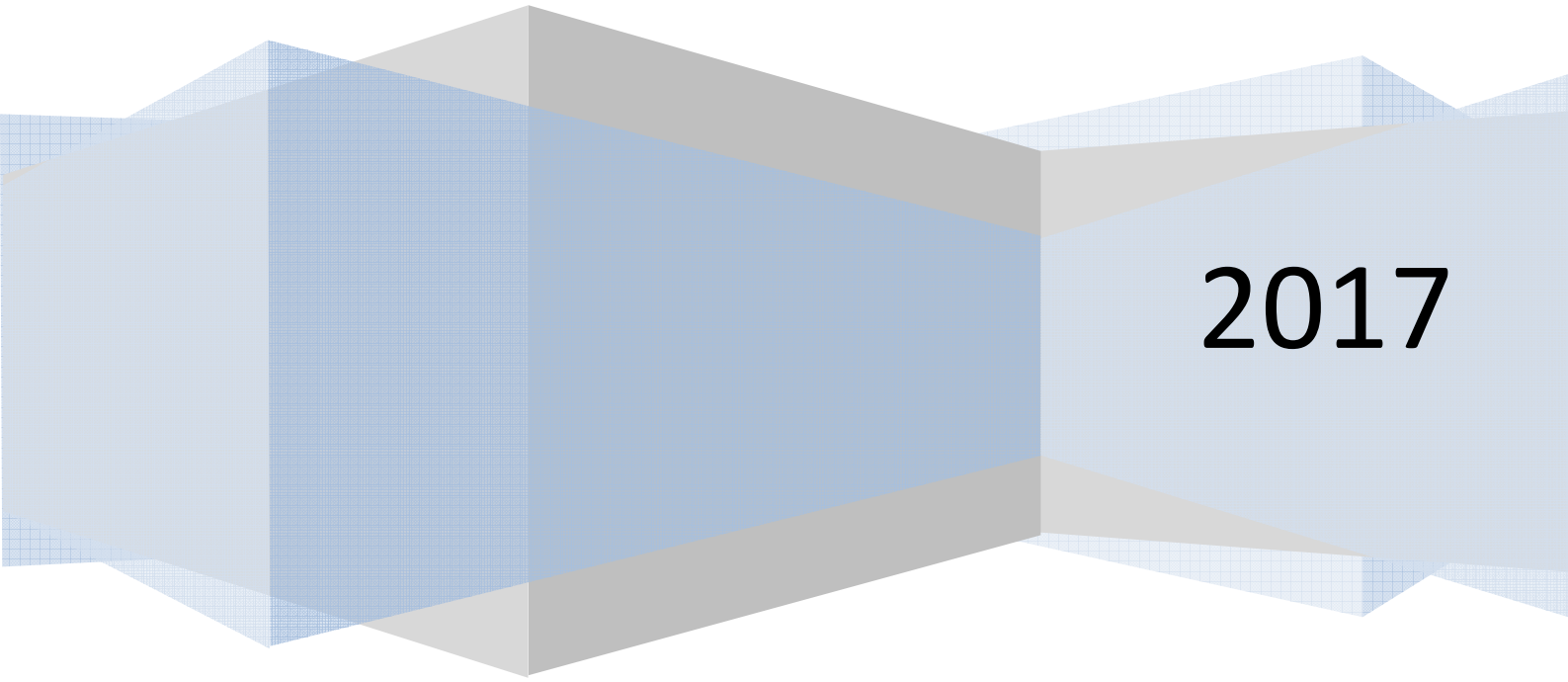
**ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ**
**Μηχανικών Πληροφορικής**

## ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

# Deep / Dark Web: Αρχιτεκτονική, Λειτουργικότητα & Τεχνικά ζητήματα

**Ανδρέας Μούτσιος**

**Επ. καθηγητής: Γεώργιος Ασημακόπουλος**



2017

# Πρόλογος

Το διαδίκτυο αποτελεί μία από τις πλέον σύγχρονες τεχνολογίες η άφιξη των οποίων έχει βοηθήσει σημαντικά το χρήστη ώστε αυτός να έχει πρόσβαση σε άπειρη πληροφορία, να ενημερώνεται τακτικά αλλά και να επικοινωνεί με οποιοδήποτε μέρος του πλανήτη.

Όπως κάθε νόμισμα έχει δύο όψεις, έτσι και το διαδίκτυο δυστυχώς παρουσιάζει και την άσχημη πλευρά του καθώς σε αυτό έχουν αναπτυχθεί συστήματα όπως το dark και το deep τα οποία αφενός εξασφαλίζουν ανωνυμία στο χρήστη, αφετέρου όμως αποκρύπτουν σειρά παράνομων ενεργειών και πράξεων που συντελούνται εντός τους.

## Περίληψη

Η ραγδαία εξέλιξη της τεχνολογίας των υπολογιστών έφερε την έλευση του διαδικτύου, μιας τεχνολογίας που κατόρθωσε να συνενώσει το σύνολο των υπολογιστών του κόσμου και να τους κάνει να επικοινωνούν μεταξύ τους με γρήγορο και άμεσο τρόπο.

Το διαδίκτυο έδωσε απεριόριστη πρόσβαση σε κάθε χρήστη σε έναν τεράστιο όγκο πληροφορίας με σχεδόν ανέξοδο τρόπο και τον οδήγησε σε μέρη που διαφορετικά δε θα είχε καμία απολύτως πρόσβαση.

Καθώς το διαδίκτυο εξελίσσεται, όπως και σε κάθε δομή της κοινωνίας, έτσι και σε αυτό συγκεντρώνεται μία σειρά κακόβουλων επιτιθέμενων που έχουν ως στόχο να προβούν σε παράνομες ενέργειες να υποκλέψουν στοιχεία των χρηστών και να τα χρησιμοποιήσουν προς ίδιο όφελος.

Η παρούσα πτυχιακή, έχει ως στόχο να παρουσιάσει τα συστήματα Deep και Dark Web τα οποία αποτελούν τα μέρη εκείνα του διαδικτύου που βρίσκονται κρυμμένα και δίνουν περιορισμένη πρόσβαση στους χρήστες τους. Μέσα στα συστήματα αυτά, διατηρείται η ανωνυμία του χρήστη αλλά την ίδια στιγμή συντελείται μία σειρά παραβάσεων και παράνομων ενεργειών που διατηρούνται κρυφές και διαφεύγουν σύλληψης.

# Περιεχόμενα

Πρόλογος ..... 2
Περίληψη ..... 3
Εισαγωγή ..... 6
Κεφάλαιο 1 – Το Διαδίκτυο ..... 7
1.1 Διαδίκτυο Αναφορά και τρόπος λειτουργίας ..... 7
1.2 Συστήματα και αλγόριθμοι ανάκλησης πληροφορίας από τον παγκόσμιο ιστό ..... 14
Κεφάλαιο 2 – Το Deep Web ..... 19
2.1 Περιγραφή ..... 19
2.2Αρχιτεκτονική και Πρωτόκολλα λειτουργίας ..... 23
2.3 Surface vs Deep Web ..... 27
2.4 Μηχανές αναζήτησης και τύποι περιεχομένου που περιλαμβάνονται στο Deep Web .. 31
2.5 Αλγόριθμοι αναζήτησης στο Deep Web ..... 34
Κεφάλαιο 3 – Το Dark Web ..... 38
3.1 Dark Web ..... 38
3.2 Αρχιτεκτονική και Πρωτόκολλα λειτουργίας ..... 41
3.3 Το TOR ..... 46
3.4 Εργαλεία του TOR ..... 50
3.5 Επιθέσεις στο Dark Net ..... 53
Κεφάλαιο 4 – Deep και Dark Web ..... 55
4.1 Ομοιότητες και Διαφορές ..... 55
Κεφάλαιο 5 - Επικινδυνότητα ..... 57
5.1 Κίνδυνοι για τον απλό χρήστη - Επιθέσεις από hackers, απώλεια προσωπικών δεδομένων ..... 57
5.2 Μαύρη αγορά στο διαδίκτυο ..... 60
Τιμοκατάλογος ..... 62
Κεφάλαιο 6 – Αντιμετώπιση ..... 63
6. Τρόποι αντιμετώπισης ..... 63
Βιβλιογραφία ..... 66

Εικόνα 1 : Το διαδίκτυο ..... 8
Εικόνα 2 : Λειτουργία του TCP ..... 10
Εικόνα 3 : Μοντέλο client – server ..... 11
Εικόνα 4 : Το Deep Web ..... 20
Εικόνα 5 : Το Deep Web Radio ..... 21
Εικόνα 6 : Web Crawler ..... 23
Εικόνα 7 : Deep Web VS Surface Web ..... 28
Εικόνα 8 : Το διαδίκτυο, το Deep και το Dark web ..... 42
Εικόνα 9 : TOR ..... 47
Εικόνα 10 : Λειτουργία του TOR ..... 48
Εικόνα 11 : Privoxy ..... 51
Εικόνα 12 : Internet, Deep Web, Dark Web ..... 56
Εικόνα 13 : Μαύρη αγορά στο διαδίκτυο ..... 60

# Εισαγωγή

Οι χρήστες που πλοηγούνται κάθε μέρα στο διαδίκτυο και εκμαιεύουν χρήσιμη πληροφορία από αυτό έχουν φτάσει πλέον τα πολλά εκατομμύρια με έναν συντριπτικό αριθμό αυτών όμως να μην έχουν καν διανοηθεί ότι μπορεί να γίνονται αντικείμενο εκμετάλλευσης ή ότι κάτω από το δίκτυο που χρησιμοποιούν υπάρχει ένα άλλο τεράστιο διαδίκτυο που λειτουργεί κανονικά χωρίς αυτοί να το γνωρίζουν αλλά και χωρίς να έχουν καμία πρόσβαση σε αυτό.

Το Deep Web όπως αυτό ονομάστηκε, ξεκίνησε αρχικά με στόχο την ανωνυμία των χρηστών μυστικών υπηρεσιών των ΗΠΑ και μετέπειτα την ανωνυμία των χρηστών σε παγκόσμιο επίπεδο ώστε να διαφυλάσσεται η ιδιωτικότητα αυτών αλλά και η ελευθερία του λόγου.

Δυστυχώς μέσα στο ίδιο το Deep Web δημιουργήθηκε ένα άλλο υποσύστημα που καλείται Dark Web και λειτουργεί κατά βάση με στόχο τη διενέργεια παράνομων πράξεων και την διεξαγωγή παράνομων συναλλαγών. Στην πλειοψηφία των περιπτώσεων οι συναλλαγές πραγματοποιούνται διαδικτυακά καθώς διαμέσου της onion δρομολόγησης που χρησιμοποιείται διατηρείται η ανωνυμία τους αλλά και η απόκρυψή τους.

Η πλέον διαδεδομένη τεχνολογία που χρησιμοποιείται ώστε να λειτουργήσει το Dark Net είναι το TOR το οποίο αποκρύπτει τα ψηφιακά ίχνη τόσο του χρήστη όσο και της ιστοσελίδας που δραστηριοποιείται στο σκοτεινό διαδίκτυο.

Η παρούσα πτυχιακή ξεκινά με το διαδίκτυο και τις βασικές λειτουργίες του και προχωρά σε μία πλήρη περιγραφή τόσο του Deep όσο και του Dark Web αλλά και των τεχνικών που χρησιμοποιούν αυτά ώστε να λειτουργήσουν. Παράλληλα παρουσιάζει το TOR, τα εγκλήματα που διαπράττονται εντός του Dark Web, τις πιθανές απειλές που δέχεται ο χρήστης αλλά και κάποιες προτεινόμενες λύσεις ώστε αυτός να προστατευθεί.

# Κεφάλαιο 1 – Το Διαδίκτυο

## 1.1 Διαδίκτυο Αναφορά και τρόπος λειτουργίας

Αυτό που σήμερα καλείται διαδίκτυο αποτελεί ιστορική εξέλιξη του δικτύου ARPANET το οποίο αναπτύχθηκε πειραματικά στα τέλη της δεκαετίας του 1960 στις ΗΠΑ. Ειδικότερα, το 1969, ερευνητές των πανεπιστημίων ξεκινούν να πειραματίζονται με την απομακρυσμένη διασύνδεση των υπολογιστών μεταξύ τους (Πανεπιστήμιο Θεσσαλίας, 1997). Το ARPANET δημιουργείται με οικονομικούς πόρους που προέρχονται από το πρόγραμμα ARPA, ένα πρόγραμμα που είχε ως στόχο την αξιόπιστη λειτουργία των δικτύων.

Στην αρχική του μορφή των πρόγραμμα είχε ως στόχο την ενασχόληση με μία πειραματική για την εποχή τεχνολογία που καλούνταν μεταγωγή πακέτων. Σε δεύτερη φάση, αποσκοπούσε στην εξασφάλιση της επικοινωνίας μεταξύ απομακρυσμένων υπολογιστών ακόμη και εάν στο πειραματικό στάδιο η επικοινωνία αυτή ήταν προβληματική (Πανεπιστήμιο Θεσσαλίας, 1997). Η μεταδιδόμενη για το δίκτυο πληροφορία βρισκόταν εντός των πακέτων τα οποία προσπαθούσαν τη μεριά του παραλήπτη να επανσυνθεθούν ώστε να του αποκαλύψουν το μήνυμα.

Οι πρώτες συνδέσεις διαδικτύου πραγματοποιήθηκαν τελικά στη δεκαετία του 1970 όπου ξεκίνησε και ένα νέο ερευνητικό πρόγραμμα με το όνομα Internetting Project. Το εν λόγω πρόγραμμα στόχευε στην αντιμετώπιση των όποιων προβλημάτων αντιμετώπιζαν οι διαφορετικοί τύποι δικτύων σχετικά με τη μεταγωγή των πακέτων τους (Πανεπιστήμιο Θεσσαλίας, 1997). Επιπλέον, μέρος των αντικειμενικών στόχων αποτελούσε φυσικά και η διασύνδεση μεταξύ ανομοιογενών δικτύων και η μεταφορά των πακέτων σε αυτά.

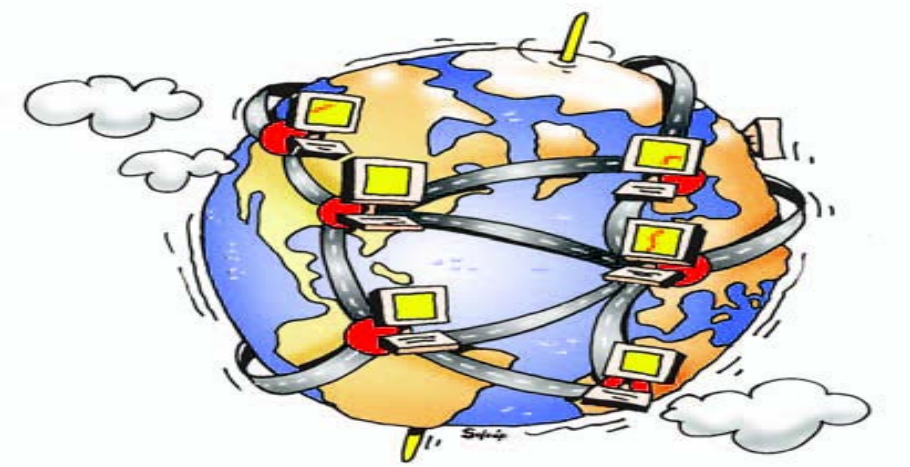
Αποτέλεσμα της έρευνας αυτής είναι η γέννηση του πρωτοκόλλου IP (Internet Protocol) από το οποίο μάλιστα παίρνει και το όνομά του το Internet (Βελούδης). Μοναδικό στοιχείο και καινοτόμο για την εποχή είναι ότι το πρωτόκολλο αυτό επιτρέπει τελικά τη σύνδεση τελικά μεταξύ των διαφορετικών δικτύων με τρόπο τέτοιο ώστε να αποτελούν ένα και μοναδικό δίκτυο. Όσοι υπολογιστές ανήκουν στο δίκτυο αυτό είναι ισοδύναμοι μεταξύ τους και έτσι μπορούν να επικοινωνήσουν

άμεσα χωρίς κανένα απολύτως πρόβλημα. Την ίδια εποχή, σχεδιάζεται και μία σειρά πρωτοκόλλων όπως το TCP και το FTP τα οποία βοήθησαν σημαντικά στην εξέλιξη του διαδικτύου (Βελούδης).

Σημαντική εξέλιξη για τη δικτύωση των υπολογιστών αποτελεί το λειτουργικό σύστημα UNIX το οποίο περιλαμβάνει μέσα του το TCP. Το 1983, εκατοντάδες πανεπιστήμια κατορθώνουν να συνδέσουν τους υπολογιστές τους στο ARPANET το οποίο επειδή με τη διασύνδεση αυτή επιβαρύνεται υπερβολικά διαχωρίζεται σε δύο μέρη (Αριστοτέλειο Πανεπιστήμιο ).

Χρονιά σταθμός για το διαδίκτυο είναι το 1985, όταν το National Science Foundation, δημιουργεί ένα ιδιαίτερα γρήγορο δικό του δίκτυο, βασισμένο στο TCP/IP, και συνδέει πέντε ιδιαίτερα μεγάλα κέντρα υπολογιστών μεταξύ τους. Τα τέλη της δεκαετίας του 1980, βρίσκουν όλε και περισσότερες χώρες διασυνδεδεμένες στο δίκτυο αυτό ενώ χιλιάδες πανεπιστήμια κάνουν προσπάθεια σύνδεσης στο παγκόσμιο αυτό δίκτυο που παίρνει το όνομα Internet (Αριστοτέλειο Πανεπιστήμιο ).

Το 1990, καταργείται πλέον επισήμως το ARPANET και τη θέση το παίρνει Internet το οποίο γίνεται πλέον διαθέσιμο για όλο τον κόσμο. Το πασίγνωστο ερευνητικό κέντρο CERN παρουσιάζει το 1993, το World Wide Web η δημιουργία του οποίου οφείλεται στον επιστήμονα Tim Berners-Lee (Πανεπιστήμιο Θεσσαλίας, 1997). Η διαφορετικότητα του συστήματος αυτού έγκειται στο ότι οι πληροφορίες μπορούν να έχουν τη μορφή πολυμέσων τα οποία βρίσκονται αποθηκευμένα σε χιλιάδες υπολογιστές σε όλο των κόσμο ενώ η παρουσίασή τους γίνεται με τη μορφή ηλεκτρονικών σελίδων στις οποίες ο χρήστης μπορεί να περιηγηθεί και να συλλέξει πληροφορίες.



Εικόνα 1 : Το διαδίκτυο

Την ίδια εποχή, το internet μπαίνει και στον ιδιωτικό χώρο, παρουσιάζονται διάφορες ιδιωτικές εταιρίες που προσφέρουν υπηρεσίες πρόσβασης σε αυτό και έτσι ανοίγουν την πρόσβαση για όλο τον κόσμο. Πλέον, οποιοσδήποτε διαθέτει έναν υπολογιστή και ένα modem, μπορεί να συνδεθεί στο internet και να έχει πρόσβαση στην παγκόσμια πληροφορία. Η υπηρεσία αυτή, προσελκύει με το πέρασμα των χρόνων, όλο και μεγαλύτερο αριθμό χρηστών, με το μεγαλύτερο ποσοστό της γης πλέον να είναι συνδεδεμένο στα διαδίκτυο.

Βασικό στοιχείο για τη λειτουργία του διαδικτύου αποτελεί το πρωτόκολλο TCP/IP. Το πρωτόκολλο αυτό αποτελεί μία σύμβαση η οποία καθορίζει τους τρόπους με τους οποίους θα διακινούνται τα δεδομένα, το πώς θα γίνεται ο έλεγχος και η παραλαβή τους στη μεριά του δέκτη (ΚΕΠΛΗΝΕΤ Φλώρινας). Όλοι οι υπολογιστές οι οποίοι είναι συνδεδεμένοι στο πρωτόκολλο αυτό είναι ταυτόχρονα συνδεδεμένοι σε δίκτυα πολύ μικρότερα του internet και κατορθώνουν να επικοινωνήσουν μεταξύ τους παρά τη διαφορετικότητά τους.

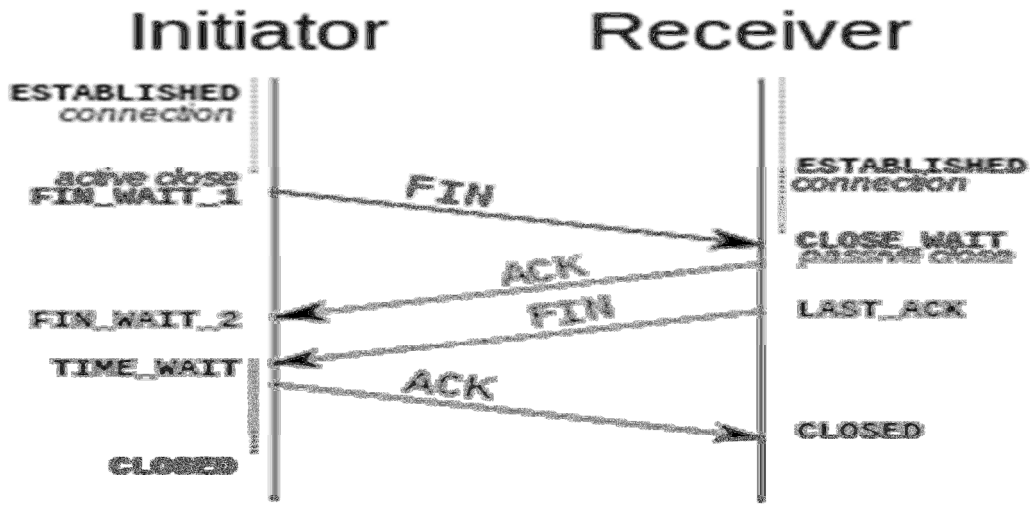
Το internet, έχοντας ως στόχο τη διατήρηση της επικοινωνίας, χρησιμοποιεί την τεχνική της μεταγωγής των πακέτων, για τη μεταφορά της πληροφορίας. Τα δεδομένα χωρίζονται σε κομμάτια τα οποία καλούνται πακέτα. Κάθε πακέτο, περιλαμβάνει πέρα από την πληροφορία, μία επικεφαλίδα η οποία φέρει τις διευθύνσεις του αποστολέα και του παραλήπτη (ΚΕΠΛΗΝΕΤ Φλώρινας). Η μεταφορά του πακέτου αυτού από τον ένα υπολογιστή στον άλλο γίνεται διαμέσου του πρωτοκόλλου IP το οποίο είναι υπεύθυνο και για τη διατήρηση των συνδέσεων.

Το TCP λειτουργεί πάνω από το IP και έχει ως στόχο να εγγυηθεί ότι τα πακέτα θα φτάσουν στον τελικό προορισμό τους, θα έχουν τη σωστή σειρά, και ότι το περιεχόμενό τους θα είναι αναλλοίωτο. Ο τρόπος με τον οποίο λειτουργεί το TCP είναι (ΚΕΠΛΗΝΕΤ Φλώρινας) :

- Αρίθμηση των πακέτων
- Παρακολούθηση της αρίθμησης από τον αποστολέα και από τον παραλήπτη
- Παραλαβή των πακέτων με τη σωστή αρίθμηση από τον παραλήπτη
- Επαναίτηση του πακέτου εάν αυτό χαθεί στη διαδρομή

Η παραπάνω μέθοδος εξασφαλίζει τόσο αξιοπιστία όσο και ταχύτητα κατά τη μεταγωγή των πακέτων καθώς οι ενδιάμεσοι υπολογιστές δε συμμετέχουν σε αυτή.

Για να γίνει η μεταγωγή των πακέτων πάνω από το TCP οι υπολογιστές χρησιμοποιούν εξαρτήματα τα οποία καλούνται routers. Αντικείμενο των routers λοιπόν είναι η δρομολόγηση των πακέτων μέσα από τα διαφορετικά δίκτυα ώστε αυτά να φτάσουν στον παραλήπτη τους.




Εικόνα 2 : Λειτουργία του TCP

Καθώς το internet αποτελείται από χιλιάδες δίκτυα συνδεδμένα μεταξύ τους και άρα εκατομμύρια υπολογιστών, ο κάθε υπολογιστής θα πρέπει να αντιστοιχιστεί με μια μοναδική διεύθυνση η οποία καλείται IP και αποτελεί την ταυτότητα με την οποία αυτός είναι αναγνωρίσιμος στο διαδίκτυο.

Κάθε διεύθυνση IP αποτελείται από 4 ομάδες αριθμών οι οποίες διαχωρίζονται μεταξύ τους με τελείες (University of Virginia). Στην πραγματικότητα, κάθε διεύθυνση IP είναι ένα δυαδικός αριθμός μεγέθους 32 bit ο οποίος για να γίνει κατανοητός στους ανθρώπους μεταφράζεται στο δεκαδικό σύστημα. Στην IP διεύθυνση βρίσκεται κρυμμένη πληροφορία σχετικά με τον αριθμό του δικτύου στο οποίο υπάγεται ο υπολογιστής και τον αριθμό του ίδιου του υπολογιστή (University of Virginia)..

Το μήκος του αριθμού του δικτύου και του υπολογιστή εξαρτάται από το δίκτυο στο οποίο ανήκουν αυτοί και για το λόγο αυτό παρουσιάζονται τρεις διαφορετικές κλάσεις δικτύων ανάλογα με το μέγεθος αυτών (University of Virginia)..

**Κλάση Α** (μεγάλα)

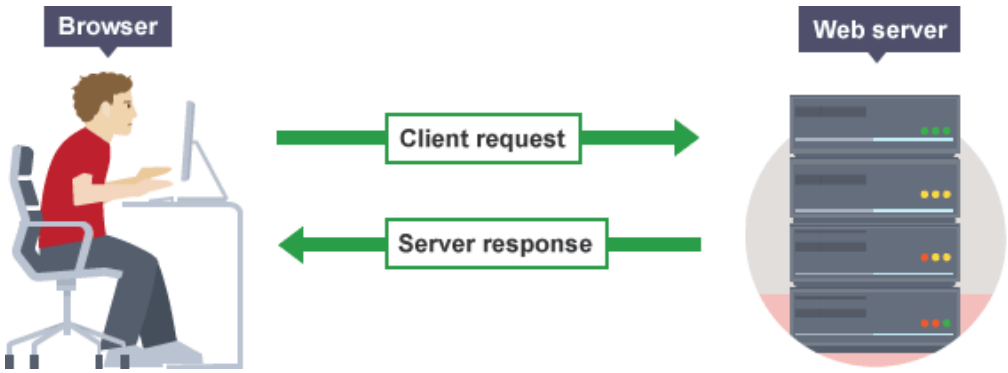
**Κλάση Β** (μεσαία)

- **Κλάση C** (μικρά)

Όπως είναι αντιληπτό, δεν είναι εύκολο για οποιοδήποτε χρήστη οποιαδήποτε στιγμή να θυμηθεί ή/και να αποστηθίσει μία τέτοια διεύθυνση. Για το λόγο αυτό λοιπόν χρησιμοποιήθηκε η τεχνική της ονοματοδοσίας των υπολογιστών διαμέσου μιας μεθόδου που καλείται DNS (Domain Name System). Ένα όνομα υπολογιστή, αποτελείται από λέξεις, από 3 έως 5, οι οποίες διαχωρίζονται μεταξύ τους με τελείες (University of Virginia)..

Βέβαια, όπως έχει προειπωθεί, η δρομολόγηση των πακέτων γίνεται με βάση την IP του παραλήπτη και άρα όταν ζητείται η επικοινωνία με έναν απομακρυσμένο υπολογιστή η IP αυτού θα πρέπει να είναι γνωστή. Έτσι εάν ο χρήστης πληκτρολογήσει πληκτρολογήσει ένα όνομα υπολογιστή όπως π.χ www.facebook.com αυτό μεταφράζεται σε μία IP διεύθυνση όπως 30.127.43.9. Για τη μετάφραση αυτή, υπεύθυνος είναι ο DNS, ο οποίος παρέχει την υπηρεσία της μετάφρασης.

Το μοντέλο πάνω στο οποίο βασίζεται το internet ώστε να παρέχει μία τεράστια γκάμα υπηρεσιών για όλους τους χρήστες καλείται μοντέλο client – server, στα ελληνικά πελάτη και εξυπηρετητή (Πανεπιστήμιο Θεσσαλίας, 1997). Προκειμένου να λειτουργήσει η υπηρεσία αυτή, θα πρέπει στο υπολογιστή να υπάρχει η υπηρεσία client η οποία ζητά την πρόσβαση σε συγκεκριμένη υπηρεσία και στη μεριά του internet να υπάρχει η υπηρεσία η οποία θα εξυπηρετήσει το αίτημα αυτό και καλείται server.




Εικόνα 3 : Μοντέλο client – server

Οι κυριότερες υπηρεσίες που παρέχει το internet είναι (Πανεπιστήμιο Θεσσαλίας, 1997):

- **E-mail** (Ηλεκτρονικό Ταχυδρομείο) : Υποστηρίζει την ανταλλαγή μηνυμάτων μεταξύ χρηστών χάρη στην προσωπική ηλεκτρονική διεύθυνση του καθενός. Το περιεχόμενο του μηνύματος μπορεί να είναι κείμενο, ήχος, εικόνα, video ή δεδομένα.

- **Mailing lists** (Λίστες E-mail) : Καθορισμένη ομάδα απομακρυσμένων μεταξύ τους χρηστών που ανταλλάσσουν μηνύματα σχετικά με κάποιο θέμα ορισμένο από κοινού, με κάποιον από αυτούς ως υπεύθυνο για την καλή λειτουργία τηςλίστας.

- **Remote Login** (Τηλεσύνδεση) : Ένας χρήστης "υπάρχει", δηλαδή έχει δικαίωμα χρήσης σε έναν ή περισσότερους υπολογιστές του δικτύου. Αν αυτοί είναι απομακρυσμένοι μεταξύ τους τότε, εργαζόμενος σε έναν από αυτούς, μπορεί να συνδεθεί με οποιονδήποτε από τους υπόλοιπους και να (τηλε-) εργαστεί σαν να ήταν παρών, δηλαδή να χρησιμοποιήσει τις δυνατότητες του απομακρυσμένου υπολογιστή σαν να βρίσκονταν στον ίδιο φυσικό χώρο με αυτόν.

- **Finger :** Αναζήτηση της ύπαρξης ενός συγκεκριμένου χρήστη σε κάποιο σημείο του δικτύου.

- **FTP** (File Transfer Protocol) : Μεταφορά αρχείων από απομακρυσμένο υπολογιστή σε τοπικό υπολογιστή και αντίστροφα.

- **Archie :** Αναζήτηση υπολογιστών στο Internet που προσφέρουν την υπηρεσία FTP και περιέχουν πληροφορίες με περιεχόμενο οριζόμενο από το χρήστη.

- **Usenet :** Ανταλλαγή μηνυμάτων οργανωμένη σε "οικογένειες ηλεκτρονικών συζητήσεων" με εξαιρετική ποικιλία θεμάτων προς συζήτηση και παγκόσμια συμμετοχή (πάνω από 10000 ηλεκτρονικές συζητήσεις).

- **Talk :** Ανταλλαγή μηνυμάτων κειμένου σε πραγματικό χρόνο μεταξύ δύο χρηστών που βρίσκονται σε απομακρυσμένα σημεία του Internet.

- **Veronica :** Αναζήτηση υπολογιστών του δικτύου που προσφέρουν την υπηρεσία Gopher, σχετικά με το θέμα που ορίζει ο χρήστης.

- **WWW** (World Wide Web) : Διαδικτυωμένες ηλεκτρονικές σελίδες με πληροφορίες σε γραφικό παραθυρικό περιβάλλον, οι οποίες αλληλοσυνδέονται μέσω λέξεων - κλειδιών. Αυτή η υπηρεσία ενοποιεί μέσα στο ίδιο λογισμικό τις FTP, Archie, Gopher, E-mail, Usenet, κλπ.

- **MUD** (Multiple User Dimension) : Παιχνίδια με σενάριο στα οποία ο χρήστης που συνδέεται παίρνει ένα προσωπικό ρόλο και αλληλεπιδρά στο περιβάλλον του παιχνιδιού με τους υπόλοιπους απομακρυσμένους παίκτες.

## 1.2 Συστήματα και αλγόριθμοι ανάκλησης πληροφορίας από τον παγκόσμιο ιστό

Η εξόρυξη δεδομένων από τον παγκόσμιο ιστό αποτελεί ένα από τα πλέον ενδιαφέροντα επιστημονικά και ερευνητικά πεδία καθώς όπως είναι γνωστό, το μέγεθος των δεδομένων του παγκόσμιου ιστού είναι σχεδόν αδύνατο να υπολογισθεί. Ο παγκόσμιος ιστός μπορεί να θεωρηθεί ως η μεγαλύτερη βάση δεδομένων η οποία μάλιστα είναι διαθέσιμη σε όλους τους χρήστες. Βέβαια, ο όρος βάση δεδομένων στην προκειμένη απλουστεύει αυτό που πραγματικά είναι ο παγκόσμιος ιστός καθώς σε αυτόν δεν υπάρχει μία συγκεκριμένη δομή.

Το παραπάνω καθιστά ακόμη πιο επιτακτική την ανάγκη εξόρυξης δεδομένων απ τον ιστό ενώ παράλληλα προσφέρει σημαντική βοήθεια σε κάθε χρήστη. Φυσικά, η εξόρυξη στον παγκόσμιο στο δεν αναφέρεται μονάχα σε ιστοσελίδες αλλά και σε δεδομένα που αφορούν στην κινήσεις των χρηστών μέσα σε αυτόν. Τα δεδομένα διαδικτύου μπορούν να χωριστούν στις ακόλουθες κατηγορίες (Diebold & Kaufmann, 2001):

- Περιεχόμενο ιστοσελίδων.
- Ενδοπληροφορία ιστοσελίδων (HTML/XML κώδικας).
- Εσωτερική δομή ιστοσελίδων, δηλαδή των πως διασυνδέονται μεταξύ τους.
- Δεδομένα χρήσης που περιγράφουν πως οι επισκέπτες προσπελαύνουν τις ιστοσελίδες.
- Προφίλ χρηστών που περιλαμβάνουν δημογραφικά δεδομένα και πληροφορίες εγγραφών (εδώ περιέχονται και πληροφορίες από cookies αρχεία).

Η εξόρυξη των δεδομένων από το διαδίκτυο μπορεί φυσικά να γίνει με διαφορετικές τεχνικές. Μία εκ των πλέον γνωστών τεχνικών είναι το web content mining, το οποίο εξετάζει τα περιεχόμενα των ιστοσελίδων αλλά και τα αποτελέσματα των αναζητήσεων (Stumme, Hotho, & Berendt, 2002). Η τεχνική αυτή εξόρυξης, έρχεται αντιμέτωπη με σειρά προκλήσεων καθώς αφενός είναι απροσδιόριστο το μέγεθος των ιστοσελίδων και αφετέρου είναι ανομοιόμορφη η δομή τους. Επιπρόσθετα, τα ίδια κείμενα υφίστανται σε πολλές και διαφορετικές εκδόσεις ενώ την ίδια στιγμή ένα

μεγάλο ποσοστό της πληροφορίας είναι ατελές. Αυτό, καθιστά ακόμη πιο επιτακτική την ανάγκη ώστε να χρησιμοποιηθούν τεχνικές που θα παράγουν ορθά και ακριβή αποτελέσματα (Stumme, Hotho, & Berendt, 2002).

Πέραν τούτου φυσικά, υφίσταται και το deep web στο οποίο δεν είναι εφικτή η πρόσβαση σε μηχανές αναζήτησης. Μεγάλο μέρος αυτού είναι ημιδομημένο γεγονός που καθιστά της τεχνικές ευρετηριών σε αυτό σχεδόν αδύνατες. Το web content mining χωρίζεται επιμέρους στο web page content mining και στο search result mining (Adomavicius & Tuzhilin). Το πρώτο είναι η παραδοσιακή αναζήτηση ιστοσελίδων σύμφωνα με το περιεχόμενο τους, ενώ το δεύτερο είναι περαιτέρω αναζήτηση σε ιστοσελίδες που είναι αποτέλεσμα προηγούμενης αναζήτησης.

Μία ακόμη εκ των γνωστών τεχνικών είναι το web structure mining, το οποίο εστιάζει στη χρήση της ανάλυσης της δομής των συνδέσμων του διαδικτύου. Οι υπερσύνδεσμοι, είναι το εργαλείο εκείνο το οποίο χρησιμοποιείται ώστε να αναγνωστούν οι σχεσεις συσχέτισης μεταξύ κειμένων, να ομαδοποιηθούν τα κείμενα

ανάλογα με τη σημαντικότητα τους και τέλος για να υπάρχει άμεση διασύνδεση με τις διάφορες παραπομπές (Adomavicius & Tuzhilin).

Το web usage mining, προβαίνει σε επεξεργασία των διαφόρων log files που περιέχουν τα δεδομένα πρόσβασης των χρηστών στις ιστοσελίδες. Η τεχνική αυτή, αποσκοπεί στην κατανόηση της συμπεριφοράς του χρήστη αλλά και στην κατανόηση του τρόπου με τον οποίο δομείται η πληροφορία (Kamdar, 2002). Τα δεδομένα των click-stream, τα cookies, τα ερωτήματα των χρηστών, και κάθε είδους δεδομένα σχετικά με τα αποτελέσματα της αλληλεπίδρασης μεταξύ ανθρώπου και διαδικτύου χρησιμοποιούνται με στόχο τονιστούν οι ανάγκες των πελατών και να βελτιωθεί η ποιότητα των παρεχόμενων προς αυτούς υπηρεσιών.

Το general access pattern tracking είναι και αυτό μία τεχνική που ανήκει στην κατηγορία των web usage mining που ασχολείται αποκλειστικά με την επισκεψιμότητα των ιστοσελίδων αλλά και το ιστορικό. Η χρήση προτυποποιείται με στόχο την εξόρυξη βασισμένη πάντα σε συγκεκριμένα πρότυπα.

Η διαδικασία της εξόρυξης δεδομένων, όπως φυσικά και κάθε άλλη διαδικασία, δε θα μπορούσε παρά να αποτελείται από συγκεκριμένα και φυσικά διαδοχικά στάδια. Τα στάδια αυτά είναι η συλλογή δεδομένων, η προεπεξεργασία αυτών, η ανακάλυψη προτύπων και η εκμετάλλευση της γνώσης (Kamdar, 2002).

Η πλέον γνωστή πηγή για τη συλλογή των δεδομένων είναι οι διακομιστές διαδικτύου οι οποίοι διαθέτουν πρόσβαση στα log files αλλά και στις βάσεις δεδομένων τους και μπορούν να συλλέξουν την πληροφορία από εκεί. Στα log files, συνήθως περιέχεται πληροφορία όπως το όνομα, η IP διεύθυνση, η ημερομηνία και ο χρόνος που έγινε το αίτημα η οποία αναπαρίσταται με μορφή common log format, extended log format, logml (Kamdar, 2002).

Ένα σημαντικό ζήτημα που καλείται να αντιμετωπιστεί σχετικά πάντα με τη χρήση δεδομένων από τα log files είναι η αναγνώριση των session των χρηστών. Αυτή η εργασία είναι πολλές φορές δύσκολη και δαπανηρή τόσο σε χρόνο όσο και σε χώρο και εξαρτάται πολύ στον τύπο της πληροφορίας που ένας διακομιστής μπορεί να συγκρατήσει..

Η πλέον χρησιμοποιούμενη τεχνική είναι αυτή των cookies, τα οποία καταγράφουν τα αιτήματα που γίνονται σε κάποιες ιστοσελίδες. Ωστόσο ακόμα και αν χρησιμοποιηθούν τα cookies, είναι μερικές φορές αδύνατο να γίνουν γνωστές οι ακριβείς κινήσεις των χρηστών σε μια ιστοσελίδα μιας και η προς τα πίσω κίνηση ενός χρήστη (back) δεν καταγράφεται στο διακομιστή.

Πέραν των log αρχείων, η συμπεριφορά του χρήστη είναι εφικτό να καταγραφεί πολλές φορές από τα πακέτα που αποστέλλονται πάνω από το TCP/IP. Η τεχνική αυτή θεωρείται αρκετά αποτελεσματική και εισάγει τα παρακάτω πλεονεκτήματα (Bounsaythip & Rinta-Runsala, 2001) :

1) Τα δεδομένα συλλέγονται σε πραγματικό χρόνο
2) Οι πληροφορίες που προέρχονται από διαφορετικούς διακομιστές μπορούν να ενωθούν σε κοινά αρχεία
3) Η χρήση ειδικών κουμπιών όπως αυτού της προς τα πίσω περιήγησης (back) μπορούν να ανιχνευθούν.

Πολλοί από τους παρόχους υπηρεσιών διαδικτύου εκχωρούν στους πελάτες τους δικαιώματα χρήσης υπηρεσιών proxy διακομιστών μεσολάβησης (proxy servers) ώστε αυτοί να είναι σε θέση να βελτιώσουν την ταχύτητα περιήγησης (Bounsaythip & Rinta-Runsala, 2001). Η συλλογή δεδομένων από το διακομιστή είναι ίδια με οποιαδήποτε συλλογή από το διαδίκτυο με την διαφορά να υφίσταται στο γεγονός ότι

στους proxy διακομιστές συλλέγονται δεδομένα από ομάδες χρηστών που έχουν πρόσβαση σε τεράστιες ομάδες από διακομιστές διαδικτύου. Και στην περίπτωση αυτή δεν είναι εύκολο να ανακατασκευαστεί οποιοδήποτε session παρόλα αυτά όμως πρόκειται για μια μέθοδο που θεωρείται σχετικά εύκολη στην αναγνώριση των συνόδων των χρηστών.

Η καταγραφή των δεδομένων μπορεί φυσικά να πραγματοποιηθεί και από τη μεριά του χρήστη με χρήση της γλώσσας Javascript ή με Java Applets (Anderson, 2002). Οι μέθοδοι αυτοί καταφέρνουν να υπερπηδήσουν τα προβλήματα που δημιουργούνται κατά την αναγνώριση των sessions καθώς και τα όποια προβλήματα δημιουργούνται από τη χρήση της κρυφής μνήμης. Παράλληλα, παρέχουν πληροφορίες σχετικά με τη συμπεριφορά του χρήστη αλλά δυστυχώς βασίζονται ξεκάθαρα σε αυτή ενώ ενέχουν θέματα ιδιωτικότητας.

Επόμενο στάδιο κατά την εξόρυξη της πληροφορίας από τον παγκόσμιο ιστό είναι η προεπεξεργασία των δεδομένων που είναι μία ιδιαίτερα χρονοβόρα και πολύπλοκη διαδικασία. Ξεκινά με τον καθαρισμό των δεδομένων, ο οποίος αφορά στη διαγραφή όλων εκείνων των δεδομένων που δεν θεωρούνται απαραίτητα. Στη συνέχεια, αναγνωρίζονται τα sessions των διαφορετικών χρηστών από τα log files και γίνεται προσπάθεια αναπαράστασης των path που ακολουθήθηκαν. Το στάδιο αυτό είναι ιδιαίτερα πολύπλοκο με τα σοβαρότερα προβλήματα να εισάγονται από την επαναποθήκευση (caching) που γίνεται είτε στους proxy διακομιστές είτε στα προγράμματα περιήγησης (browsers) (Anderson, 2002).

Το παραπάνω πρόβλημα, αντιμετωπίζεται συνήθως με χρήση των cookies, επανεγγραφή της, ή απαιτώντας από το χρήστη να κάνει σύνδεση στις σελίδες με κάποιο συνθηματικό (Anderson, 2002). Ως γνωστόν, το κάθε cookie, είναι ένα αρχείο που συλλέγει τα δεδομένα που αποστέλλονται από το διακομιστή διαδικτύου στο πρόγραμμα περιήγησης. Η εν λόγω πληροφορία, αποθηκεύεται στον υπολογιστή ως ένα αρχείο κειμένου και πολλές φορές είναι αρκετό ώστε να εξαχθεί σημαντική πληροφορία. Η πληροφορία που βρίσκεται εμφωλευμένη στο cookie, παρέχεται στο διακομιστή κάθε φορά που αυτός κάνει αίτημα ενώ παράλληλα αποθηκεύεται στο log file (Anderson, 2002).

Η πλειοψηφία των εφαρμογών που προβαίνει σε εξόρυξη δεδομένων από τον ιστό κάνει χρήση των διευθύνσεων με βάση τις οποίες έχουν γίνει τα αιτήματα ώστε να

εφαρμόσει εκεί τις τεχνικές εξόρυξης. Δυστυχώς, οι διευθύνσεις πολλές φορές, δεν προσφέρουν πληροφορία αλλά ούτε και μεταφέρουν πληροφορία σχετική πάντα με το περιεχόμενο των σελίδων. Το στάδιο της προεπεξεργασίας των δεδομένων, τελειώνει με την μορφοποίηση των δεδομένων που ακολουθεί όλες τις προηγούμενες.

Ακολουθεί η ανακάλυψη προτύπων που είναι και το σημαντικότερο στάδιο της εξόρυξης καθώς σε αυτό συντελείται η ανακάλυψη της επιθυμητής γνώσης από τα δεδομένα. Στο στάδιο αυτό, χρησιμοποιούνται μάλιστα, πολλές τεχνικές τόσο από το επιστημονικό πεδίο της μηχανικής μάθησης όσο και από το επιστημονικό πεδίο της στατιστικής (Srivastava & Deshpande, 2000).

Γνώση λοιπόν για τις εφαρμογές εξόρυξης από τον παγκόσμιο ιστό θεωρούνται τα πρότυπα εκείνα που αντικατοπτρίζουν τυποποιημένες συμπεριφορές χρηστών κατά την περιήγησή τους στον ιστό. Οι πιο πολλές από τις εμπορικές εφαρμογές στον τομέα του web usage mining εκμεταλλεύονται συνδυασμούς τεχνικών ανάλυσης ενώ αντίθετα η έρευνα σε αυτή την περιοχή περισσότερο εστιάζει στην δημιουργία τεχνικών ανακάλυψης γνώσης για την ανάλυση των web usage δεδομένων (Srivastava & Deshpande, 2000)..

Τέλος, ακολουθεί η ερμηνεία και η αξιολόγηση όλης εκείνης της γνώσεις που έχει αποκτηθεί σε μορφή που να είναι κατανοητή και να μπορεί να χρησιμοποιηθεί περαιτέρω.

# Κεφάλαιο 2 – Το Deep Web

## 2.1 Περιγραφή

Το Deep Web έγινε ιδιαίτερα γνωστό το 2013 όταν το FBI εντόπισε και έκλεισε την αγορά Silk Road η οποία διακινούσε τεράστιες ποσότητες ναρκωτικών στο διαδίκτυο. Ο Ross Ulbricht, γνωστός και ως Dread Pirate Roberts, κατηγορήθηκε για διακίνηση ναρκωτικών, για κατασκοπεία στο διαδίκτυο και για ξέπλυμα χρήματος (Σιδηρόπουλος).

Οι αναφορές στο Deep Web είναι πάρα πολλές για να καταφέρει κανείς να εισχωρήσει σε αυτό θα πρέπει να κάνει συγκεκριμένη αναζήτηση με λογισμικά όπως το TOR (Bergman, 2001). Η πρόσβαση στο Deep Web δεν είναι δυνατή από απλές μηχανές αναζήτησης. Μάλιστα, στην πραγματικότητα το TOR δημιουργήθηκε με στόχο να διασφαλίσει την επικοινωνία και την ελευθερία του λόγου, η χρήση του όμως δυστυχώς, όπως συμβαίνει και με κάθε άλλο εργαλείο, επαφίεται στην πρόθεση του χρήστη (Bergman, 2001)..

Ο όρος Deep Web αναφέρεται σε οποιοδήποτε περιεχόμενο στο διαδίκτυο που για διάφορους λόγους δεν μπορεί να εντοπιστεί από απλές μηχανές αναζήτησης όπως η Google (Ntoulas, Cho, & Olston, 2004). Εξ ορισμού, το Deep Web περιλαμβάνει, δυναμικές ιστοσελίδες, σελίδες που έχουν αποκλειστεί, σελίδες nonHTML/-contextual/-scripted περιεχομένου, και δίκτυα περιορισμένης πρόσβασης. Τα δίκτυα περιορισμένης πρόσβασης καλύπτουν όλες εκείνες τις πηγές και τις υπηρεσίες που δεν είναι διαθέσιμες στο απλό διαδίκτυο και άρα δίνουν τη δυνατότητα σε κακόβουλου χρήστες να ενεργήσουν ελεύθερα χωρίς την επίβλεψη των αντιπροσώπων του νόμου (Ntoulas, Cho, & Olston, 2004). Στα δίκτυα αυτά συγκαταλέγονται ιστοσελίδες με domain name που δεν έχει καταγραφεί στους DNS (Domain Name Service) roots και δεν είναι διαχειρίσιμες από τον ICANN (Internet Corporation for Assigned Names and Numbers) και άρα δεν μπορούν να εντοπιστούν. Παράδειγμα αποτελούν επίσης και οι ιστοσελίδες που έχουν καταγραφεί σε ένα παντελώς διαφορετικό σύστημα από το DNS όπως το .BIT (Ntoulas, Cho, & Olston, 2004). Τα παραπάνω συστήματα, όχι μόνο δεν ακολουθούν τους κανονισμούς που ορίζονται από τον ICANN αλλά επίσης κάνουν σοβαρές προσπάθειες ώστε να αποκρύψουν τα domain names. Τα δίκτυα περιορισμένης πρόσβασης ανήκουν επίσης

και τα darknets ή η ιστοσελίδες που φιλοξενούνται σε δομές που απαιτούν εξειδικευμένο λογισμικό όπως το TOR.

Το ενδιαφέρον του κοινού αναφορικά με το Deep Web εντοπίζεται κυρίως στο τί συμβαίνει εντός των darknets. Σε αντίθεση με το υπόλοιπο περιεχόμενο του Deep Web, τα δίκτυα περιορισμένης πρόσβασης τα δίκτυα αυτά δεν εντοπίζονται όπως προαναφέρθηκε από τις βασικές μηχανές αναζήτησης αλλά από gateway υπηρεσίες που επιτρέπουν στους χρήστες να έχουν πρόσβαση σε κρυμμένες υπηρεσίες.


Εικόνα 4 : Το Deep Web

Η πλέον γνωστή εικόνα αναφορικά με το Deep Web είναι αυτή του παγόβουνου όπου οτιδήποτε βρίσκεται επάνω από την επιφάνεια της θάλασσας ανήκει στο διαδίκτυο στο οποίο ο καθένας έχει πρόσβαση και ότι βρίσκεται κάτω από την επιφάνεια είναι κρυμμένο, δύσκολο να ανευρεθεί και πολλές φορές αόρατο.

Χρήστες του Deep Web είναι συνήθως άνθρωποι που θα ήθελαν να αγοράσουν ναρκωτικά σε διαδικτυακό επίπεδο χωρίς να συνδέσουν τις λέξεις αυτές με έναν κανονικό browser (Bergman, 2001). Έτσι λοιπόν, αυτοί θα έπρεπε να εισέλθουν ανώνυμα σε μία υποδομή η οποία ποτέ δε θα οδηγήσει κανέναν από τους εμπλεκόμενους στην IP ή τη φυσική τους τοποθεσία. Από την άλλη, όσοι πουλούν ναρκωτικά διαδικτυακά, δε θα ήθελαν να δημιουργήσουν ένα on line κατάστημα στο οποίο η αστυνομία θα μπορούσε να έχει άμεση πρόσβαση εντοπίζοντας και στην περίπτωση αυτή είτε την IP είτε τη φυσική τους διεύθυνση. Υπάρχουν βέβαια και

πολλοί ακόμη λόγοι για τους οποίους οι άνθρωποι θα επέλεγαν να παραμείνουν ανώνυμοι ή να στήσουν ιστοσελίδες με μη εντοπίσιμο περιεχόμενο (Bergman, 2001).

Σε οποιαδήποτε συζήτηση για το Deep Web φυσικά και εμφανίζεται και το κανονικό διαδίκτυο το οποίο φυσικά αποτελεί το ακριβώς αντίθετο, το κομμάτι δηλαδή εκείνο το διαδικτύου στο οποίο οι μηχανές αναζήτησης εντοπίζουν οποιοδήποτε περιεχόμενο το οποίο μάλιστα θεωρείται και καθαρό.

Οι έρευνες γύρω από το Deep Web είναι φυσικά πάρα πολλές και φυσικά δημιουργούν πως η πλειοψηφία των ιστοσελίδων που λαμβάνουν χώρα σε αυτό αφορούν στην πώληση παράνομων ναρκωτικών ή όπλων. Η αλήθεια όμως δεν είναι ακριβώς αυτή. Φυσικά, υπάρχει και αυτό το κομμάτι, υπάρχει όμως και ένα κομμάτι του Deep Web που ασχολείται με προσωπικά ή πολιτικά blogs, forums, ιστοσελίδες θρησκευτικού περιεχομένου ή ακόμη και ραδιοσταθμούς (Bergman, 2001).


Εικόνα 5 : Το Deep Web Radio

Λόγω της φύσης του Deep Web είναι αδύνατο να καθοριστεί ο αριθμός των ιστοσελίδων που συμμετέχουν σε αυτό αλλά και το περιεχόμενο που είναι διαθέσιμο ανά πάση στιγμή στους χρήστες. Πέραν τούτου, η φύση του δικτύου αυτό, είναι τέτοια ώστε κανείς να μη μπορεί να υποστηρίξει πως αυτό έχει εξερευνηθεί στο μέγιστο βάθος του.

Στα πλαίσια της αναζήτησης του περιεχομένου του Deep Web, η Trend Micro Forward-Looking Threat Research Team δημιούργησε ένα σύστημα που συλλέγει όλα τα URLs που συνδέονται με αυτό, συμπεριλαμβανομένων και των TOR και των I2P hidden ιστοσελίδων, αλλά και των domain χωρίς TLD και προσπάθησε να εξάγει

πληροφορία σχετική με τα domains. Το αποτέλεσμα ήταν η συλλογή 38 εκατομμυρίων γεγονότων που συνδέονται με 576,000 URLs, από τα οποία τα 244,000 έχουν πραγματικό HTML περιεχόμενο (TrendMicro, 2016).

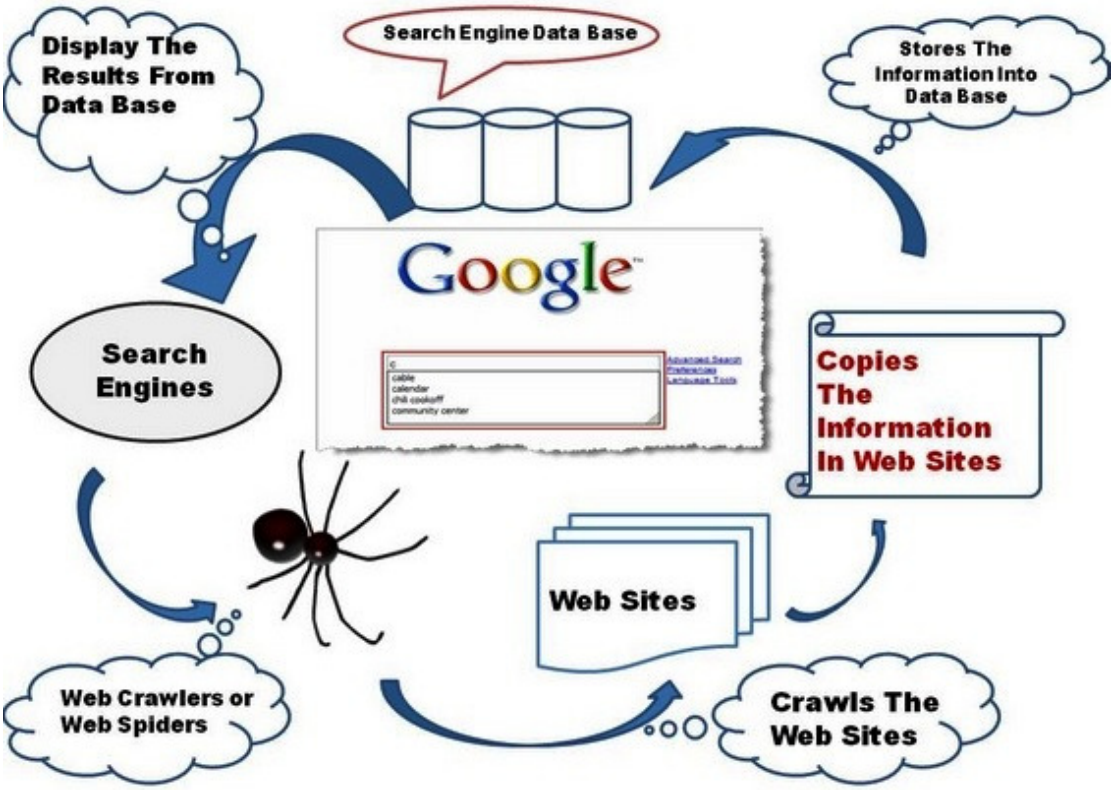
Όπως είναι αντιληπτό, είναι σχεδόν ακατόρθωτο να υποστηρίξει κάποιος ποιος ακριβώς βρίσκεται ή συμμετέχει στο Dark Web. Το χαρακτηριστικό της ανωνυμίας που το διακατέχει το κάνει ακόμη πιο ελκυστικό για τους χρήστες ή ακόμη και για τους ίδιους τους ερευνητές σε θέματα ασφάλειας. Κύρια γλώσσα που επικρατεί στο Dark Web είναι αυτή των αγγλικών, με αυτές των Ρώσικων και των Γαλλικών να ακολουθούν.

Η πλειοψηφία των συναλλαγών που γίνονται εντός του Dark Web περιλαμβάνουν αγαθά για τα οποία οι άνθρωποι θα απαιτούσαν να αποκρυφθεί στο μέγιστο δυνατό βαθμό η ταυτότητά τους.

## 2.2Αρχιτεκτονική και Πρωτόκολλα λειτουργίας

Η πρόσβαση στο περιεχόμενο του deep web γίνεται διαμέσου των crawler οι οποίοι μπορούν να αλληλεπιδράσουν με τις σελίδες αυτές και να εξάγουν πληροφορία. Βέβαια, ο κάθε crawler θα πρέπει να είναι σε να αλληλεπιδρά την ίδια στιγμή με τις φόρμες αναζήτησης που υπάρχουν στις ιστοσελίδες που φιλοξενούνται στο deep web.

Το κύριο χαρακτηριστικό των ιστοσελίδων που φιλοξενούνται στο deep web είναι οι φόρμες αναζήτησης που περιέχουν βάση των οποίων μάλιστα κατηγοριοποιούνται και οι ιστοσελίδες. Έτσι υπάρχουν ιστοσελίδες δομημένες με δομημένες φόρμες αναζήτησης, και ιστοσελίδες αδόμητες με αδόμητες φόρμες αναζήτησης.




Εικόνα 6 : Web Crawler

Η αναζήτηση στις ιστοσελίδες αυτές πραγματοποιείται διαμέσου μηχανών αναζήτησης που πραγματοποιούν τα παρακάτω βήματα (Sidiropoulou & Pohle, 2001):

- Εντοπίζουν τις πηγές του deep web περιεχομένου : Ένας άνθρωπος ή ένας crawler θα πρέπει να εντοπίσουν τις ιστοσελίδες εκείνες οι οποίες περιέχουν φόρμες αναζήτησης οι οποίες οδηγούν σε περιεχόμενο Deep Web, δηλαδή φόρμες αναζήτησης οι οποίες είναι συνδεδεμένες με κάποια βάση δεδομένων.

- Επιλέγουν τις σχετικές πηγές περιεχομένου : Για μια στοχευμένη Deeb Web αναζήτηση ( π.χ. , αναζήτηση ιατρικών άρθρων ), πρέπει να επιλεγεί ένα υποσύνολο των διαθέσιμων πηγών περιεχομένου. Στην περίπτωση του οδόμητου περιεχομένου το προβλημα είναι γνωστό ως πρόβλημα επιλογής βάσης δεδομένων ή πηγής πληροφορίας. Το πρώτο βήμα στην επιλογή πηγής περιεχόμένου είναι να μοντελοποιηθεί το διαθέσιμο περιεχόμενο σε μια μεμονωμένη Deeb Web ιστοσελίδα, π.χ. χρησιμοποιώντας δειγματολειψία βασισμένη σε ερωτήματα αναζήτησης.

- Εξάγουν το προσδοκώμενο αποτέλεσμα : Τέλος, ο crawler πρέπει να εξάγει το περιεχόμενο που βρίσκεται πίσω από την φόρμα αναζήτησης της επιλεγμένης πηγής περιεχομένου, δηλαδή να εξάγει τα αποτελέσματα που επιστρέφονται μετά την εκτέλεση της φόρμας αναζήτησης.

Η εξαγωγή του περιεχομένου αποτελεί την πλέον σύνθετη διαδικασία στο deep web και διεξάγεται με τα παρακάτω βήματα (Sidiropoulou & Pohle, 2001):

- Επιλογή ενός υποσυνόλου της φόρμα με στόχο να αφαιρεθούν τα στοιχεία εκείνα που επηρεάζουν των τρόπο παρουσίασης των αποτελεσμάτων αλλά και να μη συμπεριληφθούν στην αναζήτηση στοιχεία που θα αυξήσουν το χρόνο αυτής.

- Δημιουργία μιας βάσης δεδομένων με έγκυρες τιμές δεδομένων. Κάποιες από τις τιμές των πληροφοριών περιλαμβάνουν κάποιον διαχειριστή και περιεχόμενο περιεχόμενο το οποίο δεν προέρχεται online από το deep web, π.χ., ένα λεξικό ( με μη δομημένες λέξεις-κλειδιά ) ή τη λίστα με τους

αγαπημένους συντάκτες κάποιου, (γ) drop-down μενού για τη συμπλήρωση στοιχείων της φόρμας (π.χ., μια αναπτυσσόμενη λίστα των εκδοτών).

- Χρήση της βάσης δεδομένων που δημιουργήθηκε ώστε να συμπληρωθεί η φόρμα αναζήτησης, ανάλυση του αποτελέσματος και εξαγωγή των νέων τιμών ώστε αυτές να εισέλθουν στην βάση δεδομένων.

Κάθε ιστοσελίδα στο deep web αποτελείται από ένα ή/και περισσότερα αντικείμενα περιεχομένου τα οποία δύνανται να είναι είτε αδόμητα έγγραφα είτε δομημένες εγγραφές δεδομένων. Σε ένα αντικείμενο δεδομένων περιλαμβάνονται μεμονωμένες τιμές δεδομένων που αποτελούν είτε όρους κειμένου εάν πρόκειται για αδόμητες εγγραφές είτε εγγραφές δεδομένων όπως ονόματα και ημερομηνίες εάν πρόκειται για αδόμητες εγγραφές.

Στη φόρμα αναζήτησης, υποβάλλεται ένα ερώτημα το οποίο αποτελείται από μία μόνο τιμή δεδομένων V και επιστρέφει όλο το σύνολο των τιμών που αντιστοιχούν στο ερώτημα αυτό. Κάθε ερώτημα που υποβάλλεται, επιβαρύνει το πρόγραμμα αναζήτησης με κόστος X το οποίο προκύπτει από την επεξεργασία των αποτελεσμάτων και από το πλήθος αυτών (Sidiropoulou & Pohle, 2001).

Καθώς το μέγεθος του παγκόσμιου ιστού είναι τεράστιο και πολλές φορές αδύνατο να προσμετρηθεί ο αριθμός των προσβάσιμων εγγραφών έχει ξεπεράσει κατά πολύ το ένα δισεκατομμύριο και αυξάνεται καθημερινά με ραγδαίους ρυθμούς. Η ραγδαία αυτή αύξηση καθιστά την ανάπτυξη καινούριων μηχανών αναζήτησης μία δύσκολη διαδικασία καθώς το περιεχόμενο αυτών είναι σχεδόν αδύνατο να ενημερωθεί. Η τεχνική που έχει βρεθεί τα τελευταία χρόνια στο επίκεντρο της έρευνας όσον αφορά την αναζήτηση στο Web είναι το directed crawling ή web harvesting (Olston & Najork, 2010).

Η συγκεκριμένη τεχνική κάνει χρήση ευφυών προγραμμάτων τα οποία διατρέχουν τον ιστό με αυτοματοποιημένο τρόπο και καλούνται directed crawlers. Η αναζήτηση που πραγματοποιείται με χρήση των directed crawlers έχει ως στόχο σελίδες που συνθέτουν μια εξειδικευμένη συλλογή γνώσης και αναφέρονται σε πολύ συγκεκριμένη θεματολογία (Olston & Najork, 2010).

. Οι διαφορά των directed crawlers με τους απλούς crawlers έγκειται στο γεγονός ότι οι απλοί χρησιμοποιούνται από τις γενικές μηχανές αναζήτησης και πραγματοποιούν την αναζήτησή τους χωρίς να στοχεύουν σε ιστοσελίδες με συγκεκριμένη θεματολογία. Χαρακτηριστικά του ιστού που καθιστούν το directed crawling πολύ δύσκολο είναι (Olston & Najork, 2010):

- Ο μεγάλος όγκος του.
- Ο ταχύς ρυθμός μεταβολής του.
- Η δυναμική δημιουργία σελίδων από χρήστες του Internet.

Το σημαντικότερο εκ των προβλημάτων που καλείται να λύσει το directed crawling είναι αυτό του προσδιορισμού του βάθους στο οποίο θα γίνει το crawl. Το βάθος αναζήτησης είναι μία ιδιαίτερα σημαντική παράμετρος και πρέπει να ορίζεται εξαρχής ώστε να μην υπάρχει εξαντλητική αναζήτηση των υπερσυνδέσμων, να μην καταναλώνονται επιπλέον πόροι και να μην μεγαλώνει ο χρόνος εκτέλεσης της διαδικασίας.

Ακόμη, ιδανικό θα ήταν να λυθούν πρόβλήματα όπως η επιλογή αρχικών σελίδων αλλά και η σωστή ανάθεση των URL ώστε η αναζήτηση να είναι πιο αποδοτική.

## 2.3 Surface vs Deep Web

Έρευνες που έγιναν τα τελευταία χρόνια σχετικά πάντα με το Deep Web έδειξαν ότι το περιεχόμενο αυτού είναι ιδιαίτερα μεγάλο, πολύ μεγαλύτερο μάλιστα από ότι φανταζόντουσαν ή περίμεναν οι ερευνητές. Μάλιστα, εικάζεται πως το περιεχόμενο στον αόρατο ιστό μπορεί να είναι έως και 500 φορές μεγαλύτερο από αυτό που είναι προσβάσιμο από τις μηχανές αναζήτησης στον απλό ιστό. Ένα ακόμη σημαντικό στοιχείο σχετικά με τον αόρατο ιστό είναι ότι ο αριθμός των συνδέσεων σε αυτόν παρουσιάζεται να αυξάνεται με ραγδαίους ρυθμούς (Σιδηρόπουλος). Το παραπάνω διαφαίνεται να είναι αρκετά ανησυχητικό καθώς υπάρχει έντονη κινητικότητα στον αόρατο ιστό και σειρά αδικημάτων, ποινικών ή/και άλλων φαίνεται να συντελείται μέσα σε αυτόν. Επιπλέον στοιχεία τα οποία προέρχονται από παρόμοιες μελέτες και συγκρίνουν τον αόρατο ιστό με τον δημόσιο είναι τα εξής (Σιδηρόπουλος):

- ✓ Ο αόρατος ιστός περιέχει 7.500 terabytes πληροφορίας, ενώ ο δημόσιος ιστός μόλις 19 terabytes.

- ✓ Ο αόρατος ιστός περιέχει περίπου 500 δισεκατομμύρια διαφορετικά αρχεία, ενώ ο δημόσιος ιστός 1 δισεκατομμύριο.

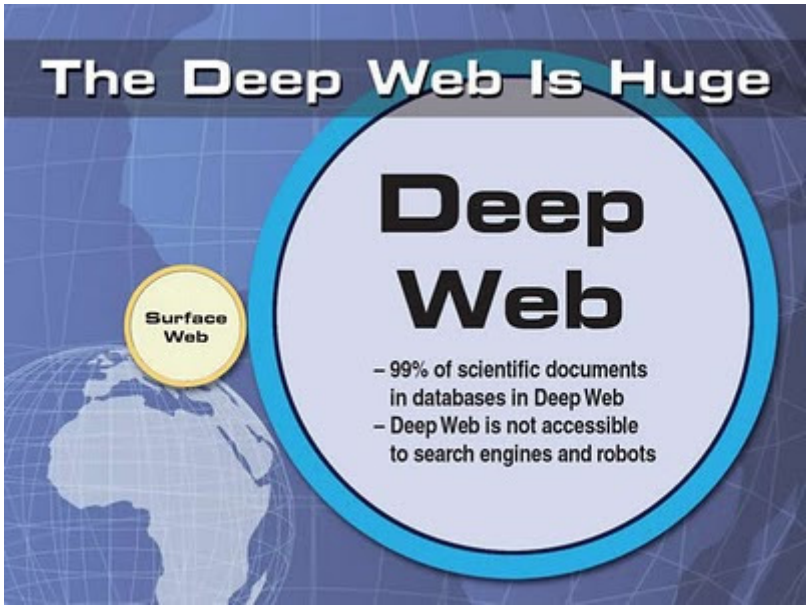
- ✓ Οι 60 μεγαλύτερες ιστοσελίδες του αόρατου ιστού περιέχουν αθροιστικά 750 terabytes πληροφορίας από μόνες τους, αριθμός ο οποίος είναι κατά 40 φορές μεγαλύτερος από το περιεχόμενο του δημόσιου ιστού στο σύνολό του.

- ✓ Κατά μέσο όρο, οι σελίδες του αόρατου ιστού δέχονται μηνιαίος 50 τις εκατό περισσότερους επισκέπτες και είναι πιο ισχυρά συνδεδεμένες με άλλες σελίδες σε σύγκριση με τις σελίδες του δημόσιου ιστού. Ωστόσο παρά το γεγονός αυτό δεν είναι ευρέως γνωστές στους χρήστες που κάνουν αναζήτηση πληροφοριών στο διαδίκτυο.

- ✓ Ο αόρατος ιστός είναι η μεγαλύτερη αυξανόμενη κατηγορία πληροφορίας στο διαδίκτυο.

- ✓ Οι σελίδες του αόρατου ιστού τείνουν να είναι πιο μικρές, με βαθύτερο περιεχόμενο.

- ✓ Το συνολικό ποιοτικό περιεχόμενο του αόρατου ιστού είναι 1.000 με 2.000 φορές μεγαλύτερο από το αντίστοιχο του δημόσιου ιστού




Εικόνα 7 : Deep Web VS Surface Web

Το περιεχόμενο το οποίο περιέχεται στον αόρατο ιστό είναι συνήθως της παρακάτω μορφής (Surface Web, Deep Web, Dark Web -- What's the Difference?, 2016) :

•**Δυναμικό περιεχόμενο**: δυναμικές σελίδες οι οποίες επιστρέφονται σαν απάντηση (response) σε ένα ερώτημα (query) ή μπορούν να προσπελαστούν μόνο μέσω μιας φόρμας.

•**Μη συνδεδεμένο περιεχόμενο (unlinked content):** web σελίδες οι οποίες δεν είναι συνδεδεμένες μέσω links με άλλες σελίδες, κάτι που μπορεί να εμποδίσει

τους crawlers των μηχανών αναζήτησης από την πρόσβαση στο περιεχόμενό τους.

•**Περιεχόμενο περιορισμένης πρόσβασης**: sites τα οποία απαιτούν registration ή εμποδίζουν εσκεμμένα τις μηχανές αναζήτησης να προσπελάσουν το περιεχόμενό τους.

•**Scripted περιεχόμενο**: σελίδες οι οποίες μπορούν να προσπελαστούν μέσω links τα οποία παράγονται από JavaScript ή Flash και χρειάζονται ειδική μεταχείριση.

•**Περιεχόμενο όχι τύπου κείμενου (non-text)**: multimedia αρχεία, έγγραφα σε μορφή διαφορετική από html format όπως pdf ή doc.

Οι σύγχρονες και πλέον χρησιμοποιούμενες μηχανές αναζήτησης έχουν την ικανότητα να ανακτούν το περιεχόμενο που παρουσιάζουν μονάχα από το δημόσιο ιστό, δηλαδή, παρουσιάζουν σελίδες οι οποίες περιέχουν links που παραπέμπουν σε άλλες σελίδες. Το βασικό στοιχείο που κάνει τα εργαλεία αυτά αναζήτησης μερικώς ανεπαρκή είναι το γεγονός ότι αδυνατούν να συμπεριλάβουν δυναμικό περιεχόμενο στις βάσεις τους καθώς δεν έχουν την ικανότητα να σκεφτούν.

Αντίθετα, ο αόρατος ιστός, λειτουργεί με έναν τρόπο πολύ διαφορετικό, τοποθετώντας το μεγαλύτερο μέρος των εγγράφων που κινούνται μέσα σε αυτόν σε αναζητήσιμες βάσεις δεδομένων. Κάθε βάση στον αόρατο ιστό περιέχει συγκεκριμένο και εξειδικευμένο περιεχόμενο το οποίο παρέχεται στο χρήστη αφού αυτός σχηματίσει κάποιο ερώτημα. Στις παραδοσιακές μηχανές αναζήτησης το παραπάνω δε μπορεί να συμβεί καθώς αυτές δεν έχουν τη δυνατότητα της μεταφοράς των λέξεων προς αναζήτηση στις βάσεις και τη λήψη αποτελεσμάτων από κει (Surface Web, Deep Web, Dark Web -- What's the Difference?, 2016) .

Η τεχνολογία με βάση την οποία λειτουργεί ο δημόσιος ιστός βασίζεται στην τεχνική των crawlers, δηλαδή στην αναζήτηση των εγγράφων μέσω τον link. Στη μηχανή αναζήτησης δίδεται μία σειρά λέξεων κλειδιών που παραπέμπουν σε έγγραφα και τα αποτελέσματα εμφανίζονται όχι από τον ίδιο τον ιστό αλλά από τη βάση δεδομένων της αναζήτησης. Άρα λοιπόν, όπως είναι αντιληπτό, η διαδικασία αναζήτησης δε

γίνεται σε πραγματικό χρόνο αλλά αντίθετα καλούνται έγγραφα τα οποία έχουν ήδη αποθηκευθεί σε κάποια βάση.

Όταν στις ιστοσελίδες δεν υπάρχουν links που να παραπέμπουν σε άλλες σελίδες τότε η λειτουργία των crawlers είναι ανεπαρκής καθώς αυτοί δε θα είναι σε θέση να εντοπίσουν κανένα link και το περιεχόμενο των σελίδων δε θα εμφανιστεί ποτέ. Ο μόνος τρόπος για να γίνει ορατή μία τέτοια ιστοσελίδα είναι εάν ο ιδιοκτήτης αυτής δηλώσει τη διεύθυνση της στην υπηρεσία αναζήτησης (Surface Web, Deep Web, Dark Web -- What's the Difference?, 2016)

Ακόμη, υπάρχει ένας ιδιαίτερα σημαντικός όγκος πληροφορίας στο διαδίκτυο ο οποίος είναι προσβάσιμος μονάχα από ορισμένο αριθμό χρηστών. Η πληροφορία αυτή ναι μεν είναι προσβάσιμη από τους crowlers από την άλλη όμως δεν υπάρχει κατηγοριοποιημένη σε μηχανές αναζήτησης. Η πληροφορία αυτής της μορφής εντοπίζεται σε (Σιδηρόπουλος):

- Σελίδες προστατευμένες με password: Ο crawler δεν μπορεί να παραβλέψει ή να προσπεράσει με κάποιο τρόπο τη φόρμα εισαγωγής username – password.

- Χρήση ειδικών προγραμμάτων: Υπάρχουν web sites τα οποία χρησιμοποιούν ειδικά προγράμματα τα οποία δεν επιτρέπουν στους crawlers να έχουν πρόσβαση σε σελίδες τους. Προσπαθούν με αυτόν τον τρόπο να ωθήσουν το χρήστη να χρησιμοποιήσει τις δικές τους φόρμες αναζήτησης και να προστατέψουν έτσι με κάποιο τρόπο τα περιεχόμενά τους.

- Ιδιοκτησιακές σελίδες: Κάποια web sites περιέχουν πληροφορίες στις οποίες για να αποκτήσει πρόσβαση ένας χρήστης θα πρέπει να συμφωνήσει σε ειδικούς όρους, να κάνει κάποιου είδους έγγραφή, να πληρώσει κάποιο χρηματικό ποσό. Οι μηχανές αναζήτησης φυσικά δεν μπορούν να ικανοποιήσουν τις απαιτήσεις αυτές.

Επιπρόσθετα, υπάρχει ένας αρκετά μεγάλος αριθμός σελίδων ο οποίος χρησιμοποιεί την HTML ως βάση αλλά ταυτόχρονα περιλαμβάνει πολλά γραφικά, τεχνικές Flash, video streaming και πολλά άλλα. Εάν ένας crawler προσπαθήσει να προσπελάσει μία

τέτοια σελίδα δε θα είναι σε θέση να αντιληφθεί το περιεχόμενό της και αυτόματα αυτή θα καταταγεί στις ιστοσελίδες που ανήκουν στον αόρατο ιστό ακριβώς γιατί το περιεχόμενό της δε θα μπορεί να καταχωρηθεί στις βάσεις των μηχανών αναζήτησης.

## 2.4 Μηχανές αναζήτησης και τύποι περιεχομένου που περιλαμβάνονται στο Deep Web

Η αρχή της ιστορίας του παγκόσμιου ιστού σήμαινε φυσικά την ύπαρξη πολύ λίγων ιστοσελίδων και ακόμη λιγότερων εγγράφων να περιέχονται σε αυτές με την πληροφορία να έχει συνήθως τη μορφή εικόνας και τη σελίδα να είναι στατική. Έτσι, οι ιστοσελίδες ήταν ιδιαίτερα εύκολο να εντοπιστούν τα διάφορα εργαλεία που χρησιμοποιούνταν την εποχή εκείνη.

Πρώτη γνωστή μηχανή αναζήτησης είναι η μηχανή Lycos, η οποία ξεκίνησε το 1994 και λειτούργησε με μόλις 54.000 έγγραφα στη βάση της (Zilman, 2017). Η εξέλιξη της τεχνολογίας σήμαινε φυσικά την ίδια στιγμή ραγδαία εξάπλωση του διαδικτύου και εισαγωγή νέων τύπων αρχείων και συνεπακόλουθα νέας μορφής πληροφορίας. Στις αρχές του 1996 συντελέστηκαν τα τρία παρακάτω φαινόμενα τα οποία άλλαξαν άρδην την πορεία του διαδικτύου (Zilman, 2017):

1. Η τεχνολογία των βάσεων -δεδομένων έκανε την εμφάνισή της στο Internet, από μεγάλες εταιρίες του χώρου.

2. Το web άρχισε να εμπορευματοποιείται, αρχικά μέσα από τα εργαλεία αναζήτησης και -καταλήγοντας στη δημιουργία του ηλεκτρονικού εμπορίου (e-commerce).

3. Υιοθετήθηκε η χρήση των web servers ώστε να επιτρέπεται η δημοσίευση «δυναμικού περιεχομένου» (χρήση ASP, PHP κ.α., εμφάνιση δυναμικών web σελίδων).

Τα παραπάνω συντέλεσαν στη δημιουργία ενός προσανατολισμού προς τις βάσεις δεδομένων, τουλάχιστον για τις ιστοσελίδες οι οποίες περιείχαν μεγαλύτερο όγκο πληροφορίας. Σήμερα, γίνεται τεράστια διακίνηση πληροφορίας διαμέσου του διαδικτύου καθώς τόσο ιδιώτες χρήστες όσο και πολλές εταιρίες το επιλέγουν για να διακινήσουν την πληροφορία που τους ενδιαφέρει διαμέσου αυτού. Αυτό το οποίο δεν έχει γίνει έως σήμερα σαφές είναι ότι τα μέσα που χρησιμοποιούνται στο διαδίκτυο ώστε να μεταφερθεί η πληροφορία είναι οι βάσεις δεδομένων (Zilman, 2017).

Οι μηχανές αναζήτησης σήμερα ανακτούν το περιεχόμενό τους από τις ιστοσελίδες που είναι προσβάσιμες στο σύνολό τους ακλουθώντας υπερκείμενες συνδέσεις (hypertext links) που περιέχονται σε προηγούμενες ιστοσελίδες, αγνοώντας τις φόρμες αναζήτησης και τις σελίδες οι οποίες απαιτούν έγκριση ή εγγραφή του χρήστη για να δώσουν πρόσβαση στο περιεχόμενο τους.

Το μεγαλύτερο μέρος των εγγράφων που βρίσκονται στο deep web είναι τοποθετημένο σε βάσεις δεδομένων που περιέχουν εξειδικευμένο περιεχόμενο στο οποίο ο χρήστης έχει πρόσβαση διαμέσου της πληκτρολόγησης λέξεων κλειδιών στη φόρμα αναζήτησης (Zilman, 2017). Η μηχανή επιστρέφει στο χρήστη μία λίστα αποτελεσμάτων που αποτελούν απάντηση στο ερώτημα που τέθηκε.

Έτσι λοιπόν, ο χρήστης πληκτρολογεί μία σειρά λέξεων κλειδών οι οποίες με την τεχνολογία των crawlers δείχνουν σε μία σειρά αρχείων. Τα αποτελέσματα που παρουσιάζονται στο χρήστη δεν προέρχονται τελικά από τον ίδιο τον ιστό αλλά από τη βάση δεδομένων στην οποία βρίσκεται αποθηκευμένη η πληροφορία. Για να υπάρξει ένα διαφορετικό αποτέλεσμα στην αναζήτηση θα έπρεπε αντίστοιχα να υπάρχει και ένας διαχειριστής ο οποίος να πληκτρολογεί χειροκίνητα ερωτήματα και

αντίστοιχα να προσθέτει αποτελέσματα στο ευρετήριο της μηχανής (Zilman, 2017). Όπως είναι αντιληπτό κάτι τέτοιο είναι πρακτικά ανέφικτο καθώς κανένας δε θα μπορούσε ούτε να προβλέψει τα ερωτήματα αλλά ούτε και να επαναλαμβάνει συνέχεια τη διαδικασία ώστε να καλύπτει τη ζήτηση.

Υπάρχουν αρκετές ιστοσελίδες οι οποίες δεν περιέχουν συνδέσμους προς άλλες σελίδες και έτσι καθιστούν ανεπαρκή την λειτουργία των crawlers. Καθώς η αναζήτηση των μηχανών γίνεται διαμέσου των υπερσυνδέσμων, όταν αυτές δεν υπάρχουν στις σελίδες γίνεται σχεδόν αδύνατη η αναζήτησή τους από τους crawlers. Ο μόνος τρόπος για να γίνει ορατή μία τέτοια ιστοσελίδα είναι να καταχωρηθεί από τον ίδιο τον ιδιοκτήτη της στην υπηρεσία αναζήτησης.

Στο διαδίκτυο, υφίσταται και ένας τεράστιος όγκος πληροφορίας που είναι ορατός υπό συγκεκριμένες προϋποθέσεις σε έναν ακόμη πιο συγκεκριμένο αριθμό χρηστών. Η πληροφορία αυτή είναι καταχωρημένη σε σελίδες που τεχνικά είναι ορατές και προσπελάσιμες από τους crawlers, αλλά εσκεμμένα έχει αποκλειστεί η δυνατότητα καταχώρησής τους στα ευρετήρια των μηχανών αναζήτησης. Οι ιστοσελίδες αυτές λοιπόν, όπως αναφέρθηκε και σε προηγούμενη ενότητα μπορεί να είναι προστατευμένες με κάποιους κωδικούς πρόσβασης και έτσι να είναι απρόσιτες στους crawlers, να χρειάζονται εξειδικευμένα προγράμματα ώστε να είναι προσπελάσιμες, να κατευθύνουν τους crawlers προς σελίδες spam, ή να είναι ιδιοκτησιακές ιστοσελίδες και έτσι να απαιτούν ειδικούς όρους ώστε να είναι προσβάσιμες (Zilman, 2017).

Πέραν των παραπάνω, υπάρχει και ένας τεράστιος αριθμός ιστοσελίδων που ναι μεν είναι γραμμένες σε HTML, αλλά το περιεχόμενο τους έχει τη μορφή γραφικών, flash, περιλαμβάνουν video, streaming και πολλά άλλα γεγονός που κάνει τον crawler να τις αντιλαμβάνεται ως σελίδες χωρίς περιεχόμενο. Τέτοιοι τύποι σελίδων ανήκουν στο Deep Web γιατί απλά δεν υπάρχει τίποτα σ' αυτές που να μπορεί να καταχωρηθεί σε ευρετήριο μιας μηχανής αναζήτησης. Στην κατηγορία αυτή μπορούμε να συμπεριλαμβάνονται επίσης και αρχεία .pdf και postscript. Αν και οι μηχανές αναζήτησης έχουν τρόπο να τα χειριστούν και να τα -εμφανίσουν σαν αποτέλεσμα σε μια αναζήτηση, οι περισσότερες δεν το κάνουν, γιατί θεωρούν πως -ποτελούν πολύ εξειδικευμένο περιεχόμενο που σπάνια θα ενδιαφέρει κάποιο χρήστη (η Google δεν ανήκει σε αυτές) (Zilman, 2017).

## 2.5 Αλγόριθμοι αναζήτησης στο Deep Web

Ο όγκος της πληροφορίας που διακινείται πλέον διαμέσου του παγκόσμιου έχει ξεπεράσει κάθε προσδοκία και όπως είναι αντιληπτό αυτός αυξάνεται καθημερινά με ιλιγγιώδεις ρυθμούς. Ο ρυθμός αύξησης του όγκου αλλά και ο ίδιος ο όγκος φυσικά είναι δύσκολο να καταχωρηθεί σε βάση αναζήτησης άμεσα καθιστώντας την ενημέρωση αυτών μία ιδιαίτερα δύσκολη διαδικασία.

Για να αντιμετωπιστεί το παραπάνω χρησιμοποιείται κατά βάση η τεχνική web harvesting. Πρόκειται για τη δημιουργία προγραμμάτων τέτοιων που διατρέχουν τον ιστό με αυτοματοποιημένο τρόπο ο οποίος κατευθύνεται από τη μηχανική μάθηση και μιμείται συμπεριφορά ανθρώπου. Τα προγράμματα αυτά καλούνται directed crawlers και έχουν ως στόχο την εξεύρεση σελίδων που αφορούν σε συγκεκριμένη θεματολογία. Δυστυχώς ο ιστός παρουσιάζει κάποια χαρακτηριστικά που καθιστούν δύσκολη τη λειτουργία των directed crawlers και αυτά είναι (LLC, 2000):

- *Ο μεγάλος όγκος του*
- *Ο ταχύς ρυθμός μεταβολής του*
- *Η δυναμική δημιουργία σελίδων από χρήστες του Internet*

Ένα βασικό χαρακτηριστικό των directed crawlers είναι το crawl frontier το οποίο δείχνει το σύνολο των URL που έχει ερευνήσει ο crawler σε μία συγκεκριμένη χρονική στιγμή t. Ένα από τα πλέον βασικά προβλήματα που καλείται να αντιμετωπίσει ο κάθε crawler είναι το βάθος στο οποίο θα πραγματοποιήσει την αναζήτηση του ώστε να μη χρειαστεί να διατρέξει σε όλους τους υπερσυνδέσμους του διαδικτύου. Το παραπάνω αποσκοπεί στην ελαχιστοποίηση της καταναλωτικής ισχύος και στη δραματική μείωση του χρόνου αναζήτησης. Ακόμη, θα πρέπει να αντιμετωπιστούν τα προβλήματα των seeds αλλά και της ανάθεσης προτεραιοτήτων στα URL ώστε το crawling να είναι όλο και πιο αποδοτικό (LLC, 2000).

Χαρακτηριστικό παράδειγμα ιστοσελίδων που περιέχουν πληροφορία που προέρχεται από τον ίδιο το χρήστη είναι τα forums. Οι σελίδες αυτές περιέχουν συλλογή θεμάτων και για το λόγο αυτό το περιεχόμενο της πληροφορίας είναι συχνά πλούσιο. Τα τελευταία χρόνια οι μηχανές αναζήτησης στρέφουν ολοένα και περισσότερο την

αναζήτησή τους σε Web forums. Ορισμένα από τα σημαντικότερα προβλήματα που αντιμετωπίζουν οι forum crawlers συνοψίζονται στα παρακάτω (LLC, 2000):

- Πολύπλοκη δομή των forums με μεγάλο βάθος και υπερβολικά μεγάλο αριθμό συνδέσμων.
- Μεγάλος αριθμός ίδιων ή πανομοιότυπων σελίδων (duplicate pages)
- Μεγάλος αριθμός άκυρων σελίδων που δημιουργούνται από αποτυχημένες απόπειρες σύνδεσης χρηστών (login)

Ένας αλγόριθμος που είθισται να χρησιμοποιείται από τους crawlers ώστε να υπάρχει άμεση πρόσβαση στο περιεχόμενο των forums είναι ο Breadth – First. Στον αλγόριθμο αυτό, η αναζήτηση γίνεται σε γράφο, ξεκινά από τον αρχικό κόμβο και διατρέχει όλους τους γειτονικούς. Κάθε ένας από τους γειτονικούς κόμβους, εξερευνά τους γειτονικούς τους ανεξευρεύνητους ώστε να βρει τον επόμενο στόχο. Η διαδικασία συνεχίζεται έως ότου να ολοκληρωθεί ο γράφος.
Η τεχνική αυτή εξερευνά διεξοδικά όλο το γράφο έως ότου να εντοπίσει το στόχο. Τα "παιδιά" του κάθε κόμβου που ερευνάται προστίθενται σε μια ουρά FIFO (First-In-First-Out) χωρίς να υπολογίζεται διαφορετική προτεραιότητα για το καθένα (LLC, 2000).
Μία ακόμη τεχνική που χρησιμοποιείται είναι η Best – First η οποία κατά τη διάσχιση του γράφου εντοπίζει αυτόν που φαίνεται να είναι ο πλέον υποσχόμενος και χαράσσει μία διαδρομή προς αυτόν. Η αναζήτηση αυτή εκτιμά την αξία του κάθε κόμβου n και παράλληλα δημιουργεί μια συνάρτηση εκτίμησης f(n) η οποία κάνει χρήση του κόμβου – στόχου με μία ευρετική μέθοδο. Η επιλογή του κατάλληλου υποψηφίου κόμβου για την επέκταση της αναζήτησης γίνεται από μια ουρά προτεραιοτήτων που χρησιμοποιείται για να αποθηκεύει προσωρινά τους "υποψηφίους" κόμβους. Είναι εμφανές ότι η τεχνική best-first χρησιμοποιεί έναν "αυτοδιδασκόμενο" αλγόριθμο προκειμένου να περιορίσει την αναζήτηση σε μέρος του γράφου και όχι να την επεκτείνει σε όλους τους κόμβους (LLC, 2000).
Η ομαδοποίηση με επαναλαμβανόμενες περιοχές ομαδοποιεί τα forum με τρόπο τέτοιο ώστε οι σελίδες να είναι πανομοιότυπες. Για κάθε σελίδα s δημιουργείται ένα μοτίβο p* για κάθε επαναλαμβανόμενη περιοχή της και στη συνέχεια ελέγχεται αν το p* ταυτίζεται με κάποιο από τα ήδη υπάρχοντα μοτίβα pi . Αν δεν υπάρξει

αντιστοίχιση τότε το p* προστίθεται στη λίστα μοτίβων του module. Οι βασικοί αλγόριθμοι που χρησιμοποιούνται για την παραπάνω διαδικασία είναι οι RepetitiveRegionDetection, AlignRecordsInRegion και TreeAlignmentCost (LLC, 2000).

Οι σελίδες οι οποίες θεωρούνται πως έχουν όμοιο περιεχόμενο συχνά περιέχουν URL με διαφορετική μορφή, ένα γεγονός το οποίο οφείλεται στο ότι πολλές σελίδες περιέχουν ίδιο εννοιολογικό περιεχόμενο αλλά διαφορετικό URL. Η ομαδοποίηση με βάση το URL βασίζεται στον υπολογισμού ενός μέτρου ομοιότητας μεταξύ δύο οποιονδήποτε URLs.

Κάθε URL μπορεί να θεωρηθεί ως μία ακολουθία χαρακτήρων που διαχωρίζονται μεταξύ τους με το χαρακτήρα "/ ". Στην περίπτωση όπου γίνεται ένα ερώτημα αναζήτησης σε ένα forum, το σύνολο των χαρακτήρων χωρίζεται σε δύο τμήματα με χρήση του χαρακτήρα "?". Το τμήμα που βρίσκεται μετά από το "?", εξάγει όλα τα ζεύγη (κλειδί, τιμή) σύμφωνα με τα σύμβολα ισότητας "=" και "&". Τα σύνολα χαρακτήρων πριν το "?" ονομάζονται διαδρομές και τα σύνολα χαρακτήρων μετά το "?" ονομάζονται παράμετροι (LLC, 2000).

Δύο URLs θεωρούνται όμοια και τοποθετούνται στην ίδια ομάδα αν έχουν :

- Τον ίδιο αριθμό και σειρά διαδρομών.
- Τις ίδιες παραμέτρους κλειδιά

Όπως είναι προφανές, η πληροφορία η οποία εμπεριέχεται σε ένα forum, δεν είναι πάντα χρήσιμη. Έτσι σε έρευνα που πραγματοποίησε ο Τζάνης (Τζάνης, 2013), κατέληξε στα εξής συμπεράσματα :

1. Μια σελίδα με χρήσιμη πληροφορία ανήκει σε μια μεγάλη ομάδα. Δηλαδή σελίδες που ανήκουν σε μια ομάδα με μεγάλο αριθμό σελίδων έχουν μεγαλύτερη πιθανότητα να είναι "πολύτιμες".

2. Οι χρήσιμες σελίδες συνήθως έχουν μεγάλο μέγεθος αρχείου. Κάτι τέτοιο είναι λογικό αφού οι χρήσιμες σελίδες είναι αυτές που δημιουργούνται δυναμικά από τους χρήστες και επομένως το αρχείο στο οποίο αποθηκεύονται έχει μεγαλύτερο μέγεθος από τα αρχεία των τυπικών σελίδων του forum.

3. Οι χρήσιμες σελίδες έχουν εννοιολογική ποικιλία αφού δημιουργούνται από χρήστες.

# Κεφάλαιο 3 – Το Dark Web

## 3.1 Dark Web

Το σκοτεινό διαδίκτυο ή dark web αποτελεί μία εκ των πλέον σκοτεινών πλευρών του διαδικτύου το οποίο έχει μία ιδιαίτερα σκληρή μορφή. Ως έτος ίδρυσής του θεωρείται η αρχή της δεκαετίας του 1990, και ως ιδρυτές του θεωρούνται οι αμερικανικές μυστικές υπηρεσίες που είχαν ως στόχο να αποκρύψουν τις διάφορες συνομιλίες στου στόλου των ΗΠΑ που ταξίδευαν σε όλη την υφήλιο (Finklea, 2015).

Χαρακτηριστικό του dark web είναι ότι αυτό έχει την ικανότητα να διαγράφει παντελώς την ταυτότητα του χρήστη και αυτός να παραμένει ανώνυμος για όσο διάστημα πλοηγείται σε αυτό. Δυστυχώς, το σκοτεινό αυτό διαδίκτυο έγινε γρήγορα αντιληπτό από σειρά παρανόμων και εγκληματιών σε ολόκληρο τον κόσμο με αποτέλεσμα να περάσει στα χέρια αυτών, στα χέρια διακινητών παράνομων εμπορευμάτων, ή ακόμη και σωματεμπόρων.

Σήμερα, το dark web αποτελεί ένα παράλληλο κόσμο στον οποίο βασιλεύει η παρανομία. Κάθε χρήστης σε αυτό διατηρείται ανώνυμος καθώς αποκρύπτεται παντελώς η ταυτότητα του υπολογιστή από τον οποίο εισήλθε. Άτομα που προέρχονται από το χώρο της ιταλικής μαφίας, ή της μαφίας των ΗΠΑ, ακόμη και πληρωμένοι εκτελεστές λειτουργούν διαμέσου του dark web και κλείνουν μεταξύ τους συμφωνίες διατηρώντας πλήρως την ανωνυμία τους (Finklea, 2015). Οι βαρόνοι των ναρκωτικών έχουν πια ειδικούς λογαριασμούς στο Darknet και κλείνουν συμφωνίες εκατομμυρίων δολαρίων για την πώληση τους, ενώ οι διακινητές παιδικού ακατάλληλου υλικού βρήκαν το απόλυτο εργαλείο για να πουλάνε το νοσηρό προϊόν τους (Finklea, 2015).

Το dark net αποτελεί τον ιστό εκείνο που λειτουργεί παράλληλα με το διαδίκτυο και προσφέρει μοναδική μυστικότητα καθώς όπως προειπώθηκε η ανίχνευση της ταυτότητας των χρηστών είναι αδύνατη. Ονομάζεται και deepnet ή hidden web ενώ υπολογίζεται πως ο όγκος των δεδομένων που κυκλοφορούν μέσα σε αυτό είναι πολύ μεγαλύτερος από αυτόν του επιφανειακού διαδικτύου (Finklea, 2015). Οι ιστότοποι που συμμετέχουν στο dark net δεν έχουν τη μορφή διευθύνσεων όπως το γνωστά σε

όλους URL αλλά αποτελούνται από φαινομενικές σειρές χαρακτήρων που χρησιμοποιούν εξειδικευμένες εφαρμογές πρόσβασης όπως είναι το TOR. Όπως είναι αντιληπτό, η επίσκεψη σε μία τέτοια ιστοσελίδα γίνεται διαμέσου μίας τυχαιοποιημένης πορείας από διάφορους υπολογιστές που κατορθώνουν να εξαφανίσουν τα ψηφιακά ίχνη του επισκέπτη. Βέβαια, αυτό μειώνει την ταχύτητα του δικτύου (Finklea, 2015).

Οι μέθοδοι εντοπισμού που χρησιμοποιούνται ώστε να εντοπιστούν οι χρήστες του dark web δεν είναι πάντα αποτελεσματικές και έτσι αυτό κατορθώνει να λειτουργεί ανεξέλεγκτο.

Η μηχανή αναζήτησης που χρησιμοποιείται για την είσοδο στο dark web είναι η Torgle η οποία χρειάζεται απλά μια πληκτρολόγηση του προς αναζήτηση αντικειμένου ώστε να προβάλλει άμεσα αποτελέσματα. Σύμφωνα με εκτιμήσεις που έγιναν σε μία μελέτη στο Πανεπιστήμιο Berkeley της Καλιφόρνια (University of California, Berkeley) το 2001, το dark web αποτελείται περίπου από 91.000 terabytes. Αντίθετα το επιφανειακό Web (που είναι εύκολα προσπελάσιμο από τις μηχανές αναζήτησης) είναι περίπου 167 terabytes (Berkeley, 2001). Η Βιβλιοθήκη του Αμερικάνικου Κονγκρέτου, υπολογίστηκε πως το 1997 είχε 3.000 terabytes. Το 2011, το YouTube υπολογίζεται ότι είχε αποθηκευμένα περίπου 200 εκατομμύρια βίντεο, συνολικού μεγέθους 5 petabytes ή 5000 terabytes. Ο υπολογισμός του μεγέθους του web διαφέρει από πηγή σε πηγή και έτσι υπάρχει ένα μεγάλο περιθώριο λάθους και κανένας αριθμός δε μπορεί να θεωρηθεί ως ακριβής (Sui, Caverlee, & Rudesill).

Οι πληροφορίες του Dark Web ανήκουν σε μία ή περισσότερες από τις παρακάτω κατηγορίες (Sui, Caverlee, & Rudesill).:

- Δυναμικά παραγόμενο περιεχόμενο: δυναμικές ιστοσελίδες οι οποίες δημιουργούνται ως αποτέλεσμα της εκτέλεσης κάποιας επερώτησης (query) ή προσπελαύνονται μόνο μέσω κάποιας φόρμας.
- Μη συνδεδεμένο περιεχόμενο: ιστοσελίδες οι οποίες δεν περιέχουν συνδέσμους από άλλες ιστοσελίδες, εμποδίζοντας έτσι τα προγράμματα που κάνουν Web crawling να επισκεφθούν το περιεχόμενό τους.
- Ιδιωτικό Web: ιστότοποι που απαιτούν εγγραφή (registration) και κωδικό πρόσβασης.

- Περιεχόμενο περιορισμένης πρόσβασης: ιστότοποι που περιορίζουν την πρόσβαση στις σελίδες τους με τεχνικό τρόπο (π.χ. χρησιμοποιώντας το Robots Exclusion Standard, CAPTCHAs, ή το no-cache Pragma στις επικεφαλίδες του πρωτοκόλλου HTTP, τα οποία απαγορεύουν στις μηχανές αναζήτησης να πλοηγούνται στις ιστοσελίδες τους).
- Περιεχόμενο που δεν είναι σε μορφή HTML: κείμενα που συμπεριλαμβάνονται σε multimedia αρχεία (εικόνες ή video) ή που έχουν συγκεκριμένη μορφή την οποία δεν μπορούν να χειριστούν οι μηχανές αναζήτησης.
- Κείμενα που χρησιμοποιούν το παλαιότερο πρωτόκολλο Gopher και αρχεία που βρίσκονται σε διακομιστές FTP και τα οποία δεν μπορούν να εντοπιστούν από τις περισσότερες μηχανές αναζήτησης.

Ο Brad Chacos σημείωσε πως : «Το Σκοτεινό Διαδίκτυο έχει και μία εξαιρετικά σημαντική "φωτεινή" πλευρά, εξασφαλίζοντας την ελευθερία της έκφρασης σε ανθρώπους που ζουν σε απολυταρχικά καθεστώτα και βοηθώντας να έρθουν στο φως συνταρακτικά ντοκουμέντα, χωρίς τον φόβο όσων τα διέρρευσαν πως θα διωχθούν ποινικά» (Chacos, 2013). Σε αυτό έχουν φιλοξενηθεί κατά καιρούς αντίγραφα αρχείων από το Wikileaks ενώ μάλιστα το περιοδικό New Yorker, έχει δημιουργήσει στο dark web μία υπηρεσία η οποία εγγυάται ανωνυμία σε όσους θελήσεις να επικοινωνήσουν με τους συντάκτες του και να προβούν σε ανώνυμες καταγγελίες ή να δώσουν απόρρητες πληροφορίες.

Ακόμη, η διοργάνωση «Δημοσιογράφοι χωρίς Σύνορα» συμβουλεύει τα μέλη της ώστε να χρησιμοποιούν το dark web για να έρχονται σε επαφή με τις διάφορες πηγές τους που επιθυμούν να παραμείνουν ανώνυμες (Finklea, 2015). Σε υπηρεσίες που βασίζονται στο TOR καταφεύγουν και απλοί χρήστες οι οποίοι θέλουν να αποφύγουν τη λογοκρισία ή είναι ακτιβιστές.

Για να εισέλθει κανείς στο dark web χρειάζεται να χρησιμοποιήσει το λογισμικό Tor Browser Bundle και διαμέσου αυτού να εισέλθει στο Onionland. Ασφαλώς για τον απλό χρήστη ο οποίος διακατέχεται από άγνοια το παραπάνω ελλοχεύει σειρά κινδύνων. Το dark web αποτελείται από μία σειρά ιστοσελίδων οι οποίες δεν είναι ορατές στις κοινές μηχανές αναζήτησης ενώ είναι προσπελάσιμες μονάχα διαμέσου του TOR ο οποίος αποκρύπτει και τις IP τους (Sui, Caverlee, & Rudesill).

Από τη στιγμή που θα εισέλθει κάποιος εντός του dark web, δε θα είναι σε θέση να προσδιορίσει ποιος άλλος ή τι άλλο βρίσκεται στο ίδιο δίκτυο με αυτόν. Αυτό ακριβώς είναι και το χαρακτηριστικό που κάνει το dark web ιδιαίτερα δημοφιλές. Εικάζεται πως πέραν από τη σειρά παράνομων πράξεων που συντελούνται εντός του dark web, μέσα σε αυτό μπορεί κανείς να παρακολουθήσει ζωντανά βιασμούς ή ακόμη και δολοφονίες ανθρώπων.

Η δημοσιογράφος της Washington Post αναφέρει ότι σε ιστοσελίδες όπως το Reddit, το 4Chan και το Hidden Wiki συμβαίνει ένα ιδιότυπο αλισβερίσι λογαριασμών από δεύτερο, τρίτο ή τέταρτο χέρι από φερόμενα ως ενεργά (ή και ανενεργά) δωμάτια μέσα στα οποία εκτελούνται άνθρωποι και διαπράττονται βιασμοί. Πειστικότερο όλων είναι το άρθρο ενός Βρετανού ερασιτέχνη κρυπτογράφου, ονόματι Cthulhu, ο οποίος αναλύοντας τους σερβερς φερόμενων ως πληρωμένων χάκερ και φονιάδων διαπίστωσε ότι είναι διάτρητοι από άποψη ασφάλειας — γεγονός που συντείνει στην άποψη ότι πρόκειται περί απατεώνων που εισπράττουν ποσά για τις υποτιθέμενες υπηρεσίες τους και στη συνέχεια γίνονται καπνός.

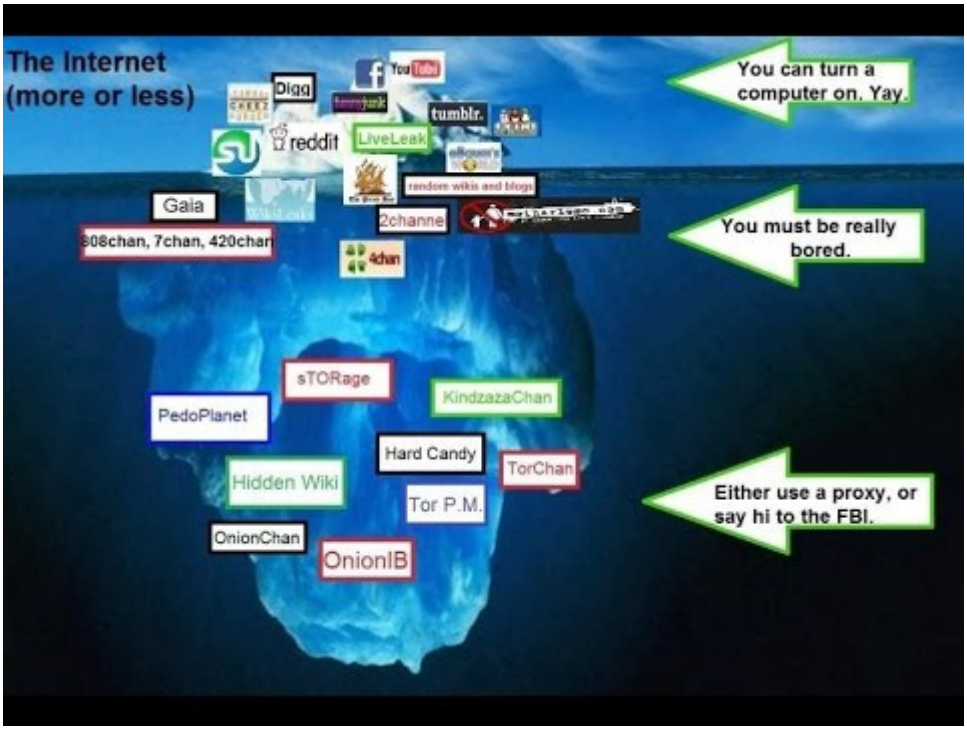
## 3.2 Αρχιτεκτονική και Πρωτόκολλα λειτουργίας

Ένας πρώτος ορισμός γύρω από το dark net χαρακτηρίζει αυτό ως το σύνολο των μη προσβάσιμων ιστοσελίδων στο χώρο του διαδικτύου. Με τον όρο πρόσβαση, εννοείται φυσικά η οποιαδήποτε επικοινωνία μεταξύ ενός υπολογιστή και ενός δικτυακού κόμβου που αντιστοιχεί σε μία ηλεκτρονική διεύθυνση. Ο παραπάνω ορισμός ελλοχεύει μία ασάφεια καθώς το dark net δεν είναι κοινό για όλους του υπολογιστές που ανήκουν σε ένα δίκτυο. Έτσι λοιπόν, για κάθε υπολογιστή, το dark net αποτελεί το σύνολο των ιστοσελίδων – κόμβων στις οποίες δεν μπορεί να εισέλθει.

Έτσι λοιπόν, εάν υπάρχει ένα σύνολο διευθύνσεων A, και το συνολικό μέγεθος των διευθύνσεων στο διαδίκτυο είναι B, τότε το dark net αποτελεί για τον υπολογιστή στο σύνολο A – B. Οι λόγοι για τους οποίους κάποιος υπολογιστής ενδέχεται να μην έχει πρόσβαση σε κάποια σελίδα ενδέχεται να ποικίλουν (Finklea, 2015). Το dark net

αποτελεί εδώ και χρόνια αντικείμενο μελέτης καθώς κρύβει σειρά απειλών που προβληματίζουν τις αστυνομικές αρχές σήμερα.

Για το διαδίκτυο, το dark net αποτελεί ένα σύνολο κόμβων που δεν είναι προσβάσιμοι από το μέσο χρήστη. Ένα χαρακτηριστικό παράδειγμα των ιστοσελίδων που ανήκουν στο dark web είναι αυτές του στρατού των ΗΠΑ. Κύριος λόγος για τον οποίο οι σελίδες που ανήκουν στο dark net δεν είναι προσπελάσιμες είναι το γεγονός ότι αυτές είναι δυσεύρετες.




**Εικόνα 8 : Το διαδίκτυο, το Deep και το Dark web**

Καθώς το διαδίκτυο είναι χωρισμένο σε επίπεδα, στο υψηλότερο επιπεδο αυτού απαντώνται οι υπολογιστές των χρηστών ενώ στο χαμηλότερο οι κεντρικοί δρομολογητές των παρόχων που ουσιαστικά αποτελούν τη βάση του δικτύου. Όπως αντιληπτό, σε όσο χαμηλότερο επίπεδο βρίσκεται ένας δρομολογητής, τόσο μεγαλύτερη κίνηση δικτύου διέρχεται από αυτόν, και άρα αυτός λειτουργεί ως μία πύλη που εξυπηρετεί σειρά διευθύνσεων.

Εάν λοιπόν δεν έχει γίνει η ανάλογη ρύθμιση στο δρομολογητή, η κίνηση που διέρχεται από αυτόν είναι προβληματική και τον καθιστά άμεσα μέρος του dark net. Το παραπάνω συναντάται σε πολλούς δρομολογητές του δικτύου καθιστώντας το

dark net ένα ιδιαίτερα πολύπλοκο φαινόμενο. Έρευνα που έγινε από την εταιρία Arbor Networks, έδειξε ότι στο dark net ανήκουν το 5% των κόμβων του ιστού, δηλαδή περιπου 100.000.000 διακομιστές (Arbornetworks).

Ένα τέτοιο φαινόμενο, δε θα μπορούσε φυσικά παρά να αποτελεί αντικείμενο εκμετάλλευσης πολλών επιτηδείων που χρησιμοποιούν το dark net ώστε να διαπράξουν παράνομες ενέργειες. Πολλοί από τους hackers, εκμεταλλεύονται τους δρομολογητές του dark net ώστε να πραγματοποιήσουν επιθέσεις από τους υπολογιστές τους προς το παγκόσμιο δίκτυο καθώς αυτοί δεν είναι δυνατό να ανιχνευθούν λόγω του ότι η διεύθυνσή τους δεν εμφανίζεται σε κανέναν από τους πίνακες δρομολόγησης. Μάλιστα, πολλές φορές γίνεται αναδρομολόγηση των πακέτων που αποστέλλονται διαμέσου του dark net ώστε να μην υπάρξει σε καμία περίπτωση εντοπισμός της πηγής τους.

Το dark net, κάνει χρήση της δρομολόγησης onion που χρησιμοποιείται με βάση την ανώνυμη επικοινωνία μεταξύ των υπολογιστώνν που βρίσκονται σε ένα δίκτυο. Αποτελεί μία υποδομή ιδιωτικών συνδέσεων οι οποίες διατηρούνται ανώνυμες χρησιμοποιώντας διαφορετικα επίπεδα κρυπτογράφησης τα οποία είναι ανεκτικά σε επιθέσεις διαφόρων τύπων. Οι συνδέσεις είναι δικατευθυντήριες, σχεδόν πραγματικού χρόνου και μπορούν να χρησιμοποιηθούν είτε για κινήσεις προσανατολισμένες σε σύνδεση, είτε για κινήσεις άνευ εγκατάστασης σύνδεσης (Reed, Syverson, & Goldschlag).

Η ιδέα της δημιουργίας του Onion Routing προήλθε από την ανάμιξη των συνδέσεων χρήστων και εφαρμογών με στόχο να αποκρυφθεί η ταυτότητά τους. Βασική ιδιότητα του Onion Routing είναι ότι δεν επιτρέπει σε εκείνους που έχουν πρόσβαση στο μέσο να αναγνωρίσουν όσους συμμετέχουν σε μία επικοινωνία. Το μόνο που κατορθώνουν είναι να κατανοήσουν ότι η επικοινωνία διεξάγεται κανονικά και όχι να αναγνωρίσουν ποιοι συμμετέχουν σε αυτή.

Το Onion routing αποτελείται από δύο κύρια μέρη (Reed, Syverson, & Goldschlag):

• Τη δικτυακή υποδομή που εξυπηρετεί τις ανώνυμες συνδέσεις και περιλαμβάνει τους δρομολογητές Onion

• Τους πληρεξούσιους που μεσολαβούν στις εφαρμογές του χρήστη και στις συνδέσεις στο Internet.

Η δρομολόγηση onion έχει στηριχτεί πάνω στην ιδέα των Mix Networks τα οποία δημιουργήθηκαν στις αρχές του 1980 από τον David Chaum που είναι και ο

εφευρέτης διαφόρων κρυπτογραφικών πρωτοκόλλων στα οποία περιλαμβάνονται οι ψηφιακές υπογραφές (digital signature), η ηλεκτρονική ψηφοφορία (voting systems) και το ψηφιακό χρήμα (digital cash) (Reed, Syverson, & Goldschlag).

Τα mix networks, τα δίκτυα δηλαδή που χρησιμοποιούν δρομολόγηση onion, δημιουργούν μία κατάσταση σχεδόν αδύνατης ανίχνευσης κάνοντας χρήση των proxy servers. Τα μηνύματα που διακινούνται στα δίκτυα αυτά, κρυπτογραφούνται με ένα δημόσιο κλειδί σε κάθε proxy server. Έτσι, ο κάθε Proxy που συμμετέχει στο δίκτυο, απογυμνώνει το μήνυμα από ε΄να επιπεδο κρυπτογράφησης ώστε να το στείλει προς τον παραλήπτη. Βασική έννοια στα mixnets είναι ο MIX, ένας proxy που αποδέχεται τα κρυπτογραφημένα μηνύματα με το public key τους τα αποκωδικοποιεί τα ταξινομεί και τα προωθεί στον τελικό αποδέκτη τους διαγράφοντας όλες τις πληροφορίες για την πηγή τους (Reed, Syverson, & Goldschlag).

Πλεονέκτημα της παραπάνω τεχνικής αποτελεί το γεγονός ότι η ανώνυμη επικοινωνία πραγματοποιείται σε οποιαδήποτε περίπτωση. Το παραπάνω συμβαίνει λόγω του ότι κάθε δρομολογητής λαμβάνει τα μηνύματα, τα κρυπτογραφεί εκ νέου και στη συνέχεια τα προωθεί. Η δρομολόγηση αυτή, δε μπορεί φυσικά να εγγυηθεί την πλήρη μυστικότητα καθώς το επίπεδο γνώσεων του υποκλοπεά δεν είναι σε καμία περίπτωση γνωστό. Ασφαλώς καυ υπάρχει ένα επίπεδο προστασίας ο οποίος εξαρτάται πάντα όμως από τον αριθμό και τη γνώση των κακόβουλων επιτιθέμενων.

Τα μηνύματα σε μία δρομολόγηση onion, κρυπτογραφούνται κατ επανάληψη και αποστέλλονται διαμέσου διαφορετικών κόμβων του δικτύου οι οποίοι έχουν δημιουργήσει μοναδικές συνδέσεις με τους γειτονικούς δρομολογητές. Κάθε δρομολογητής που συμμετέχει σε ένα δίκτυο onion, γνωρίζει τα δημόσια κλειδιά αλλά και την ταυτότητα των λοιπών δρομολογητών, αφαιρεί ένα στρώμα της κρυπτογράφησης για να αποκαλύψει τις οδηγίες δρομολόγησης και στέλνει το μήνυμα στον επόμενο δρομολογητή, όπου αυτό επαναλαμβάνεται έως ότου το μήνυμα φτάσει στον τελικό προορισμό του (Hooks & Miles, 2006).

Το μονοπάτι που θα ακολουθήσει το μήνυμα περιλαμβάνει σειρά δρομολογητών ενώ για κάθε δρομολογητή δημουργείται το αντίστοιχο πακέτο το οποίο περιλαμβάνει την IP του επόμενου δρομολογητή αλλά και όλες τις απαραίτητες για την κρυπτογράφηση πληροφορίες. Έτσι λοιπόν, ο κόμβος ο οποίος ξεκινά την αποστολή ενός μηνύματος, κρυπτογραφεί τα στοιχεία που θέλει να επικοινωνήσει σε διάφορα επίπεδα με δημόσιο πάντα κλειδί και τα στέλνει προς τον πρώτο κόμβο της αλυσίδας. Ο πρώτος κόμβος, θα αφαιρέσει το στρώμα της κρυπτογράφησης ενώ ο δεύτερος θα μεταλάξει

και πάλι τα στοιχεία, θα αφαιρέσει ένα στρώμα κρυπτογράφησης και θα προωθήσει το μήνυμα προς τον τρίτο κόμβο. Η παραπάνω διαδικασία ακολουθείται έως ότου το μήνυμα να φτάσει στον οριστικό παραλήπτη τουν (Hooks & Miles, 2006).

Οι δρομολογητές που συμμετέχουν σε ένα onion δίκτυο, λειτουργούν ως μεσάζοντες και για άλλες ανώνυμες συνδέσεις. Ένας δρομολογητής onion, δημιουργεί και διαχειρίζεται σειρά ανώνυμων συνδέσεων και άρα αποτελεί το πλέον αξιόπιστο μέρος σε όλο το σύστημα. Αυτός αποτελείται από τα παρακάτω επίπεδα (Hooks & Miles, 2006):

- Ένα μη απαραίτητο φίλτρο, εξειδικευμένο σε εφαρμογές, που ελέγχει και διαμορφώνει κατάλληλα τις ροές των δεδομένων.
- Ένα φίλτρο, εξειδικευμένο για εφαρμογές onion, το οποίο μεταφράζει τις ροές των δεδομένων σε μορφή ανεξάρτητη από την εκάστοτε εφαρμογή η οποία είναι αποδεκτή από το δίκτυο του Onion Routing.
- Έναν onion rooter, ο οποίος οικοδομεί και διευθύνει ανώνυμες συνδέσεις και αποτελεί το πιο αξιόπιστο τμήμα του συστήματος.

Η δημιουργία ενός onion, ξεκινά από το δρομολογητή που βρίσκεται στην κεφαλή ο οποίος επιλέγει ένα αριθμό δρομολογητών και παράγει ένα μήνυμα για τον καθένα. Το μήνυμα αυτό προωθείται με συμμετρικά κλειδιά ενώ σε όλο το μονοπάτι υπάρχει πληροφορία για το ποιος θα είναι ο επόμενος δρομολογητής. Τα μηνύματα αυτά είναι κρυπτογραφημένα, γεγονός το οποίο εξασφαλίζει μία επίπεδο δομή στο σύστημα.

Απαραίτητο φυσικά για τον δρομολογητή είναι να γνωρίζει την τοπολογία του δικτύου, την κατάσταση αυτού, αλλά και τα δημόσια πιστοποιητικά ώστε να μπορεί να διαχεριστεί τη ροή των πληροφοριών στο δίκτυο.

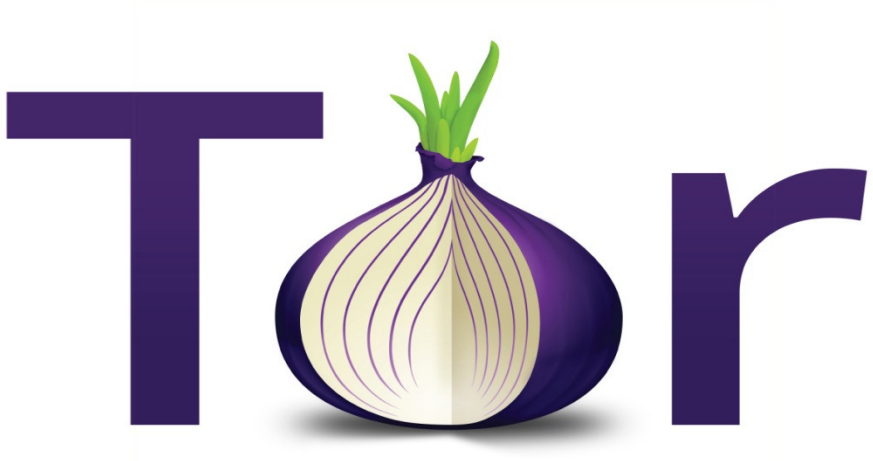
Οι socket συνδέσεις βρίσκονται σε χαμηλότερο επίπεδο από αυτό της εφαρμογής του πρωτοκόλλου, συνεπώς είναι ανεξάρτητες εφαρμογής. Προκειμένου να χρησιμοποιηθεί η υποδομή των των Onion δρομολογητών, αρκεί οι δικτυακές εφαρμογές του χρήστη να υποστηρίζουν onion. Οι συνδέσεις που δημιουργούνται είναι (Hooks & Miles, 2006):

- Socket σύνδεση μεταξύ αποστολέα και Onion rooter.
- Ανώνυμη σύνδεση μεταξύ Onion rooter αποστολέα και παραλήπτη.

Η λειτουργία του onion rooting ξεκινά με την προυπόθεση ότι κάθε δρομολογητής αναγνωρίζει τα στοιχεία, την ταυτότητα αλλά και τα δημόσια κλειδιά του επόμενου στην αλυσίδα δρομολογητή. Καθώς το μήνυμα προωθείται μέσα στο δίκτυο, τα διάφορα επίπεδα που έχει αυτό ξετυλίγονται και όταν το πακέτο φτάσει στον τελευταίο δρομολογητή, φτάνει τελικά και στον παραλήπτη του. Οι απαντήσεις ακολουθούν το ίδιο μονοπάτι, από τον παραλήπτη προς τον αποστολέα ώστε να ενημερώσουν για την επιτυχημένη πορεία του μηνύματος.

## 3.3 Το TOR

Η πιο σημαντική εφαρμογή δρομολόγησης onion είναι το TOR, το οποίο δεν αποτελεί τίποτε άλλο από ένα πρόγραμμα λογισμικού που έχει ως στόχο να προστατέψει τις επικοινωνίες στο διαδίκτυο από οποιαδήποτε παρακολούθηση απειλεί την ελευθερία και την ιδιωτικότητα (Dingledine, Mathewson, & Syverson).
Το παραπάνω συντελείται διαμέσου της σύνδεσης του υπολογιστή σε ένα δίκτυο υπολογιστών όπου χρησιμοποιείται κρυπτογράφηση ώστε να διασφαλιστεί η μυστικότητα των πακέτων κατά την επικοινωνία. Η υλοποίηση του TOR κάνει χρήση διαφόρων proxy servers οι οποίοι ελέγχουν την κυκλοφορία του δικτύου και κρυπτογραφούν τα δεδομένα που κινούνται εντός του (Dingledine, Mathewson, & Syverson). .



Εικόνα 9 : TOR

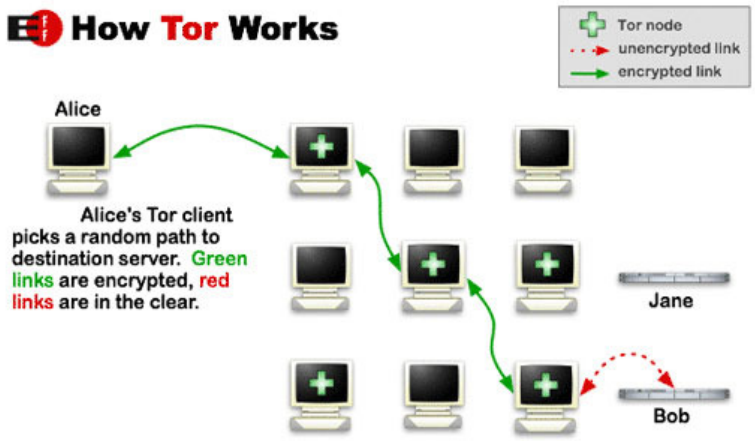
Υπάρχει σειρά εφαρμογών που υποστηρίζουν το TOR και όλες βασίζονται στο πρωτόκολλο TCP. Οι χρήστες του TOR είναι έως σήμερα άγνωστοι στον αριθμό και μεταξύ τους συγκαταλέγονται δημοσιογράφοι, ακτιβιστές, κακοποιοί, στρατιωτικοί, απλοί πολίτες, hackers και εγκληματίες.

Η διασφάλιση της ανωνυμίας από το TOR δεν είναι σε καμία περίπτωση σίγουρη εάν αυτό δε χρησιμοποιηθεί σωστά. Πολλοί από τους χρήστες του, αναγνωρίζουν την χρησιμότητά του καθώς έχουν ως στόχο να προβούν σε αξιόποινες πράξεις εντός του. Υπάρχει βέβαια και ένας αρκετά σημαντικός αριθμός χρηστών που κάνουν χρήση του TOR όχι για να εγκληματήσουν αλλά για να διαφυλάξουν την ανωνυμία τους.

Το TOR χρησιμοποιείται και από διάφορου κυβερνητικούς φορείς ώστε αυτοί να είναι σε θέση να εντοπίσουν εντός του dark net σειρά κακοποιών.

Η λειτουργία του TOR ξεκινά με την ύπαρξη ενός δικτύου που φυσικά αποτελείται από διάφορους servers. Το TOR συνδέεται στο δίκτυο αυτό και μεταφέρει τα πακέτα μεταξύ των κόμβων ανώνυμα ώστε να μειώσει τον οποιοδήποτε κίνδυνο υποκλοπής των στοιχείων από το δίκτυο (Dingledine, Mathewson, & Syverson).

Αντί για τη λήψη μιας άμεσης διαδρομής από την πηγή στον προορισμό, τα πακέτα δεδομένων παίρνουν μια τυχαία ροή μέσω διαφορετικών relays (αναμεταδότες), που καλύπτουν τις διαδρομές μας, με αποτέλεσμα κανένας παρατηρητής σε κανένα σημείο να μη μπορεί να ξεχωρίσει τα δεδομένα από πού προέρχονται και ποιος είναι ο προορισμός τους (NATO, 2015).




Εικόνα 10 : Λειτουργία του TOR

Κάθε ιδιωτικό μονοπάτι στο TOR δημιουργείται από ένα κύκλωμα κρυπτογραφημένων συνδέσεων εντός του δικτύου. Το κύκλωμα αυτό επεκτείνεται σταδιακά ενώ κάθε relay που βρίσκεται σε αυτό γνωρίζει από πού παρέλαβε τα δεδομένα και που θα πρέπει να τα προωθήσει (NATO, 2015). Ο client διαπραγματεύεται ένα ξεχωριστό ζευγάρι κλειδιών για κάθε βήμα κατά μήκους του κυκλώματος, έτσι ώστε να σιγουρέψει ότι κάθε βήμα δεν μπορεί να εντοπιστεί όσο το πακέτο δρομολογείται στο κύκλωμα.

Το TOR λειτουργεί αποκλειστικά και μόνο πάνω από το TCP και χρησιμοποιεί συνδέσεις που πραγματοποιούνται με μέγιστο το χρόνο των 10 λεπτών. Οποιαδήποτε άλλη σύνδεση αποκόπτεται. Οι χρήστες που θέλουν να εισέλθουν στο TOR, εκτελούν στην υπολογιστή τους έναν onion proxy server ο οποίος του συνδέει με το TOR δίκτυο (NATO, 2015). Εκεί, πραγματοποιείται κρυπτογράφηση υψηλού επιπέδου ώστε να διασφαλιστεί η ανωνυμία. Τα δεδομένα, αποστέλλονται κρυπτογραφημένα από τον ένα δρομολογητή στον άλλο ώστε να φτάσουν ανέπαφα στον προορισμό τους.

Το Tor ξεχωρίζει από τα άλλα δίκτυα ανωνυμίας γιατί λειτουργεί στο TCP επίπεδο. Οι ανώνυμες εφαρμογές που περιλαμβάνει είναι το IRC (Internet Relay Chat), τα άμεσα μηνύματα και η πρόσβαση στο διαδίκτυο.

Το μοναδικό συγκριτικό πλεονέκτημα που εισάγει το TOR είναι η ανωνυμία που παρέχει στους servers του, γεγονός που τους καθιστά σχεδόν αδύνατους στο να εντοπιστούν. Φυσικά για να υπάρχει πρόσβαση σε συγκεκριμένη υπηρεσία, χρήση του TOR, θα πρέπει να κάνουν και οι χρήστες οι οποίοι φυσικά μπορούν να

συνδεθούν σε οποιαδήποτε υπηρεσία χωρίς ο ένας να γνωρίζει την ταυτότητα του άλλου.

Κάθε υπηρεσία, δε θα μπορούσε να είναι κάτι άλλο από έναν υπολογιστή ο οποίος επιτρέπει σε σειρά χρηστών να συνδεθούν με αυτών χωρίς την αποκάλυψη της ταυτότητά τους. Ο κεντρικός υπολογιστής που συμμετέχει στο δίκτυο, δημιουργεί ένα κύκλωμα και διαχέει την υπηρεσία εντός του δικτύου αυτού. Όταν ένας χρήστης επιθυμεί να συνδεθεί με την υπηρεσία, στέλνει αίτημα προς το TOR το οποίο με τη σειρά του ορίζει το Rendezvous Point, δηλαδή τον κόμβο διαμέσου του οποίου θα επικοινωνούν οι δύο πλευρές (NATO, 2015).

Πέρα από τα μοναδικά του πλεονεκτήματα, το TOR δε θα μπορούσε παρά να παρουσιάζει και κάποια αρνητικά στοιχεία. Ένα από τα προβλήματά του είναι η διαρροή του DNS και άρα οι χρήστες που είναι συνδεδεμένοι σε αυτό. Το TOR στην περίπτωση αυτή δεν είναι σε θέση να κρύψει το ποιος χρησιμοποιεί το δίκτυο, αλλά κατά βάση το λόγο για τον οποίο το χρησιμοποιεί. Έτσι λοιπόν, εάν οι επιτιθέμενοι είναι σε θέση να εντοπίσουν τα πακέτα που διακινούνται σε ένα δίκτυο, ενδεχόμενα να εντοπίσουν και ποιοι είναι οι χρήστες εκείνοι που κάνουν χρήση του δικτύου αυτού. Μία προτεινόμενη λύση είναι η χρήση ενός Privoxy, ενός λογισμικού δηλαδή που περιλαμβάνει φίλτρα προστασίας της ιδιωτικότητας.

Πολλές φορές η κρυπτογράφηση των πακέτων εντός του TOR δεν επαρκεί ώστε να προσφέρει επαρκή προστασία από τους επιτιθέμενους που κατασκοπεύουν μέρη του διαδικτύου και χρησιμοποιούν διάφορα εργαλεία για να αποσπάσουν πληροφορία. Έτσι, καθίσταται σχετικά δύσκολο για το TOR ώστε αυτό να ελέγξει την κυκλοφορία καθώς αυτή μεταφέρεται από το ένα άκρο της σύνδεσης προς το άλλο. Μάλιστα, εάν κάποιος από τους κόμβους δεν χρησιμοποιήσει κρυπτογράφηση από άκρο σε άκρο τότε είναι εύκολο να αποκρυπτογραφηθεί η πληροφορία και φυσικά να διαρρεύσει. Ένας ωτακουστής είναι ένας επιτιθέμενος ικανός να παρακολουθεί όλες τις πληροφορίες που είτε αποστέλλονται, είτε λαμβάνονται από κάποιο συγκεκριμένο συμμετέχοντα, με σκοπό να ανιχνευθεί είτε ο ιδρυτής, είτε ο παραλήπτης για κάθε επικοινωνία. Οι ωτακουστές αντιμετωπίζονται δύσκολα, ακριβώς επειδή μπορούν να καταγράφουν και να συγκρίνουν όλα τα εισερχόμενα και εξερχόμενα μηνύματα.

Δεν είναι λίγες οι φορές εκείνες όπου οι κόμβοι προσπαθούν να εκμεταλλευθούν τις συνδέσεις διαμέσου TOR ώστε να διασφαλίσουν την ανωνυμία τους. Για την περίπτωση αυτή το TOR χρησιμοποιεί πολιτικές αστυνόμευσης που ελέγχουν που

παραχωρείται η πρόσβαση και που όχι. Οι περισσότερες εκμεταλλεύσεις δικτύου έχουν χαρακτηριστικά όπως (NATO, 2015):

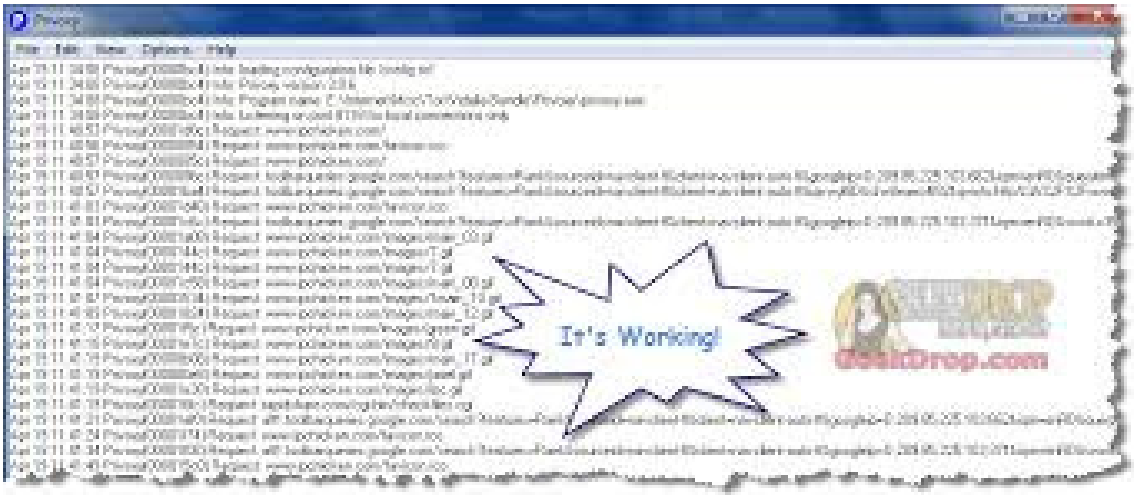
- Εκμετάλλευση εύρους ζώνης με μεταφορά μεγάλου όγκου δεδομένων.
- Λόγω της χρησιμοποίησης περισσότερου εύρους ζώνης δεν προτείνονται και μεταφορές Bit torrent.
- Αποστολή spam μηνυμάτων.
- Κάποιοι ανώνυμοι χρήστες δεν δέχονται τις διαφορετικές πολιτικές πρόσβασης σελίδων του web.

Τέλος, μεταξύ του τεράστιου αριθμού των χρηστών που χρησιμοποιούν το TOR βρίσκεται και ένας αριθμός κακοποιών που εκμεταλλεύονται την ανωνυμία για να προβούν σε παράνομες ενέργειες. Έχει παρατηρηθεί ότι το ποσοστό αυτών που το χρησιμοποιούν για παράνομες χρήσεις είναι μεγαλύτερο από αυτούς που το χρησιμοποιούν καθαρά για ανωνυμία και ιδιωτικότητα της επικοινωνίας τους.

## 3.4 Εργαλεία του TOR

Η εισαγωγή στο TOR απαιτεί τη χρήση μιας σειράς εργαλείων τα οποία περιγράφονται στην ενότητα που ακολουθεί. Το πρώτο εργαλείο το οποίο καλείται να χρησιμοποιήσει κανείς είναι το Privoxy. Πρόκειται για ένα πρόγραμμα το οποίο αποτελεί ιδανική επιλογή για όσους θέλουν πάση θυσία να διαφυλάξουν την ιδιωτικότητά τους. Το συγκεκριμένο λογισμικό βασίζεται στον proxy server κα χρησιμοποιείται πάντα σε συνδυασμό με το TOR. Στόχος των δύο είναι η απόκρυψη της IP του χρήστη. Αυτό επιτυγχάνεται διαμέσου της δρομολόγησης μέσω από onion δρομολογητές (Privoxy).

Το Privoxy είναι ένας web server με απεριόριστες επιλογές και υψηλή λειτουργικότητα. Μάλιστα έχει την ιδιότητα να προωθεί όλα τα HTTP αιτήματα, κάτι στο οποίο αποτυγχάνει το TOR καθώς αυτό προωθεί μονάχα την κίνηση.




Εικόνα 11 : Privoxy

Χαρακτηριστικά Privoxy (Privoxy):

- Μπορεί να κρατά ενεργές τις εξερχόμενες συνδέσεις και να τις επαναχρησιμοποιήσει αργότερα.
- Υποστηρίζει την αλλαγή συμπεριφοράς η οποία αφορά τις επικεφαλίδες των client και servers.
- Μπορεί να «τρέχει» ως ενδιάμεσος proxy server, ο οποίος προλαμβάνει την ανάγκη να χωριστούν οι μηχανές αναζήτησης χωριστά.
- Περιέχει περίπλοκες ενέργειες και φίλτρα για να χειρίζεται τις επικεφαλίδες των server και client.
- Μπορεί να συνεργαστεί και με άλλους proxy servers.
- Μπορεί να φιλτράρει τις ιστοσελίδες, δηλαδή κάνει αντικατάσταση κειμένου, αφαιρεί τα bunner, διαθέτει ορατά εργαλεία για διαχείριση, αγνοεί ταεπανεμφανιζόμενα παράθυρα και τα κομμάτια κώδικα javascript και html κλπ.
- Βελτιώνει τη διαχείριση των cookies.
- Υποστηρίζει πολυμέσα.

Μία άλλη εφαρμογή η οποία χρησιμοποιείται στο TOR είναι το Vidalia GUI, το οποίο χρησιμοποιείται ουσιαστικά για τη διαχείριση του TOR. Καθώς το λογισμικό εγκαθίσταται στον υπολογιστή, εγκαθίσταται μαζί με αυτό και ένας Privoxy ο οποίος επιτυγχάνει τη σύνδεση με το δίκτυο TOR. Άρα λοιπόν, το Vidalia, αποτελεί τον

client εκείνο που θα εγκαταστήσει στον υπολογιστή το TOR και τον Privoxy και την ίδια στιγμή θα κάνει ιδιαίτερα εύκολο το χειρισμό τους (UbuntuGeek).

Μέσα από τον πίνακα ελέγχου του λογισμικού ο χρήστης μπορεί να ελέγξει τη σύνδεση με το TOR, να την εκκινήσει και να τη σταματήσει όπως επιθυμεί αυτός.

Το Torcap, αποτελεί και αυτό ένα εργαλείο το οποίο επιτρέπει σε οποιαδήποτε διαδικτυακή εφαρμογή να συνδεθεί με το TOR. Η λειτουργία του εισάγει ένα DLL ( Dynamic Link Library ) σε κάθε διαδικασία που τρέχει στον υπολογιστή του χρήστη. Αυτό το DLL δουλεύει με το Winsock API (Winsock Application Programming Interface ) το οποίο συνεργάζεται τη παροχή DNS ( Domain Name System ) και τις συνδέσεις TCP/IP (UbuntuGeek). Η απόφαση για το πότε θα συνδέεται ο χρήστης μέσω του Tor ή απευθείας στο internet, εξαρτάται από την μάσκα της εφαρμογής ( application mask ) και τη μάσκα του εξυπηρετητή (host mask). Όταν αυτά τα δύο ταιριάξουν τότε η σύνδεση θα γίνει μέσω του Tor Socks server (UbuntuGeek).

Το Torcap 2 είναι και αυτό ένα ιδιαίτερα ελαφρύ λογισμικό το οποίο βασίζεται επάνω στο TOR και εκτελεί ακριβώς την ίδια λειτουργία. Ανάλογη είναι και η λειτουργία του Freecap το οποίο ασχολείται με τις συνδέσεις των socks με τους servers (UbuntuGeek).

Το OperaTor αποτελεί το λογισμικό εκείνο το οποίο εγκαθίσταται στις φορητές μνήμες και επιτρέπει την ανώνυμη πλοήγηση στο διαδίκτυο όταν ο χρήστης βρίσκεται σε δημόσιους και όχι μόνο χώρους. Το OperaTor έχει μία σειρά ιδιαίτερων χαρακτηριστικών μεταξύ των οποίων τη μη αποθήκευση των πληροφοριών πλοήγησης στον υπολογιστή που έχει χρησιμοποιηθεί (UbuntuGeek).

Ανάλογο λογισμικό είναι και ο Xerobank browser ο οποίος επιτρέπει την ανώνυμη πλοήγηση στο διαδίκτυο. Σε αυτό περιλαμβάνονται oxB Browser, το xB Mail και το xB VPN. Ο xB Browser χρησιμοποιείται για πρόσβαση στον web browser μέσω φορητής μνήμης και μπορεί να προσφέρει υψηλού επιπέδου ανωνυμία. Ακόμη, κρυπτογραφεί κάθε δραστηριότητα του χρήστη και αποτρέπει τον εντοπισμό αυτού από οποιονδήποτε κακόβουλο εισβολέα (UbuntuGeek).

Τέλος, ένα χρήσιμο εργαλείο TOR είναι το TorChat, ένας peer to peer instant messenger ο οποίο παρέχει πλήρη ανωνυμία κατά τη χρήση του και μπορεί να εγκατασταθεί σε οποιοδήποτε λειτουργικό σύστημα.

## 3.5 Επιθέσεις στο Dark Net

Τα τελευταία χρόνια παρατηρείται μία ραγδαία αύξηση στη συχνότητα των διαδικτυακών επιθέσεων αλλά και των εγκλημάτων που συντελούνται μέσω του διαδικτύου. Στόχος των επιθέσεων αυτών είναι κυρίως τα προσωπικά δεδομένα των χρηστών με απώτερο σκοπό την αρπαγή των χρημάτων, την χρήση των πιστωτικών καρτών και φυσικά την αγοραπωλησία των στοιχείων αυτών.

Η πλέον γνωστή μορφή κυβερνοεπίθεσης είναι αυτή των Δούρειων Ίππων που εμφανίζονται ως ένα απλό καλόβουλο λογισμικό και τελικά κατορθώνει να υποκλέψει τα δεδομένα του χρήστη. Προς την κατεύθυνση αυτή κατευθύνονται και οι ψεύτικοι τραπεζικοί ιστότοποι οι οποίοι φαινομενικά είναι ίδιοι με τους πραγματικούς αλλά έχουν φτιαχτεί ώστε οι χρήστες να καταχωρούν τα στοιχεία τους και οι κακόβουλοι να τα συλλέγουν προς όφελός τους. Δυστυχώς για τον απλό χρήστη είναι σχεδόν αδύνατο να διαχωρίσει τη διαφορά ανάμεσα στον πραγματικό και τον ψεύτικο ιστότοπο και έτσι πολλές φορές γίνεται θύμα αυτών των επιθέσεων και χάνει σημαντικά προσωπικά δεδομένα που αφορούν σε τραπεζικές συναλλαγές (Gareth & Nick, 2015).

Η κοινωνική μηχανική είναι μία μορφή κυβερνοεγκλήματος η οποία αναπτύσσεται τα τελευταία χρόνια ιδιαίτερα ραγδαία και βασίζεται εξ ολοκλήρου στην ανθρώπινη επικοινωνία. Ο επιτιθέμενος συχνά προσποιείται έναν άλλο σχεδόν αξιόπιστο ρόλο ώστε να εκμαιεύσει στοιχεία από το χρήστη και να τα χρησιμοποιήσει κατά βούληση.

Μεταβαίνοντας στον τομέα των λογισμικών θα μπορούσε κανείς να υποστηρίξει πως οι sniffers είναι πλέον μεταξύ των πιο επικίνδυνων λογισμικών υποκλοπής στοιχείων. Πρόκειται για ένα λογισμικό το οποίο ανιχνεύει την κίνηση στο δίκτυο και αποσκοπεί στην υποκλοπή στοιχείων. Ακόμη, έχει τη δυνατότητα να διαμορφώνει υπολογιστές μέσα σε ένα δίκτυο ώστε να δέχονται μηνύματα, οι χρήστες τους να απαντούν σε αυτά τα μηνύματα και να συλλέγει πληροφορίες από αυτές τις απαντήσεις (Gareth & Nick, 2015).

Ιδιαίτερα γνωστή τα τελευταία χρόνια είναι και η τεχνική του phising η οποία λειτουργεί με την αποστολή e mail σε χρήστες. Το e mail υποτίθεται ότι αποστέλλεται από μία τράπεζα ή έναν έγκυρο οργανισμό και έχει ως στόχο την υποκλοπή των δεδομένων. Μάλιστα, το e mail προτρέπει το χρήστη ώστε να αυτός να μεταβεί σε μία ιστοσελίδα και εκεί να συμπληρώσει τα στοιχεία του ώστε να υποκλαπούν.

Μία ακόμη γνωστή απειλή για το διαδίκτυο είναι οι key loggers οι οποίοι καταγράφουν τις πληκτρολογήσεις του χρήστη χωρίς αυτός να το γνωρίσει. Πρόκειται φυσικά για μια τεράστια απειλή για τη διαρροή των προσωπικών δεδομένων καθώς καταγράφει οποιαδήποτε ενέργεια έκανε ο χρήστης ή οποιαδήποτε λέξη αυτός πληκτρολόγησε. (Gareth & Nick, 2015).

# Κεφάλαιο 4 – Deep και Dark Web

## 4.1 Ομοιότητες και Διαφορές

Οι όροι dark & deep web δεν πρέπει σε καμία περίπτωση να συγχέονται καθώς δεν πρόκειται για το ίδιο ακριβώς πράγμα. Το λεξικό Cambridge ορίζει το deep web ως «το τμήμα του Internet που είναι κρυμμένο από τις συμβατικές μηχανές αναζήτησης, όπως με την κρυπτογράφηση, ως το άθροισμα των μη ευρετηριασμένων ιστοσελίδων» ενώ το dark web, «ως το τμήμα του Internet που σκόπιμα κρύβεται από τις μηχανές αναζήτησης, χρησιμοποιεί αποκρυμμένες διευθύνσεις IP, και είναι προσβάσιμο μόνο με ένα ειδικό πρόγραμμα περιήγησης στο web: μέρος του deep web (Amores & Motta, 2012).

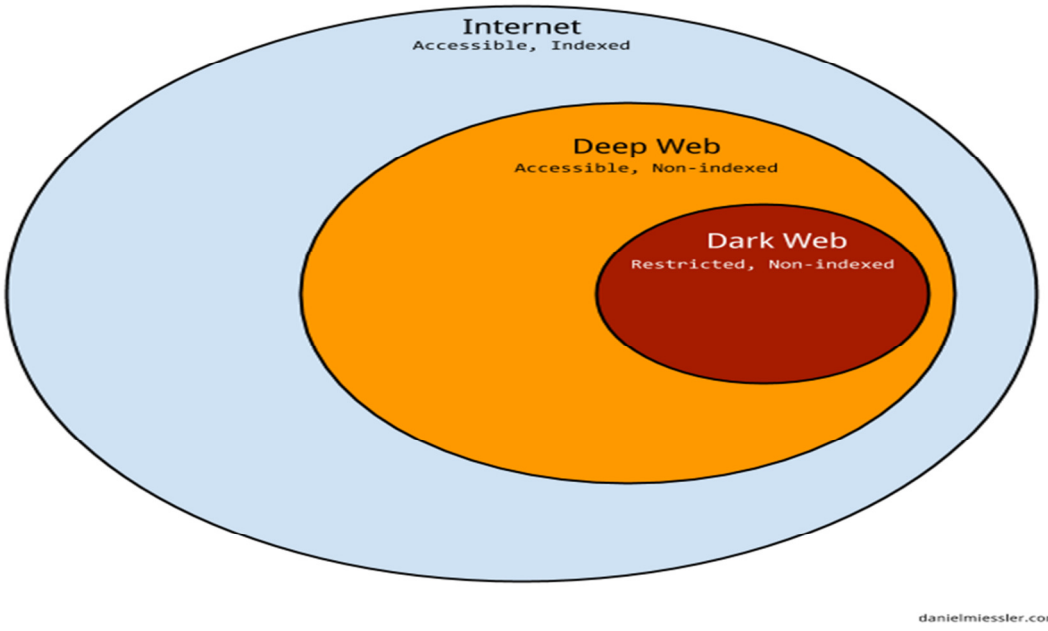
Αυτό λοιπόν το οποίο θα πρέπει να γίνει άμεσα κατανοητό είναι ότι το dark web αποτελεί μέρος του deep web. Κοινό χαρακτηριστικό και των δύο είναι ότι αυτά δεν είναι εντοπίσιμα από τις γνωστές μηχανές αναζήτησης. Δηλαδή, εάν κάποιος πραγματοποιήσει μία αναζήτηση είτε στη Google είτε στη Bing, δεν θα έχει πρόσβαση σε αυτά.
Ο Greenberg υποστηρίζει ότι το deep web είναι τεράστιο και αναλογεί στο 90% και επιπλέον του Διαδικτύου, ενώ το dark web πιθανότατα αντιπροσωπεύει μονάχα το 0,01%. Η πρόσβαση στο dark web που πολλές φορές καλείται και ως dark net γίνεται μονάχα διαμέσου το TOR ώστε τόσο οι χρήστες όσο και οι ιδιοκτήτες των ιστοσελίδων να διατηρήσουν την ανωνυμία τους (Amores & Motta, 2012).
Ο Max Eddy για το PC Magazine αναφέρει ότι «το TOR αρχικά αναπτύχθηκε από το Υπουργείο Άμυνας των ΗΠΑ», και ενώ τώρα είναι μια μη κερδοσκοπική επιχείρηση από εθελοντές, χρηματοδοτείται από την κυβέρνηση των ΗΠΑ και το Εθνικό Ίδρυμα Επιστημών. Στο Vox, ο Timothy B. Lee εξηγεί: «Η κυβερνητική υποστήριξη για το TOR συνεχίστηκε κατά τα τελευταία χρόνια, ως μέρος της ατζέντας της ελευθερίας του διαδικτύου του Στέιτ Ντιπάρτμεντ, η οποία επιδιώκει να βοηθήσει τους ανθρώπους σε καταπιεστικά καθεστώτα να αποκτήσουν πρόσβαση σε πληροφορίες που λογοκρίνονται από τις κυβερνήσεις τους». Για παράδειγμα, το Facebook ξεκίνησε πρόσφατα μια έκδοση του site του στο dark web για να «καταστήσει ευκολότερη την πρόσβαση στην ιστοσελίδα από χώρες που περιορίζουν την υπηρεσία, όπως η Κίνα και το Ιράν» (Ban, et al, 2012).

Και τα δύο συστήματα, βασίζουν τη λειτουργία τους στην κρυπτογράφηση της κίνησης που συντελείται εντός τους, αφαιρώντας ένα επίπεδο κρυπτογράφησης καθώς το μήνυμα μεταφέρεται από τον ένα δρομολογητή στον άλλο. Ο χρήστης του TOR καταφέρνει να αποκρύψει την IP του ενώ από την άλλη ιστοσελίδα χρησιμοποιεί την υπηρεσία του TOR ώστε να είναι επισκέψιμη μόνο από χρήστες που το χρησιμοποιούν. Η κίνηση στο δίκτυο δρομολογείται από τυχαίους υπολογιστές έως ότου τα δύο μέρη να συναντηθούν σε ένα σημείο και να αλλάξουν την πληροφορία.

Ένα από τα πρώτα site σε επισκεψιμότητα εντός του dark web είναι το wikileaks το οποίο δημιουργήθηκε με στόχο να συλλέγει πληροφορία από ανώνυμες πηγές. Η ιστοσελίδα αυτή κάνει χρήση του SecureDrop, ενός λογισμικού που ενσωματώνει όλες τις υπηρεσίες του TOR ώστε να λαμβάνει ανώνυμες υποβολές ειδησεογραφικού υλικού (Amores & Motta, 2012).

Ακόμη και το ίδιο το Facebook έχει ξεκινήσει μία νέα ιστοσελίδα στο dark web ώστε να είναι σε θέση να προστατέψει τους χρήστες της και αυτοί να μην αποτελούν αντικείμενο λογοκρισίας.

Τόσο το deep όσο και το dark web κάνουν χρήση των crawlers ώστε να πραγματοποιήσουν αναζήτηση στις βάσεις δεδομένων που χρησιμοποιούν και να εξάγουν πληροφορία.




Εικόνα 12 : Internet, Deep Web, Dark Web

# Κεφάλαιο 5 - Επικινδυνότητα

## 5.1 Κίνδυνοι για τον απλό χρήστη - Επιθέσεις από hackers, απώλεια προσωπικών δεδομένων

Η ιδιωτική ζωή των ανθρώπων είναι ένα ιδιαίτερο και πολύ σημαντικό καθώς περιλαμβάνει δεδομένα που οι άνθρωποι δεν θέλουν να διαμοιράζονται με τρίτους. Η ανωνυμία λοιπόν, και η διασφάλιση των δεδομένων αυτών προσφέρει μία αίσθηση ασφάλειας η οποία συνδυάζεται πάντα με το γεγονός ότι οι άνθρωποι αισθάνονται πως κανείς δεν παρακολουθεί τη ζωή τους και τις δραστηριότητές τους.
Τι είναι όμως τα προσωπικά δεδομένα ; Ως προσωπικό δεδομένο καλείται κάθε πληροφορία που χαρακτηρίζει έναν άνθρωπο όπως για παράδειγμα το όνομα, η διεύθυνση, το τηλέφωνο, οι επιδόσεις, τα ενδιαφέρονται, οι σχέσεις, οι δραστηριότητες, οι συναναστροφές και πολλά άλλα. Ακόμη σε αυτά μπορούν να συμπεριληφθούν δεδομένα όπως το θρήσκευμα, η πολιτική πεποίθηση ή ακόμη και η ερωτική ζωή.
Τα προσωπικά δεδομένα των ανθρώπων καταγράφονται καθημερινά σε φόρμες στις οποίοι οι άνθρωποι συναινούν να συμπληρώσουν όπως (ΑΠΠΔ) :

- Η συμμετοχή σε διαγωνισμούς και παιχνίδια
- Η εγγραφή σε on line καταστήματα
- Τα ιατρικά αρχεία
- Τα αρχεία αθλητικών και διάφορων άλλων συλλόγων
- Τα προφίλ στα κοινωνικά δίκτυα

Τα προσωπικά λοιπόν αυτά και ταυτόχρονα πολύ ευαίσθητα δεδομένα ενδέχεται να χρησιμοποιηθούν από επιτήδειους με στόχο την δυσφήμιση ή τον εκβιασμό. Σε πολλές περιπτώσεις κάποιος μπορεί να γίνει θύμα υποκλοπής ταυτότητας και τα στοιχεία του να χρησιμοποιηθούν εν αγνοία του ώστε να συντελεστούν παράνομες πράξεις.

Η νομοθεσία τόσο της χώρας όσο και της Ευρωπαϊκής Ένωσης, έχει θεσπίσει ανάλογους νόμους ώστε να προστατέψει τους ανθρώπους από ην ανεξέλεγκτη χρήση των προσωπικών τους δεδομένων. Η Αρχή Προστασίας Δεδομένων είναι ο αρμόδιος

φορέας για την εφαρμογή αυτής της νομοθεσίας (νόμοι 2472/1997 και 3471/2006) (ΑΠΠΔ) :

Βασική αρχή με βάση την οποία χρησιμοποιούνται τα προσωπικά δεδομένα είναι η συγκατάθεση των ίδιων των ατόμων ενώ σε περίπτωση ανηλίκων η συγκατάθεση των γονιών. Η συγκατάθεση είναι ο γενικός κανόνας, αλλά υπάρχουν και εξαιρέσεις. Για παράδειγμα κάποιοι οργανισμοί, όπως π.χ. ο δήμος ή το σχολείο, μπορούν να επεξεργάζονται συγκεκριμένα προσωπικά δεδομένα χωρίς συγκατάθεσή. Το παραπάνω είναι απαραίτητο ώστε να εκτελεστεί το έργο τους και συνήθως ορίζεται από κάποιο συγκεκριμένο νόμο.

Στην Ευρώπη, το 9% των παιδιών ηλικίας 11-16 ετών έπεσαν θύματα παράνομης χρήσης των προσωπικών τους δεδομένων στο διαδίκτυο. Στη Νορβηγία, το 23% των νέων μεταξύ 8 και 18 έτυχε να δουν φωτογραφίες και άλλα προσωπικά τους δεδομένα στο διαδίκτυο χωρίς την έγκρισή τους. Το 41% αυτών ήταν μεταξύ 17 και 18 ετών. Το 25% των νέων της ίδιας κατηγορίας (17 – 18 ετών) παραδέχεται ότι έχει δημοσιοποιήσει φωτογραφίες και βίντεο στα οποία εμφανίζονται άλλα πρόσωπα, χωρίς την έγκρισή τους. Σε άλλη έρευνα μεταξύ νέων, το 2007 περισσότεροι από ένας στους οκτώ βρήκαν «κακή» πληροφορία για αυτούς στο διαδίκτυο. Το 29% των παιδιών ηλικίας 9-16 που χρησιμοποιούν το διαδίκτυο στην Ευρώπη, έχουν στο παρελθόν επικοινωνήσει από κοντά με κάποιον που γνώρισαν στο διαδίκτυο (ΑΠΠΔ).

Η δημιουργία ενός προφίλ σε ένα κοινωνικό δίκτυο είναι μία ιδιαίτερα εύκολη και απλοποιημένη διαδικασία κατά την οποία αυτός που το δημιουργεί αποφασίζει ποιος θέλει να είναι. Μπορεί λοιπόν είτε να χρησιμοποιήσει την κανονική του ταυτότητα είτε να χρησιμοποιήσει μία άλλη ανάλογα με το πώς αυτός επιθυμεί. Ακραίο αλλά και χαρακτηριστικό παράδειγμα της κατηγορίες αυτής είναι οι ενήλικοι παιδεραστές οι οποίοι αποσκοπούν να πλησιάσουν ανυποψίαστα ανήλικά θύματα.
Το 8% των παιδιών έχουν συναντήσει από κοντά κάποιον που γνώρισαν στο διαδίκτυο κατά τη διάρκεια του προηγούμενου έτους. Το 1% των παιδιών αυτών (ή 1 στα 7 από αυτά που είχαν τέτοια συνάντηση) ενοχλήθηκαν από αυτού του είδους την εμπειρία. Το 9% των παιδιών ηλικίας 11-16 ετών έπεσαν θύματα παράνομης χρήσης των προσωπικών τους δεδομένων (7% παραβίαση κωδικών και 5% παραβίαση άλλων προσωπικών δεδομένων), ή θύματα οικονομικής απάτης (2%) (ΑΠΠΔ)

Μάλιστα, τα κοινωνικά δίκτυα παρουσιάζουν σοβαρά κενά ασφάλειας καθώς ο οποιοσδήποτε μπορεί να αποθηκεύσει στο προφίλ των διαφόρων χρηστών και να το χρησιμοποιήσει. Η υπηρεσία κοινωνική δικτύωσης παρέχεται δωρεάν στους χρήστες και έτσι αναζητά πόρους υποστήριξης από τις διάφορες διαφημιστικές εταιρίες η οποίες προβαίνουν σε στοχευμένες διαφημιστικές καμπάνιες βάση του προφίλ των χρηστών τους. Το παραπάνω αποτελεί και αυτό μία μορφή παραβίασης προσωπικών δεδομένων.

Ακόμη και εάν ένα προφίλ διαγραφεί, η συνολική πληροφορία δε διαγράφεται εντελώς από το κοινωνικό δίκτυο αλλά παραμένει ανενεργή σε κάποιο server. Το παραπάνω σημαίνει άμεσα ότι μπορεί να αποτελέσει αντικείμενο υποκλοπής για οποιονδήποτε επίδοξο.

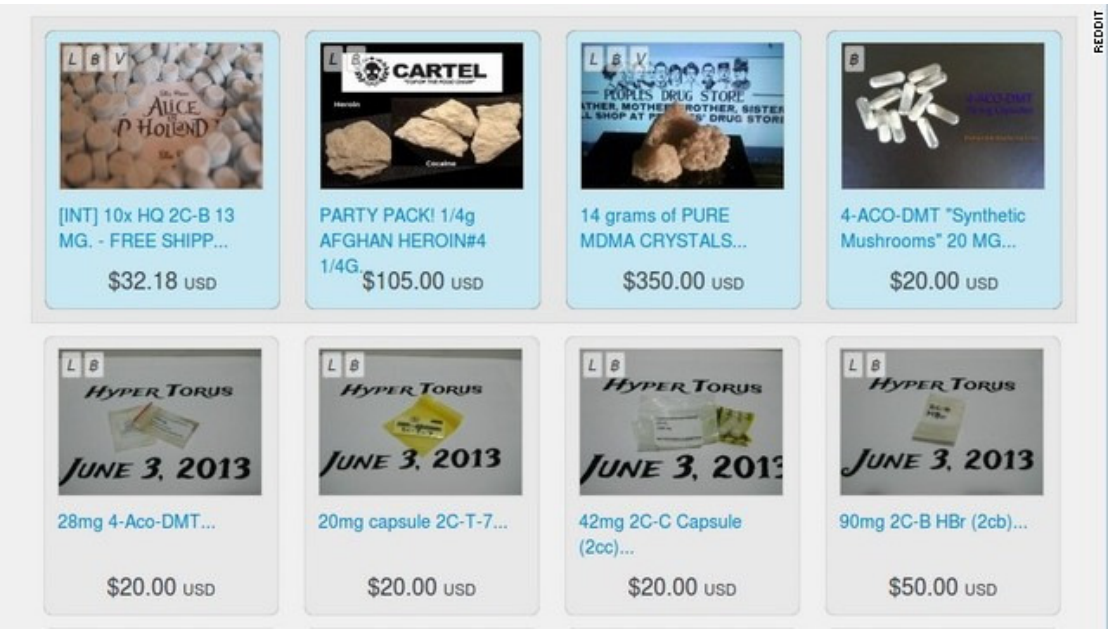
Μία ανάλογη ενέργεια υποκλοπής προσωπικών δεδομένων είναι και αυτή του ψαρέματος ή phising όπου κάποιος αποστέλλει στους χρήστες ένα ψεύτικο παραπλανητικό μήνυμα ώστε να υποκλέψει τα προσωπικά τους δεδομένα. Ο πραγματικός αποστολέας δεν είναι σε καμία περίπτωση αυτός που φαίνεται και έχει ως στόχο να πείσει το χρήστη να του παραχωρήσει ευαίσθητα προσωπικά δεδομένα για να τα χρησιμοποιήσει για δικό του όφελος.

Το κακόβουλο λογισμικό, ανήκει και αυτό στις μεθόδους που χρησιμοποιούν οι διάφοροι επιτήδειοι ώστε να υποκλέψουν τα δεδομένα των χρηστών. Διαμέσου διαφόρων εφαρμογών που εγκαθίστανται στον υπολογιστή, του γνωστού δηλαδή κακόβουλου λογισμικού, επιδιώκουν να υποκλέψουν τα προσωπικά δεδομένα του χρήστη και να τα χρησιμοποιήσουν προς ίδιο όφελος. Η εγκατάσταση των προγραμμάτων αυτών είναι πολύ εύκολη και την ίδια στιγμή πολύ επικίνδυνη.

## 5.2 Μαύρη αγορά στο διαδίκτυο

Ναρκωτικά, όπλα, πλαστές ταυτότητες και διαβατήρια, άνθρωποι αλλά και ότι άλλο παράνομο μπορεί να φανταστεί κανείς μπορεί είτε να το παραγγείλει είτε να το διακινήσει διαδικτυακά. Η μαύρη αγορά έχει εισέλθει για τα καλά στην ψηφιακή εποχή και μάλιστα έχει κατορθώσει σχεδόν να κυριαρχήσει.

Η ιστοσελίδα η οποία έκανε γνωστή στο ευρύ κοινό τη δραστηριότητα αυτή ονομάστηκε Silk Road και διακίνησε τεράστιους όγκους ναρκωτικών, παράνομων προϊόντων και παράνομου χρήματος. Όπως σχολίασε και ένας βρετανός δημοσιογράφος «Αντί για βιβλία και ηλεκτρονικά είδη, μπορεί κάποιος να αγοράσει LSD ή ένα ημιαυτόματο όπλο».




Εικόνα 13 : Μαύρη αγορά στο διαδίκτυο

Όπως προαναφέρθηκε για να προβεί κανείς σε τέτοιου είδος ενέργειες χρειάζεται να εισέλθει στο σκοτεινό διαδίκτυο και αυτό επιτυγχάνεται μονάχα διαμέσου του TOR και όχι των απλών μηχανών αναζήτησης. Ακριβώς όπως ισχύει για τις πλέον γνωστές διαδικτυακές αγορές όπως το e-bay ή το Amazon, έτσι και στο dark web οι διάφοροι πωλητές λαμβάνουν βαθμολογίες γεγονός που εξασφαλίζει ότι οι πελάτες θα εξυπηρετούνται όπως πρέπει. Τα πάντα πακετάρονται άψογα και παραδίδονται

κατευθείαν στην πόρτα του σπιτιού του αγοραστή από τον ταχυδρόμο, με εγγυημένη διακριτικότητα.

Οι διευθύνσεις στις οποίες γίνεται η διακίνηση στο dark web αλλάζουν καθημερινά ώστε να αποφευχθεί ο εντοπισμός τους ενώ οι αγορές γίνονται με το ψηφιακό νόμισμα bitcoin ώστε να μην είναι δυνατός ο εντοπισμός των συναλλαγών. Μοναδική αδυναμία του συστήματος είναι ότι η αποστολή του προϊόντος θα πρέπει να γίνει στην πραγματική διεύθυνση του φυσικού προσώπου. Ακόμη και γι αυτό έχουν εφευρεθεί διαφορετικοί τρόποι συσκευασίας και πακεταρίσματος ενώ η μεταφορά πολλές φορές γίνεται με εταιρίες ταχυμεταφοράς.

Η μαύρη αγορά στο διαδίκτυο βασιλεύει και στην ελληνική επικράτεια ενώ διάφοροι hackers κατά καιρούς γίνονται στόχοι των ελληνικών αρχών. Στη μαύρη αγορά του Ιντερνετ διακινούνται καθημερινά χιλιάδες λίστες που περιέχουν ευαίσθητα προσωπικά δεδομένα εκατοντάδων χιλιάδων Ελλήνων πολιτών. Προσωπικά e-mails, διεύθυνση κατοικίας, προσωπικοί αριθμοί κινητών και σταθερών τηλεφώνων, ακόμα και ΑΦΜ, καταλήγουν στα χέρια επικίνδυνων κακοποιών (NewsIt).

Όπως υποστήριξε γνωστός hacker μετά τη σύλληψή του, οι λίστες για την Ελλάδα είναι τρεις. Η πρώτη περιλαμβάνει περίπου 50.000 επιχειρηματικά e-mails και τα ακόλουθα στοιχεία: ονοματεπώνυμο, διεύθυνση, σταθερό τηλέφωνο, φαξ και αριθμό κινητού τηλεφώνου. Η δεύτερη λίστα περιέχει περίπου 150.000 e-mails και η τρίτη προσφέρει περίπου 290.000 e-mails. Η πληρωμή γίνεται μέσω PayPal (NewsIt).

Τα παραπάνω στοιχεία, πωλούνται συνήθως σε εταιρίες προώθηση προϊόντων ώστε να υπάρχει συγκεκριμένη στόχευση σε καταναλωτές. Ανάλογες λίστες υπάρχουν για τις περισσότερες ευρωπαϊκές χώρες, τις ΗΠΑ, τον Καναδά, την Αυστραλία, τη Νότια Αμερική, τη Μέση Ανατολή, την Ανατολική Ασία κτλ. Ο γνωστός hacker υποστήριξε πως οι λίστες αυτές ενημερώνονται τακτικά ενώ σε αυτές δεν υπάρχουν διπλότυπα (NewsIt).

"Τα emails αποτελούν προσωπικό δεδομένο, είναι έμμεσος τρόπος αναγνώρισης φυσικού προσώπου. Αποτελούν "Κερκόπορτα" για να διεισδύσουν οι χάκερ στους υπολογιστές μιας εταιρείας ή ενός προσώπου. Απαξ και το πετύχουν, θα μπορούν να βρουν όλα τα στοιχεία μας, κωδικούς σε τράπεζες , μέσα κοινωνικής δικτύωσης, συναλλαγές, φορολογικά στοιχεία και στη συνέχεια να αδειάσουν τους τραπεζικούς

λογαριασμούς ή να τα χρησιμοποιήσουν σε άλλες παράνομες δραστηριότητες...Στις μέρες μας το email μας είναι συνδεδεμένο με τον υπολογιστή και το κινητό μας. Η παραβίασή του οδηγεί σε μεγάλο κίνδυνο" εξηγούν αξιωματικοί της Δίωξης ηλεκτρονικού Εγκλήματος (NewsIt).

**Τιμοκατάλογος**

| | |
|---|---|
| **200.000 επιχειρηματικά e-mails** | **40 ευρώ** |
| **350.000 προσωπικά e-mails** | 50 ευρώ |
| **Επώνυμο, διεύθυνση, τηλέφωνο** | 70 ευρώ |
| **Όλα τα παραπάνω μαζί** | 80 ευρώ |
| **Δεδομένα 22 χωρών από 350 ευρώ** | 300 ευ |

# Κεφάλαιο 6 – Αντιμετώπιση

## 6. Τρόποι αντιμετώπισης

Η εκτεταμένη χρήση του διαδικτύου έχει κάνει ιδιαίτερα ευάλωτα σε κλοπή τα προσωπικά δεδομένα των χρηστών καθώς ο διαμοιρασμός αυτών έχει αυξήσει σημαντικά την πιθανότητα για διαρροή τους και έκθεσή τους σε κίνδυνο. Τα δεδομένα πλέον, μεταφέρονται εύκολα, διακινούνται εύκολα και αντιγράφονται με τους ίδιους έντονους ρυθμούς.

Η επικράτηση των διαφόρων μικρών κινητών συσκευών και η μεταφορά των δεδομένων σε όλες αυτές τις συσκευές είναι μία ιδιαίτερα ευπαθής διαδικασία. Καθώς τα δεδομένα διαμοιράζονται, έχουν πλέον σε αυτά πρόσβαση όλες οι συσκευές και αυτό συντελεί άμεσα στην απώλεια της ασφάλειας. Δυστυχώς, ακόμη και σήμερα δεν υπάρχει επαρκής ενημέρωση του χρήστη για τους διάφορους κινδύνους που συντελούνται από την απώλεια των δεδομένων καθώς και πόσο εύκολο είναι να απολεσθούν τα δεδομένα όταν αυτά υπάρχουν σε διάφορες συσκευές ή εάν δεν λαμβάνονται τα ανάλογα μέτρα προστασίας.

Έτσι λοιπόν οι χρήστες, θα πρέπει στο σύνολό τους να είναι ενημερωμένοι και να αυξήσουν τα επίπεδα προστασίας των δεδομένων τους όπου αυτό κρίνεται αναγκαίο. Έτσι λοιπόν ο κάθε χρήστης θα πρέπει να λαμβάνει υπόψη του τα εξής (Finklea, 2015) :

- Που βρίσκονται τα ευαίσθητα προσωπικά του δεδομένα και σε ποιον έχει εξασφαλίσει την πρόσβαση σε αυτά
- Ποιοι κανονισμοί και πρότυπα αρμόζουν στα δεδομένα του
- Εάν έχει υπάρξει στο παρελθόν θύμα κλοπής δεδομένων
- Εάν κατά τη χρήση του διαδικτύου θέτει άμεσα τα δεδομένα του σε κίνδυνο
- Εάν τα μέτρα που έχει λάβει για την προστασία των δεδομένων του τα έχει λάβει για όλες τις συσκευές που χρησιμοποιεί

Μία σωστή στρατηγική ώστε να αντιμετωπίσει ο χρήστης τους διάφορους κινδύνους που προέρχονται από το διαδίκτυο είναι :

- Να ταξινομεί τα δεδομένα του ανάλογα με το πόσο σημαντικά είναι
- Να γνωρίζει που βρίσκονται αποθηκευμένα αυτά αλλά και που μεταφέρονται
- Να κατανοήσει τις διάφορες απειλές που δέχονται
- Να ενημερώνει συχνά τα λογισμικά προστασίας ιών που έχει εγκατεστημένα στους υπολογιστές του

Η κατάλληλη στρατηγική προστασίας των προσωπικών δεδομένων είναι μία ιδιαίτερα σημαντική διαδικασία και είναι καλό να εφαρμόζεται από όλους τους χρήστες ώστε να μην υποκλαπούν τα προσωπικά τους δεδομένα και να μην αποτελέσουν αντικείμενο χρηματικών συναλλαγών στο dark web. Σε περίπτωση που κάποιος διαπιστώσει οποιαδήποτε κλοπή των στοιχειών του θα πρέπει να επικοινωνήσει άμεσα με την Αρχή Προστασίας Δεδομένων αλλά και με τη Διώξη Ηλεκτρονικού Εγκλήματος ώστε να επιληφθούν οι ανάλογες αρχές του θέματος (ΑΠΠΔ).

Ως λύση στις διάφορες απειλές που δέχεται ο μέσος χρήστης έχει προταθεί το σύστημα IP Viking το οποίο έχει ως στόχο την πρόληψη και την αποφυγή επιθέσεων ανιχνεύοντας όσες από αυτές συντελούνται τόσο σε τοπικό όσο και σε παγκόσμιο επίπεδο. Πρόκειται για έναν online χάρτη σε πραγματικό χρόνο που ενημερώνει για το ποιοι πραγματοποιούν επιθέσεις, ποιοι τις δέχονται και τι είδους επιθέσεις συμβαίνουν (IP Viking).

Το σύστημα λειτουργεί με τη χρήση ευφυών πρακτόρων που οποίοι ελέγχουν καθημερινά τεράστιο όγκο πληροφορίας ο οποίος σχετίζεται άμεσα με τα όσα συντελούνται στο διαδίκτυο και με τις διάφορες κυβερνοεπιθέσεις. Η εταιρία που παρήγαγε το σύστημα αυτό, η Norse, κάνει χρήση διαφόρων honeypots ώστε να εντοπίσει και να εξαπατήσει διάφορα κακόβουλα εργαλεία και να εντοπίσεις IP διευθύνσεις από τις οποίες συντελούνται κακόβουλες ενέργειες (IP Viking).

Η μορφή των επιθέσεων που συντελούνται στο dark net αλλάζει συνεχώς κα αποτελεί αδιαμφισβήτητα ένα τεράστιο πρόβλημα που λαμβάνει το μορφή πολέμου που έχει μεταβεί όμως στην ηλεκτρονική του μορφή. Οι ανυποψίαστοι χρήστες δυστυχώς

πολλές φορές καταχωρούν οικειοθελώς τα στοιχεία τους και γίνονται θύματα υποκλοπών εν αγνοία τους. Όπως έχει αναφέρει και ο διευθυντής marketing της Norse «Υπάρχει ένα σημαντικό ποσοστό των χρηστών του διαδικτύου σήμερα που αποτελούν μέρος ενός botnet και δεν το ξέρουν. Και σκοπεύουμε να βοηθήσουμε να μειώσουμε κατά πολύ αυτόν τον αριθμό.» (IP Viking).

Στην ιστοσελίδα της εταιρίας που έχει δημιουργήσει το λογισμικό αυτό μπορεί κανείς να δει το χάρτη με τις επιθέσεις που συντελούνται σε πραγματικό χρόνο. Ο χρήστης είναι σε θέση διαμέσου της ιστοσελίδας να δει τις χώρες που πραγματοποιούν επιθέσεις, τις χώρες που δέχονται επιθέσεις, τη μορφή των επιθέσεων αλλά και ένα συγκεντρωτικό πίνακα με όλες τις παραπάνω πληροφορίες.

# Βιβλιογραφία


Adomavicius, G., & Tuzhilin, A. (n.d.). *Extending recommender systems: A multidimensional approach.*

Amores, R., & Motta, G. (2012). *The Deep Dark Web: the hidden world.* Paganini - Amores.

Anderson, C. (2002). A machine learning approach to web personalization. *Ph.D. thesis*. University of Washington.

*Arbornetworks*. (n.d.). Ανάκτηση από https://www.arbornetworks.com/

Bergman, M. (2001). *The Deep Web: Surfacing Hidden Value.* Ανάκτηση από <http://www.press.umich.edu/jep/07-01/bergman.html>

*Berkeley*. (2001). Ανάκτηση από http://www.berkeley.edu/news/archives/2001.html

Bounsaythip, C., & Rinta-Runsala, E. (2001). Overview of data mining for customer behavior modeling. *Technical Report TTE1-2001-18*. VTT Information Technology.

Chacos, B. (2013). *Meet Darknet, the hidden, anonymous underbelly of the searchable Web*. Ανάκτηση από http://www.pcworld.com/article/2046227/meet-darknet-the-hidden-anonymous-underbelly-of-the-searchable-web.html

Diebold, B., & Kaufmann, M. (2001). Usage-based visualization of web localities. *Australian symposium on information visualisation*, 159-164.

Dingledine, R., Mathewson, N., & Syverson, P. (n.d.). *Tor: The Second-Generation Onion Router.* Ανάκτηση από https://svn.torproject.org/svn/projects/design-paper/tor-design.pdf

Finklea, C. (2015). *Dark Web.* Ανάκτηση από https://fas.org/sgp/crs/misc/R44101.pdf

Gareth, O., & Nick, S. (2015). *The Tor Dark Net.* Ανάκτηση από https://www.ourinternet.org/sites/default/files/publications/no20_0.pdf

Hooks, M., & Miles, J. (2006). *Onion Routing and Online Anonymity.* Ανάκτηση από http://static.cs.brown.edu/people/jadrian/docs/papers/old/20060430%20Hooks%20and%20Miles%20-%20Onion%20Routing.pdf

*IP Viking*. (n.d.). Ανάκτηση από http://hp.ipviking.com/

Kamdar, T. (2002). Creating adaptive web servers using incremental web log mining. *Master Thesis*. Baltimore County: Computer Science Department, University of Maryland.

LLC, B. (2000). *Deep Web Surfacing Hidden Value.* Ανάκτηση από http://resources.mpi-inf.mpg.de/d5/teaching/ws01_02/proseminarliteratur/deepwebwhitepaper.pdf

NATO. (2015). *Technical and Legal Overview of the Tor Anonymity Network.* Ανάκτηση από https://ccdcoe.org/sites/default/files/multimedia/pdf/TOR_Anonymity_Network.pdf

*NewsIt*. (n.d.). Ανάκτηση από http://www.newsit.gr/egklhma/S-O-S-Sti-mayri-agora-toy-internet-ta-prosopika-dedomenon-olon-Poloyntai-synomilies-meso-email-AFM-arithmoi-kiniton-Ti-prepei-na-gnorizete/626096

Ntoulas, A., Cho, J., & Olston, C. (2004). What's New on the Web? The Evolution of the Web from a Search Engine Perspective. *Proceedings of the 13th international conference on World Wide Web*, 1-12.

Olston, C., & Najork, M. (2010). *Web Crawling.* Ανάκτηση από http://infolab.stanford.edu/~olston/publications/crawling_survey.pdf

*Privoxy*. (n.d.). Ανάκτηση από https://www.privoxy.org/

Reed, M., Syverson, P., & Goldschlag, D. (n.d.). *Anonymous Connections and Onion Routing.* Ανάκτηση από https://www.onion-router.net/Publications/JSAC-1998.pdf

Sidiropoulou, M., & Pohle, C. (2001). Data mining for measuring and improving the success of web sites. σσ. 85-114.

Srivastava, J., & Deshpande, M. (2000). Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explorations, 2*(1), σσ. 12-23.

Stumme, G., Hotho, A., & Berendt, B. (2002). Usage mining for and on the semantic web. *National Science Foundation Workshop on Next Generation Data Mining*.

Sui, D., Caverlee, J., & Rudesill, D. (n.d.). *THE DEEP WEB AND THE DARKNET: A LOOK INSIDE THE INTERNET'S MASSIVE BLACK BOX.* Ανάκτηση από https://www.wilsoncenter.org/sites/default/files/stip_dark_web.pdf

*Surface Web, Deep Web, Dark Web -- What's the Difference?* (2016). Ανάκτηση από https://www.cambiaresearch.com/articles/85/surface-web-deep-web-dark-web----whats-the-difference

*TrendMicro*. (2016). Ανάκτηση από https://www.trendmicro.com/vinfo/us/security/research-and-analysis

*UbuntuGeek*. (n.d.). Ανάκτηση από http://www.ubuntugeek.com/howto-install-torprivoxy-and-tor-gui-programs-vidaliatork-and-torbuttonin-ubuntu.html

*University of Virginia.* (n.d.). Ανάκτηση 2017, από www.cs.virginia.edu: http://www.cs.virginia.edu/~cs458/slides/module03-ipaddrV3.pdf

Zilman, M. (2017). *Deep Web Research and Discovery Resources 2017.* Ανάκτηση από http://whitepapers.virtualprivatelibrary.net/DeepWeb.pdf

ΑΠΠΔ. (n.d.). Προσωπικά Δεδομένα και Διαδίκτυο.

*Αριστοτέλειο Πανεπιστήμιο .* (n.d.). Ανάκτηση από pacific.jour.auth.gr: http://pacific.jour.auth.gr/totsidou/The_Internet.htm

Βελούδης, Χ. (n.d.). *ΤΕΙ Σερρων.* Ανάκτηση 2016, από ftp://teiser.gr/pliroforiki/Diktia%20II/Veloudis/02.%20Communication%20Protocols%20-%20IP%20Protocol.pdf

*ΚΕΠΛΗΝΕΤ Φλώρινας*. (n.d.). Ανάκτηση 2017, από http://dide.flo.sch.gr/Plinet/Tutorials/Tutorials-TCP-IP-Introduction.html

*Πανεπιστήμιο Θεσσαλίας*. (1997). Ανάκτηση 2016, από http://www.uth.gr/main/help/help-desk/internet/internet3.html

Σιδηρόπουλος, Χ. (n.d.). *Το Deep Web.*

Τζάνης, Ά. (2013). DEEP WEB DYNAMICS: ΕΝΤΟΠΙΣΜΟΣ, ΠΡΟΣΒΑΣΗ ΚΑΙ ΚΑΤΑΓΡΑΦΗ ΕΞΕΛΙΚΤΙΚΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ " ΑΟΡΑΤΟ ΙΣΤΟ " ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗ ΣΤΗ PubMed. *Μεταπτυχιακή Διατριβή*. Πανεπτιστήμιο Θεσσαλίας.