ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Τεχνικές εξόρυξης και ανάλυσης δεδομένων: από τα κοινωνικά δίκτυα στο σκοτεινό ιστό

Διπλωματική Εργασία
του
Σωτήρη Α. Νικολέτου

Επιβλέπουσα:
Παρασκευή Ραυτοπούλου
ΕΔΙΠ

Τρίπολη, Μάρτιος 2022

Περίληψη

Με την αυξανόμενη παρακολούθηση και την ενίσχυση των κανονισμών-νόμων από τις πλατφόρμες, οι κοινότητες με εγκληματικά συμφέροντα μετακινούνται στον σκοτεινό ιστό, ο οποίος φιλοξενεί περιεχόμενο που ποικίλλει από καταγγελίες (whistle-blowing) και προστασία της ιδιωτικής ζωής, έως ναρκωτικά, τρομοκρατία και πειρατεία. Χρησιμοποιώντας δεδομένα μετά συζήτησης από πέντε φόρουμ σκοτεινού ιστού, κατασκευάζουμε πέντε γραφήματα αλληλεπίδρασης και χρησιμοποιούμε εργαλεία ανάλυσης κοινωνικών δικτύων για να μελετήσουμε αυτές τις υπόγειες κοινότητες. Παρατηρούμε τη δομή κάθε δικτύου για να επισημάνουμε δομικά μοτίβα και να εντοπίσουμε σημαντικούς κόμβους μέσω της ανάλυσης κεντρικότητας δικτύου. Τα ευρήματά μας υποδηλώνουν ότι στην πλειονότητα των φόρουμ ορισμένα μέλη είναι πολύ συνδεδεμένα και σχηματίζουν κεντρικούς κόμβους ιδιαίτερης σημασίας, ενώ τα περισσότερα μέλη έχουν μικρότερο αριθμό συνδέσεων. Κατά την εξέταση των δραστηριοτήτων ανάρτησης των κεντρικών κόμβων, διαπιστώσαμε ότι οι περισσότεροι από αυτούς δημοσιεύουν σε υποφόρουμ με ευρύτερα θέματα, όπως γενικές συζητήσεις και σεμινάρια/εγχειρίδια χρήσης (tutorials).

Αυτά τα μέλη παίζουν διαφορετικούς ρόλους στα διάφορα φόρουμ όπου εμφανίζονται, ενώ για κάθε φόρουμ διαπιστώσαμε οτι συμμετέχουν διαφορετικά προφίλ χρηστών.

Λέξεις Κλειδιά μεγάλα δεδομένα, θεωρία γράφων, ανάλυση κοινωνικών δικτύων, ψηφιακές κοινότητες, σκοτεινός ιστός

Abstract

With increasing monitoring and regulation by platforms, communities with criminal interests are moving to the dark web, which hosts content ranging from whistle-blowing and privacy, to drugs, terrorism, and hacking. Using post discussion data we construct six interaction graphs from six dark web forums and use social network analysis tools to study these underground communities. We observe the structure of each network to highlight structural patterns and identify nodes of importance through network centrality analysis. Our findings suggest that in the majority of the forums some members are highly connected and form hubs, while most members have a lower number of connections. When examining the posting activities of central nodes we found that most of these nodes post in sub-forums with broader topics, such as general discussions and tutorials.

These members play different roles in the different forums, and within each forum we identified diverse user profiles.

Keywords big data, graph theory, social network analysis, virtual communities, dark web

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω θερμά την κ. Παρασκευή Ραυτοπούλου που μου έδωσε τη δυνατότητα να πραγματοποιήσω την παρούσα διπλωματική εργασία υπό την επίβλεψη και καθοδήγησή της.

Θα ήθελα ακόμα να ευχαριστήσω την οικογένεια μου και τους φίλους μου για την υποστήριξη και τη συμπαράσταση τους αυτά τα χρόνια φοίτησης μου στο Πανεπίστημιο Πελοποννήσου.

Κατάλογος Σχημάτων

2.1 Σχέσεις μεταξύ ιστοσελίδων 10
2.2 Παράδειγμα κοινωνικού δικτύου (από πραγματικά δεδομένα) 11
2.3 Σύνδεση κόμβων 11
2.4 Διεπαφή χρήστη του Gephi 16
2.5 Παράδειγμα ενός query 16
2.6 Γράφος του igraph 17
2.7 Γράφος του networkx 17
2.8 Περιγραφή του πρωτόκολλου robots.txt με δύο παραδείγματα 20
2.9 Ένα από τα captcha του δημοφιλέστερου φόρουμ στο dark web 22
2.10 Σύνθεση user-agent και session cookies από .onion ιστότοπο 23

3.1 Διεπαφή χρήστη της εικονικής μηχανής του Tails OS 29
3.2 Διεπαφή χρήστη της μηχανής αναζήτησης ahmia 29
3.3 Τα αποτελέσματα της ahmia όταν ψάχνουμε τον όρο "forum" 30
3.4 Διεπαφή χρήστη του hidden wiki 31
3.5 Διεπαφή χρήστη του dread 32
3.6 Μερικά από τα υποφόρουμ του Dread 33
3.7 Το φόρουμ Hidden Answers 34
3.8 Το φόρουμ CryptBB 35
3.9 Το φόρουμ Tape 36
3.10 Το φόρουμ Helium 37
3.11 Μορφή εγγράφου MongoDB 38
3.12 Οι συλλογές μας στο MongoDB 39
3.13 Παράδειγμα σε json από υποφόρουμ του dread στο MongoDB 39

4.1 Cookies του φόρουμ 41
4.2 Παράδειγμα ενός post 42
4.3 Πηγαίος κώδικας και επιθεώρηση στοιχείου στην HTML 44
4.4 Παράδειγμα διεύθυνσης μιας δημοσίευσης 45
4.5 Παράδειγμα ενός αντικειμένου json χωρίς σχολιαστές 46

5.1 Έναρξη του Neo4j 49
5.2 Δημιουργία σχέσεων με cypher 49
5.3 Δημιουργία εικονικού γράφου με cypher 50
5.4 Αριθμός υποφόρουμ ανα φόρουμ 53
5.5 Μέγιστες τιμές κεντρικότητας με βάση το βαθμό για το Dread 53
5.6 Μέσες τιμές κεντρικότητας με βάση την εγγύτητα για το Dread 54
5.7 Μέσες τιμές κεντρικότητας με βάση την διαμεσολαβικότητα για το Dread 54
5.8 Μέγιστες τιμές κεντρικότητας με βάση το ιδιοδιάνυσμα για το Dread . . 55

5.9 Αριθμός κοινοτήτων υποφόρουμ του Dread 55

Κατάλογος Πινάκων

5.1 Αριθμός κόμβων, σχέσεων και πυκνότητα 52
5.2 Μέσες τιμές κεντρικότητας των υποφόρουμ του CryptBB 56
5.3 Μέγιστες τιμές κεντρικότητας των υποφόρουμ του CryptBB 56
5.4 Μέγιστες τιμές κεντρικότητας των υποφόρουμ του Helium 57
5.5 Μέσες τιμές κεντρικότητας των υποφόρουμ του Helium 57
5.6 Μέγιστες τιμές κεντρικότητας των υποφόρουμ του Hidden Answers . . . 58
5.7 Μέσες τιμές κεντρικότητας των υποφόρουμ του Hidden Answers 58
5.8 Μέγιστες τιμές κεντρικότητας των υποφόρουμ του Tape 59
5.9 Μέσες τιμές κεντρικότητας των υποφόρουμ του Tape 59

Περιεχόμενα

1 Εισαγωγή **7**
1.1 Προσέγγιση του θέματος 8
1.2 Δομή της πτυχιακής εργασίας 9

2 Θεωρητικό υπόβαθρο **10**
2.1 Ανάλυση κοινωνικών δικτύων 10
2.1.1 Με τι μοιάζει ένας γράφος SNA 11
2.1.2 Κόμβοι και ακμές 11
2.1.3 Μετρήσεις αναφορικά με τους γράφους 12
2.1.4 Κεντρικότητα με βάση το ιδιοδιάνυσμα (eigenvector) 14
2.1.5 Ρόλοι μέσα στο δίκτυο 15
2.1.6 Εργαλεία για οπτικοποίηση δεδομένων SNA 15
2.2 Web crawling, web scraping 18
2.2.1 Web crawlers 18
2.2.2 Ευρετηρίαση 18
2.2.3 Πώς λειτουργεί το crawling 19
2.2.4 Διαφορές web crawling, web scraping 20
2.2.5 Εργαλεία για Web Scraping 21
2.2.6 Εμπόδια κατά την διαδικασία του scraping 22
2.3 TOR Browser και Dark Web 24

3 Ανάλυση εργαλείων και φόρουμ **28**
3.1 Tails OS 28
3.2 Ahmia 29
3.3 Hidden Wiki 30
3.4 Τα 5 τελικά φόρουμ 31
3.4.1 Dread 31
3.4.2 Hidden Answers 33
3.4.3 CryptBB 34
3.4.4 Tape 35
3.4.5 Helium 36
3.4.6 Γενικά εμπόδια και δυσκολίες 37
3.4.7 MongoDB 38

4 Ανάλυση κώδικα συγκομιδής δεδομένων **40**
4.1 Δημιουργία σύνδεσης 40
4.2 Πρώτα βήματα crawling 42
4.3 Τελικά βήματα 46

5 Μετρικές κοινωνικών δικτύων και αποτελέσματα **48**
5.1 Το πρόγραμμα Neo4j 48
5.2 Μετρικές κοινωνικών δικτύων 50
5.2.1 Κεντρικότητα με βάση το βαθμό 50
5.2.2 Κεντρικότητα με βάση την εγγύτητα 50
5.2.3 Κεντρικότητα με βάση την διαμεσολαβικότητα 51
5.2.4 Κεντρικότητα με βάση το ιδιοδιάνυσμα 51
5.2.5 Αριθμός κοινοτήτων γράφων 51
5.3 Αποτελέσματα εντολών και οπτικοποιήσεις 52
5.4 Συμπεράσματα 59

6 Επίλογος και μελλοντική έρευνα **61**

Α΄ Ακρωνύμια και συντομογραφίες **63**

Βιβλιογραφία **64**

Κεφάλαιο 1

Εισαγωγή

'Το διαδίκτυο είναι συχνά ένα παράξενο και μυστήριο μέρος. Κάτω από την επιφάνειά του βρίσκεται ένα τεράστιο κουβάρι αόρατων ιστοσελίδων που είναι πολύ μεγαλύτερο από αυτό που ξέρουμε ή μπορούμε να φανταστούμε. Αυτός είναι ο Βαθύς Ιστός (Deep Web) και αποτελεί το μεγαλύτερο μέρος του Διαδικτύου.'[1]

Όταν κάνουμε μια αναζήτηση στο διαδίκτυο έχουμε πρόσβαση στον Ορατό Ιστό (Surface Web) και περιλαμβάνει όλες τις ιστοσελίδες που μπορούν να βρουν οι κλασικές μηχανές αναζήτησης όπως είναι η Google, το Bing ή το Yahoo και αποτελεί το 4% του συνολικού διαδικτύου. Το 2001, ο Michael Bergman, ιδρυτής της εταιρίας Bright Planet, μια εξειδικευμένης εταιρείας αναζήτησης ευρετηρίασης επινόησε τον όρο Deep Web (Βαθύς Ιστός) αναφερόμενος στους ιστότοπους που δεν μπορούν να βρουν οι μηχανές αναζήτησης και που αποτελεί το 96% του Διαδικτύου.[2] Κοινωνικά δίκτυα σε διαδικτυακά υπόγεια (underground) φόρουμ έχουν λάβει ερευνητική προσοχή με ενδιαφέροντα που κυμαίνονται από αγορές, ψηφιακές underground οικονομίες έως την ανάλυση βασικών παραγόντων (actors). Στον πυρήνα των underground φόρουμ είναι μέλη που αλληλεπιδρούν το ένα με το άλλο. Η ανάλυση κοινωνικών δικτύων (Social Network Analysis) παρέχει πολύτιμες πληροφορίες για τον τρόπο λειτουργίας αυτών των κοινοτήτων. Οι δομές, λειτουργίες και αλληλεπιδράσεις αυτών των κοινοτήτων σκοτεινού Ιστού δεν έχουν ακόμη ανακαλυφθεί σε μεγάλο βαθμό λόγω δυσκολιών που συνδέονται με τη συλλογή δεδομένων. Επομένως, κατευθύνουμε την προσοχή μας σε ένα συγκεκριμένο υποσύνολο αυτών των κοινοτήτων, που λειτουργούν αποκλειστικά στο σκοτεινό ιστό.

Ο σκοτεινός ιστός (dark web) είναι ένα τμήμα του Διαδικτύου στο οποίο είναι δυνατή μόνο η πρόσβαση μέσω εργαλείων και δικτύων ανωνυμίας, όπως το Tor [3], και δεν ευρετηριάζεται από τις τυπικές μηχανές αναζήτησης. Ο σκοτεινός ιστός έχει γίνει γνωστός κυρίως επειδή επιτρέπει την ανωνυμία και κάθε είδους παράνομης δραστηριότητας. Κρυφές υπηρεσίες (hidden services) φιλοξενούν μια ευρεία γκάμα περιεχομένου, συμπεριλαμβανομένων των καταγγελιών, της ιδιωτικότητας, ναρκωτικά, τρομοκρατία και πειρατεία. Προηγούμενη έρευνα για τα φόρουμ του σκοτεινού Ιστού έχουν επικεντρωθεί κυρίως σε τρομοκρατικό περιεχόμενο και εξτρεμιστικές ομάδες. Ωστόσο, παρουσιάζει

[1] https://www.ertnews.gr/eidiseis/epistimi/vathys-ke-skotinos-istos-astikos-mythos-pragmatikotita/

[2] https://www.historyofinformation.com/detail.php?entryid=3599

[3] https://www.torproject.org/

ένα πλούσιο οικοσύστημα των κοινοτήτων και οι ερευνητικές προσπάθειες θα έπρεπε να περιλαμβάνουν και άλλες κοινότητες σκοτεινού Ιστού. Για παράδειγμα, υπάρχει μεγάλη ανησυχία για τις ψηφιακές αγορές (κρυπταγορές) ναρκωτικών, πυροβόλων όπλων και το έγκλημα στον κυβερνοχώρο (cybercrime) καθώς αρχίζουν να 'μεταναστεύουν' προς τον σκοτεινό ιστό.

Στην παρούσα ερευνητική εργασία, πραγματοποιούμε ανάλυση και σύγκριση 5 σκοτεινών διαδικτυακών φόρουμ διαφόρων γλωσσών που όμως ύστερα μεταφράστηκαν στα Αγγλικά, που σχετίζονται με το έγκλημα στον κυβερνοχώρο, την χρήση και διακίνηση όλων των ειδών ναρκωτικών ουσιών, προσωπικές εμπειρίες, πολιτικά κ.α. για να αντλήσουμε πληροφορίες για τις οργανωτικές τους δομές που αποτελούν τη βάση των ροών πληροφοριών και πόρων εντός αυτών των κοινοτήτων. Είναι σημαντικό να σημειωθεί ότι αυτά τα δίκτυα συνεχώς εξελίσσονται, ωστόσο αυτή η μελέτη εστιάζει σε μια στατική πλευρά των αλληλεπιδράσεων, που αφορούν δηλαδή στη χρονική στιγμή που αντλήθηκαν τα δεδομένα. Πιο συγκεκριμένα, στην παρούσα εργασία χρησιμοποιήσαμε ένα στιγμιότυπο αναρτήσεων του τελευταίου 5μηνου-6μηνου. Κατά τη διεξαγωγή αυτής της μελέτης, τα δεδομένα μας συμβάλλουν στη βαθύτερη κατανόηση αυτών των κοινοτήτων, κάτι που μπορεί να θέσει τα θεμέλια για μελλοντική έρευνα.

1.1 Προσέγγιση του θέματος

Όπως προαναφέρθηκε παραπάνω, η παρούσα πτυχιακή έχει στόχο μέσω ανάλυσης δεδομένων, να αποφανθεί ποιοι είναι οι βασικοί actors των φόρουμ του dark web, πως συνδέονται κοινωνικά με τα υπόλοιπα άτομα κατά βάση ανώνυμα και εάν μέσα σε αυτά τα φόρουμ που συζητούν κυριολεκτικά για τα πάντα χωρίς ενδιασμούς μπορούμε να διακρίνουμε ψηφιακές κοινότητες και τέλος, αν παρατηρείται το φαινόμενο του μικρόκοσμου (small world effect) σε αυτό το τεράστιο δίκτυο. Για να απαντήσουμε σε αυτά τα ερωτήματα καταλήγουμε να χρησιμοποιούμε Ανάλυση Κοινωνικών Δικτύων (Social Network Analysis) που χρησιμοποιεί τη Θεωρία Γράφων και Πολύπλοκων Δικτύων για να βρει τις κοινωνικές σχέσεις, συνδέοντας κόμβους με ακμές. Στην εργασία μας, θεωρούμε κοινωνικές σχέσεις μεταξύ χρηστών εάν στο post ενός χρήστη κάποιος άλλος χρήστης έχει σχολιάσει.

Συγκεκριμένα, για κάθε ένα από τα 5 φόρουμ, και για κάθε ένα υπο-φόρουμ αυτού, τα δεδομένα που ανακτήθηκαν είναι σε μορφή διατεταγμένη ως:

- Όνομα χρήστη συγγραφέα
- Τίτλος του post
- Σώμα του post
- Όνομα χρήστη σχολιαστή
- Σώμα του σχολίου

Αρχικά, για την συγκομιδή των δεδομένων, δημιουργήσαμε ένα πρόγραμμα ανίχνευσης ιστού (web crawler), λεπτομέρειες της υλοποίησης του θα αναλυθούν στο επόμενο κεφάλαιο, που δίνοντας του το url του εκάστοτε φόρουμ και παραμετροποιώντας το κατάλληλα, επιστρέφει τα δεδομένα στη μορφή που αναφέραμε παραπάνω. Τα δεδομένα αυτά αποθηκεύτηκαν σε μια NOSQL βάση δεδομένων και ύστερα χρησιμοποιήθηκε

κατάλληλο πρόγραμμα οπτικοποίησης δεδομένων όπου με χρήση κατάλληλων ερωτημάτων (queries) κατασκευάστηκαν οι τελικοί γράφοι για το κάθε φόρουμ.

Με τη χρήση αυτών των γράφων, γίνεται μια χαρτογράφηση του πυρήνα των φόρουμ του dark-deep web και καταλαβαίνουμε πόσο δυνατές κοινωνικές σχέσεις έχουν δημιουργηθεί σε αυτά τα φόρουμ και τα μοτίβα που ακολουθούνται.

1.2 Δομή της πτυχιακής εργασίας

Στο Κεφάλαιο 2, παρουσιάζεται το θεωρητικό υπόβαθρο καθώς και μια μικρή εισαγωγή για τις περιοχές της Ανάλυσης Κοινωνικών Δικτύων και του Σκοτεινού Ιστού. Στο Κεφάλαιο 3, δείχνουμε με λεπτομέρεια όλα τα εργαλεία και τις βιβλιοθήκες που χρησιμοποιήθηκαν για τη συγκεκριμένη εργασία από το στήσιμο του Virtual Machine μέχρι την οπτικοποίηση των δεδομένων (datasets). Στο Κεφάλαιο 4 αναλύεται ο κώδικας της εργασίας που είναι εξ ολοκλήρου σε Python. Στο Κεφάλαιο 5, παρουσιάζεται αναλυτικά το πρόγραμμα που χρησιμοποίησαμε για την δημιουργία υπεργράφων καθώς και όλες οι οπτικοποιήσεις που δημιουργήσαμε και τα συμπεράσματα που βγάζουμε από αυτές. Τέλος, στο Κεφάλαιο 6 συγκεντρώνουμε όλα τα συμπεράσματα-πορίσματα μας για τις κοινωνικές σχέσεις των χρηστών και ορίζουμε μελλοντικές προοπτικές.

Κεφάλαιο 2

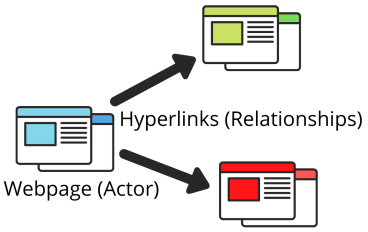
Θεωρητικό υπόβαθρο

Στο παρόν κεφάλαιο δίνουμε τους ορισμούς και εξηγούμε την μεθοδολογία που χρησιμοποιείται για την ανάλυση κοινωνικών δικτύων καθώς και παρουσιάζουμε όλα τα εργαλεία που απαιτούνταν για την επιτυχή εκπόνηση της εργασίας.

2.1 Ανάλυση κοινωνικών δικτύων

Τα δίκτυα είναι παντού γύρω μας, παραδείγματα είναι τα οδικά δίκτυα, τα δίκτυα Διαδικτύου και τα διαδικτυακά κοινωνικά δίκτυα όπως το Facebook.

Η ανάλυση κοινωνικών δικτύων (SNA), επίσης γνωστή ως επιστήμη δικτύων, είναι ένα πεδίο ανάλυσης δεδομένων που χρησιμοποιεί δίκτυα και θεωρία γραφημάτων για την κατανόηση των κοινωνικών δομών. Οι τεχνικές SNA μπορούν επίσης να εφαρμοστούν σε δίκτυα εκτός των κοινωνικών. Προκειμένου να δημιουργηθούν γραφήματα SNA, απαιτούνται δύο βασικά στοιχεία: ηθοποιοί (actors) και σχέσεις (relationships). Μια κοινή εφαρμογή των τεχνικών SNA είναι με το ίδιο το διαδίκτυο. Οι ιστοσελίδες στο Διαδίκτυο συχνά συνδέονται με άλλες ιστοσελίδες (βλ. Σχήμα 2.1) — είτε στον δικό τους ιστότοπο είτε σε άλλο ιστότοπο. Αυτοί οι σύνδεσμοι μπορούν να θεωρηθούν σχέσεις μεταξύ των ηθοποιών, που είναι οι ίδιες οι ιστοσελίδες σε αυτή τη περίπτωση. Αυτό που μόλις περιγράψαμε είναι στην πραγματικότητα ένα βασικό συστατικό της αρχιτεκτονικής των μηχανών αναζήτησης. [1], [2], [3]



Σχήμα 2.1: Σχέσεις μεταξύ ιστοσελίδων

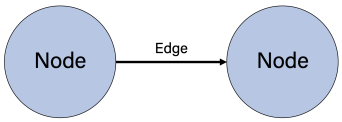
2.1.1 Με τι μοιάζει ένας γράφος SNA

Ένα γράφημα κοινωνικού δικτύου περιέχει σημεία και γραμμές που συνδέουν αυτές τις κουκκίδες — παρόμοιο με ένα παζλ Connect-the-dot. Τα σημεία αντιπροσωπεύουν τους ηθοποιούς και οι γραμμές τις σχέσεις. Ένα παράδειγμα γραφήματος κοινωνικού δικτύου μπορείτε να δείτε στο Σχήμα 2.2.

Σχήμα 2.2: Παράδειγμα κοινωνικού δικτύου (από πραγματικά δεδομένα)

2.1.2 Κόμβοι και ακμές

Μέχρι τώρα αναφερόμασταν σε ηθοποιούς και σε σχέσεις. Στην επιστήμη των γράφων, οι ηθοποιοί λέγονται κόμβοι (οι τελείες στο Σχήμα 2.2) και οι σχέσεις λέγονται ακμές (οι γραμμές στο Σχήμα 2.2). Στο Σχήμα 2.3 φαίνεται ένα μικρό δίκτυο με δύο κόμβους που συνδέονται με μια κατεθυνόμενη ακμή.



Σχήμα 2.3: Σύνδεση κόμβων

Οι κόμβοι μπορούν να αντιπροσωπεύουν μια ποικιλία από «actors». Για παράδειγμα, στα δίκτυα Διαδικτύου οι κόμβοι μπορούν να αντιπροσωπεύουν ιστοσελίδες, στα κοινωνικά δίκτυα οι κόμβοι μπορούν να αντιπροσωπεύουν ανθρώπους, στα δίκτυα εφοδιαστικής αλυσίδας μπορούν να αντιπροσωπεύουν οργανισμούς, στα δίκτυα εξωτερικών

σχέσεων μπορούν να αντιπροσωπεύουν χώρες. Οι ακμές αντίστοιχα μπορούν να αντιπροσωπεύουν μια ποικιλία συσχετίσεων μεταξύ των κόμβων. Στα δίκτυα Διαδικτύου, οι άκρες μπορούν να αντιπροσωπεύουν υπερσυνδέσμους, στα κοινωνικά δίκτυα οι ακμές μπορούν να αντιπροσωπεύουν συνδέσεις, στα δίκτυα εφοδιαστικής αλυσίδας οι ακμές μπορούν να αντιπροσωπεύουν τη μεταφορά αγαθών, στα δίκτυα εξωτερικών σχέσεων, οι ακμές μπορούν να αντιπροσωπεύουν πολιτικές.

Σημειώστε οτι σε κάποιες περιπτώσεις, οι σχέσεις μεταξύ των κόμβων ενδέχεται να είναι αμφίδρομες, ενώ άλλοτε κατευθυνόμενες. Για παράδειγμα, σε γράφο που μοντελοποιεί οδικό δίκτυο, όπου οι κόμβοι είναι οι πόλεις και οι ακμές είναι οι δρόμοι μεταξύ τους, οι σχέσεις θεωρούνται αμφίδρομες (μη-κατευθυνόμενες). Από την άλλη, στο Instagram οι κόμβοι είναι οι χρήστες και οι ακμές μεταξύ τους είναι η σχέση που πιθανώς έχουν, όπως φιλία, σχόλιο, like.Στην περίπτωση αυτή μιλάμε για κατευθυνόμενες σχέσεις. [1], [2], [3]

2.1.3 Μετρήσεις αναφορικά με τους γράφους

Προκειμένου να μελετήσουμε τους γράφους, χρησιμοποιούμε διάφορες μετρικές, τις οποίες παρουσιάζουμε παρακάτω.

Συνοχή

Ο βαθμός κατά τον οποίο τα μέλη του δικτύου συνδέονται μεταξύ τους με συνεκτικούς δεσμούς. Οι ομάδες χαρακτηρίζονται κλίκες όταν κάθε μέλος τους συνδέεται κατευθείαν με όλα τα άλλα μέλη της ομάδας, διαφορετικά ονομάζονται κοινωνικοί κύκλοι. [4]

Δομική συνοχή

Ο ελάχιστος αριθμός των μελών που εάν μετακινηθούν από το δίκτυο θα διακόψουν τη σύνδεση του δικτύου.

Πυκνότητα

Ο λόγος του αριθμού των υπαρχόντων δεσμών σε ένα δίκτυο προς τον συνολικό αριθμό των δεσμών που θα μπορούσαν να υπάρχουν. [4] Ο τύπος που την υπολογίζει είναι:

$$\frac{2T}{N(N-1)}$$

όπου N ο αριθμός κόμβων του δικτύου και T οι υπάρχουσες συνδέσεις (ties).[1]

Μήκος μονοπατιού

Η απόσταση μεταξύ ενός ζεύγους κόμβων στο δίκτυο. Μέσο μήκος μονοπατιού είναι ο μέσος όρος των αποστάσεων μεταξύ όλων των ζευγαριών του δικτύου [4] Ο τύπος που το υπολογίζει είναι:

$$\frac{\sum d(v_i, v_j)}{N(N-1)}$$

[1] https://www.the-vital-edge.com/what-is-network-density/

όπου N ο αριθμός των κόμβων στο δίκτυο και d η απόσταση μεταξύ 2 κόμβων i και j.[2]

Κεντρικότητα με βάση το βαθμό

Υπάρχουν οι εξής δύο προσεγγίσεις:

- προσέγγιση κατά Freeman: οι κόμβοι στους οποίους εισέρχονται πολλοί δεσμοί είναι οι προεξέχοντες ή αυτοί που έχουν μεγαλύτερο κύρος.[3]
- προσέγγιση κατά Bonacich: η κεντρικότητα ενός κόμβου είναι συνάρτηση των άμεσων σχέσεων που διατηρεί, αλλά και των συνδέσμων που τα άτομα στη γειτονιά του με τη σειρά τους έχουν. Το να είσαι συνδεδεμένος με καλά συνδεδεμένους άλλους ναι μεν σε κάνει κεντρικό, αλλά όχι και ισχυρό. [5], [4]

Ψπολογίζεται με τον τύπο:

$$\sum a(p_i, p_j)$$

όπου εάν θεωρήσουμε ότι τα δεδομένα μας μπορούν να αναπαριστούν στοιχεία δυαδικού πίνακα γειτνίασης, τότε

$$ap_{ij} = ap_{ji} = 1$$

εάν τα 2 στοιχεία αυτά σχετίζονται, αλλιώς

$$ap_{ij} = ap_{ji} = 0$$

Κεντρικότητα με βάση την εγγύτητα

Δίνει έμφαση στις αποστάσεις κάθε κόμβου από όλους τους υπόλοιπους. Αν κάποιος κόμβος συνδέεται άμεσα με πολλούς άλλους, οι οποίοι όμως είναι σχετικά αποκομμένοι από το σύνολο, ο κόμβος αυτός ναι μεν θα είναι αρκετά κεντρικός αλλά μόνο σε μια τοπική γειτονιά / τμήμα του δικτύου [5]

Υπολογίζεται με τον τύπο:

$$\frac{N-1}{\sum d(p_i, p_j)}$$

όπου

$$d(p_i, p_j)$$

η απόσταση μεταξύ των κορυφών i,j για ένα γράφο με N κόμβους.

Κεντρικότητα με βάση τη διαμεσολαβικότητα

Εξετάζει εάν ένας κόμβος εξαρτάται από άλλους ενδιαμέσους κόμβους στην προσπάθεια του να προσεγγίσει έναν άλλο κόμβο. Καθένας από αυτούς μπορεί (ηθελημένα ή μη) να καθυστερήσει ή και να αποτρέψει εντελώς τα αιτήματα για επικοινωνία. Όσο λιγότεροι ενδιάμεσοι τόσο πιο κεντρική θέση. [5] Στην ανάλυση των κοινωνικών δικτύων, μπορεί να έχει διαφορετικές επιπτώσεις. Από μακροσκοπική σκοπιά, οι θέσεις γεφύρωσης ή οι «δομικές τρύπες» (που υποδεικνύονται με υψηλή ενδιάμεση κεντρικότητα) αντανακλούν δύναμη, επειδή επιτρέπουν στο άτομο στη θέση γεφύρωσης να ασκεί έλεγχο

[2] http://qasimpasta.info/data/uploads/sina-2015/calculating-average-path-length.pdf
[3] https://faculty.ucr.edu/ hanneman/nettext/C10$_{C}$*entrality.html*

(π.χ., να αποφασίζει αν θα μοιραστεί πληροφορίες ή όχι) στα άτομα που συνδέονται μαζί του. Ωστόσο, από τη μικροσκοπική προοπτική των εγωκεντρικών δικτύων (ego networks) (δηλαδή, λαμβάνοντας υπόψη μόνο τις συνδέσεις πρώτου βαθμού) [6], στα διαδικτυακά κοινωνικά δίκτυα μια υψηλή τιμή κεντρικότητας συμπίπτει με τις τιμές των στενότερων φίλων (δηλαδή, ισχυρούς διαπροσωπικούς δεσμούς), επειδή αντανακλά τις επενδύσεις κοινωνικού κεφαλαίου στη σχέση όταν γεφυρώνονται μακρινοί κοινωνικοί κύκλοι (π.χ., οικογένεια και πανεπιστήμιο) (που συχνά προκύπτουν από μια εισαγωγή από το εγώ).[4]

Υπολογίζεται με τον τύπο:

$$\frac{\sum s_{st}(u)}{\sum s_{st}}$$

όπου

$$s_{st}(u)$$

η μέση απόσταση μονοπατιού από το s στο t για κάποιον κόμβο u.

2.1.4 Κεντρικότητα με βάση το ιδιοδιάνυσμα (eigenvector)

Η κεντρικότητα με βάση το ιδιοδιάνυσμα [5], είναι ένας τρόπος να υπολογίσουμε την επιρροή των κόμβων σε ένα γράφο ή δίκτυο. Εκχωρούνται βαθμοί σε όλους τους κόμβους με βάση την ιδεά ότι οι συνδέσεις με κόμβους υψηλής βαθμολογίας, συνεισφέρουν περισσότερο στη βαθμολογία του εν λόγω κόμβου, παρά ίσες συνδέσεις με κόμβους χαμηλότερης βαθμολογίας. Ο αλγόριθμος της Google, Page Rank [6] έχει βασιστεί σε αυτή την κεντρικότητα. Υπολογίζεται με τον τύπο:

$$X_v = \frac{\sum a(v,t) * X_t}{\lambda}$$

όπου λ σταθερά, και

$$X_t, X_v$$

η κεντρικότητα της κορυφής v και t.

Ισχύς με βάση το βαθμό

Σε κόμβους με μεγάλο βαθμό παρουσιάζονται περισσότερες ευκαιρίες γιατί έχουν περισσότερες επιλογές. Εξαρτώνται επομένως λιγότερο από κάποιον άλλο κόμβο και είναι περισσότερο ισχυροί. [1]

Ισχύς με βάση την εγγύτητα

Η ισχύς έχει να κάνει και με το εάν κάποιος βρίσκεται σε θέση που να του επιτρέπει οι απόψεις του να ακούγονται από πολλούς άλλους ή πολλοί άλλοι λόγω του ότί είναι κοντά του να τον χρησιμοποιούν ως σημείο αναφοράς ώστε να επιβεβαιώνουν και να διασώζουν τις ιδεές τους. [1]

[4] https://www.sciencedirect.com/science/article/abs/pii/S1094996820301390?via%3Dihub
[5] https://www.sci.unich.it/ francesc/teaching/network/eigenvector.html
[6] https://www.semrush.com/blog/pagerank/

Ισχύς με βάση την διαμεσολαβικότητα

Ένας κόμβος είναι ισχυρότερος εάν βρίσκεται στο ενδιάμεσο περισσότερων ζευγών κόμβων.Αυτό δίνει τη δυνατότητα στο κόμβο να παίξει το ρόλο μεσάζοντα με ό,τι αυτό συνεπάγεται (μίσθωση των υπηρεσιών του, προώθηση των απαφών που επιθυμεί και τα λοιπά). [1]

2.1.5 Ρόλοι μέσα στο δίκτυο

Οι κόμβοι ενός γράφου μπορεί να έχουν διαφορετικούς ρόλους. Παρουσιάζουμε μερικούς παρακάτω.

Κεντρικός ρόλος

Πρόκειται για κόμβους με μεγάλο αριθμό συνδέσμων (σχέσεων). Όταν ο γράφος αφορά σε κοινωνικό δίκτυο, πριν αποφασίσουμε εάν και κατά πόσο θα επέμβουμε στην δράση ενός κεντρικού κόμβου-ατόμου, είναι σημαντικό να καταλάβουμε τον τρόπο με τον οποίο αυτό επηρεάζει την συνολική απόδοση της ομάδας. [1]

Αφανείς Ήρωες είναι άτομα που, προς έκπληξη των ανωτέρων τους, τείνουν να παίζουν πολύ κεντρικότερο ρόλο από αυτόν που είχε αρχικά προβλεφθεί. [1]

Στένωση προκαλούν τα άτομα που με διάφορους τρόπους παρακωλύουν την εργασία ολόκληρης της ομάδας. [1]

Ρόλος μεσάζοντα

Πρόκειται για κόμβους που βοηθούν να κρατηθεί ο γράφος συνεκτικός. Σε ένα κοινωνικό δίκτυο, πρόκειται για άτομα που κρατούν την ομάδα ενωμένη. [1]

Περιφερειακός ρόλος

Αφορά κόμβους που είναι χαλαρά συνδεδεμένοι με τους υπόλοιπους ή ακόμα και απομονωμένοι. Στη περίπτωση των κοινωνικών δικτύων, περιφεριακό ρόλο έχει ένα άτομο που είναι απομονωμένο από μια συγκεκριμένη ομάδα. Είναι βέβαια σημαντικό να διαπιστωθεί πώς το άτομο αυτό επηρεάζει το σύνολο. [1]

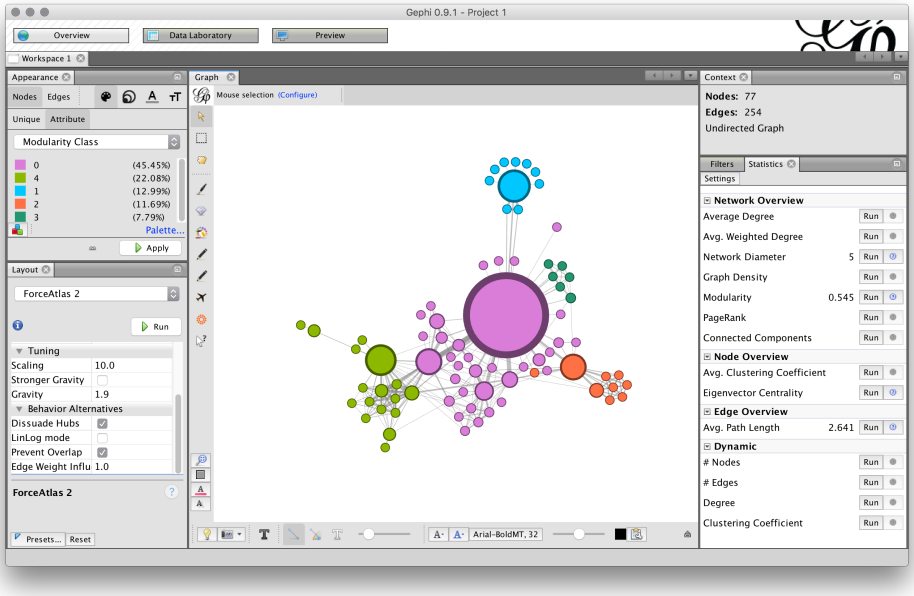
2.1.6 Εργαλεία για οπτικοποίηση δεδομένων SNA

Σε αυτό το κεφάλαιο, παρουσιάζουμε μερικά γνωστά και ευρέως διαδεδομένα εργαλεία, τα οποία χρησιμοποιούνται για οπτικοποίηση δεδομένων που προέρχονται από κοινωνικά δίκτυα.

Gephi

Το Gephi είναι ένα λογισμικό ανοιχτού κώδικα για ανάλυση γραφημάτων και δικτύου. Χρησιμοποιεί μια μηχανή απόδοσης (render engine) 3D για να εμφανίζει μεγάλα δίκτυα σε πραγματικό χρόνο και να επιταχύνει την εξερεύνηση. Μια ευέλικτη και πολλαπλών εργασιών αρχιτεκτονική φέρνει νέες δυνατότητες εργασίας με πολύπλοκα σύνολα δεδομένων και παράγει πολύτιμα οπτικά αποτελέσματα. Παρέχει εύκολη

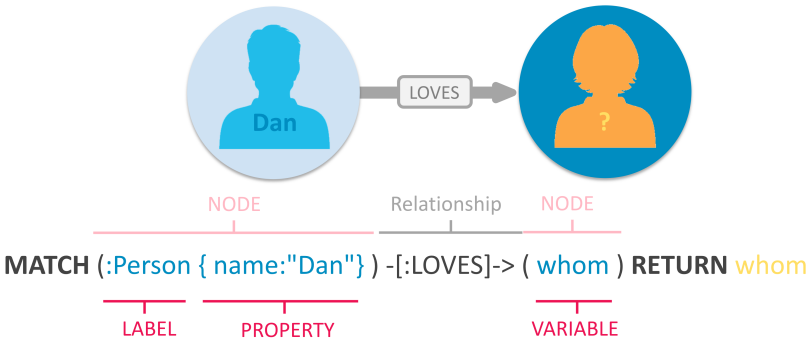
και ευρεία πρόσβαση σε δεδομένα δικτύου και επιτρέπει χωροταξία (spatializing), φιλτράρισμα (filtering), πλοήγηση (navigating), χειρισμό (manipulating) και ομαδοποίηση (clustering). [7]. Στο Σχήμα 2.4 βλέπουμε τη διεπαφή χρήστη του Gephi.



Σχήμα 2.4: Διεπαφή χρήστη του Gephi

Neo4J

Το Neo4j είναι το μεγαλύτερο σύστημα διαχείρησης δεδομένων γραφημάτων που επιτρέπει την αποθήκευση μεγάλων δεδομένων και την επεξεργασία υπεργράφων. Χρησιμοποιεί τη γλώσσα προγραμματισμού Java αλλά έχει ενσωματομένη μια δική της γλώσσα (query language), την Cypher. [8]. Στο Σχήμα 2.5 φαίνεται η σύνταξη ενός ερωτήματος στη γλώσσα Cypher.

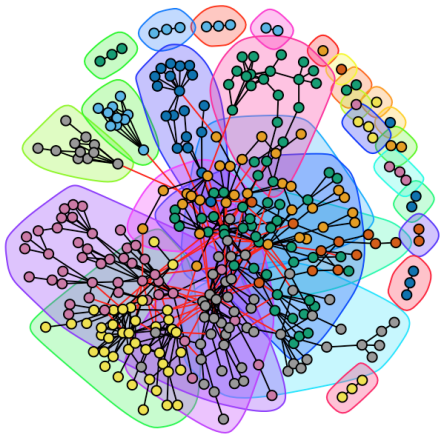


Σχήμα 2.5: Παράδειγμα ενός query

IGraph

Το IGraph είναι μια συλλογή από εργαλεία ανάλυσης δικτύων με έμφαση στην αποτελεσματικότητα, φορητότητα και εύκολη χρήση.Μπορεί να προγραμματιστεί και να

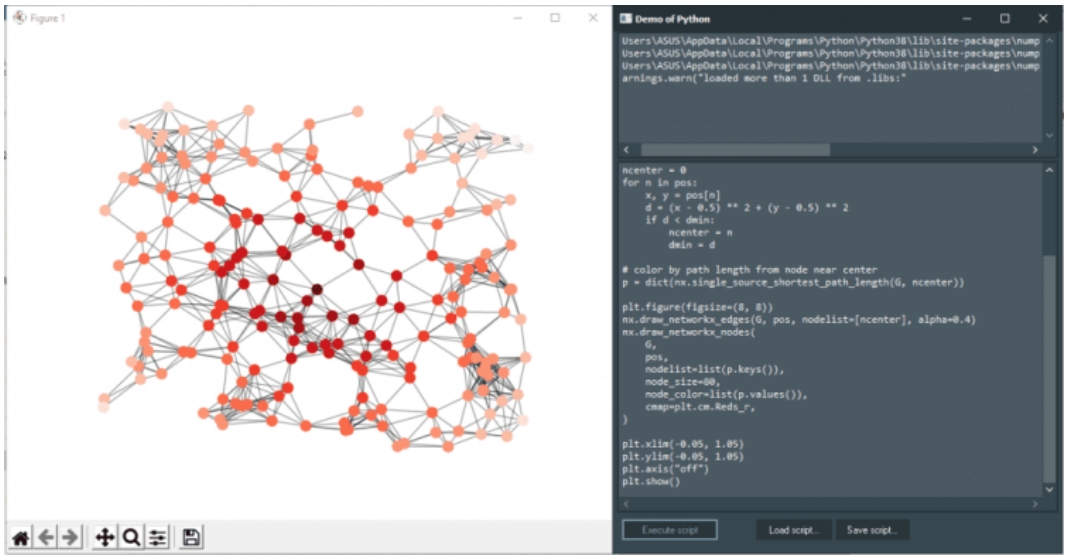
ενσωματωθεί σε γλώσσες όπως Python, R, C/C++ και Mathematica. [9]. Στο Σχήμα 2.6 φάινεται ένας γράφος με μικρές κοινότητες που δημιουργήθηκε από το igraph.



Σχήμα 2.6: Γράφος του igraph

NetworkX

Το NetworkX είναι ένα πακέτο της γλώσσας Python για τη δημιουργία, το χειρισμό και τη μελέτη της δομής, της δυναμικής και των λειτουργιών πολύπλοκων δικτύων.Είναι κατάλληλο για λειτουργία σε μεγάλα γραφήματα πραγματικού κόσμου: π.χ. γραφήματα που υπερβαίνουν τους 10 εκατομμύρια κόμβους και 100 εκατομμύρια ακμές.Είναι ένα αρκετά αποδοτικό, πολύ επεκτάσιμο, εξαιρετικά φορητό πλαίσιο για ανάλυση δικτύων και κοινωνικών δικτύων (SNA). [10]. Στο Σχήμα 2.7 βλέπουμε ένα παράδειγμα χρήσης της βιβλιοθήκης αυτής.



Σχήμα 2.7: Γράφος του networkx

Εμείς, για την οπτικοποίηση δεδομένων και την δημιουργία γράφων, χρησιμοποιήσαμε

το πρόγραμμα **Neo4j** το οποίο αντιληφθήκαμε ότι είναι το πιο έυκολο να μάθεις και να χρησιμοποιείς αφού έχει τη δική του γλώσσα Cypher. Έχει μια μεγάλη κοινότητα από προγραμματιστές που συνεισφέρουν σε αυτό το οικοσύστημα και είναι αρκετά γρήγορο στο read/write των δεδομένων ενώ παράλληλα προστατεύτει την ακεραιότητα τους. Τέλος, έχει μια κλιμακούμενη αρχιτεκτονική και είναι η μοναδική βάση δεδομένων γράφων που είναι και ACID compliant [7].

2.2 Web crawling, web scraping

Στα προηγούμενα κεφάλαια, παρουσιάσαμε την ορολογία και τις μετρικές που χρησιμοποιούνται στους γράφους γενικότερα και στα κοινωνικά δίκτυα ειδικότερα, και συζητήσαμε ερευνητικές κατευθύνσεις που αφορούν σε διαφορετικούς τύπους κοινωνικών δικτύων.

Στο παρόν κεφάλαιο, παρουσιάζονται οι μηχανισμοί με τους οποίους μπορούμε να ανακτήσουμε δεδομένα από ιστότοπους, είτε πρόκειται για το Διαδίκτυο που όλοι γνωρίζουμε (Surface Web), είτε για το Βαθύ Ιστό (Deep Web).

2.2.1 Web crawlers

Ένα πρόγραμμα ανίχνευσης ιστού, ένα spider ή ένα bot μηχανής αναζήτησης πραγματοποιεί λήψη και ευρετηρίαση περιεχομένου από όλο Διαδίκτυο. Ο στόχος ενός τέτοιου μηχανισμού είναι να εξάγει με αυτόματο τρόπο, μέσω κατάλληλου λογισμικού, κάθε ιστοσελίδα στον ιστό, έτσι ώστε οι πληροφορίες να μπορούν να ανακτηθούν όταν χρειάζονται. Ονομάζονται "web crawlers" επειδή η ανίχνευση είναι ο τεχνικός όρος για την αυτόματη πρόσβαση σε έναν ιστότοπο και τη λήψη δεδομένων μέσω ενός προγράμματος λογισμικού. [11], [12]

Οι μηχανισμοί εξόρυξης (web crawlers) λειτουργούν σχεδόν πάντα μέσω των μηχανών αναζήτησης. Εφαρμόζοντας έναν αλγόριθμο αναζήτησης στα δεδομένα που συλλέγονται από προγράμματα ανίχνευσης ιστού, οι μηχανές αναζήτησης μπορούν να παρέχουν σχετικούς συνδέσμους ως απάντηση σε ερωτήματα αναζήτησης χρηστών, δημιουργώντας μια λίστα ιστοσελίδων αντίστοιχη με αυτή που εμφανίζονται αφού ο χρήστης πληκτρολογήσει μια αναζήτηση στο Google ή στο Bing (ή σε άλλη μηχανή αναζήτησης).

Για να γίνει κατανοητό πως λειτουργεί ένας web crawler μηχανισμός, θα το παραλληλίσουμε με κάποιον που διαβάζει όλα τα βιβλία σε μια αποδιοργανωμένη βιβλιοθήκη και συνθέτει έναν κατάλογο καρτών, ώστε όποιος επισκέπτεται τη βιβλιοθήκη να μπορεί γρήγορα και εύκολα να βρει τις πληροφορίες που χρειάζεται. Για να βοηθήσει στην κατηγοριοποίηση και ταξινόμηση των βιβλίων της βιβλιοθήκης ανά θέμα, ο άνθρωπος μας θα διαβάσει τον τίτλο, την περίληψη και μέρος του εσωτερικού κειμένου κάθε βιβλίου για να καταλάβει περί τίνος πρόκειται. [11], [12]

2.2.2 Ευρετηρίαση

Η ευρετηρίαση αναζήτησης είναι σαν να δημιουργείται ένας κατάλογος καρτών βιβλιοθήκης για το Διαδίκτυο, έτσι ώστε μια μηχανή αναζήτησης να γνωρίζει πού στο Διαδίκτυο

[7] https://retool.com/blog/whats-an-acid-compliant-database/

πρέπει να ανακτήσει πληροφορίες όταν κάποιος τις αναζητά. Μπορεί επίσης να συγκριθεί με το ευρετήριο στο πίσω μέρος ενός βιβλίου, το οποίο παραθέτει όλα τα μέρη του βιβλίου όπου αναφέρεται ένα συγκεκριμένο θέμα ή φράση. [11], [12]

Η ευρετηρίαση εστιάζει κυρίως στο κείμενο που εμφανίζεται στη σελίδα και στα μεταδεδομένα[8] για τη σελίδα που δεν βλέπουν οι χρήστες. Όταν οι περισσότερες μηχανές αναζήτησης ευρετηριάζουν μια σελίδα, προσθέτουν όλους τους όρους που εμφανίζονται στη σελίδα στο ευρετήριο – εκτός από τις συχνά χρησιμοποιούμενες λέξεις όπως 'a', 'an' και 'the' στην περίπτωση της Google. Όταν οι χρήστες αναζητούν δεδομένα πληκτρολογώντας λέξεις ή φράσεις, η μηχανή αναζήτησης περνά από το ευρετήριό της με όλες τις σελίδες όπου εμφανίζονται αυτοί οι όροι και επιλέγει τις πιο σχετικές.

2.2.3 Πώς λειτουργεί το crawling

Το Διαδίκτυο αλλάζει και επεκτείνεται συνεχώς. Επειδή δεν είναι δυνατό να γνωρίζουμε πόσες συνολικά ιστοσελίδες υπάρχουν στο Διαδίκτυο, οι αυτόματοι μηχανισμοί ανίχνευσης ιστού ξεκινούν από μια αρχή ή μια λίστα γνωστών διευθύνσεων URL. Ανιχνεύουν πρώτα τις ιστοσελίδες σε αυτές τις διευθύνσεις URL. Καθώς ανιχνεύουν αυτές τις ιστοσελίδες, θα βρουν υπερσυνδέσμους προς άλλες διευθύνσεις URL και τις προσθέτουν στη λίστα των σελίδων για ανίχνευση στη συνέχεια. [11], [12]

Δεδομένου του τεράστιου αριθμού ιστοσελίδων στο Διαδίκτυο που θα μπορούσαν να ευρετηριαστούν για αναζήτηση, αυτή η διαδικασία θα μπορούσε να συνεχιστεί σχεδόν επ' αόριστον. Ωστόσο, ένας ανιχνευτής ιστού θα ακολουθήσει ορισμένες πολιτικές που το καθιστούν πιο επιλεκτικό σχετικά με το ποιες σελίδες θα ανιχνεύονται, με ποια σειρά θα τις ανιχνεύουν και πόσο συχνά θα πρέπει να τις ανιχνεύουν ξανά για να ελέγχει για ενημερώσεις περιεχομένου.

Όπως θα περίμενε κανείς, τα περισσότερα προγράμματα ανίχνευσης ιστού δεν ανιχνεύουν ολόκληρο το δημόσια διαθέσιμο Διαδίκτυο. Αντίθετα, αποφασίζουν ποιες σελίδες θα ανιχνευτούν πρώτα με βάση τον αριθμό των άλλων σελίδων που συνδέονται με αυτήν τη σελίδα, τον αριθμό των επισκεπτών που λαμβάνει αυτή η σελίδα και άλλους παράγοντες που υποδηλώνουν την πιθανότητα η σελίδα να περιέχει σημαντικές πληροφορίες.Η ιδέα είναι ότι μια ιστοσελίδα που αναφέρεται από πολλές άλλες ιστοσελίδες και έχει πολλούς επισκέπτες είναι πιθανό να περιέχει υψηλής ποιότητας, έγκυρες πληροφορίες, επομένως είναι ιδιαίτερα σημαντικό μια μηχανή αναζήτησης να τις έχει ευρετηριάσει – ακριβώς όπως μπορεί να κάνει μια βιβλιοθήκη η οποία σίγουρα θα κρατήσει πολλά αντίτυπα ενός βιβλίου που ελέγχεται από πολλούς ανθρώπους.

Επίσης σημειώνουμε πως το περιεχόμενο στον Ιστό ενημερώνεται, αφαιρείται ή μετακινείται συνεχώς σε νέες τοποθεσίες. Τα προγράμματα ανίχνευσης ιστού θα πρέπει περιοδικά να επισκέπτονται ξανά σελίδες για να βεβαιωθούν ότι η πιο πρόσφατη έκδοση του περιεχομένου έχει ευρετηριαστεί.

Τέλος, οι ανιχνευτές Ιστού αποφασίζουν επίσης ποιες σελίδες θα ανιχνεύσουν με βάση το πρωτόκολλο robots.txt (γνωστό και ως πρωτόκολλο εξαίρεσης ρομπότ). Πριν από την ανίχνευση μιας ιστοσελίδας, θα ελέγξουν το αρχείο robots.txt που φιλοξενείται από τον

[8]Στο πλαίσιο της ευρετηρίασης αναζήτησης, τα μεταδεδομένα είναι δεδομένα που λένε στις μηχανές αναζήτησης τι είναι μια ιστοσελίδα. Συχνά ο τίτλος και η μετα-περιγραφή είναι αυτά που εμφανίζονται στις σελίδες αποτελεσμάτων της μηχανής αναζήτησης, σε αντίθεση με το περιεχόμενο από την ιστοσελίδα που είναι ορατό στους χρήστες.

διακομιστή ιστού αυτής της σελίδας. Ένα αρχείο robots.txt είναι ένα αρχείο κειμένου που καθορίζει τους κανόνες για τυχόν βοτς που έχουν πρόσβαση στον φιλοξενούμενο ιστότοπο ή εφαρμογή. Αυτοί οι κανόνες ορίζουν ποιες σελίδες μπορούν να ανιχνεύουν τα bots και ποιους συνδέσμους μπορούν να ακολουθήσουν. [11], [12] Στο Σχήμα 2.8 φαίνονται δύο αρχεία-παραδείγματα του πρωτόκολλο robots.txt.

Σχήμα 2.8: Περιγραφή του πρωτόκολλου robots.txt με δύο παραδείγματα

Όλοι αυτοί οι παράγοντες που αναφέρθηκαν, σταθμίζονται διαφορετικά στους αλγόριθμους που κάθε μηχανή αναζήτησης ενσωματώνει στους μηχανισμούς ανίχνευσης και ευρετηρίασης ιστοσελίδων τους οποίους χρησιμοποιεί. Τα προγράμματα ανίχνευσης ιστού από διαφορετικές μηχανές αναζήτησης θα συμπεριφέρονται ελαφρώς διαφορετικά, αν και ο τελικός στόχος είναι ο ίδιος: λήψη και ευρετηρίαση περιεχομένου από ιστοσελίδες.

2.2.4 Διαφορές web crawling, web scraping

Σε αυτό το κεφάλαιο θα προσπαθήσουμε να εξηγήσουμε τις διαφορές των όρων Web Crawling και Web Scraping. Το γεγονός ότι υπάρχουν μεγάλες ομοιότητες μεταξύ των δύο διαδικασιών αποτελεί τον λόγο για τον οποίο οι δύο όροι συχνά χρησιμοποιούνται με τον ίδιο τρόπο. Παρ' όλα αυτά, υπάρχουν σημαντικές διαφορές μεταξύ τους. Ο όρος Web Crawling απευθύνεται στην διαδικασία χρήσης αυτοματοποιημένου λογισμικού (γνωστό ως bot ή spider) για την ανάγνωση και την αποθήκευση ολόκληρου του περιεχομένου μίας ιστοσελίδας, με σκοπό την αρχειοθέτηση και την ευρετηρίαση των πληροφοριών της. Αρκετές μηχανές αναζήτησης όπως η Bing και η Google, χρησιμοποιούν Web Crawlers για να εξάγουν όλες τις διαθέσιμες πληροφορίες από την κάθε ιστοσελίδα και, στη συνέχεια, να δημιουργήσουν ένα ευρετήριο για αυτές στις μηχανές αναζήτησης τους. Η λειτουργία ενός web crawler συνηθίζεται να παρομοιάζεται με την διαδικασίας ύφανσης του ιστού μίας αράχνης, εξού και η εναλλακτική του ονομασία Web Spider. Όπως προαναφέρθηκε, το Web Crawling εφαρμόζεται σε όλες τις σελίδες ενός ιστότοπου και όχι σε ένα υποσύνολο των περιεχομένων του, γεγονός που ενισχύει την καταλληλότητα του για εφαρμογές μεγαλύτερης κλίμακας. Το μέγεθος των εφαρμογών αυτών απαιτεί, επιπλέον, την αποτροπή συλλογής διπλότυπων πληροφοριών (duplication), με σκοπό την αποφυγή επιβάρυνσης του συστήματος και, παράλληλα, την εξοικονόμηση αποθηκευτικού χώρου. [11], [12]

Το Web Scraping είναι η διαδικασία χρήσης αυτοματοποιημένου λογισμικού για την ανάγνωση και την αποθήκευση περιεχομένου μίας σελίδας. Ωστόσο, στην προκειμένη

περίπτωση, συγκεντρώνεται ένας στοχευμένο υποσύνολο δεδομένων και όχι όλο το περιεχόμενο της ιστοσελίδας. Σε αντίθεση με το Web Crawling, το Web Scraping, δεν περιορίζεται στα όρια του διαδικτύου, αλλά μπορεί να εφαρμοστεί σε οποιαδήποτε δομή δεδομένων όπως καθορίζεται από τον χρήστη του μηχανισμού. Η χρήση του Web Scraping, συνηθίζεται σε εφαρμογές μικρότερης κλίμακας, με αποτέλεσμα ο έλεγχος για την ύπαρξη διπλοτύπων να είναι λιγότερο απαιτητικός, όμως αυτό δεν αποτρέπει την χρήση του σε μεγαλύτερου μεγέθους δεδομένα. Εν ολίγοις, παρατηρείται πως ένα Web Scraper λειτουργεί αρκετά πιο επιλεκτικά, σε σχέση με ένα Web Crawler. [11], [12]

2.2.5 Εργαλεία για Web Scraping

Scrapy

Το Scrapy είναι ένα Framework της γλώσσας προγραμματισμού Python, το οποίο χρησιμοποιείται προκειμένου να αυτοματοποιηθεί και να πραγματοποιηθεί η διαδικασία του web scraping διάφορων ιστοσελίδων. Το εργαλείο αυτό έχει ως στόχο την εξαγωγή δομημένων δεδομένων τα οποία μπορούν εν συνεχεία, να επεξεργαστούν και να χρησιμοποιηθούν για διάφορους σκοπούς. [13], [14]

BeautifulSoup

Η Beautiful Soup είναι μια βιβλιοθήκη της γλώσσας προγραμματισμού Python για την ανάλυση εγγράφων HTML και XML. Δημιουργεί ένα δέντρο ανάλυσης (parse tree) για σελίδες που μπορεί να χρησιμοποιηθεί για την εξαγωγή δεδομένων από HTML, το οποίο είναι χρήσιμο για την λειτουργεία του web scraping. [15]

Requests

Η Requests είναι μία βιβλιοθήκη της γλώσσας προγραμματισμού Python, που επιτρέπει την αποστολή HTTP αιτημάτων με εύκολο τρόπο, καθώς η πολυπλοκότητα που χαρακτηρίζει την σύνταξη HTTP αιτημάτων αντικαθίσταται από ένα API. Οι δυνατότητες της βιβλιοθήκης αυτής περιορίζονται μόνο στην ανάκτηση του πηγαίου κώδικα ενός ιστότοπου και για αυτό τον λόγο η χρήση της συνοδεύεται από επιπλέον εργαλεία, όπως το beautiful soup με σκοπό την εξαγωγή πληροφοριών. [16]

Selenium

Το Selenium αποτελεί ένα σύστημα εργαλείων που ειδικεύονται στην αυτοματοποιημένη διαδικτυακή πλοήγηση και είναι συμβατό με αρκετές γλώσσες προγραμματισμού. Τα εργαλεία αυτά χρησιμοποιούν τεχνικές για τον απομακρυσμένο έλεγχο ενός προγράμματος περιήγησης και την προσομοίωση της ανθρώπινης αλληλεπίδρασης με αυτό. Το συγκεκριμένο σύστημα επιτρέπει στους χρήστες να αυτοματοποιούν συχνές δραστηριότητες τους, όπως τη συμπλήρωση ψηφιακών φορμών, την επιλογή συνδέσμων ή ακόμα και την κίνηση του κέρσορα σε διάφορα σημεία της σελίδας.Συμπερασματικά, μπορεί να συμβάλλει στη συλλογή δεδομένων από το διαδίκτυο, δρομολογώντας μια ακολουθία διαδοχικών βημάτων, που τελικά οδηγούν στην επιλογή των επιθυμητών στοιχείων μιας ιστοσελίδας. [17]

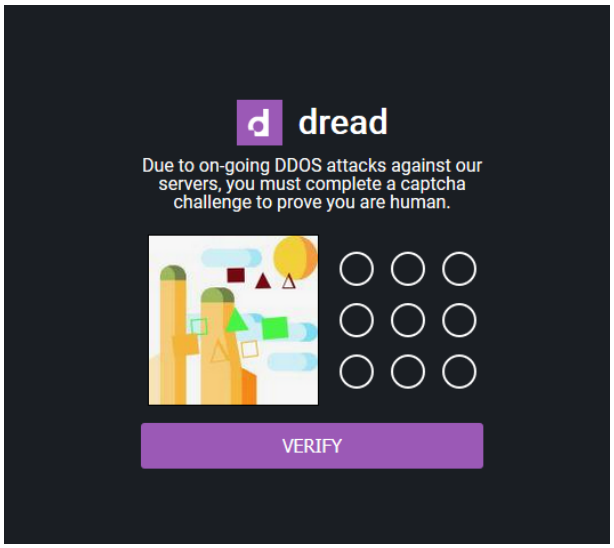
2.2.6 Εμπόδια κατά την διαδικασία του scraping

Παρ'όλα τα διαθέσιμα εργαλεία, η διαδικασία εξόρυξης δεδομένων από τον Ιστό δεν είναι πάντα εύκολη υπόθεση. Παρακάτω εξηγούμε τους συνήθεις λόγους που μπορούν να κάνουν έναν αυτοματοποιημένο μηχανισμό να αποτύχει.

CAPTCHA

Το CAPTCHA (ελλ. Τούρινγκ Τεστ για το διαχωρισμό ανθρώπου-υπολογιστή) (βλ. Σχήμα 2.9) χρησιμοποιείται συχνά για τον διαχωρισμό των ανθρώπων από τα εργαλεία απόξεσης με την εμφάνιση εικόνων ή λογικών προβλημάτων που οι άνθρωποι βρίσκουν εύκολο να λύσουν, αλλά οι web crawlers όχι. [12]

Πολλοί CAPTCHA solvers μπορούν να ενσωματωθούν σε bot για να διασφαλίσουν συνεχόμενα ασταμάτητα scrapes. Αν και οι τεχνολογίες για την υπέρβαση του CAPTCHA μπορούν να βοηθήσουν στην απόκτηση συνεχών ροών δεδομένων, θα μπορούσαν να επιβραδύνουν λίγο τη διαδικασία του scraping.



Σχήμα 2.9: Ένα από τα captcha του δημοφιλέστερου φόρουμ στο dark web

Παγίδες honeypots στον πηγαίο κώδικα

Το Honeypot είναι μια παγίδα που βάζει ο ιδιοκτήτης του ιστότοπου στη σελίδα για να πιάσει scrapers. Οι παγίδες μπορεί να είναι σύνδεσμοι που είναι αόρατοι στον άνθρωπο αλλά ορατές στους scrapers. Μόλις ένα scraper πέσει στην παγίδα, ο ιστότοπος μπορεί να χρησιμοποιήσει τις πληροφορίες που λαμβάνει (π.χ, τη διεύθυνση IP του) για να αποκλείσει αυτό το scraper. [12]

Αναγκαστικό Login

Ορισμένες προστατευμένες πληροφορίες ενδέχεται να απαιτήσουν να συνδεθούμε πρώτα. Αφού υποβάλουμε τα διαπιστευτήρια σύνδεσής μας, το πρόγραμμα περιήγησής προσθέτει αυτόματα την τιμή του cookie σε πολλαπλά αιτήματα που κάνουμε όπως οι περισσότεροι ιστότοποι, έτσι ώστε ο ιστότοπος να γνωρίζει ότι είμαστε το ίδιο άτομο που

μόλις συνδέθηκε νωρίτερα (βλ. Σχήμα 2.10). Επομένως, όταν κάνουμε scrape ιστότοπους που απαιτούν σύνδεση, πρέπει να βεβαιωθούμε ότι έχουν σταλεί τα κατάλληλα cookies μαζί με τα αιτήματα. [12]

```
cookies = {
    'XSRF-TOKEN': 'eyJpdiI6ImI4cmg1dFRZUmxxdFhhd1ZMTzdSQnc9PSIsInZhbHVlIjoiVldZUHB1bEhONG1tazFF
    'laravel_session': 'eyJpdiI6Inl1U3N0NU92YWN6YXgySDJDSllLZkE9PSIsInZhbHVlIjoiNHFEM0dUcDZ1Nkl
}

headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; rv:91.0) Gecko/20100101 Firefox/91.0',
    'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
    'Accept-Language': 'en-US,en;q=0.5',
    'Referer': 'http://fahue6hb7odzns36vfoi2dqfvqvjq4btt7vo52a67jivmyz6a6h3vzqd.onion/login',
    'Connection': 'keep-alive',
    'Upgrade-Insecure-Requests': '1',
    'Sec-Fetch-Dest': 'document',
    'Sec-Fetch-Mode': 'navigate',
    'Sec-Fetch-Site': 'same-origin',
    'Sec-Fetch-User': '?1',
    'Cache-Control': 'max-age=0',
}
```

Σχήμα 2.10: Σύνθεση user-agent και session cookies από .onion ιστότοπο

Στο Σχήμα 2.10 βλέπουμε ποια δεδομένα στέλνουμε στον ιστότοπο του Σκοτεινού Ιστού για να μας επιτρέψει να συνδεθούμε και στην συνέχεια να τραβήξουμε δεδομένα. Σαν user-agent βλέπουμε ότι είναι ο Mozilla ο οποίος είναι στην περίπτωση μας ο default του TOR, αλλά θα μπορούσαμε εύκολα να βάλουμε στον κώδικα μας κάποιον άλλον. Τα XSRF TOKEN Laravel Session είναι 64 ψηφία τυχαίας ακολουθίας γραμμάτων τα οποία όρισε ο ιστότοπος που θέλουμε να επισκεφθούμε, το όνομα του οποίου φαίνεται στο πεδίο 'Referer'.

Δυναμική εναλλαγή της IP

Κατά την διαδικασία εξαγωγής πληροφοριών και δεδομένων από μία σελίδα, η διεύθυνση IP του χρήστη είναι ορατή προς τον εξυπηρετητή της σελίδας. Το συγκεκριμένο γεγονός οφείλεται στο ότι ο ο ιστότοπος στον οποίο πραγματοποιείται η επίσκεψη, έχει τη δυνατότητα να συλλέγει πληροφορίες σχετικά με τους χρήστες και να ελέγχει αν κάποιος από αυτούς συμπεριφέρεται περίεργα, δηλαδή αν συλλέγει μεγάλο όγκο δεδομένων, αν χρησιμοποιεί κάποιο μοτίβο κατά τη διάρκεια της συλλογής δεδομένων ή αν στέλνει πολλά αιτήματα, ενέργειες που κάποιος άνθρωπος δεν θα ήταν δυνατό να πραγματοποιήσει χωρίς τη χρήση αυτοματοποιημένου μηχανισμού. Σε περίπτωση που παρατηρηθεί κάποιο από τα προηγούμενα, γίνεται αντιληπτό ότι ο επισκέπτης υπέπεσε σε παραβίαση των όρων προστασίας των δεδομένων και, ως εκ τούτου, αποκλείεται η διεύθυνση του από την ιστοσελίδα. Μια από τις μεθόδους για την αποτροπή του αποκλεισμού της διεύθυνσης IP είναι η δημιουργία ενός ηλεκτρονικού αποθετηρίου, το οποίο θα απαρτίζεται από μία λίστα με διαφορετικά IPs (proxy servers). Κάθε φορά που ο χρήστης θα μαζεύει έναν αριθμό από δεδομένα θα πρέπει να προσέχει για την γρήγορη και αυτόματη εναλλαγή της διεύθυνσης του.

Παρόμοιες λύσεις στο παραπάνω πρόβλημα προσφέρει και χρήση των VPN [9] και TOR [10]. Όμως, η σύνδεση σε εξυπηρετητή/μεσολαβητή επιβαρύνει την ταχύτητα αποστολής και λήψης αιτημάτων, κάτι το οποίο μπορεί να αποβεί μοιραίο για τη συνολική διαδικασία, προκαλώντας σφάλματα. [12]

Δυναμικό περιεχόμενο

Κάποιες ιστοσελίδες, χρησιμοποιούν AJAX για το δυναμικό περιεχόμενό τους. Μερικά παραδείγματα είναι εικόνες που φορτώνουν αργά, το συνεχές scrolling για να εμφανιστεί νέο περιεχόμενο και το κουμπί ¨περισσότερα¨. Για τους χρήστες είναι μια απλή διαδικασία, όχι όμως για τους scrapers. [12]

2.3 TOR Browser και Dark Web

Αφού συζητήσαμε, τι είναι ανάλυση κοινωνικών δικτύων, τι σημαίνει web scraping και τα εργαλεία που χρησιμοποιούνται ευρέως για την απόξεση περιεχομένου από ιστοσελίδες, θα αναφερθούμε στον πυρήνα της εργασίας, τον σκοτεινό ιστό.

Σκοτεινός ιστός ή Βαθύς ιστός

Αν και οι 2 όροι χρησιμοποιούνται το ίδιο όταν θέλουμε να μιλήσουμε για υπηρεσίες και ιστότοπους που δεν είναι ορατές στον επιφανειακό ιστό (surface web). Από τη μια πλευρά, ο **Βαθύς Ιστός** περιλαμβάνει στο μεγαλύτερο βαθμό βάσεις δεδομένων, οι οποίες περιέχουν τόσο δημόσια όσο και ιδιωτικά προστατευμένες συλλογές αρχείων που δεν είναι συνδεδεμένες με άλλες περιεχές του Ιστού, αλλά μόνο για αναζήτηση εντός της ίδιας βάσης δεδομένων.Επίσης περιλαμβάνει intranets, εσωτερικά δίκτυα δηλαδή, που χρησιμοποιούνται από κυβερνήσεις, οργανισμούς και επιχειρήσεις. Καθώς εμβαθύνουμε στο δεεπ ωεβ, το περιεχόμενο εκεί, παραδόξως, είναι πιο 'καθαρό' και ασφαλές.Μπορεί να περιλαμβάνει αναρτήσεις σε ιστολόγια (blog posts) και σελίδες τις οποίες επισκεπτόμαστε όταν κάνουμε κάποια τραπεζική συναλλαγή διαδικτυακά.Οι σελίδες αυτές δεν αποτελούν καμία απειλή και ο μόνος λόγος που είναι 'κρυφές' είναι για την ασφάλεια και την προστασία του απορρήτου των χρηστών, όπως για παράδειγμα χρηματοοικονομικοί λογαριασμοί (τραπεζικοί, συνταξιοδοτικοί), λογαριασμοί κοινωνικής δικτύωσης ή/και ηλεκτρονικού ταχυδρομίου, βάσεις δεδομένων ιδιωτικών επιχειρήσεων και νομικά αρχεία [18]

Όπως προείπαμε, το να εισέρθουμε περισσότερο στον Βαθύ Ιστό φέρνει λίγο περισσότερο φως. Για ορισμένους χρήστες, τμήματα του Βαθύ ιστού προσφέρουν την ευκαιρία να παρακάμψουν τους τοπικούς περιορισμούς και να αποκτήσουν πρόσβαση σε υπηρεσίες τηλεόρασης ή ταινίας που ενδέχεται να μην είναι διαθέσιμες στις τοπικές τους περιοχές. Άλλοι πηγαίνουν κάπως πιο βαθιά για να κατεβάσουν πειρατική μουσική ή να 'κλέψουν' ταινίες που δεν έχουν βγει ακόμα στις αίθουσες.

Στο σκοτεινό άκρο του ιστού, βρίσκουμε το πιο επικίνδυνο περιεχόμενο και δραστηριότητα. Οι ιστότοποι Tor βρίσκονται σε αυτό το άκρο του βαθύ ιστού, οι οποίοι θεωρούνται ο «σκοτεινός ιστός» και είναι προσβάσιμοι μόνο από ένα ανώνυμο πρόγραμμα περιήγησης.

[9] https://www.forbes.com/advisor/business/software/what-is-a-vpn-and-how-does-it-work/
[10] https://www.torproject.org/

Η ασφάλεια στο Deep Web είναι πιο σχετική με τον μέσο χρήστη του διαδικτύου από την ασφάλεια του σκοτεινού ιστού, καθώς θα μπορούσαμε να καταλήξουμε σε επικίνδυνες περιοχές κατά λάθος: πολλά τμήματα του Βαθύ Ιστού μπορούν ακόμα να είναι προσβάσιμα σε κανονικά προγράμματα περιήγησης στο Διαδίκτυο. Αυτός είναι ο τρόπος με τον οποίο οι χρήστες μπορούν να ταξιδέψουν μέσα από αρκετά εφαπτομενικά μονοπάτια και να καταλήξουν σε έναν ιστότοπο πειρατείας, ένα πολιτικά ριζοσπαστικό φόρουμ ή να βλέπουν βίαιο περιεχόμενο.

Από την άλλη, ο **Σκοτεινός Ιστός** (Dark Web αναφέρεται σε ιστότοπους που δεν είναι ευρετηριασμένοι και είναι προσβάσιμοι μόνο μέσω εξειδικευμένων προγραμμάτων περιήγησης ιστού.[11] Σημαντικά μικρότερος από τον μικροσκοπικό ιστό επιφάνειας, ο Σκοτεινός Ιστός θεωρείται μέρος του Βαθύ Ιστού. Χρησιμοποιώντας την οπτική μας για τον ωκεανό και το παγόβουνο, ο σκοτεινός ιστός θα ήταν η κάτω άκρη του βυθισμένου παγόβουνου. [18]

Ο σκοτεινός ιστός, ωστόσο, είναι ένα πολύ κρυφό τμήμα του Βαθύ Ιστού με το οποίο λίγοι θα αλληλεπιδράσουν ή ακόμη και θα δουν. Με άλλα λόγια, ο Βαθύς Ιστός καλύπτει τα πάντα κάτω από την επιφάνεια που είναι ακόμα προσβάσιμα με το κατάλληλο λογισμικό, συμπεριλαμβανομένου του σκοτεινού ιστού.

Η ανάλυση της κατασκευής του σκοτεινού ιστού αποκαλύπτει μερικά βασικά επίπεδα που τον καθιστούν ένα ανώνυμο καταφύγιο:

1. Δεν υπάρχει ευρετηρίαση ιστοσελίδων από μηχανές αναζήτησης ιστού επιφάνειας. Το Google και άλλα δημοφιλή εργαλεία αναζήτησης δεν μπορούν να ανακαλύψουν ή να εμφανίσουν αποτελέσματα για σελίδες στον σκοτεινό ιστό.

2. Δεν είναι προσβάσιμο από τα παραδοσιακά προγράμματα περιήγησης λόγω του μοναδικού χειριστή μητρώου. Επίσης, κρύβεται περαιτέρω από διάφορα μέτρα ασφάλειας δικτύου, όπως τείχη προστασίας και κρυπτογράφηση. Η φήμη του σκοτεινού ιστού έχει συχνά συνδεθεί με εγκληματικές προθέσεις ή παράνομο περιεχόμενο και ιστότοπους «εμπορίου» όπου οι χρήστες μπορούν να αγοράσουν παράνομα αγαθά ή υπηρεσίες. Ωστόσο, και τα νομικά μέρη έκαναν χρήση αυτού του πλαισίου.

3. Όσον αφορά την ασφάλεια του σκοτεινού ιστού, οι κίνδυνοι του Βαθύ Ιστού είναι πολύ διαφορετικοί από τους κινδύνους του Σκοτεινού Ιστού. Η παράνομη διαδικτυακή δραστηριότητα δεν μπορεί απαραιτήτως να συναντηθεί εύκολα, αλλά τείνει να είναι πολύ πιο ακραία και απειλητική εάν την αναζητήσουμε. [18]

Μαύρες αγορές

Η μαύρη αγορά είναι μια οικονομική δραστηριότητα που λαμβάνει χώρα εκτός των 'ορίων' που έχουν εγκριθεί από την κυβέρνηση. Οι παράνομες συναλλαγές στην αγορά γίνονται συνήθως «κάτω από το τραπέζι» για να επιτρέψουν στους συμμετέχοντες να αποφύγουν τους κρατικούς ελέγχους τιμών ή τους φόρους. Τα αγαθά και οι υπηρεσίες

[11] Η πρόσβαση σε αυτούς τους ιστότοπους γίνεται μέσω του Tor (The Onion Router), ένα ειδικό πρόγραμμα περιήγησης που συγκαλύπτει την IP διεύθυνση του χρήστη. Αυτό το ανώνυμο δίκτυο αποτελείται από χιλιάδες διακομιστές ή κόμβους που βρίσκονται σε όλο τον κόσμο και που τους λειτουργούν εθελοντές και υποστηρικτές της ιδιωτικότητας. Το Tor αναπτύχθηκε αρχικά από τον στρατό των ΗΠΑ, προκειμένου να βοηθήσει στην ασφαλή μεταφορά επικοινωνιών και η χρήση του δεν είναι παράνομη. Άλλα δυο δίκτυα που παρέχουν πρόσβαση στο Βαθύ Ιστό είναι το Freenet και το I2P.

που προσφέρονται σε μια μαύρη αγορά μπορεί να είναι παράνομα, που σημαίνει ότι η αγορά και η πώλησή τους απαγορεύονται από το νόμο ή μπορεί να είναι νόμιμα αλλά να γίνονται συναλλαγές για την αποφυγή φόρων. [19]

Οι παράνομες αγορές είναι επίσης γνωστές ως σκιώδεις αγορές ή υπόγειες αγορές. Μερικά σημεία-κλειδία που πρέπει να σημειώσουμε είναι τα εξής. Παραδείγματα υπόγειων αγορών περιλαμβάνουν την πώληση παράνομων ναρκωτικών, όπλων, εμπορίας ανθρώπων και το παράνομο εμπόριο άγριων ζώων. Έχουν αρνητικό αντίκτυπο στην οικονομία επειδή η δραστηριότητα δεν αναφέρεται και δεν εισπράττονται φόροι επί των συναλλαγών. Παρέχουν ορισμένα οφέλη, όπως η δημιουργία θέσεων εργασίας για όσους ενδέχεται να μην μπορούν να βρουν δουλειά σε παραδοσιακές αγορές και η πρόσβαση στην ιατρική και την υγειονομική περίθαλψη σε εκείνα τα άτομα που ενδέχεται να μην είχαν πρόσβαση διαφορετικά.

Παραδοσιακά, η δραστηριότητα της υπόγειας αγοράς διεξαγόταν σε μετρητά, μια από τις καθοριστικές πτυχές της. Αυτό έγινε για να αποφευχθεί η ανίχνευση χρηματικής δραστηριότητας (paper trail). Με την άνοδο του Διαδικτύου, πολλές υπόγειες συναλλαγές στην αγορά γίνονται πλέον διαδικτυακά, όπως στον σκοτεινό ιστό, και συχνά πραγματοποιούνται με ψηφιακά νομίσματα όπως bitcoins, monero coins, ethereum coins. [19]

Εξαιτίας αυτής της φύσης, οι παράνομες αγορές μπορούν να επηρεάσουν αρνητικά μια οικονομία, καθώς είναι σκιώδεις αγορές όπου δεν καταγράφεται οικονομική δραστηριότητα και δεν καταβάλλονται φόροι. Συχνά θεωρείται ότι το ακαθάριστο εγχώριο προϊόν (ΑΕΠ) μιας χώρας δεν είναι το πραγματικό της ΑΕΠ επειδή δεν λαμβάνει υπόψη όλη την επιχειρηματική δραστηριότητα που διεξάγεται σε υπόγειες αγορές.

Στα πολλά μειονεκτήματα της υπόγειας αγοράς περιλαμβάνονται ο κίνδυνος απάτης, η πιθανότητα βίας και η επιβάρυνση με πλαστά προϊόντα ή νοθευμένα προϊόντα, κάτι που είναι ιδιαίτερα επικίνδυνο στην περίπτωση των φαρμάκων. Μιας από τις πιο γνωστές μαύρες διαδικτυακές αγορές ήταν και το Silk Road όπου μπορούσε κανείς να βρει ναρκωτικά, διαβατήρια, όπλα και ένα σωρό άλλα παράνομα προϊόντα και υπηρεσίες. Οι αγορές αυτές γίνονταν παράνομα με χρήση του Bitcoin. Το ψηφιακό αυτό νόμισμα χρησιμοποιεί το σύστημα Peer to Peer και οι χρήστες του μπορούν να κάνουν τις αγορές τους ανώνυμα.

Το 2013, το FBI συνέλαβε τον ιδρυτή του Silk Road για διακίνηση ναρκωτικών, ξέπλυμα μαύρου χρήματος και χάκινγκ. Κατά το διάστημα της λειτουργίας του, οι πωλήσεις ανέρχονταν στο 1 δις δολάρια. Από τότε όμως ξεπήδησαν πολλές παρόμοιες ιστοσελίδες. Θεωρητικά αυτές οι εικονικές αγορές απαγορεύουν την παιδική πορνογραφία και τις αγγελίες για πληρωμένους δολοφόνους, αποδέχονται όμως την πώληση όπλων, την πορνογραφία, χακαρισμένα λογισμικά, πλαστά χρήματα, διαβατήρια και ναρκωτικά.[12]

Σήμερα, ο Βαθύς Ιστός (Deep Web) είναι αρκετά πιο αποδεκτός από ό,τι ήταν πριν από λίγα χρόνια και πολλές διαδικτυακές υπηρεσίες πειραματίζονται με αυτήν την πλατφόρμα. Τον Οκτώβριο του 2014, το Facebook δημιούργησε μια άλλη διεύθυνση στην οποία μπορεί να μπει κάποιος μέσω του TOR. Σε δελτίο τύπου του, το δημοφιλές κοινωνικό δίκτυο ανακοίνωσε την δημιουργία της διεύθυνσής αυτής αναφέροντας μεταξύ άλλων

[12]https://www.facebook.com/notes/facebook-over-tor/1-million-people-use-facebook-over-tor/865624066877648

ότι «Είναι σημαντικό για εμάς στο Facebook να παρέχουμε μεθόδους στους χρήστες που θέλουν να χρησιμοποιήσουν την ιστοσελίδα μας με ασφάλεια». Τον Απρίλιο του 2016 η σελίδα αυτή είχε πάνω από ένα εκατομμύριο χρήστες.[13]

[13]http://www.whoishostingthis.com/blog/2013/12/17/tor-deep-web/

Κεφάλαιο 3

Ανάλυση εργαλείων και φόρουμ

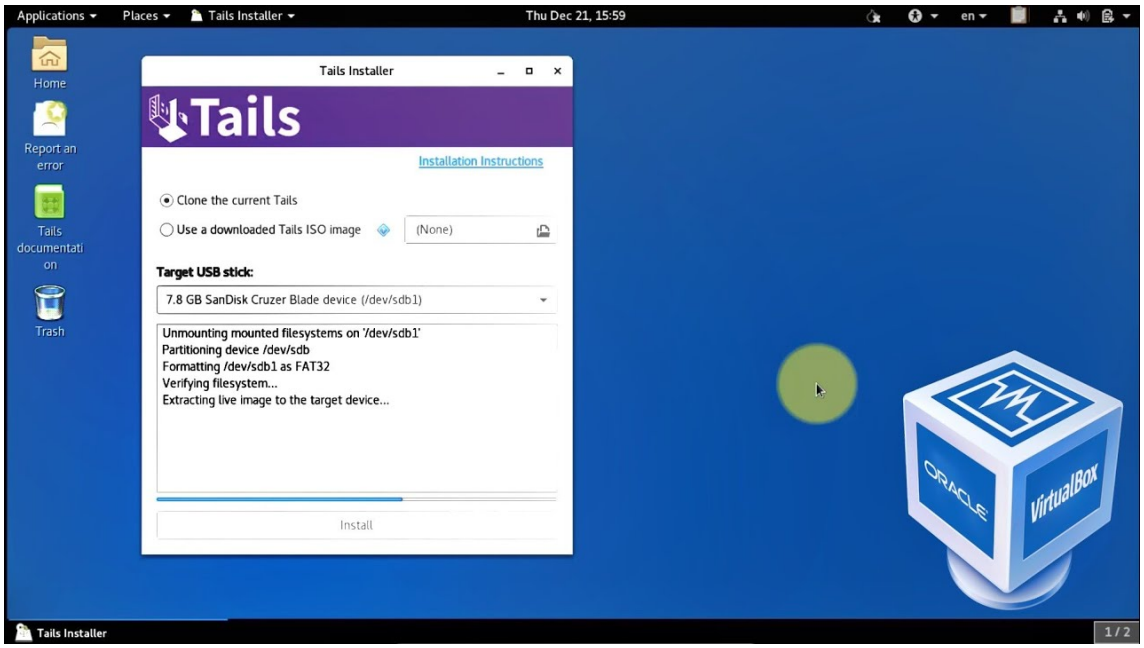
Στο παρόν κεφάλαιο θα αναλύσουμε εκτενώς όλα τα εργαλεία που χρησιμοποιήθηκαν με τη σειρά για την επιτυχή εκπόνηση της εργασίας και το τρόπο που προσεγγίσαμε το θέμα, έτσι ώστε να λειτουργήσει σαν γέφυρα για το Κεφάλαιο 4 που θα αναλύσουμε τον κώδικα μας.

3.1 Tails OS

Θα ξεκινήσουμε με το πρώτο πρόγραμμα-εργαλείο που ίσως φαίνεται προαιρετικό για όσους χρήστες θέλουν απλά μια γρήγορη εξερεύνηση του σκοτεινού ιστού. Παρολαυτά για τις ανάγκες της εργασίας, χρησιμοποιήθηκε για να είμαστε απόλυτα σίγουροι οτι δεν θα κινδυνέψει το σύστημα μας.

Το λειτουργικό σύστημα Tails είναι μια Debian-based διανομή του Linux που επιτρέπει απόλυτη ασφάλεια και ανωνυμία για σερφάρισμα στο διαδίκτυο. Συνδέεται μέσω του Tor οπότε ήταν οτι χρειαζόμασταν για να ξεκινήσουμε την αναζήτηση κοινωνικών δικτύων και φόρουμ. Για να αποφύγουμε την εγκατάσταση του λειτουργικού στο δίσκο μας, το τρέξαμε μέσω εικονικής μηχανής[1] (virtual machine) (βλ. Σχήμα 3.1).

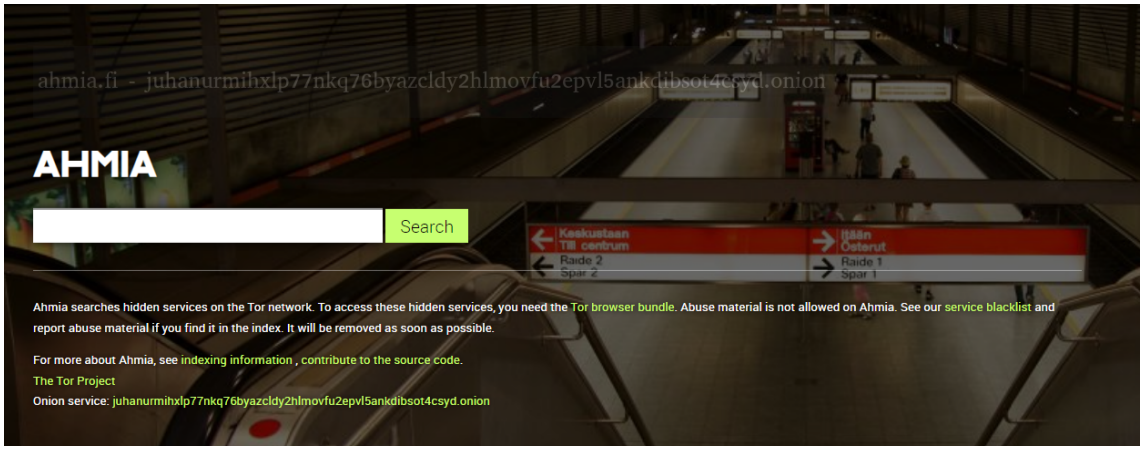
[1] https://tails.boum.org/doc/advanced$_{t}$*opics/virtualization/index.en.html*



Σχήμα 3.1: Διεπαφή χρήστη της εικονικής μηχανής του Tails OS

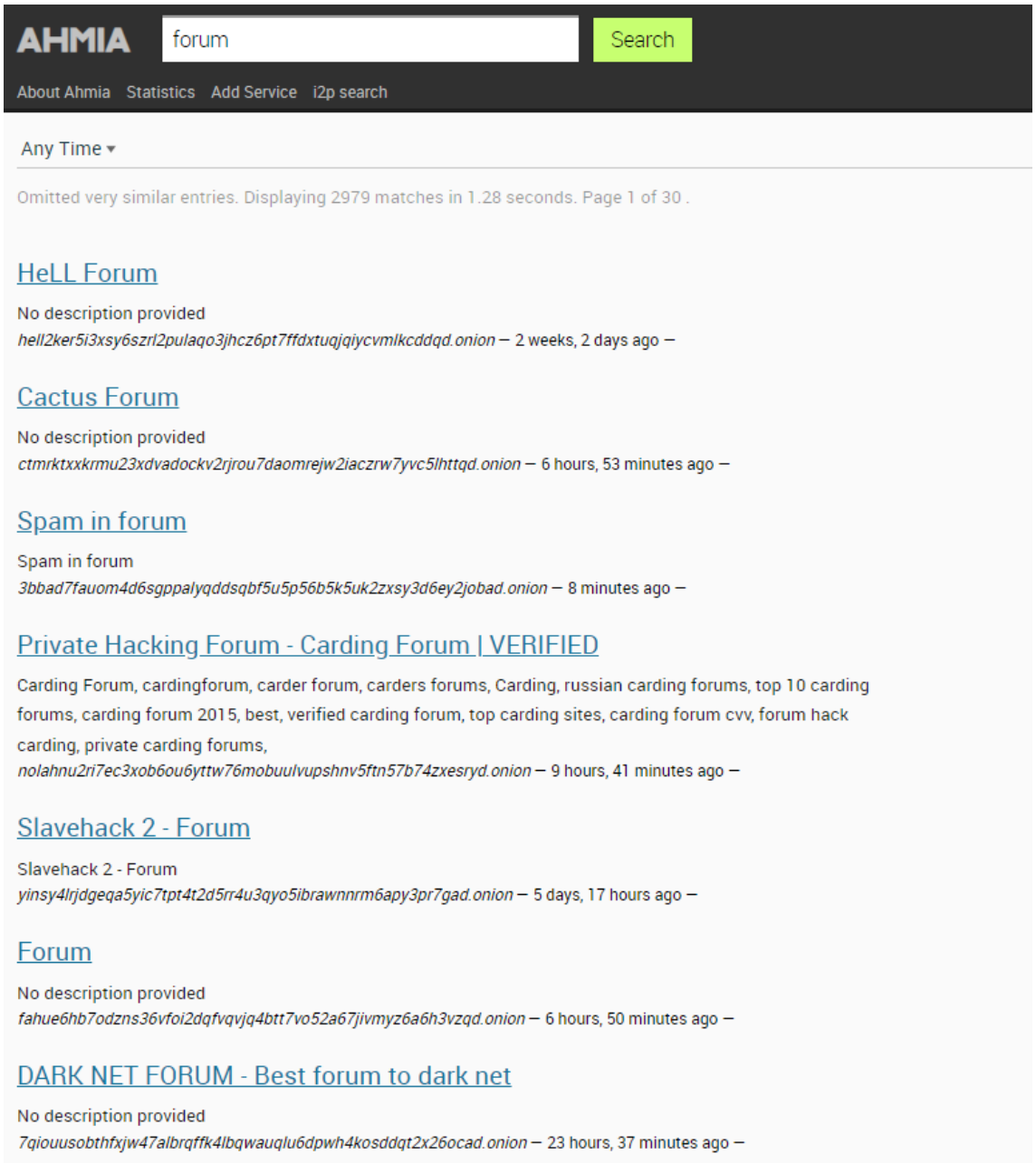
3.2 Ahmia

Δημιουργήσαμε μια σύνδεση μέσω TOR και είμαστε έτοιμοι να ξεκινήσουμε για αναζήτηση ιστότοπων. Ο τρόπος που σχεδόν όλοι θα σκεφτούν είναι μέσω μηχανών αναζήτησης. Η πιο γνωστή μηχανή αναζήτησης που μπορεί κάποιος να χρησιμοποιήσει ακόμα και από το surface web είναι η Ahmia[2] (βλ. Σχήμα 3.2). Βέβαια, τα αποτελέσματα που θα βγάλει είναι σε .onion, δηλαδή ιστότοποι που μόνο με το TOR μπορούν να ανοίξουν. Στο Σχήμα 3.3 φάινονται μερικά αποτελέσματα που εμφανίζονται όταν πατάμε τη λέξη-κλειδί "forum".



Σχήμα 3.2: Διεπαφή χρήστη της μηχανής αναζήτησης ahmia

[2] ahmia.fi



Σχήμα 3.3: Τα αποτελέσματα της ahmia όταν ψάχνουμε τον όρο "forum"

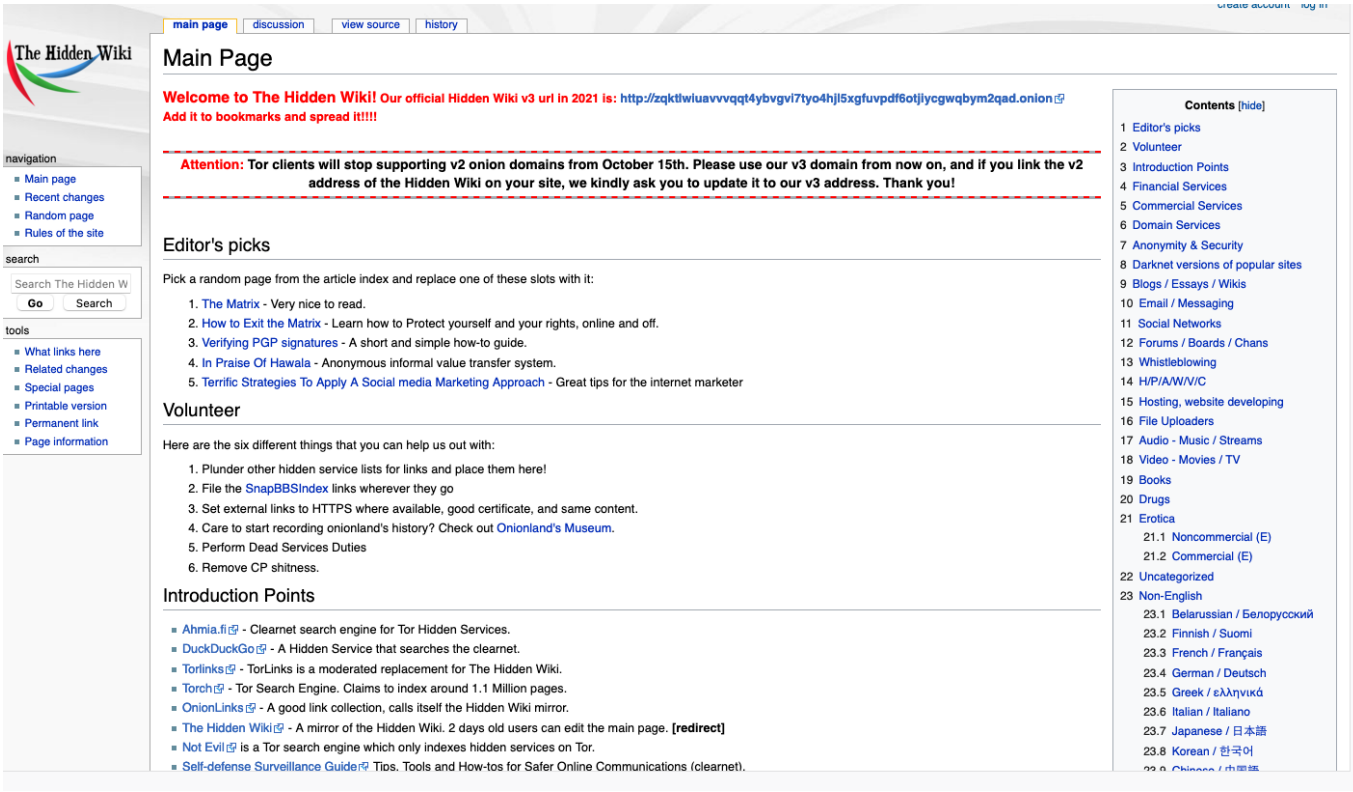
Άλλες γνωστές μηχανές αναζήτησης του σκοτεινού ιστού είναι η torch, η recon και η duckduckgo.

Οι μηχανές αναζήτησης στο σκοτεινό ιστό όμως δεν λειτουργούν όπως οι παραδοσιακές, αφού πρέπει να ακολουθήσεις πολλούς υπερσυνδέσμους για να φτάσεις σε ένα επιθυμητό αποτελέσμα, όπως ένα φόρουμ με χιλιάδες χρήστες και αρκετό περιεχόμενο για scrape.

3.3 Hidden Wiki

Όπως είπαμε παραπάνω, η χρήση των μηχανώς αναζήτησης για να βρούμε αυτό που ψάχνουμε μπορεί να είναι χρονοβόρα διαδικασία ακόμα και επικίνδυνη, αφού αρκετοί

ιστότοποι μπορεί να είναι scams.[3] Ο πιο αποτελεσματικός τρόπος για να βρούμε ότι ψάχνουμε πιο άμεσα είναι η χρήση του hidden wiki που δεν είναι άλλο από μια αποθήκη ιστότοπων που συνεχώς ενημερώνεται.Αυτή η ¨βιβλιοθήκη¨ έχει δημιουργηθεί έτσι ώστε αρχάριοι χρήστες του TOR να βρούν εύκολα αυτό που ψάχνουν και να ενημερώνονται εάν ο ιστότοπος που θέλουν να επισκεφθούν λειτουργεί ή άλλαξε url (βλ. Σχήμα 3.4).



Σχήμα 3.4: Διεπαφή χρήστη του hidden wiki

Στο hidden wiki βρίσκονται ήδη κάποια από τα φόρουμ που εν τέλει αναλύσαμε (π.χ., το Dread) αλλά και συνδέσεις με άλλα κρυμμένα που μας οδήγησαν στα υπόλοιπα φόρουμ για να καταλήξουμε στα 5 τελικά.

3.4 Τα 5 τελικά φόρουμ

3.4.1 Dread

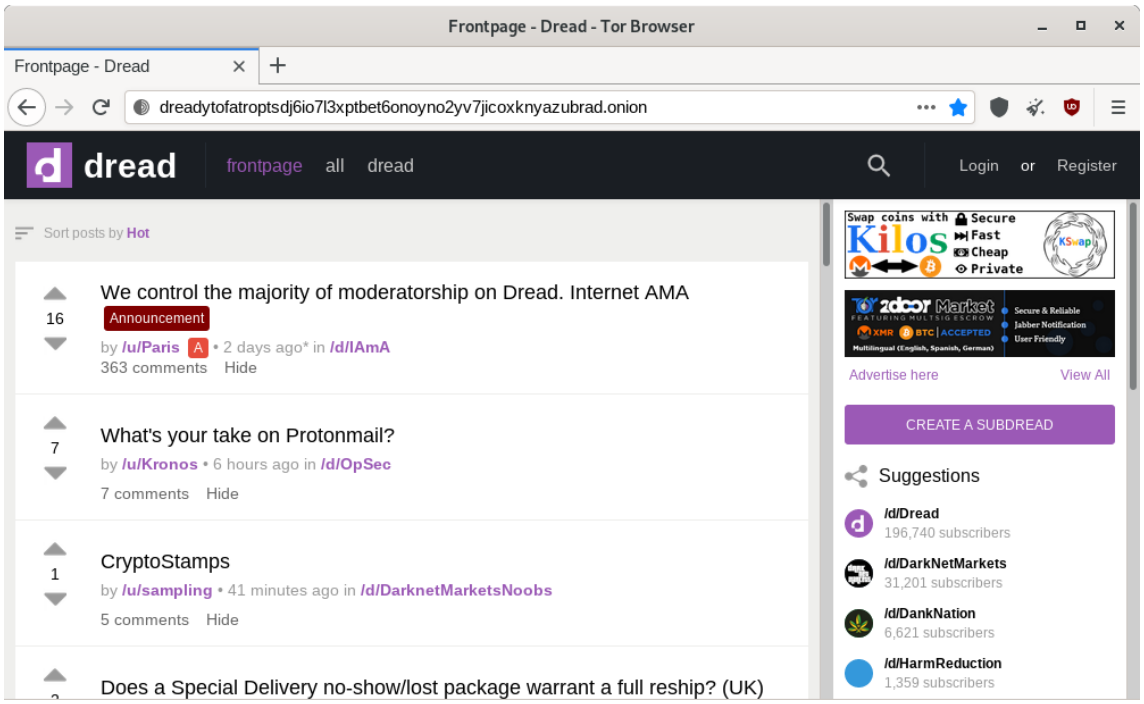
Το Dread (βλ. Σχήμα 3.5) είναι αυτή τη στιγμή το μεγαλύτερο φόρουμ-αγορά στο σκοτεινό ιστό, και ίσως το πιο δύσκολο να αντλήσεις δεδομένα με αυτοματοποιημένο τρόπο μιας και ύστερα από αλλεπάλληλες επιθέσεις ddos [4], τα captcha είναι παγκοσμίως τα πιο δύσκολα να επιλύσεις (βλ. Σχήμα 2.9). Είναι ένας κλώνος του reddit

[3]Οι ιστότοποι απάτης (scams) είναι οποιοσδήποτε παράνομος ιστότοπος στο Διαδίκτυο που χρησιμοποιείται για να παραπλανήσει τους χρήστες σε απάτη ή κακόβουλες επιθέσεις. Οι απατεώνες καταχρώνται την ανωνυμία του Διαδικτύου για να κρύψουν την πραγματική τους ταυτότητα και τις προθέσεις τους πίσω από διάφορες μεταμφιέσεις. Αυτά μπορεί να περιλαμβάνουν ψευδείς ειδοποιήσεις ασφαλείας, δώρα και άλλες παραπλανητικές μορφές που δίνουν την εντύπωση νομιμότητας.

[4]https://www.imperva.com/learn/ddos/ddos-attacks/

με εκατοντάδες υποφόρουμ όπως τεχνολογία, υποκλοπή καρτών, χάκινγκ , χρήση και παραγωγή ναρκωτικών και άλλα.

Επίσης, ενώ δεν προυποθέτει να έχεις λογαριασμό για να κάνεις μια απλή περιήγηση, απαιτείται όταν θες να στείλεις ένα αίτημα (http request) [5] και να τραβήξεις δεδομένα (μέσω session cookies που θα μιλήσουμε στο επόμενο κεφάλαιο).

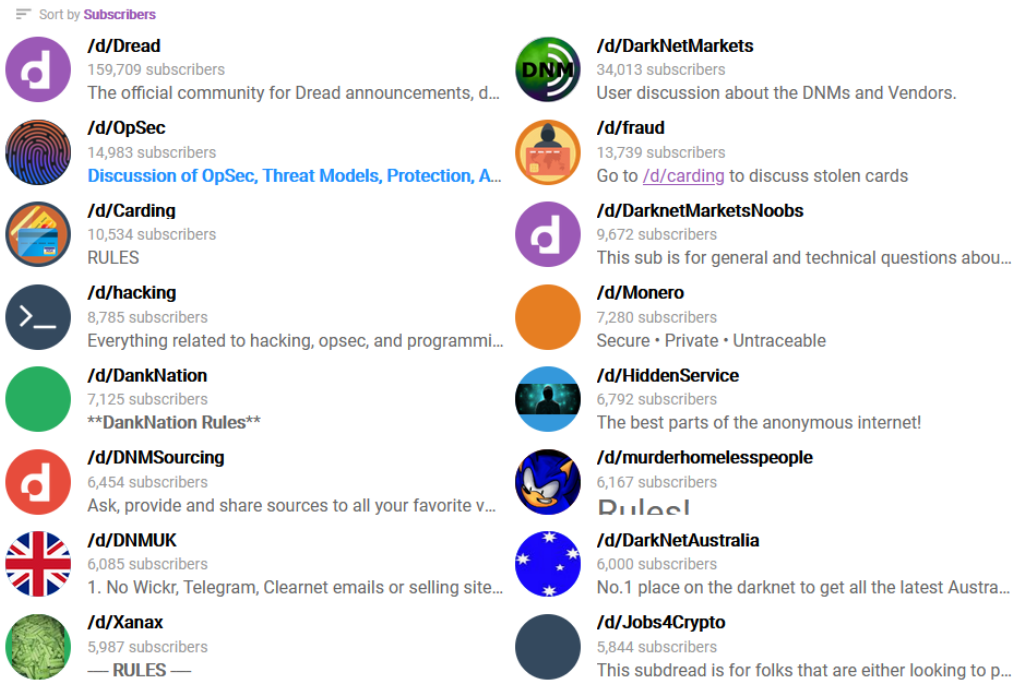


Σχήμα 3.5: Διεπαφή χρήστη του dread

Όπως είπαμε, αφού το Dread είναι το μεγαλύτερο φόρουμ του σκοτεινού ιστού, είναι λογικό να έχει αρκετό υλικό όπως αναρτήσεις ή/και σχόλια άλλα εμάς μας αφορούν μόνο τα πιο πρόσφατα (5-6 μήνες). Γιαυτό το λόγο διαχωρίσαμε 23 από τα 100 συνολικά υποφόρουμ που υπάρχουν, διότι στα υπόλοιπα η τελευταία ανάρτηση ήταν πριν 1 με 2 χρόνια. Μερικά από τα υποφόρουμ (βλ. Σχήμα 3.6) αυτά φέρουν τίτλους όπως opiates, cyber-security, cannabis, hacking, crypto, fraud, report vendors, dark0dereborn.[6]

[5] https://developer.mozilla.org/en-US/docs/Web/HTTP/Overview

[6] Μετά το κλείσιμο του Silk Road, το Dark0deReborn πήρε τη σκυτάλη του εμπορίου ναρκωτικών.



Σχήμα 3.6: Μερικά από τα υποφόρουμ του Dread

Όπως γίνεται αντιληπτό από την παραπάνω εικόνα, τα περισσότερα ποστ και οι περισσότεροι χρήστες ανήκουν στο ομώνυμο υποφόρουμ. Αξιοσημείωτο είναι επίσης ότι στο Dread, του οποίου ο αριθμός χρηστών πολλαπλασιάζεται συνεχώς, έχει κατά μέσο όρο 500 ποστ την ημέρα.

3.4.2 Hidden Answers

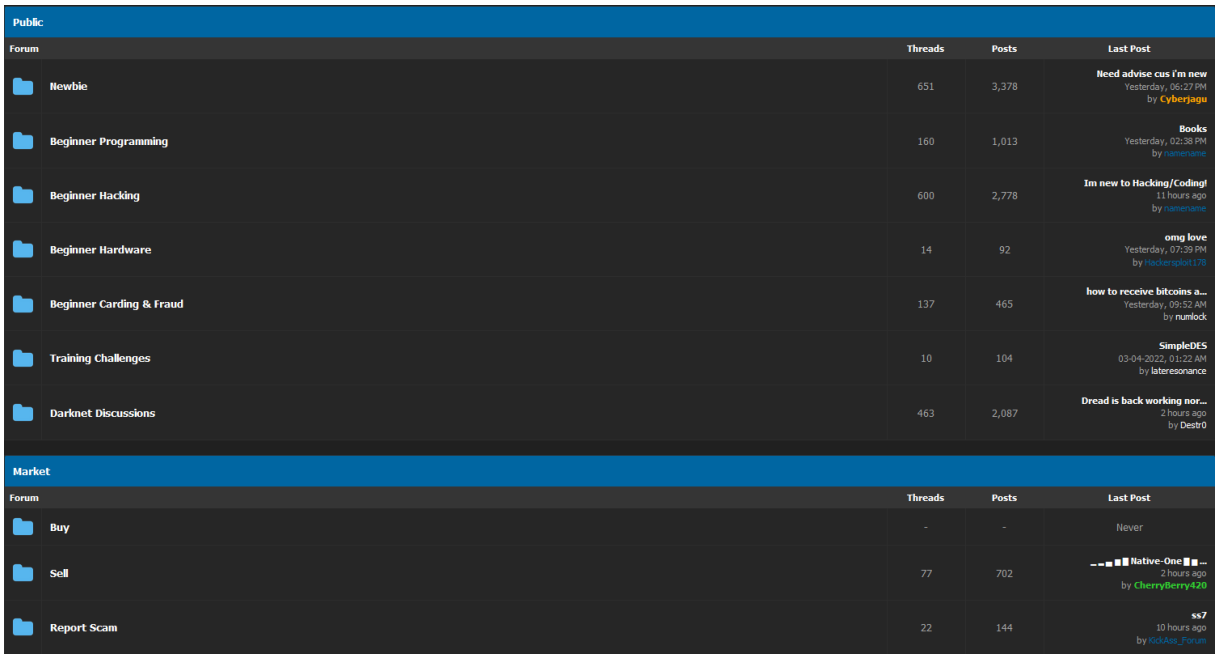
Το Hidden Answers (ισπ. Respostas Ocultas) (βλ. Σχήμα 3.7) είναι ένα ισπανόφωνο φόρουμ και ίσως το πιο απομακρυσμένο από τις μηχανές αναζήτησης και τα hidden wiki του σκοτεινού ιστού. Με πάνω από 2000 χρήστες και 4000 συνολικά ποστ, τα υπόφορουμ του αναφέρονται σε θέματα όπως επιστήμη, πολιτική, μαύρες αγορές, προσωπικές στιγμές-εμπειρίες και γενικά είναι το φόρουμ που εναγκαλίζει όλων των ειδών τις 'περίεργες' ερωτήσεις που κάποιος χρήστης μπορεί να έχει και να δυσκολεύεται να μοιραστεί αλλού (εξού και το όνομα του φόρουμ).

Είναι ένας ιστότοπος-αντίγραφο ενός παλαιότερου δημοφιλή αγγλόφωνου ιστότοπου με το ίδιο όνομα που έκλεισε ύστερα από αρκετό καιρό λειτουργίας. Για τις ανάγκες τις πτυχιακής, έγινε μετάφραση όλων των ποστ και απαντήσεων αυτών στην αγγλική γλώσσα για καλύτερη ανάλυση. Να σημειώσουμε ότι μαζί με το Dread τα αμέσως μεγαλύτερα σε σειρά φόρουμ είναι μη-Αγγλικά όπως το γερμανικό Deutchland im deep web, το πολωνικό Cebulka και το ρωσικό XSS. Το καθένα από αυτά είχε τις δικές του δυσκολίες και εμπόδια, οπότε αποφασίσαμε να συμπεριλάβουμε μόνο το ισπανικό Hidden Answers μεταξύ των υπόλοιπων αγγλόφωνων του οποίου τραβήξαμε δεδομένα χωρίς να χρειάζεται δημιουργία λογαριασμού.

Σχήμα 3.7: Το φόρουμ Hidden Answers

3.4.3 CryptBB

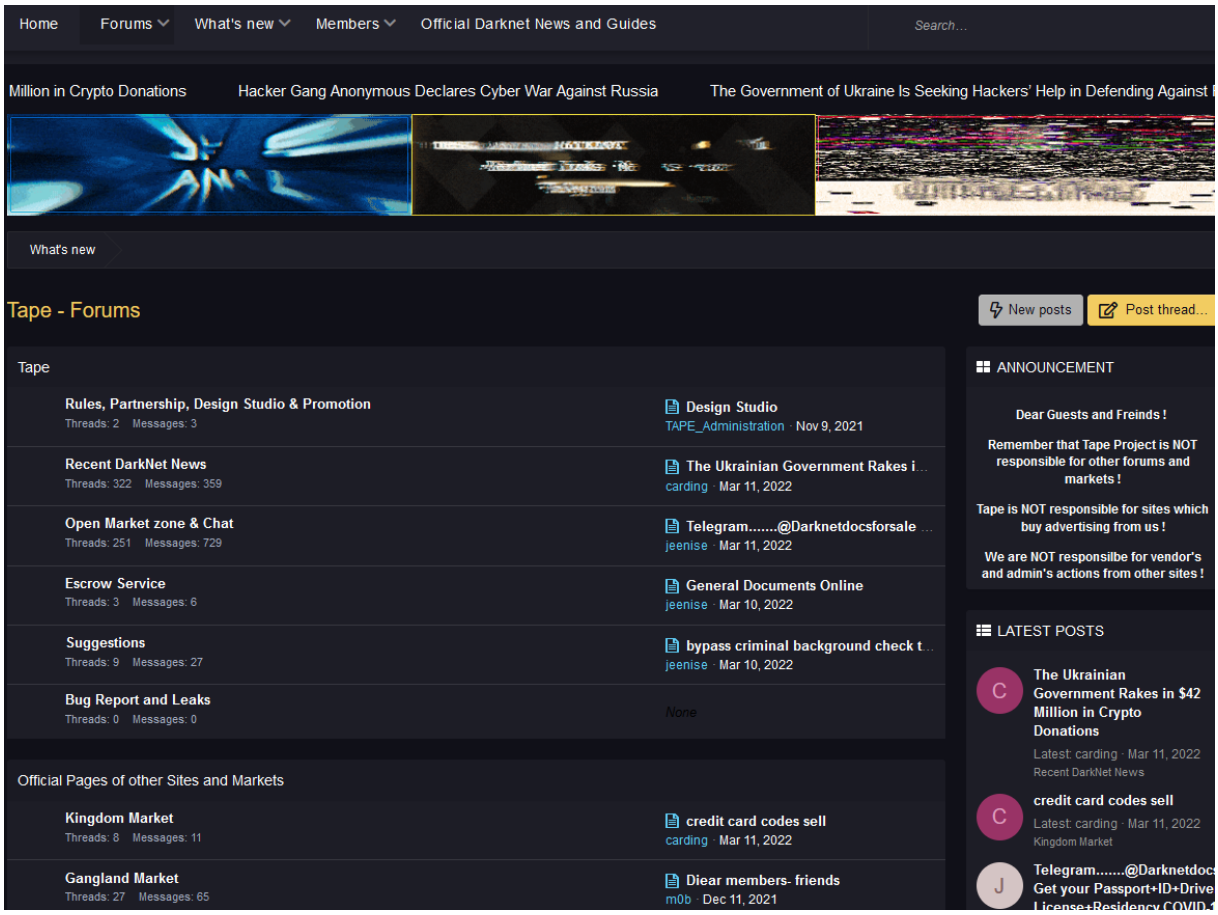
Το CryptBB (βλ. Σχήμα 3.8) είναι ένα καθαρά χάκινγκ φόρουμ με θέματα όπως προγραμματισμός, software, hardware και υποκλοπή καρτών. Αυτά είναι τα μοναδικά πεδία-υποφόρουμ που βλέπει ο μέσος χρήστης-επισκέπτης. Για να ξεκλειδωθούν τα υπόλοιπα υποφόρουμ όπως μαθήματα χάκινγκ ο χρήστης πρέπει να περάσει από μερικές αποστολές κρυπτογράφησης αποκρυπτογράφησης αρχείων έτσι ώστε να τον εμπιστευθεί ο διαχειριστής ότι πρόκειται για άτομο που όντως ενδιαφέρεται για το χώρο του hacking και ότι δεν είναι κάποιος απλός θεατής (bystander).



Σχήμα 3.8: Το φόρουμ CryptBB

3.4.4 Tape

Το Tape (Σχήμα 3.9) είναι ένα μικρό φόρουμ που έχει κύριο σκοπό την προώθηση ναρκωτικών, οδηγίες παρασκευής αυτών και γενικά είναι πιο πολύ online marketplace και λιγότερο φόρουμ, παρόλαυτα υπάρχουν σχετικά ποστ με θέματα που μόλις αναφέραμε. Επίσης λειτουργεί και σαν γέφυρα για μεγαλύτερα φόρουμ και διαφημίζει γνωστούς προμηθευτές του σκοτεινού ιστού.

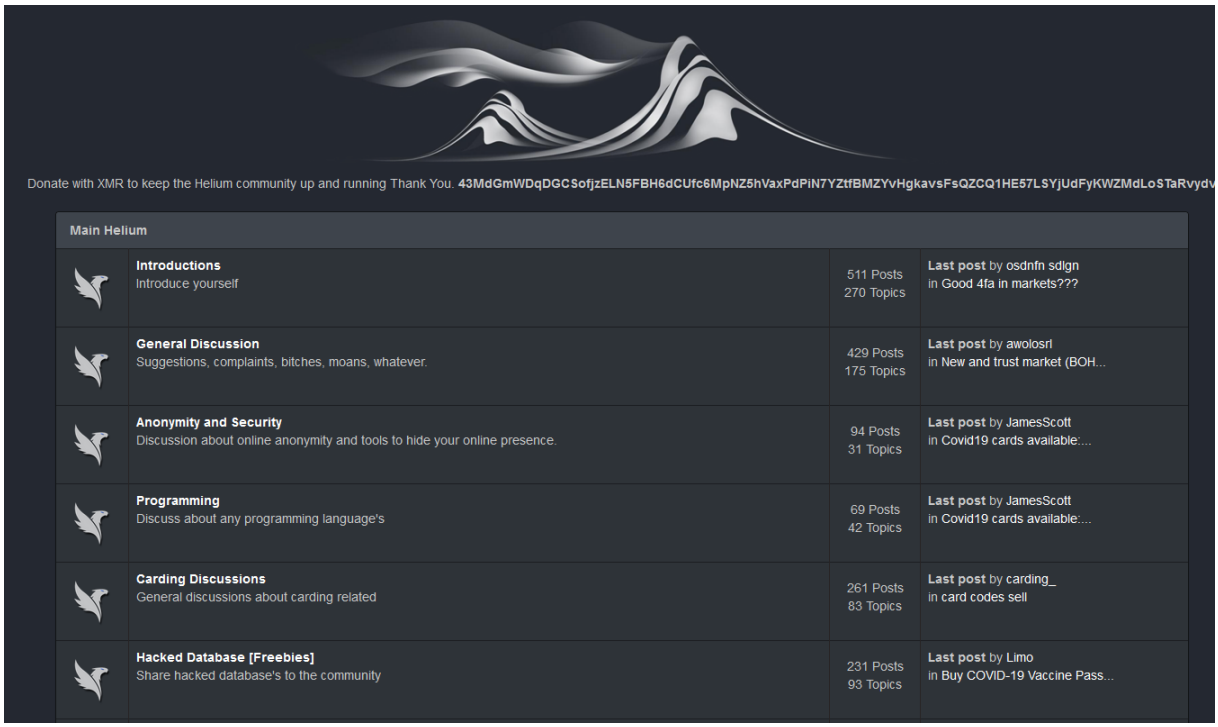


Σχήμα 3.9: Το φόρουμ Tape

3.4.5 Helium

Το πέμπτο και τελευταίο φόρουμ που τραβήξαμε δεδομένα, είναι το Helium (Σχήμα 3.10 το οποίο είναι και αυτό σχετικά μικρό και καινούργιο και κατά κόρον αφορά θέματα όπως χάκινγκ, ασφάλεια, ανωνυμία κ.α. Εκτός από φόρουμ, ο διαχειριστής της σελίδας ενημερώνει με αυτόματο μήνυμα κάθε νέο χρήστη, ότι είναι ελεύθερος να αγοράσει και να ανταλάξει οτιδήποτε επιθυμεί με κάποιον άλλον χρήστη, είτε αυτό είναι υπηρεσία χάκινγκ είτε κάποιο κλεμμένο αντικείμενο ή κωδικός. Όμως υπογραμμίζει ότι απαγορεύεται η διανομή, αγορά, ανταλλαγή υλικού που έχει να κάνει με παιδική πορνογραφία. Όλες οι συναλλαγές γίνονται μέσω της υπηρεσίας escrow.[7]

[7] https://www.escrow.com/



Σχήμα 3.10: Το φόρουμ Helium

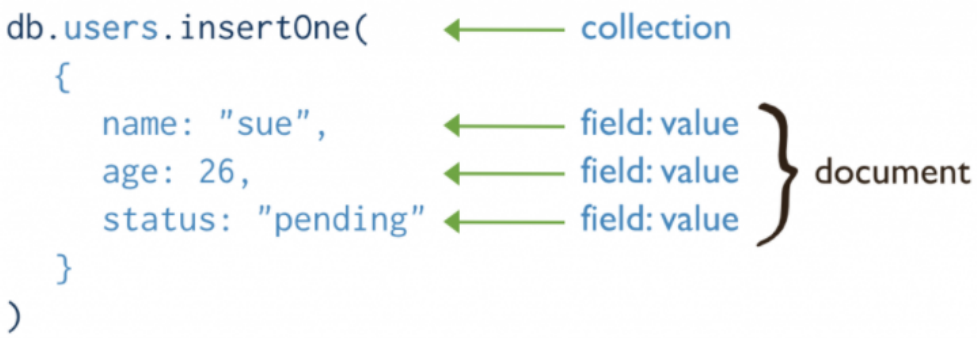
3.4.6 Γενικά εμπόδια και δυσκολίες

Στο σημείο αυτό να αναφέρουμε μερικά σημεία κλειδιά της όλης διαδικασίας εύρεσης υλικού στο σκοτεινό ιστό για όποιον ενδιαφέρεται μελλοντικά. Αν εξαιρέσουμε το Dread, το οποίο όπως είπαμε είναι τεράστιο και προωθείται-διαφημίζεται από παντού, συναντήσαμε και άλλα φόρουμ εξίσου μεγάλα και γνωστά, που ήταν πιο πολύ κοινωνικά δίκτυα όπως το facebook και το twitter παρά φόρουμ. Εκεί θα βρίσκαμε τεράστιο υλικό και ίσως πιο ενδιαφέροντα θέματα συζήτησης από αυτά που εν τέλει βρήκαμε. Όμως το 99% αυτών απαιτούσαν καταβολή χρηματικού αντίτιμου για να αποκτήσεις πρόσβαση, πράγμα το οποίο ξέφευγε από τα πλαίσια της πτυχιακής μας και ίσως να αφορούσε περισσότερο ερευνητές παράνομης δραστηριότητας, αντιτρομοκρατικές υπηρεσίες ή/και αστυνομία. Επίσης κάποια από τα φόρουμ και κοινωνικά δίκτυα που βρήκαμε ύστερα από αρκετές εβδομάδες αναζήτησης, αφορούσαν θέματα που ναι μεν είχαμε πρόσβαση και μπορούσαμε να τραβήξουμε αλλά δεν θα μπορούσαμε σε καμία περίπτωση να αναλύσουμε και να παρουσιάσουμε στα πλαίσια μιας εργασίας για ευνόητους λόγους. Μερικά από τα φόρουμ αυτά είναι τα παρακάτω:

1. The HUB (επί πληρωμή)
2. HackTown (επί πληρωμή)
3. Guanji (κοινωνικό δίκτυο που αναρτείς ελεύθερα)
4. Cebulka (επί πληρωμή)

3.4.7 MongoDB

Με σκοπό την αποθήκευση των δεδομένων που συλλέχθηκαν με τη χρήση των παραπάνω εργαλείων, χρησιμοποιήθηκε η MongoDB βάση δεδομένων.[8] Η MongoDB αποτελεί μία NoSQL βάση δεδομένων εγγράφων [9] και προσφέρει μία μεγάλη ποικιλία υπηρεσιών, που εξυπηρετούν τόσο την αποθήκευση και τη διαχείριση των δεδομένων όσο και την αναπαράστασή τους. Ο όρος 'βάση δεδομένων εγγράφων' αντικατοπτρίζει την ιδιότητα της βάσης να αποθηκεύει τα δεδομένων σε μορφή που μοιάζει με ένα JSON έγγραφο. Όσον αφορά στη MongoDB, κάθε εγγραφή αντιπροσωπεύει ένα έγγραφο, δηλαδή μια δομή δεδομένων που απαρτίζεται από ζεύγη κλειδιού-τιμής. Τα έγγραφα της είναι πανομοιότυπα με τα αντικείμενα JSON. Οι τιμές των πεδίων μπορεί να περιλαμβάνουν άλλα έγγραφα, πίνακες ή ακόμα και πίνακες εγγράφων. Οι εγγραφές δεδομένων που αποθηκεύονται ως έγγραφα συγκροτούν ένα ευρύτερο σύνολο δεδομένων, το οποίο ονομάζεται συλλογή (Collection). Μία βάση δεδομένων συνίσταται από μία ή περισσότερες συλλογές εγγράφων. Στο Σχήμα 3.11 παρουσιάζεται η τυπική μορφή ενός εγγράφου MongoDB.



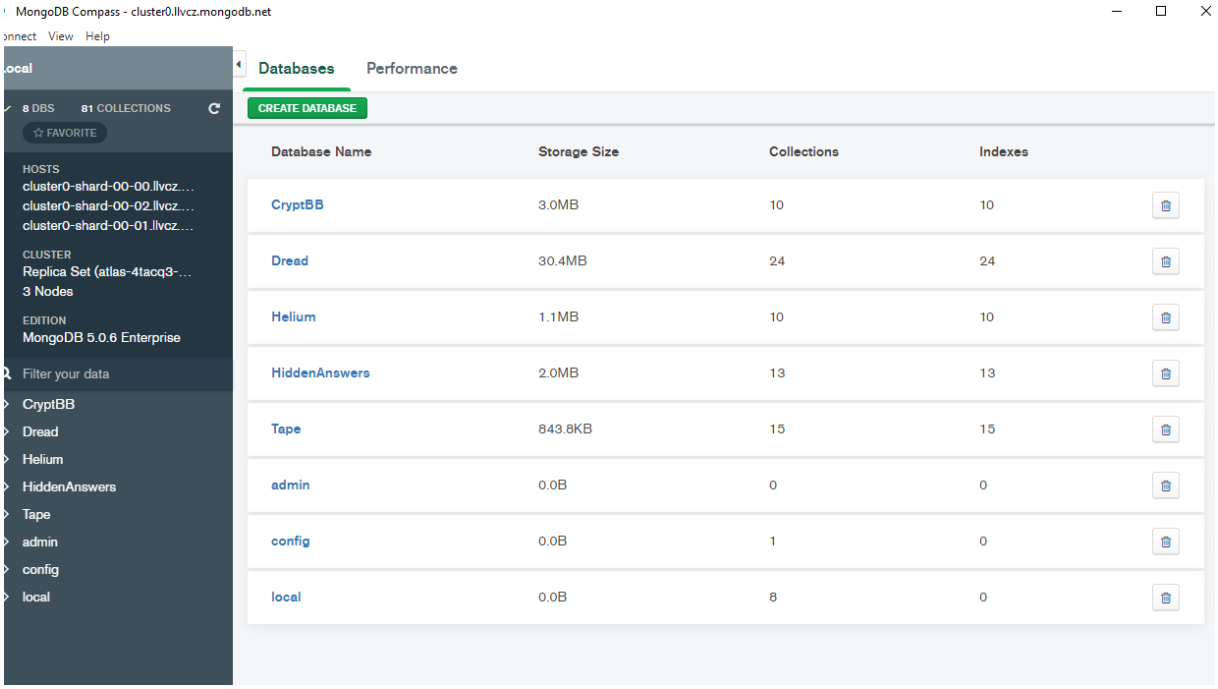
Σχήμα 3.11: Μορφή εγγράφου MongoDB

MongoDB Compass

Το MongoDB Compass [10] είναι το γραφικό περιβάλλον διεπαφής χρήστη (GUI) που παρέχεται από την MongoDB. Το συγκεκριμένο εργαλείο επιτρέπει στον χρήστη να επεξεργάζεται, να αναλύει και να εκτελεί στα δεδομένα της βάσης του, χωρίς να κρίνεται απαραίτητη η γνώση της σύνταξης MongoDB ερωτημάτων. Το MongoDB Compass ενσωματώνει λειτουργίες μέσω των οποίων καθίστανται δυνατά η βελτιστοποίηση της απόδοσης των ερωτημάτων, η διαχείριση των ευρετηρίων, καθώς και ο ορισμός ενός σχήματος (schema) για τα υπό αποθήκευση έγγραφα.

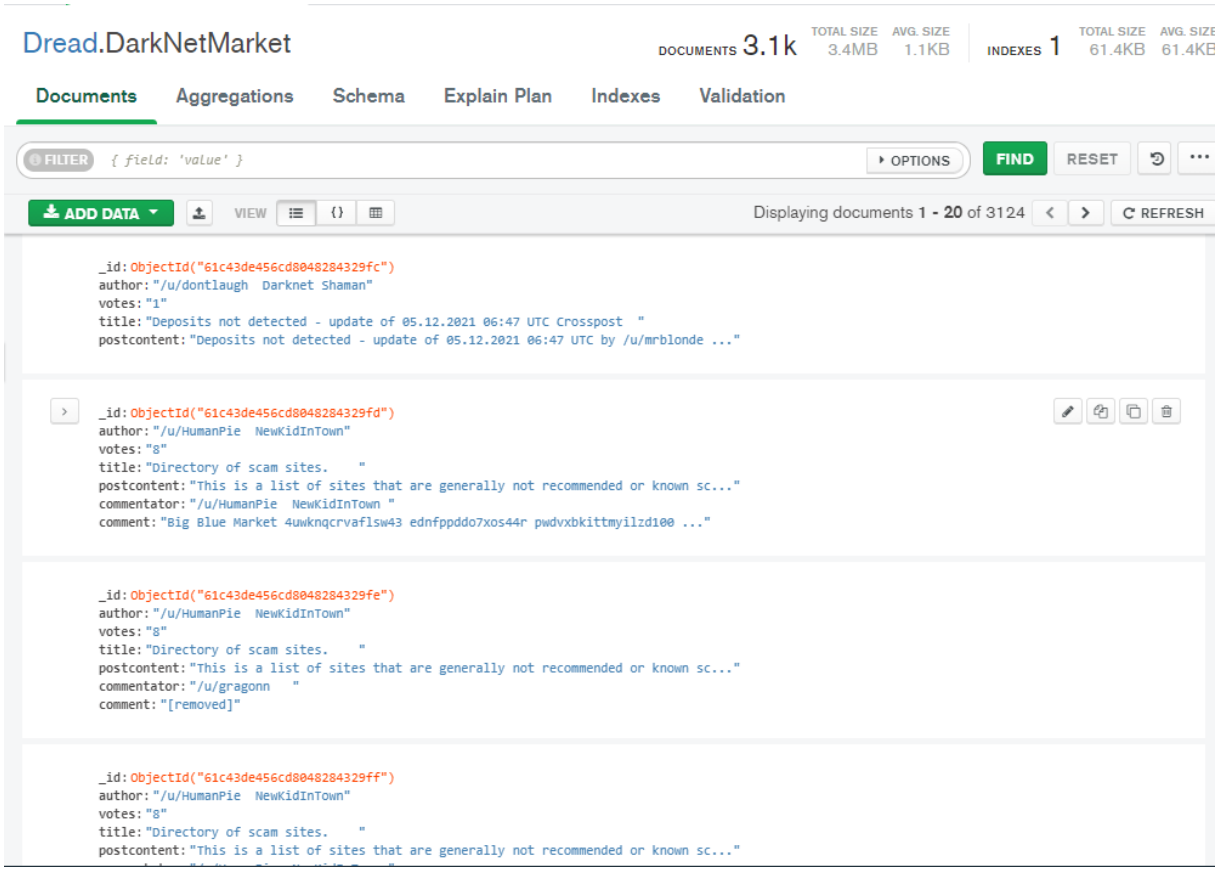
Στο Σχήμα 3.12 βλέπουμε πώς είναι οι συλλογές μας, δηλαδή όλα τα αρχεία json που φτιάξαμε από όλα τα φόρουμ.

[8] https://www.mongodb.com/
[9] https://www.mongodb.com/nosql-explained
[10] https://www.mongodb.com/products/compass



Σχήμα 3.12: Οι συλλογές μας στο MongoDB

Στο Σχήμα 3.13 βλέπουμε ένα παράδειγμα εγγράφου σε json από τα δεδομένα που συλλέξαμε.



Σχήμα 3.13: Παράδειγμα σε json από υποφόρουμ του dread στο MongoDB

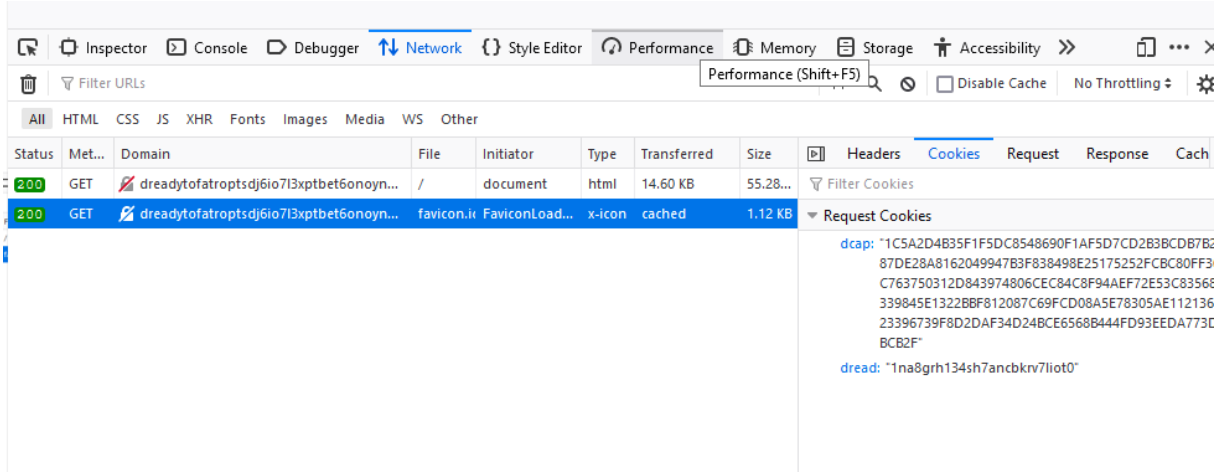
Κεφάλαιο 4

Ανάλυση κώδικα συγκομιδής δεδομένων

Στα προηγούμενα κεφάλαια μιλήσαμε για τον σκοτεινό ιστό σαν σύνολο και τις μάυρες αγορές, αναλύσαμε τα εργαλεία και τις βιβλιοθήκες που χρειαζόμαστε, καθώς και τα τελικά φόρουμ/κοινωνικά δίκτυα που συναντήσαμε. Στο παρόν κεφάλαιο θα δείξουμε πώς προσεγγίσαμε προγραμματιστικά την εργασία για να αντλήσουμε τα δεδομένα από αυτά τα πέντε φόρουμ.

4.1 Δημιουργία σύνδεσης

Η διαδικασία είναι η ίδια για όλα τα φόρουμ. Αρχικά φτιάχνουμε ένα απλό proxy με τη χρήση του πρωτόκολλου **socks5** και της βιβλιοθήκης **requests** [16] για να στείλουμε αίτημα για σύνδεση στο TOR το οποίο δέχεται αιτήματα μόνο στη θύρα 9150. Αυτό γίνεται διότι όπως είπαμε οι onion ιστότοποι ανοίγουν μόνο μέσω του TOR. Όπως αναφέραμε στην Υποενότητα 3.4 για να τραβήξουμε δεδομένα από τα φόρουμ, δεν μπορούμε απλά να στείλουμε ένα αίτημα. Χρειάζεται να συμπεριλάβουμε στο κώδικα μας τα cookies που μας δίνει η σελίδα, τα οποία είναι διαφορετικά κάθε φορά που συνδεόμαστε. Στο Σχήμα 4.1 φαίνεται ο τρόπος που παίρνουμε αυτή τη πληροφορία (καρτέλα network στο πρόγραμμα περιήγησης) και ύστερα την βάζουμε στο κώδικα μας όπως φαίνεται παρακάτω, στη μεταβλητή cookies.



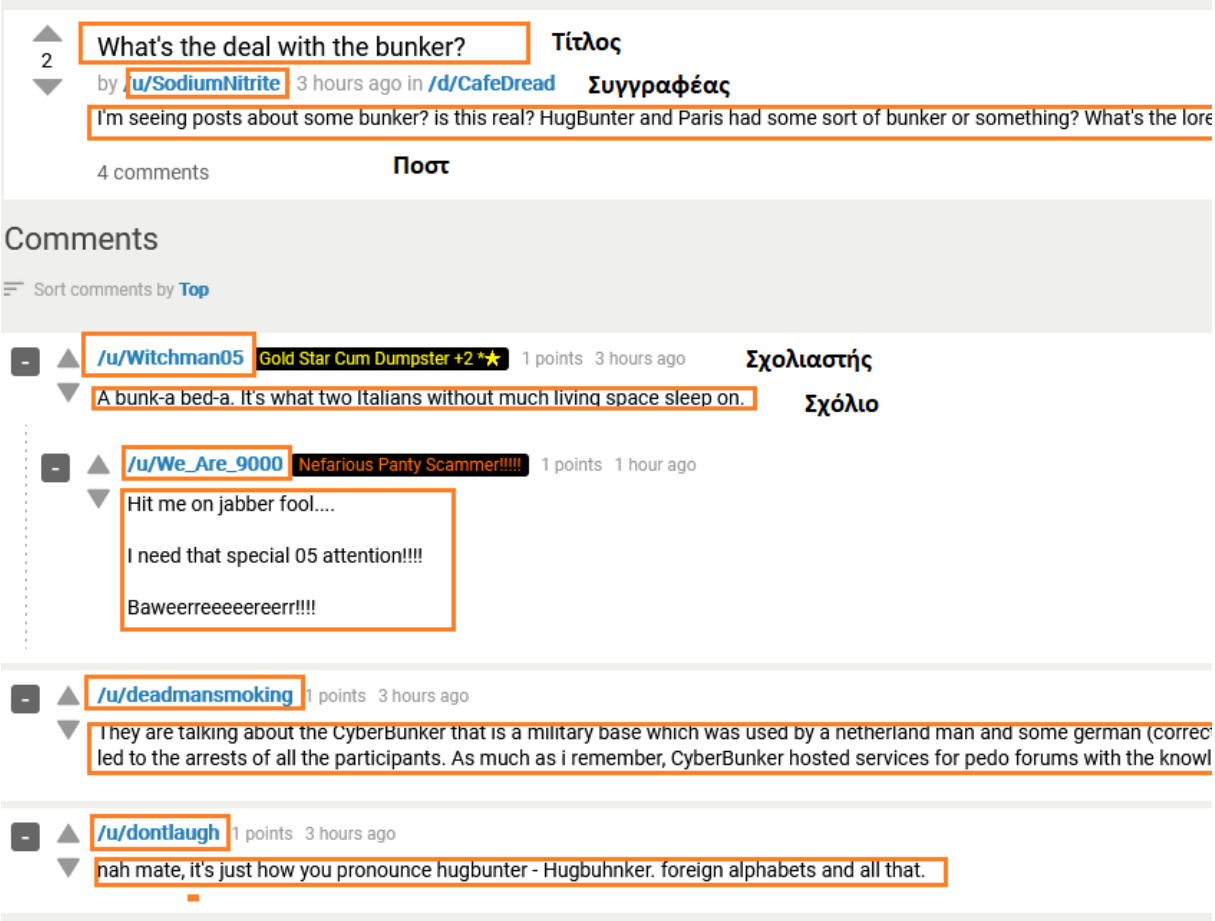
Σχήμα 4.1: Cookies του φόρουμ

Παρακάτω βλέπουμε τον κώδικα που γίνεται ακριβώς αυτό που περιγράφουμε. Αρχικά δημιουργείται ένα proxy με τη θύρα 9150 και μετά γίνεται αρχικοποίηση των μεταβλητών με τις πληροφορίες που πήραμε από το πρόγραμμα περιήγησης. Η μεταβλητή **headers** περιέχει πληροφορίες που επίσης χρειάζονται για το αίτημα και μπορούμε έυκολα να επεξεργαστούμε για να αποφύγουμε πιθανούς αποκλεισμούς. Για παράδειγμα μπορούμε να φτιάξουμε μια λίστα με ψεύτικους user agents [1] και να τους εναλλάσουμε ανά τακτά διαστήματα για να φαίνεται ότι στέλνουμε αιτήματα από διαφορετικά συστήματα στο server. Στη προκειμένη περίπτωση χρησιμοποιούμε απλώς αυτόν που μας παρέχει το TOR (firefox 91.0).

```
session = requests.session()
session.proxies = {'http':  'socks5h://localhost:9150',
                   'https': 'socks5h://localhost:9150'}


cookies = {
    'dcap': '1C5A2DFBB9E....2C577DB8301D8D66D151AB048CBC2DEF91A7819E1AC83C',
    'dread': '1vnu4bkklk9s0uulk5vgc0kvgb',
}

headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; rv:91.0) Gecko/20100101 Firefox/91.0',
    'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
    'Accept-Language': 'en-US,en;q=0.5',
    'Referer':
        'http://dreadytofatroptsdj6io7l3xptbet6onoyno2yv7jicoxknyazubrad.onion/auth/login',
    'Connection': 'keep-alive',
    'Upgrade-Insecure-Requests': '1',
    'Sec-Fetch-Dest': 'document',
    'Sec-Fetch-Mode': 'navigate',
    'Sec-Fetch-Site': 'same-origin',
    'Sec-Fetch-User': '?1',
    'Cache-Control': 'max-age=0',
}
```

[1] https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/User-Agent

Στο σημείο αυτό να υπενθυμίσουμε (Υποενότητα 1.1) πως τα δεδομένα που θέλουμε να αντλήσουμε είναι ο τίτλος, το όνομα του συγγραφέα, το κείμενο (post), το όνομα των σχολιαστών και τα σχόλια αυτών. Έτσι, δημιουργείται μια σχέση μεταξύ του συγγραφέα και των σχολιαστών του, κόμβοι και ακμές (Υποενότητα 2.1.2). Ολά αυτά επισημαίνονται στο Σχήμα 4.2.



Σχήμα 4.2: Παράδειγμα ενός post

Βέβαια, σε ορισμένα ποστ, ειδικά στα μικρότερα φόρουμ και υποφόρουμ, δεν υπάρχουν σχολιαστές οπότε δεν δημιουργείται καμία σχέση μεταξύ των χρηστών για τη συγκεκριμένη δημοσίευση.

4.2 Πρώτα βήματα crawling

Αφού δημιουργήσαμε τη σύνδεση και παρείχαμε στο πρόγραμμα τα κατάλληλα διαπιστευτήρια, είμαστε έτοιμοι να δώσουμε μέσω της βιβλιοθήκης **Beautiful Soup** το url του φόρουμ και να του πούμε που να ψάξει και τι, αφού προτίστως έχουμε αποθηκεύσει όλον τον πηγαίο κώδικα (content) της σελίδας (μεταβλητή **soup**), όπως φαίνεται στο παρακάτω τμήμα κώδικα.

```
def posts_dread():

    file =open('posts_replies_of_dread.json','w',encoding='UTF-8')
```

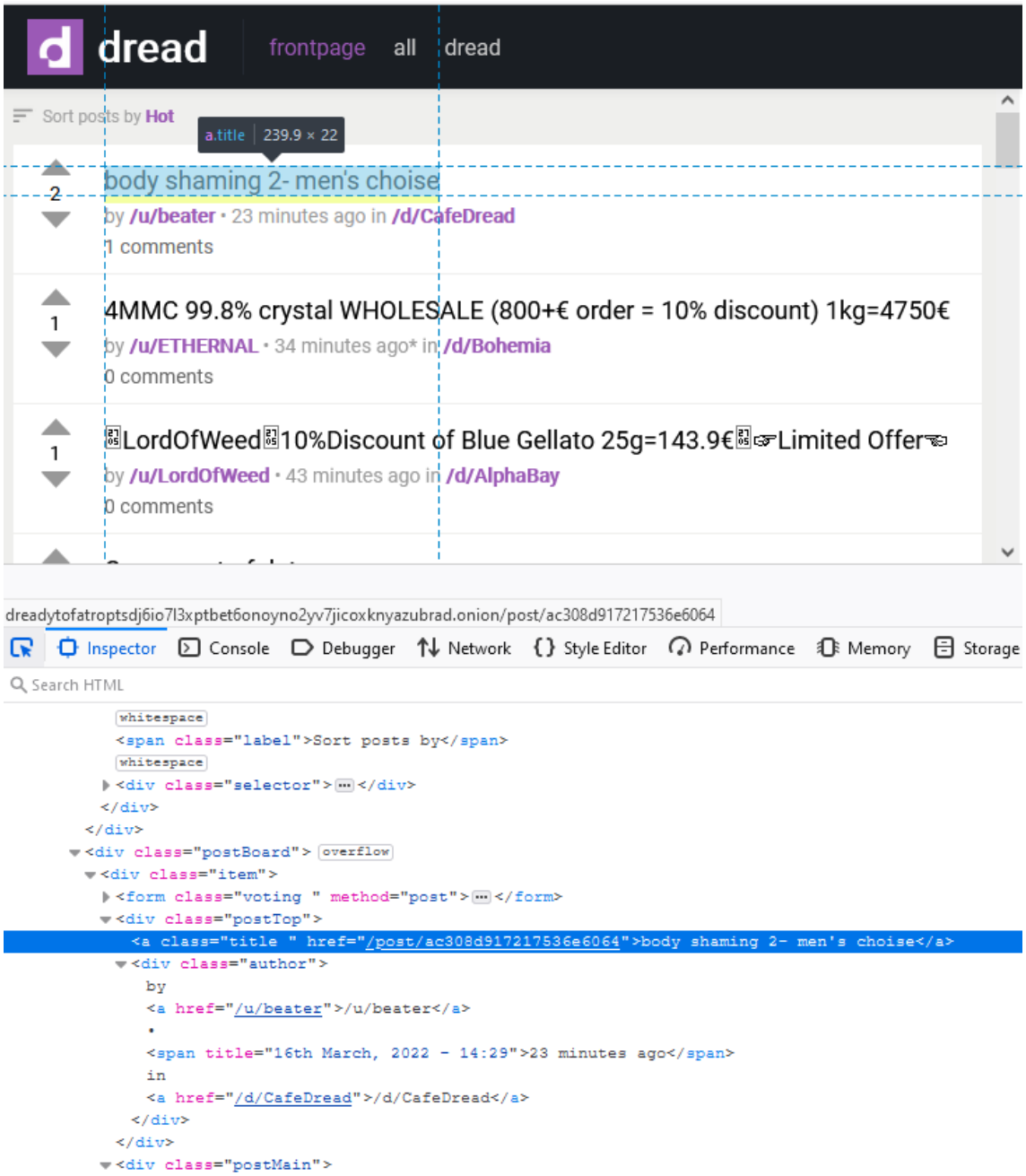
```
    j=0
    links=[]
    for i in range(12):
        r=session.get(f'http://dreadytofatroptsdj6io7l3xptbet6onoyno2yv7jicoxknyazubrad.onion/d/Dread?p=
        {i+1}',headers=headers,cookies=cookies,verify=False)

        soup=BeautifulSoup(r.content,'lxml')

        for href_get in soup.find_all(class_='title',href=True):


            links.append(href_get['href'])
}
```

Για κάθε υποφόρουμ, ακολουθούμε την ίδια διαδικασία φτιάχνοντας μια συνάρτηση με το όνομα κάθε υποφόρουμ και μέσα σε αυτή, δημιουργούμε ένα αρχείο json για να βάλουμε τα δεδομένα που τραβάμε. Ύστερα, λέμε στον crawler πόσες φόρες θέλουμε να επισκεφθεί το url που του παρέχουμε. Στην προκειμένη περίπτωση είναι το νούμερο 12 αλλά αυτό ποικίλλει ανάλογα το μέγεθος του υποφόρουμ. Ο αριθμός αυτός ουσιαστικά είναι επιλεγμένος από εμάς για να πούμε στον crawler να μαζέψει δεδομένα (posts) μόνο του τελευταίου 3μήνου.

Η γραμμή

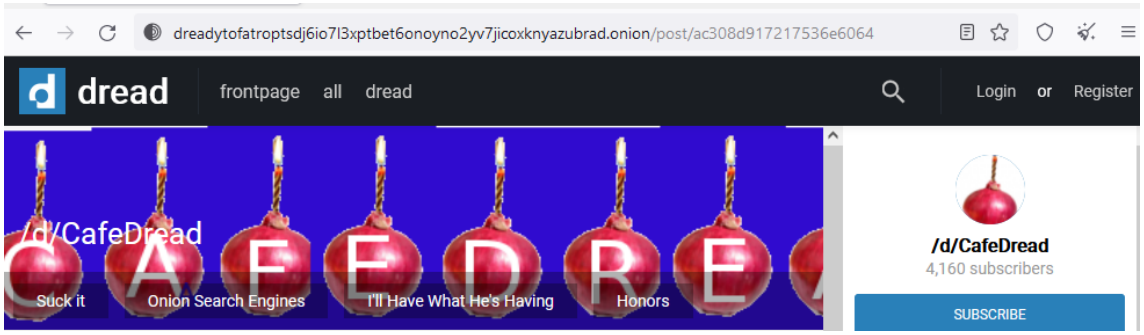
```
for href_get in soup.find_all(class_='title',href=True)
```

λέει στον μηχανισμό απόξεσης δεδομένων, μέσω του πηγαίου κώδικα της σελίδας που προηγουμένως έχουμε μελετήσει, να βρεί όλες τις κλάσεις που έχουν όνομα **"title"** και στη συνέχεια να αποθηκεύσει τα ονόματα των υπερσυνδέσμων αυτών των κλάσεων σε έναν πίνακα αλφαριθμητικών. Ο πηγαίος κώδικας και το παράδειγμα μιας HTML σελίδας που ανακτήθηκε, φαίνονται στο Σχήμα 4.3 και καθιστούν σαφές τον μηχανισμό που μόλις περιγράψαμε.



Σχήμα 4.3: Πηγαίος κώδικας και επιθεώρηση στοιχείου στην HTML

Πατώντας δεξί κλικ πάνω σε οποιοδήποτε στοιχείο, στη προκειμένη περίπτωση, σε έναν τίτλο, παίρνουμε το όνομα που έχει η κλάση στο οποίο ανήκει. Αυτή η διαδικασία γίνεται διότι στον κώδικα **HTML**, και συγκεκριμένα στο tag **"href"** βρίσκεται το κομμάτι του url κάθε δημοσίευσης που θα χρειαστούμε στη συνέχεια για να πούμε στον crawler ποια σελίδα ακριβώς να επισκευθεί. Παράδειγμα διεύθυνσης url μιας δημοσίευσης φαίνεται στο Σχήμα 4.4, όπου οι χαρακτήρες μετά το .onion είναι αυτοί που πήραμε από τη κλάση title και την ετικέτα (tag) **href**.



Σχήμα 4.4: Παράδειγμα διεύθυνσης μιας δημοσίευσης

Στο παρακάτω τμήμα κώδικα, κάνουμε αυτό που περιγράψαμε. Έχοντας αποθηκεύσει όλα τα τμήματα των urls κάθε δημοσίευσης στο πίνακα **links** κάνουμε μια συνένωση αλφαριθμητικών, η μια πλευρά περιέχει το base url το οποίο είναι το αρχικό url του ιστότοπου που επισκεπτόμαστε, και από την άλλη πλευρά είναι το post url το οποίο πήραμε από τον πίνακα links. Δημιουργώντας μια επαναληπτική διαδικασία, ο crawler επισκέπτεται ένα-ένα τα urls που ενώνονται κάθε φορά μέχρις ότου να φτάσουμε στο τελικό στοιχείο του πίνακα **links**, δηλαδή στη τελευταία δημοσίευση της σελίδας. Υπενθυμίζουμε ότι η διαδικασία αυτή είναι εμφωλευμένη με την παραπάνω, που ορίσαμε ως αριθμό σελίδων του φόρουμ προς επίσκεψη, τις 12.

```
table=[]
title1 =[]
post1=[]
authors1=[]
names1=[]
comments1=[]
for i in range(len(links)):
        baseurl='http://dreadytofatroptsdj6io7l3xptbet6onoyno2yv7jicoxknyazubrad.onion'
        posturl=links[i]
        fullurl=urljoin(baseurl,posturl)
        r=session.get(fullurl,headers=headers,cookies=cookies)

        soup=BeautifulSoup(r.text.replace("\n","").replace("\r",""),'html.parser')

        title =soup.find('a',{'class' : 'title'}) .text
        title1 .append(title)
        author=soup.find('div',{'class' : 'author'})
        name=author.find('a').text
        votecount=soup.find('div',{'class' : 'voteCount'}).text

        content=soup.find('div',{"class" :"postContent"}).text.strip()
        commentsection = soup.find('div',{"class" :"postComments"})
```

Κατόπιν, με τη βοήθεια ξανά της **Beautiful Soup**, αποθηκεύουμε σε μεταβλητές όλα όσα χρειαζόμαστε από το πηγαίο κώδικα της σελίδας, δηλαδή το όνομα του συγγραφέα της δημοσίευσης, τη δημοσίευση, τους σχολιαστές και τα σχόλια αυτών. Τα ονόματα των κλάσεων τα πήραμε με τον ίδιο τρόπο που δείχνουμε μέσω του Σχήματος 4.3.

```
        if (soup.find('div',{"class":"postComments"})):
            names=commentsection.find_all('a',{"class":"username"})
```

```
            comments=soup.find_all('div',{"class" : "commentBody"})
            if names!=None or comments!=None:

                for j in range(len(names)):
                    comm={
                        'author':name,
                        'votes':votecount,
                        ' title ': title ,
                        'postcontent':content,
                        'commentator':names[j].text,
                        'comment':comments[j].text,
                    }

                    json.dump(comm,file,ensure_ascii=False)

        else:
            comm={
                'author':name,
                'votes':votecount,
                ' title ': title ,
                'postcontent':content
            }
            json.dump(comm,file,ensure_ascii=False)
```

Τέλος, ελέγχουμε εάν η κάθε δημοσίευση περιέχει σχολιαστές και σχόλια με το να δώσουμε εντολή στον crawler του Beautiful Soup μέσω του πηγαίου κώδικα να ελέγξει αν στη σελίδα περιέχεται η κλάση **postComments** στην οποία συνδέονται οι κλάσεις **username** για τον κάθε σχολιαστή και **commentBody** για το σχόλιο αυτού. Εάν υπάρχει, δημιουργούμε για όσους σχολιαστές βρούμε (μεταβλητή len(names)) ένα Python dictionary με ζευγάρια κλειδιά-τιμές έτσι ώστε να είμαστε όσο κοντά γίνεται στη μορφή που θα χρειαστούμε μετέπειτα, την JSON. Συγκεκριμένα δημιουργούμε ένα dictionary of dictionaries[2] όπου θα έχει τη μορφή που φαίνεται στο Σχήμα 3.13. Εάν δεν υπάρχουν σχολιαστές άρα και σχόλια στη δημοσίευση, δημιουργούμε ένα απλό dictionary με τη μορφή που φαίνεται στο Σχήμα 4.5.

```
_id: ObjectId("61c442bd56cd804828435a12")
title: "  Looking For Valid Bank Logs  "
author: "chase2244"
body: " Hello guys, can someone please share some tested and trusted recommen..."
```

Σχήμα 4.5: Παράδειγμα ενός αντικειμένου json χωρίς σχολιαστές

4.3 Τελικά βήματα

Αφού έχουμε μαζέψει με αυτό το τρόπο όλα τα δεδομένα από όλα τα φόρουμ, πρέπει να τα αποθηκεύσουμε στη MongoDB (Υποενότητα 3.4.7). Όμως, οι μορφές αρχείων που είναι δεκτές είναι είτε json είτε csv. Θα χρειαστεί να κάνουμε έναν τελευταίο μετασχηματισμό των αρχείων που δημιουργήσαμε, δηλαδή των dictionaries, έτσι ώστε να έχουν την τυπική μορφή json δηλαδή να ξεχωρίζουν τα αντικείμενα μεταξύ τους με

[2] https://linuxhint.com/python$_d ictionary_o f_d ictionaries/$

κόμμα. Αυτό πραγματοποιείται με το παρακάτω τμήμα του κώδικα το οποίο τρέξαμε για όλα τα αρχεία από όλα τα υποφόρουμ.

```
 f = open('dread/posts_replies_of_whitehouse2.json','r',encoding='UTF-8')
filedata = f.read()
f.close()

newdata = filedata.replace("}{","} , {")

f = open('dread/posts_replies_of_whitehouse2.json','w',encoding='UTF-8')
f.write(newdata)
f.close()
```

Η μεταβλητή **newdata** θα περιέχει τη νέα μορφή αρχείου από όπου αντικαταστήσαμε τα άγκυστρα του dictionary που φτιάξαμε, με άγκυστρα και κόμμα που χρειάζεται για να γίνει ένα έγκυρο json αρχείο, έτοιμο να αποθηκευτεί στη MongoDB.

Κεφάλαιο 5

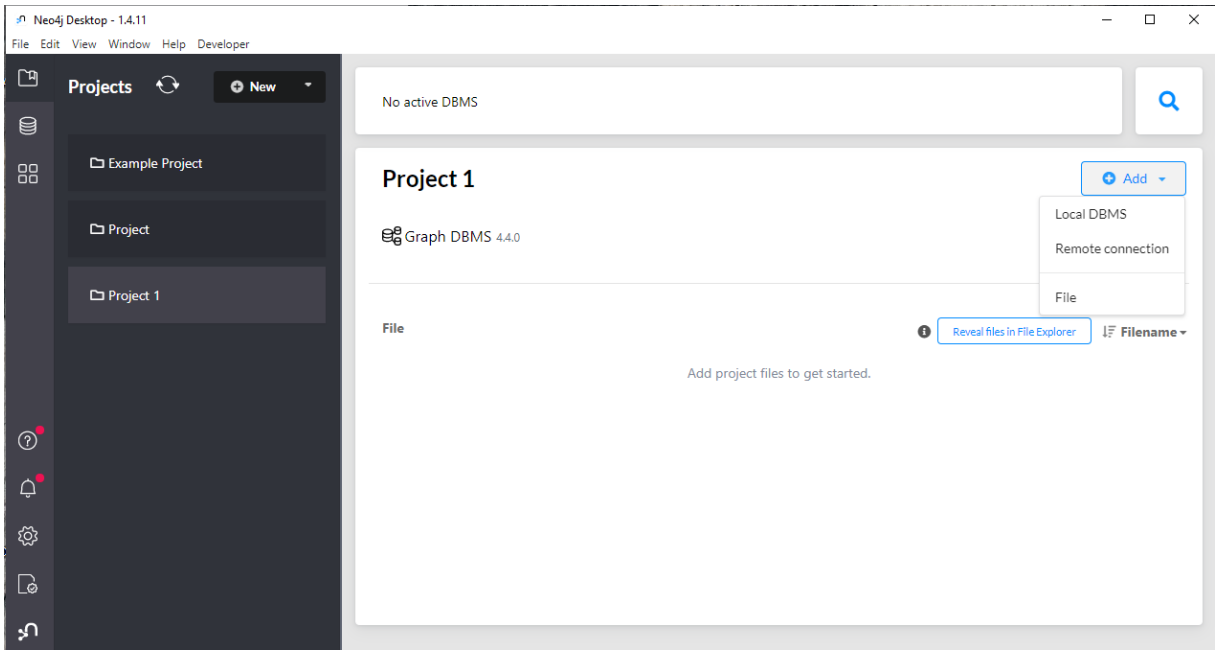
Μετρικές κοινωνικών δικτύων και αποτελέσματα

Στο Κεφάλαιο 4, παρουσιάσαμε και αναλύσαμε τον κώδικα που χρησιμοποιήσαμε για την συλλογή των δεδομένων και την αποθήκευση στη βάση δεδομένων. Το πιο σημαντικό κομμάτι αυτής της εργασίας, παρουσιάζεται στο παρόν κεφάλαιο όπου δείχνουμε πως χρησιμοποιήθηκε το πρόγραμμα Neo4j και όλες τις οπτικοποιήσεις. Πήραμε ένα ένα τα αρχεία των υποφόρουμ και δημιουργήσαμε γράφους, υπολογίσαμε μετρικές όπως κεντρικότητα με βάση το βαθμό, την διαμεσολαβικότητα, το ιδιοδιάνυσμα και την εγκύτητα. Τέλος, βρήκαμε πόσες σχέσεις δημιουργούνται μεταξύ των κόμβων και πόσες μικρές κοινότητες (communities) δημιουργούνται μέσα σε αυτές τις σχέσεις, χρησιμοποιώντας τον αλγόριθμο Louvain[1].

5.1 Το πρόγραμμα Neo4j

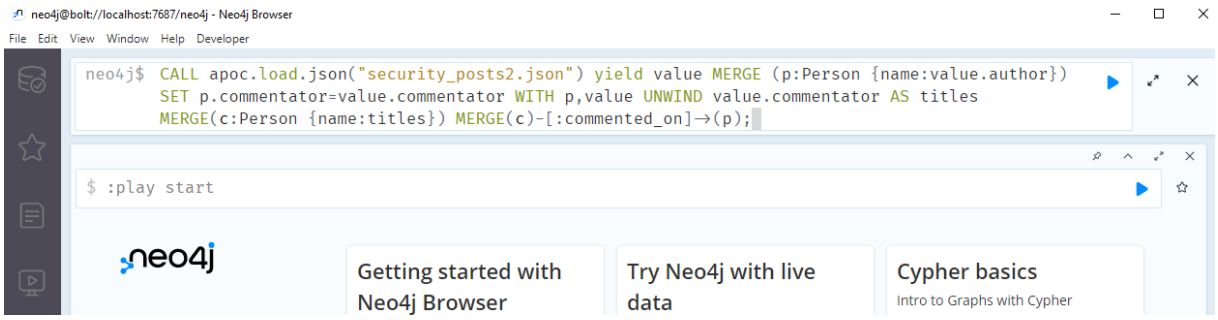
Έχοντας αναφέρει τους λόγους που καταλήξαμε στο Neo4j στην Ενότητα 2.1.6, ξεκινάμε δείχνοντας πως συνδεόμαστε στο πρόγραμμα, πώς ανεβάζουμε τα αρχεία και πώς μέσω ενός Cypher query (βλ. Σχήμα 2.5) παίρνουμε όλα τα τελικά αποτελέσματα που επιθυμούμε. Αρχικά, δημιουργούμε ένα project και μια βάση τοπικά στον υπολογιστή μας, για να αποθηκεύσουμε τα αρχεία json και όλα τα αποτελέσματα που θα παραχθούν, όπως φαίνεται στο Σχήμα 5.1.

[1] https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/



Σχήμα 5.1: Έναρξη του Neo4j

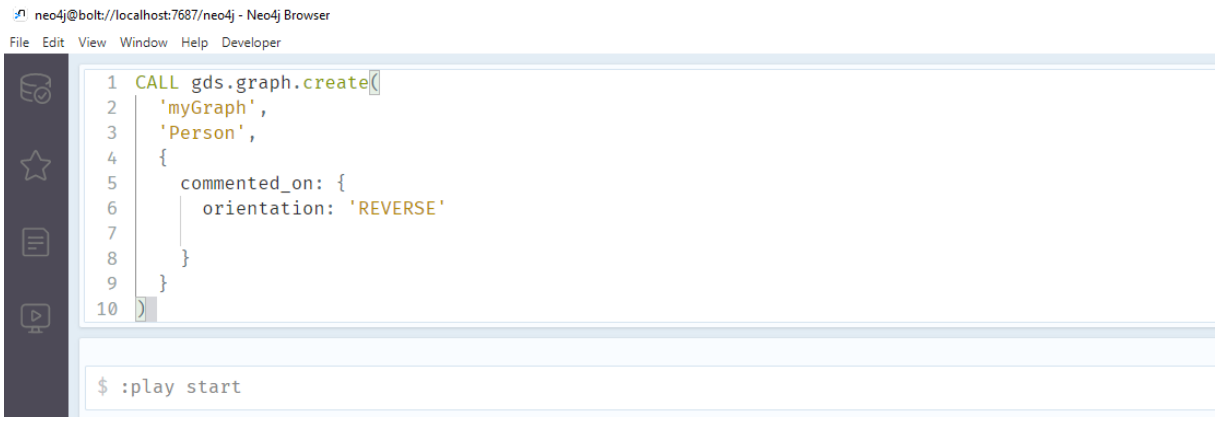
Αφού γίνει αυτό, ξεκινάμε τη βάση μας και είμαστε έτοιμοι να τρέξουμε το πρώτο query, δηλαδή να φορτώσουμε ένα αρχείο json. Το Neo4j μας παρέχει δύο τρόπους για την οπτικοποίηση των δεδομένων, το Neo4j bloom[2] και το Neo4j browser[3]. Αν και δεν υπάρχουν σοβαρές διαφορές στα δύο εργαλεία, προτιμήσαμε το Neo4j Browser επειδή οι γράφοι φαίνονταν καθαρότεροι από το Bloom και όχι συμπυκνωμένοι, όταν οι γράφοι είχαν πάνω από έναν ορισμένο αριθμό κόμβων. Η πρώτη εντολή που τρέχουμε για να φορτώσουμε τα δεδομένα και να δημιουργήσουμε σχέσεις μεταξύ συγγραφέα και σχολιαστών φαίνεται στο Σχήμα 5.2.



Σχήμα 5.2: Δημιουργία σχέσεων με cypher

Αφού έγινε το πρώτο βήμα με επιτυχία, τώρα πρέπει να φτιάξουμε έναν εικονικό γράφο στο Cypher που να συνδέεται με το αρχείο που ανεβάσαμε, για να μπορέσουμε πάνω σε αυτόν να τρέξουμε τις μετρικές. Το query για την δημιουργία γράφου φαίνεται στο Σχήμα 5.3.

[2]https://neo4j.com/docs/bloom-user-guide/current/about-bloom/
[3]https://neo4j.com/developer/neo4j-browser/



Σχήμα 5.3: Δημιουργία εικονικού γράφου με cypher

Περιέχει τις πληροφορίες που εξαρχής δώσαμε στο πρώτο query δηλαδή το όνομα της σχέσης **commented on** και το όνομα της μεταβλητής **Person** για κάθε συγγραφέα-σχολιαστή που υπάρχει στο αρχείο.

Είμαστε πλέον έτοιμοι να τρέξουμε όλες τις εντολές που θέλουμε για να βρούμε αριθμό κόμβων και σχέσεων, κεντρικότητα και αριθμό κοινοτήτων σε κάθε υποφόρουμ.

5.2 Μετρικές κοινωνικών δικτύων

Παρακάτω παρουσιάζουμε τις μετρικές που θα χρησιμοποιήσουμε για την αξιολόγηση των φόρουμ. Για κάθε μετρική δίνεται το cypher query, το πεδίο τιμών αλλά και η διαισθητική ερμηνεία του.

5.2.1 Κεντρικότητα με βάση το βαθμό

Η πρώτη μετρική που μας αφορά, είναι η κεντρικότητα με βάση το βαθμό, δηλαδή πόσο δημοφιλής είναι κάθε κόμβος. Όσο μεγαλύτερο το **score** τόσο πιο δημοφιλής ο κόμβος και κατά συνέπεια πιο ισχυρά συνδεδεμένος.

```
CALL gds.degree.stream(
   'myGraph',
   { orientation: 'NATURAL' }
)
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).name AS name, score AS followees
ORDER BY followees DESC, name DESC
```

5.2.2 Κεντρικότητα με βάση την εγγύτητα

Στη συνέχεια, τρέχουμε το παρακάτων query το οποίο θα μας δώσει με αύξουσα σειρά τους κόμβους που είναι πιο κοντά με τους υπόλοιπους, δηλαδή που έχουν το μικρότερο μήκος μονοπατιού.

```
CALL gds.alpha.closeness.stream({
```

```
  nodeProjection: 'Person',
  relationshipProjection: 'commented_on'
})
YIELD nodeId, centrality
RETURN gds.util.asNode(nodeId).name AS user, centrality
ORDER BY centrality DESC
```

όπου **node projection** είναι το όνομα της μεταβλητής που εξαρχής δώσαμε όταν φτιάξαμε τις σχέσεις για τους συγγραφείς-σχολιαστές και **relationship projection** το όνομα της σχέσης καθαυτής, **commented on**.

5.2.3 Κεντρικότητα με βάση την διαμεσολαβικότητα

Το παρακάτω query θα μας επιστρέψει με αύξουσα σειρά τους κόμβους που είναι πιο δημοφιλείς με γνώμονα όμως κατά πόσο δρούν σαν γέφυρα διάδοσης πληροφορίας για τους υπόλοιπους κόμβους.

```
CALL gds.betweenness.stream('myGraph')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).name AS name, score
ORDER BY name ASC
```

5.2.4 Κεντρικότητα με βάση το ιδιοδιάνυσμα

Η τελευταία μετρική κεντρικότητας, αφορά στο ιδιοδιάνυσμα και η εκτέλεση της παρακάτω εντολής στον cypher θα μας επιστρέψει με αύξουσα σειρά τους κόμβους με το μεγαλύτερο βαθμό ιδιοδιανύσματος, έναν βαθμό που όπως αναφέραμε στην Υποενότητα 2.1.3 αφορά στον αριθμό συνδέσεων κάθε κόμβου, συνδιασμένο με τον βαθμό που έχει και ο γειτονικός κόμβος με τον οποίο συνδέεται.

```
CALL gds.eigenvector.stream('myGraph')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).name AS name, score
ORDER BY score DESC, name ASC
```

5.2.5 Αριθμός κοινοτήτων γράφων

Η τελευταία μετρική αφορά στον αριθμό των κοινοτήτων που δημιουργούνται με βάση τον αλγόριθμο Louvain[4], ο οποίος είναι ένας αλγόριθμος εύρεσης κοινοτήτων. Η δουλειά του αλγόριθου είναι να βελτιστοποιεί την δομικότητα (modularity) του γράφου-δικτύου κατά την διάρκεια εκτέλεσης του. Η δομικότητα είναι μια τιμή εντός του [-5,1] που αφορά στην πυκνότητα των σχέσεων εντός των κοινοτήτων, λαμβάνοντας υπόψιν όμως και τις σχέσεις εκτός κοινοτήτων. Η βελτιστοποίηση της δομικότητας έχει ως αποτέλεσμα την καλύτερη δυνατή συσταδοποίηση[5]. Τον αριθμό των κοινοτήτων μπορούμε να τον βρούμε εάν εκτελέσουμε το παρακάτω query στο Neo4j.

[4]https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/
[5]https://perso.uclouvain.be/vincent.blondel/research/louvain.html

```
CALL gds.louvain.stream('myGraph')
YIELD nodeId, communityId
RETURN gds.util.asNode(nodeId).name AS name, communityId
ORDER BY name ASC
```

5.3 Αποτελέσματα εντολών και οπτικοποιήσεις

Όσα αποτελέσματα και γραφήματα δημιουργήσαμε, αφορούν μόνο τις παραπάνω σχέσεις και όχι όσες αναλύσαμε στην Υποενότητα 2.1.3. Αφού τρέξουμε το πρώτο query που φαίνεται στο Σχήμα 5.2 και δημιουργήσουμε τις συνδέσεις, το Neo4j θα μας επιστρέψει ένα μήνυμα επιτυχίας και τον αριθμό των κόμβων και σχέσεων που δημιούργησε για το υποφόρουμ. Στον Πίνακα 5.1, συμπεριλάβαμε για όλα τα υποφόρουμ τους αριθμούς αυτούς για να έχουμε μια εικόνα του μεγέθους τους.

Φόρουμ	Αριθμός Κόμβων	Αριθμός Σχέσεων	Πυκνότητα
Dread	14,845	37,427	0.00033
Hidden Answers	1,188	3,637	0.0051
CryptBB	1,493	3,712	0.0033
Helium	1,001	767	0.0015
Tape	747	621	0.0022

Πίνακας 5.1: Αριθμός κόμβων, σχέσεων και πυκνότητα

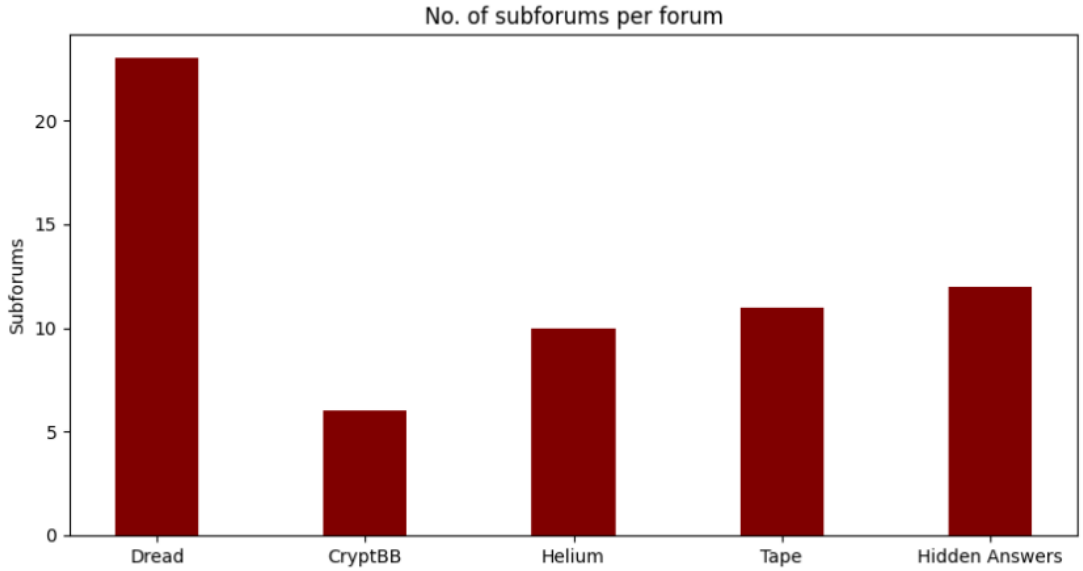
Αποδεικνύεται ότι το Dread είναι το μεγαλύτερο από τα υπόλοιπα, τα οποία μεν έχουν σχετικά ίδιο αριθμό κόμβων αλλά όχι σχέσεων. Η πυκνότητα (ή συνοχή)[6], αν και δεν είναι μετρική για ανάλυση κοινωνικών δικτύων, υπολογίστηκε γιατί έχουμε να κάνουμε με γράφους-δίκτυα οπότε μας παρέχει μια ακόμη σημαντική πληροφορία. Υπολογίζεται με τον τύπο:

$$\frac{2 * E}{N * (N - 1)}$$

όπου E ο αριθμός σχέσεων και N ο αριθμός κόμβων. Η πυκνότητα δείχνει τις πιθανές σχέσεις που μπορούν να δημιουργηθούν σε ένα δίκτυο με πραγματικές συνδέσεις-σχέσεις.

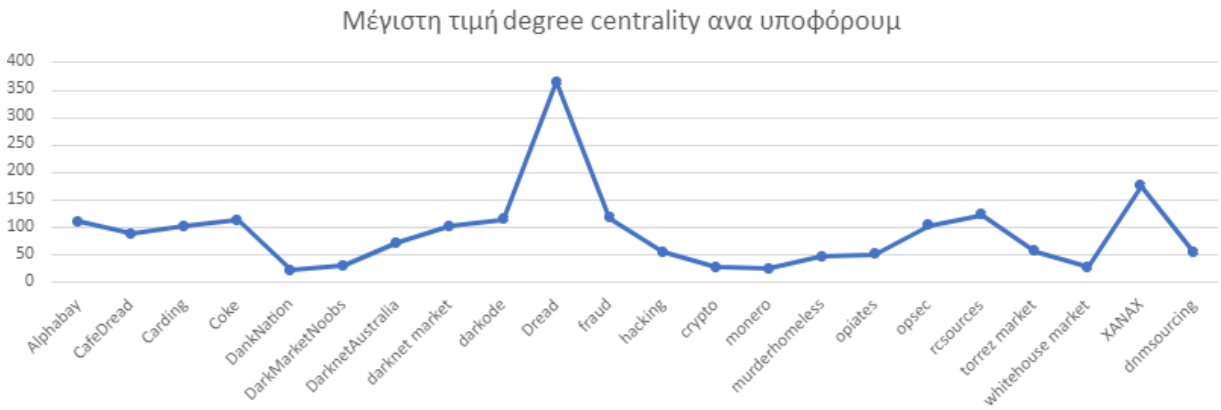
Πριν αναδείξουμε τα αποτελέσματα των μετρικών, στο παρακάτω γράφημα φάινεται ο αριθμός των υποφόρουμ που επιλέξαμε για ανάλυση (Σχήμα 5.4).

[6]https://www.the-vital-edge.com/what-is-network-density/



Σχήμα 5.4: Αριθμός υποφόρουμ ανα φόρουμ

Έπειτα, εκτελώντας το δεύτερο query για το φόρουμ Dread, βρίσκουμε την κεντρικότητα με βάση το βαθμό για όλα τα υποφόρουμ του. Τα αποτελέσματα φαίνονται στο Σχήμα 5.5.

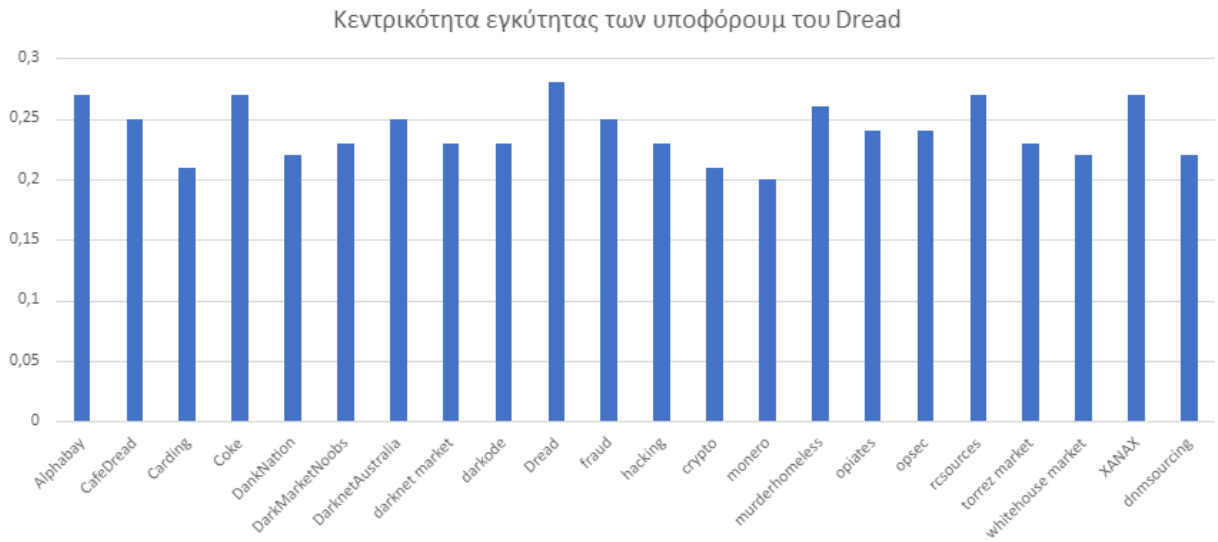


Σχήμα 5.5: Μέγιστες τιμές κεντρικότητας με βάση το βαθμό για το Dread

Όπως είπαμε, η μετρική αυτή μας δείχνει το πόσο δημοφιλής είναι ένας κόμβος με βάση το πόσοι άλλοι συνδέονται σε αυτόν ή για να μιλήσουμε με όρους κοινωνικών δικτύων, πόσους ακόλουθους έχει. Ο μεγαλύτερος αριθμός βρέθηκε στο ομώνυμο υποφόρουμ με τιμή κεντρικότητας 350, ενώ σε όλα τα υπόλοιπα κυμαίνεται από 15 έως 100. Εννοείται ότι σε όλα τα υπόφορουμ υπήρχε ως ελάχιστη τιμή το 0, που σημαίνει ότι υπήρχε έστω μια δημοσίευση χωρίς απαντήσεις.

Εκτελώντας το επόμενο query, βρίσκουμε το βαθμό της κεντρικότητας εγγύτητας, που οι τιμές του κυμαίνονται εντός του 0 και του 1. Όσο πιο κοντά στο 1, τόσο πιο 'κεντρικός' κόμβος και κατά επέκταση κοντά στους περισσότερους υπόλοιπους κόμβους. Οι

μέσες τιμές κεντρικότητας με βάση την εγγύτητα για το φόρουμ Dread φαίνονται στο Σχήμα 5.6.



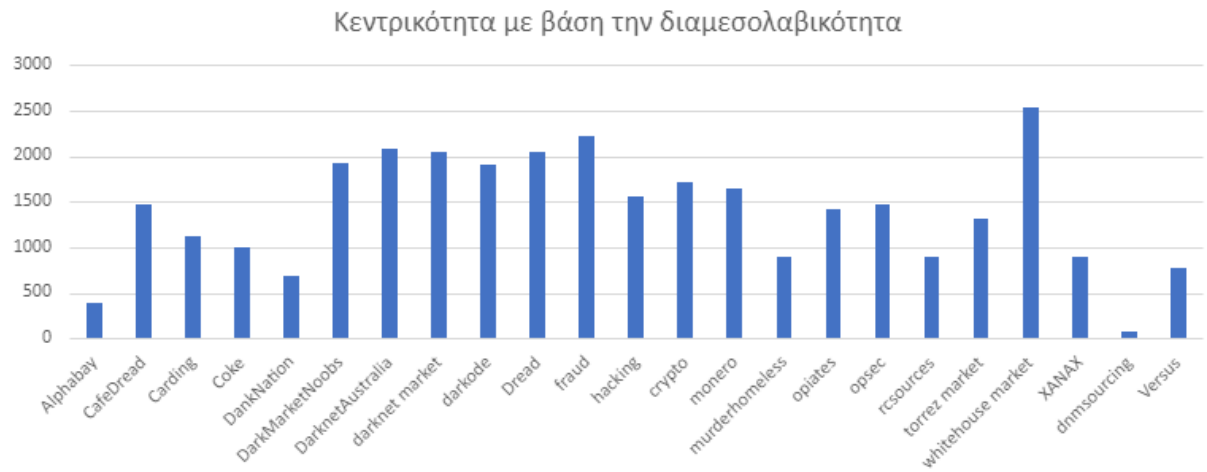
Σχήμα 5.6: Μέσες τιμές κεντρικότητας με βάση την εγγύτητα για το Dread

Όπως βλέπουμε από τα αποτελέσματα, οι τιμές δεν ξεπερνούν το 0.3 που σημαίνει ότι δεν υπάρχουν αρκετοί κεντρικοί κόμβοι που είναι ισχυρά συνδεδεμένοι με τους άλλους. Περιμέναμε αυτό το αποτέλεσμα από τη στιγμή που η συνολική πυκνότητα του γράφου είναι αρκετά χαμηλή (Πίνακας 5.1).

Στη συνέχεια, θέλουμε να υπολογίσουμε τις τιμές κεντρικότητας με βάση την διαμεσολαβικότητα. Η μέγιστη τιμή που μπορεί να πάρει ένας κόμβος δεν μπορεί να ξεπερνά τον αριθμό

$$N^2 - N - 1$$

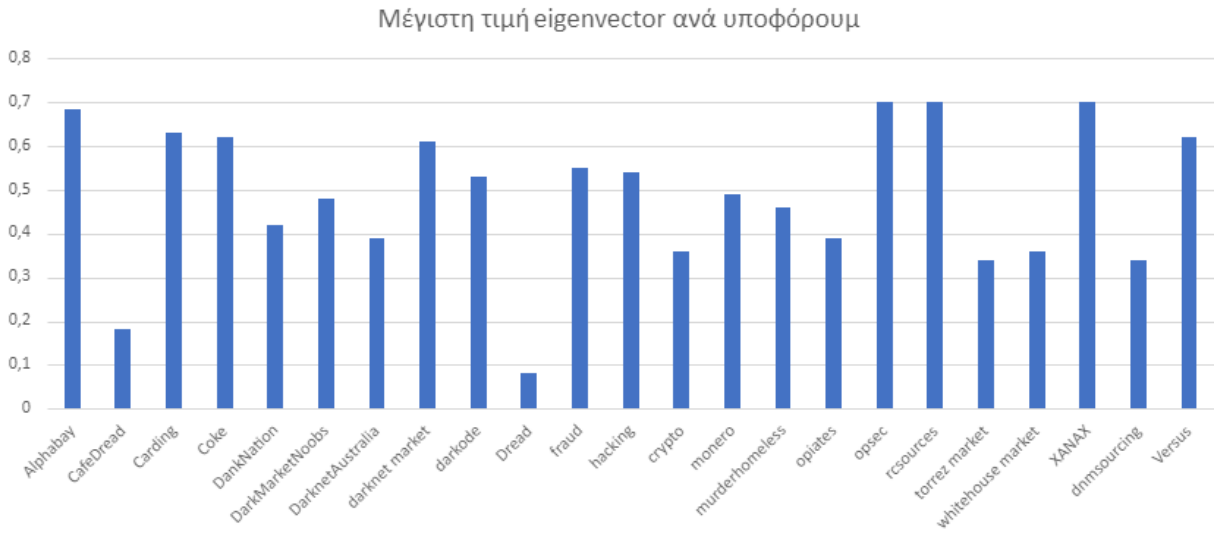
Για να έχουμε καλύτερη εικόνα, παρουσιάζουμε τις μέσες τιμές ανα υποφόρουμ στο Σχήμα 5.7.



Σχήμα 5.7: Μέσες τιμές κεντρικότητας με βάση την διαμεσολαβικότητα για το Dread

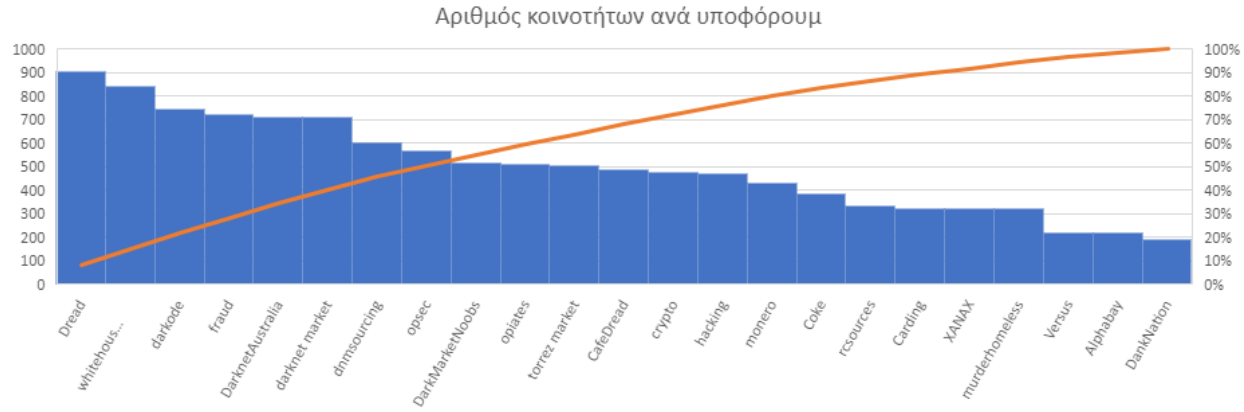
Διαπιστώνουμε ότι τα περισσότερα υποφόρουμ σχετικά με ναρκωτικά και μάυρες αγορές (darkode, whitehouse, dark market of australia) έχουν μεγάλο βαθμό κεντρικότητας, δηλαδή μεγαλύτερη και γρηγορότερη διάδοση πληροφοριών.

Στη συνέχεια, υπολογίζουμε τις κεντρικότητες με βάση το ιδιοδιάνυσμα για κάθε υποφόρουμ που μπορούν να λάβουν τιμή μόνο εντός του 0 και 1. Είναι σημαντικό να σημειώσουμε επίσης ότι ενώ ένας κόμβος μπορεί να έχει πολλές συνδέσεις δεν σημαίνει ότι θα έχει υψηλό βαθμό κεντρικότητας, αφού ενδέχεται όλοι οι κόμβοι που συνδέονται μαζί του να έχουν χαμηλό βαθμό. Αντίστοιχα και για έναν κόμβο με ελάχιστες συνδέσεις. Στο Σχήμα 5.8 παρουσιάζουμε τις μέγιστες τιμές που βρήκαμε.



Σχήμα 5.8: Μέγιστες τιμές κεντρικότητας με βάση το ιδιοδιάνυσμα για το Dread

Στο τελευταίο Σχήμα 5.9 για το φόρουμ Dread, βλέπουμε τον αριθμό των κοινοτήτων που μας δίνει ο αλγόριθμος Louvain. Η κόκκινη γραμμή μας δείχνει της διακύμανση των τιμών αυτών.



Σχήμα 5.9: Αριθμός κοινοτήτων υποφόρουμ του Dread

Όπως βλέπουμε, το υποφόρουμ στο οποίο δημιουργούνται οι περισσότερες κοινότητες, είναι το ομώνυμο το οποίο όπως είδαμε στις προηγούμενες μετρικές είχε τη μεγαλύτερη τιμή degree centrality και closeness centrality.

Συνεχίζοντας και στα υπόλοιπα φόρουμ, βλέπουμε στους Πίνακες 5.2 και 5.3 τις μέσες και μέγιστες τιμές αντίστοιχα για τις τέσσερις μετρικές κεντρικότητας. Στο υποφόρουμ hacking βρέθηκε ένας κόμβος με 54 συνδέσεις, αρκετά μεγαλύτερος από το μέγιστο των υπολοίπων. Παρ'όλα αυτά όπως διαπιστώνουμε και παρακάτω σε όλα τα φόρουμ, ένας μικρός αριθμός χρηστών-κόμβων έχει αρκετές συνδέσεις αλλά η πλειοψηφία έχει ελάχιστες. Αυτό μας δείχνει πως ακόμα και στα underground φόρουμ υπάρχει κάποια ιεραρχία, δηλαδή κάποια άτομα, είτε απλοί χρήστες είτε διαχειριστές, που έχουν τον έλεγχο στο δίκτυο. Χρήστες που είναι παλαιότεροι από άλλους ή έχουν φήμη λόγω της θέσης τους τείνουν να έχουν περισσότερες αλληλεπιδράσεις από έναν μέσο χρήστη.

CryptBB				
Υποφόρουμ	Degree	Eigenvector	Betweeness	Closeness
Hacking	2.13	0.0062	840	0.269
Programming	1.46	0.015	479	0.228
Discussions on darknet	1.77	0.013	557	0.253
Newbies	1.74	0.011	774	0.255
Sell	1.35	0.040	206	0.27
Support	1.33	0.032	110	0.325

Πίνακας 5.2: Μέσες τιμές κεντρικότητας των υποφόρουμ του CryptBB

Το CryptBB, όπως το Dread και το Hidden Answers, σύμφωνα με τον Πίνακα 5.3 έχουν τριπλάσιο αριθμό σχέσεων έναντι κόμβων, οπότε είναι φυσιολογικό η κεντρικότητα με βάση τη διαμεσολαβικότητα (betweeness να έχει υψηλό βαθμό διότι είναι λογικό ότι γίνεται άμεση διάδοση πληροφορίας μεταξύ των γειτονικών κόμβων. Όσον αφορά στο ιδιοδιάνυσμα, οι μέγιστες τιμές που βρήκαμε δεν ξεπερνούν το 0.54 ένταντι του 0.7 στο φόρουμ Dread.

CryptBB				
Υποφόρουμ	Degree	Eigenvector	Betweeness	Closeness
Hacking	54	0.39	99894	1
Programming	13	0.54	5050	1
Discussions on darknet	21	0.38	17494	1
Newbies	29	0.404	47047	1
Sell	21	0.35	3857	1
Support	8	0.508	2113	1

Πίνακας 5.3: Μέγιστες τιμές κεντρικότητας των υποφόρουμ του CryptBB

Στο Helium, που μαζί με το Tape είναι τα μικρά φόρουμ της παρούσας έρευνας, ο μέγιστος βαθμός degree είναι το 14, ενώ είναι αξιοσημείωτο πως οι μέσες τιμές ανά υποφόρουμ δεν ξεπερνούν το 1. Οι μέγιστες τιμές κεντρικότητας με βάση το ιδιοδιάνυσμα είναι παρόμοιες με τα προηγούμενα φόρουμ γιατί ο αλγόριθμος αυτός δεν λαμβάνει υπόψη το μέγεθος του δικτύου. Η κεντρικότητα με βάση τη διαμεσολαβικότητα έχει σχετικά χαμηλές τιμές σε όλα τα υποφόρουμ εκτός του Introduction ίσως διότι σε αυτό υπάρχει μεγάλος αριθμός δημοσιέυσεων και σχολίων γιατί αφορά νέους

χρήστες του φόρουμ που γίνονται μέλη στη κοινότητα. Όλα τα αποτελέσματα φαίνονται στους Πίνακες 5.5 και 5.4.

Helium				
Υποφόρουμ	Degree	Eigenvector	Betweeness	Closeness
Anonymity	7	0.48	14	1
Carding	11	0.59	10	1
General Discussion	8	0.499	103	1
Tutorials	10	0.31	3.6	1
Hacked Data	14	0.4	11	1
Hacking Tools	8	0.34	6	1
Hacking Tutorials	7	0.37	15	1
Introduction	9	0.74	2929	1
Paid hacking services	4	0.4	2	1
Programming	5	0.37	15	1

Πίνακας 5.4: Μέγιστες τιμές κεντρικότητας των υποφόρουμ του Helium

Helium				
Υποφόρουμ	Degree	Eigenvector	Betweeness	Closeness
Anonymity	0.91	0.067	1.83	0.41
Carding	0.86	0.022	0.25	0.37
General Discussion	0.79	0.017	1.68	0.31
Tutorials	0.64	0.126	0.17	0.5
Hacked Data	0.83	0.022	0.22	0.3
Hacking Tools	0.71	0.058	0.12	0.39
Hacking Tutorials	0.67	0.11	0.79	0.72
Introduction	0.7	0.0068	58	0.28
Paid hacking services	0.71	0.033	0.061	0.33
Programming	0.54	0.085	1.32	0.52

Πίνακας 5.5: Μέσες τιμές κεντρικότητας των υποφόρουμ του Helium

Στο ισπανικό φόρουμ Hidden Answers, το οποίο σε αριθμό χρηστών δεν είναι μεγάλο αλλά έχει αρκετές αλληλεπιδράσεις, συναντάμε και την μεγαλύτερη πυκνότητα σε σχέση με τα υπόλοιπα τέσσερα φόρουμ (Πίνακας 5.1). Αν και ο μέγιστος αριθμός κεντρικότητας με βάση το βαθμό που βρήκαμε κυμαίνεται σε φυσιολογικά πλαίσια, είναι σημαντικό να υπογραμμίσουμε τις μέσες τιμές degree οι οποίες είναι μεταξύ 2 και 3. Αυτό σημαίνει ότι οι περισσότεροι κόμβοι είχαν πάνω από δύο συνδέσεις και ελάχιστοι είχαν μία ή καμία.

Hidden Answers				
Υποφόρουμ	Degree	Eigenvector	Betweeness	Closeness
Crypto	11	0.52	65	0.63
Darknet	21	0.445	792	0.6
Erotic Discussion	20	0.35	402	0.57
Hacking	15	0.37	102	0.6
Health	42	0.38	536	0.64
Knowledge Information	42	0.45	2709	1
Money	17	0.44	375	0.59
Other	43	0.4	3704	0.67
Programming	11	0.49	57	0.588
SadTimes	20	0.45	398	0.59
Technology	24	0.48	509	0.58
WTF	29	0.43	690	0.59

Πίνακας 5.6: Μέγιστες τιμές κεντρικότητας των υποφόρουμ του Hidden Answers

Επίσης, εξαιρώντας τα υποφόρουμ crypto, hacking, programming, σε όλα τα υπόλοιπα η μέση τιμή διαμεσολαβικότητας είναι ικανοποιητική και αρκετά υψηλότερη σε σχέση με τα μικρά φόρουμ με λίγες συνδέσεις όπως το Helium προηγουμένως και το Tape παρακάτω. Τα ακριβή αποτελέσματα για το Hidden Answers φαινονται στους Πίνακες 5.6 και 5.7.

Hidden Answers				
Υποφόρουμ	Degree	Eigenvector	Betweeness	Closeness
Crypto	2.16	0.043	1.89	0.38
Darknet	2.7	0.04	24.5	0.35
Erotic Discussion	2.3	0.078	14.81	0.37
Hacking	2.21	0.04	3.85	0.35
Health	2.35	0.075	12.25	0.38
Knowledge Information	3.24	0.036	77	0.39
Money	2.51	0.049	10.20	0.35
Other	5.26	0.035	127	0.40
Programming	1.9	0.068	2.15	0.36
SadTimes	2.87	0.073	24.98	0.38
Technology	2.35	0.06	11.11	0.37
WTF	3.14	0.04	30.6	0.37

Πίνακας 5.7: Μέσες τιμές κεντρικότητας των υποφόρουμ του Hidden Answers

Για το τελευταίο φόρουμ, τα αποτελέσματα των μετρικών παρουσιάζονται στους Πίνακες 5.8 και 5.9. Και σε αυτό το φόρουμ διακρίνουμε πολύ μικρό αριθμό σε degree διότι είναι το φόρουμ με τις λιγότερες σχέσεις. Όμως σε αυτό το φόρουμ, έχουμε μέσες τιμές

closeness κοντά στο 0.5 οι οποίες είναι λίγο μεγαλύτερες από τις τιμές των μεγαλύτερων φόρουμ και αυτό εξηγείται λόγω του μεγέθους του δικτύου, καθώς ένας κόμβος βρίσκεται πιο έυκολα από τους υπόλοιπους. Υπενθυμίζουμε ότι όσο μεγαλύτερος ο βαθμός closeness, τόσο μικρότερο το μήκος μονοπατιού.

Tape				
Υποφόρουμ	Degree	Eigenvector	Betweeness	Closeness
Bankdrops	15	0.59	16	1
Cannabis	5	0.59	6	1
Cocaine	6	0.55	6	1
Coding	3	0.37	3	1
Credit Cards	17	0.79	10	1
Crypto	8	0.34	1	1
Gangland Market	7	0.47	10	0.75
Guides	7	0.4	41	1
Openmarket Zone	21	0.62	315	1
Paypal	14	0.74	30	1
Security	13	0.74	6	1

Πίνακας 5.8: Μέγιστες τιμές κεντρικότητας των υποφόρουμ του Tape

Tape				
Υποφόρουμ	Degree	Eigenvector	Betweeness	Closeness
Bankdrops	0.87	0.034	0.29	0.4
Cannabis	0.67	0.03	0.192	0.404
Cocaine	0.8	0.031	0.131	0.52
Coding	0.67	0.055	0.06	0.61
Credit Cards	0.88	0.018	0.315	0.415
Crypto	0.55	0.088	0.02	0.47
Gangland Market	1.27	0.25	1.63	0.46
Guides	0.67	0.049	2	0.53
Openmarket Zone	0.86	0.009	0.46	0.23
Paypal	1	0.046	1.3	0.46
Security	0.97	0.022	0.178	0.46

Πίνακας 5.9: Μέσες τιμές κεντρικότητας των υποφόρουμ του Tape

5.4 Συμπεράσματα

Μέσω των μετρικών ανάλυσης κοινωνικών δικτύων, μας δίνεται η δυνατότητα να έχουμε μια εικόνα για τους ρόλους των χρηστών μέσα στο δίκτυο (Υποενότητα 2.1.5). Από τη μία έχουμε το αρκετά μεγάλο φόρουμ Dread, το οποίο λόγω του μεγέθους, έχει αρκετά μικρή πυκνότητα. Στο φόρουμ αυτό, διαπιστώσαμε ότι σε κάθε υποφόρουμ, υπάρχει

ένας ή περισσότεροι χρήστες που έχουν κεντρικό ρόλο. Ανάλογα με τον τύπο του υποφόρουμ, είτε αγορά ναρκωτικών, είτε hacking οι χρήστες αυτοί μπορεί να είναι είτε διαχειριστές, δηλαδή υπεύθυνοι των δημοσιεύσεων, είτε επαγγελματίες στον εκάστοτε χώρο που γίνεται η συζήτηση, είτε χρήστες υπέυθυνοι για την ενσωμάτωση άλλων μελών και το χτίσιμο περισσότερων κοινοτήτων στο δίκτυο. Η μεγάλη διακύμανση τιμών degree μας δείχνει πως υπάρχει η δυνατότητα διάδοσης πληροφορίας και πόρων μεταξύ ανόμοιων χρηστών, για παράδειγμα νέοι χρήστες έναντι παλαιότερων και πιο έμπειρων χρηστων. Έρευνες δείχνουν πως αυτή η κατάσταση μπορεί να επηρεάσει τις συμπεριφορές και να διαμορφώσει ιδέες για μελλοντικούς νέους χρήστες[7].Από την άλλη πλευρά, τα φόρουμ Tape και Helium είναι συγκεντρωμένα περισσότερο γύρω από τις κρυπταγορές, δίνοντας μικρότερη έμφαση στην αλληλεπίδραση με νέους χρήστες. Στο φόρουμ CryptBB, λόγω του ότι είναι ένα hacking φόρουμ, που δίνει τη δυνατότητα σε νέους χρήστες μέσω δοκιμασιών, να αποκτήσουν πρόσβαση σε σημαντικά εγχειρίδια εκμάθησης εργαλείων, υπάρχει αρκετή αλληλεπίδραση και γρηγόρετη διάδοση πληροφορίας στα σχετικά υποφόρουμ και μικρότερη στα άλλα όπως κρυπταγορές ή απλός προγραμματισμός. Τέλος, στο Hidden Answers το οποίο είναι έυκολα προσβάσιμο και υπέρμαχο του free speech, βρήκαμε πως έχει τη μεγαλύτερη πυκνότητα ακριβώς για αυτό το λόγο. Υπάρχουν ελάχιστοι κανόνες για τις δημοσιέυσεις και κυρίως δεν υπάρχει ιεραρχία χρηστών. Μπορούν όλοι ισάξια να έχουν κεντρικό ή όχι ρόλο μέσα στο δίκτυο.

[7]https://cybercrimejournal.com/holtetal2012janijcc.pdf

Κεφάλαιο 6

Επίλογος και μελλοντική έρευνα

Όταν κάνουμε μια αναζήτηση στο διαδίκτυο έχουμε πρόσβαση στον ορατό ιστό (surface web). Τι συμβαίνει όμως στον σκοτεινό ιστό (dark web), στο μέρος εκείνο του διαδικτύου όπου η πρόσβαση είναι δυνατή μόνο μέσω εργαλείων ανωνυμίας·

Η παρούσα εργασία έχει ως στόχο να ανακαλύψει και να μελετήσει φόρουμ του σκοτεινού ιστού. Χρησιμοποιήσαμε κατάλληλα εργαλεία για να αποκτήσουμε πρόσβαση στο dark web και να ανακτήσουμε δεδομένα όπως χρήστες, συζητήσεις, σχόλια από πέντε φόρουμ που σχετίζονται με παραβατικές δραστηριότητες όπως αγοραπωλησία και κατασκευή ναρκωτικών ουσιών, μαθήματα και υπηρεσίες hacking επί πληρωμή, καθώς και συζητήσεις περί κοινωνικών και πολιτικών ζητημάτων. Τα δεδομένα αυτά αφορούν σε αναρτήσεις των τελευταίων 3 μηνών, συγκεκριμένα από Σεπτέμβριο μέχρι Δεκέμβριο. Στη συνέχεια, χρησιμοποιήσαμε ανάλυση κοινωνικών δικτύων για να μελετήσουμε τις συσχετίσεις των χρηστών σε αυτά τα φόρουμς.

Τα αποτελέσματα μας, αρχικά για το πρώτο φόρουμ Dread δείχνουν πως σε κάθε υποφόρουμ, υπάρχει ένας ή περισσότεροι χρήστες που έχουν κεντρικό ρόλο. Ανάλογα με τον τύπο του υποφόρουμ, είτε αγορά ναρκωτικών, είτε hacking οι χρήστες αυτοί μπορεί να είναι είτε διαχειριστές, δηλαδή υπεύθυνοι των δημοσιεύσεων, είτε επαγγελματίες στον εκάστοτε χώρο που γίνεται η συζήτηση, είτε χρήστες υπέυθυνοι για την ενσωμάτωση άλλων μελών και το χτίσιμο περισσότερων κοινοτήτων στο δίκτυο. Η μεγάλη διακύμανση τιμών degree μας δείχνει πως υπάρχει η δυνατότητα διάδοσης πληροφορίας και πόρων μεταξύ ανόμοιων χρηστών, για παράδειγμα νέοι χρήστες έναντι παλαιότερων και πιο έμπειρων χρηστων. Έρευνες δείχνουν πως αυτή η κατάσταση μπορεί να επηρεάσει τις συμπεριφορές και να διαμορφώσει ιδέες για μελλοντικούς νέους χρήστες[1]. Από την άλλη πλευρά, τα φόρουμ Tape και Helium είναι συγκεντρωμένα περισσότερο γύρω από τις κρυπταγορές, δίνοντας μικρότερη έμφαση στην αλληλεπίδραση με νέους χρήστες. Στο φόρουμ CryptBB, λόγω του ότι είναι ένα hacking φόρουμ, που δίνει τη δυνατότητα σε νέους χρήστες μέσω δοκιμασιών, να αποκτήσουν πρόσβαση σε σημαντικά εγχειρίδια εκμάθησης εργαλείων, υπάρχει αρκετή αλληλεπίδραση και γρηγόρετη διάδοση πληροφορίας στα σχετικά υποφόρουμ και μικρότερη στα άλλα όπως κρυπταγορές ή απλός προγραμματισμός. Τέλος, στο Hidden Answers το οποίο είναι έυκολα προσβάσιμο και υπέρμαχο του free speech, βρήκαμε πως έχει τη μεγαλύτερη πυκνότητα ακριβώς για αυτό το λόγο. Υπάρχουν ελάχιστοι κανόνες για τις δημοσιέυσεις και κυρίως δεν υπάρ-

[1] https://cybercrimejournal.com/holtetal2012janijcc.pdf

χει ιεραρχία χρηστών. Μπορούν όλοι ισάξια να έχουν κεντρικό ή όχι ρόλο μέσα στο δίκτυο.

Ως πιθανές βελτιώσεις της δουλειάς μας, προτείνονται αρχικά η ανάλυση δημοσιεύσεων από τα φόρουμς που επιλέξαμε περισσότερο των 3 μηνών έτσι ώστε να καταλάβουμε καλύτερα με τη πάροδο του χρόνου τη συμπεριφορά των χρηστών. Επίσης, θα μπορέσουμε να αξιολογήσουμε τις σχέσεις των χρηστών και πώς αλλάζει η ιεραρχική τους θέση ανάλογα με το activity τους.

Τέλος, η δουλειά μας θα μπορούσε να επεκταθεί με την περαιτέρω έρευνα και ανάλυση φόρουμς τα οποία δεν είναι στην αγγλική γλώσσα. Όπως αναφέραμε στην Υποενότητα 3.4, υπάρχει αρκετό υλικό σε γερμανικά, πολωνικά και ρώσσικα φόρουμς, τα οποία με την απαραίτητη μετάφραση θα μας δώσουν αρκετό υλικό να αξιολογήσουμε τις κοινωνικές σχέσεις στα underground forums, άλλων χωρών.

Παράρτημα Α΄

Ακρωνύμια και συντομογραφίες

SNA Social Network Analysis

HTTP Hyper Text Transfer Protocol

HTML Hyper Text Markup Language

CAPTCHA Completely Automated Public Turing Test to tell Computers Humans Apart

API Application Programming Interface

IP Internet Protocol

VPN Virtual Private Network

TOR The Onion Router

AJAX Asynchronous Javascript and XML

Bibliography

[1] "Social networks and sna." [Online]. Available: http://imu.ntua.gr/static/courses/strategicISmanagement/lectures/11-SocialNetworks.pdf

[2] "Getting started with sna." [Online]. Available: https://towardsdatascience.com/how-to-get-started-with-social-network-analysis-6d527685d374

[3] "Moving towards smarter data." [Online]. Available: https://morioh.com/p/0af0fca7a490

[4] R. A. Hanneman, "Introduction to social network methods." [Online]. Available: https://faculty.ucr.edu/~hanneman/nettext/C10_Centrality.html

[5] J. Najera, "Measuring network centrality." [Online]. Available: https://www.cantorsparadise.com/measuring-network-centrality-2a76b0045410

[6] "Predicting tie strength with ego network structures." [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1094996820301390?via%3Dihub

[7] "Gephi." [Online]. Available: https://gephi.org/

[8] "Neo4j." [Online]. Available: https://neo4j.com/

[9] "Igraph." [Online]. Available: https://igraph.org/

[10] "Networkx." [Online]. Available: https://networkx.org/

[11] "What is a web crawler? how web spiders work." [Online]. Available: https://www.cloudflare.com/learning/bots/what-is-a-web-crawler/

[12] "9 web scraping challenges." [Online]. Available: https://www.octoparse.com/blog/9-web-scraping-challenges

[13] "Web scraping with scrapy." [Online]. Available: https://www.analyticsvidhya.com/blog/2017/07/web-scraping-in-python-using-scrapy/

[14] "Web scraping in python with scrapy." [Online]. Available: https://www.analyticsvidhya.com/blog/2017/07/web-scraping-in-python-using-scrapy/

[15] "The beautiful soup documentation." [Online]. Available: https://beautiful-soup-4.readthedocs.io/en/latest/

[16] "Get started with requests library in python." [Online]. Available: https://www.digitalocean.com/community/tutorials/how-to-get-started-with-the-requests-library-in-python

[17] "A deeper look at selenium." [Online]. Available: https://www.selenium.dev/documentation/overview/details/

[18] "Is the dark web dangerous?" [Online]. Available: https://www.kaspersky.com/resource-center/threats/deep-web

[19] "Black market definition." [Online]. Available: https://www.investopedia.com/terms/b/blackmarket.asp#toc-what-is-a-black-market

[20] W. Park, "A study on analytical visualization of deep web." [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9061283

[21] E. Phillips, J. R. C. Nurse, M. Goldsmith, and S. Creese, "Applying social network analysis to security." [Online]. Available: https://www.researchgate.net/publication/274663734_Applying_Social_Network_Analysis_to_Security

[22] Y. Yang, H. Yu, L. Yang, M. Y. L. Chen, G. Zhu, and L. Wen, "Hadoop-based dark web threat intelligence analysis framework." [Online]. Available: https://ieeexplore.ieee.org/document/8984106

[23] L. Yang, F. Liu, and J. M. K. R. K. Ege, "Discovering topics from dark websites." [Online]. Available: https://ieeexplore.ieee.org/document/4925106

[24] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and L. Fernández-Roblese, "Torank: Identifying the most influential suspicious domains in the tor network." [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419300296

[25] H. Zhang, B. Qiut, C. L. Giles, H. C. Foley, and J. Yen, "An lda-based community structure discovery approach for large-scale social networks." [Online]. Available: https://ieeexplore.ieee.org/document/4258697

[26] M. M. Morshed, "An empirical study of deep web based on graph analysis." [Online]. Available: https://www.researchgate.net/publication/345657288_An_Empirical_Study_of_Deep_Web_based_on_Graph_Analysis

[27] A. Tirados, "Dark networks: Social network analysis of dark web communities." [Online]. Available: https://www.recordedfuture.com/dark-web-networks/

[28] E. Phillips, J. R. Nurse, M. Goldsmith, and S. Creese, "Extracting social structure from darkweb forums." [Online]. Available: https://www.researchgate.net/publication/283908783_Extracting_Social_Structure_from_DarkWeb_Forums

[29] G. L'Huillier, H. Alvarez, S. A. Ríos, and F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the dark web." [Online]. Available: https://www.researchgate.net/publication/220025840_Topic-Based_Social_Network_Analysis_for_Virtual_Communities_of_Interests_in_the_Dark_Web_ABSTRACT

[30] B. Monk, J. Mitchell, R. Frank, and G. Davies, "Uncovering tor: An examination of the network structure." [Online]. Available: https://www.hindawi.com/journals/scn/2018/4231326/

[31] S. Toral, R. Martinez-Torres, F. BarreroF, and F. Cortés, "An empirical study of the driving forces behind online communities." [Online]. Avail-

able: https://www.researchgate.net/publication/220146806_An_empirical_study_of_the_driving_forces_behind_online_communities

[32] T. Anwar and M. Abulaish, "Identifying cliques in dark web forums – an agglomerative clustering approach." [Online]. Available: https://www.researchgate.net/publication/261048739_Identifying_cliques_in_dark_web_forums_-_An_agglomerative_clustering_approach

[33] N. Tavabi, N. Bartley1, A. Abeliuk, S. Soni, E. Ferrara, and K. Lerman, "Characterizing activity on the deep and dark web." [Online]. Available: https://www.researchgate.net/publication/331485647_Characterizing_Activity_on_the_Deep_and_Dark_Web

[34] H. Alnabulsi and R. Islam, "Identification of illegal forum activities inside the dark net." [Online]. Available: https://ieeexplore.ieee.org/document/8613998

[35] G. Gadek, S. Brunessaux, and A. Pauche, "Application of ai techniques to deep web social network analysis."

[36] B. RV, E.Grace, M. Kanaga, and S. Kundu, "Application of machine learning in the social network."

[37] "Python's request library." [Online]. Available: https://realpython.com/python-requests/

[38] "Tor project." [Online]. Available: https://www.torproject.org/

[39] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks." [Online]. Available: https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.74.47

[40] M. Newman, "Networks: An introduction. oxford university press." [Online]. Available: ISBN:978-0-19-920665-0.

[41] R. Cohen and S. Havlin, "'complex networks: Structure, robustness and function". cambridge university press." [Online]. Available: ISBN: 978-0-521-84156-6

[42] D. S. Bassett and O. Sporns, "Network neuroscience." [Online]. Available: https://www.nature.com/articles/nn.4502

[43] A. Fornito, "An introduction to network neuroscience: How to build, model, and analyse connectomes." [Online]. Available: https://www.pathlms.com/ohbm/courses/12238/sections/15846/video_presentations/137536

[44] M. Saberi, R. Khosrowabadi, A. Khatibi, B. Misic, and G. Jafari, "Topological impact of negative links on the stability of resting-state brain network." [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7838299/

[45] A. Barabasi and E. Bonabeau, "Scale-free networks." [Online]. Available: https://www.scientificamerican.com/article/scale-free-networks/

[46] S. H. Strogatz and D. J. Watts, "Collective dynamics of small-world networks." [Online]. Available: https://www.nature.com/articles/30918

[47] C. Thompson, "Beyond google: Everything you need to know about the hidden internet." [Online]. Available: https://www.businessinsider.com/difference-between-dark-web-and-deep-web-2015-11?IR=T

[48] "Tor deep web." [Online]. Available: http://www.whoishostingthis.com/blog/2013/12/17/tor-deep-web/