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

Μία μέθοδος δειγματοληψίας με διευθύνσεις πρωτοκόλλου Ίντερνετ για την εξόρυξη δεδομένων από τον Παγκόσμιο Ιστό και το Διαδίκτυο

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Παρασκευάς Β. Λεκέας

Αθήνα, Ιανουάριος 2004

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

Μία μέθοδος δειγματοληψίας με διευθύνσεις πρωτοκόλλου Ίντερνετ για την εξόρυξη δεδομένων από τον Παγκόσμιο Ιστό και το Διαδίκτυο

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Παρασκευάς Β. Λεκέας

Συμβουλευτική επιτροπή : Φώτω Αφράτη

Γεώργιος Παπακωνσταντίνου

Μιλτιάδης Αναγνώστου

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 12η Ιανουαρίου 2004

Φ. Αφράτη Καθηγήτρια Ε.Μ.Π.	Γ. Παπακωνσταντίνου Καθηγητής Ε.Μ.Π.	Μ. Αναγνώστου Καθηγητής Ε.Μ.Π.
Τ. Σελλής Καθηγητής Ε.Μ.Π.	Β. Λούμος Καθηγητής Ε.Μ.Π.	Γ. Κολέτσος Αν. Καθηγητής Ε.Μ.Π.

Μ. Γεργατσούλης
Επικ. Καθ. Ιονίου Παν.

Αθήνα, Ιανουάριος 2004

Π α ρ α σ κ ε υ ά ς Β. Λ ε κ έ α ς

Διδάκτορας Ε.Μ.Π.

ΠΕΡΙΛΗΨΗ

Ο παγκόσμιος ιστός (web) και η τεχνολογική του πλατφόρμα, το διαδίκτυο (Internet), είναι δύο μεγάλα και πολύπλοκα δίκτυα τα οποία δεν μπορούν να μελετηθούν με άλλο τρόπο παρά μόνο με παρατηρήσεις και μετρήσεις. Για το λόγο αυτό υπάρχει ανάγκη εύρεσης μεθόδων εξαγωγής στατιστικών δειγμάτων από τα πολύπλοκα αυτά δίκτυα (κεφάλαιο 1).

Δύο μέθοδοι κυρίως υπάρχουν για εξαγωγή δειγμάτων. Η πρώτη μέθοδος ονομάζεται δειγματοληψία με random walk και βασίζεται στην έννοια των τυχαίων περιπάτων (random walk). Αυτή η μέθοδος, χρησιμοποιώντας τη συνεκτικότητα του web γράφου, κατασκευάζει σχεδόν ομοιόμορφα και τυχαία δείγματά του βάσει της κατανομής ισορροπίας του περιπάτου. Η δεύτερη μέθοδος, η οποία είναι και το κύριο αντικείμενο της διατριβής, ονομάζεται δειγματοληψία με IP (IP sampling) και σύμφωνα με αυτήν ένα δείγμα του web προκύπτει εάν πάρουμε ένα δείγμα από IP διευθύνσεις και κρατήσουμε όσες από αυτές ανήκουν σε web hosts (κεφάλαιο 2).

Συνήθως η δειγματοληψία με IP εφαρμόζεται σε όλο το χώρο διευθύνσεων του Internet (IPv4 - Internet Protocol version 4), οπότε και προκύπτει ένα αντιπροσωπευτικό δείγμα του. Σε αυτή τη διατριβή εφαρμόσαμε την πιο πάνω μέθοδο για συγκεκριμένα domains του Internet (π.χ. .gr, .uk) συμβουλευόμενοι τις βάσεις δεδομένων των ηπειρωτικών ληξίαρχων (RIR - Regional Internet Registries) που είναι υπεύθυνοι για το Internet των αντίστοιχων γεωγραφικών περιοχών. Για το σκοπό αυτό υλοποιήσαμε ένα δειγματολήπτη ο οποίος παίρνει σαν είσοδο το "χάρτη" με τις IP διευθύνσεις στις οποίες θέλουμε να κάνουμε δειγματοληψία, επιλέγει το δείγμα των IP διευθύνσεων και το "φιλτράρει" κρατώντας μόνο τις web σελίδες. Δοκιμάσαμε το δειγματολήπτη σε διάφορα domains και είδαμε ότι είναι αρκετά αξιόπιστος, π.χ. κάνοντας δειγματοληψία στο .gr υπολογίσαμε το μέγεθος του ελληνικού web και το βρήκαμε σε συμφωνία με τρίτες πηγές (κεφάλαιο 3).

Στη συνέχεια χρησιμοποιήθηκε ο δειγματολήπτης για εξαγωγή δειγμάτων και επεξεργασία τους. Έτσι, έγινε δειγματοληψία στο .uk και από το δείγμα αποδείχθηκε ότι ο τρόπος γραφής των hostname συνδέεται με αναπαράσταση χωρικής και χρονικής πληροφορίας. Συγκεκριμένα υπολογίστηκε η γεωγραφική κατανομή της υποδομής, η κατανομή της κυκλοφορίας Internet και ο ρυθμός ανάπτυξης διαφόρων ISPs (Internet Service Providers) που δραστηριοποιούντο στην περιοχή της δειγματοληψίας από δείγματα hostname (κεφάλαιο 4).

Δοκιμάστηκε, επίσης, ο δειγματολήπτης έτσι ώστε να μην εξάγει δείγματα από τους IP χάρτες αλλά να τους διατρέχει εξαντλητικά. Η πιο πάνω δοκιμή έγινε στο .jo domain όπου και υπολογίστηκε ο αριθμός των web server του. Τέλος, ο δειγματολήπτης χρησιμοποιήθηκε σαν crawler διατρέχοντας ολόκληρα web site και επαληθεύοντας κατανομές power law για τους out-degree αυτών (κεφάλαιο 5).

Λέξεις Κλειδιά

Δειγματοληψία Παγκόσμιου Ιστού, Δειγματοληψία Ίντερνετ

An IP Sampling method for Web and Internet mining

ABSTRACT

The world wide web (web) and its technological platform, the Internet, are two large and complex networks that can only be studied through observations and measurements. For that reason there is a need for developing methods that provide us with statistical samples of these large and complex networks (chapter 1).

There exist two main methods for sampling these networks. The first method is called "sampling with random walks". This method uses the web graph connectivity and constructs almost uniform random samples according to the stationary distribution of the random walk. The second method, which is the main objective of this dissertation, is called "IP sampling". This method samples the IP address space and converts the obtained samples to their web representation (chapter 2).

Usually IP sampling is applied to the whole IPv4 (Internet Protocol version 4) address space drawing representative samples. In this work we applied IP sampling in specific domains (e.g. .gr, .uk) consulting the Regional Internet Registries of the corresponding regions. We implemented an IP sampler that gets as input maps of IP addresses and samples them. We tested the sampler in various domains and we saw that it is reliable; e.g. our estimation for the size of .gr web is in agreement with third sources (chapter 3).

We used the sampler as a tool for extracting a representative sample of hostnames from .uk domain. From this sample we proved that the hostname representation can give us spatial and temporal information. Therefore we calculated the geographical distribution of the infrastructure and the internet traffic of a British ISP (Internet Service Provider) and we also calculated the growth rate of another British ISP (chapter 4).

We modified the sampler and used it as a crawler who crawls IP maps and we calculated the number of web hosts for .jo domain. Finally we used the crawler to completely crawl web sites calculating their out-degrees and proving that power law distributions holds also for the web site abstraction of the web (chapter 5).

Keywords

IP sampling, web sampling

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τόσους πολλούς ανθρώπους που με βοήθησαν στις διδακτορικές μου σπουδές που ίσως θα έπρεπε να γράψω ένα ολόκληρο κεφάλαιο.

Θέλω πρώτα, όμως, να ευχαριστήσω το Θεό, την Αγία Τριάδα, η οποία με αξίωσε να τελειώσω αυτές τις σπουδές. Στη Δόξα Της αφιερώνω αυτή την εργασία.

Θέλω στη συνέχεια να ευχαριστήσω την επιβλέπουσα, καθηγήτρια του Ε.Μ.Π., κυρία Φώτω Ν. Αφράτη, η οποία μου έχει συμπαρασταθεί πάρα πολύ τα τελευταία 6 χρόνια των σπουδών μου. Από το μεταπτυχιακό Μ.Π.Λ.Α. και τη διπλωματική εργασία που έκανα μαζί της, μέχρι την ολοκλήρωση των διδακτορικών μου σπουδών. Πιστεύω έδειξε μεγάλη υπομονή για να με ανεχτεί όλα αυτά τα χρόνια. Κυρία Αφράτη σας ευχαριστώ πολύ!

Ευχαριστώ τους γονείς μου, Βασίλειο και Ιωάννα, την αδερφή μου Ιωάννα - Βιργινία και το θείο μου Νικόλαο Λεκέα, που βοήθησαν απλόχερα παρέχοντάς μου κάθε είδους υλικής και ψυχολογικής συμπαράστασης.

Ευχαριστώ πολύ τον π. Κωνσταντίνο Καραϊσαρίδη, ο οποίος με βοήθησε να περάσω κρίσιμες καμπές στα 4 χρόνια εκπόνησης της εργασίας, αλλά και τους φίλους μου της Αγίας Μαρίνας ευχαριστώ, οι οποίοι δεν έχουν μεν καμία σχέση με τη διατριβή αυτή αλλά έδειξαν σεβασμό και μεγάλη κατανόηση στο χρόνο που τους διέθετα. Επίσης, ευχαριστώ τον π. Δημήτριο Μούτσελο και τη σύζυγό του Κατερίνα που πολλές φορές με την καλή τους φιλοξενία εξισορροπούσαν τις δυσκολίες των σπουδών μου. Ακόμα, ευχαριστώ και τον π. Βασίλειο Γεωργόπουλο, που μου έδωσε την ευκαιρία να προσφέρω σε συνανθρώπους μου βοήθεια, και να μάθω πώς να μοιράζομαι τη γνώση που αποκτώ μαζί τους.

Ένα μεγάλο ευχαριστώ στους υπαλλήλους του Ε.Μ.Π. που όλα αυτά τα χρόνια μας πρόσφεραν τις υπηρεσίες τους, χωρίς τις οποίες ίσως να μην είχε ολοκληρωθεί η παρούσα εργασία. Ευχαριστώ, λοιπόν, την κυρία Χριστούλα, που κάθε μέρα μας καθάριζε το εργαστήριο, ευχαριστώ τους υπαλλήλους του Εστιατορίου που μας έκαναν και νιώθαμε σαν το σπίτι μας, ευχαριστώ τους υπαλλήλους της Βιβλιοθήκης (Κεντρικής και Ηλεκτρολόγων) για την άμεση εξυπηρέτηση που πάντοτε είχαμε. Μεγάλο ευχαριστώ και στη μεταπτυχιακή γραμματεία των Ηλεκτρολόγων, αλλά και στους φύλακες των κτιρίων που πρωί βράδυ ξεκλείδωναν τους διαδρόμους.

Ευχαριστώ τα παιδιά με τα οποία συνεργάστηκα και δουλέψαμε μαζί κατά διαστήματα στο εργαστήριο και στο γραφείο. Μιλάω για τους Τίμο Ασλανίδη, Nidal Al-Said, Βάσια Παυλάκη και Θοδωρή Μανουηλίδη. Εύχομαι ο Θεός να τους δίνει φώτιση για να τελειώσουν και αυτοί τις σπουδές τους. Ευχαριστώ επίσης τον Αχιλλέα Μάτζιο και το Θοδωρή τον Ανδρόνικο.

Θα ήθελα να ευχαριστήσω τους καθηγητές κυρίους Γ. Παπακωνσταντίνου και Μ. Αναγνώστου που συμμετείχαν στην τριμελή μου επιτροπή, στην επιτροπή ενδιάμεσης κρίσης αλλά και στη τελική κρίση αυτής της διατριβής.

Ευχαριστώ επίσης τους καθηγητές κυρίους Μ. Θεολόγου, Β. Λούμο, Ε. Ζάχο, Φ. Κωνσταντίνου, Γ. Μέντζα, Ι. Βενιέρη, Τ. Σελλή, Γ. Κολέτσο και Μ. Γεργατσούλη.

Γενικά ευχαριστώ το ίδρυμα που μου παρείχε αυτή την ευκαιρία να εργαστώ στους χώρους του και να ασχοληθώ με την έρευνα.

Ευχαριστώ το φίλο Δημήτρη Κουβέλη που μου είπε μερικές ωραίες ιδέες για τη διατριβή, τον ξάδερφό μου Γιώργο Λεκέα για κάποιες ωραίες συζητήσεις που κάναμε, αλλά και το φίλο Αριστείδη Δοκουμετζίδη ο οποίος με βοήθησε να εργαστώ για ένα διάστημα αυτών των σπουδών και να εξασφαλίσω αναγκαίους πόρους.

Ευχαριστώ τους τεχνικούς και το helpdesk του ripe ncc, του ntua noc και του κτιρίου κεντρικού Η/Υ του Ε.Μ.Π. για τη βοήθειες που μου παρείχαν. Ευχαριστώ επίσης και τον κύριο Σάσσαλο για τις βιβλιοδεσίες των δεκάδων βιβλίων κατά τη διάρκεια των τελευταίων 8 ετών.

Τέλος ελπίζω να μην ξέχασα κανέναν, αλλά και αν ξέχασα τον ευχαριστώ πολύ για τη βοήθεια που μου έδωσε.

Παρασκευάς Β. Λεκέας

Αθήνα, Δεκέμβριος 2003

Περιεχόμενα

Φιγούρες

Πίνακες

1 ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΜΕΛΕΤΗΣ ΤΟΥ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ

Σύμφωνα με τον Καθηγητή κ. Χρήστο Παπαδημητρίου το Διαδίκτυο (Internet) είναι το πρώτο ανθρώπινο δημιούργημα το οποίο μπορεί να μελετηθεί μόνο με παρατηρήσεις, μετρήσεις και με την ανάπτυξη θεωριών, όπως ακριβώς γίνεται με το σύμπαν, τον ανθρώπινο εγκέφαλο και τα κύτταρα [Pap03]. Το ίδιο φαίνεται να συμβαίνει και με τον Παγκόσμιο Ιστό (web) του οποίου τεχνολογική πλατφόρμα ανάπτυξης είναι το Internet. Στη μελέτη τέτοιων πολύπλοκων συστημάτων μία μέθοδος για παρατηρήσεις και μετρήσεις είναι η δειγματοληψία. Η μέθοδος αυτή εφαρμόζεται και στη μελέτη του web, δημιουργώντας έτσι ένα καινούργιο όρο, αυτόν του web sampling. Το θεωρητικό υπόβαθρο που χρειάζεται να έχει κανείς για να ασχοληθεί με το web sampling συνήθως είναι η θεωρία γράφων και οι πιθανότητες-στατιστική.

Στο κεφάλαιο αυτό περιγράφουμε το πρόβλημα της μελέτης του web γράφου το οποίο βρίσκει μία λύση στη δειγματοληψία.

1.1 Ο Παγκόσμιος Ιστός και το Διαδίκτυο

Η ανάγκη να συνδεθούν μεταξύ τους απομακρυσμένοι υπολογιστές, ώστε οι επιστήμονες να μπορούν να έχουν πρόσβαση σε δεδομένα και προγράμματα που βρίσκονται μακριά τους, οδήγησε στην ανάπτυξη της τεχνολογίας του Ίντερνετ. Σύμφωνα με αυτή κάθε επιμέρους δίκτυο μπορεί να διασυνδεθεί (interconnected) με τα υπόλοιπα, και μάλιστα με τέτοιο τρόπο ώστε να μην αποτελεί συνιστώσα κάποιου άλλου δικτύου αλλά μία ισότιμη (peer) δικτυακή οντότητα. Το πιο πάνω χαρακτηριστικό της τεχνολογίας Ίντερνετ ονομάζεται ανοιχτή αρχιτεκτονική (open architecture) και οδήγησε στη δημιουργία ενός πρωτοκόλλου, του TCP/IP (Transmission Control Protocol/Internet Protocol), το οποίο διέπει τον τρόπο επικοινωνίας των επιμέρους δικτύων.

Το Ίντερνετ μάλλον θα πρέπει να ειδωθεί σαν μία πλατφόρμα πάνω στην οποία προστίθενται νέες τεχνολογίες και προϊόντα, τα οποία άλλες φορές δημιουργούνται από ανάγκη για εξέλιξη υπαρχουσών υπηρεσιών και άλλες φορές για καθαρά εμπορικούς σκοπούς. Έτσι, στην πλατφόρμα του Ίντερνετ έχουν προστεθεί τεχνολογίες όπως τρόποι διαχείρισης της ροής της πληροφορίας (routers), τεχνολογίες διασύνδεσης δικτύων (Ethernet), τρόποι διευθυνσιοδότησης των δικτύων (DNS), ηλεκτρονικό ταχυδρομείο (e-mail). Μία νέα υπηρεσία που χρησιμοποίησε την τεχνολογική πλατφόρμα του Ίντερνετ είναι και το World Wide Web (Παγκόσμιος Ιστός ή www ή αλλιώς web). Την υπηρεσία αυτή την εισήγαγε ο Tim Berners Lee, ο οποίος και θεωρείται ο ιδρυτής του web, στις αρχές της δεκαετίας του ’90. Ο Lee έφτιαξε ένα πρωτόκολλο που το ονόμασε http (hypertext transfer protocol – πρωτόκολλο μεταφοράς υπερκειμένου), το οποίο επέτρεπε στους χρήστες του Ίντερνετ να έχουν μεν πρόσβαση στην πληροφορία αλλά επιπλέον να μπορούν να επικοινωνούν με οπτικό τρόπο (με γραφικά δηλαδή) μέσα από την οθόνη του υπολογιστή τους ανταλλάσσοντας ήχο και εικόνα. Επιπλέον, ήταν δυνατόν με τη χρήση του http (και της γλώσσας που το υποστήριζε, της html – hypertext markup language) εκτός από ανάγνωση της πληροφορίας να μπορεί να κάνει κανείς και διόρθωση ή και διαγραφή της. Με αυτό τον τρόπο έγινε ακόμα πιο εύκολη η πρόσβαση στην πληροφορία, ώστε ο κάθε ένας με το κατάλληλο λογισμικό (Browser)

να μπορεί να δει πληροφορίες σε απομακρυσμένους από αυτόν υπολογιστές μέσω του δικτύου.

Τα πιο πάνω λογισμικά (προγράμματα δηλαδή) παρουσιάζουν στην οθόνη του υπολογιστή "σελίδες" (web pages – ιστοσελίδες) πληροφοριών και σε κάθε σελίδα υπάρχουν σύνδεσμοι (hyperlinks – υπερσύνδεσμοι ή απλά links) τους οποίους αν ακολουθήσει κανείς οδηγείται σε άλλες σελίδες κ.ο.κ. Με αυτό τον τρόπο μπορεί κάποιος χρήστης να πλοηγείται στο web χρησιμοποιώντας κάποια σελίδα σαν αφετηρία και ακολουθώντας τα links που επιθυμεί.

1.2 Προβλήματα που παρουσιάζονται στη μελέτη του web

Η εισαγωγή της υπηρεσίας web στο Internet και η δημιουργία ενός παγκόσμιου ιστού με πληροφορία σε κείμενο, ήχο και εικόνα άνοιξε νέα μέτωπα έρευνας. Ένα από αυτά είναι η μελέτη του παγκόσμιου ιστού και η προσπάθεια κατανόησής του. Πιο κάτω περιγράφουμε κάποια από τα προβλήματα που παρουσιάζονται σε αυτή την προσπάθεια.

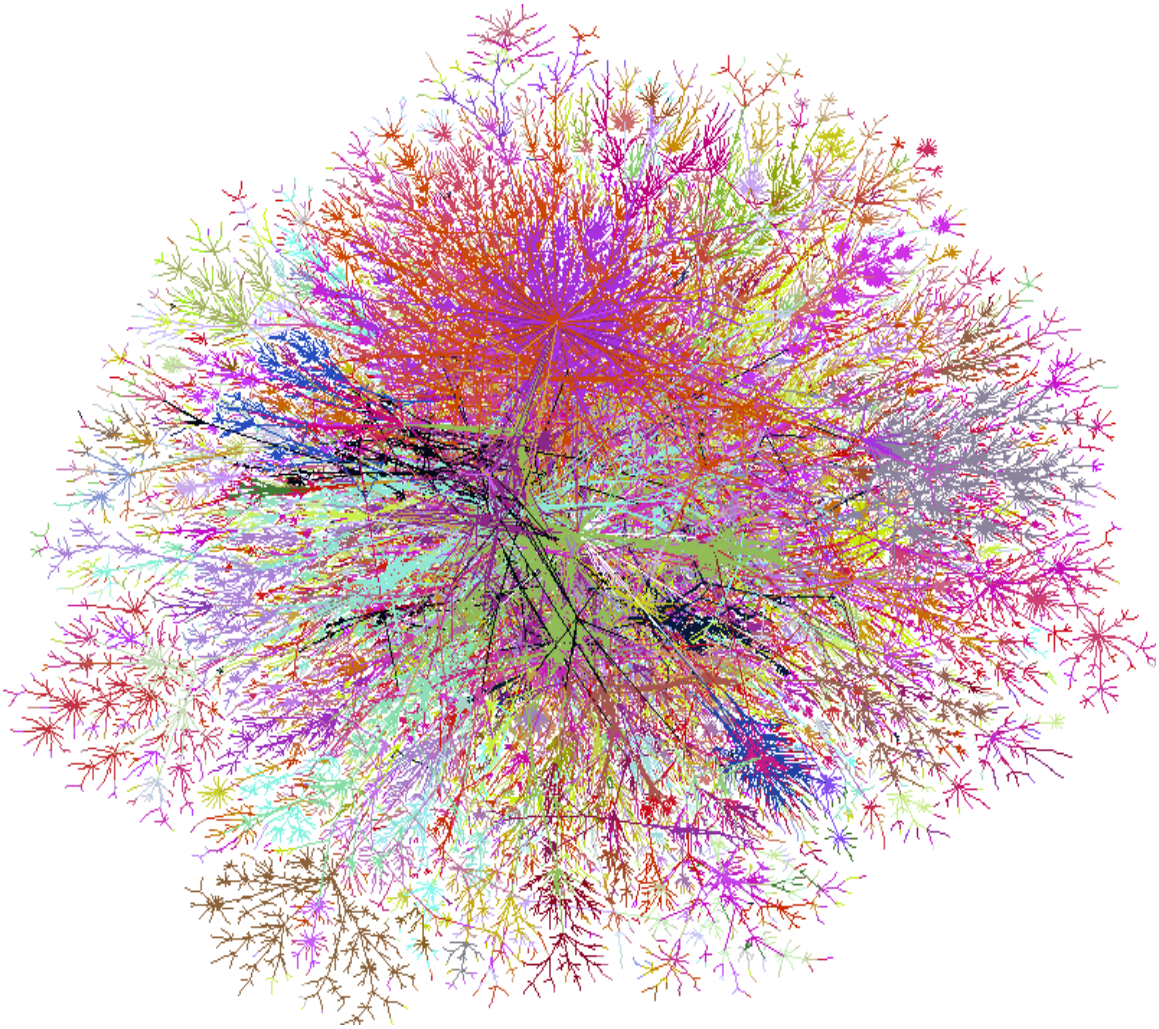
1.2.1 Το web είναι ένα μεγάλο και πολύπλοκο δίκτυο

Στη φιγούρα 1.1 βλέπουμε μία απεικόνιση του Ίντερνετ (σε επίπεδο router) όπως ήταν το 1998, κάτι αντίστοιχο (ίσως και πιο πολύπλοκο) είναι και το δίκτυο του web. Οι γραμμές (ακμές) αναπαριστούν links και οι άκρες των links, που ονομάζονται κόμβοι, είναι οι web σελίδες. Από τη φιγούρα παρατηρούμε την εξαιρετικά πολύπλοκη δικτύωση.

Άλλα τέτοια δίκτυα με πολύπλοκη τοπολογία είναι, εκτός από το Ίντερνετ με κόμβους τους δρομολογητές (routers) και ακμές τις μεταξύ τους φυσικές συνδέσεις, τα *έμβια συστήματα* των οποίων οι κόμβοι είναι οι πρωτεΐνες και τα γονίδια, ενώ οι ακμές τους αναπαριστούν τις χημικές αντιδράσεις μεταξύ τους, το *νευρικό σύστημα* (κόμβοι-νευρικά κελιά, που ενώνονται με άξονες), τα *κοινωνικά δίκτυα* (κόμβοι-άνθρωποι ή οργανισμοί, ακμές-κοινωνικές αλληλεπιδράσεις μεταξύ τους). Δίκτυα όπως τα πιο πάνω περιγράφονται στη βιβλιογραφία σαν μεγάλα δίκτυα με πολύπλοκη τοπολογία (complex topology). Αυτά τα δίκτυα λόγω του μεγάλου μεγέθους τους και της πολυπλοκότητάς τους είναι κατά πολύ ανεξερεύνητα.

Το web ανήκει στην κατηγορία των μεγάλων δικτύων με πολύπλοκη τοπολογία και σύμφωνα με τους [LG99] το 1999 είχε μέγεθος περίπου 800 εκατομμύρια κόμβους, ενώ στην [GL02] αναφέρεται ότι τον Ιούλιο του 2000 περιείχε περίπου 2.1 δισεκατομμύρια κόμβους και το Φεβρουάριο του 2002 περισσότερους από 6 δισεκατομμύρια κόμβους. Στην ίδια εργασία αναφέρεται ότι ο ρυθμός με τον οποίο νέες σελίδες προστίθενται στο web είναι περίπου 7.3 εκατομμύρια κάθε ημέρα (βέβαια αρκετές από τις ήδη υπάρχουσες τροποποιούνται ή και καταργούνται).

Τα δύο πιο πάνω χαρακτηριστικά του web, το μέγεθος και ο μεγάλος ρυθμός ανάπτυξής του, είναι από τα πιο μεγάλα προβλήματα που συναντά κανείς όταν θελήσει να μελετήσει το πολύπλοκο αυτό δίκτυο.



Φιγούρα 1.1: Το δίκτυο του Internet (1998, [CB98])

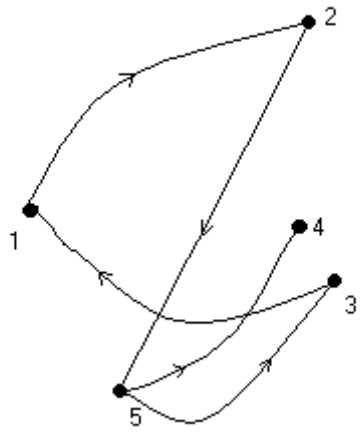
1.2.2 Γράφοι και web

Όπως είδαμε προηγουμένως το web είναι ένα μεγάλο και πολύπλοκο δίκτυο. Στα μαθηματικά την έννοια του δικτύου πολλές φορές τη μελετούμε με μία μαθηματική οντότητα που ονομάζεται γράφος (graph). Ένας γράφος G αποτελείται από κόμβους (nodes) και ακμές (edges). Όταν, επιπλέον, οι ακμές του γράφου έχουν και κατευθύνσεις τότε ο γράφος λέγεται κατευθυνόμενος (directed). Στη φιγούρα 1.2 φαίνεται ένας κατευθυνόμενος γράφος με 5 κόμβους και 5 ακμές.

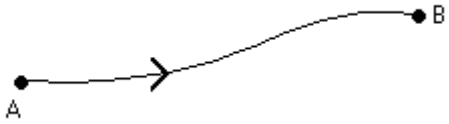
Η έννοια του γράφου στη μελέτη του web παίρνει την αυτονόητη ερμηνεία: οι κόμβοι να είναι οι web σελίδες και οι ακμές να είναι τα links. Αν λοιπόν η σελίδα Α περιέχει ένα link προς τη σελίδα Β τότε αυτό μπορούμε να το απεικονίσουμε στη φιγούρα 1.3. Σε αυτή τη φιγούρα το βέλος δηλώνει ότι η ακμή είναι κατευθυνόμενη που σημαίνει ότι από τη σελίδα Α επιτρέπεται να μεταβώ στη σελίδα Β (όχι όμως το αντίστροφο)[1]. Θα δούμε και πιο κάτω ότι για τη σελίδα Α το συγκεκριμένο link ονομάζεται και out-link (έξω-link) διότι εξέρχεται από την Α, ενώ για την Β

[1] Φυσικά αν η ακμή δεν είναι κατευθυνόμενη επιτρέπεται και η αντίστροφη πορεία.

ονομάζεται in-link (μέσα-link) διότι εισέρχεται σε αυτή. Επίσης, σε ένα κατευθυνόμενο γράφο ορίζουμε το μονοπάτι από το Α στο Β να είναι η διαδοχική σειρά των κατευθυνόμενων ακμών που οδηγούν από τον κόμβο Α στον κόμβο Β. Για παράδειγμα στη φιγούρα 1.2 το μονοπάτι 3,1,2,5,4 οδηγεί από τον κόμβο 3 στον κόμβο 4 μέσω των κατευθυνόμενων ακμών (3,1), (1,2), (2,5), (5,4).



Φιγούρα 1.2: Ένας κατευθυνόμενος γράφος με 5 κόμβους και 5 ακμές



Φιγούρα 1.3: Μία κατευθυνόμενη ακμή

Ένας διαφορετικός τρόπος να απεικονίζουμε γράφους, ο οποίος είναι πολύ χρήσιμος ιδιαίτερα στην αναπαράστασή τους σε υπολογιστές, είναι ο Πίνακας Γειτονικών Κόμβων (Adjacency Matrix) [ΑΠ93]. Ο πίνακας αυτός αποτελείται από 0 και 1, και για ένα γράφο με k κόμβους η διάστασή του είναι $k \times k$. Ένα στοιχείο του πίνακα γειτονικών κόμβων Α, A[i, j] (i, j=1,2,… k) είναι 0 αν δεν υπάρχει ακμή μεταξύ των i και j, ενώ είναι 1 αν υπάρχει. Για παράδειγμα ο πίνακας γειτονικών κόμβων για το γράφο της φιγούρας 1.2 φαίνεται στη φιγούρα 1.4.

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Φιγούρα 1.4: Ο Πίνακας Γειτονικών Κόμβων για το γράφο της φιγούρας 1.2

Επειδή το web είναι ένα μεγάλο και πολύπλοκο δίκτυο το ίδιο συμβαίνει και με το γράφο του. Ο γράφος του web είναι τεράστιος, μεγαλώνει με ραγδαίο ρυθμό και επιπλέον η αποθήκευση αλλά και η αλγοριθμική μεταχείρισή του από υπολογιστές είναι πολύ δύσκολο να επιτευχθεί. Ας φανταστούμε και μόνο πράξεις με πίνακες διαστάσεων $10^9 \times 10^9$ που υπολογίζεται ότι είναι η τάξη μεγέθους του web!

1.2.3 Crawlers και web

Όπως είδαμε, για να μελετήσει κανείς το web πρέπει να μελετήσει ολόκληρο το γράφο του, πράγμα το οποίο είναι αδύνατο, διότι ο γράφος αυτός είναι τρομακτικά μεγάλος και ραγδαία αυξανόμενος[2]. Ακόμα όμως και εάν κατορθώναμε να τον συλλέξουμε, πάλι θα συναντούσαμε πρόβλημα στην αποθήκευση και στην αλγοριθμική μεταχείρισή του.

Αφού, λοιπόν, είναι ακατόρθωτο να συλλέξουμε ολόκληρο το γράφο του web προσπαθούμε να συλλέξουμε μεγάλα κομμάτια του. Η μέθοδος με την οποία συλλέγουμε κομμάτια του web ονομάζεται crawling. Η μέθοδος αυτή υλοποιείται από ειδικά προγράμματα τα οποία έχουν διάφορα ονόματα, όπως crawlers, spiders, robots, harvesters, συνήθως όμως αποκαλούνται robots ή crawlers[3]. Όταν χρησιμοποιούμε ένα τέτοιο πρόγραμμα αυτό που γίνεται είναι να ξεκινούμε τον crawler από κάποια σελίδα του web (από μία αφετηρία δηλαδή) και να τον αφήνουμε να εξερευνήσει το web συλλέγοντας πληροφορίες για κάθε σελίδα που επισκέπτεται. Ένας crawler συνήθως ξεκινά από την αφετηρία και πραγματοποιεί μία BFS αναζήτηση (Breadth First Search – Αναζήτηση κατά Πλάτος (ή κατά επίπεδα)) επισκεπτόμενος τα links της αφετηρίας (1^o επίπεδο). Στη συνέχεια επισκέπτεται όλα τα links των σελίδων που δείχνουν τα link της αφετηρίας (2^o επίπεδο) κ.ο.κ. (βλ. 5.2.1).

Οι crawlers στην προσπάθειά τους να συλλέξουν web pages συναντούν διάφορα προβλήματα. Ένα πρόβλημα είναι τα broken links. Broken links είναι τα σπασμένα-ανενεργά links, αυτά δηλαδή που δείχνουν σε διεύθυνση που δεν υπάρχει πια. Πρόβλημα επίσης δημιουργούν οι κύκλοι και τα αδιέξοδα από τα οποία είναι γεμάτο το web. Άλλο πρόβλημα είναι η σύνδεση με το δίκτυο που πολλές φορές δεν είναι αξιόπιστη. Άλλο πρόβλημα είναι οι servers (εξυπηρετητές) στους οποίους φυλάσσονται οι web σελίδες. Οι servers συνήθως δεν ικανοποιούν αιτήματα πάνω από ένα συγκεκριμένο αριθμό για να αποφύγουν την υπερφόρτωση, πράγμα που σημαίνει ότι ο crawler δεν θα μπορέσει ποτέ να κάνει περίπατο σε όλες τις web σελίδες του συγκεκριμένου server, άρα ένα τμήμα του web θα μείνει ανεξερεύνητο από το συγκεκριμένο crawler.

Οι πιο πάνω λόγοι είναι κάποιοι από τους οποίους σύμφωνα με τους [GL02] είναι δύσκολο να συλλέξει κανείς περισσότερες από 300 εκατομμύρια web pages το μήνα. Και επιπλέον, αφού το web αλλάζει συνέχεια, αυτό που θα έχει συλλέξει κάποιος δεν

[2] Ενδιαφέρον είναι το project "Internet Archive" (http://www.archive.org) στο οποίο αποθηκεύεται ένα μεγάλο κομμάτι του web δίνοντας σε οποιονδήποτε τη δυνατότητα να έχει πρόσβαση σε οποιαδήποτε πληροφορία του web όπως ήταν αποθηκευμένη κάποια στιγμή στο παρελθόν (1996 -).

[3] Η λέξη crawl έχει διάφορες ερμηνείες που σχετίζονται με κίνηση, και η ερμηνεία που ταιριάζει στην περίπτωσή μας είναι "περίπατος". Ένας crawler λοιπόν στα Ελληνικά θα μπορούσε να ονομαστεί "περιπατητής". Σε όλες τις μεταφράσεις των αγγλικών όρων σε αυτή την εργασία χρησιμοποιήθηκαν τα λεξικά [Hor89] και [SH85].

θα είναι πραγματικά ένα κομμάτι όλου του γράφου σε μία δεδομένη χρονική στιγμή αλλά θα μοιάζει σαν τις εικόνες που παίρνουμε από τα ραντάρ (radar scan) για το χρονικό διάστημα του crawling. Βλέπουμε λοιπόν πόσο δύσκολο είναι ακόμα και να συλλέξει κανείς μεγάλα τμήματα του web γράφου.

Ο πιο πάνω περιορισμός των 300 εκατομμυρίων σελίδων το μήνα μας δίνει ένα μέγεθος του τί θεωρούμε αντιπροσωπευτικό κομμάτι του web την εποχή που γράφτηκε αυτή η διατριβή (Δεκ. 2003). Έτσι, οι μελέτες πάνω στο web και το γράφο του περιορίζονται συνήθως σε κομμάτια του που το μέγεθός τους κυμαίνεται από μερικά εκατομμύρια μέχρι μερικές εκατοντάδες εκατομμύρια σελίδες.

Το ότι αποφασίσαμε να μελετάμε τμήματα του γράφου του web αντί για ολόκληρο το γράφο έκανε πιο εύκολο το πρόβλημα, όπως όμως θα δούμε παρακάτω υπάρχουν και άλλες δυσκολίες, οι οποίες αυτή τη φορά σχετίζονται με τη φύση του πολύπλοκου αυτού δικτύου, με τον τρόπο δηλαδή που έχει δημιουργηθεί, και με την τοπολογία του.

1.2.4 Η δομή του web γράφου

Είδαμε ότι το web είναι ένα μεγάλο δίκτυο με πολύπλοκη τοπολογία. Τέτοια δίκτυα περιγράφτηκαν για πρώτη φορά από τη θεωρία των τυχαίων γράφων (random graphs) των Erdős και Renyi (ER). Σύμφωνα με την πιο πάνω θεωρία, αν υποθέσουμε ότι έχουμε N κόμβους και συνδέουμε ζεύγη από αυτούς με πιθανότητα p, τότε η πιθανότητα ένας κόμβος να έχει k ακμές είναι μία κατανομή Poisson $P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ όπου $\lambda = N\binom{N-1}{k}p^k(1-p)^{N-1-k}$.

Η θεωρία ER προτάθηκε το 1959, πολύ πριν δημιουργηθεί το Internet και το web. Τότε δεν υπήρχαν δεδομένα από μεγάλα και πολύπλοκα δίκτυα για να δοκιμαστεί η θεωρία, οπότε η εμφάνιση του web και των crawler έδωσε τη δυνατότητα να συλλεγούν μεγάλα τμήματα του web και να δοκιμαστούν στην πιο πάνω θεωρία.

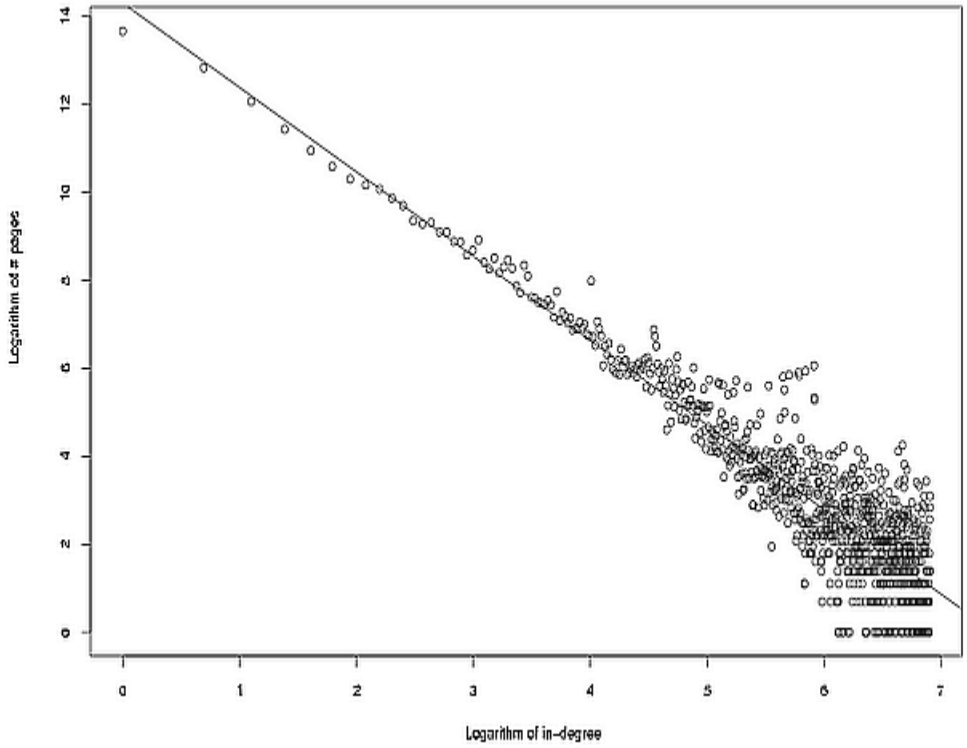
Από τους πρώτους που μελέτησαν ένα "μεγάλο" τμήμα του web ήταν οι Broder et al. [BKM+00]. Αυτοί επαλήθευσαν προηγούμενες μετρήσεις [BA99], [KKR+99] που έλεγαν ότι ο γράφος του web έχει την ιδιότητα: οι συνδέσεις των ακμών του να ακολουθούν κατανομή power law[4].

Μία κατανομή λέμε ότι είναι power law αν για την τυχαία μεταβλητή X ισχύει $\Pr[X = k] \sim \frac{1}{k^\alpha}$, όπου α πραγματικός αριθμός και k ένα διάστημα. Η power law κατανομή λέγεται καμιά φορά και zipfian, όμως οι δύο αυτές κατανομές διαφέρουν. Η μεν zipf είναι αντίστροφη πολυωνυμική της τάξης (rank) ενώ η power law του μεγέθους (magnitude). Οι πιο πάνω κατανομές λέγονται και κατανομές με βαριά ουρά (heavy tail) διότι δεν ελαττώνονται με εκθετικό ρυθμό αλλά με πολυωνυμικό. Αυτές είναι ειδικές περιπτώσεις της κατανομής $\Pr[X > x] = k^\alpha x^{-\alpha} L(x)$, όπου k θετικός

[4] Πρώτοι που παρατήρησαν τέτοιους νόμους στο Internet ήταν οι Faloutsos et al. [FFF99].

πραγματικός και $L(x)$ μία αργά μεταβαλλόμενη συνάρτηση ($\lim_{t\to\infty}\frac{L(tx)}{L(x)}=1$). Για $\Pr[X>x]=k^{\alpha}x^{-\alpha}$ έχουμε την κατανομή Pareto, η οποία είναι η γενική περίπτωση των κατανομών power law και zipf.

Στη φιγούρα 1.5 βλέπουμε την power law κατανομή που ακολουθεί ένα τμήμα του γράφου του web για τον αριθμό των in-link (in-degree).



Φιγούρα 1.5: Κατανομή in-degree για το web [BCSV02]

Ο αριθμός των link που περιέχει μία web σελίδα ονομάζεται out-degree (έξω-βαθμός) αυτής της σελίδας και φανερώνει πόσες κατευθυνόμενες ακμές φεύγουν έξω από αυτήν. Αντίστοιχα ο αριθμός των link που δείχνουν σε μία σελίδα ονομάζεται in-degree (μέσα-βαθμός) και φανερώνει πόσες κατευθυνόμενες ακμές του web γράφου οδηγούν προς αυτήν. Για παράδειγμα στη φιγούρα 1.2 ο κόμβος 5 έχει out-degree = 2 και in-degree = 1, ενώ ο κόμβος 4 έχει out-degree = 0 και in-degree = 1.

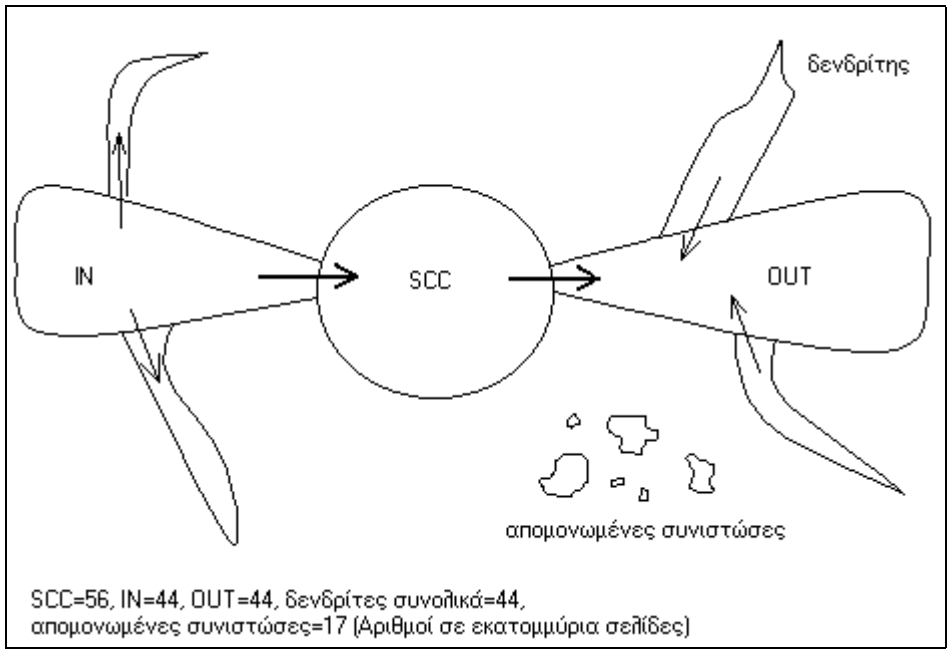
Από τη γραφική παράσταση της φιγούρας 1.5 παρατηρούμε ότι οι πιο πολλές σελίδες έχουν μικρό in-degree, όμως είναι αρκετές σελίδες με μεγάλο in-degree. Το πιο πάνω είναι χαρακτηριστικό των power law κατανομών, διότι αυτές οι κατανομές έχουν μικρή αλλά σημαντική πιθανότητα απόκλισης από τη μέση τιμή. Έτσι, ενώ ο μέσος in-degree του web γράφου είναι γύρω στο 8 υπάρχει σημαντική πιθανότητα μία σελίδα να έχει π.χ. 1000 in-links[5].

Από τα πιο πάνω παρατηρούμε ότι η παραδοσιακή θεωρία των τυχαίων γράφων (ER) δεν εξηγεί το γράφο του web, διότι η ER προβλέπει Poisson κατανομή ενώ στο

[5] Και ο μέσος out-degree είναι γύρω στο 8, αφού κάθε ακμή συνεισφέρει ισοδύναμα στο συνολικό out- και στο συνολικό in-degree.

web παρατηρείται power law κατανομή. Αυτό σημαίνει ότι ο γράφος του web δεν είναι τυχαίος (δεν είναι στατιστικά ομογενής όπως προβλέπει η ER), ή με άλλα λόγια το web δεν είναι μία τυχαία κατασκευή. Εάν ο web γράφος ήταν τυχαίος τότε κάθε κόμβος του θα είχε σχεδόν τον ίδιο αριθμό ακμών και η συχνότητα των υψηλά συνεκτικών κόμβων θα μειωνόταν εκθετικά. Αντίθετα, το web είναι ένα δίκτυο χωρίς κλίμακα (scale free) που σημαίνει ότι δεν είναι ομογενές ή αλλιώς, όπως λέμε, έχει fractal δομή. Αυτή η συμπεριφορά του web συναντιέται και σε άλλα φυσικά δίκτυα όπως στους νευρώνες του εγκεφάλου και στα οικονομικοκοινωνικά δίκτυα, και είναι η αιτία που κάνει δύσκολη τη συλλογή μεγάλων κομματιών του. Ας δούμε γιατί γίνεται αυτό.

Οι Broder et al. [BKM+00] εκτός από τους power laws παρατήρησαν ακόμα ότι το 90% του web[6] σχηματίζει μία συνεκτική συνιστώσα[7] η οποία χωρίζεται σε 4 κομμάτια. Το πρώτο κομμάτι είναι ο κεντρικός πυρήνας του web. Σε αυτόν όλες οι σελίδες μπορούν να προσεγγιστούν η μία με την άλλη (είτε απευθείας είτε μέσω τρίτων σελίδων που ανήκουν στον πυρήνα). Αυτό το κομμάτι ονομάζεται γιγαντιαία ισχυρά συνεκτική συνιστώσα (SCC – Strongly Connected Component). Το δεύτερο και το τρίτο κομμάτι ονομάζονται IN και OUT. Το IN αποτελείται από σελίδες που μπορούν να προσεγγίσουν την SCC αλλά όχι το αντίστροφο, δηλ. σελίδες από την SCC δε μπορούν να φτάσουν σελίδες του IN. Το OUT αποτελείται από σελίδες που προσεγγίζονται από την SCC χωρίς όμως να συμβαίνει το αντίστροφο. Το τέταρτο κομμάτι αποτελείται από τους TENDRILLS (δενδρίτες) και περιέχει σελίδες που είναι απομονωμένες, δηλ. ούτε μπορούν να προσεγγίσουν την SCC αλλά ούτε και να προσεγγιστούν από αυτήν. Το υπόλοιπο τμήμα του web αποτελείται από άλλες συνιστώσες που δε συνδέονται μεταξύ τους και υπάρχουν απομονωμένες (disconnected components). Στη φιγούρα 1.6 βλέπουμε τη μορφή του web και τα κομμάτια που το αποτελούν.



Φιγούρα 1.6: Μακροσκοπική δομή του web γράφου

[6] Οι παρατηρήσεις τους έγιναν σε τμήμα του web και μετά γενίκευσαν τα αποτελέσματα για ολόκληρο το web.

[7] Συνεκτική συνιστώσα ενός κατευθυνόμενου γράφου είναι το τμήμα εκείνο του γράφου στο οποίο για κάθε κόμβο του υπάρχει και ένα κατευθυνόμενο μονοπάτι προς οποιοδήποτε άλλο κόμβο του.

Η πιο πάνω φιγούρα έχει τη μορφή του παπιγιόν και γι' αυτό στη βιβλιογραφία ονομάζεται και bow-tie. Οι Dill et al. [DKM+02] δείχνουν ότι το web στην πραγματικότητα περιέχει πολλά τέτοια bow-ties αν χωρίσουμε το γράφο σε υπογράφους ως προς σελίδες με ίδιο περιεχόμενο, ή ως προς σελίδες που βρίσκονται στον ίδιο server ή ως προς σελίδες που βρίσκονται στην ίδια γεωγραφική περιοχή (συμπεριφορά δηλαδή καθαρά fractal).

Όπως περιγράφουν οι Broder et al. [BKM+00] στην εργασία τους: "…κατά μία έννοια το web είναι σαν ένας πολύπλοκος οργανισμός στον οποίο η τοπική δομή σε μικροσκοπική κλίμακα μοιάζει με τα βιολογικά κύτταρα, αλλά η όλη του εικόνα φανερώνει ενδιαφέροντα μορφολογικά στοιχεία (σώμα, άκρα) τα οποία δεν είναι προφανή σε μικροσκοπικό επίπεδο (τοπικά δηλαδή). Γι' αυτό ενώ είναι ελκυστικό να εξάγει κανείς συμπεράσματα για τη δομή του web γράφου έχοντας μελετήσει μία τοπική του εικόνα, τέτοια συμπεράσματα μπορεί να είναι παραπλανητικά". Η προηγούμενη παρατήρηση μας λέει ότι ένας crawler όσο και να εργαστεί και όσο "μεγάλο" τμήμα του web και να "περπατήσει" θα έχει εξερευνήσει ένα τοπικό κομμάτι του web από το οποίο τα συμπεράσματα που μπορούν να βγουν όταν γενικευτούν για όλο το γράφο ίσως και να μην είναι σωστά. Επιπλέον, από τη φιγούρα 1.6 παρατηρούμε ότι αναλόγως την περιοχή που θα ξεκινήσουμε τον crawler θα έχουμε και το αντίστοιχο αποτέλεσμα. Για παράδειγμα αν η αφετηρία του βρίσκεται στις μη συνεκτικές συνιστώσες τότε ένα πολύ μικρό μέρος του web θα εξερευνήσουμε. Ακόμη, αν η αφετηρία βρίσκεται στο OUT τμήμα τότε μόνο ένα μέρος του OUT θα μπορούμε να εξερευνήσουμε (αφού από το OUT δε μπορούμε να προσεγγίσουμε το SCC) κ.ο.κ. Επιπλέον, δεν είμαστε σε θέση πάντα να γνωρίζουμε σε πιο κομμάτι ανήκει η αφετηρία από την οποία ξεκινούμε τον crawler.

Βλέπουμε, λοιπόν, ότι η φύση του web και η τοπολογία του μας δυσκολεύει και να συλλέξουμε ένα μεγάλο τμήμα του web αλλά και να βγάλουμε συμπεράσματα για όλο το web από το τμήμα που συλλέξαμε.

1.3 Η Δειγματοληψία στο web

Είδαμε ότι η μελέτη ολόκληρου ή ακόμα και μεγάλου μέρους του web και του γράφου του είναι δύσκολη και προβληματική. Μία λύση σε αυτό το πρόβλημα είναι, αντί να κάνουμε εξαντλητική συλλογή μεγάλων τμημάτων του web, να αναπτύξουμε τεχνικές δειγματοληψίας με τις οποίες θα μπορούμε να παίρνουμε αντιπροσωπευτικά δείγματά του. Τα δείγματα αυτά θα αποτελούνται από web σελίδες και θέλουμε να μπορούν εύκολα να αναπαραχθούν από οποιονδήποτε επιθυμεί να επαναλάβει τη δειγματοληψία (πράγμα το οποίο είναι δύσκολο να συμβεί στην περίπτωση που μαζεύουμε μεγάλα τμήματα του web με crawlers). Επιπλέον, θέλουμε τα δείγματα αυτά να είναι ομοιόμορφα, δηλαδή κάθε web σελίδα να έχει την ίδια πιθανότητα να συμμετέχει στο δείγμα. Δυστυχώς όμως δεν έχει ανακαλυφθεί μέχρι στιγμής καμία τέτοια τεχνική δειγματοληψίας, διότι παραμένει ανοιχτό πρόβλημα το πώς μπορεί να επιλέξει κάποιος μία τυχαία (uniformly at random) σελίδα στο web. Αν υπήρχε τέτοια μέθοδος τότε επαναλαμβάνοντάς την θα μπορούσαμε να δημιουργήσουμε ένα ομοιόμορφο δείγμα από web σελίδες και μάλιστα όσο μεγάλο θέλουμε.

Οι μελετητές του web έχουν ασχοληθεί πάρα πολύ τα τελευταία χρόνια με το πρόβλημα της ομοιόμορφης δειγματοληψίας του web γράφου [LG99], [HHMN00], [BBC+00], [RPLG01] και σε γενικές γραμμές υπάρχουν δύο προσεγγίσεις. Η πρώτη

προσέγγιση είναι αυτή του τυχαίου περιπάτου[8] (random walk). Σύμφωνα με αυτήν εκτελούμε τυχαίο περίπατο στο γράφο του web και όταν ο περίπατος φθάσει σε κατανομή ισορροπίας από εκεί και πέρα η δειγματοληψία γίνεται με αυτή την κατανομή. Η αποτελεσματικότητα αυτής της τεχνικής περιορίζεται, κυρίως από την ιδιαιτερότητα και την πολύπλοκη δομή του web γράφου.

Σύμφωνα με τη δεύτερη προσέγγιση[9] (IP Sampling) επιλέγονται στην τύχη IP διευθύνσεις και ελέγχονται εάν αυτές φιλοξενούν / εξυπηρετούν (host) ένα web site. Web site είναι μία συλλογή από web σελίδες οι οποίες δικτυώνονται με τέτοιο τρόπο ώστε να αποτελούν ομάδα με συγκεκριμένο πληροφοριακό αντικείμενο [wcp]. Η αρχική σελίδα ενός website, αυτή δηλαδή γύρω από την οποία οργανώνονται οι θεματικές ενότητες του αντικειμένου, συνήθως ονομάζεται home page και πολλές φορές η πρόσβαση σε αυτή είναι ένας εύκολος τρόπος για να πλοηγηθεί κανείς στο web site. Εάν ο host για τη συγκεκριμένη IP διεύθυνση που ελέγχεται εξυπηρετεί κάποιο web site, τότε η τεχνική προσπαθεί να πάρει ομοιόμορφα δείγματα σελίδων από το συγκεκριμένο web site. Ο πιο σημαντικός περιορισμός αυτής της μεθόδου είναι το ότι δεν έχει βρεθεί ικανοποιητικός τρόπος για ομοιόμορφη δειγματοληψία web σελίδων από ένα web site εάν κάποιος δεν διαθέτει μία αναλυτική λίστα με τις web σελίδες που περιέχονται σε αυτό. Για να δημιουργηθεί μία τέτοια λίστα χρειάζεται να διατρέξουμε όλο το web site για να βρούμε όλες τις web σελίδες που περιέχει. Κάτι τέτοιο όμως θα μας οδηγούσε ξανά στο πρόβλημα με τους crawlers, για το οποίο έγινε λόγος πιο πριν. Γι' αυτό το λόγο θέλουμε να κάνουμε δειγματοληψία χωρίς την ύπαρξη μίας τέτοιας λίστας.

Σε αυτή την εργασία ασχολούμαστε με τη δεύτερη προσέγγιση (IP Sampling) και μελετούμε πώς είναι δυνατόν να εφαρμοστεί η μέθοδος για συγκεκριμένα υποσύνολα του Internet. Τις δύο πιο πάνω μεθόδους θα τις δούμε αναλυτικότερα στο 2^{ο} κεφάλαιο στο οποίο θα περιγράψουμε τα βασικά χαρακτηριστικά τους μαζί με μία αναδρομή στη βιβλιογραφία.

1.4 Επίλογος

Σε αυτό το κεφάλαιο είδαμε ότι το πρόβλημα της μελέτης του web και του γράφου του είναι δύσκολο για τους εξής λόγους:

- Ο web γράφος έχει πολύ μεγάλο μέγεθος με αποτέλεσμα να υπάρχει δυσκολία στην αποθήκευση και αλγοριθμική μεταχείρισή του από Η/Υ.

- Ο υψηλός ρυθμός ανάπτυξης του web έχει σαν αποτέλεσμα κατά τη διάρκεια της συλλογής των δεδομένων τα ίδια τα δεδομένα να αλλάζουν.

- Τα προγράμματα συλλογής (crawlers) του web γράφου συναντούν δυσκολίες από το ίδιο το δίκτυο και τις υπηρεσίες του.

- Η τοπολογία του web είναι τέτοια ώστε να επιτρέπει στους crawlers μόνο τοπική εξερεύνησή του.

[8] Βλέπε κεφ. 2
[9] Βλέπε κεφ. 2, 3

Μία προσέγγιση στο πιο πάνω πρόβλημα είναι αυτή της δειγματοληψίας. Υπάρχουν 2 μέθοδοι για δειγματοληψία στο web:

- Random Walks. Σε αυτή τη μέθοδο εκτελούμε τυχαίους περιπάτους στο web και από τις σελίδες που επισκεπτόμαστε εκλέγουμε το δείγμα.

- IP Sampling. Σε αυτή τη μέθοδο δοκιμάζουμε τυχαία IP και ελέγχουμε αν βρίσκονται web sites πίσω από αυτά.

Οι δύο πιο πάνω μέθοδοι δε λύνουν τελείως το πρόβλημα της δειγματοληψίας των web σελίδων, διότι για την πρώτη η άγνωστη φύση και η πολύπλοκη δικτύωση του web δρουν περιοριστικά, ενώ για τη δεύτερη δεν έχει βρεθεί τρόπος ομοιόμορφης δειγματοληψίας για web site. Το πρόβλημα, λοιπόν, της δειγματοληψίας web σελίδων παραμένει ακόμα ανοιχτό και μία λύση του σύμφωνα με την Henzinger [Hen03] θα ήταν "…να βρεθεί μία τεχνική η οποία αποδεδειγμένα να παράγει ομοιόμορφα τυχαία δείγματα web σελίδων και επιπλέον να δουλεύει και στην πράξη".

2 ΜΕΘΟΔΟΙ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΤΟΥ WEB

Στο προηγούμενο κεφάλαιο είδαμε ότι ένας τρόπος για να μελετήσουμε το web και το γράφο του είναι να χρησιμοποιήσουμε τεχνικές δειγματοληψίας (web sampling). Είπαμε, επίσης, ότι υπάρχουν 2 βασικές μέθοδοι web sampling, οι random walks και το IP sampling. Σε αυτό το κεφάλαιο θα δούμε πώς λειτουργούν αυτές οι τεχνικές κάνοντας μία αναδρομή στη σχετική βιβλιογραφία. Επίσης, θα δούμε ποιά είναι τα πλεονεκτήματα και τα μειονεκτήματά τους.

2.1 Μέθοδος δειγματοληψίας με random walks

2.1.1 Μαρκοβιανές Αλυσίδες[10]

Ας υποθέσουμε ότι έχουμε ένα σύνολο από S καταστάσεις και ένα πίνακα πιθανοτήτων P ανάμεσα σε αυτές τις καταστάσεις. Ο πίνακας P έχει διαστάσεις $|S|\times|S|$, όπου $|S|$ είναι ο αριθμός των καταστάσεων. Μία μαρκοβιανή αλυσίδα πάνω στα S, P είναι μία στοχαστική διαδικασία σύμφωνα με την οποία αν X_t είναι η κατάστασή της τη χρονική στιγμή t, τότε η πιθανότητα μετάβασης από την κατάσταση i (τη χρονική στιγμή t) στην κατάσταση j (τη χρονική στιγμή $t+1$), $(i, j \in S)$ να ορίζεται σαν το στοιχείο P_{ij} του πίνακα P. Το P_{ij} δηλαδή είναι η δεσμευμένη πιθανότητα η αλυσίδα από την κατάσταση i τη χρονική στιγμή t $(X_t = i)$ να πάει στην κατάσταση j τη χρονική στιγμή $t+1$ $(X_{t+1} = j)$. Το τελευταίο μπορεί να γραφτεί και σαν $P_{ij} = \Pr[X_{t+1} = j \mid X_t = i]$.

Κάθε φορά η επόμενη κατάσταση της αλυσίδας εξαρτάται μόνο από την κατάσταση στην οποία η αλυσίδα βρίσκεται (παρούσα κατάσταση) και όχι από τις προηγούμενες. Αυτός είναι ο λόγος που καμιά φορά λέμε ότι η αλυσίδα ξεχνά το παρελθόν της. Λαμβάνοντας αυτό υπ' όψιν, αν γνωρίζουμε τον πίνακα P και την κατανομή της αρχικής κατάστασης $X_0 = i$, τότε είμαστε σε θέση να γνωρίζουμε όλη την ακολουθία των καταστάσεων πάνω στις οποίες θα πάει η αλυσίδα. Έστω $q_i(t)$ η πιθανότητα η αλυσίδα να βρίσκεται τη χρονική στιγμή t στην κατάσταση i $(\forall i)$. Τότε, αν το πλήθος όλων των καταστάσεων είναι n, ορίζουμε ένα διάνυσμα $q^{(t)} = (q_1^{(t)}, q_2^{(t)}, ..., q_n^{(t)})$ το οποίο το ονομάζουμε διάνυσμα πιθανοτήτων κατάστασης. Θα λέμε ότι ένα διάνυσμα πιθανοτήτων κατάστασης π είναι μία ευσταθής κατανομή (stationary distribution) για τη μαρκοβιανή αλυσίδα αν, $\pi = \pi P$ (πολλαπλασιασμός πινάκων).

Όμως $\forall i, j$

$$q_i^{(t)} P_{ij} = \Pr[X_t = i]\Pr[X_{t+1} = j \mid X_t = i] = \Pr[X_t = i]\frac{\Pr[X_{t+1} = j]\Pr[X_t = i]}{\Pr[X_t = i]} = q_j^{(t+1)}$$

δηλαδή $q^{(t+1)} = q^{(t)} P$. Αυτό μας επιτρέπει να πούμε ότι για την ευσταθή κατανομή π της μαρκοβιανής αλυσίδας θα ισχύει ότι $\pi^{(t+1)} = \pi^{(t)}$. Αυτό σημαίνει ότι αν μία μαρκοβιανή αλυσίδα βρίσκεται στην ευσταθή κατανομή τη χρονική στιγμή t, τότε θα παραμένει στην ίδια κατανομή και για όλες τις υπόλοιπες χρονικές στιγμές.

[10] [ΝΣ96]

2.1.2 Random walks και Μαρκοβιανές Αλυσίδες

Ας υποθέσουμε τώρα ότι έχουμε ένα συνεκτικό γράφο[11] $G=(V,E)$, όπου με V συμβολίζουμε το σύνολο των κορυφών του γράφου και με E το σύνολο των ακμών του. Έστω ότι ο G έχει n κόμβους ($|V|=n$) και m ακμές ($|E|=m$). Ένας random walk (τυχαίος περίπατος / βηματισμός) στο γράφο G σε κάθε βήμα του επιλέγει τυχαία και ισοπίθανα τον επόμενο κόμβο που θα επισκεφθεί ανάμεσα από τους γείτονες[12] του κόμβου που ήδη βρίσκεται. Έστω λοιπόν ένας random walk στον G με αφετηρία τον κόμβο ν_0. Κάποια στιγμή μετά από "χρόνο" t έστω ότι ο random walk βρίσκεται στον κόμβο ν_t (ο χρόνος στον random walk σημαίνει πόσα βήματα έχουμε κάνει, π.χ. χρόνος 0 σημαίνει ότι βρισκόμαστε στον ν_0 δηλ. κανένα βήμα, χρόνος 10 σημαίνει ότι κάναμε 10 βήματα και βρισκόμαστε στον ν_{10} κ.λπ.). Αν ο βαθμός[13] του κόμβου ν_t είναι $d(\nu_t)$, τότε κάθε γείτονας του ν_t έχει ίση πιθανότητα $\frac{1}{d(\nu_t)}$ να είναι ο επόμενος κόμβος του περιπάτου. Οι κόμβοι που έχει επισκεφθεί ο random walk μπορούν να γραφούν σαν μία ακολουθία $\nu_0,\nu_1,\nu_2,\nu_3,...$ η οποία ακολουθία είναι μία μαρκοβιανή αλυσίδα με πίνακα μετάβασης P και $P_{ij}=\begin{cases}\frac{1}{d(\nu_i)}, \text{αν (i, j) ακμή στο E}\\ 0, \text{αλλιώς}\end{cases}$. Ένας τέτοιος random walk λέγεται και μαρκοβιανός διότι κάθε του βήμα είναι ανεξάρτητο από τα προηγούμενα και εξαρτάται μόνο από την παρούσα κατάσταση.

Σύμφωνα με όσα είπαμε για τις μαρκοβιανές αλυσίδες, ο random walk που εξετάζουμε κάποτε θα φτάσει[14] σε μία κατανομή ισορροπίας π. Αυτή η κατανομή ισορροπίας καθορίζει (για κάθε κόμβο) το ποσοστό των βημάτων που ο random walk ξόδεψε σε κάθε κόμβο. Επίσης, κατά ένα άλλο τρόπο η κατανομή ισορροπίας μας δίνει την πιθανότητα να βρούμε τον random walk σε ένα ορισμένο κόμβο μετά από άπειρο αριθμό βημάτων.

2.1.3 Random walks και δειγματοληψία στο web

Αν στην προηγούμενη ενότητα 2.1.2 υποθέσουμε ότι ο γράφος G είναι ο γράφος του web και τα σύνολα V και E είναι αντίστοιχα οι web σελίδες και τα links, τότε αυτόματα ο random walk εκτελείται στο web. Αυτός ο random walk μπορεί να μας βοηθήσει να πάρουμε δείγματα του web ως εξής: Αφού, όπως είπαμε, ο random walk συγκλίνει στην κατανομή ισορροπίας από εκεί και πέρα όποιες σελίδες επισκέπτεται θα τις επισκέπτεται με την κατανομή αυτή. Αυτό είναι μία πολύ καλή ιδιότητα, διότι έτσι μπορούμε να επισκεπτόμαστε σελίδες του web βάσει μίας συγκεκριμένης κατανομής, πράγμα που σημαίνει ότι μπορούμε να κάνουμε δειγματοληψία πάνω στο

[11] Ένας γράφος (κατευθυνόμενος ή μη) ονομάζεται συνεκτικός (connected) όταν για οποιοδήποτε ζευγάρι {x, y} διαφορετικών κόμβων του υπάρχει μονοπάτι από το x στο y.
[12] Οι γείτονες ενός κόμβου είναι όσοι κόμβοι συνδέονται με ακμές μαζί του.
[13] Ο βαθμός ενός κόμβου είναι ο αριθμός των γειτονικών του κόμβων.
[14] Θεωρούμε ότι η μαρκοβιανή αλυσίδα που εξετάζουμε συγκλίνει σε κατανομή ισορροπίας. Υπάρχουν και μαρκοβιανές αλυσίδες που δε συγκλίνουν σε κατανομή ισορροπίας (αυτές ονομάζονται περιοδικές μαρκοβιανές αλυσίδες).

web βάσει αυτής της κατανομής. Αν π.χ. η κατανομή ισορροπίας είναι ομοιόμορφη, τότε ο random walk θα κάνει ομοιόμορφη δειγματοληψία στο web.

2.2 Εργασίες σχετικά με random walks και web sampling

Στις υποενότητες που ακολουθούν παρουσιάζουμε κάποιες αντιπροσωπευτικές προσπάθειες χρήσης random walk για δειγματοληψία στο web που συναντούμε στη διεθνή βιβλιογραφία.

2.2.1 Μέτρηση της ποιότητας των web καταλόγων με random walks

Στην εργασία [HHMN99] χρησιμοποιούνται random walks για να μετρηθεί η ποιότητα των καταλόγων (index) που υπάρχουν στο web. Οι κατάλογοι είναι μεγάλες συλλογές από web pages, τις οποίες τις διαχειρίζονται μηχανές αναζήτησης (search engines). Οι μηχανές αναζήτησης κάθε μέρα εξερευνούν με crawlers το web και αποθηκεύουν τις σελίδες που εξερεύνησαν σε καταλόγους. Ένας από τους βασικούς στόχους πολλών μηχανών αναζήτησης είναι να καταλογοποιήσουν όσο το δυνατόν μεγαλύτερο τμήμα του web, ώστε στη συνέχεια επεξεργαζόμενες τους καταλόγους να μπορούν να απαντούν σε ερωτήματα που τους απευθύνουν οι χρήστες. Γι' αυτό το λόγο παλαιότερα το μέγεθος του καταλόγου που διέθετε μία μηχανή αναζήτησης αποτελούσε και το πρωταρχικό κριτήριο πληρότητας περιεχομένου για αυτήν. Όμως, τα τελευταία χρόνια με τη συνεχή εξέλιξη των αλγορίθμων αναζήτησης (αλγόριθμος PageRank) κατανοήθηκε η αξία της ποιότητας των απαντήσεων που αυτές επιστρέφουν και αμφισβητήθηκε αρκετά ο όγκος των καταλόγων τους. Σε αυτή την εργασία αναδεικνύεται το πιο πάνω γεγονός με τη χρήση random walks.

Έστω ότι σε κάθε web σελίδα p αντιστοιχεί και κάποιο βάρος (αξία) $w(p)$, έτσι ώστε $\sum_p w(p) = 1$. Αν S είναι ένα σύνολο από σελίδες (ο κατάλογος), τότε ορίζεται η ποιότητα ή το βάρος του S σαν το άθροισμα των βαρών των σελίδων που περιέχονται στο S, δηλ. $w(S) = \sum_{p \in S} w(p)$ $(0 \le w(S) \le 1)$.

Το βάρος που ανατέθηκε σε κάθε σελίδα είναι μία συνάρτηση η οποία κατά κάποιο τρόπο θα δηλώνει την αξία της σελίδας. Δηλαδή, αν $w(p_1) > w(p_2)$ τότε σύμφωνα με τη συνάρτηση βάρους που διαλέξαμε θα λέμε ότι η αξία ή η ποιότητα της σελίδας p_1 είναι μεγαλύτερη από αυτήν της p_2. Έστω τώρα ότι έχουμε δύο μηχανές αναζήτησης οι οποίες χρησιμοποιούν τους καταλόγους S_1 και S_2. Αν υποθέσουμε ότι ο κατάλογος της πρώτης περιέχεται στον κατάλογο της δεύτερης ($S_1 \subseteq S_2$) τότε το βάρος του S_2 θα είναι μεγαλύτερο ή ίσο από το αντίστοιχο του S_1. Μπορεί, λοιπόν, να οριστεί και η μέση ποιότητα σελίδων ενός καταλόγου $a(S)$ σαν την ποσότητα $a(S) = \frac{w(S)}{|S|}$ ($|S|$ είναι το μέγεθος του καταλόγου).

Για να οριστεί η ποιότητα ενός καταλόγου θα πρέπει να έχουμε ορίσει και μία συνάρτηση για την ποιότητα των web σελίδων. Αυτό γίνεται με τον ακόλουθο τρόπο. Αν έχουμε δύο σελίδες p_1, p_2 και η p_1 έχει ένα link προς την p_2, τότε αυτό

σημαίνει ότι η p_1 προτείνει την p_2, δηλαδή με άλλα λόγια ένας τρόπος μέτρησης της ποιότητας της p_2 είναι να μετρήσουμε τον in-degree της. Η πιο πάνω ιδέα είχε προταθεί από τους Carriere et al. [CK97] και μία πολύ δημοφιλής παραλλαγή της είναι ο αλγόριθμος PageRank [BP98] στον οποίο βασίζεται η λειτουργία της μηχανής αναζήτησης Google[15]. Σύμφωνα με τον πιο πάνω αλγόριθμο το PageRank μίας σελίδας (η ποιότητα μίας σελίδας) είναι υψηλό εάν προτείνεται από άλλες σελίδες που έχουν και αυτές υψηλό PageRank. Επιπλέον, μία σελίδα με λίγα links συνεισφέρει πιο πολύ βάρος στις σελίδες που προτείνει παρά μία σελίδα με πολλά links.

Έστω λοιπόν ότι έχουμε συνολικά T σελίδες στο web και έστω ότι οι σελίδες $p_1,...,p_k$ δείχνουν προς την p. Έστω τώρα ότι ο out-degree της p είναι $C(p)$, τότε ο PageRank της σελίδας ορίζεται να είναι ο $R(p) = d\frac{1}{T} + (1-d)\sum_{i=1}^{h}\frac{R(p_i)}{C(p)}$ (όπου $\sum_{p} R(p) = 1$ και d μία παράμετρος, $0 < d < 1$). Ο PageRank μπορεί να ειδωθεί σαν μία συνάρτηση βάρους για τις σελίδες.

Ο random walk που χρησιμοποιείται σε αυτή την εργασία κινείται ανάμεσα στις web σελίδες ακολουθώντας τα links. Όταν ο random walk πάει από μία σελίδα σε άλλη αυτό θεωρείται ένα βήμα. Επειδή όμως υπάρχει κίνδυνος ο random walk να εγκλωβιστεί σε κάποιο κύκλο ή σε κάποιο αδιέξοδο[16], περιστασιακά πηδά σε μία τυχαία σελίδα του web. Αν λοιπόν ο random walk βρίσκεται στη σελίδα p, τότε με πιθανότητα d πηδά σε μία τυχαία σελίδα του web ή με πιθανότητα $1-d$ διαλέγει ένα νέο link από αυτά που δείχνει η παρούσα σελίδα. Ο πιο πάνω random walk έχει κατανομή ισορροπίας την $R(p)$.

Η πιο πάνω ανάλυση μας λέει ότι θεωρητικά μπορούμε να συλλέγουμε web σελίδες με την κατανομή του PageRank τους. Όπως όμως είχαμε αναφέρει και στο πρώτο κεφάλαιο, δεν υπάρχει μία μέθοδος με την οποία να μπορούμε να επιλέγουμε web σελίδες με τυχαίο και ομοιόμορφο τρόπο. Αν μπορούσε να γίνει αυτό τότε όταν θα ερχόταν η ώρα ο random walk να επιλέξει μία τυχαία σελίδα του web θα μπορούσε να γίνει αυτή η επιλογή με τυχαίο και ομοιόμορφο τρόπο. Επαναλαμβάνοντας την πιο πάνω διαδικασία πολλές φορές ο random walk τελικά θα έπαιρνε ένα δείγμα από web σελίδες οι οποίες θα ήταν κατανεμημένες σε αυτό σύμφωνα με το PageRank τους.

Ένα επιπλέον πρόβλημα χρησιμοποιώντας την πιο πάνω μέθοδο είναι ότι η αφετηρία του random walk εισάγει μία μεροληψία υπέρ των σελίδων που είναι ισχυρά συνδεδεμένες (strongly connected) με αυτήν. Η μεροληψία αυτή και ο έλεγχός της δεν είναι γνωστό πώς μπορούν να αντιμετωπιστούν (π.χ. δεν είναι γνωστό πόσα βήματα πρέπει να κάνει ο random walk για να εξαλείψει την πιο πάνω μεροληψία). Η θεωρία, ωστόσο, των random walk μας ενημερώνει ότι αρκεί κάποιος να κάνει random walk σε ένα μικρό αλλά υψηλά συνεκτικό υπογράφο του web για να χειριστεί το πρόβλημα της μεροληψίας της αρχικής σελίδας (υπάρχουν γνωστά όρια για το

[15] http://www.google.com

[16] Το web είναι γεμάτο από αδιέξοδα και κύκλους. Για παράδειγμα, στη φιγούρα 1.2 η διαδρομή 5,3,1,2,5 είναι κύκλος ενώ ο κόμβος 4 είναι αδιέξοδο.

πόσα βήματα χρειάζεται ένας random walk σε υψηλά συνεκτικό γράφο προκειμένου να φτάσει στην κατανομή ισορροπίας). Αν, για παράδειγμα, ο γράφος του web ήταν πλήρης (δηλ. αν κάθε σελίδα του web συνδεόταν με οποιαδήποτε άλλη), τότε μόνο 2 βήματα θα ήταν αρκετά για να αφαιρεθεί σχεδόν όλη η μεροληψία της αφετηρίας. Όπως, όμως, είδαμε στο κεφάλαιο 1 ο γράφος του web κάθε άλλο παρά πλήρης είναι. Οπότε είτε θα πρέπει να περιορίζουμε τον random walk στην SCC του web γράφου ή θα πρέπει να βρούμε έναν άλλο τρόπο να λύσουμε το πρόβλημα της μεροληψίας.

Βλέπουμε ότι η μεροληψία της αφετηρίας και η τυχαία επιλογή μίας σελίδας του web είναι τα δύο πιο σημαντικά προβλήματα της πιο πάνω τεχνικής. Μία προσεγγιστική λύση θα ήταν ο random walk αντί να πηδά με πιθανότητα d σε μία καινούργια σελίδα του web να πηδά σε μία σελίδα από αυτές που έχει ήδη επισκεφθεί, πράγμα το οποίο δεν είναι αποτελεσματικό, όπως έδειξαν τα πειραματικά αποτελέσματα. Αντί αυτού επιλέγεται τυχαία ένας host από αυτούς που ήδη έχει επισκεφθεί ο random walk και στη συνέχεια επιλέγεται τυχαία μία από τις ανακαλυφθείσες σελίδες του.

Τελικά, με την πιο πάνω τεχνική επιτυγχάνεται ο random walk να προσεγγίσει την κατανομή PageRank. Έχοντας έτσι δείγματα σύμφωνα με τη συγκεκριμένη κατανομή μετριέται η ποιότητα διαφόρων καταλόγων που εμφανίζονται στο web από την εκτίμηση της μέσης ποιότητας του $a(S)$.

2.2.2 Σχεδόν ομοιόμορφη δειγματοληψία URL[17] με random walks[18]

Στην προηγούμενη εφαρμογή δειγματοληψίας με random walk είδαμε ότι υπάρχει το πρόβλημα της μεροληψίας. Αυτό έχει σαν αποτέλεσμα σελίδες κοντά στην αφετηρία να επιλέγονται πιο συχνά. Η μεροληψία, όμως, επιπλέον ενεργεί και υπέρ των σελίδων που βρίσκονται σε υψηλά συνεκτικές συνιστώσες. Έτσι, ένας random walk τείνει να επιλέγει σελίδες που είναι υψηλά συνδεδεμένες μεταξύ τους παρά σελίδες που είναι πιο φτωχές σε συνδέσμους. Αυτό, όμως, έχει σαν αποτέλεσμα το δείγμα να μην είναι ομοιόμορφο. Σε αυτή την εργασία προτείνεται μία τροποποίηση του random walk της υποενότητας 2.2.1 ώστε να ελαττωθεί λίγο το πιο πάνω πρόβλημα.

Έστω ότι θέτουμε τον random walk σε λειτουργία με σκοπό να μαζέψουμε ένα τμήμα του web και στη συνέχεια σε αυτό το δείγμα να κάνουμε ομοιόμορφη δειγματοληψία. Η πιθανότητα μία σελίδα X να βρίσκεται στο τελικό δείγμα ισούται με την πιθανότητα να προσπελαστεί (crawled) από τον random walk επί την πιθανότητα να επιλεχθεί στο δείγμα, δηλ.:

$$\Pr[X \text{ στο δείγμα}] = \Pr[X \text{ crawled}]\Pr[X \text{ στο δείγμα} \mid X \text{ crawled}]$$

Μία προσέγγιση του δεξιού μέρους της ισότητας βρίσκεται ως εξής: Αν ο random walk είναι μακρύς ώστε να έχει φτάσει στην κατανομή ισορροπίας, τότε κάθε σελίδα X θα δέχεται επισκέψεις από αυτόν ανάλογα με το PageRank της $R(X)$, διότι η κατανομή ισορροπίας του random walk είναι ανάλογη του PageRank. Αν, λοιπόν, L είναι το μήκος του random walk, τότε η ποσότητα $L \cdot R(X)$ μας δίνει τη

[17] URL σημαίνει Uniform Resource Locator και είναι η επίσημη ονομασία των υπερσυνδέσμων (RFC 1738 [RFC]).
[18] [HHMN00]

μέση τιμή των επισκέψεων του random walk στη σελίδα X, δηλ. $\mathrm{E}[\text{αριθμός επισκέψεων στην X}] \approx L \cdot \mathrm{R}(X)$. Άρα, αν κατορθώναμε να κάνουμε τη δειγματοληψία του κομματιού που παίρνουμε από το random walk με πιθανότητα $\Pr[X \text{ στο δείγμα} \mid \text{X crawled}]$ αντιστρόφως ανάλογη του $R(X)$, τότε η δειγματοληψία θα πλησίαζε στην ομοιόμορφη, επειδή $\Pr[X \text{ στο δείγμα}] \propto \frac{1}{R(X)}$. Με αυτό τον τρόπο θα ελαττώναμε τη μεροληψία της πρώτης σελίδας.

Το πρόβλημα που συναντούμε τώρα είναι πώς θα βρίσκουμε το $R(x)$ κάθε σελίδας καθώς εξελίσσεται ο random walk. Προσεγγιστικά μπορούμε να βρούμε το $R(x)$ αν υπολογίζουμε το κλάσμα των εμφανίσεων της σελίδας X στο random walk προς το μήκος του random walk για μεγάλους περιπάτους. Ένας άλλος τρόπος θα ήταν να πραγματοποιήσουμε τον random walk και από το γράφο που θα προκύψει να υπολογίζουμε το PageRank κάθε κόμβου (σελίδας). Αυτός είναι ένας τρόπος με περισσότερο υπολογιστικό κόστος, διότι απαιτείται η αποθήκευση και αλγοριθμική μεταχείριση μεγάλων ποσοτήτων πληροφορίας (όλες οι προσπελασθείσες ακμές και σελίδες θα πρέπει να αποθηκεύονται και να επεξεργάζονται).

Η πιο πάνω τεχνική παράγει σχεδόν ομοιόμορφα δείγματα από web σελίδες αλλά έχει τους εξής περιορισμούς: i) δε μπορεί να εφαρμοστεί για όχι καλά συνεκτικές συνιστώσες του web (αυτό είναι ένα γενικότερο πρόβλημα στους random walks). ii) επειδή ο random walk κατά καιρούς πηδά σε τυχαίες σελίδες που έχουν ήδη προσπελαστεί, δεν είναι δυνατόν να ανακαλύψει σελίδες στις οποίες οδηγούμαστε μόνο μέσα από μεγάλου μήκους μονοπάτια. iii) σελίδες με δυναμικό περιεχόμενο δεν μπορούν να προσπελαστούν από τον random walk. Εδώ πρέπει να αναφέρουμε ότι οι σελίδες με δυναμικό περιεχόμενο είναι μία κατηγορία web σελίδων που δημιουργούνται ύστερα από αλληλεπίδραση του χρήστη με σελίδες-φόρμες (forms). Για παράδειγμα, η συμπλήρωση μίας φόρμας ή η απάντηση σε ένα ερωτηματολόγιο μπορεί να οδηγήσει σε διαφορετικού περιεχομένου σελίδες οι οποίες φτιάχνονται εκείνη την ώρα και είναι προσαρμοσμένες στις απαντήσεις του χρήστη. iv) η μεροληψία δεν εξαφανίζεται τελείως και η αφετηρία του random walk επηρεάζει την πιθανότητα εμφάνισης κάποιας άλλης σελίδας. Το ίδιο γίνεται και με τα τυχαία άλματα του random walk σε τυχαίες σελίδες που έχουν ήδη προσπελαστεί κάνοντας αυτές να εμφανίζονται περισσότερες φορές από ό,τι ίσως θα έπρεπε. v) η προσέγγιση ότι η μέση τιμή των επισκέψεων του random walk στη σελίδα X είναι $L \cdot R(X)$ δεν ισχύει για μεγάλους περιπάτους και για σελίδες με μεγάλο βαθμό PageRank. iv) δεν γνωρίζουμε πόσα βήματα πρέπει να κάνει ο random walk για να προσεγγίσει αρκετά την κατανομή ισορροπίας.

2.2.3 Προσέγγιση αθροιστικών ερωτήσεων για web σελίδες με random walks

Οι δύο προηγούμενες εργασίες χρησιμοποίησαν random walks και πήραν δείγματα του web σχεδόν ομοιόμορφα. Μία ακόμα καλύτερη προσέγγιση της ομοιομορφίας στα δείγματα σημειώνεται στη εργασία [BBC+00]. Η αφετηρία για να γίνει αυτό είναι η ανάγκη να απαντούμε σε ερωτήματα που αφορούν μεγάλο μέρος από web σελίδες. Τέτοια ερωτήματα ονομάζονται αθροιστικά ερωτήματα (aggregate queries) και είναι της μορφής: τί ποσοστό των web σελίδων τελειώνει σε .com; πόσες σελίδες βρίσκονται στον κατάλογο μίας συγκεκριμένης μηχανής αναζήτησης; κ.λπ.

Απαντήσεις σε τέτοια ερωτήματα μπορούμε να δώσουμε αν πάρουμε ομοιόμορφα δείγματα web σελίδων και τα επεξεργαστούμε.

Από τη θεωρία των random walk είναι γνωστό ότι ένας random walk σε ένα κανονικό[19] και κατευθυνόμενο γράφο μας δίνει ένα σχεδόν ομοιόμορφο δείγμα από κόμβους. Επειδή, όμως, ο γράφος του web δεν είναι ούτε κατευθυνόμενος ούτε κανονικός, καθώς ο random walk που προτείνεται σε αυτή την εργασία προχωρά, κατασκευάζει ένα "ιδανικό" γράφο G ο οποίος να είναι κανονικός και μη κατευθυνόμενος. Για κάθε link, δηλαδή, που ακολουθείται από το random walk προστίθεται στο γράφο G ένα δεύτερο link με αντίθετη κατεύθυνση, ώστε να μπορεί ο random walk να ακολουθήσει και την αντίθετη πορεία (έτσι ο G γίνεται μη κατευθυνόμενος), και επιπλέον αν d είναι ο μέγιστος βαθμός κορυφής που έχουμε συναντήσει προστίθεται σε κάθε κόμβο και κατάλληλος αριθμός ακμών που δείχνουν στον εαυτό τους (self loops) έτσι ώστε κάθε κόμβος να έχει τον ίδιο βαθμό (έτσι ο G γίνεται κανονικός). Επειδή, επιπλέον, ο γράφος G πρέπει να είναι και συνεκτικός θεωρείται ότι ο random walk κινείται στο SCC κομμάτι του web (και φυσικά μπορεί να μπει και στο OUT). Σύμφωνα, λοιπόν, με τα πιο πάνω ο γράφος G είναι συνεκτικός, κανονικός και μη κατευθυνόμενος, άρα ο random walk στον G θα πλησιάζει στην ομοιόμορφη κατανομή.

Ο πιο πάνω γράφος G είναι ιδανικός και μόνο σε αυτόν ο random walk θα παράγει σχεδόν ομοιόμορφα δείγματα. Όταν ο random walk γίνεται στον πραγματικό γράφο του web δεν υπάρχει γνωστή τεχνική ή τρόπος ώστε να κάνουμε όλο το γράφο του web συνεκτικό, κανονικό και μη κατευθυνόμενο. Για παράδειγμα, συνεκτικό είναι μόνο το SCC κομμάτι του web. Επίσης, δεν υπάρχει αποτελεσματική τεχνική για να βρίσκει κανείς τον in-degree[20] κάθε κόμβου (έτσι ώστε να υπολογίζει το βαθμό κάθε κόμβου). Ένας τρόπος να προσεγγίζουμε τον in-degree είναι να ρωτάμε τις μηχανές αναζήτησης σχετικά με το πόσα link δείχνουν σε μία συγκεκριμένη σελίδα. Όμως, ακόμα και αυτός ο τρόπος έχει περιορισμούς, διότι οι κατάλογοι των μηχανών αναζήτησης δεν είναι πλήρεις αλλά και διότι η διαδικασία των ερωτήσεων-απαντήσεων από και προς τις μηχανές αναζήτησης για κάθε κόμβο που συναντά ο random walk είναι χρονοβόρα. Επιπλέον, οι μηχανές αναζήτησης επιστρέφουν περιορισμένο αριθμό απαντήσεων (π.χ. 1000 απαντήσεις) για κάθε ερώτηση, οπότε κόμβοι με μεγαλύτερο in-degree από κάποιο όριο δεν μπορούν να εντοπιστούν.

Παρόλα τα πιο πάνω προβλήματα και περιορισμούς η μέθοδος καταφέρνει και πλησιάζει την ομοιόμορφη δειγματοληψία web σελίδων και δίνει αρκετά καλές προσεγγιστικές απαντήσεις στα πιο πάνω αθροιστικά ερωτήματα.

2.2.4 Συμπεράσματα από δειγματοληψία με random walks

Είδαμε μέχρι στιγμής 3 προσπάθειες ομοιόμορφης δειγματοληψίας του web με random walks. Υπάρχουν, βέβαια, και άλλες όπως π.χ. η [RPLG01] στην οποία τροποποιούνται οι τεχνικές των υποενοτήτων 2.2.1, 2.2.2, 2.2.3, οπότε και πάλι στο όριο (δηλ. για πολύ μεγάλους περιπάτους) παίρνουμε σχεδόν ομοιόμορφα δείγματα web σελίδων (πράγμα το οποίο αμφισβητείται στην [Hen03]).

[19] Ένας γράφος (μη κατευθυνόμενος) είναι κανονικός εάν κάθε κόμβος του έχει τον ίδιο βαθμό.

[20] Αντίθετα, ο out-degree υπολογίζεται πολύ εύκολα αν εξάγουμε όλα τα links που αναγράφονται στον html κώδικα της σελίδας.

Συνοπτικά, από τα πιο πάνω θα λέγαμε ότι η μέθοδος δειγματοληψίας web σελίδων με random walks έχει:

Πλεονεκτήματα:

- Απλή στην ιδέα μέθοδος με καλά θεμελιωμένο μαθηματικό υπόβαθρο. Η αρχή της βασίζεται σε μία στοχαστική διαδικασία (stochastic process) μέσω της οποίας παράγονται τυχαίες κατασκευές (random generations) που στην περίπτωσή μας είναι δείγματα από web σελίδες.

- Δυνατότητα εφαρμογής προσεγγιστικών μεθόδων (αλγορίθμων)[21].

Μειονεκτήματα:

- Η δομή του web γράφου είναι τέτοια ώστε μόνο στην SCC συνιστώσα να μπορούν οι random walks να εκμεταλλεύονται πλήρως τις δυνατότητές τους. Αυτό σημαίνει ότι δεν μπορεί να εφαρμοστεί δειγματοληψία σε όλο το web γράφο παρά μόνο στις καλώς συνεκτικές συνιστώσες του (π.χ. ο random walk δεν φτάνει ποτέ σελίδες στις οποίες οδηγούμαστε μέσα από μεγάλου μήκους μονοπάτια). Επίσης, δεν υπάρχει αξιόπιστη μέθοδος που να μετατρέπει το γράφο του web σε συνεκτικό.

- Η μεροληψία που εισάγεται από τη σελίδα αφετηρίας του random walk, διότι όπως είπαμε δεν έχει βρεθεί τρόπος ομοιόμορφης επιλογής μίας τυχαίας σελίδας του web. Η μεροληψία είναι έντονη ιδίως για περιοχές του web γράφου με υψηλή συνεκτικότητα. Μεροληψία, επίσης, εισάγει η προσπάθεια ο random walk να αποφεύγει τους κύκλους και τα αδιέξοδα από τα οποία είναι γεμάτο το web.

- Δεν είναι γνωστός ο ακριβής αριθμός των βημάτων που πρέπει να εκτελέσουν οι random walks για να φτάσουν στην κατανομή ισορροπίας.

- Δε γνωρίζουμε πώς μπορούμε ικανοποιητικά να υπολογίζουμε τον in-degree κάθε web σελίδας και ούτε υπάρχει αξιόπιστη τεχνική που να δουλεύει στην πράξη ώστε ο web γράφος να μετατρέπεται σε κανονικό και μη κατευθυνόμενο.

[21] Για ένα survey βλέπε [Afr02].

2.3 Μέθοδος δειγματοληψίας με IP

2.3.1 Η βασική ιδέα[22]

Όπως είδαμε το web αποτελείται από σελίδες οι οποίες συνδέονται μεταξύ τους με links. Είπαμε, επίσης, ότι μία συλλογή από web σελίδες που δικτυώνονται έτσι ώστε να αποτελούν ομάδα με συγκεκριμένο πληροφοριακό αντικείμενο ονομάζεται web site [wcp]. Συνήθως, ένα web site οργανώνεται κάτω από ένα URL το οποίο το ονομάζουμε βασικό URL (base URL). Για παράδειγμα, το web site του Εθνικού Μετσόβιου Πολυτεχνείου έχει base URL το http://www.ntua.gr . Παραδείγματα υποκαταλόγων και web σελίδων που ανήκουν στο ίδιο web site (και έχουν το ίδιο base URL) είναι τα: http://www.ntua.gr/gr_announce/index.shtm, http://www.ntua.gr/doy/, http://www.ntua.gr/gr_academics/index.htm. Συνήθως, το base URL που είναι και η αρχική σελίδα του web site ονομάζεται και home page.

Σύμφωνα με τους [BCHR01] οι web pages και τα web sites θεωρούνται στοιχειώδεις μονάδες (σε επίπεδο αφαιρετικό)[23] όταν θέλουμε να μελετήσουμε το web. Στην περίπτωσή μας θέλουμε να παίρνουμε ομοιόμορφα δείγματα από web pages, κάτι που όπως είδαμε μέχρι τώρα δεν είναι και τόσο εύκολο. Μία διαφορετική προσέγγιση είναι η εξής: Αφού κάθε web site αποτελείται από web pages, αν μπορούσαμε να πάρουμε ένα τυχαίο δείγμα από web sites τότε θα είχαμε και ένα δείγμα από web pages. Η πιο πάνω τεχνική ονομάζεται single stage cluster sampling (cluster – συστάδα) και σύμφωνα με αυτήν έχουμε δύο μονάδες δειγματοληψίας, την άμεση και την έμμεση. Στην περίπτωσή μας η άμεση μονάδα είναι το web site και η έμμεση η web page. Σύμφωνα με αυτή την τεχνική παίρνουμε ένα δείγμα της άμεσης μονάδας (web site) και μία συστάδα – δείγμα (cluster sample) της έμμεσης μονάδας (web page). Θα πρέπει, άρα, να βρεθεί μία τεχνική η οποία θα παράγει τυχαία δείγματα από web site. Αν υπήρχε κάποια λίστα όλων των web server τότε μέσω αυτής θα μπορούσαμε να πάρουμε ένα δείγμα από web sites. Τέτοια λίστα όμως δεν υπάρχει.

Ένας άλλος τρόπος προσέγγισης του προβλήματος είναι ο εξής: Έστω H το σύνολο όλων των web server που υπάρχουν στο web. Αν υπάρχει ένας γνωστός πληθυσμός I που να περιέχει τον H ($H \subset I$) τότε δειγματοληπτώντας σε αυτό τον πληθυσμό μπορούμε να πάρουμε δείγμα από τον άγνωστο H και άρα να έχουμε ένα δείγμα από web servers και άρα από web sites. Ένας τέτοιος πληθυσμός I υπάρχει και είναι ο IPv4 χώρος διευθύνσεων του Internet.

2.3.2 Ο IPv4 χώρος και οι IP διευθύνσεις

Στην εισαγωγή είχαμε αναφέρει ότι το Internet είναι η τεχνολογική πλατφόρμα του web. Σύμφωνα με τη θεωρία δικτύων [PD96] το Internet είναι ένα σύνολο από ετερογενή δίκτυα, τα οποία όμως για να λειτουργήσουν χρειάζονται ένα πρωτόκολλο. Αυτό το πρωτόκολλο ονομάζεται IP (Internet Protocol) υπάρχει σε όλους τους κόμβους του δικτύου (δηλαδή σε όλους τους host (εξυπηρετητές) και τους routers (δρομολογητές)) και καθορίζει πώς τα διάφορα ετερογενή δίκτυα θα λειτουργούν σαν ένα ενιαίο δίκτυο (διαδίκτυο). Το IP πρωτόκολλο διαθέτει ένα μοντέλο υπηρεσιών

[22] [OML97]
[23] Μία ακόμα μονάδα αφαίρεσης είναι και αυτή του domain την οποία θα συναντήσουμε πιο κάτω.

(service model) το οποίο καθορίζει το ποιές υπηρεσίες θέλουμε να υπάρχουν μεταξύ των host. Το μοντέλο υπηρεσιών αποτελείται από δύο μέρη. Το πρώτο μέρος είναι το σχήμα διευθυνσιοδότησης (addressing scheme) το οποίο παρέχει ένα τρόπο για να ταυτοποιούνται όλοι οι host στο δίκτυο. Το δεύτερο μέρος είναι ένα μοντέλο μεταφοράς των δεδομένων (datagram model). Αυτό που άμεσα μας αφορά σε αυτή την εργασία είναι το πρώτο μέρος του μοντέλου υπηρεσιών του IP πρωτοκόλλου, το σχήμα διευθυνσιοδότησης του Internet.

Το σχήμα διευθυνσιοδότησης του Internet αποτελείται από τις IP διευθύνσεις. Μία IP διεύθυνση (IP address) αποτελείται από 2 μέρη, ένα μέρος που αναφέρεται στο δίκτυο (network part) και ένα μέρος που αναφέρεται στον host (host part). Το μέρος που αναφέρεται στο δίκτυο ταυτοποιεί το τοπικό δίκτυο με το οποίο ένας host είναι συνδεδεμένος. Με αυτό τον τρόπο όλοι οι host που βρίσκονται στο ίδιο δίκτυο έχουν το ίδιο network part στην IP διεύθυνση. Το host part τώρα καθορίζει μονοσήμαντα κάθε host μέσα σε ένα συγκεκριμένο δίκτυο. Το μήκος μίας IP διεύθυνσης είναι 32 bit πράγμα που σημαίνει ότι όλες οι IP διευθύνσεις που υπάρχουν είναι 2^{32} = 4,294,967,296. Το πιο πάνω σύνολο αυτών των IP διευθύνσεων αποτελεί τον IPv4 χώρο διευθύνσεων του Internet[24].

Οι IP διευθύνσεις έχουν ένα συγκεκριμένο τρόπο γραφής. Όπως είπαμε, κάθε IP διεύθυνση έχει μήκος 32 bits. Όμως, 8 bit = 1 byte άρα μία IP διεύθυνση είναι 4 bytes. Κάθε byte μίας IP διεύθυνσης το αναπαριστούμε με έναν ακέραιο γραμμένο στο δεκαδικό σύστημα ξεκινώντας από το πιο σημαντικό (το πιο αριστερό δηλαδή) ψηφίο της IP διεύθυνσης. Αφού 1 byte = 8 bit κάθε τέτοιος ακέραιος θα μπορεί να πάρει 2^8 = 256 τιμές. Αν ξεκινήσουμε την αρίθμηση από το 0 τότε κάθε ακέραιος παίρνει τιμές από 0 – 255. Όταν τώρα γράφουμε μία IP διεύθυνση οι 4 αυτοί ακέραιοι χωρίζονται μεταξύ τους με τελείες. Για παράδειγμα μία IP διεύθυνση είναι η 147.102.222.210 .

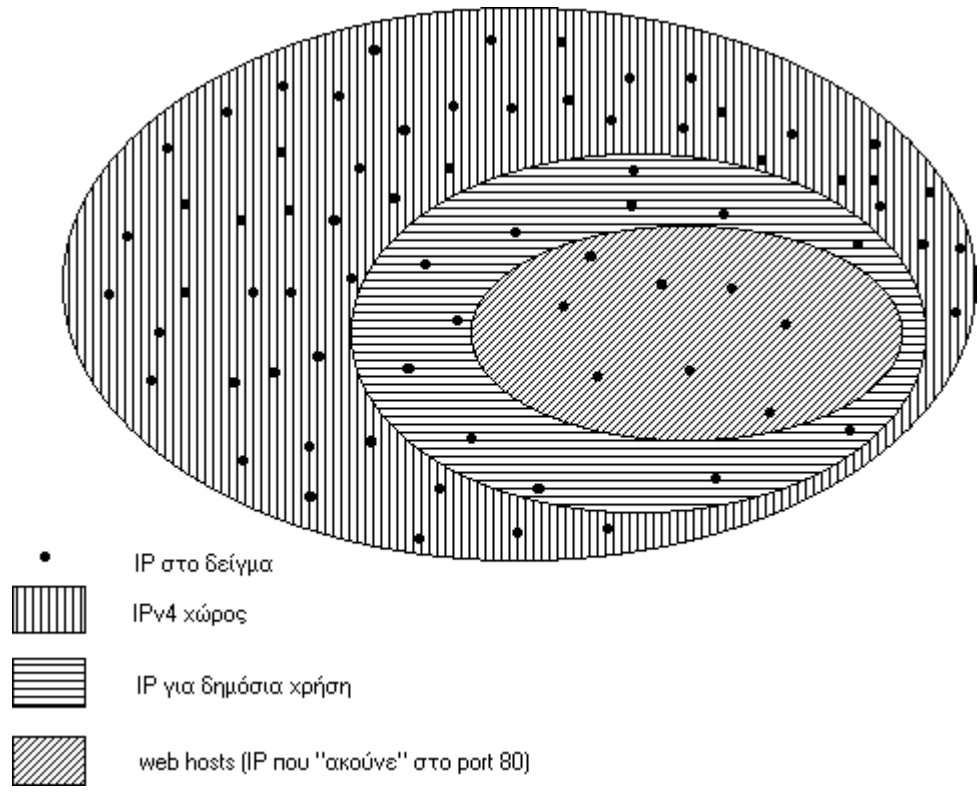
Για να αποκτήσει ένας χρήστης πρόσβαση στις διάφορες υπηρεσίες του Internet το πρώτο πράγμα που έχει να κάνει είναι να "βάλει" Internet αποκτώντας μία IP διεύθυνση. Όταν, τώρα, αυτός ο χρήστης θελήσει να χρησιμοποιήσει μία συγκεκριμένη υπηρεσία θα πρέπει να το δηλώνει μαζί με τη διεύθυνσή του. Διάφορες υπηρεσίες του Internet είναι οι http, ftp κ.λπ. και κάθε μία από αυτές έχει συνήθως και ένα port (πόρτα) στο οποίο "ακούει". Παρόλο που δεν είναι υποχρεωτικό κάθε υπηρεσία να ακούει σε συγκεκριμένο port, έχει γίνει συνήθεια στο Internet να υπάρχουν προκαθορισμένες (default) πόρτες για κάθε υπηρεσία ώστε, αν δεν καθορίζει ο χρήστης κάτι διαφορετικό, να γίνεται έλεγχος σε αυτές για δεδομένα. Για παράδειγμα η default πόρτα για την υπηρεσία http (για το web δηλαδή) είναι η 80. Έτσι, όταν ένας χρήστης θελήσει να μεταβεί στην home page του Εθνικού Μετσόβιου Πολυτεχνείου (Ε.Μ.Π.) τότε μπορεί να γράψει στον Browser http://www.ntua.gr:80 . Επειδή όμως, όπως είπαμε, η default port είναι η 80 μπορεί απλά να γράψει http://www.ntua.gr . Έρευνες [WAB+96] έχουν δείξει ότι με ποσοστό 93.6% όλες οι html σελίδες του web (δηλ. οι web pages) εξυπηρετούνται από port 80. Εάν, λοιπόν, μπορούσαμε να πάρουμε ένα δείγμα από IP διευθύνσεις και κοιτούσαμε ποιές από αυτές απαντούν στο port 80 τότε θα μαθαίναμε ποιές αντιστοιχούν σε host που

[24] Η έκδοση του Internet που χρησιμοποιείται στις μέρες μας (Δεκ. 2003) είναι η 4 (IPv4 – Internet Protocol version 4). Η 5 έκδοση δεν θα μπει σε λειτουργία αλλά υπάρχει για πειραματικούς σκοπούς. Το νέο Internet θα είναι η έκδοση 6 (IPv6) και οι διευθύνσεις του θα έχουν μήκος 128 bits.

εξυπηρετούν web σελίδες[25] (web hosts). Δηλαδή έμμεσα από το δείγμα των IP διευθύνσεων έχουμε ένα δείγμα από διευθύνσεις web host (οι οποίες είναι και web σελίδες).

Η δειγματοληψία, λοιπόν, γίνεται ως εξής: Από όλο τον IPv4 χώρο διευθύνσεων με τις 2^{32} διευθύνσεις παίρνουμε ένα δείγμα από IP διευθύνσεις. Στη συνέχεια, αυτό το δείγμα το "φιλτράρουμε" κρατώντας μόνο όσα IP "ακούνε" στο port 80. Αυτά τα IP μας καθορίζουν τους web host που φιλοξενούν web σελίδες.

Όλες οι διευθύνσεις του IPv4 χώρου διαχειρίζονται από μία αρχή που ονομάζεται IANA (Internet Assigned Numbers Authority) [IANA]. Η IANA έχει καθορίσει ότι κάποιες από τις IP διευθύνσεις δεν θα ανατίθενται σε τελικούς χρήστες αλλά θα κρατιούνται για ειδικούς σκοπούς[26] (π.χ. ιδιωτικά δίκτυα που δεν επιτρέπουν πρόσβαση στο κοινό κ.λπ.). Έτσι, στην πραγματικότητα δεν κάνουμε δειγματοληψία μέσα από όλες τις 2^{32} διευθύνσεις αλλά από ένα υποσύνολό τους (IP για δημόσια χρήση). Τα πιο παραπάνω φαίνονται σχηματικά στην πιο κάτω φιγούρα.



Φιγούρα 2.1: IP δειγματοληψία

[25] Μπορεί ένας host να έχει το port 80 ανοιχτό αλλά να μην εξυπηρετεί web σελίδες, ίσως επειδή αυτές δεν έχουν "στηθεί" ακόμα.
[26] Στη διεύθυνση http://www.iana.org/assignments/ipv4-address-space μπορεί κάποιος να δει ολόκληρη την κατανομή του IPv4 χώρου διευθύνσεων.

2.4 Εργασίες σχετικά με IP sampling

Στις υποενότητες που ακολουθούν παρουσιάζουμε κάποιες αντιπροσωπευτικές χρήσεις της μεθόδου IP sampling που συναντούμε στη διεθνή βιβλιογραφία.

2.4.1 Προσδιορισμός της ποσότητας και της κατανομής της πληροφορίας στο web με IP sampling[27]

Σε αυτή την εργασία έγινε επεξεργασία στο περιεχόμενο ενός τυχαίου δείγματος από web servers ώστε να διερευνηθεί η ποσότητα και η κατανομή της πληροφορίας στο web. Το δείγμα πάρθηκε με τη μέθοδο IP sampling από όλο τον IPv4 χώρο που όπως είδαμε περιέχει περίπου 4.3 δισεκατομμύρια διευθύνσεις IP. Η δειγματοληψία έγινε με επανατοποθέτηση, που σημαίνει ότι: i) η πιθανότητα επιλογής είναι η ίδια για κάθε IP και ii) ένα IP μπορεί να εμφανιστεί περισσότερες από μία φορές στο δείγμα. Σε αυτή την εργασία δεν εξαιρέθηκαν όσες IP διευθύνσεις δεν έχουν ανατεθεί ακόμα, ούτε όσες χρησιμοποιούνται για άλλους σκοπούς εκτός της web υπηρεσίας. Το δείγμα που προέκυψε από τη δειγματοληψία περιείχε 3.6 εκατομμύρια IP διευθύνσεις.

Η επεξεργασία του δείγματος έγινε ως εξής: Δοκιμάζεται κάθε IP και αν εντοπίζεται κάποιος web server το IP φυλάσσεται, αλλιώς εξαιρείται. Ο χρόνος αναμονής για απάντηση κάθε server (request time) ορίζεται στα 30 δευτερόλεπτα και αν τα υπερβούμε τότε το IP θεωρείται ότι δεν αντιστοιχεί σε web server. Επειδή επιπλέον, πολλά web sites είναι προσωρινώς μη διαθέσιμα, εξαιτίας των συνδέσεων του Internet αλλά και διάφορων παροδικών προβλημάτων των web server (π.χ. web server downtime), οι δοκιμές επαναλήφθηκαν για όλο το δείγμα μία εβδομάδα αργότερα.

Τα αποτελέσματα από τις πιο πάνω δοκιμές ήταν ότι για κάθε 269 δοκιμές ένα IP αντιστοιχούσε σε web server. Αυτό σημαίνει ότι η εκτίμηση για όλους τους web servers είναι γύρω στα 16 εκατομμύρια (Φεβ. 1999). Η πιο πάνω εκτίμηση όμως είναι παραπλανητική, διότι δεν αντιστοιχεί στο "δημόσιο" όπως λέμε web (publicly indexable web)[28]. Για παράδειγμα, πολλοί από τους web servers που ανακαλύφθηκαν χρειάζονται αναγνώριση στοιχείων (authentication) από το χρήστη για να τους χρησιμοποιήσει (π.χ. φόρμες με username και passwords, όπως web-based emails). Άλλοι, πάλι, IP αριθμοί αντιστοιχούσαν σε printers ή σε διάφορα web interfaces. Τελικά, εκτίμησαν ότι από τα 16 εκατομμύρια μόνο 2.8 ήταν την εποχή εκείνη ο αριθμός των web server. Από αυτούς επιλέχθηκαν 2,500 τυχαίοι servers και χρησιμοποιήθηκαν crawlers για να προσπελαστούν όλες οι σελίδες τους. Κατά μέσο όρο βρέθηκαν 289 web σελίδες σε κάθε server, που σημαίνει ότι κατ' εκτίμηση το μέγεθος του web υπολογίστηκε περίπου ως $8 \cdot 10^8$ σελίδες. Επίσης, μετρώντας και το μέσο μέγεθος των πιο πάνω σελίδων (18,7 kbytes καθαρού κειμένου ανά σελίδα) εκτίμησαν τον όγκο του web σε 6Tbytes. Ακόμα, χρησιμοποιώντας το πιο πάνω δείγμα υπολόγισαν τί ποσοστό του web καλύπτουν γνωστές μηχανές αναζήτησης και, τέλος, εξήγαγαν το συμπέρασμα ότι οι τελευταίες καταλογοποιούν ένα πολύ μικρό τμήμα του web.

[27] [LG99]
[28] Περισσότερα για το publicly indexable web θα δούμε στην 2.4.2

Σύμφωνα με τα πιο πάνω βλέπουμε ότι έγινε δειγματοληψία σε όλο τον IPv4 χώρο, χωρίς να λαμβάνονται υπ' όψιν οι διάφοροι περιορισμοί του. Αυτό είναι ένα μειονέκτημα για την "καθαρότητα" του δείγματος. Επίσης, ένα άλλο πρόβλημα είναι το ότι δεν έχει βρεθεί τρόπος ομοιόμορφης και τυχαίας δειγματοληψίας web σελίδων από web sites. Γι' αυτό το λόγο αναγκαστικά προσπελάστηκαν όλες οι web σελίδες των 2,500 web server. Ένα άλλο στοιχείο είναι ότι με τη μέθοδο αυτή ανακαλύπτουμε και web server οι οποίοι με τους random walk δεν θα μπορούσαν ποτέ να ανακαλυφθούν. Αυτό, όπως θα δούμε, είναι και ένα μεγάλο πλεονέκτημα της μεθόδου. Τέλος, θα μπορούσαμε να παρατηρήσουμε ότι το IP sampling είναι μία μέθοδος πρακτική και όχι θεωρητική.

2.4.2 Deep web και IP sampling[29]

Όπως είδαμε στην υποενότητα 2.4.1 με τον όρο publicly indexable web εννοούμε το σύνολο των web sites που είναι προσβάσιμα από όλους τους χρήστες του web. Αυτό το τμήμα του web ονομάζεται καμιά φορά και surface web (επιφανειακό web) και είναι αυτό που συνήθως καταλογοποιούν οι μηχανές αναζήτησης. Υπάρχει, όμως, και ένα άλλο μέρος του web που ονομάζεται deep ή non-indexable web. Συνήθως, με τον πιο πάνω όρο εννοούμε τις ογκώδεις βάσεις δεδομένων που βρίσκονται στο Internet. Η πληροφορία που περιέχεται σε αυτές τις βάσεις κρύβεται πίσω από φόρμες ερωτήσεων που πρέπει να συμπληρώσει κανείς για να τις ψάξει. Γι' αυτό το λόγο μερικοί ονομάζουν το τμήμα αυτό του web και κρυμμένο (hidden) ή αόρατο (invisible) web. Επειδή η πληροφορία σε αυτές τις βάσεις δεν μπορεί να προσπελαθεί απευθείας με κάποιο URL και επειδή οι διάφοροι crawlers δεν μπορούν αποτελεσματικά να συμπληρώνουν τις διάφορες φόρμες ερωτήσεών τους δεν τις εξερευνούν. Ένας τρόπος να παρακαμφθεί το πιο πάνω πρόβλημα είναι η μέθοδος IP sampling.

Στην [CHLZ03] θεωρείται ότι οι web server είναι ομοιόμορφα κατανεμημένοι στον IPv4 χώρο και εξαιρώντας από αυτόν όλες τις μη διαθέσιμες IP διευθύνσεις λαμβάνεται τυχαίο δείγμα από 1,000,000 IP. Αφαιρώντας από αυτές τις διπλο-εμφανίσεις λόγω δειγματοληψίας με επανατοποθέτηση προκύπτουν 999,789 διευθύνσεις από τις οποίες οι 4705 ήταν πιθανοί web server διότι είχαν το port 80 ανοιχτό. Από αυτές οι 1004 αντιστοιχούσαν σε προσπελάσιμους web server, διότι επέστρεφαν http σελίδες όταν τους ζητήθηκε, ενώ οι υπόλοιπες αντιστοιχούσαν είτε σε ιδιωτικούς, είτε σε προσωρινούς, είτε σε server που δεν χρησιμοποιούσαν το http πρωτόκολλο. Από τα web sites των 1004 server εξαιρέθηκαν όσοι είχαν αρχικές σελίδες με μηνύματα λάθους (π.χ. access denied, page not found), μηνύματα υπό κατασκευή (under construction), κενές από περιεχόμενο σελίδες, σελίδες τεστ για έλεγχο επιτυχημένης εγκατάστασης web server κ.λπ. Τελικά, στο δείγμα έμειναν 527 web sites από τα οποία εξετάσθηκε η συχνότητα εμφάνισης των πιό πάνω βάσεων δεδομένων. Η τάξη μεγέθους του deep web που υπολογίστηκε είναι 10^5 web sites.

Μία παρατήρηση στα πιο πάνω είναι ότι τα 527 web sites που τελικά έμειναν στο δείγμα δεν αντιστοιχούν κατ' ανάγκη σε 527 web servers. Αυτό συμβαίνει λόγω του virtual hosting[30] (ή multihosting). Το virtual hosting είναι μία τεχνική σύμφωνα με την οποία ένας server φιλοξενεί περισσότερα από ένα web site. Αυτό συνήθως γίνεται

[29] [CHLZ03]
[30] Περισσότερα για το virtual hosting θα δούμε στην 3.5.1 .

για τους εξής λόγους: i) ο διαχειριστής (administrator) ενός server δε χρειάζεται κάθε φορά που προστίθεται ένας νέος χρήστης και θέλει να δημιουργήσει ένα web site να ξαναρυθμίσει τον server, και ii) λόγοι οικονομικοί. Υπάρχουν δύο τεχνικές για virtual hosting στον ίδιο server [Wai99]. Η μία ονομάζεται IP-based virtual hosting και απαιτεί την τοποθέτηση πολλαπλών network interface στο server (ένα για κάθε IP) ή ενός πολυπλέκτη (multiplexer) σε ένα interface. Η άλλη μέθοδος ονομάζεται name-based virtual hosting και δίνει τη δυνατότητα πολλά web sites να μοιράζονται το ίδιο IP. Σύμφωνα, λοιπόν, με την name-based virtual hosting μέθοδο είναι δυνατόν τα 527 πιο πάνω IP που τελικά έμειναν στο δείγμα να μην αντιστοιχούν σε 527 web sites, αλλά σε αρκετά περισσότερα.

2.4.3 Συμπεράσματα από δειγματοληψία με IP

Στις υποενότητες 2.4.1 και 2.4.2 είδαμε 2 εφαρμογές του IP sampling στη μελέτη του web. Όμως, η μέθοδος αυτή, επειδή βασίζει τη λειτουργία της όχι στο web αλλά στο Internet, μπορεί να εφαρμοστεί και για τη μελέτη του τελευταίου. Ένα παράδειγμα αποτελεί η [GT00] στην οποία εφαρμόζεται το IP sampling για να εξερευνηθεί ο IPv4 χώρος και να κατασκευαστούν χάρτες απεικόνισης του Internet σε επίπεδο router. Το πιο πάνω δεν μπορεί να το κάνει ούτε κάποιος crawler αλλά ούτε και ένας random walk, διότι αυτοί βασίζουν τη λειτουργία τους στα links μεταξύ των web σελίδων[31].

Συνοπτικά για τη μέθοδο IP sampling θα λέγαμε ότι υπάρχουν:

Πλεονεκτήματα:

- Είναι μία ντετερμινιστική μέθοδος η οποία προαπαιτεί τη γνώση του πληθυσμού της δειγματοληψίας.
- Δυνατότητα εφαρμογής και στο web και στο Internet.
- Μπορεί να εφαρμοστεί στο non indexable τμήμα του web.
- Απλή σε μάθηματικά μέθοδος με ξεκάθαρα στατιστικά συμπεράσματα (clean statistics).
- Εύκολη μέθοδος για επανάληψη της δειγματοληψίας από τρίτους.

Μειονεκτήματα:

- Οι διάφοροι περιορισμοί του IPv4 χώρου (τμήματα διευθύνσεων που δεν έχουν ανατεθεί ή είναι για ιδιωτική χρήση).
- Δεν είναι γνωστό πώς μπορούμε να πάρουμε δείγματα web pages από web sites, οπότε τα δείγματα web pages που παράγει συνήθως αποτελούνται από τις home pages των αντίστοιχων sites. Επιπλέον, το virtual hosting περιορίζει τη δυνατότητα της μεθόδου σε ανακάλυψη ενός web site ανά web server.

[31] Το IP sampling είναι μέθοδος πρακτική που αρκετές φορές χρειάζεται ευρετικές (heuristic) τεχνικές, πράγμα το οποίο άλλες φορές αποτελεί πλεονέκτημα και άλλες μειονέκτημα.

2.5 Επίλογος

Σε αυτό το κεφάλαιο είδαμε δύο βασικές μεθόδους δειγματοληψίας για το web, τους random walks και το IP sampling. Είδαμε διάφορες εφαρμογές των μεθόδων και αναφέραμε τα πλεονεκτήματα και μειονεκτήματά τους.

Στα επόμενα κεφάλαια θα ασχοληθούμε ειδικότερα με τη μέθοδο IP sampling βλέποντας πώς αυτή μπορεί να υλοποιηθεί και να χρησιμοποιηθεί για την εξόρυξη δεδομένων από το web και το Internet.

3 ΥΛΟΠΟΙΗΣΗ ΕΝΟΣ IP ΔΕΙΓΜΑΤΟΛΗΠΤΗ

Όπως είδαμε στα προηγούμενα κεφάλαια ένας τρόπος μελέτης του web είναι η δειγματοληψία και μία από τις μεθόδους δειγματοληψίας είναι το IP sampling. Σε αυτό το κεφάλαιο θα παρουσιάσουμε έναν IP δειγματολήπτη ο οποίος υλοποιήθηκε στο εργαστήριο Λογικής και Αλγορίθμων[32] του Ε.Μ.Π. περιγράφοντας, επίσης, και την πρώτη του δειγματοληψία η οποία έγινε στο Ελληνικό web (.gr).

3.1 Hostnames και Domain Name System (DNS)[33]

Είχαμε δει στην 2.3.2 ότι το Internet έχει ένα σχήμα διευθυνσιοδότησης που αποτελείται από IP διευθύνσεις οι οποίες χρησιμοποιούνται για να ταυτοποιούν τους διάφορους host. Επειδή, όμως, οι IP διευθύνσεις, σαν αριθμοί που είναι, δεν είναι και τόσο φιλικοί προς τους χρήστες, συνήθως σε κάθε host δίνουμε και ένα μοναδικό όνομα μέσα σε ένα δίκτυο. Αυτά τα ονόματα αποκαλούνται hostnames και διαφέρουν από τις IP διευθύνσεις στο ότι: i) είναι μεταβλητού και όχι σταθερού μήκους και ii) είναι "μνημονικά" (mnemonic - δηλαδή ευκολότερα στους ανθρώπους να τα θυμούνται).

Όταν ένας χρήστης ζητήσει ένα hostname τότε υπάρχει ένας μηχανισμός ανάλυσης (resolution mechanism) του hostname στη διεύθυνση που κρύβεται από πίσω. Τη δουλειά αυτή συνήθως την κάνουν ειδικοί υπολογιστές που ονομάζονται name servers. Με αυτό τον τρόπο κάθε hostname έχει και μία τιμή συνδεδεμένη μαζί του, η οποία συνήθως είναι μία IP διεύθυνση. Το πιο πάνω σύστημα ονομάζεται Domain Name System ή DNS. Το DNS σύστημα είναι ιεραρχικό και τα ονόματα σε αυτό επεξεργάζονται ιεραρχικά από τα δεξιά προς τα αριστερά χρησιμοποιώντας τελείες (.) σαν διαχωριστικά. Για παράδειγμα, ένα hostname είναι το achilles.noc.ntua.gr . Στο πρώτο επίπεδο της ιεραρχίας είναι 6 μεγάλα domains (χώροι αρμοδιότητας – περιοχές), .edu, .com, .gov, .mil, .org, .net, και ένα domain για κάθε κράτος, π.χ. .uk (United Kingdom), .gr (Greece), .ar (Argentina) κ.λπ. Το domain ενός κράτους ονομάζεται και country code Top Level Domain (ccTLD).

3.2 Δειγματοληψία μέσα σε ένα ccTLD[34]

Μέχρι στιγμής, στα παραδείγματα IP sampling που συναντήσαμε η δειγματοληψία γινόταν σε ολόκληρο τον IPv4 χώρο, χωρίς μερικές φορές να εξαιρούνται από αυτόν όποιες διευθύνσεις δεν έχουν ανατεθεί από την IANA, η οποία όπως έχουμε πει τις διαχειρίζεται.

Έστω ότι θέλουμε να κάνουμε δειγματοληψία μέσα σε ένα συγκεκριμένο ccTLD. Αν προσπαθήσουμε να κάνουμε τη δειγματοληψία με random walk θα πρέπει από πριν να γνωρίζουμε το γράφο του συγκεκριμένου domain. Είπαμε, βέβαια, στο κεφ. 1 ότι ο γράφος του web έχει fractal δομή, οπότε το πιο πιθανό είναι ο γράφος του συγκεκριμένου domain να είναι ένα bowtie [DKM+02] του οποίου όμως δε γνωρίζουμε ποιές σελίδες ανήκουν στις διάφορες συνιστώσες (π.χ. SCC, IN, OUT κ.λπ.). Ένα επιπλέον εμπόδιο που συναντούμε είναι το γεγονός ότι ο random walk

[32] Το Εργαστήριο Λογικής και Αλγορίθμων είναι ένα μη θεσμοθετημένο εργαστήριο που ανήκει στον τομέα επικοινωνιών, ηλεκτρονικής και συστημάτων πληροφορικής του Ε.Μ.Π., http://www.softlab .ece.ntua.gr/facilities/public/AD/page.html .
[33] [PD96]
[34] [Lek03b]

μπορεί καθώς θα τον ξεκινήσουμε από κάποια σελίδα μέσα στο ccTLD να βγει εκτός. Αυτό, όμως σημαίνει ότι ο random walk βγαίνει εκτός πληθυσμού ενδιαφέροντος και έτσι η δειγματοληψία θεωρείται όχι αξιόπιστη. Από την άλλη πλευρά αν προσπαθήσουμε να επέμβουμε, και, μόλις φτάσει η στιγμή ο random walk να βγει εκτός, τον επαναφέρουμε, χαλάμε τη στοχαστικότητα της διαδικασίας, διότι αν η επαναφορά γίνει σε μία σελίδα που έχει ήδη επισκεφθεί μέσα στο ccTLD τότε μεροληπτούμε υπέρ αυτής. Επίσης, αν η επαναφορά γίνει σε μία τυχαία σελίδα του ccTLD πέφτουμε στο πρόβλημα της τυχαίας επιλογής μίας web σελίδας από ένα ccTLD, πρόβλημα παρόμοιο με το άλυτο πρόβλημα της τυχαίας επιλογής μίας web σελίδας από ολόκληρο το web (βλ. κεφ. 1, 2).

Αν προσπαθήσουμε να εφαρμόσουμε τη μέθοδο IP sampling όπως τη συναντήσαμε μέχρι στιγμής συναντούμε τα εξής προβλήματα: Το να κάνουμε δειγματοληψία μέσα σε όλο το χώρο για να πάρουμε δείγματα από το συγκεκριμένο ccTLD είναι κάτι που απαιτεί δείγματα μεγάλου μεγέθους. Επιπλέον, η άνιση κατανομή των IP σε διάφορα domain θα δυσκολεύει την ακριβή ανάδειξη των ιδιαιτεροτήτων του συγκεκριμένου ccTLD.

Μία λύση στο πιο πάνω πρόβλημα είναι να κάνουμε μία IP δειγματοληψία μόνο στο συγκεκριμένο ccTLD. Αυτό μπορεί να γίνει εάν μπορούμε να μάθουμε ποιές IP διευθύνσεις ανήκουν στο ccTLD ή, όπως υποδεικνύουμε σε αυτή την εργασία, εάν μπορούμε να γνωρίζουμε το "χάρτη" των IP διευθύνσεων του ccTLD.

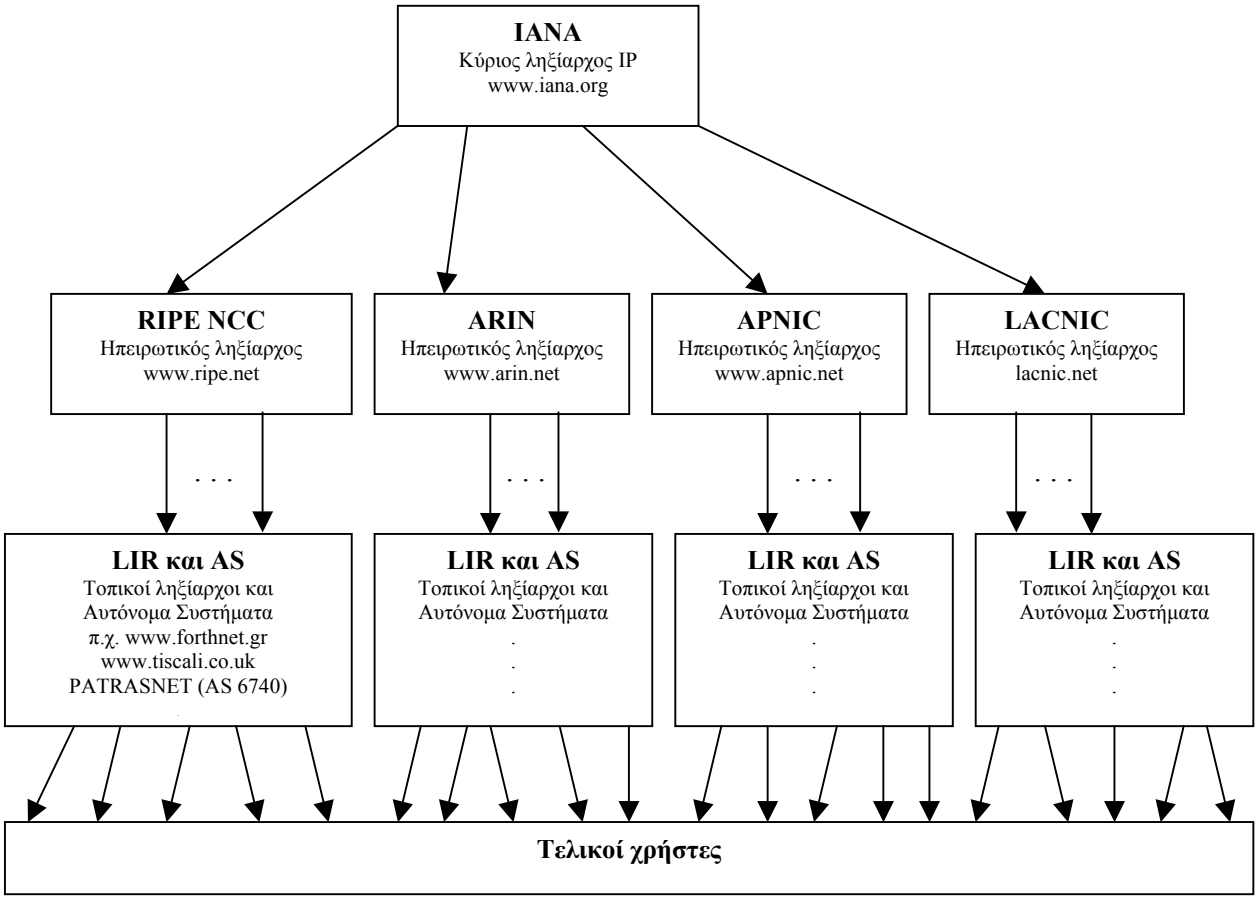
3.2.1 Το Internet Registry System

Όπως είδαμε, υπάρχει ανάγκη να γνωρίζουμε το χάρτη με τις IP διευθύνσεις ενός ccTLD. Αυτό μπορεί να γίνει μέσα από το ληξιαρχικό σύστημα του Ίντερνετ (Internet Registry System). Στο RFC 2050 [RFC] περιγράφεται η ιεραρχία του Internet Registry System η οποία αποτελείται από 3 επίπεδα: Ο οργανισμός που ευθύνεται για την ανάθεση των IP διευθύνσεων, όπως έχουμε πει, είναι η IANA [IANA]. Η IANA (κύριος ληξίαρχος) βρίσκεται στην κορυφή της ιεραρχίας και μοιράζει τις IP διευθύνσεις στους RIR (Regional Internet Registries – ηπειρωτικοί ληξίαρχοι). Οι RIR βρίσκονται στο δεύτερο επίπεδο της ιεραρχίας και, μέχρι στιγμής (10/2003), υπάρχουν 4 RIR: η ARIN (American Registry for Internet Numbers) για τη Βόρεια και Νότια Αμερική την Καραϊβική και τις Αφρικανικές χώρες Νότια του Ισημερινού, η APNIC (Asia Pacific Network Information Centre) για την περιφέρεια της Ασίας από την πλευρά του Ειρηνικού, η LACNIC (Regional Latin-American and Caribbean IP Address Registry) για τη Λατινική Αμερική και κάποια νησιά της Καραϊβικής, και ο RIPE NCC (Reseaux IP Europeens Network Coordination Centre) για την Ευρώπη την Κεντρική Ασία τη Μέση Ανατολή και τις Αφρικανικές Χώρες Βόρεια του Ισημερινού. Οι RIR με την σειρά τους κατανέμουν τις IP διευθύνσεις στους LIR (Local Internet Registries – τοπικοί ληξίαρχοι) και στα AS (Autonomous Systems – αυτόνομα συστήματα[35]). Ένα παράδειγμα LIR είναι ένας ISP (π.χ. forthnet) ενώ ένα παράδειγμα AS είναι το Ε.Μ.Π. Τέλος, οι LIR και τα AS παρέχουν IP σε τελικούς χρήστες. Τα πιο πάνω επίπεδα τα παρατηρούμε και στη φιγούρα 3.1.

Ας υποθέσουμε, τώρα, ότι θέλουμε να κάνουμε δειγματοληψία στο .gr ccTLD. Αυτό σημαίνει ότι πρέπει να γνωρίζουμε το χάρτη των IP διευθύνσεων του .gr, δηλαδή του Ελληνικού domain. Ο Ηπειρωτικός ληξίαρχος (RIR) που είναι υπεύθυνος

[35] βλ. 3.3.1

για την περιοχή της Ελλάδας είναι ο RIPE-NCC[36] [RIPE]. Ο RIPE, όπως είπαμε, κατανέμει τις IP διευθύνσεις στους τοπικούς ληξίαρχους (LIR) και στα αυτόνομα συστήματα (AS), που με τη σειρά τους κατανέμουν τις διευθύνσεις σε web χρήστες στο ccTLD.



Φιγούρα 3.1: Το ληξιαρχικό σύστημα του Ίντερνετ

Ο IP χάρτης του .gr, όπως ήταν διαμορφωμένος την 26η Νοεμβρίου 2002, φαίνεται στο παράρτημα Β και μπορεί κάποιος να προμηθευτεί παρόμοιους χάρτες από τη διεύθυνση: ftp://ftp.ripe.net/ripe/stats .

3.2.2 Απλή τυχαία δειγματοληψία[37]

Αφού έχουμε στα χέρια μας τον IP χάρτη του .gr το επόμενο βήμα είναι να επιλέξουμε τον τρόπο με τον οποίο θα κάνουμε δειγματοληψία πάνω σε αυτόν. Η πιο απλή μέθοδος συλλογής δειγμάτων από ένα γνωστό πληθυσμό είναι η απλή τυχαία δειγματοληψία (simple random sampling). Μία δειγματοληψία ονομάζεται απλή τυχαία δειγματοληψία αν εκλέξουμε ένα δείγμα μεγέθους *n* από ένα πληθυσμό μεγέθους *N*, έτσι ώστε κάθε στοιχείο του *n* να έχει την ίδια πιθανότητα να επιλεγεί. Στην περίπτωση αυτή το δείγμα ονομάζεται απλό τυχαίο δείγμα (simple random sample).

Στην περίπτωσή μας, για τη δειγματοληψία του .gr που μελετάμε, έστω ότι ο πληθυσμός που περιέχει το .gr ccTLD είναι *N* IP διευθύνσεις. Στον πιο πάνω πληθυσμό κάποιες IP διευθύνσεις αντιστοιχούν σε web σελίδες, ενώ κάποιες άλλες

[36] Στη διεύθυνση ftp://ftp.ripe.net μπορεί να βρει κανείς όλο το documentation του RIPE.
[37] [SMO79]

όχι. Επιλέγουμε το *i*th στοιχείο του *N* και ορίζουμε μία μεταβλητή y_i έτσι ώστε: $y_i = \begin{cases} 0, \text{αν } \textit{το } i\text{th στοιχείο δεν είναι web page} \\ 1, \text{αλλιώς} \end{cases}$. Τότε, ο συνολικός αριθμός των IP διευθύνσεων που περιέχονται στο *N* και είναι και web σελίδες είναι $\sum_{i=1}^{N} y_i$. Αν, τώρα, το ποσοστό (αναλογία) των web σελίδων στο *N* είναι *p* τότε μία εκτιμήτρια του *p* είναι η $\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n} \equiv \bar{y}$ με διακύμανση $\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right)$ όπου $\hat{q} = 1 - \hat{p}$, και όριο σφάλματος για την εκτίμηση $2\sqrt{\hat{V}(\hat{p})} = 2\sqrt{\frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right)}$. Το μέγεθος *n* του δείγματος που χρειάζεται, έτσι ώστε να εκτιμήσουμε το ποσοστό *p* με σφάλμα στην εκτίμηση μεγέθους *B*, δίνεται από τη σχέση $n = \frac{Npq}{(N-1)\frac{B^2}{4} + pq}$. Στην προηγούμενη σχέση θα πρέπει να γνωρίζουμε την αναλογία *p* έτσι ώστε να υπολογίσουμε το μέγεθος του δείγματος που θα πάρουμε. Αν την αναλογία αυτή (ή ίσως μία εκτίμησή της από προηγούμενη δειγματοληψία) δεν την έχουμε, τότε παίρνουμε μία υπερεκτίμηση του δείγματος βάζοντας *p*=0.5.

Συνοψίζοντας τα πιο πάνω έχουμε τον πίνακα 3.1 που μας δίνει η θεωρία της απλής τυχαίας δειγματοληψίας.

(Πληθυσμός, μέγεθος δείγματος):
(*N*, *n*)
Εκτιμήτρια αναλογίας *p* στον πληθυσμό:

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n} \equiv \bar{y} \qquad \textbf{(2.1)}$$

Εκτιμούμενη διακύμανση του $\hat{p}$:

$$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right) \quad (\hat{q} = 1 - \hat{p}) \qquad \textbf{(2.2)}$$

Όριο σφάλματος εκτίμησης:

$$2\sqrt{\hat{V}(\hat{p})} = 2\sqrt{\frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right)} \qquad \textbf{(2.3)}$$

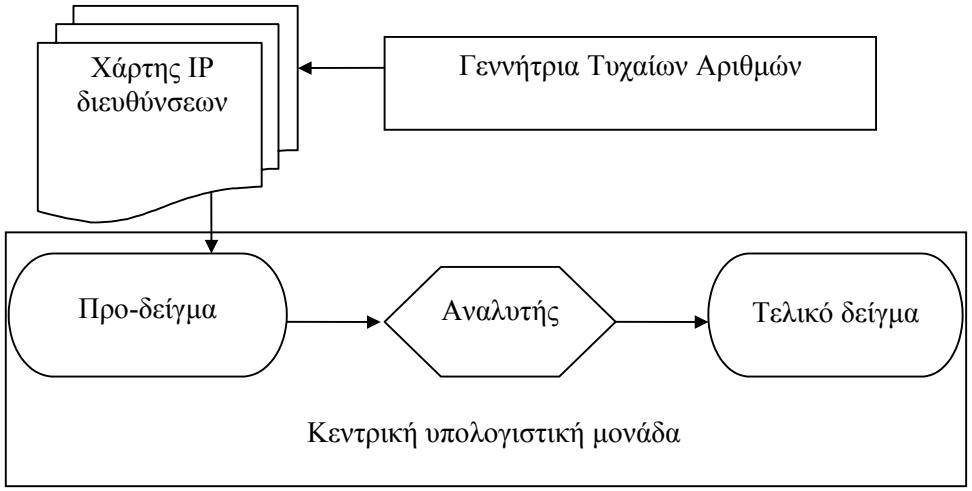
Απαιτούμενο μέγεθος δείγματος για εκτίμηση αναλογίας *p* με σφάλμα στην εκτίμηση μεγέθους *B*:

$$n = \frac{Npq}{(N-1)\frac{B^2}{4} + pq} \qquad \textbf{(2.4)}$$

Πίνακας 3.1: Απλή τυχαία δειγματοληψία

3.3 Η αρχιτεκτονική του δειγματολήπτη

Μέχρι στιγμής έχουμε δει από πού μπορούμε να βρούμε το χάρτη των IP διευθύνσεων και με ποιά μέθοδο θα γίνει η δειγματοληψία. Αυτό που τώρα χρειαζόμαστε είναι ο δειγματολήπτης ο οποίος θα πραγματοποιήσει τη δειγματοληψία. Σε αυτή την ενότητα περιγράφουμε τα διάφορα στάδια υλοποίησης ενός δειγματολήπτη η αρχιτεκτονική του οποίου φαίνεται στη φιγούρα 3.2. Ταυτόχρονα παρουσιάζουμε και την πρώτη δειγματοληψία του που έγινε στο .gr ccTLD κατά το χρονικό διάστημα 26/11 - 2/12/2002.



Φιγούρα 3.2: Η αρχιτεκτονική του δειγματολήπτη

Η δομή του δειγματολήπτη είναι πολύ απλή και αποτελείται από 3 μέρη: Το πρώτο περιέχει το χάρτη με τις IP διευθύνσεις. Το δεύτερο είναι η γεννήτρια των τυχαίων αριθμών και το τρίτο η κεντρική υπολογιστική μονάδα.

Ο δειγματολήπτης σε γενικές γραμμές λειτουργεί ως εξής: Αποθηκεύεται ο χάρτης των IP διευθύνσεων πάνω στον οποίο θα κάνουμε τη δειγματοληψία και με τη βοήθεια της γεννήτριας τυχαίων αριθμών επιλέγεται το επιθυμητού μεγέθους τυχαίο δείγμα από το χάρτη. Στη συνέχεια αυτό μετατρέπεται στο τελικό δείγμα στην κεντρική υπολογιστική μονάδα.

3.3.1 Χάρτης IP διευθύνσεων

Το τμήμα αυτό του δειγματολήπτη περιέχει το χάρτη με τις IP διευθύνσεις. Το σύνολο αυτών των διευθύνσεων είναι ο πληθυσμός της δειγματοληψίας και προέρχεται από τη διεύθυνση ftp://ftp.ripe.net/ripe/stats στην οποία κάθε μέρα φυλάσσονται όποιες νέες αναθέσεις σε IP διευθύνσεις γίνονται στην περιοχή ευθύνης του RIPE. Επειδή η δειγματοληψία ξεκίνησε στις 26/11/2002 χρησιμοποιήθηκε ο IP χάρτης του RIPE, όπως ήταν διαμορφωμένος την 26^{η} Νοεμβρίου 2002 (δηλαδή το αρχείο ripencc.20021126 από την πιο πάνω διεύθυνση). Στο αρχείο αυτό περιέχονται όλοι οι τοπικοί ληξίαρχοι (LIR) και τα αυτόνομα συστήματα (autonomous systems -

AS[38]) που μέχρι εκείνη τη χρονική στιγμή χρησιμοποιούσαν διευθύνσεις IP από τον RIPE. Ένα τμήμα του πιο πάνω αρχείου φαίνεται στον πίνακα 3.2:

```
.
.
.
ripencc|GR|ipv4|62.1.0.0|65536|2000-02-16|allocated
ripencc|SA|ipv4|62.3.32.0|8192|2002-01-09|allocated
.
.
.
ripencc|RU|asn|5467|1|1995-11-05|allocated
ripencc|GR|asn|5470|1|1995-11-24|allocated
ripencc|UK|asn|5490|1|1995-12-07|allocated
ripencc|DK|asn|5491|1|1995-12-07|allocated
.
.
.
```

Πίνακας 3.2: Μικρό τμήμα ενός IP χάρτη

Από τον πιο πάνω πίνακα μπορούμε να δούμε, για παράδειγμα, πως σε κάποιο τοπικό ληξίαρχο που δραστηριοποιείται στην Ελλάδα (GR) ανατέθηκε (allocated) από τον RIPE στις 16 Φεβρουαρίου 2000 (2000-02-16) ένα πακέτο από 2^{16}=65536 συνεχόμενες διευθύνσεις με αρχικό IP 62.1.0.0. Επίσης, παρατηρούμε ότι οι αναθέσεις στους ληξίαρχους περιέχουν το block των IP διευθύνσεων που τους έχει ανατεθεί (π.χ. στις 16 Φεβρουαρίου 2000 (2000-02-16)), αλλά στα AS μία τέτοια πληροφορία δεν υπάρχει. Για παράδειγμα από τον πιο πάνω χάρτη δε μπορούμε να γνωρίζουμε τί IP έχει πάρει ο AS με αριθμό 5470[39]. Για το λόγο αυτό από το πιο πάνω αρχείο κρατήσαμε τις εγγραφές που περιείχαν μόνο GR και οι οποίες επιπλέον αφορούσαν τοπικούς ληξίαρχους (εξαιρέσαμε δηλαδή τα AS για τα οποία δεν υπήρχε δυνατότητα μαζικής επεξεργασίας). Θεωρούμε, λοιπόν, ότι ο δειγματολήπτης δέχεται σαν είσοδο το χάρτη που περιέχει όλες τις αναθέσεις που έχουν γίνει σε ληξίαρχους. Ο χάρτης αυτός, όπως είπαμε, υπάρχει στο παράρτημα Β και περιέχει N=1,216,512 IP διευθύνσεις.

3.3.2 Γεννήτρια τυχαίων αριθμών

Κάθε δειγματολήπτης πρέπει με κάποιο τρόπο να έχει πρόσβαση σε τυχαίους αριθμούς οι οποίοι του επιτρέπουν να επιλέξει από τον πληθυσμό. Απαιτείται, λοιπόν, μία γεννήτρια τυχαίων αριθμών. Όπως, όμως, αναφέρουν οι Park και Miller [PM88], η εκλογή μίας γεννήτριας τυχαίων αριθμών δεν είναι εύκολη υπόθεση, αφού λίγες είναι αυτές που περνάνε κάποια βασικά τεστ. Στο πιο πάνω άρθρο προτείνουν μία minimal random number generator της οποίας τις προδιαγραφές (τουλάχιστον) πρέπει να πληρούν όλες οι υπόλοιπες. Η γεννήτρια που χρησιμοποιήθηκε σε αυτό το

[38] Αυτόνομο σύστημα (AS) είναι μία περιοχή του Ίντερνετ που είναι υπό τη διαχείριση μίας και μοναδικής οντότητας. Συνήθως η έννοια του AS συμπίπτει με αυτή του domain.
[39] Το AS με αριθμό 5470 είναι το Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

δειγματολήπτη βασίζεται σε μία βελτιωμένη έκδοση της minimal random number generator των Park και Miller [PFTV02][40].

Η γεννήτρια αυτή είναι έτσι κατασκευασμένη ώστε δίνοντάς της σαν σπόρο (seed) έναν αρνητικό ακέραιο και τον αριθμό των τυχαίων που επιθυμούμε, να παράγει τυχαίους που ανήκουν στο υποσύνολο των πραγματικών (0,1). Αφού στο πείραμα που κάνουμε ο χάρτης του πληθυσμού περιέχει N=1,216,512 IP διευθύνσεις σύμφωνα με τον πίνακα 3.1 το μέγεθος του δείγματος που θα πρέπει να πάρουμε είναι $n = \frac{Npq}{(N-1)\frac{B^2}{4} + pq} = \frac{1{,}216{,}512 \cdot 0.5 \cdot 0.5}{(1{,}216{,}512-1)\frac{0.01^2}{4} + 0.5 \cdot 0.5} \approx 10{,}000$. Εδώ θα πρέπει να σημειώσουμε ότι επειδή δε γνωρίζαμε την αναλογία των web σελίδων στο .gr πήραμε $p = q = 0.5$, οπότε και θα επιλέξουμε μεγαλύτερο μέγεθος δείγματος. Αφού το δείγμα είχε μέγεθος 10,000 τροφοδοτήσαμε τη γεννήτρια με seed = -10,000[41] και της ζητήσαμε να παράγει 10,000 τυχαίους. Ας σημειωθεί ότι ο σπόρος μπορεί να είναι ένας οποιοσδήποτε αρνητικός ακέραιος, όμως διαφορετικοί αρνητικοί ακέραιοι θα δώσουν διαφορετικές ακολουθίες τυχαίων αριθμών. Συνεπώς εάν κάποιος ήθελε να επαναλάβει το πιο πάνω πείραμα θα έπρεπε να τροφοδοτήσει τη γεννήτρια με τον ίδιο σπόρο (-10,000), ώστε να πάρει την ίδια ακολουθία τυχαίων.

3.3.3 Κεντρική υπολογιστική μονάδα του δειγματολήπτη

Το μέρος αυτό του δειγματολήπτη είναι το πιο βασικό και κάνει την περισσότερη δουλειά. Είναι υλοποιημένο με τεχνολογία java[42] [Sun] και αποτελείται από τρία τμήματα.

3.3.3.1 Προ-δείγμα

Το πρώτο τμήμα της κεντρικής υπολογιστικής μονάδας το ονομάσαμε προ-δείγμα. Το τμήμα αυτό διαβάζει έναν-έναν τους τυχαίους και επιλέγει τον IP που αντιστοιχεί στον τυχαίο από το χάρτη. Όπως είδαμε, η πιο πάνω γεννήτρια είναι έτσι κατασκευασμένη ώστε να βγάζει αριθμούς που ανήκουν στο ανοικτό διάστημα (0,1). Οι αριθμοί αυτοί έχουν 6 σημαντικά ψηφία (π.χ. 0.189623) και θα πρέπει με κάποιο τρόπο να αντιστοιχούν στο χάρτη με τις IP διευθύνσεις, ώστε να ξέρει ο δειγματολήπτης πως μόλις συναντήσει τον τάδε τυχαίο να πάει να ψάξει την IP διεύθυνση που του αντιστοιχεί. Έστω, λοιπόν, πως στο χάρτη με τις IP διευθύνσεις περιέχονται N συνολικά IP. Ο δειγματολήπτης εφαρμόζει ένα γραμμικό μετασχηματισμό (από τους πραγματικούς στους φυσικούς) $M : (0,1) \to (1, N)$ μετατρέποντας τον τυχαίο αριθμό από το υποσύνολο των πραγματικών (0,1) στο υποσύνολο των φυσικών $(1, N)$. Ο γραμμικός μετασχηματισμός είναι αυτός του πίνακα 3.3:

---

[40] Σε αυτό το σημείο πρέπει να ευχαριστήσουμε τον πολύ καλό φίλο και συνάδελφο κ. Δημήτρη Κουβέλη, ο οποίος βοήθησε στον προγραμματισμό σε C της γεννήτριας των τυχαίων αριθμών [KR88].

[41] Είναι απλή σύμπτωση ότι στη συγκεκριμένη περίπτωση η τιμή του σπόρου είναι κατά απόλυτη τιμή ίση με τον αριθμό των τυχαίων που θέλουμε να πάρουμε.

[42] Χρησιμοποιήθηκε το java ™ 2 software development kit (j2sdk), standard edition, version 1.4.0 και τα βιβλία [Fla00], [Fla02].

1. Πολλαπλασίασε τον τυχαίο με το *N* 2. Στρογγυλοποίησε στον πλησιέστερο φυσικό

Πίνακας 3.3: Γραμμικός μετασχηματισμός

Για παράδειγμα στη δειγματοληψία του .gr ο πρώτος τυχαίος που έδωσε η γεννήτρια ήταν ο 0.189623. Εφαρμόζοντας τον πιο πάνω γραμμικό μετασχηματισμό ο τυχαίος 0.189623 αντιστοιχίζεται στον 230,679, διότι έχουμε: 0.189623·1,216,512 = 230,678.65 ≈ 230,679 με στρογγυλοποίηση στον κοντινότερο φυσικό. Αυτό, δηλαδή, σημαίνει ότι ο τυχαίος 0.189623 είναι ο $230,679^{ος}$ IP από την αρχή της λίστας του IP χάρτη. Στη συνέχεια ο δειγματολήπτης επιλέγει το IP που βρίσκεται στη $230,679^{η}$ θέση της λίστας (βλέπε και παράρτημα Β) σύμφωνα με τον εξής αλγόριθμο:

Βρες σε ποιά γραμμή του χάρτη ανήκει η σειρά 230,679 (ανήκει στην $6^{η}$ γραμμή του χάρτη). Αυτό εύκολα βρίσκεται αν αρχίζουμε και προσθέτουμε τα στοιχεία της $5^{ης}$ στήλης του χάρτη μέχρι να φτάσουμε ή να ξεπεράσουμε τη θέση που θέλουμε να βρούμε (65,536 + 65,536 + 8,192 + 8,192 + 65,536 + 32,768= 245,760 > 230,679). Όσα στοιχεία της $5^{ης}$ στήλης του χάρτη προσθέσαμε τόσες γραμμές πρέπει να κατεβούμε για να βρούμε τη θέση.

Βρες σε ποιά θέση x της $6^{ης}$ γραμμής του χάρτη ανήκει η σειρά 230,679
x = 230,679 - 65,536 - 65,536 - 8,192 - 8,192 - 65,536 = 17,687

Θέσε α=62, β=75, γ=0, δ=0 και εκτέλεσε:

x = x -1
α = α + [x/2^{24}] &[43] x = x - 2^{24}[x/2^{24}]-1
β = β + [x/2^{16}] & x = x - 2^{16}[x/2^{16}]
γ = γ + [x/2^{8}] & x = x - 2^{8}[x/2^{8}]
δ= δ + x

αν
δ > 255 τότε (δ = δ - 256, γ = γ + 1)
αν
γ > 255 τότε (γ = γ - 256, β = β + 1)
αν
β > 255 τότε (β = β - 256, α = α + 1)
αν
α > 255 σφάλμα (δεν επιτρέπεται α>255)

τελικά, ο αλγόριθμος εκτελείται μία φορά και δίνει:

α=**62**, β=**75**, γ=**69**, δ=**22**

Δηλαδή, στη $230,670^{η}$ θέση του χάρτη βρίσκεται ο IP: 62.75.69.22 και, άρα, σε αυτό το IP αντιστοιχεί ο τυχαίος 0.189623. Η πιο πάνω διαδικασία γίνεται για κάθε έναν από τους τυχαίους αριθμούς που έχει παράγει η γεννήτρια, και όλοι οι IP αριθμοί που προκύπτουν αποτελούν ένα προ-δείγμα. Το ονομάζουμε προ-δείγμα διότι, όπως

[43] [x] σημαίνει ακέραιο μέρος του x.

είδαμε, στην 2.3.2 δεν αποτελούν όλοι από αυτούς έγκυρες web σελίδες. Αυτό το προ-δείγμα[44] θα πρέπει με κάποιο τρόπο να αναλυθεί ώστε να προκύψει το τελικό δείγμα με τις web σελίδες.

3.3.3.2 Αναλυτής

Την πιο πάνω δουλειά της ανάλυσης (φιλτραρίσματος) του προ-δείγματος, ώστε να προκύψει το τελικό δείγμα που θα περιέχει κόμβους του web, την κάνει το δεύτερο τμήμα της κεντρικής υπολογιστικής μονάδας, ο αναλυτής. Ο αναλυτής ελέγχει κάθε IP που περιέχεται στο προ-δείγμα αν είναι IP που αντιστοιχεί σε έγκυρη web σελίδα. Ο πιο πάνω έλεγχος βασίζεται, όπως είπαμε, στο γεγονός ότι οι web σελίδες συνήθως "ακούνε" στο port 80 των host.

Ο αναλυτής παίρνει το IP και προσπαθεί να κάνει μία http σύνδεση με το port 80 του host ζητώντας από αυτόν μία απάντηση. Οι απαντήσεις που μπορεί να επιστρέψει ο host είναι πολλές. Για παράδειγμα, μπορεί να επιστρέψει τον κωδικό 200 που σημαίνει ότι "είναι όλα εντάξει και περιμένω να εξυπηρετήσω τα αιτήματά σου", μπορεί να επιστρέψει κωδικό 404 που σημαίνει ότι "δεν εξυπηρετώ στο port 80 τίποτα". Μπορεί ακόμα να μην επιτευχθεί και καθόλου σύνδεση, οπότε θα έχουμε timeout και ο αναλυτής θα προχωρήσει στον επόμενο IP αποκλείοντας τον προηγούμενο από το τελικό δείγμα. Στον πίνακα 3.4 βλέπουμε μερικούς από τους πιο συχνούς κωδικούς που μπορεί να επιστρέψει ένας host και την ερμηνεία τους[45] για την έκδοση 1.1 του πρωτοκόλλου http.

http κωδικός	**Ερμηνεία**
200	request succeeded
400	bad request
401	unauthorized
403	forbidden
404	not found
407	proxy authendication required
408	request timeout
500	internal server error
502	bad gateway
503	service unavailable
504	gateway timeout

Πίνακας 3.4: Συχνές απαντήσεις πρωτοκόλλου http v.1.1

Αφού ο αναλυτής φιλτράρει όλο το προ-δείγμα τότε αποθηκεύει τους IP που επέλεξε και δημιουργεί το τελικό δείγμα, το οποίο είναι και η έξοδος του δειγματολήπτη.

Το τελικό δείγμα μπορεί να υποστεί μία επιπλέον επεξεργασία διότι είναι στη μορφή IP διευθύνσεων και όχι στη μορφή που γράφουμε συνήθως τις web σελίδες (π.χ. www.ntua.gr, www.hotmail.com κ.λπ.). Για παράδειγμα, στο τελικό δείγμα

[44] Το προ-δείγμα φαίνεται στον Πίνακα Π.γ.1 στο παράρτημα Γ και στην http://users.ntua.gr/plekeas/Data_sets/ Sampling_a_web_subgraph/presample_gr.txt
[45] RFC 2616 [RFC].

ανήκει το IP 212.251.52.104 το οποίο αντιστοιχεί στην www.air-conditioners.gr . Θα βοηθούσε, λοιπόν, μία επιπλέον επεξεργασία η οποία θα μας έδινε τα ονόματα (γνωστά και ως aliases) των IP. Κάτι τέτοιο, όμως, όπως θα δούμε και πιο κάτω δεν είναι εύκολο να γίνει και ένας από τους λόγους είναι, όπως αναφέρθηκε στην 2.4.2, το virtual hosting. Προς το παρόν, όμως, σώζει την κατάσταση η παρατήρηση ότι: πρόσβαση στις web σελίδες του τελικού δείγματος μπορεί να γίνει και μέσω των IP αριθμών και όχι αποκλειστικά μόνο με τα aliases αυτών των IP[46].

Τελικά, το δείγμα που προέκυψε από τη δειγματοληψία του .gr φαίνεται στον Πίνακα Π.γ.2 στο παράρτημα Γ και υπάρχει και στη διεύθυνση http://users.ntua.gr /plekeas/Data_sets/ Sampling_a_web_subgraph/finalsample_gr.txt .

3.4 Δοκιμάζοντας την αξιοπιστία του δειγματολήπτη

Σε αυτή την ενότητα ελέγχεται κατά πόσο το δείγμα που προέκυψε από τη δειγματοληψία του .gr είναι αξιόπιστο. Ο έλεγχος γίνεται με δύο τεστ. Το πρώτο τεστ σχετίζεται με τα στατιστικά στοιχεία που μπορούν να εξαχθούν από το δείγμα σχετικά με το .gr ccTLD. Το δεύτερο τεστ σχετίζεται με την κατανομή power-law που ακολουθούν οι out-degree του τελικού δείγματος.

3.4.1 Τα αποτελέσματα ενός project

Η ερευνητική μας ομάδα συμμετείχε μαζί με άλλες από το Πανεπιστήμιο Πατρών και το Οικονομικό Πανεπιστήμιο Αθηνών σε ένα project στο οποίο έγινε χρήση του AVS SDK (Alltavista Software Developers Kit) INDEXER, το οποίο παίρνοντας συγκεκριμένα queries δημιουργεί index files, που χρησιμοποιούνται για να εξαχθεί ολόκληρος ο κατάλογος του .gr. Εμείς θα χρησιμοποιήσουμε αυτό τον κατάλογο για να ελέγξουμε το δείγμα μας[47]. Μπορεί κάποιος να κατεβάσει αυτό τον κατάλογο (ουσιαστικά πρόκειται για όλο το ελληνικό web όπως ήταν στα μέσα περίπου του 2001) από τη διεύθυνση http://users.ntua.gr/plekeas/Pened/gr_domain . Σύμφωνα με τον πιο πάνω χάρτη στα μέσα του 2001 ο πληθυσμός των web sites που τελείωναν σε .gr ήταν 6,858. Αν, τώρα, ανατρέξουμε στις βιβλιοθήκες του RIPE θα δούμε ότι όλο το Ελληνικό web την εποχή εκείνη περιείχε περίπου 491,520 IP σε τοπικούς ληξίαρχους, 912,976 IP σε AS και συνολικά 1,404,496 IP. Αυτό σημαίνει ότι σύμφωνα με το project το ποσοστό των .gr web sites ήταν περίπου $\frac{6,858}{1,404,496} \approx 4.88 \cdot 10^{-3}$ web sites.

3.4.2 Στατιστικά του τελικού δείγματος

Στη δειγματοληψία του .gr ο πληθυσμός ήταν 1,216,512 ενώ το δείγμα περιείχε 10,000 IP, και από αυτά τα 110 μόνο απάντησαν στο port 80. Αν υπολογίσουμε την αναλογία p θα δούμε ότι $\hat{p} = 1.1000 \cdot 10^{-2}$ με

[46] Για παράδειγμα, είτε γράψουμε σε ένα browser http://195.167.36.46 είτε http://195.167.36.46:80 είτε http://www.athensmap.gr είτε http://www.athensmap.gr:80 έχουμε πρόσβαση στην ίδια web σελίδα.

[47] Το project αυτό ήταν ένα ΠΕΝΕΔ χρηματοδοτούμενο από τη ΓΓΕΤ με τίτλο "Υποστήριξη αποφάσεων μικροοικονομικής διαχείρισης μέσω τεχνικών εξόρυξης και βελτιστοποίησης", και εμείς συμμετείχαμε σε αυτό για το διάστημα 1/11/2000 – 31/3/2001. Αν λάβουμε υπ΄ όψιν ότι το ΠΕΝΕΔ έληξε στις 31/3/2001, ο χάρτης του .gr που θα χρησιμοποιήσουμε για να ελέγξουμε το τελικό δείγμα τοποθετείται χρονολογικά στα μέσα του 2001.

$V(\hat{p}) = \frac{0.011 \cdot 0.989}{10{,}000 - 1}\left(\frac{1{,}216{,}512 - 10{,}000}{1{,}216{,}512}10{,}000\right) \approx 1.1 \cdot 10^{-6} = 0.0001 \cdot 10^{-2}$ και

$2\sqrt{V(\hat{p})} = 2 \cdot \sqrt{0.0001 \cdot 10^{-2}} = 0.002$ Άρα, $\hat{p} = (1.10 \pm 0.20) \cdot 10^{-2}$ η αναλογία web σελίδων στο .gr στα τέλη του 2002.

Αν προσπαθήσουμε, όμως, να συνδεθούμε με κάποιο Browser με τα 110 IP του τελικού δείγματος θα πάρουμε τις πιο κάτω απαντήσεις οι οποίες συνοψίζονται στον πίνακα 3.5.

α/α	**Πλήθος IP από το τελικό δείγμα (12/2002)**
1	40 IP ήταν web pages που τελείωναν σε .gr
2	33 IP απάντησαν με Can't find server or DNS error
3	13 IP απάντησαν με http 404 File not Found
4	6 IP απάντησαν με http 403 forbidden
5	5 IP είχαν domain names με ψευδώνυμο (alias) που έληγε σε .com
6	5 IP απάντησαν με http 401 unauthorised
7	3 IP απάντησαν με access denied (firewall)
8	2 IP απάντησαν με web page under construction
9	2 IP ήταν server test pages
10	1 IP απάντησε με asp error

Πίνακας 3.5: Απαντήσεις server σε αιτήματα σύνδεσης με τα IP του τελικού δείγματος (12/2002)

Αν, επιπλέον, υπολογίσουμε και τις αναθέσεις που έχουν γίνει και στα AS[48] (περίπου 1,057,872 IP για τη χρονική περίοδο που συζητάμε), τότε ο πραγματικός πληθυσμός θα είναι $N' = 2{,}274{,}384$ και αναλογικά από το $\hat{p}$ βρίσκουμε ότι στα τέλη του 2002 υπήρχαν 25,018 web pages που τελείωναν σε .gr. Τα πιο πάνω αποτελέσματα φαίνονται συνοπτικά στον πίνακα 3.6.

Αριθμός IP που ανατέθηκαν στην Ελλάδα (οριζόντια) **Για το εξάμηνο** (κάθετα)	**Σε**		**Συνολικός πληθυσμός IP**	**Αριθμός web sites**
	Ληξίαρχους	**AS**		
Πρώτο 2001	491,520	912,976	1,404,496	6,858
Δεύτερο 2002	1,216,512	1,057,872	2,274,384	25,018

Πίνακας 3.6: Συνοπτική απεικόνιση IP αριθμών και εκτίμηση των web sites για την περίπτωση του .gr ccTLD

Σχολιάζοντας τα πιο πάνω θα μπορούσαμε να πούμε τα εξής:

Το ποσοστό $\hat{p}$ που υπολογίσαμε βλέπουμε ότι είναι μεγαλύτερο από το ποσοστό που έδωσε το project της εξαντλητικής αναζήτησης. Κάτι τέτοιο είναι

[48] βλ. 3.5.3

φυσιολογικό να συμβαίνει διότι, θεωρούμε ότι παντού (άρα και στην Ελλάδα) το web αυξάνει το μέγεθός του. Επίσης, το ποσοστό που υπολογίσαμε είναι αξιόπιστο δεδομένου ότι αν το συγκρίνουμε με τις τιμές που δίνει η στατιστική υπηρεσία του RIPE[49] θα δούμε ότι είναι πολύ κοντά (ο RIPE για την περίοδο που μελετάμε δίνει για το .gr ccTLD αριθμό www sites: 22,996).

Παρατηρώντας τον πίνακα 3.5 θα δούμε ότι το 33% του δείγματος ήταν IP που όταν ζητήσαμε τις http σελίδες τους απάντησαν με το μύνημα "can't find server or DNS error". Αυτό εξηγείται αν σκεφτούμε ότι, ουσιαστικά, η πιο πάνω δειγματοληψία ήταν μία δειγματοληψία στους ληξίαρχους που δραστηριοποιούνται στο .gr ccTLD. Αυτό σημαίνει πολλά, αν σκεφτούμε ότι οι ληξίαρχοι είναι κυρίως ISPs (Internet Service Providers – π.χ. otenet, acn, forthnet, hol) οι οποίοι προσφέρουν και υπηρεσίες, dial up σύνδεσης με το internet. Πράγματι, παρατηρώντας το προ-δείγμα θα δούμε ότι το 1/3 ήταν ppp (dial up) συνδέσεις μέσω διαφόρων ελληνικών ISPs. Αν, τώρα, στη δειγματοληψία είχαμε συμπεριλάβει και το χάρτη αναθέσεων στους AS, τότε τα πράγματα θα ήταν διαφορετικά, αφού στα AS περιλαμβάνονται Πανεπιστήμια, Ερευνητικά Ιδρύματα, Σχολεία κ.α. τα οποία παρέχουν μεν αλλά όχι τόσες πολλές dial up συνδέσεις. Δηλαδή, με αυτό τον τρόπο πιστεύουμε ότι η αναλογία $\hat{p}$ θα ήταν μεγαλύτερη.

Όπως είχαμε πει στην 1.2.3 ένα από τα προβλήματα που συναντούν οι crawlers είναι και το πρόβλημα της προσπέλασης των διαφόρων host, είτε λόγω όχι καλής σύνδεσης είτε γιατί απαιτείται αναγνώριση του χρήστη κ.λπ. Πράγματι, αυτό το πρόβλημα το παρατηρήσαμε και εδώ, αφού όπως βλέπουμε και στον πίνακα 3.5 περίπου το 10% του δείγματος δεν μπορούσε να προσπελαστεί από τον Browser (File not Found). Για να ελέγξουμε το πιο πάνω ποσοστό προσπαθήσαμε να ξανασυνδεθούμε με κάθε ένα από τα 110 IP του τελικού δείγματος περίπου 1 μήνα μετά από την πρώτη μας προσπάθεια. Τα αποτελέσματα τα συνοψίζουμε στον πιο κάτω πίνακα:

α/α	**Πλήθος IP από το τελικό δείγμα (1/2003)**
1	35 IP ήταν web pages που τελείωναν σε .gr
2	39 IP απάντησαν με Can't find server or DNS error
3	12 IP απάντησαν με http 404 File not Found
4	3 IP απάντησαν με http 403 forbidden
5	5 IP είχαν domain names με ψευδώνυμο (alias) που έληγε σε .com
6	5 IP απάντησαν με http 401 unauthorised
7	3 IP απάντησαν με access denied (firewall)
8	5 IP απάντησαν με web page under construction
9	2 IP ήταν server test pages
10	1 IP απάντησε με asp error

Πίνακας 3.7: Απαντήσεις server σε αιτήματα σύνδεσης με τα IP του τελικού δείγματος (1/2003)

[49] Η στατιστική υπηρεσία του RIPE ονομάζεται hostcount και βρίσκεται στη διεύθυνση http://www.ripe.net/ripencc/pub-services/stats/hostcount.

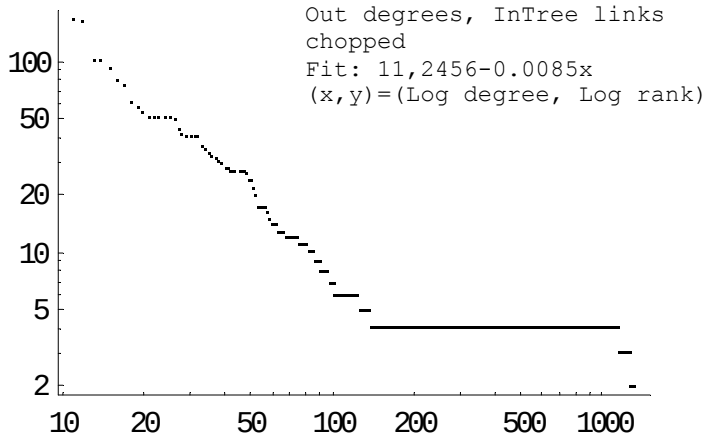
Βλέπουμε, λοιπόν, ότι το ίδιο δείγμα σε διάστημα 1 μηνός έδωσε διαφορετική συμπεριφορά.

3.4.3 Κατανομή power law για τους out-degree του τελικού δείγματος

Παρατηρήσαμε από τα προηγούμενα ότι το τελικό δείγμα μας έδωσε στατιστικές πληροφορίες οι οποίες συμφωνούν με τρίτες ανεξάρτητες πηγές. Αυτό είναι ένδειξη ότι η συμπεριφορά του δείγματος είναι καλή και άρα ότι ο δειγματολήπτης λειτουργεί σωστά.

Ένα επιπλέον τεστ το οποίο φανερώνει φυσιολογική συμπεριφορά ενός δείγματος του web γράφου είναι το power law τεστ. Επειδή, όπως είπαμε και στην 1.2.4, οι power law φαίνεται να είναι μία ενδογενής ιδιότητα του web γράφου [BKM+00] θα πρέπει ένα αντιπροσωπευτικό τμήμα του να έχει παρόμοια συμπεριφορά. Θεωρήσαμε, λοιπόν, καλό να υποβάλουμε σε έλεγχο το τελικό δείγμα για να δούμε αν εμφανίζει αυτή τη συμπεριφορά.

Για το σκοπό αυτό κάναμε crawl όλες τις web pages που ανήκαν σε όλα τα web sites του τελικού δείγματος και υπολογίσαμε τα out-degree τους[50]. Στη φιγούρα 3.3 παρατηρούμε την power law κατανομή (log degree, log rank) για τα πιο πάνω out-degree. Για τον υπολογισμό αυτής της φιγούρας χρησιμοποιήθηκε η προσέγγιση ότι ένα web site μπορεί να θεωρηθεί σαν ένα δέντρο με ρίζα τη home page του και φύλλα τις υπόλοιπες web σελίδες κάτω από αυτήν. Σαν out-degree υπολογίστηκαν μόνο όσα link κάθε σελίδας έβγαιναν εκτός του πιο πάνω δέντρου και όχι όσα έδειχναν σε άλλες σελίδες εντός.



Φιγούρα 3.3: Κατανομή out-degree για τα web site του τελικού δείγματος

Η γραφική παράσταση παρατηρούμε ότι κυρτώνει πάρα πολύ στο y = 4, πράγμα το οποίο οφείλεται σε ένα πορνογραφικό web site με εκατοντάδες επαναλήψεις των ίδιων link στις σελίδες του.

[50] βλ. και κεφ. 5 για το πώς έγινε το crawl και το πώς υπολογίστηκαν τα out-degree.

3.5 Περιορισμοί και προβλήματα

Μέχρι στιγμής είδαμε πώς μπορεί να εφαρμοστεί η μέθοδος IP sampling σε ένα συγκεκριμένο ccTLD. Οι περιορισμοί και τα προβλήματα που συναντήσαμε κατά την εκτέλεση των πιο πάνω πειραμάτων είναι τα ίδια με αυτά της εφαρμογής της μεθόδου για όλο τον IPv4 χώρο. Τα πιο σημαντικά προβλήματα είναι το virtual hosting και η δειγματοληψία web σελίδων από ένα web site.

3.5.1 Virtual Hosting

Όπως είχαμε αναφέρει και στην 2.4.2, η τεχνική του virtual hosting εφαρμόζεται από τους διαχειριστές των web server με σκοπό να φιλοξενούνται πολλά web sites μέσα στον ίδιο server και πολλές φορές κάτω από το ίδιο IP (name based virtual hosting).

Ας πάρουμε, για παράδειγμα, ένα server Apache [Wai99]. Όταν ο διαχειριστής του εφαρμόσει το name based virtual hosting τότε κάθε εικονικός (virtual) host δεν έχει το δικό του IP αλλά όλοι βρίσκονται κάτω από το ίδιο IP. Όταν ένας Apache ρυθμίζεται για name based virtual hosting τότε ο διαχειριστής του δηλώνει το IP που θα φιλοξενεί τα διάφορα web site με τη δήλωση *NameVirtualHost*. Κάθε web site κάτω από αυτόν τον virtual host μπορούμε να το δηλώσουμε με τον κώδικα της φιγούρας 3.4:

```
<VirtualHost IPnumber>

ServerName .....

.....virtual host directives.....

</VirtualHost>
```

Φιγούρα 3.4: Δήλωση virtual host σε Apache server

Στη φιγούρα 3.5 βλέπουμε το virtual hosting των web site www.websiteA.com και web2.lab.ece.ntua.gr κάτω από το IP 104.7.212.214[51]. Όταν, τώρα, ο Apache πάρει ένα αίτημα για το IP 104.7.212.214 βλέπει ότι πρόκειται για name based virtual host και ελέγχει αν το πρόθεμα Host που του στέλνει αυτός που ζητά το αίτημα (ο client δηλαδή), περιέχει ένα έγκυρο ServerName. Αν, π.χ., στείλουμε αίτημα για το IP 104.7.212.214 με Host: www.websiteA.com τότε θα συνδεθούμε με το συγκεκριμένο web site και όχι με κάποιο άλλο (π.χ. το web2.lab.ece.ntua.gr). Αν, όμως, ο client στείλει στο αίτημα μόνο το IP και δεν δίνει πληροφορία για τον Host, τότε ο Apache δε γνωρίζει τί web site πρέπει να επιστρέψει, οπότε επιστρέφει κάποιο μύνημα λάθους. Είναι βέβαια δυνατόν να μην επιστρέφεται μύνημα λάθους αλλά ένα default web site (συνήθως αυτό της πρώτης δήλωσης). Έτσι, όμως, ενώ υπάρχουν περισσότερα από ένα web site στο ίδιο IP εμείς ανακαλύπτουμε μόνο το default. Επιπλέον, δεν υπάρχει κάποια τεχνική που να σου επιστρέφει όλα τα web site ενός

[51] Το IP και τα web site είναι υποθετικά.

virtual host[52]. Βλέπουμε, λοιπόν, ότι η πιο πάνω τεχνική του virtual hosting εισάγει μία μεροληψία στο δείγμα υπέρ των default web site για όσα IP του τελικού δείγματος κάνουν virtual hosting.

```
<VirtualHost 104.7.212.214>

ServerName  www.websiteA.com

.....virtual host directives.....

</VirtualHost>


<VirtualHost 104.7.212.214>

ServerName  web2.lab.ece.ntua.gr

.....virtual host directives.....

</VirtualHost>
```

Φιγούρα 3.5: Δηλώσεις virtual host για το IP: 104.7.212.214

3.5.2 Πρόβλημα δειγματοληψίας web σελίδων από web site

Όπως έχουμε δει, η τεχνική του IP sampling δημιουργεί τα δείγματα των web σελίδων έμμεσα παίρνοντας πρώτα ένα μεγαλύτερο δείγμα από IP διευθύνσεις, κρατώντας ύστερα όποιες από αυτές φιλοξενούν web sites και επιλέγοντας, τελικά, web σελίδες από αυτά τα web sites. Όμως, είχαμε αναφέρει στην 1.3 και στην 2.4.3 ότι ένα από τα μειονεκτήματα του IP sampling είναι το ότι δεν έχει βρεθεί τεχνική για ομοιόμορφη δειγματοληψία web σελίδων μέσα από ένα web site.

Μία προσέγγιση του πιο πάνω προβλήματος είναι να αρκεστούμε στο δείγμα των web site που έχουμε και να θεωρήσουμε ότι οι web σελίδες του τελικού δείγματός μας είναι οι αρχικές σελίδες (οι home pages δηλαδή) των web site. Έτσι, όμως, το δείγμα μας θα περιέχει μόνο home pages. Μία άλλη προσέγγιση είναι η εξαντλητική αναζήτηση όλων των web site που υπάρχουν στο δείγμα[53]. Αυτή, βέβαια, η προσέγγιση ξεφεύγει από την έννοια της δειγματοληψίας.

Πάντως, μέχρι στιγμής το πρόβλημα της ομοιόμορφης δειγματοληψίας web σελίδων από web site παραμένει άλυτο.

[52] Στην πραγματικότητα υπάρχει μία τεχνική η οποία λέγεται zone transfer που σου επιτρέπει να δεις ποιά web sites έχουν γίνει configured σε ένα IP. Παρόλο που ο RIPE επικροτεί τη χρήση του zone transfer είναι πάγια τακτική των διαχειριστών να την απαγορεύουν για λόγους ασφαλείας.
[53] βλ. κεφ. 5.

3.5.3 Προσδιορισμός IP αναθέσεων σε αυτόνομα συστήματα

Στην 3.3.1 είδαμε ότι ο χάρτης των IP διευθύνσεων, όπως δίνεται από τον RIPE, δεν αναφέρει τί αναθέσεις έχουν γίνει στα διάφορα αυτόνομα συστήματα ενός συγκεκριμένου ccTLD. Στη δειγματοληψία που περιγράψαμε προηγουμένως παραλήφθηκαν τα AS με αποτέλεσμα ο χάρτης που χρησιμοποιήσαμε για τη δειγματοληψία του .gr να μην είναι ακριβής.

Ένας τρόπος για να υπολογίσει κανείς τις αναθέσεις στα διάφορα AS είναι ο εξής: Ο RIPE φυλάσσει όλα τα δεδομένα που σχετίζονται με το internet στην περιοχή ευθύνης του σε μία βάση δεδομένων που ονομάζεται whois[54]. Η βάση αυτή περιέχει αντικείμενα (objects) τα οποία περιέχουν διάφορες πληροφορίες. Ένα αντικείμενο χρήσιμο για τον υπολογισμό των αναθέσεων IP σε AS ονομάζεται route. Έστω ότι έχουμε το AS με αριθμό 6744. Μπορούμε να ρωτήσουμε τη βάση για το τί IP έχουν ανατεθεί στο συγκεκριμένο AS με το ερώτημα: *-r -T route -i origin AS6744*. Η απάντηση που θα πάρουμε φαίνεται στη φιγούρα 3.6 και μας λέει ότι στο AS6744 έχουν ανατεθεί $2^{32-16} = 2^{16}$ συνεχόμενες IP διευθύνσεις με αρχή την 150.140.0.0 .

```
route:          150.140.0.0/16
descr:          PATRASNET
origin:         AS6744
remarks:        The core of PATRASnet is
                CTInet (the CTI Local Area
                Network)
remarks:        which provides the
                gateways to national and
                international
remarks:        computer networks.
mnt-by:         CTINET-NOC
changed:        netmgr@cti.gr 19990510
source:         RIPE
```

Φιγούρα 3.6: Απάντηση του RIPE whois server σε ερώτημα σχετικά με ένα αντικείμενο route

Ο πιο πάνω τρόπος ερώτησης της βάσης μπορεί να αυτοματοποιηθεί ώστε να ερωτάται αυτόματα η βάση από κάποιο πρόγραμμα και να επιστρέφεται το αντικείμενο route.

3.6 Επίλογος και γενική αποτίμηση του δειγματολήπτη

Στο κεφάλαιο αυτό περιγράψαμε την υλοποίηση και τη λειτουργία ενός IP δειγματολήπτη. Αυτός ο δειγματολήπτης, αντίθετα με τις μέχρι τώρα εφαρμογές σε όλο τον IPv4 χώρο της μεθόδου IP Sampling, κάνει δειγματοληψία σε συγκεκριμένα ccTLD του Internet χρησιμοποιώντας χάρτες IP διευθύνσεων από τους αντίστοιχους

[54] Η whois βάση χρησιμοποιεί τη γλώσσα RPSL (http://www.ripe.net/ripencc/pub-services/db/rpsl/index.html) και το documentation της βάσης βρίσκεται στην http://www.ripe.net/ripe/docs/databaseref-manual.html.

ηπειρωτικούς ληξίαρχους (RIR). Σε αυτό το σημείο ο δειγματολήπτης πλεονεκτεί έναντι ενός random walk, διότι ο τελευταίος δεν μπορει να περιοριστεί σε ένα ccTLD χωρίς την εισαγωγή μεροληψίας υπέρ κάποιων web σελίδων. Ο δειγματολήπτης δίνει αξιόπιστα αποτελέσματα τα οποία συμβαδίζουν με τρίτες έγκυρες πηγές μέτρησης. Είναι, επίσης, υλοποιημένος σε java και τρέχει σε οποιαδήποτε πλατφόρμα χωρίς να απαιτεί ιδιαίτερη υπολογιστική ισχύ (π.χ. η δειγματοληψία έγινε σε απλό PC, ο κώδικας είναι μόλις 4.2KB και απαιτείται μόνο μία απλή σύνδεση με το Internet) και, επιπλέον, τα αποτελέσματα μπορούν να χωρέσουν σε μία απλή δισκέττα 1.44".

Στα μειονεκτήματα του δειγματολήπτη μπορούμε να αναφέρουμε όλα αυτά που έχει η μέθοδος IP Sampling, δηλ.: i) το virtual hosting, ii) το πρόβλημα δειγματοληψίας web σελίδων από web site, και iii) το γεγονός ότι η μορφή του τελικού δείγματος είναι στο μεγαλύτερο ποσοστό IP αριθμοί και δεν είναι στη μορφή που συνήθως γράφουμε τις web σελίδες.

Κλείνουμε αυτό το κεφάλαιο με τις εξής 2 παρατηρήσεις: 1) ο δειγματολήπτης που περιγράψαμε βγάζει ομοιόμορφα και τυχαία δείγματα web σελίδων τα οποία μπορούν να δωθούν σαν σελίδες αφετηρίας σε δειγματολήπτες με random walk ελαττώνοντας κάπως τη μεροληψία της πρώτης σελίδας που αναφέραμε στην 2.2.4 . 2) ο δειγματολήπτης από τη φύση του μπορεί να χρησιμεύσει όχι μόνο για εξαγωγή δειγμάτων του web αλλά και του Internet. Αυτή η τελευταία παρατήρηση αποτελεί και μία από τις κεντρικές ιδέες του επόμενου κεφαλαίου.

4 ΧΡΗΣΗ ΤΟΥ IP ΔΕΙΓΜΑΤΟΛΗΠΤΗ ΓΙΑ ΕΞΑΓΩΓΗ ΧΩΡΙΚΩΝ ΚΑΙ ΧΡΟΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΩΝ ΑΠΟ HOSTNAMES[55]

Στο κεφάλαιο 3 είδαμε την υλοποίηση ενός IP δειγματολήπτη και τη χρησιμοποίησή του για την παραγωγή ενός ομοιόμορφου δείγματος από web σελίδες. Στο κεφάλαιο αυτό θα χρησιμοποιήσουμε το δειγματολήπτη όχι για να εξάγουμε πληροφορίες που αφορούν το web αλλά πληροφορίες που αφορούν το Internet. Συγκεκριμένα, θα πάρουμε ένα δείγμα από hostnames που ανήκουν στο .uk ccTLD και επεξεργάζοντάς το θα βγάλουμε χωρικές (spatial) και χρονικές (temporal) πληροφορίες για τους ιδιοκτήτες αυτών των hostname.

4.1 Εισαγωγή

Όπως είδαμε στα προηγούμενα είναι δυνατόν να κάνουμε δειγματοληψία μέσα σε ένα συγκεκριμένο ccTLD. Όπως, επίσης, έχουμε πει για κάθε κράτος υπάρχει και το αντίστοιχο ccTLD. Θα μπορούσε, λοιπόν, κάποιος να πει ότι έχοντας τις IP διευθύνσεις ενός ccTLD ουσιαστικά έχει και τον IP χάρτη της αντίστοιχης γεωγραφικής περιοχής. Η πιο πάνω προσέγγιση, όμως, δεν είναι ακριβώς σωστή, διότι ένας host που χρησιμοποιεί, π.χ., το Γαλλικό domain .fr (France) δεν είναι απαραίτητο να έχει φυσική παρουσία στη Γαλλία (δηλ. δεν είναι απαραίτητο οι server του να είναι στη Γαλλία). Για παράδειγμα, σύμφωνα με τον Paltridge [Pal99] περίπου το 1% των περιεχομένων του .gr ccTLD βρίσκεται σε μηχανήματα στον Καναδά. Παρά την πιο πάνω παρατήρηση σύμφωνα με τον Paltridge το να προσεγγίσουμε τις IP διευθύνσεις μίας γεωγραφικής περιοχής με τις IP αναθέσεις που έχουν γίνει στο αντίστοιχο ccTLD της περιοχής μπορεί να γίνει με αρκετά μεγάλη εμπιστοσύνη.

Δεχόμενοι, λοιπόν, την πιο πάνω προσέγγιση έστω ότι θέλουμε να κάνουμε μία δειγματοληψία στη γεωγραφική περιοχή της Μεγάλης Βρετανίας. Επιλέγουμε το .uk ccTLD και σχηματίζουμε τον αντίστοιχο χάρτη από τον RIPE. Επειδή στη γεωγραφική περιοχή που μελετούμε με την πάροδο του χρόνου νέα δίκτυα μπορεί να συνδεθούν με τα ήδη υπάρχοντα (π.χ. ένα Πανεπιστήμιο επεκτείνει τη δικτυακή του υποδομή) ή κάποια δίκτυα μπορεί να σταματήσουν να λειτουργούν (π.χ. ένας παροχέας υπηρεσιών Ίντερνετ -ISP[56]- σταματά να προσφέρει τις υπηρεσίες του), υπάρχει ένα είδος χρονικής πληροφορίας η οποία μεταβάλλεται και η οποία αφορά τα δίκτυα της γεωγραφικής περιοχής. Επιπλέον, υπάρχει και ένα είδος αντίστοιχης χωρικής πληροφορίας μιας και οι δικτυακές οντότητες της περιοχής είναι συνήθως εντοπισμένες σε συγκεκριμένες γεωγραφικές συντεταγμένες. Αυτές οι χωρικές και χρονικές πληροφορίες θα πρέπει με κάποιο τρόπο να αντικατοπτρίζονται και στο χάρτη των IP διευθύνσεων της περιοχής και κατ' επέκταση και στο όποιο δείγμα προκύψει από αυτόν. Όπως θα δούμε και παρακάτω αυτές οι χωρικές και χρονικές πληροφορίες είναι φανερές όχι απευθείας στις IP διευθύνσεις αλλά στην πιο φιλική τους αναπαράσταση στα hostnames. Αυτό γίνεται διότι στα hostnames επιτρέπεται η χρήση αλφαριθμητικών χαρακτήρων δίνοντας τη δυνατότητα στους διάφορους διαχειριστές των δικτύων να διατυπώνουν στα hostnames, έστω και κωδικοποιημένα, διάφορες τοποθεσίες της γεωγραφικής περιοχής.

[55] [Lek03a], [Lek04]
[56] Internet Service Provider

4.2 Η επιλογή της γεωγραφικής περιοχής και ο IP χάρτης της

'Οπως είπαμε στην εισαγωγή ως γεωγραφική περιοχή που θα εργαστούμε επιλέγουμε τη Μεγάλη Βρετανία (.uk). Επειδή η Μεγάλη Βρετανία βρίσκεται στην Ευρώπη πέφτουμε στην περιοχή ευθύνης του RIPE. Στον Πίνακα 4.1 μπορούμε να δούμε ένα μικρό τμήμα (την αρχή και το τέλος) του .uk IP χάρτη, όπως αυτός είχε διαμορφωθεί το Δεκέμβριο του 2002. Ο πλήρης χάρτης μπορεί να προσπελαστεί στην http://users.ntua.gr/plekeas/Data_sets/metadata_in_hostnames/IP_map_of_uk.txt .

62	3	64	0	16384
62	6	0	0	65536
62	7	0	0	65536
62	8	96	0	8192
62	12	64	0	8192
62	13	128	0	8192
62	18	0	0	131072
62	24	128	0	32768
62	25	0	0	32768
62	25	128	0	32768
62	28	0	0	65536
62	30	0	0	65536
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
217	196	0	0	4096
217	196	224	0	4096
217	197	32	0	4096
217	197	192	0	4096
217	198	32	0	4096
217	199	160	0	4096
217	199	176	0	4096
217	204	0	0	262144

Πίνακας 4.1: Μικρό τμήμα του .uk IP χάρτη (Δεκ. 2002)

Η πρώτη σειρά του πίνακα 4.1 δείχνει πως ξεκινώντας από το IP 62.3.64.0, 2^{14} = 16384 συνεχείς IP διευθύνσεις έχουν ανατεθεί σε κάποια δικτυακή οντότητα στο Ηνωμένο Βασίλειο. Κάτι αντίστοιχο ισχύει και για τις υπόλοιπες γραμμές.

Η επιλογή της συγκεκριμένης γεωγραφικής περιοχής δεν ήταν τυχαία. Δύο ήταν οι λόγοι που μας ανάγκασαν να επιλέξουμε το .uk. Ο πρώτος ήταν πως, επειδή δεν θα εργαστούμε απευθείας σε IP αριθμούς αλλά στα αντίστοιχα hostname, θα έχουμε να αντιμετωπίσουμε πληροφορία η οποία έχει αναπαρασταθεί σε κείμενο. Γι' αυτό το λόγο θα θέλαμε να εξαλείψουμε το γλωσσικό παράγοντα στον τρόπο που παριστάνεται η πληροφορία. Αυτός ο παράγοντας δείχνει να είναι πολύ σημαντικός στην ανακάλυψη ιδιοτήτων και στην εξόρυξη σχημάτων (patterns) και πληροφορίας από κείμενα. Ένα γραφικό παράδειγμα είναι το ακόλουθο: Όλοι οι Έλληνες αναγνώστες αυτού του κεφαλαίου μάλλον θα αναγνώριζαν πως η μνημονική αναπαράσταση του www.kathimerini.gr αναφέρεται σε μία ημερήσια ελληνική

εφημερίδα. Αλλά αν υποθέταμε ότι κάποιος δεν το γνώριζε τότε και πάλι θα μπορούσε να αναγνωρίσει από το συλλαβισμό των γραμμάτων ότι πρόκειται για κάποια ελληνική λέξη. Αντίθετα, ένας αναγνώστης που δε μιλάει ελληνικά θα ήταν απίθανο να καταλάβει από την αρχή τί σημαίνει αυτό το περίεργο όνομα. Εάν, βέβαια επέλεγε τη διεύθυνση θα έβλεπε ότι δείχνει σε μία ελληνική εφημερίδα αλλά και πάλι ίσως να αγνοούσε το ακριβές νόημα της λέξης (εκτός και αν κάποιος του το μαρτυρούσε ή αν ξεκινούσε να μαθαίνει ελληνικά). Το πιο πάνω φαινόμενο δε θα θέλαμε να συμβεί και σε εμάς, οπότε επιλέξαμε να συλλέξουμε το δείγμα από μία περιοχή στην οποία οι άνθρωποι σκέφτονται, μιλούν και γράφουν μία γλώσσα που και εμείς μιλάμε και μία τέτοια γλώσσα ήταν η εγγλέζικη. Ένας δεύτερος λόγος για την επιλογή της συγκεκριμένης περιοχής ήταν αυτός των υψηλών ποσοστών host/κάτοικο [Eur], ο οποίος θα μας έδινε πλούσιο σε πληροφορία δείγμα (άλλα δείγματα τα οποία συλλέξαμε από άλλα κράτη - Ελλάδα, Ιορδανία - ήταν πολύ πιο φτωχά σε τέτοιου είδους πληροφορία).

4.2.1 Κάνοντας δειγματοληψία στο χάρτη

Όπως είπαμε στον Πίνακα 4.1 φαίνεται ένα μικρό τμήμα του IP χάρτη. Όλος ο χάρτης αποτελείται από 21,771,520 IP διευθύνσεις. Χρησιμοποιώντας το δειγματολήπτη του προηγούμενου κεφαλαίου παίρνουμε ένα δείγμα μεγέθους 0.1% το οποίο αποτελείται από 21,772 IP.

4.3 Η μορφή του δείγματος

Ένα μικρό μέρος του δείγματος που προέκυψε από τη δειγματοληψία φαίνεται στον πίνακα 4.2 (όλο το δείγμα μπορεί κάποιος να το δει στην http://users.ntua.gr/plekeas/Data_sets/metadata_in_hostnames/uk_data_set.txt καθώς και διάφορους άλλους πίνακες και αποτελέσματα που αφορούν αυτό το κεφάλαιο). Παρατηρώντας το δείγμα μπορούμε να διαπιστώσουμε ότι σχηματίζονται συστάδες (clusters) από hostnames και σε κάθε συστάδα η πληροφορία κωδικοποιείται με τον ίδιο τρόπο. Στο δείγμα υπάρχει ένα τμήμα όπου όλα τα στοιχεία δεν είναι σε μορφή hostname. Αυτό συμβαίνει διότι είτε στους συγκεκριμένους IP δεν έχουν ανατεθεί hostname είτε διότι οι υπολογιστές με τα συγκεκριμένα IP για κάποιο λόγο δεν επέστρεψαν hostname, παρόλο που τους ζητήθηκε. Αυτό, λοιπόν, το τμήμα του δείγματος παραμένει σε IP αναπαράσταση και δε μας είναι χρήσιμο για την ανακάλυψη χωρικών και χρονικών δεδομένων.

Υπάρχει, επίσης, ένα άλλο τμήμα του δείγματος στο οποίο όλα τα hostname ξεκινούν με το πρόθεμα "www" (βλέπε και πίνακα 4.3). Αυτή είναι η www συστάδα και τα στοιχεία της είναι base-URLs[57]. Παρατηρούμε ότι η κωδικοποίηση των πιο πάνω hostname γίνεται με τέτοιο τρόπο ώστε να μην περιέχεται πληροφορία για φυσικές συντεταγμένες των κωδικοποιημένων αντικειμένων. Αντίθετα, η κωδικοποίηση που παρατηρούμε μας προσανατολίζει στον υπερχώρο (hyperspace), δηλαδή στο web. Δοκιμάζοντας όλα τα hostname αυτής της συστάδας, αν όντως εξυπηρετούν μία web σελίδα, είδαμε ότι όλες οι http συνδέσεις τους ήταν επιτυχείς. Όμως, το web δεν είναι χωρικό με την έννοια της φυσικής εμπειρίας [Sve98], άρα ούτε η www συστάδα μας είναι τόσο χρήσιμη για την ανακάλυψη χωρικών και χρονικών δεδομένων.

[57] βλ. 2.3.1

. . .
modem-3575.orangutan.dialup.pol.co.uk
modem-3719.hyena.dialup.pol.co.uk
modem-972.barrelled.dialup.pol.co.uk
ns.ecat-tech.com
ns2.avantdns.co.uk
ns2.qtec.uk.net
ntfm382-10.facility.pipex.com
pc1-cmbg3-5-cust115.cmbg.cable.ntl.com
pc1-papw1-6-cust77.cmbg.cable.ntl.com
80-194-58-94.cable.ubr08.bf.blueyonder.co.uk
adsl.r-t-f-m.co.uk
213.219.21.165
213.249.159.75
217.13.131.167
www.heavens-above.org
www.jacksonspower.co.uk
www.kayto.co.uk
www.learning-opportunities.co.uk
fp03-347.web.dircon.net
ftp.dip.co.uk
host213-1-133-95.in-addr.btopenworld.com
. . .

Πίνακας 4.2: Ένα μικρό μέρος του δείγματος

...
www.focus21.co.uk
www.continentalresources.org
www.easyshopping4u.co.uk
www.oouid.man.ac.uk
www.sundogs.co.uk
www.patricksofcamelon.com
www.quays.co.uk
www.issltech.net
www.vinesgroup.co.uk
...

Πίνακας 4.3: Μέρος της www συστάδας

Το πιο μεγάλο μέρος του δείγματος αποτελείται από μικρότερες συστάδες των οποίων τα στοιχεία δεν έκαναν επιτυχημένες http συνδέσεις (εκτός από κάποιες περιπτώσεις mail server ή login prompt). Αυτό σημαίνει ότι τα πιο πάνω στοιχεία των συστάδων δεν κωδικοποιήθηκαν έχοντας το web κατά νου αφού δεν προσανατολίζουν στον υπερχώρο. Η κωδικοποίηση αυτών των hostname έγινε για να δηλώσει χωρική και χρονική πληροφορία, όπως θα δούμε και πιο κάτω. Σε αυτές τις συστάδες αυτό που παρατηρείται είναι ότι όλα τα στοιχεία τους έχουν την ίδια ουρά

(tail). Εάν σκεφτούμε ότι τα hostname είναι αλφαριθμητικές ακολουθίες τμημάτων (το κάθε τμήμα επιτρέπεται να έχει μέχρι 63 χαρακτήρες), τα οποία χωρίζονται από τελείες (.), με μέγιστο αριθμό χαρακτήρων 255, τότε η ουρά ενός hostname θα είναι κάποια ακολουθία των τελευταίων τμημάτων. Για παράδειγμα, η κοινή ουρά των pc3-hudd3-3-cust138.hudd.cable.ntl.com και pc1-bagu4-3-cust54.mant.cable.ntl.com είναι η cable.ntl.com. Στο δείγμα παρατηρούμε ότι συνήθως οι ουρές προδίδουν τον ιδιοκτήτη του hostname (βλ. ενότητα 4.4). Έτσι, μπορούμε να πούμε με μεγάλη πιθανότητα επιτυχίας εάν ένα hostname ανήκει σε έναν παροχέα υπηρεσιών Internet ή σε κάποιο εκπαιδευτικό ίδρυμα κ.λπ.

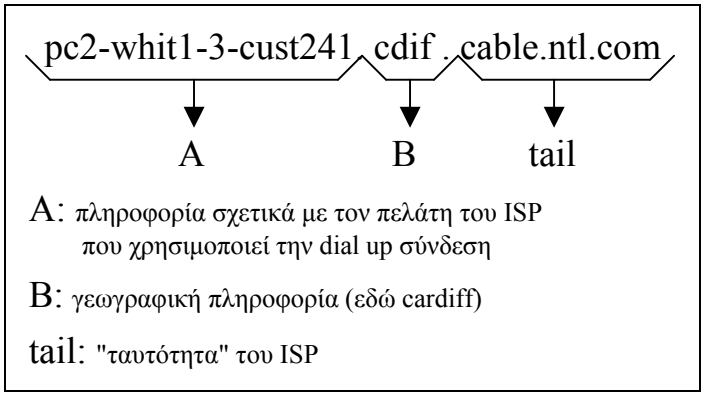
4.4 Εξόρυξη πληροφορίας σχετικά με την υποδομή και την κυκλοφορία Internet (Internet traffic) ενός Βρετανικού ISP

Όπως αναφέραμε στην προηγούμενη παράγραφο το δείγμα είναι έτσι χωρισμένο ώστε τα hostname με τις ίδιες ουρές να ανήκουν στις ίδιες συστάδες. Αρκετές από τις συστάδες έχουν στοιχεία που κωδικοποιούν ονόματα πόλεων της Βρετανίας. Η μεγαλύτερη τέτοια συστάδα είναι η συστάδα 5 η οποία ανήκει σε έναν από τους πιο μεγάλους ISP της Βρετανίας. Σε αυτή τη συστάδα τα στοιχεία είναι όπως στον πίνακα 4.4 .

.
.
.
pc2-whit1-3-cust241.cdif.cable.ntl.com
pc3-hitc1-6-cust172.lutn.cable.ntl.com
pc2-cmbg4-5-cust140.cmbg.cable.ntl.com
.
.
.

Πίνακας 4.4: Ένα μικρό τμήμα της συστάδας 5

Η ουρά των hostname έχει 3 τμήματα (cable.ntl.com) και κάθε hostname μπορεί να γραφτεί σαν A.B.tail. Για παράδειγμα, στο hostname pc2-whit1-3-cust241.cdif. cable.ntl.com, A = pc2-whit1-3-cust241 και B = cdif. Τα πιο πάνω φαίνονται και στη φιγούρα 4.1.

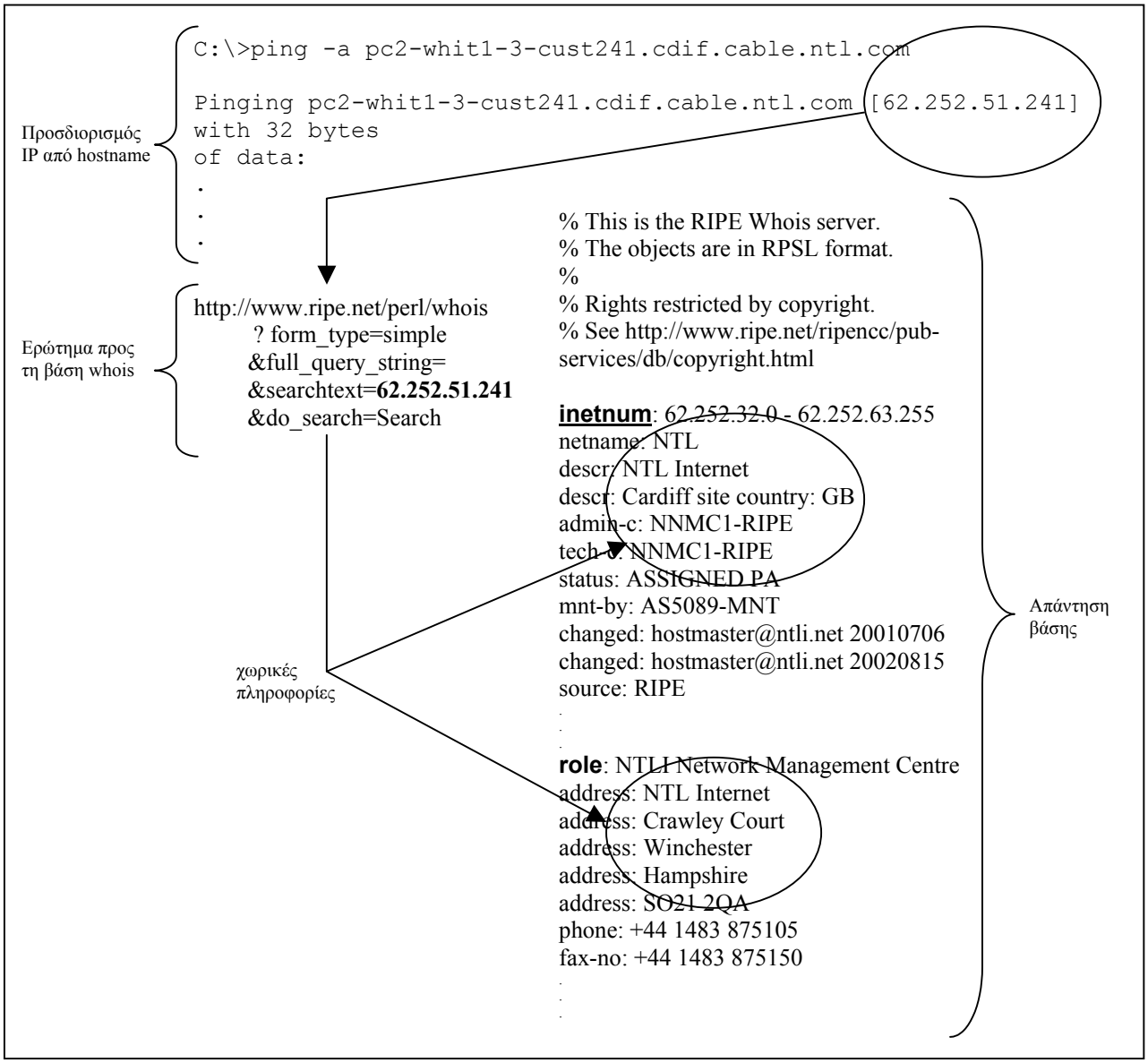


Φιγούρα 4.1: Τα 3 τμήματα των hostname της συστάδας 5

Απλή επιθεώρηση της συστάδας 5 δείχνει ότι το τμήμα του ενδιαφέροντος το οποίο κωδικοποιεί ονόματα πόλεων είναι το Β, οπότε χωρίζουμε το τμήμα Β από τους γείτονές τους και ελέγχουμε αν ταιριάζει με τα πρώτα 3 γράμματα Βρετανικών πόλεων, όπως αυτές αναφέρονται στην τελευταία (2001) απογραφή του πληθυσμού της Μεγάλης Βρετανίας και Ουαλίας [UKNS01]. Θεωρούμε ότι έχουμε επιτυχημένο ταίριασμα εάν τα πρώτα 3 γράμματα του τμήματος Β ταιριάζουν απόλυτα με τα πρώτα 3 γράμματα κάποιας Βρετανικής πόλης, (αν δύο ή περισσότερες πόλεις υπάρχουν με ίδια τα 3 πρώτα γράμματα τότε υπολογίζουμε αυτή με το μεγαλύτερο πληθυσμό όπως καταγράφτηκε στην απογραφή).

Για να επαληθεύσουμε τα αποτελέσματά μας, επιπλέον, ρωτάμε και τη whois βάση δεδομένων του RIPE δίνοντας σαν ερώτημα το hostname. Η πιο πάνω βάση περιέχει πληροφορίες σχετικά με τις αναθέσεις IP διευθύνσεων, τις τακτικές δρομολόγησης κ.λπ. στην περιοχή ευθύνης του RIPE και όχι μόνο. Οι εγγραφές αυτής της βάσης είναι γνωστές σαν αντικείμενα (βλ. 3.5.3) και στην περίπτωσή μας τα αντικείμενα που μας ενδιαφέρουν είναι 2. Το πρώτο είναι το "inetnum", το οποίο περιέχει πληροφορίες σχετικά με τις αναθέσεις IP διευθύνσεων του IPv4 χώρου. Το δεύτερο αντικείμενο είναι το "role", το οποίο περιέχει πληροφορίες σχετικά με το ρόλο που παίζει ο κάτοχος των IP διευθύνσεων (π.χ. ISP, university) ή δίνει πληροφορίες σχετικά με τη δράση ενός ή περισσότερων φυσικών προσώπων (π.χ. το διαχειριστή ενός συστήματος ή τον υπεύθυνο τεχνικό με τον οποίον ο RIPE θα επικοινωνήσει σε έκτακτη περίπτωση). Πολλές φορές το αντικείμενο role περιέχει και χωρικές πληροφορίες οι οποίες αφορούν τον κάτοχο των IP διευθύνσεων (στην περίπτωση της συστάδας 5, του ISP). Στη φιγούρα 4.2 βλέπουμε πώς μπορούμε να μάθουμε από τη βάση whois, χωρικές πληροφορίες που αφορούν ένα hostname. Παίρνουμε την IP διεύθυνση του hostname[58] και δίνουμε ένα ερώτημα που περιέχει το IP στη βάση whois η οποία μας επιστρέφει την απάντηση. Στα αντικείμενα "role" και "inetnum" βλέπουμε ότι υπάρχει η ζητούμενη χωρική πληροφορία. Π.χ. για το hostname της φιγούρας 4.2 βλέπουμε ότι ο υπεύθυνος φορέας είναι ο ISP "NTL Internet" με έδρα την περιοχή του Hampshire, ενώ ο host στον οποίο ανήκει το hostname βρίσκεται στην περιοχή του Cardiff[59].

[58] Οι εντολές που μετατρέπουν hostname στα αντίστοιχα IP είναι οι ping (MSDOS) και nslookup (UNIX), π.χ. ping achilles.noc.ntua.gr μας δίνει το αντίστοιχο IP.
[59] Επειδή έχει περάσει αρκετός χρόνος από το πείραμα οι αντιστοιχίες μπορεί να μην ισχύουν πια.



Φιγούρα 4.2: Προσδιορίζοντας από τη whois βάση του RIPE χωρικές πληροφορίες για hostname

Στον πίνακα 4.5 συνοψίζουμε τα αποτελέσματα των πιο πάνω συγκρίσεων με την τελευταία απογραφή και με τις απαντήσεις της βάσης whois του RIPE.

Πρώτα 3 γράμματα του τμήματος B	ταίριασμα (απογραφή 2001)	ταίριασμα (whois database)	Συνολικά ταιριάσματα στη συστάδα 5
blf	x	**Bel**fast	3
brh	x	**Bir**ming**h**am	5
bro	Bromley	**Bro**mley	1
cdi	x	**C**ar**di**ff	5
cmb	x	**C**a**mb**ridge	6
col	Colchester	**Col**chester	2
glf	x	**G**u**ild**ford	1
hud	x	**Hud**dersfield	1
lds	x	**L**ee**ds**	4
lut	Luton	**Lut**on	5
man	Manchester	Baguley	1
mid	Middlesbrough	**Mid**dlesbrough	7
not	Nottingham	**Not**tingham	18
nrt	x	**N**or**th**ampton	2
oxf	Oxford	**Oxf**ord	3
pop	x	Waltham Park	1
ren	x	**Ren**frew	9
swa	Swansea	**Swa**nsea	3
win	Windsor	Brentford	2

Πίνακας 4.5: Συγκρίνοντας τα 3 πρώτα γράμματα του τμήματος B της συστάδας 5

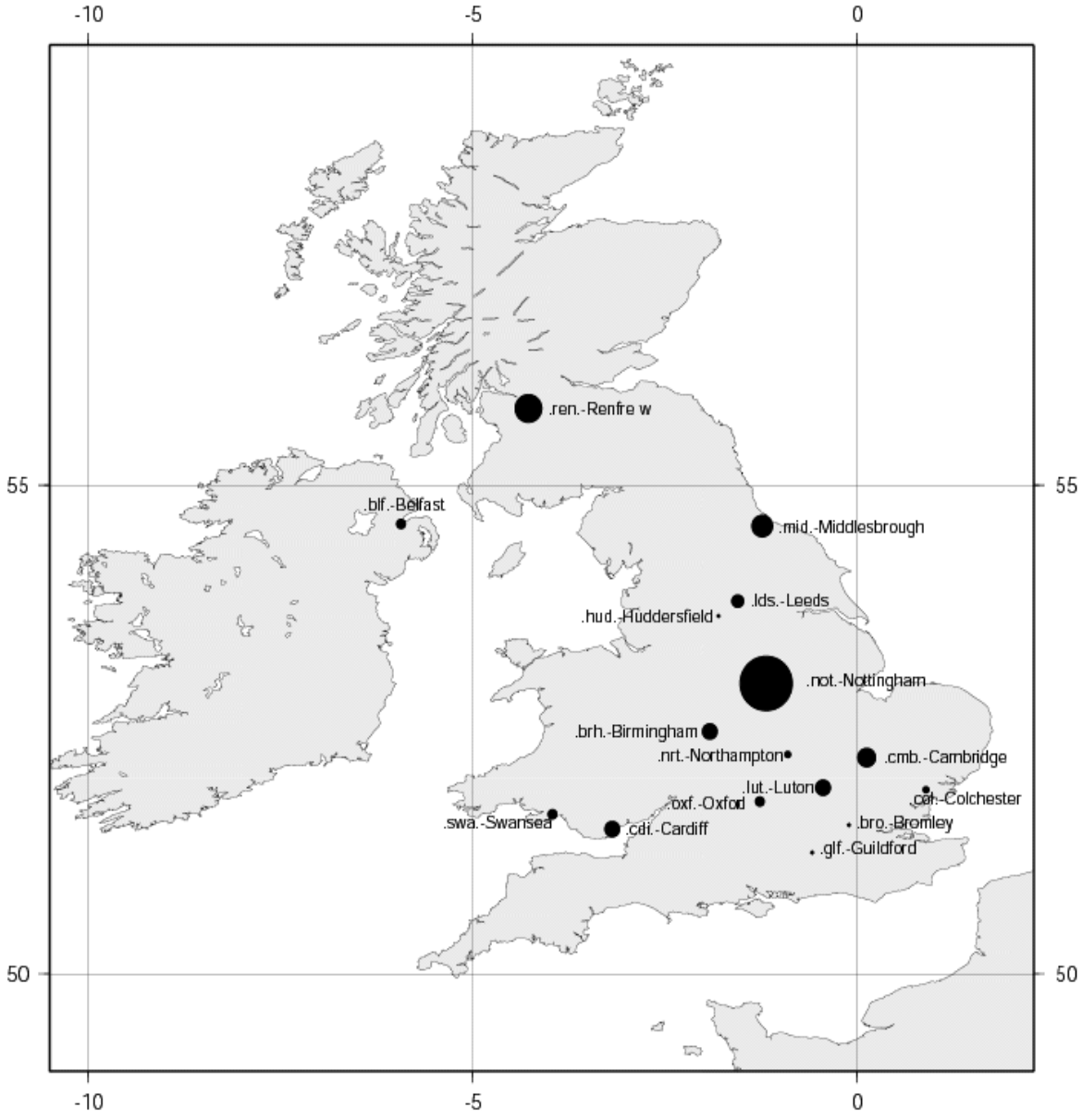
Από τον πιο πάνω πίνακα μπορούμε να δούμε ότι υπάρχει ποσοστό 43% ταιριάσματος με τα δεδομένα της απογραφής και ποσοστό ταιριάσματος 62% με τα δεδομένα του RIPE. Οι πιο πάνω τιμές δείχνουν ότι υπάρχει μία πατέντα (pattern) σύμφωνα με την οποία Βρετανικές πόλεις κωδικοποιούνται στα συγκεκριμένα hostname.

Αν, τώρα, θεωρήσουμε ότι τα στοιχεία της συστάδας 5 είναι δρομολογητές (routers) τότε έχουμε μία εικόνα της υποδομής του ISP στη γεωγραφική περιοχή της Μεγάλης Βρετανίας. Επιπλέον, το "-cust"[60] κομμάτι του τμήματος A δηλώνει ότι τα πιο πάνω hostname μπορούν να θεωρηθούν και σαν συνδέσεις Internet μέσω του συγκεκριμένου ISP.

Τα πιο πάνω μας βοηθούν να απεικονίσουμε γεωγραφικά την υποδομή του συγκεκριμένου ISP και την κίνηση των router του (Internet traffic) για τη συγκεκριμένη χρονική περίοδο που κράτησε η δειγματοληψία. Στην επόμενη σελίδα μπορεί κάποιος να δει τον τελικό γεωγραφικό χάρτη (φιγούρα 4.3). Σε αυτό το χάρτη τα μεγέθη των κύκλων αντιπροσωπεύουν τον αριθμό των εμφανίσεων κωδικοποιημένων Βρετανικών πόλεων στα hostname της συστάδας 5. Επιπλέον, τα μεγέθη των κύκλων μπορούν να ειδωθούν και σαν το φόρτο (κίνηση) του δικτύου του

[60] Θεωρούμε ότι πρόκειται για τη λέξη customer.

ISP δείχνοντας έτσι την κίνησή του για το αντίστοιχο χρονικό διάστημα (12/2002) της συλλογής του δείγματος. Η φιγούρα 4.3 δημιουργήθηκε με τη χρήση του προγράμματος GMT [GMTa], [GMTb] και των βιβλιοθηκών του (coastlines), και τα κέντρα των κύκλων αναπαριστούν τα πραγματικά γεωγραφικά μήκη και πλάτη των συγκεκριμένων πόλεων.



Φιγούρα 4.3: Χάρτης υποδομής και κυκλοφορίας ενός Βρετανικού ISP (Δεκ. 2002)

4.5 Προσδιορίζοντας το ρυθμό ανάπτυξης ενός Βρετανικού ISP

Μέχρι τώρα είδαμε μία περίπτωση στην ποία χωρικά δεδομένα μπορούν να εξαχθούν από τα hostname. Σε αυτή την ενότητα θα δούμε πώς είναι δυνατό να εξαχθούν χρονικά δεδομένα από τα hostname.

Στο δείγμα και σχεδόν σε όλες τις συστάδες του παρατηρούμε ότι εμφανίζονται κωδικοί στα hostname οι οποίοι έχουν κάποιον αύξοντα αριθμό. Αυτό δηλώνει αμέσως κάποια μορφή απαρίθμησης. Για παράδειγμα, ας δούμε τα hostname modem-2266.giraffe.dialup.pol.co.uk και modem-785.giraffe.dialup.pol.co.uk της συστάδας 3 (άλλος ένας μεγάλος Βρεταννικός ISP). Χρησιμοποιώντας την κοινή λογική θα δούμε ότι η πιο πάνω κωδικοποίηση δηλώνει πως στον εξυπηρετητή (server) giraffe υπάρχουν δύο modem με ονόματα 785 και 2266. Άλλα παραδείγματα είναι τα hostname ppp-0-162.birm-a-2.access.uk.tiscali.com, ppp-1-189.birm-a-2.access.uk.tiscali.com, ppp-1-98.birm-a-1.access.uk.tiscali.com από τη συστάδα 7, ή τα hostname, web151.pavilion.net και web411.pavilion.net από τη συστάδα 13.

Ξαναγυρνώντας στη συστάδα 3, (σε αυτή την παράγραφο επεξεργαζόμαστε τη συστάδα 3 μικρό μέρος της οποίας φαίνεται στη φιγούρα 4.4), μπορούμε βάσει των πιο πάνω να υποθέσουμε ότι η πληροφορία που πιθανόν να μεταβληθεί με το χρόνο είναι ο αριθμός των modem στο συγκεκριμένο server, (π.χ. ο αριθμός μπορεί είτε να αυξηθεί λόγω προσθήκης επιπλέον modem είτε να ελαττωθεί διότι προέκυψε κάποιο πρόβλημα και αφαιρέθηκαν κάποια). Παρά το γεγονός ότι ο πληθυσμός των modem που είναι συνδεδεμένα στο συγκεκριμένο server μπορεί να αλλάξει με την πάροδο του χρόνου, ο πληθυσμός αυτός είναι εντοπισμένος χωρικά. Αυτό συμβαίνει διότι όταν κάπου γίνεται συστηματικό dial up (όπως σε έναν ISP) τότε όλα τα modem τοποθετούνται σε ειδικές μεταλλικές κατασκευές που ονομάζονται ρακς (racks) κοντά στον server για γρήγορη επιθεώρηση και συντήρηση.

```
...
modem-3975.llama.dialup.pol.co.uk
modem-4008.giraffe.dialup.pol.co.uk
modem-4092.gorilla.dialup.pol.co.uk
modem-4095.giraffe.dialup.pol.co.uk
modem-449.porcupine.dialup.pol.co.uk
modem-459.giraffe.dialup.pol.co.uk
modem-47.argonath.dialup.pol.co.uk
modem-475.hyena.dialup.pol.co.uk
modem-510.tiger.dialup.pol.co.uk
modem-521.gorilla.dialup.pol.co.uk
modem-543.gorilla.dialup.pol.co.uk
modem-584.hyena.dialup.pol.co.uk
modem-601.clefairy.dialup.pol.co.uk
modem-645.jaguar.dialup.pol.co.uk
modem-647.alakazam.dialup.pol.co.uk
...
```

Φιγούρα 4.4: Μικρό μέρος της συστάδας 3

Πώς, όμως, μπορούμε να εισάγουμε την έννοια του χρόνου στη συστάδα που μελετούμε; Αυτό το κάνουμε με τη βοήθεια μίας μερικής διάταξης $\leq_t$. Επιπλέον, μπορούμε να ορίσουμε και ένα μερικά διατεταγμένο σύνολο στο οποίο θα υπάρχει αυτή η διάταξη και στο οποίο δύο hostname θα μπορούν μεταξύ τους να συγκριθούν ως προς το χρόνο.

Έστω, λοιπόν, C3 το σύνολο των hostname της συστάδας 3. Το C3 είναι ένας μερικά διατεταγμένος χώρος με τη σχέση $\leq_t$ διότι είναι μη κενό σύνολο, και για κάθε (a, b, c$\in$ C3) οι ιδιότητες ανακλαστική ($a \leq_t a$), αντισυμμετρική ($a \leq_t b,\ b \leq_t a \Rightarrow a=b$) και μεταβατική ($a \leq_t b,\ b \leq_t c \Rightarrow a \leq_t c$) ικανοποιούνται. Οι τρεις πιο πάνω ιδιότητες ικανοποιούνται διότι κάθε hostname έχει το αντίστοιχό του IP. Επειδή, όμως, οι IP αριθμοί μπορούν να αναπαριστάνουν ακέραιους αριθμούς μπορούν να συγκριθούν με τη συνηθισμένη σχέση ($\leq$) στους ακέραιους. Η σχέση που εμείς χρησιμοποιούμε ($\leq_t$) διαφέρει λίγο από τη συνηθισμένη ανισότητα στο ότι συγκρίνει ακέραιους που περιέχουν χρονικά γεγονότα (γι' αυτό υπάρχει και ο δείκτης t). Η ιδέα πίσω από όλα αυτά είναι ότι, καθώς ο ISP θα αυξάνει την υποδομή του, όλο και πιο πολλοί πελάτες θα θέλουν να χρησιμοποιούν τις υπηρεσίες του (π.χ. να κάνουν dial up) και άρα ο ISP θα αναγκάζεται να αυξάνει τον αριθμό των modem στους server του (μέχρι ένα όριο φυσικά). Επίσης, πολύ πιθανό είναι να αυξάνει και τον αριθμό των server του ώστε να ανταποκριθεί στην αυξανόμενη ζήτηση. Αλλά, αφού κάθε χρήστης πρέπει να έχει ένα ξεχωριστό modem όταν συνδέεται με το Internet, κάθε νέο modem που προστίθεται θα πρέπει να έχει ένα μοναδικό αριθμό IP. Όπως, όμως, είπαμε στην 3.2.1 ένας ISP είναι LIR (τοπικός ληξίαρχος) ο οποίος παίρνει τις IP διευθύνσεις από τον RIR και τις αναθέτει περαιτέρω. Ο ISP θα αναθέσει κάποια IP στα νέα modem που θα προσθέσει, όμως τα IP αυτά δεν θα είναι τυχαία. Αυτό γίνεται για να μην υπάρχει κερματισμός του χώρου διευθύνσεων που διαθέτει ο ISP. Είναι φυσικό, για παράδειγμα, αν ο ISP προσθέσει 100 καινούργια modem να αναθέσει 100 στατικές IP διευθύνσεις (στατικές σημαίνει κάθε modem να έχει συγκεκριμένο IP πάντα). Αυτές οι 100 διευθύνσεις λογικό είναι να είναι συνεχόμενες από το χώρο διευθύνσεων που διαθέτει ο ISP. Όμως, με αυτό τον τρόπο εισάγεται η έννοια του χρόνου που χρειαζόμαστε, διότι για παράδειγμα τα modem με ονόματα modem-99 και modem-100 θα έχουν συνεχόμενα IP. Υποθέτοντας, τώρα, ότι το modem-99 ονομάστηκε πιο πριν από το modem-100 τότε μπορούμε να συγκρίνουμε τα αντίστοιχα IP τους στο χρόνο (π.χ. θα ισχύει $IP_{modem\text{-}99} \leq_t IP_{modem\text{-}100}$)[61]. Όμως, αφού κάθε IP έχει και ένα αντίστοιχο hostname τότε μπορούμε να διατάξουμε και τα hostname στο χρόνο, δηλαδή μπορούμε να τα συγκρίνουμε.

Στον πίνακα 4.6 παρουσιάζουμε τη συστάδα 3 με τα αντίστοιχα IP, τα ονόματα των server και τους αριθμούς των modem. Για παράδειγμα, στον server gazelle το modem-3967 έχει IP: 81.78.79.127. Στον πίνακα αυτό μπορούμε να δούμε ότι μερικοί server συμμετέχουν στη συστάδα με ένα μόνο modem. Αυτό το γεγονός εμποδίζει τη διμελή σχέση $\leq_t$ να γίνει ολική σχέση στο C3, διότι θα πρέπει να υπάρχουν τουλάχιστον 2 στοιχεία στον ίδιο server για να τα διατάξουμε στο χρόνο.

[61] Η πιο πάνω σχέση δε σημαίνει απαραίτητα ότι ο ακέραιος που αντιστοιχεί στον $IP_{modem\text{-}99}$ είναι μικρότερος (με την ανισότητα στους ακέραιους) από τον ακέραιο που αντιστοιχεί στον $IP_{modem\text{-}100}$. Μπορεί π.χ. $IP_{modem\text{-}99}$=100.100.100.100 και $IP_{modem\text{-}100}$=100.100.100.99, αλλά την πιο πάνω περίπτωση ποτέ δεν τη συναντήσαμε στα hostame της συστάδας 3.

Όνομα server	Αριθμός modem	Αντίστοιχο IP
jaguar	144, 645, 1080, 1102, 1450, 1603, 1689, 2125, 2539, 2669, 2756, 2778, 3192, 3606	**81.76.**(176.144-178.133- 180.56-180.78-181.170-182.67-182.153-184.77-185.235-186.109-186.196-186.218-188.120-190.22)
hyena	214, 279, 475, 584, 1237, 1346, 1629, 1672, 1955, 2391, 2456, 2586, 3044, 3087, 3218, 3719, 3828	**81.78.**(112.214-113.23-113.219-114.72-116.213-117.66-118.93-118.136-119.163-121.87-121.152-122.26-123.228-124.15-124.146-126.135-126.144)
gorilla	216, 303, 521, 543, 652, 761, 957, 1131, 1218, 1240, 1566, 1675, 2002, 2176, 2633, 3134, 3221, 3308, 3591, 3765, 4092	**81.78.**(96.216-97.47-98.9-98.31-98.140-98.249-99.189-100.107-100.194-100.216-102.30-102.139-103.210-104.128-106.73-108.62-108.149-108.236-110.7-110.181-111.252)
giraffe	459, 698, 785, 1177, 1330, 1351, 1656, 1765, 1787, 1918, 2070, 2266, 2549, 2919, 2941, 3137, 3158, 3202, 3376, 3398, 3550, 4008, 4095	**81.78.**(81.203-82.186-83.17-84.153-85.50-85.71-86.120-86.229-86.251-87.126-88.22-88.218-89.245-91.103-91.125-92.65-92.86-92.130-93.48-93.70-93.222-95.168-95.255)
gazelle	201, 397, 745, 897, 984, 1093, 1311, 1485, 2182, 2225, 2509, 2617, 2682, 3314, 3336, 3444, 3684, 3793, 3858, 3967	**81.78.**(64.201-65.141-66.233-67.129-67.216-68.69-69.31-69.205-72.134-72.177-73.205-74.57-74.122-76.242-77.8-77.116-78.100-78.209-79.18-79.127)
alakazam	647	**217.135.**13.135
antelope	1833	**217.134.**23.41
argonath	47	**62.136.**124.47
barrelled	972	**62.25.**143.204
blue-streak-damsel	74	**62.136.**241.74
buffalo	1257	**217.134.**68.233
charizard	1149	**217.135.**73.125
cheetah	2013	**217.134.**103.221
clefairy	601	**217.135.**91.89
cougar	3099	**217.134.**236.27
dragon-wrasse	115	**62.137.**1.115
elephant	3489	**217.134.**253.161
elk	1801	**81.76.**167.9
grommet	163	**62.25.**156.163
hodad	272	**62.25.**161.16
lemur	2483	**217.135.**137.179
leopard	2197	**217.135.**152.149
lion	3303	**217.135.**172.231
monkey	1293	**217.135.**213.13
new-mexico	170	**62.137.**82.170
orangutan	3575	**217.135.**237.247
parrotfish	27	**62.137.**46.155
porcupine	449	**217.134.**193.193
tiger	510	**62.136.**209.254
wolf	1176	**81.76.**132.152

Πίνακας 4.6: Server, modem και αντίστοιχα IP στη συστάδα 3

Με απλή επισκόπηση του πίνακα 4.6 βλέπουμε ότι για όλα τα στοιχεία a, b $\in$C3 τα οποία μπορούν να συγκριθούν (όσα δηλαδή ανήκουν στον ίδιο server) ισχύει ότι a $\leq_t$ b $\Rightarrow$ $IP_a \leq_t IP_b$, όπου IP_x είναι το IP του hostname x. Επίσης, ισχύει ότι για όλα τα στοιχεία a, b $\in$C3 που μπορούν να συγκριθούν, $|m_a - m_b| = |IP_a - IP_b|$, όπου m_x

δηλώνει τον αριθμό modem του hostname x και |x| είναι η απόλυτη τιμή του x. Το πιο πάνω επιβεβαιώνει την υπόθεση που είχαμε κάνει προηγουμένως, ότι δηλαδή οι διευθύνσεις που αναθέτει ο ISP είναι συνεχόμενες για τα καινούργια modem που προστίθενται.

Η πιο πάνω απλή δομή του μερικά διατεταγμένου συνόλου C3 μας βοηθά πολύ στο να ανακαλύψουμε προς ποιά κατεύθυνση μπορούμε να επεκτείνουμε το data set έτσι ώστε να συμπεριλάβουμε στην επέκταση μόνο στοιχεία χρήσιμα. Στο C3 μπορούμε να ορίσουμε μία server-αλυσίδα να είναι της μορφής $IP_{sn}k \leq_{t+k+1} IP_{sn}(k+1) \leq_{t+k+2} \ldots \leq_{t+k+m} IP_{sn}(k+m)$ (όπου $k, m \in \mathrm{N}$, $IP_{sn}k$ δηλώνει το IP του $k^{ου}$ modem στον server με όνομα sn και $\leq_{t+k+l}$ σημαίνει ότι ξεκινώντας από την $t+k+0$ χρονική στιγμή, όπου αναθέσαμε το $IP_{sn}k$, βρισκόμαστε τώρα στην $t+k+l$ χρονική στιγμή αναθέτοντας το $IP_{sn}(k+l)$). Εάν, τώρα, γνωρίζουμε ένα μικρό κομμάτι από μία server-αλυσίδα, τότε μπορούμε να προβλέψουμε όλη την έκτασή της βρίσκοντας απλά τα ελάχιστα και μέγιστα της αλυσίδας. Μπορούμε να πάρουμε τα ελάχιστα και μέγιστα στοιχεία κάθε αλυσίδας που εμφανίζεται στη συστάδα 3, να βρούμε τα αντίστοιχα IP τους και να ξεκινήσουμε να επεκτείνουμε την αλυσίδα προς τα πάνω και προς τα κάτω, για να εντοπίσουμε πότε θα έχουμε κάποια αλλαγή στο όνομα του server. Μία τέτοια αλλαγή θα σημαίνει ότι αυτά τα σημεία στα οποία φτάσαμε είναι τα πιο ακραία της αλυσίδας και αυτά είναι που μπορούν να μας βοηθήσουν να καθορίσουμε τον πληθυσμό των modem που έχουν προσαρτηθεί στο συγκεκριμένο server.

Ας δώσουμε, όμως, ένα παράδειγμα: στον πίνακα 4.6 βλέπουμε ότι η server-αλυσίδα για το *jaguar* είναι $81.76.176.144 \leq_{t+144} 81.76.178.133 \leq_{t+645} \ldots \leq_{t+3191} 81.76.188.120 \leq_{t+3192} 81.76.190.22 \leq_{t+3606}$. Η αρχή της αλυσίδας πρέπει να είναι το IP *81.76.176.0*, δηλαδή 144 μονάδες πιο κάτω από την τιμή *81.76.176.144* που υπάρχει στη συστάδα μας. Πράγματι, ζητώντας το hostname του *81.76.176.0* παίρνουμε τιμή modem-0 για το συγκεκριμένο server. Επίσης ζητώντας το hostname του *81.76.175.255* (που είναι το αμέσως προηγούμενο IP από το *81.76.176.0*) παίρνουμε διαφορετικό όνομα server, που σημαίνει ότι όντως έχουμε βρει την αρχή της αλυσίδας. Αν, τώρα, αρχίζουμε και προχωρούμε προς τα πάνω θα δούμε ότι το μέγιστο στοιχείο της αλυσίδας είναι το *81.76.191.255*. Άρα ολόκληρη η server-αλυσίδα για το server *jaguar* είναι η $81.76.176.0 \leq_{t+1} 81.76.176.1 \leq_{t+2} \ldots \leq_{t+4094} 81.76.191.254 \leq_{t+4095} 81.76.191.255$. Μπορούμε να διαπιστώσουμε από την πιο πάνω αλυσίδα ότι στο συγκεκριμένο server έχουν συνδεθεί 2^{12} modem.

Την παραπάνω μέθοδο μπορούμε να την εφαρμόσουμε και σε server που εμφανίζονται στη συστάδα με μόνο ένα modem. Για παράδειγμα, ο server cougar εμφανίζεται στη συστάδα 3 με μόνο ένα IP (217.134.236.27) και με αριθμό modem 3099. Ανασκευάζοντας την server-αλυσίδα του προκύπτει ότι το ελάχιστο στοιχείο της έχει IP *217.134.224.0* και το μέγιστο έχει IP *217.134.239.255*. Αν κατασκευάσουμε όλες τις server-αλυσίδες της συστάδας 3 έχουμε τον πίνακα 4.7.

Μέχρι εδώ η υπόθεσή μας ήταν ότι μόνο στοιχεία του C3 που ανήκουν στον ίδιο server μπορούν να συγκριθούν με την $\leq_t$. Θα ήταν σωστό να συγκρίναμε και στοιχεία που ανήκουν σε διαφορετικούς server; Γενικά αυτό δεν είναι σωστό, διότι ο RIPE αναθέτει συχνότητες στους ληξίαρχους βάσει των αναγκών τους και βάσει των διαθέσιμων IP. Στον πίνακα 4.8 μπορούμε να δούμε τον πλήρη IP χάρτη του ISP της

συστάδας 3 (http://www.ripe.net/ripencc/mem-services/general/allocs4.html). Από τον πίνακα 4.8 μπορούμε να δούμε ότι αναθέσεις IP συμβαίνουν σε διαφορετικές χρονικές στιγμές, και δεν είναι απαραίτητο μεγαλύτερα IP να έχουν ανατεθεί αργότερα μέσα στο χρόνο. Για παράδειγμα, στις 2001/1/25 ο RIPE ανάθεσε $2^{32-15} = 2^{17}$ IP (217.134.0.0 – 217.135.255.255) στο συγκεκριμένο ISP ενώ στις 2002/03/06 του ανάθεσε $2^{32-14} = 2^{18}$ IP (81.76.0.0 – 81.78.255.255). Έτσι, από τον πίνακα 4.7 θα είναι λάθος να πούμε ότι οι server lion και elk μπορούν να συγκριθούν με την $\leq_t$. Εάν μπορούσαν θα λέγαμε ότι ο elk είναι αρχαιότερος από τον lion, πράγμα που φαίνεται να είναι λάθος βάσει του πίνακα 4.8. Αυτό που είναι σωστό να πούμε είναι ότι οι 5 διαφορετικές[62] περιοχές του πίνακα 4.7 αντιπροσωπεύουν 5 διαφορετικές περιόδους ανάπτυξης στην ιστορία του ISP. Χρησιμοποιώντας τους πίνακες 4.7 και 4.8 μπορούμε να σχεδιάσουμε τη φιγούρα 4.1, η οποία δείχνει το ρυθμό ανάπτυξης του συγκεκριμένου ISP.

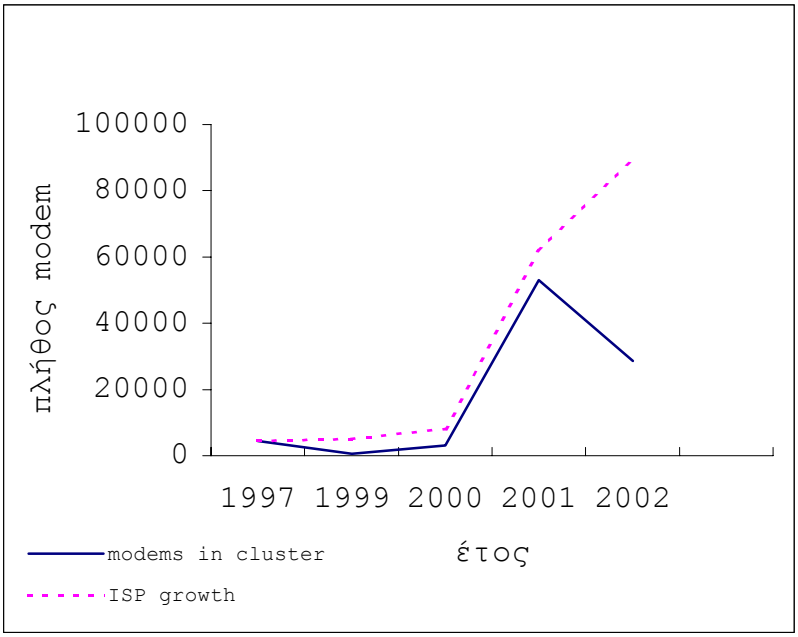
Όνομα server	IP πρώτου modem	IP τελευταίου modem	Πλήθος modems
antelope	217.134.16.0	217.134.31.255	/ 4096
porcupine	217.134.192.0	217.134.208.255	4096
cougar	217.134.224.0	217.134.239.255	4096
elephant	217.134.240.0	217.134.255.255	4096
buffalo	217.134.64.0	217.134.78.255	4096
cheetah	217.134.96.0	217.134.111.255	4096
alakazam	217.135.11.0	217.135.15.255	1280
lemur	217.135.128.0	217.135.143.255	4096
leopard	217.135.144.0	217.135.159.255	4096
lion	217.135.160.0	217.135.175.255	4096
llama	217.135.176.0	217.135.191.255	4096
monkey	217.135.208.0	212.135.223.255	4096
orangutan	217.135.224.0	217.135.239.255	4096
charizard	217.135.69.0	217.135.73.255	1280
clefairy	217.135.89.0	217.135.93.255	/ 1280
argonath	62.136.124.1	62.136.124.126	~ 126
tiger	62.136.208.0	62.136.223.255	4096
blue-streak-damsel	62.136.241.1	62.136.241.254	~ 254
dragon-wrasse	62.137.1.1	62.137.1.254	* 254
parrotfish	62.137.46.129	62.137.46.254	126
new-mexico	62.137.82.1	62.137.82.254	* 254
barrelled	62.25.140.0	62.25.143.255	# 1024
grommet	62.25.156.0	62.25.159.255	1024
hodad	62.25.160.0	62.25.164.255	# 1024
wolf	81.76.128.0	81.76.143.255	+ 4096
elk	81.76.160.0	81.76.175.255	4096
jaguar	81.76.176.0	81.76.191.255	4096
hyena	81.78.112.0	81.78.127.255	4096
gazelle	81.78.64.0	81.78.79.255	4096
giraffe	81.78.80.0	80.78.95.255	4096
gorilla	81.78.96.0	81.78.111.255	+ 4096
Συνολικός αριθμός modem[63]	/ 52992	~ 4476	* 634
		# 3072	+ 28672

Πίνακας 4.7: Οι server-αλυσίδες που αντιστοιχούν στη συστάδα 3

[62] Οι περιοχές οριοθετούνται με τα σύμβολα /, ~, *, #, +
[63] Τα σύμβολα +, * κ.λπ. οριοθετούν τα αθροίσματα.

Ημερομηνία ανάθεσης	IP που ανατέθηκαν
20000329	62.25.64/18
20001123	62.25.128/17
19970506	62.136/16
19990521	62.137/16
20020306	81.76/14
19960119	194.152.64/19
19970116	195.80.64/19
19960612	195.92/16
20010125	217.134/15

Πίνακας 4.8: Ο πλήρης IP χάρτης του ISP της συστάδας 3



Φιγούρα 4.5: Η ανάπτυξη του ISP

Από την πιο πάνω φιγούρα μπορούμε να δούμε τον υψηλό ρυθμό ανάπτυξης του ISP ο οποίος εξηγεί και τις μεγάλες αναθέσεις που έχουν γίνει σε αυτόν από τον RIPE, (ο RIPE ζητά αποδείξεις όπως ρυθμοί ανάπτυξης, οικονομικά στοιχεία κ.λπ. πριν κάνει τόσο μεγάλες αναθέσεις IP διευθύνσεων).

4.6 Παραπλήσιες εργασίες

Η εργασία που παρουσιάζεται σε αυτό το κεφάλαιο [Lek03a], [Lek04] μπορεί να συγκριθεί με τις τεχνικές που χαρτογραφούν εξυπηρετητές (hosts) σε γεωγραφικές περιοχές. Στην [PS01] παρουσιάζονται 3 τέτοιες τεχνικές. Δύο από αυτές είναι πολύ κοντά με αυτό που παρουσιάσαμε. Η πρώτη τεχνική ονομάζεται GeoCluster, και αυτό που κάνει είναι να κατασκευάζει χάρτες γεωγραφικών συντεταγμένων για μεγάλα υποσύνολα του IPv4. Υποθέτει ότι μεγάλα κομμάτια από συνεχόμενους IP αριθμούς σχηματίζουν γεωγραφικές περιοχές, οπότε εάν κάποιος γνωρίζει τις γεωγραφικές

συντεταγμένες κάποιων host μέσα από τη συστάδα, τότε μπορεί να βγάλει κάποια συμπεράσματα για όλη τη συστάδα. Η δεύτερη τεχνική ονομάζεται GeoTrack και προσπαθεί να συμπεράνει τις χωρικές συντεταγμένες των host από τα hostname. Στην [ZFRD02] host γνωστοί για τις γεωγραφικές τους συντεταγμένες κατανέμονται και μετρώντας τις καθυστερήσεις του δικτύου (network delay) υπολογίζεται η γεωγραφική θέση των υπόλοιπων host. Στην [MPDC00] προτείνεται το εργαλείο NetGeo το οποίο ρωτά whois server (όπως η whois βάση του RIPE που χρησιμοποιήσαμε) και βγάζει συμπεράσματα σχετικά με τις γεωγραφικές συντεταγμένες των host. Τέλος, ένα άλλο εργαλείο, το οποίο δουλεύει παρόμοια, είναι και το IP2LL που παρουσιάζεται στην [IP2LL].

4.7 Επίλογος - Συμπεράσματα

Στην ενότητα 4.6 παρουσιάσαμε κάποιες τεχνικές παραπλήσιες με αυτή που προτείνουμε σε αυτό το κεφάλαιο. Όλες αυτές οι τεχνικές αυτό που προσπαθούν να κάνουν είναι να εξάγουν γεωγραφικές συντεταγμένες έχοντας ως δεδομένο ένα IP ή ένα hostname. Όμως εμείς σε αυτό το κεφάλαιο είχαμε για στόχο να παρουσιάσουμε κάποια στοιχεία για το ότι χωρικά και χρονικά δεδομένα σχετίζονται με την αναπαράσταση των hostname. Επιπλέον, χρησιμοποιήσαμε τον IP δειγματολήπτη του κεφαλαίου 3 και είδαμε ότι η μέθοδος IP sampling μπορεί να εφαρμοστεί όχι μόνο στο web αλλά και στο Internet.

Πιο συγκεκριμένα, παρουσιάσαμε 2 τεχνικές οι οποίες μπορούν να εφαρμοστούν σε αναπαραστάσεις hostname για να εξαχθούν χωρικά και χρονικά δεδομένα. Για τα χωρικά δεδομένα θεωρήσαμε ότι μία συστάδα από hostname αναπαριστά δρομολογητές (routers) και συνδέσεις Internet. Αυτή η θεώρηση δεν ήταν αυθαίρετη αλλά βασίστηκε στην ικανότητά μας να καταλάβουμε τη γλωσσική αναπαράσταση των hostname. Επειδή, όμως, δεν υπάρχει ένας στάνταρ τρόπος για να κωδικοποιούμε τα hostname, ο κάθε διαχειριστής δικτύου που είναι υπεύθυνος για την ονομασία των hostname δίνει αυθαίρετα τη δική του κωδικοποίηση. Αυτό έχει σαν αποτέλεσμα να απαιτούνται ευρετικές (heuristic) τεχνικές και να γίνονται πολλές υποθέσεις για την ανακάλυψη των κωδικοποιήσεων. Επίσης, ένα άλλο αποτέλεσμα είναι ότι μεταδεδομένα εισέρχονται στην αναπαράσταση των hostname. Για παράδειγμα στη συστάδα 3 ανιχνεύσαμε ονόματα άγριων ζώων σαν ονόματα των διάφορων server, αλλά δεν ήταν όλα τα ονόματα των server της συστάδας 3 ονόματα άγριων ζώων. Αυτό είναι ένα μεταδεδομένο το οποίο ίσως έχει σχέση με τους διαχειριστές του ISP που ονόμασαν τους server (μπορεί, για παράδειγμα, τα ονόματα να μη δόθηκαν όλα από το ίδιο άτομο). Το γεγονός, τώρα, της χρήσης heuristic τεχνικών δημιουργεί προβλήματα, διότι τεχνικές που δουλεύουν για κάποιες συστάδες μπορεί να μην δουλεύουν για κάποιες άλλες. Για παράδειγμα, στις συστάδες 8, 9, 13, 14, 17, 18, 19, 25, 26, 28 δεν υπάρχει η παραμικρή ένδειξη για χωρική πληροφορία.

Για τα χρονικά δεδομένα θεωρήσαμε ότι μία συστάδα από hostname αναπαριστά διάφορα modem ενός ISP. Ορίσαμε ένα μερικά διατεταγμένο σύνολο C3 και χρησιμοποιώντας μία διμελή σχέση $\leq_t$ επεκταθήκαμε έξω από τα όρια του data set που διαθέταμε προβλέποντας την ύπαρξη στοιχείων που θα έπρεπε να ανήκουν στη συστάδα. Παρατηρούμε, όμως, ότι η πιο πάνω τεχνική βασίζεται πολύ στην αναπαράσταση κειμένου των hostname. Για παράδειγμα, η αλλαγή στο όνομα ενός server είναι πολύ σημαντική στην τεχνική μας για να εντοπίσουμε το μέγιστο και ελάχιστο αριθμό modem που είναι συνδεμένα επάνω του. Επίσης, η ίδια τεχνική

αποτυγχάνει σε άλλες συστάδες του data set (π.χ. 12, 25, 31, 34) που έχουν διαφορετική κωδικοποίηση κειμένου. Δηλαδή, συμπεραίνουμε ότι παρόλο που στην αναπαράσταση των hostname χρονικά δεδομένα υπάρχουν η εξαγωγή τους απαιτεί ευρετικές (heuristic) τεχνικές.

5 ΤΡΟΠΟΠΟΙΗΜΕΝΕΣ ΜΟΡΦΕΣ ΤΟΥ IP ΔΕΙΓΜΑΤΟΛΗΠΤΗ

Σε αυτό το κεφάλαιο δοκιμάζουμε δύο τροποποιημένες μορφές του δειγματολήπτη. Στην πρώτη ο δειγματολήπτης χρησιμοποιείται σαν ένας crawler διατρέχοντας όλο τον IP χάρτη της εισόδου του, ενώ στη δεύτερη ο δειγματολήπτης (crawler) ρυθμίζεται έτσι ώστε να υπολογίζει τα out-degree του web site που διατρέχει.

5.1 Ο δειγματολήπτης σαν crawler

Μέχρι τώρα ο IP δειγματολήπτης που υλοποιήσαμε έπαιρνε σαν είσοδο έναν IP χάρτη και, δοθέντος του επιθυμητού μεγέθους του δείγματος, έκανε απλή τυχαία δειγματοληψία επάνω στο χάρτη. Ενδιαφέρον, τώρα, θα είχε να τροποποιούσαμε το δειγματολήπτη ώστε να μην παίρνει τυχαία δείγματα από το χάρτη αλλά να τον εξερευνά ολόκληρο. Με αυτό τον τρόπο ο δειγματολήπτης δεν θα είναι πλέον ένα εργαλείο παραγωγής δειγμάτων αλλά θα μπορεί να χρησιμοποιηθεί για crawling που, όπως είπαμε στην ενότητα 1.2.3, είναι μία μέθοδος συλλογής τμημάτων του web.

Η τροποποίηση του δειγματολήπτη γίνεται ως εξής: Αφαιρείται η γεννήτρια των τυχαίων αριθμών και ρυθμίζεται η κεντρική υπολογιστική μονάδα έτσι ώστε να ελέγχει σειριακά όλες τις IP διευθύνσεις του χάρτη και να αναφέρει ποιές αντιστοιχούν σε web host. Με αυτό τον τρόπο ο δειγματολήπτης μετατρέπεται σε ένα crawler[64] ο οποίος παίρνει σαν είσοδο μία λίστα από IP και βγάζει έξοδο μία λίστα από hostnames.

Βέβαια, μία λίστα από IP διευθύνσεις μπορεί να περιέχει εκατομμύρια στοιχεία και εξαντλητική αναζήτησή της μπορεί να διαρκέσει πολύ καιρό. Για παράδειγμα ας υποθέσουμε ότι θέλουμε να σαρώσουμε όλο τον IPv4 χώρο, που έχει 2^{32} διευθύνσεις, και έστω ότι ο έλεγχος κάθε διεύθυνσης διαρκεί 2^3 δευτερόλεπτα. Τότε ένας υπολογιστής θα χρειαζόταν 2^{35} δευτερόλεπτα να κάνει τον έλεγχο, ενώ 128 υπολογιστές μαζί[65] θα χρειάζονταν 2^{28} δευτερόλεπτα (~8.63 χρόνια). Αν, όμως, περιοριστούμε σε πολύ μικρότερους χάρτες (~2^{18}) τότε ο έλεγχος μπορεί να γίνει σε ικανοποιητικό χρόνο (π.χ. μερικές μέρες).

5.1.1 Ο IP χάρτης της Ιορδανίας

Ένα παράδειγμα μικρού IP χάρτη είναι αυτό της Ιορδανίας, ο οποίος φαίνεται στον πίνακα Π.β.1 του παραρτήματος Β. Αυτός ο χάρτης αποτελείται από 133,632 IP διευθύνσεις.

Για να σαρώσουμε όλες τις διευθύνσεις του χάρτη τροποποιούμε το δειγματολήπτη της φιγούρας 3.2 έτσι ώστε να παρακάμπτεται η γεννήτρια των τυχαίων αριθμών και να ελέγχονται σειριακά όλα τα IP του χάρτη. Ο έλεγχος που γίνεται είναι για το ποιές IP διευθύνσεις αντιστοιχούν σε host που έχουν το port 80 ανοιχτό[66]. Από τον πιο πάνω έλεγχο (Φεβ. 2003) προέκυψαν 1950 hostname πιθανών

[64] προς το παρόν δεν είναι ακριβώς crawler, διότι δεν ελέγχει εξαντλητικά το περιεχόμενο κάθε web host που συναντά. Στην 5.2.1 θα δούμε περισσότερα γι' αυτό.

[65] $128=2^7$. Επειδή η τεχνολογία java που χρησιμοποιεί ο δειγματολήπτης είναι πολυνηματική (multithreaded) είναι δυνατόν πολλά νήματα του ίδιου δειγματολήπτη να κάνουν τον έλεγχο "παράλληλα" μειώνοντας το χρόνο εκτέλεσης.

[66] βλ. 2.3.2

web host [Πειρ]. Στη συνέχεια, από αυτά τα hostname ζητήθηκαν να επιστραφούν web σελίδες και οι απαντήσεις που πήραμε φαίνονται στον πίνακα 5.1.

πλήθος server	**http κωδικός**
222	200
801	400
107	401
774	403
36	404
6	500
1	502
3	503

Πίνακας 5.1: Http απαντήσεις από servers του Ιορδανικού IP χάρτη

Από τον πιο πάνω πίνακα παρατηρούμε ότι μόνο τα 222 hostname επέστρεψαν έγκυρο http κωδικό (βλ. Πίνακα 3.4), που σημαίνει ότι πρόκειται για web σελίδες. Αυτά τα 222 hostname αποτελούν το Ιορδανικό web, αν αυτό οριστεί σαν το σύνολο των web host που βρίσκονται μέσα στα γεωγραφικά όρια της Ιορδανίας. Στον Πίνακα Π.γ.3 του παραρτήματος Γ βλέπουμε τους πιο πάνω 222 web host.

5.1.2 Οι web server του Ιορδανικού web

Όπως είδαμε πριν, το Μάρτιο του 2003 το Ιορδανικό web περιείχε 222 web server. Σε αυτή την ενότητα θα ασχοληθούμε με το ποιά πλατφόρμα "έτρεχε" σε αυτούς κατά την περίοδο του πειράματος. Για την εξαγωγή αυτής της πληροφορίας θα χρησιμοποιήσουμε μία γλώσσα προγραμματισμού που ονομάζεται WebL [WebL].

Η WebL είναι μία γλώσσα προγραμματισμού ειδικά σχεδιασμένη για το web (**Web L**anguage). Είναι γραμμένη σχεδόν εξ ολοκλήρου σε java και περιέχει βιβλιοθήκες με έτοιμες αυτοματοποιημένες εργασίες (functions) που πραγματοποιούνται συχνά στο web (π.χ. συμπλήρωση φόρμας, μεταχείριση URL, μεταχείριση web σελίδων κ.λπ.).

Χρησιμοποιώντας την WebL ζητάμε από τους 222 web host να μας επιστρέψουν πληροφορίες σχετικά με το είδος της πλατφόρμας που χρησιμοποιούν. Για παράδειγμα ρωτώντας τον web host *http://193.188.95.178* παίρνουμε την απάντηση *Apache/1.3.20 (Unix) PHP/4.1.2*, που μας πληροφορεί για τον τύπο της πλατφόρμας, την έκδοσή της κ.λπ. Στον πίνακα 5.2 βλέπουμε συνοπτικά τις απαντήσεις που πήραμε.

Από τον πίνακα 5.2 παρατηρούμε την κυριαρχία της Microsoft IIS πλατφόρμας.

πλατφόρμα	πλήθος server
Apache	17
Boa	4
First Class	4
Foundry Networks	2
GeoHttpServer	1
MDN Server	2
Microsoft IIS	133
IBM Http Server	3
Netscape	24
NCSA	1
Oracle	15
Resin	2
Sambar	1
SCOL	1
Sun	1
Virata	1
Λοιπά	10

Πίνακας 5.2: Οι πλατφόρμες των 222 web server του .jo (Φεβ. 2003)

5.2 Κατανομές power law για τους out-degree web site

Στην προηγούμενη ενότητα μετατρέψαμε το δειγματολήπτη ώστε να λειτουργεί περίπου σαν crawler. Δεν ήταν ακριβώς crawler, διότι κάθε web site που συναντούσε δεν το εξερευνούσε ακολουθώντας τα link των σελίδων του. Σε αυτή την ενότητα θα μετατρέψουμε το δειγματολήπτη σε κανονικό crawler, ο οποίος θα μπορεί να εξερευνεί οποιοδήποτε web site συναντά υπολογίζοντας, ταυτόχρονα, το out-degree κάθε σελίδας του (βλ. 1.2.2 και 1.2.4).

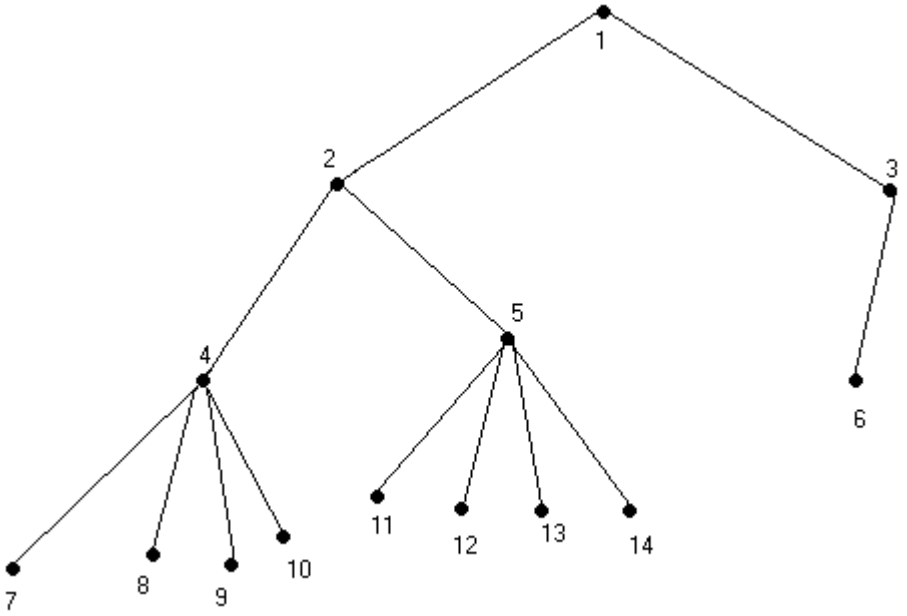
5.2.1 Η δεντρική προσέγγιση ενός web site

Στην 1.3 είχαμε δει ότι μία συλλογή από web σελίδες, οι οποίες δικτυώνονται έτσι ώστε να αποτελούν μία ομάδα με συγκεκριμένο πληροφοριακό αντικείμενο, ονομάζεται web site. Συνήθως, τα web site είναι έτσι οργανωμένα ώστε πρόσβαση στην αρχική τους σελίδα (home page) να μας επιτρέπει να επισκεφτούμε οποιοδήποτε σημείο τους.

Σύμφωνα με την [ZYD01] ένα ρεαλιστικό μοντέλο για την αναπαράσταση ενός web site είναι ένα δέντρο με ρίζα (rooted tree). Μία δομή δέντρου με ρίζα φαίνεται στη φιγούρα 5.1. Από αυτή τη φιγούρα παρατηρούμε ότι η αντιστοιχία με ένα web site είναι προφανής (βλ. πίνακα 5.3). Η ρίζα του δέντρου είναι η home page ενώ τα κλαδιά είναι οι υπερσύνδεσμοι που εξέρχονται από τις web σελίδες. Στη δομή ενός δέντρου υπάρχει ιεραρχία και έτσι το δέντρο μπορεί να χωριστεί σε επίπεδα. Όσο προχωράμε προς τη ρίζα τόσο ανεβαίνουμε στην ιεραρχία. Στο δέντρο της φιγούρας 5.1 παρατηρούμε ότι κάθε κόμβος του δέντρου[67] έχει τόσα παιδιά όσοι είναι και οι

[67] Κάθε δέντρο είναι ταυτόχρονα και γράφος με κόμβους και ακμές.

κόμβοι των πιο κάτω επιπέδων με τους οποίους ενώνεται. Για παράδειγμα, στη φιγούρα 5.1 η ρίζα (που βρίσκεται στο επίπεδο 0 πάντοτε) ενώνεται με 2 κόμβους (επίπεδο 1), οπότε και λέμε ότι η ρίζα έχει 2 παιδιά. Το πιο πάνω μπορούμε να το πούμε, βέβαια, και για κάθε κόμβο του δέντρου. Όταν, τώρα, κάποιος ξεκινά από τη ρίζα ενός δέντρου και το εξερευνά κατά επίπεδα, τότε λέμε ότι κάνει μία BFS (Breadth First Search) αναζήτηση ή αλλιώς αναζήτηση κατά πλάτος.



Φιγούρα 5.1: Ένα δέντρο με ρίζα

Κόμβος	Υπερσύνδεσμος
1	http://users.ntua.gr/plekeas
2	http://users.ntua.gr/plekeas/Data_sets
3	http://users.ntua.gr/plekeas/Pened
4	http://users.ntua.gr/plekeas/Data_sets/Sampling_a_web_subgraph
5	http://users.ntua.gr/plekeas/Data_sets/metadata_in_hostnames/
6	http://users.ntua.gr/plekeas/Pened/gr_domain
7	http://users.ntua.gr/plekeas/Data_sets/Sampling_a_web_subgraph/finalsample_gr.txt
8	http://users.ntua.gr/plekeas/Data_sets/Sampling_a_web_subgraph/IP_map_of_gr.txt
9	http://users.ntua.gr/plekeas/Data_sets/Sampling_a_web_subgraph/presample_gr.txt
10	http://users.ntua.gr/plekeas/Data_sets/Sampling_a_web_subgraph/ sampling_a_web_subgraph.ppt
11	http://users.ntua.gr/plekeas/Data_sets/metadata_in_hostnames/uk_data_set.txt
12	http://users.ntua.gr/plekeas/Data_sets/metadata_in_hostnames/ http_connections_uk_data_set.txt
13	http://users.ntua.gr/plekeas/Data_sets/metadata_in_hostnames/ clusters_uk_data_set_without_duplications.txt
14	http://users.ntua.gr/plekeas/Data_sets/metadata_in_hostnames/IP_map_of_uk.txt

Πίνακας 5.3: Αντιστοιχίες κόμβων με υπερσυνδέσμους web site για το δέντρο της Φιγούρας 5.1

Ξαναγυρνώντας πάλι στα web site, έστω ότι βρισκόμαστε στην home page ενός site. Αυτή περιέχει έναν αριθμό από out link. Κάποια από αυτά δείχνουν σε σελίδες που ανήκουν στο web site ενώ τα υπόλοιπα σε web σελίδες εκτός. Ένας crawler,

τώρα, που πρόκειται να κάνει BFS αναζήτηση στο πιο πάνω site θα πρέπει να αποφύγει τα link που τον οδηγούν έξω από το δέντρο και να λάβει υπ' όψιν μόνο όσα πηγαίνουν σε σελίδες εντός. Αυτό, δηλαδή, που πρέπει να κάνει ο crawler είναι να βρει ποιά out link της home page τον οδηγούν σε σελίδες εντός του δέντρου. Το πιο πάνω μπορεί να γίνει πολύ εύκολα αν ο crawler παρατηρεί τη δομή των link που πρόκειται να ακολουθήσει.

Όπως είπαμε στην 2.3.1, κάθε web site συνήθως οργανώνεται κάτω από ένα URL το οποίο ονομάζεται βασικό URL και δείχνει την home page. Αυτό που κάνει ο crawler είναι να σπάει το κάθε link που συναντά και να βλέπει αν το βασικό URL του link συμπίπτει με το βασικό URL του δέντρου που εξερευνά. Αν ναι, το ακολουθεί, αν όχι, το απορρίπτει. Με αυτό τον τρόπο ο crawler δε θα βγει ποτέ εκτός του δέντρου και κάποια στιγμή θα εξερευνήσει όλο το web site.

Ερχόμαστε, τώρα, στα out-degree του web site. Μία σελίδα[68] ενός web site, όπως είπαμε, περιέχει κάποια out link που δείχνουν εντός και εκτός του web site. Τα link που δείχνουν εντός είναι μεν χρήσιμα για το crawling αλλά, όπως θα δούμε και πιο κάτω, "αλλοιώνουν" τους out-degree σε τέτοιο βαθμό που οι power law παραμορφώνονται. Αυτό το φαινόμενο εμφανίζεται διότι out links που "δείχνουν" μέσα στο ίδιο το web site μπορεί να επαναλαμβάνονται πολλές φορές σε διάφορες σελίδες[69] συνεισφέροντας παραπάνω από ό,τι ίσως θα έπρεπε στον υπολογισμό του out-degree. Στην ενότητα 5.2.2 θα δούμε περισσότερα γι' αυτό. Προς το παρόν η πιο πάνω παρατήρηση μας οδηγεί στο να διαχωρίσουμε την ανάλυση των συνδέσμων ενός web site σε inter- και intra-site ανάλυση[70].

5.2.2 Υπολογισμός power law για web site

Σύμφωνα με τα παραπάνω μπορούμε να τροποποιήσουμε το δειγματολήπτη ώστε να κάνει crawl ολόκληρα web site και ταυτόχρονα να εξάγει τα out-degree των σελίδων που συναντά.

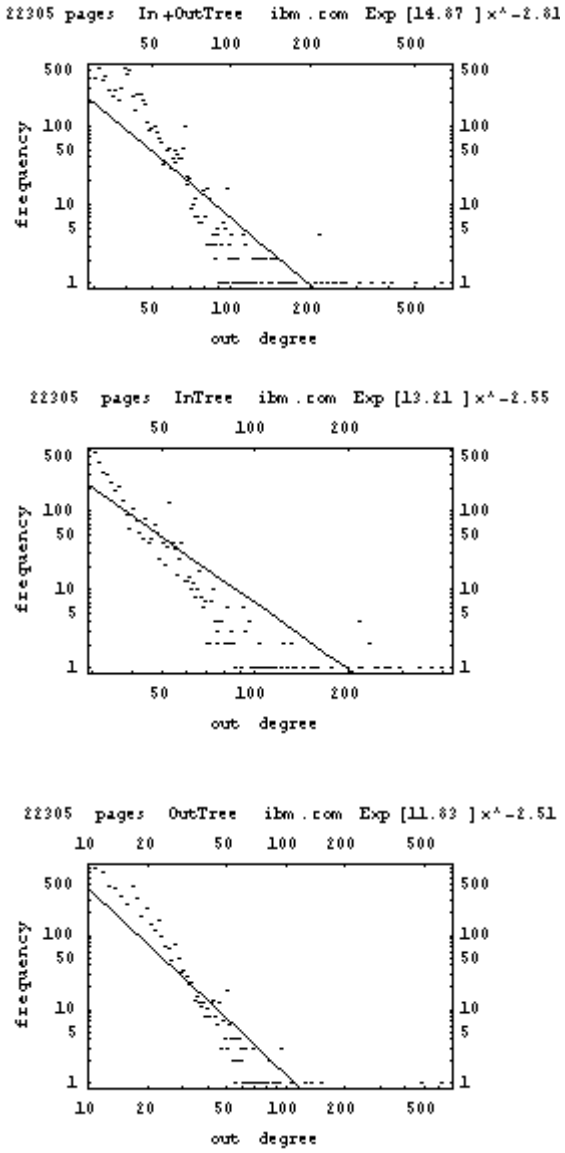
Η τροποποίηση του δειγματολήπτη έγινε με τη χρήση της γλώσσας WebL [WebL] και διάφορα αποτελέσματα φαίνονται στις φιγούρες που ακολουθούν.

Στη φιγούρα 5.2 παρατηρούμε τις power law κατανομές για τα out-degree από 23,305 σελίδες του web site www.ibm.com. Η intra- και inter-link ανάλυση μας δείχνει ότι στις δύο πρώτες γραφικές παραστάσεις (In+OutTree και InTree) φαίνεται η όχι τόσο καλή ταύτιση των ευθειών προσαρμογής[71]. Στην περίπτωση, όμως, που κρατήσουμε μόνο τα link εκτός web site (OutTree) παρατηρούμε μία αρκετά καλή προσαρμογή.

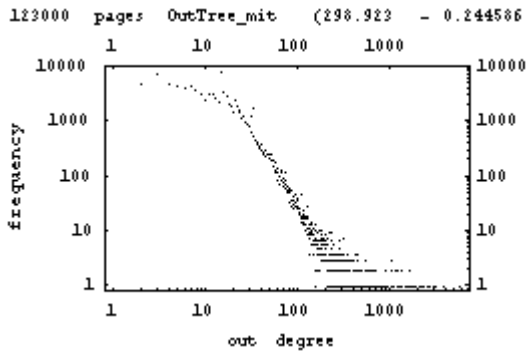
Στη φιγούρα 5.3 παρατηρούμε τις power law κατανομές για τα out-degree από 123,000 σελίδες του web site www.mit.edu

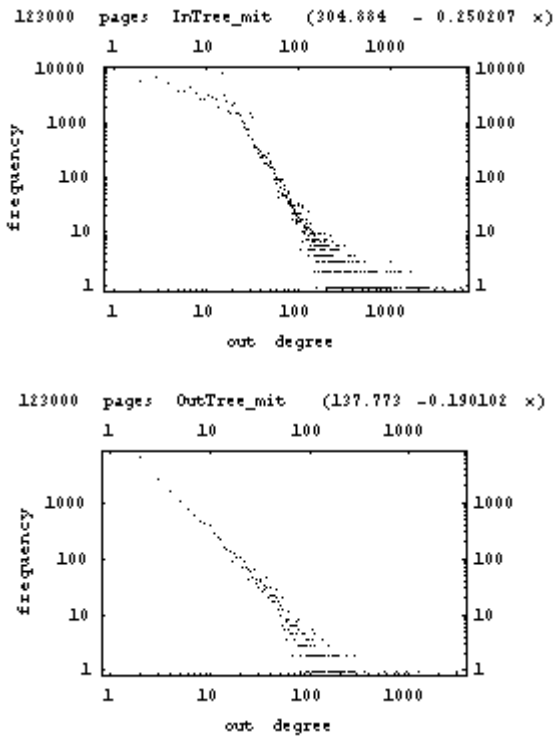
[68] Αναφερόμαστε σε σελίδες με text/html περιεχόμενο.
[69] π.χ. το out link που δείχνει πάντα την home page σε πολλά web site εμφανίζεται σχεδόν σε όλες τις σελίδες τους.
[70] Αντίστοιχος διαχωρισμός συναντιέται και στην [BCHR01].
[71] Η προσαρμογή έγινε με τη μέθοδο των ελαχίστων τετραγώνων και οι γραφικές παραστάσεις με το πρόγραμμα mathematica [Mat96].

Τέλος στη φιγούρα 5.4 παρατηρούμε την power law κατανομή για τα out-degree δύο web site (www.mit.edu, www.ntua.gr).

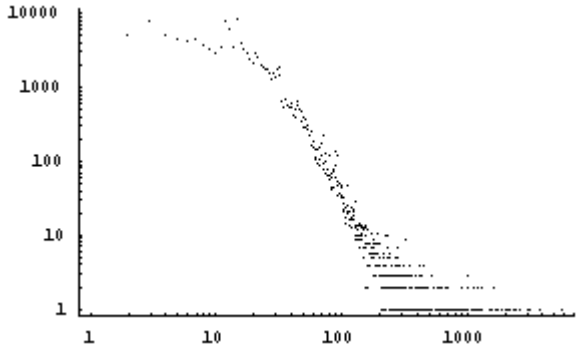


Φιγούρα 5.2: Power law κατανομές για τα out-degree του web site www.ibm.com





Φιγούρα 5.3: Power law κατανομές για τα out-degree του web site web.mit.edu



Φιγούρα 5.4: Power law κατανομή για τα out-degree των web site web.mit.edu και www.ntua.gr

Από τα πιο πάνω παρατηρούμε ότι επαληθεύονται κάποια χαρακτηριστικά της fractal δομής του web, αφού και σε επίπεδο μελέτης web site ισχύει η power law κατανομή για τους out-degree.

5.3 Επίλογος

Σε αυτό το κεφάλαιο είδαμε δύο διαφορετικές χρήσεις του IP δειγματολήπτη.

Η πρώτη ήταν ο δειγματολήπτης να εξερευνά όλα τα IP του χάρτη βλέποντας αν ανήκουν σε web server και παίρνοντας πληροφορίες γι' αυτούς. Με αυτή τη χρήση

εντοπίστηκε ο αριθμός των server μέσα στο .jo domain και εξακριβώθηκε η πλατφόρμα που αυτοί χρησιμοποιούν.

Η δεύτερη ήταν ο δειγματολήπτης να δουλεύει σαν crawler εξερευνώντας ολόκληρα web site και υπολογίζοντας out-degree. Με αυτή τη χρήση κάναμε crawl διάφορα web site και επαληθεύτηκε ότι οι out-degree ακολουθούν κατανομή power law και σε επίπεδο web site.

6 ΣΥΝΕΙΣΦΟΡΕΣ ΤΗΣ ΔΙΑΤΡΙΒΗΣ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Σε αυτό το κεφάλαιο αναφέρονται οι συνεισφορές της παρούσας διατριβής και προτείνονται επιπλέον ερευνητικά θέματα στη συγκεκριμένη περιοχή.

6.1 Συνεισφορές της Διατριβής

Οι κύριες συνεισφορές της παρούσας διατριβής είναι οι ακόλουθες:

- Υλοποιείται ένας νέος δειγματολήπτης ο οποίος βελτιώνει τη μέθοδο δειγματοληψίας με IP. Συγκεκριμένα, η δειγματοληψία δεν είναι ανάγκη να γίνει σε όλο τον IPv4 χώρο [LG99, CHLZ03], αλλά μπορεί να περιοριστεί και σε υποσύνολά του, όπως τα ccTLD. Η συνεισφορά αυτή είναι σημαντική, διότι τα δείγματα που χρειάζεται να ληφθούν δεν είναι απαραίτητο να είναι μεγάλα σε μέγεθος. Επιπλέον, υπάρχει δυνατότητα η δειγματοληψία να γίνεται και σε συγκεκριμένα γεωγραφικά υποσύνολα του διαδικτύου μιας και όπως αναφέρουμε στο κεφ. 5, σύμφωνα με την [Pal99] υπάρχει μεγάλη αντιστοιχία μεταξύ μίας γεωγραφικής περιοχής και ενός ccTLD. Τέλος αν λάβουμε υπ' όψιν οτι η μέθοδος δειγματοληψίας με random walk σε ένα ccTLD δεν είναι εφικτή, μιας και είναι αδύνατον να περιορίσουμε τον περίπατο μέσα στο domain χωρίς να επηρεάσουμε το δείγμα (βλ. παραγράφους 3.2 και 3.6), ο δειγματολήπτης που υλοποιήθηκε από οτι γνωρίζουμε είναι ο πρώτος που κάνει δειγματοληψία σε ccTLD.

- Μέχρι στιγμή η δειγματοληψία με IP [LG99, CHLZ03] περιοριζόταν στην αυθαίρετη επιλογή του μεγέθους του δείγματος χωρίς τη χρήση κάποιων στατιστικών μετρικών. Στο δειγματολήπτη που υλοποιήθηκε, εφαρμόστηκε η θεωρία της απλής τυχαίας δειγματοληψίας και το μέγεθος του δείγματος επιλέγεται βάσει της υπό εκτίμηση αναλογίας σε όλον τον πληθυσμό. Επιπλέον, λαμβάνονται υπ' όψιν και τα ανάλογα σφάλματα με τις αντίστοιχες διασπορές (βλ. παράγραφο 3.2.2).

- Ο δειγματολήτης που υλοποιήθηκε συνεισφέρει στη λύση του προβλήματος της μεροληψίας που εισάγει η αυθαίρετη επιλογή σελίδας αφετηρίας σε τυχαίους περιπάτους στο web. Συγκεκριμένα, όπως αναφέραμε και στο κεφ. 2, ως αφετηρία ενός random walk, είτε επιλέγονται ισχυρά συνεκτικές σελίδες (οι οποίες με μεγάλη πιθανότητα βρίσκονται στον πυρήνα του web γράφου), είτε επιλέγονται σύνολα από πιθανές αφετηρίες από τις οποίες ο random walk επιλέγει μία [HHMN99, HHMN00]. Η συνεισφορά του δειγματολήπτη που υλοποιήθηκε βρίσκεται στο γεγονός οτι η έξοδός του μπορεί να αποτελεί το σύνολο από πιθανές αφετηρίες ενός random walk. Με αυτό τον τρόπο μειώνεται η αυθαίρετη επιλογή του συνόλου των πιθανών αφετηριών.

- Στην παρούσα διατριβή αποδεικνύεται οτι υπάρχει συσχέτιση χωρικής και χρονικής πληροφορίας με τον τρόπο γραφής των hostname. Συγκεκριμένα, με τη βοήθεια του δειγματολήπτη, επιλέχτηκε ένα δείγμα από hostnames από το .uk ccTLD. Συστάδες από αυτό το δείγμα μας έδωσαν χωρικές και χρονικές πληροφορίες σχετικά με Βρετανικούς παροχείς Ίντερνετ. Ειδικότερα

υπολογίστηκε η χωρική κατανομή της υποδομής ενός Βρετανικού ISP και ο ρυθμός ανάπτυξης ενός άλλου (βλ. παραγράφους 4.4 και 4.5). Οι πιο πάνω πληροφορίες αποδεικνύονται πολύ χρήσιμες διότι μπορούν να χρησιμεύσουν σαν αμερόληπτες εκτιμήτριες των συγκεκριμένων εταιρειών. Για παράδειγμα, με αυτόν τον τρόπο μπορεί να εκτιμηθεί το μέγεθος της υποδομής των δρομολογητών ενός ISP χωρίς να βασιζόμαστε σε πληροφορίες από την ίδια την εταιρεία αλλά σε πληροφορίες που παίρνουμε μόνο από το δίκτυο. Επιπλέον, είναι δυνατόν με την πιο πάνω μέθοδο να έχουμε αμερόληπτη εκτίμηση που αφορά το ρυθμό ανάπτυξης συγκεκριμένων ISP ως προς το πόσους εξυπηρετητές "ρίχνουν" στο δίκτυο, κάτι που από όσο γνωρίζουμε δεν έχει γίνει ποτέ χρησιμοποιώντας μόνο πληροφορίες από το δίκτυο. Η εκτίμηση αυτή μπορεί να μας βοηθήσει να βγάλουμε ανεξάρτητα συμπεράσματα για τον αριθμό των πελατών της κάθε εταιρείας, τα οποία δεν θα βασίζονται σε υποκειμενικές πηγές (π.χ. διαφημιστικές καμπάνιες). Επιπλέον με τη χρήση της πιο πάνω μεθόδου είναι δυνατόν να εκτιμηθεί και η κυκλοφορία Ίντερνετ (Internet traffic) των ISP για τη συγκεκριμένη ημέρα της δειγματοληψίας. Τέλος η ίδια μέθοδος μπορεί να αποτελέσει και ένα τρόπο χαρτογράφησης εξυπηρετητών (host) σε γεωγραφικές περιοχές διαφορετικό από αυτόν που προτείνεται στις εργασίες [PS01, ZFRD02].

Δευτερεύουσες συνεισφορές της παρούσας διατριβής είναι οι ακόλουθες:

- Από όσο γνωρίζουμε για πρώτη φορά εκτιμήθηκε με αρκετά καλή ακρίβεια το μέγεθος του Ελληνικού web με τεχνικές δειγματοληψίας (βλ. παράγραφο 3.4.2). Άλλη μέθοδος που εφαρμόστηκε για τον ίδιο σκοπό (βλ. παράγραφο 3.4.1) υπολόγισε το ίδιο αποτέλεσμα (λαμβανομένου και του ρυθμού ανάπτυξης του web) αλλά, σε πολύ περισσότερο χρόνο, αφού έκανε εξαντλητική αναζήτηση του .gr, και με πολύ περισσότερα χρήματα.

- Εξερευνήθηκε εξαντλητικά ο Ιορδανικός χάρτης του Ίντερνετ και υπολογίστηκε ο αριθμός των web server του. Στην ίδια περίπτωση καταγράφτηκε επίσης η πλατφόρμα κάθε server και ανακαλύφθηκε η κυριαρχία συγκεκριμένης εταιρείας λειτουργικών συστημάτων (βλ. παράγραφο 5.2).

- Τροποποιημένη μορφή του δειγματολήπτη χρησιμοποιήθηκε ώστε να γίνεται εξαντλητική αναζήτηση ολόκληρων web site. Με αυτό τον τρόπο επαληθεύτηκε η κατανομή αντιστρόφου πολυωνύμου για τους out degree και σε επίπεδο web site.

6.2 Μελλοντική έρευνα

Σε αυτή την παράγραφο σχολιάζουμε μερικά από τα προβλήματα που προκύπτουν από την παρούσα διατριβή και κάποια άλλα που παραμένουν άλυτα.

- Ανοιχτά παραμένουν τα δύο σημαντικά προβλήματα των μεθόδων δειγματοληψίας που αναφέραμε στα προηγούμενα κεφάλαια, δηλ. 1) του προβλήματος της τυχαίας και ομοιόμορφης επιλογής (uniformly at random) μίας web σελίδας και, 2) του προβλήματος της ομοιόμορφης και τυχαίας δειγματοληψίας ενός web site. Σημειώνεται πάντως ότι, η μέθοδος που

προτείναμε δίνει μία μερική λύση στο πρόβλημα της τυχαίας επιλογής μίας web σελίδας από ένα ccTLD υπό την προϋπόθεση ότι, αναφερόμαστε σε επίπεδο αρχικών σελίδων (home page), και ότι εφαρμόζεται η τεχνική zone transfer (βλ. Παράγραφο 3.5.2).

- Ενδιαφέρον θα είχε η εφαρμογή διαφορετικών μεθόδων δειγματοληψίας π.χ. στρωματοποιημένη τυχαία δειγματοληψία (stratified random sampling) και δειγματοληψία κατά συστάδες (cluster sampling). Σύμφωνα με την πρώτη μέθοδο δειγματοληψίας, ο χάρτης μπορεί να χωριστεί σε στρώματα (strata) και να επιλέγονται IP από εκεί με ομοιόμορφο τρόπο. Σύμφωνα με τη δεύτερη μέθοδο, ο χάρτης μπορεί να χωριστεί σε συστάδες (clusters) και η δειγματοληψία να γίνει σε αυτές. Η διαφορά των δυό μεθόδων είναι οτι ο δεύτερος τρόπος απαιτεί επιπλέον γνώση για τις δικτυακές οντότητες του IP χάρτη πριν γίνει η συσταδοποίησή του.

- Περαιτέρω έρευνα χρειάζεται να γίνει για το Deep web, με τη μέθοδο IP sampling μιας και όπως φαίνεται από την εργασία [CHLZ03] υπάρχει μεγάλο και αυξανόμενο ενδιαφέρον για την πληροφορία που κρύβεται πίσω από αυτό. Όπως είδαμε και στην παράγραφο 2.4.2, η μέθοδος IP sampling είναι η μόνη που μπορεί να χρησιμοποιηθεί για επιλογή δειγμάτων από το Deep web μιας και η μέθοδος των τυχαίων περιπάτων αποτυγχάνει.

- Ενδιαφέρον θα είχε η προσπάθεια συνδυασμού των δύο μεθόδων δειγματοληψίας (IP sampling και random walk), π.χ. η έξοδος του IP δειγματολήπτη να είναι είσοδος του random walk. Αυτό όπως αναφέραμε και πιο πριν είναι σημαντικό διότι έτσι θα πετύχουμε ακόμα καλύτερη προσέγγιση της ομοιομορφίας του δείγματος με random walk αφού θα ελαττώσουμε τη μεροληψία της αφετηρίας. Επιπλέον με την πιο πάνω προσπάθεια θα έχουμε τη δυνατότητα να κάνουμε δειγματοληψία με random walk και στο Deep web το οποίο μέχρι σήμερα είναι κατά πολύ ανεξερεύνητο.

- Βελτιστοποίηση της μεθόδου δειγματοληψίας θα μπορούσε να προκύψει στα εξής θέματα:

    - Πλήρης αυτοματοποίηση της δημιουργίας του IP χάρτη των συγκεκριμένων domain. Αυτό μπορεί να γίνει αν δεν ξεχωρίζουμε τους τοπικούς ληξίαρχους από τα αυτόνομα συστήματα, υπολογίζοντας με διαφορετικό τρόπο τις αναθέσεις που τους γίνονται, αλλά απευθύνουμε ερωτήσεις για όλες τις δικτυακές οντότητας (LIR και AS) στη whois βάση καταρτίζοντας απευθείας το χάρτη.

    - Δυνατότητα σχηματισμού IP χαρτών όχι μόνο σε ccTLD αλλά και στα υπόλοιπα 6 domain (.edu, .com, .gov, .mil, .org, .net). Αυτό το πρόβλημα είναι πολύ σημαντικό διότι θα μας βοηθήσει να καταλάβουμε περισσότερο τη φύση του πολύπλοκου δικτύου του web αφού θα μπορούμε πλέον να κάνουμε δειγματοληψία σε οποιοδήποτε κομμάτι του. Η δυσκολία του πιο πάνω προβλήματος επικεντρώνεται κυρίως σε 2 ζητήματα: 1) τα 6 domain στα οποία θέλουμε να κάνουμε δειγματοληψία δεν έχουν κάποιο γεωγραφικό περιορισμό, με την έννοια ότι, οι υπολογιστές που συνεισφέρουν στο περιεχόμενό τους

δεν βρίσκονται (σε μεγάλο ποσοστό) εντοπισμένοι σε συγκεκριμένη γεωγραφική περιοχή. Αυτό κάνει δύσκολη τη σύνταξη IP χαρτών. 2) Οι LIR και τα AS που αποτελούν αυτά τα domain ανήκουν στη δικαιοδοσία διαφορετικών ηπειρωτικών ληξίαρχων, οι οποίοι αρκετές φορές υιοθετούν διαφορετικές πολιτικές για τον τρόπο διαχείρισης των δεδομένων τους. Για παράδειγμα ο RIR ARIN δεν διαθέτει πληροφορίες σχετικά με το τι IP έχουν ανατεθεί σε AS της περιοχής ευθύνης του.

- Προσαρμογή της μεθόδου δειγματοληψίας για το νέο Ίντερνετ (IPv6). Το νέο Ίντερνετ (έκδοση 6) επειδή έχει διευθύνσεις μήκους 128 bit, θα έχει ασύγκριτα μεγαλύτερο αριθμό διευθύνσεων (2^{128}) πράγμα που θέτει την τεχνική IP sampling σε μεγάλη δοκιμασία ως προς την κλιμάκωση (scaling) της όλης διαδικασίας σε τόσους μεγάλους χώρους. Ενδιαφέρον λοιπόν πρόβλημα θα ήταν η αποτελεσματική προσαρμογή της μεθόδου στη νέα έκδοση του Ίντερνετ. Δύο προβλήματα τα οποία εμποδίζουν προς το παρόν μία τέτοια προσπάθεια είναι: 1) το γεγονός οτι παρόλο που πολλοί ηπειρωτικοί ληξίαρχοι έχουν αρχίσει και κάνουν αναθέσεις του IPv6 χώρου, δεν διατηρούν συστηματικά αρχεία από τα οποία μπορούμε να πάρουμε την αντίστοιχη πληροφορία, 2) η μετάβαση από το παλιό στο νέο Ίντερνετ δεν έχει ακόμα επιτευχθεί. Αυτό που μέχρι τώρα εφαρμόζεται είναι η σταδιακή μετάβαση από την έκδοση 4 στην έκδοση 6[72] χωρίς να υπάρχει κάποια "flag day" στην οποία όλα τα μηχανήματα θα γυρίσουν στη νέα έκδοση.

- Δημιουργία καλύτερων φίλτρων ώστε στην περίπτωση που θέλουμε να πάρουμε δείγμα από web σελίδες, τη διεύθυνση κάθε σελίδας να την παίρνουμε μετά από επεξεργασία του περιεχομένου της σελίδας και όχι μόνο από τον web host που την φιλοξενεί. Αυτό είναι ένα αρκετά δύσκολο πρόβλημα που απαιτεί εκτός από γνώσεις πολλών προγραμματιστικών εργαλείων και γλωσσών προγραμματισμού, γνώσεις τεχνικών εξόρυξης πληροφορίας (data mining) από ημιδομημένα δεδομένα (semistructured data) όπως η πληροφορία που περιέχεται στις web σελίδες.

- Βελτίωση της ταχύτητας δειγματοληψίας η οποία μπορεί να γίνει με βελτίωση του κώδικα. Ο κώδικας θα μπορεί να επανασχεδιαστεί σε γλώσσα WebL (βλ. παραγράφους 5.1.2 και 5.2.2) η οποία παρότι είναι γραμμένη σχεδόν εξ ολοκλήρου σε java κάνει χρήση έτοιμων ρουτίνων, τύπων και συναρτήσεων, οι οποίες μπορούν να ενσωματωθούν, αυτούσιες, στον κώδικα αυξάνοντας την απόδοσή του.

[72] Η έκδοση 5, όπως έχουμε πει, δεν θα χρησιμοποιηθεί ποτέ αλλά αναπτύχθηκε για ερευνητικούς σκοπούς.

Βιβλιογραφία

[Afr02] Afrati, F. N. On Approximation Algorithms for Data Mining Applications. To be considered for publication in the book: *Approximation Algorithms* (to be edited by E. Bampis, K. Jansen and C. Kenyon).

[BA99] Barabasi, A., R. Albert. *Emergence of scaling in random networks*. Science, 286:509, October 1999.

[BBC+00] Bar-Yossef, Z. B., A. Berg, S. Chien, J. Fakcharoenphol, D. Weitz. Approximating Aggregate Queries about Web Pages via Random Walks. *Proceedings of the 26th VLDB Conference*, Cairo, Egypt, 2000.

[BCHR01] Bharat, K., B.-W. Chang, M. Henzinger, M. Ruhl. Who Links to Whom: Mining Linkage between Web Sites. *Proceedings of the IEEE International Conference on Data Mining*, November 2001.

[BCSV02] Boldi, P., B. Codenotti, M. Santini, S. Vigna. Structural Properties of the African Web. *Proceedings of the 11th International World Wide Web Conference*, Hawai, USA, 2002.

[BKM+00] Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener. Graph Structure in the Web. *Proceedings of the 9th International World Wide Web Conference*. Also in *Computer Networks 33*, 1-6, 309-320, 2000.

[BP98] Brin, S., L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, 107-117, 1998.

[CB98] Cheswick, B., H. Burch. The Internet Mapping Project. http://research.lumeta.com/ches/map/index.html

[CHLZ03] Chang, K. C.-C., B. He, C. Li, and Z. Zhang. Structured Databases on the Web: Observations and Implications. *Technical Report UIUCDCS-R-2003-2321*, Department of Computer Science, University of Illinois at Urbana-Champaign (*UIUC*), February 2003.

[CK97] Carriere, J., R. Kazman. Webquery: Searching and visualizing the web through connectivity. *Proceedings of the 6th International World Wide Web Conference*, Santa Clara, California, pp. 701-711, 1997.

[DKM+02] Dill, S., R. Kumar, K. S. McCurley, S. Rajagopalan, D. Sivakumar, A. Tomkins. Self-Similarity In the Web. *ACM Transactions on Internet Technology*, Vol. 2, No. 3, August 2003, Pages 205-223.

[Eur] Eurostat, Statistical Office of the European Communities. http://www.europa.eu.int/comm/eurostat

[FFF99] Faloutsos, M., P. Faloutsos, C. Faloutsos. On power law relationships of the internet topology. *Proceedings of the ACM SIGCOMM Conference*, 251-262, 1999.

[Fla00] Flanagan David, Java Examples in a Nutshell, 2nd. ed. O' Reilly & Associates, United States of America, 2000.

[Fla02] Flanagan David, Java in a Nutshell, 4th. ed. O' Reilly & Associates, United States of America, 2002.

[GL02] Guillaume, J.-L., M. Latapy. The Web graph: an overview. *In 4èmes rencontres francophones sur les Aspects Algorithmiques des Telecommunications (AlgoTel)*, INRIA, 2002.

[GMTa] The Generic Mapping Tools, GMT Version 3.4.2 Technical Reference Cookbook by P. Wessel and W. H. F. Smith, Laboratory for Satellite Altimetry, NOAA/NESDIS/NODC, October 2002.

[GMTb] The Generic Mapping Tools, GMT Version 3.4.2 A Map Making Tutorial by P. Wessel and W. H. F. Smith, Laboratory for Satellite Altimetry, NOAA/NESDIS/NODC, October 2002.

[GT00] Govidan, R., H. Tangmunarunkit. Heuristics for Internet Map Discovery, *IEEE INFOCOM* 2000.

[Hen03] Henzinger, M. R. Algorithmic Challenges in Web Search Engines. *Internet mathematics* Vol. 1, No 1, 2003, pages 115-126, A. K. Peters, Ltd.

[HHMN00] Henzinger, M., R. A. Heydon, M. Mitzenmacher, M. Najork. On near-uniform URL sampling. *Proceedings of the 9th International World Wide Web Conference*, pp. 295-308, Amsterdam. Also in *Computer Networks*, 33 (2000), pages 295- 308, Elsevier.

[HHMN99] Henzinger, M., R. A. Heydon, M. Mitzenmacher, M. Najork. Measuring Index Quality using Random Walks on the Web. *Proceedings of the 8th International World Wide Web Conference*, Toronto, Canada, pages 213-225. Elsevier Science, May 1999.

[Hor89] Hornby, A. S., Oxford Advanced Learner's Dictionary (Of Current English), 4th. ed. Oxford University Press, Oxford, 1989.

[IANA] Internet Assigned Numbers Authority, http://www.iana.org

[IP2LL] IP2LL, IP to Latitude/Longitude tool, http://cello.cs.uiuc.edu/cgi-bin/slamm/ip2ll

[KKR+99] Kleinberg, J., R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins. The web as a graph: measurements, models and methods, *Lecture Notes in Computer Science*, 1627, (1999), pp. 1-17.

[KR88] Kernighan B. W., D. M. Ritchie. *Η Γλώσσα Προγραμματισμού C*. Μετάφραση από το Πρωτότυπο, Εκδόσεις Κλειδάριθμος, Αθήνα, 2001.

[Lek03a] Lekeas, P. Extracting Information for Internet Service Providers from Hostname Samples. *Proceedings of the International Computational Management Science Conference*, May 2003, Chania, Greece.

[Lek03b] Lekeas, P. Sampling a Web Subgraph. *Proceedings of the 5th Algorithms, Scientific Computing, Modelling and Simulation web conference (5th ASCOMS)*, New York, USA, Sept. 15-17, 2003. Also in WSEAS Transactions on Computers, Issue 3, Vol. 2, July 2003, ISSN 1109-2750, pp. 618-622.

[Lek04] Lekeas, P. Observing spatial and Temporal Metadata in Hostnames. (To appear in) *Journal of Information and Knowledge Management (JIKM)*, March Issue 2004, World Scientific, ISSN: 0219-6492.

[LG99] Lawrence, S., C. L. Giles. Accessibility of information on the web. *Nature*, vol. 400, 8 July 1999.

[Mat96] Mathematica® 3.0 . *Wolfram Research* 1996. http://www.wolfram.com

[MPDC00] Moore, D., R. Periakaruppan, J. Donohoe, K. Claffy, "Where in the world is netgeo.caida.org?", *Proceedings of the* INET'2000, Yokohama, Japan, July 2000.

[OML97] O' Neil, E. T., P. D. McClain and B. F. Lavoie. A Methodology for Sampling the World Wide Web. *Technical report, OCLC Annual Review of Research,* 1997, http://wcp.oclc.org

[Pal99] Paltridge, S. OECD regulatory and statistical update, *Telecommunication Policy* 23 (1999), 683-686, Elsevier.

[Pap03] Papadimitriou C., The Emergent Science of the Internet and the Web. The 2003 Lectures In Computer Science: Internet and Web: Crawling the Algorithmic Foundations. *The Onassis Foundation Science Lecture Series*, Heraklion Crete, July 7-11, 2003, http://www.forth.gr/onassis

[PD96] Peterson, L. L., B. S. Davie. Computer Networks: A Systems Approach. *Morgan Kaufmann*, San Francisco, CA, 1996.

[PFTV02][73] Press, W. V., B. P. Flannery, S. A. Teukolsky, W. T. Vetterling. Numerical Recipes in C : The Art of Scientific Computing. *Cambridge University Press*, New York, 2002.

[73] Με την άδεια του Cambridge University Press, ηλεκτρονική μορφή του βιβλίου υπάρχει δωρεάν στη διεύθυνση: http://www.nr.com

[PM88] Park, S. K., K. W. Miller. Random Number Generators: Good ones are hard to find. *Communications of the ACM* 31 (10), October 1988.

[PS01] Padmanabhan V. N., L. Subramanian, An investigation of geographic mapping techniques for Internet hosts, in *Proceedings of the ACM SIGCOMM'2001*, San Diego, CA, USA, Aug. 2001.

[RFC] Request For Comments, (π.χ. http://www.faqs.org/rfcs ή http://www.rfc -editor.org/rfcsearch.html)

[RIPE] RIPE NCC, Reseaux IP Europeens Network Coordination Centre, http://www.ripe.net

[RPLG01] Rusmevichientong, P., D. M. Pennock, S. Lawrence, C. L. Giles. Methods for Sampling Pages Uniformly from the World Wide Web. *Proceedings of the AAAI Fall Symposium on Using Uncertainty within Computation*, pp. 121-128, Menlo Park: AAAI Press, 2001.

[SH85] Stavropoulos, D. N., A. S. Hornby, *Oxford English-Greek Learner's Dictionary*, 2nd impression, Oxford University Press, Oxford, 1985.

[SMO79] Scheaffer, R. L., W. Mendenhall, L. Ott. Elementary Survey Sampling, 2nd ed. *PWS Publishers*, United States of America, 1979.

[Sun] Sun Microsystems, Inc. 901 San Antonio, Road, Palo Alto, California 94303, http://www.sun.com

[Sve98] Svensson, M. Social Navigation. Chapter 6 from Dahlback, Nils (ed.) Exploring Navigation: Towards a Framework for Design and Evaluation of Navigation in Electronic Spaces, *Swedish Institute of Computer Science* TR98:01, 1998.

[UKNS01] UK's National Statistics online, 2001 Census of population of England and Wales, http://www.statistics.gov.uk

[WAB+96] Woodruff, A., P. M. Aoki, E. Brewer, P. Gauthier, L. P. Rowe. An Investigation of Documents from the World Wide Web. *Proceedings of the 5th International World Wide Web Conference*, May 6-10, 1996, Paris, France.

[Wai99] Wainwright, P. Professional Apache, *Wrox Press*, Birmingham, 1999.

[wcp] Web Characterization Project, http://wcp.oclc.org

[WebL] Compaq's Web Language. A programming Language for the Web. Reference manual for ver 3.0, http://research.compaq.com/SRC/WebL

[ZFRD02] Ziviani, A., S. Fdida , J. F. Rezende, O. C. M. B. Duarte. Placing Landmarks to Locate Internet Hosts. *Workshop on Quality of Service*

and Mobility, WQoSM 2002, Angra dos Reis, RJ, Brazil, November 2002.

[ZYD01] Zhu, X., J. Yu, D. John. Heavy Tails, Generalized Coding, and Optimal Web Layout. *IEEE INFOCOM* 2001.

[ΑΠ93] Αφράτη, Φ., Γ. Παπαγεωργίου. Αλγόριθμοι: Μέθοδοι Σχεδίασης και Ανάλυση Πολυπλοκότητας. *Εκδόσεις Συμμετρία*, Αθήνα, 1993.

[ΝΣ96] Νικολετσέας, Σ., Π. Σπυράκης. Στοιχεία της Πιθανοτικής Μεθόδου. Μαθηματικές Θεμελιώσεις της Επιστήμης των Υπολογιστών, Τόμος Ι. *Εκδόσεις Gutenberg*, Αθήνα, 1996.

[Πειρ] Πειραματικά δεδομένα, http://users.ntua.gr/plekeas/

Παραρτήματα

Παράρτημα Α: Λίστα Συντμήσεων

APNIC	Asia Pacific Network Information Centre
ARIN	American Registry for Internet Numbers
AS	Autonomous System
bit	binary digit
DNS	Domain Name System
html	hypertext markup language
http	hypertext transfer protocol
IANA	Internet Assigned Numbers Authority
IP	Internet Protocol
IPv4(6)	Internet Protocol version 4(6)
ISP	Internet Service Provider
LACNIC	Regional Latin American and Caribbean IP Address Registry
LIR	Local Internet Registry
RFC	Request For Comments
RIPE NCC	Reseaux IP Europeens Network Coordination Centre
RIR	Regional Internet Registry
SCC	Strong Connected Component
TCP/IP	Transmission Control Protocol/Internet Protocol
TLD	Top Level Domain
ccTLD	country code Top Level Domain
URL	Uniform Resource Locator
WebL	Web Language
www	world wide web

Παράρτημα Β: Χάρτες διευθύνσεων

- Πιο κάτω βλέπουμε το χάρτη αναθέσεων των IP διευθύνσεων σε τοπικούς ληξίαρχους (LIR) που δραστηριοποιούνται στην Ελλάδα όπως ήταν διαμορφωμένος την 26η Νοε. 2002 (πηγή: ftp://ftp.ripe.net/ripe/stats). Επίσης, βλέπουμε μία εκτέλεση του αλγορίθμου επιλογής IP από το συγκεκριμένο χάρτη.

62	1	0	0	65536
62	38	0	0	65536
62	68	64	0	8192
62	68	128	0	8192
62	74	0	0	65536
62	75	0	0	32768
62	103	0	0	65536
62	169	192	0	16384
62	216	32	0	8192
80	76	32	0	4096
80	106	0	0	131072
80	245	160	0	4096
193	92	0	0	65536
194	30	192	0	8192
194	30	224	0	8192
194	177	192	0	8192
194	219	0	0	65536
195	46	0	0	8192
195	66	96	0	8192
195	74	224	0	8192
195	97	0	0	32768
195	130	64	0	16384
195	167	0	0	32768
195	170	0	0	8192
195	190	32	0	8192
195	200	64	0	8192
195	242	128	0	8192
195	251	0	0	65536
212	30	32	0	8192
212	37	64	0	8192
212	54	192	0	8192
212	70	192	0	8192
212	75	224	0	8192
212	89	160	0	8192
212	104	64	0	8192
212	107	0	0	8192
212	114	96	0	8192
212	120	192	0	8192
212	152	64	0	16384
212	205	0	0	65536
212	251	0	0	32768
213	5	0	0	65536
213	16	128	0	32768
213	140	128	0	8192
213	142	128	0	8192
213	152	0	0	8192
213	170	192	0	8192
213	249	0	0	16384
217	19	64	0	4096
217	19	80	0	4096
217	30	160	0	4096
217	69	0	0	4096
217	78	224	0	4096
217	150	224	0	4096
217	170	192	0	4096
217	195	128	0	4096
217	199	192	0	4096

Οι 4 πρώτες στήλες συμβολίζουν την αρχική IP διεύθυνση που ανατέθηκε και η 5η στήλη δείχνει τον αριθμό των διευθύνσεων της ανάθεσης. Για παράδειγμα, η 12η ανάθεση είναι μία ανάθεση 2^{12}=4096 διευθύνσεων (η μικρότερη που μπορεί να γίνει από τον RIPE). Το κομμάτι αυτό των 4096 διευθύνσεων ξεκινάει από το IP: 80.245.160.0

Ο διπλανός χάρτης περιέχει 1,216,512 διευθύνσεις IP. Εφαρμόζοντας το γραμμικό μετασχηματισμό που περιγράφεται στην 3.3.3.1 ο τυχαίος για παράδειγμα 0.189623 αντιστοιχεί στην 230,679η διεύθυνση (0.189623*1,216,512 = 230,678.65 = 230,679 με στρογγυλοποίηση στον κοντινότερο φυσικό). Η 230,679η θέση της λίστας από την αρχή αντιστοιχεί στον IP:62.75.69.22 . Η πιο πάνω αντιστοίχιση γίνεται με τον εξής αλγόριθμο:

- Βρες σε ποιά γραμμή του χάρτη ανήκει η σειρά 230,679 (ανήκει στην 6η γραμμή). Αυτό εύκολα βρίσκεται αν αρχίζουμε και προσθέτουμε τα στοιχεία της 5ης στήλης μέχρι να φτάσουμε ή να ξεπεράσουμε τη θέση που θέλουμε να βρούμε (65,536 + 65,536 + 8,192 + 8,192 + 65,536 + 32,768 = 245,760 > 230,679). Όσα στοιχεία της 5ης στήλης προσθέσαμε τόσες γραμμές πρέπει να κατεβούμε για να βρούμε τη θέση.
- Βρες σε ποιά θέση x της 6ης γραμμής ανήκει η σειρά 230,679
  x = 230,679 - 65536 - 65536 - 8192 - 8192 - 65536 = 17,687
- Θέσε α=62, β=75, γ=0, δ=0 και εκτέλεσε:

  x = x - 1
  α = α + [x/2^{24}] & x = x - 2^{24}[x/2^{24}]
  β = β + [x/2^{16}] & x = x - 2^{16}[x/2^{16}]
  γ = γ + [x/2^{8}] & x = x - 2^{8}[x/2^{8}]
  δ= δ + x

  αν δ > 255 τότε (δ = δ - 256, γ = γ + 1)

  αν γ > 255 τότε (γ = γ - 256, β = β + 1)

  αν β > 255 τότε (β = β - 256, α = α + 1)

  αν α > 255 σφάλμα (δεν επιτρέπεται α>255)

  τελικά ο αλγόριθμος εκτελείται μία φορά και δίνει:

  α=**62**, β=**75**, γ=**69**, δ=**22**

Δηλαδή, στη 230,670η θέση του χάρτη βρίσκεται ο IP: 62.75.69.22 και άρα σε αυτό τον IP αντιστοιχεί ο τυχαίος 0.189623.

(Σημείωση: [x] σημαίνει ακέραιο μέρος του x)

- Πιο κάτω βλέπουμε το χάρτη των αναθέσεων IP διευθύνσεων για την Ιορδανία όπως ήταν διαμορφωμένος το Φεβ. 2003 (πηγή: ftp://ftp.ripe.net/ripe/stats). Ο πίνακας διαβάζεται ως εξής: 8192 συνεχόμενα IP με αρχή το 62.68.96.0 έχουν ανατεθεί σε κάποια δικτυακή οντότητα που δραστηριοποιείται στην Ιορδανία, κ.ο.κ.

62.68.96.0	8192
80.69.128.0	4096
80.77.160.0	4096
80.89.192.0	4096
80.90.160.0	4096
193.188.64.0	8192
194.165.128.0	8192
195.90.96.0	8192
195.90.104.0	2048
195.158.192.0	8192
195.248.64.0	8192
196.27.0.0	256
196.27.1.0	256
212.33.192.0	8192
212.34.0.0	8192
212.35.64.0	8192
212.38.128.0	8192
212.118.0.0	8192
213.139.32.0	8192
213.186.160.0	8192
217.23.32.0	4096
217.144.0.0	4096

Πίνακας Π.β.1: Ο πλήρης IP χάρτης της Ιορδανίας (.jo) (Φεβ. 2003)

Παράρτημα Γ: Δείγματα δειγματοληψίας

Πιο κάτω ακολουθεί το προ-δείγμα και το τελικό δείγμα που προέκυψαν από τη δειγματοληψία του .gr ccTLD στο διάστημα 26/11 – 1/12/2002. Αυτά βρίσκονται και στην http://users.ntua.gr/plekeas/Data_sets/Sampling_a_web_subgraph .

Πρέπει, επίσης, να σημειώσουμε ότι λόγω του χρόνου που έχει περάσει από τη δειγματοληψία πολλά IP του δείγματος ίσως να μην αντιστοιχούν σε web σελίδες ή να αντιστοιχούν σε διαφορετικές από ό,τι αντιστοιχούσαν.

Unresolved IP-hostnames		Unresolved IP-hostnames	
1	http://212.251.73.110	56	http://athe530-t188.otenet.gr
2	http://193.92.172.171	57	http://pc003.anosis.spark.net.gr
3	http://212.205.191.111	58	http://ppp94-188.salonica.access.acn.gr
4	http://212.37.64.183	59	http://212.205.117.140
5	http://213.249.53.196	60	http://212.205.42.223
6	http://www.sacrohn.gr	61	http://213.142.137.0
7	http://www.printec.gr	62	http://athe530-t160.otenet.gr
8	http://212.37.64.43	63	http://koza364-a044.otenet.gr
9	http://217.19.84.63	64	http://polyi36-a008.otenet.gr
10	http://194.219.241.218	65	http://ppp28-200-gw06.athens.access.acn.gr
11	http://194.219.91.53	66	http://ppp72-011.patras.access.acn.gr
12	http://212.104.84.36	67	http://194.30.197.172
13	http://212.37.64.181	68	http://athe530-f077.otenet.gr
14	http://195.190.34.112	69	http://athe720-b-multi-119.otenet.gr
15	http://www.combank.gr	70	http://athe720-b-multi-121.otenet.gr
16	http://194.219.23.47	71	http://athei53-b-261.otenet.gr
17	http://212.251.52.104	72	http://ppp1-101.kal.forthnet.gr
18	http://hermes.aegean.gr	73	http://ppp28-040-gw06.athens.access.acn.gr
19	http://sanisere2.sb-clients.otenet.gr	74	http://thes530-b009.otenet.gr
20	http://mgn.mgnonset.gr	75	http://vdp186.ath06.cas.hol.gr
21	http://mailgate.tzioras.gr	76	http://vdp211.ath03.cas.hol.gr
22	http://hnodc.ncmr.gr	77	http://194.30.197.108
23	http://gamesdomain.hol.gr	78	http://athe530-g091.otenet.gr
24	http://frw-srv.guardiantrust.gr	79	http://lar-1-39.dialup.tee.gr
25	http://eshop.alphanet.gr	80	http://trip366-isdn-a023.otenet.gr
26	http://ermis.lib.uth.gr	81	http://vdp009-the01.cal.internet.gr
27	http://ds.grnet.gr	82	http://vdp01496-noc01.cos.internet.gr
28	http://draw.papiotis-draw.gr	83	http://ath11-ppp78.compulink.gr
29	http://callback0.acci.gr	84	http://athe530-t178.otenet.gr
30	http://alsinco.ath.forthnet.gr	85	http://ppp123.ser.forthnet.gr
31	http://217.19.77.248	86	http://ppp24-089-gw01.athens.access.acn.gr
32	http://212.251.73.47	87	http://212.205.155.130
33	http://195.251.210.2	88	http://athe530-h127.otenet.gr
34	http://195.167.36.46	89	http://patr364-a04.otenet.gr
35	http://194.30.228.13	90	http://213.140.133.167
36	http://194.219.183.96	91	http://www.tecnoman.gr
37	http://194.219.136.77	92	http://vdp003.krp01.gwc.hol.gr
38	http://212.251.52.90	93	http://194.30.196.10
39	http://212.251.38.231	94	http://194.30.197.124
40	http://themail.netfiles.gr	95	http://emphaze-213-5-234-80.athens.customers.acn.gr
41	http://mail.kristenmarine.com	96	http://pat-nce01.hol.gr
42	http://194.219.154.216	97	http://212.205.171.9
43	http://adultdvdcritic.com	98	http://h-212-89-166-124.ellinogermaniki.gr
44	http://212.251.24.109	99	http://217.78.225.242
45	http://193.92.119.63	100	http://217.78.224.76
46	http://poly364-a13.otenet.gr	101	http://217.78.225.130
47	http://194.219.156.49	102	http://212.89.169.32
48	http://212.205.138.195	103	http://athe-hcsd-net.customers.otenet.gr
49	http://212.205.171.32	104	http://serial01-00.pat01.acl.hol.gr
50	http://212.37.64.189	105	http://212.205.95.14
51	http://212.37.64.52	106	http://testanite.galileo.gr
52	http://212.37.65.136	107	http://212.251.52.85
53	http://213.249.53.88	108	http://dnsp.vernet.gr
54	http://62.1.210.101	109	http://mail.optimum.gr
55	http://62.1.210.99	110	http://194.219.29.10

Πίνακας Π.γ.1: Προ-δείγμα .gr (26/11-1/12/2002)

Resolved IP-hostnames		Resolved IP-hostnames	
1	http://www.sonne.gr	29	http://callback0.acci.gr
2	http://www.villagalini.gr	30	http://alsinco.ath.forthnet.gr
3	http://www.valuenet.gr	31	http://217.19.77.248
4	http://www.siotos.gr	32	http://212.251.73.47
5	http://www.schneider-electric.com.gr	33	http://195.251.210.2
6	http://www.sacrohn.gr	34	http://www.athensmap.gr
7	http://www.printec.gr	35	http://194.30.228.13
8	http://www.omilosdp.gr	36	http://194.219.183.96
9	http://www.nereusart.gr	37	http://194.219.136.77
10	http://www.magdalinos.gr	38	http://212.251.52.90
11	http://www.konsoulas.gr	39	http://212.251.38.231
12	http://www.haci.gr	40	http://themail.netfiles.gr
13	http://www.gradio.gr	41	http://www.kristenmarine.com
14	http://www.forexhellas.gr	42	http://www.iremma.com
15	http://www.combank.gr	43	http://adultdvdcritic.com
16	http://www.alexandris.gr	44	http://www.medhotels.com
17	http://www.air-conditioners.gr	45	http://www.grcreplay.com
18	http://www.aegean.gr	46 - 58	HTTP 404 - File not found
19	http://sanisere2.sb-clients.otenet.gr	59 - 90	Cannot find server or DNS Error
20	http://mgn.mgnonset.gr	91 - 96	forbidden 403
21	http://mailgate.tzioras.gr	97	Unauthorized 401
22	http://hnodc.ncmr.gr	98	HTTP 401.2 - Unauthorized
23	http://gamesdomain.hol.gr	99 - 101	firewall access denied
24	http://frw-srv.guardiantrust.gr	102 - 104	Authorization required
25	http://eshop.alphanet.gr	105 - 106	under construction
26	http://ermis.lib.uth.gr	107 - 108	server test page
27	http://ds.grnet.gr	109	asp error
28	http://draw.papiotis-draw.gr	110	Can't find server or DNS error (http://noc.sae.gr

Πίνακας Π.γ.2: Τελικό δείγμα .gr (26/11-1/12/2002)

1 http://mail.itp.com.jo/index.html
2 http://193.188.90.67
3 http://213.139.63.194
4 http://217.23.32.64
5 http://62.68.96.2
6 http://213.139.63.196
7 http://193.188.95.178
8 http://80.90.164.145
9 http://212.33.201.124
10 http://193.188.87.172
11 http://195.158.196.3
12 http://212.33.192.244
13 http://212.33.201.117
14 http://80.90.164.153
15 http://mms.fastlink.jo/mms/
16 http://194.165.136.40
17 http://194.165.142.114
18 http://194.165.159.163
19 http://212.34.2.129
20 http://212.34.2.146
21 http://212.34.2.154
22 http://212.118.1.107
23 http://212.35.70.146
24 http://212.35.71.50
25 http://212.118.0.12
26 http://212.33.195.203
27 http://217.144.0.136
28 http://212.35.70.165
29 http://194.165.142.150
30 http://212.118.8.78
31 http://193.188.66.3
32 http://194.165.134.115
33 http://80.90.162.134/mailaccess.nsf
34 http://mail.go.com.jo
35 http://webmail.cyberia.jo
36 http://193.188.90.194
37 http://194.165.145.146
38 http://212.35.70.53
39 http://193.188.73.41
40 http://193.188.95.34
41 http://194.165.145.178/onetools/
42 http://194.165.145.190/eRoomASP/DlgUnsupportedBrowser.asp
43 http://194.165.149.254
44 http://212.118.1.66
45 http://212.118.8.58
46 http://212.33.196.254
47 http://212.35.70.18
48 http://212.35.70.82
49 http://212.35.70.84
50 http://212.38.130.54
51 http://212.38.133.210
52 http://212.38.134.234
53 http://217.23.35.10
54 http://217.23.35.34
55 http://193.188.64.134
56 http://193.188.67.77
57 http://193.188.67.79
58 http://193.188.67.83
59 http://193.188.68.40
60 http://193.188.68.41
61 http://193.188.68.42
62 http://193.188.68.43
63 http://193.188.68.44
64 http://193.188.68.45
65 http://193.188.68.46
66 http://193.188.68.47
67 http://193.188.76.67
68 http://193.188.80.163
69 http://193.188.80.229
70 http://193.188.80.34
71 http://193.188.84.130
72 http://193.188.86.91
73 http://193.188.87.138
74 http://193.188.87.140
75 http://193.188.87.51
76 http://193.188.87.74
77 http://193.188.93.100
78 http://194.165.129.2
79 http://194.165.129.244
80 http://194.165.129.250
81 http://194.165.134.116
82 http://194.165.134.203
83 http://194.165.135.168/alaydi/
84 http://194.165.135.175/HIPAACertified/
85 http://194.165.135.200
86 http://194.165.141.50
87 http://194.165.145.136
88 http://194.165.145.50
89 http://194.165.145.51
90 http://194.165.145.60
91 http://194.165.146.200
92 http://194.165.148.146
93 http://194.165.148.148
94 http://194.165.148.152
95 http://194.165.148.154
96 http://194.165.148.155
97 http://194.165.148.72
98 http://194.165.148.73
99 http://194.165.149.123
100 http://194.165.149.179
101 http://194.165.149.61
102 http://194.165.149.62
103 http://194.165.151.83
104 http://194.165.151.85
105 http://194.165.152.114
106 http://194.165.153.106
107 http://194.165.153.18
108 http://194.165.153.42
109 http://194.165.155.18
110 http://194.165.155.98
111 http://194.165.159.164
112 http://194.165.159.27
113 http://212.118.0.31
114 http://212.118.1.115
115 http://212.118.1.226
116 http://212.118.1.250
117 http://212.118.1.253
118 http://212.118.1.40
119 http://212.118.10.150
120 http://212.118.10.82
121 http://212.118.2.140
122 http://212.118.31.194
123 http://212.118.31.78
124 http://212.118.8.235
125 http://212.118.8.99
126 http://212.33.192.199
127 http://212.33.193.85/default/default.asp
128 http://212.33.201.120
129 http://212.33.201.126
130 http://212.34.2.74
131 http://212.34.2.77/pc/default.asp
132 http://212.34.2.93
133 http://212.35.67.179/epicenter/
134 http://212.35.69.42
135 http://212.35.70.122
136 http://212.38.130.82
137 http://212.38.132.122
138 http://212.38.132.202
139 http://212.38.133.200
140 http://212.38.133.202
141 http://212.38.133.207
142 http://212.38.133.228
143 http://212.38.134.146
144 http://212.38.141.17
145 http://212.38.147.81
146 http://212.38.149.90
147 http://212.38.150.3
148 http://217.144.6.13
149 http://217.144.6.15/NetPerfMon/Login.asp
150 http://217.144.6.22
151 http://217.144.7.156
152 http://217.23.32.20
153 http://217.23.37.108
154 http://80.90.160.103
155 http://80.90.160.60
156 http://80.90.162.62
157 http://mail.spotcell.com
158 http://mail.telefono.com.jo
159 http://rtgsjo.cbj.gov.jo
160 http://signup.link.net.jo/content/default.asp
161 http://www.daysinn.com.jo
162 http://www.jmts-fastlink.com
163 http://www.spotcell.com
164 http://212.35.66.83
165 http://212.38.149.120
166 http://80.90.161.65
167 http://www.ccs.com.jo
168 http://www.ccs.com.jo
169 http://212.35.70.52
170 http://dinar.cbj.gov.jo
171 http://193.188.71.6
172 http://193.188.80.1
173 http://193.188.84.5
174 http://194.165.134.106
175 http://195.158.192.1
176 http://193.188.67.81
177 http://193.188.81.7
178 http://212.38.128.10
179 http://mail.nets1.com.jo
180 http://mailbk.nets1.com.jo
181 http://web.nets.com.jo
182 http://www.just.edu.jo
183 http://212.35.67.227
184 http://194.165.129.200
185 http://196.27.0.17
186 http://196.27.0.19
187 http://coolmail.jo
188 http://mars.rss.gov.jo
189 http://193.188.87.202
190 http://193.188.87.203
191 http://212.118.1.254
192 http://ns2.nets1.com.jo
193 http://212.118.0.13
194 http://193.188.93.5
195 http://212.118.8.210
196 http://212.35.66.85
197 http://212.35.71.250
198 http://212.38.147.16
199 http://213.139.63.148
200 http://217.144.6.29
201 http://217.144.6.4
202 http://193.188.65.35
203 http://193.188.65.53
204 http://193.188.68.10
205 http://193.188.68.12
206 http://193.188.88.66
207 http://212.118.1.36
208 http://212.33.195.37
209 http://212.33.195.37
210 http://212.38.130.122
211 http://217.23.37.135
212 http://amon.nic.gov.jo
213 http://193.188.65.188
214 http://212.118.10.179
215 http://212.118.3.117
216 http://www.jordangps.com
217 http://213.139.63.199
218 http://194.165.142.122
219 http://212.35.70.58
220 http://212.38.149.147
221 http://193.188.80.66
222 http://80.90.164.160

Πίνακας Π.γ.3: Το Ιορδανικό (.jo) web (Φεβ. 2003)

Παράρτημα Δ: CD ROM

Στο CD ROM που συνοδεύει τη διατριβή υπάρχουν οι πιο πολλές πηγές αναφοράς (π.χ. papers, άρθρα) καθώς και διάφορα πειραματικά αποτελέσματα (δείγματα, IP χάρτες κ.λπ.).

Παράρτημα Ε: Δείκτες (index)