Τίτλος Έργου	Προηγμένη ανίχνευση εισβολών και παράνομου ηλεκτρονικού περιεχομένου για τη διασφάλιση επιχειρησιακής και λειτουργικής ασφάλειας σε κατανεμημένα συστήματα συστημάτων		
Ακρωνύμιο Έργου	ΜΥΡΙΩΠΟΣ		
Κωδικός Έργου	Τ2ΕΔΚ-04665		
Πρόσκληση / Θέμα	ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ - ΚΑΙΝΟΤΟΜΩ Β' ΚΥΚΛΟΣ / ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ, ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ & ΚΑΙΝΟΤΟΜΙΑ		
Ημερομηνία Εκκίνησης	16/06/2020	Διάρκεια	30 Μήνες

Π2.1 Συλλογή αποδείξεων και κλιμακούμενη διαδικτυακή ανίχνευση κυβερνοκινδύνων

Ενότητα Εργασίας	**ΕΕ2 Κατανεμημένη Ευφυΐα Ιστού και Σκοτεινού Δικτύου**
Επικεφαλής	**MAGGIOLI**
Συνεισφορές	**UBITECH**

Περιεχόμενα

Συντομογραφίες

API	Application Programming Interfaces
BOM	Byte-Order Mark
BMP	Bitmap Image File
CAPTCHA	Completely Automated Public Turing Test to Tell Computers and Humans Apart
DNS	Domain Name Server
DDoS	Distributed Denial of Service
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
I2P	Invisible Internet Project
IP	Internet Protocol
ISP	Internet Service Provider

IP address	Internet Protocol Address
JPG / JPEG	Joint Photographic Experts Group
JS	JavaScript
NIST	National Institute of Standards and Technology
MISP	Malware Information Sharing Platform
OpenCV	Open-Source Computer Vision Library
PNG	Portable Network Graphics
TCP	Transmission Control Protocol
TOR	The Onion Router
UDP	User Datagram Protocol
URL	Uniform Resource Locator
VPN	Virtual Private Network
WWW	World Wide Web

Λίστα Εικόνων

Περίληψη

Το έργο ΜΥΡΙΩΠΟΣ[1] έχει ως στόχο να υλοποιήσει μια ολοκληρωμένη λύση ασφάλειας, ικανή να περιορίσει τις κυβερνοεπιθέσεις, να προσδιορίσει και να κατηγοριοποιήσει εγκαίρως παραβατικές συμπεριφορές στο διαδίκτυο και στο σκοτεινό ιστό, μέσω της κατάλληλης μοντελοποίησης των διαφορετικών εννοιών του συστήματος, της παρακολούθησης περιουσιακών πόρων, της προηγμένης αξιολόγησης και επιβολής πολιτικών ασφάλειας σε πραγματικό χρόνο. Το συγκεκριμένο παραδοτέο με τίτλο «Π2.1 Συλλογή αποδείξεων και κλιμακούμενη διαδικτυακή ανίχνευση κυβερνοκινδύνων» είναι άμεσο αποτέλεσμα της δράσης «Δ2.1 Κλιμακούμενη διαδικτυακή ανίχνευση, πολλαπλά επίπεδα επεξεργασίας και συλλογή αποδείξεων». Το βασικό του αντικείμενο είναι να περιγράψει με λεπτομέρειες το υφιστάμενο εννοιολογικό και τεχνολογικό υπόβαθρο καθώς και το λογισμικό που αναπτύχθηκε αναφορικά με τις διεργασίες συλλογής αποδείξεων και τη κλιμακούμενη ανίχνευση κυβερνοκινδύνων από τον σκοτεινό δίκτυο/ιστό. Η υλοποίηση των προαναφερθέντων, πραγματοποιήθηκε με ανάπτυξη λειτουργικών στοιχείων λογισμικού ανίχνευσης περιεχομένου ιστού, καθαρισμού κειμένου από μη σχετικό περιεχόμενο και όρους που δε συνεισφέρουν στην κειμενική ανάλυση, εξόρυξης νέων όρων, γνώσης, κλιμακούμενων μεθόδων συλλογής δεδομένων, καθώς και εντοπισμού όλων των πιθανών κινδύνων που μπορούν να εντοπιστούν στο σκοτεινό δίκτυο και δημόσιο ιστό.

[1] ΜΥΡΙΩΠΟΣ: ὁ ἔχων ἀναριθμήτους ὀφθαλμούς, ἐπὶ τοῦ Ἄργου, Αἰσχύλ. Πρ. 569.

1. Εισαγωγικά Στοιχεία

Στο κεφάλαιο αυτό εισάγουμε το πεδίο εφαρμογής και τους στόχους του εν λόγω παραδοτέου. Επίσης, παρουσιάζουμε τη συσχέτισή του με τις δράσεις και τις ενότητες εργασίας του έργου ΜΥΡΙΩΠΟΣ και τέλος περιγράφουμε προς διευκόλυνση του αναγνώστη τη δομή που θα ακολουθηθεί κατά την αναφορά των επιμέρους στοιχείων.

1.1 Πεδίο Εφαρμογής και Στόχοι Παραδοτέου

Σε έναν ολοένα και πιο ψηφιακό κόσμο με θεωρητικά αναρίθμητες αλληλεπιδράσεις υπάρχει επιτακτική ανάγκη παρακολούθησης και εντοπισμού στοχευμένου περιεχομένου από το σκοτεινό δίκτυο και δημόσιο ιστό. Κυρίως στο σκοτεινό δίκτυο, το εν λόγω περιεχόμενο σχετίζεται με παράνομες δραστηριότητες και συγχρονισμό ομάδων για μαζικές κυβερνοεπιθέσεις. Ενδεικτικά, αναφέρουμε ως παράδειγμα την υποκλοπή δεδομένων εφαρμογών που πωλούνται στο σκοτεινό δίκτυο, κλεμμένα διαπιστευτήρια που πωλούνται σε σκοτεινές αγορές, παραβιασμένοι λογαριασμοί email, κ.λπ.. Ορισμένοι από τους χρήστες του σκοτεινού δικτύου εκμεταλλευόμενοι την ανωνυμία που προσφέρει το λογισμικό TOR [1], διαδικτυακοί τόποι που δραστηριοποιούνται στη παροχή παράνομου περιεχομένου στο σκοτεινό δίκτυο και το γεγονός ότι οι ιστοσελίδες δεν μπορούν να δεικτοδοτηθούν από συμβατικές μηχανές αναζήτησης βρίσκουν πρόσφορο έδαφος για παράνομες δραστηριότητες. Ειδικότερα χρησιμοποιούν αυτό το λογισμικό για να κρύψουν τις πραγματικές τους διευθύνσεις (IP) και τις φυσικές τους τοποθεσίες. Ταυτόχρονα, η χρήση του λογισμικού TOR καθιστά γνωστούς αλγόριθμους δεικτοδότησης όπως το PageRank [2] κ.λπ. αδύνατο να εκτελεστούν αποτελεσματικά επειδή οι διαδικτυακοί τόποι στο TOR αλλάζουν συνεχώς και δεν υιοθετούν ένα δομημένο Uniform Resource Locator (URL) μοτίβο, όπως οι ιστοσελίδες στο δημόσιο ιστό. Η ύπαρξη ενός συστήματος για την αρχειοθέτηση και παρακολούθηση περιεχομένου που είναι διαθέσιμο στο σκοτεινό δίκτυο, ως μέρος ενός αναπτυσσόμενου συνόλου εργαλείων στο ΜΥΡΙΩΠΟ, αντισταθμίζει το πλεονέκτημα της ανωνυμίας που προσφέρει το TOR και δρα σαν γραμμή άμυνας έναντι κακόβουλων ενεργειών στο σκοτεινό δίκτυο, ενισχύοντας την επίγνωση των οργανισμών και των τελικών χρηστών σχετικά με τα δεδομένα τους από εφαρμογές επαφών ή κοινωνικής δικτύωσης που χρησιμοποιούν[2, 3], τα διαπιστευτήρια τους και το ρίσκο που υπάρχει από τις εκτεθειμένες υπηρεσίες, τις εφαρμογές που χρησιμοποιούν ή/και τις Διεπαφές Προγραμματισμού Εφαρμογών (Application Programming Interfaces (API)) τους.

Στο παραδοτέο Π2.1 παρουσιάζουμε μια διεξοδική βιβλιογραφική επισκόπηση και αξιολόγηση των υπαρχουσών προσεγγίσεων, οι οποίες στοχεύουν στο να αναλύσουν και να εκμεταλλευτούν περιστατικά ασφάλειας και πληροφορίες που σχετίζονται με κινδύνους και απειλές, τα οποία είναι ενσωματωμένα σε περιεχόμενο που δημιουργείται από χρήστες και ανταλλάσσεται ή πωλείται στο σκοτεινό δίκτυο. Στη συνέχεια αναφέρουμε βασικές έννοιες για τη κατανόηση της δομής του σκοτεινού δικτύου και των προβλημάτων που χρειάζεται να αντιμετωπιστούν. Τέλος, εισάγουμε τις μεθόδους και τις τεχνολογίες που χρησιμοποιεί ο ΜΥΡΙΩΠΟΣ για την εξόρυξη γνώσης από το σκοτεινό δίκτυο.

1.2 Συσχέτιση με Δράσεις και Ενότητες Εργασίας του ΜΥΡΙΩΠΟΥ

Το παραδοτέο Π2.1 είναι άμεσο αποτέλεσμα και αποτελεί αναφορά της δράσης «Δ2.1 Κλιμακούμενη διαδικτυακή ανίχνευση, πολλαπλά επίπεδα επεξεργασίας και συλλογή αποδείξεων». Σχετίζεται επίσης με τη

[2] https://covve.com/
[3] https://businessplus.ie/tech/social-media-lost-user-data/

δράση «Δ1.2 Αρχιτεκτονική πλαισίου ΜΥΡΙΩΠΟΣ και ορισμός νέου προϊόντος» που όρισε την Αρχιτεκτονική Συστήματος αφού τη χρησιμοποιεί ως είσοδο και ως εκ τούτου ως βασικό πρότυπο σχεδιασμού της τεχνολογικής λύσης που αναπτύχθηκε. Ως παραδοτέο, τροφοδοτεί και δίνει είσοδο στις υπόλοιπες δράσεις της ΕΕ2 Κατανεμημένη Ευφυΐα Ιστού και Σκοτεινού Δικτύου και συγκεκριμένα των «Δ2.2 Αναλυτική και εξαγωγή γνώσης από τα δεδομένα ιστού και σκοτεινό δικτύου» και «Δ2.3 Πολυτροπική ειδοποίηση και διαλειτουργικότητα με τρίτα εργαλεία».

1.3 Δομή Παραδοτέου

Το παραδοτέο Π2.1 περιγράφει στο **Κεφάλαιο 1 – Εισαγωγή** το γενικό σκοπό του παραδοτέου. Τεκμηριώνει επίσης τη θέση του παραδοτέου στο έργο, δηλαδή τη σχέση του τρέχοντος παραδοτέου με τα άλλα παραδοτέα, δράσεις και ενοτήτων εργασίας, καθώς και πώς η γνώση που παράγεται στα άλλα παραδοτέα και ενότητες εργασίας χρησιμεύουν ως είσοδο στο συγκεκριμένο παραδοτέο.

Στη συνέχεια στο **Κεφάλαιο 2 – Επισκόπηση Βιβλιογραφίας** παρουσιάζουμε γενικές πληροφορίες αναφορικά με: α) ανιχνευτές σκοτεινού δικτύου και πολυνημάτωση για να μην μπλοκάρουν οι διεργασίες αναζήτησης, β) επικύρωση και εξαγωγή συνδέσμων από το σκοτεινό δίκτυο, γ) τα εμπόδια στην ανακάλυψη περιεχομένου, δ) τις διαφορετικές κωδικοποιήσεις που χρησιμοποιούνται στα κείμενα και πώς το διαχειριζόμαστε και ε) τα διαφορετικά πρωτόκολλα πρόσβασης που χρησιμοποιούνται.

Το **Κεφάλαιο 3 – Υπηρεσίες ΜΥΡΙΩΠΟΥ για Εξόρυξη Δεδομένων από το Σκοτεινό Δίκτυο και Δημόσιο Ιστό**, παρουσιάζει πως πραγματοποιείται στα πλαίσια του ΜΥΡΙΩΠΟΥ, η: α) ανάκτηση και η συλλογή δεδομένων και β) επιμέλεια, καθαρισμός, εμπλουτισμός και η εναρμόνιση των δεδομένων.

Τέλος, το παραδοτέο ολοκληρώνεται με το **Κεφάλαιο 4 – Επίλογος**, το οποίο περιγράφει τα κύρια συμπεράσματα του παραδοτέου και παρέχει ένα πλάνο που θα καθοδηγήσει τις μελλοντικές ερευνητικές και τεχνολογικές προσπάθειες της σύμπραξης.

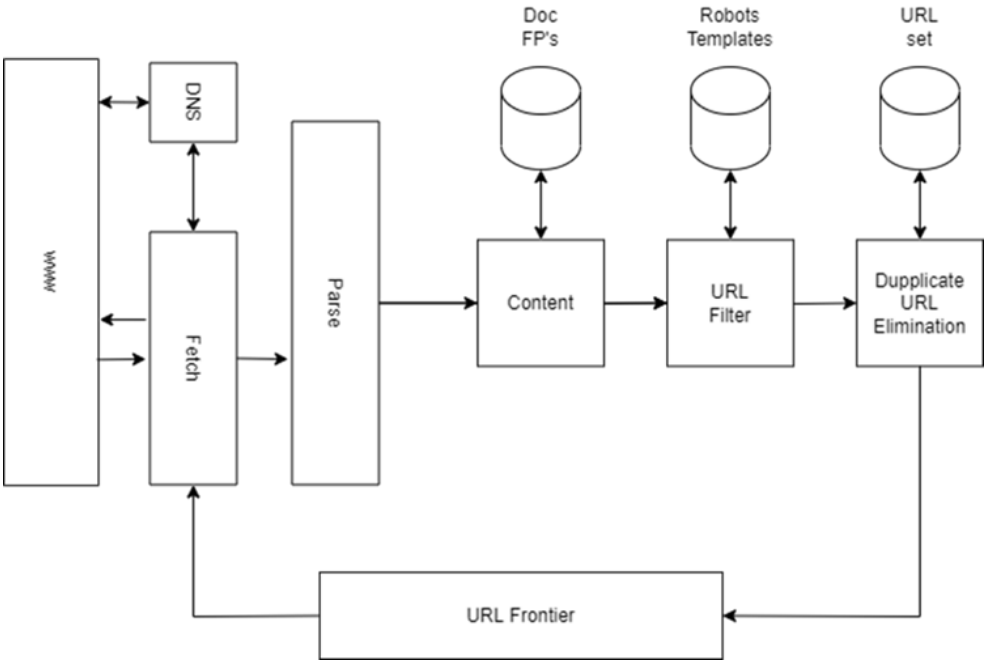
2. Επισκόπηση Βιβλιογραφίας σχετικά με Εξόρυξη Δεδομένων από το Σκοτεινό Δίκτυο και Δημόσιο Ιστό

Το διαδίκτυο αποτελείται από το δημόσιο ιστό και το σκοτεινό δίκτυο. Ο δημόσιος ιστός είναι μια συλλογή από ιστοτόπους, ευρετηριασμένους από μηχανές αναζητητής όπως η Google [3], Yahoo [4], και Bing [5]. Παράλληλα, είναι προσβάσιμος μέσω κλασσικών πρωτοκόλλων και προγραμμάτων περιήγησης, όπως το Google Chrome [6], Firefox [7], Opera [8] κλπ.. Παρόλα αυτά, αυτό αποτελεί μόνο την κορυφή του παγόβουνου συγκριτικά με το τι περιεχόμενο υπάρχει διαθέσιμο στον ιστό. Υπολογίζεται ότι το σκοτεινό δίκτυο είναι σχεδόν 4500 φορές μεγαλύτερο από το δημόσιο [9]. Για έναν μέσο χρήστη, είναι κοινή πεποίθηση ότι το σκοτεινό δίκτυο συγκεντρώνει μυστηριώδεις και παράνομες δραστηριότητες. Ωστόσο, αυτό είναι αληθινό κατά το ήμισυ, καθώς η πρόσβαση στο σκοτεινό δίκτυο γίνεται συνήθως από ιστοτόπους και εφαρμογές κοινωνικής δικτύωσης όπως το Facebook [10], Twitter [11], Instagram [12] ή Snapchat [13]. Η πεποίθηση αυτή δημιουργήθηκε από ορισμένες δραστηριότητες που λαμβάνουν χώρα στο σκοτεινό δίκτυο, οι οποίες περιλαμβάνουν την πώληση δεδομένων που έχουν κλαπεί, οργανωμένα εγκλήματα όπως μαζικές και συγχρονισμένες κλήσεις κοινοτήτων χρηστών για μαζικές και κατανεμημένες επιθέσεις DDoS (Distributed Denial of Service, βλέπε Π1.1), κλπ.. Το περιεχόμενο του σκοτεινού δικτύου δεν είναι ευρετηριασμένο από μηχανές αναζητητής όπως η Google, καθώς το περιεχόμενο σε αυτό το μέρος του διαδικτύου δεν είναι εύκολα προσβάσιμο, και τα URLs δεν ακολουθούν ένα δομημένο Uniform Resource Locator (URL) μοτίβο, όπως οι ιστότοποι στο δημόσιο ιστό. Συγκεκριμένα, η πρόσβαση στο σκοτεινό δίκτυο πραγματοποιείται μέσω του λογισμικού TOR [1], που βασίζεται στη δρομολόγηση μέσω κρυφών δρομολογητών (Onion Routing) και μέσω των Invisible Internet Project (I2P) διακομιστών που λειτουργούν ως ενδιάμεσοι διαμεσολαβητές (proxies) [14]. Το TOR [1] είναι ταυτόχρονα και λογισμικό και δίκτυο, που βοηθά στη διατήρηση της ανωνυμίας στο διαδίκτυο. Το I2P είναι ένα αποκεντρωμένο δίκτυο ανωνυμοποίησης παρόμοιο με το TOR. Το σκοτεινό δίκτυο έχει ορισμένες ιδιότητες, όπως είναι η αποφυγή των συμβατικών μηχανών αναζήτησης και ο τρόπος πρόσβασης σε αυτές, καθώς και η χρήση μη δομημένων URLs. Το σκοτεινό δίκτυο χρησιμοποιείται για νόμιμες δραστηριότητες που απαιτούν την ανωνυμία και περιλαμβάνει ιστοτόπους που δεν μπορούν να δεικτοδοτηθούν / ευρετηριαστούν. Το σκοτεινό δίκτυο είναι αρκετά μικρό και δυσπρόσιτο υποσύνολο του δημόσιου ιστού και πληροί συγκεκριμένες προϋποθέσεις. Σύμφωνα με την εργασία των Owen και Savage [15], υπολογίζεται ότι υπάρχουν περίπου 45000 ιστότοποι στο σκοτεινό δίκτυο και ο μόνος τρόπος πρόσβασης σε αυτούς είναι μέσω ειδικών προγραμμάτων περιήγησης ή/και διακομιστών μεσολάβησης που κάνουν χρήση του δικτύου TOR [1]. *Συνεπώς, ένα εργαλείο εξόρυξης δεδομένων από το σκοτεινό δίκτυο παραμετροποιημένο ώστε να συλλέγει στοχευμένο περιεχόμενο σχετικό με δραστηριότητες κυβερνοασφάλειας και όχι μη σχετικές δραστηριότητες (βλέπε εμπορία παράνομων υλικών, συλλογές ακατάλληλου περιεχομένου, κ.λπ.) απαιτεί την ανάπτυξη λειτουργιών λογισμικού για τη συλλογή, επιμέλεια, εναρμόνιση, ανάλυση και εντοπισμό εγκληματικών ενεργειών.*

Η ιδέα του *ανιχνευτή ιστού* δεν είναι καινούργια, έχουν γίνει ποίκιλες υλοποιήσεις και όλες μοιράζονται τον ίδιο στόχο: να επισκεφτούν ιστοτόπους, να εξάγουν χρήσιμη πληροφορία και να τους ευρετηριάσουν, δημιουργώντας αντίγραφο του κειμένου όλων αυτών. Η βασική λειτουργία τους έγκειται στην ευρετηρίαση του περιεχομένου των ιστοτόπων ώστε να διευκολύνουν τις μηχανές αναζήτησης στην αρχειοθέτηση αυτού για μεταγενέστερη επεξεργασία. Η διαδικασία αρχειοθέτησης και περεταίρω ανάλυσης καλείται εξόρυξη.

Βέβαια, η τεχνολογική εξέλιξη και τα νέα εργαλεία που βασίζονται στο διαδίκτυο έχουν καταστήσει τη δημιουργία αποτελεσματικών ανιχνευτών ιστού μια δύσκολη και απαιτητική διαδικασία. Σίγουρα, εξαιτίας

της μεγάλης ποικιλίας των τεχνολογιών που χρησιμοποιούνται για την υλοποίηση των ιστοσελίδων, η κατασκευή του τέλειου ανιχνευτή ιστού δεν είναι μια εύκολη υπόθεση [16]. Εντοπίζονται πολλά τεχνολογικά εμπόδια που πρέπει να ξεπεραστούν, όπως η ικανότητα ανίχνευσης ιστού και εξόρυξης, αλλάζοντας δυναμικά διεπαφές για να μην μπλοκάρει η αναζήτηση, IPs και θύρες υιοθετώντας νέες τεχνολογίες όπως HTML5[4], AJAX[5], και JavaScript (JS)[6].



Εικόνα 1. Βασική Αρχιτεκτονική Ανιχνευτή[7]

Στην Εικόνα 1 παρουσιάζεται η αρχιτεκτονική ενός κλασσικού ανιχνευτή ιστού. Συγκεκριμένα, η μονάδα «URL Frontier», περιέχει τα URLs που πρέπει να επισκεφθεί (να κάνει Fetch) ο ανιχνευτής. Το URL frontier διατηρεί ένα σύνολο από υψηλού επιπέδου επισκέψιμων URLs προκειμένου να μπορεί να ξεκινήσει η διαδικασία ανίχνευσης. Το DNS χρησιμοποιείται για τον προσδιορισμό της IP διεύθυνσης, του διακομιστή ιστού του URL που δόθηκε, μαζί με μια μονάδα «Fetch» για την ανάκτηση της ιστοσελίδας χρησιμοποιώντας το HTTP πρωτόκολλο. Η εξαγωγή του κειμένου και του συνόλου των συνδέσμων που περιέχονται στην ληφθείσα ιστοσελίδα, επιτυγχάνεται με τη μονάδα «Parse». Τέλος, η μονάδα «Duplicate URL elimination» είναι υπεύθυνη για την αποφυγή λήψης και συνεπώς επίσκεψης ιστοσελίδων, που έχουν ήδη ληφθεί από το «URL Frontier».

Όπως αναφέρθηκε, ένα υποσύνολο του σκοτεινού δικτύου [17] δεν είναι εύκολα προσβάσιμο χρησιμοποιώντας ένα κλασσικό πρόγραμμα περιήγησης. Αντίθετα, ο χρήστης που θέλει να έχει πρόσβαση σε

[4] https://developer.mozilla.org/en-US/docs/Glossary/HTML5
[5] https://en.wikipedia.org/wiki/Ajax_(programming)
[6] https://www.javascript.com/
[7] https://nlp.stanford.edu/IR-book/pdf/20crawl.pdf

αυτό, θα πρέπει να χρησιμοποιήσει το δίκτυο TOR χρησιμοποιώντας ένα εξειδικευμένο πρόγραμμα περιήγησης. Οι τοποθεσίες στο σκοτεινό δίκτυο κρύβονται πίσω από πολλά πρωτόκολλα κρυπτογράφησης, όπως το TOR ή το I2P και ο τομέας τους (domain) λήγει σε «.onion». Τέτοιοι ιστότοποι συνήθως δεν μπορούν να ανακτηθούν από έναν κανονικό ανιχνευτή ιστού και ο μόνος τρόπος είναι να χρησιμοποιηθούν εξωτερικοί διακομιστές μεσολάβησης HTTP ειδικά ρυθμισμένοι για το δίκτυο TOR [18].

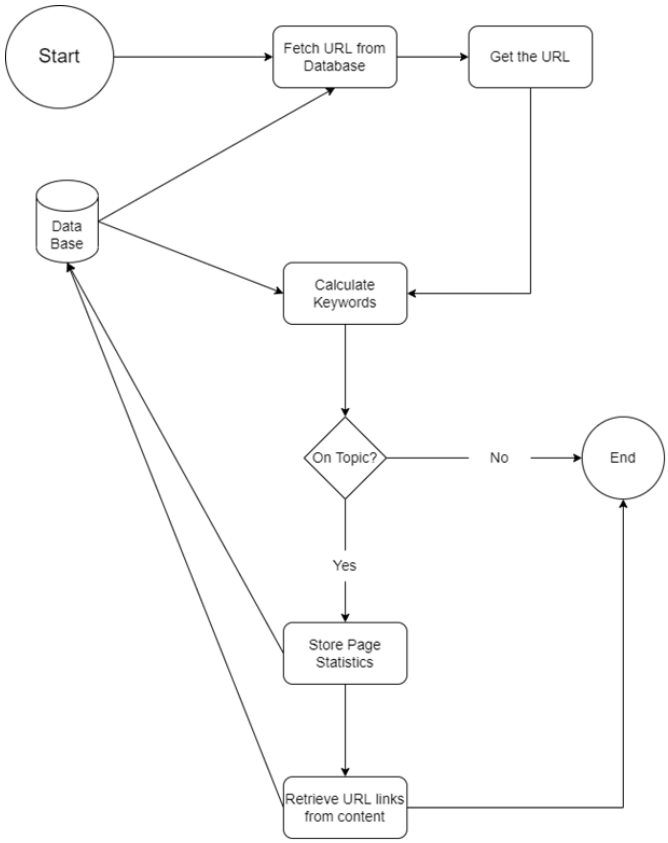
Μια πρόκληση για την ανίχνευση στοχευμένου περιεχομένου στο σκοτεινό δίκτυο είναι η αποδοτικότητα και η χρηστικότητα των κειμένων που ανακαλύπτει κανείς. Αυτό αφορά στο πλήθος και στην ποιότητα των κειμένων που μπορεί ο ανιχνευτής ιστού να επισκεφθεί και να ανακτήσει, εξαιτίας της δομής του σκοτεινού δικτύου. Το πρόβλημα αυτό, αντιμετωπίζεται χρησιμοποιώντας τη μέθοδο που εφαρμόζει δειγματοληψία και χιονοστιβάδα, για τον εντοπισμό κρυφών υπηρεσιών και τη συλλογή δεδομένων. Η μέθοδος αυτή λειτουργεί επαναληπτικά, συλλέγοντας όλους τους εξωτερικούς συνδέσμους προς άλλους ιστοτόπους, από ένα τομέα (domain) URL με προκαθορισμένο βάθος. Το αποτέλεσμα είναι μια εκτενής λίστα από URLs του σκοτεινού δικτύου. Επιπλέον, αυτή η μέθοδος επιτυγχάνει σημαντικά αποτελέσματα όταν εφαρμόζεται σε κοινότητες (forum) του σκοτεινού δικτύου, καθώς σύνδεσμοι προς άλλες πιο σχετικές κοινότητες μπορούν γρήγορα να ανακαλυφθούν αφού οδηγούν σε συναφές περιεχόμενο. Συνοψίζοντας, ένας αποτελεσματικός ανιχνευτής για το σκοτεινό δίκτυο πρέπει να υλοποιεί http διακομιστές μεσολάβησης (proxies), όπως και την μέθοδο δειγματοληψίας [19] για να εξασφαλίσει πιο αποτελεσματική και εστιασμένη αναζήτηση.

Οι επόμενες ενότητες παρουσιάζουν μια βιβλιογραφική επισκόπηση των ανιχνευτών σκοτεινού δικτύου και της πολυνημάτωσης, της επικύρωσης και της εξαγωγής έγκυρων συνδέσμων, των εμπόδιων που πρέπει να παρακαμφθούν για την πρόσβαση σε περιεχόμενου του σκοτεινού δικτύου καθώς και μεθόδους αποφυγής αυτών, όπως και η παραμετροποίηση του κώδικα για τη διαχείριση διαφορετικών κωδικοποιήσεων και πρωτοκόλλων.

2.1 Ανιχνευτές Σκοτεινού Δικτύου και Πολυνημάτωση

Γενικά, υπάρχουν αρκετές υλοποιήσεις ανιχνευτών για το σκοτεινού δίκτυο. Ένας ικανός ανιχνευτής πρέπει να είναι ανεξάρτητος από τη πλατφόρμα, γραμμικά επεκτάσιμος και αποκεντρωμένος. Όπως ήδη έχει αναφερθεί, οι κλασσικοί ανιχνευτές ιστού, δεν είναι ιδανικοί υποψήφιοι για το σκοτεινό δίκτυο [17], εξαιτίας της ανωνυμίας του δικτύου TOR. Ένας ανιχνευτής σκοτεινού δικτύου πρέπει να ξεπεράσει τους περιορισμούς που θέτει το δίκτυο TOR. Επιπλέον, πρέπει να διαθέτει ικανότητες μηχανικής μάθησης για να βοηθάει τις υπηρεσίες εντοπισμού περιεχομένου να ανακαλύπτουν στοχευμένα στοιχεία και πληροφορίες που σχετίζονται πραγματικά με δραστηριότητες που αφορούν στην κυβερνοασφάλεια. Αυτό απαιτείται γιατί το σκοτεινό δίκτυο έχει υψηλό ποσοστό ανάκλησης πληροφορίας («recall») για αρκετές δραστηριότητες δίχως αυτό να είναι ακριβές λόγω υψηλής εκφραστικότητας και αμφισημίας. Προκειμένου να συλλέγει σχετικό περιεχόμενο για δραστηριότητες κυβερνοασφάλειας (π.χ. παραβίαση δεδομένων εφαρμογών, κλεμμένα διαπιστευτήρια, παραβιασμένοι λογαριασμοί email, μαζικές κλήσεις για επιθέσεις DDoS) , χρησιμοποιούνται οντολογίες πληροφορίας απειλών, όπως η πλατφόρμα κοινοποίησης απειλών MISP Threat Sharing [20]. Αυτό είναι σε συμμόρφωση με τις οδηγίες του Ινστιτούτου NIST [36] που ορίζει πρότυπα τεχνολογίας. Δυστυχώς, δεν υπάρχει επαρκής αριθμός εργαλείων για την αναζήτηση στο σκοτεινό δίκτυο αναφορικά με κυβερνοεπιθέσεις. Ακόμα, ένας σημαντικός αριθμός από ιστοτόπους στο σκοτεινό δίκτυο παραμένουν κρυμμένοι σκόπιμα, καθώς μπλοκάρουν τους ανιχνευτές και δεν περιλαμβάνονται σε ιστοτόπους όπως το

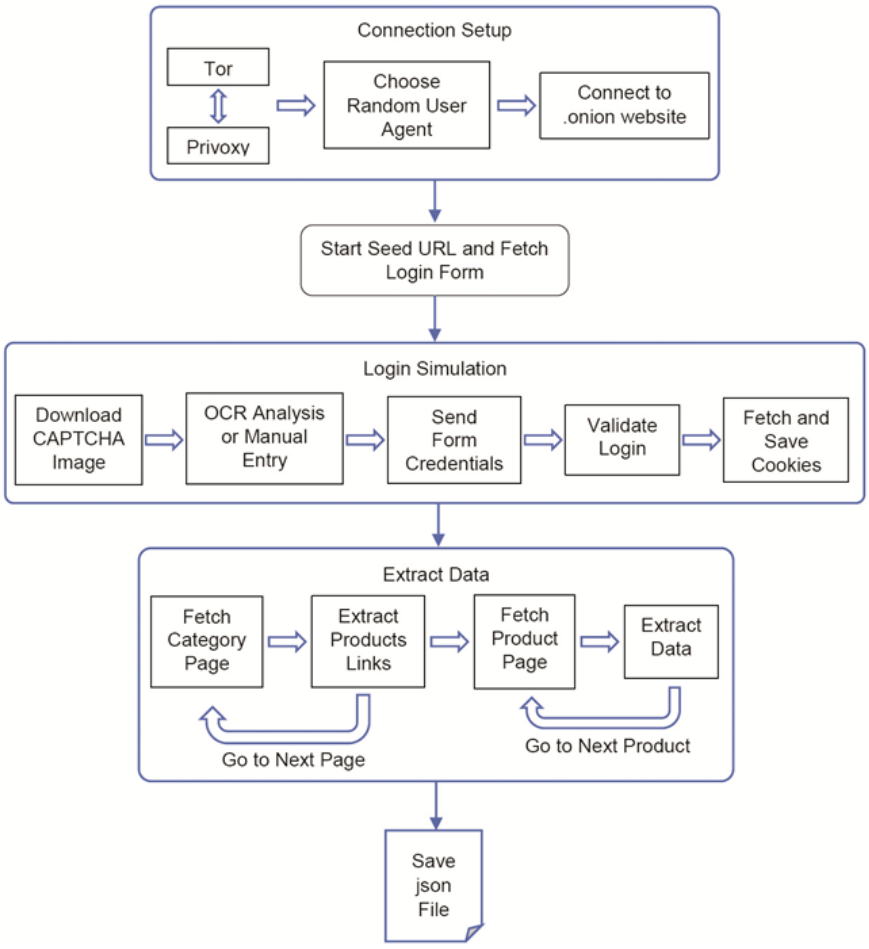
«hidden wiki»[8]. Λαμβάνοντας υπόψη αυτές τις συνθήκες και συνδυάζοντας το γεγονός ότι αυτοί οι ιστότοποι αλλάζουν πολύ συχνά URLs για να μη γίνονται προσβάσιμοι, η διαδικασία ανίχνευσης γίνεται αρκετά δύσκολη. Στη συνέχεια, ακολουθούν ορισμένες υλοποιήσεις ανιχνευτών.



Εικόνα 2. Διάγραμμα Ροής Εργασιών σε ένα Νήμα του Ανιχνευτή Σκοτεινού Δικτύου

Οι Zulkarnine κ.ά. [18] έχουν προτείνει έναν ανιχνευτή, τον «Σκοτεινό Ανιχνευτή», ο οποίος χρησιμοποίει το «Privoxy»[9] για να συνδεθεί στο δίκτυο TOR, ενώ η αναζήτηση πραγματοποιείται σύμφωνα με μια προκαθορισμένη λίστα από λέξεις-κλειδιά. Οι συγγραφείς, επισημαίνουν την ανάγκη για έναν δυναμικό ανιχνευτή, που θα είναι ικανός να επιταχύνει τη διαδικασία συλλογής, αποθηκεύοντας αυτόματα υψηλής αξίας περιεχόμενο. Η Εικόνα 2 αποτυπώνει τη συνήθη ροή εργασιών ενός ανιχνευτή σκοτεινού ιστού. Αυτή περιλαμβάνει την παραμετροποίηση του ανιχνευτή ώστε ανακτώντας ιστοσελίδες του σκοτεινού δικτύου από μια βάση δεδομένων να χρησιμοποιούνται μόνο αυτές που τις κατάλληλες λέξεις-κλειδιά και είναι σχετικές με τη στοχευμένη αναζήτηση που θέλουμε να εξασφαλίσουμε.

[8] https://thehiddenwiki.org/
[9] https://www.privoxy.org/



Εικόνα 3. Αρχιτεκτονική Ανιχνευτή Σκοτεινού Ιστού [21]

Οι Alkhatib κ.ά. [21] έχουν προτείνει έναν ανιχνευτή, τον λεγόμενο «Darky», χρησιμοποιώντας τη βιβλιοθήκη Scrapy[10] της Python και το Privoxy. Πρότειναν επίσης τις βασικές απαιτήσεις που πρέπει να περιλαμβάνει κάθε ανιχνευτής σκοτεινού δικτύου. Ο προτεινόμενος ανιχνευτής ξεκινά την αναζήτηση από μια καθορισμένη λίστα ιστοτόπων με παράνομες δραστηριότητες. Υπάρχουν ωστόσο πολλά προβλήματα που πρέπει να ξεπεραστούν, όπως οι παγίδες για τους ανιχνευτές που τους οδηγούν σε ατέρμονες επαναλήψεις. Η βασική αρχιτεκτονική του προτεινόμενου συστήματος παρουσιάζεται στην Εικόνα 3.

Οι Pantelis κ.ά. [22] έχουν επίσης προτείνει έναν βελτιωμένο και στοχευμένο ανιχνευτή σκοτεινού δικτύου, ο οποίος θα παρουσιαστεί στις επόμενες ενότητες του παραδοτέου Π2.1 και στο παραδοτέο Π2.2 με περισσότερες λεπτομέρειες, ως κομμάτι της τεχνικής υλοποίησης της ΕΕ2 Κατανεμημένη Ευφυΐα Ιστού και Σκοτεινού Δικτύου.

Τέλος, στο [23], έχει προταθεί ένας ανιχνευτής σκοτεινού δικτύου ανοιχτού κώδικα, ο «ACHE»[11]. Ο ανιχνευτής χρησιμοποιεί το «Privoxy» για να συνδεθεί στο δίκτυο TOR, ενώ στο [24], προτείνεται ο πιο πρόσφατος

[10] https://scrapy.org/
[11] https://ache.readthedocs.io/

ανιχνευτής σκοτεινού ιστού, με το όνομα «TorBot»[12]. Το TorBot έχει τρεις βασικές μονάδες, τον «Crawler», το «Intelligence Extractor» και το «Visualizer». Σχεδιάστηκε με βάση την αναζήτηση κατά πλάτος. Επιπλέον, μπορεί να αναγνωρίσει και να συλλέξει μηνύματα ηλεκτρονικού ταχυδρομείου με μεταδεδομένα, κείμενο, «scripts», «bots», αρχεία, «fuzzy» διευθύνσεις URL και βασικές ευπάθειες που εντοπίζονται στον ιστό. Δυστυχώς, η ανάπτυξη και βελτίωσή του μόλις πρόσφατα σταμάτησε.

Η απόδοση/επεκτασιμότητα είναι κρίσιμο στοιχείο που πρέπει να λαμβάνεται υπόψη σε κάθε υλοποίηση ανιχνευτή. Επηρεάζεται όχι μόνο από την ίδια την υλοποίηση αλλά και από τον αριθμό των ιστοτόπων που μπορεί να ανακαλύψει και να επισκεφθεί. *Οπότε, η λύση για να βελτιωθεί η απόδοση ενός ανιχνευτή είναι να ανιχνεύει ιστοτόπους μέσω πολλαπλών διακομιστών ταυτόχρονα και να επεξεργάζεται τα δεδομένα παράλληλα.*

Η παράλληλη επεξεργασία δεδομένων δεν είναι εύκολη, ωστόσο με τις κατάλληλες τεχνικές μπορεί να υλοποιηθεί βέλτιστα, ειδικά όταν έχουμε ανεξάρτητες συλλογές δεδομένων. Συγκεκριμένα, ένας ανιχνευτής διατηρεί μια ουρά από διευθύνσεις URL. Επιλέγει το πρώτο στοιχείο και το αφαιρεί από την ουρά, επισκέπτεται την αντίστοιχη ιστοσελίδα, ανακτά όλες τις εξερχόμενες διευθύνσεις URL και τις προσθέτει στο τέλος της ουράς. Στη συνέχεια, επιλέγει το επόμενο στοιχείο της ουράς και ούτω καθεξής. Συνεπώς, αν θέλουμε να παραλληλίσουμε αυτή τη διαδικασία πρέπει να έχουμε *πάντα συγχρονισμένη τη λίστα με τις ιστοσελίδες που έχουν ήδη συλλεχθεί*, καθώς και με τις πρόσφατα εντοπισμένες ιστοσελίδες, έτσι ώστε να αποφύγουμε κάποια απώλεια. Για παράδειγμα, μπορούμε να ρυθμίσουμε τον ανιχνευτή να λειτουργεί με έναν ασύγχρονο τρόπο για να το επιτύχουμε αυτό, ή μπορούμε να χρησιμοποιήσουμε πολλαπλά νήματα έχοντας όμως μια κοινή δομή δεδομένων (για παράδειγμα, ούρα προτεραιότητας) για τους ιστοτόπους που ανακαλύπτονται.

Υπάρχουν δυο διαφορετικοί τρόποι πρόσβασης και αποθήκευσης του περιεχομένου σχετικά με τον παράλληλο χειρισμό των δεδομένων: σύγχρονος και ασύγχρονος. Ο σύγχρονος κλείνει το νήμα και αναμένει μέχρι να υπάρξει απάντηση, ενώ ο ασύγχρονος ελευθερώνει το νήμα ανεξάρτητα από την ύπαρξη απάντησης, και το αίτημα μπαίνει αμέσως στην ουρά.

Ένας πολυνηματικός ανιχνευτής, δεν πρέπει να επισκέπτεται την ίδια ιστοσελίδα πολλές φορές και μια ουρά δεν επαρκεί για να λύσει αυτό το ζήτημα. Για να λυθεί, απαιτείται ένα ενημερωμένο σύνολο με όλες τις ανιχνευμένες διευθύνσεις URL. Ένα νέο URL προστίθεται στο σύνολο, μόνο και μόνο αν είναι μοναδικό, δηλαδή δεν έχει ξανά-ανακαλυφθεί / επισκεφθεί στο παρελθόν. Ένας συλλέκτης (collector) είναι υπεύθυνος για τη δημιουργία και διαχείριση / τερματισμό των νημάτων. Τα νήματα λειτουργούν μέχρι η ουρά να αδειάσει. Συνεπώς, η διεργασία του συλλέκτη τερματίζει όλα τα νήματα, αν δεν υπάρχουν άλλα δεδομένα για επεξεργασία. Αυτή είναι μια σημαντική πτυχή της διαδικασίας ανίχνευσης καθώς πηγαίνει βαθύτερα προς το προκαθορισμένο βάθος.

2.2 Επικύρωση και Εξαγωγή Έγκυρων Συνδέσμων

Οι σύνδεσμοι του σκοτεινού δικτύου δεν ακολουθούν τη συμβατική δομή URL του δημοσίου ιστού. Αντίθετα, περιέχουν αλφαριθμητικές συμβολοσειρές με την επέκταση «.onion» ώστε να παραμένουν κρυφοί. Ενδεικτικά, ένας σύνδεσμος διεύθυνσης URL σκοτεινού ιστού έχει την εξής μορφή:

[12] https://torbot.com/

http://zqktlwiuavvvqqt4ybvgvi7tyo4hjl5xgfuvpdf6otjiycgwqbym2qad.onion/

Εξάλλου, αυτές οι αλφαριθμητικές συμβολοσειρές αλλάζουν περιοδικά ακόμα και όταν πρόκειται για την ίδια τοποθεσία του ιστοτόπου, ώστε να μην είναι εύκολα ανιχνεύσιμες. Ως αποτέλεσμα, απαιτείται ένα πρόγραμμα περιήγησης TOR που λειτουργεί ως διακομιστής μεσολάβησης για την πρόσβαση σε αυτούς τους ιστοτόπους, ενώ ορισμένα εμπόδια πρέπει να ξεπεραστούν, συμπεριλαμβανομένων εμποδίων περιεχομένου όπως π.χ. τα CAPTCHA, διαφορετικών κωδικοποιήσεων χαρακτήρων και διαφορετικών πρωτοκόλλων πρόσβασης, όπως παρουσιάζονται εκτενέστερα στις ενότητες που ακολουθούν.

Λόγω του πλήθους των άγνωστων διευθύνσεων URL που έχουν συλλεχθεί, απαιτείται ένα πρωτόκολλο για τον εντοπισμό των πιο «υποσχόμενων» από αυτών, ώστε να μετριαστεί η ανίχνευση των μη χρηστικών συνδέσμων. Αυτό επιτυγχάνεται μέσω ενός πρωτοκόλλου εξαγωγής συνδέσμων, με στόχο να οδηγήσει σε μια πιο αποτελεσματική, στοχευμένη και πιο υποσχόμενη ανίχνευση στο σκοτεινό δίκτυο. Ως αποτέλεσμα, οι ιστοσελίδες που μπορούν να ανακαλυφθούν είναι λιγότερες αλλά πιο έγκυρες κειμενικά, καθώς η εστιασμένη ανίχνευση στοχεύει διευθύνσεις URL προσανατολισμένες στο επιθυμητό περιεχόμενο ή/και σε πιο συναφείς διευθύνσεις URL [25].

Λόγω της ανωνυμίας που χρησιμοποιούν οι διευθύνσεις URL του σκοτεινού δικτύου, είναι δύσκολο να βρεθούν οι επιθυμητοί κρυφοί ιστότοποι. Εάν είναι διαθέσιμη η αρχική σελίδα για την έναρξη της ανίχνευσης, η διαδικασία είναι σημαντικά ευκολότερη. Ωστόσο, αυτό δεν συμβαίνει συνήθως, καθώς αυτές οι πληροφορίες μεταφέρονται από άτομο σε άτομο, είτε αυτοπροσώπως είτε διαδικτυακά. Αυτό το ζήτημα αντιμετωπίζεται με την αναζήτηση σε ορισμένους καταλόγους που περιέχουν συνδέσμους προς κρυφές υπηρεσίες ή χρησιμοποιώντας τη μέθοδο δειγματοληψίας χιονοστιβάδας όπως περιγράφηκε παραπάνω [26].

Η δειγματοληψία χιονοστιβάδας σταματά όταν επιτευχθεί το προκαθορισμένο βάθος αναζήτησης. Έτσι, για την εξάλειψη των συνδέσμων URL που δημιουργούνται τυχαία από μηχανές, ο αλγόριθμος ανίχνευσης συνήθως υπολογίζει αρκετούς δείκτες ακρίβειας για να χαρακτηρίσει τις δομές γράφων και την αποτελεσματικότητά τους [26], [27], όπως οι παρακάτω:

- Τον βαθμό (αριθμός εξερχόμενων συνδέσμων) κάθε κόμβου, συμπεριλαμβανομένου του βαθμού-εισόδου (ο αριθμός των εισερχόμενων συνδέσμων) και του βαθμού-εξόδου (ο αριθμός των εξερχόμενων συνδέσμων).
- Η κεντρικότητα εγγύτητας ενός κόμβου, μετρά τη μέση απόσταση από όλους τους άλλους κόμβους. Με άλλα λόγια, είναι ένας τρόπος ανίχνευσης κόμβων που μπορούν να διαδώσουν εύκολα τις πληροφορίες στο γράφο.
- Η κεντρικότητα διαμέσου ενός κόμβου υπολογίζει τις συντομότερες διαδρομές μεταξύ οποιωνδήποτε ζευγών κόμβων στο γράφο. Χρησιμοποιείται για να ανιχνεύσει το μέγεθος της επιρροής που έχει ένας κόμβος για να διαδώσει τις πληροφορίες σε ένα γράφο.
- PageRank κάθε κόμβου [26]: μια επαναληπτική μέτρηση της κεντρικότητας, η οποία υπολογίζεται με βάση τον αριθμό των συνδέσμων που δείχνουν σε έναν κόμβο και στην κεντρικότητα αυτών. Χρησιμοποιείται για τη μέτρηση της σημασίας ενός κόμβου μέσα σε ένα γράφο.
- Το επίπεδο Assortative Mixing by Degree κάθε κόμβου μετρά το επίπεδο ομοιότητας μεταξύ ενός δεδομένου κόμβου και των γειτόνων του [26].

2.3 Εμπόδια στην Ανακάλυψη Περιεχομένου

Κατά τη διάρκεια ανίχνευσης στον ιστό, συναντώνται ορισμένα εμπόδια περιεχομένου και προκλήσεις για να διαπιστωθεί ότι ο τελικός χρήστης είναι άνθρωπος και όχι αυτόματη μηχανή. Αυτά είναι για παράδειγμα τα CAPTCHA και χρειάζεται να επιλυθούν προκειμένου να προχωρήσει ο ανιχνευτής στην επόμενη φάση (που είναι η πρόσβαση στο πραγματικό περιεχόμενο της ιστοσελίδας). Η επίλυση των CAPTCHA, η οποία συνήθως περιλαμβάνεται σε πολλές σελίδες για τον εντοπισμό αυτόματων μηχανών ανακάλυψης περιεχομένου, είναι ένα από αυτά τα προβλήματα και είναι μια χρονοβόρα διαδικασία. Ως αποτέλεσμα, οι ανιχνευτές ιστού αντιμετωπίζουν προβλήματα απόδοσης, όπως είναι αργή ταχύτητα συλλογής δεδομένων, η συλλογή μη ενημερωμένων δεδομένων από ιστοσελίδες που είναι αποθηκευμένα σε βάσεις δεδομένων εξυπηρετητών, κ.λπ.

Το απλό CAPTCHA [28] είναι μια εικόνα που περιέχει ορισμένα συνήθως αλφαριθμητικά μοτίβα που πρέπει να αναγνωριστούν από έναν χρήστη / άνθρωπο και χρησιμοποιούνται ως επαλήθευση. Υπάρχουν πολλοί τύποι CAPTCHA, όπως CAPTCHA ήχου, CAPTCHA σταθερής εικόνας, CAPTCHA κινούμενης οθόνης, CAPTCHA με μαθηματικού υπολογισμούς, κ.λπ. Επίσης, τα διαφορετικά CAPTCHA έχουν διαφορετικό αριθμό χαρακτήρων, αριθμών και ειδικών σύμβολων, με διαφορετική γραμματοσειρά και μέγεθος, διαφορετικά χρώματα χαρακτήρων καθώς και διαφορετικά φόντα. Η επίλυση τέτοιων CAPTCHA με αυτοματοποιημένες διαδικασίες απαιτεί την επεξεργασίας μιας πολύπλοκης εικόνας. Τα λυμένα CAPTCHA, π.χ. η αναγνώριση χαρακτήρων, τροφοδοτούνται στη διαδικασία ανίχνευσης και η διαδικασία συλλογής δεδομένων συνεχίζεται [27], [29]. Ο αυτόματος εντοπισμός CAPTCHA με χρήση της βιβλιοθήκης ανοικτού κώδικα OpenCV [30] έχει τα ακόλουθα βήματα:

- Η εφαρμογή πρέπει να ξεχωρίσει τα CAPTCHA και τις εικόνες που περιέχονται.
- Θα πρέπει να αναγνωρίζει χαρακτήρες εντός εύρους a-z, A-Z, 0–9.
- Το CAPTCHA πρέπει να είναι σε μορφή png, jpg, jpeg ή bmp για να μπορεί να διαβαστεί ψηφιακά.

Ένα άλλο εμπόδιο στην ανίχνευση περιεχομένου στον ιστό είναι η απαγόρευση πρόσβασης σε ιστοσελίδες σε μη εγγεγραμμένους χρήστες. Η σύνδεση σε ένα διακομιστή μεσολάβησης είναι ένας τρόπος παράκαμψης των περιορισμών. Ένας διακομιστής μεσολάβησης μπορεί να αποκρύψει τον περιορισμένο ιστότοπο από τον πάροχο της υπηρεσίας (ISP) και μπορεί να ενεργοποιήσει την ανώνυμη δραστηριότητα του τελικού χρήστη. Ωστόσο, οι διακομιστές μεσολάβησης δεν κρυπτογραφούν την ανταλλαγή των δεδομένων όπως τα VPN, επομένως είναι λιγότερο ασφαλείς. Μια εναλλακτική σε αυτό είναι η δημιουργία εικονικών ψηφιακών χαρακτήρων ως Personas. Τα Personas είναι bot που μπορούν να οριστούν με συγκεκριμένα διαπιστευτήρια σύνδεσης, και μπορούν να ανακτήσουν περιεχόμενο κατόπιν ελέγχου της ταυτότητάς τους. Σε αυτό το πλαίσιο, προσομοιάζουμε τον τρόπο με τον οποίο ένας πραγματικός χρήστης αποκτά πρόσβαση σε ιστοτόπους σκοτεινού δικτύου ώστε να συλλέξει και να ανακτήσει πληροφορίες. Τα Personas μπορεί να έχουν συγκεκριμένες συνδέσεις με συγκεκριμένα χαρακτηριστικά (για παράδειγμα, υπολογιστής με macOS και πρόγραμμα περιήγησης Safari, ή υπολογιστής με Windows που διαθέτει Chrome, κ.λπ.), όνομα, «user agent» και περιγραφές.

2.4 Διαφορετική Κωδικοποίηση Κειμένων

Η κωδικοποίηση διαφορετικών χαρακτήρων ήταν πάντα ένα ιδιαίτερο πρόβλημα στην επικοινωνία των διακομιστών ιστού. Αυτό συμβαίνει επειδή οι υπολογιστές χρειάζεται να κωδικοποιούν τους χαρακτήρες μεταξύ διαφορετικών γλωσσών και συμβόλων και αφού εμείς, ως άνθρωποι, δεν χρησιμοποιούμε μόνο μία γλώσσα για την επικοινωνία, έχουν προκύψει πολλά προβλήματα κωδικοποίησης. Έτσι, καθιερώθηκαν

αντιστοιχήσεις χαρακτήρων σε υπολογιστή, που είναι γνωστές ως συστήματα κωδικοποίησης χαρακτήρων. Υπάρχουν πολλές κωδικοποιήσεις χαρακτήρων, όπως τα Windows-1250 για τις γλώσσες της Κεντρικής Ευρώπης, τα Windows 1252 για τις δυτικές γλώσσες, τα Windows-1251 για τα κυριλλικά αλφάβητα, τα Windows-1253 για τα ελληνικά, κ.λπ. Επιπλέον, το Unicode είναι μια καθολική κωδικοποίηση που μπορεί να χειριστεί πολλές κωδικοποιήσεις χαρακτήρων [31].

Η κωδικοποίηση χαρακτήρων σε ένα έγγραφο HTML συνήθως δηλώνεται στο «meta» HTML στοιχείο με το «content» ως ετικέτα. Μια τέτοια δήλωση μπορεί να οριστεί στα HTTP «headers», αλλά αυτό δεν συμβαίνει πάντα. Τις περισσότερες φορές, το πρόγραμμα περιήγησης ή ο ανιχνευτής προσπαθεί να μαντέψει την κωδικοποίηση χρησιμοποιώντας ένα Byte Oder Mark (BOM), το οποίο είναι μια ακολουθία από bytes. Επιπλέον, η κωδικοποίηση μπορεί να βρεθεί φιλτράροντας τη γλώσσα του ιστότοπου, αλλά σε κάποιες γλώσσες όπως τα ιαπωνικά δεν λειτουργεί επαρκώς και αποτελεσματικά.

2.5 Διαφορετικά Πρωτόκολλα Πρόσβασης

Ο Παγκόσμιος Ιστός (WWW) είναι ένας από τους πολλούς τρόπους ανάκτησης πληροφοριών από το Διαδίκτυο. Αυτοί οι διαφορετικοί τύποι σύνδεσης στο Διαδίκτυο είναι γνωστοί ως πρωτόκολλα. Το «Hypertext Transfer Protocol» (HTTP) προτάθηκε σαν βασικό πρωτόκολλο επικοινωνίας για τον internet και εξελίσσεται συνεχώς. Για παράδειγμα, το νέο HTTP/3[13] πρωτόκολλο, που χρησιμοποιείται ευρέως στις μέρες μας, άλλαξε το υποκείμενο πρωτόκολλο από TCP σε UDP. Ως αποτέλεσμα, κάθε λογισμικό που λειτουργεί στον ιστό, συμπεριλαμβανομένων και των προγραμμάτων ανίχνευσης ιστού, πρέπει να προσαρμοστεί στην πιο πρόσφατη έκδοση.

Ένα άλλο απαραίτητο πρωτόκολλο για τη διαδικασία ανίχνευσης, είναι το πρωτόκολλο μεσολάβησης (proxy-protocol). Εκτός από την απόκρυψη της IP διεύθυνσης ενός μηχανήματος, οι διακομιστές μεσολάβησης (proxies) μπορεί να ξεπεράσουν τα όρια αιτημάτων που μπορεί να θέτει ο στοχευμένος ιστότοπος. Οι ιστότοποι για να αποφύγουν την κατάρρευση της εφαρμογής από πλημμύρα αιτημάτων (ενδεχομένως και τις DDoS επιθέσεις) διαχειρίζονται το ανώτερο πλήθος των αιτημάτων που μπορούν να δεχτούν και τερματίζουν τις ανενεργές συνδέσεις. Πολλοί ιστότοποι προσπαθούν να αποτρέψουν τα bots και την αυτοματοποιημένη πρόσβαση στο περιεχόμενό τους αποκλείοντας τη μαζική συλλογή περιεχομένου ή οποιαδήποτε ύποπτη δραστηριότητα. Μια τέτοια δραστηριότητα μπορεί να είναι ένας τεράστιος αριθμός αιτημάτων HTTP από την ίδια διεύθυνση IP σε σύντομο χρονικό διάστημα. Αυτός ο περιορισμός μπορεί να αποφευχθεί, όσο γίνεται, με την εφαρμογή ενός πρωτοκόλλου διακομιστή μεσολάβησης (proxy-protocol) στον ανιχνευτή [32] ή με την υιοθέτηση πολιτικών «ευγένειας» (ορίζοντας δευτερόλεπτα αναμονής για την απάντηση από την ιστοσελίδα – χαρακτηριστικό που χρειάζεται στο σκοτεινό δίκτυο γιατί είναι ιδιαιτέρως αργό, αλλάζοντας εκ περιτροπής τη διεύθυνση πρόσβασης, κ.α.) κατά την πρόσβαση σε ιστοσελίδες.

[13] https://www.cloudflare.com/learning/performance/what-is-http3/

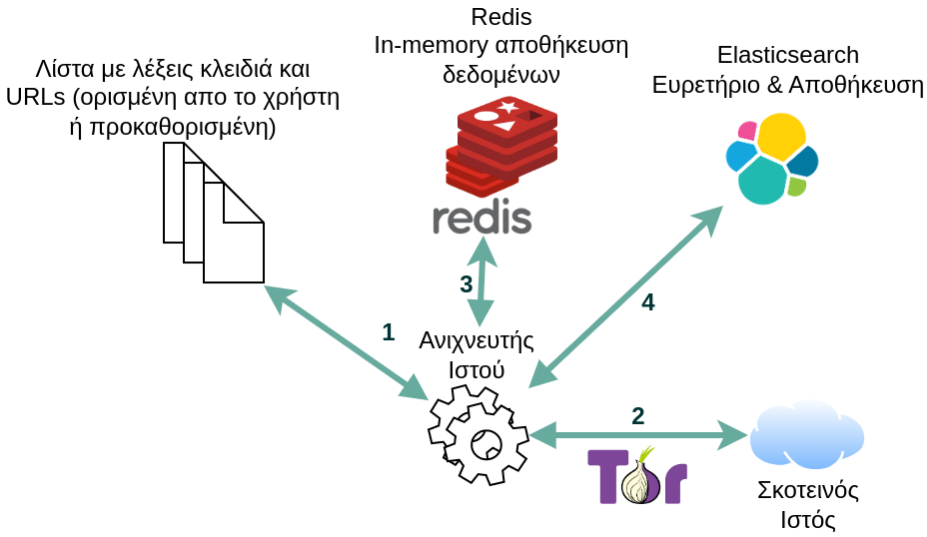
3. Υπηρεσίες ΜΥΡΙΩΠΟΥ για Εξόρυξη Δεδομένων από το Σκοτεινό Δίκτυο και Δημόσιο Ιστό

Οι υπηρεσίες του ΜΥΡΙΩΠΟΥ για την εξόρυξη δεδομένων από το σκοτεινό δίκτυο, αποτελούνται από διαφορετικές τεχνολογίες και εργαλεία, τα οποία έχουν ενοποιηθεί σε ένα εξελιγμένο, δομημένο, υψηλά παραμετροποιήσιμο και κλιμακούμενο λογισμικό εξόρυξης πληροφορίας από τον ιστό και επιμέλειας περιεχομένου για την υποβοήθηση αναλυτικών πάνω από κείμενα, γράφους και τη δημιουργία ευφυΐας. Στο παραδοτέο Π2.1 παρουσιάζουμε το υποσύστημα που εντοπίζει στοχευμένα ιστοσελίδες στο σκοτεινό δίκτυο, παρακάμπτει εμπόδια όπως παρουσιάστηκαν στην επισκόπηση της βιβλιογραφίας και επιμελείται τα δεδομένα για περεταίρω ανάλυση. Οι αναλυτικές για την εξαγωγή γνώσης και η δημιουργία αναφορών και ειδοποιήσεων παρουσιάζεται εκτενέστερα στο παραδοτέο Π2.2. Τα χαρακτηριστικά του υποσυστήματος συλλογής και επεξεργασίας μεγάλων δεδομένων από συλλογές κειμένων, άρθρων και περιεχομένου που εντοπίζεται στο σκοτεινό δίκτυο, αναλύονται σε:

- Υποσύστημα ανίχνευσης και συλλογής:
  - Κλιμακούμενη ανίχνευση ιστού που δεν μπλοκάρεται
  - Πολυεπίπεδη ροή επεξεργασίας
  - Συλλογή αποδεικτικών στοιχείων

3.1 Αρχιτεκτονική Αναφοράς

Η αρχιτεκτονική αναφοράς του υποσυστήματος ανίχνευσης και συλλογής περιεχομένου από το σκοτεινό δίκτυο είναι η ακόλουθη:



Εικόνα 4. Αρχιτεκτονική Υποσυστήματος Συλλογής και Ανίχνευσης

Η λίστα με τα URLs ενημερώνεται σε τακτική βάση είτε για να ξεκινήσει η ανίχνευση με ένα προκαθορισμένο σύνολο από «.onion» URLs, η οποία θα εκτελείται στο παρασκήνιο, είτε για να επιλέξει ο χρήστης τα «.onion» URLs που επιθυμεί, από τα οποία θα ξεκινήσει η ανίχνευση στο σκοτεινό δίκτυο (Βήμα 1). Η λειτουργικότητα που αφορά στο χρήστη έχει οριστεί στις απαιτήσεις και προδιαγραφεί στην Αρχιτεκτονική Συστήματος (Π1.1).

Στη συνέχεια, τα επιλεγμένα URLs μαζί με το βάθος της αναζήτησης που επιθυμεί ο χρήστης, σηματοδοτούν την έναρξη της ανίχνευσης. Ο ανιχνευτής ιστού, μέσω του διακομιστή μεσολάβησης TOR συλλέγει το περιεχόμενο των URLs που επιλέχθηκαν στο πρώτο βήμα (Βήμα 2). Τα καινούργια URLs/σύνδεσμοι που συλλέγονται από το ανακαλυφθέν περιεχόμενο, περνάνε μια διαδικασία «καθαρισμού» και επιμέλειας και στη συνέχεια αποθηκεύονται σε μια in-memory βάση δεδομένων που εξασφαλίζει γρήγορη ανάκτηση και αποθήκευση (Βήμα 3, Redis [34]). Αυτό, χρειάζεται ώστε ο ανιχνευτής ιστού να είναι σε θέση να γνωρίζει ποιους συνδέσμους έχει επισκεφθεί, έτσι ώστε να αποφευχθούν επαναλήψεις. Το περιεχόμενο που συλλέγεται ευρετηριάζεται και αποθηκεύεται (Βήμα 4, Elasticsearch [35]).

3.2 Ανάκτηση και Συλλογή Δεδομένων

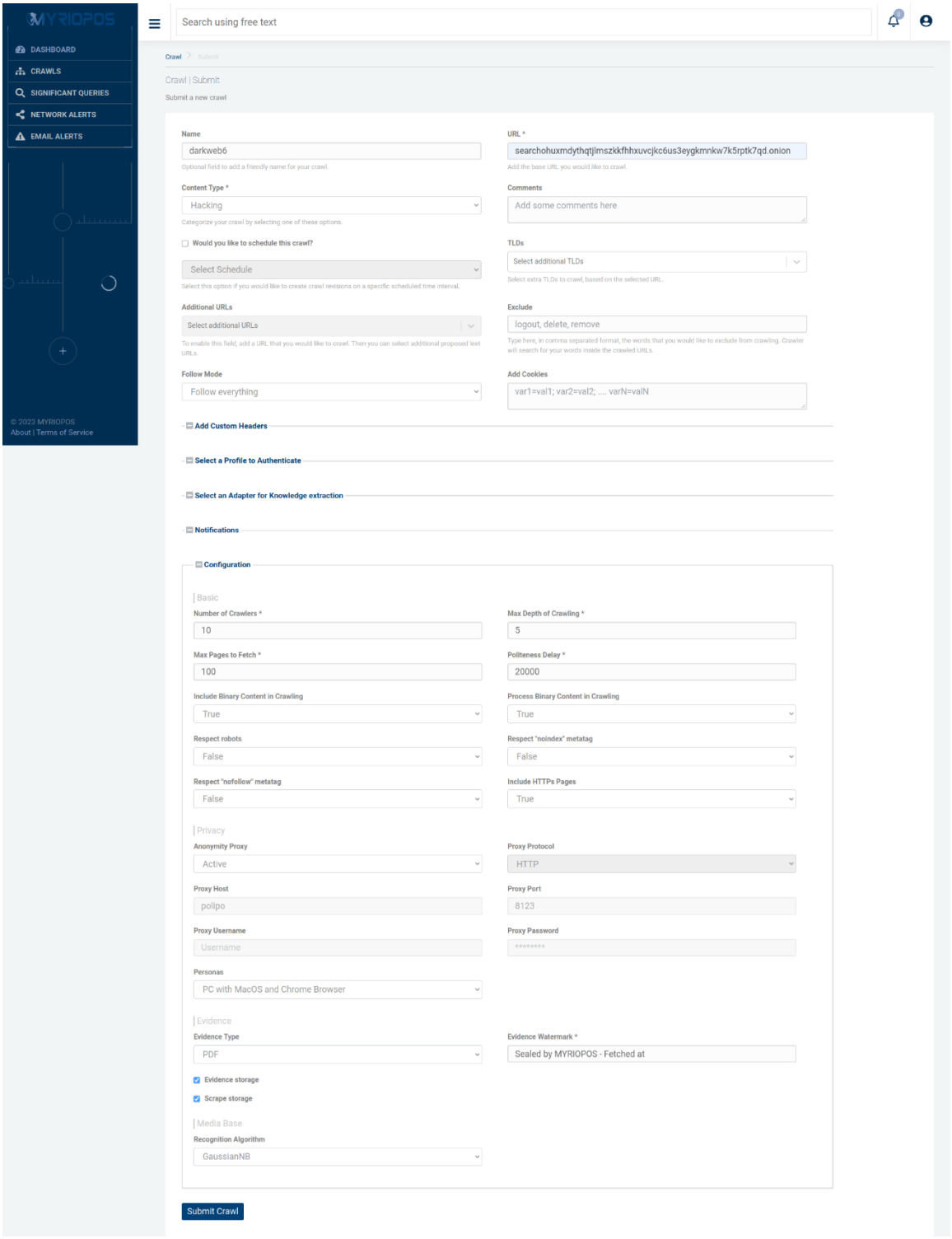
Η διαδικασία συλλογής δεδομένων από το σκοτεινό δίκτυο, φόρουμ και παράνομων αγορών, που σχετίζονται με έννοιες σχετικές στα πλαίσια του ΜΥΡΙΩΠΟΥ, προϋποθέτει κάποια βήματα αρχικοποίησης όπως παρουσιάζονται στην Εικόνα 5. Η εικόνα παρουσιάζει τον τρόπο που έχει ενοποιηθεί το σύνολο των υπηρεσιών για το σκοτεινό δίκτυο και δημόσιο ιστό στο ολοκληρωμένο σύστημα του ΜΥΡΙΩΠΟΣ.

Το πρώτο βήμα, διαθέσιμο και στους τελικούς χρήστες του συστήματος, είναι η αρχικοποίηση της λίστας των συνδέσμων από τους οποίους θα ξεκινήσει ο ανιχνευτής την αναζήτησή του. Η αρχικοποίηση αυτή μπορεί να γίνει είτε από τον τελικό χρήστη, είτε από το σύστημα του ΜΥΡΙΩΠΟΥ αυτόματα και αθόρυβα. Στη συνέχεια πρέπει να γίνουν ορισμένες ρυθμίσεις στις υπηρεσίες του ΜΥΡΙΩΠΟΥ, να παραμετροποιηθεί ο διακομιστής μεσολάβησης και να αρχικοποιηθεί η λίστα με τις λέξεις κλειδιά (η οποία εμπλουτίζεται συνεχώς όσο λαμβάνουν χώρα νέες ανακαλύψεις περιεχομένου), έτσι ώστε η αναζήτηση να είναι εστιασμένη σε περιεχόμενο που αναφέρεται σε περιστατικά κυβερνοασφάλειας (συλλογές από παραβιασμένα δεδομένα, SQL Injection, phishing, επιθέσεις DDoS, διαπιστευτήρια εισόδου που πωλούνται στο σκοτεινό δίκτυο, διαβατήρια, διπλώματα οδήγησης κ.λπ.), υποκλοπή λογαριασμών email (παραβιασμένους λογαριασμούς email, τίτλοι εργασίας, φυσικές διευθύνσεις, προφίλ κοινωνικής δικτύωσης, ονόματα, τηλέφωνα, κ.λπ.), υποκλοπές σε εταιρικούς διακομιστές email, εντοπισμός διακομιστών email που βρίσκονται στη μαύρη λίστα. Καθώς συλλέγουμε το περιεχόμενο σε μορφή html, εκτελούμε ορισμένες λειτουργίες καθαρισμού δεδομένων για να επιμεληθούμε το κείμενο που συλλέγεται, να υποβοηθήσουμε την ανάλυση που θα ακολουθήσει σε επόμενο βήμα ή τις διευθύνσεις URL που περιέχουν μόνο εικόνες.

Η διαδικασία της ανάκτησης και συλλογής δεδομένων μπορεί να αρχικοποιηθεί είτε από τον τελικό χρήστη μέσω της γραφικής διεπαφής στην οποία μπορεί να επιλέξει τα URLs που επιθυμεί, το βάθος της αναζήτησης, τις πολιτικές «ευγένειας» για να μην καταρρεύσει ο διακομιστής εξυπηρέτησης καθώς και τις λέξεις κλειδιά, είτε από τις back-end υπηρεσίες του ΜΥΡΙΩΠΟΥ, όπου χρησιμοποιείται μια προκαθορισμένη λίστα από «.onion» URLs για την εκκίνηση της ανίχνευσης. Στο ολοκληρωμένο σύστημα του ΜΥΡΙΩΠΟΣ, αυτή η λειτουργία είναι προσβάσιμη μέσω της γραφικής διεπαφής, η οποία επιτρέπει τον ορισμό των URLs από τον χρήστη, πριν από την έναρξη της ανίχνευσης. Η πρόσβαση και η επεξεργασία των διαφορετικών διευθύνσεων URL γίνεται μέσα σε 2-3 δευτερόλεπτα, ενώ κάθε μέρα συλλέγουμε και αποθηκεύουμε περίπου 20.000 νέες σκοτεινές ιστοσελίδες.

Όπως φαίνεται και μπορεί να αποτελέσει μέρος επίδειξης της εφαρμογής, η υπηρεσία συλλογής δεδομένων, μπορεί να παραμετροποιηθεί με τις στοχευμένες λέξεις-κλειδιά που αναζητά στο περιεχόμενο του ο χρήστης, με την επιθυμία του αν θέλει να προγραμματίσει τη συλλογή δεδομένων ώστε να εκτελεστεί μια δεδομένη

χρονική στιγμή. Επίσης, ο χρήστης μπορεί να καθορίσει το βάθος αναζήτησης και τη μετρική «ευγένειας» καθώς και τα χαρακτηριστικά των εικονικών ψηφιακών χαρακτήρων, Personas.



Εικόνα 5. Αρχικοποίηση Ανίχνευσης

Σημειώνεται ότι τα θέματα κωδικοποίησης και επίλυσης βασικών CAPTCHAs είναι μέρος εσωτερικής διαχείρισης σε επίπεδο κώδικα και δεν παραμετροποιούνται από το χρήστη. Για μεγαλύτερη ανάκληση περιεχομένου σε επίπεδο κωδικοποίησης διαχειριζόμαστε κυρίως ιστοσελίδες σε UTF-8 και UTF-16. Αναφορικά με τα CAPTCHAs χρησιμοποιούμε τη βιβλιοθήκη ανοικτού κώδικα OpenCV για να διαχειριστούμε επαρκώς απλά αλφαριθμητικά και όχι πιο σύνθετα όπως έχουν εξελιχθεί συνδυάζοντας κουίζ με εικόνες, ή φωνητικά μηνύματα τα τελευταία χρόνια. Από τη συνολική συλλογή των 84037 σελίδων που έχουμε συγκεντρώσει μέχρι στιγμής, το 3% αυτών περιείχαν αλφαριθμητικό CAPTCHA το οποίο έχουμε διαχειριστεί αποτελεσματικά. Όσον αφορά στα πιο σύνθετα κουίζ, αγνοούμε την αναζήτηση δίχως να τερματίζει η διαδικασία καθώς είναι πολυνηματική. Ενδεικτικά αναφέρουμε ότι μέσω αυτόματης συλλογής περιεχομένου ή οποία διεξάγεται κατά τη διάρκεια της νύχτας που η κίνηση στο σκοτεινό δίκτυο και το δημόσιο ιστό είναι λιγότερη, αρχικοποιούμε έναν παραμετροποιήσιμο αριθμό, για παράδειγμα 10, από παράλληλα νήματα τα οποία τερματίζουν όταν φτάσουν το επιθυμητό βάθος αναζήτησης. Αν κάποιο νήμα τερματίσει απότομα δίχως να φτάσει το επιθυμητό βάθος αναζήτησης το επανεκκινούμε αρχικοποιώντας με νέο URL.

Τέλος, αναφορικά με τη δημιουργία αναφορών σχετικά με το αν κάποιο email (χρήστη ή εταιρικό) έχει υποκλαπεί, χρησιμοποιούμε την υπηρεσία δημόσιου ιστού «have I been pawned» [33] προκειμένου να δημιουργήσουμε μια ειδοποίηση στην Κεντρική Διεπαφή Χρήστη του ΜΥΡΙΩΠΟΣ. Περισσότερες τεχνικές λεπτομέρειες θα δοθούν στο Π2.2 που συγκεντρώνει όλες τις αναλυτικές μεθόδους και τις πολυτροπικές ειδοποιήσεις που υλοποιούνται.

4. Επίλογος

Ο στόχος του Π2.1 ήταν να τεκμηριώσει την πρώτη επανάληψη υλοποίησης (υπενθυμίζουμε ότι θα πραγματοποιηθούν 2 κύκλοι υλοποίηση στα πλαίσια του ΜΥΡΙΩΠΟΣ) στα πλαίσια της ΕΕ2, με σκοπό να περιγραφούν οι αντίστοιχες μέθοδοι, τεχνολογίες και εργαλεία που χρησιμοποιούνται για την υποστήριξη της κατανεμημένης ευφυΐας ιστού και σκοτεινού δικτύου.

Για αυτό το σκοπό, το συγκεκριμένο παραδοτέο έχει διττό ρόλο. Από τη μία πλευρά, στοχεύει στη διεξαγωγή μιας ενδελεχούς βιβλιογραφικής επισκόπησης που υποβοηθά στον καθορισμό της τρέχουσας κατάστασης των ερευνών στα πλαίσια ανίχνευσης ιστού, εμποδίων περιεχομένου, εξαγωγής έγκυρων συνδέσμων, διαφορετικών κωδικοποιήσεων και πρωτοκόλλων. Ταυτόχρονα, περιγράφει τις διαδικασίες και λειτουργίες λογισμικού του ΜΥΡΙΩΠΟΥ για την υλοποίηση της συλλογής και ανίχνευσης περιεχομένου από το σκοτεινό δίκτυο και δημόσιο ιστό.

Τέλος, στο παραδοτέο Π2.2 θα εξηγηθεί το υποσύστημα που εκπαιδεύει τα μοντέλα και παράγει τις αναλυτικές μεθόδους από το περιεχόμενο του σκοτεινού δικτύου, τις ειδοποιήσεις και αναφορές που εκθέτει στην Κεντρική Διεπαφή Χρήστη και πληροφορίες σχετικά με τη διαλειτουργικότητα με εργαλεία τρίτων.

5. Αναφορές

[1] Tor Browser. Διαθέσιμο στο Διαδίκτυο: https://www.torproject.org/
[2] Facts about Google and Competition. Διαθέσιμο στο Διαδίκτυο: https://web.archive.org/web/20111104131332/https://www.google.com/competition/howgooglesearchworks.html
[3] Google Search Engine. Διαθέσιμο στο Διαδίκτυο: https://www.google.gr/
[4] Yahoo Search Engine. Διαθέσιμο στο Διαδίκτυο: https://www.yahoo.com/
[5] Bing Search Engine. Διαθέσιμο στο Διαδίκτυο: https://www.bing.com/
[6] Google Chrome Browser. Διαθέσιμο στο Διαδίκτυο: https://www.google.com/chrome/
[7] Mozilla Firefox Browser. Διαθέσιμο στο Διαδίκτυο: https://www.mozilla.org/en-US/firefox
[8] Opera Browser: Διαθέσιμο στο Διαδίκτυο: https://www.opera.com/
[9] Darknet Markets Are Not beyond the Reach of Law. 2016. Accessed August 30, 2016. Διαθέσιμο στο Διαδίκτυο: https://darkwebnews.com/darknet-markets/darknet-not-beyond-law/
[10] Facebook Application. Διαθέσιμο στο Διαδίκτυο: https://www.facebook.com/
[11] Twitter Application. Διαθέσιμο στο Διαδίκτυο: https://twitter.com/
[12] Instagram Application. Διαθέσιμο στο Διαδίκτυο: https://www.instagram.com/
[13] Snapchat Application. Διαθέσιμο στο Διαδίκτυο: https://www.snapchat.com/
[14] Sabillon, R., Cavaller, V., Cano, J., & Serra-Ruiz, J. (2016, June). Cybercriminals, cyberattacks and cybercrime. In 2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF) (pp. 1-9). IEEE.
[15] Owen, Gareth, and Nick Savage. 2015. “The Tor Dark Net.” Accessed December 13, 2016. Διαθέσιμο στο Διαδίκτυο: https://www.ourinternet.org/sites/default/files/publications/no20_0.pdf
[16] Croft, W. Bruce, Donald Metzler, and Trevor Strohman. Search engines: Information retrieval in practice. Vol. 520. Reading: Addison-Wesley, 2010.
[17] Chen, Hsinchun. Dark web: Exploring and data mining the dark side of the web. Vol. 30. Springer Science & Business Media, 2011.
[18] Zulkarnine, Ahmed T., et al. "Surfacing collaborated networks in dark web to find illicit and criminal content." 2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, 2016.
[19] Fu, Tianjun, Ahmed Abbasi, and Hsinchun Chen. "A focused crawler for Dark Web forums." Journal of the American Society for Information Science and Technology 61.6 (2010): 1213-1231.
[20] The MISP Threat Sharing Platform. Διαθέσιμο στο Διαδίκτυο: https://www.misp-project.org/objects.html
[21] AlKhatib, RB Bassel, and Randa Basheer. "Crawling the Dark Web: A Conceptual Perspective, Challenges and Implementation." Journal of Digital Information Management (JDIM) 17.2 (2019): 51-60.
[22] Pantelis, G., Petrou, P., Karagiorgou, S., & Alexandrou, D. (2021, August). On Strengthening SMEs and MEs Threat Intelligence and Awareness by Identifying Data Breaches, Stolen Credentials, and Illegal Activities on the Dark Web. In The 16th International Conference on Availability, Reliability and Security (pp. 1-7).
[23] ACHE Crawler Documentation. Διαθέσιμο στο Διαδίκτυο: https://ache.readthedocs.io/en/latest/

[24] Narayanan, P. S., R. Ani, and Akeem TL King. "TorBot: Open-Source Intelligence Tool for Dark Web." Inventive Communication and Computational Technologies. Springer, Singapore, 2020. 187-195.
[25] Sharma, D. K., & Sharma, A. K. (2011). A Novel architecture for deep web crawler. International Journal of Information Technology and Web Engineering (IJITWE), 6(1), 25-48.
[26]Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. Comput. Networks, 56, 3825-3833.
[27] Wolfe, A. W. (1997). Social network analysis: Methods and applications. American Ethnologist, 24(1), 219-220.
[28] Mane, S., Lokare, M., & Gite, B. B. Solving CAPTCHAs Automatically for Web Crawling.
[29] Automation of resolving CAPTCHAs for web crawling. Διαθέσιμο στο Διαδίκτυο: https://medium.com/@abhay_bhagat/automation-of-resolving-captchas-for-web-crawling-4026495eeea4
[30] OpenCV. Διαθέσιμο στο Διαδίκτυο: https://opencv.org/
[31] Nazah, S., Huda, S., Abawajy, J., & Hassan, M. M. (2020). Evolution of dark web threat analysis and detection: A systematic approach. IEEE Access, 8, 171796-171819.
[32] Khosrow, M. (2018). The Dark Web: Breakthroughs in Research and Practice.
[33] Have I been pawned? Διαθέσιμο στο Διαδίκτυο: https://haveibeenpwned.com/
[34] Redis. Διαθέσιμο στο Διαδίκτυο: https://redis.io/
[35] Elasticsearch. Διαθέσιμο στο Διαδίκτυο: https://www.elastic.co/
[36] National Institute of Standards & Technology (NIST) Special Publication (SP) 800-150 (2016). Guide to Cyber Threat Information Sharing. Διαθέσιμο στο Διαδίκτυο: http://dx.doi.org/10.6028/NIST.SP.800-150.