού, αρχικά μελετάται το spamming ως προς τη φύση του, περιγράφοντας τόσο τους τρόπους με τους οποίους δημιουργείται όσο και αυτούς με τους οποίους αντιμετωπίζεται. Στη συνέχεια, περιγράφεται ο σχεδιασμός και η υλοποίηση του συστήματος χρησιμοποιώντας ελεύθερα διαθέσιμες τεχνολογίες και πόρους του Διαδικτύου, δίνοντας έμφαση στον χωρο-χρονικό προσδιορισμό, στην ανάκτηση, στην επεξεργασία και στους τρόπους απεικόνισης (π.χ. χάρτες, διαγράμματα) της πληροφορίας σχετικά με τις IP διευθύνσεις που αποστέλλουν spam μηνύματα. Βασικό δε πρόβλημα αποτελεί η διαχείριση και απεικόνιση του μεγάλου όγκου πληροφορίας, και για το λόγο αυτό μελετώνται διάφορες μέθοδοι συσταδοποίησης (clustering) δεδομένων, οι οποίες περιορίζουν-συμπιέζουν την πληροφορία και βελτιώνουν τόσο την απόδοση του συστήματος όσο και την ποιότητα της απεικόνισης της πληροφορίας στον ιστότοπο. Ύστερα από τη σύγκριση των μεθόδων αυτών, διακρίνονται εκείνα τα χαρακτηριστικά τους που ταιριάζουν περισσότερο στις προδιαγραφές απεικόνισης του συστήματος, και τελικά υλοποιείται και ενσωματώνεται στο σύστημα ένας μηχανισμός που εγγυάται σχετικά γρήγορη και ακριβή ενημέρωση των δεδομένων κατά την περιήγηση του χρήστη στον ιστότοπο.

Summary

The goal of the M.Sc. dissertation is the development of a prototype web-site – Web-based system, which presents the spatiotemporal distribution of the IP addresses sending spam messages from all over the world over the Electronic Mail (e-mail) service infrastructure, so that a potential user can study and understand the spamming phenomenon in terms of its source, its intensity and the way it evolves throughout time. According to this, at first, spamming is studied in terms of its nature by describing both the ways it is invoked and the ones it is encountered. Then, the design and the implementation of the system is presented, which is based on open-source or generally available for free Internet technologies and resources, focusing on the spatiotemporal determination, the acquisition, the processing and the ways of representation (e.g. maps, diagrams) of the corresponding information to the IP addresses sending spam messages. A key issue is the control and representation of the corresponding large amount of information, and for this reason various clustering methods are studied considering the compression of information they provide and consequently the improvement of both the system efficiency and the quality of representation which can be achieved. After the comparison of these methods, some of their features which provide better performance with respect to the representation requirements of the system are extracted and are incorporated in the design of a system's mechanism which promises relatively good performance for the user interaction with the web-site.

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής μου διατριβής θα ήθελα να ευχαριστήσω τον κ. Ξενοφώντα Δημητρόπουλο που μου πρότεινε το συγκεκριμένο θέμα, καθώς και για την σημαντική καθοδήγησή του ως επιβλέπων κατά τη διάρκεια της εκπόνησης της διατριβής. Επίσης, θα ήθελα να ευχαριστήσω τον κ. Θανάση Χατζηλάκο για την υποστήριξή του σε διάφορα θέματα κατά τη διάρκεια της φοίτησής μου στο Ανοιχτό Πανεπιστήμιο Κύπρου. Και τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την αμέριστη συμπαράστασή τους.

Περιεχόμενα

Κεφάλαιο 1
Εισαγωγή

Ένα φαινόμενο το οποίο αποτελεί ενδιαφέρον μελέτης είναι εκείνο της αποστολής spam ηλεκτρονικών μηνυμάτων (μέσω ηλεκτρονικού ταχυδρομείου) ή spamming, το οποίο βρίσκεται σε έξαρση τα τελευταία χρόνια αγγίζοντας το ποσοστό του 90% των ηλεκτρονικών μηνυμάτων που αποστέλλονται συνολικά σε κάποια χρονική περίοδο. Για να αντιμετωπισθεί αυτό το φαινόμενο έχουν αναπτυχθεί διάφορες τεχνικές αντιμετώπισης οι οποίες εστιάζουν είτε στην αποθάρρυνση-«πρόληψη» από την αποστολή τέτοιων μηνυμάτων είτε στην προστασία-«θεραπεία» από τα τελευταία εντοπίζοντας και αποκλείοντάς τα. Πιο συγκεκριμένα, το ενδιαφέρον που παρουσιάζει το spamming έγκειται στην κατανόηση της προέλευσής του, στον προσδιορισμό της έντασης με την οποία εμφανίζεται, καθώς και στην παρατήρηση του τρόπου που εξελίσσεται στο χρόνο. Έτσι, στόχος της μεταπτυχιακής διατριβής είναι η δημιουργία ενός συστήματος που θα αποσκοπεί στην απεικόνιση αυτής της πληροφορίας με τέτοιο τρόπο ώστε να διευκολύνει την παρατήρηση του φαινομένου αυτού και την περαιτέρω κατανόησή του για οποιονδήποτε ενδιαφερόμενο.

Δεδομένων των παραπάνω, σχεδιάζεται και υλοποιείται ένα διαδικτυακό σύστημα (ιστότοπος) που θα είναι δημόσια προσπελάσιμο και θα δίνει τη δυνατότητα σε κάποιο χρήστη της

περιήγησης σε πληροφορία που αφορά το spamming φαινόμενο, ενώ παράλληλα θα επιτυγχάνει την καλύτερη δυνατή παρουσίασή του. Εξαιτίας της χωροχρονικής φύσης του φαινομένου αυτού, κρίνεται απαραίτητη η απεικόνιση της πληροφορίας, εκτός από τα συμβατικά διαγράμματα και γραφικές παραστάσεις, και με την τοποθέτηση αυτής σε χάρτες οι οποίοι θα προσφέρουν τη δυνατότητα επισκόπησης του spamming ως προς την προέλευσή του ανά τον κόσμο ή μια συγκεκριμένη περιοχή. Τα δεδομένα που χρησιμοποιούνται ανακτούνται και αποθηκεύονται περιοδικά από το σύστημα, ώστε να περιγράφεται σε επίπεδο πραγματικού χρόνου το spamming φαινόμενο. Αυτά είναι IP διευθύνσεις που μαζί με τον χρόνο εντοπισμού τους περιλαμβάνονται στον κατάλογο αποκλεισμού («μαύρη λίστα») ενός παροχέα (www.heise.de/ix/nixspam/nixspam.blackmatches) ο οποίος βασίζεται στην τεχνική αντιμετώπισης του spamming μέσω εντοπισμού της προέλευσης των ηλεκτρονικών μηνυμάτων. Επειδή δε η απεικόνιση σε χάρτες απαιτεί γεωγραφικές συντεταγμένες, γίνεται γεωπροσδιορισμός αυτών των IP διευθύνσεων μέσω μιας κατάλληλης βάσης δεδομένων που διατηρείται στον εξυπηρετητή του ιστότοπου. Η βάση αυτή παρέχεται ελεύθερα από τον παροχέα HostIP.com. Όσον αφορά την απεικόνιση σε χάρτη των χωροχρονικών πια δεδομένων που δημιουργούνται, για να επιτευχθεί η αναπαράσταση της χρονικής εξέλιξης του spamming χρησιμοποιούνται είτε περισσότεροι από έναν χάρτες σε παράθεση είτε εναλλαγή εικόνων (animation) για κάποιο χρονικό διάστημα που θα επιλέξει ο χρήστης, κάνοντας έτσι ευδιάκριτες τις αλλαγές που παρουσιάζει το φαινόμενο ανά χρονική βαθμίδα (π.χ. ώρες, ημέρες, εβδομάδες, μήνες, χρόνια).

Τέλος, βασικό αλγοριθμικό πρόβλημα αποτελεί η διαχείριση και απεικόνιση σε χάρτες του μεγάλου όγκου πληροφορίας που πραγματεύεται ο υλοποιηθής ιστότοπος. Συγκεκριμένα, μελετούνται διάφορες μέθοδοι συσταδοποίησης (clustering) των δεδομένων, οι οποίες περιορίζουν-συμπιέζουν την πληροφορία που ζητάει ο χρήστης προς απεικόνιση και κατά συνέπεια τον όγκο των δεδομένων που μεταφέρονται από τον εξυπηρετητή στον υπολογιστή του, ενώ παράλληλα η ποιότητα της πληροφορίας διατηρείται σε επίπεδο τέτοιο που να εκφράζει με αρκετή πληρότητα την κατάσταση του spamming φαινομένου. Από αυτές τις μεθόδους επιλέγονται εκείνα τα χαρακτηριστικά τους που είναι ικανά για να παράγουν μια ικανοποιητική συσταδοποίηση για το σύστημα, σύμφωνα με τις προδιαγραφές απεικόνισης που το διέπουν.

Η μεταπτυχιακή διατριβή οργανώνεται ως εξής:

- Το *Κεφάλαιο 2* περιγράφει το φαινόμενο της ανεπιθύμητης ηλεκτρονικής αλληλογραφίας, spam. Συγκεκριμένα, γίνεται μια παρουσίαση των βασικών στοιχείων και μηχανισμών που διέπουν την ηλεκτρονική αλληλογραφία και των πλεονεκτημάτων της, δίνεται ο ορισμός της έννοιας spam, γίνεται περιγραφή των μεθόδων με τις οποίες οι spammers εξασφαλίζουν-συλλέγουν τις ηλεκτρονικές διευθύνσεις στις οποίες αποστέλλουν μηνύματα spam, παρουσιάζονται οι μέθοδοι που υιοθετούνται για την αντιμετώπιση του φαινομένου, και τέλος γίνεται παρουσίαση του σχετικού νομικού υπόβαθρου.

- Το *Κεφάλαιο 3*, αρχικά παρουσιάζει γενικότερα την αρχιτεκτονική ενός συστήματος απεικόνισης χωρικών δεδομένων, και στη συνέχεια γίνεται μια εξέταση της συγκεκριμένης περίπτωσης του spamming για τα δεδομένα ενός τέτοιου συστήματος παρουσιάζοντας και τη συναφή εργασία σε σχέση με το υλοποιηθέν σύστημα της μεταπτυχιακής διατριβής. Στο τέλος, γίνεται μια παρουσίαση των βασικών («περιφερειακών») στοιχείων που απαιτούνται από ένα τέτοιο σύστημα (και χρησιμοποιούνται στο υλοποιηθέν) όσον αφορά τη συλλογή διευθύνσεων IP που αποστέλλουν μηνύματα spam και τον γεωπροσδιορισμό τους.

- Στο *Κεφάλαιο 4,* αρχικά γίνεται παρουσίαση της έννοιας της χωρικής συσταδοποίησης και στη συνέχεια γίνεται μελέτη των αντιπροσωπευτικότερων μεθόδων συσταδοποίησης ανά κατηγορία ως προς τα βήματα που ακολουθούν. Στο τέλος, γίνεται μια σύγκριση αυτών ως προς την απόδοσή τους, ώστε να διακριθούν εκείνα τα χαρακτηριστικά τους που υπόσχονται μια ικανοποιητική συσταδοποίηση όσον αφορά τις προδιαγραφές απεικόνισης δεδομένων του υλοποιηθέντος συστήματος.

- Στο *Κεφάλαιο 5* γίνεται παρουσίαση της αρχιτεκτονικής του υλοποιηθέντος συστήματος, και περιγράφονται η βάση δεδομένων που διαχειρίζεται, οι λειτουργίες που το διέπουν, καθώς και η διεπαφή με την οποία απεικονίζει τη σχετική με το spamming πληροφορία και αλληλεπιδρά με το χρήστη.

- Τέλος, στο *Κεφάλαιο 6* παρουσιάζονται συμπεράσματα που αφορούν την μεταπτυχιακή διατριβή, καθώς και προτάσεις για μελλοντική εργασία με σκοπό τη βελτίωση ή διόρθωση των όσων έχουν πραγματοποιηθεί στα πλαίσια αυτής.

Κεφάλαιο 2

Ανεπιθύμητη Ηλεκτρονική Αλληλογραφία (SPAM) – Ορισμός, Τεχνικές Αντιμετώπισης, Νομοθεσία

Αρχικά, η γνωστοποίηση και προώθηση προϊόντων ή υπηρεσιών των εταιριών στους καταναλωτές χρησιμοποιώντας τη διαφήμιση, διαδικασία η οποία αποτελεί μέρος της ευρύτερης έννοιας του μάρκετινγκ, βασιζόταν κυρίως κατά ένα μέρος στη χρησιμοποίηση των μέσων μαζικής ενημέρωσης και κατά ένα μέρος στη διανομή εντύπων (αποστολή φυλλαδίων, επιστολών, κ.α.) είτε στις οικίες και στα γραφεία μέσω αλληλογραφίας είτε σε δημόσιους χώρους. Όσον αφορά το πρώτο μέσο, ήταν και συνεχίζει να είναι ο πιο εύκολος και αξιόπιστος τρόπος διαφήμισης, ενώ το δεύτερο χρησιμοποιείται για ενημέρωση συγκεκριμένων ομάδων καταναλωτών καθώς και όταν σκοπός είναι η μείωση του κόστους διαφήμισης. Παρόλο που η αποστολή εντύπων ήταν οικονομικότερη, οι εταιρίες συνέχιζαν να επωμίζονται ένα αρκετά υψηλό κόστος συγκριτικά με το αντίκτυπο σε προσέλκυση καταναλωτών. Με την ανάπτυξη και

προώθηση της διαφήμισης μέσω της ηλεκτρονικής αλληλογραφίας (email), το συγκεκριμένο κόστος ουσιαστικά μειώθηκε στο ελάχιστο δυνατό, αποτελώντας τον επικρατέστερο τρόπο διαφήμισης όχι μόνο ως προς το οικονομικό όφελος, αλλά και ως προς το αντίκτυπο στην προσέλκυση πελατών, δεδομένης και της ολοένα αυξανόμενης χρήσης της ηλεκτρονικής αλληλογραφίας έναντι της συμβατικής κατά το μεγαλύτερο μέρος της καθημερινής «γραπτής» επικοινωνίας.

Η ανάπτυξη του τελευταίου τρόπου διαφήμισης σε συνδυασμό με την αδυναμία του πρωτοκόλλου ηλεκτρονικής αλληλογραφίας *Simple Mail Transfer Protocol* (SMTP) [51] να προστατεύσει τον δέκτη από την συγκεκριμένη αποστολή μηνυμάτων, ουσιαστικά μετατόπισε το «φόρτο» στον καταναλωτή. Έτσι, σήμερα κατακλύζεται καθημερινά από πληθώρα τέτοιων μηνυμάτων όσον αφορά την ηλεκτρονική του αλληλογραφία, που είτε δυσχεραίνουν την εύρεση της επιθυμητής αλληλογραφίας είτε αποπροσανατολίζουν από την άντληση των σημαντικών πληροφοριών είτε, τέλος, παραπλανούν. Αυτός ο τύπος ανεπιθύμητης αλληλογραφίας μπορεί γενικά να ονομασθεί *spam*.

Υπάρχουν διάφορες διαθέσιμες αναλύσεις που παρουσιάζουν την κατάσταση του προβλήματος της αποστολής μηνυμάτων spam ή αλλιώς *spamming*, μερικές από τις οποίες δείχνουν ότι μέχρι και περίπου το 90% των αποστελλόμενων μηνυμάτων ηλεκτρονικής αλληλογραφίας αναγνωρίζονται ως ενοχλητικά ανεπιθύμητα μηνύματα, δηλαδή spam [64, 95]. Αυτή η κατάσταση, δεν απασχολεί μόνο τους παραλήπτες των μηνυμάτων αυτών, αλλά επίσης και τους παροχείς ηλεκτρονικής αλληλογραφίας οι οποίοι χρησιμοποιούν μεθόδους για την ελάττωση του αριθμού των μηνυμάτων spam που κατακλύζουν τους αποθηκευτικούς χώρους των εξυπηρετητών τους.

Στο συγκεκριμένο κεφάλαιο, αρχικά, η Ενότητα 2.1 περιγράφει τους βασικούς ορισμούς και τους σχετικούς μηχανισμούς όσον αφορά την ηλεκτρονική αλληλογραφία, καθώς και τα πλεονεκτήματα που προσφέρει ως μέσο επικοινωνίας. Έπειτα, στην Ενότητα 2.2 αναφέρονται διάφοροι ορισμοί για το spam, ενώ στην επόμενη ενότητα γίνεται μια σύντομη ιστορική αναδρομή στην εξέλιξή του. Τρόποι συλλογής διευθύνσεων ηλεκτρονικής αλληλογραφίας από τους αποστολείς μηνυμάτων spam και τρόποι αντιμετώπισης του προβλήματος παρουσιάζονται στις Ενότητες 2.4 και 2.5, αντίστοιχα, και τέλος, το σχετικό νομικό πλαίσιο που ισχύει σήμερα, παρουσιάζεται στην Ενότητα 2.6.

2.1 Ηλεκτρονική Αλληλογραφία

Για να μελετηθεί πιο εκτενώς το spamming, στη συγκεκριμένη ενότητα γίνεται μια σύντομη αναφορά στα βασικά χαρακτηριστικά του ηλεκτρονικού ταχυδρομείου και στους μηχανισμούς που το διέπουν, καθώς και μια παρουσίαση των πλεονεκτημάτων που έχουν οι χρήστες του, είτε αυτοί είναι εταιρίες είτε κοινωνικές και άλλες ομάδες είτε μεμονωμένα πρόσωπα.

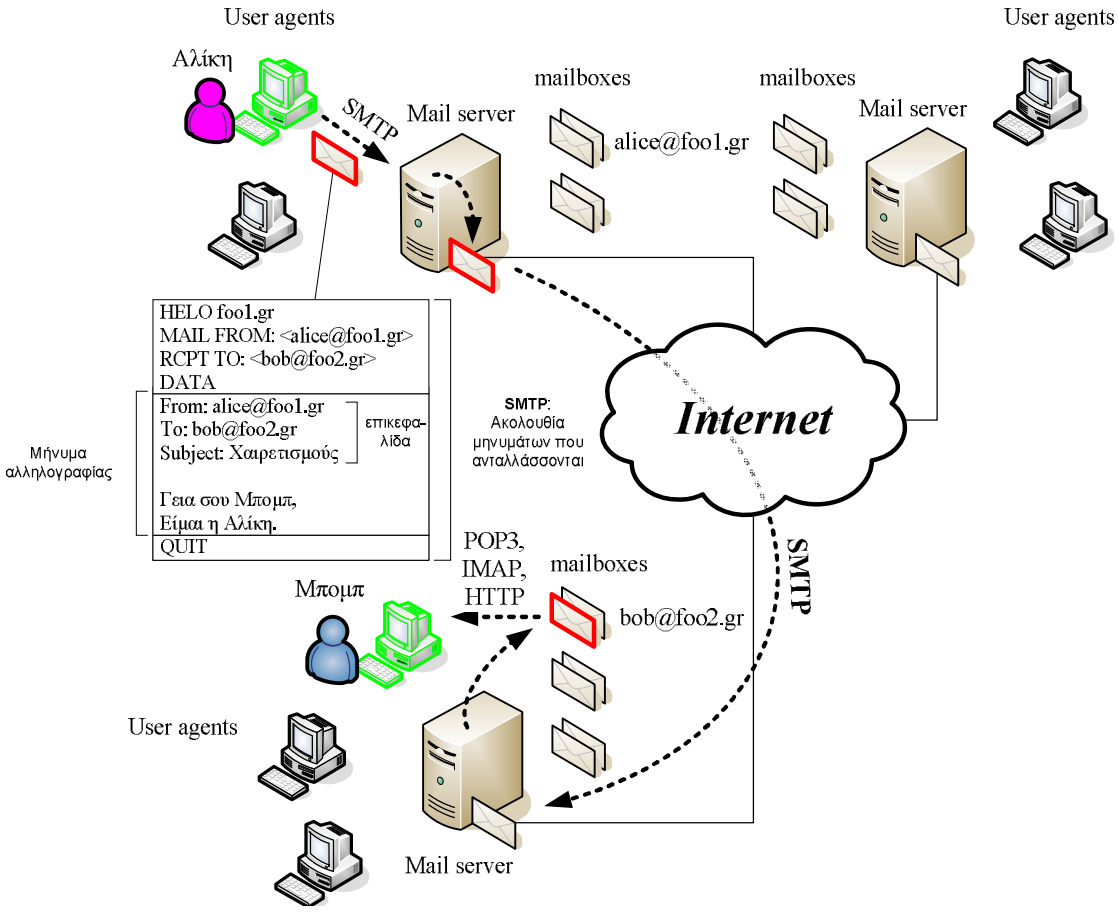
2.1.1 Βασικά Χαρακτηριστικά

Στην Εικόνα 2.1, φαίνεται μια τυπική διαδρομή ενός ηλεκτρονικού μηνύματος από τον αποστολέα στον παραλήπτη χρησιμοποιώντας το Διαδίκτυο. Πιο συγκεκριμένα, η αποστολή του ηλεκτρονικού μηνύματος βασίζεται στο *πρωτόκολλο SMTP* και στη λογική πελάτη-εξυπηρετητή, και τα κύρια στοιχεία που συμμετέχουν στην επικοινωνία είναι το *λογισμικό χρήστη-πελάτη* (user agent), π.χ. προγράμματα διαχείρισης ηλεκτρονικής αλληλογραφίας, όπως Mozilla Thunderbird, Microsoft Outlook κ.α., και ο *εξυπηρετητής ηλεκτρονικού ταχυδρομείου* (mail server), ο οποίος ανήκει σε κάποιον παροχέα, όπως Google, Yahoo, Hotmail κ.α.. Το μήνυμα δε που μεταφέρεται αποτελείται από το *φάκελο* (envelope), ο οποίος περιλαμβάνει τις ηλεκτρονικές διευθύνσεις του αποστολέα και του παραλήπτη, καθώς και προαιρετικές επεκτάσεις του SMTP πρωτοκόλλου, και από το περιεχόμενο (content), το οποίο αποτελείται από δύο μέρη, την *επικεφαλίδα* (email head) και το *σώμα* (email body) του μηνύματος. Μια τυπική διαδικασία αποστολής ηλεκτρονικού μηνύματος μέσω του SMTP πρωτοκόλλου (χρησιμοποιούνται τα ευρέως διαδεδομένα ονόματα για χρήστες σε διαδικτυακά θέματα, Αλίκη/Alice και Μπομπ/Bob) είναι:

1. Αρχικά, η Αλίκη χρησιμοποιεί το λογισμικό πελάτη στον υπολογιστή της, για να συνθέσει το μήνυμα που θέλει να στείλει στον Μπομπ, παρέχοντας επίσης την ηλεκτρονική του διεύθυνση (π.χ. bob@foo2.gr). Μετά τη σύνθεση του μηνύματος, η Αλίκη δίνει τελικά την εντολή αποστολής του.

2. Το λογισμικό πελάτη στον υπολογιστή της Αλίκης στέλνει το ηλεκτρονικό μήνυμα [85] στον εξυπηρετητή ηλεκτρονικού ταχυδρομείου στον οποίο ανήκει η ηλεκτρονική της διεύθυνση, και ο τελευταίος το τοποθετεί στην ουρά αποστολής μηνυμάτων του.

3. Το λογισμικό του εξυπηρετητή (client πλευρά του πρωτοκόλλου SMTP), στη συνέχεια, βλέπει το ηλεκτρονικό μήνυμα στην ουρά αποστολής μηνυμάτων και συνδέεται με το



Εικόνα 2.1: Τυπική διαδρομή ενός ηλεκτρονικού μηνύματος από τον αποστολέα (Αλίκη) στον παραλήπτη (Μπομπ)

λογισμικό (server πλευρά του πρωτοκόλλου SMTP) στον εξυπηρετητή ηλεκτρονικού ταχυδρομείου του Μπομπ, δημιουργώντας μια σύνδεση με βάση το πρωτόκολλο μεταφοράς του Διαδικτύου, *Transmission Control Protocol* (TCP) [80].

4. Μετά από κάποια ανταλλαγή SMTP μηνυμάτων, το λογισμικό του εξυπηρετητή ηλεκτρονικού ταχυδρομείου της Αλίκης προωθεί το ηλεκτρονικό μήνυμα (μετά την εντολή DATA) στην TCP σύνδεση που δημιουργήθηκε.

5. Ο εξυπηρετητής ηλεκτρονικού ταχυδρομείου του Μπομπ λαμβάνει το μήνυμα και το τοποθετεί στο χώρο αποθήκευσης των μηνυμάτων (mailbox) του Μπομπ, ώστε να μπορεί να προσπελάσει το μήνυμα όποτε αυτός συνδεθεί στο ηλεκτρονικό του ταχυδρομείο.

6. Τέλος, ο Μπομπ χρησιμοποιεί το λογισμικό πελάτη στον υπολογιστή του για να προσπελάσει το χώρο αποθήκευσης μηνυμάτων που αντιστοιχεί στην ηλεκτρονική του

διεύθυνση και να διαβάσει το μήνυμα της Αλίκης. Η προσπέλαση γίνεται με χρήση πρωτοκόλλων πρόσβασης στο ηλεκτρονικό ταχυδρομείο (mail access protocols), όπως τα *Post Office Protocol* (POP3) [65], *Internet Message Access Protocol* (IMAP) [18] και *HyperText Transfer Protocol* (HTTP) [30], με το τελευταίο να χρησιμοποιείται στην περίπτωση που γίνεται πρόσβαση με τον φυλλομετρητή ιστού (web browser).

2.1.2 Πλεονεκτήματα της Χρήσης του Ηλεκτρονικού Ταχυδρομείου

Γενικά, το ηλεκτρονικό ταχυδρομείο θεωρείται ως το μέσο εκείνο που μείωσε σημαντικά τον απαιτούμενο χρόνο επικοινωνίας και μετάδοσης δεδομένων, ανεξάρτητα από γεωγραφικές αποστάσεις, αλλά και το κόστος, με το μόνο πράγμα που είναι απαραίτητο να προηγηθεί να είναι η δήλωση μιας μοναδικής ανά τον κόσμο ηλεκτρονικής διεύθυνσης και ενός κωδικού πρόσβασης σε έναν από τους πολλούς παροχείς ηλεκτρονικού ταχυδρομείου που υπάρχουν σήμερα. Επίσης, προσέφερε επιπλέον δυνατότητες, όπως αυτοματοποιημένες αποστολές σε πολλαπλούς παραλήπτες, αποστολή υλικού πολυμέσων, οργάνωση και αποθήκευση των επαφών και της πληροφορίας σε κάθε συμβάν επικοινωνίας, και πρόσβαση στην αλληλογραφία από οπουδήποτε, αρκεί να υπάρχει πρόσβαση στο Διαδίκτυο. Όλα αυτά τα πλεονεκτήματα προσέλκυσαν και το εμπόριο, μετατρέποντας το συγκεκριμένο μέσο επικοινωνίας σε ένα εξαιρετικό εργαλείο μάρκετινγκ, προφέροντας ευκολία, ευελιξία, αμεσότητα, εξοικονόμηση χρόνου και χρήματος. Πιο συγκεκριμένα, τα κυριότερα πλεονεκτήματα όσον αφορά το ηλεκτρονικό ταχυδρομείο στον τομέα του μάρκετινγκ (email marketing) θα μπορούσαν να συνοψισθούν ως εξής:

1. *Χαμηλό κόστος, υψηλή ποιότητα και ευελιξία διαφήμισης*: Δεδομένου ότι οι επικοινωνία γίνεται ηλεκτρονικά, το κόστος της διαφήμισης επαφίεται μόνο στο σχεδιασμό των γραφικών, την εισαγωγή πολυμέσων και γενικότερα των εμπλουτισμένων στοιχείων που συνοδεύουν ένα ηλεκτρονικό μήνυμα, χωρίς να εξαρτάται από τον αριθμό των παραληπτών-καταναλωτών, ενώ παράλληλα παρέχεται η δυνατότητα τροποποίησης και επέκτασης των στοιχείων αυτών ανάλογα με την στρατηγική προώθησης και την διαθεσιμότητα των προϊόντων του διαφημιζόμενου προσώπου.

2. *Σωστή στόχευση και προσωπική επικοινωνία*: Μέσω χρήσης ειδικών λιστών διευθύνσεων αλληλογραφίας είτε από το ίδιο το διαφημιζόμενο πρόσωπο είτε από άλλους φορείς (π.χ. περιοδικά, ομάδες συζητήσεων) ομαδοποιούνται οι καταναλωτές σύμφωνα με διάφορα κριτήρια, όπως τα ενδιαφέροντά τους, οι προτιμήσεις τους, η γεωγραφική τους

τοποθεσία κ.α., πράγμα το οποίο κατευθύνει και ενισχύει την αποδοτικότητα τού μάρκετινγκ συμπεριλαμβάνοντας το προσωπικό στοιχείο στην επικοινωνία με τον καταναλωτή.

3. *Άμεση και αμφίδρομη επικοινωνία*: Με την ηλεκτρονική αλληλογραφία, σε αντίθεση με κάθε άλλο μέσο μάρκετινγκ (π.χ. έντυπο υλικό, τηλεόραση, ραδιόφωνο), η επικοινωνία με τους παραλήπτες καταναλωτές-πελάτες είναι άμεση και γρήγορη. Ο καταναλωτής έχει τη δυνατότητα της γρήγορης ανταπόκρισης στην λήψη ενός διαφημιστικού μηνύματος είτε καταθέτοντας μια παραγγελία είτε κάνοντας αίτηση για περισσότερες λεπτομέρειες για ένα προϊόν που τον ενδιαφέρει μέσω φορμών και υπερσυνδέσμων. Επίσης, ενθαρρύνεται ο διάλογος μέσω της ανταλλαγής απόψεων και σχολίων πάνω στα διατιθέμενα προϊόντα. Η πληροφορία που λαμβάνεται από κάθε ανταλλαγή τέτοιων μηνυμάτων δε, βοηθάει στην περεταίρω κατανόηση των αναγκών των καταναλωτών, πράγμα το οποίο μπορεί να καθοδηγήσει τις μετέπειτα ενέργειες του διαφημιζόμενου προσώπου (π.χ. εταιρεία, ελεύθερος επαγγελματίας) όσον αφορά τις προωθητικές του ενέργειες (π.χ. διαφήμιση) και τα προϊόντα που διαθέτει.

4. *Υψηλή και μετρήσιμη απόδοση διαφήμισης*: Εκτός από την υψηλή ποιότητα, δεδομένου του μεγάλου αριθμού παραληπτών τόσο των μηνυμάτων που στέλνονται από το διαφημιζόμενο πρόσωπο ή εταιρεία όσο και αυτών που προωθούνται από τους ίδιους τους παραλήπτες, η απόδοση της διαφήμισης βελτιώνεται αυξάνοντας την αναγνωρισιμότητα του διαφημιζόμενου προσώπου. Επίσης, από την ανταπόκριση των παραληπτών, μπορεί να μετρηθεί το πόσο αποδοτική ήταν μια διαφήμιση, μετρώντας για παράδειγμα τα μηνύματα που λήφθηκαν και αναγνώστηκαν από τους παραλήπτες, καθώς και πιο περιεχόμενο ήταν το πιο δημοφιλές, ώστε να καθοδηγηθούν οι μετέπειτα «κινήσεις» μάρκετινγκ που θα οδηγήσουν σε ακόμη μεγαλύτερη αύξηση της απόδοσης.

Από την άλλη μεριά, τα σημαντικά αυτά πλεονεκτήματα της χρήσης της ηλεκτρονικής αλληλογραφίας δεν θα μπορούσαν να μην οδηγήσουν και σε προβλήματα. Για παράδειγμα στο εμπόριο, παρόλο που πολλές εταιρίες χρησιμοποιούν το ηλεκτρονικό ταχυδρομείο για να επικοινωνήσουν με υπάρχοντες πελάτες, υπάρχουν άλλες που στέλνουν πληθώρα ηλεκτρονικών μηνυμάτων χωρίς τη συγκατάθεση των παραληπτών, με στόχο την όσον το δυνατό εκτενέστερη προώθηση των προϊόντων τους. Τέτοια μηνύματα, τα οποία μπορούν να προέρχονται είτε από εταιρίες είτε από άλλες κοινωνικές ομάδες και πρόσωπα, είναι γνωστά ως spam, έννοια η οποία παρουσιάζεται στη συνέχεια.

2.2 Ορισμός της Έννοιας Spam

Η έννοια *spamming* θεωρείται γενικότερα ως η χρήση συστημάτων ή μέσων μετάδοσης ηλεκτρονικής πληροφορίας για αποστολή πληθώρα αυτόκλητων μηνυμάτων χωρίς να απευθύνεται σε συγκεκριμένους παραλήπτες. Έτσι, εκτός από τον ευρέως διαδεδομένο τύπο μηνύματος spam ο οποίος σχετίζεται με το ηλεκτρονικό ταχυδρομείο, και που αποτελεί το αντικείμενο της μεταπτυχιακής διατριβής, άλλα συστήματα ή μέσα μετάδοσης ηλεκτρονικής πληροφορίας στα οποία αναφέρεται η συγκεκριμένη έννοια αποτελούν η αποστολή μηνυμάτων στα κινητά τηλέφωνα (mobile phone messaging spam, ή αλλιώς m-spam), η συγχρονισμένη ανταλλαγή μηνυμάτων (Instant Messaging (IM) spam, ή αλλιώς spim), η χρήση μηχανών αναζήτησης στο Διαδίκτυο (web search engine spam, ή αλλιώς spamdexing), κ.α..

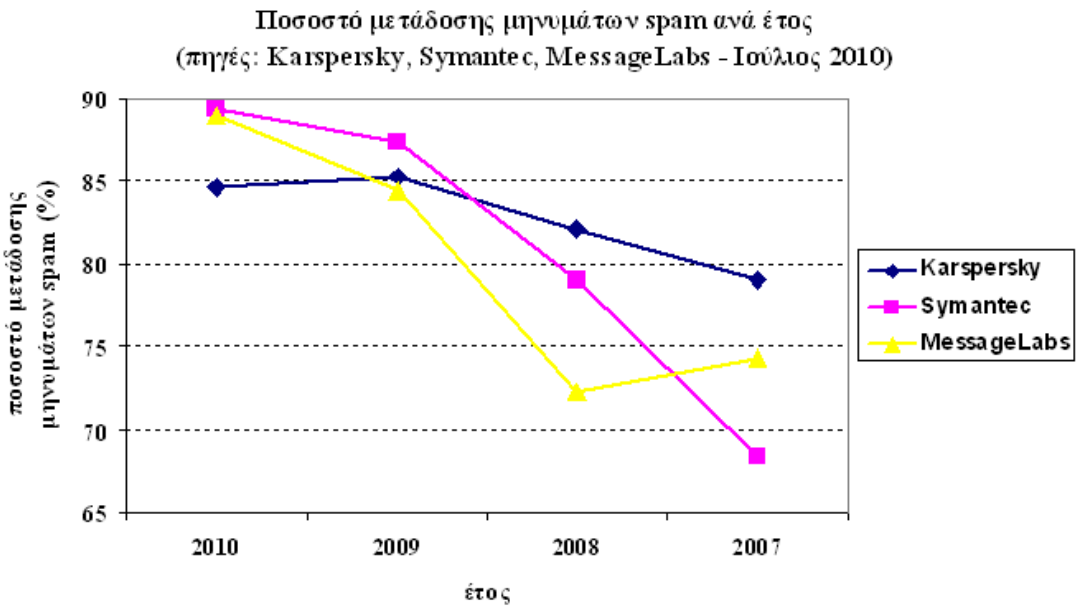
Σχετικά με το spam στο ηλεκτρονικό ταχυδρομείο, ή αλλιώς *email spam* ή *junk email*, υπάρχουν διάφοροι ορισμοί στη βιβλιογραφία για την συγκεκριμένη έννοια [31, 43, 64, 92]. Γενικά, *αυτόκλητα μηνύματα εμπορικού περιεχομένου* (Unsolicited Commercial Email - UCE) και *αυτόκλητα μηνύματα μαζικής αποστολής σε παραλήπτες* (Unsolicited Bulk Email - UBE) είναι δύο ορισμοί που χρησιμοποιούνται συνήθως για να περιγράψουν την έννοια του μηνύματος spam και τα χαρακτηριστικά του. Ο πρώτος αναφέρεται σε εκείνα τα μηνύματα που η πληροφορία που περιέχουν έχει να κάνει με το εμπόριο, όπως για παράδειγμα τη γνωστοποίηση μιας λίστας προϊόντων και των ενδεχόμενων προσφορών μιας εταιρίας, ενώ ο δεύτερος αναφέρεται σε μηνύματα που στέλνονται σε μεγάλο αριθμό παραληπτών. Η διαφορά μεταξύ των δύο αυτών ορισμών, ουσιαστικά, είναι ότι χρησιμοποιούν διαφορετικές μετρικές για να χαρακτηρίσουν το κατά πόσο ένα μήνυμα αλληλογραφίας είναι spam, με τον πρώτο να υπολογίζει τη σχέση της παρεχόμενης πληροφορίας με το εμπόριο και με τον δεύτερο να υπολογίζει τον αριθμό των παραληπτών στους οποίους αποστέλλεται μια συγκεκριμένη χρονική περίοδο το συγκεκριμένο μήνυμα. Παρόλα αυτά, και οι δύο ορισμοί περιλαμβάνουν το χαρακτηριστικό των μηνυμάτων αυτών ότι αποστέλλονται στους παραλήπτες χωρίς προηγουμένως να το έχουν ζητήσει, όπως προκύπτει και από το χαρακτηρισμό «αυτόκλητα». Βέβαια, ο συνδυασμός όλων των προηγουμένων χαρακτηριστικών είναι αυτός που θα καθορίσει τελικά αν ένα μήνυμα είναι spam, καθώς και το κατά πόσο το ληφθέν μήνυμα είναι επιθυμητό από τους παραλήπτες.

2.3 Ιστορία του Spam

Σύμφωνα με τον Templeton [96], η έννοια spam αποδίδεται σε μια βρετανική σατιρική παράσταση καλούμενη Monty Python, η οποία περιλάμβανε μια σκηνή σε ένα εστιατόριο όπου ορισμένοι Βίκινγκ τραγουδούσαν «Spam, spam, spam, spam, spam, spam, spam, spam, lovely spam! Wonderful spam!», μέχρι που τους είπαν να σταματήσουν εξαιτίας το εκνευριστικού τους ήχου. Αυτή είναι άλλωστε και η σημασία της έννοιας του spam, δηλαδή κάτι που επαναλαμβάνεται συνεχώς (π.χ. επαναλαμβανόμενη αποστολή ενοχλητικής ηλεκτρονικής αλληλογραφίας) προκαλώντας την δυσανασχέτηση του παραλήπτη.

Το πρώτο spam ως αυτόκλητο εμπορικό ηλεκτρονικό μήνυμα καταγράφηκε το 1978 όταν η εταιρεία DEC (Digital Equipment Corporation) διαφήμισε το προϊόν της DEC-20 στο τότε περιορισμένο σχετικά Διαδίκτυο που ονομαζόταν ARPANET. Έπειτα, μετά την τυποποίηση του SMTP πρωτοκόλλου το 1982, στο Usenet, ένα κατανεμημένο παγκοσμίως σύστημα συζητήσεων που αποτελούταν από ομάδες χρηστών με διάφορες θεματικές ενότητες, άρχισαν ορισμένες ομάδες να λαμβάνουν μηνύματα που είτε ζητούσαν χρήματα είτε παρουσίαζαν τρόπους γρήγορης απολαβής χρημάτων («make money fast»). Παρόλα αυτά, το συγκεκριμένο συμβάν δεν θεωρήθηκε σημαντικό μέχρι που το 1994 μια θρησκευτική ομάδα και οι δικηγόροι Laurence Canter και Martha Siegel έστειλαν πληθώρα μηνυμάτων («Global Alert for All: Jesus is Coming Soon» και «Green Card Lottery – Final One?», αντίστοιχα) σε κάθε ομάδα χρηστών του Usenet. Από το συγκεκριμένο συμβάν και μετά, η έννοια του spam έγινε δημοφιλής, αποτελώντας σημαντικό πρόβλημα ιδιαίτερα κατά την δεκαετία του ’90 με την έξαρση της χρήσης του Διαδικτύου, θέτοντας την ανάγκη για υιοθέτηση μέτρων αντιμετώπισης. Ο Paul Vixie ήταν ένας από τους πρώτους που δημιούργησαν «μαύρη λίστα» γνωστών αποστολέων spam. Το πρόβλημα του spamming έγινε ακόμα πιο έντονο μέχρι το 2002, οπότε και δόθηκε ιδιαίτερη προσοχή στην δημιουργία φίλτρων αλληλογραφίας (π.χ. Bayesian φίλτρα). Τέλος, κατά το 2005 το ενδιαφέρον στράφηκε σε τεχνικές πιστοποίησης (authentication mechanisms), ώστε να βελτιωθεί το σχετικά απλό πρωτόκολλο ηλεκτρονικού ταχυδρομείου SMTP ως προς την πιστοποίηση του αποστολέα της ηλεκτρονικής αλληλογραφίας, δεδομένης και της αυξημένης μετάδοσης μηνυμάτων spam σήμερα, περίπου το 90% της αποστελλόμενης ηλεκτρονικής αλληλογραφίας [64].

Η μεταβολή της μετάδοσης spam ως ποσοστό της αποστελλόμενης ηλεκτρονικής αλληλογραφίας σύμφωνα με τις αναφορές που παρουσιάζονται στους παροχείς λογισμικού ασφάλειας Διαδικτύου Kaspersky [46], Symantec [94] και MessageLabs [63] για τα έτη 2007-



Εικόνα 2.2: Απεικόνιση του ποσοστού μετάδοσης μηνυμάτων spam ετησίως

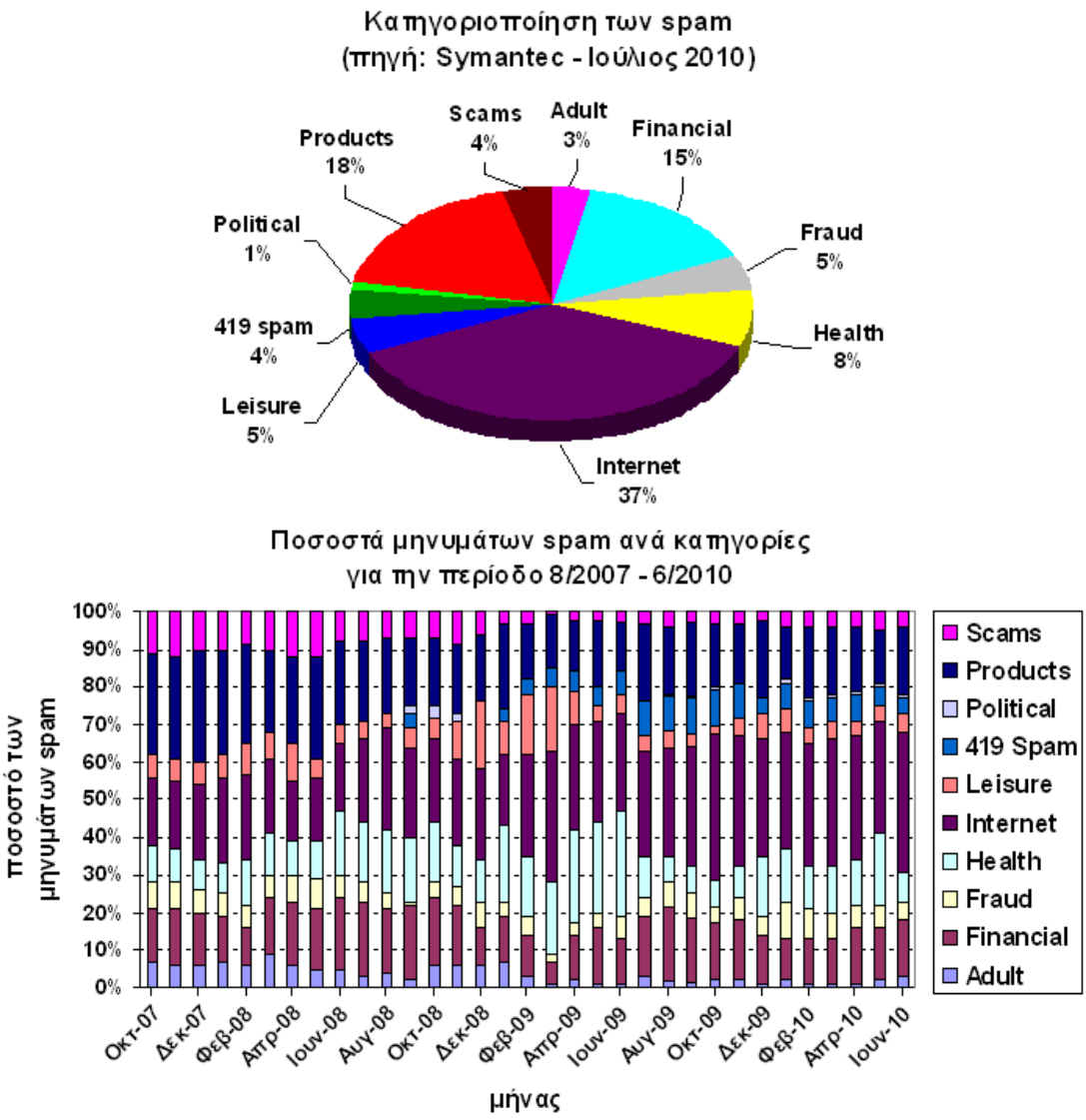
2010 φαίνεται στην Εικόνα 2.2. Επίσης, οι διάφορες κατηγορίες στις οποίες μπορούν να διαχωριστούν ανάλογα με το περιεχόμενό τους τα μηνύματα spam, σύμφωνα με τον παροχέα Symantec για τον Ιούλιο 2010 φαίνονται στην Εικόνα 2.3α, ενώ η μεταβολή τους στο χρόνο στην Εικόνα 2.3β. Οι κατηγορίες αυτές παρουσιάζονται παρακάτω.

- *Αποστολή μηνυμάτων σχετιζόμενα με τον ελεύθερο χρόνο (**Leisure**)*, τα οποία παρουσιάζουν τιμές και προσφορές διαφόρων ειδών απασχόλησης στον ελεύθερο χρόνο και στις διακοπές, π.χ. ταξιδιωτικές προσφορές.

- *Διαφήμιση διαδικτυακών ή σχετιζόμενων με υπολογιστές υπηρεσιών και προϊόντων (**Internet**)*, π.χ. φιλοξενία ιστοσελίδων, σχεδιασμός ιστοσελίδων.

- *Διαφήμιση προϊόντων ή υπηρεσιών σχετιζόμενες με την υγεία (**Health**)*, π.χ. φάρμακα, θεραπευτικά βότανα.

- *Αποστολή μηνυμάτων εξαπάτησης (**Fraud**)*, τα οποία παρουσιάζουν ως αποστολέα μια γνωστή εταιρεία και στόχο έχουν να ξεγελάσουν τους παραλήπτες αποκαλύπτοντας προσωπικές πληροφορίες, π.χ. ειδοποιήσεις σχετικά με λογαριασμούς τραπέζης και πιστωτικές κάρτες.



Εικόνα 2.3: α) Κατηγοριοποίηση μηνυμάτων spam και β) μεταβολή των ποσοστών των κατηγοριών μηνυμάτων spam ως προς το χρόνο

- *Αποστολή μηνυμάτων σχετιζόμενων με οικονομικά (**Financial**)*, τα οποία περιλαμβάνουν αναφορές ή προσφορές σχετιζόμενες με χρήματα, χρηματιστήριο και άλλες οικονομικές δραστηριότητες, π.χ. επενδύσεις, δάνεια, καταθέσεις σε τράπεζες.

- *Διαφήμιση προϊόντων ή υπηρεσιών για ενήλικες (**Adult**)*.

- *Αποστολή μηνυμάτων εξαπάτησης με στόχο την ευρεία προώθησή τους (**Scam**)*, τα οποία βασίζονται στη δημιουργία εμπιστοσύνης στο πρόσωπο του αποστολέα-spammer με διάφορα μέσα εξαπάτησης και παραπλάνησης, π.χ. αλυσιδωτά μηνύματα (chain letters), προώθηση διαφημιστικού υλικού τύπου πυραμίδας (pyramid schemes).

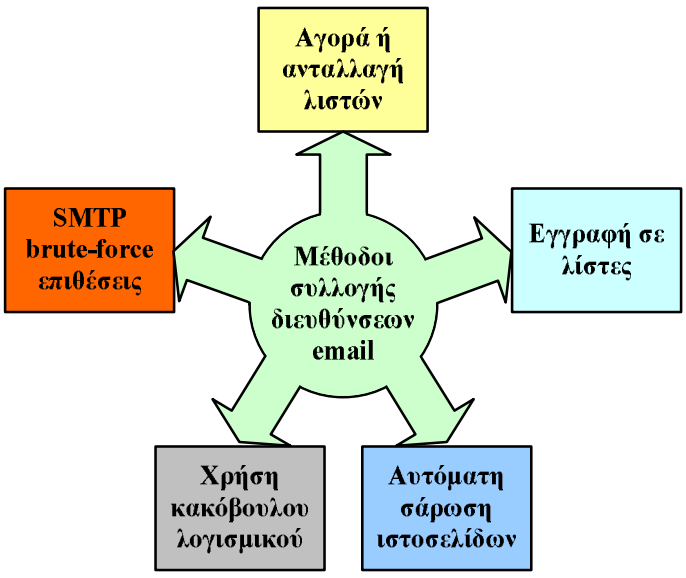
- *Διαφήμιση προϊόντων και υπηρεσιών γενικότερα (**Products**)*, π.χ. συσκευές, είδη ρουχισμού, δακτυλογραφήσεις.

- *Αποστολή πολιτικών μηνυμάτων (**Political**)*, π.χ. διαφήμιση πολιτικών προσώπων υποψήφιων σε εκλογές, αιτήσεις χορήγησης χρημάτων σε πολιτικά κόμματα ή πρόσωπα.

- *Ειδική κατηγορία μηνυμάτων εξαπάτησης (**419 spam** ή advance fee fraud)*, τα οποία συνήθως περιλαμβάνουν ειδοποιήσεις για απολαβή χρημάτων προερχόμενα από διαγωνισμούς-τυχερά παιχνίδια (λοταρίες). Το όνομά της προήλθε από σχετικό άρθρο του ποινικού κώδικα της Νιγηρίας όσον αφορά τα μηνύματα εξαπάτησης [69].

2.4 Μέθοδοι Συλλογής Ηλεκτρονικών Διευθύνσεων για Αποστολή Spam

Για να μπορέσουν οι αποστολείς, γνωστοί ως *spammers*, να στείλουν spam σε όσο το δυνατό περισσότερους παραλήπτες, ώστε να ικανοποιήσουν τους στόχους τους, αναφορικά με τα πλεονεκτήματα τις επικοινωνίας μέσω αλληλογραφίας έχουν αναπτύξει πολλαπλές μεθόδους συλλογής ηλεκτρονικών διευθύνσεων (email address harvesting), καθώς και τρόπους αποστολής και επιβεβαίωσης της λήψης των μηνυμάτων τους από τους παραλήπτες. Οι μέθοδοι αυτές παρουσιάζονται παρακάτω, ενώ συνοψίζονται στην Εικόνα 2.4.



Εικόνα 2.4: Μέθοδοι συλλογής email διευθύνσεων για αποστολή spam

2.4.1 Αγορά ή Ανταλλαγή Λιστών

Ο πιο εύκολος τρόπος απόκτησης λιστών ενεργών ηλεκτρονικών διευθύνσεων είναι η αγορά ή η ανταλλαγή τους με άλλες. Ανάλογα με τον παροχέα των λιστών αυτών, δύο τύποι αγοράς ή ανταλλαγής μπορούν να διακριθούν:

- Απόκτηση λιστών ηλεκτρονικών διευθύνσεων, οι οποίες συλλέχθηκαν είτε με άλλες μεθόδους (από τις επόμενες) είτε νόμιμα (για παράδειγμα μέσω εγγραφής πελατών σε λίστες αλληλογραφίας), και μεταπωλούνται σε οποιαδήποτε εταιρεία ή άλλο πρόσωπο τις επιθυμεί, για να διαφημισθεί μέσω ηλεκτρονικού ταχυδρομείου.

- Απόκτηση λιστών ηλεκτρονικών διευθύνσεων με νόμιμες διαδικασίες, όπου για παράδειγμα, μια εταιρία που έχει δημιουργήσει κάποιες λίστες επίσης με νόμιμες διαδικασίες (π.χ. με την εγγραφή πελατών σε κάποια λίστα για περαιτέρω ενημέρωση για κάποιο περιοδικό) τις πουλάει με σκοπό το κέρδος.

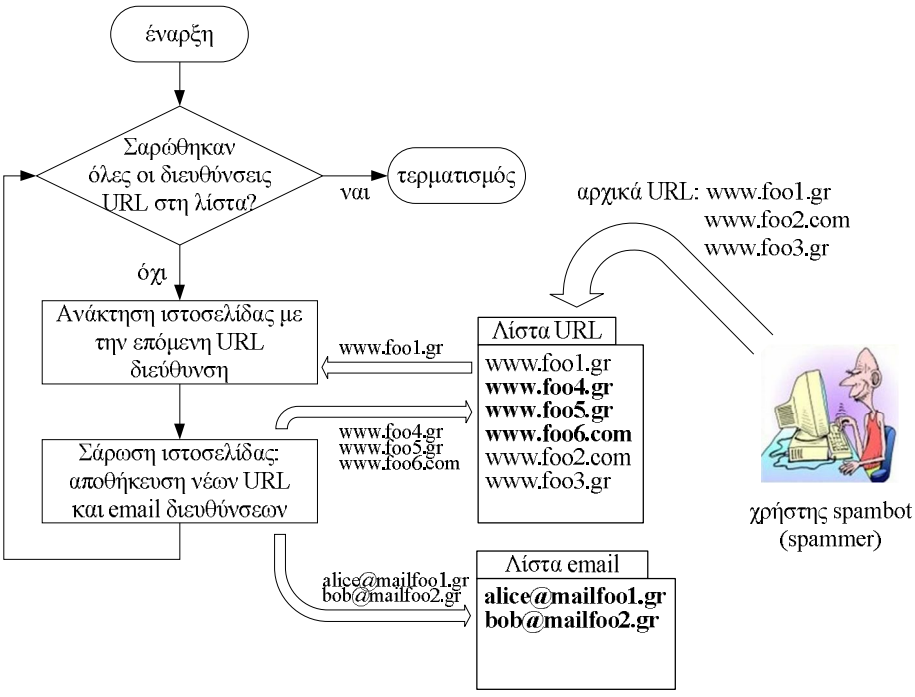
2.4.2 Εγγραφή σε Λίστες

Από παλιότερα υπήρχε η μέθοδος όπου οι spammers προσπαθούσαν να αποτελούν μέρος μιας ή περισσότερων λιστών χρηστών σε ομάδες ενημέρωσης (newsgroups) ή ανταλλαγής ιδεών και απόψεων (forum), όπως έγινε στην περίπτωση του Usenet, όπου γινόταν αποστολή μηνυμάτων σε όλους τους χρήστες που ήταν εγγεγραμμένοι στις συγκεκριμένες ομάδες. Η μέθοδος αυτή συνέχισε να εφαρμόζεται, με τους spammers τώρα να εγγράφονται σε λίστες ηλεκτρονικής αλληλογραφίας ή λίστες επικοινωνίας (mailing lists) ελεύθερες προς το κοινό. Στην περίπτωση αυτή, τα μέλη μιας λίστας στέλνουν ηλεκτρονικά μηνύματα σε όλα τα υπόλοιπα μέλη δημοσιεύοντας επίσης και την ηλεκτρονική τους διεύθυνση αλληλογραφίας, με αποτέλεσμα να είναι πολύ εύκολο για τους παραλήπτες spammers να ανακτούν από τις επικεφαλίδες των μηνυμάτων αυτών τις ηλεκτρονικές διευθύνσεις των αποστολέων. Έτσι, το μόνο που πρέπει να κάνει ένας spammer είναι αφού εγγραφεί στη λίστα να περιμένει, και δεδομένου της ιδιαίτερα υψηλής κίνησης αλληλογραφίας που υπάρχει σε πολλές τέτοιες λίστες, μέσα σε μικρό χρονικό διάστημα να συλλέξει πληθώρα ηλεκτρονικών διευθύνσεων οι οποίες μάλιστα είναι έγκυρες, αφού προέρχονται από εγγεγραμμένα πραγματικά μέλη.

2.4.3 Αυτόματη Σάρωση Ιστοσελίδων

Μια επίσης αρκετά εύκολη και αποτελεσματική μέθοδος αποτελεί η χρήση ειδικού λογισμικού το οποίο ψάχνει τον παγκόσμιο ιστό για ηλεκτρονικές διευθύνσεις αλληλογραφίας που περιέχονται στις διάφορες ιστοσελίδες, και είναι γνωστό ως *spambot*. Ουσιαστικά, το λογισμικό αυτό αποτελεί ειδική κατηγορία λογισμικού το οποίο ψάχνει τον παγκόσμιο ιστό με μεθοδικό και αυτόματο τρόπο για οτιδήποτε πληροφορίες (π.χ. χρήση μηχανών αναζήτησης για εύρεση λέξεων και φράσεων, όπως Google, Yahoo), διαδικασία η οποία είναι γνωστή ως *web crawling*. Η μέθοδος εύρεσης των ηλεκτρονικών διευθύνσεων από τα spambots είναι σχετικά απλή και βασίζεται σε μια αναδρομική διαδικασία σάρωσης διευθύνσεων του παγκόσμιου ιστού ή URL (Uniform Resource Locator), όπως φαίνεται στην Εικόνα 2.5. Τα αποτελέσματα είναι ιδιαίτερα ικανοποιητικά, λόγω του αυτοματοποιημένου τρόπου με τον οποίο τελικά σαρώνεται μεγάλος αριθμός ιστοσελίδων.

Πιο συγκεκριμένα, αρχικά, η όλη διαδικασία ξεκινάει με μια λίστα από URL συνήθως ιστοσελίδων με πολλές υπερσυνδέσεις (άλλα URL). Έτσι, καθώς το spambot επισκέπτεται τις ιστοσελίδες των URL που περιλαμβάνονται στη λίστα, εντοπίζεται στο σχετικό κώδικα σήμανσης (για παράδειγμα μέσω των ετικετών υπερσύνδεσης της HTML - HyperText Markup Language) μεγάλος αριθμός από νέα URL άλλων ιστοσελίδων, τα οποία και αποθηκεύονται στη λίστα με τα υπόλοιπα URL προς προσπέλαση (αναδρομή). Κατά τη σάρωση δε της κάθε ιστοσελίδας



Εικόνα 2.5: Διαδικασία συλλογής ηλεκτρονικών διευθύνσεων αλληλογραφίας

εντοπίζονται και οι ηλεκτρονικές διευθύνσεις αλληλογραφίας αναζητώντας στοιχεία που υποδηλώνουν την ύπαρξή τους (π.χ. το σύμβολο «@» ή την ετικέτα ορισμού μιας ηλεκτρονικής διεύθυνσης αλληλογραφίας της HTML, «mailto:»), οι οποίες στη συνέχεια αποθηκεύονται, διαμορφώνοντας τις λίστες που μπορούν να χρησιμοποιηθούν για αποστολή spam. Τελικά, η διαδικασία ολοκληρώνεται αν δεν υπάρχουν άλλα URL να προσπελασθούν στη σχετική λίστα.

2.4.4 Χρήση Κακόβουλου Λογισμικού

Πολλοί από τους spammers χρησιμοποιούν κακόβουλο λογισμικό, όπως ιούς (viruses) και email «σκουλήκια» (worms), προκειμένου να συλλέξουν ηλεκτρονικές διευθύνσεις αλληλογραφίας, δεδομένης της γρήγορης εξέλιξής τους και του γεγονότος ότι μέχρι να εντοπισθούν (να γίνουν γνωστά, ώστε να μπορούν να εντοπισθούν από κάποιο λογισμικό καταπολέμησης κακόβουλου λογισμικού - π.χ. antivirus) το χρονικό διάστημα που περνάει είναι αρκετό (λίγες εβδομάδες) ώστε να φέρουν ικανοποιητικά αποτελέσματα.

Πιο συγκεκριμένα, η λειτουργία-δράση του κακόβουλου λογισμικού συνίσταται στη δημιουργία αδυναμιών ασφαλείας του συστήματος που προσβάλλει, και στη γρήγορη μετάδοσή του και σε άλλα συστήματα πριν εντοπισθεί από κάποιο λογισμικό καταπολέμησής του. Η αδυναμίες αυτές στην ασφάλεια του συστήματος έχουν να κάνουν με τον εντοπισμό και την παρακολούθηση του βιβλίου διευθύνσεων και της δραστηριότητας της ηλεκτρονικής αλληλογραφίας. Έτσι, γίνεται ενεργοποίηση κατάλληλων θυρών επικοινωνίας, ώστε να δοθεί στον spammer πλήρης πρόσβαση για συλλογή των παρεχόμενων ηλεκτρονικών διευθύνσεων, καθώς και να γνωρίζει ποιοι είναι οι επόμενοι στόχοι που προσβάλλονται από το λογισμικό, για να τους παρακολουθήσει. Όσον αφορά δε την μετάδοση του κακόβουλου λογισμικού από σύστημα σε σύστημα, ένας τρόπος με τον οποίο επιτυγχάνεται είναι με αποστολή του μέσω της ηλεκτρονικής αλληλογραφίας σε διευθύνσεις που περιέχονται στο σχετικό βιβλίο διευθύνσεων του προσβεβλημένου από το λογισμικό κάθε φορά σύστημα.

2.4.5 SMTP Brute-Force Επιθέσεις

Δεδομένου ότι το πρωτόκολλο SMTP που χρησιμοποιείται στο ηλεκτρονικό ταχυδρομείο είναι ιδιαίτερα απλό, υπάρχει μια μέθοδος που χρησιμοποιούν οι spammers η οποία βασίζεται στην έλλειψη μηχανισμών πιστοποίησης του αποστολέα (authentication mechanisms). Πιο συγκεκριμένα, με βάση το πρωτόκολλο SMTP, υπάρχουν μόνο βασικές εντολές ανταλλαγής

μηνυμάτων μεταξύ δύο υπολογιστών του Διαδικτύου (hosts) για τη μεταφορά της ηλεκτρονικής αλληλογραφίας, όπως HELO, MAIL FROM, RCPT TO και DATA, με αποτέλεσμα ένας spammer να μπορεί να στείλει κάποιο ηλεκτρονικό μήνυμα σε οποιονδήποτε παραλήπτη, αρκεί βέβαια να διαθέτει την ηλεκτρονική διεύθυνση αλληλογραφίας του. Έτσι, το μόνο που πρέπει να κάνει ο spammer, δεδομένου ότι γνωρίζει ή «μαντεύει» σωστά το όνομα τομέα (domain name - π.χ. «yahoo.gr», «hotmail.com») του εξυπηρετητή ηλεκτρονικού ταχυδρομείου ώστε να μπορεί να προσδιοριστεί η διεύθυνση IP του τελευταίου με το σύστημα ονομάτων τομέα ή DNS (Domain Name System), είναι να δοκιμάζει διάφορους συνδυασμούς αλφαριθμητικών ως πρόθεμα στην μορφή της ηλεκτρονικής διεύθυνσης <χρήστης>@<όνομα τομέα> και να κάνει πολλαπλή αποστολή ενός μηνύματος. Ο συγκεκριμένη μέθοδος είναι ιδιαίτερα αποτελεσματική για περιπτώσεις ονομάτων τομέα όπου υπάρχει πληθώρα ηλεκτρονικών διευθύνσεων, όπως στην περίπτωση των Google, Yahoo και Hotmail, οι οποίοι είναι παροχείς δωρεάν υπηρεσιών ηλεκτρονικού ταχυδρομείου.

Συνήθως, το λογισμικό που χρησιμοποιείται για την σύνταξη των ηλεκτρονικών διευθύνσεων και αποστολή σε αυτές μηνυμάτων βασίζεται σε λίστες ονοματεπωνύμων. Έτσι, συνδυάζονται τα ονόματα και τα επώνυμα ανθρώπων σε διάφορες συνήθεις μορφές, όπως <όνομα>.<επώνυμο>@<όνομα τομέα>, <1ο γράμμα ονόματος>.<επώνυμο>@<όνομα τομέα> κ.α.. Όσον αφορά δε την εξακρίβωση για το αν μια ηλεκτρονική διεύθυνση είναι έγκυρη, οπότε και καταχωρείται στη λίστα των ηλεκτρονικών διευθύνσεων για περαιτέρω αποστολή spam, δύο τρόποι με τους οποίους αυτή γίνεται είναι:

- Χρήση εντολών του SMTP πρωτοκόλλου, όπως «Return-Receipt-To:» και «X-Confirm-Reading-To:» για αίτηση αποδεικτικού μηνύματος παράδοσης ή ανάγνωσης του απεσταλμένου μηνύματος από τον εξυπηρετητή ή από το λογισμικό πελάτη του παραλήπτη, αντίστοιχα.

- Ενσωμάτωση HTML κώδικα στο περιεχόμενο του αποστελλόμενου μηνύματος, ο οποίος περιέχει υπερσύνδεση προς εφαρμογή (π.χ. CGI - Common Gateway Interface) επιβεβαίωσης ανάγνωσης μηνύματος και καταχώρησης της ηλεκτρονικής διεύθυνσης στην οποία στάλθηκε το μήνυμα. Έτσι, σε περίπτωση που ο παραλήπτης διαβάσει το μήνυμα και το λογισμικό πελάτη εκτελέσει την επισυναπτόμενη εφαρμογή (συνήθως εμφανιζόμενη ως ενσωματωμένη εικόνα) μέσω του παρεχόμενου URL, ο spammer μπορεί να επιβεβαιώσει ελέγχοντας τη λίστα καταχωρήσεων που διατηρείται από τον

σχετικό εξυπηρετητή αν για κάποια ηλεκτρονική διεύθυνση έγινε ανάγνωση του spam, όποτε και είναι έγκυρη.

Σε περίπτωση, βέβαια, που ο spammer λάβει μήνυμα σφάλματος από την αποστολή μηνύματος σε κάποια ηλεκτρονική διεύθυνση, η τελευταία απορρίπτεται ως μη έγκυρη. Η όλη διαδικασία που περιγράφθηκε ονομάζεται διαδικασία *brute force*.

2.4.6 Τρόποι Αποστολής Spam

Τόσο οι χρήστες όσο και οι διαχειριστές συστημάτων του Διαδικτύου χρησιμοποιούν πληθώρα τρόπων, ώστε είτε να παρεμποδίζονται τα spam μηνύματα από το να φτάνουν στο χώρο αποθήκευσης μηνυμάτων (mailbox) των διαφόρων ηλεκτρονικών διευθύνσεων αλληλογραφίας είτε να φιλτράρονται και να ξεχωρίζουν από τα υπόλοιπα επιθυμητά μηνύματα. Παρόμοια, και οι παροχείς πρόσβασης στο Διαδίκτυο (ISPs) δεν επιτρέπουν τη χρήση των υπηρεσιών και πόρων τους για αποστολή μηνυμάτων spam. Παρόλα αυτά δε, οι spammers, χρησιμοποιώντας διάφορες μεθόδους, εξακολουθούν να αποστέλλουν επιτυχώς μηνύματα σε χρήστες που δεν επιθυμούν να τα λάβουν καθώς και να χρησιμοποιούν υπηρεσίες και πόρους των παροχέων, αν και οι τελευταίοι το απαγορεύουν. Στόχος των spammers, οι οποίοι θέλουν να αποκρύψουν την ταυτότητά τους, είναι η χρήση της ανωνυμίας και η δημιουργία δυσχέρειας στον εντοπισμό σχετιζόμενων με αυτούς στοιχείων, κρατώντας, βέβαια, το κόστος όσο το δυνατό πιο χαμηλό. Μερικές από τις μεθόδους αποστολής μηνυμάτων spam που χρησιμοποιούνται είναι οι εξής:

- *Χρήση πολλαπλών λογαριασμών ηλεκτρονικού ταχυδρομείου*: Μια πολύ συνηθισμένη τεχνική που ακολουθείται από τους spammers είναι η δημιουργία πολλαπλών λογαριασμών σε παροχείς δωρεάν υπηρεσιών ηλεκτρονικού ταχυδρομείου (π.χ. Google, Hotmail) έτσι ώστε είτε να αποστέλλουν μηνύματα spam είτε να λαμβάνουν απαντητικά μηνύματα από εν δυνάμει πελάτες ή ενδιαφερόμενους γενικά παραλήπτες. Για να εξασφαλίσουν την αποστολή όσο το δυνατόν περισσότερων μηνυμάτων, ο αριθμός των λογαριασμών αυτών που θα δημιουργηθούν θα πρέπει να είναι αρκετά μεγάλος, πράγμα το οποίο γίνεται εφικτό με τη χρήση ειδικού λογισμικού (web bots) το οποίο αυτοματοποιεί και επιταχύνει τη συγκεκριμένη διαδικασία. Παρά δε το γεγονός ότι έχουν υιοθετηθεί από τους παροχείς ηλεκτρονικού ταχυδρομείου διάφορα μέτρα για την αντιμετώπιση της συγκεκριμένης μεθόδου των spammers (π.χ. χρήση γραφικών ή ήχου - captcha), οι τελευταίοι βρίσκουν νέους τρόπους να τα αποφεύγουν (π.χ. χρήση τεχνικών επεξεργασίας εικόνας ή ήχου).

- *Χρήση δυναμικών διευθύνσεων Διαδικτύου (IP)*: Ένας εύκολος τρόπος αποστολής πολλαπλών μηνυμάτων από τους spammers είναι η χρήση εξυπηρετητών ηλεκτρονικού ταχυδρομείου οι οποίοι συνδέονται δυναμικά στο Διαδίκτυο μέσω modem (π.χ. DSL, dial-up). Πιο συγκεκριμένα, για κάθε σχετική σύνδεση που δημιουργείται μεταξύ του modem και ενός παροχέα πρόσβασης στο Διαδίκτυο (ISP) εξασφαλίζεται ότι ο spammer θα έχει στη διάθεσή του μια νέα IP διεύθυνση μεταξύ πολλών άλλων που διαθέτει ο παροχέας για αυτό το σκοπό. Έτσι, είναι πολύ δύσκολο να εντοπισθεί η ταυτότητα του spammer και ο ίδιος να αποκλεισθεί από τον παροχέα, ενώ πολλές μέθοδοι που βασίζονται στις IP διευθύνσεις για την καταπολέμηση του spamming (π.χ. «μαύρες λίστες») κρίνονται λιγότερο αποτελεσματικές. Σαν μέτρο αντιμετώπισης του συγκεκριμένου προβλήματος, πολλοί εξυπηρετητές ηλεκτρονικού ταχυδρομείου αποκλείουν μηνύματα που προέρχονται από συνδέσεις με δυναμικές IP διευθύνσεις, καθώς θεωρούνται ότι είναι πολύ πιθανό να πρόκειται για αποστολείς spam.

- *Υπογραφή «ροζ συμβολαίων»*: Από πολύ νωρίς, οι spammers παρατήρησαν ότι η αποστολή πληθώρα μηνυμάτων χρησιμοποιώντας τους λογαριασμούς τους στους παροχείς πρόσβασης στο Διαδίκτυο προκαλούσε τον αποκλεισμό τους από τους τελευταίους ύστερα από παράπονα παραληπτών και άλλων παροχέων. Έτσι, για να μπορούν να συνεχίσουν να στέλνουν spam, πολλοί από αυτούς έρχονται σε συνεννόηση με κάποιον παροχέα ο οποίος θέλει να αυξήσει τα έσοδά του, συμφωνώντας σε ειδικές υπηρεσίες όπου επιτρέπεται η αποστολή μεγάλου αριθμού μηνυμάτων με αντίτιμο συνήθως επιπλέον χρεώσεις ανάλογα με την κίνηση που προκαλείται στο δίκτυο. Αυτού του είδους οι συμφωνίες ονομάζονται «ροζ συμβόλαια» (pink contracts).

- *Χρήση υπολογιστών τρίτων*: Παρόλο που οι spammers μπορούν να κρύψουν στοιχεία σχετιζόμενα με την ταυτότητά τους υπογράφοντας συμφωνίες (ροζ συμβόλαια) με κάποιους παροχείς πρόσβασης στο Διαδίκτυο, οι τελευταίοι βρίσκονται σε κίνδυνο να εκτεθούν στους πελάτες τους, για αυτό και λίγοι είναι εκείνοι που το επιχειρούν. Έτσι οι spammers έχουν στραφεί σε μεθόδους αποστολής μηνυμάτων από υπολογιστές τρίτων που είναι συνδεδεμένοι στο Διαδίκτυο εξασφαλίζοντας τόσο την ανωνυμία τους, με την δυσχέρεια που δημιουργείται στο να εντοπισθούν (π.χ. δεν εμφανίζεται πουθενά η διεύθυνση IP που έχει ανατεθεί στους προσωπικούς υπολογιστές που χρησιμοποιούν), όσο και την αυτοματοποιημένη και σε μεγάλη κλίμακα αποστολή μηνυμάτων spam. Κατά ένα μέρος αυτό επιτυγχάνεται με τη χρήση των open proxies, οι οποίοι είναι υπολογιστές-εξυπηρετητές που επιτρέπουν την ελεύθερη δημιουργία οποιωνδήποτε

έμμεσων συνδέσεων σε μια επικοινωνία πελάτη-εξυπηρετητή. Πιο συγκεκριμένα, ο spammer συνδέεται με έναν open proxy εξυπηρετητή στον οποίο δίνει εντολή να συνδεθεί με έναν εξυπηρετητή ηλεκτρονικού ταχυδρομείου, ώστε να μπορεί έπειτα να στείλει μηνύματα spam χωρίς να φαίνεται η διεύθυνση IP του υπολογιστή που χρησιμοποιεί. Παρόλα αυτά, βέβαια, δεδομένου ότι οι IP διευθύνσεις των open proxy εξυπηρετητών είναι εμφανείς, μέθοδοι αντιμετώπισης του spam, όπως οι «μαύρες λίστες», περιόρισαν τη συγκεκριμένη διέξοδο των spammers με τον αποκλεισμό των IP διευθύνσεων των εξυπηρετητών αυτών. Για το λόγο αυτό, κατά κύριο μέρος, οι spammers αντί να αναζητούν ελεύθερες υπηρεσίες από open proxy εξυπηρετητές, δημιουργούν ουσιαστικά τους δικούς τους proxy εξυπηρετητές χρησιμοποιώντας κακόβουλο λογισμικό το οποίο διαδίδεται από σύστημα σε σύστημα, όπως αναφέρθηκε στην υποενότητα 2.4.4. Το κάθε ένα από τα προσβεβλημένα από το λογισμικό αυτό συστήματα ονομάζεται *zombie*, και το σύνολο αυτών, τα οποία αποτελούν ένα ευρύ «δίκτυο εξυπηρετητών» διαθέσιμο προς τον spammer, *botnet*. Το botnet, εκτός του ότι εξασφαλίζει υψηλό εύρος ζώνης για την αποστολή μεγάλου αριθμού μηνυμάτων spam, δυσκολεύει ακόμα περισσότερο τον εντοπισμό του spammer με τη δημιουργία πολλαπλών ενδιάμεσων συνδέσεων από proxy εξυπηρετητές-zombies μέχρι τον τελικό εξυπηρετητή ηλεκτρονικού ταχυδρομείου.

2.4.7 Τρόποι Επιβεβαίωσης Ηλεκτρονικών Διευθύνσεων

Μετά τη συλλογή με τις διάφορες μεθόδους που παρουσιάστηκαν παραπάνω ηλεκτρονικών διευθύνσεων από τους spammers, είναι επιθυμητή η επιβεβαίωση για το αν τα μηνύματα spam που αποστέλλονται, τελικά διαβάζονται. Σε περίπτωση που συμβαίνει το αντίθετο, πολύ πιθανόν η συγκεκριμένη ηλεκτρονική διεύθυνση να μη χρησιμοποιείται πια, οπότε και δεν χρειάζεται να στέλνονται περαιτέρω μηνύματα σε αυτή. Υπάρχουν πολλοί τρόποι για να επιτευχθεί αυτή η επιβεβαίωση εκ των οποίων μερικοί είναι:

- Χρήση εντολών του SMTP πρωτοκόλλου για αίτηση αποδεικτικών παράδοσης και ανάγνωσης (βλ. Ενότητα 2.4.5).

- Αποστολή μηνύματος στην ηλεκτρονική διεύθυνση προς επιβεβαίωση και αναμονή για τυχόν μήνυμα σφάλματος, πράγμα το οποίο θα δηλώνει ότι έγινε επιτυχής αποστολή ή απώλεια μηνύματος.

- Αποστολή απαντητικού μηνύματος από τον παραλήπτη του spam στον spammer επηρεαζόμενος από το παραπλανητικό περιεχόμενο του spam.

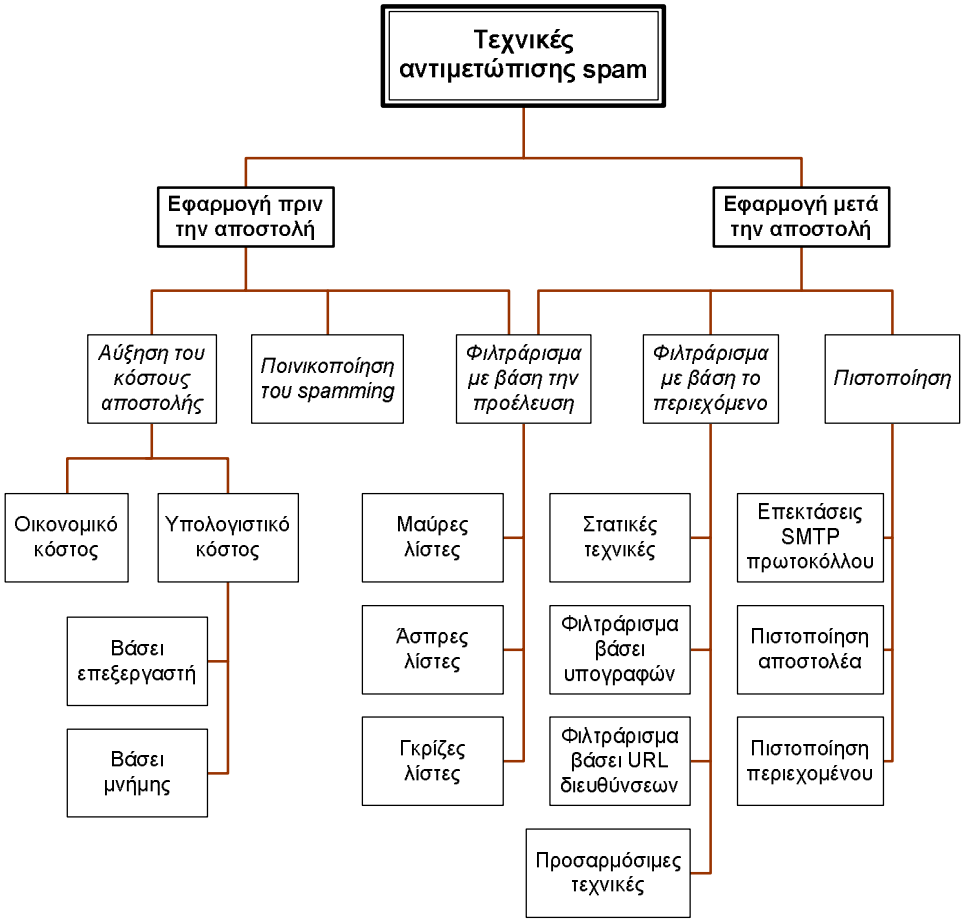
- Χρήση HTML κώδικα στο περιεχόμενο του αποστελλόμενου από τον spammer μηνύματος, περιέχοντας υπερσύνδεση προς εφαρμογή επιβεβαίωσης ανάγνωσης του μηνύματος και κατά συνέπεια ύπαρξης της σχετικής ηλεκτρονικής διεύθυνσης (βλ. Ενότητα 2.4.5).

- Χρήση κακόβουλου λογισμικού, το οποίο με την είσοδό του σε κάποιο σύστημα μέσω της ηλεκτρονικής αλληλογραφίας να ενημερώνει τον spammer παρέχοντας την ηλεκτρονική διεύθυνση από την οποία απεστάλη (βλ. Ενότητα 2.4.4).

2.5 Τεχνικές Αντιμετώπισης του Spamming

Για να αντιμετωπισθεί το spamming, τόσο οι τελικοί χρήστες-παραλήπτες όσο και οι διαχειριστές συστημάτων ηλεκτρονικού ταχυδρομείου χρησιμοποιούν διάφορες τεχνικές – *anti-spam τεχνικές* – με πολλές από αυτές να ενσωματώνονται σε προϊόντα και υπηρεσίες λογισμικού προς διευκόλυνσή τους. Δεδομένου και του μεγάλου αριθμού των τεχνικών που υπάρχουν, καμία από αυτές δεν αποτελεί από μόνη της μια ολοκληρωμένη λύση στην αντιμετώπιση του spamming, έχοντας πλεονεκτήματα και μειονεκτήματα ως προς τον σωστό διαχωρισμό των ηλεκτρονικών μηνυμάτων σε spam και σε «υγιή», καθώς και ως προς το σχετιζόμενο κόστος σε χρόνο και επεξεργασία. Το γεγονός αυτό απαιτεί οι διάφορες τεχνικές να συνδυάζονται μεταξύ τους συμπληρώνοντας η μία την άλλη στα στοιχεία και συνθήκες όπου υστερεί η καθεμιά, ώστε να επιτυγχάνεται η όσο το δυνατό πληρέστερη αντιμετώπιση του προβλήματος.

Γενικά, οι τεχνικές που χρησιμοποιούνται μπορούν να χωρισθούν σε εκείνες που εφαρμόζονται πριν αποσταλεί ένα ηλεκτρονικό μήνυμα (email) και σε εκείνες που εφαρμόζονται μετά την αποστολή του μηνύματος, περιλαμβάνοντας έτσι το συνολικό εύρος των δυνατών τρόπων αντιμετώπισης οι οποίοι βασίζονται είτε στη μείωση του αριθμού των μηνυμάτων spam που αποστέλλονται στο Διαδίκτυο είτε στην προστασία του παραλήπτη ύστερα από επεξεργασία και κατηγοριοποίηση των ληφθέντων μηνυμάτων. Όσον αφορά την πρώτη κατηγορία, αυτή ουσιαστικά περιλαμβάνει τεχνικές οι οποίες έχουν ως στόχο κυρίως την αποθάρρυνση-«πρόληψη» της αποστολής μηνυμάτων spam. Αντίθετα, οι τεχνικές της δεύτερης κατηγορίας έχουν ως στόχο περισσότερο την προστασία-«θεραπεία» από τα αποστελλόμενα μηνύματα



Εικόνα 2.6: Ταξινόμηση των τεχνικών αντιμετώπισης του spam

spam εντοπίζοντας και αποκλείοντάς τα, βασιζόμενες είτε στα στοιχεία προέλευσης (π.χ. ηλεκτρονική διεύθυνση αποστολέα) είτε στο περιεχόμενο του κάθε ληφθέντος μηνύματος. Μερικές από τις κυριότερες τεχνικές παρουσιάζονται παρακάτω και συνοψίζονται στην Εικόνα 2.6. Μια αναλυτικότερη ταξινόμηση των τεχνικών αυτών, συνοδευόμενη και από περαιτέρω ειδικές πληροφορίες, παρέχεται από την ASRG - Anti-Spam Research Group [5] (μέλος της IRTF - Internet Research Task Force [44]).

2.5.1 Αύξηση του Κόστους Αποστολής Μηνυμάτων

Η συγκεκριμένη τεχνική βασίζεται στην επιβολή κάποιας μορφής επιβάρυνσης των spammers στην αποστολή πολλαπλών ηλεκτρονικών μηνυμάτων είτε αυτή σχετίζεται με το χρόνο και τη χρήση πόρων του υλικού είτε με το οικονομικό κόστος της αποστολής (μπορεί βέβαια να επεκταθεί και σε άλλα είδη κόστους).

Όσον αφορά την *οικονομική επιβάρυνση*, αυτή περιλαμβάνει την κοστολόγηση κάποιου που θέλει να στείλει ένα μήνυμα αλληλογραφίας, υιοθετώντας διάφορους τρόπους, όπως

περιγράφονται στη σχετική βιβλιογραφία [6, 53, 99]. Έτσι, δεδομένου του μεγάλου αριθμού μηνυμάτων που στέλνονται από κάποιον spammer, το παραγόμενο κόστος είναι αποθαρρυντικό για τον τελευταίο, ενώ για τους νόμιμους χρήστες, η σχετική οικονομική επιβάρυνση είναι πολύ μικρή. Ωστόσο, με την επιβολή οικονομικής επιβάρυνσης στην αποστολή μηνυμάτων αλληλογραφίας, μπορούν τελικά να αποθαρρύνονται και οι νόμιμοι χρήστες της υπηρεσίας του ηλεκτρονικού ταχυδρομείου, καθώς επίσης υπάρχει και το πρόβλημα ότι οι σχετικοί μηχανισμοί που απαιτούνται να υλοποιηθούν είναι σχετικά πολύπλοκοι και υπάρχει περαιτέρω επιβάρυνση από πλευράς διαχείρισης.

Σχετικά με το δεύτερο τρόπο επιβάρυνσης, ο οποίος είναι πιο τεχνικός, όπως και στην περίπτωση της οικονομικής επιβάρυνσης, σκοπός είναι κατά κάποιο τρόπο να «πληρώσει» ο spammer για την αποστολή μηνυμάτων, μόνο που αυτή τη φορά η σχετική «πληρωμή» έγκειται σε έναν υπολογισμό κάποιας μέτρια πολύπλοκης συνάρτησης – *συνάρτηση τιμολόγησης* (pricing function) – πριν την αποστολή κάθε μηνύματος. Με αυτόν τον τρόπο επιβάλλεται τελικά μια μορφή καθυστέρησης. Δεδομένου ότι γενικά το ηλεκτρονικό ταχυδρομείο δεν αποτελεί μέσο επικοινωνίας πραγματικού-χρόνου (real-time), για τον μέσο νόμιμο χρήστη ο οποίος στέλνει μερικές δεκάδες μηνύματα αλληλογραφίας την ημέρα, η σχετική επιβαλλόμενη καθυστέρηση δεν είναι πολύ σημαντική. Αντίθετα, για τον spammer, ο χρόνος που απαιτείται για κάθε μήνυμα είναι πολύ σημαντικός, αφού ουσιαστικά μειώνει τον αριθμό των παραληπτών spam στο χρόνο.

Σε σχέση με την οικονομική επιβάρυνση, η *υπολογιστική επιβάρυνση* (POW – Proof-of-work), όπως θα μπορούσε να ονομασθεί ο δεύτερος τρόπος αύξησης του κόστους αποστολής μηνυμάτων αλληλογραφίας, έχει το πλεονέκτημα ότι οι σχετικοί μηχανισμοί που απαιτούνται είναι πιο εύκολα υλοποιήσιμοι, μιας και απαιτείται η ενσωμάτωση στο σύστημα μιας συνάρτησης τιμολόγησης η οποία βασίζεται στο υλικό (hardware – π.χ. επεξεργαστής, μνήμη). Το γεγονός αυτό, βέβαια, αποτελεί και τη βάση του κύριου μειονεκτήματος του συγκεκριμένου τρόπου επιβάρυνσης, δεδομένου ότι υπάρχει ασυμμετρία του συμμετέχοντος στο σύστημα του ηλεκτρονικού ταχυδρομείου υλικού.

Πιο αναλυτικά, δύο κύριοι τρόποι «τιμολόγησης» αποτελούν η χρήση *βασιζόμενων στον επεξεργαστή* (CPU-based pricing functions) [10, 25] και *βασιζόμενων στην μνήμη* (Memory-bound pricing functions) [1, 16, 24] συναρτήσεων τιμολόγησης, αντίστοιχα. Όσον αφορά τον πρώτο, αυτός περιλαμβάνει μόνο τη χρήση του επεξεργαστή για τον απαιτούμενο υπολογισμό πριν την αποστολή κάποιου μηνύματος. Οπότε, δεδομένης της ραγδαίας εξέλιξης των επεξεργαστών τα τελευταία χρόνια, υπάρχει μεγάλη ποικιλία ταχυτήτων για διαφορετικές

συσκευές (π.χ. προσωπικοί υπολογιστές, κινητά τηλέφωνα) με αποτέλεσμα μια σχετικά περίπλοκη συνάρτηση τιμολόγησης για μια συσκευή να είναι σχετικά απλή για κάποια άλλη. Από την άλλη μεριά δε, η εξέλιξη της ταχύτητας της μνήμης δεν ήταν τόσο ραγδαία σαν αυτή του επεξεργαστή, με αποτέλεσμα η ποικιλία της ταχυτήτων της μνήμης των συσκευών που μπορούν να συμμετέχουν στο σύστημα του ηλεκτρονικού ταχυδρομείου να μην είναι τόσο εκτενής. Αυτό μπορεί να οδηγήσει σε επιλογή πιο αποδοτικής συνάρτησης τιμολόγησης που να καλύπτει το μεγαλύτερο εύρος των ταχυτήτων μνήμης, αντιμετωπίζοντας καλύτερα περιπτώσεις όπου οι spammers διαθέτουν το καλύτερο δυνατό σε απόδοση υλικό.

Γενικά, η τεχνική της αύξησης του κόστους αποστολής των μηνυμάτων αλληλογραφίας είτε με τον έναν είτε με τον άλλον τρόπο, για να είναι επιτυχής, θα πρέπει να υπάρχει κάποιος συντονισμός μεταξύ των παροχέων υπηρεσιών ηλεκτρονικού ταχυδρομείου στην εφαρμογή της. Σε αντίθετη περίπτωση, οι spammers μπορούν να αποφεύγουν τη σχετική επιβάρυνση που επιβάλλεται από κάποιους παροχείς απλά επιλέγοντας εκείνους που δεν υιοθετούν τέτοιες τεχνικές.

2.5.2 Ποινικοποίηση του Spamming

Σε αντίθεση με την επιβολή επιβάρυνσης στην αποστολή ηλεκτρονικών μηνυμάτων, η οποία «πλήττει» τόσο τους spammers όσο και τους νόμιμους χρήστες του ηλεκτρονικού ταχυδρομείου, μια πιο συγκεντρωτική ως προς το στόχο μέθοδος αντιμετώπισης του spamming αποτελεί η επιβολή νομικών μέτρων κατά της δραστηριότητας των spammers. Τέτοια μέτρα συνιστούν τη θέσπιση νέων νομικών διατάξεων, καθώς και την ανάπτυξη σχετικής υποδομής και διαδικασιών για την εφαρμογή τους. Λεπτομέρειες σχετικά με τα μέτρα αυτά, όπως ορίζουν οι νομοθεσίες της Ευρώπης και των Ηνωμένων Πολιτειών Αμερικής, παρουσιάζονται στην Ενότητα 2.6.

Δεδομένων των στατιστικών που προκύπτουν για την κατάσταση της αποστολής spam μηνυμάτων ανά τον κόσμο (περίπου το 90% των αποστελλόμενων μηνυμάτων ηλεκτρονικής αλληλογραφίας – βλ. Εικόνα 2.2), καθώς και του γεγονότος ότι σχετικές νομοθεσίες έχουν εδώ και κάποια χρόνια αναπτυχθεί, συμπεραίνεται ότι οι τελευταίες θέτουν απλά ένα εμπόδιο στους spammers, χωρίς να είναι σε θέση να αντιμετωπίσουν πλήρως το πρόβλημα. Ουσιαστικά, οι spammers δεν επηρεάζονται τόσο πολύ από τις πιθανές επιπτώσεις που μπορούν να έχουν από την νομοθεσία, αφού μπορούν να αποκρύψουν εύκολα την ταυτότητά τους ή ακόμα και να μεταφέρουν την δραστηριότητά τους σε άλλες χώρες (αποστολή μηνυμάτων από αυτές) όπου δεν υπάρχει σχετικό νομικό πλαίσιο. Έτσι, όπως και στην περίπτωση της τεχνικής αύξησης του

κόστους αποστολής ηλεκτρονικών μηνυμάτων, το αποτέλεσμα της συγκεκριμένης τεχνικής βασίζεται σε μεγάλο βαθμό (ακόμα μεγαλύτερο) στον συντονισμό και συνεργασία μεταξύ των κρατών ανά τον κόσμο.

2.5.3 Φιλτράρισμα των Μηνυμάτων με Βάση το Περιεχόμενό τους

Όταν επιτευχθεί η σύνδεση με το πρωτόκολλο SMTP και το ηλεκτρονικό μήνυμα αλληλογραφίας παραδοθεί στον εξυπηρετητή ηλεκτρονικού ταχυδρομείου (mailbox παραλήπτη) και στη συνέχεια στον τελικό παραλήπτη μέσω ενός πρωτοκόλλου πρόσβασης (π.χ. POP3), το περιεχόμενο του μηνύματος μπορεί να σαρωθεί και αναλυθεί ψάχνοντας για στοιχεία που γενικότερα προσδιορίζουν τα spam μηνύματα. Μερικές από τις σχετικές τεχνικές εφαρμόζονται κατά τη διάρκεια της παράδοσης ενός μηνύματος, αλλά γενικά οι έλεγχοι που χρειάζεται να γίνουν απαιτούν αρκετό χρόνο για να μπορούν να ολοκληρωθούν έγκαιρα. Οι μέθοδοι που χρησιμοποιούνται ποικίλουν, με τις πιο απλές να περιλαμβάνουν ταίριασμα λέξεων και φράσεων με συγκεκριμένα πρότυπα που χαρακτηρίζουν τα spam μηνύματα, και με τις πιο σύνθετες να στηρίζονται σε πιο «έξυπνους» τρόπους (εξόρυξη γνώσης - artificial intelligence) για το φιλτράρισμα των μηνυμάτων. Μερικές από τις κατηγορίες των τεχνικών αυτών είναι οι στατικές τεχνικές, οι τεχνικές που βασίζονται σε «υπογραφές» των μηνυμάτων, οι τεχνικές που βασίζονται στο φιλτράρισμα των συμπεριλαμβανομένων διευθύνσεων URL, και οι προσαρμόσιμες τεχνικές.

Στατικές Τεχνικές

Οι συγκεκριμένες τεχνικές, για το φιλτράρισμα της ηλεκτρονικής αλληλογραφίας, βασίζονται στην αναζήτηση στο περιεχόμενο του κάθε ληφθέντος μηνύματος, είτε στο σώμα είτε στην επικεφαλίδα, για λέξεις ή φράσεις οι οποίες εμφανίζονται σε πληθώρα μηνυμάτων spam ενώ σπάνια εμφανίζονται σε συνηθισμένα νόμιμα μηνύματα. Τέτοιες λέξεις-φράσεις συνήθως αφορούν συγκεκριμένα φάρμακα, δωρεάν προσφορές προϊόντων και άλλα, όπως «Viagra» και «get for free». Αν εντοπισθεί κάποιο από αυτά, χρησιμοποιείται ανάλογα με την κάθε τεχνική ως ένδειξη για την πιθανότητα ένα ληφθέν μήνυμα να είναι spam. Τρεις κύριοι τύποι τεχνικών που υιοθετούν τη συγκεκριμένη μέθοδο είναι:

- *Βασιζόμενες σε μεμονωμένες λέξεις-φράσεις κλειδιά* (keywords) τεχνικές, όπου η αναζήτηση απαιτεί την ακριβή ταύτιση των λέξεων-φράσεων που συναντώνται στο περιεχόμενο ενός μηνύματος με εκείνες που είναι αποθηκευμένες σε μια σχετική λίστα,

ώστε να χαρακτηρισθεί τελικά το μήνυμα ως spam. Για παράδειγμα, αν υπάρχει στη σχετική λίστα η λέξη «Viagra», θα πρέπει και κατά την αναζήτηση να βρεθεί η μεμονωμένη λέξη «Viagra» ως έχει, χωρίς τυχόν μορφοποιήσεις, όπως «Vi@gra» και «Vaigra», οι οποίες αποτελούν και τέχνασμα των spammers για αντιμετώπιση της συγκεκριμένης τεχνικής.

- *Βασιζόμενες σε πρότυπα αντιστοίχησης λέξεων-φράσεων* (pattern matching) τεχνικές, όπου στην αναζήτηση περιλαμβάνονται εκτός από μεμονωμένες λέξεις-φράσεις, και παραλλαγές αυτών οι οποίες παράγονται με διάφορους τρόπους μορφοποίησης κειμένου, όπως μη διάκριση πεζών-κεφαλαίων γραμμάτων (π.χ. «viagra», «Viagra»), παράβλεψη επαναλαμβανόμενων γραμμάτων και συμβόλων (π.χ. «viagra», «vviaggraaa») κ.α.. Έτσι, αντιμετωπίζονται τεχνάσματα τροποποίησης των λέξεων-φράσεων κλειδιών από τους spammers, αλλά παραμένει η μεγάλη πιθανότητα εσφαλμένου χαρακτηρισμού των νόμιμων μηνυμάτων ως spam.

- *Βασιζόμενες σε κανόνες* (rules) τεχνικές, όπου κάθε κανόνας, ο οποίος χρησιμοποιείται κατά τη διαδικασία αναζήτησης λέξεων-φράσεων στο περιεχόμενο ενός ηλεκτρονικού μηνύματος, ορίζει μια σύνθετη σύνταξη βασικών λέξεων-φράσεων κλειδιών (regular expressions [34]), επεκτείνοντας τις δύο προηγούμενες μεθόδους. Για παράδειγμα, μια τέτοια σύνταξη αποτελεί το κατά πόσο μια λέξη σαν τη «Viagra» είναι το κύριο θέμα σε ένα συγκεκριμένο μήνυμα. Για το λόγο αυτό, χρησιμοποιείται κάποιο σύστημα βαθμολόγησης του κάθε κανόνα (π.χ. καταμέτρηση εμφανίσεων, ταίριασμα σε πρότυπα), και το σύνολο των βαθμολογιών ορίζει τελικά το κατά πόσο ένα μήνυμα είναι spam, με τις πιο υψηλές βαθμολογίες να ενισχύουν αυτήν την πιθανότητα. Παρά δε την μείωση των εσφαλμένων εκτιμήσεων όσον αφορά το χαρακτηρισμό των μηνυμάτων ως spam, εξακολουθεί να υπάρχει το πρόβλημα της συχνής ενημέρωσης μιας εκτενούς λίστας λέξεων-φράσεων κλειδιών. Αυτό, διότι δεδομένου του στατικού χαρακτήρα των τεχνικών αυτών, οι spammers εφευρίσκουν νέες εκφράσεις ώστε να μπορούν να διατηρούν τις σχετικές βαθμολογίες σε χαμηλά επίπεδα.

Φιλτράρισμα με Βάση τις Υπογραφές των Μηνυμάτων

Η αποθήκευση του συνόλου των περιεχομένων μηνυμάτων spam για σύγκριση με τα ληφθέντα κάθε φορά μηνύματα, ώστε να εντοπισθούν ποια από τα τελευταία είναι spam, απαιτεί πολύ αποθηκευτικό χώρο και χρόνο επεξεργασίας. Για το λόγο αυτό, οι συγκεκριμένες τεχνικές

περιλαμβάνουν το φιλτράρισμα-κατηγοριοποίηση των ηλεκτρονικών μηνυμάτων με βάση το ταίριασμα των *υπογραφών* τους (υπολογιζόμενες αριθμητικές τιμές που απαιτούν κάποια bytes χώρου αποθήκευσης) με εκείνες γνωστών μηνυμάτων spam (signature-based filtering), οι οποίες μπορούν να είναι διαφόρων τύπων, όπως:

- *Ψηφιακό ίχνος* (digital fingerprint), το οποίο αποτελεί μια τιμή που προκύπτει από την εφαρμογή ενός αλγόριθμου κατακερματισμού (hash algorithm) στο περιεχόμενο του ληφθέντος μηνύματος, όπως MD5 [87] και SHA [26, 27].

- *Άθροισμα ελέγχου* (checksum), το οποίο αποτελεί μια τιμή που προκύπτει από το άθροισμα όλων των αριθμών που δίνονται ως είσοδος. Ουσιαστικά, είναι μια πιο απλή μορφή του ψηφιακού ίχνους, όπου, για παράδειγμα, ακόμα και με την αναδιάταξη των αριθμών που δίνονται ως είσοδος, η τιμή checksum παραμένει η ίδια.

- *Κυκλικός έλεγχος πλεονασμού* (Cyclic Redundancy Check - CRC), το οποίο αποτελεί έναν τύπο αθροίσματος ελέγχου αλλά βασισμένο σε κάποιο πολυώνυμο. Ο τύπος υπογραφής αυτός είναι πιο αξιόπιστος από το απλό άθροισμα ελέγχου, ικανός να διακρίνει ακόμα και πολύ μικρές αλλαγές στα δεδομένα που δίνονται ως είσοδος, αλλά παραμένει σχετικά εύκολο να παραχθεί η ίδια τιμή με διαφορετικά δεδομένα.

Γενικά, οι τεχνικές φιλτραρίσματος με υπογραφές περιλαμβάνουν ουσιαστικά την παρακολούθηση της συχνότητας με την οποία εμφανίζονται σε ένα σύνολο χρηστών ηλεκτρονικού ταχυδρομείου ενός παροχέα τα ληφθέντα μηνύματα, βασιζόμενες σε ένα σύστημα πελάτη-εξυπηρετητή. Έτσι, ο κάθε χρήστης υπολογίζει την υπογραφή ενός ληφθέντος μηνύματος, τη στέλνει στον εξυπηρετητή που ψάχνει για τυχόν ταιριάσματα με προηγούμενα μηνύματα των οποίων οι υπογραφές έχουν σταλθεί από το σύνολο των χρηστών, και τέλος επιστρέφεται στο χρήστη ο αριθμός των προηγούμενων εμφανίσεων του συγκεκριμένου μηνύματος (της υπογραφής του). Ανάλογα δε με τα κριτήρια του κάθε χρήστη, ο συγκεκριμένος αριθμός-μετρική καθορίζει αν το ληφθέν ηλεκτρονικό μήνυμα είναι spam.

Παραδείγματα εφαρμογής τέτοιων τεχνικών αποτελούν η μέθοδος Distributed Checksum Clearinghouse – DCC [86] και η Vipul's Razor [81], οι οποίες διαφέρουν στον τρόπο με τον οποίο οι χρήστες αναφέρουν στον εξυπηρετητή τα ληφθέντα μηνύματα. Η πρώτη μέθοδος περιλαμβάνει την αναφορά οποιουδήποτε μηνύματος λαμβάνεται από κάποιον χρήστη αποστέλλοντας τις σχετικές υπογραφές. Το μειονέκτημά της είναι ότι δεδομένου του γεγονότος

ότι μπορεί να υπάρχουν και επιθυμητοί αποστολείς οι οποίοι στέλνουν μεγάλο αριθμό μηνυμάτων, δεν μπορεί να διακριθεί από την υπολογιζόμενη συχνότητα εμφάνισης αν ένα μήνυμα είναι spam ή όχι. Για το λόγο αυτό, απαιτείται από τους χρήστες η διατήρηση σχετικών λιστών αποστολέων (whitelists) οι οποίες να διακρίνουν τέτοιες περιπτώσεις. Όσον αφορά δε τη μέθοδο Vipul's Razor (καθώς και την εκδοχή ανοιχτού λογισμικού Pyzor [97]), ο χρήστης συμμετέχει πιο ενεργά με το να αναφέρει στον εξυπηρετητή μόνο μηνύματα τα οποία ο ίδιος θεωρεί ότι είναι spam. Έτσι, στη σχετική βάση που διατηρεί ο εξυπηρετητής περιέχονται μόνο υπογραφές μηνυμάτων τα οποία έχουν αναγνωρισθεί ως spam. Για να ενισχυθεί δε η αξιοπιστία του συστήματος, υιοθετείται ένας μηχανισμός ο οποίος μετράει την αντίστοιχη αξιοπιστία του κάθε χρήστη όσον αφορά τις αναφορές του για τα ληφθέντα από αυτόν μηνύματα.

Φιλτράρισμα με Βάση τις Συμπεριλαμβανόμενες Διευθύνσεις URL

Δεδομένου ότι τα περισσότερα μηνύματα spam περιέχουν διευθύνσεις URL στο περιεχόμενό τους οι οποίες παραπέμπουν τον παραλήπτη σε σχετικές με τους spammers ιστοσελίδες, οι συγκεκριμένες τεχνικές βασίζονται στο εντοπισμό αυτών των διευθύνσεων ώστε να κατηγοριοποιήσουν τα ληφθέντα μηνύματα. Ουσιαστικά, αποτελούν μια μορφή δημιουργίας μαύρης λίστας διευθύνσεων URL οι οποίες διατηρούνται σε μια βάση δεδομένων και συγκρίνονται με εκείνες που συμπεριλαμβάνονται σε ένα ληφθέν μήνυμα, ώστε να προκύψει τελικά μια ενδεικτική τιμή που θα δηλώνει το κατά πόσο αφορά περίπτωση spam (όπως και στις στατικές τεχνικές). Τα αποτελέσματα φαίνονται να είναι πολλές φορές καλύτερα από άλλες παρόμοιες τεχνικές [36], ενώ υπάρχουν και υλοποιήσεις οι οποίες επεκτείνουν τη συγκεκριμένη μέθοδο με το συνδυασμό διαφόρων τεχνικών ανάλυσης των URL διευθύνσεων [52].

Προσαρμόσιμες Τεχνικές

Πολλές από τις προηγούμενες τεχνικές περιλαμβάνουν τη χρήση στατικών λιστών-κανόνων, οι οποίες είναι αποθηκευμένες σε μια βάση δεδομένων, ώστε να επιτυγχάνεται ο εντοπισμός στοιχείων του περιεχόμενου των λαμβανόμενων μηνυμάτων που ταιριάζουν με εκείνα των spam. Το βασικό μειονέκτημά τους είναι ότι η κατηγοριοποίηση μεταξύ spam και νόμιμων-επιθυμητών μηνυμάτων γίνεται σχεδόν κατά απόλυτο τρόπο όσον αφορά το ταίριασμα των εν λόγω στοιχείων, οδηγώντας πολλές φορές σε λανθασμένες αποφάσεις (π.χ. για τους κοινούς χρήστες – οι οποίοι είναι και οι περισσότεροι – η λέξη «Viagra» στο περιεχόμενο ενός ληφθέντος μηνύματος αποτελεί στοιχείο spam, ενώ για έναν βιοχημικό μπορεί να αποτελέσει κύριο θέμα της αλληλογραφίας του, με αποτέλεσμα εσφαλμένα να χαρακτηρισθούν κάποια μηνύματα του

τελευταίου ως spam). Για το λόγο αυτό, αναπτύχθηκαν οι προσαρμόσιμες (ή στατιστικές) τεχνικές οι οποίες χαρακτηρίζουν ένα ληφθέν μήνυμα ως spam βασιζόμενες: α) στην ομοιότητα του συνόλου του περιεχομένου του με εκείνο των προηγουμένως ληφθέντων μηνυμάτων spam, και β) στην καθοδήγηση από τους εμπλεκόμενους χρήστες του συστήματος (το οποίο υλοποιεί την προσαρμόσιμη τεχνική). Ουσιαστικά, οι τεχνικές αυτές προσαρμόζουν τα στοιχεία που χρησιμοποιούν κατά τον εντοπισμό των spam στις σχετικές αναφορές των χρηστών – «μαθαίνουν» από τους χρήστες, οπότε και από την στιγμή που θα στηθεί ένα τέτοιο σύστημα να μην χρειάζεται περαιτέρω εκτενείς ενέργειες διατήρησης.

Μια από τις πιο διαδεδομένες προσαρμόσιμες τεχνικές, και κατά συνέπεια αντιπροσωπευτική αυτών, είναι η χρήση του Bayesian φιλτραρίσματος (Bayesian filtering) [4, 13, 22]. Τα φίλτρα Bayes κατηγοριοποιούν τα ληφθέντα ηλεκτρονικά μηνύματα με βάση τις λέξεις-αλφαριθμητικά που εμφανίζονται στο περιεχόμενό τους, από τα οποία υπολογίζεται η πιθανότητα να πρόκειται για μηνύματα spam. Τα αλφαριθμητικά αυτά ονομάζονται *κουπόνια* (tokens). Ο υπολογισμός δε της σχετικής (spam) πιθανότητας στηρίζεται κατά μεγάλο ποσοστό στη «μάθηση» (training) που εφαρμόζεται στα φίλτρα Bayes χρησιμοποιώντας δύο σύνολα μηνυμάτων: ένα που να περιέχει μόνο spam και ένα μόνο νόμιμα-επιθυμητά μηνύματα. Με βάση αυτά τα δύο σύνολα αρχικοποιούνται δύο λίστες, με καθεμιά να αποτελείται από ένα συγκεκριμένο αριθμό κουπονιών τα οποία εμφανίζονται με τη μεγαλύτερη συχνότητα στο αντίστοιχο σύνολο. Για καθένα από το κουπόνια αυτά υπολογίζονται οι πιθανότητες εμφάνισής σε spam και μη μηνύματα των συνόλων, τιμές οι οποίες ουσιαστικά «εκπαιδεύουν» το φίλτρο για επίτευξη σωστής κατηγοριοποίησης, με τον καθορισμό της προκύπτουσας (spam) πιθανότητας από το σύνολο το περιλαμβανόμενων κουπονιών στο περιεχόμενο κάθε ληφθέντος μηνύματος.

Το γεγονός ότι τα φίλτρα Bayes μπορούν και προσαρμόζονται στις παραλλαγές των spam μηνυμάτων (π.χ. παραλλαγές όσον αφορά τη χρήση λέξεων-φράσεων από τους spammers) κάνει τις αντίστοιχες τεχνικές που τα χρησιμοποιούν ιδιαίτερα αποδοτικές, πράγμα το οποίο δικαιολογεί και τη δημοτικότητά τους. Παρόλα αυτά, η αποδοτικότητά τους μειώνεται στο χρόνο, αφού υπάρχει το πρόβλημα της πολλές φορές μη-έγκυρης εισχώρησης γνώσης από τους χρήστες – παρόμοιο πρόβλημα με εκείνο της περίπτωσης φιλτραρίσματος με υπογραφές – απαιτώντας την περιοδική επανεκπαίδευσή τους με έγκυρα σύνολα μηνυμάτων.

2.5.4 Φιλτράρισμα των Μηνυμάτων με Βάση την Προέλευσή τους

Εκτός από την εξέταση του περιεχομένου των μηνυμάτων, ο εντοπισμός εκείνων που αποτελούν spam μπορεί να επιτευχθεί (ή ενισχυθεί) με την εξέταση της προέλευσής τους. Η σχετική πληροφορία που εκφράζει την προέλευση των μηνυμάτων αφορά στις περισσότερες περιπτώσεις IP διευθύνσεις ή ονόματα τομέων, τα οποία λαμβάνονται είτε από τα ανταλλασσόμενα μηνύματα-πακέτα που ορίζονται από το SMTP πρωτόκολλο είτε από το ίδιο το αποστελλόμενο μήνυμα αλληλογραφίας. Οι πιο διαδεδομένες τεχνικές που περιλαμβάνουν φιλτράρισμα με βάση την προέλευση χρησιμοποιούν μαύρες λίστες (blacklists), άσπρες λίστες (whitelists), ή κάτι ενδιάμεσο των δύο, όπως στην περίπτωση των γκρίζων λιστών (greylists).

Μαύρες Λίστες

Οι *μαύρες λίστες* περιλαμβάνουν πληροφορίες για υπολογιστές δικτύου (hosts) οι οποίοι είναι γνωστοί για την παραβίαση συγκεκριμένων κανόνων καλής πρακτικής όσον αφορά την ηλεκτρονική αλληλογραφία και γενικότερα τη «διαδικτυακή τους συμπεριφορά». Το πιο αξιόπιστο στοιχείο αυτών των λιστών αποτελεί η διεύθυνση IP κάθε υπολογιστή δικτύου που παρουσιάζει «κακή» συμπεριφορά, ένα από τα στοιχεία εκείνα τα οποία δεν μπορούν να παραποιηθούν (εκτός αν πρόκειται για proxies και παρόμοιες τεχνικές απόκρυψης του αποστολέα), οπότε και πάντα η πληροφορία που προσδιορίζουν να είναι έγκυρη. Οι λίστες αυτές χρησιμοποιούνται σε μια διαδικασία εντοπισμού και στη συνέχεια μπλοκαρίσματος των σχετικών SMTP συνδέσεων που δημιουργούνται από spammers, αποτρέποντας την αποστολή spam και αποφεύγοντας τις ενδεχόμενες αδυναμίες και μειονεκτήματα που παρουσιάζουν οι τεχνικές φιλτραρίσματος βασιζόμενες στο περιεχόμενο.

Οι τεχνικές που περιλαμβάνουν μαύρες λίστες διακρίνονται σε διάφορους τύπους, όπως *Domain Name System Blacklist* (DNSBL), *Right Hand Side Blacklist* (RHSBL) και *Uniform Resource Identifier Blacklist* (URIBL). Οι δύο τελευταίοι τύποι λιστών περιλαμβάνουν τη σάρωση του περιεχομένου ενός αποστελλόμενου μηνύματος, με τον πρώτο να χρησιμοποιείται στον εντοπισμό μηνυμάτων spam μέσω του ονόματος τομέα της ηλεκτρονικής διεύθυνσης του αποστολέα που βρίσκεται στην επικεφαλίδα του μηνύματος, και με τον δεύτερο να χρησιμοποιείται κατά τη σάρωση του σώματος του μηνύματος για τον εντοπισμό είτε κάποιων ονομάτων τομέα είτε κάποιων IP διευθύνσεων αποστολέων. Αντίθετα, στην περίπτωση της DNSBL, αυτό που παίζει ρόλο κατά τον εντοπισμό spam μηνυμάτων είναι η διεύθυνση IP του αποστολέα που εμφανίζεται κατά την

δημιουργία της SMTP σύνδεσης. Έτσι, δε χρειάζεται να μεταφερθεί ολόκληρο το μήνυμα αν διαπιστωθεί ότι η προέλευσή του αφορά κάποιον spammer.

Οι λίστες IP διευθύνσεων που χρησιμοποιούνται στην περίπτωση της DNSBL, μπορεί να είναι είτε συγκεντρωμένες ή κατανεμημένες σε κάποιους εξυπηρετητές εκτός τομέα του εξυπηρετητή αλληλογραφίας του παραλήπτη είτε να διατηρούνται και ανανεώνονται από κάποιον διαχειριστή στον ίδιο τον τομέα του παραλήπτη. Γενικά, η διαδικασία που ακολουθείται περιλαμβάνει την αποστολή ερωτημάτων στο σχετικό εξυπηρετητή, ο οποίος είναι υπεύθυνος να αναζητήσει κάποια εγγραφή στη διατηρούμενη μαύρη λίστα που να ταιριάζει με την επισυναπτόμενη IP διεύθυνση, και τελικά να απαντήσει με το αν πρόκειται για spammer ή όχι. Ένα παράδειγμα της διαδικασίας αυτής παρουσιάζεται στην Εικόνα 2.7.

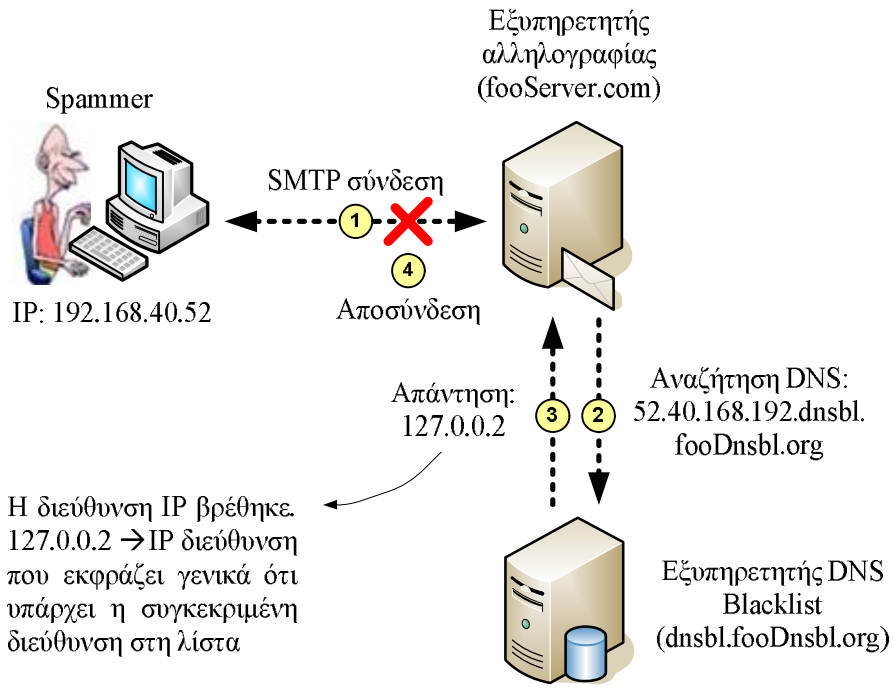
Σύμφωνα και με το συγκεκριμένο παράδειγμα, τα βήματα που ακολουθούνται για τον εντοπισμό των spammers με DNS μαύρες λίστες είναι τα εξής:

- Αρχικά, ένας εξυπηρετητής ηλεκτρονικής αλληλογραφίας δέχεται μια αίτηση για δημιουργία σύνδεσης SMTP από έναν πελάτη-αποστολέα (εδώ τον υπολογιστή ενός spammer), ώστε να επιτευχθεί η μεταφορά ενός ηλεκτρονικού μηνύματος.

- Ο εξυπηρετητής ηλεκτρονικής αλληλογραφίας στη συνέχεια ξεχωρίζει τη διεύθυνση IP του αποστολέα (εδώ 192.168.40.52), αντιστρέφει τη σειρά των bytes διεύθυνσης και προσθέτει το όνομα τομέα του εξυπηρετητή με την DNS μαύρη λίστα (εδώ 52.40.168.192.dnsbl.fooDnsbl.org). Το όνομα που προκύπτει το αποστέλλει ως ερώτημα στον τελευταίο για να αναζητήσει την συγκεκριμένη διεύθυνση στη μαύρη λίστα διευθύνσεων IP.

- Ο εξυπηρετητής με τη μαύρη λίστα, ανάλογα με την ύπαρξη ή όχι της ερωτηθείσας διεύθυνσης, απαντάει είτε με μια IP διεύθυνση που υποδηλώνει κάποια πληροφορία [84] σχετικά με την τελικά εντοπισμένη διεύθυνση (εδώ 127.0.0.2) είτε με έναν κωδικό που δηλώνει ότι δεν υπάρχει η συγκεκριμένη διεύθυνση.

- Τελικά, ο εξυπηρετητής ηλεκτρονικής αλληλογραφίας, ανάλογα με την απάντηση σχετικά με τον εντοπισμό ή όχι της IP διεύθυνσης αποστολέα στη μαύρη λίστα, απορρίπτει ή αποδέχεται τη μεταφορά του ηλεκτρονικού μηνύματος, αντίστοιχα.



Εικόνα 2.7: Διαδικασία εντοπισμού IP διευθύνσεων που αποστέλλουν spam με DNS μαύρες λίστες

Γενικά, οι τεχνικές αυτές που περιλαμβάνουν τη χρήση μαύρων λιστών, και συγκεκριμένα οι DNSBL τεχνικές, αποτελούν μια σχετικά εύκολα υλοποιήσιμη και αξιόπιστη λύση στην αντιμετώπιση του spamming. Παρόλα αυτά δε, το γεγονός ότι δεν εξετάζεται το περιεχόμενο των μηνυμάτων μπορεί να οδηγήσει πολλές φορές σε εσφαλμένα μπλοκαρίσματα μηνυμάτων. Για το λόγο αυτό, αρκετά σημαντική είναι η χρησιμοποίηση περισσότερων του ενός μαύρων λιστών από διάφορους παροχείς, οι οποίες θα καθορίζουν με πλειοψηφικό τρόπο αν μια δεδομένη για παράδειγμα διεύθυνση IP του αποστολέα ενός μηνύματος αποτελεί πηγή spam μηνυμάτων. Μερικοί παροχείς DNS μαύρων λιστών είναι οι SORBS [89], SPAMHAUS [91], UCEPROTECT [100], NJABL [71], SPAMCOP [90], CBL [14], και HEISE-NIXSPAM [70] (ο οποίος χρησιμοποιείται και στην υλοποίηση του συστήματος που περιγράφεται στην μεταπτυχιακή διατριβή).

Άσπρες Λίστες

Οι άσπρες λίστες περιλαμβάνουν διευθύνσεις αποστολέων οι οποίες θεωρούνται επιβεβαιωμένες ότι δεν στέλνουν spam μηνύματα, οπότε και παρακάμπτονται συνήθως οι περισσότεροι από τους περαιτέρω μηχανισμούς εντοπισμού που μεσολαβούν μέχρι τον τελικό παραλήπτη. Αυτές οι διευθύνσεις μπορεί να είναι είτε διευθύνσεις αλληλογραφίας είτε ονόματα τομέα είτε IP διευθύνσεις, οπότε και μπορούν να αφορούν τόσο συγκεκριμένους αποστολείς όσο και ένα ευρύτερο σύνολο αυτών. Οι σχετικές δε τεχνικές που περιλαμβάνουν τη χρήση τέτοιων

λιστών μπορούν να χρησιμοποιηθούν σε τοπικό επίπεδο τόσο στην πλευρά του εξυπηρετητή όσο και στην πλευρά του πελάτη του ηλεκτρονικού ταχυδρομείου, αλλά και σε ευρύτερο επίπεδο όπως και στην περίπτωση των μαύρων λιστών. Όμως, κατά κύριο λόγο αποτελούν μια συμπληρωματική λύση σε άλλες πιο αποδοτικές τεχνικές σχετικά με την αντιμετώπιση του spam, μιας και περιλαμβάνουν την κατηγοριοποίηση μηνυμάτων που στέλνονται μόνο από συγκεκριμένες διευθύνσεις.

Στην περίπτωση που οι άσπρες λίστες χρησιμοποιούνται από την πλευρά του εξυπηρετητή σε τοπικό επίπεδο, συνήθως κάποιος διαχειριστής είναι υπεύθυνος για την εκχώρηση νέων διευθύνσεων σε μια τέτοια λίστα. Αυτό βέβαια είναι εφικτό μόνο σε μια εταιρεία ή σε έναν εξυπηρετητή ηλεκτρονικού ταχυδρομείου με σχετικά περιορισμένο αριθμό λογαριασμών. Στην περίπτωση δε της πλευράς του πελάτη, οι «άσπρες» διευθύνσεις αποτελούν το σύνολο των διευθύνσεων που είναι αποθηκευμένες στον κατάλογο επαφών του σχετικού χρήστη. Όσον αφορά τη χρήση άσπρων λιστών σε ευρύτερο επίπεδο, υπάρχουν διάφοροι παροχείς που παρέχουν τέτοιες λίστες βασιζόμενες στο DNS (Domain Name System Whitelists – DNSWL), οι οποίοι παρέχουν πληροφορίες σχετικά με διευθύνσεις που συμπεριλαμβάνονται σε παρόμοια ερωτήματα, όπως και στην περίπτωση των μαύρων λιστών. Μερικοί από αυτούς είναι οι Swiss Whitelist [93], Habeas [38] και Dnswl.org [23], οι οποίοι περιλαμβάνουν διευθύνσεις διαφόρων εγγεγραμμένων οργανισμών.

Γκρίζες Λίστες

Ενώ οι τεχνικές που χρησιμοποιούν μαύρες και άσπρες λίστες είναι απόλυτες σχετικά με την αποδοχή ή μπλοκάρισμα-απόρριψη των αποστελλόμενων μηνυμάτων αλληλογραφίας, αντίστοιχα, οι τεχνικές που χρησιμοποιούν γκρίζες λίστες [40, 54] περιλαμβάνουν συνδυασμένα στοιχεία λειτουργίας και διαφοροποιούνται σε σχέση με τον προκαθορισμένο χαρακτήρα των πρώτων. Το βασικό δε στοιχείο στο οποίο στηρίζεται η επιτυχία των τεχνικών χρήσης γκρίζων λιστών αποτελεί το γεγονός ότι τα spam μηνύματα αποστέλλονται κυρίως από ειδικό λογισμικό πολλαπλής αποστολής μηνυμάτων (spambots) που δεν τηρεί πιστά τα πρότυπα σχετικά με την επικοινωνία στο ηλεκτρονικό ταχυδρομείο (RFC 5321 [51]). Πιο συγκεκριμένα, σημαντικότερο αποτελεί το γεγονός ότι τα μηνύματα που δεν φτάνουν στον τελικό παραλήπτη τελικά δεν ξαναστέλνονται όπως υπαγορεύει το σχετικό πρότυπο, και αυτό το χαρακτηριστικό είναι που εκμεταλλεύονται οι τεχνικές χρήσης γκρίζων λιστών.

Για να επιτύχουν στο στόχο τους, οι τεχνικές αυτές χρησιμοποιούν τα εξής στοιχεία για κάθε μήνυμα αλληλογραφίας:

1. Τη διεύθυνση IP του αποστολέα του μηνύματος.

2. Την ηλεκτρονική διεύθυνση του αποστολέα που βρίσκεται στην επικεφαλίδα του μηνύματος.

3. Την ηλεκτρονική διεύθυνση του παραλήπτη που βρίσκεται στην επικεφαλίδα του μηνύματος.

Έτσι, όταν ένας εξυπηρετητής λαμβάνει ένα μήνυμα από έναν αποστολέα από τον οποίο σε προηγούμενο χρονικό διάστημα δεν έχει σταλεί κανένα άλλο μήνυμα, δημιουργεί μια εγγραφή με τα παραπάνω στοιχεία του μηνύματος, την αποθηκεύει σε μια "γκρίζα" λίστα, και στη συνέχεια απορρίπτει το μήνυμα. Ο εξυπηρετητής- αποστολέας ο οποίος είναι σύμφωνος με τα πρότυπα επικοινωνίας του ηλεκτρονικού ταχυδρομείου θα ξαναστείλει το συγκεκριμένο μήνυμα μετά από συγκεκριμένο διάστημα (π.χ. μεγαλύτερο από μισή ώρα και μικρότερο από 4 με 5 μέρες, σύμφωνα με το RFC 5321 [51]), οπότε ο εξυπηρετητής που θα το λάβει θα αναζητήσει τη σχετική εγγραφή στη λίστα, και εφόσον υπάρχει, το μήνυμα θα σημειωθεί ως ασφαλές και θα προωθηθεί στον παραλήπτη. Αντίθετα, στην περίπτωση ενός spambot, το μήνυμα δε θα ξανασταλθεί όπως πρέπει και δε θα φτάσει ποτέ στον τελικό παραλήπτη. Ο λόγος δε που δεν υποστηρίζονται τέτοιοι (σύμφωνα με τα πρότυπα) μηχανισμοί από τους spammers, είναι ότι εκτός από τους επιπλέον πόρους αποθήκευσης που απαιτούνται για την επανα-αποστολή των μηνυμάτων, απαιτείται και επιπλέον χρόνος, μέσα στον οποίο μπορούν να αποσταλούν πληθώρα μηνυμάτων σε πλήθος άλλων παραληπτών.

2.5.5 Πιστοποίηση των Μηνυμάτων

Όπως έχει αναφερθεί και σε προηγούμενες ενότητες, το ηλεκτρονικό ταχυδρομείο δημιουργήθηκε σε μια περίοδο όπου το δίκτυο υποδομής περιλάμβανε μόνο αξιόπιστους χρήστες, με αποτέλεσμα να υπάρχει έλλειψη μηχανισμών πιστοποίησης των αποστελλόμενων μηνυμάτων και των αποστολέων τους. Έτσι, σε μια προσπάθεια να μην μεταβληθεί η υπάρχουσα υποδομή και αρχιτεκτονική πελάτη-εξυπηρετητή, δημιουργήθηκαν διάφορες τεχνικές πιστοποίησης είτε του αποστολέα μέσω του φακέλου ή της επικεφαλίδας ενός αποστελλόμενου μηνύματος καθώς και επεκτάσεων του SMTP πρωτοκόλλου είτε του

περιεχομένου μέσω του σώματος του μηνύματος, ώστε να αναγνωρίζονται τα «νόμιμα» ηλεκτρονικά μηνύματα. Τρεις κατηγορίες τεχνικών πιστοποίησης αποτελούν λοιπόν οι εξής:

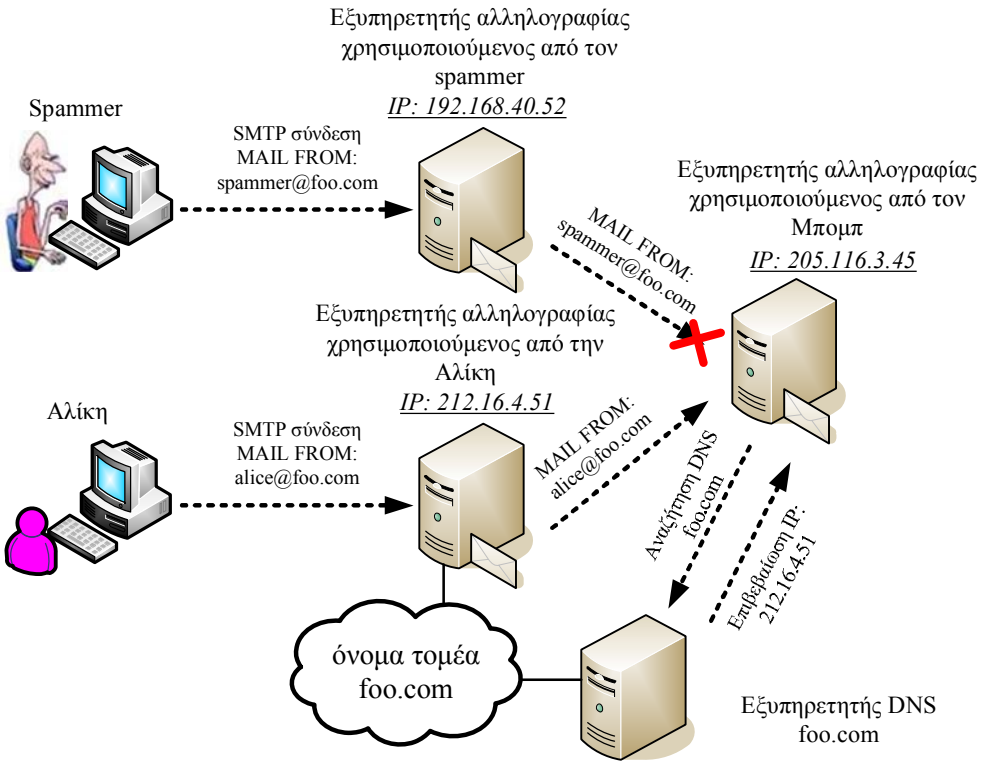
- Τεχνικές που περιλαμβάνουν επεκτάσεις του πρωτοκόλλου SMTP.
- Τεχνικές που περιλαμβάνουν πιστοποίηση αποστολέα με βάση το όνομα τομέα προέλευσης ενός αποστελλόμενου μηνύματος.
- Τεχνικές που περιλαμβάνουν πιστοποίηση περιεχομένου ενός αποστελλόμενου μηνύματος με βάση υπογραφές.

Επεκτάσεις Πρωτοκόλλου SMTP

Το *SMTP-AUTH* [88] (το οποίο υπόκειται στο πρότυπο του Simple Authentication and Security Layer – SASL [61]) και το *POP/IMAP before SMTP* αποτελούν επεκτάσεις του πρωτοκόλλου SMTP σε μια προσπάθεια εφαρμογής κάποιου είδους πιστοποίησης κατά την αποστολή-μεταφορά ηλεκτρονικών μηνυμάτων. Ένας βασικός τρόπος πιστοποίησης που υιοθετούν αποτελεί η χρήση κάποιου ονόματος χρήστη και κωδικού πρόσβασης για την εξασφάλιση της σχετικής σύνδεσης (SMTP ή POP/IMAP), ώστε τελικά να επιτραπεί η ανταλλαγή μηνυμάτων μεταξύ υπολογιστών χρηστών και εξυπηρετητών για ένα συγκεκριμένο χρονικό διάστημα. Παρόλα αυτά δε, από τη στιγμή που τα ονόματα χρήστη και οι κωδικοί πρόσβασης είναι εκτεθειμένα σε ενδεχόμενους ιούς (π.χ. zombies), μπορεί οι συγκεκριμένες τεχνικές να παρακαμφτούν από τους spammers.

Πιστοποίηση Αποστολέα των Μηνυμάτων

Σκοπός της πιστοποίησης αποστολέα – γνωστή και ως *πιστοποίηση βασιζόμενη στη διαδρομή* (path-based authentication) – είναι ουσιαστικά να εντοπίζονται εκείνα τα μηνύματα τα οποία έχουν ψεύτικους αποστολείς, τακτική η οποία υιοθετείται ως επί το πλείστον από τους spammers. Το βασικό πλεονέκτημα δε στη συγκεκριμένη περίπτωση είναι ότι δεδομένης της διαδικασίας πιστοποίησης πριν αποσταλεί ολόκληρο το μήνυμα (π.χ. το σώμα του μηνύματος) μπορεί να εξοικονομηθεί κάποιο εύρος ζώνης στην επικοινωνία καθώς και να μειωθεί ο χρόνος επεξεργασίας που θα απαιτούταν αν το μήνυμα περνούσε ολόκληρο σε περαιτέρω μηχανισμούς αντιμετώπισης του spamming, όπως αυτοί που αναφέρθηκαν στις προηγούμενες υποενότητες. Μέθοδοι πιστοποίησης αποστολέα αποτελούν οι Sender Policy Framework (SPF) [103] και



Εικόνα 2.8: Διαδικασία πιστοποίησης αποστολέα μηνυμάτων

Sender ID [58], οι οποίες είναι μέλη της ευρύτερης οικογένειας μεθόδων πιστοποίησης Lightweight MTA Authentication Protocol (LMAP) [19].

Όσον αφορά την SPF, η οποία είναι και η πιο ευρέως χρησιμοποιούμενη μέθοδος για την πιστοποίηση αποστολέα, αυτή βασίζεται στην εξέταση του φακέλου ενός αποστελλόμενου μηνύματος για το αν τα πεδία ονόματος στις HELO και MAIL FROM εντολές του SMTP πρωτοκόλλου ταιριάζουν με την IP διεύθυνση του αποστολέα. Για το λόγο αυτό, απαραίτητο στοιχείο αποτελεί η δυνατότητα των εξυπηρετητών DNS να διαθέτουν μια λίστα με όλους τους εξυπηρετητές ηλεκτρονικού ταχυδρομείου που ανήκουν στο συγκεκριμένο κάθε φορά πεδίο ονομάτων. Η διαδικασία που ακολουθείται για την συγκεκριμένη πιστοποίηση παρουσιάζεται στην Εικόνα 2.8. Η κύρια διαφορά δε που παρουσιάζει η μέθοδος Sender ID, η οποία προτάθηκε από τη Microsoft και βασίστηκε σε μεγάλο βαθμό στην SPF, είναι ότι στη διαδικασία εξέτασης που ακολουθείται λαμβάνονται υπόψη και τα πεδία στην επικεφαλίδα ενός μηνύματος.

Πιστοποίηση Περιεχομένου των Μηνυμάτων

Η πιστοποίηση περιεχομένου των μηνυμάτων περιλαμβάνει τη χρήση *ψηφιακών υπογραφών* (digital signatures) οι οποίες προστίθενται στα αποστελλόμενα μηνύματα ώστε να μπορούν οι

παραλήπτες να επιβεβαιώσουν την έγκυρη ταυτότητα του αποστολέα. Οι υπογραφές αυτές βασίζονται κυρίως στην *κρυπτογραφία δημόσιου κλειδιού* (public-key cryptography), όπου γίνεται χρήση δύο διαφορετικών κλειδιών (*κλειδί* - αλφαριθμητική σειρά χαρακτήρων), ενός *ιδιωτικού κλειδιού* (private key) για κωδικοποίηση και ενός *δημόσιου κλειδιού* (public key) για αποκωδικοποίηση. Το βασικό δε μειονέκτημα στη συγκεκριμένη περίπτωση είναι ότι θα πρέπει να ληφθεί ολόκληρο το μήνυμα ώστε να μπορεί να ξεκινήσει η διαδικασία πιστοποίησης. Τρεις αντιπροσωπευτικές τεχνικές πιστοποίησης περιεχομένου αποτελούν εκείνες των *DomainKeys Identified Mail* (DKIM) [2, 3], *Secure/Multipurpose Internet Mail Extensions* (S/MIME) [82] και *Pretty Good Privacy* (PGP) [12].

Σχετικά με την DKIM τεχνική, αυτή αποτελεί συνδυασμό των DomainKeys [20] και Identified Internet Mail (IIM) [28] ενσωματώνοντας περαιτέρω βελτιώσεις όσον αφορά την πιστοποίηση τόσο του αποστολέα όσο και του περιεχομένου ενός ηλεκτρονικού μηνύματος. Πιο συγκεκριμένα, για το σκοπό αυτό η DKIM περιλαμβάνει τον καθορισμό ενός αλγορίθμου κατακερματισμού (π.χ. MD5, SHA) και ενός αλγορίθμου κωδικοποίησης με δημόσιο κλειδί. Η όλη διαδικασία έγκειται στα εξής βήματα:

1. Παράγεται μια *σύνοψη* (hash) από την επικεφαλίδα και το σώμα του αποστελλόμενου μηνύματος.

2. Η σύνοψη κωδικοποιείται με χρήση ενός ιδιωτικού κλειδιού και το αποτέλεσμα ενσωματώνεται στην ψηφιακή υπογραφή που περιλαμβάνεται στην επικεφαλίδα του μηνύματος.

3. Όταν ληφθεί το μήνυμα από τον παραλήπτη, το αντίστοιχο δημόσιο κλειδί ζητείται από τον DNS εξυπηρετητή του ονόματος τομέα της διεύθυνσης του αποστολέα, ώστε να αποκωδικοποιηθεί η σύνοψη του μηνύματος.

4. Το αποτέλεσμα συγκρίνεται με την υπολογισθείσα σύνοψη που προκύπτει από το ληφθέν μήνυμα και αν συμπίπτει, τότε πρόκειται για αυθεντικό μήνυμα.

Όσον αφορά τις S/MIME και PGP, τεχνικές οι οποίες επίσης περιλαμβάνουν κωδικοποίηση και υπογραφές του περιεχομένου των αποστελλόμενων μηνυμάτων για πιστοποίηση, αυτές αποτελούν λιγότερο αξιόπιστες τεχνικές επικεντρωνόμενες μόνο σε πιστοποίηση μηνυμάτων αποστολέων που αφορούν χρήστες, σε αντίθεση με την DKIM που αφορά και πιστοποίηση

μηνυμάτων από εξυπηρετητές. Σχετικά δε με την επιβεβαίωση των κλειδιών αποστολέων που απαιτούνται για την κωδικοποίηση-αποκωδικοποίηση, αυτή στηρίζεται εξολοκλήρου σε ξεχωριστούς παροχείς οι οποίοι περιλαμβάνουν τα κλειδιά από δηλωμένους αποστολείς.

2.6 Νομοθεσία

Από τη στιγμή που το διαδίκτυο άρχισε να διευρύνεται και να παίρνει τη σημερινή του μορφή, όπου θεωρείται ένα από τα περισσότερο χρησιμοποιούμενα μέσα επικοινωνίας, άρχισε και το spamming να διευρύνεται φτάνοντας στα σημερινά υψηλά ποσοστά ηλεκτρονικών μηνυμάτων. Έτσι, κρίθηκε αναγκαίο να υπάρξει κάποια παρέμβαση από τις κυβερνήσεις των χωρών παγκοσμίως, ώστε με κατάλληλα νομοθετικά μέτρα να περιορίσουν το συγκεκριμένο φαινόμενο. Ωστόσο, παρά την πληθώρα μέτρων που έχουν παρθεί από πολλές χώρες, η ισχύουσα γενικά νομοθεσία δεν μπόρεσε να καταπολεμήσει το spamming, πράγμα το οποίο φαίνεται και από τα ποσοστά μηνυμάτων spam που συνεχίζουν να παραμένουν υψηλά. Κύριοι λόγοι είναι η αδυναμία των σχετικών μέτρων να προσαρμοσθούν στην εξελισσόμενη φύση του spamming, καθώς και το γεγονός ότι υιοθετείται διαφορετική πολιτική από κάθε χώρα σχετικά με τη συγκεκριμένη νομοθεσία ή σε πολλές περιπτώσεις είναι σχεδόν ανύπαρκτη. Έτσι, δεδομένου ότι το διαδίκτυο δεν έχει σύνορα, θα πρέπει και η νομοθεσία σε κάθε χώρα να ακολουθεί ένα παγκόσμιο ενιαίο πλαίσιο. Αυτό δεν σημαίνει ότι οι ισχύοντες νόμοι δεν είναι αποδοτικοί, αλλά απλώς δεν είναι τόσο αποδοτικοί όσο θα ήταν επιθυμητό. Δύο κύρια καθεστώτα πάνω στα οποία βασίζονται οι ισχύοντες νόμοι για την καταπολέμηση του spamming είναι:

- «*προγενέστερης συγκατάθεσης*» (opt-in): επιτρέπεται η αποστολή (εμπορικών) μηνυμάτων μόνο σε εκείνους τους χρήστες Διαδικτύου που έχουν προηγουμένως συναινέσει ρητά να τα λαμβάνουν, και

- «*κατά προαίρεση αυτοεξαίρεσης*» (opt-out): επιτρέπεται η αποστολή (εμπορικών) μηνυμάτων, ακόμα και αυτόκλητων, σε χρήστες Διαδικτύου που δεν έχουν προηγουμένως συναινέσει ρητά σε κάτι τέτοιο, αρκεί να τους δίδεται η δυνατότητα να αφαιρεθούν από τη σχετική λίστα του αποστολέα και να σταματήσουν την περαιτέρω αποστολή μηνυμάτων όποτε αυτοί θελήσουν.

2.6.1 Νομοθεσία στην Ελλάδα

Στην Ελλάδα, το νομικό πλαίσιο που αφορά το spamming έγκειται στους νόμους 2472/1997 [72] και 3471/2006 [73]. Και οι δύο προέρχονται από την ενσωμάτωση Ευρωπαϊκών Κοινοτικών οδηγιών, και συγκεκριμένα των 2000/31/ΕΚ [74] και 2002/58/ΕΚ [75], αντίστοιχα.

Σχετικά με τον πρώτο, αυτός αφορά την προστασία του ατόμου από την επεξεργασία προσωπικών δεδομένων και ορίζει την Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα [8]. Η τελευταία είναι αρμόδια της τήρησης μητρώου όπου σύμφωνα με το άρθρο 13, παρ. 3 αποθηκεύονται στοιχεία ατόμων που δεν επιθυμούν προσωπικά δεδομένα να γίνουν αντικείμενο επεξεργασίας για λόγους προώθησης πώλησης αγαθών ή παροχής υπηρεσιών, καθώς και μητρώου όπου σύμφωνα με το άρθρο 19, παρ. 4 μπορούν να καταχωρηθούν όσοι χρήστες Διαδικτύου το επιθυμούν δηλώνοντας με μια σχετική αίτηση ότι αρνούνται να λαμβάνουν οποιαδήποτε ηλεκτρονική αλληλογραφία εμπορικού-διαφημιστικού τύπου.

Όσον αφορά το δεύτερο νόμο, σύμφωνα με το άρθρο 11 με τίτλο «Μη ζητηθείσα επικοινωνία», υιοθετείται κατά κύριο λόγο το καθεστώς της προγενέστερης συγκατάθεσης (opt-in) με μερικές εξαιρέσεις που περιλαμβάνουν στοιχεία του κατά προαίρεση αυτοεξαίρεσης (opt-out) καθεστώς. Πιο συγκεκριμένα, ισχύουν τα εξής:

- απαγορεύεται η αποστολή κάθε ηλεκτρονικού μηνύματος εμπορικού-διαφημιστικού τύπου σε κάποιον χρήστη Διαδικτύου χωρίς την εκ των προτέρων ρητή συγκατάθεση του τελευταίου,

- εξαιρούνται οι περιπτώσεις όπου τα στοιχεία επαφής του χρήστη αποκτήθηκαν νομίμως στο πλαίσιο κάποιας συναλλαγής, με την προϋπόθεση ότι μπορεί ο χρήστης να εκφράσει την αντίρρηση του στη λήψη περαιτέρω μηνυμάτων,

- απαγορεύεται η αποστολή κάθε ηλεκτρονικού μηνύματος εμπορικού-διαφημιστικού τύπου σε κάποιον χρήστη που έχει δηλώσει στην Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα ότι δεν δέχεται τέτοια μηνύματα,

- οι αποστολείς των παραπάνω μηνυμάτων υποχρεούνται να περιλαμβάνουν έγκυρα στοιχεία και διευθύνσεις που τους αφορούν, και

- οι χρήστες-παραλήπτες αποτελούν τόσο φυσικά όσο και νομικά πρόσωπα.

2.6.2 Νομοθεσία στην Ευρωπαϊκή Ένωση

Η Ευρωπαϊκή Ένωση, το 2002, εξέδωσε την οδηγία 2002/58/ΕΚ που αφορά γενικότερα την προστασία των δεδομένων προσωπικού χαρακτήρα και της ιδιωτικής ζωής στον τομέα των ηλεκτρονικών επικοινωνιών, η οποία στη συνέχεια τροποποιήθηκε με την οδηγία 2009/136/ΕΚ [76]. Κύριος στόχος ήταν οι χώρες-μέλη να υιοθετήσουν ένα κοινό νομικό πλαίσιο όσον αφορά τον συγκεκριμένο τομέα προσαρμόζοντας τις εθνικές τους νομοθεσίες. Ένα από τα περιλαμβανόμενα ζητήματα αποτελούσε και το spamming, για το οποίο γίνεται αναφορά στο Άρθρο 13 της σχετικής οδηγίας με τίτλο «Αυτόκλητες κλήσεις». Σύμφωνα με αυτό, κατά κάποιο τρόπο υιοθετούνται και τα δύο καθεστώτα (opt-in και opt-out) για τον καθορισμό των παράνομων μηνυμάτων spam ως εξής:

- opt-in – απαγορεύεται η χρήση συστημάτων αυτόματης αποστολής ηλεκτρονικών μηνυμάτων με σκοπούς απευθείας εμπορικής προώθησης σε χρήστες του Διαδικτύου που δεν έχουν δώσει εκ των προτέρων τη συγκατάθεσή τους, και

- opt-out – εκτός αν οι τελευταίοι υπήρξαν πελάτες στο πλαίσιο πώλησης ενός προϊόντος ή μιας υπηρεσίας, με την προϋπόθεση ότι τους παρέχεται η δυνατότητα να αντιτάσσονται δωρεάν και εύκολα στην περαιτέρω αποστολή μηνυμάτων γνωρίζοντας στοιχεία επικοινωνίας του αποστολέα.

Επίσης, συγκροτήθηκε μια ενιαία Αρχή-ομάδα Αρχών, ονομαζόμενη «the Contact Network of Spam Authorities» (CSNA), η οποία αποτελείται από επιμέρους Αρχές σε εθνικό επίπεδο των χωρών-μελών, και έχει ως στόχο τη μεταξύ τους συνεργασία όσον αφορά θέματα αντιμετώπισης της διασυνοριακής δραστηριότητας του spamming. Σχετικά δε με την ποινικοποίηση, αυτό έγκειται στην καθεμιά εθνική νομοθεσία, οπότε και οι αποστολείς μηνυμάτων spam τιμωρούνται ανάλογα με τη χώρα που βρίσκονται, ανεξαρτήτως της τοποθεσίας των θιγόμενων χρηστών-παραληπτών. Η Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα στην Ελλάδα αποτελεί μέλος της συγκεκριμένης ομάδας.

2.6.3 Νομοθεσία στις Ηνωμένες Πολιτείες Αμερικής

Στις Ηνωμένες Πολιτείες Αμερικής, αρχικά, κάθε πολιτεία είχε τη δικιά της νομοθεσία για την αντιμετώπιση του spamming, η οποία ήταν ανεξάρτητη από εκείνες των άλλων πολιτειών. Έτσι, υπήρχαν νομοθεσίες ιδιαίτερα αυστηρές (π.χ. πολιτεία της Βιρτζίνια) που είτε οδηγούσαν τους

αποστολείς μηνυμάτων spam στη φυλακή είτε τους επέβαλλαν σημαντικά πρόστιμα, καθώς και ανύπαρκτες (π.χ. πολιτεία της Φλόριντας) που είχαν ως αποτέλεσμα τη δημιουργία «ισχυρών πηγών» spamming οι οποίες προστατευμένες πλέον μπορούσαν να δρουν ανενόχλητες παγκοσμίως.

Το 2003, σε μια προσπάθεια να αντιμετωπιστεί ο ολοένα και αυξανόμενος ρυθμός spamming, ιδιαίτερα δε σε μια χώρα όπου ο αριθμός των χρηστών του Διαδικτύου είναι ο μεγαλύτερος κατά αναλογία συνολικού πληθυσμού, ψηφίστηκε ο ομοσπονδιακός νόμος «Controlling the Assault of Non-Solicited Pornography And Marketing Act of 2003» (CAN-SPAM Act of 2003) [17]. Με βάση αυτό το νομικό πλαίσιο πλέον, κάθε πολιτεία θα έπρεπε να προσαρμόσει τη νομοθεσία της ακόμα και αν ήταν προηγουμένως περισσότερο αυστηρή με το spamming, ενώ αρμόδια Αρχή για την εφαρμογή του ορίστηκε η Ομοσπονδιακή Επιτροπή Εμπορίου (Federal Trade Commission - FTC). Όσον αφορά το καθεστώς το οποίο υιοθετήθηκε, ήταν το opt-out, σύμφωνα με το οποίο επιτρέπεται η αποστολή αυτόκλητων εμπορικών μηνυμάτων αρκεί να τηρούνται οι εξής κύριες προϋποθέσεις:

- να περιλαμβάνεται στο αποστελλόμενο μήνυμα είτε μια έγκυρη ηλεκτρονική διεύθυνση επικοινωνίας με τον αποστολέα είτε κάποιος εναλλακτικός ηλεκτρονικός μηχανισμός ώστε να μπορεί ο παραλήπτης να σταματήσει την περαιτέρω αποστολή μηνυμάτων για ορισμένο χρονικό διάστημα 30 ημερών,

- να υπάρχει ευδιάκριτη ένδειξη ότι το μήνυμα είναι εμπορικό, και

- να περιλαμβάνεται μια έγκυρη (πραγματική) διεύθυνση του αποστολέα.

Σύμφωνα με τις προϋποθέσεις αυτές, λοιπόν, ο νόμος CAN-SPAM δείχνει ότι είναι ιδιαίτερα ανεκτικός στο spamming, πράγμα το οποίο δικαιολογεί τις έντονες κριτικές που έχει δεχθεί από τη στιγμή που ψηφίστηκε.

Κεφάλαιο 3

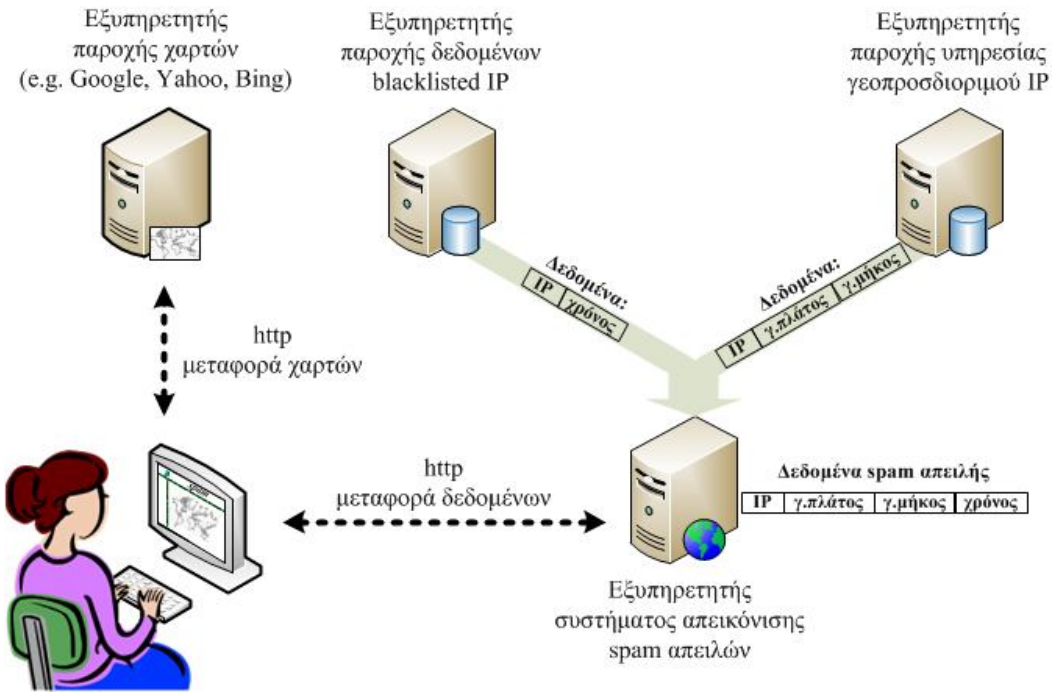
Σύστημα Απεικόνισης Χωρικών Δεδομένων – η Περίπτωση του Spamming

Στο προηγούμενο κεφάλαιο έγινε μια παρουσίαση του φαινομένου της αποστολής ανεπιθύμητης ηλεκτρονικής αλληλογραφίας - spam ως προς τον τρόπο με τον οποίο πραγματοποιείται καθώς και τις τεχνικές που έχουν αναπτυχθεί για την αντιμετώπισή του. Δεδομένης της σημαντικότητάς του ως φαινομένου στον τομέα των διαδικτυακών εφαρμογών, θα μπορούσε να ειπωθεί ότι μια παρουσίαση της όλης εικόνας της απειλής αυτής, όπως θα μπορούσε να θεωρηθεί το spamming, ανά τον κόσμο θα ήταν ιδιαίτερα ενδιαφέρουσα για την περαιτέρω μελέτη του φαινομένου και την εξαγωγή χρήσιμων συμπερασμάτων. Για να επιτευχθεί κάτι τέτοιο, απαιτείται κυρίως η χρήση ενός συστήματος απεικόνισης δεδομένων-πληροφοριών, όπως είναι τα *Γεωγραφικά Συστήματα Πληροφοριών* (Geographic Information Systems - GIS), όπου θα συσχετίζονται δεδομένα σχετικά με τη γεωγραφική εμφάνιση των απειλών αυτών με την εξέλιξή τους στο χρόνο, ενώ η απεικόνισή τους θα γίνεται πάνω σε χάρτες όπως αυτούς που παρέχονται από σχετικούς παροχείς, όπως οι Google, Yahoo, Bing κ.α. Βασική

δε προϋπόθεση για τη ζητούμενη «χωρο-χρονική τοποθέτηση» των απειλών αυτών, αποτελεί φυσικά η ανάκτηση και χρησιμοποίηση δεδομένων που να σχετίζονται με τον τόπο και τη χρονική στιγμή που εντοπίστηκε μια απειλή ως αποστολή μηνυμάτων spam.

Αναφορικά με την ανάκτηση τέτοιων χωρο-χρονικών δεδομένων, λαμβάνεται ουσιαστικά υπόψη η φύση της επικοινωνίας που πραγματοποιείται στην εφαρμογή του ηλεκτρονικού ταχυδρομείου, ώστε να επιλεχθεί ο κατάλληλος τρόπος δημιουργίας και αποθήκευσής τους. Έτσι, δεδομένου του γεγονότος ότι βασικό στοιχείο της συγκεκριμένης επικοινωνίας αποτελεί η ύπαρξη ενός αποστολέα κάποιου ηλεκτρονικού μηνύματος ο οποίος συνδέεται μέσω ενός υπολογιστή στο Διαδίκτυο (βλ. Ενότητα 2.1.1) και συνάμα του δίνεται η δυνατότητα πρόσβασης στη συγκεκριμένη υπηρεσία επικοινωνίας, θα μπορούσε να χρησιμοποιηθεί η σχετική ηλεκτρονική διεύθυνση IP για τον εντοπισμό της θέσης μιας ενδεχόμενης απειλής spam. Το τελευταίο στηρίζεται στο γεγονός ότι οι διευθύνσεις IP είναι κατανεμημένες σε διάφορες ομάδες ανά τον κόσμο σχηματίζοντας κατάλληλα δίκτυα το εύρος των οποίων καθώς και η θέση τους έχει ανατεθεί σε καθέναν από τους εκάστοτε παροχείς πρόσβασης στο Διαδίκτυο. Το μόνο που απομένει είναι η εύρεση ενός τρόπου με τον οποίο θα εντοπίζεται η αποστολή μηνυμάτων spam από κάποια διεύθυνση IP σε κάποια χρονική στιγμή, και σίγουρα αυτό θα γίνεται με τη λήψη αυτών των μηνυμάτων, αφού μόνο έτσι μπορούν να εξετασθούν τα στοιχεία του αποστολέα καθώς και κάθε μήνυμα για το αν πρόκειται για spam. Ανατρέχοντας στο προηγούμενο κεφάλαιο και συγκεκριμένα στην Ενότητα 2.5, όπου γίνεται αναφορά στις τεχνικές αντιμετώπισης του spamming, παρατηρείται ότι μια συγκεντρωτική λύση της παραπάνω αναζήτησης αποτελεί η περίπτωση των τεχνικών φιλτραρίσματος με βάση την προέλευση των μηνυμάτων (Ενότητα 2.5.4). Σύμφωνα με τις τεχνικές αυτές, παρέχονται λίστες IP διευθύνσεων (π.χ. μαύρες λίστες) που αντιπροσωπεύουν πηγές spam μηνυμάτων, και συνήθως συνοδεύονται από τη χρονική στιγμή (timestamp) που χαρακτηρίστηκαν ως τέτοιες πηγές απειλής. Έτσι, για να δημιουργηθούν πλέον χωρο-χρονικά δεδομένα θα πρέπει να μετατραπούν οι διευθύνσεις IP σε γεωγραφικές συντεταγμένες, πράγμα το οποίο επιτυγχάνεται με την χρησιμοποίηση κατάλληλων υπηρεσιών ή βάσεων δεδομένων που παρέχονται από σχετικούς παροχείς που επιτελούν γεωπροσδιορισμό IP (IP geolocation). Μια ενδεικτική παρουσίαση της λειτουργίας ενός διαδικτυακού συστήματος γεωγραφικής απεικόνισης spam απειλών παρουσιάζεται στην Εικόνα 3.1, ενώ καθένα από τα τμήματα που το απαρτίζουν περιγράφεται πιο αναλυτικά σε καθεμιά από τις επόμενες ενότητες του συγκεκριμένου κεφαλαίου.

Αρχικά, στην Ενότητα 3.1 περιγράφονται οι γενικές προδιαγραφές και λειτουργίες ενός συστήματος απεικόνισης χωρικών δεδομένων και παρουσιάζεται συναφής με το αντικείμενο της



Εικόνα 3.1: Γενική λειτουργία ενός συστήματος γεωγραφικής απεικόνισης spam απειλών

μεταπτυχιακής διατριβής εργασία, δηλαδή η απεικόνιση της απειλής του φαινομένου spamming ανά τον κόσμο. Στη συνέχεια, στην Ενότητα 3.2 γίνεται αναφορά στα χρησιμοποιούμενα δεδομένα τα οποία περιλαμβάνουν διευθύνσεις IP που αποστέλλουν μηνύματα spam, και παρουσιάζεται ο σχετικός παροχέας από τον οποίο αυτά ανακτούνται για την περίπτωση του υλοποιηθέντος συστήματος. Τέλος, στην Ενότητα 3.3 παρουσιάζεται ο τρόπος με τον οποίο γίνεται γεωπροσδιορισμός των IP διευθύνσεων που παρέχονται από τους παροχείς της προηγούμενης ενότητας.

3.1 Αρχιτεκτονική Συστημάτων Απεικόνισης Χωρικών Δεδομένων

Ένα σύστημα απεικόνισης δεδομένων-πληροφοριών γενικά αποτελεί ένα σύστημα με το οποίο συνήθως μεγάλος όγκος δεδομένων συσχετίζεται και παρουσιάζεται με διάφορους τρόπους απεικόνισης, όπως είναι οι γραφικές παραστάσεις, τα ραβδογράμματα, οι πίνακες-λίστες δεδομένων, οι σταθερές και κινούμενες (animations) εικόνες κ.α. Απώτερος δε σκοπός ενός τέτοιου συστήματος είναι η διευκόλυνση καθώς και η επιτάχυνση της διαδικασίας για την εξαγωγή συμπερασμάτων και κατά συνέπεια κατανόηση διαφόρων καταστάσεων και φαινομένων που στην περίπτωση των απλών ή «ακατέργαστων» δεδομένων θα ήταν δύσκολο να επιτευχθεί. Τα συμπεράσματα αυτά στη συνέχεια οδηγούν τις περισσότερες φορές στη λήψη

κατάλληλων περαιτέρω αποφάσεων συναφείς με το προς εξέταση αντικείμενο. Στη συγκεκριμένη δε περίπτωση που τα δεδομένα τα οποία εξετάζονται έχουν μια χωρική υπόσταση, βασικός τρόπος απεικόνισής τους αποτελεί η συσχέτισή τους με δεδομένα χαρτών. Τα συστήματα GIS αναφέρονται σε αυτήν την περίπτωση, και είναι εκείνα ουσιαστικά που καθορίζουν τις προδιαγραφές και τη δομή ενός συστήματος αναπαράστασης χωρικών δεδομένων.

3.1.1 Γεωγραφικά Συστήματα Πληροφοριών

Ένα σύστημα GIS μπορεί να ορισθεί ως ένα ισχυρό σύνολο από εργαλεία για τη συλλογή, αποθήκευση, ανάκτηση κατά βούληση, μετασχηματισμό και απεικόνιση χωρικών δεδομένων του πραγματικού κόσμου για κάποιους συγκεκριμένους σκοπούς [11]. Ωστόσο, για να αποσυνδεθεί από την στενή έννοια των εργαλείων - προγραμμάτων (συνήθως εμπορικών), όπως είναι τα MapGuide, ArcGIS κ.α., θα μπορούσε να θεωρηθεί ως μια αρχιτεκτονική συστήματος η οποία γενικά περιλαμβάνει εισαγωγή, αποθήκευση, επεξεργασία και εξαγωγή χωρικής πληροφορίας. Το βασικότερο δε χαρακτηριστικό αποτελεί η χρήση μιας εκτενούς βάσης δεδομένων που αποθηκεύει όλες αυτές τις πληροφορίες και τις διαθέτει συνδυάζοντάς τες αναμεταξύ τους προς προβολή ή επεξεργασία σε οποιαδήποτε από τις διεργασίες που αποτελούν το ευρύτερο σύστημα GIS. Όσον αφορά την απεικόνιση της χωρικής πληροφορίας, επεξεργασμένης και μη, βασικό στοιχείο αποτελεί η χρήση επιπέδων χαρτών οι οποίοι περιλαμβάνουν είτε γεωγραφική πληροφορία (π.χ. γεωλογικοί χάρτες, οριοθέτηση χωρών, πόλεις κ.α.) είτε πληροφορία σχετικά με κάποιο φαινόμενο προς ανάλυση η οποία έχει χωρική υπόσταση (π.χ. πληθυσμός, ξενοδοχεία, μόλυνση περιβάλλοντος κ.α.). Συνολικά, οι διεργασίες που αποτελούν ένα σύστημα GIS μπορούν να συνοψισθούν ως εξής:

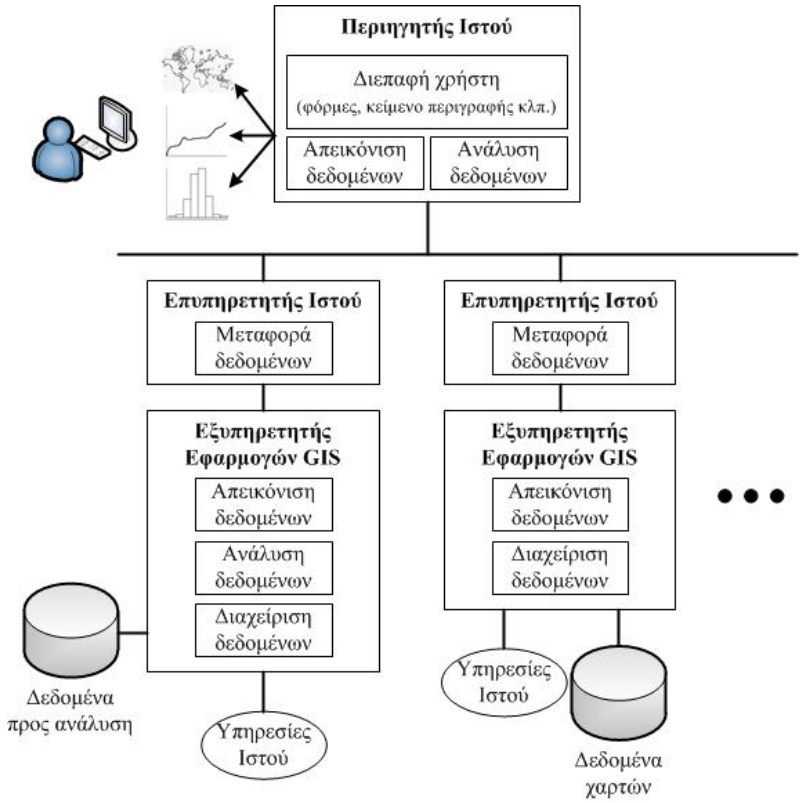
- *Απεικόνιση δεδομένων*, όπου εκτός από γραφικές παραστάσεις και ραβδογράμματα συνήθως δημιουργούνται εικόνες από τα δεδομένα που παρουσιάζονται σε συνδυασμό με κατάλληλους χάρτες, ενώ παρέχονται δυνατότητες ελέγχου της τελικής απεικόνισης στο χρήστη μέσω διεπαφών που επιτρέπουν από απλές λειτουργίες, όπως μεγέθυνση/σμίκρυνση, μέχρι και πιο σύνθετες, όπως τρισδιάστατη απεικόνιση και εφέ κίνησης.

- *Μεταφορά δεδομένων*, όπου ανάλογα με την υποδομή του συστήματος GIS που χρησιμοποιείται (π.χ. διαδικτυακή διεπαφή χρήστη ή συγκεκριμένο λογισμικό) καθορίζεται και ο τρόπος με τον οποίο μεταφέρονται τα δεδομένα στον τελικό χρήστη.

- *Ανάλυση δεδομένων*, όπου εξετάζονται τα δεδομένα που είναι αποθηκευμένα μέσω της υποβολής κατάλληλων επερωτήσεων αναφορικά είτε με τα χωρικά χαρακτηριστικά τους είτε με εκείνα που τα περιγράφουν, ώστε να προκύψουν χρήσιμα συμπεράσματα.

- *Διαχείριση δεδομένων*, όπου περιλαμβάνονται όλες εκείνες οι λειτουργίες που αφορούν την αποθήκευση, την ανάκτηση και τη διαχείριση των βάσεων δεδομένων, οι οποίες εξασφαλίζουν την εγκυρότητα των τελευταίων ως προς τα δεδομένα που περιλαμβάνουν, καθώς και την απόδοσή τους ως προς την αναζήτηση των δεδομένων αυτών.

Όταν σε ένα σύστημα GIS χρησιμοποιείται μια διαδικτυακή διεπαφή χρήστη (π.χ. μέσω ενός περιηγητή ιστού (web browser)) για την απεικόνιση των δεδομένων, όπως συμβαίνει και με την περίπτωση του συστήματος που παρουσιάζεται στη συγκεκριμένη μεταπτυχιακή διατριβή, τότε πρόκειται για ένα *διαδικτυακό σύστημα GIS* (Web-based GIS). Μια βασική αρχιτεκτονική ενός τέτοιου συστήματος, λαμβάνοντας υπόψη και τις γενικές διεργασίες που περιγράφθηκαν παραπάνω, παρουσιάζεται στην Εικόνα 3.2, όπου φαίνεται ότι οι τελευταίες μπορούν να κατανέμονται σε περισσότερα του ενός σημεία του συστήματος.



Εικόνα 3.2: Βασική αρχιτεκτονική διαδικτυακού συστήματος GIS

3.1.2 Η Περίπτωση της Απεικόνισης των Spamming Δεδομένων

Όπως αναφέρθηκε και παραπάνω, το υλοποιηθέν σύστημα που παρουσιάζεται στη μεταπτυχιακή διατριβή αποτελεί στην ουσία μια περίπτωση διαδικτυακού συστήματος GIS όπου τα δεδομένα προς απεικόνιση είναι IP διευθύνσεις από τις οποίες γίνεται αποστολή μηνυμάτων spam. Μια αναλυτική εικόνα της αρχιτεκτονικής του συστήματος αυτού, περιγράφοντας επακριβώς τα λειτουργικά στοιχεία που χρησιμοποιούνται, δίνεται στην Ενότητα 5.1. Μέχρι αυτό το σημείο αρκούν η δομή του συστήματος όπως παρουσιάζεται στην Εικόνα 3.1, καθώς και τα χαρακτηριστικά που διέπουν ένα σύστημα GIS, όπως αυτά περιγράφηκαν στην προηγούμενη ενότητα, ώστε να προσδιορισθούν οι απαιτήσεις σε θέματα απεικόνισης των spamming δεδομένων προς καλύτερη κατανόηση του αντίστοιχου φαινομένου.

Αρχικά, αναφορικά με τα χωρικά δεδομένα που προκύπτουν από τον γεωπροσδιορισμό των διευθύνσεων IP που αποστέλλουν μηνύματα spam (βλ. Ενότητα 3.3), οι οποίες και πολύ πιθανόν να παρέχονται από τις μαύρες λίστες κάποιου παροχέα όπως αναφέρθηκε στην αρχή του κεφαλαίου, θα είναι πολύ σημαντικό να παρουσιάζονται πάνω σε γεωγραφικούς χάρτες, παγκόσμιους ή μη, ώστε να μπορεί να προσδιορισθεί η προέλευση των spam απειλών. Το τελευταίο, για να γίνει βέβαια πιο αποδοτικά θα πρέπει να δίνεται η δυνατότητα στο χρήστη να κάνει μεγέθυνση/σμίκρυνση στα δεδομένα του χάρτη και να περιηγείται σε κάποιον τόπο που τον ενδιαφέρει είτε αυτός αναφέρεται σε επίπεδο ηπείρου είτε χώρας είτε πόλης χωρίς να γίνεται αλλοίωση της ευκρίνειας της πληροφορίας. Το γεγονός δε ότι τα δεδομένα μπορούν να παρουσιάζουν διάφορες πυκνότητες όσον αφορά την κατανομή τους στο χώρο, πράγμα το οποίο μπορεί να προέρχεται είτε από την παρατήρηση συγκεντρωμένων απειλών spam σε κάποιο σημείο είτε από τη συγχώνευση/διαίρεσή τους κατά τη σμίκρυνση/μεγέθυνση στο σχετικό χάρτη, απαιτεί την αναπαράστασή τους με κατάλληλο τρόπο ώστε να επιτρέπεται η σχετική σύγκριση της έντασης του φαινομένου spamming κατά τόπους, όπως για παράδειγμα με τη χρήση κατάλληλα χρωματισμένων περιοχών, κλιμακούμενου μεγέθους και έντασης χρώματος κυκλικών σημείων (markers) κ.α.. Σε αυτήν την ποιότητα της απεικόνισης καθώς και στην απόδοση του τρόπου με τον οποίο γίνεται, ιδιαίτερα σε τέτοιες περιπτώσεις μεγάλου μεγέθους βάσεων δεδομένων, κυρίαρχο ρόλο παίζουν οι τεχνικές συσταδοποίησης, και για αυτό το λόγο γίνεται εκτενής αναφορά σε αυτές στο επόμενο κεφάλαιο. Εκτός βέβαια από την απεικόνιση σε χάρτες, σημαντική είναι και η εικόνα που δίνεται από την αναπαράσταση των δεδομένων σε ραβδογράμματα ή γραφικές παραστάσεις όπου παρουσιάζεται η ένταση του φαινομένου spamming με περισσότερο εποπτικό τρόπο για διάφορες περιοχές ενδιαφέροντος.

Επίσης, όπως προκύπτει από το περιεχόμενο των spamming δεδομένων καθώς και από τη φύση του αντίστοιχου φαινομένου που μελετάται, βασικό στοιχείο αποτελεί, εκτός από τον τόπο εμφάνισης μιας spam απειλής που προσδιορίζεται μέσω της IP διεύθυνσης, και ο χρόνος που εντοπίστηκε η τελευταία και καταχωρήθηκε στην μαύρη λίστα του αντίστοιχου παροχέα. Αυτή η χρονική πληροφορία στα δεδομένα προσδίδει μια επιπλέον διάσταση που θα πρέπει να αναπαρασταθεί στις προαναφερθείσες απεικονίσεις για την πληρότητα της προκύπτουσας κατανόησης του φαινομένου. Αυτό μπορεί να επιτευχθεί είτε με μια επιλεκτική απεικόνιση των χωρικών δεδομένων φιλτράροντάς τα ως προς το χρόνο εμφάνισής τους με το να δίνεται η δυνατότητα στο χρήστη επιλογής του επιθυμητού χρονικού διαστήματος, είτε με την πολλαπλή απεικόνιση σε περισσότερα του ενός χάρτες, ραβδογράμματα ή γραφικές παραστάσεις για τη σε παράθεση εξέταση της χρονικής εξέλιξης του φαινομένου spamming, είτε τέλος με την εναλλαγή εικόνων της χρονικής εξέλιξης αυτής ως εφέ κίνησης (animation).

Τέλος, συνήθης είναι και η προβολή στοιχείων σε λίστες των μεγαλύτερων απειλών όσον αφορά τη συχνότητα εμφάνισής τους στη μαύρη λίστα των διευθύνσεων IP που αποστέλλουν μηνύματα spam, αλλά προσφέρουν κυρίως ειδική ή συμπληρωματική πληροφορία και δεν συμβάλλουν ουσιαστικά στον γενικό χαρακτήρα της κατανόησης του φαινομένου του spamming. Όλες αυτές οι απαιτήσεις απεικόνισης όσον αφορά τα συγκεκριμένα δεδομένα ικανοποιούνται από διάφορα υλοποιημένα συστήματα, όπως εκείνα που αναφέρονται στην επόμενη ενότητα, και μπορούν να συνοψισθούν ως εξής:

- Απεικόνιση των εντοπισμένων spam απειλών σε γεωγραφικό χάρτη.

- Κατάλληλη αναπαράσταση της πυκνότητας κατανομής των δεδομένων στο χάρτη μέσω χρωματισμών και ποικίλων μεγεθών κυκλικών σημείων, υιοθετώντας κάποια τεχνική συσταδοποίησής τους ώστε να διατηρείται όσο το δυνατόν περισσότερο η ποιότητα της προσφερόμενης πληροφορίας κατά την περιήγηση του χρήστη.

- Απεικόνιση των δεδομένων σε ραβδογράμματα ή γραφικές παραστάσεις, ώστε να παρουσιάζεται στο χρήστη μια περισσότερο εποπτική εικόνα της κατάστασης του spamming φαινομένου.

- Παροχή στο χρήστη δυνατότητας επιλογής των δεδομένων που απεικονίζονται με βάση τον χρόνο εντοπισμού των διευθύνσεων IP που αντιπροσωπεύουν.

- Πολλαπλή απεικόνιση (σε χάρτες, ραβδογράμματα, γραφικές παραστάσεις) των δεδομένων για σύγκριση του spamming φαινομένου διαφόρων χρονικών διαστημάτων.

- Απεικόνιση της χρονικής εξέλιξης των δεδομένων σε κάποια περιοχή μέσω κατάλληλης εναλλαγής εικόνων (animation).

- Ενδεχομένως, παρουσίαση σε λίστες των στοιχείων των διευθύνσεων IP με τη μεγαλύτερη ένταση όσον αφορά την αποστολή μηνυμάτων spam.

3.1.3 Συναφής Εργασία

Στη συγκεκριμένη ενότητα παρουσιάζονται διάφορα υλοποιημένα συστήματα απεικόνισης δεδομένων τα οποία συγκαταλέγονται στην κατηγορία των διαδικτυακών συστημάτων GIS, και όπως και το υλοποιημένο σύστημα που περιγράφεται στην συγκεκριμένη μεταπτυχιακή διατριβή, σχετίζονται με τη διαχείριση, επεξεργασία και απεικόνιση δεδομένων που αφορούν το φαινόμενο spamming. Ανάμεσα στους παροχείς αυτών των συστημάτων περιλαμβάνονται δημοφιλείς παροχείς λογισμικού ασφαλείας και δικτυακών συστημάτων, πράγμα το οποίο επιβεβαιώνει το ενδιαφέρον που παρουσιάζει η μελέτη του spamming φαινομένου. Από ότι φαίνεται δε από την παρακάτω παρουσίαση των συστημάτων, οι απαιτήσεις που περιγράφηκαν στην προηγούμενη ενότητα ικανοποιούνται σε κάποιο βαθμό στα περισσότερα από αυτά.

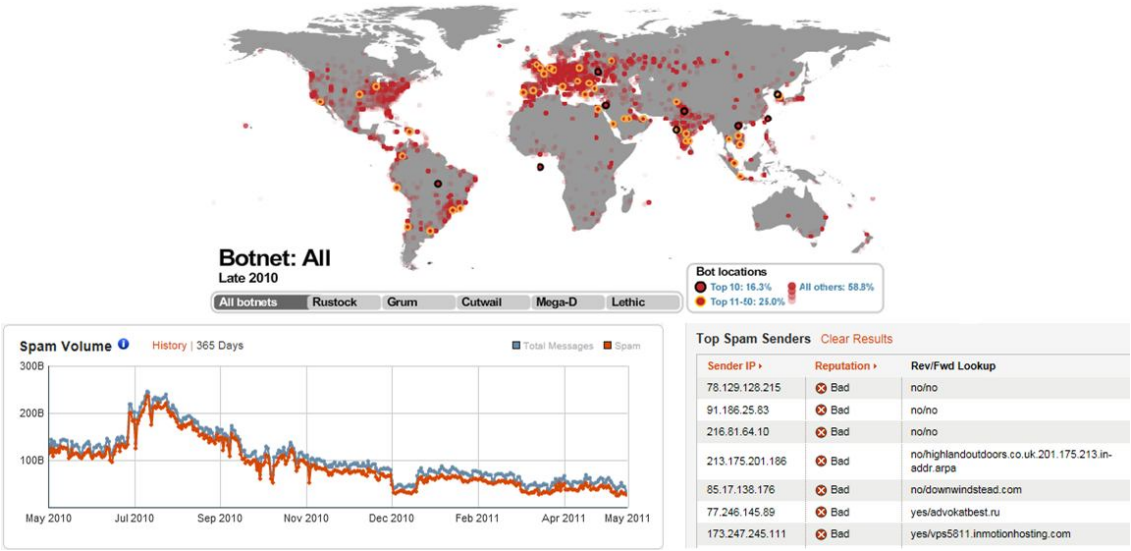
us.trendmicro.com

Πρόκειται για τον δημοφιλή παροχέα λογισμικού και γενικά συστημάτων ασφαλείας Trend Micro Inc., ο οποίος έχει υλοποιήσει χρησιμοποιώντας την τεχνολογία Flash ένα σύστημα που απεικονίζει σε χάρτη (Yahoo) spamming δεδομένα που αφορούν την τελευταία ημέρα, καθώς και σχετικές λίστες με τα ποσοστά σε μηνύματα spam ανά χώρα και ήπειρο (Εικόνα 3.3). Η αναπαράσταση αυτή σε χάρτη περιορίζεται στην απλή σημείωση των χωρών από τις οποίες έχουν σταλεί μηνύματα spam μόνο τη συγκεκριμένη ημέρα, και δεν δίνεται περαιτέρω πληροφορία σχετικά με την κατανομή των IP διευθύνσεων - αποστολέων σε αυτές.

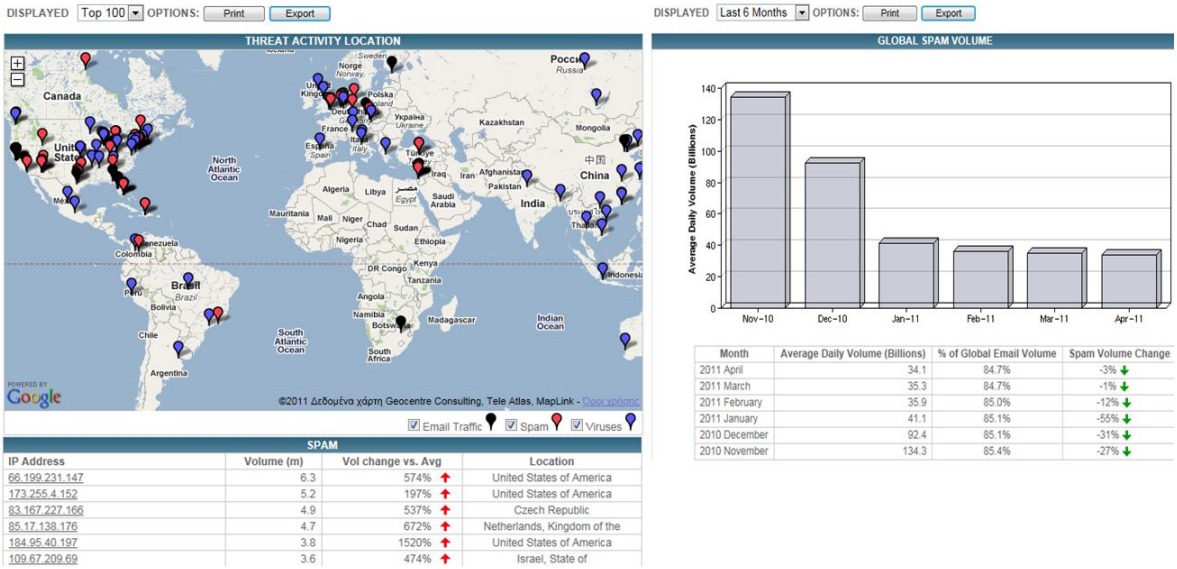
Εικόνα 3.3: Σύστημα απεικόνισης spamming δεδομένων της Trend Micro Inc.

www.messagelabs.com

Πρόκειται για τον επίσης πολύ δημοφιλή παροχέα λογισμικού ασφαλείας Symantec Co., ο οποίος απεικονίζει σε χάρτη spamming δεδομένα που σχετίζονται με τη δραστηριότητα των botnets (βλ. Ενότητα 2.4.6) σε κάποιο συγκεκριμένο χρονικό διάστημα, παρουσιάζει λίστες των σχετικών IP διευθύνσεων με την πιο έντονη αποστολή μηνυμάτων spam, καθώς περιγράφει και τη χρονική εξέλιξη της spamming δραστηριότητας σε σχετικές γραφικές παραστάσεις (Εικόνα 3.4). Στην απεικόνιση σε χάρτη, οι θέσεις των spam απειλών προσδιορίζονται με κατάλληλα κυκλικά σημεία τα οποία υποδεικνύουν και την ένταση της κάθε απειλής μέσω της έντασης του χρώματος, αλλά δεν υπάρχει δυνατότητα περιήγησης ώστε να αναζητηθεί πληροφορία σε μεγαλύτερη λεπτομέρεια.



Εικόνα 3.4: Σύστημα απεικόνισης spamming δεδομένων της Symantec Co.



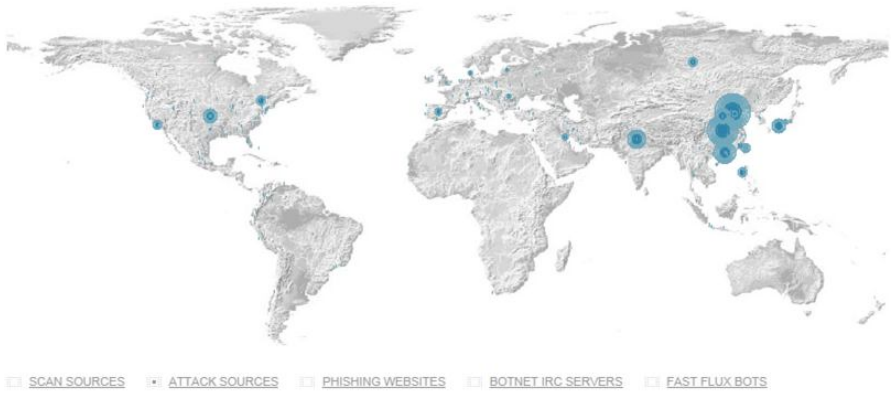
Εικόνα 3.5: Σύστημα απεικόνισης spamming δεδομένων του SenderBase

www.senderbase.org

Το SenderBase αποτελεί ένα δίκτυο πληροφοριών για την παροχή υπηρεσιών ελέγχου της φήμης ηλεκτρονικών μηνυμάτων (email reputation service) και ανήκει στην εταιρεία Cisco IronPort Systems LLC. Το σύστημα που παρουσιάζεται απεικονίζει σε χάρτη (Google) έναν αριθμό από IP διευθύνσεις με την εντονότερη αποστολή μηνυμάτων spam για τις τελευταίες 24 ώρες, οι οποίες και παραθέτονται σε σχετική λίστα, ενώ επιπλέον παρουσιάζονται σχετικά ραβδογράμματα που αφορούν τη χρονική εξέλιξη της spamming δραστηριότητας για διάφορα χρονικά διαστήματα τα οποία επιλέγει ο χρήστης (Εικόνα 3.5). Παρόλο δε που χρησιμοποιούνται οι χάρτες και οι σχετικές υπηρεσίες της Google τα οποία προσδίδουν εκτενείς δυνατότητες περιήγησης στο χρήστη, τα δεδομένα που απεικονίζονται είναι σχετικά μικρού όγκου οπότε και η προσφορά της περιήγησης είναι περιορισμένη.

atlas.arbor.net

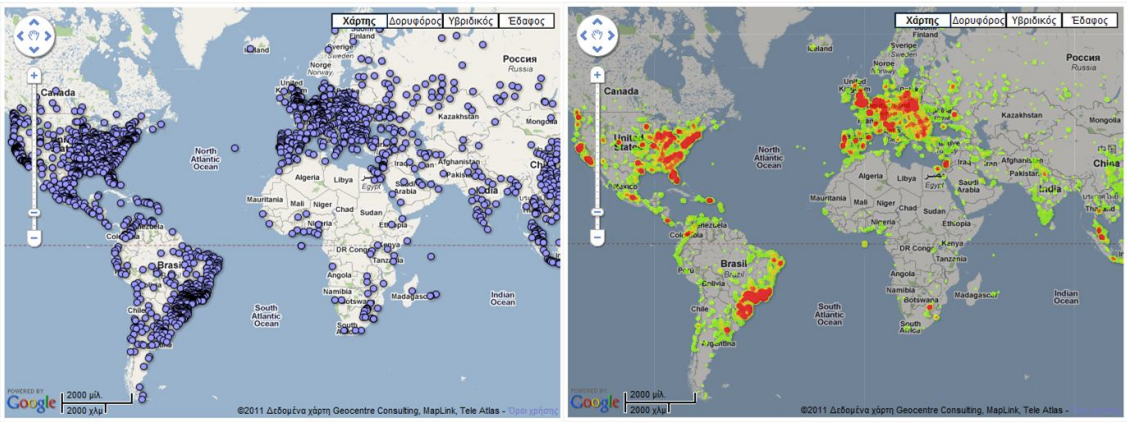
Το συγκεκριμένο σύστημα της εταιρείας Arbor Networks Inc. απεικονίζει σε χάρτη την γενικότερη κατάσταση ασφαλείας που επικρατεί στο διαδίκτυο για κάποια συγκεκριμένη χρονική περίοδο (Εικόνα 3.6). Ενώ δεν παρέχει καθόλου δυνατότητες περιήγησης στον χρήστη, αξίζει να σημειωθεί η χρήση διαφορετικού μεγέθους κυκλικών σημείων για την αναπαράσταση της έντασης των απειλών όπως αυτές κατανέμονται στο χώρο.



Εικόνα 3.6: Σύστημα απεικόνισης δεδομένων απειλών ασφαλείας στο Διαδίκτυο της Arbor Networks

www.google.com/postini/

Η Postini είναι μια υπηρεσία παροχής ασφάλειας όσον αφορά τη λήψη ηλεκτρονικών μηνυμάτων και την περιήγηση στο διαδίκτυο, και έχει εξαγορασθεί από την Google Inc.. Το σχετικό δε σύστημα που παρουσιάζεται αποτελεί μια απεικόνιση σε χάρτη (Google) των IP διευθύνσεων που αποστέλλουν μηνύματα spam σε κάποια χρονική περίοδο επιτρέποντας την πλήρη περιήγηση του χρήστη σε όλο το χάρτη (Εικόνα 3.7). Σχετικά δε με την αναπαράσταση των δεδομένων, χρησιμοποιούνται δύο τρόποι: α) ισομεγέθη κυκλικά σημεία, τα οποία δυσχεραίνουν την ευκρίνεια της πληροφορίας στα υψηλότερα επίπεδα μεγέθυνσης στο χάρτη, και β) χρωματισμένες επιφάνειες, οι οποίες εκφράζουν την ένταση του spamming φαινομένου ως προς το χώρο και προσαρμόζονται κατάλληλα στο επίπεδο μεγέθυνσης/σμίκρυνσης που επιλέγεται κάθε φορά για το χάρτη.



Εικόνα 3.7: Σύστημα απεικόνισης δεδομένων απειλών ασφαλείας στο Διαδίκτυο από www.google.com/postini/

3.2 Ανάκτηση των Spamming Δεδομένων

Όπως αναφέρθηκε και παραπάνω, ένας τρόπος με τον οποίο μπορεί να επιτευχθεί η ανάκτηση δεδομένων IP διευθύνσεων που αποστέλλουν spam, είναι μέσω της χρήσης τεχνικών αντιμετώπισης που βασίζονται στο φιλτράρισμα με βάση την προέλευση των ηλεκτρονικών μηνυμάτων και συγκεκριμένα εκείνων φυσικά των τεχνικών που χρησιμοποιούν μαύρες λίστες. Ωστόσο, όπως παρουσιάστηκε στην Ενότητα 2.5.4, ο τρόπος με τον οποίο παρέχονται οι πληροφορίες των λιστών αυτών στηρίζεται κυρίως σε μια διαδικασία ερώτησης-απάντησης σε σχετικό εξυπηρετητή του παροχέα για τη φήμη μιας IP διεύθυνσης αποστολέα. Αυτό που απαιτείται στη συγκεκριμένη περίπτωση της υλοποίησης ενός διαδικτυακού συστήματος GIS, είναι η περιοδική ανάκτηση ενός συνόλου από δεδομένα IP διευθύνσεων μιας μαύρης λίστας, και όχι μεμονωμένων, όπου κάθε εγγραφή εκτός από την ίδια την IP διεύθυνση θα πρέπει να περιλαμβάνει και σχετική πληροφορία για το χρόνο εντοπισμού και εκχώρησης της τελευταίας στη μαύρη λίστα. Μια τέτοια δυνατότητα δίνεται από τον παροχέα Heise, ο οποίος διαθέτει την μαύρη λίστα Nix Spam, και παρουσιάζεται στη συνέχεια.

3.2.1 Ο Παροχέας Heise και η Μαύρη Λίστα Nix Spam

Η μαύρη λίστα Nix Spam ουσιαστικά αποτελεί μια ενημερωμένη συλλογή από IP διευθύνσεις που αποστέλλουν μηνύματα spam η οποία χρησιμοποιείται από το φίλτρο ηλεκτρονικής αλληλογραφίας του εξυπηρετητή ηλεκτρονικού ταχυδρομείου της γερμανικής εταιρείας Heise Medien Gruppe GmbH & Co. KG [41]. Η ευρύτερη πρόσβασή της γίνεται ελεύθερα μέσω DNS από τον εξυπηρετητή ix.dnsbl.manitu.net. Βασικό δε χαρακτηριστικό της συγκεκριμένης λίστας στο οποίο διαφοροποιείται σε σχέση με εκείνες που δημιουργούνται από τους περισσότερους από τους υπόλοιπους παροχείς, πράγμα το οποίο και την κάνει ιδιαίτερα κατάλληλη για την περίπτωση του διαδικτυακού συστήματος GIS που υλοποιείται στην μεταπτυχιακή διατριβή, είναι η συνεχής ενημέρωσή της με εγγραφές που αποτελούνται από IP διευθύνσεις και χρόνους εντοπισμού των τελευταίων κατά την αποστολή μηνυμάτων spam, στοιχεία τα οποία μπορούν να περιγράφουν πλήρως την εξέλιξη του spamming φαινομένου σε πραγματικό χρόνο.

Πιο συγκεκριμένα, η μαύρη λίστα Nix Spam αποτελείται από έναν μεταβαλλόμενο αριθμό εγγραφών (μέχρι και πάνω από 500.000) οι οποίες παράγονται αυτόματα σε καθημερινή βάση, χωρίς να γίνεται διάκριση μεταξύ των διαφόρων τύπων πηγών προέλευσης μηνυμάτων spam, όπως είναι οι proxy εξυπηρετητές, οι αναμεταδότες (relays), οι δυναμικές συνδέσεις μέσω

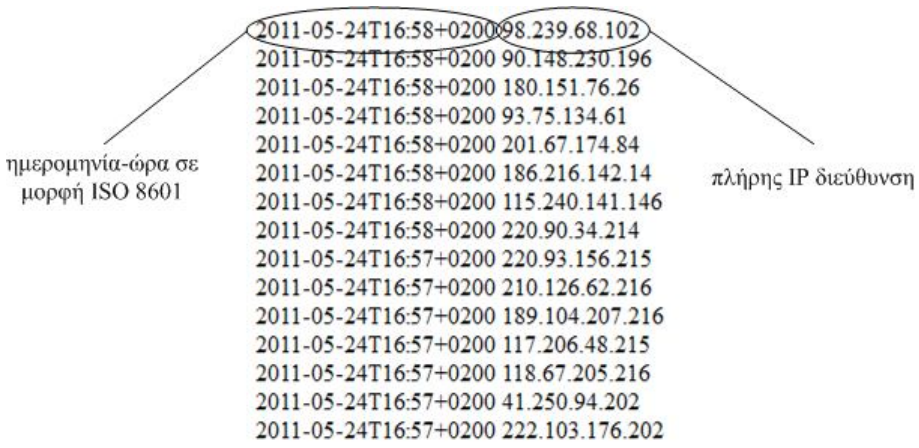
τηλεφώνου (dialup) κ.α.. Καθεμιά δε από τις εγγραφές αυτές αφαιρείται από τη λίστα αν δεν εντοπισθεί μέσα σε 12 ώρες κάποια νέα απειλή αποστολής μηνύματος spam με αποστολέα την IP διεύθυνση στην οποία αναφέρεται η συγκεκριμένη εγγραφή. Αυτός είναι και ένας από τους λόγους που οι IP διευθύνσεις που καταχωρούνται είναι μεμονωμένες και δεν αφορούν κάποιο υποδίκτυο (π.χ. /24). Με αυτόν τον τρόπο, ουσιαστικά κάθε IP διεύθυνση που περιέχεται στην μαύρη λίστα Nix Spam αποτελεί πηγή μηνυμάτων spam η οποία είναι ενεργή το πολύ τις τελευταίες 12 ώρες.

Αυτή η ιδιαιτερότητα που παρουσιάζει η συγκεκριμένη λίστα ως προς τον τρόπο ενημέρωσης και διατήρησής της οφείλεται στο γεγονός ότι βασικός στόχος του παροχέα είναι η όσο το δυνατό γρηγορότερη επεξεργασία των περιλαμβανόμενων εγγραφών. Έτσι, αντί να αποθηκεύονται εκατομμύρια εγγραφές οι οποίες απαιτούνται ώστε να θεωρηθεί ως μια πλήρως αποδοτική μαύρη λίστα spammming διευθύνσεων IP, και δεδομένου ότι οι περισσότερες από αυτές τις διευθύνσεις αναθέτονται δυναμικά στους αποστολείς spam αλλάζοντας επί το πλείστον καθημερινά, η πολιτική που ακολουθείται περιορίζει το μέγεθος της λίστας με το να αποθηκεύονται εγγραφές διευθύνσεων IP που εντοπίζονται το τελευταίο χρονικό διάστημα διαγράφοντας τις παλαιότερες. Αναφορικά δε με την απόδοση της συγκεκριμένης μαύρης λίστας ως προς τον εντοπισμό spam απειλών, φαίνεται ότι δικαιώνεται κατά κάποιο τρόπο δεδομένου του σχετικά περιορισμένου όγκου δεδομένων που περιλαμβάνει, μιας και σύμφωνα με τα στατιστικά που παρουσιάζονται στην επόμενη ενότητα καταφέρνει να εντοπίσει έως και περισσότερο από το 50% της συνολικής κίνησης spam μηνυμάτων.

Επίσης, για να περιορισθεί η πιθανότητα λανθασμένου αποκλεισμού IP διευθύνσεων που αντιστοιχούν σε κάποιο έμπιστο αποστολέα, όπως εξυπηρετητές αλληλογραφίας των Google, Yahoo, Hotmail κ.α., από κάποιον εξυπηρετητή αλληλογραφίας ο οποίος χρησιμοποιεί τη συγκεκριμένη λίστα για αναγνώριση spam απειλών, ο παροχέας έχει λάβει επιπλέον μέτρα ως εξής:

- Διατηρείται μια εσωτερική άσπρη λίστα, η οποία περιέχει IP διευθύνσεις από πολλούς κατά κάποιο τρόπο έμπιστους εξυπηρετητές αλληλογραφίας, έτσι ώστε να μην μπορούν να εκχωρηθούν οι συγκεκριμένες διευθύνσεις στην μαύρη λίστα ακόμα και αν από αυτές περιστασιακά αποστέλλονται μηνύματα spam.

- Η εκχώρηση μιας IP διεύθυνσης στην μαύρη λίστα προϋποθέτει ότι το φίλτρο αλληλογραφίας έχει εντοπίσει είτε την αποστολή περισσότερων του ενός μηνυμάτων



Εικόνα 3.8: Τμήμα του περιεχομένου του ανακτήσιμου αρχείου της μαύρης λίστας Nix Spam

spam από τη συγκεκριμένη διεύθυνση IP είτε ένα ήδη γνωστό μέχρι εκείνη τη στιγμή spam μήνυμα μέσω της παρατήρησης του αντίστοιχου αθροίσματος ελέγχου (checksum) του μηνύματος.

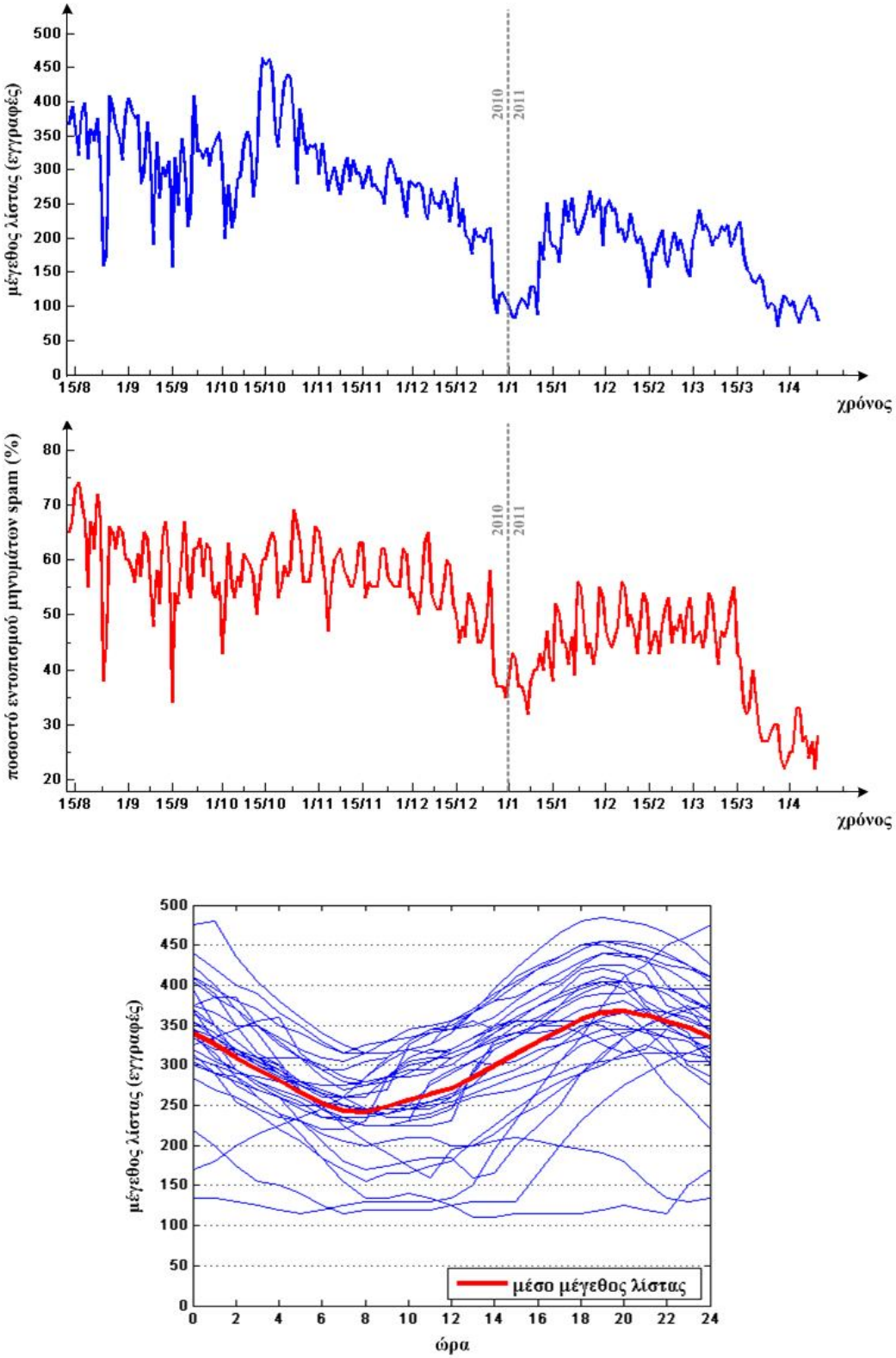
Τέλος, ο παροχέας δίνει τη δυνατότητα ανάκτησης μέσω του συνδέσμου http://www.heise.de/ix/nixspam/nixspam.blackmatches τμήματος της μαύρης λίστας στο οποίο περιλαμβάνονται 40.000 εγγραφές IP διευθύνσεων που καταχωρήθηκαν στη λίστα τις τελευταίες ώρες. Η ενημέρωση του αρχείου αυτού γίνεται κάθε 15 λεπτά, και είναι εκείνο που χρησιμοποιείται από το υλοποιηθέν σύστημα της μεταπτυχιακής διατριβής για να δημιουργήσει μια ολοκληρωμένη βάση δεδομένων με όλες εκείνες τις IP διευθύνσεις που ανιχνεύονται ως spam απειλές κατά το πέρασμα του χρόνου. Στην Εικόνα 3.8 παρουσιάζεται τμήμα του περιεχομένου του συγκεκριμένου αρχείου.

3.2.2 Στατιστικά Στοιχεία των Δεδομένων της Μαύρης Λίστας Nix Spam

Η απόδοση της χρήσης μιας μαύρης λίστας ως προς τον εντοπισμό της αποστολής μηνυμάτων spam μπορεί να μετρηθεί μέσω του αριθμού ή του ποσοστού των μηνυμάτων spam που ανιχνεύονται (spam hits) στο σύνολο εκείνων που αποστέλλονται και περνούν από το σχετικό φίλτρο ηλεκτρονικής αλληλογραφίας που διαθέτει ένας εξυπηρετητής ηλεκτρονικού ταχυδρομείου. Για την περίπτωση της Nix Spam μαύρης λίστας που εξετάζεται, ανακτήθηκαν (προσεγγιστικά, λόγω της γραφικής αναπαράστασής τους) μετρήσεις σχετικά με την εξέλιξη ως προς το χρόνο του μεγέθους της λίστας και του αντίστοιχου ποσοστού των εντοπιζόμενων μηνυμάτων spam από τον παροχέα της λίστας και συγκεκριμένα τη διεύθυνση www.dnsbl.manitu.net. Οι μετρήσεις αυτές οι οποίες αφορούν μέσες τιμές για κάθε ημέρα για το

ενδεικτικό χρονικό διάστημα 12/8/2010 έως 9/4/2011 παρουσιάζονται στην Εικόνα 3.9α. Σύμφωνα με αυτή, είναι εμφανής η συνεχής μεταβολή του μεγέθους της συγκεκριμένης μαύρης λίστας, πράγμα το οποίο οφείλεται στην παροδικότητα των εγγραφών διάρκειας 12 ωρών που υιοθετείται από τον παροχέα Heise ως πολιτική για τη διαμόρφωση του περιεχομένου της. Επίσης, φαίνεται η σχέση που υπάρχει μεταξύ του μεγέθους της μαύρης λίστας και του ποσοστού των εντοπιζόμενων μηνυμάτων spam, σύμφωνα με την οποία όταν αυξάνεται το μέγεθος της λίστας, τότε αυξάνεται και η απόδοσή της μιας και εκτός από το γεγονός ότι οι καταχωρημένες IP διευθύνσεις είναι περισσότερες, το ποσοστό των μηνυμάτων spam που αποστέλλονται και αφορά την τρέχουσα κίνηση των μηνυμάτων αυτών (εντός 12 ωρών) είναι σχετικά υψηλό.

Τέλος, από τις μετρήσεις αυτές στο σύνολό τους θα μπορούσαν να βγουν και επιπλέον συμπεράσματα σχετικά με ορισμένα χαρακτηριστικά του spamming φαινομένου και της λειτουργίας της μαύρης λίστας Nix Spam, δεδομένης της παροδικότητας και συνάμα της ακρίβειας-πιστότητας της εικόνας του spamming για κάθε χρονική στιγμή που προσφέρει η τελευταία. Ενδεικτικά, αναφέρεται η σχετική μείωση που παρατηρείται στην Εικόνα 3.9α του μεγέθους της λίστας κατά τη διάρκεια της περιόδου των Χριστουγέννων, ενώ σε επίπεδο ημέρας, όπως φαίνεται από την Εικόνα 3.9β για τον Σεπτέμβριο του 2010, η διακύμανση που υπάρχει στο μέγεθος της λίστας κατά τη διάρκεια ενός 24ώρου ενδεχομένως να μπορεί να συνδεθεί με τη γεωγραφική προέλευση της πλειοψηφίας των αποστελλόμενων μηνυμάτων spam (π.χ. ήπειρος, χώρα).



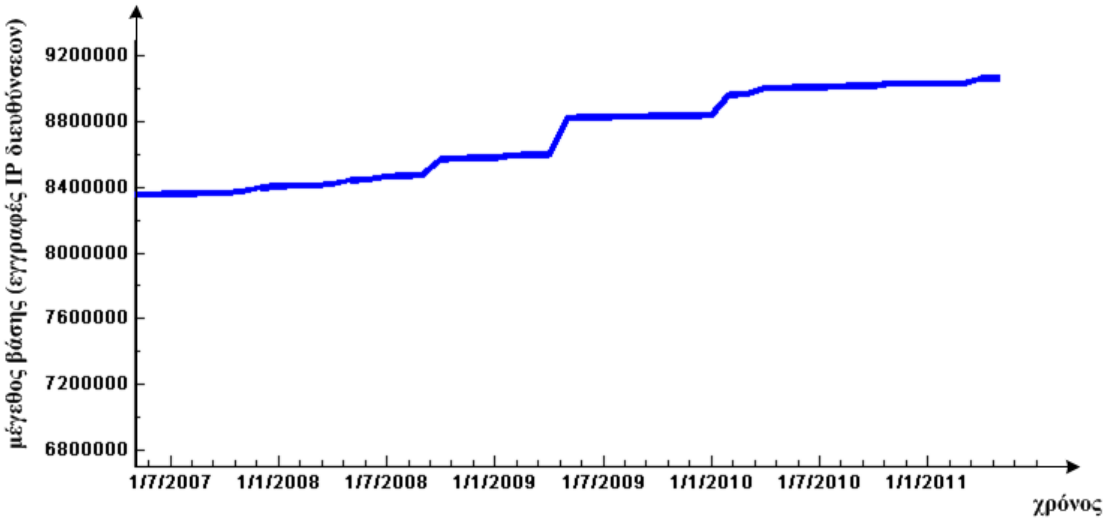
Εικόνα 3.9: Χρονική εξέλιξη της μαύρης λίστας Nix Spam: α) ως προς το μέγεθος και την απόδοσή της για το χρονικό διάστημα 12/8/2010 έως 9/4/2011, και β) ως προς το μέγεθός της ανά ημέρα για τον Σεπτέμβριο του 2010

3.3 Γεωπροσδιορισμός των Spamming Δεδομένων

Όπως αναφέρθηκε και στην αρχή του κεφαλαίου, ο γεωπροσδιορισμός των δεδομένων που παρέχονται από την μαύρη λίστα ενός παροχέα είναι απαραίτητος ώστε να δοθεί εκείνη η χωρική υπόσταση η οποία θα καθορίσει τις αναφερόμενες IP διευθύνσεις που αποστέλλουν μηνύματα spam γεωγραφικά εντοπίσιμες σε κάποιο χάρτη. Έτσι, τα δεδομένα θα μετατραπούν σε χωρο-χρονικά δεδομένα και κατά συνέπεια θα είναι ικανά να περιγράψουν την αντίστοιχη χωρο-χρονική φύση του spamming φαινομένου.

Ο γεωπροσδιορισμός των IP διευθύνσεων χρησιμοποιείται σε αρκετές εφαρμογές, μεταξύ άλλων στην προσαρμογή του περιεχομένου των ιστοσελίδων ανάλογα με την τοποθεσία από την οποία γίνεται η πρόσβαση ενός χρήστη, στην μελέτη και αξιολόγηση υπηρεσιών και προϊόντων σχετικά με θέματα μάρκετινγκ, σε περιπτώσεις θεμάτων ασφαλείας διαδικτυακών συναλλαγών κ.α.. Για το λόγο αυτό, πολλοί είναι οι παροχείς που διατηρούν μια βάση αντιστοίχισης IP διευθύνσεων με πληροφορίες που σχετίζονται με κάποια τοποθεσία, όπως χώρα, πόλη, ταχυδρομικό κώδικα, γεωγραφικές συντεταγμένες και ζώνη ώρας, και είτε την παρέχουν μέσω υποβολής HTTP ερωτημάτων για γεωπροσδιορισμό μεμονωμένων IP διευθύνσεων είτε διαθέτουν ολόκληρη τη σχετική βάση δεδομένων η οποία μπορεί να ενημερώνεται περιοδικά. Για την τελευταία δε περίπτωση όπου διατίθεται ολόκληρη η βάση δεδομένων, ανάλογα με το προσφερόμενο επιτευχθέν ποσοστό κάλυψης των IP διευθύνσεων, πολλοί είναι εκείνοι οι παροχείς οι οποίοι συνήθως απαιτούν κάποιο ποσό πληρωμής για την πλήρη διάθεση των δεδομένων της βάσης, διαφορετικά οι διατιθέμενες πληροφορίες είναι σημαντικά περιορισμένες (π.χ. αντιστοίχιση IP διεύθυνσης μόνο με χώρα, σημαντικά μειωμένο εύρος κάλυψης IP διευθύνσεων κ.α.). Τέτοιοι παροχείς είναι οι MaxMind Inc. [60] και IP2Location [45].

Στη μεταπτυχιακή διατριβή, ο παροχέας που χρησιμοποιείται για τον γεωπροσδιορισμό των IP διευθύνσεων που ανακτούνται από τη μαύρη λίστα Nix Spam, η οποία παρουσιάστηκε στην προηγούμενη ενότητα, είναι ο HostIP (www.hostip.info). Πρόκειται για έναν παροχέα ο οποίος προσφέρει δυνατότητες γεωπροσδιορισμού τόσο μέσω υποβολής HTTP ερωτημάτων όσο και με την ελεύθερη διάθεση ολόκληρης της βάσης δεδομένων αντιστοίχισης IP διευθύνσεων με γεωγραφική πληροφορία που χρησιμοποιεί, με το τελευταίο να είναι ο κύριος λόγος επιλογής του παροχέα HostIP στην μεταπτυχιακή διατριβή. Αναφορικά δε με τον τρόπο με τον οποίο ενημερώνεται η συγκεκριμένη βάση δεδομένων που διατηρεί ο παροχέας, αυτό γίνεται εν μέρει από τους χρήστες αυτής της παρεχόμενης υπηρεσίας, για αυτό και σε πολλές περιπτώσεις IP διευθύνσεων ο συγκεκριμένος παροχέας ενδεχομένως να είναι περισσότερο αξιόπιστος από



Εικόνα 3.10: Μεταβολή του μεγέθους της HostIP βάσης δεδομένων με το πέρασμα του χρόνου

άλλους. Σχετικά με την δομή των αποθηκευμένων δεδομένων, η αντιστοίχιση των IP διευθύνσεων με χωρικά στοιχεία και συγκεκριμένα με κάποια πόλη, το οποίο είναι και το θεμελιώδες στοιχείο κβαντοποίησης του χώρου, γίνεται σε ομάδες-υποδίκτυα /24, ενώ συνολικά οι πληροφορίες που συνοδεύουν κάθε IP διεύθυνση είναι: πόλη, χώρα, κωδικός χώρας, γεωγραφικό πλάτος, γεωγραφικό μήκος και πολιτεία. Στην Εικόνα 3.10 παρουσιάζεται ο ρυθμός αύξησης του αριθμού των εγγραφών γεωπροσδιορισμού IP διευθύνσεων της βάσης δεδομένων σύμφωνα με τη σχετική πληροφορία που συνοδεύει κάθε εγγραφή για τον χρόνο εισαγωγής της στη βάση.

Κεφάλαιο 4
Απεικόνιση Δεδομένων – Συσταδοποίηση

Η διαδικασία κατά την οποία επιχειρείται η ομαδοποίηση ενός συνόλου από αντικείμενα δεδομένων, φυσικά και μη, σε ομάδες ή *συστάδες* (clusterss) από συναφή μεταξύ τους αντικείμενα ως προς κάποια χαρακτηριστικά τους ονομάζεται *συσταδοποίηση* (clustering). Η διαδικασία αυτή αποτελεί ουσιαστικά μια από τις περισσότερο σημαντικές τεχνικές εξόρυξης γνώσης από δεδομένα (data-mining techniques), όπως μεταξύ άλλων η *κατηγοριοποίηση* (classification) και η αναζήτηση *κανόνων συσχέτισης* (association rules). Θα μπορούσε να ειπωθεί ότι η συσταδοποίηση μοιάζει αρκετά με την κατηγοριοποίηση μιας και οι δύο τεχνικές κατανέμουν τα αντικείμενα των δεδομένων σε διαφορετικές ομάδες. Ωστόσο, σε αντίθεση με την τελευταία, στην περίπτωση της συσταδοποίησης οι ομάδες αυτές δεν είναι προκαθορισμένες, πράγμα το οποίο είναι επιθυμητό τις περισσότερες φορές μιας και με τον τρόπο αυτό πρώτα αναζητούνται τα περισσότερο όμοια αντικείμενα μεταξύ τους μέσω κατάλληλων *μετρικών ομοιότητας* (similarity measures), και έπειτα περιγράφονται από τα χαρακτηριστικά τους σαν ομάδες πλέον. Επιπλέον, δεδομένης της ραγδαίας ανάπτυξης συστημάτων όπου επιχειρείται εξόρυξη γνώσης από δεδομένα ποικίλων εφαρμογών, όπως στη γεωλογία, τη βιολογία, το

περιβάλλον, το μάρκετινγκ, τη στατιστική κ.α., τα οποία κατά κύριο λόγο προέρχονται από μεγάλες βάσεις δεδομένων, η χρησιμοποίηση της τεχνικής της συσταδοποίησης είναι ιδιαίτερα συνήθης λόγω της συμπίεσης των αντικειμένων των δεδομένων σε πιο σύνθετα αντικείμενα που προσφέρει. Στην περίπτωση δε που τα δεδομένα τα οποία συμμετέχουν σε μια συσταδοποίηση περιγράφονται από χωρική-γεωμετρική πληροφορία (χωρικά δεδομένα), τότε η συσταδοποίηση ονομάζεται *χωρική συσταδοποίηση* (spatial clustering), αντικείμενο το οποίο πραγματεύεται η μεταπτυχιακή διατριβή.

Πιο συγκεκριμένα, στο παρόν κεφάλαιο μελετώνται διάφορες κατηγορίες μεθόδων συσταδοποίησης περιγράφοντας τα βήματα που ακολουθούνται από ορισμένες αντιπροσωπευτικές μεθόδους των κατηγοριών αυτών. Στόχος δεν είναι η εκτεταμένη ανάλυση και εμβάθυνση στις μεθόδους συσταδοποίησης, αλλά η ανάδειξη ορισμένων στοιχείων που χαρακτηρίζουν τις μεθόδους αυτές και συμβάλλουν στην επίτευξη της συσταδοποίησης δεδομένων, ώστε να χρησιμοποιηθούν για τη δημιουργία ενός μηχανισμού συσταδοποίησης που να ταιριάζει στις ανάγκες απεικόνισης του υλοποιηθέντος συστήματος. Έτσι, αρχικά, η Ενότητα 4.1 αναφέρεται στην έννοια της χωρικής συσταδοποίησης, καθώς και στις μετρικές ομοιότητας που χρησιμοποιούνται για τη συσταδοποίηση των αντικειμένων των δεδομένων. Έπειτα, παρουσιάζεται στην Ενότητα 4.2 μια κατηγοριοποίηση των μεθόδων συσταδοποίησης που έχουν αναπτυχθεί, ενώ σε καθεμιά από τις επόμενες ενότητες (4.3, 4.4, 4.5, 4.6) περιγράφονται ως προς τον τρόπο που γίνεται η συσταδοποίηση διάφορες αντιπροσωπευτικές μέθοδοι των παραπάνω κατηγοριών. Τέλος, στην Ενότητα 4.7 μελετάται και συγκρίνεται η απόδοση των περιγραφόμενων μεθόδων ως προς τη συσταδοποίηση δεδομένων με αναφορά στις ανάγκες απεικόνισης του υλοποιηθέντος συστήματος.

4.1 Χωρική Συσταδοποίηση

Όπως αναφέρθηκε και παραπάνω, η έννοια της χωρικής συσταδοποίησης αναφέρεται στην κατανομή αντικειμένων χωρικών δεδομένων. Η δε αντιπροσωπευτική κατηγορία συστημάτων που πραγματεύονται κατά κύριο λόγο χωρικά δεδομένα και κατά συνέπεια σχετίζονται με την έννοια της χωρικής συσταδοποίησης είναι τα *Γεωγραφικά Συστήματα Πληροφοριών* (Geographic Information Systems - GIS) που αναφέρθηκαν και στο προηγούμενο κεφάλαιο. Σκοπός των συστημάτων αυτών είναι η γρήγορη και ουσιαστική γραφική απεικόνιση των συνήθως πολυπληθών χωρικών δεδομένων με τέτοιο τρόπο ώστε να διευκολύνεται η εξαγωγή

συμπερασμάτων και γνώσης για τα χαρακτηριστικά και τη συμπεριφορά των ευρύτερων φαινομένων που περιγράφονται.

Δεδομένων των χωρικών χαρακτηριστικών αυτών των δεδομένων, η σχετική συσταδοποίηση που χρησιμοποιείται ανάγει το πρόβλημα της αναζήτησης των περισσότερο όμοιων μεταξύ τους αντικειμένων των δεδομένων σε αναζήτηση των γεωμετρικά κοντινότερων. Κάτι ανάλογο συμβαίνει και με τις μετρικές που ορίζουν την ομοιότητα δύο αντικειμένων, οι οποίες βασίζονται στην έννοια της απόστασης που υπολογίζεται χρησιμοποιώντας ως συντεταγμένες-διαστάσεις τα μέτρα των χαρακτηριστικών των αντικειμένων των δεδομένων. Για το λόγο αυτό και για τις μετρικές ομοιότητας αυτές ταιριάζει περισσότερο η έννοια των *δεικτών εγγύτητας* (proximity indexes). Πιο συγκεκριμένα, για δύο αντικείμενα $\hat{x}$ και $\hat{y}$, και n το πλήθος των χαρακτηριστικών τους, τέτοιοι δείκτες είναι:

- *Απόσταση Minkowski*, όπου q είναι φυσικός αριθμός:

$$d(\hat{x}, \hat{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^q \right)^{\frac{1}{q}} \quad (1)$$

- Ευκλείδεια απόσταση:

$$d(\hat{x}, \hat{y}) = \sqrt{\left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)} \quad (2)$$

- Απόσταση Manhattan (ή City-block):

$$d(\hat{x}, \hat{y}) = \sum_{i=1}^{n} |x_i - y_i| \quad (3)$$

- Μέγιστη απόσταση (ή Chebychev):

$$d(\hat{x}, \hat{y}) = max_{i=1}^{n} |x_i - y_i| \quad (4)$$

4.2 Κατηγοριοποίηση Μεθόδων Συσταδοποίησης

Υπάρχουν ποικίλες μέθοδοι που χρησιμοποιούνται για τη συσταδοποίηση πολυπληθών συνόλων δεδομένων. Αυτές διαφοροποιούνται ως προς την ποιότητα και την μορφή των τελικών παραγόμενων συστάδων, την ταχύτητα ολοκλήρωσης της διαδικασίας συσταδοποίησης, καθώς και τον τρόπο και τις μετρικές (δείκτες) που χρησιμοποιούνται για την εύρεση της ομοιότητας μεταξύ των αντικειμένων των δεδομένων. Για το λόγο αυτό, η εφαρμογή κάθε μεθόδου ταιριάζει συνήθως σε σύνολα δεδομένων με συγκεκριμένα κάθε φορά μεγέθη και χαρακτηριστικά. Όπως υποδεικνύεται και στη σχετική βιβλιογραφία [39], τέσσερις βασικές κατηγορίες μεθόδων συσταδοποίησης αποτελούν οι εξής:

- *Μέθοδοι διαμερισμού* (partitioning methods): Για ένα σύνολο δεδομένων n αντικειμένων, μια μέθοδος διαμερισμού ουσιαστικά δημιουργεί k επιμέρους συστάδες αντικειμένων, όπου για κάθε συστάδα ικανοποιείται με το βέλτιστο τρόπο ένα συγκεκριμένο κριτήριο συσταδοποίησης. Το κριτήριο αυτό αφορά τη σχέση-ομοιότητα των αντικειμένων που αποτελούν την κάθε συστάδα, και βέβαια, θα πρέπει να προσδιορίζει μια κατάσταση όπου τα συσταδοποιημένα αντικείμενα θα πρέπει να είναι όσο το δυνατό πιο «όμοια». Έτσι, όταν χρησιμοποιείται η απόσταση ως δείκτης εγγύτητας, ουσιαστικά η ομοιότητα αυτή εκφράζεται με το πόσο γεωμετρικά κοντά είναι δύο αντικείμενα, δεδομένων των χαρακτηριστικών τους ως συντεταγμένες διαστάσεων (βλ. Ενότητα 4.3).

  Το πρόβλημα δε που αντιμετωπίζουν οι συγκεκριμένες μέθοδοι είναι η υψηλή πολυπλοκότητα που απαιτούν για να καταλήξουν σε μια βέλτιστη λύση, δεδομένου ότι πρέπει να εξαντλήσουν όλους τους δυνατούς συνδυασμούς αντικειμένων που μπορούν να σχηματίσουν συστάδες. Για το λόγο αυτό, τα αντικείμενα προς συσταδοποίηση είναι περιορισμένα σε πλήθος, ενώ για να επιταχυνθεί η όλη διαδικασία συνήθως επιλέγεται μια αρχική κατάσταση συσταδοποιήσεων και στη συνέχεια εφαρμόζονται οι μέθοδοι αυτές ώστε να γίνουν οι απαραίτητες διορθώσεις.

- *Ιεραρχικές μέθοδοι* (hierarchical methods): Οι μέθοδοι αυτές δημιουργούν μια ιεραρχική δομή των αντικειμένων των δεδομένων, και ανάλογα με τη φορά που εκτελείται η σχετική διαδικασία δημιουργίας της δομής αυτής, χωρίζονται σε δύο επιμέρους κατηγορίες (βλ. Ενότητα 4.4):

i. *Συσσωρευτικές μέθοδοι* (agglomerative methods), όπου αρχικά κάθε αντικείμενο αποτελεί από μόνο του μια ξεχωριστή συστάδα, και στη συνέχεια συγχωνεύονται επαναληπτικά μεταξύ τους (bottom-up προσέγγιση) δημιουργώντας νέες συστάδες με κριτήριο την ομοιότητά τους (ή την απόστασή τους) είτε μέχρι να συγχωνευθούν όλα τα αντικείμενα σε μία συστάδα είτε μέχρι να ικανοποιηθεί μια συγκεκριμένη συνθήκη τερματισμού της διαδικασίας.

ii. *Διαιρετικές μέθοδοι* (divisive methods), όπου αρχικά όλα τα αντικείμενα αποτελούν την ίδια συστάδα, η οποία στη συνέχεια διαιρείται σε μικρότερες συστάδες (top-down προσέγγιση) μέχρι τελικά κάθε αντικείμενο να αποτελεί από μόνο του μια συστάδα, ή μέχρι να ικανοποιηθεί μια αντίστοιχη συνθήκη τερματισμού.

Το βασικό πρόβλημα των μεθόδων αυτών είναι ότι από τη στιγμή που γίνει μια συγχώνευση ή διαχωρισμός, τα τελευταία δεν αναιρούνται ούτε διορθώνονται. Έτσι, το γεγονός αυτό ναι μεν οδηγεί στην μείωση των απαιτούμενων υπολογισμών μιας και περιορίζεται ο αριθμός των συνδυασμών διαφορετικών λύσεων που μπορούν να επιλεγούν, αλλά ενδεχόμενες λανθασμένες αποφάσεις-συνδυασμοί συγχώνευσης ή διαχωρισμού να μη διορθώνονται και η τελική λύση να μην είναι βέλτιστη.

- *Μέθοδοι βασισμένες στην πυκνότητα* (density-based methods): Οι μέθοδοι αυτές χρησιμοποιούν την πυκνότητα της κατανομής των αντικειμένων, όπως αυτή προσδιορίζεται από την ομοιότητά τους με αντικείμενα στη γύρω περιοχή («γειτονιά» τους), για να δημιουργήσουν συστάδες. Έτσι, μια συστάδα ξεκινάει από ένα αντικείμενο και συνεχίζει να διευρύνεται διαρκώς απορροφώντας νέα αντικείμενα των οποίων η πυκνότητα υπερβαίνει κάποιο όριο, ή πιο συγκριμένα, σε μια δεδομένη «ακτίνα ομοιότητας» γύρω από αυτά περιέχεται τουλάχιστον ένας ελάχιστος αριθμός από άλλα αντικείμενα (βλ. Ενότητα 4.5).

- *Μέθοδοι βασισμένες σε πλέγμα* (grid-based methods): Βασικός στόχος των μεθόδων αυτών είναι η κβαντοποίηση του συνόλου των δεδομένων σε ένα συγκεκριμένο αριθμό από κελιά (cells), τα οποία σχηματίζουν μια δομή πλέγματος. Έτσι, δε γίνονται ανταλλαγές αντικειμένων μεταξύ των κελιών, και τα τελευταία αντιμετωπίζονται ως ξεχωριστά σύνθετα αντικείμενα τα οποία μπορούν να συμμετέχουν σε περαιτέρω συσταδοποιήσεις με μεθόδους άλλων κατηγοριών. Το κύριο δε πλεονέκτημα αυτής της προσέγγισης είναι ότι μειώνεται ο χρόνος των απαιτούμενων υπολογισμών, δεδομένου

ότι εξαρτάται από τον αριθμό των δημιουργηθέντων κελιών και όχι από τον αριθμό των αρχικών αντικειμένων των δεδομένων (βλ. Ενότητα 4.6).

- *Μέθοδοι βασισμένες σε μοντέλο* (model-based methods): Οι μέθοδοι της κατηγορίας αυτής χρησιμοποιούν κάποιο μοντέλο για να περιγράψουν καθεμιά από τις επιθυμητές τελικές συστάδες αντικειμένων, και προσπαθούν να βρουν μια βέλτιστη προσαρμογή των δεδομένων στο συγκεκριμένο αυτό μοντέλο. Για το σκοπό αυτό, χρησιμοποιούνται είτε στατιστικές προσεγγίσεις είτε προσεγγίσεις παρόμοιες με εκείνες άλλων κατηγοριών μεθόδων συσταδοποίησης, όπως η χρήση της πυκνότητας για την εύρεση της κατανομής των αντικειμένων ως προς τις διαστάσεις-χαρακτηριστικά τους. Παραδείγματα τέτοιων μεθόδων αποτελούν οι EM [21], AUTOCLASS [15] και COBWEB [32].

Στις επόμενες ενότητες παρουσιάζονται αναλυτικότερα ορισμένες μέθοδοι συσταδοποίησης που ανήκουν στις πρώτες τέσσερις προαναφερθείσες κατηγορίες μιας και ταιριάζουν πιο πολύ στις απαιτήσεις που θέτει η εφαρμογή τους στην περίπτωση της χωρικής συσταδοποίησης που εξετάζεται στην μεταπτυχιακή διατριβή.

4.3 Μέθοδοι Διαμερισμού

Όπως αναφέρθηκε και στην προηγούμενη ενότητα, οι μέθοδοι διαμερισμού λαμβάνουν ως είσοδο ένα σύνολο δεδομένων από n αντικείμενα καθώς και έναν αριθμό k ($k \leq n$) ο οποίος δηλώνει από πόσες συστάδες θα πρέπει να αποτελείται η τελική λύση της συσταδοποίησης. Έπειτα, βασιζόμενες σε κάποιο κριτήριο ομοιότητας και συγκεκριμένα της απόστασης μεταξύ των αντικειμένων των δεδομένων, οι μέθοδοι διαμερισμού, προσπαθούν με επαναληπτική επανατοποθέτηση να επιτύχουν μια βέλτιστη κατανομή των τελευταίων σε αυτές τις k συστάδες, έτσι ώστε σε καθεμιά από αυτές τα αντικείμενα να είναι περισσότερο όμοια μεταξύ τους από ότι με τα αντικείμενα άλλων συστάδων.

Μερικές από αυτές τις μεθόδους, οι οποίες εφαρμόζονται στην περίπτωση της απεικόνισης σημείων πάνω σε χάρτη που πραγματεύεται η συγκεκριμένη μεταπτυχιακή διατριβή, παρουσιάζονται στη συνέχεια.

4.3.1 K-means

Η μέθοδος **k-means** [59] αποτελεί μια από τις πιο γνωστές και απλές μεθόδους που επιλύουν το πρόβλημα της συσταδοποίησης. Σαν μέθοδος είχε προηγουμένως παρουσιασθεί με ορισμένες διαφορές σε σχετικές αναφορές [33, 56]. Σκοπός της είναι να σχηματίσει συγκεκριμένο αριθμό k συστάδων από ένα σύνολο δεδομένων από n αντικείμενα, έτσι ώστε να ελαχιστοποιηθεί η παρακάτω συνάρτηση τετραγωνικού σφάλματος εκλαμβανόμενη ως ένδειξη της ομοιότητας των αντικειμένων ανά συστάδα στο σύνολό τους:

$$e = \sum_{i=1}^{k} \sum_{j=1}^{size(C_i)} \left\| \hat{p}_{ij} - \hat{c}_i \right\|^2 \quad (5)$$

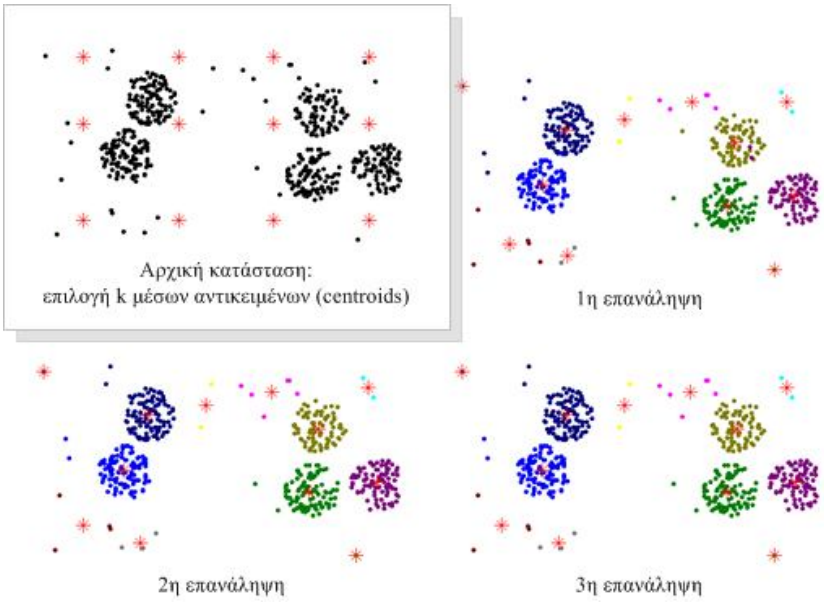
, όπου C_i είναι η i συστάδα από τις k, $\hat{p}_{ij}$ είναι το j αντικείμενο της συστάδας C_i, $\hat{c}_i$ είναι το μέσο αντικείμενο (κέντρο μάζας ή centroid) ως προς τις διαστάσεις-χαρακτηριστικά των αντικειμένων που περιλαμβάνονται στη C_i συστάδα, και $\left\| \hat{p}_{ij} - \hat{c}_i \right\|$ είναι η απόσταση ως το μέτρο ομοιότητας των δύο αντικειμένων, $\hat{p}_{ij}$ και $\hat{c}_i$. Πιο συγκεκριμένα, το μέσο αντικείμενο $\hat{c}_i$ της συστάδας C_i είναι ένα διάνυσμα που περιέχει τις μέσες τιμές των χαρακτηριστικών των αντικειμένων της συστάδας, δηλαδή $\hat{c}_i = \sum_{j=1}^{size(C_i)} \left(\hat{p}_{ij} / size(C_i) \right)$, ενώ ως απόσταση μπορεί να χρησιμοποιηθεί οποιοσδήποτε από τους δείκτες εγγύτητας στην Ενότητα 4.1.

Η εξέλιξη της συγκεκριμένης μεθόδου παρουσιάζεται στην Εικόνα 4.1 και τα βήματα που ακολουθούνται είναι τα εξής:

1. Αρχικά, επιλέγονται k αντικείμενα ως τα αρχικά μέσα αντικείμενα των k συστάδων, χωρίς να είναι απαραίτητο να αποτελούν και πραγματικά αντικείμενα των δεδομένων.

2. Κάθε αντικείμενο των δεδομένων ανατίθεται στη συστάδα εκείνη όπου το αντίστοιχο μέσο αντικείμενο βρίσκεται στην μικρότερη απόσταση σε σχέση με τα μέσα αντικείμενα των άλλων συστάδων. Για κάθε δε αντικείμενο που ανατίθεται, υπολογίζεται για κάθε συστάδα το αντίστοιχο νέο μέσο αντικείμενο από το σύνολο αυτών που της ανήκουν.

3. Επαναλαμβάνεται το βήμα 2 μέχρι να μην υπάρχει αλλαγή στις διαστάσεις-χαρακτηριστικά των k μέσων αντικειμένων.



Εικόνα 4.1: Εξέλιξη της μεθόδου k-means για k=12 συστάδες

Η k-means μέθοδος τερματίζει πάντα, αλλά οι τελικές k συστάδες (εκφραζόμενες από τα τελικά μέσα αντικείμενα) δεν αποτελούν πάντα τη βέλτιστη λύση. Αυτό οφείλεται στο γεγονός ότι οι τελικές συστάδες βασίζονται τόσο στην αρχική επιλογή για τον αριθμό των συστάδων k όσο και στα αρχικά επιλεχθέντα μέσα αντικείμενα. Στην περίπτωση δε όπου τα τελευταία δεν αποτελούν αντικείμενα δεδομένων, μπορεί ακόμα να είναι και άδειες μερικές από τις συστάδες. Επίσης, σημαντική είναι και η επιρροή των ακραίων αντικειμένων (outliers) στην τελική λύση, διότι όντας αντικείμενα απομακρυσμένα σε σχέση με τα υπόλοιπα σε μια συστάδα, συμβάλλουν αρνητικά στον υπολογισμό των μέσων αντικειμένων (η επιρροή οξύνεται δεδομένης της χρήσης των τετραγώνων στο κριτήριο ελαχιστοποίησης της συνάρτησης τετραγωνικού σφάλματος της σχέσης (5)).

Δεδομένων των παραπάνω προβλημάτων που παρουσιάζει η μέθοδος k-means, διάφορες παραλλαγές της μεθόδου έχουν προταθεί στη βιβλιογραφία. Αυτές είτε βελτιώνουν την αρχική επιλογή των μέσων αντικειμένων και του αριθμού συστάδων [7, 9, 47, 55, 77, 83] είτε αυξάνουν την ταχύτητα της k-means μεθόδου [66, 79] είτε βελτιώνουν την ακρίβειά της [78] είτε ακόμα ελέγχουν την πολυπλοκότητά της με το να χρησιμοποιούν διαφορετικά μέτρα ομοιότητας (δείκτες εγγύτητας) [101] και τρόπους υπολογισμού των μέσων αντικειμένων [62], ενώ ενδέχεται να περιορίζουν και την εφαρμογή της σε χαμηλό αριθμό συστάδων.

4.3.2 K-medoids και PAM

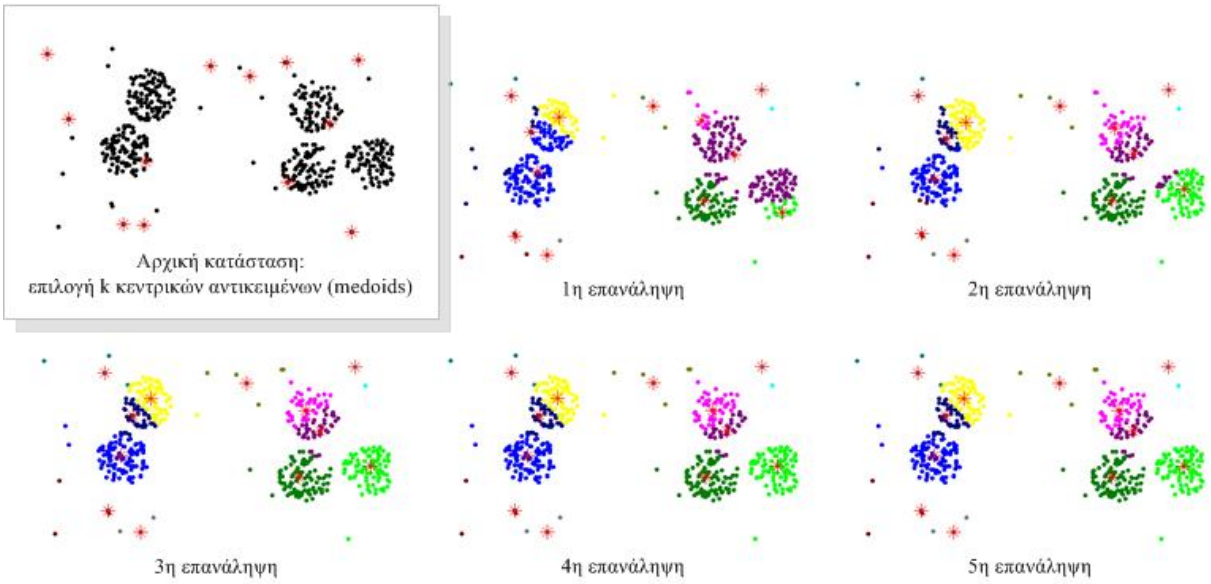
Μια παραλλαγή ουσιαστικά της k-means μεθόδου αποτελούν και οι **k-medoids** μέθοδοι [48, 49]. Η βασική διαφορά έγκειται στο γεγονός ότι τα αντικείμενα που αντιπροσωπεύουν τις συστάδες των τελευταίων δεν είναι υπολογιζόμενα, όπως είναι τα μέσα αντικείμενα στην k-means μέθοδο, αλλά πραγματικά αντικείμενα των δεδομένων των συστάδων, τα *medoids*. Πιο συγκεκριμένα, κάθε ένα από αυτά αποτελεί το πιο κεντρικά τοποθετημένο αντικείμενο για κάθε συστάδα, με αποτέλεσμα η επιρροή των ακραίων αντικειμένων (outliers) στη δημιουργία των συστάδων να περιορίζεται βελτιώνοντας την ποιότητα της συσταδοποίησης. Για το λόγο αυτό, οι k-medoids μέθοδοι είναι περισσότερο κατάλληλοι για την ειδική περίπτωση της χωρικής συσταδοποίησης που μελετάται.

Όσον αφορά την εύρεση των medoids από το σύνολο των αντικειμένων των δεδομένων, η γενική διαδικασία που ακολουθείται βασίζεται στην μεγιστοποίηση της συνολικής ομοιότητας τού κάθε αντικειμένου με το αντίστοιχο medoid στην κάθε συστάδα. Αυτό εκφράζεται μέσω της ελαχιστοποίησης της συνάρτησης απόλυτου σφάλματος:

$$e = \sum_{i=1}^{k} \sum_{j=1}^{size(C_i)} \left\| \hat{p}_{ij} - \hat{m}_i \right\| \qquad (6)$$

, όπου C_i είναι η i συστάδα από τις k, $\hat{p}_{ij}$ είναι το j αντικείμενο της συστάδας C_i, $\hat{m}_i$ είναι το medoid αντικείμενο ως προς τις διαστάσεις-χαρακτηριστικά των αντικειμένων που περιλαμβάνονται στη C_i συστάδα, και $\left\| \hat{p}_{ij} - \hat{m}_i \right\|$ είναι η απόσταση ενός αντικειμένου μιας συστάδας, $\hat{p}_{ij}$, με το αντίστοιχο medoid της, $\hat{m}_i$.

Τα βήματα γενικά που ακολουθούνται, σε μια απλή εκδοχή, είναι τα ίδια με την περίπτωση της k-means μεθόδου (και συγκεκριμένα της εκδοχής κατά Forgy [33]) με τη διαφορά ότι αντί για τα υπολογιζόμενα μέσα αντικείμενα, εκείνα που συμμετέχουν στους υπολογισμούς και στο τελικό κριτήριο τερματισμού της διαδικασίας συσταδοποίησης είναι τα medoids αντικείμενα. Πιο συγκεκριμένα, για κάθε επανάληψη, στο πρώτο βήμα ανατίθενται τα αντικείμενα των δεδομένων σε συστάδες βρίσκοντας για κάθε ένα από αυτά το πιο κοντινό medoid, ενώ στο επόμενο βήμα, εφόσον έχει γίνει η ανάθεση όλων των αντικειμένων σε συστάδες, επαναϋπολογίζονται τα medoids αναζητώντας για κάθε συστάδα (C_i) εκείνο το αντικείμενο ($\hat{p}_m$) του οποίου το σύνολο των αποστάσεων από τα υπόλοιπα αντικείμενα $\left(\hat{p}_{ij}\right)$ της συστάδας,



Εικόνα 4.2: Εξέλιξη μιας απλής μεθόδου k-medoids για k=12 συστάδες

$d = \sum_{j=1}^{size(C_i)} \|\hat{p}_{ij} - \hat{p}_m\|$, ελαχιστοποιείται. Αυτός ο επαναϋπολογισμός των medoids είναι βέβαια και η αιτία που οι k-medoids μέθοδοι παρουσιάζουν μεγάλο αριθμό υπολογισμών. Ένα παράδειγμα της εξέλιξης της διαδικασίας συσταδοποίησης με βάση τα βήματα αυτά φαίνεται στην Εικόνα 4.2.

Μια από τις χρονικά πρώτες k-medoids μεθόδους αποτελεί η **PAM** (Partitioning Around Medoids) [49]. Σύμφωνα με αυτή, ο επαναϋπολογισμός των k medoids αντικειμένων βασίζεται σε μια απλή διαδικασία επαναληπτικής εναλλαγής αντικειμένων στο ρόλο των medoids (μεταξύ αυτών που είναι medoids και ορισμένων από τα υπόλοιπα αντικείμενα των δεδομένων που δεν είναι), η οποία αποσκοπεί στον να επιφέρει ολοένα και καλύτερα αποτελέσματα όσον αφορά την συσταδοποίηση. Όπως ισχύει γενικά για τις k-medoids μεθόδους, κριτήριο για την καλύτερη επιλογή αντικειμένων ως medoids είναι η ελαχιστοποίηση του αθροίσματος των αποστάσεων του κάθε αντικειμένου από το πιο κοντινό του medoid που εκφράζεται από τη σχέση (6). Με βάση αυτό το κριτήριο ορίζεται μια έννοια κόστους εναλλαγής το οποίο καθορίζει για το αν μια συγκεκριμένη επιλογή αντικειμένων ως medoids βελτιώνει ή χειροτερεύει τη συσταδοποίηση.

Δεδομένου του μεγάλου αριθμού υπολογισμών που απαιτείται για την εύρεση των medoids, η PAM, σχετικά με το πρώτο βήμα των k-medoids μεθόδων, για να επιταχύνει τη διαδικασία δημιουργεί μια αρχική συσταδοποίηση με το να βρίσκει ένα-ένα εκείνα τα k αντικείμενα που ικανοποιούν το κριτήριο της ελαχιστοποίησης της σχέσης (6). Έτσι, το πρώτο αντικείμενο που επιλέγεται είναι το medoid όλων των αντικειμένων των δεδομένων, το δεύτερο είναι εκείνο που

ελαχιστοποιεί τη σχέση (6) για την περίπτωση δύο συστάδων δεδομένου του πρώτου medoid, κοκ μέχρι την εύρεση και του k-οστού medoid. Έπειτα, η PAM προσπαθεί να βελτιώσει τις αρχικές επιλογές ώστε να επιτευχθεί καλύτερη συσταδοποίηση, εναλλάσσοντας κάθε φορά ένα medoid με ένα μη medoid αντικείμενο, εκ των οποίων το τελευταίο να ελαχιστοποιεί το κόστος της εναλλαγής αυτής ως προς τη σχέση (6). Στην Εικόνα 4.3 παρουσιάζεται ένα παράδειγμα της εξέλιξης της μεθόδου όσον αφορά την αρχική εύρεση k medoids αντικειμένων, τη βελτίωση της συσταδοποίησης με την εναλλαγή medoids με μη medoids αντικειμένων, και την μεταβολή του ελάχιστου κόστους εναλλαγής από μια επανάληψη στην επόμενη. Συνολικά, τα βήματα που ακολουθούνται είναι τα εξής:

1. Αρχικά, επιλέγονται ένα-ένα k αντικείμενα για να αποτελέσουν τα αρχικά medoids, όπου το πρώτο είναι το medoid του συνόλου των αντικειμένων των δεδομένων και για την εύρεση του m-στού ακολουθούνται τα εξής βήματα:

   1.1. Για κάθε αντικείμενο $\hat{p}_i$ που δεν είναι medoid:

   - Υπολογίζεται η διαφορά D_{ji} μεταξύ της απόστασης κάθε αντικειμένου $\hat{p}_j$ $(i \neq j)$ από το κοντινότερο medoid και της απόστασης $\|\hat{p}_j - \hat{p}_i\|$ από το αντικείμενο $\hat{p}_i$.

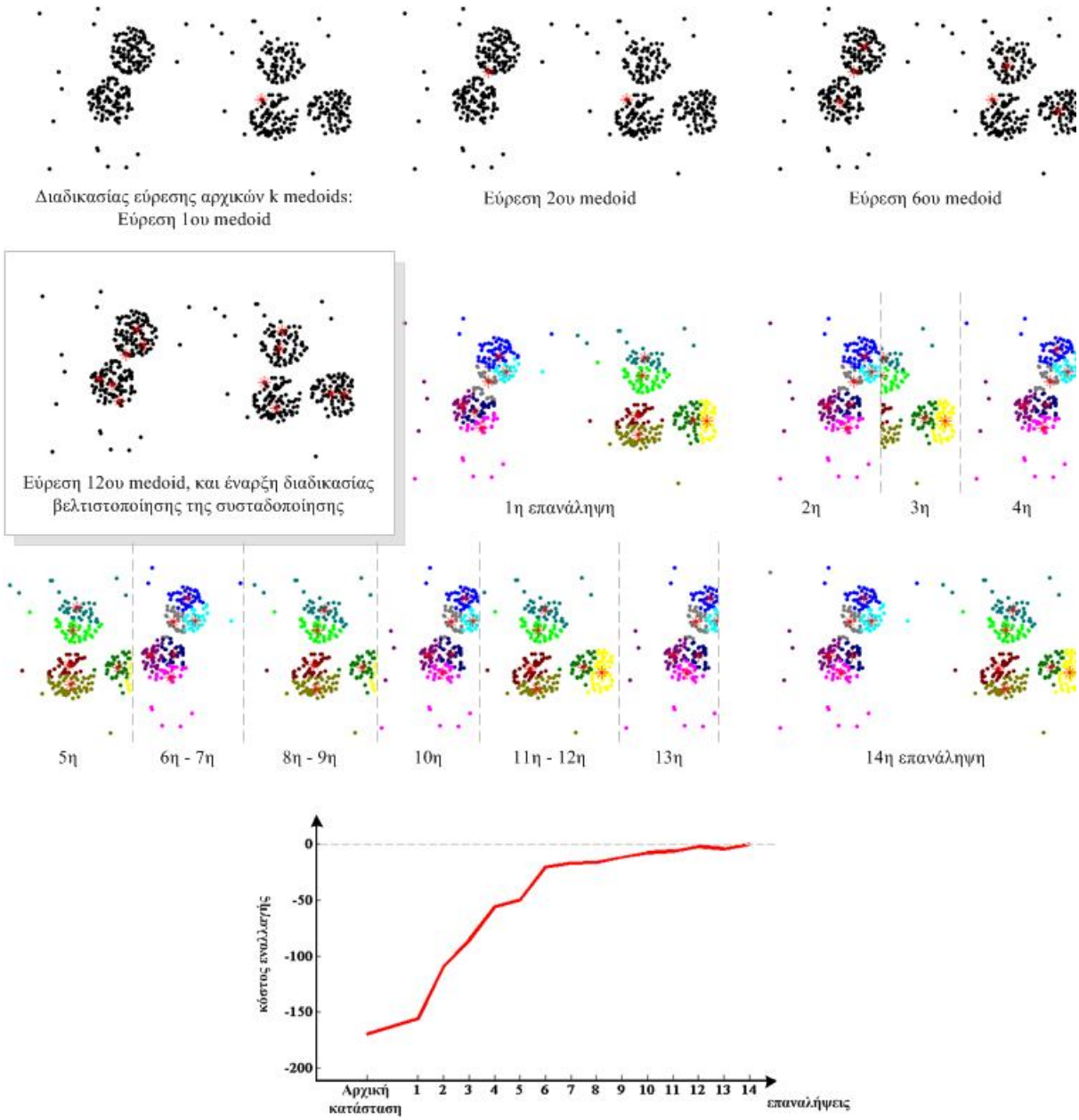
   - Υπολογίζεται το άθροισμα $\sum D_{ji}$ των θετικών διαφορών D_{ji} (> 0) κάθε αντικειμένου $\hat{p}_j$.

   1.2. Επιλέγεται ως medoid εκείνο το αντικείμενο $\hat{p}_i$ για το οποίο το άθροισμα των διαφορών $\sum D_{ji}$ μεγιστοποιείται.

2. Για κάθε αντικείμενο $\hat{p}_i$ που δεν είναι medoid υπολογίζεται η επίδραση στην ποιότητα της συσταδοποίησης μιας ενδεχόμενης εναλλαγής του με κάποιο medoid αντικείμενο $\hat{p}_m$:

   2.1. Υπολογίζεται το μερικό κόστος C_{jmi} τα οποίο επιφέρει η εναλλαγή αυτή ως προς κάθε άλλο μη medoid αντικείμενο $\hat{p}_j$ ως εξής:

   - Αν υπάρχει κάποιο άλλο medoid $\hat{p}_o$ $(\neq \hat{p}_m)$ για το οποίο $\|\hat{p}_j - \hat{p}_o\| < \|\hat{p}_j - \hat{p}_m\|$ και $\|\hat{p}_j - \hat{p}_o\| < \|\hat{p}_j - \hat{p}_i\|$, τότε $C_{jmi} = 0$.



Εικόνα 4.3: Εξέλιξη της PAM μεθόδου ως προς την αρχική επιλογή k=12 medoids, τη βελτίωση της συσταδοποίησης με εναλλαγή αντικειμένων, και την μεταβολή του ελάχιστου κόστους εναλλαγής σε κάθε επανάληψη

- Αν για κάθε άλλο medoid $\hat{p}_o$ $(\neq \hat{p}_m)$ ισχύει $\|\hat{p}_j - \hat{p}_o\| \leq \|\hat{p}_j - \hat{p}_m\|$, και για το δεύτερο πιο κοντινό στο $\hat{p}_j$ medoid $\hat{p}_{sec}$ ισχύει $\|\hat{p}_j - \hat{p}_{sec}\| > \|\hat{p}_j - \hat{p}_i\|$, τότε $C_{jmi} = \|\hat{p}_j - \hat{p}_i\| - \|\hat{p}_j - \hat{p}_m\|$.

- Αν για κάθε άλλο medoid $\hat{p}_o$ $(\neq \hat{p}_m)$ ισχύει $\|\hat{p}_j - \hat{p}_o\| \leq \|\hat{p}_j - \hat{p}_m\|$, και για το δεύτερο πιο κοντινό στο $\hat{p}_j$ medoid $\hat{p}_{sec}$ ισχύει $\|\hat{p}_j - \hat{p}_{sec}\| \leq \|\hat{p}_j - \hat{p}_i\|$, τότε $C_{jmi} = \|\hat{p}_j - \hat{p}_{sec}\| - \|\hat{p}_j - \hat{p}_m\|$.

- Αν υπάρχει medoid $\hat{p}_o$ $(\neq \hat{p}_m)$ ώστε $\|\hat{p}_j - \hat{p}_o\| < \|\hat{p}_j - \hat{p}_m\|$, και για κάθε medoid $\hat{p}_e$ ισχύει $\|\hat{p}_j - \hat{p}_i\| \leq \|\hat{p}_j - \hat{p}_e\|$, τότε $C_{jmi} = \|\hat{p}_j - \hat{p}_i\| - \|\hat{p}_j - \hat{p}_o\|$.

2.2. Υπολογίζεται το συνολικό κόστος $TC_{mi} = \sum C_{jmi}$ της εναλλαγής του αντικειμένου $\hat{p}_i$ με το $\hat{p}_m$ ως medoid.

3. Επιλέγεται εκείνη η εναλλαγή η οποία ελαχιστοποιεί το συνολικό κόστος TC_{mi}.

4. Επαναλαμβάνονται τα βήματα 2 και 3 μέχρι το ελάχιστο συνολικό κόστος TC_{mi} της βέλτιστης εναλλαγής να είναι θετικό ή μηδέν.

4.3.3 CLARA και CLARANS

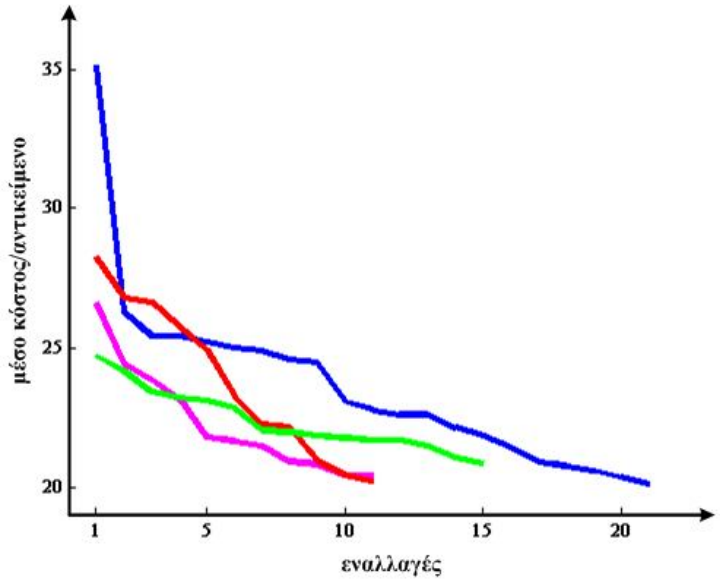
Η PAM μέθοδος προσφέρει υψηλής ποιότητας συσταδοποίηση δεδομένου ότι συγκρίνει κάθε δυνατό συνδυασμό από k medoids αντικείμενα από τους οποίους επιλέγει τον καλύτερο. Η απόδοσή της, όμως, είναι καλή μόνο στην περίπτωση όπου το σύνολο των αντικειμένων των δεδομένων είναι σχετικά μικρό, και αυτό λόγω των υψηλών επιπέδων χρήσης μνήμης και του μεγάλου αριθμού υπολογισμών που απαιτείται για την εύρεση του καλύτερου συνδυασμού. Για να ελαττωθούν αυτές οι απαιτήσεις σε μνήμη και υπολογιστική πολυπλοκότητα, ιδιαίτερα όσον αφορά μεγάλα σύνολα αντικειμένων των δεδομένων, διάφορες άλλες μέθοδοι βασιζόμενες στην PAM έχουν αναπτυχθεί. Δύο από τις πιο γνωστές, οι οποίες στηρίζονται στη *δειγματοληψία* (sampling) του αρχικού συνόλου δεδομένων, παρουσιάζονται στην συνέχεια.

Η μέθοδος **CLARA** (Clustering LARge Applications) [48, 50], για να αντιμετωπίσει τα προβλήματα της PAM, αντί να συμπεριλάβει στους υπολογισμούς των k medoids ολόκληρο το σύνολο των αντικειμένων των δεδομένων, επιλέγει κάποια αντιπροσωπευτικά δείγματα. Σε αυτά τα περιορισμένα σε μέγεθος δείγματα στη συνέχεια εφαρμόζεται η απαιτητική μέθοδος PAM για την εύρεση των medoids, ενώ σχετικά με την μεγιστοποίηση της ομοιότητας μεταξύ αντικειμένων και των αντίστοιχων medoids που επιχειρείται, αυτή αφορά ολόκληρο το σύνολο των αντικειμένων των δεδομένων. Σαν τελική λύση επιλέγεται μεταξύ των συσταδοποιήσεων που προκύπτουν από κάθε δείγμα εκείνη με την καλύτερη ποιότητα, η οποία μετριέται μέσω της μέσης απόστασης κάθε αντικειμένου του συνόλου των δεδομένων (και όχι αποκλειστικά του κάθε δείγματος) από το κοντινότερό του medoid.

Δεδομένης της δειγματοληψίας στα αντικείμενα των δεδομένων που χρησιμοποιεί η CLARA, ενδέχεται κάποια από τα αντικείμενα που θα μπορούσαν να χαρακτηρισθούν ως βέλτιστες medoid επιλογές να μην περιλαμβάνονται στα δείγματα. Αυτό έχει ως αποτέλεσμα να είναι σχετικά δύσκολο να «τύχει» να βρεθεί η καλύτερη δυνατή συσταδοποίηση, η οποία τελικά εξαρτάται σε μεγάλο βαθμό από το μέγεθος του κάθε δείγματος (μεγαλύτερα δείγματα⇒ μεγαλύτερη πιθανότητα εύρεσης βέλτιστης λύσης⇒ αύξηση υπολογιστικής πολυπλοκότητας και χρήσης μνήμης). Σύμφωνα με τη σχετική αναφορά [50], ικανοποιητική επιλογή αποτελεί η χρήση πέντε δειγμάτων των $40 + 2k$ αντικειμένων.

Ως μια λύση στα προβλήματα μεταξύ ποιότητας και απόδοσης της συσταδοποίησης που παρουσιάζει η CLARA, έχει αναπτυχθεί η μέθοδος **CLARANS** (Clustering Large Applications based on RANdomized Search) [67, 68] η οποία ταιριάζει αρκετά στις ανάγκες της εξόρυξης χωρικών δεδομένων. Γενικά, η διαδικασία συσταδοποίησης που ακολουθείται μπορεί να αναπαρασταθεί ως μια επαναληπτική αναζήτηση σε ένα γράφο όπου κάθε κόμβος αποτελεί μια λύση από k medoids, ενώ δύο κόμβοι που είναι γειτονικοί (συνδέονται με μια ακμή) μεταξύ τους διαφέρουν κατά ένα αντικείμενο. Τα δε δείγματα του συνόλου των αντικειμένων των δεδομένων που χρησιμοποιούνται δεν είναι συγκεκριμένα, όπως στην CLARA, αλλά επιλέγονται ουσιαστικά τυχαία σε κάθε βήμα της αναζήτησης. Πιο συγκεκριμένα, για κάθε δεδομένη συσταδοποίηση-κόμβο αναζητούνται με τυχαίο τρόπο γειτονικοί κόμβοι οι οποίοι βελτιώνουν τη συσταδοποίηση που προκύπτει καταλήγοντας (πιθανώς) σε ένα τοπικό ελάχιστο όσον αφορά το σχετικό «κόστος», όπως ονομάζεται στη συγκεκριμένη μέθοδο το αποτέλεσμα της σχέσης (6) δεδομένου ενός επιλεγμένου συνόλου από k medoids. Υπάρχουν, βέβαια, δύο παράμετροι που καθορίζουν τη διάρκεια της αναζήτησης αυτής: ο *αριθμός τοπικών ελαχίστων* προσδιορίζει πόσες φορές θα γίνει η αναζήτηση για κάποιο τοπικό ελάχιστο, ενώ ο *μέγιστος αριθμός εξεταζόμενων γειτόνων* προσδιορίζει τον μέγιστο αριθμό των γειτονικών κόμβων που μπορούν να εξετασθούν ως προς το κόστος της συσταδοποίησης που παρέχουν μέχρι να βρεθεί κάποιο τοπικό ελάχιστο.

Λαμβάνοντας υπόψη την περίπτωση του συνόλου των αντικειμένων των δεδομένων των προηγούμενων παραδειγμάτων (Εικόνες 4.1-4.3), στην Εικόνα 4.4 παρουσιάζεται ένα παράδειγμα της εξέλιξης της μεθόδου όσον αφορά τέσσερις επαναλήψεις (αριθμός τοπικών ελαχίστων = 4) ως προς το μέσο κόστος ανά αντικείμενο για κάθε εναλλαγή κάποιου medoid με κάποιο από τα υπόλοιπα αντικείμενα. Όπως φαίνεται, εξαιτίας της τυχαίας επιλογής δειγμάτων αντικειμένων για εναλλαγή με τα τρέχοντα medoids, επηρεάζεται τόσο η εξέλιξη της εύρεσης μιας τοπικά βέλτιστης λύσης όσο και η ποιότητα (κόστος για την CLARANS) της τελικής συσταδοποίησης για κάθε επανάληψη. Σχετικά με τα βήματα της μεθόδου, αυτά είναι τα εξής:



Εικόνα 4.4: Εξέλιξη 4 επαναλήψεων της CLARANS με k=12 συστάδες και μέγιστο αριθμό εξεταζόμενων γειτόνων = 16

1. Αρχικά, καθορίζονται ο αριθμός τοπικών ελαχίστων και ο μέγιστος αριθμός εξεταζόμενων γειτόνων.

2. Επιλέγεται μια αρχική τυχαία λύση-κόμβος από k medoids.

3. Επιλέγεται τυχαία μια από τις γειτονικές λύσεις-κόμβους i της τρέχουσας λύσης m, και υπολογίζεται το κόστος εναλλαγής TC_{mi} όπως στην περίπτωση της PAM μεθόδου.

4. Αν το κόστος εναλλαγής είναι αρνητικό, επιλέγεται η γειτονική λύση ως τρέχουσα (τοπικό ελάχιστο), ενημερώνεται το κόστος της αντίστοιχης συσταδοποίησης και επαναλαμβάνεται το βήμα 3.

5. Αν το κόστος εναλλαγής είναι θετικό ή μηδέν, επαναλαμβάνεται το βήμα 3 εκτός και αν έχει εξετασθεί ο μέγιστος αριθμός εξεταζόμενων γειτόνων για την τρέχουσα λύση.

6. Επιλέγεται η λύση εκείνης της επανάληψης με την καλύτερη ποιότητα συσταδοποίησης (⇒ ελάχιστο κόστος).

7. Επαναλαμβάνονται τα βήματα 1 έως 6 τόσες φορές όσες υποδεικνύει ο αριθμός τοπικών ελαχίστων.

4.3.4 Πλησιέστερων Γειτόνων

Η μέθοδος **πλησιέστερων γειτόνων** (nearest neighbor clustering) [57], ενώ χρησιμοποιεί την τεχνική απλού συνδέσμου (βλ. Ενότητα 4.4.1) για τον καθορισμό της απόστασης μεταξύ δύο συστάδων και συνεπώς θα μπορούσε να αποτελέσει μια ιεραρχική μέθοδο (προσέγγιση bottom-up), παρ' όλα αυτά, συγκαταλέγεται στην κατηγορία των μεθόδων διαμερισμού δεδομένου ότι παρουσιάζει μια *ενός-περάσματος διαδικασία συσταδοποίησης* (single-pass clustering) ή διαφορετικά ενός επιπέδου ιεραρχική συσταδοποίηση. Πιο συγκεκριμένα, επιχειρείται με επαναληπτικό τρόπο μια φορά για κάθε αντικείμενο του συνόλου των δεδομένων η ανάθεσή του στην πλησιέστερη με βάση την τεχνική απλού συνδέσμου συστάδα από τις προηγουμένως δημιουργηθείσες. Ωστόσο, η ανάθεση αυτή προϋποθέτει ότι η αντίστοιχη απόσταση μεταξύ των πλησιέστερων αντικειμένων είναι μικρότερη από κάποιο προκαθορισμένο κατώφλι (threshold), διαφορετικά το αντικείμενο ανατίθεται σε μια νέα συστάδα της οποίας αρχικά αποτελεί και μοναδικό περιεχόμενο. Αυτό το κατώφλι ουσιαστικά καθορίζει και τον αριθμό των τελικών δημιουργηθέντων συστάδων, όπου η αύξησή του προκαλεί τη μείωση του αριθμού των τελευταίων, όπως φαίνεται στην Εικόνα 4.5 για την περίπτωση της συγκεντρωμένης κατανομής των αντικειμένων των προηγούμενων παραδειγμάτων καθώς και μιας πλήρως ομοιόμορφης κατανομής των ίδιων αντικειμένων στην ίδια περιοχή (εδώ 500x300). Σχετικά με τα βήματα που ακολουθούνται, αυτά είναι τα εξής:

1. Αρχικά, καθορίζεται το κατώφλι t για την ελάχιστη απόσταση μεταξύ δύο αντικειμένων ώστε να ανήκουν στην ίδια συστάδα.

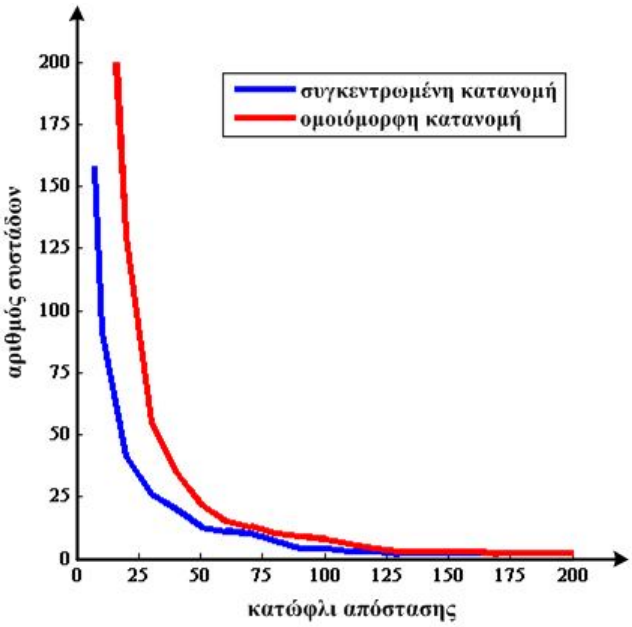
2. Επιλέγεται κάποιο αντικείμενο $\hat{p}_1$ το οποίο αποτελεί και την πρώτη συστάδα.

3. Για κάθε αντικείμενο $\hat{p}_i$ από τα υπόλοιπα, αναζητείται το πλησιέστερο αντικείμενο $\hat{p}_j$ το οποίο ανήκει σε κάποια συστάδα C_j.

4. Αν $\|\hat{p}_i - \hat{p}_j\| \leq t$, τότε το $\hat{p}_i$ ανατίθεται στην συστάδα C_j.

5. Αν $\|\hat{p}_i - \hat{p}_j\| > t$, τότε το $\hat{p}_i$ ανατίθεται σε μια νέα συστάδα C_{new}.

6. Επαναλαμβάνονται τα βήματα 3 έως 5 μέχρι να εξετασθούν όλα τα αντικείμενα $\hat{p}_i$ στο αντίστοιχο σύνολο των δεδομένων.



Εικόνα 4.5: Αριθμός των τελικών συστάδων σε σχέση με το κατώφλι απόστασης για την περίπτωση της μεθόδου συσταδοποίησης των πλησιέστερων γειτόνων για δύο κατανομές αντικειμένων

Στη συγκεκριμένη μέθοδο εκτός από την τεχνική απλού συνδέσμου, θα μπορούσε να χρησιμοποιηθεί οποιαδήποτε άλλη από τις τεχνικές συνδέσμου που χρησιμοποιούνται στις ιεραρχικές μεθόδους (βλ. Ενότητα 4.4.1). Για παράδειγμα, οι τεχνικές συνδέσμου βασισμένες είτε στην απόσταση των κέντρων βάρους είτε στην απόσταση των medoids παράγουν συσταδοποιήσεις παρόμοιες με αυτές των k-means και k-medoids, αντίστοιχα, οι οποίες ταιριάζουν πιο πολύ στην περίπτωση της απεικόνισης χωρικών δεδομένων (π.χ. σφαιρικές συστάδες). Επίσης, δεδομένης της μικρής σχετικά πολυπλοκότητας που προσφέρει η μέθοδος αυτή ως ενός-περάσματος διαδικασία συσταδοποίησης, καθώς και του γεγονότος ότι επηρεάζεται σε μεγάλο βαθμό από τη σειρά εξέτασης των αντικειμένων, δίνεται η δυνατότητα περαιτέρω βελτίωσης της ποιότητας συσταδοποίησης μέσω παραλλαγών που περιλαμβάνουν:

- Επανάληψη της ανάθεσης των αντικειμένων στις προηγουμένως δημιουργηθείσες συστάδες.

- Επαναπροσδιορισμός των συστάδων των αντικειμένων.

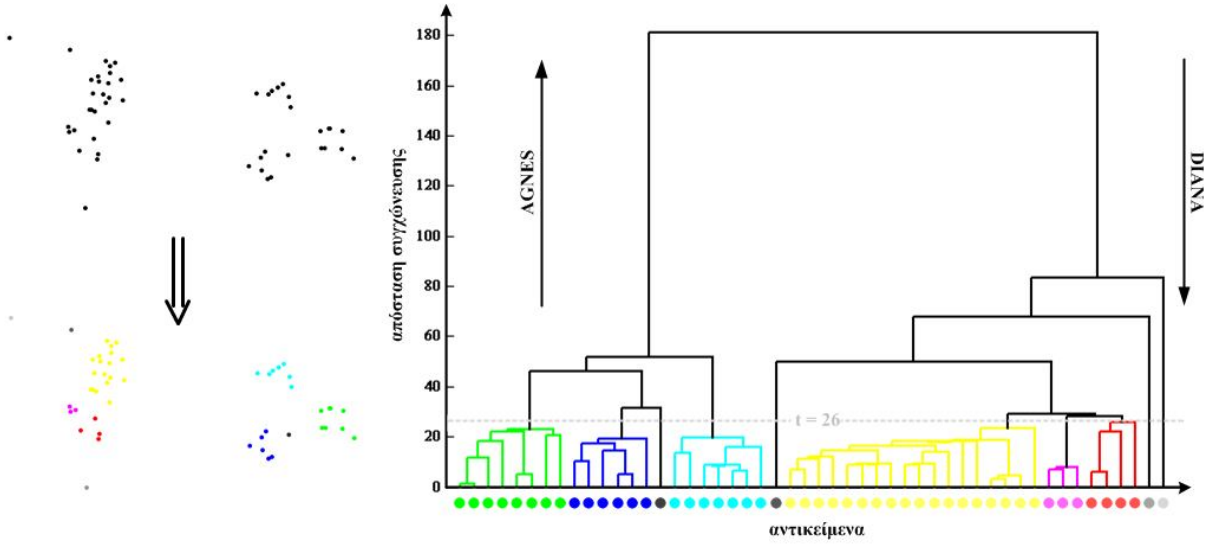
4.4 Ιεραρχικές Μέθοδοι

Σε αντίθεση με τις μεθόδους διαμερισμού, οι ιεραρχικές μέθοδοι δεν αναθέτουν τα αντικείμενα των δεδομένων σε k συστάδες από τα πρώτα βήματα, αλλά ακολουθούν μια σειρά από συνεχόμενες συσταδοποιήσεις δημιουργώντας μια ιεραρχία από επίπεδα στο καθένα από τα οποία ο αριθμός των συστάδων που υπάρχουν εκφράζει και το επίπεδο ομοιότητας της συσταδοποίησης. Τα επίπεδα αυτά συνήθως αποτυπώνονται γραφικά χρησιμοποιώντας ένα διάγραμμα δένδρου, το *δενδρόγραμμα* (dendrogram), το οποίο παρουσιάζει την εξέλιξη της ιεραρχικής αυτής συσταδοποίησης. Όπως δε αναφέρθηκε και παραπάνω, ανάλογα με τη φορά της διαδικασίας δημιουργίας της ιεραρχίας, οι ιεραρχικές μέθοδοι διακρίνονται σε συσσωρευτικές αν η συσταδοποίηση ξεκινάει με κάθε αντικείμενο να αποτελεί μια ξεχωριστή συστάδα και καταλήγει σε λιγότερες (k ή 1) συστάδες, και σε διαιρετικές αν η συσταδοποίηση ξεκινάει με όλα τα αντικείμενα να ανήκουν σε μια συστάδα η οποία έπειτα να διαιρείται συνεχώς ώστε οι συστάδες να γίνουν περισσότερες (k ή ένα αντικείμενο ανά συστάδα). Μερικές από τις πιο γνωστές ιεραρχικές μεθόδους παρουσιάζονται στη συνέχεια.

4.4.1 AGNES και DIANA

Οι μέθοδοι **AGNES** (AGglomerative NESting) και **DIANA** (DIvisive ANAlysis) [50] αποτελούν δύο απλές και συνάμα αντιπροσωπευτικές ιεραρχικές μεθόδους συσταδοποίησης. Η AGNES ακολουθεί τη συσσωρευτική προσέγγιση, όπου αρχικά κάθε αντικείμενο αποτελεί από μόνο του μια ξεχωριστή συστάδα και στη συνέχεια με επαναλαμβανόμενη συγχώνευση των πλησιέστερων συστάδων δημιουργούνται ολοένα και μεγαλύτερες συστάδες μέχρι είτε όλα τα αντικείμενα να ανήκουν σε μια συστάδα είτε να ικανοποιηθεί μια συγκεκριμένη συνθήκη τερματισμού της διαδικασίας συσταδοποίησης (π.χ. αριθμός συστάδων = επιθυμητός αριθμός συστάδων k ή απόσταση μεταξύ των δύο κοντινότερων συστάδων > κάποιο προκαθορισμένο κατώφλι t). Αντίθετα, η DIANA ακολουθεί τη διαιρετική προσέγγιση με ακριβώς την αντίστροφη διαδικασία συσταδοποίησης. Στην Εικόνα 4.6 φαίνεται ένα παράδειγμα εφαρμογής των μεθόδων αυτών για μια περιοχή 500x300, καθώς και το αντίστοιχο δενδροδιάγραμμα για την περίπτωση όπου υπάρχει ένα προκαθορισμένο κατώφλι απόστασης t.

Όσον αφορά τον τρόπο με τον οποίο μετριέται η απόσταση μεταξύ δύο συστάδων C_i και C_j, ώστε να δημιουργηθούν (μέσω συσσώρευσης ή διαίρεσης) οι κατάλληλες νέες συστάδες, διακρίνονται διάφορες τεχνικές από τις οποίες ορισμένες αντιπροσωπευτικές είναι:



Εικόνα 4.6: Συσταδοποίηση απλού συνδέσμου με τις μεθόδους AGNES και DIANA για 50 αντικείμενα

- *Απλού συνδέσμου* (single link): Χρησιμοποιείται η ελάχιστη απόσταση μεταξύ οποιουδήποτε αντικειμένου $\hat{p}_{ik}$ της μιας συστάδας και οποιουδήποτε αντικειμένου $\hat{p}_{jl}$ της άλλης, δηλαδή

$$dist(C_i, C_j) = \min_{1 \le k \le size(C_i), 1 \le l \le size(C_j)} (\|\hat{p}_{ik} - \hat{p}_{jl}\|) \quad (7)$$

- *Πλήρους συνδέσμου* (complete link): Χρησιμοποιείται η μέγιστη απόσταση μεταξύ οποιουδήποτε αντικειμένου $\hat{p}_{ik}$ της μιας συστάδας και οποιουδήποτε αντικειμένου $\hat{p}_{jl}$ της άλλης, δηλαδή

$$dist(C_i, C_j) = \max_{1 \le k \le size(C_i), 1 \le l \le size(C_j)} (\|\hat{p}_{ik} - \hat{p}_{jl}\|) \quad (8)$$

- *Μέσου συνδέσμου* (average link): Χρησιμοποιείται η μέση απόσταση μεταξύ οποιουδήποτε αντικειμένου $\hat{p}_{ik}$ της μιας συστάδας και οποιουδήποτε αντικειμένου $\hat{p}_{jl}$ της άλλης, δηλαδή

$$dist(C_i, C_j) = \frac{1}{\mathrm{size}(C_i)\mathrm{size}(C_j)} \sum_{k=1}^{size(C_i)} \sum_{l=1}^{size(C_j)} \|\hat{p}_{ik} - \hat{p}_{jl}\| \quad (9)$$

- *Centroid συνδέσμου* (centroid link): Χρησιμοποιείται η απόσταση μεταξύ των μέσων αντικειμένων $\hat{c}_i$ και $\hat{c}_j$ των συστάδων, δηλαδή

$$dist(C_i, C_j) = \|\hat{c}_i - \hat{c}_j\| \tag{10}$$

- *Medoid συνδέσμου* (medoid link): Χρησιμοποιείται η απόσταση μεταξύ των medoid αντικειμένων $\hat{m}_i$ και $\hat{m}_j$ των συστάδων, δηλαδή

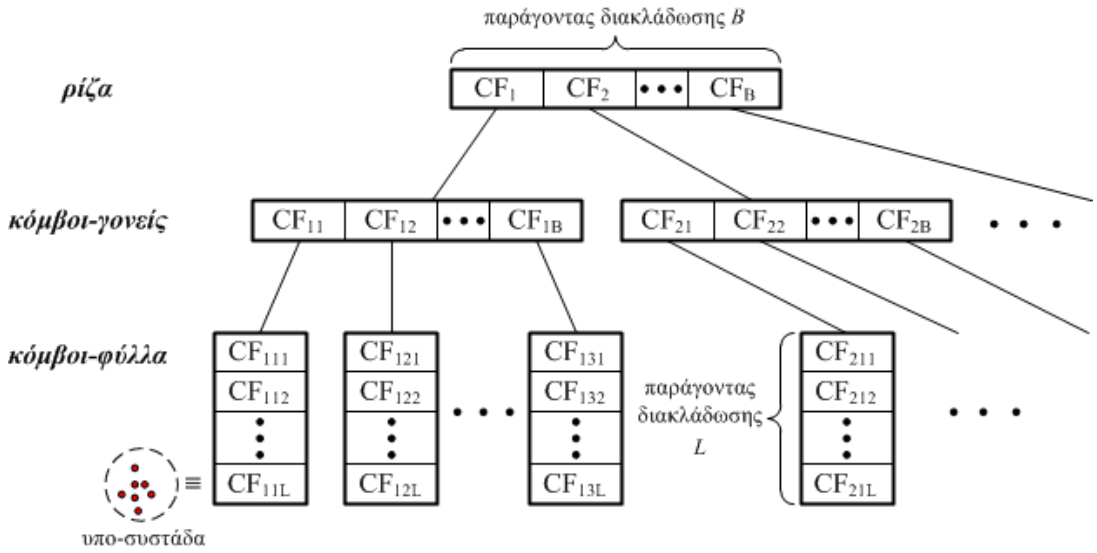
$$dist(C_i, C_j) = \|\hat{m}_i - \hat{m}_j\| \tag{11}$$

Δεδομένης της απλούστευσης ως προς τη διαδικασία εύρεσης των συστάδων που θα συγχωνευθούν ή διαιρεθούν, σε συνδυασμό με την έλλειψη δυνατότητας επιστροφής και διόρθωσης κάποιας λανθασμένης επιλογής σε κάποιο σημείο της εξέλιξης της συσταδοποίησης, οι μέθοδοι AGNES και DIANA είναι πολύ πιθανό να οδηγήσουν σε χαμηλής ποιότητας τελικές συστάδες. Για το λόγο αυτό, έχουν προταθεί διάφορες άλλες ιεραρχικές μέθοδοι που βελτιώνουν την απόδοση της συσταδοποίησης εκ των οποίων δύο πιο αντιπροσωπευτικές παρουσιάζονται στη συνέχεια.

4.4.2 BIRCH

Η μέθοδος **BIRCH** (Balanced Iterative Reducing and Clustering using Hierarchies) [105] έχει προταθεί για την περίπτωση της συσταδοποίησης μεγάλων συνόλων δεδομένων και συνδυάζει την ιεραρχική συσταδοποίηση, ώστε να επιτευχθεί μείωση των απαιτήσεων σε υπολογιστική πολυπλοκότητα και χρήση μνήμης, με κάποια άλλη μέθοδο συσταδοποίησης, όπως διαμερισμού. Βασική ιδέα αποτελεί η ανάθεση των αντικειμένων των δεδομένων σε υπο-συστάδες και έπειτα η εφαρμογή κάποιας διαδικασίας συσταδοποίησης στις υπο-συστάδες αυτές, οι οποίες είναι πολύ λιγότερες σε πλήθος από το σύνολο των αντικειμένων των δεδομένων. Για να επιτευχθεί αυτό χρησιμοποιούνται οι έννοιες των χαρακτηριστικών συστάδας (Clustering Feature - CF) και δένδρου χαρακτηριστικών συστάδας (Clustering Feature Tree - CF-tree), οι οποίες δίνουν μια εικόνα των χαρακτηριστικών των συστάδων που έχουν δημιουργηθεί ως προς το πλήθος και την κατανομή των αντικειμένων που περιλαμβάνουν.

Πιο συγκεκριμένα, κάθε συστάδα περιγράφεται από ένα CF το οποίο αποτελεί μια τριάδα χαρακτηριστικών:



Εικόνα 4.7: Δομή του CF-tree δένδρου το οποίο χρησιμοποιείται από τη μέθοδο BIRCH

$$CF = \langle n, \widehat{LS}, SS \rangle \tag{12}$$

, όπου n είναι το πλήθος των αντικειμένων $\hat{p}_i$ που περιλαμβάνει η συστάδα, $\widehat{LS}$ είναι το άθροισμα των αντικειμένων της συστάδας ως προς τις διαστάσεις τους, $\widehat{LS} = \sum_{i=1}^{n} \hat{p}_i$, και SS είναι το άθροισμα των τετραγώνων των διαστάσεων των αντικειμένων, $SS = \sum_{i=1}^{n} \hat{p}_i^2$. Καθένα δε από τα χαρακτηριστικά αυτά είναι απόλυτα συναφή με τις χαρακτηριστικές τιμές μιας συστάδας, όπως είναι το κέντρο βάρους (centroid), η ακτίνα και η διάμετρος. Αυτές οι τριάδες CF τοποθετούνται σε ένα δένδρο CF-tree, όπως φαίνεται στην Εικόνα 4.7, το οποίο είναι ένα σταθμισμένο δένδρο αναζήτησης (height-balanced tree) και χρησιμοποιείται για την ιεραρχική συσταδοποίηση των υπο-συστάδων. Οι κόμβοι-φύλλα του δένδρου περιλαμβάνουν τα CF των αρχικών υπο-συστάδων των αντικειμένων, ενώ κάθε κόμβος-γονέας περιλαμβάνει το άθροισμα των CF των αντίστοιχων κόμβων-παιδιών. Το μέγεθος του δέντρου και συνάμα η απόδοση της συσταδοποίησης καθορίζονται από δύο παραμέτρους:

- *παράγοντας διακλάδωσης* (branching factor) B ή L, ο οποίος ορίζει τον μέγιστο αριθμό κόμβων-παιδιών για κάθε κόμβο-γονέα ή τον μέγιστο αριθμό από CF των υπο-συστάδων που περιλαμβάνονται σε έναν κόμβο-φύλλο, αντίστοιχα.

- προκαθορισμένο κατώφλι (threshold) T, το οποίο καθορίζει τη μέγιστη επιτρεπτή διάμετρο των υπο-συστάδων στους κόμβους-φύλλα του δέντρου.

Όσον αφορά την όλη διαδικασία συσταδοποίησης της μεθόδου BIRCH, διακρίνονται τέσσερις φάσεις εκ των οποίων οι δύο είναι προαιρετικές, και μπορούν να συνοψισθούν ως εξής:

1. Αρχικά, τα αντικείμενα των δεδομένων τοποθετούνται ένα-ένα στο αρχικό CF-tree ως εξής:

    1.1. Για οποιοδήποτε αντικείμενο αναζητείται η πλησιέστερη υπο-συστάδα σε κάποιον από τους κόμβους-φύλλα ξεκινώντας από τη ρίζα του δέντρου.

    1.2. Ελέγχεται αν η διάμετρος της συγκεκριμένης υπο-συστάδας στην οποία επιλέχθηκε να ανατεθεί ένα αντικείμενο ξεπερνάει το προκαθορισμένο κατώφλι T.

    1.3. Αν όχι, ανατίθεται το αντικείμενο στην υπο-συστάδα και ενημερώνεται το CF της καθώς και τα αντίστοιχα CF των κόμβων-γονέων στους οποίους υπάγεται.

    1.4. Αν ναι, σε περίπτωση που οι υπο-συστάδες στον επιλεγμένο κόμβο-φύλλο είναι λιγότερες από L, τότε το αντικείμενο τοποθετείται στον κόμβο αυτό ως ξεχωριστή υπο-συστάδα, διαφορετικά ο κόμβος διαιρείται σε δύο επιμέρους κόμβους οι οποίοι μοιράζονται τις υπο-συστάδες.

2. (Προαιρετικό) Αν το CF-tree που δημιουργήθηκε κατά την πρώτη φάση είναι αρκετά μεγάλο σε μέγεθος (π.χ. υπερβαίνει το μέγεθος της διατιθέμενης μνήμης), τότε δημιουργείται ένα μικρότερο CF-tree επιλέγοντας μεγαλύτερη τιμή για το κατώφλι T, πράγμα το οποίο συνεπάγεται συγχώνευση μερικών από τις υπο-κλάσεις των κόμβων-φύλλων.

3. Εφαρμόζεται μια μέθοδος συσταδοποίησης, όπως διαμερισμού, για τις υπο-συστάδες που περιλαμβάνονται στους κόμβους-φύλλα χρησιμοποιώντας τα χαρακτηριστικά των CF τους.

4. (Προαιρετικό) Χρησιμοποιούνται τα μέσα αντικείμενα (centroids) των συστάδων που δημιουργήθηκαν στην φάση 3, ώστε να ανατεθούν για ακόμη μία φορά τα αντικείμενα των δεδομένων στις πλησιέστερες από αυτές συστάδες βελτιώνοντας περαιτέρω τη συσταδοποίηση.

4.4.3 CURE

Οι περισσότερες από τις μεθόδους που παρουσιάστηκαν παραπάνω βασίζονται στην επιλογή ενός μόνο αντιπροσωπευτικού αντικειμένου για κάθε συστάδα, όπως centroid, medoid κ.α. με αποτέλεσμα η ποιότητα της συσταδοποίησης ενδεχομένως να μην είναι πολύ καλή. Για το λόγο αυτό προτάθηκε η μέθοδος **CURE** (Clustering Using REpresentatives) [37], η οποία χρησιμοποιεί περισσότερα του ενός κατάλληλα διασκορπισμένα αντικείμενα ώστε να αντιπροσωπευθεί κάθε συστάδα, γεγονός το οποίο προσφέρει τη δυνατότητα εντοπισμού συστάδων διαφόρων σχημάτων και όχι αποκλειστικά σφαιρικών. Δεδομένων αυτών των πολλαπλών αντιπροσωπευτικών αντικειμένων, η απόσταση μεταξύ δύο συστάδων καθορίζεται από την απόσταση των δύο πλησιέστερων από αυτά. Τα βήματα της μεθόδου συνοψίζονται παρακάτω:

1. Αρχικά, κάθε αντικείμενο των δεδομένων αποτελεί μια ξεχωριστή συστάδα, οπότε το καθένα από αυτά είναι αντιπροσωπευτικό αντικείμενο της συστάδας στην οποία ανήκει.

2. Συγχωνεύονται οι πλησιέστερες μεταξύ τους συστάδες σε μία συστάδα C_m.

3. Για τη νέα συστάδα C_m, επιλέγονται ένας προκαθορισμένος αριθμός από c καλά διασκορπισμένα αντιπροσωπευτικά αντικείμενα ως εξής:

   3.1. Ως πρώτο αντιπροσωπευτικό αντικείμενο επιλέγεται εκείνο του οποίου η απόσταση από το μέσο αντικείμενο (centroid) της συστάδας C_m είναι η μεγαλύτερη.

   3.2. Για καθένα από τα υπόλοιπα $c - 1$ αντιπροσωπευτικά αντικείμενα, επιλέγεται εκείνο το αντικείμενο των δεδομένων του οποίου η απόσταση από οποιοδήποτε άλλο αντιπροσωπευτικό αντικείμενο που έχει προηγουμένως επιλεγεί είναι η μεγαλύτερη.

4. Η απόσταση κάθε αντιπροσωπευτικού αντικειμένου $\hat{p}_{ij}$ από το μέσο αντικείμενο $\hat{c}_i$ της συστάδας C_i μειώνεται με την επιβολή ενός συντελεστή συρρίκνωσης (shrink factor) α στις διαστάσεις των πρώτων, δημιουργώντας «εικονικά» αντιπροσωπευτικά αντικείμενα $\hat{p}'_{ij}$: $\hat{p}'_{ij} = \hat{p}_{ij} + a(\hat{c}_i - \hat{p}_{ij})$.

5. Επαναλαμβάνονται τα βήματα 2 έως 4 έως ότου ο αριθμός των συστάδων είναι ο προκαθορισμένος k.

Επίσης, για να διαχειρισθεί μεγάλα σύνολα δεδομένων, η CURE υιοθετεί τη διαδικασία της δειγματοληψίας και της τμηματοποίησης της συσταδοποίησης περιορίζοντας την υπολογιστική πολυπλοκότητα. Έτσι, αρχικά ένα τυχαίο δείγμα από το σύνολο των αντικειμένων των δεδομένων επιλέγεται και τμηματοποιείται, έπειτα κάθε τμήμα συσταδοποιείται ξεχωριστά σε υπο-συστάδες, και τέλος εφαρμόζεται η διαδικασία συσταδοποίησης που περιγράφθηκε παραπάνω για τις υπο-συστάδες που δημιουργήθηκαν.

4.5 Μέθοδοι Βασισμένες στην Πυκνότητα

Οι μέθοδοι βασιζόμενες στην πυκνότητα της κατανομής των αντικειμένων των δεδομένων προτάθηκαν κυρίως για την εύρεση συστάδων διαφόρων σχημάτων. Η δε χρήση του μεγέθους της πυκνότητας για την συσταδοποίηση των αντικειμένων, ως εναλλακτικό της απόστασης μεταξύ αντικειμένων η οποία χρησιμοποιείται από τις περισσότερες από τις προαναφερθείσες στις προηγούμενες υποενότητες μεθόδους, ουσιαστικά ορίζει τις συστάδες ως πυκνές περιοχές από αντικείμενα στο «χώρο» των δεδομένων οι οποίες οριοθετούνται από περιοχές μικρότερης πυκνότητας (π.χ. περιοχές με outliers, ή αλλιώς περιοχές θορύβου). Η πιο γνωστή τέτοια μέθοδος, αλλά και αντιπροσωπευτική των υπολοίπων, αποτελεί η **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) [29].

Σύμφωνα με την μέθοδο DBSCAN, η πυκνότητα των αντικειμένων μετριέται μέσω του αριθμού των γειτονικών αντικειμένων σε μια περιοχή. Έτσι, αναζητούνται εκείνα τα αντικείμενα τα οποία έχουν έναν ελάχιστο αριθμό $MinPts$ από γειτονικά αντικείμενα, σε μια προκαθορισμένη ακτίνα Eps, ώστε να εντοπισθούν οι αντίστοιχες περιοχές που έχουν μια ικανοποιητική πυκνότητα και μπορούν να αποτελέσουν τμήματα συστάδων. Στο τέλος της μεθόδου τα αντικείμενα των δεδομένων είτε θα ανήκουν σε κάποια συστάδα που σχηματίζεται από την συνένωση των περιοχών υψηλής πυκνότητας είτε θα αποτελούν «θόρυβο» και δε θα λαμβάνονται υπόψη σε κάποια ενδεχόμενη απεικόνιση της συσταδοποίησης. Όλα αυτά εκφράζονται μέσω κατάλληλων εννοιών που αφορούν τη συγκεκριμένη μέθοδο ως εξής:

- Η *Eps-γειτονιά* (Eps-neighborhood) ενός αντικειμένου $\hat{p}$ περιλαμβάνει όλα τα αντικείμενα $\hat{q}$ των δεδομένων για τα οποία ισχύει $\|\hat{p} - \hat{q}\| \leq Eps$.

- Αν η Eps-γειτονιά ενός αντικειμένου $\hat{p}$ περιλαμβάνει το λιγότερο $MinPts$ αντικείμενα, τότε το $\hat{p}$ ονομάζεται *σημείο πυρήνα* (core point), ενώ κάθε άλλο αντικείμενο $\hat{q}$ που

περιλαμβάνεται στην Eps-γειτονιά του $\hat{p}$ είναι *άμεσα πυκνά-προσεγγίσιμο* (directly density-reachable).

- Ένα αντικείμενο $\hat{p}$ είναι *πυκνά-προσεγγίσιμο* (density-reachable) από κάποιο αντικείμενο $\hat{q}$ αν υπάρχει μια αλληλουχία $\hat{p}_1,\ldots,\hat{p}_n$, με $\hat{p}_1 = \hat{p}$ και $\hat{p}_n = \hat{q}$, όπου το αντικείμενο $\hat{p}_{i+1}$ είναι άμεσα πυκνά-προσεγγίσιμο από το $\hat{p}_i$.

- Ένα αντικείμενο $\hat{p}$ είναι *πυκνά-συνδεδεμένο* (density-connected) με κάποιο αντικείμενο $\hat{q}$ αν υπάρχει κάποιο $\hat{o}$ από το οποίο είναι πυκνά-προσεγγίσιμα τόσο το $\hat{p}$ όσο και το $\hat{q}$.

- Μια συστάδα C ορίζεται από τις εξής συνθήκες:

  - Αν ένα αντικείμενο $\hat{p}$ είναι σημείο πυρήνα και ανήκει στη συστάδα C, και $\hat{q}$ είναι ένα πυκνά-προσεγγίσιμο αντικείμενο από το $\hat{p}$, τότε και το $\hat{q}$ ανήκει στη συστάδα C. Αν δε το $\hat{q}$ δεν είναι σημείο πυρήνα, τότε είναι *συνοριακό σημείο* (border point).

  - Αν δύο αντικείμενα $\hat{p}$ και $\hat{q}$ ανήκουν στη συστάδα C, τότε είναι πυκνά-συνδεδεμένα μεταξύ τους.

- Τα αντικείμενα που δεν ανήκουν σε κάποια συστάδα ονομάζονται *σημεία θορύβου* (noise points).

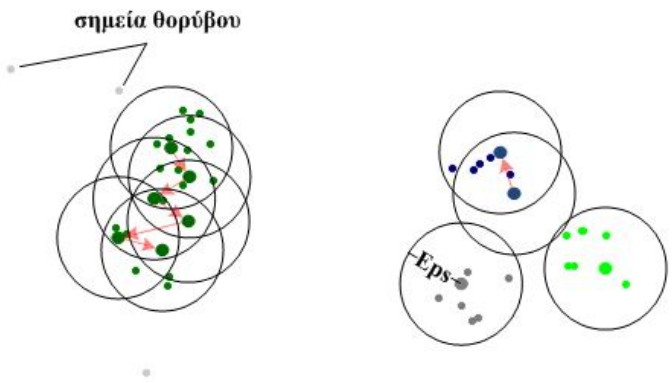
Στην Εικόνα 4.8 παρουσιάζεται ένα παράδειγμα της εξέλιξης της μεθόδου DBSCAN, ενώ τα βήματα που ακολουθούνται συνοψίζονται ως εξής:

1. Αρχικά, καθορίζονται οι παράμετροι Eps και $MinPts$.

2. Επιλέγεται τυχαία ένα αντικείμενο $\hat{p}$ που δεν έχει ανατεθεί σε κάποια συστάδα ή δεν είναι σημείο θορύβου, και υπολογίζεται ο αριθμός n των αντικειμένων στην Eps-γειτονιά του.

3. Αν $n < MinPts$, τότε το $\hat{p}$ χαρακτηρίζεται ως σημείο θορύβου και εκτελείται το βήμα 7.

4. Αν $n \geq MinPts$, τότε το $\hat{p}$ καθώς και όλα τα αντικείμενα της Eps-γειτονιάς του που δεν έχουν ανατεθεί προηγουμένως σε κάποια συστάδα ανατίθενται σε μια νέα συστάδα C.



Εικόνα 4.8: Εξέλιξη της μεθόδου DBSCAN για 50 αντικείμενα, Eps=46 και MinPts=5

5. Για κάθε μη προηγουμένως ελεγμένο ως προς την Eps-γειτονιά του αντικείμενο $\hat{q}$ το οποίο περιλαμβάνεται στην Eps-γειτονιά ενός σημείου πυρήνα της συστάδας C, υπολογίζεται ο αριθμός n' των Eps-γειτονικών αντικειμένων.

   - Αν $n' \geq MinPts$, τότε όλα τα αντικείμενα της Eps-γειτονιάς του $\hat{q}$ τα οποία δεν έχουν ανατεθεί προηγουμένως σε κάποια συστάδα ανατίθενται στη C.

6. Επαναλαμβάνεται το βήμα 5 μέχρι να μην υπάρχουν άλλα αντικείμενα της συστάδας C που να μην έχουν ελεγχθεί για την Eps-γειτονιά τους.

7. Επαναλαμβάνεται το βήμα 2 μέχρι να μην υπάρχουν άλλα αντικείμενα στο σύνολο των δεδομένων που να μην έχουν ελεγχθεί για την Eps-γειτονιά τους.

4.6 Μέθοδοι Βασισμένες σε Πλέγμα

Οι μέθοδοι της συγκεκριμένης κατηγορίας, εξαιτίας του δομικού τους στοιχείου, το πλέγμα, απευθύνονται κυρίως στην περίπτωση της χωρικής συσταδοποίησης. Κύριο πλεονέκτημά τους αποτελεί το γεγονός ότι είναι πολύ γρήγοροι όσον αφορά το χρόνο των απαιτούμενων υπολογισμών, το οποίο οφείλεται στη μείωση ουσιαστικά του συνόλου των αντικειμένων που συμμετέχουν στη συσταδοποίηση χρησιμοποιώντας, αντί για τα αρχικά μεμονωμένα αντικείμενα του συνόλου των δεδομένων, σύνθετα αντικείμενα που ορίζονται από τα κελιά του πλέγματος. Η δε ποιότητα της συσταδοποίησης που παρέχεται εξαρτάται σε μεγάλο βαθμό από το μέγεθος των κελιών (granularity) του πλέγματος. Μια από τις αντιπροσωπευτικές τέτοιες μεθόδους είναι η **STING** (STatistical INformation Grid) [102] και παρουσιάζεται στη συνέχεια.

Η STING είναι μια μέθοδος η οποία απαντάει ουσιαστικά σε ερωτήματα που αφορούν χωρικά δεδομένα επιστρέφοντας περιοχές - συστάδες όπου τα περιλαμβανόμενα αντικείμενα ικανοποιούν κάποιες συνθήκες που τίθενται από το κάθε ερώτημα. Για το σκοπό αυτό, χρησιμοποιεί μια ιεραρχική δομή από πλέγματα με καθένα να αποτελεί ένα διαφορετικό επίπεδο ανάλυσης ως προς το μέγεθος των ορθογώνιων κελιών που το απαρτίζουν. Έτσι, τα κελιά των ανώτερων επιπέδων αποτελούνται από έναν αριθμό από κελιά των κατώτερων επιπέδων σχηματίζοντας ένα ιεραρχικό δένδρο παρόμοιο με την περίπτωση του CF-tree που εξετάστηκε στην Ενότητα 4.4.2. Για καθένα από τα κελιά υπολογίζεται και αποθηκεύεται πληροφορία σχετικά με τον αριθμό και την κατανομή των περιλαμβανόμενων αντικειμένων, η οποία χρησιμοποιείται κατά τη διαδικασία της συσταδοποίησης. Σε αυτήν την πληροφορία περιλαμβάνονται, εκτός από τον αριθμό των αντικειμένων, παράμετροι για κάθε χαρακτηριστικό-διάσταση των αντικειμένων αυτών σε κάποιο κελί, και συγκεκριμένα, η μέση τιμή, η τυπική απόκλιση, η ελάχιστη τιμή, η μέγιστη τιμή και το είδος της κατανομής των τιμών. Οι τελικές δε συστάδες προκύπτουν από τη συνένωση κελιών από διάφορα επίπεδα πλεγμάτων που ικανοποιούν της συνθήκες που τίθενται από το κάθε ερώτημα.

Ουσιαστικά, υπάρχουν δύο φάσεις από τις οποίες αποτελείται η μέθοδος. Στην πρώτη φάση (η οποία εκτελείται μια φορά και όχι επανειλημμένα για κάθε ερώτημα) δημιουργούνται τα επίπεδα πλεγμάτων και υπολογίζονται οι σχετικές στατιστικές πληροφορίες που περιγράφουν το περιεχόμενο των κελιών, ενώ στη δεύτερη φάση ακολουθείται μια top-down προσέγγιση όπου αναζητούνται τα κατάλληλα κελιά που θα απαρτίζουν τις τελικές συστάδες. Πιο αναλυτικά τα βήματα που ακολουθούνται μπορούν να συνοψισθούν ως εξής:

1. (*1η φάση*) Αρχικά, δεδομένου του μεγέθους ενός κελιού του χαμηλότερου επιπέδου πλέγματος, ορίζονται τα επίπεδα των πλεγμάτων και τα κελιά του καθενός, με κάθε κελί του i επιπέδου να αποτελείται από έναν αριθμό κελιών του $i + 1$ επιπέδου.

2. Υπολογίζονται και αποθηκεύονται οι πληροφορίες σχετικά με τα κελιά, όπου στην περίπτωση του χαμηλότερου επιπέδου προκύπτουν απευθείας από τα αντικείμενα των δεδομένων, ενώ στα υψηλότερα επίπεδα προκύπτουν από τις πληροφορίες των κελιών των αμέσως επόμενων επιπέδων.

3. (*2η φάση*) Επιλέγεται ένα προκαθορισμένο υψηλό επίπεδο από το οποίο ξεκινάει η όλη διαδικασία για την εύρεση των κατάλληλων κελιών για τη συσταδοποίηση.

4. Για κάθε κελί του τρέχοντος επιλεγμένου επιπέδου υπολογίζεται το διάστημα εμπιστοσύνης (confidence interval) της πιθανότητας να σχετίζεται το κελί με το υποβληθέν ερώτημα, σύμφωνα με το οποίο καθορίζεται αν το κελί είναι σχετικό (relevant) ή όχι.

    - Αν είναι σχετικό, τότε το κελί προστίθεται στη λίστα των κελιών προς συσταδοποίηση και περαιτέρω έλεγχο για το αν τα υπαγόμενα σε αυτό κελιά του αμέσως επόμενου επιπέδου είναι σχετικά.

5. Επαναλαμβάνεται το βήμα 4 για τα κελιά του επόμενου επιπέδου που έχουν χαρακτηρισθεί ως σχετικά του ερωτήματος, εκτός και αν το τρέχον επίπεδο είναι το κατώτερο.

6. Συγχωνεύονται τα κοντινά μεταξύ τους κελιά ώστε να προκύψουν οι τελικές περιοχές-συστάδες.

4.7 Σύγκριση Μεθόδων Συσταδοποίησης

Έχοντας περιγράψει στις προηγούμενες ενότητες τους μηχανισμούς που διέπουν τις διαφορές κατηγορίες μεθόδων συσταδοποίησης καθώς και τα χαρακτηριστικά τους με αναφορά σε ορισμένες αντιπροσωπευτικές μεθόδους, η συγκεκριμένη ενότητα παρουσιάζει μια σύγκριση των μεθόδων αυτών ως προς την απόδοση της συσταδοποίησης που επιτυγχάνουν. Όπως αναφέρθηκε και στην αρχή του παρόντος κεφαλαίου, στόχος δεν είναι η εκτεταμένη και «αυστηρή» ανάλυση της συσταδοποίησης που επιτελούν οι διάφορες κατηγορίες τέτοιων μεθόδων, αλλά η βασική και «ποιοτική» ανάλυση των μηχανισμών και χαρακτηριστικών ορισμένων εξ αυτών, ώστε να αναδειχθούν στοιχεία που συμβάλλουν σε μια καλή συσταδοποίηση δεδομένων. Ο χαρακτηρισμός «καλή» όσον αφορά τη συσταδοποίηση θα μπορούσε να ειπωθεί ότι έχει μια ερμηνεία «τριών διαστάσεων»:

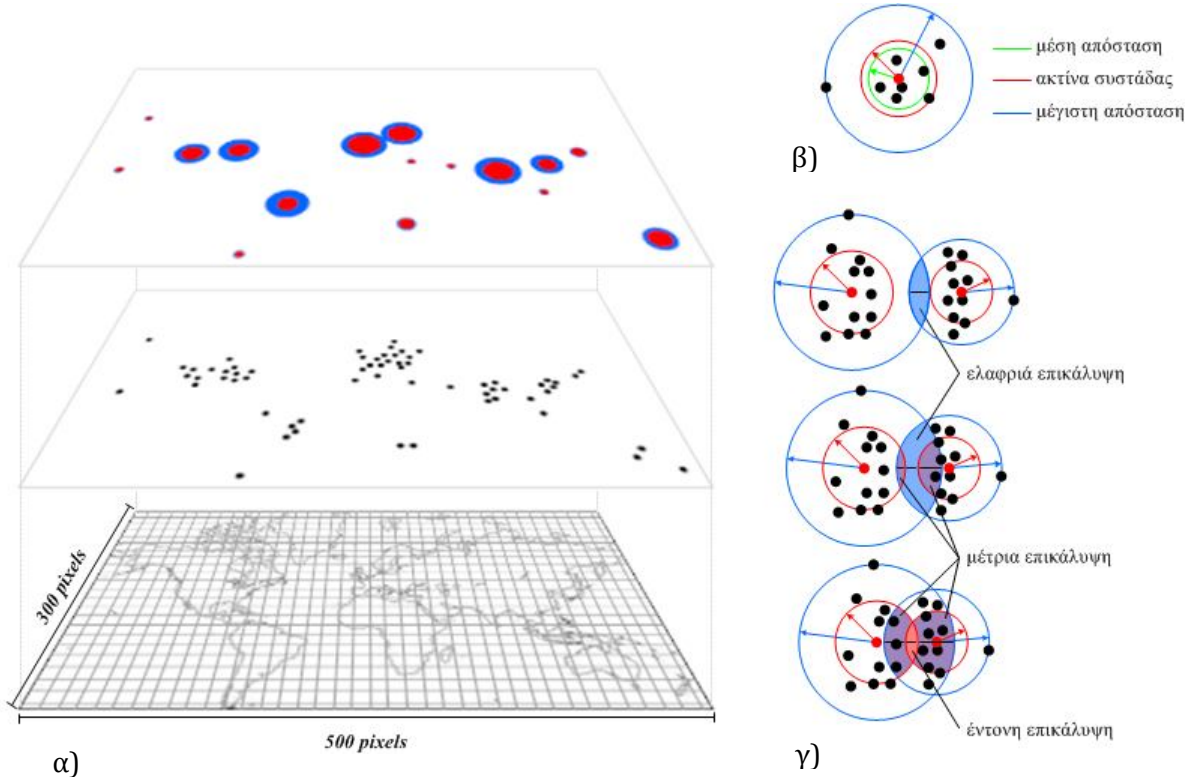
- *αντιπροσωπευτικότητα* των συστάδων, η οποία εκφράζει το πόσο καλά οι απεικονιζόμενες συστάδες αναπαριστούν-περιγράφουν τα αντικείμενα από τα οποία αποτελούνται,

- *ευκρίνεια απεικόνισης* των συστάδων, η οποία εκφράζει το πόσο ευδιάκριτες είναι οι απεικονιζόμενες συστάδες (π.χ. αν υπάρχουν επικαλύψεις μεταξύ των σχημάτων που αναπαριστούν τις συστάδες), και

- *ταχύτητα* συσταδοποίησης, η οποία εκφράζει το πόσο γρήγορα ολοκληρώνεται η διαδικασία εύρεσης συστάδων από ένα σύνολο αντικειμένων δεδομένων χρησιμοποιώντας κάποια μέθοδο συσταδοποίησης.

Καθεμιά από αυτές εξετάζεται όσο το δυνατό ξεχωριστά για κάθε μέθοδο, και στο τέλος συμβάλλει στην προκύπτουσα «συνισταμένη», την *απόδοση* της μεθόδου, με σημείο αναφοράς, βέβαια, την τελική απεικόνιση των συστάδων όπως υπαγορεύουν οι ανάγκες απεικόνισης του υλοποιηθέντος συστήματος.

Δεδομένου ότι η απεικόνιση των συστάδων αποτελεί βασικό σημείο αναφοράς για την περαιτέρω ανάλυση και σύγκριση των μεθόδων συσταδοποίησης, αρχικά γίνεται μια περιγραφή των βασικών χαρακτηριστικών στοιχείων τα οποία ανταποκρίνονται στις προδιαγραφές απεικόνισης που ακολουθούνται από το υλοποιηθέν σύστημα. Στην Εικόνα 4.9 γίνεται μια



Εικόνα 4.9: Απεικόνιση συστάδων: α) αντιστοίχιση συστάδων με αντικείμενα δεδομένων σε πλαίσιο διαστάσεων 500x300 pixels, β) χαρακτηριστικές ακτίνες συστάδας, και γ) περιπτώσεις επικάλυψης μεταξύ δύο συστάδων

παρουσίαση των στοιχείων αυτών. Τα δεδομένα αποτελούν αντικείμενα (μαύρες κουκίδες) με συντεταγμένες σε pixels πάνω σε ένα επίπεδο πλαίσιο διαστάσεων 500x300 pixels (ίδιες διαστάσεις με το πλαίσιο του υλοποιηθέντος συστήματος όπου προβάλλονται οι spam απειλές πάνω σε χάρτη). Οι συστάδες αναπαριστώνται με ομόκεντρους κύκλους (κόκκινους-μπλε) οι οποίοι καθορίζονται από δύο χαρακτηριστικές ακτίνες:

1. την ακτίνα συστάδας, r, η οποία, αν $\hat{p}_i$ είναι ένα από τα n αντικείμενα της συστάδας και $\hat{c}$ είναι το κέντρο της (centroid), δίνεται από τον τύπο:

$$r = \sqrt{\frac{\sum_{i=1}^{n}(\hat{p}_i - \hat{c})^2}{n}} \tag{13}$$

2. και την μέγιστη απόσταση κάποιου αντικειμένου της συστάδας από το κέντρο της.

Η αντιπροσωπευτικότητα μιας συστάδας μετριέται κυρίως βάσει της μέσης απόστασης από το κέντρο της των αντικειμένων που την αποτελούν, καθώς και από την μέγιστη απόσταση κάποιου αντικειμένου της. Όσον αφορά δε την ευκρίνεια απεικόνισης των συστάδων, αυτή για δύο συστάδες μετριέται μέσω της επικάλυψης των δύο χαρακτηριστικών ακτινών τους στη διεύθυνση της ευθείας που ενώνει τα δύο κέντρα τους, με την τελευταία να μπορεί να χαρακτηρισθεί ως έντονη, μέτρια ή ελαφριά, ανάλογα με το ποιος συνδυασμός ακτινών επικαλύπτεται. Πιο συγκεκριμένα, για τις διάφορες περιπτώσεις επικάλυψης δύο συστάδων, όπως εκείνες που παρουσιάζονται στην Εικόνα 4.9γ, και με r_i και m_i να είναι η ακτίνα και η μέγιστη απόσταση αντικειμένου της συστάδας i από το κέντρο της, αντίστοιχα, καθώς και $d = \|\hat{c}_i - \hat{c}_j\|$ η απόσταση μεταξύ των κέντρων των δύο συστάδων, i και j, κάθε τύπος επικάλυψης σε μονάδες απόστασης (π.χ. pixels) μετριέται ως εξής:

$$\begin{matrix}\text{έντονη} \\ \text{επικάλυψη} \\ (O_s)\end{matrix} = \begin{cases} 0, & r_i + r_j \le d \\ r_i + r_j - d, & max(r_i, r_j) - min(r_i, r_j) \le d < r_i + r_j \\ 2 \times min(r_i, r_j), & 0 \le d < max(r_i, r_j) - min(r_i, r_j) \end{cases}$$

$$\text{μέτρια επικάλυψη}_{i \to j}\left(O_{m:i \to j}\right) =$$

$$\begin{cases} 0, & m_i + r_j \le d \\ m_i + r_j - d - O_s, & max(m_i, r_j) - min(m_i, r_j) \le d < m_i + r_j \\ 2 \times min(m_i, r_j) - O_s, & 0 \le d < max(m_i, r_j) - min(m_i, r_j) \end{cases}$$

$$\underset{(O_m)}{\text{μέτρια επικάλυψη}} = \underset{(O_{m:i\to j})}{\text{μέτρια επικάλυψη}} + \underset{(O_{m:j\to i})}{\text{μέτρια επικάλυψη}}$$

$$\text{ελαφριά επικάλυψη } (O_i) =$$

$$\begin{cases} 0, & m_i + m_j \le d \\ m_i + m_j - d - O_m - O_s, & \max(m_i, m_j) - \min(m_i, m_j) \le d < m_i + m_j \\ 2 \times min(m_i, m_j) - O_m - O_s, & 0 \le d < max(m_i, m_j) - min(m_i, m_j) \end{cases}$$

Δεδομένων των στοιχείων παραπάνω, τα οποία ουσιαστικά αποτελούν τις μετρικές απόδοσης της απεικόνισης των συστάδων, γίνεται ανάλυση και σύγκριση των μεθόδων συσταδοποίησης που παρουσιάσθηκαν στις προηγούμενες ενότητες. Η υλοποίηση για καθεμιά από αυτές τις μεθόδους ακολουθεί όσο το δυνατόν περισσότερο τα σχετικά βήματα που περιγράφονται, με κάποιες διαφοροποιήσεις να υπάρχουν σε περιπτώσεις που επιχειρείται είτε να γίνουν κάποιες απλουστεύσεις όσον αφορά την πολυπλοκότητα της υλοποίησης είτε να ικανοποιηθούν οι προδιαγραφές απεικόνισης που παρουσιάσθηκαν παραπάνω. Συγκεκριμένα, για κάθε μέθοδο ισχύουν τα εξής ως προς την υλοποίηση και την επιλογή παραμέτρων κατά την εκτέλεσή τους:

- **K-means**: Υλοποιούνται δύο εκδοχές τις μεθόδου, οι κατά **Forgy** [33] και η κατά **MacQueen** [59], οι οποίες διαφοροποιούνται ως προς την ανανέωση των μέσων αντικειμένων των συστάδων. Η πρώτη ανανεώνει τα τελευταία αφού έχει ολοκληρωθεί η διαδικασία ανάθεσης του συνόλου των αντικειμένων σε συστάδες, η οποία πραγματοποιείται σε κάθε επανάληψη, ενώ η δεύτερη τα ανανεώνει για κάθε ανάθεση αντικειμένου σε συστάδα.

- **K-medoids** και **PAM**: Υλοποιούνται σύμφωνα με τα σχετικά βήματα στην Ενότητα 4.3.2.

- **CLARA** και **CLARANS**: Υλοποιούνται σύμφωνα με τα σχετικά βήματα στην Ενότητα 4.3.3, με την πρώτη μέθοδο να εκτελείται με χρήση 5 δειγμάτων των k + 1 αντικειμένων, όπου k ο επιλεγμένος κάθε φορά αριθμός συστάδων, και με τη δεύτερη να εκτελείται για τιμές των παραμέτρων *αριθμός τοπικών ελαχίστων* και *μέγιστος αριθμός εξεταζόμενων γειτόνων*, 2 και 5, αντίστοιχα.

- **CURE**: Υλοποιείται σύμφωνα με τα σχετικά βήματα στην Ενότητα 4.4.3, και εκτελείται για 5 αντιπροσωπευτικά αντικείμενα ανά συστάδα και συντελεστή συρρίκνωσης 0,3.

- **Πλησιέστερων Γειτόνων (Nearest Neighbors)**: Υλοποιείται σύμφωνα με τα σχετικά βήματα στην Ενότητα 4.3.4, με τη διαφορά ότι εκτός από το κατώφλι απόστασης αντικειμένων t, χρησιμοποιείται και ένα κατώφλι μέγιστης διαμέτρου συστάδας $t_d = 2t$ κατά την ανάθεση των αντικειμένων στις δημιουργούμενες συστάδες, ώστε να περιορισθούν οι τελευταίες σε όσο το δυνατόν κυκλικά σχήματα.

- **DBSCAN**: Υλοποιείται σύμφωνα με τα σχετικά βήματα στην Ενότητα 4.5, με τη διαφορά ότι χρησιμοποιείται επιπλέον ένα κατώφλι μέγιστης απόστασης t_p των αντικειμένων που ελέγχονται περαιτέρω ως προς την Eps γειτονιά τους. Πιο συγκεκριμένα, για κάθε αντικείμενο στο βήμα 5 της μεθόδου το οποίο προορίζεται για έλεγχο της Eps-γειτονιάς του προηγείται ένας έλεγχος της μέγιστης απόστασής του από προηγουμένως ελεγχθέντα σημεία της δημιουργούμενης συστάδας. Αν υπάρχει υπέρβαση του κατωφλίου t_p, τότε το αντικείμενο αυτό δεν ελέγχεται περαιτέρω ως προς την Eps-γειτονιά του, με αποτέλεσμα να μην επεκτείνεται περισσότερο η συστάδα και να διατηρεί ένα όσο το δυνατό κυκλικό σχήμα. Σχετικά με την εκτέλεση της μεθόδου δε, η τιμή του κατωφλίου αυτού τίθεται ίση με την ακτίνα Eps, ενώ ο ελάχιστος αριθμός γειτονικών αντικειμένων MinPts τίθεται ίσος με 1 δεδομένου ότι όταν υπάρχουν γειτονικά μεταξύ τους αντικείμενα θα πρέπει να συσταδοποιηθούν. Τα σημεία θορύβου που προκύπτουν στο τέλος θεωρούνται ξεχωριστές συστάδες.

- **STING**: Για τη συγκεκριμένη μέθοδο δεν υλοποιείται ο μηχανισμός αναζήτησης σε πολλαπλά επίπεδα όπως περιγράφθηκε στην Ενότητα 4.6, αλλά χρησιμοποιείται μόνο ένα επίπεδο, ώστε να μελετηθεί κυρίως η συμβολή του πλέγματος στη συσταδοποίηση. Κατά τα άλλα η συγχώνευση των κελιών (βήμα 6) ακολουθεί την ίδια διαδικασία με την μέθοδο DBSCAN, θεωρώντας κάθε κελί ως ξεχωριστό αντικείμενο. Για την εκτέλεση δε της μεθόδου οι παράμετροι για τη συγχώνευση είναι οι ίδιες με αυτές της DBSCAN, ενώ χρησιμοποιούνται κελιά των 5 και των 20 pixels (**STING-5**, **STING-20**).

- **BIRCH**: Όπως και στην περίπτωση της STING, έτσι και στη συγκεκριμένη μέθοδο χρησιμοποιείται για την εκτέλεσή της ουσιαστικά μόνο ένα επίπεδο συσταδοποίησης θέτοντας τους παράγοντες διακλάδωσης, B και L, θεωρητικά στο άπειρο. Επίσης, εκτελείται μόνο η πρώτη φάση από αυτές που παρουσιάσθηκαν στην Ενότητα 4.4.2 δεδομένου ότι οι υπόλοιπες τρεις μπορούν να εφαρμοσθούν και στις άλλες μεθόδους.

- **AGNES/DIANA**: Οι συγκεκριμένες μέθοδοι δεν περιλαμβάνονται στη διαδικασία της σύγκρισης, δεδομένων των ομοιοτήτων τους με τις μεθόδους Πλησιέστερων Γειτόνων και CURE.

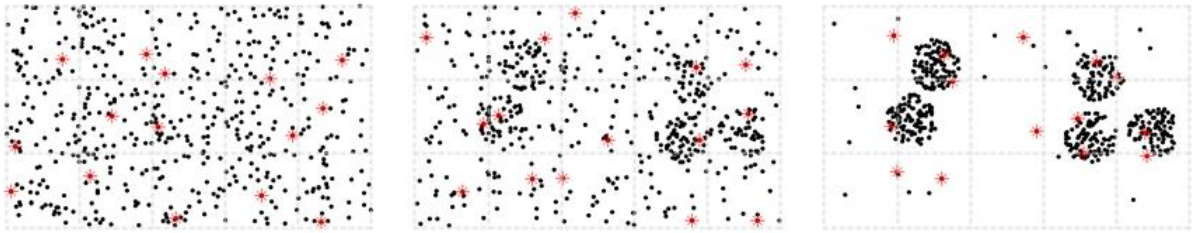
Για να επιτευχθεί όσο το δυνατό πιο «δίκαιη» σύγκριση των μεθόδων, επιχειρείται για κάθε εξεταζόμενη περίπτωση συσταδοποίησης αντικειμένων οι μέθοδοι να παράγουν περίπου τον ίδιο αριθμό συστάδων. Για τις μεθόδους Πλησιέστερων Γειτόνων, DBSCAN, STING και BIRCH, αυτό επιτυγχάνεται με την κατάλληλη επιλογή των παραμέτρων κατώφλι απόστασης t, ακτίνα Eps, ακτίνα Eps και κατώφλι ακτίνας συστάδας T, αντίστοιχα. Για τις υπόλοιπες μεθόδους οι οποίες απαιτούν ως είσοδο τον επιθυμητό αριθμό συστάδων, προκειμένου ο τελευταίος να υπολογισθεί, καθώς και να ανταποκρίνεται στο συγκεκριμένο πρόβλημα της δυναμικής προσαρμογής του στον αριθμό και την κατανομή των αντικειμένων προς συσταδοποίηση, ακολουθείται η εξής προεργασία:

1. Δημιουργείται ένα πλέγμα πάνω από το επίπεδο απεικόνισης των αντικειμένων των δεδομένων με συγκεκριμένο μέγεθος κελιού το οποίο να ταιριάζει στις επιθυμητές προδιαγραφές απεικόνισης των τελικών συστάδων.

2. Κάθε κελί ελέγχεται για το αν περιλαμβάνει αντικείμενα στο εσωτερικό του, και αν ναι, τότε επιλέγεται ένα από αυτά τυχαία, ώστε να αποτελέσει ένα αρχικό σημείο συστάδας.

3. Τελικά, προκύπτει τόσο ο αριθμός συστάδων, όσο και τα αρχικά σημεία που είναι απαραίτητα στην περίπτωση των μεθόδων K-means και K-medoids.

Με τον τρόπο αυτό, τα αποτελέσματα κάθε μεθόδου καθορίζονται με την επιλογή μιας παραμέτρου τύπου απόστασης που λαμβάνει τιμές σε pixels. Όσον αφορά δε τις περιπτώσεις κατανομών των αντικειμένων δεδομένων με βάση τις οποίες αξιολογούνται οι μέθοδοι, αυτές αφορούν τις: α) τυχαία, β) συγκεντρωμένη κατά 50% και γ) συγκεντρωμένη κατά 95% σε απόσταση 35 pixels γύρω από συγκεκριμένα σημεία του επιπέδου (120,145), (150,220), (365,120), (375,205) και (450,130), ενώ οι αριθμοί αντικειμένων που μελετώνται είναι 50, 500 και 5000. Για κάθε περίπτωση δε, επαναλαμβάνεται η διαδικασία παραγωγής αντικειμένων και συσταδοποίησής τους από κάποια μέθοδο 100 φορές, ώστε να προκύψει ένας μέσος όρος των τιμών των μετρικών απόδοσης της απεικόνισης συστάδων που παρουσιάσθηκαν παραπάνω στα αποτελέσματα. Στην Εικόνα 4.10 παρουσιάζονται οι συγκεκριμένες κατανομές για 500 αντικείμενα, καθώς και ένα παράδειγμα καθορισμού των αρχικών συστάδων για τις



Εικόνα 4.10: Περιπτώσεις κατανομών για 500 αντικείμενα (αριστερά προς δεξιά): τυχαία, συγκεντρωμένη κατά 50% και συγκεντρωμένη κατά 95%, καθώς και αντίστοιχος καθορισμός των αρχικών συστάδων

περιπτώσεις των μεθόδων που το απαιτούν με μέγεθος κελιού 100 pixels, ενώ στην Εικόνα 4.11 και στην Εικόνα 4.12 παρουσιάζονται τα αποτελέσματα που προκύπτουν για την κάθε μέθοδο όσον αφορά των υπό μελέτη μετρικών απόδοσης.

Σύμφωνα με τα αποτελέσματα των μετρικών απόδοσης της συσταδοποίησης, μπορούν να προκύψουν διάφορα συμπεράσματα για το ποια χαρακτηριστικά των μεθόδων μπορούν να συμβάλλουν τελικά σε μια «καλή» συσταδοποίηση όσον αφορά τις υπό μελέτη προδιαγραφές απεικόνισης. Σχετικά δε με τις τιμές των αποτελεσμάτων, αυτές ενδέχεται να διαφέρουν ανάλογα με τον τρόπο υλοποίησης των μεθόδων, αλλά σε γενικές γραμμές θα καταλήγουν στα ίδια συμπεράσματα.

Αρχικά, παρατηρείται ότι οι μέθοδοι που βασίζονται στα medoids, PAM, CLARA και CLARANS, παρουσιάζουν μια πολύ μεγάλη καθυστέρηση σε σχέση με τις υπόλοιπες μεθόδους, πράγμα το οποίο δικαιολογείται λόγω των ιδιαίτερων απαιτήσεων σε σχετικούς υπολογισμούς του εντοπισμού κάθε φορά των medoids των συστάδων (αυτός είναι και ο λόγος που δεν υπάρχουν αποτελέσματα για τη μέθοδο PAM σε περιπτώσεις συσταδοποίησης που περιλαμβάνουν μεγάλο αριθμό αντικειμένων με ενδεικτική περίπτωση εκείνης της τυχαίας κατανομής 500 αντικειμένων όπου φαίνεται ότι η PAM είναι πιο αργή από την K-means Forgy κατά 147559 φορές!). Αυτό το πρόβλημα του χρονοβόρου εντοπισμού των medoids σε σχέση με των υπολογισμό των centroids αναδεικνύεται και στην απλούστερη περίπτωση της μεθόδου K-medoids, η οποία εξαιτίας της συγκεκριμένης διαφοροποίησής της από τις K-means μεθόδους παρουσιάζει αυξημένους χρόνους συσταδοποίησης σε σχέση με τις τελευταίες ιδιαίτερα για μεγάλους αριθμούς αντικειμένων. Επίσης, σχετικά με τη δειγματοληψία που περιλαμβάνουν οι CLARA και CLARANS, φαίνεται ότι βελτιώνεται με τον τρόπο αυτό η ταχύτητα συσταδοποίησης σε σχέση με την PAM που αναφέρεται στο σύνολο των αντικειμένων των δεδομένων, αλλά αντίθετα χειροτερεύει η αντιπροσωπευτικότητα και εντείνεται η επικάλυψη των συστάδων.

Στη συνέχεια, αναφορικά με τις K-means μεθόδους, φαίνεται ότι υπερτερούν των υπολοίπων τόσο στην αντιπροσωπευτικότητα των συστάδων όσο και στην ταχύτητα συσταδοποίησης για την περίπτωση σχετικά μικρού αριθμού αντικειμένων (~50), διατηρώντας την επικάλυψη των συστάδων σε καλό επίπεδο. Η απόδοσή τους συνεχίζει να είναι καλή σε σχέση με τις υπόλοιπες μεθόδους και για μεγαλύτερους αριθμούς αντικειμένων, καθιστώντας τες έτσι (και επιβεβαιώνοντας το λόγο της ευρείας χρήση τους) ως μια ικανοποιητική λύση στα πρόβλημα της συσταδοποίησης. Όσον αφορά δε τη σύγκριση των δύο εκδοχών της K-means που παρουσιάζονται, φαίνεται ότι η συνεχόμενη προσαρμογή των centroids για κάθε ανάθεση αντικειμένου σε συστάδα (MacQueen) υπερτερεί της διατήρησης των centroids σταθερών σε κάθε επανάληψη ανάθεσης του συνόλου των αντικειμένων (Forgy).

Εξ ίσου ικανοποιητική προσέγγιση αποτελεί και η μέθοδος Πλησιέστερων Γειτόνων παρουσιάζοντας μια καλή απόδοση όσον αφορά όλες της σχετικές μετρικές που αφορούν τη συσταδοποίηση. Το κύριο δε πλεονέκτημα αυτής σε σχέση με τις μεθόδους K-means είναι ότι μπορεί να προσαρμόζεται καλύτερα στις υπό συσταδοποίηση κατανομές αντικειμένων δημιουργώντας τον απαιτούμενο αριθμό συστάδων δυναμικά κατά τη διάρκεια εκτέλεσης της μεθόδου. Το ίδιο βέβαια πλεονέκτημα έχουν και οι υπόλοιπες μέθοδοι που δεν απαιτούν ως είσοδο των αριθμό των συστάδων. Η BIRCH, όντας μία από αυτές, φαίνεται να σχετίζεται περισσότερο με την Πλησιέστερων Γειτόνων (όσον αφορά τη συγκεκριμένη υλοποίηση και τις επιλεγμένες τιμές για τις παραμέτρους της BIRCH που μελετώνται) με κύρια διαφοροποίησή της τον τρόπο με τον οποίο γίνεται ο έλεγχος για το αν ένα αντικείμενο θα πρέπει να ανατεθεί σε κάποια συστάδα. Όπως παρουσιάζεται και στα σχετικά αποτελέσματα συσταδοποίησης, το γεγονός ότι χρησιμοποιούνται οι τιμές των χαρακτηριστικών συστάδας (CF), κάνει πολύ εύκολο από άποψη υπολογισμών και συνάμα γρήγορο τον έλεγχο για την ανάθεση των αντικειμένων σε συστάδες, και σε συνδυασμό με το γεγονός ότι απαιτείται μια μόνο σάρωση των αντικειμένων, καθίσταται η συγκεκριμένη μέθοδος ως η καλύτερη από άποψη ταχύτητας συσταδοποίησης διατηρώντας παράλληλα τις άλλες μετρικές απόδοσης σε ικανοποιητικά επίπεδα.

Ως επίσης πολύ γρήγορη μέθοδος αποδεικνύεται ότι είναι και η STING, αρκεί το μέγεθος των κελιών του πλέγματος να είναι ικανοποιητικό, ώστε να αντισταθμίζεται η σχετική καθυστέρηση που δημιουργείται για την επιπλέον διαδικασία ομαδοποίησης των αντικειμένων στα κελιά. Ο τρόπος με τον οποίο δημιουργεί τις συστάδες είναι ο ίδιος με εκείνον της περίπτωσης της DBSCAN με τη διαφορά ότι αντί για τα αρχικά αντικείμενα των δεδομένων χρησιμοποιεί τα πιο σύνθετα αντικείμενα των κελιών στα οποία οδηγεί η χρήση του πλέγματος. Έτσι, ενώ η συγκριμένη εκδοχή της DBSCAN που μελετάται μπορεί να θεωρηθεί σύμφωνα με τα

αποτελέσματα ως μια σχετικά μέτρια μέθοδος συσταδοποίησης, με τη χρήση του πλέγματος τελικά επιταχύνεται (όπως φαίνεται με τη STING), ενώ παράλληλα συνεχίζει να διατηρεί κατά κάποιο τρόπο σταθερές της μετρικές απόδοσης της αντιπροσωπευτικότητας και της ευκρίνειας των συστάδων. Η δε επιτάχυνση αυτή, ανάλογα με το μέγεθος των κελιών που ερευνώνται στη συγκεκριμένη περίπτωση, φαίνεται ότι επιτυγχάνεται για μέγεθος κελιών 5 pixels (STING-5) όταν μεγάλος αριθμός αντικειμένων πρέπει να συσταδοποιηθεί (~5000), και για μέγεθος 20 pixels ακόμα και για λιγότερα αντικείμενα (~500 και άνω).

Τέλος, φαίνεται ότι η χρήση των αντιπροσωπευτικών αντικειμένων που περιλαμβάνει η μέθοδος CURE παρέχει πάρα πολύ καλά αποτελέσματα όσον αφορά την ευκρίνεια των συστάδων (πολύ χαμηλές επικαλύψεις), ιδιαίτερα μάλιστα όταν τα αντικείμενα είναι συγκεντρωμένα. Ωστόσο, τόσο η αναζήτηση των κοντινότερων μεταξύ τους συστάδων όσο και η εύρεση των αντιπροσωπευτικών αντικειμένων απαιτούν σημαντικό χρόνο, ο οποίος αυξάνεται με την αύξηση των αντικειμένων προς συσταδοποίηση, πράγμα το οποίο την υποβαθμίζει σε σχέση με άλλες που αναφέρθηκαν παραπάνω ως προς τις επιθυμητές προδιαγραφές απεικόνισης.

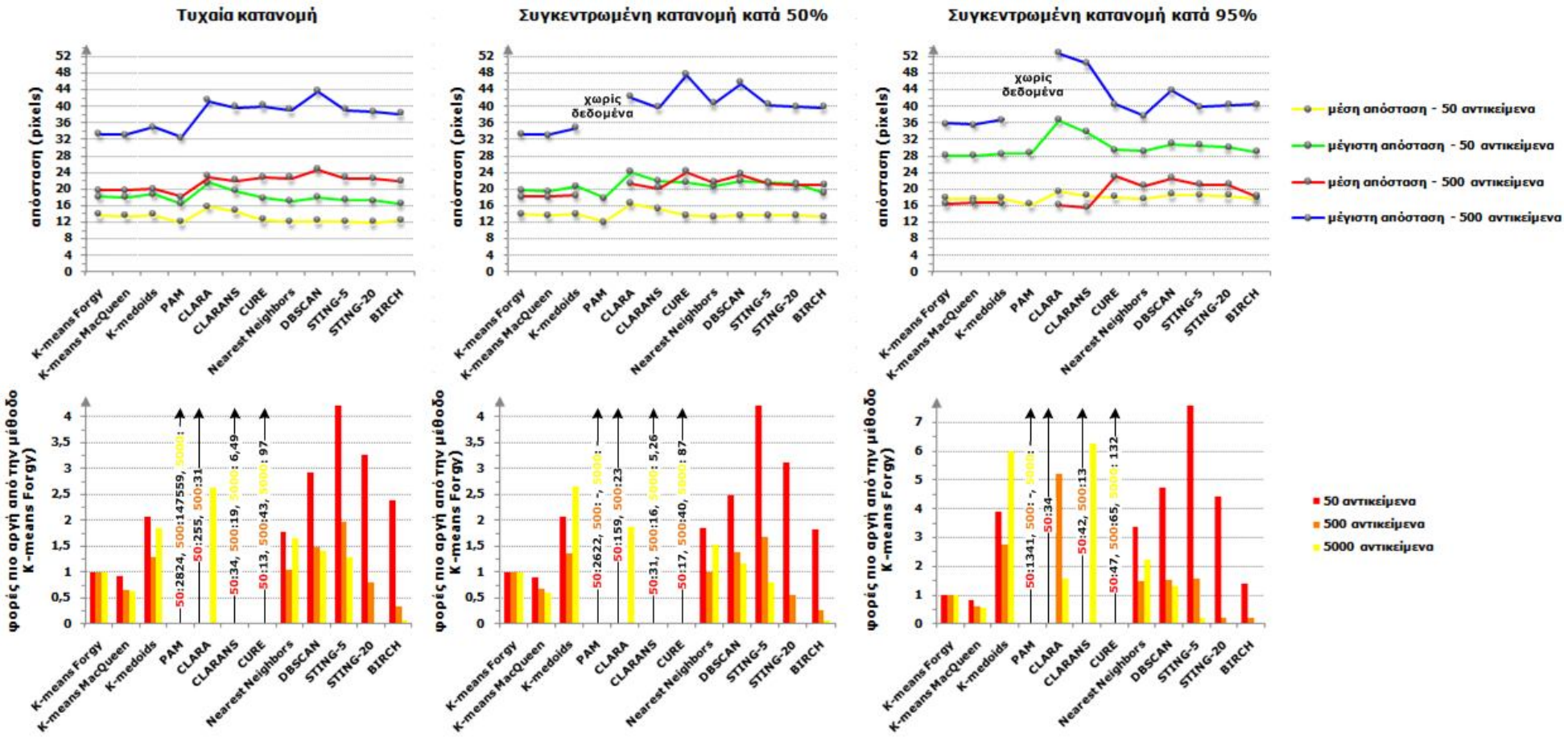
Λαμβάνοντας υπόψη τα παραπάνω, με βάση τα χαρακτηριστικά των μεθόδων που μπορούν να συμβάλλουν σε μια «καλή» συσταδοποίηση θα μπορούσαν να συνοψισθούν τα εξής:

- Προτίμηση των centroids αντί των medoids για την αναπαράσταση των συστάδων λόγω της σχετικά χαμηλής υπολογιστικής πολυπλοκότητας που απαιτούν.

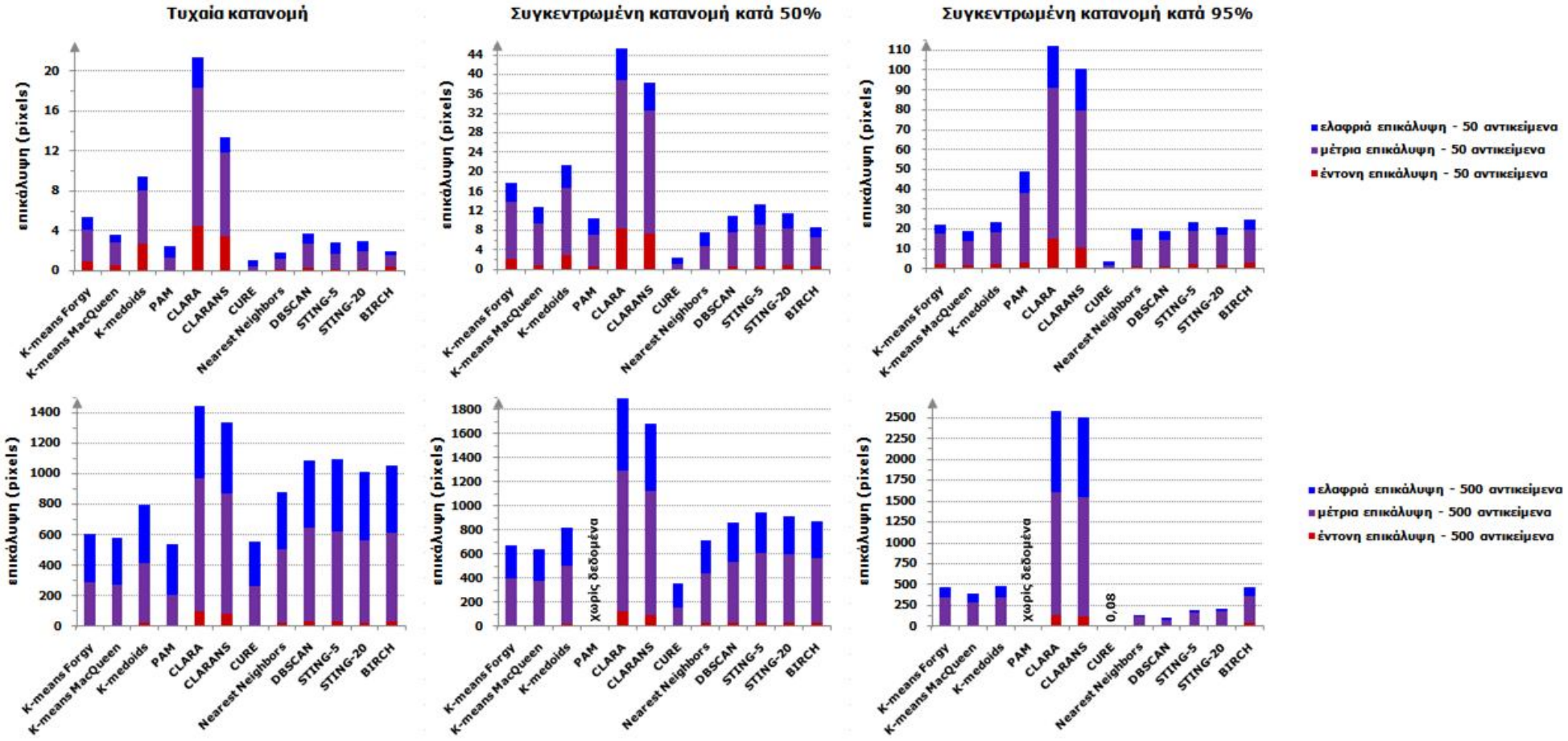
- Ανάθεση των αντικειμένων σε συστάδες υιοθετώντας τη σχετική διαδικασία που ακολουθείται στη BIRCH μέθοδο, χρησιμοποιώντας τα χαρακτηριστικά συστάδας (CF), λόγω της υψηλής ταχύτητας συσταδοποίησης η οποία επιτυγχάνεται.

- Χρήση πλέγματος για την επιτάχυνση ακόμα περισσότερο της διαδικασίας συσταδοποίησης χρησιμοποιώντας τα σύνθετα αντικείμενα των κελιών, τα οποία ουσιαστικά μειώνουν τον αριθμό αντικειμένων προς συσταδοποίηση.

- (Προαιρετικά) «Επανασυσταδοποίηση» των δημιουργηθέντων συστάδων και ανάθεση ξανά των αντικειμένων στις νέες συστάδες που προκύπτουν, όπως υπαγορεύουν η 3η και η 4η φάση της BIRCH μεθόδου, ώστε να βελτιωθεί η αντιπροσωπευτικότητα και η ευκρίνεια των τελικών συστάδων.



Εικόνα 4.11: Μέση και μέγιστη απόσταση των αντικειμένων των τελικών συστάδων (πάνω) και χρόνοι συσταδοποίησης σε σχέση με εκείνους της μεθόδους K-means Forgy (κάτω), για καθεμιά από τις υλοποιηθείσες μεθόδους συσταδοποίησης υπό διαφορετικές κατανομές αντικειμένων



Εικόνα 4.12: Μέγεθος και τύπος επικάλυψης των τελικών συστάδων για 50 (πάνω) και 500 (κάτω) αντικείμενα για καθεμιά από τις υλοποιηθείσες μεθόδους συσταδοποίησης υπό διαφορετικές κατανομές αντικειμένων

Κεφάλαιο 5
Σχεδιασμός, Υλοποίηση και Αποτελέσματα

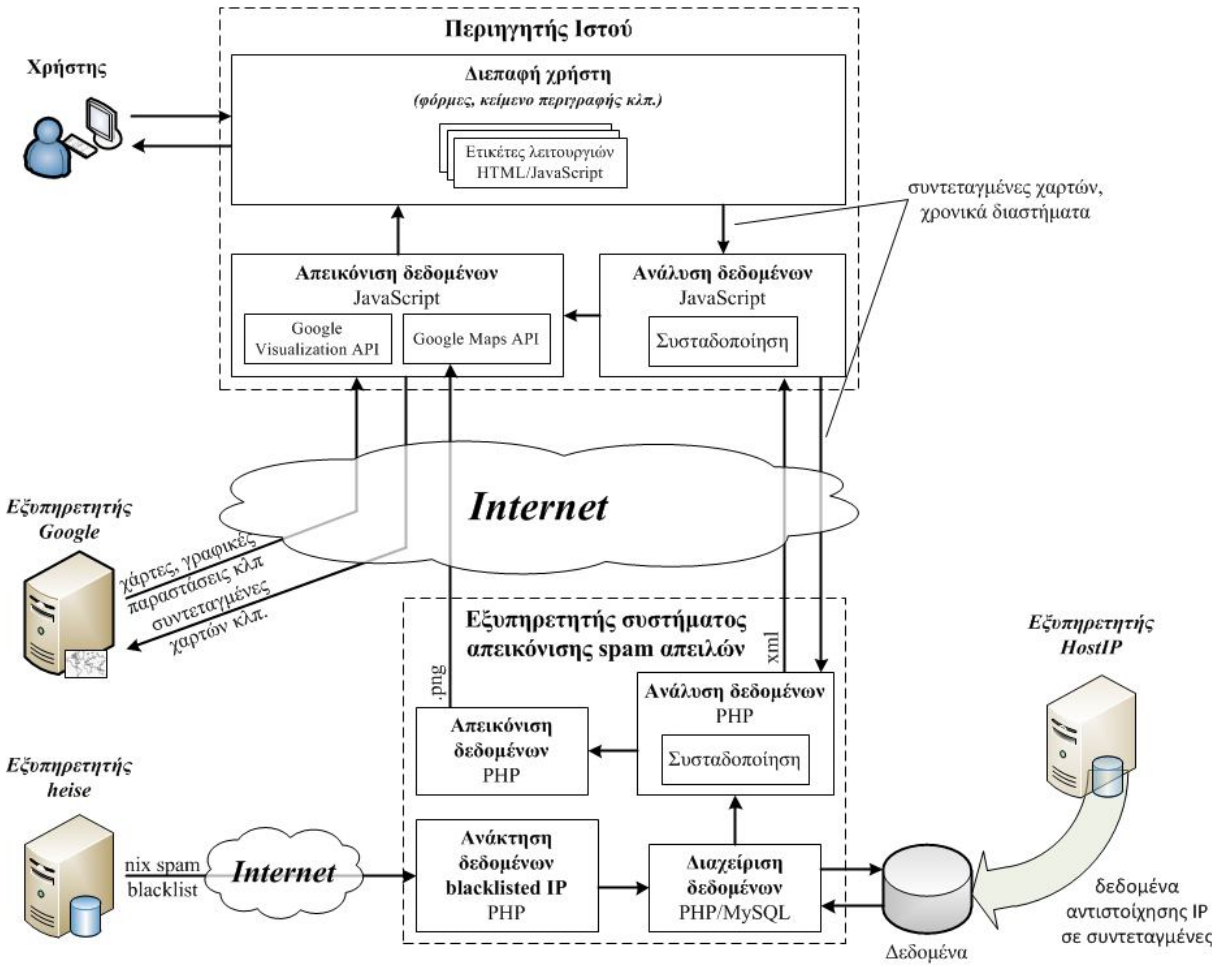
Στα προηγούμενα κεφάλαια έγινε γενικά αναφορά στη φύση του φαινομένου spamming και περιγράφηκαν τα βασικά στοιχεία που θα πρέπει να έχει ένα σύστημα απεικόνισης χωρο-χρονικών δεδομένων που σχετίζονται με την εμφάνιση των απειλών spam ανά τον κόσμο, ώστε να διευκολυνθεί κάποιος χρήστης του συστήματος στην περαιτέρω κατανόηση του φαινομένου αυτού. Μάλιστα έγινε και περιγραφή του τρόπου με τον οποίο μπορεί να επιτευχθεί το τελευταίο παρουσιάζοντας και τους σχετικούς παροχείς από τους οποίους μπορούν να αντλούνται τα απαραίτητα δεδομένα και να τοποθετούνται γεωγραφικά σε χάρτη. Λαμβάνοντας υπόψη όλα αυτά, σχεδιάστηκε και υλοποιήθηκε ένα τέτοιο σύστημα στα πλαίσια της μεταπτυχιακής διατριβής, το οποίο παρέχει τη δυνατότητα σε οποιονδήποτε ενδιαφερόμενο χρήστη πρόσβασης στη συγκεκριμένη οργανωμένη πληροφορία μέσω του Διαδικτύου. Έτσι, σκοπός του παρόντος κεφαλαίου είναι να γίνει μια παρουσίαση του συστήματος αυτού ως προς την αρχιτεκτονική και γενικότερα τις λειτουργίες που το διέπουν.

Πιο συγκεκριμένα, αναφορικά με τη δομή του κεφαλαίου, αρχικά, στην Ενότητα 5.1 παρουσιάζεται η αρχιτεκτονική του υλοποιηθέντος συστήματος περιγράφοντας τη βάση δεδομένων που χρησιμοποιείται και τις λειτουργίες που διέπουν το σύστημα αναφορικά με το κάθε τμήμα από αυτά που αποτελείται. Στη συνέχεια, στην Ενότητα 5.2 περιγράφεται η διεπαφή χρήστη του συστήματος με την οποία δίνεται η δυνατότητα πρόσβασης στη σχετική πληροφορία, ενώ στην Ενότητα 5.3 αναλύεται το σύστημα ως προς θέματα απόδοσης της απεικόνισης και της ανάκτησης δεδομένων. Τέλος, στην Ενότητα 5.4 γίνεται ανάλυση των spamming δεδομένων χρησιμοποιώντας τη διεπαφή χρήστη, ώστε να μελετηθεί η κατάσταση του φαινομένου spamming ανά τον κόσμο για διάφορα χρονικά διαστήματα.

5.1 Αρχιτεκτονική Συστήματος

Το σύστημα απεικόνισης των απειλών spam ανά τον κόσμο το οποίο υλοποιείται στα πλαίσια της μεταπτυχιακής διατριβής, όπως αναφέρθηκε και στο Κεφάλαιο 3, αποτελεί ένα σύστημα διαδικτυακού GIS. Τα βασικά δε χαρακτηριστικά τα οποία το διακρίνουν είναι η χωρο-χρονική φύση των δεδομένων τα οποία πραγματεύεται, που για το λόγο αυτό σημαντική κρίνεται η απεικόνισή τους σε χάρτες, καθώς και το γεγονός ότι κάθε είδους σχετική πληροφορία είναι δημόσια προσβάσιμη μέσω του Διαδικτύου. Μια απλουστευμένη εικόνα της αρχιτεκτονικής ενός τέτοιου συστήματος με αναφορά στην απεικόνιση δεδομένων spam παρουσιάσθηκε στην Εικόνα 3.1, περιγράφοντας τα βασικά μέρη από τα οποία θα πρέπει να αποτελείται, το ρόλο τους, την επικοινωνία τους και τα δεδομένα που ανταλλάσουν. Λαμβάνοντας υπόψη την εικόνα της αρχιτεκτονικής αυτή, καθώς και εκείνη γενικότερα των διαδικτυακών συστημάτων GIS (Εικόνα 3.2), μια λεπτομερής παρουσίαση της αρχιτεκτονικής του υλοποιηθέντος συστήματος γίνεται στην Εικόνα 5.1. Η αντιστοίχηση και κατανομή σε αρχεία κώδικα κάθε στοιχείου που παρουσιάζεται φαίνεται στον Πίνακα 5.1, ενώ μια σύντομη περιγραφή για καθένα από αυτά τα στοιχεία παρουσιάζεται παρακάτω:

- *Διεπαφή χρήστη*: Αποτελεί το γραφικό περιβάλλον του συστήματος με το οποίο καθίσταται δυνατή η αλληλεπίδραση με το χρήστη. Ο τελευταίος ανάλογα με τις τιμές των παραμέτρων που επιβάλλει στο σύστημα, όπως χρονικά διαστήματα, γεωγραφικές συντεταγμένες κ.α., με έμμεσο ή με άμεσο τρόπο, παίρνει ως αποκρίσεις από το σύστημα κατάλληλες απεικονίσεις της σχετικής πληροφορίας που αφορά το φαινόμενο spamming (βλ. Ενότητα 5.2).



Εικόνα 5.1: Αρχιτεκτονική υλοποιηθέντος συστήματος

- *Απεικόνιση δεδομένων (περιηγητής ιστού)*: Αναφέρεται στην κατάλληλη χρησιμοποίηση βιβλιοθηκών συναρτήσεων που παρέχει η Google Inc. από τον κώδικα σε JavaScript που σχετίζεται με τη διεπαφή χρήστη του συστήματος, ώστε να προκύψουν οι επιθυμητές κάθε φορά απεικονίσεις της πληροφορίας που σχετίζεται με το spamming είτε αυτές αφορούν χάρτες είτε γραφικές παραστάσεις.

- *Ανάλυση δεδομένων (περιηγητής ιστού)*: Αναφέρεται στους μηχανισμούς εκείνους με τους οποίους διαχειρίζεται το σύστημα τη βάση δεδομένων όπου αποθηκεύει τη σχετική πληροφορία για το spamming είτε ανακτώντας συγκεκριμένα δεδομένα για κάθε περίπτωση αιτούσης από το χρήστη απεικόνισης είτε ενημερώνοντάς την περιοδικά με νέα δεδομένα που παρέχονται από τους εξυπηρετητές των σχετικών παροχέων που χρησιμοποιούνται από το σύστημα. Βέβαια, οι μηχανισμοί αυτοί περιλαμβάνονται ως επί το πλείστον στο σύστημα διαχείρισης βάσεων δεδομένων (ΣΔΒΔ) της MySQL που χρησιμοποιείται, με τους μηχανισμούς του κυρίου συστήματος να περιορίζονται στην επιβολή επερωτήσεων (queries).

στοιχείο αρχιτεκτονικής \ αρχείο κώδικα	header.html/ menu.html/ footer/.html	index.php	mapSpamApi.php	progress.php	animationApi.php	graphPresentation.php	graphApi.php	genClusterData.php	genAnimationData.php	genGraphData.php	updated.php
Διεπαφή χρήστη	√	√	√	√	√	√	√				
Απεικόνιση δεδομένων (περιηγητής ιστού)			√		√		√				
Ανάλυση δεδομένων (περιηγητής ιστού)			√		√		√				
Διαχείριση δεδομένων								√	√	√	√
Ανάλυση δεδομένων (εξυπηρετητής)								√	√	√	
Απεικόνιση δεδομένων (εξυπηρετητής)										√	
Ανάκτηση δεδομένων blacklisted IP											√

Πίνακας 5.1: Αντιστοίχιση στοιχείων της αρχιτεκτονικής του συστήματος σε αρχεία κώδικα της υλοποίησης

- *Ανάλυση δεδομένων (εξυπηρετητής)*: Περιλαμβάνει την προετοιμασία των ανακτώμενων δεδομένων από τη βάση που προκύπτουν σύμφωνα με τις παραμέτρους που θέτει ο χρήστης, όπου ανάλογα με τον επιθυμητό τρόπο απεικόνισής τους είτε συσταδοποιούνται πρώτα και ύστερα στέλνονται πίσω στον περιηγητή ιστού με την μορφή xml αρχείου είτε στέλνονται κατευθείαν με την μορφή xml αρχείου είτε τροφοδοτούνται στο στοιχείο απεικόνισης δεδομένων.

- *Απεικόνιση δεδομένων (εξυπηρετητής)*: Μετατρέπει τα δεδομένα που τροφοδοτούνται από το στοιχείο ανάλυσης δεδομένων σε εικόνες (.png) τις οποίες και στέλνει πίσω στον εξυπηρετητή ιστού του χρήστη για απεικόνιση των δεδομένων με τη μορφή εναλλαγής εικόνων (animation).

- *Ανάκτηση δεδομένων blacklisted IP*: Ανακτά περιοδικά το αρχείο με τις αποκλεισμένες διευθύνσεις IP της μαύρης λίστας nix spam που παρέχεται από τον παροχέα Heise (βλ. Ενότητα 3.2.1) και ενημερώνει τη βάση δεδομένων του συστήματος με νέες εγγραφές spam απειλών θέτοντας τις κατάλληλες επερωτήσεις.

Στη συνέχεια της συγκεκριμένης ενότητας περιγράφονται αναλυτικά η βάση δεδομένων που χρησιμοποιείται από το σύστημα, καθώς και οι λειτουργίες οι οποίες διέπουν το τελευταίο. Αξίζει στο σημείο αυτό δε να παρουσιασθεί η υποδομή στην οποία φιλοξενείται το υλοποιηθέν

Λογισμικό	Έκδοση
Linux	REDHAT EL5 64bi
Apache HTTP Server	2.2.3
MySQL	5.0.77
PHP	5.3.3

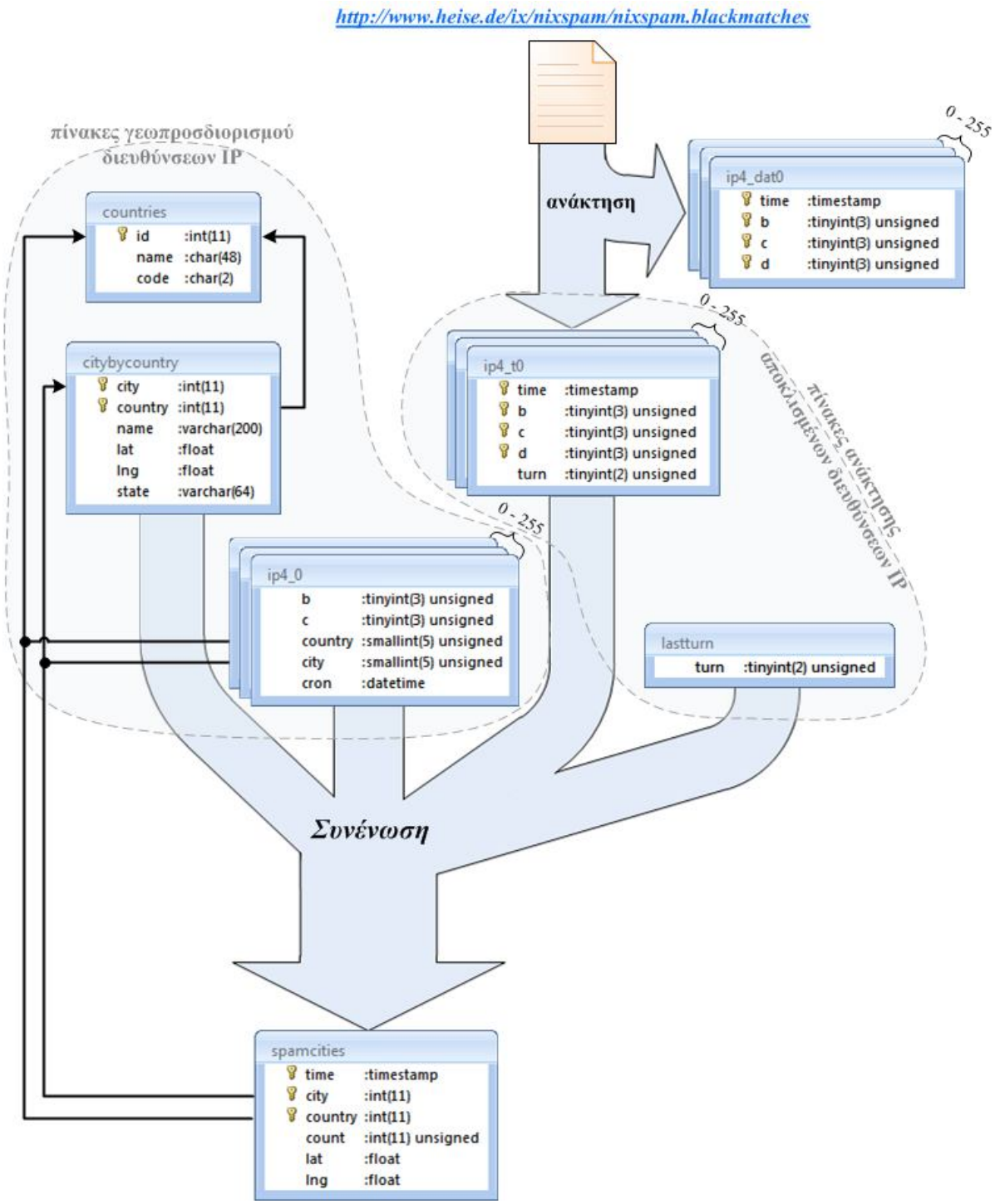
Πίνακας 5.2: Το πακέτο λογισμικών της υποδομής όπου φιλοξενείται το υλοποιηθέν σύστημα

σύστημα. Συγκεκριμένα, πρόκειται για ένα πακέτο ανοιχτού κώδικα (open source) λογισμικών γνωστού και ως LAMP, η ονομασία του οποίου προέρχεται από τα αρχικά των Linux, Apache HTTP Server, MySQL και PHP (ή Perl/PHP/Python). Οι εκδόσεις για κάθε ένα από αυτά παρουσιάζονται στον Πίνακα 5.2.

5.1.1 Βάση Δεδομένων

Όπως παρουσιάσθηκε και στην αρχιτεκτονική του υλοποιηθέντος συστήματος, διατηρείται μια βάση δεδομένων όπου είναι αποθηκευμένες οι σχετικές πληροφορίες περί των εντοπιζόμενων spam απειλών, η οποία ενημερώνεται περιοδικά με νέες εγγραφές που προκύπτουν από το ανακτώμενο αρχείο των αποκλεισμένων διευθύνσεων IP της μαύρης λίστας nix spam. Το στοιχείο της αρχιτεκτονικής του συστήματος που είναι υπεύθυνο για την τελευταία ενέργεια, όπως φαίνεται και από τον Πίνακα 5.1, περιλαμβάνει την περιοδική εκτέλεση ενός PHP αρχείου, του updatadb.php. Για τη δημιουργία αυτών των εγγραφών συνδυάζονται μεταξύ τους η πληροφορία που περιλαμβάνεται στον αρχείο των αποκλεισμένων διευθύνσεων IP και η σχετική αντιστοίχιση διευθύνσεων με γεωγραφικές συντεταγμένες και άλλου τέτοιου είδους χωρική πληροφορία (π.χ. πόλη, χώρα κλπ) που παρέχει η βάση δεδομένων του παροχέα HostIP, της οποίας τα στοιχεία αντιγράφονται (ή συγχρονίζονται) στη βάση δεδομένων του συστήματος επίσης περιοδικά, αλλά ανά πολύ μεγαλύτερα χρονικά διαστήματα (π.χ. μια φορά το μήνα). Παρακάτω γίνεται μια περιγραφή της δομής της βάσης δεδομένων του συστήματος και στη συνέχεια παρουσιάζεται ο τρόπος με τον οποίο ενημερώνεται.

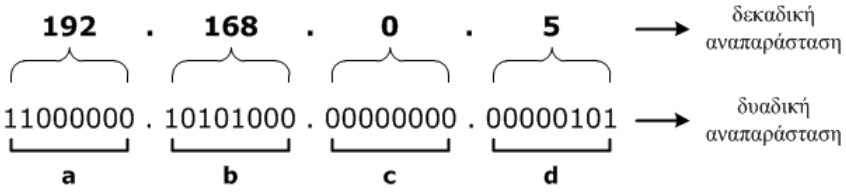
Στην Εικόνα 5.2 παρουσιάζεται η δομή της χρησιμοποιούμενης βάσης δεδομένων. Όπως φαίνεται, τα δεδομένα κατανέμονται σε τρεις κατηγορίες πινάκων: α) οι πίνακες που ουσιαστικά αποτελούν αντίγραφα των αντίστοιχων πινάκων της βάσης δεδομένων του παροχέα HostIP, β) οι πίνακες που χρησιμοποιούνται ως βοηθητικοί κατά την επεξεργασία των δεδομένων που προκύπτουν από την ανάκτηση της λίστας αποκλεισμένων διευθύνσεων IP, και γ) οι πίνακες στους οποίους καταχωρείται η πληροφορία για το spamming.



Εικόνα 5.2: Δομή της βάσης δεδομένων του υλοποιηθέντος συστήματος

Σχετικά με την πρώτη κατηγορία πινάκων της βάσης δεδομένων, αυτήν την αποτελούν οι εξής πίνακες:

- *countries*: Περιλαμβάνει στοιχεία των χωρών όπως όνομα (*name*) και κωδικό χώρας (*code*) σύμφωνα με το πρότυπο δύο γραμμάτων ISO 3166-1 alpha-2 [42], ενώ η μοναδικότητα κάθε εγγραφής καθορίζεται από έναν αύξοντα αριθμό (*id*).



Εικόνα 5.3: Δυαδική και δεκαδική αναπαράσταση διεύθυνσης IP (v4) με bytes και αριθμούς,

- *cityByCountry*: Περιλαμβάνει στοιχεία των πόλεων όπως όνομα (*name*), γεωγραφικό πλάτος (*lat*) και μήκος (*lng*), χώρα μέσω του ξένου κλειδιού (*country*) στο αντίστοιχο πεδίο του πίνακα *countries*, και πολιτεία (*state*), ενώ η μοναδικότητα κάθε εγγραφής καθορίζεται από το συνδυασμό του ξένου κλειδιού της χώρας και ενός αριθμού που ξεχωρίζει τις πόλεις μιας χώρας μεταξύ τους (*city*).

- *ip4_0 – ip4_255*: Πρόκειται για 256 πίνακες οι οποίοι αντιστοιχούν στον πρώτο αριθμό της αναπαράστασης (dot-decimal notation) της διεύθυνσης IP, όπως φαίνεται στην Εικόνα 5.3, και τα πεδία τους περιλαμβάνουν τους δύο επόμενους αριθμούς της αναπαράστασης της διεύθυνσης (*b* και *c*), τα ξένα κλειδιά των χωρών και των πόλεων (*country* και *city*), και τη χρονική στιγμή που έγινε η καταχώρηση της κάθε εγγραφής (*cron*). Ουσιαστικά, μέσω των πινάκων αυτών γίνεται η αντιστοίχιση μεταξύ διευθύνσεων IP (ανά υποδίκτυα /24) και πόλεων, και σε συνδυασμό με τα στοιχεία της κάθε πόλης δύναται να προσφέρουν μια αντιστοίχιση διευθύνσεων IP με γεωγραφικές συντεταγμένες – γεωπροσδιορισμός IP διευθύνσεων – που είναι και το ζητούμενο για το υλοποιηθέν σύστημα.

Όσον αφορά τη δεύτερη κατηγορία πινάκων, αυτή περιλαμβάνει τους εξής πίνακες:

- *ip4_t0 – ip4_t255*: Πρόκειται για 256 πίνακες οι οποίοι στα πεδία τους περιλαμβάνουν και τους τρεις επόμενους αριθμούς της αναπαράστασης της διεύθυνσης IP (*b*, *c* και *d*), τη χρονική στιγμή εντοπισμού μιας διεύθυνσης IP ως spam απειλή (*time*) με ακρίβεια ώρας και έναν αριθμό (*turn*) που χαρακτηρίζει κάθε εγγραφή ως τρέχουσα ή όχι ανάλογα με το αν εκχωρήθηκε κατά την διάρκεια της τελευταίας ανάκτησης της λίστας των αποκλεισμένων διευθύνσεων IP ή όχι. Ουσιαστικά το μέγεθος των πινάκων αυτών παραμένει πάντοτε περιορισμένο, όπως περιγράφεται και παρακάτω, στην παρουσίαση του τρόπου ενημέρωσης της βάσης δεδομένων, αφού περιλαμβάνει πάντοτε τις διευθύνσεις IP της τελευταίας ανάκτησης, και για το λόγο αυτό θεωρούνται ως

βοηθητικοί πίνακες. Ο δε ρόλος τους αφορά την επιτάχυνση αυτής της διαδικασίας ανάκτησης.

- *lastturn*: Αποτελεί ουσιαστικά μια μεταβλητή η οποία ενημερώνεται με το πέρας κάθε ανάκτησης αποκλεισμένων διευθύνσεων IP, διατηρώντας στο σχετικό πεδίο (*turn*) έναν αριθμό που σχετίζεται με την επόμενη ανάκτηση.

Τέλος, σχετικά με τους πίνακες που διατηρούν την πληροφορία για το spamming:

- *ip4_dat0 – ip4_dat255*: Πρόκειται για 256 πίνακες που διατηρούν ιστορικό όλων των διευθύνσεων IP που έχουν εντοπισθεί ως spam απειλές, επισυνάπτοντας τα δεδομένα που προκύπτουν από την σχετική ανάκτηση. Για το λόγο αυτό και τα πεδία αυτών των πινάκων είναι τα ίδια με εκείνα των ip4_t0 - ip4_t255, εξαιρώντας το ότι η ακρίβεια του χρονικής στιγμής εντοπισμού των διευθύνσεων IP είναι σε λεπτά, και βέβαια το πεδίο *turn*.

- *spamcities*: Ουσιαστικά αποτελεί το βασικό πίνακα δεδομένων του υλοποιηθέντος συστήματος μιας και οποιαδήποτε πληροφορία ζητείται από το χρήστη μέσω της σχετικής διεπαφής με το σύστημα αντλείται από τον συγκεκριμένο πίνακα. Εξαιτίας του γεγονότος αυτού, τα πεδία που περιλαμβάνει είναι άμεσα συνδεδεμένα με τις ανάγκες απεικόνισης του spamming φαινομένου, τις οποίες ικανοποιεί και η διεπαφή χρήστη. Πιο συγκεκριμένα, τα δεδομένα ενημερώνονται κατά τη διάρκεια της διαδικασίας ανάκτησης των αποκλεισμένων διευθύνσεων IP από τον παροχέα Heise, συνενώνοντας για κάθε διεύθυνση IP που εισάγεται σε κάποιον από τους πίνακες ip4_t0 - ip4_t255 τα στοιχεία συνολικά τεσσάρων πινάκων, όπως φαίνεται στην Εικόνα 2, ανάλογα με τον πρώτο αριθμό της αναπαράστασης της διεύθυνσης IP (π.χ. αν η IP είναι της μορφής 185.xxx.xxx.xxx, τότε στη συνένωση συμμετέχουν οι πίνακες cityByCountry, ip4_185, ip4_t185 και lastturn). Αναφορικά δε με τα πεδία του συγκεκριμένου πίνακα, αυτά αφορούν τα ξένα κλειδιά των χωρών και των πόλεων (*country* και *city*), τις γεωγραφικές συντεταγμένες (*lat* και *lng*) που χαρακτηρίζουν κάθε πόλη, και τον αριθμό (*count*) των spam απειλών που εμφανίζονται στη συγκεκριμένη κάθε φορά πόλη (για αυτό και το όνομα του πίνακα «spamcities») για κάποια συγκεκριμένη ώρα (*time*).

Όπως αναφέρθηκε και στην αρχή της συγκεκριμένης υποενότητας, ο τρόπος με τον οποίο ενημερώνεται η βάση δεδομένων του υλοποιηθέντος συστήματος ως προς την πληροφορία του

spamming έγκειται στην περιοδική εκτέλεση του αρχείου updatedb.php, η οποία σύμφωνα με τις ανάγκες απεικόνισης γίνεται κάθε μια ώρα. Για να επιτευχθεί αυτή η περιοδική εκτέλεση του συγκεκριμένου αρχείου χρησιμοποιείται το πρόγραμμα *Cron* που παρέχεται από το λειτουργικό σύστημα Linux του εξυπηρετητή. Πρόκειται για ένα βοηθητικό πρόγραμμα που χρησιμοποιείται για το χρονοπρογραμματισμό διεργασιών το οποίο τρέχει στο παρασκήνιο ελέγχοντας συνεχώς το αρχείο *crontab* όπου αποθηκεύονται οι προς εκτέλεση διεργασίες. Για την περίπτωση του συγκεκριμένου συστήματος και την περιοδική εκτέλεση του updatedb.php πρέπει να εκχωρηθεί στο αρχείο *crontab* η εξής εντολή:

```
7 * * * * /usr/bin/php /**path**/updatedb.php
```

Τα πέντε πρώτα πεδία που αφορούν το κάθε πότε θα εκτελείται το συγκεκριμένο αρχείο updatedb.php, δηλώνουν ότι θα πρέπει αυτό να γίνεται κάθε 7ο λεπτό της ώρας, για κάθε ώρα, ημέρα, μήνα, εβδομάδα και ημέρα της εβδομάδας που εκφράζουν οι αστερίσκοι (*), αντίστοιχα.

Όσον αφορά δε τη σχετική λειτουργία που υλοποιεί το αρχείο updatedb.php, θα μπορούσε να περιγραφεί με τα εξής βήματα:

1. Ανακτάται το αρχείο αποκλεισμένων διευθύνσεων IP της μαύρης λίστας nix spam από το σχετικό εξυπηρετητή του παροχέα Heise μέσω του συνδέσμου http://www.heise.de/ix/nixspam/nixspam.blackmatches.

2. Επιστρέφεται μέσω κατάλληλης επερώτησης από τη βάση δεδομένων ο αριθμός της τρέχουσας ανάκτησης, *turn*, του πίνακα lastturn.

3. Για κάθε εγγραφή διεύθυνσης IP του αρχείου της μαύρης λίστας που διαβάζεται:

   3.1. Μετατρέπεται η χρονική στιγμή εντοπισμού σε ζώνη ώρας GMT0.

   3.2. Εισάγονται τα στοιχεία της εγγραφής στον κατάλληλο πίνακα από το σύνολο των ip4_dat0 - ip4_dat255.

   3.3. Εισάγονται τα στοιχεία της εγγραφής στον κατάλληλο πίνακα από το σύνολο των ip4_t0 - ip4_t255 με τη χρονική στιγμή εντοπισμού να είναι σε ακρίβεια ώρας και τον αριθμό ανάκτησης *turn* να είναι ίσος με τον αριθμό της τρέχουσας ανάκτησης. Με αυτόν τον τρόπο φιλτράρονται οι εγγραφές με απαλοιφή διπλότυπων (duplicates)

θεωρώντας ότι μια διεύθυνση IP του αρχείου που εμφανίζεται περισσότερες από μια φορές σε κάποια ώρα της ημέρας αποτελεί μία μόνο απειλή spam.

3.4. Αν πρόκειται για διπλότυπη εγγραφή, αυξάνεται ο αριθμός διπλότυπων εγγραφών σε σειρά κατά 1, διαφορετικά μηδενίζεται.

3.5. Αν δεν υπάρχουν άλλες εγγραφές να διαβαστούν από το αρχείο ή αν ο αριθμός των διπλότυπων εγγραφών σε σειρά υπερβεί κάποιο όριο (π.χ. 50), τότε γίνεται μετάβαση στο βήμα 4, διαφορετικά συνεχίζεται το βήμα 3. Με αυτόν τον τρόπο, βελτιώνεται ο συνολικός χρόνος της διαδικασίας ενημέρωσης της βάσης δεδομένων του συστήματος, αφού περιορίζονται σε αριθμό οι χρονοβόρες εντολές εισαγωγής σχεδόν σε εκείνες που είναι απαραίτητες.

4. Αν έχει γίνει έστω και μια ενημέρωση στις εγγραφές των πινάκων παραπάνω, τότε η διαδικασία συνεχίζεται στο επόμενο βήμα, διαφορετικά ολοκληρώνεται.

5. Διαγράφονται οι εγγραφές από τους πίνακες ip4_t0 - ip4_t255 που έχουν αριθμό ανάκτησης *turn* διαφορετικό από αυτόν της τρέχουσας ανάκτησης θέτοντας κατάλληλες εντολές διαγραφής.

6. Ενημερώνεται ο πίνακας lastturn και συγκεκριμένα ο αριθμός ανάκτησης, *turn*, μέσω μιας εντολής ενημέρωσης, ώστε να περιλαμβάνει την τιμή της τρέχουσας ανάκτησης αυξημένη κατά 1.

7. Για κάθε ζεύγος πινάκων μεταξύ των συνόλων ip4_t0 - ip4_t255 και ip4_0 - ip4_255 γίνεται τριπλή συνένωση με τον πίνακα cityByCountry ως προς τα πεδία *b*, *c*, *country* και *city*, και με ομαδοποίηση ως προς τα *time*, *city* και *country*, ώστε να απαριθμηθούν οι απειλές spam (*count*) που αφορούν μια συγκεκριμένη πόλη σε μια συγκεκριμένη ώρα της ημέρας στον πίνακα spamcities. Σε περίπτωση δε που υπάρχει ήδη μια τέτοια εγγραφή όσον αφορά το σύνθετο κλειδί *time-city-country* του συγκεκριμένου πίνακα, τότε εκτελείται μια εντολή ενημέρωσης της συγκεκριμένης εγγραφής ως προς την απαρίθμηση των απειλών spam (*count*). Ένα παράδειγμα τέτοιας εντολής για το ζεύγος πινάκων ip4_t185 και ip4_185 είναι:

```
INSERT INTO spamcities (time, city, country, lat, lng, count)
  SELECT db2.time AS Time, db1.city AS City, db1.country AS Country,
    CityByCountry.lat, CityByCountry.lng, COUNT(*)
  FROM ip4_t185 AS db2, ip4_185 AS db1,
    cityByCountry AS CityByCountry
  WHERE db2.b=db1.b AND db2.c=db1.c AND
    db1.country=CityByCountry.country AND
    db1.city=CityByCountry.city
  GROUP BY Time, City, Country
ON DUPLICATE KEY UPDATE count=count+VALUES(count)
```

5.1.2 Λειτουργίες

Έχοντας παρουσιασθεί η βάση δεδομένων του συστήματος και ο τρόπος με τον οποίο αυτή ενημερώνεται, η συγκεκριμένη υποενότητα προβαίνει στην παρουσίαση των υπόλοιπων μερών του συστήματος. Αυτά αναφέρονται στις λειτουργίες οι οποίες διέπουν το κυρίως σύστημα και στηρίζονται κατεξοχήν στην ανάκτηση των δεδομένων της βάσης που σχετίζονται με το φαινόμενο spamming.

Αρχικά, θα μπορούσε να ειπωθεί ότι υπάρχουν τρεις γενικευμένες λειτουργίες οι οποίες αντιστοιχούν στις επίσης τρεις δυνατότητες απεικόνισης των δεδομένων του spamming που προσφέρονται από το σύστημα στο χρήστη μέσω της σχετικής διεπαφής. Έτσι, οι λειτουργίες και συνάμα τα σχετικά με αυτές αρχεία κώδικα που υλοποιούν το σύστημα χωρίζονται στις εξής κατηγορίες:

i. Λειτουργίες συσταδοποίησης και απεικόνισης των δεδομένων σε χάρτες.

ii. Λειτουργίες δημιουργίας εικόνων για απεικόνιση των δεδομένων με εφέ κίνησης πάνω σε χάρτη.

iii. Λειτουργίες απεικόνισης δεδομένων με γραφικές παραστάσεις και ραβδογράμματα.

Βασική δε δομή όσον αφορά την υλοποίηση για καθεμιά από αυτές της κατηγορίες αποτελεί η συνεργασία μεταξύ λειτουργιών που εκτελούνται στην πλευρά του περιηγητή ιστού και εκείνων στην πλευρά του εξυπηρετητή του συστήματος. Για το λόγο αυτό, κάθε κατηγορία λειτουργιών περιλαμβάνει μια τριάδα βασικών αρχείων κώδικα που αναφέρονται σε τρεις βασικές τεχνολογίες διαδικτύου διαφορετικού σκοπού, HTML, JavaScript και PHP. Βέβαια, υπάρχουν και αρχεία κώδικα τα οποία είναι κοινά και για τις τρεις κατηγορίες λειτουργιών. Αυτά

περιλαμβάνουν κυρίως τιμές σταθερών και παραμέτρων του συστήματος, καθώς και σχετικούς ορισμούς συναρτήσεων που χρησιμοποιούνται σε περισσότερα του ενός αρχεία κώδικα. Αυτά είναι τα εξής:

- *dbinfo.php*: Περιλαμβάνει ορισμούς μεταβλητών στις οποίες αποθηκεύεται σημαντική πληροφορία, όπως όνομα χρήστη και κωδικός πρόσβασης για σύνδεση με τη βάση δεδομένων MySQL, όνομα βάσης δεδομένων και κλειδί χρήσης της βιβλιοθήκης συναρτήσεων της Google Inc (Google Maps API και Google Visualization API).

- *includings.php*: Περιλαμβάνει σταθερές και παραμέτρους που αφορούν χαρακτηριστικά των χαρτών και τη διαδικασία συσταδοποίησης, διάφορες διαδρομές αρχείων (paths) (π.χ. εικόνες), και τέλος ορισμούς συναρτήσεων που αφορούν την επεξεργασία χρονικών δεδομένων (π.χ. μετατροπή ζώνης ώρας).

- *header.html, menu.html, footer.html, style.css*: Ουσιαστικά αποτελούν εκείνα τα αρχεία που διαμορφώνουν το γραφικό κυρίως περιβάλλον που παραμένει συνήθως σταθερό (ανάλογα βέβαια με τον αναπτυσσόμενη ιστοσελίδα) γύρω από το πλαίσιο όπου γίνονται οι διάφορες απεικονίσεις δεδομένων spamming.

Λειτουργίες Συσταδοποίησης και Απεικόνισης των Δεδομένων σε Χάρτες

Αδιαμφισβήτητα, η πρώτη κατηγορία λειτουργιών είναι και η σημαντικότερη όσον αφορά την μεταπτυχιακή διατριβή, διότι εκτός από την απεικόνιση των δεδομένων περιλαμβάνει και τους μηχανισμούς που αφορούν τη συσταδοποίησή τους, ιδιαίτερα εκτενής αναφορά για την οποία έγινε στο προηγούμενο κεφάλαιο. Η τριάδα αρχείων η οποία την αποτελεί είναι: *index.php*, *mapSpamApi.php* και *genClusterData.php*.

Σχετικά με το ***index.php***, πρόκειται για το αρχείο που εκτελείται κατά την αρχική προσπέλαση της διεύθυνσης του ιστότοπου του συστήματος, και υλοποιεί το περιβάλλον της διεπαφής χρήστη σε εκείνο το πλαίσιο της ιστοσελίδας όπου προβάλλονται οι απεικονίσεις δεδομένων για τη συγκεκριμένη κατηγορία λειτουργιών. Βασικά στοιχεία αυτού αποτελούν: α) η παρουσίαση στο χρήστη της δυνατότητας επιλογής του συνολικού χρονικού διαστήματος που προσδιορίζει κάθε φορά τα επιστρεφόμενα από το σύστημα δεδομένα, μέσω των επιμέρους επιλογών της χρονικής βαθμίδας (π.χ. ώρα, ημέρα, εβδομάδα κλπ) που μελετάται ανά χάρτη καθώς και του πλήθους αυτών (π.χ. 1 ώρα, 5 εβδομάδες κλπ), β) η οριοθέτηση και κατανομή πολλαπλών

χαρτών και των αντίστοιχων σε αυτούς περιγραφικών στοιχείων ανάλογα με την επιλογή του πλήθους των χρονικών βαθμίδων, και γ) οι συνδέσεις τόσο με το σχετικό αρχείο του συστήματος που περιλαμβάνει τον κώδικα JavaScript, mapSpamApi.php, όσο και με το αντίστοιχο αρχείο της βιβλιοθήκης συναρτήσεων Google Maps API v2 της Google Inc. για τη δυνατότητα προβολής χαρτών. Ο λόγος δε που τόσο το πρώτο όσο και το δεύτερο αρχείο της τριάδας έχουν κατάληξη .php, και όχι .html και js, αντίστοιχα, είναι ότι αποτελούν κάθε φορά στιγμιότυπα των επιλογών του χρήστη όσον αφορά: α) τις χρονικές περιόδους (τύπος χρονικής βαθμίδας και πλήθος χρονικών βαθμίδων) οι οποίες προσδιορίζουν χρονικά τα δεδομένα, β) τη χρονική στιγμή (ώρα, ημέρα, μήνας, έτος) που υποβάλλονται οι συγκεκριμένες επιλογές η οποία λειτουργεί ως σημείο χρονικής αναφοράς, και γ) τις έμμεσες επιλογές που αφορούν την κατάσταση των χαρτών (ζουμ, γεωγραφικές συντεταγμένες του κέντρου των χαρτών) μέσω της περιήγησης σε αυτούς οι οποίες προσδιορίζουν χωρικά τα δεδομένα.

Από την άλλη μεριά, του εξυπηρετητή, το αρχείο ***genClusterData.php*** είναι υπεύθυνο για την ανάκτηση δεδομένων από τη βάση του συστήματος, τη συσταδοποίηση των δεδομένων αυτών και τη δημιουργία ενός XML αρχείου που να τα περιγράφει. Οι παράμετροι δε που καθορίζουν το ποια δεδομένα θα λάβουν μέρος σε αυτήν τη διαδικασία είναι το ζουμ και οι γεωγραφικές συντεταγμένες του «αιτούντος» χάρτη, καθώς και η χρονική περίοδος κατά τη διάρκεια της οποίας εντοπίζονται τα σχετικά δεδομένα spamming.

Όσον αφορά το αρχείο ***mapSpamApi.php***, αυτό είναι που προσδίδει τη λειτουργικότητα γενικότερα στον ιστότοπο και ειδικότερα στη δυναμική αλλαγή της απεικόνισης των δεδομένων στους χάρτες με την περιήγηση του χρήστη χρησιμοποιώντας το Google Maps API. Η λειτουργικότητα γενικότερα στον ιστότοπο αφορά μεταξύ άλλων την προσαρμογή των κειμένων και γραφικών της ιστοσελίδας ανάλογα με την κατάσταση απεικόνισης των δεδομένων και τις επιλογές του χρήστη, μέρος της οποίας παρουσιάζεται στην Εικόνα 5.4 καθώς και στις εικόνες της επόμενης ενότητας που αφορούν τη διεπαφή χρήστη. Ιδιαίτερη έμφαση, ωστόσο, δίνεται στην προσφερόμενη λειτουργικότητα των χαρτών που αφορούν την απεικόνιση των δεδομένων spamming, βασικό στοιχείο της οποίας αποτελεί η χρήση της AJAX (Asynchronous JavaScript and XML) [35]. Πρόκειται για μια προγραμματιστική τεχνική που συνδυάζει διάφορες διαδικτυακές τεχνολογίες, όπως HTML, XML, Cascading Style Sheets (CSS) και JavaScript, με τέτοιο τρόπο ώστε να επιταχύνεται η διαδικτυακή επικοινωνία πελάτη-εξυπηρετητή και συνάμα να βελτιώνεται η ποιότητα περιήγησης του χρήστη στις ιστοσελίδες. Κυριότερο στοιχείο στο οποίο οφείλεται η επιτάχυνση αυτή αποτελεί η ανταλλαγή μικρού μεγέθους μηνυμάτων με τον εξυπηρετητή ανεξάρτητα από τη δραστηριότητα του χρήστη, ώστε

ο τελευταίος όταν πρόκειται τελικά να ζητήσει δεδομένα από τον εξυπηρετητή να μη χρειάζεται να φορτωθεί εξ ολοκλήρου το περιεχόμενο της ιστοσελίδας στην οποία περιηγείται (ασύγχρονη επικοινωνία). Το XMLHttpRequest είναι εκείνο το (προγραμματιστικό) αντικείμενο το οποίο υλοποιεί τη συγκεκριμένη επικοινωνία με το δεδομένα των μηνυμάτων που μεταφέρονται να είναι σε μορφή XML. Αναφορικά δε με την περίπτωση συγκεκριμένα του Google Maps API, παρέχεται ένα παρόμοιο αντικείμενο, το GXmlHttp, το οποίο υλοποιεί το πρωτόκολλο επικοινωνίας που περιλαμβάνεται στο XMLHttpRequest με τέτοιο τρόπο ώστε να αντιμετωπίζονται προβλήματα ασυμβατότητας μεταξύ των διαφόρων περιηγητών ιστού.

Γενικότερα, λαμβάνοντας τα παραπάνω υπόψη, η λειτουργία που επιτυγχάνεται με αυτήν την τριάδα αρχείων θα μπορούσε να περιγραφεί ως προς τα βασικά βήματα που ακολουθούνται ως εξής:

1. Αν ο χρήστης επιλέξει το συγκεκριμένο τρόπο απεικόνισης των δεδομένων spamming από την αντίστοιχη ετικέτα στη διεπαφή, τότε φορτώνεται ολόκληρη η σελίδα (index.php, mapSpamApi.php) στον περιηγητή ιστού με συγκεκριμένες αρχικές τιμές παραμέτρων: «χρονική βαθμίδα» = ώρα, «αριθμός χρονικών βαθμίδων» = 1, «χρονική στιγμή» = τρέχουσα ώρα (τρέχουσες τιμές για ώρα, ημέρα, μήνα και έτος), «ζουμ» = 0, «γεωγραφικό πλάτος» = 0.0, «γεωγραφικό μήκος» = 0. Μετάβαση στο βήμα 6.

2. Αν ο χρήστης έχει επιχειρήσει να κάνει αλλαγή του χρονικού διαστήματος που αφορά τα δεδομένα προς απεικόνιση είτε μέσω αλλαγής της χρονικής βαθμίδας είτε μέσω αλλαγής του πλήθους αυτών, τότε φορτώνεται ολόκληρη η σελίδα στον περιηγητή ιστού με τιμές για τις υπόλοιπες παραμέτρους όπως αυτές είχαν καθορισθεί από προηγούμενη δραστηριότητα του χρήστη. Μετάβαση στο βήμα 6.

3. Αν ο χρήστης επιχειρήσει την αλλαγή της κατάστασης (*Zoom*, *Move*) κάποιου χάρτη (μεγέθυνση/σμίκρυνση ή μετακίνηση) μέσω της περιήγησής του σε αυτόν, τότε ενημερώνονται (*setZoom*, *setCenter*) οι καταστάσεις και των υπόλοιπων χαρτών.

4. Ελέγχεται αν η νέα κατάσταση των χαρτών υπερβαίνει τα σχετικά περιθώρια που ορίζονται για την τιμή του ζουμ και για τις συντεταγμένες του κέντρου τους.

5. Αν τα υπερβαίνει, τότε εκτελείται το βήμα 7, διαφορετικά το βήμα 9.

6. Μόλις ολοκληρωθεί η φόρτωση της σελίδας, ξεκινάει μια διαδικασία (*load*) κατά την οποία φορτώνονται χάρτες της Google Inc. σε πλαίσια μεγέθους 500x300 ο αριθμός των οποίων είναι ίσος με αυτόν των χρονικών βαθμίδων, ενώ η τοποθεσία που απεικονίζεται καθορίζεται από τις σχετικές παραμέτρους που συνοδεύουν τη σελίδα.

7. Για κάθε χάρτη:

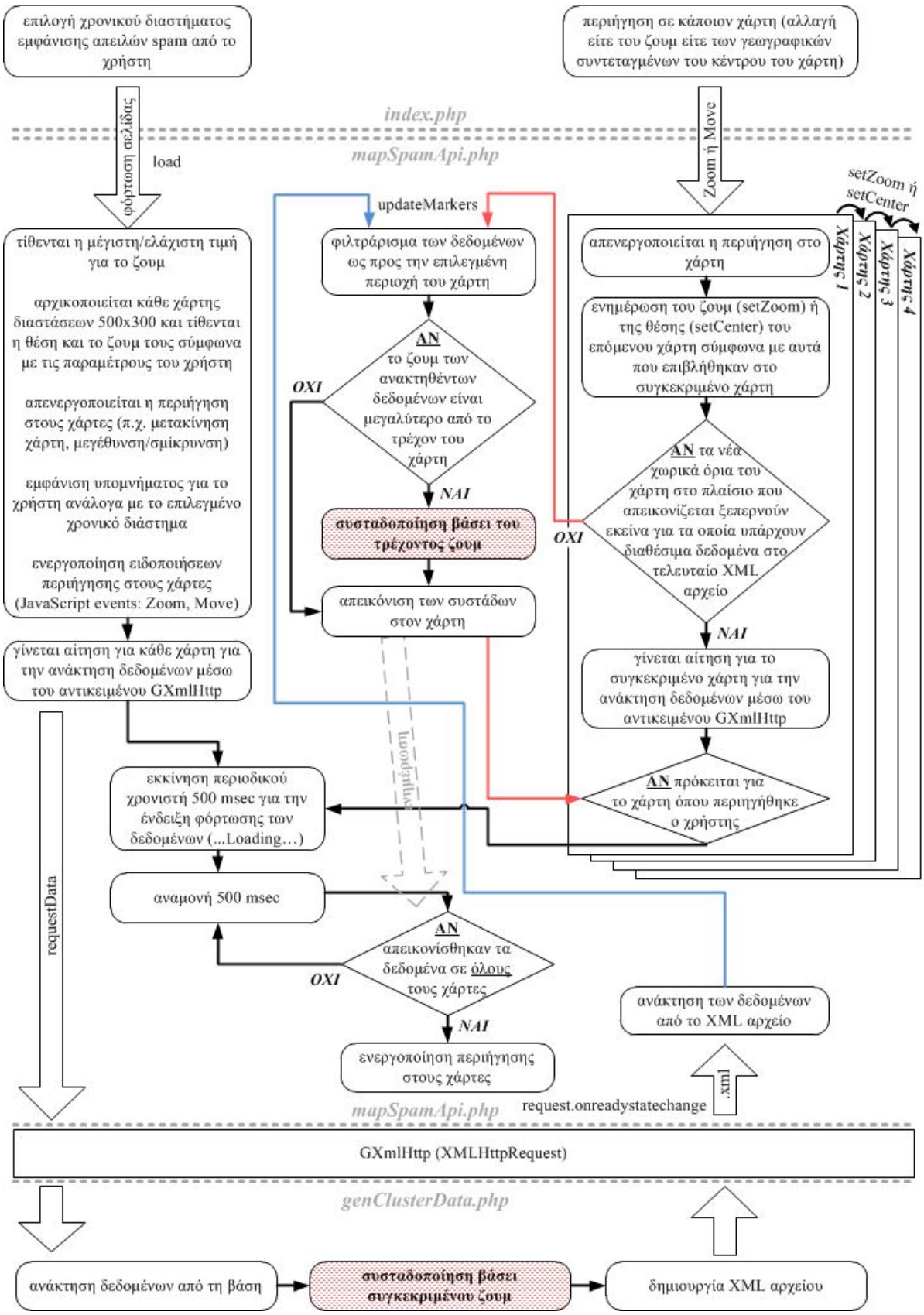
    7.1. Υπολογίζεται η χρονική περίοδος των δεδομένων spamming που αντιπροσωπεύει.

    7.2. Αρχικοποιείται μια διαδικασία (*requestData*) ανάκτησης των σχετικών με το χάρτη δεδομένων με μορφή XML από τον εξυπηρετητή, ζητώντας ουσιαστικά μέσω ενός αντικειμένου GXmlHttp το αρχείο genClusterData.php. Οι παράμετροι δε που συνοδεύουν το τελευταίο είναι οι ίδιες με τις τρέχουσες της ιστοσελίδας με εξαίρεση εκείνη της χρονικής περιόδου που αντιπροσωπεύει ο χάρτης.

8. Για κάθε ολοκλήρωση (*request.onreadystatechange*) μεταφοράς δεδομένων (XML) μέσω του αντικειμένου GXmlHttp, γίνεται ανάλυση του σχετικού XML αρχείου και ανάκτηση των σχετικών δεδομένων που περιλαμβάνονται.

9. Ενημερώνονται (*updateMarkers*) τα δεδομένα κάθε χάρτη (ενδεχομένως, αφού προηγηθεί συσταδοποίηση – βλ. παρακάτω).

Όπως αναφέρθηκε και προηγουμένως, βασικό στοιχείο της παρούσας κατηγορίας λειτουργιών είναι η (χωρική) συσταδοποίηση που διενεργείται στα δεδομένα με σκοπό την καλύτερη αναπαράστασή τους πάνω σε χάρτες. Βέβαια, για λόγους ποιότητας ή απόδοσης γενικότερα της απεικόνισης (βλ. Κεφάλαιο 4), αυτό επιτυγχάνεται με μια δομή πελάτη-εξυπηρετητή, όπως φαίνεται και στην Εικόνα 5.4 (οι λειτουργίες συσταδοποίησης φαίνονται με κόκκινη σκίαση), με την έννοια ότι η συσταδοποίηση διενεργείται τόσο στον περιηγητή ιστού όσο και στον εξυπηρετητή, και κατά συνέπεια υλοποιείται τόσο σε JavaScript (mapSpamApi.php) όσο και σε PHP (genClusterData.php), αντίστοιχα. Η επικοινωνία δε μεταξύ των δύο τμημάτων της συσταδοποίησης, όπως περιγράφθηκε και νωρίτερα, στηρίζεται στη χρήση του αντικειμένου GXmlHttp μέσω του οποίου το τμήμα του εξυπηρετητή τροφοδοτεί με δεδομένα-συστάδες σε XML μορφή το τμήμα που εκτελείται στον περιηγητή ιστού.



Εικόνα 5.4: Λειτουργίες συσταδοποίησης και απεικόνισης των δεδομένων σε χάρτες

Όσον αφορά την μέθοδο η οποία χρησιμοποιείται για τη συσταδοποίηση, λαμβάνονται υπόψη τα χαρακτηριστικά εκείνα που συμβάλλουν σε μια «καλή» συσταδοποίηση όπως υποδείχθηκαν στην Ενότητα 4.7, και ενσωματώνονται στην υλοποίηση των αντίστοιχων τμημάτων του συστήματος. Πιο συγκεκριμένα, χρησιμοποιούνται τα centroids για την αναπαράσταση των συστάδων, τα αντικείμενα των δεδομένων ανατίθενται σε συστάδες σύμφωνα με τη διαδικασία της πρώτης φάσης της BIRCH, όπως επίσης και σε κελιά, γίνεται «επανασυσταδοποίηση» χρησιμοποιώντας μια απλή μέθοδο συνένωσης των συστάδων, και τέλος ανατίθενται τα αντιπροσωπευτικά κελιά πλέον των αντικειμένων στις διαμορφωμένες συστάδες για την παραγωγή των τελικών συστάδων. Επιπλέον αυτών, σημαντικό ρόλο στη σχετική ποιότητα υπηρεσίας παίζει και ο τρόπος με τον οποίο κατανέμεται η συσταδοποίηση στον εξυπηρετητή και στον περιηγητή ιστού. Για το σκοπό αυτό υιοθετείται μια διαδικασία κατά την οποία για μεμονωμένα ζουμ ο εξυπηρετητής ανακτά από τη βάση δεδομένων αντικείμενα (απειλές spam) που εντοπίζονται σε τετραπλάσια επιφάνεια από εκείνη που καθορίζει ο χρήστης μέσω της περιήγησής του σε κάποιο χάρτη, ώστε αυτά να είναι διαθέσιμα σε περίπτωση μετακίνησης του χάρτη σε γειτονικές θέσεις και να μην χρειάζεται εκ νέου επικοινωνία με τον εξυπηρετητή. Επίσης, λόγω αυτής της ανάκτησης παραπάνω δεδομένων από τη βάση, ορίζονται συγκεκριμένα ζουμ στα οποία δεν χρειάζεται να υπάρξει καθόλου επικοινωνία με τον εξυπηρετητή, αναλαμβάνοντας τη συσταδοποίηση μόνο η πλευρά του περιηγητή ιστού. Δεδομένης της άμεσης σύνδεσης μεταξύ της απόδοσης της συσταδοποίησης και των μηχανισμών που υιοθετούνται, περισσότερο λεπτομερής παρουσίαση όλων αυτών γίνεται στην Ενότητα 5.3.1. Σχετικά με τα βήματα της διαδικασίας συσταδοποίησης, αυτά μπορούν να συνοψισθούν ως εξής:

1. Αρχικά, ο χρήστης προσδιορίζει τα δεδομένα του spamming προς απεικόνιση σε κάποιο χάρτη επιλέγοντας τόσο το χρονικό διάστημα αναφοράς μέσω των σχετικών επιλογών της διεπαφής όσο και την περιοχή ενδιαφέροντος στον χάρτη (γεωγραφικές συντεταγμένες του κέντρου του χάρτη και ζουμ) μέσω της περιήγησής του σε αυτόν.

2. Οι επιλογές αυτές μεταφέρονται μέσω του αντικειμένου GXmlHttp στον εξυπηρετητή (μέθοδος μεταφοράς «GET», π.χ. genClusterData.php?zoomDefined=0&latDefined=0.0 &lngDefined=0.0&yearDefined=2011&monthDefined=04).

3. Υπολογίζονται τα όρια της περιοχής ενδιαφέροντος του χρήστη και ανακτώνται τα συναφή με αυτή δεδομένα του spamming (αντικέιμενα) από τη βάση και συγκεκριμένα από τον πίνακα *spamcities* στη μορφή:

(*γεωγραφικό πλάτος, γεωγραφικό μήκος, αριθμός απειλών*)

4. Καθορίζεται ένα πλέγμα συγκεκριμένου μεγέθους (π.χ. 10 pixels στον χάρτη) που θα οδηγήσει στη δημιουργία σύνθετων αντικειμένων τα οποία θα επιταχύνουν τη διαδικασία της δεύτερης ανάθεσης αντικειμένων σε συστάδες που περιγράφεται στο βήμα 14 και μετά.

5. (Πρώτη ανάθεση) Το πρώτο αντικείμενο που έχει ανακτηθεί αποτελεί την πρώτη συστάδα και ανατίθεται επίσης στο αντίστοιχο με τις συντεταγμένες του κελί. Ενημερώνονται δε τα CF τόσο της συστάδας όσο και του κελιού.

6. Για κάθε επόμενο αντικείμενο γίνεται ανάθεσή του στο σχετικό κελί ενημερώνοντας το CF του καθώς και αναζητείται η πλησιέστερη σε αυτό βάσει του centroid της συστάδα.

7. Ελέγχεται για το αν η απόσταση του αντικειμένου από το centroid της συστάδας που επιλέχθηκε να ανατεθεί είναι μικρότερη από ένα ορισμένο κατώφλι $maxDist$ (π.χ. 40 pixels στον χάρτη), καθώς και αν η προκύπτουσα ακτίνα συστάδας συνεχίζει να είναι μικρότερη από ένα κατώφλι T (π.χ. $2 * maxDist/3$).

8. Αν ναι, τότε το αντικείμενο ανατίθεται στη συστάδα και ενημερώνεται το σχετικό CF της.

9. Αν όχι, μια νέα συστάδα δημιουργείται που περιλαμβάνει το συγκεκριμένο αντικείμενο.

10. Επαναλαμβάνεται το βήμα 6 για όλα τα αντικείμενα που έχουν ανακτηθεί.

11. («Επανασυσταδοποίηση») Κάθε συστάδα που δημιουργήθηκε ελέγχεται ως προς την απόσταση του centroid της με το centroid καθεμιάς από τις υπόλοιπες συστάδες που δεν έχουν συγχωνευθεί για το αν είναι μικρότερη από το κατώφλι $maxDist$.

12. Αν ναι, τότε γίνεται συγχώνευση των συστάδων σε μία ενημερώνοντας το σχετικό CF.

13. Επαναλαμβάνεται το βήμα 11 μέχρι να μην υπάρχουν άλλες συγχωνεύσεις.

14. (Δεύτερη ανάθεση) Για κάθε σύνθετο πλέον αντικείμενο των κελιών αναζητείται η πλησιέστερη σε αυτό συστάδα με βάση την απόσταση των centroid τους (όπως προκύπτουν από τα CF).

15.Το σύνθετο αντικείμενο ανατίθεται στη συστάδα χωρίς όμως να ενημερώνεται το CF της τελευταίας.

16.Επαναλαμβάνεται το βήμα 14 μέχρι να ανατεθούν όλα τα σύνθετα αντικείμενα.

17.Για κάθε συστάδα που δημιουργήθηκε υπολογίζεται το CF της από τα σύνθετα αντικείμενα που περιλαμβάνει.

18.(Δημιουργία XML αρχείου) Για κάθε συστάδα υπολογίζεται η μέγιστη απόσταση αντικειμένου από το centroid της, και μαζί με αυτή τα στοιχεία του CF της συστάδας (κανονικοποιημένα ως προς τον αριθμό των απειλών spam) ενσωματώνονται ως πεδία κάθε εγγραφής σε XML αρχείο, όπως για παράδειγμα:

```
<markers>
  <marker lat="46.9" lng="10.49" spams="89780" SS="16.1"
    maxDistance="48"/>
  <marker lat="38.9" lng="-119.03" spams="3929" SS="11.5"
    maxDistance="39"/>
  ... ... ...
</markers>
```

19.Αποστέλλεται το XML αρχείο με τα αποτελέσματα των συστάδων στον περιηγητή ιστού ως απάντηση.

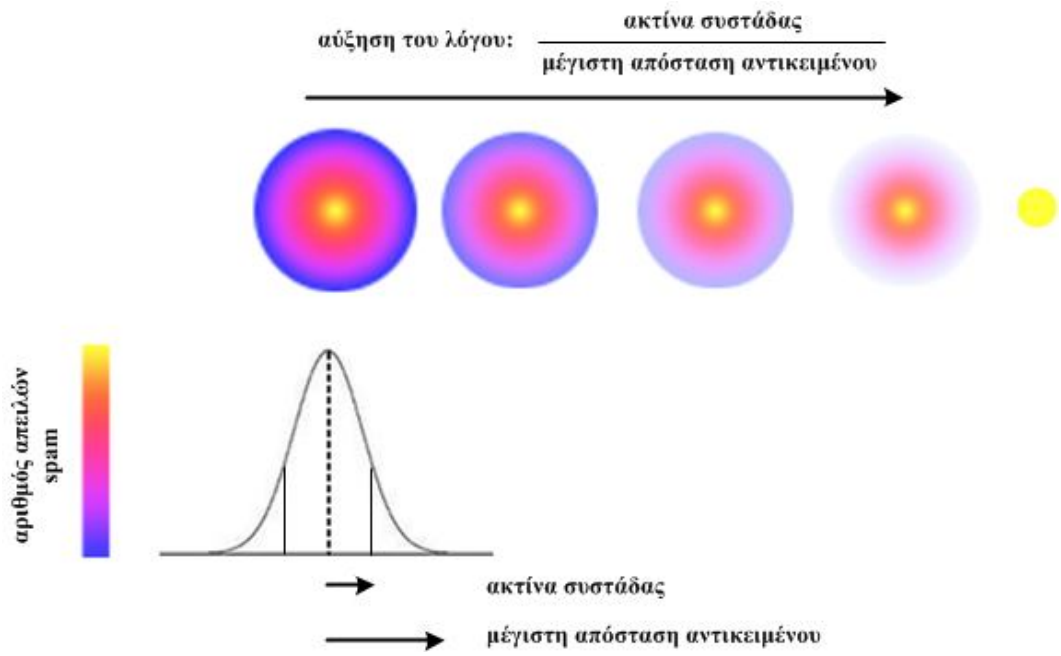
20.(Συσταδοποίηση από την πλευρά του περιηγητή ιστού) Ανάλογα με το ποιο ζουμ έχει επιλέξει ο χρήστης επιτελείται ή όχι η συσταδοποίηση (βλ. Ενότητα 5.3.1) των συστάδων που ανακτούνται από το XML αρχείο.

21.Αν το ζουμ δεν είναι κατάλληλο, τότε εκτελείται το βήμα 25.

22.Αν το ζουμ είναι κατάλληλο, κάθε συστάδα ελέγχεται ως προς την απόσταση του centroid της με το centroid καθεμιάς από τις υπόλοιπες συστάδες που δεν έχουν συγχωνευθεί για το αν είναι μικρότερη από το κατώφλι $maxDist$.

23.Αν ναι, τότε γίνεται συγχώνευση των συστάδων σε μία ενημερώνοντας το σχετικό CF.

24.Επαναλαμβάνεται το βήμα 22 μέχρι να μην υπάρχουν άλλες συγχωνεύσεις.



Εικόνα 5.5: Αναπαράσταση συστάδων απειλών spam

25.(Απεικόνιση) Για κάθε συστάδα υπολογίζεται η ακτίνα της, και γίνεται απεικόνισή της σε χάρτη στις συντεταγμένες του centroid της με χρήση κυκλικών κλιμακούμενα χρωματισμένων σχημάτων τα οποία προκύπτουν από τις χαρακτηριστικές ακτίνες: ακτίνα συστάδας και μέγιστη απόσταση κάποιου αντικειμένου από το centroid της. Συγκεκριμένα, υπολογίζεται το ποσοστό της ακτίνας συστάδας έναντι της μέγιστης απόστασης αντικειμένου για τον καθορισμό της διαφάνειας του σχήματος στις μεγαλύτερες αποστάσεις από το centroid, ενώ ο αριθμός των απειλών spam καθορίζει το χρώμα το οποίο θα έχει το κέντρο του σχήματος. Στην Εικόνα 5.5 παρουσιάζεται η συγκεκριμένη αναπαράσταση με κυκλικά σχήματα.

Λειτουργίες Δημιουργίας Εικόνων για Απεικόνιση των Δεδομένων με Εφέ Κίνησης Πάνω σε Χάρτη

Δεδομένου του γεγονότος ότι τόσο η προηγούμενη όσο και η συγκεκριμένη κατηγορία λειτουργιών περιλαμβάνουν τη χρήση του Google Maps API για την απεικόνιση των δεδομένων spamming σε χάρτη(ες), οι δύο κατηγορίες αυτές παρουσιάζουν αρκετές ομοιότητες ως προς τις υλοποιήσεις τους. Η βασικές διαφορές έγκεινται στον τρόπο απεικόνισης των ανακτώμενων δεδομένων και την μορφή με την οποία αυτά μεταφέρονται από τον εξυπηρετητή στον περιηγητή ιστού του χρήστη. Τα αρχεία που αποτελούν τη συγκεκριμένη κατηγορία είναι: *progress.php*, *animationApi.php* και *genAnimationData.php*.

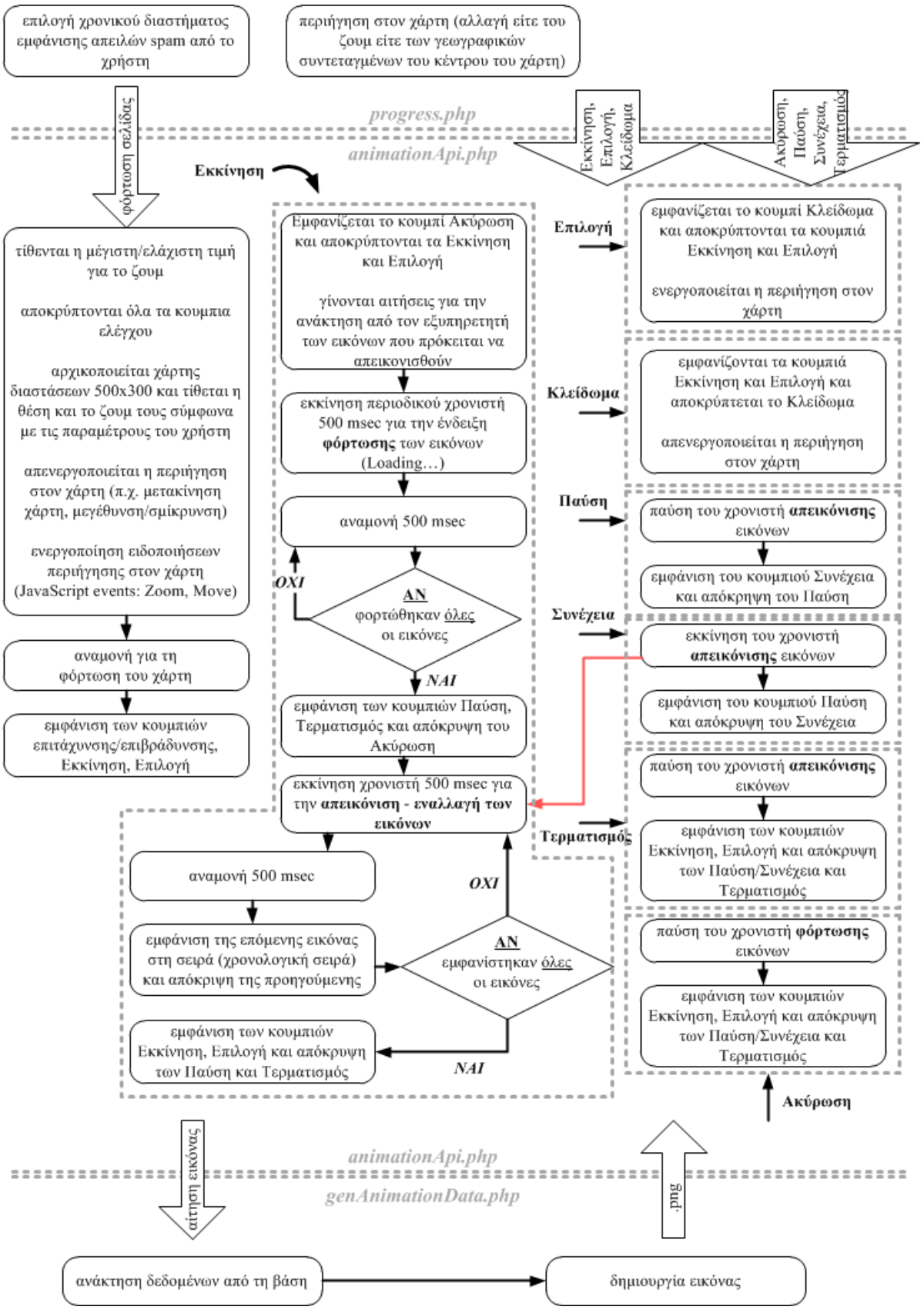
Αναφορικά με το ***progress.php***, πρόκειται για το αρχείο που επεκτείνει τη διεπαφή χρήστη στην περίπτωση του πλαισίου της ιστοσελίδας όπου προβάλλεται η απεικόνιση δεδομένων για τη συγκεκριμένη κατηγορία λειτουργιών. Οι χρονικές επιλογές οι οποίες προσδιορίζουν τα ανακτώμενα δεδομένα είναι οι ίδιες με εκείνες του αρχείου index.php, όπως επίσης και οι συνδέσεις με τα αρχεία JavaScript, μόνο που αυτή τη φορά χρησιμοποιείται το αρχείο animationApi.php αντί του mapSpamApi.php. Νέο στοιχείο αποτελεί η παρουσίαση στον χρήστη στοιχείων ελέγχου (κουμπιά εκκίνησης, τερματισμού, επιτάχυνσης κ.α.) του εφέ κίνησης των δεδομένων πάνω σε έναν μόνο χάρτη αυτή τη φορά, το οποίο στηρίζεται στην εναλλαγή εικόνων των δεδομένων για συνεχόμενες χρονικές περιόδους ίσες σε πλήθος με το αριθμό των χρονικών βαθμίδων που επιλέγονται. Όσον αφορά δε τις παραμέτρους που ορίζουν τόσο στιγμιότυπα του αρχείου αυτού όσο και του animationApi.php, αυτές είναι ακριβώς οι ίδιες με εκείνες της πρώτης κατηγορίας για τα αντίστοιχα αρχεία.

Σε αντίθεση με το genClusterData.php, το αρχείο ***genAnimationData.php***, το οποίο είναι και αυτό υπεύθυνο για την ανάκτηση δεδομένων από τη βάση του συστήματος, δεν προβαίνει σε καμιά συσταδοποίηση των δεδομένων, ενώ το αρχείο που παράγει αποτελεί μια εικόνα (.png) της κατανομής των απειλών spam ανά τον κόσμο ως προς την πυκνότητα της προέλευσής τους. Όσον αφορά τον τρόπο με τον οποίο επιτυγχάνεται αυτό, γίνεται μια σάρωση των δεδομένων spamming, και ανάλογα με τις συντεταγμένες τους τοποθετείται ένα κυκλικό σημείο με ορισμένη διαφάνεια στο πλαίσιο της εικόνας (500x300), αλλά και με μεταβλητό μέγεθος ανάλογα με το επίπεδο του ζουμ. Οι παράμετροι δε που καθορίζουν την επιλογή των δεδομένων προς απεικόνιση παραμένουν παρόλα αυτά οι ίδιες.

Ως το αρχείο που προσδίδει την κατεξοχήν λειτουργικότητα στο σύστημα αναφορικά με τη συγκεκριμένη κατηγορία λειτουργιών, το ***animationApi.php*** αναλαμβάνει την ανάκτηση των εικόνων των δεδομένων του αρχείου genAnimationData.php οι οποίες προκύπτουν από τις σχετικές χωρο-χρονικές παραμέτρους που επιβάλλει ο χρήστης, καθώς και τον χρονοπρογραμματισμό των εικόνων αυτών ώστε να επιτυγχάνεται η παρουσίαση της εξέλιξης του φαινομένου spamming στο χρόνο. Αντίθετα με την περίπτωση του mapSpamApi.php που χρησιμοποιεί το αντικείμενο GXmlHttp ώστε να ανακτήσει τα δεδομένα σε μορφή XML, το αρχείο animationApi.php χρησιμοποιεί τη δυνατότητα που προσφέρει το Google Maps API για τη δημιουργία προσαρμοσμένων στις ανάγκες των εφαρμογών αντικειμένων, ώστε να δημιουργηθεί ένα αντικείμενο που να μπορεί να φιλοξενήσει τις εικόνες που ανακτώνται και τελικά να τις παρουσιάσει πάνω από χάρτη. Συγκεκριμένα, χρησιμοποιείται το αντικείμενο GOverlay, το οποίο προσαρμόζεται ώστε να περιλαμβάνει μια δομή πλαισίου ( )

διαστάσεων ίδιων με εκείνων του χάρτη (500x300) στης οποίας το φόντο ανατίθεται η ανακτημένη κάθε φορά εικόνα του genAnimationData.php. Όσον αφορά δε τον χρονοπρογραμματισμό αυτών των εικόνων, βασικό στοιχείο αποτελεί όπως είναι η φυσικό η χρήση ενός χρονιστή (timer). Δεδομένου αυτού, για να ελέγξουν το εφέ κίνησης που δημιουργείται μέσω της εναλλαγής των εικόνων spamming, σχεδόν όλες οι επιλογές ελέγχου που παρουσιάζει το αρχείο progress.php ενεργούν μέσω των αντίστοιχων συναρτήσεων του animationApi.php πάνω σε αυτόν τον χρονιστή.

Δεδομένου του γεγονότος ότι οι λειτουργίες που επιτελούνται είναι άμεσα συνδεδεμένες με τη διεπαφή χρήστη, μια καλύτερη παρουσίαση αυτών επιτυγχάνεται με την Εικόνα 5.6 όπου περιγράφεται η συγκεκριμένη κατηγορία λειτουργιών, καθώς και με τις αντίστοιχες εικόνες της διεπαφής χρήστη στην Ενότητα 5.2.



Εικόνα 5.6: Λειτουργίες δημιουργίας εικόνων για απεικόνιση των δεδομένων με εφέ κίνησης πάνω σε χάρτη

Λειτουργίες Απεικόνισης Δεδομένων με Γραφικές Παραστάσεις και Ραβδογράμματα

Σε αντίθεση με τις δύο προηγούμενες, η συγκεκριμένη κατηγορία λειτουργιών δεν περιλαμβάνει τη χρήση μόνο του Google Maps API, αλλά και του Google Visualization API το οποίο μεταξύ άλλων προσφέρει δυνατότητες απεικόνισης στατιστικών στοιχείων με συγκεκριμένες παραστάσεις. Επίσης, λόγω σχετικών περιορισμών στον τρόπο με τον οποίο γίνονται οι απεικονίσεις όπως φαίνεται και από τις σχετικές εικόνες στην Ενότητα 5.2, καθώς και για λόγους πληρότητας των απεικονιζόμενων δεδομένων spamming, τα τελευταία ανακτούνται σε ομάδες ανά χώρα και όχι ανά πόλη, πράγμα το οποίο επωφελείται και από το γεγονός ότι στην συγκεκριμένη περίπτωση δεν απαιτείται η γνώση γεωγραφικών συντεταγμένων. Η τριάδα αρχείων που περιλαμβάνονται στην παρούσα κατηγορία λειτουργιών είναι: *graphPresentation.php*, *graphApi.php* και *genGraphData.php*.

Το αρχείο ***graphPresentation.php*** είναι εκείνο που καθορίζει πως θα κατανεμηθούν στην ιστοσελίδα οι σχετικές απεικονίσεις που αφορούν τη συγκεκριμένη κατηγορία λειτουργιών. Βέβαια, όπως και τα προηγουμένως αναφερθέντα συναφή αρχεία των δύο πρώτων κατηγοριών, παρουσιάζει τις χρονικές επιλογές με τις οποίες ο χρήστης προσδιορίζει τα ανακτώμενα δεδομένα, όπως επίσης και δημιουργεί τις συνδέσεις με τα αρχεία JavaScript που απαιτούνται, το graphApi.php και εκείνα που αντιστοιχούν στα Google Maps API και Google Visualization API. Ωστόσο, οι σχετικές παράμετροι που ορίζουν τόσο στιγμιότυπα του αρχείου αυτού όσο και του graphApi.php, δεν περιλαμβάνουν εκείνες του ζουμ και των γεωγραφικών συντεταγμένων μιας και δεν είναι απαραίτητες για τις απεικονίσεις που πραγματεύονται στη συγκεκριμένη περίπτωση.

Από την άλλη μεριά, το αρχείο ***genGraphData.php*** είναι υπεύθυνο για την ανάκτηση των δεδομένων spamming της βάσης με τη διαφορά σε σχέση με τα αντίστοιχα αρχεία των άλλων κατηγοριών ότι θα πρέπει να ομαδοποιηθούν ως προς τη χώρα προέλευσής τους. Η μορφή δε με την οποία τα δεδομένα αποστέλλονται πίσω στον περιηγητή ιστού είναι η XML, ενώ οι παράμετροι που τα προσδιορίζουν περιορίζονται σε εκείνες που καθορίζουν μια χρονική περίοδο.

Σχετικά με το αρχείο ***graphApi.php***, είναι εκείνο που προσδίδει την κυρίως λειτουργικότητα στο σύστημα σε συνδυασμό βέβαια και με το Google Visualization API. Το τελευταίο παρέχει τις κατάλληλες απεικονίσεις τις οποίες το graphApi.php φορτώνει στα σχετικά πλαίσια της

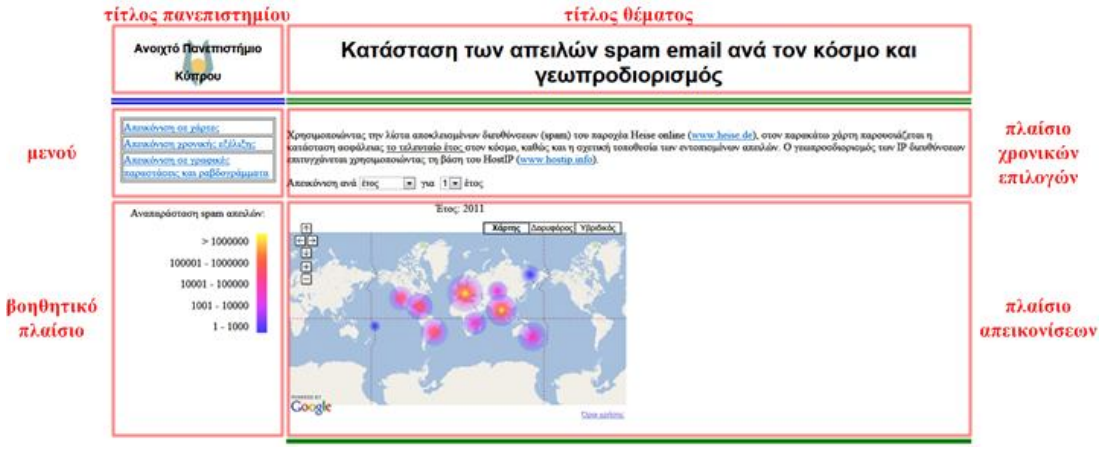
ιστοσελίδας που έχουν καθορισθεί από το graphPresentation.php αφού πρώτα έχουν ανακτηθεί και έχουν διαμορφωθεί στη συγκεκριμένη μορφή που απαιτείται από τις εν λόγω απεικονίσεις τα σχετικά δεδομένα spamming. Όπως και στην πρώτη κατηγορία λειτουργιών, έτσι και σε αυτή χρησιμοποιείται το αντικείμενο GXmlHttp ώστε να ανακτηθούν τα XML δεδομένα που επιστρέφει το αρχείο genGraphData.php και αφορούν τις χώρες ανά τον κόσμο για το χρονικό διάστημα που έχει προσδιορίσει ο χρήστης μέσω των αντίστοιχων παραμέτρων που καθορίζουν το παρόν κάθε φορά στιγμιότυπο του αρχείου. Επίσης, χρησιμοποιώντας τη σχετική διεπαφή που προσφέρει μια από τις απεικονίσεις αυτές (χάρτης με χρωματισμένες τις χώρες ανά τον κόσμο ανάλογα με τον αριθμό των απειλών spam που περιλαμβάνουν) επιτυγχάνεται διασύνδεση των απεικονίσεων, ώστε να προσαρμόζονται στις επιλογές του χρήστη όσον αφορά συγκεκριμένες χώρες ανά τον κόσμο.

Σχετικά με την υλοποίηση των λειτουργιών αυτών, όπως αναφέρθηκε ο τρόπος με τον οποίο ανακτούνται τα δεδομένα είναι ο ίδιος με εκείνον της πρώτης κατηγορίας λειτουργιών με τη διαφορά ότι δεν υπάρχει συσταδοποίηση, ενώ το υπόλοιπο της υλοποίησης αναφέρεται στην διαμόρφωση των σχετικών απεικονίσεων που προσφέρει η Google Inc. Συνεπώς, μια καλύτερη παρουσίαση του συνόλου των λειτουργιών γίνεται στις εικόνες της επόμενης ενότητας, δεδομένου ότι είναι στενά συνδεδεμένες με τη διεπαφή χρήστη.

5.2 Διεπαφή Χρήστη – Παρουσίαση Συστήματος

Μέχρι στιγμής έχει παρουσιασθεί η αρχιτεκτονική του υλοποιηθέντος συστήματος καθώς και οι λειτουργίες που το διέπουν. Σα συμπλήρωμα αυτών, στη συγκεκριμένη ενότητα παρουσιάζεται η διεπαφή χρήστη του συστήματος, δηλαδή ο τρόπος με τον οποίο τα αποτελέσματα των λειτουργιών εμφανίζονται μέσω του περιηγητή ιστού στο χρήστη, καθώς και ο τρόπος με τον οποίο ο τελευταίος αλληλεπιδρά με το σύστημα, ώστε να προκύψουν οι επιθυμητές απεικονίσεις που αφορούν τα δεδομένα για το φαινόμενο spamming.

Όπως περιγράφθηκε στην Ενότητα 5.1.2, οι λειτουργίες που αφορούν το σύστημα κατανέμονται σε τρεις κατηγορίες, οι οποίες αναφέρονται ουσιαστικά σε τρεις διαφορετικές εκδοχές απεικόνισης των δεδομένων του spamming. Καθεμιά δε από αυτές επιλέγεται μέσω της διεπαφής χρήστη, πράγμα το οποίο οδηγεί στη διάκριση επίσης τριών περιπτώσεων σελίδων τις οποίες περιλαμβάνει ο σχετικός ιστότοπος του συστήματος που επιλέγονται από αντίστοιχες ετικέτες στη δομή τους:



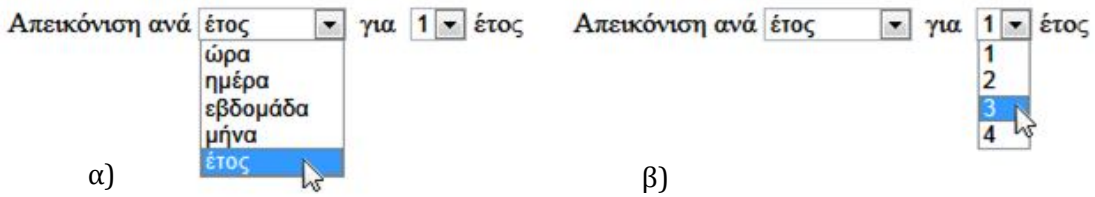
Εικόνα 5.7: Βασική δομή κάθε σελίδας του ιστότοπου του συστήματος

i. Απεικόνιση των απειλών spam σε πολλαπλούς χάρτες με παράμετρο το χρόνο. (Ετικέτα: *Απεικόνιση σε χάρτες*)

ii. Απεικόνιση σε χάρτη της χρονικής εξέλιξης της κατανομής των απειλών spam. (Ετικέτα: *Απεικόνιση χρονικής εξέλιξης*)

iii. Απεικόνιση των απειλών spam ως προς τον χρόνο για κάθε χώρα ανά τον κόσμο σε γραφικές παραστάσεις και ραβδογράμματα. (Ετικέτα: *Απεικόνιση σε γραφικές παραστάσεις και ραβδογράμματα*)

Στην Εικόνα 5.7 παρουσιάζεται η βασική δομή κάθε σελίδας του ιστότοπου. Συγκεκριμένα, διακρίνονται οι περιοχές του τίτλου του θέματος του ιστότοπου, του τίτλου του Ανοιχτού Πανεπιστημίου Κύπρου, του μενού επιλογών σελίδας, του βοηθητικού πλαισίου όπου φιλοξενούνται επεξηγήσεις των απεικονίσεων (λεζάντα), του πλαισίου χρονικών επιλογών, και τέλος του πλαισίου απεικονίσεων. Το τελευταίο είναι το σημαντικότερο τμήμα της δομής των σελίδων, γι' αυτό και είναι κατά κύριο λόγο το μόνο που διαφοροποιείται μεταξύ τους και που θα μπορούσε να συμπεριληφθεί σαν ξεχωριστό κομμάτι σε οποιονδήποτε άλλον ιστότοπο παρέχοντας τις λειτουργίες του υλοποιηθέντος συστήματος. Επίσης σημαντικό δε για το σύστημα είναι το πλαίσιο χρονικών επιλογών, όπως φαίνεται στην Εικόνα 5.8, το οποίο όμως είναι σχεδόν το ίδιο για κάθε σελίδα. Μέσω αυτού δίνεται η δυνατότητα στο χρήστη επιλογής της χρονικής βαθμίδας (ώρα, ημέρα, εβδομάδα, μήνας, έτος) υπό την οποία εξετάζονται χρονικά τα δεδομένα του spamming, καθώς και το πλήθος αυτών, το οποίο αντιστοιχεί στην απεικόνιση σε ισάριθμους χάρτες (μέχρι 4) ή σε ισάριθμες εναλλαγές εικόνων (μέχρι 10) ή σε αντίστοιχου μεγέθους γραφικές παραστάσεις και ραβδογράμματα (μέχρι 10). Οι χρονικές περίοδοι που

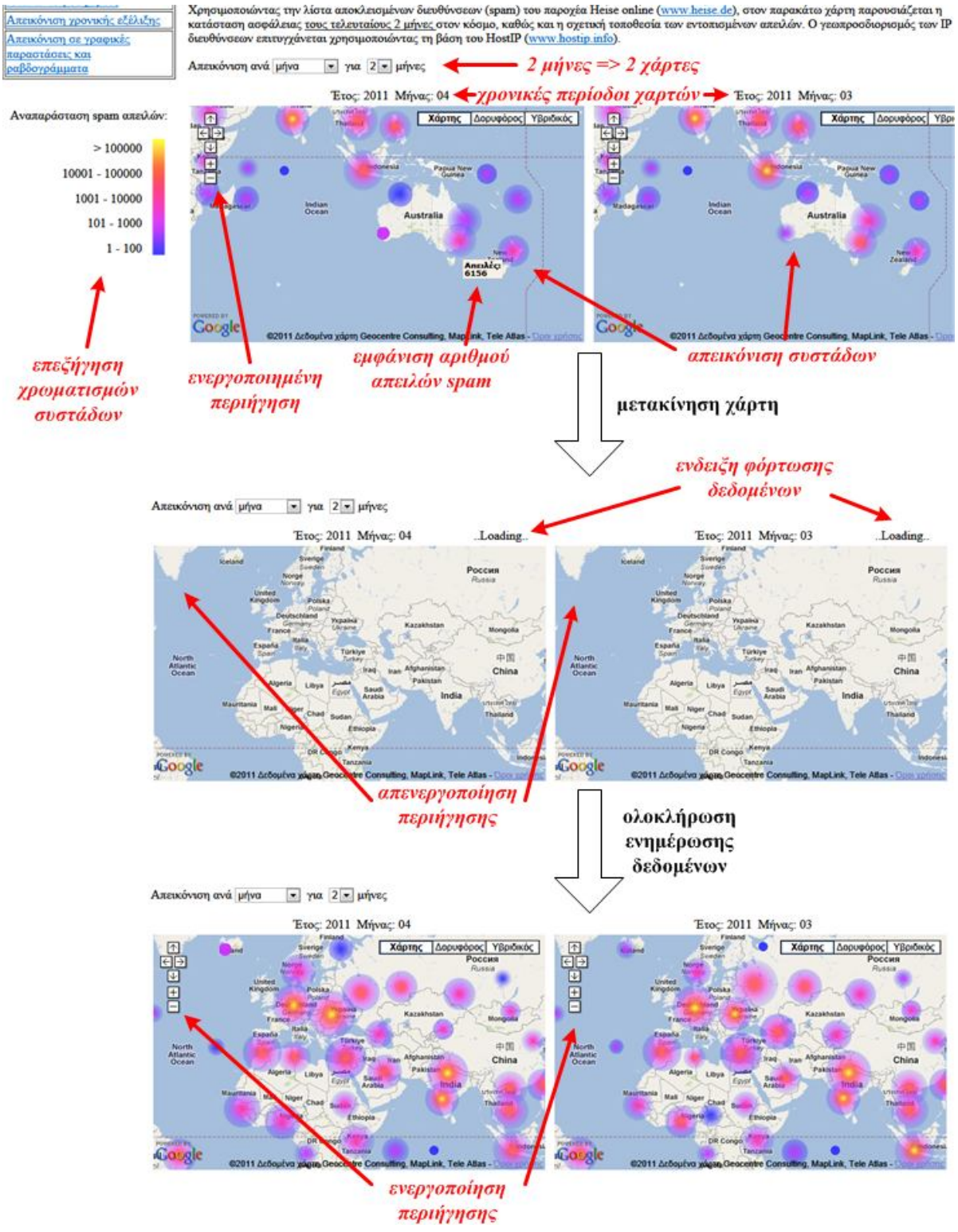


Εικόνα 5.8: Πλαίσιο χρονικών επιλογών του ιστότοπου του συστήματος: α) επιλογή χρονικής βαθμίδας, β) επιλογή πλήθους χρονικών βαθμίδων

αντιπροσωπεύονται με αυτές τις επιλογές έχουν ως σημείο αναφοράς την χρονική στιγμή (με ακρίβεια ώρας και ζώνη ώρας GMT0) που ο χρήστης έκανε τις συγκεκριμένες επιλογές. Για παράδειγμα, αν τη χρονική στιγμή 7/7/2011 12:34:45 GMT0 έγιναν οι επιλογές: χρονική βαθμίδα = ώρα, πλήθος χρονικών βαθμίδων = 4, τότε η χρονικές περίοδοι που μελετώνται είναι διάρκειας μιας ώρας και αφορούν τις τελευταίες 4 ώρες στον κόσμο, δηλαδή οι χρονικές περίοδοι είναι: α) 7/7/2011 11:00:00 - 7/7/2011 11:59:59, β) 7/7/2011 10:00:00 - 7/7/2011 10:59:59, γ) 7/7/2011 09:00:00 - 7/7/2011 09:59:59, και δ) 7/7/2011 08:00:00 - 7/7/2011 08:59:59

Απεικόνιση των Απειλών Spam σε Πολλαπλούς Χάρτες με Παράμετρο το Χρόνο

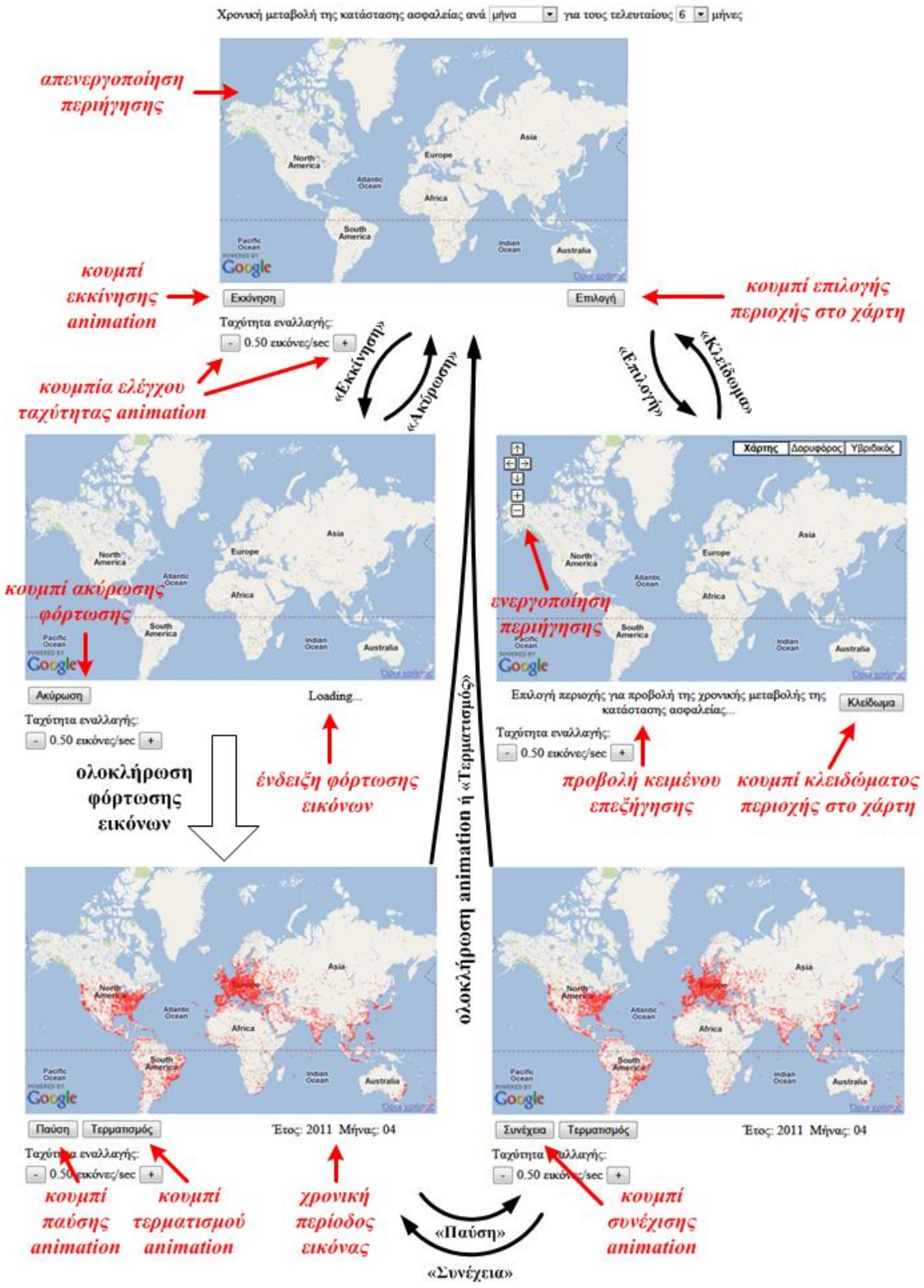
Στο πλαίσιο απεικόνισης της συγκεκριμένης περίπτωσης σελίδας του ιστότοπου παραθέτονται τόσοι χάρτες όσους υποδεικνύει η επιλογή του χρήστη για το πλήθος των χρονικών βαθμίδων. Έτσι, κάθε χάρτης απεικονίζει δεδομένα του spamming που αφορούν συνεχόμενες χρονικές περιόδους, με καθεμιά να αντιστοιχεί σε μια χρονική βαθμίδα ώστε να μπορεί ο χρήστης να συγκρίνει εύκολα τη χρονική εξέλιξη της κατάστασης του φαινομένου σε μια περιοχή. Όσον αφορά την απεικόνιση των δεδομένων, αυτή γίνεται μέσω συστάδων αυτών, αναπαριστώμενες με κυκλικά σχήματα όπως παρουσιάσθηκε στην Εικόνα 5.5, οι οποίες μεταβάλλονται ανάλογα με την περιοχή που απεικονίζεται στους χάρτες και το ζουμ της. Σχετική επεξήγηση των χρωματισμών των συστάδων όσον αφορά το πλήθος των απειλών spam που εκφράζουν παρουσιάζεται στο βοηθητικό πλαίσιο, ενώ επιπλέον το πλήθος των απειλών spam εμφανίζεται σε περίπτωση υπέρθεσης του δείκτη του ποντικιού πάνω από κάποια συστάδα. Πάνω δε από κάθε χάρτη υπάρχει μια ένδειξη που δηλώνει σε ποια χρονική περίοδο αναφέρεται, ενώ εμφανίζεται επίσης ένα κινούμενο κείμενο «…Loading…» στην περίπτωση που ενημερώνεται με νέα δεδομένα απενεργοποιώντας την ίδια στιγμή τη δυνατότητα περιήγησης του χρήστη. Εκτός από τις χρονικές επιλογές, λοιπόν, ο χρήστης έχει τη δυνατότητα περιήγησης στους χάρτες για τη μετάβαση σε κάποια περιοχή ενδιαφέροντος, επιλογές οι οποίες προκαλούν την ενημέρωση των δεδομένων των χαρτών. Στην Εικόνα 5.9 παρουσιάζεται η δομή της συγκεκριμένης σελίδας, καθώς και μια περίπτωση ενημέρωσης των δεδομένων.



Εικόνα 5.9: Δομή της σελίδας του ιστότοπου του συστήματος που αφορά την απεικόνιση απειλών spam σε πολλαπλούς χάρτες με παράμετρο το χρόνο

Απεικόνιση σε Χάρτη της Χρονικής Εξέλιξης της Κατανομής των Απειλών Spam

Αναφορικά με τη συγκεκριμένη απεικόνιση, σκοπός είναι να παρουσιασθεί με μορφή εναλλαγής εικόνων η χρονική εξέλιξη του φαινομένου spamming ως προς την κατανομή των πηγών προέλευσης των σχετικών απειλών (κόκκινα ημι-διαφανή κυκλικά σημεία) πάνω σε χάρτη. Έτσι, αυτή τη φορά το πλήθος των χρονικών βαθμίδων ουσιαστικά καθορίζει τον αριθμό των εικόνων που εναλλάσσονται, με κάθε εικόνα να περιλαμβάνει δεδομένα spamming για συνεχόμενες περιόδους ίσες με την επιλεγμένη χρονική βαθμίδα. Ο χρήστης έχει τη δυνατότητα να επιλέξει μέσω του κουμπιού «Επιλογή» οποιαδήποτε περιοχή τον ενδιαφέρει την οποία στη συνέχεια «κλειδώνει» με το κουμπί «Κλείδωμα», ώστε να εκκινήσει την εναλλαγή εικόνων (animation). Αυτό επιτυγχάνεται με το πάτημα του κουμπιού «Εκκίνηση». Πριν όμως γίνει αυτό εφικτό, οι εικόνες φορτώνονται από τον εξυπηρετητή στον περιηγητή ιστού, διαδικασία κατά τη διάρκεια της οποίας μια ένδειξη φόρτωσης «…Loading…» εμφανίζεται. Αν ο χρήστης αλλάξει γνώμη κατά τη διάρκεια της φόρτωσης και θέλει να κάνει νέα επιλογή περιοχής, τότε πατάει το κουμπί «Ακύρωση». Κατά τη διάρκεια δε της εναλλαγής των εικόνων, του δίνεται η δυνατότητα επιτάχυνσης/επιβράδυνσής της μέσω σχετικών κουμπιών (ταχύτητες εναλλαγής: 0,20 εικόνες/sec → 2,00 εικόνες/sec), ενώ μπορεί ακόμα είτε να τη σταματήσει προσωρινά ώστε να δει καλύτερα την κατανομή των πηγών προέλευσης των απειλών spam για μια ορισμένη χρονική περίοδο (και έπειτα να την συνεχίσει πατώντας το κουμπί «Συνέχεια») είτε να την τερματίσει με σκοπό την απεικόνιση νέων δεδομένων. Κάτω από τον χάρτη εμφανίζεται πάντα η χρονική περίοδος που αφορά την τρέχουσα κάθε φορά εικόνα που παρουσιάζεται. Στην Εικόνα 5.10 παρουσιάζονται οι παραπάνω δυνατότητες ελέγχου ως προς την απεικόνιση των δεδομένων spamming που παρέχει στο χρήστη η συγκεκριμένη σελίδα του ιστότοπου.



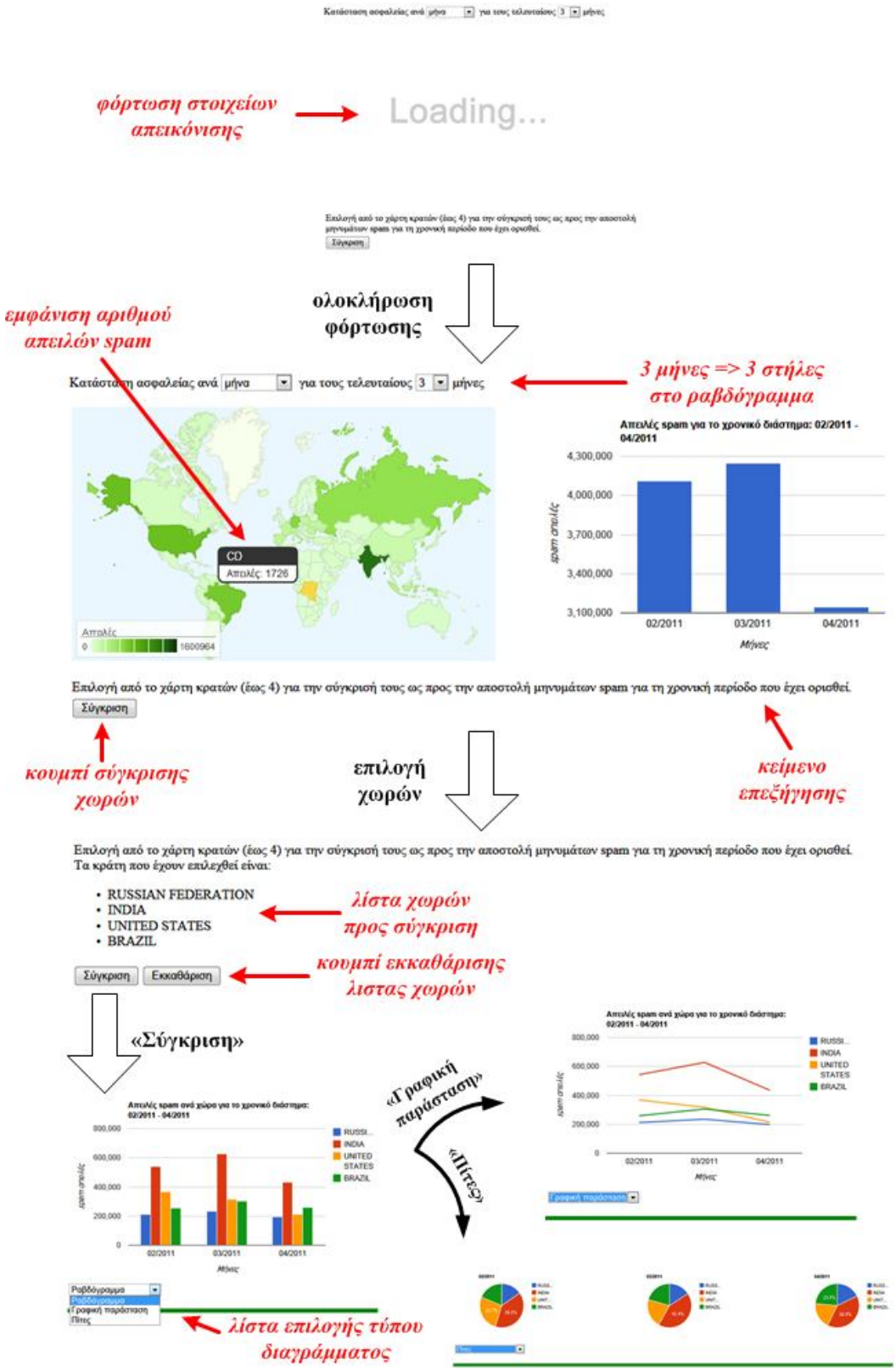
Εικόνα 5.10: Δομή της σελίδας του ιστότοπου του συστήματος που αφορά την απεικόνιση σε χάρτη της χρονικής εξέλιξης της κατανομής των απειλών spam

Απεικόνιση των Απειλών Spam ως προς τον Χρόνο για κάθε Χώρα ανά τον Κόσμο σε Γραφικές Παραστάσεις και Ραβδογράμματα

Τέλος, η παρούσα απεικόνιση έρχεται να συμπληρώσει τις άλλες δύο αποσκοπώντας στην παρουσίαση του αριθμού των απειλών spam ανά χώρα και χρονική περίοδο μέσω κατάλληλων γραφικών παραστάσεων και ραβδογραμμάτων. Πιο συγκεκριμένα, το πλαίσιο απεικόνισης της σελίδας αποτελείται από τα εξής στοιχεία: α) παγκόσμιος χάρτης κλιμακούμενα χρωματισμένων χωρών ανάλογα με τον αριθμό των απειλών spam που τις χαρακτηρίζει για το επιλεγμένο συνολικό χρονικό διάστημα από το χρήστη, β) ραβδόγραμμα του συνολικού αριθμού απειλών όλων των χωρών του κόσμου ανά χρονική περίοδο ίση με την χρονική βαθμίδα που επιλέγει ο χρήστης, και γ) σύγκριση (έως και 4) χωρών που επιλέγει ο χρήστης ως προς τον αριθμό των απειλών spam και ανά χρονική περίοδο ίση με μια χρονική βαθμίδα μέσω σχετικού ραβδογράμματος ή γραφικής παράστασης ή διαγραμμάτων πιτών.

Όσον αφορά το πρώτο στοιχείο, πρόκειται για έναν χάρτη όπου οι χώρες είναι χρωματισμένες έτσι ώστε με σκούρο χρώμα να αναπαρίστανται εκείνες που τις χαρακτηρίζει σχετικά μεγάλος αριθμός απειλών spam, ενώ με ανοιχτό χρώμα εκείνες που τις χαρακτηρίζει σχετικά μικρός αριθμός απειλών spam. Εκτός βέβαια της συγκεκριμένης απεικόνισης που παρέχει ο χάρτης, χρησιμοποιείται και ως διεπαφή με την οποία ο χρήστης επιλέγει τις χώρες που επιθυμεί να συγκριθούν, μέσω απεικόνισης των σχετικών αριθμών απειλών spam στα διαγράμματα που αποτελούν την τρίτη κατηγορία στοιχείων απεικόνισης της σελίδας. Συγκεκριμένα, με απλό «κλικ» πάνω σε κάποια χρωματισμένη χώρα του χάρτη, αυτή προστίθεται σε μια λίστα με τις χώρες που πρόκειται να συγκριθούν. Αν επιλεχθεί δε ξανά με αυτόν τον τρόπο η συγκεκριμένη χώρα, τότε αφαιρείται από τη λίστα. Μόλις επιλεχθούν οι χώρες προς απεικόνιση, ο χρήστης με το κουμπί «Σύγκριση» εμφανίζει σε ραβδόγραμμα τους αριθμούς των απειλών spam ανά χώρα και ανά χρονική περίοδο. Μέσω κατάλληλης επιλογής από σχετική λίστα η απεικόνιση της σύγκρισης αυτής μπορεί να γίνει επίσης είτε με γραφική παράσταση είτε με πολλαπλά διαγράμματα πιτών με το καθένα να αφορά μια ξεχωριστή χρονική περίοδο. Για διευκόλυνση δε του χρήστη ως προς την επιλογή κυρίως μικρών χωρών προς σύγκριση, επιτρέπεται με διπλό «κλικ» η μεγέθυνση στην περιοχή γύρο από μια χώρα. Από την άλλη μεριά, αναφορικά με την αφαίρεση χωρών από τη λίστα σύγκρισης, υπάρχει το κουμπί «Εκκαθάριση» με το οποίο αφαιρούνται όλες οι χώρες από τη λίστα. Τέλος, αξίζει να σημειωθεί ότι με την τοποθέτηση του δείκτη του ποντικιού πάνω από κάποιο στοιχείο οποιασδήποτε από τις απεικονίσεις (π.χ. χώρα του χάρτη, στήλη ραβδογράμματος, κομμάτι διαγράμματος πίτας) εμφανίζεται ο αντίστοιχος

αριθμός απειλών spam. Στην Εικόνα 5.11 παρουσιάζονται οι παραπάνω δυνατότητες ελέγχου της διεπαφής χρήστη για τις συγκεκριμένες απεικονίσεις στο σχετικό πλαίσιο της σελίδας.



Εικόνα 5.11: Δομή της σελίδας του ιστότοπου του συστήματος που αφορά την απεικόνιση σε γραφικές παραστάσεις και ραβδογράμματα της χρονικής εξέλιξης των απειλών spam

5.3 Ανάλυση Συστήματος – Χαρακτηριστικά Απόδοσης

Στη συγκεκριμένη ενότητα παρουσιάζεται μια βασική ανάλυση του συστήματος όσον αφορά την απόδοσή του στην απεικόνιση των δεδομένων που πραγματεύεται και στην ανάκτησή τους από τη βάση που διατηρεί. Δεδομένου ότι οποιαδήποτε υποβάθμιση της απόδοσης του συστήματος λόγω καθυστερήσεων στην μεταφορά των απαραίτητων κάθε φορά αρχείων (π.χ. αρχεία JavaScript, εικόνες κλπ.) δεν οφείλεται στο ίδιο το σύστημα, αλλά σε εξωγενείς παράγοντες που αφορούν το Διαδίκτυο, η απόδοση του συστήματος μελετάται ως προς λειτουργίες που διενεργούνται τοπικά στον εξυπηρετητή ή στον περιηγητή ιστού.

5.3.1 Απεικόνιση Δεδομένων

Αναφορικά με την απεικόνιση των δεδομένων του spamming, ουσιαστικά, οι λειτουργίες που θα μπορούσαν να μελετηθούν ως προς την απόδοση που προσφέρουν είναι αυτές που σχετίζονται με την πρώτη κατηγορία λειτουργιών από εκείνες που περιγράφηκαν στην Ενότητα 5.1.2, μιας και υπάρχει έντονο το στοιχείο της δυναμικής ενημέρωσης των δεδομένων στους χάρτες κατά την περιήγηση του χρήστη σε αυτούς. Από την άλλη μεριά, οι χρονικές επιλογές του χρήστη, οι οποίες αποτελούν βασικό τρόπο προσδιορισμού των δεδομένων για όλες τις κατηγορίες των λειτουργιών και απεικονίσεων του συστήματος, περιλαμβάνουν ενημερώσεις δεδομένων που δεν είναι κρίσιμες αρκετά ως προς των χρόνο ολοκλήρωσής τους, δεδομένου ότι κάθε φορά που γίνονται αυτές ο χρήστης αποσκοπεί στην περαιτέρω ανάλυση των δεδομένων μέσω είτε της περιήγησης σε χάρτες είτε της παραγωγής διαγραμμάτων (ανά χώρα).

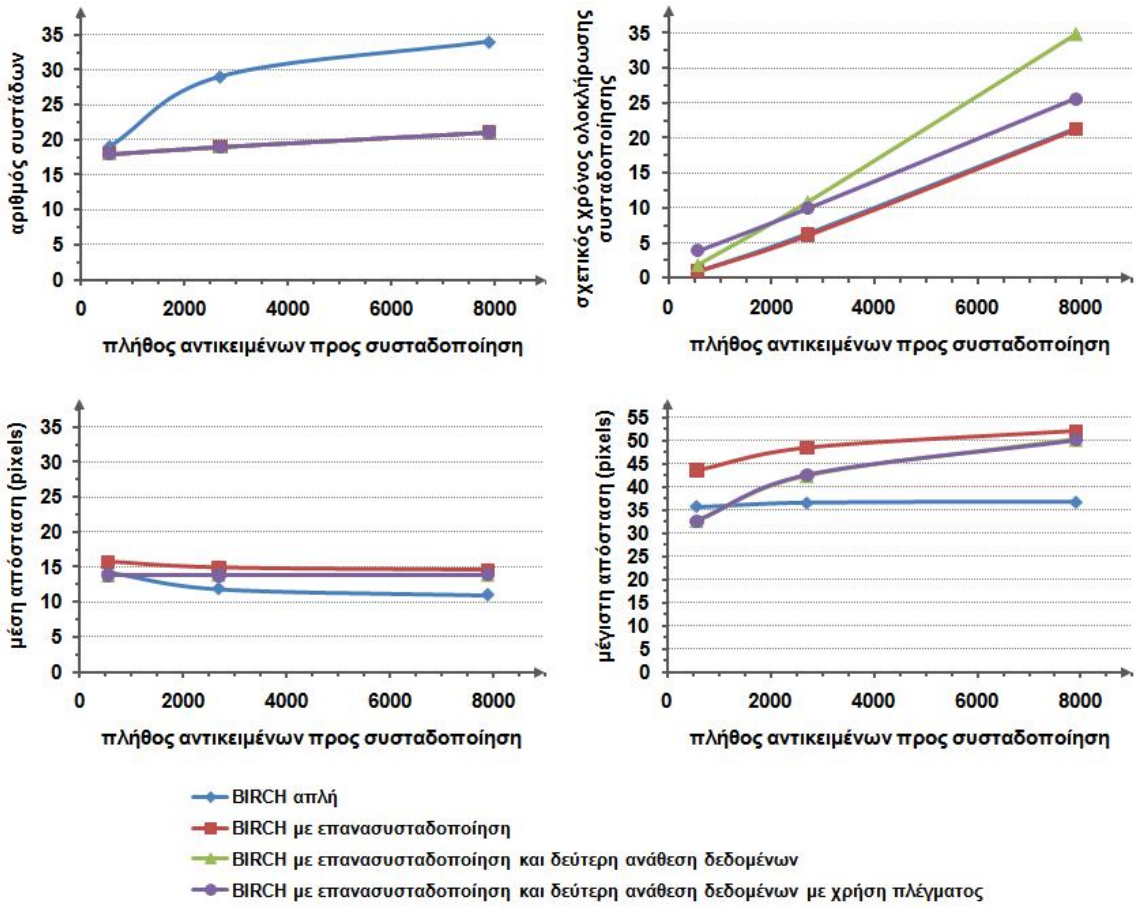
Σημαντική παρατήρηση (και συνάμα προϋπόθεση) αποτελεί το γεγονός ότι ο χρήστης, κατά την περιήγησή του στους χάρτες, αρκείται συνήθως σε μικρές μετακινήσεις του κέντρου τους στο πλαίσιο της σελίδας που φιλοξενείται ο καθένας, ώστε να παρατηρήσει το σχετικό φαινόμενο που απεικονίζεται σε γειτονικές περιοχές γύρω από μια συγκεκριμένη που τον ενδιαφέρει. Αυτή η περιήγηση πολλές φορές συνοδεύεται και από μια παλινδρομική μεγέθυνση/σμύκρυνση των χαρτών, η οποία αντιστοιχεί σε μεταβολή του ζουμ μεταξύ συνήθως δύο επιπέδων, ώστε να διευκολυνθεί ο χρήστης στην αναζήτηση των γειτονικών περιοχών που τον ενδιαφέρουν. Αυτό θα μπορούσε να ειπωθεί ότι αποτελεί ένα «κβάντο» περιήγησης-δραστηριότητας του χρήστη το οποίο θα πρέπει να διακόπτεται με την ενημέρωση των δεδομένων όσο το δυνατόν λιγότερο. Προς αυτήν την κατεύθυνση κινείται και το παρόν σύστημα χρησιμοποιώντας την διαδικασία

της προ-ανάκτησης δεδομένων που αφορούν μεγαλύτερη περιοχή από ότι ζητάει ο χρήστης μέσω της περιήγησής του.

Πιο συγκεκριμένα, σε περίπτωση αίτησης για ενημέρωση των δεδομένων σε μια περιοχή του χάρτη τα όρια (συντεταγμένες (x_{11}, y_{11}) αριστερού κάτω και (x_{12}, y_{12}) δεξιού πάνω σημείου) της οποίας καθορίζονται από το σχετικό πλαίσιο διαστάσεων 500x300 pixels που τον φιλοξενεί και για συγκεκριμένο ζουμ z_1, το καλούμενο αρχείο genClusterData.php ανακτά δεδομένα που αφορούν περιοχή διπλάσια ή και τετραπλάσια εκείνης που αιτείται. Αυτή δε η διαφοροποίηση στο μέγεθος της περιοχής για την οποία τα δεδομένα ανακτούνται οφείλεται στο γεγονός ότι η ανάκτηση δεδομένων έχει ορισθεί να πραγματοποιείται σε συγκεκριμένα ζουμ ώστε να γίνεται εκμετάλλευση της ιδιότητας της «λανθάνουσας αποθήκευσης» - *cashing* που παρέχει ο περιηγητής ιστού. Έτσι, γενικός κανόνας, τον οποίο υπαγορεύει και η πελάτη-εξυπηρετητή δομή της διαδικασίας συσταδοποίησης, αποτελεί η ανάκτηση συσταδοποιημένων δεδομένων από τον εξυπηρετητή σε ένα-παρά-ένα ζουμ από αυτά που είναι ενεργοποιημένα για την περιήγηση, ενώ σε κάθε ενδιάμεσο από αυτά διενεργείται συσταδοποίηση σε τοπικό επίπεδο (περιηγητής ιστού) στα δεδομένα-συστάδες που έχουν ανακτηθεί προηγουμένως από τον εξυπηρετητή και αφορούν μεγαλύτερο ζουμ.

Όσον αφορά τη μέθοδο συσταδοποίησης που εφαρμόζεται στα δεδομένα με σκοπό την καλύτερη απεικόνισή τους στους χάρτες, όπως περιγράφθηκε και στην Ενότητα 5.1.2, στηρίζεται κατά κύριο λόγο στα βήματα και τις σχετικές φάσεις που περιλαμβάνει η μέθοδος BIRCH. Στην Ενότητα 4.7, τα σχετικά αποτελέσματα έδειξαν ότι πρόκειται για μια πολύ γρήγορη μέθοδο συσταδοποίησης πράγμα το οποίο βασίζεται στη διαδικασία μιας μόνο (για την 1η φάση) σάρωσης των δεδομένων που διενεργείται. Οι υπόλοιπες φάσεις, οι οποίες και επιφέρουν μια επιπλέον καθυστέρηση στην ολοκλήρωση της συσταδοποίησης, αποσκοπούν κυρίως στη βελτίωση των διαμορφούμενων συστάδων, ώστε να μπορεί να προσεγγισθεί η απόδοση στο συγκεκριμένο τομέα που επιτυγχάνουν ορισμένες από τις υπόλοιπες μεθόδους.

Στην Εικόνα 5.12 παρουσιάζονται τα αποτελέσματα της σύγκρισης διαφόρων εκδοχών της BIRCH μεθόδου, οι οποίες υλοποιήθηκαν σε PHP και ενσωματώθηκαν στο σύστημα για τον σκοπό αυτό. Όπως φαίνεται η απλή εκδοχή της BIRCH που περιλαμβάνει μόνο την 1η φάση συσταδοποίησης δημιουργεί αυξημένο αριθμό συστάδων, πράγμα το οποίο δυσχεραίνει την απεικόνισή τους πάνω σε χάρτη. Η επανασυσταδοποίηση έρχεται να λύσει το πρόβλημα αυτό, χειροτερεύοντας όμως την αντιπροσωπευτικότητα των συστάδων στην περίπτωση που χρησιμοποιηθεί αποκλειστικά και μόνο αυτή σαν επιπλέον μηχανισμός. Η ανάθεση για δεύτερη

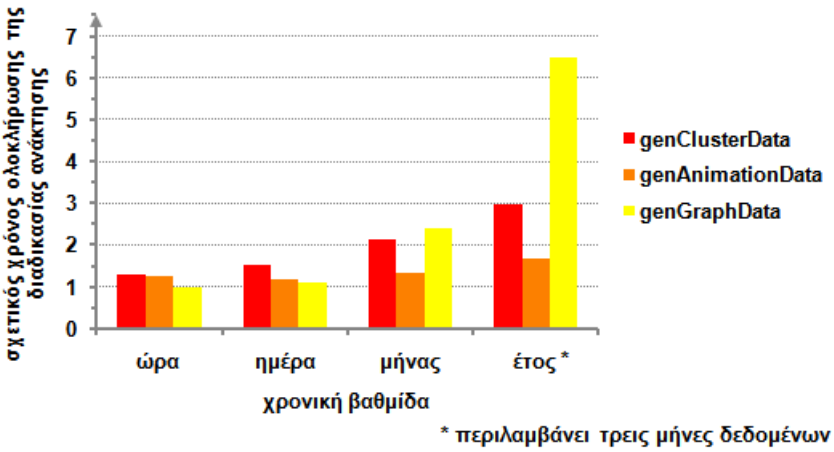


Εικόνα 5.12: Αριθμός συστάδων, σχετικός χρόνος ολοκλήρωσης συσταδοποίησης, μέση απόσταση αντικειμένων από το centroid των συστάδων τους, και μέση μέγιστη απόσταση αντικειμένου ανά συστάδα για διαφορετικό πλήθος αντικειμένων των δεδομένων για διάφορες εκδοχές της BIRCH που υλοποιούνται για το υλοποιηθέν σύστημα

φορά των αντικειμένων των δεδομένων στις δημιουργηθείσες συστάδες βελτιώνει τις νέες συστάδες ως προς αυτό το χαρακτηριστικό τους, επιβάλλοντας όμως μια καθυστέρηση στην ολοκλήρωση της συσταδοποίησης λόγω της δεύτερης σάρωσης των δεδομένων. Τέλος, υιοθετώντας τη χρήση πλέγματος (μεγέθους 10 pixels) για τη δημιουργία σύνθετων αντικειμένων κατά την πρώτη σάρωση των δεδομένων, δίνεται η δυνατότητα επιτάχυνσης της δεύτερης σάρωσης διατηρώντας παράλληλα σταθερά τις υπόλοιπες μετρικές απόδοσης, επιβεβαιώνοντας το λόγο της επιλογής της συγκεκριμένης εκδοχής για τη διαδικασία της συσταδοποίησης που χρησιμοποιεί το παρόν σύστημα.

5.3.2 Ανάκτηση Δεδομένων

Τα αρχεία που πραγματεύονται την ανάκτηση των δεδομένων του spamming από τη βάση του συστήματος είναι τα genClusterData.php, genAnimationData.php και genGraphData.php. Καθένα



Εικόνα 5.13: Σύγκριση χρόνων ολοκλήρωσης της διαδικασίας ανάκτησης δεδομένων που πραγματοποιεί το υλοποιηθέν σύστημα μέσω των PHP αρχείων του

από αυτά αφορά την επιστροφή στον περιηγητή ιστού των δεδομένων με συγκεκριμένη μορφή, εξυπηρετώντας έτσι τις ανάγκες απεικόνισης για τις οποίες χρησιμοποιούνται. Δεδομένης λοιπόν της διαφοροποίησης των απεικονίσεων, οι λειτουργίες που συνοδεύουν κάθε ανάκτηση είναι επίσης διαφορετικές, με βασικό κοινό σημείο την εξάρτηση που υπάρχει από τις χρονικές επιλογές του χρήστη, οι οποίες αποστέλλονται με την κλήση των αρχείων αυτών (μέθοδος «GET») ως παράμετροι. Θα μπορούσε να θεωρηθεί ότι η κλήση των αρχείων συνοδευόμενες από τις συγκεκριμένες παραμέτρους είναι σαν να καλείται μια συνάρτηση με συγκεκριμένα ορίσματα και συγκεκριμένες εξόδους, αυτές των επιστρεφόμενων μορφών δεδομένων. Έτσι, το σύνολο των λειτουργιών των αρχείων PHP αυτών μπορεί να ειπωθεί ότι αποτελούν μέρος γενικότερα της διαδικασίας ανάκτησης δεδομένων την οποία καλεί το αντίστοιχο αρχείο JavaScript που βρίσκεται στον περιηγητή ιστού. Με αυτή τη γενικότερη πλέον έννοια της ανάκτησης δεδομένων, στην Εικόνα 5.13 παρουσιάζεται μια σύγκριση των χρόνων επιστροφής των κατάλληλων κάθε φορά δεδομένων από την εκτέλεση του κώδικα κάθε αρχείου PHP στον εξυπηρετητή, από εκείνα που αναφέρθηκαν παραπάνω, ως προς διαφορετικές χρονικές βαθμίδες που μπορεί να επιλέξει ο χρήστης, οι οποίες και αντιστοιχούν σε διαφορετικό πλήθος αντικειμένων των δεδομένων Για τις περιπτώσεις των αρχείων που δέχονται παραμέτρους χώρου (π.χ. ζουμ, γεωγραφικές συντεταγμένες), αυτές επιλέγονται έτσι ώστε τα δεδομένα που ανακτούνται να αφορούν όλον τον κόσμο (δηλ. να μην υπάρχουν χωρικοί περιορισμοί).

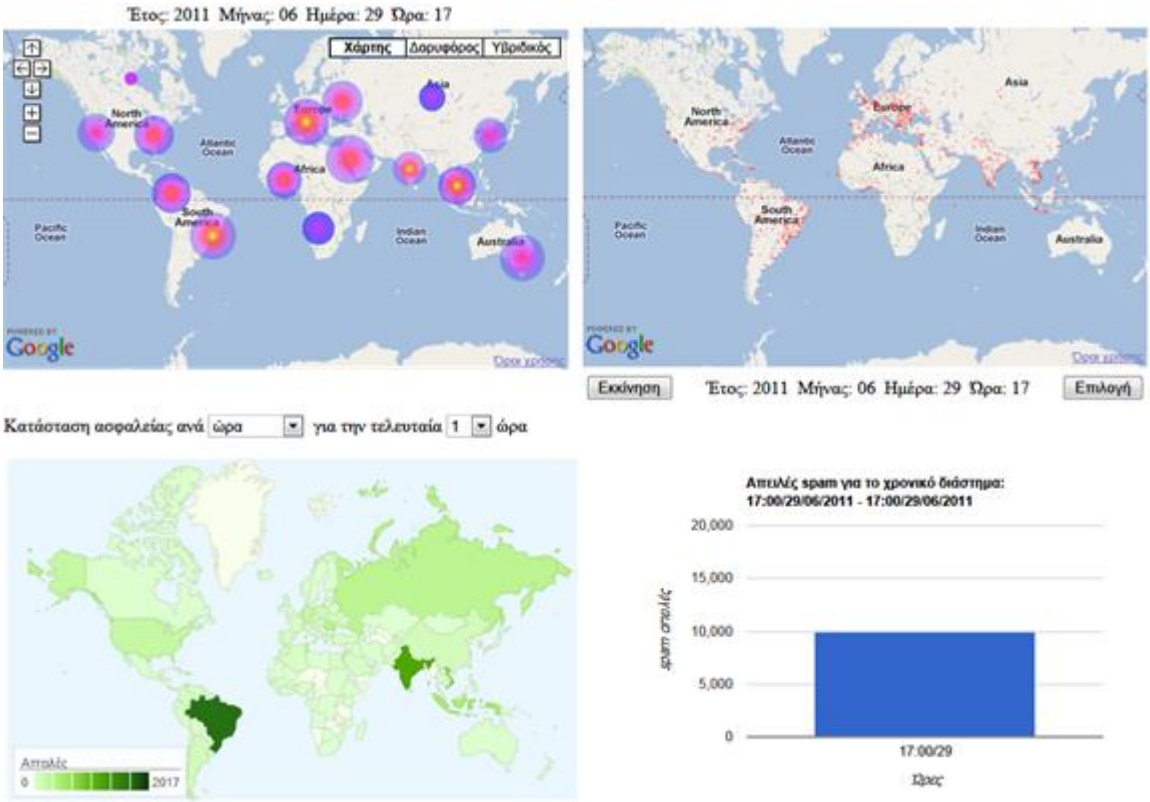
Όπως φαίνεται, με την αύξηση της χρονικής βαθμίδας, αυξάνονται και οι χρόνοι ανάκτησης των δεδομένων διότι όλο και περισσότερα αντικείμενα αυτών ικανοποιούν τους χρονικούς περιορισμούς κάθε φορά. Επίσης, οι λειτουργίες που αφορούν την ανάκτηση δεδομένων για τη δημιουργία εικόνων στον εξυπηρετητή και διέπουν τον κώδικα του αρχείου

genAnimationData.php, φαίνεται ότι είναι αρκετά περιορισμένες, αφού αρκούνται στην απλή ανάκτηση-σαρωση των αντικειμένων και τη δημιουργία-«τύπωση» σχετικής εικόνας. Αντίθετα, τα αρχεία genClusterData.php και genGraphData.php περιλαμβάνουν σχετικά απαιτητικές λειτουργίες που επηρεάζονται με την αύξηση του αριθμού των αντικειμένων των δεδομένων που ανακτούνται. Βέβαια, σχετικά με την περίπτωση του δεύτερου αρχείου, αυτό παρουσιάζει μεγαλύτερους χρόνους ανάκτησης ακόμα και από την εκτέλεση της μεθόδου συσταδοποίησης, δεδομένου ότι περιλαμβάνει κατά την ανάκτηση από τη βάση συνένωση των πληροφοριών πινάκων της βάσης, ώστε να προκύψουν τα απαραίτητα δεδομένα που σχετίζονται με στοιχεία χωρών.

5.4 Μελέτη Δεδομένων και Εξαγωγή Συμπερασμάτων

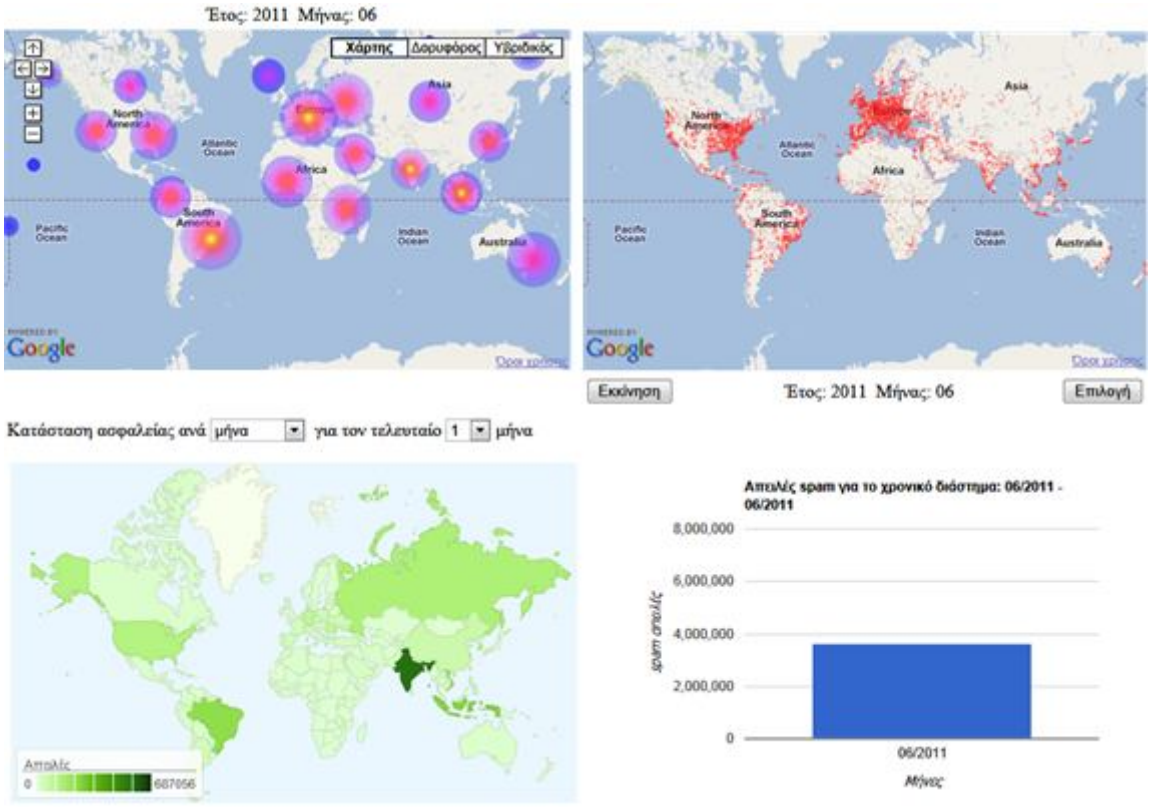
Έχοντας περιγραφθεί το υλοποιηθέν σύστημα ως προς την αρχιτεκτονική, τις λειτουργίες που το διέπουν και τη διεπαφή με την οποία αλληλεπιδρά με αυτό ο χρήστης, στην παρούσα ενότητα παρουσιάζεται μια διαδικασία μελέτης των δεδομένων που παρέχει το σύστημα για το φαινόμενο spamming, ώστε να προκύψουν σχετικά χρήσιμα συμπεράσματα. Για να επιτευχθεί αυτό, γίνεται χρήση της διεπαφής χρήστη του συστήματος, ενώ τα δεδομένα που είναι διαθέσιμα στη βάση αφορούν τη χρονική περίοδο 25/1/2011 – 29/6/2011.

Μια πρώτη εικόνα περί της κατάστασης του φαινομένου spamming ανά τον κόσμο φαίνεται με την πρώτη σελίδα που εμφανίζεται κατά την πρόσβαση στην ιστοσελίδα του συστήματος και αφορά την απεικόνιση απειλών spam συσταδοποιημένων για λόγους ευκρίνειας για την τελευταία ώρα (Εικόνα 5.14). Όπως διακρίνεται, υπάρχουν έντονες εμφανίσεις απειλών spam στις περιοχές της Ευρώπης, της νοτιοανατολικής Ασίας, της Βραζιλίας (και γενικότερα της κεντρικής Νότιας Αμερικής), των ΗΠΑ, της Καραϊβικής και της βόρειας πλευράς της Νότιας Αμερικής, ενώ πολύ μικρότερης έντασης απειλές εμφανίζονται σε άλλες περιοχές όπως της Νότιας Αφρικής, της κεντρικής Ασίας και της Αυστραλίας. Αυτή είναι μια «χοντρική» εικόνα της έντασης και παράλληλα κατανομής του φαινομένου spamming ανά τον κόσμο την τελευταία ώρα, περιγράφοντας σε γενικές γραμμές ποια είναι η τρέχουσα κατάσταση. Απεικονίσεις που περιγράφουν με κάποια περισσότερη σαφήνεια αυτή τη γενική εικόνα αποτελούν εκείνες των δύο άλλων σελίδων με περισσότερο σαφή αυτή που παρουσιάζει τις απειλές spam ανά χώρα. Όπως φαίνεται, και οι τρεις απεικονίσεις συμπίπτουν ως προς την παρεχόμενη πληροφορία.



Εικόνα 5.14: Κατάσταση του spamming φαινομένου στις 29/6/2011 και ώρα 17:00 GMT0

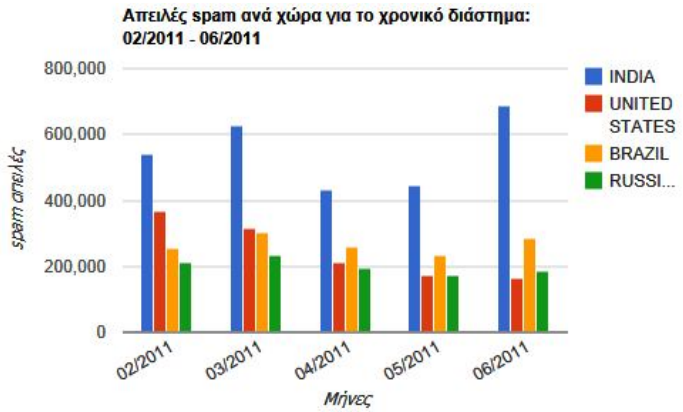
Σε περίπτωση που το ενδιαφέρον προσανατολίζεται σε αναζήτηση των πιο απειλητικών περιοχών για μια μεγαλύτερη τρέχουσα χρονική περίοδο (π.χ. ημέρα, εβδομάδα, μήνας), μιας και με δεδομένα μιας ώρας δεν μπορεί να δοθεί με σαφήνεια μια τέτοια απάντηση, τότε κάνοντας τις κατάλληλες αλλαγές στην επιλογή χρονικής βαθμίδας αναγνωρίζονται οι περιοχές αυτές ανά τον κόσμο. Όπως φαίνεται, κυρίως από τους δύο πρώτους χάρτες της Εικόνας 5.15, σχεδόν ολόκληρη η Ευρώπη αποστέλλει μηνύματα spam, θέτοντάς την στο σύνολό της ως μια από τις απειλητικές εκείνες περιοχές του κόσμου. Άλλες τέτοιες περιοχές - κράτη αποτελούν οι ΗΠΑ (κυρίως οι ανατολικές), η Βραζιλία (οι ανατολικές ακτές), η Ινδία, η Ινδονησία και η Νότια Κορέα., πράγμα το οποίο συνάδει και με το γεγονός ότι ο πληθυσμός στις συγκεκριμένες περιοχές είναι ιδιαίτερα μεγάλος. Συγκρίνοντάς δε τις τέσσερις πιο απειλητικές από αυτές για τους πέντε τελευταίους μήνες, προκύπτει ότι η πιο απειλητική είναι η Ινδία, όπως φαίνεται και στο σχετικό διάγραμμα της Εικόνας 5.16.



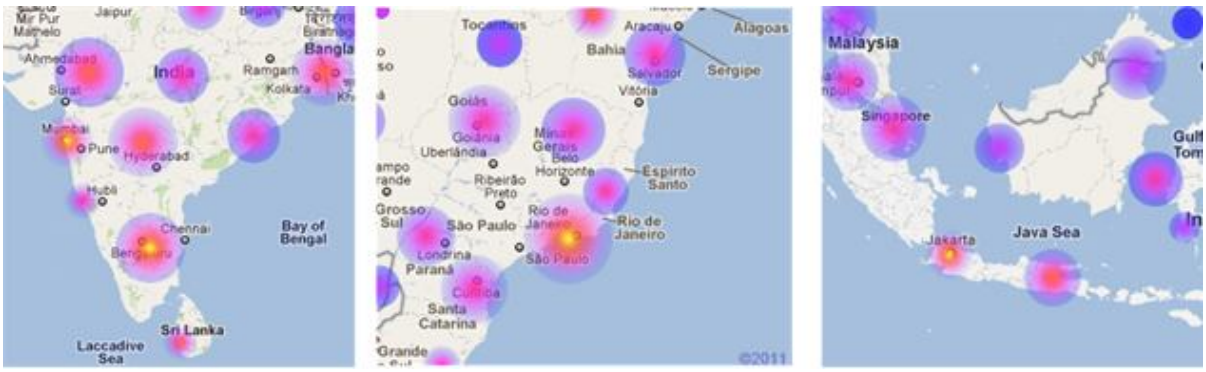
Εικόνα 5.15: Κατάσταση του spamming φαινομένου για τον μήνα Ιούνιο του 2011

Αξιοσημείωτο είναι δε το γεγονός ότι οι περισσότερες από αυτές τις απειλητικές περιοχές-κράτη συγκεντρώνουν τη δραστηριότητά τους αυτή σε αστικά κέντρα που είναι περισσότερο ανεπτυγμένα σε σχέση με το υπόλοιπο των περιοχών. Τέτοιες περιπτώσεις αποτελούν, όπως φαίνεται και στην Εικόνα 5.17, οι εξής:

- η Ινδία, που συγκεντρώνει μεγάλη κίνηση μηνυμάτων spam σε πόλεις όπως οι Mumbai, Bengaluru και Chennai,



Εικόνα 5.16: Σύγκριση ως προς τον αριθμό των απειλών spam τεσσάρων «απειλητικών» χωρών ανά τον κόσμο για την περίοδο 2/2011 – 6/2011

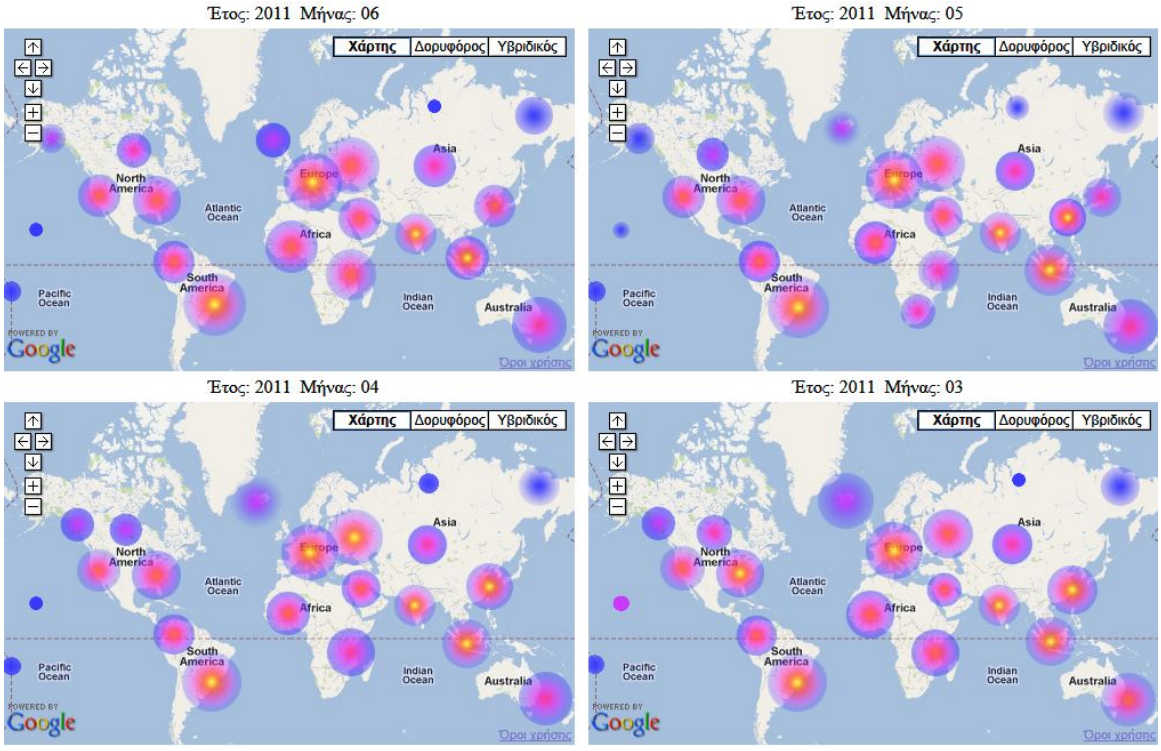


Εικόνα 5.17: Συγκέντρωση του μεγαλύτερου μέρους των απειλών spam κρατών σε συγκεκριμένες περιοχές γύρω από αστικά κέντρα. Περιπτώσεις της Ινδίας, της Βραζιλίας και της Ινδονησίας (από αριστερά προς τα δεξία)

- η Βραζιλία, που εμφανίζει μεγάλη ένταση του φαινομένου spamming γύρω από το Rio de Janeiro, και

- η Ινδονησία, όπου η πρωτεύουσα Jakarta ξεχωρίζει ως ισχυρή πηγή μηνυμάτων spam στην περιοχή γύρω από αυτήν, αλλά και γενικότερα ολόκληρης της Ινδονησίας.

Τέλος, με τη δυνατότητα παράθεσης χαρτών που προσφέρει η σχετική σελίδα του ιστότοπου, οι οποίοι αφορούν διαδοχικές χρονικές περιόδους παρατήρησης του φαινομένου spamming,



Εικόνα 5.18: Χρονική εξέλιξη του φαινομένου spamming για τέσσερις συνεχόμενους μήνες (3/2011 – 6/2011)

μπορεί να μελετηθεί η εξέλιξη του τελευταίου τόσο ως προς την ένταση όσο και ως προς την κατανομή. Συγκεκριμένα, παραθέτοντας τέσσερις χάρτες με καθέναν να απεικονίζει τις απειλές spam κάθε μήνα, όπως φαίνεται στην Εικόνα 5.18, μπορεί να ειπωθεί σε γενικές γραμμές ότι με μια μικρή διακύμανση ως προς το πλήθος των εμφανιζόμενων απειλών, οι χώρες που είναι γενικά απειλητικές ως προς την αποστολή μηνυμάτων spam βρίσκονται στην ίδια κατάσταση για κάθε χρονική περίοδο που μπορεί να εξετασθεί (εδώ μήνες).

Κεφάλαιο 6
Συμπεράσματα και Μελλοντική Εργασία

Στα συγκεκριμένο κεφάλαιο παρουσιάζονται τα συμπεράσματα που προκύπτουν από την μεταπτυχιακή διατριβή, καθώς και η μελλοντική εργασία που θα μπορούσε να πραγματοποιηθεί.

6.1 Συμπεράσματα

Στη μεταπτυχιακή διατριβή μελετήθηκε το φαινόμενο spamming ως προς τη φύση του, δηλαδή πως λειτουργεί πάνω από τη δομή του ηλεκτρονικού ταχυδρομείου και με ποιον τρόπο αυτό επιτυγχάνεται, τους τρόπους με τους οποίους αυτό μπορεί να αντιμετωπισθεί παρουσιάζοντας τις σχετικές τεχνικές αντιμετώπισης που έχουν αναπτυχθεί, και τη νομοθεσία που βοηθάει προς αυτήν την κατεύθυνση. Βάσει δε αυτών, σχεδιάστηκε και αναπτύχθηκε ένα σύστημα χρησιμοποιώντας της διαθέσιμες τεχνολογίες του Διαδικτύου που είναι δημόσια προσπελάσιμο μέσω σχετικού ιστότοπου, παρέχοντας έτσι τη δυνατότητα σε κάποιον χρήστη πρόσβασης στη σχετική με το spamming πληροφορία, ώστε να μπορέσει να το κατανοήσει ως προς την

προέλευσή του, την ένταση με την οποία αυτό εμφανίζεται, καθώς και τον τρόπο με τον οποίο αυτό εξελίσσεται στο χρόνο. Για να επιτευχθεί αυτός ο χωρο-χρονικός προσδιορισμός του φαινομένου, η σχεδίαση του συστήματος βασίστηκε στην τεχνική αντιμετώπισης του spamming η οποία στηρίζεται στον αποκλεισμό της προέλευσης τέτοιων μηνυμάτων μέσω σχετικών λιστών αποκλεισμού, και αυτό, διότι η προέλευση στη συγκεκριμένη περίπτωση καθορίζεται από την IP διεύθυνση που έχει κάθε συσκευή συνδεδεμένη στο Διαδίκτυο και αποτελεί γεωγραφικά προσδιορίσιμο στοιχείο. Ο δε γεωπροσδιορισμός των αποκλεισμένων αυτών IP διευθύνσεων ή απειλές spam που ανακτούνται από τη σχετική μαύρη λίστα την οποία διαθέτει δημόσια ο παροχέας Heise, επιτεύχθηκε με τη χρήση της βάσης δεδομένων αντιστοίχισης διευθύνσεων IP σε συντεταγμένες και άλλα χωρικά στοιχεία που παρέχει ελεύθερα ο παροχέας HostIP.

Βασικό στοιχείο στο οποίο επικεντρώθηκε το συγκεκριμένο σύστημα αποτέλεσε ο τρόπος απεικόνισης της σχετικής με το spamming πληροφορίας, ώστε να επιτευχθεί η όσο το δυνατό καλύτερη παρουσίαση του φαινομένου. Για το λόγο αυτό μελετήθηκαν διάφορες απεικονίσεις οι οποίες χρησιμοποιούνται σε συναφείς εργασίες που περιλαμβάνουν επεξεργασία τέτοιων δεδομένων, και έτσι κρίθηκαν απαραίτητοι και υλοποιήθηκαν οι εξής τρεις τρόποι απεικόνισης:

- Απεικόνιση των απειλών spam σε πολλαπλούς περιηγήσιμους χάρτες οι οποίοι αναφέρονται σε διαδοχικές χρονικές περιόδους, όπου πραγματοποιείται και χωρική συσταδοποίηση αυτών για την καλύτερη παρουσίαση τόσο της κατανομής όσο και της έντασης του φαινομένου spamming.

- Απεικόνιση της χρονικής εξέλιξης της κατανομής της προέλευσης των απειλών spam χρησιμοποιώντας εναλλαγή εικόνων (animation).

- Απεικόνιση με διαγράμματα και γραφικές παραστάσεις κυρίως της έντασης του φαινομένου spamming ανά χώρα και για διάφορες συνεχόμενες χρονικές περιόδους.

Τέλος, λόγω των ιδιαίτερων απαιτήσεων περιήγησης του πρώτου τρόπου απεικόνισης σε συνδυασμό με τον μεγάλο όγκο πληροφορίας που πρέπει να διαχειρισθεί το σύστημα, βασικό αλγοριθμικό πρόβλημα αποτέλεσε η χωρική συσταδοποίηση των δεδομένων. Για το λόγο αυτό μελετήθηκαν και συγκρίθηκαν διάφορες τέτοιες μέθοδοι που έχουν αναπτυχθεί για το σκοπό αυτό, θέτοντας ως «περιβάλλον» σύγκρισης τις προδιαγραφές απεικόνισης του υλοποιηθέντος συστήματος. Από τα σχετικά αποτελέσματα που παράχθηκαν, διακρίθηκαν μεταξύ άλλων

ορισμένα χαρακτηριστικά των μεθόδων που συμβάλλουν περισσότερο σε μια καλή συσταδοποίηση. Αυτά περιλαμβάνουν τη χρήση στους σχετικούς υπολογισμούς των κέντρων βάρους (centroids) των αντικειμένων των δεδομένων που συσταδοποιούνται, τη χρήση της διαδικασίας ανάθεσης των αντικειμένων σε συστάδες με τον τρόπο που υπαγορεύει η μέθοδος BIRCH για λόγους ταχύτητας, και τέλος τη χρήση πλέγματος, όπως υπαγορεύει για παράδειγμα η μέθοδος STING, ώστε να επιταχυνθεί ακόμα περισσότερο η διαδικασία της συσταδοποίησης. Λαμβάνοντας υπόψη όλα αυτά, υλοποιήθηκε ένας μηχανισμός συσταδοποίησης που να ταιριάζει στις συγκεκριμένες ανάγκες απεικόνισης του συστήματος, ώστε να επιτευχθεί η όσο το δυνατό γρηγορότερη και ακριβέστερη ενημέρωση των δεδομένων κατά την περιήγηση κάποιου χρήστη στου χάρτες.

6.2 Μελλοντική Εργασία

Το σύστημα που αναπτύχθηκε βασίστηκε σε συγκεκριμένες δυνατότητες που παρέχονταν τη συγκεκριμένη χρονική περίοδο υλοποίησής του αναφορικά με την τεχνολογία διαδικτυακού προγραμματισμού που χρησιμοποιήθηκε, τα δεδομένα που περιλήφθηκαν στη σχετική βάση για την περιγραφή της χωρο-χρονικής φύσης του φαινομένου spamming και τις δυνατότητες που παρέχονταν για την περίπτωση της απεικόνισης των δεδομένων αυτών. Επί αυτών, αρχικά, θα μπορούσε να χρησιμοποιηθεί κάποια άλλη τεχνολογία διαδικτυακού προγραμματισμού πλούσιων εφαρμογών διαδικτύου (Rich Internet Application client technologies), όπως είναι για παράδειγμα οι Adobe Flash και Microsoft Silverlight, ώστε να βελτιωθεί ποιοτικά η περιήγηση του χρήστη στην ιστοσελίδα του συστήματος. Προς την κατεύθυνση αυτή, θα μπορούσαν να χρησιμοποιηθούν και ενημερωμένες εκδόσεις των Google Maps και Visualization API, οι οποίες παρέχουν νέα χαρακτηριστικά που βοηθούν στη βελτίωση της απεικόνισης των δεδομένων. Εκτός δε από τους τρεις τρόπους απεικόνισης των δεδομένων που παρέχονται από το σύστημα, θα ήταν χρήσιμη για την καλύτερη περιγραφή του φαινομένου spamming και η προβολή σχετικών πινάκων οι οποίοι θα περιλαμβάνουν αναλυτικά στοιχεία τόσο στατιστικά για την περίπτωση χωρών και πόλεων όσο και περισσότερο περιγραφικά που μπορούν να αφορούν λεπτομέρειες για τις απειλές spam που εντοπίζονται και συγκεκριμένα τις αντίστοιχες διευθύνσεις IP. Για να επιτευχθεί δε κάτι τέτοιο απαιτείται ο συνδυασμός περισσότερων βάσεων δεδομένων και από άλλους παροχείς, οι οποίοι παρέχουν επιπλέον τέτοιας μορφής πληροφορία, πράγμα το οποίο μπορεί να βοηθήσει και στον εμπλουτισμό ή συμπλήρωση της απαραίτητης πληροφορίας.

Αναφορικά με τη διαδικασία συσταδοποίησης, ο μηχανισμός που υλοποιήθηκε για το σύστημα δεν λαμβάνει υπόψη του τις γεωγραφικές ιδιαιτερότητες των χαρτών που αφορούν για παράδειγμα θαλάσσιες περιοχές, βουνά και άλλης μορφής τέτοια «εμπόδια». όπως θα μπορούσαν να χαρακτηρισθούν για τη περίπτωση της συσταδοποίησης. Για το λόγο αυτό ίσως η ενσωμάτωση κάποιων χαρακτηριστικών ορισμένων μεθόδων συσταδοποίησης [98, 104] που λαμβάνουν υπόψη τέτοιας μορφής εμπόδια θα μπορούσε να βελτιώσει τις συστάδες που δημιουργούνται αποφεύγοντας για παράδειγμα την τοποθέτηση των κέντρων τους σε περιοχές που δεν είναι δυνατή η ύπαρξη spam απειλής.

Βιβλιογραφία

[001] M. Abadi, M. Burrows, M. Manasse, T. Wobber. «Moderately Hard, Memory-Bound Functions». In 10th Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, February 2003

[002] E. Allman, J. Callas, M. Delany, M. Libbey, J. Fenton, M. Thomas. «DomainKeys Identified Mail (DKIM) Signatures». RFC 4871, May 2007 [Online]. Available at: http://tools.ietf.org/pdf/rfc4871.pdf

[003] E. Allman, J. Fenton, M. Delany, J. Levine. «DomainKeys Identified Mail (DKIM) Author Domain Signing Practices (ADSP)». RFC 5617, August 2009 [Online]. Available at: http://tools.ietf.org/pdf/rfc5617.pdf

[004] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, C. Spyropoulos. «An Evaluation of Naive Bayesian Anti-Spam Filtering». In Proceedings of the Workshop on Machine Training in the New Information Age: 11th European Conference on Machine Learning (ECML), pp. 9-17, 2000

[005] Anti-Spam Research Group (ASRG) [Online]. Available at: http://asrg.sp.am

[006] S. Arrison. «Canning Spam: An Economic Solution to Unwanted Email». Pacific Research Institute, 2004

[007] D. Arthur, S. Vassilvitskii. «k-means++: The Advantages of Careful Seeding». In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027-1035, 2007

[008] Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα, http://www.dpa.gr

[009] G. Babu, M. Murty. «A Near-Optimal Initial Seed Value Selection in k-Means Algorithm Using a Genetic Algorithm». Pattern Recognition Letters, vol. 14 (10), pp. 763-769, 1993

[010] A. Back. «Hashcash - A Denial of Service Counter-Measure». Technical report, August 2002 [Online]. Available at: http://www.hashcash.org/papers/hashcash.pdf

[011] P. Burrough. «Principles of Geographical Information Systems for Land Resources Assessment». Oxford University Press, 1986

[012] J. Callas, L. Donnerhacke, H. Finney, R. Thayer. «OpenPGP Message Format». RFC 4880, November 2007 [Online]. Available at: http://tools.ietf.org/pdf/rfc4880.pdf

[013] J. Carpinter, R. Hunt. «Tightening the Net: A Review of Current and Next Generation Filtering Tools». Computers & Security, Volume25, pp. 566-578, 2006

[014] CBL, The Composite Blocking List, http://cbl.abuseat.org

[015] P. Cheeseman, J. Stutz. «Bayesian Classification (AutoClass): Theory and Results». Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Ed.), AAAI/MIT Press, Cambridge, MA, pp. 153-180, 1996

[016] F. Coelho. «Exponential Memory-Bound Functions for Proof of Work Protocols». Cryptology ePrint Archive, Report 2005/356, 2005 [Online]. Available at: http://eprint.iacr.org/2005/356.pdf

[017] Controlling the Assault of Non-Solicited Pornography And Marketing Act of 2003 (CAN-SPAM Act of 2003) [Online]. Available at: http://www.ftc.gov/os/caselist/0723041/canspam.pdf

[018] M. Crispin. «Internet Message Access Protocol - Version 4rev1». RFC 3501, March 2003 [Online]. Available at: http://tools.ietf.org/pdf/rfc3501.pdf

[019] A. Dekok. «Lightweight MTA Authentication Protocol (LMAP) Discussion and Applicability Statement». Internet draft, April 2004 [Online]. Available at: http://tools.ietf.org/pdf/draft-irtf-asrg-lmap-discussion-01.pdf

[020] M. Delany. «Domain-Based Email Authentication Using Public Keys Advertised in the DNS (DomainKeys)». RFC 4870, May 2007 [Online]. Available at: http://tools.ietf.org/pdf/rfc4870.pdf

[021] A. P. Dempster, N. M. Laird, D. B. Rubin. «Maximum Likelihood from Incomplete Data Via the EM Algorithm». Journal of the Royal Statistical Society, vol. 39, pp. 1-38, 1977

[022] V. Deshpande, R. Erbacher, C. Harris. «An Evaluation of Naive Bayesian Anti-Spam Filtering Techniques». In Proceedings of the IEEE Information Assurance Workshop, West Point, NY, pp. 333-340, June 2007

[023] Dnswl.org, http://www.dnswl.org

[024] C. Dwork, A. Goldberg, M. Naor. «On Memory-Bound Functions for Fighting Spam». In Advances in Cryptology - CRYPTO 2003, volume 2729 of Lecture Notes in Computer Science, pp. 426-444. Springer, 2003

[025] C. Dwork, M. Naor. «Pricing via Processing or Combating Junk Mail». In Advances in Cryptology - Crypto 1992, Lecture notes in Computer Science 740, pp. 139-147, Springer Verlag, 1993

[026] D. Eastake 3rd, T. Hansen. «US Secure Hash Algorithms (SHA and HMAC-SHA)». RFC 4634, July 2006 [Online]. Available at: http://tools.ietf.org/pdf/rfc4634.pdf

[027] D. Eastlake 3rd, P. Jones. «US Secure Hash Algorithm 1 (SHA1)». RFC 3174, September 2001 [Online]. Available at: http://tools.ietf.org/pdf/rfc3174.pdf

[028] J. Fenton, M. Thomas, «Identified Internet Mail draft-fenton-identified-mail-02». Internet draft, May 2005, [Online]. Available at: http://tools.ietf.org/pdf/draft-fenton-identified-mail-02.pdf

[029] M. Ester, H.-P. Kriegel, J. Sander, X. Xu. «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise». In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, (KDD), Portland, OR, USA, AAAI Press, pp. 226–231, 2–4 August 1996

[030] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee. «Hypertext Transfer Protocol -- HTTP/1.1». RFC 2616, June 1999 [Online]. Available at: http://tools.ietf.org/pdf/rfc2616.pdf

[031] P. M. Figliola. «Spam: An Overview of Issues Concerning Commercial Electronic Mail». CRS report for Congress, May 2008 [Online]. Available at: http://www.ipmall.info/hosted_resources/crs/RL31953_080514.pdf

[032] D. Fisher. «Improving Inference through Conceptual Clustering». In Proceedings of the National Conference on Artificial Intelligence (AAAI'87), Seattle, WA, pp. 461-465, July 1987

[033] E. Forgy. «Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications». Biometrics, vol. 21, pp. 768-780, 1965

[034] J. Friedl. «Mastering Regular Expressions: Powerful Techniques for Perl and Other Tools». (Nutshell Handbook) O'Reilly and Associates, 1997

[035] J. J. Garret. «AJAX: A New Approach to Web Applications». Article on the Adaptive Path Website, 2005 [Online]. Available at: http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications

[036] T. Green. «How URL Spam Filtering Beats Bayesian/Heuristics Hands Down». Greenview Data, Inc, 2005 [Online]. Available at: http://www.greenviewdata.com/documents/white papers/ssh url filtering white paper.pdf

[037] S. Guha, R. Rastogi, K. Shim. «CURE: An Efficient Clustering Algorithm for Large Databases». In Proceeding of the ACM SIGMOD International Conference on Management of Data, pp. 73-84, 1998

[038] Habeas, http://www.habeas.com

[039] J. Han, M. Kamber. «Data Mining: Concepts and Techniques». Morgan Kaufmann, 2001

[040] E. Harris. «The Next Step in the Spam Control War: Greylisting». Article, August 2003, http://www.greylisting.org/articles/whitepaper.shtml

[041] Heise Medien Gruppe GmbH & Co. KG [Online]. Available at: http://www.heise-medien.de

[042] International Organization for Standardization [Online]. Available at: http://www.iso.org/iso/english country names and code elements

[043] International Telecommunication Union, «Report on Spam». Report by the ITU Secretary General on Spam to ITU Council 2005, Geneva, July 2005 [Online]. Available at: http://www.itu.int/osg/spu/spam/itu-spam-council-report.pdf

[044] Internet Research Task Force (IRTF) [Online]. Available at: http://www.irtf.org

[045] IP2Location.com [Online]. Available at: http://www.ip2location.com

[046] Spam evolution reports by Karspersky Corporation [Online]. Available at: http://www.securelist.com/en/analysis/spam

[047] I. Katsavounidis, C. Kuo, Z. Zhang. «A New Initialization Technique for Generalized Lloyd Iteration». IEEE Signal Processing Letters, vol. 1 (10), pp. 144-146, 1994

[048] L. Kaufman and, P.J. Rousseeuw. «Clustering Large Data Sets (with discussion)». In Pattern Recognition in Practice II, edited by E. S. Gelsema and L. N. Kanal, Elsevier/North-Holland, Amsterdam, pp. 405-416, 1986

[049] L. Kaufman, P.J. Rousseeuw. «Clustering by Means of Medoids». In Statistical Data Analysis Based on the Norm, Y. Dodge, Ed., North Holland Elsevier, Amsterdam, pp. 405-416, 1987

[050] L. Kaufman and, P. J. Rousseeuw. «Finding Groups in Data: An Introduction to Cluster Analysis». John Wiley & Sons, Brussels, Belgium, 1990

[051] J. Klensin. «Simple Mail Transfer Protocol». RFC 5321, October 2008 [Online]. Available at: http://tools.ietf.org/pdf/rfc5321.pdf

[052] O. Kolesnikov, W. Lee, R. Lipton. «Filtering Spam Using Search Engines». Technical report, GIT-CC-03-58, Georgia Tech, College of Computing, Georgia Institute of Technology, Atlanta, 2003 [Online]. Available at: ftp://ftp.cc.gatech.edu/pub/coc/tech_reports/2003/GIT-CC-03-58.pdf

[053] R. Kraut, S. Sunder, R. Telang, J. Morris. «Pricing Electronic Mail to Solve the Problem of Spam». Human-computer Interaction 20, pp. 195-223, 2005

[054] J. Levine, «Experiences with Greylisting». Procceding of the Second Conference on Email and Anti-spam (CEAS 2005), July 2005 [Online]. Available at: http://ceas.cc/2005/papers/120.pdf

[055] Y. Linde, A. Buzo, R. M. Gray. «An Algorithm for Vector Quantizer Design». IEEE Transactions on Communications, vol. COM-28, pp. 84-95, 1980

[056] S. P. Lloyd. «Least-Squares Quantization in PCM». IEEE Transactions on Information Theory, vol. IT-28, pp. 129-137, March 1982

[057] S. Lu, K. Fu. «A Sentence-to-Sentence Clustering Procedure for Pattern Analysis». IEEE Transaction on Systems, Man and Cybernetics, vol. 8, pp. 381-389, 1978

[058] L. Lyon, M. Wong. «Sender ID: Authenticating E-Mail». RFC 4406, April 2006 [Online]. Available at: http://tools.ietf.org/pdf/rfc4406.pdf

[059] J. B. MacQueen. «Some Methods for Classification and Analysis of Multivariate Observations». In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297, 1967

[060] MaxMind Inc. [Online]. Available at: http://www.maxmind.com

[061] A. Melnikov, K. Zeilenga. «Simple Authentication and Security Layer (SASL)». RFC 4422, July 2006 [Online]. Available at: http://tools.ietf.org/pdf/rfc4422.pdf

[062] N. Memarsadeghi, D. M. Mount, N. S. Netanyahu, J. L. Moigne. «A Fast Implementation of the Isodata Clustering Algorithm». International Journal of Computational Geometry & Applications, vol. 17 (1), pp. 71-103, 2007

[063] MessageLabs, «MessageLabs Intelligent Reports» [Online]. Available at: http://www.messagelabs.com/resources/mlireports

[064] Messaging Anti-Abuse Working Group (MAAWG), «Email Metrics Program: The Network Operators' Perspective». Report #12, Third and Fourth Quarter 2009, March 2010 [Online]. Available at: http://www.maawg.org/sites/maawg/files/news/MAAWG_2009-Q3Q4_Metrics_Report_12.pdf

[065] J. Myers, M. Rose. «Post Office Protocol - Version 3». STD 53, RFC 1939, May 1996 [Online]. Available at: http://tools.ietf.org/pdf/rfc1939.pdf

[066] N. Netanyahu, T. Kanungo, D. Mount, C. Piatko, R. Silverman, A. Wu. «An Efficient K Means Clustering Algorithm: Analysis and Implementation». IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, p. 887-892, 2002

[067] R. T. Ng, J. Han. «Efficient and Effective Clustering Methods for Spatial Data Mining». In Proceedings of the 20th International Conference on Very Large Data Bases, (VLDB), Santiago, Chile, Morgan Kaufmann, pp. 144–155, 12–15 September 1994

[068] R.T. Ng, J. Han. «CLARANS: A Method for Clustering Objects for Spatial Data Mining». IEEE Transactions on Knowledge and Data Engineering, vol. 14 (5), pp. 1003-1016, 2002

[069] Nigerian Criminal Code. «Obtaining Property by False Pretences; Cheating». chapter 38, section 419 [Online]. Available at: http://www.nigeria-law.org

[070] The Nix Spam DNS-based blacklist, http://www.dnsbl.manitu.net

[071] NJABL, Not Just Another Bogus List, http://www.njabl.org

[072] Νόμος 2472. «Προστασία του Ατόμου από την Επεξεργασία Δεδομένων Προσωπικού Χαρακτήρα». 1997 [Online]. Available at: http://www.dpa.gr/pls/portal/docs/PAGE/APDPX/LAW/NOMOTHESIA%20PROSOPIKA%20DEDOMENA/2472_97_APR_10_FINAL.PDF

[073] Νόμος 3471. «Προστασία Δεδομένων Προσωπικού Χαρακτήρα και της Ιδιωτικής Ζωής στον Τομέα των Ηλεκτρονικών Επικοινωνιών». 2006 [Online]. Available at: http://www.dpa.gr/pls/portal/docs/PAGE/APDPX/LAW/NOMOTHESIA%20PROSOPIKA%20DEDOMENA/3471_2006.PDF

[074] Οδηγία 2000/31/ΕΚ [Online]. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2000:178:0001:0016:EL:PDF

[075] Οδηγία 2002/58/ΕΚ [Online]. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:201:0037:0047:EL:PDF

[076] Οδηγία 2009/136/ΕΚ [Online]. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:337:0011:0036:EL:PDF

[077] J. M. Pena, J. A. Lozano, P. Larranaga. «An Empirical Comparison of Four Initialization Methods for the k-means Algorithm». Pattern Recognition Letters, vol. 20, pp. 1027-1040, 1999

[078] D. T. Pham, S. S. Dimov, C. D. Nguyen. «An Incremental K-means Algorithm». In Proceedings of the Institution of Mechanical Engineers, Part C, Journal of Mechanical Engineering Science. vol. 218 (7), pp.783-795, July 2004

[079] S. Phillips. «Acceleration of K-Means and Related Clustering Algorithms». In Proceedings of the 4th International Workshop on Algorithm Engineering and Experiments, pp. 166-177, 2002

[080] J. Postel. «Transmission Control Protocol». STD 7, RFC 793, September 1981 [Online]. Available at: http://tools.ietf.org/pdf/rfc793.pdf

[081] V. V. Prakash. «Vipul's Razor». [Online]. Available at: http://razor.sourceforge.net/

[082] B. Ramsdell, S. Turner. «Secure/Multipurpose Internet Mail Extensions (S/MIME) Version 3.2 Message Specification». RFC 5751, January 2010 [Online]. Available at: http://tools.ietf.org/pdf/rfc5751.pdf

[083] S. J. Redmond, C. Heneghan. «A Method for Initializing the K-means Clustering Algorithm Using kd-Trees». Pattern Recognition Letters, vol. 28, pp. 965-973, 2007

[084] P. Resnick. «Internet Message Format». RFC 5782, February 2008 [Online]. Available at: http://tools.ietf.org/pdf/rfc5782.pdf

[085] P. Resnick. «Internet Message Format». RFC 5322, October 2008 [Online]. Available at: http://tools.ietf.org/pdf/rfc5322.pdf

[086] Rhyolite Software. «Distributed Checksum ClearingHouse». [Online]. Available at: http://www.rhyolite.com/dcc/

[087] R. Rivest. «The MD5 Message-Digest Algorithm». RFC 1321, April 1992 [Online]. Available at: http://tools.ietf.org/pdf/rfc1321.pdf

[088] R. Siemborski, A. Melnikov. «SMTP Service Extension for Authentication». RFC 4954, July 2007 [Online]. Available at: http://tools.ietf.org/pdf/rfc4954.pdf

[089] SORBS, Spam and Open Relay Blocking System, http://www.sorbs.net

[090] SpamCop, http://www.spamcop.net

[091] The Spamhaus project, http://www.spamhaus.org

[092] Spamhaus. «The Definition of Spam» [Online]. Available at: http://www.spamhaus.org/definition.html

[093] Swiss whitelist, http://www.swisswhitelist.ch

[094] Symantec Corporation. «The State of Spam & Phishing» [Online]. Available at: http://www.symantec.com/business/theme.jsp?themeid=state_of_spam

[095] Symantec Corporation. «State of Spam: A Monthly Report». Report #32, August 2009 [Online]. Available at: http://eval.symantec.com/mktginfo/enterprise/otherresources/b-state_of_spam_report_08-2009.en-us.pdf

[096] B. Templeton. «Origin of the Term "spam" to Mean Net Abuse» [Online]. Available at: http://www.templetons.com/brad/spamterm.html

[097] F. Tobin. «Pyzor», [Online]. Available at: http://sourceforge.net/apps/trac/pyzor/

[098] A. K. H. Tung, J. Hou, J. Han. «Spatial Clustering in the Presence of Obstacles». In Proceedings 2001 International Conference on Data Engineering (ICDE'01), Heidelberg, Germany, April 2001

[099] D. Turner, D. Havey, «Controlling Spam through Lightweight Currency». In Proceedings of the 37th Annual Hawaii International Conference on System Sciences, Honolulu, USA, 2004

[100] The UCEPROTECT network project, http://www.uceprotect.net

[101] M. Vrahatis, B. Boutsinas, P. Alevizos, G. Pavlides. «The New k-Windows Algorithm for Improving the k-Means Clustering Algorithm». Journal of Complexity, Academic Press, vol. 18, pp. 375-391, 2002

[102] W. Wang, J. Yang, R. Muntz. «STING: A Statistical Information Grid Approach to Spatial Data Mining». In Proceedings of the twenty-third International Conference on Very Large Data Bases, Athens, Greece, pp. 186-195, 1997

[103] M. Wong, W. Schlitt. «Sender Policy Framework (SPF) for Authorizing Use of Domains in E-Mail, Version 1». RFC 4408, April 2006 [Online]. Available at: http://tools.ietf.org/pdf/rfc4408.pdf

[104] O. R. Zaiane, C.-H. Lee. «Clustering Spatial Data in the Presence of Obstacles: a Density-Based Approach». Sixth International Database Engineering and Applications Symposium (IDEAS 2002), Edmonton, Alberta, Canada, July 2002

[105] T. Zhang, R. Ramakrishnan, M. Livny. «BIRCH: An Efficient Data Clustering Method for Very Large Databases». In: Jagadish, H.V., Mumick, I.S., eds. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Quebec: ACM Press, pp. 103-114, 1996