ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΚΑΤΑΣΚΕΥΗ ΒΑΣΗΣ ΔΕΔΟΜΕΩΝ ΜΕ ΙΣΤΟΤΟΠΟΥΣ ΠΟΥ ΠΕΡΙΕΧΟΥΝ ΚΑΚΟΒΟΥΛΟ ΛΟΓΙΣΜΙΚΟ, ΜΕ ΧΡΗΣΗ ΤΗΣ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗΣ ΓΛΩΣΣΑΣ PYTHON

Διπλωματική Εργασία

Παναγιώτης Κλωνής

Επιβλέπων: Γεώργιος Σταμούλης

Βόλος 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΚΑΤΑΣΚΕΥΗ ΒΑΣΗΣ ΔΕΔΟΜΕΩΝ ΜΕ ΙΣΤΟΤΟΠΟΘΣ ΠΟΥ ΠΕΡΙΕΧΟΥΝ ΚΑΚΟΒΟΥΛΟ ΛΟΓΙΣΜΙΚΟ, ΜΕ ΧΡΗΣΗ ΤΗΣ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗΣ ΓΛΩΣΣΑΣ PYTHON

Διπλωματική Εργασία

Παναγιώτης Κλωνής

Επιβλέπων: Γεώργιος Σταμούλης

Βόλος 2020

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

CONSTRUCTION OF A DATABASE OF DOMAINS, CONTAINING MALICIOUS SOFTWARE, USING THE PYTHON PROGRAMMING LANGUAGE

Diploma Thesis

Panagiotis Klonis

Supervisor: Georgios Stamoulis

Volos 2020

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων	**Γεώργιος Σταμούλης** Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανι- κών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Ασπασία Δασκαλοπούλου** Επίκουρος Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανι- κών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Παναγιώτης Κίκιρας** EDA Research and Technology Coordinator – Head of Unit Technology and Innovatio at European Defence Agency, Τμήμα Ηλεκτρολόγων Μηχανι- κών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ημερομηνία έγκρισης: 08-10-2020

ΕΥΧΑΡΙΣΤΙΕΣ

Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή αυτής της διπλωματικής εργασίας Γεώργιο Σταμούλη, ο οποίος μου επέτρεψε να εργαστώ πάνω στο θέμα χωρίς περιορισμούς και άλογες απαιτήσεις.

Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένειά μου για την οικονομική και ψυχολογική στήριξη που μου παρείχε καθ' όλη τη διάρκεια της φοίτησής μου στο τμήμα, η οποία συνοδεύτηκε από πολλές δυσκολίες.

Τέλος, θα ήθελα να ευχαριστήσω τον εγκάρδιο φίλο μου Γεώργιο Αϊβάτογλου, ο οποίος συνέδραμε στην οριστικοποίηση του θέματος αυτής της εργασίας μέσω πολύωρων συζητήσεων.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο/Η Δηλών/ούσα

(Υπογραφή)
Ονοματεπώνυμο Φοιτητή/ήτριας
Ημερομηνία

ΠΕΡΙΛΗΨΗ

Η ασφάλεια είναι από τα πιο σημαντικά συστατικά για την διατήρηση των υπολογιστικών συστημάτων και της τεχνολογίας της δικτύωσης. Αυτό συμβαίνει, διότι με τους ηλεκτρονικούς υπολογιστές και την δικτύωση διαχειριζόμαστε πληροφορίες που είναι ένα από τα βασικότερα συστατικά λειτουργίας του σύγχρονου κόσμου. Υπάρχουν πληροφορίες που επιθυμούμε να διαμοιράζονται εύκολα και γρήγορα. Υπάρχουν όμως και τέτοιες πληροφορίες, οι οποίες θέλουμε να παραμένουν γνωστές μεταξύ μικρών ομάδων ατόμων και άλλες που επιθυμούμε, να παραμένουν κρυφές. Είναι λοιπόν εύλογο, να προσπαθούμε, να διατηρούμε τα υπολογιστικά συστήματα και τη δικτύωση ακέραια ως προς την καλή λειτουργία και την ασφάλεια. Αυτό όμως δεν είναι πάντα εφικτό. Η δομή και η σχεδίαση των ηλεκτρονικών υπολογιστών και των δικτύων επιτρέπουν την δημιουργία απειλών από επιτήδειες πλευρές, που αφορούν την ασφάλεια και κατ΄ επέκταση την απειλή πάνω στην πληροφορία. Τέτοιες απειλές αφορούν στην παραποίηση ή άρση της λειτουργίας των υπολογιστικών συστημάτων αφενός και αφετέρου την κλοπή, παραποίηση πληροφοριών και φυσικά τη μη εξουσιοδοτημένη χρήση αυτών. Ιστορικά οι απειλές αλλάζουν μορφή. Στην αρχή της ανάπτυξης των υπολογιστικών συστημάτων αρκούσε μια απλή διακοπή ρεύματος για την πρόκληση ζημιάς σε ένα σύστημα. Με την πάροδο των χρόνων και φτάνοντας στη σύγχρονη ψηφιακή εποχή, όπου ο προγραμματισμός των συστημάτων γίνεται σε πολύ υψηλό επίπεδο, οι απειλές έχουν αναβαθμιστεί κι αυτές και εκφέρονται εν μέρει ως κακόβουλο λογισμικό. Σε αυτή τη διπλωματική εργασία αρχικώς, θα εξετάσουμε το πρόβλημα της ασφάλειας η οποία απειλείται από κακόβουλο λογισμικό και θα αναλύσουμε κάποια βασικά συστατικά των υπολογιστικών συστημάτων τα οποία είναι σχετικά με το θέμα αυτό. Στη συνέχεια, θα παρουσιάσουμε τα εργαλεία που χρησιμοποιήθηκαν για την ανάπτυξη του κώδικα για τη δημιουργία μίας βάσης δεδομένων με ιστοσελίδες όπου κρύβεται κακόβουλο λογισμικό και τέλος κάποια στατιστικά στοιχεία που προκύπτουν από τις πληροφορίες που μπήκαν στη βάση δεδομένων μετά από επεξεργασία.

ABSTRACT

Security is one of the most important components in maintaining computer systems and networking technology. This is because by using computers and networking we manage information, and this is one of the key ways of how the modern world operates. The world wants information to be shared easily and fast. However, there is the kind of information that we want to be shared only among small groups of people and other types of information that we want it to remain secret and private. It is, therefore, reasonable to try and keep computer systems and networking intact as far as the functionality and security are concerned. But this is not always possible. The structure and design of computers and networks allow cunning parties to create threats which concern security and consequently the safety of information. Such threats are about both the falsification or the decommissioning of the functionality of computer systems, and the theft and falsification of information, as well as, of course, its unauthorized use. Throughout computer history threats have changed forms. At the beginning of the development of computer systems, a simple power outage was enough to cause damage to a system. Over the years and especially during the modern digital age, where systems programming is done at a very high level, the threats have also been upgraded and are partly delivered as malware. In this dissertation I will firstly examine the security problems posed by malware and analyze some key components of computer systems that are relevant to this topic. Next, I will present the tools used to develop the code in order to create a database of malicious websites, and finally some derived statistics from the information inserted into the database after processing it.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ……vii

ABSTRACT……viii

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ……ix

ΕΙΣΑΓΩΓΗ……1

ΚΕΦΑΛΑΙΟ 1……3

ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΑΣΦΑΛΕΙΑΣ……3

1.1 Εισαγωγή……3

1.2 Κακόβουλο λογισμικό……5

1.2.1 Ορισμός – Συμπτώματα μολυσμένου συστήματους……5

1.2.2 Πώς ένα σύστημα μολύνεται από κακόβουλο λογισμικό……6

1.2.3 Μορφές και είδη κακόβουλου λογισμικού……7

1.3 Επισκόπηση του προβλήματος……9

1.4 Καθολική Βάση Δεδομένων Κακόβουλου Λογισμικού……11

ΚΕΦΑΛΑΙΟ 2……14

ΕΡΓΑΛΕΙΑ ΚΑΙ ΜΕΘΟΔΟΙ ΠΟΥ ΣΥΝΘΕΤΟΥΝ ΤΗΝ ΕΠΙΘΥΜΗΤΗ ΠΛΗΡΟΦΟΡΙΑ ΚΑΙ ΟΙ ΤΡΟΠΟΙ ΑΝΤΛΗΣΗΣ ΤΗΣ……14

2.1 Εισαγωγή……14

2.2 TCP/IP……14

2.3 Βάσεις δεδομένων……18

2.4 Unstrurctured Data – Structured Data……20

2.4.1 Unstructured Data – Μη δομημένα δεδομένα……20

2.4.2 Structured Data – Δομημένα δεδομένα……22

2.5 Web Scraping……23

2.6 STIX – TAXII……26

2.6.1 STIX……26

2.6.2 TAXII……29

2.7 Google Safe Browsing……30

ΚΕΦΑΛΑΙΟ 3……32

ΥΛΟΠΟΙΗΣΗ – ΑΠΟΤΕΛΕΣΜΑΤΑ……32

3.1.1 Python 2……32

3.1.2 Beautiful Soup……32

3.1.3 urlib2……34

3.2 Τα τέσσερα scripts του κώδικα……34

3.2.1 Το πρώτο script……34

3.2.1.1 Εξόρυξη από τον HTML κώδικα της σελίδας……35

3.2.1.2 Εξόρυξη με API……38

3.2.1.3 Προβολή της βάσης δεδομένων σε γραφικό περιβάλλον……42

3.2.2 Το δεύτερο Script……44

3.2.3 Το τρίτο script……46

3.2.5 Το τέταρτο script - γραφήματα……46

ΒΙΒΛΙΟΓΡΑΦΙΑ……50

ΕΙΣΑΓΩΓΗ

Η περιέργεια του ανθρώπου σε συνδυασμό με τις ανάγκες της καθημερινότητας τον παρακινεί να ερευνά, να ανακαλύπτει, να πειραματίζεται και εν τέλει να φτάνει σε επιστημονικά επιτεύγματα. Οι άνθρωποι χρησιμοποιούν ως βάση αυτά που ήδη ξέρουν, και με ατομικές προσπάθειες, αλλά κυρίως συντονισμένες προσπάθειες ομάδων ατόμων καταλήγουν να δημιουργούν-εφευρίσκουν καινούρια εργαλεία τα οποία ως βασικό στόχο έχουν την επίλυση προβλημάτων.

Μία βασική ανάγκη που έπρεπε να εκπληρώσει ο άνθρωπος προκειμένου να μπορέσει να εξελίξει τις επιστήμες, ήταν αυτή των γρήγορων και αξιόπιστων υπολογισμών. Αρχικά, ήθελε να επιτύχει ταχύτητα στους απλούς μαθηματικούς υπολογισμούς προκειμένου να μειωθεί ο απαιτούμενος χρόνος εκτέλεσης αυτών. Με την πάροδο των χρόνων και την ανάπτυξη της επιστήμης η μείωση του χρόνου περάτωσης των υπολογισμών δεν ήταν αρκετή, καθώς οι υπολογισμοί άρχισαν να περιπλέκονται και να απαιτούν περισσότερους πόρους. Από τις αρχές της δεκαετίας του 1940 εμφανίζονται οι πρώτοι υπολογιστές που εκτελούσαν αποκλειστικά δύσκολες μαθηματικές πράξεις πολλές φορές γρηγορότερα από τον άνθρωπο και πολύ σύντομα κατασκευάστηκε μηχάνημα που έτρεξε το πρώτο πρόγραμμα γραμμένο για υπολογιστή[1].

Οι υπολογιστές για κάποιο καιρό είχαν πολύ μεγάλο μέγεθος. Μετέπειτα, κατασκευάστηκαν μηχανήματα που μπορούσαν να εκτελούν επιπλέον εργασίες πέρα από μαθηματικούς υπολογισμούς. Στα μέσα της δεκαετίας του 1970 εμφανίστηκαν στο εμπόριο οι πρώτοι υπολογιστές με μέγεθος ενός μικρού θρανίου, και από τις αρχές του 1980 και μετά κατασκευάζονται μηχανήματα περίπου με τη μορφή που έχουν σήμερα οι προσωπικοί επιτραπέζιοι ηλεκτρονικοί υπολογιστές, φορητοί υπολογιστές(laptop) και κινητά τηλέφωνα ικανά πλέον να προσφέρουν σε επιστήμονες και ερευνητές γενναία υπολογιστική δύναμη. Διαθέτουν επεξεργαστή και μνήμη και ο προγραμματισμός ανθίζει. Στις αρχές της δεκαετίας του 1990 διάφορα μοντέλα υπολογιστικών μηχανών άρχισαν να διατίθενται μαζικά στο εμπόριο, που φυσικά μπορούσαν να διεκπεραιώσουν πολυσύνθετες εργασίες. Σημείο αναφοράς για την έξαρση της αγοράς ηλεκτρονικών

υπολογιστών από ιδιώτες είναι τα μέσα της δεκαετίας του 1990 όπου πλέον εμπορικοποιείται το internet.

Η άλλη βασική ανάγκη που δημιουργήθηκε ήταν αυτή της γρήγορης επικοινωνίας και ανταλλαγής πληροφοριών. Η έρευνα στις επιστήμες γίνεται και σε ακαδημαϊκό επίπεδο πέρα από τις ιδιωτικές επιχειρήσεις. Η αρχική ιδέα της δικτύωσης υλοποιήθηκε αρχικώς από κυβερνητικές υπηρεσίες των ΗΠΑ οι οποίες δημιούργησαν δίκτυα υπολογιστών για να χρησιμοποιηθούν από τις ένοπλες δυνάμεις[2]. Στη συνέχεια, η ιδέα της δικτύωσης αξιοποιείται σε διάφορες χώρες ικανές να επενδύσουν σε αυτό το θέμα, κυρίως για κυβερνητικές υπηρεσίες αλλά και από ιδιωτικές εταιρίες και πανεπιστήμια που χρησιμοποιούν τοπικά δίκτυα χάριν γρηγορότερης επικοινωνίας και ανταλλαγής πληροφοριών. Το 1982 επισημοποιείται το πρωτόκολλο TCP/IP το οποίο είναι πρωτόκολλο επικοινωνίας πάνω από δίκτυο και προσφέρει επικοινωνία μεταξύ δύο άκρων. Το 1983 δημιουργείται το DNS(Domain Name System) το οποίο είναι ένα σύστημα ονοματολογίας υπολογιστών, υπηρεσιών ή άλλων πηγών που είναι διασυνδεδεμένοι με ένα ιδιωτικό δίκτυο ή το internet.

Το internet λοιπόν δημιουργείται στις αρχές του 1980 και είναι απόρροια διασύνδεσης των σημαντικότερων δικτύων που είχαν δημιουργηθεί έως τότε. Ως το 1995 δημιουργούνται τεχνολογίες στις οποίες θα βασιστεί η καλή λειτουργία του και η εμπορικοποίηση του, η οποία σηματοδοτεί την έναρξη αγοράς προσωπικών ηλεκτρονικών υπολογιστών σε μεγάλη κλίμακα.

Σε αυτή τη φάση που η χρήση των ηλεκτρονικών υπολογιστών και του ίντερνετ αρχίζει σταδιακά και γιγαντώνεται σε συνδυασμό με την ανάπτυξη τεχνολογιών για αποθήκευση και διαχείριση πληροφοριών online, η εγκληματικότητα στον κυβερνοχώρο αρχίζει να παίρνει σημαντικές διαστάσεις. Έτσι λοιπόν, δημιουργείται η ανάγκη για εντατικοποίηση των προσπαθειών για λήψη μέτρων ασφαλείας εναντίον της εγκληματικότητας στον κυβερνοχώρο και οι θεμελιώνονται σιγά σιγά οι κλάδοι της ασφάλειας δικτύων και κυβερνοασφάλειας(cyber security) στο χώρο της επιστήμης των υπολογιστών.

ΚΕΦΑΛΑΙΟ 1

ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΑΣΦΑΛΕΙΑΣ

1.1 Εισαγωγή

Το πρόβλημα της ασφάλειας είναι άρρηκτα συνδεδεμένο με την εγκληματικότητα στον κυβερνοχώρο. Η εγκληματικότητα στον κυβερνοχώρο είναι ο όρος που εμπεριέχει κάθε παράνομη δραστηριότητα για την οποία απαιτείται τουλάχιστον ένας ηλεκτρονικός υπολογιστής και το διαδίκτυο[3]. Το εύρος των παράνομων δραστηριοτήτων είναι μεγάλο. Εμπεριέχει την παράνομη διείσδυση σε ένα δίκτυο, τη διάδοση κακόβουλου λογισμικού μέσω του δικτύου, όπως επίσης και τις παραλλαγές αυτών των εγκλημάτων που απορρέουν σαφώς από το είδος του κακόβουλου λογισμικού που διανέμεται και λειτουργεί παρά τη θέληση των κατόχων των ηλεκτρονικών υπολογιστών και πολλές φορές εν αγνοία τους. Μερικές από τις βασικές παρανομίες που διαπράττονται είναι η κλοπή ταυτότητας, η παρακολούθηση φυσικών προσώπων μέσω της δραστηριότητάς τους στο διαδίκτυο(stalking), ο εκφοβισμός, η τρομοκρατία, η πώληση παράνομων προϊόντων, ο εκβιασμός αλλά και η διακοπή δραστηριότητας σημαντικών επιχειρήσεων των οποίων το ιδιωτικό δίκτυο έχει υποστεί εισβολή από εξωτερική επίθεση.

Οι οργανισμοί που παρέχουν υπηρεσίες μέσω του διαδικτύου αναβαθμίζουν συνεχώς τα μέτρα ασφαλείας τους, ώστε να διατηρούν ασφαλείς τις δραστηριότητές τους αλλά και τα δεδομένα των πελατών τους. Οι επιθέσεις για την κλοπή προσωπικών δεδομένων είναι αναμενόμενες, και για αυτό το λόγο οι επιχειρήσεις δαπανούν μεγάλα χρηματικά ποσά για την σχεδίαση της ασφάλειας, ώστε να μπορούν να αντιμετωπίζουν επιθέσεις με αποτελεσματικότητα χωρίς όμως να είναι αυτό πάντα εφικτό. Οι εγκληματίες που δραστηριοποιούνται στο χώρο στη σύγχρονη εποχή, στοχεύουν κυρίως στην κλοπή προσωπικών δεδομένων από ιστοσελίδες κοινωνικής δικτύωσης αλλά και κλοπή στοιχείων που αφορούν τραπεζικές συναλλαγές μέσω του διαδικτύου. Τα είδη των επιθέσεων είναι πολύ περισσότερα και θα δοθεί ιδιαίτερη προσοχή σε επόμενη παράγραφο.

Οι απειλές στο διαδίκτυο θα συνεχίσουν να είναι ένα μείζον θέμα για όσο το διαδίκτυο χρησιμοποιείται ως μέσον για τη διατήρηση, πρόσβαση και μετάδοση

πληροφοριών. Έχουν επινοηθεί διάφοροι αμυντικοί μηχανισμοί αντιμετώπισης τέτοιων απειλών. Μερικοί από τους πιο διαδεδομένους είναι οι εξής[4]:

- Συστήματα κρυπτογράφησης: Η κρυπτογραφία χρησιμοποιείται ευρέως και είναι ένα εργαλείο το οποίο μετατρέπει τις πληροφορίες σε δεδομένα που δεν βγάζουν νόημα για κανέναν παρά μόνο για αυτούς που έχουν το κλειδί αποκρυπτογράφησης προκειμένου να εκτελέσουν την αντίστροφη διαδικασία μετατροπής των πληροφοριών από άχρηστα δεδομένα σε δεδομένα που βγάζουν νόημα.
- Τοίχος προστασίας: Το τοίχος προστασίας είναι ένας αμυντικός μηχανισμός ενάντια στους εισβολείς που θέλουν να εισχωρήσουν σε ένα σύστημα μέσω του δικτύου χωρίς εξουσιοδότηση. Τέτοιοι αμυντικοί μηχανισμοί μπορούν να εφαρμοστούν και στο software και στο hardware. Ένα τοίχος προστασίας κατά βάση φιλτράρει την κίνηση στο δίκτυο και αποκλείει αυτή που θεωρεί επικίνδυνη.
- Λογισμικό ανίχνευσης κακόβουλου λογισμικού: Υπάρχουν διάφοροι τύποι κακόβουλου λογισμικού. Έχουν αναπτυχθεί λοιπόν λογισμικά τα οποία μπορούν να ανιχνεύσουν και να αποκλείσουν κακόβουλα λογισμικά από το να τρέξουν σε ένα σύστημα.
- Βιομετρικός Έλεγχος στις Έξυπνες Συσκευές: Πλέον για την πιστοποίηση ότι η χρήση των υπηρεσιών που προσφέρει μια έξυπνη συσκευή θα γίνει από άτομο που έχει εξουσιοδότηση, έχουν αναπτυχθεί βιομετρικοί μηχανισμοί πιστοποίησης της ταυτότητας. Τέτοιοι μηχανισμοί είναι η αναγνώριση δακτυλικών αποτυπωμάτων και η αναγνώριση της ίριδας των ματιών. Ο βιομετρικός έλεγχος δεν είναι καινούρια τεχνολογία μιας και χρησιμοποιείται εδώ και κάποιες δεκαετίες από τράπεζες και άλλες επιχειρήσεις με πολλά επίπεδα ασφαλείας. Παρόλα ταύτα, θεωρείται καινοτομία το γεγονός ότι τέτοια συστήματα έχουν βελτιστοποιηθεί για να χρησιμοποιηθούν σε φορητούς υπολογιστές και έξυπνες συσκευές που προορίζονται για ιδιωτική χρήση.

Τέλος, η ασφάλεια είναι ένα ζωτικής σημασίας θέμα στον χώρο των ηλεκτρονικών υπολογιστών και του διαδικτύου. Υπάρχουν πολλές οπτικές γωνίες όσον αφορά τις πολιτικές ασφαλείας. Το κλειδί για την ανάπτυξη ενός αποτελεσματικού συστήματος ασφαλείας είναι η καταγραφή των απαιτήσεων ασφαλείας συναρτήσει του είδους των υπηρεσιών που προσφέρονται.

1.2 Κακόβουλο λογισμικό

1.2.1 Ορισμός – Συμπτώματα μολυσμένου συστήματος

Με τον όρο κακόβουλο λογισμικό(malicious software ή malware) καλείται οποιοδήποτε πρόγραμμα που είναι γραμμένο για υπολογιστή και είναι επιζήμιο για τα υπολογιστικά συστήματα. Υπάρχουν αρκετές υποκατηγορίες κακόβουλου λογισμικού. Τέτοιου είδους προγράμματα επιδιώκουν να εισβάλουν, να καταστρέψουν, να απενεργοποιήσουν υπολογιστικά συστήματα, δίκτυα, προσωπικούς υπολογιστές, φορητούς υπολογιστές και έξυπνα τηλέφωνα. Αυτό γίνεται συχνά αποκτώντας μερικό έλεγχο πάνω σε αυτές τις συσκευές ή τα δίκτυα και παραποιώντας τη λειτουργία τους. Ο απώτερος σκοπός των προγραμματιστών που γράφουν κακόβουλο λογισμικό είναι συνήθως το χρηματικό κέρδος. Είναι δυνατό να γίνει κλοπή, διαγραφή, καταστροφή, κρυπτογράφηση αποθηκευμένων πληροφοριών, η κατασκοπία του κατόχου μιας προσβεβλημένης συσκευής αλλά και η αλλαγή βασικών λειτουργιών ενός συστήματος προς όφελος άλλων μερών.

Υπάρχουν πολλοί τρόποι με τους οποίους μπορεί να αντιληφθεί κανείς ότι το σύστημα του έχει προσβληθεί από κακόβουλο λογισμικό. Κάποια από τα συμπτώματα ενός μολυσμένου συστήματος είναι φανερά και κάποια όχι. Παρακάτω μπορείτε να δείτε μερικές από τις ενδείξεις που παραπέμπουν σε μολυσμένο σύστημα[5].

- Η απόκριση του συστήματος έχει γίνει πιο αργή από το συνηθισμένο είτε ο χρήστης χρησιμοποιεί το διαδίκτυο μέσω ενός web browser είτε χρησιμοποιεί τοπικά εγκατεστημένες εφαρμογές

- Πολλές διαφημίσεις και αναδυόμενα παράθυρα κατά τη διάρκεια εργασιών που δεν έπρεπε να εμφανίζονται.
- Πολλές ανεξήγητες καταρρεύσεις του συστήματος, παγώματα, συχνή εμφάνιση μπλε οθόνης(blue screen of death) συνοδευόμενη από μήνυμα για κάποιο σοβαρό σφάλμα του λειτουργικού συστήματος
- Περίεργη αύξηση ή μείωση του αποθηκευτικού χώρου του σκληρού δίσκου
- Αυξημένη δραστηριότητα του συστήματος στο διαδίκτυο
- Η αυξημένη χρήση των πόρων του συστήματος είναι σημάδι ότι κάποιο κακόβουλο λογισμικό τρέχει στο παρασκήνιο
- Η αρχική σελίδα ενός web browser έχει αλλάξει χωρίς να το έχει κάνει ο χρήστης
- Εμφάνιση νέων γραμμών εργαλειών στον web browser χωρίς να τις έχει προσθέσει ο χρήστης
- Άρνηση εξυπηρέτησης όταν πρόκειται για σύστημα επιχείρησης που προσφέρει διαδικτυακές υπηρεσίες

1.2.2 Πώς ένα σύστημα μολύνεται από κακόβουλο λογισμικό

Οι πιο κοινοί τρόποι μετάδοσης κακόβουλου λογισμικού είναι το διαδίκτυο και η λήψη ηλεκτρονικού ταχυδρομείου[ρεφερενς στο μαλγουερ μπαιτσ]. Το σύστημα μπορεί να μολυνθεί αν ο χρήστης επισκεφτεί μία ιστοσελίδα με κακόβουλο λογισμικό, αν κατεβάσει μολυσμένα αρχεία μουσικής ή εικόνες, αν κλικάρει μια κακόβουλη διαφήμιση αγνώστου προελεύσεως και πρακτικά από οποιοδήποτε μολυσμένο περιεχόμενο μπορεί ο χρήστης να αποθηκεύσει στον υπολογιστή του που προέρχεται από το διαδίκτυο ή μεταφέρεται στον υπολογιστή του από μια εξωτερική μονάδα αποθήκευσης περιεχομένου όπως εξωτερικοί σκληροί δίσκοι usb flash memories και άλλα. Στην τελευταία περίπτωση, η πηγή του κακού είναι και πάλι το ίντερνετ εκτός αν το μολυσμένο περιεχόμενο προέρχεται απευθείας από τον προγραμματιστή που έγραψε το κακόβουλο λογισμικό.

Τα τελευταία χρόνια παρατηρείται ότι, κακόβουλο λογισμικό μπορεί να είναι κρυμμένο και σε νόμιμα προγράμματα και εφαρμογές όταν χρησιμοποιούνται ανεπίσημες εκδόσεις αυτών και προμηθεύονται από ιστοσελίδες εκτός των εταιριών

παραγωγής του εκάστοτε προγράμματος ή εφαρμογής. Ο «κακόβουλος» κώδικας λειτουργεί στην αφάνεια και ο χρήστης τις περισσότερες φορές δεν μπορεί να αντιληφθεί ότι το σύστημά του έχει μολυνθεί πριν είναι πολύ αργά.

Ο καλύτερος τρόπος πρόληψης είναι ο ίδιος ο χρήστης. Αρχικά θα πρέπει να κατεβάζει περιεχόμενο, προγράμματα και εφαρμογές από έμπιστες πηγές και να μην να κατεβάζει αρχεία ή να ανοίγει μηνύματα ηλεκτρονικού ταχυδρομείου που φαίνονται ύποπτα. Επιπλέον, μια καλή λύση είναι η χρήση λογισμικού ανίχνευσης απειλών με δυνατότητα αποκλεισμού αυτών ή και επαναφοράς του συστήματος μετά από μια σοβαρή μόλυνση. Γενικά η ευαισθητοποίηση του κόσμου σχετικά με τον τρόπο προστασίας από ηλεκτρονικές απειλές είναι θεμιτή και ο κόσμος θα πρέπει να δίνει ιδιαίτερη σημασία σε περιεχόμενο που ως σκοπό έχει την ενημέρωση σχετικά με το διαδίκτυο και το ηλεκτρονικό έγκλημα.

1.2.3 Μορφές και είδη κακόβουλου λογισμικού

Πριν προχωρήσουμε στην παράθεση των σημαντικότερων ειδών κακόβουλου λογισμικού, κρίνεται απαραίτητο σε αυτό το σημείο, να ξεκαθαριστεί ότι τα είδη του κακόβουλου λογισμικού δε θα πρέπει να συνδέονται με το εκάστοτε λειτουργικό σύστημα που τρέχει μία συσκευή. Πρέπει, να γίνει κατανοητό ότι τα κακόβουλα λογισμικά δεν γράφονται μόνο για συσκευές που τρέχουν το λειτουργικό σύστημα Windows. Τα Mac συστήματα είναι εξίσου ευάλωτα σε απειλές και έχουν ευπάθειες. Το ίδιο ισχύει και για τα συστήματα που τρέχουν κάποια διανομή του λειτουργικού συστήματος linux. Βέβαια, δε βρίσκονται στο απυρόβλητο και οι κινητές συσκευές που τρέχουν android ή ios. Οι έξυπνες συσκευές προσφέρουν πλέον στους χρήστες ιδίου επιπέδου υπηρεσίες με τους υπολογιστές σε πολλούς τομείς και η χρήση τους αποτελεί ανοιχτή πρόσκληση για τους προγραμματιστές κακόβουλου λογισμικού προκειμένου να αναπτύξουν τις εγκληματικές δραστηριότητές τους, ειδικότερα όταν οι χρήστες δεν έχουν μεγάλη εμπειρία και δεν είναι υποψιασμένοι, με αποτέλεσμα άθελά τους γίνονται θύματα κάποιας απάτης που σχετίζεται με κακόβουλο λογισμικό.

Κάθε είδος κακόβουλου λογισμικού από τα παρακάτω μπορεί να παραμετροποιηθεί λοιπόν, για να τρέξει σε οποιοδήποτε λειτουργικό σύστημα και κατ' επέκταση σε οποιαδήποτε συσκευή. Τα σημαντικότερα είδη κακόβουλου λογισμικού που μαστίζουν τον κυβερνοχώρο είναι τα εξής[5]:

- ***Adware***: Τα adwares είναι ανεπιθύμητο λογισμικό και σχεδιασμένο να προωθεί διαφημίσεις στην οθόνη μας, συχνά με εκνευριστικό τρόπο. Πολλές φορές τέτοιο λογισμικό είναι κρυμμένο πίσω από αναδυόμενες διαφημίσεις στο διαδίκτυο και μπορεί να εγκατασταθεί σε ένα σύστημα αν ο χρήστης επιλέξει τη διαφήμιση για να τη δει. Επιπλέον, μπορεί να είναι κρυμμένο στα αρχεία εγκατάστασης ενός εγκεκριμένου προγράμματος ή εφαρμογής. Αυτό συμβαίνει όταν ο χρήστης κατεβάζει λογισμικό από ανεπιβεβαίωτες πηγές. Μαζί με την εγκατάσταση του λογισμικού που επιθυμεί, παράλληλα εγκαθιστά χωρίς να το γνωρίζει και το καλά κρυμμένο adware.
- ***Spyware***: Είναι λογισμικό που παρακολουθεί μυστικά τις δραστηριότητες του χρήστη χωρίς εξουσιοδότηση και στέλνει αναφορές για αυτές τις δραστηριότητες στο άτομο που έγραψε το λογισμικό.
- ***Virus(ιός)***: Λογισμικό που εγκαθίσταται σε ένα σύστημα με παρόμοιο τρόπο όπως τα adwares. Βασικός στόχος του είναι να αναπαράγεται μέσα στο σύστημα, να μολύνει στοχευμένα ή μη πολλά αρχεία και προγράμματα του συστήματος.
- ***Worms***: Λογισμικό παρόμοιο με τους ιούς με τη διαφορά ότι αναπαράγεται σε πολλούς υπολογιστές χρησιμοποιώντας ένα δίκτυο με σκοπό να καταστρέφει αρχεία και σημαντικά δεδομένα.
- ***Trojan***: Ή αλλιώς Trojan Horse, είναι ένα από τα πιο επικίνδυνα είδη malware. Εμφανίζεται συνήθως σαν ένα χρήσιμο πρόγραμμα ή εφαρμογή για να ξεγελάσει τους χρήστες να το βάλουν στο σύστημά τους. Χρησιμοποιείται συνήθως για να δώσει απομακρυσμένη πρόσβαση στο σύστημα σε αυτούς που το έγραψαν, με απώτερο σκοπό την κλοπή προσωπικών δεδομένων ή για την εγκατάσταση άλλων ειδών malware.
- ***Ransomware***: Είναι λογισμικό το οποίο αφαιρεί την πρόσβαση του χρήστη από σημαντικές πληροφορίες ή πολλές φορές τον κλειδώνει εντελώς έξω από το σύστημά του. Τα άτομα πίσω από τέτοια λογισμικά συνήθως εκβιάζουν τους χρήστες για να τους δώσουν πίσω την πρόσβαση στο σύστημά τους έναντι χρηματικής αμοιβής, με τρόπο που δε μπορεί κάποιος να ανιχνεύσει το άκρο που

λαμβάνει την χρηματική κατάθεση. Η αντιμετώπιση τέτοιου λογισμικού είναι πάρα πολύ δύσκολη έως αδύνατη.

- ***Keylogger***: Αυτό το λογισμικό καταγράφει όλη την πληροφορία που εισάγει ο χρήστης μέσω του πληκτρολογίου και την στέλνει σε αυτόν που παρακολουθεί. Με αυτό τον τρόπο ο επιτιθέμενος μπορεί να μάθει κωδικούς ασφαλείας, ή να αποκτήσει πρόσβαση σε ευαίσθητο περιεχόμενο όπως συζητήσεις που εκτελεί ο χρήστης μέσω μηνυμάτων.
- ***Rootkit***: Λογισμικό που δίνει δικαιώματα διαχειριστή στον επιτιθέμενο, που αφορούν το λειτουργικό σύστημα της μολυσμένης συσκευής και μένει κρυμμένο μέσα στο σύστημα.
- ***Malicious cryptomining***: Λογισμικό που εισβάλει στο σύστημα μέσω trojan. Επιτρέπει σε κάποιον να κάνει εξόρυξη κρυπτονομισμάτων χρησιμοποιώντας πόρους του υπολογιστή που έχει μολύνει. Χρησιμοποιεί τον επεξεργαστή ή την κάρτα γραφικών ενός υπολογιστή και μειώνει τη λειτουργικότητα που ο χρήστης περιμένει να έχει.

1.3 Επισκόπηση του προβλήματος

Ο κυβερνοχώρος ή αλλιώς ο κόσμος του διαδικτύου είναι ένα οικοσύστημα και περιλαμβάνει τους υπολογιστές, τους χρήστες, το διαδίκτυο, όλα τα επιμέρους δίκτυα που είναι διασυνδεδεμένα με το διαδίκτυο και όλες τις δραστηριότητες που διεκπεραιώνονται με αυτά τα εργαλεία. Ο σύγχρονος κόσμος είναι δομημένος με τέτοιο τρόπο κατά τον οποίο κάποιες από τις διαδικτυακές δραστηριότητες των ατόμων, να αποτελούν συνάμα βασικές δραστηριότητες της καθημερινότητάς τους. Αντιλαμβάνεται κανείς εύκολα, ότι οι αδυναμίες του κυβερνοχώρου αλλά και των ανθρώπων προσφέρουν εύπορο έδαφος για την ευδοκίμηση μίας μεγάλης γκάμας εγκλημάτων αυτών που συνθέτουν τον όρο έγκλημα στον κυβερνοχώρο(cybercrime).

Όπως έχει αναφερθεί σε προηγούμενη παράγραφο το έγκλημα στον κυβερνοχώρο έρχεται σε πολλές μορφές και είναι συνήθως συνδεδεμένο με κακόβουλο ή παράνομο λογισμικό. Η δραστηριότητα των εγκληματιών στον κυβερνοχώρο σχετίζεται

κατά μεγάλο ποσοστό με οικονομικά εγκλήματα άμεσα ή έμμεσα. Ωστόσο, δεν είναι ο μόνος τομέας στον οποίο επικεντρώνονται οι δραστηριότητες αυτές.

Τα εγκλήματα αυτά ταξινομούνται ως εξής[3]:

- ***Financial Frauds(Εγκλήματα οικονομικής απάτης)***: Στηρίζονται στην ηλεκτρονική απάτη και κατ΄επέκταση στην τροποποίηση περιεχομένου με μη εξουσιοδοτημένο τρόπο. Απαιτεί εμπειρία πάνω στον τομέα που επιχειρείται η απάτη. Οδηγεί σε ενέργειες ή καθόλου ενέργειες λόγω παραποιημένων δεδομένων από άτομα σε σημαντικές θέσεις οι οποίες επιφέρουν οικονομικές ζημιές. Επιπλέον, τέτοια εγκλήματα μπορεί να προέλθουν από απάτες σε τράπεζες, κλοπή ηλεκτρονικής ταυτότητας και κάρτας και εκβιασμό.
- ***Cyberterrorism(Κυβερνοτρομοκρατία)***: Συμπεριλαμβάνονται όλες οι ενέργειες οι οποίες οδηγούν κυβερνήσεις ή οργανισμούς στο να παρεκκλίνουν από πολιτικούς ή κοινωνικούς στόχους με αποτέλεσμα την πρόκληση φόβου. Ξεκινούν με μια επίθεση από υπολογιστή προς άλλους υπολογιστές, δίκτυα ή τις πληροφορίες που είναι αποθηκευμένες σε αυτά.
- ***Cyberextortion(Εκβιασμός)***: Τέτοια εγκλήματα αφορούν επιθέσεις με κατανεμημένο τρόπο σε έναν ιστότοπο, διακομιστή ηλεκτρονικού ταχυδρομείου με σκοπό την διαρκή άρνηση υπηρεσίας. Αυτοί που επιτίθενται σε τέτοια συστήματα(Hackers-Χάκερς) αναστέλλουν την ικανότητα των συστημάτων να προσφέρουν υπηρεσίες, μέχρις ότου ικανοποιηθούν οι απαιτήσεις τους για πληρωμές.
- ***Σωματεμπορία/Σεξουαλικός Εκβιασμός(Cybersex trafficking)***: Αφορά την απαγωγή, απειλή, μεταφορά ατόμων και εξαναγκασμό αυτών σε σεξουαλικές πράξεις μπροστά σε κάμερα με σκοπό το διαμοιρασμό του βιντεοσκοπημένου υλικού. Προκειμένου οι δράστες να προσεγγίσουν τα ανυποψίαστα θύματά τους, χρησιμοποιούν ιστότοπους και άλλες διαδικτυακές πλατφόρμες ενώ διεκπεραιώνουν τις πληρωμές τους με διαδικτυακά συστήματα πληρωμών και κρυπτονομίσματα που δεν είναι ανιχνεύσιμα.

Τα εγκλήματα αυτά διαπράττονται συνήθως συντονισμένα από ομάδες εγκληματιών. Οι δράστες εκμεταλλεύονται τη ραγδαία ανάπτυξη της τεχνολογίας για να

επινοήσουν εξυπνότερα εγκλήματα. Γι’ αυτό το λόγο ο κόσμος είναι απροετοίμαστος για την καταπολέμηση αυτών των εγκλημάτων. Οι δράστες αξιοποιούν την τεχνολογική άγνοια των χρηστών αλλά και τις αδυναμίες τους. Η ιχνηλάτηση των υπολογιστώ/εργαλείων των εγκληματιών είναι συνήθως αδύνατη και οι αρχές περιορίζονται στη διαχείριση των ζημιών που προκαλούνται. Η επιτυχία των εγκλημάτων στηρίζεται και στην επινοητικότητα των εγκληματιών ενώ η αντιμετώπισή τους είναι εξ’ αρχής δύσκολη διότι απαιτεί την ενημέρωση και συνεργασία πολλών κυβερνήσεων αφού τα εγκλήματα είναι και διασυνοριακά εκτός από εγχώρια.

1.4 Καθολική Βάση Δεδομένων Κακόβουλου Λογισμικού

Το διαδίκτυο επιτρέπει τον καθορισμό στόχων από διάφορες τοποθεσίες του πλανήτη ενώ ταυτόχρονα μπορεί να μεγενθύνει την κλίμακα των βλαβών που μπορεί να προκληθούν όπως για παράδειγμα την στοχοποίηση περισσοτέρων από ένα άτομα ή έναν οργανισμό με κάθε επίθεση.

Υπάρχουν κάποιοι ενδεδειγμένοι τρόποι σύμφωνα με τους οποίους ξεκινά μια έρευνα για κάποιο ηλεκτρονικό έγκλημα. Συνήθως τέτοιες έρευνες ξεκινούν με ένα ίχνος μιας διεύθυνσης IP(internet protocol) ως στοιχείο ωστόσο, δεν είναι πάντα καλή βάση ώστε να εξιχνιαστεί ένα έγκλημα. Από την άλλη πλευρά είναι σημαντικό να είναι ευρέως γνωστό ό,τι μια συγκεκριμένη IP κρύβει κινδύνους. Γενικώς, οι μέθοδοι της αστυνομικές εγκληματικότητας συνεχώς βελτιώνονται είτε από κλειστές αστυνομικές ομάδες, είτε από ινστιτούτα που ασχολούνται με την εγκληματικότητα στον κυβερνοχώρο και την κυβερνοασφάλεια, είτε από κρατικούς φορείς. Παρόλου που έχουν θεσπιστεί διεθνείς συνεργασίες με στόχο την συνεχή αναβάθμιση της κυβερνοασφάλειας, υπάρχουν ακόμα κωλύματα τα οποία είναι δύσκολο να ξεπεραστούν για διαφόρους λόγους που δεν κρίνεται απαραίτητο να αναλυθούν όλοι εκτενώς σε αυτή την εργασία.

Κατ’ αρχάς όπως είναι φυσικό, η έρευνα πάνω στο θέμα της κυβερνοασφάλειας απαιτεί πολλά χρήματα και μεγάλο και εκπαιδευμένο ανθρώπινο δυναμικό. Κάθε κράτος που σέβεται στο ελάχιστο το λαό και τις υποδομές του, φροντίζει να έχει ένα τμήμα έρευνας και αντιμετώπισης ηλεκτρονικού εγκλήματος είτε ως αυτόνομο οργανισμό είτε υπό την αιγίδα της αστυνομίας. Η έρευνα όμως αυτή είναι πολυεπίπεδη και απαιτούνται ιδιαιτέρως μεγάλα χρηματικά κονδύλια. Οι εύπορες ανεπτυγμένες χώρες σαφώς έχουν τη δυνατότητα, να επενδύουν χρήματα για αυτό το σκοπό. Δε συμβαίνει όμως το ίδιο και με τις αναπτυσσόμενες χώρες. Αυτό έχει ως αποτέλεσμα, την ύπαρξη κρατών των οποίων η προσπάθεια για πρόληψη και αντιμετώπιση ηλεκτρονικών εγκλημάτων είναι φτωχή. Οι εγκληματίες στον κυβερνοχώρο χρησιμοποιούν λοιπόν αναπτυσσόμενες χώρες στις οποίες η νομοθεσία είναι αδύναμη λόγω της φτωχής γνώσης σε αυτό τον τομέα, ώστε να μπορούν ανενόχλητοι να δρουν, να μένουν απαρατήρητοι και ατιμώρητοι. Παρόλα αυτά, επειδή όπως έχει αναφερθεί ήδη, οι επιθέσεις μπορεί κάλλιστα να είναι διασυνοριακές, επιφέρουν ζημιές σε οργανισμούς και άτομα εκτός των χωρών δράσης τους χωρίς τη δυνατότητα τιμωρίας από την αντίπερα όχθη.

Μια αλυσίδα για να παραμείνει ισχυρή, πρέπει όλοι οι κρίκοι να είναι δυνατοί και να κρατούν την αλυσίδα ενωμένη. Κατ’ αναλογία, κάτι παρόμοιο πρέπει να ισχύει και με την νοητή αλυσίδα κυβερνοασφάλειας που συνθέτουν τα κράτη του κόσμου. Το θέμα της διπλωματικής εργασίας αυτής πηγάζει από αυτή την ιδέα. Δηλαδή, τη δημιουργία ενός εργαλείου που θα εξοικονομήσει χρηματικά κονδύλια από ένα επίπεδο της έρευνας τα οποία θα μπορούν να διατεθούν σε άλλα επίπεδα της έρευνας, γεγονός που θα επιτρέψει στις αναπτυσσόμενες χώρες να παρακάμψουν μερικά κατώτερα επίπεδα της έρευνας επενδύοντας στα ανώτερα. Μία καθολική βάση δεδομένων που θα περιέχει domain names(ηλεκτρονικές διευθύνσεις) και τις αντίστοιχες IP διευθύνσεις, οι οποίες κρύβουν κακόβουλο λογισμικό. Μια τέτοια βάση θα μπορεί να διατεθεί ελεύθερα σε όλες τις ενδιαφερόμενες πλευρές παρακάμπτοντας οποιαδήποτε προβλήματα που εμποδίζουν την διεθνή συνεργασία που αφορά την κυβερνοασφάλεια.

Το πρώτο βήμα για τη δημιουργία μιας τέτοιας βάσης απαιτεί αρχικώς τη διάθεση του ήδη υπάρχοντος ερευνητικού υλικού από αντίστοιχους φορείς που προσπαθούν να σημαδέψουν ως κακόβουλους ιστότοπους και διευθύνσεις IP. Κατά τη διάρκεια της αναζήτησης πηγών για άντληση πληροφοριών και εισαγωγή τους στη βάση δεδομένων έγινε γνωστό ότι υπάρχουν ήδη συντονισμένες προσπάθειες προς αυτή την

κατεύθυνση. Κάποιοι φορείς επιλέγουν να διαθέτουν τη γνώση τους ελεύθερα και κάποιοι υπό προϋποθέσεις και περιορισμούς. Η εργασία αυτή στηρίχθηκε σε τέτοιους φορείς.

ΚΕΦΑΛΑΙΟ 2

ΕΡΓΑΛΕΙΑ ΚΑΙ ΜΕΘΟΔΟΙ ΠΟΥ ΣΥΝΘΈΤΟΥΝ ΤΗΝ ΕΠΙΘΥΜΗΤΗ ΠΛΗΡΟΦΟΡΙΑ ΚΑΙ ΟΙ ΤΡΟΠΟΙ ΑΝΤΛΗΣΗΣ ΤΗΣ

2.1 Εισαγωγή

Πριν ξεκινήσει το τεχνικό μέρος της αυτής της εργασίας του οποίου η ανάλυση θα γίνει στο επόμενο κεφάλαιο, κρίνεται απαραίτητο να αναλυθούν κάποιες τεχνολογίες οι οποίες έπρεπε να κατανοηθούν πριν τη χρήση τους. Στις επόμενες παραγράφους του κεφαλαίου, ο αναγνώστης θα διαβάσει για τη σουίτα πρωτοκόλλου του διαδικτύου TCP/IP το οποίο περιλαμβάνει το σύνολο των πρωτοκόλλων επικοινωνίας που χρησιμοποιούνται στο διαδίκτυο και κάποιες εισαγωγικές πληροφορίες για τις βάσεις δεδομένων και γιατί είναι βασικό συστατικό αυτής της εργασίας. Επιπλέον, θα δοθούν οι ορισμοί τα δομημένα δεδομένα και μη-δομημένα δεδομένα, δύο κατηγορίες μορφοποίησης δεδομένων οι οποίες χρησιμοποιούνται παράλληλα με τις μεθόδους άντλησης πληροφοριών με API και Web Scraping αντίστοιχα. Κατά παρόμοιο τρόπο, στις δύο τελευταίες παραγράφους του κεφαλαίου για τους σκοπούς της καλύτερης κατανόησης του τεχνικού μέρους θα αναφερθούν πληροφορίες σχετικά με τα API's και το Web Scraping.

2.2 TCP/IP

Η σουίτα πρωτοκόλλου του διαδικτύου(internet protocol suite) είναι το μοντέλο και το σύνολο των πρωτοκόλλων επικοινωνίας που χρησιμοποιούνται στο διαδίκτυο και σε παραπλήσια δίκτυα υπολογιστών[6]. Είναι ευρέως γνωστό ως TCP/IP επειδή τα βασικά πρωτόκολλα που συνθέτουν τη σουίτα είναι το πρωτόκολλο ελέγχου μετάδοσης (TCP) και το πρωτόκολλο διαδικτύου (IP-internet protocol). Στις αρχές της ανάπτυξής του

χρηματοδοτήθηκε από το υπουργείο άμυνας των ΗΠΑ και ήταν γνωστό με την ονομασία μοντέλο υπουργείου άμυνας(DOD-Department Of Defense).

Το TCP/IP εξασφαλίζει επικοινωνία δεδομένων από άκρο σε άκρο, τρόπος που καθορίζει το πακετάρισμα των δεδομένων, τη μετάδοση τους, το δρομολόγιο που ακολουθούν, σε ποιον απευθύνονται και πώς λαμβάνονται. Αυτή η λειτουργικότητα είναι οργανωμένη σε τέσσερα επίπεδα. Τα επίπεδα αυτά ιεραρχικά από το κατώτερο προς το ανώτερο είναι:

- ***Link layer(επίπεδο συνδέσμου)***: Αποτελεί την ομάδα μεθόδων και πρωτοκόλλων επικοινωνίας που περιορίζονται στη σύνδεση με την οποία συνδέεται ο φυσικός υπολογιστής. Είναι το φυσικό και το λογικό συστατικό του δικτύου που χρησιμοποιείται για τη διασύνδεση κεντρικών υπολογιστών ή κόμβων στο δίκτυο και επίσης καθορίζει την τοπολογία του τοπικού δικτύου υπολογιστών όπου δεν απαιτούνται ενδιάμεσα δρομολογητές(routers)[7].

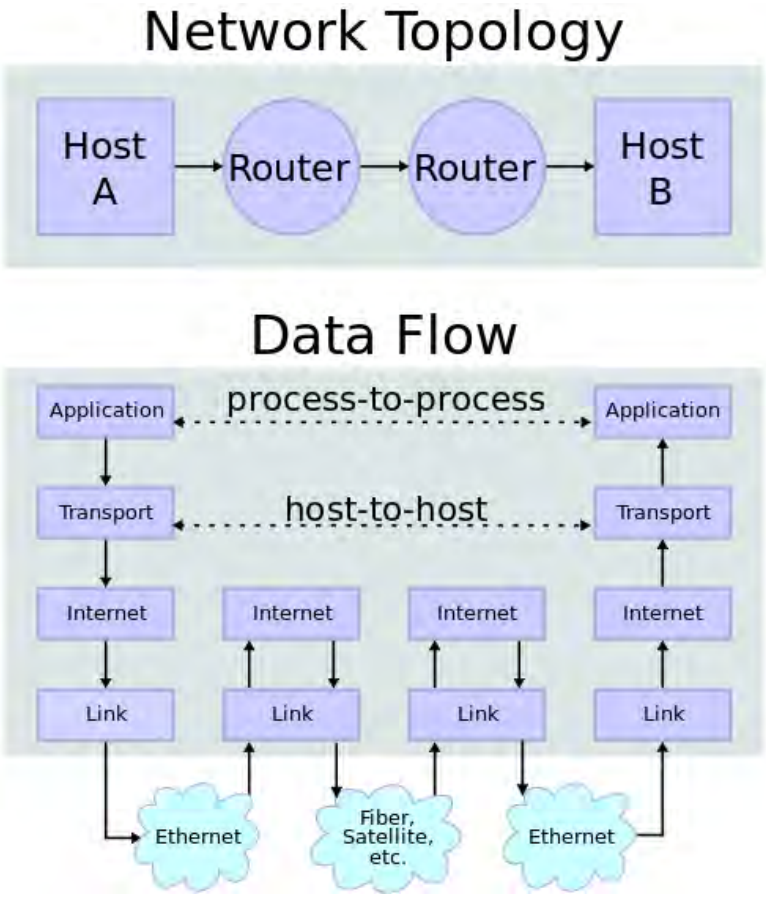
- ***Internet Layer(επίπεδο του διαδικτύου)***: Είναι μια ομάδα μεθόδων διασύνδεσης, πρωτοκόλλων και προδιαγραφών στο TCP/IP που χρησιμοποιούνται για τη μεταφορά πακέτων δικτύου από τον αρχικό κεντρικό υπολογιστή εκτός των ορίων του δικτύου σε έναν άλλο υπολογιστή/προορισμού ο οποίος καθορίζεται από μία διεύθυνση IP. Με το internet layer διεκπεραιώνονται διαδικτυακές εργασίες δλδ η διασύνδεση πολλών υπολογιστών μεταξύ τους μέσω πυλών(gateways). Το κύριο πρωτόκολλο σε αυτό το επίπεδο είναι το πρωτόκολλο διαδικτύου το οποίο καθορίζει τις διευθύνσεις IP. Μεταφέρει δεδομένα στον επόμενο κεντρικό υπολογιστή που λειτουργεί ως δρομολογητής, ο οποίος δύναται να συνδεθεί με ένα δίκτυο κοντύτερα στον τελικό προορισμό των δεδομένων[8].

- ***Transport layer(επίπεδο μεταφοράς)***: Αποτελεί μια εννοιολογική ταξινόμηση μεθόδων στην αρχιτεκτονική πρωτοκόλλων στο TCP/IP. Τα πρωτόκολλα του επιπέδου αυτού παρέχουν host-to-host υπηρεσίες

επικοινωνίας για εφαρμογές είτε εντός του τοπικού δικτύου είτε προς απομακρυσμένα δίκτυα που συνδέονται με δρομολογητές. Το πιο γνωστό πρωτόκολλο μεταφοράς του της σουίτας πρωτοκόλλου διαδικτύου είναι το TCP. Χρησιμοποιείται για μεταδόσεις με γνώμονα τη σύνδεση και παρέχει αξιοπιστία στη μετάδοση δεδομένων, ενώ υπάρχει και το πρωτόκολλο UDP που χρησιμοποιείται για μετάδοση δεδομένων χωρίς σύνδεση σε περιπτώσεις απλούστερων μεταδόσεων[9].

- ***Application Layer(επίπεδο εφαρμογής)***: Καθορίζει τα κοινά πρωτόκολλα επικοινωνίας και τις μεθόδους διεπαφής που χρησιμοποιούνται από κεντρικούς υπολογιστές σε ένα δίκτυο επικοινωνιών. Καθιερώνει την επικοινωνία και εξαρτάται από τα υποκείμενα πρωτόκολλα του transport layer για τη δημιουργία καναλιών μεταφοράς δεδομένων από έναν κεντρικό υπολογιστή σε έναν άλλο κεντρικό υπολογιστή και διαχειρίζεται την ανταλλαγή δεδομένων στο μοντέλο δικτύωσης client-server(πελάτη-διακομιστή) ή peer to peer(αρχιτεκτονική κατανεμημένων εφαρμογών). Είναι το επίπεδο εντός του οποίου οι εφαρμογές δημιουργούν δεδομένα και τα μεταδίδουν σε άλλες εφαρμογές του ίδιου ή σε άλλον κεντρικό υπολογιστή. Οι εφαρμογές χρησιμοποιούν τις υπηρεσίες που παρέχονται από τα κατώτερα στρώματα του TCP/IP[10].

Στην παρακάτω εικόνα φαίνεται η τοπολογία του δικτύου δύο υπολογιστών σύμφωνα με την host-to-host επικοινωνία, όπου δύο κεντρικοί υπολογιστές συνδέονται πάνω από το δίκτυο μέσω άλλων κεντρικών υπολογιστών οι οποίοι χρησιμοποιούντα ως δρομολογητές. Επιπλέον, στην εικόνα φαίνεται και η ροή των δεδομένων κατά τη διάρκεια της επικοινωνίας όπως αυτή γίνεται σύμφωνα με τα επίπεδα του TCP/IP.



Πηγή: wikipedia

Το 1985 ξεκίνησε η αυξανόμενη εμπορική χρήση του TCP/IP στη βιομηχανία των υπολογιστών αφού πρώτα είχε υιοθετηθεί από το υπουργείο άμυνας των ΗΠΑ ως πρότυπο για τη στρατιωτική του δικτύωση. Η περεταίρω εξάπλωσή του άρχισε των Ιούνιο του 1989 και εταιρίες όπως η IBM το συμπεριέλαβαν σε εμπορικές εκδόσεις λογισμικού με αποκορύφωμα τη Microsoft και το λειτουργικό σύστημα Windows 95. Είναι σημαντικό πρωτόκολλο και μας ενδιαφέρει σε αυτή την εργασία διότι εισάγει την έννοια της διευθυνσιοδότησης των υπολογιστών μέσω των διευθύνσεων IP, ζωτικής σημασίας συστατικό των εργασιών που γίνονται με τον κώδικα που θα παρατεθούν σε επόμενο κεφάλαιο.

2.3 Βάσεις δεδομένων

Στη σύγχρονη εποχή η έννοια της πληροφορίας είναι πολυδιάστατη και εξετάζεται σε πολλά πλαίσια όπως αυτά της επικοινωνίας, του ελέγχου, της γνώσης, της σημασίας, της εντροπίας, των δεδομένων και άλλων. Για τις επιχειρήσεις και όχι μόνο η σημασία της πληροφορίας και η μορφή της έχει κεφαλαιώδης αξία. Και όταν αναφερόμαστε σε τέτοιες πληροφορίες συνήθως γίνεται με την έννοια των δεδομένων. Η αποτελεσματική διαχείριση των δεδομένων, η επεξεργασία τους και η εξαγωγή νέων πληροφοριών από αυτά είναι μία δύσκολη και επίπονη διαδικασία, ενώ ο όγκος της πληροφορίας που συγκεντρώνεται αποτελεί ένα μεγάλο βάρος η αντιμετώπιση του οποίου κοστίζει πολύ και επιβαρύνει ποικιλοτρόπως μια επιχείρηση, έναν οργανισμό και φυσικά τον άνθρωπο που κλήθηκε εξ' αρχής στο καθήκον της διαχείρισης. Για την αντιμετώπιση των παραπάνω δυσκολιών αλλά και την καλύτερη και βέλτιστη εξυπηρέτηση που μπορεί να προσφέρει η πληροφορία επινοήθηκαν οι βάσεις δεδομένων.

Η **βάση δεδομένων(database)** είναι μια οργανωμένη συλλογή σχετικών δεδομένων, τα οποία αποθηκεύονται και προσπελάσονται ηλεκτρονικά από ένα σύστημα υπολογιστή και προσδιορίζει τον τρόπο οργάνωσής τους[11]. Ανάλογα με το είδος των δεδομένων μια βάση δεδομένων μπορεί να είναι απλή αλλά και πολύ περίπλοκη ειδικά όταν έχουν να κάνουν με δεδομένα που αφορούν δραστηριότητες ενός οργανισμού ή μιας επιχείρησης. Για αυτό το λόγο αναπτύσσονται χρησιμοποιώντας τεχνικές σχεδιασμού και μοντελοποίησης.

Το **συστημα διαχείρησης βάσεων δεδομένων(DBMS)** είναι το λογισμικό που αλληλεπιδρά με τους τελικούς χρήστες, εφαρμογές αλλά και την ίδια τη βάση δεδομένων για τη διαχείριση και την ανάλυση των δεδομένων[12]. Επιτρέπει στους χρήστες να έχουν πρόσβαση σε όλα τα δεδομένα που περιέχονται στη βάση. Προσφέρουν λειτουργίες που επιτρέπουν τον **ορισμό των δεδομένων**, δηλαδή τη δημιουργία την τροποποίηση και αφαίρεση κανόνων που ορίζουν την οργάνωσή τους, την **ενημέρωση** δηλαδή την εισαγωγή, την τροποποίηση και τη διαγραφή πραγματικών δεδομένων, την **ανάκτηση** των δεδομένων κατά την οποία τα δεδομένα μπορούν να επιστρέφονται όπως ακριβώς είναι αποθηκευμένα στη βάση ή σε μια άλλη μορφή έπειτα από τροποποίηση ή το συνδυασμό των πραγματικών δεδομένων και τέλος τη **διαχείριση**

που προβλέπει μεταξύ άλλων την παρακολούθηση των χρηστών του DBMS, την εγγραφή νέων χρηστών, την ενίσχυση της ασφάλειας των δεδομένων, τη διατήρηση της ακεραιότητας των δεδομένων και την αντιμετώπιση των ενδεχόμενων αστοχιών του συστήματος με ό,τι αυτό συνεπάγεται.

Ανάλογα με τον τρόπο που μοντελοποιεί ο χρήστης την πραγματική επιχείρηση ή οργανισμό για το οποίο θα σχεδιαστεί μια βάση δεδομένων, χρησιμοποιεί κατάλληλο DBMS για την καταχώριση των δεδομένων. Υπάρχουν πολλοί τύποι DBMS αλλά, μέχρι στιγμής το επικρατέστερο είναι το **σχεσιασκό σύστημα διαχείρισης βάσεων δεδομένων(RelationalDBMS).**

Η χρήση βάσεων δεδομένων και συστημάτων διαχείρισης βάσεων δεδομένων παρουσιάζει ένα πλήθος πλεονεκτημάτων σε σχέση με την παραδοσιακή αρχειοθέτηση και μερικά από τα σημαντικότερα είναι τα εξής[11]:

- **Διαχείριση των δεδομένων**: Αποφυγή διπλοτύπων, ρύθμιση της φυσικής καταχώτησης των δεδομένων ώστε να μεγιστοποιούνται οι επιδόσεις του συστήματος κατά την ανάκτηση καταχωρημένων πληροφοριών.
- **Ταυτόχρονη πρόσβαση και επαναφορά από βλάβη**: Το DBMS χρονοπρογραμματίζει την ταυτόχρονη πρόσβαση στα δεδομένα ώστε οι χρήστες να θεωρούν ότι μόνον έναν από αυτούς έχει πρόσβαση σε μια καταχωρημένη πληροφορία κάθε μια χρονική στιγμή της επεξεργασίας. Επιπλέον, το DBMS παρέχει εργαλεία στο χρήστη ώστε να διασφαλίζει την ασφάλεια των δεδομένων σε περίπτωση που υποστεί το σύστημα βλάβη.
- **Ασφάλεια και ακεραιότητα δεδομένων**: Όταν η προσπέλαση δεδομένων γίνεται μέσω ενός DBMS, μπορούν να επιβληθούν συγκεκριμένοι περιορισμοί που αφορούν την ακεραιότητά τους, όπως για παράδειγμα αν αναφερόμαστε σε μισθούς υπαλλήλων το DBMS ελέγχει σχετικά ώστε να μην εισαχθούν στο σύστημα πληροφορίες που δε συμβαδίζουν με την πραγματικότητα. Επίσης το DBMS μπορεί να καθορίσει επίπεδα ασφαλείας στους χρήστες ώστε αυτοί να έχουν πρόσβαση μόνο σε πληροφορίες που πρέπει σύμφωνα με το διαχειριστή του συστήματος.
- **Ταχεία πρόσβαση στα δεδομένα**: Το DBMS εφαρμόζει μία πληθώρα τεχνικών για την καταχώρηση και ανάκτηση των δεδομένων ώστε αυτά να γίνονται με παραγωγικό τρόπο. Οι τεχνικές αυτές είναι σωτήριες

ειδικότερα όταν τα δεδομένα καταχωρούνται σε εξωτερικές περιφερειακές μονάδες.

- **Μείωση χρόνου ανάπτυξης εφαρμογών**: Το DBMS υποστηρίζει πλήθος υποσυστημάτων και έτοιμων συναρτήσεων για τους χρήστες που επιθυμούν να χτίσουν εφαρμογές που αλληλεπιδρούν με τη βάση δεδομένων. Κατ' αυτό τον τρόπο εξασφαλίζεται η ασφάλεια των δεδομένων σε περίπτωση που κάποιος προγραμματιστής γράψει κακόβουλη εφαρμογή για να προκαλέσει καταστροφές ή αλλοίωση δεδομένων στη βάση, αλλά και με τις έτοιμες συναρτήσεις και τα υποσυστήματα που προσφέρει δίνει τη δυνατότητα στο χρήστη να περιορίσει τον κώδικά του μόνο στις λειτουργίες που τον ενδιαφέρουν σε επίπεδο εφαρμογής. Οι εφαρμογές που αναπτύσσονται μπορεί να μην είναι εντελώς αυτόνομες αλλά εξασφαλίζονται ως προς την αξιοπιστία και τη λειτουργικότητα όσον αφορά τα δεδομένα που αντλούν από τη βάση δεδομένων.

Η δημιουργία μιας βάσης δεδομένων για τα δεδομένα που συγκεντρώνονται για αυτή την εργασία είναι μονόδρομος αν αναλογιστεί κανείς όλα τα παραπάνω. Σε συνδυασμό με τη χρήση της γλώσσας python η ανάκτηση των δεδομένων ανά κατηγορίες για λόγους επεξεργασίας καθιστά τη βάση δεδομένων πολύτιμο εργαλείο για τους σκοπούς της εργασίας.

2.4 Unstructured Data – Structured Data

2.4.1 Unstructured Data – Μη Δομημένα Δεδομένα

Με τον όρο **μη δομημένα δεδομένα** ή **μη δομημένες πληροφορίες** εννοούμε τις πληροφορίες που δεν είναι οργανωμένες με προκαθορισμένο τρόπο ή δεν έχουν κάποιο

προκαθορισμένο μοντέλο[13]. Τέτοια δεδομένα είναι συνήθως κείμενο, το οποίο μπορεί να περιέχει αριθμούς που ορίζουν κάτι γενικό, ημερομηνίες, γεγονότα, διευθύνσεις και άλλες πληροφορίες. Υπάρχουν πολύ περιορισμένες ενδείξεις για τον τύπο των δεδομένων και τέτοια δεδομένα εμπεριέχονται σε ιστοσελίδες σε HTML. Ο όρος αυτός δεν είναι απόλυτα ακριβής διότι κατά περιπτώσεις τα δεδομένα που χαρακτηρίζονται ως μη δομημένα μπορεί να έχουν κάποια μορφή δομής η οποία παρόλα αυτά να μην είναι χρήσιμη για τις εργασίες επεξεργασίας των δεδομένων που απαιτούνται για ένα πολύ συγκεκριμένο σκοπό. Επιπλέον, υπάρχουν δομημένες πληροφορίες οι οποίες μπορεί να είναι πολύ δομημένες αλλά με τρόπους που είναι απρόβλεπτοι ή δεν ακολουθούν κάποιο σταθερό μοντέλο. Τέλος, ο όρος χαρακτηρίζεται ως ανακριβής διότι η έννοια της δομής που εμπεριέχεται στον όρο «**μη δομημένα**» δε μπορεί να οριστεί ρητά και πολλές φορές μπορεί ακόμα να υπονοείται.

Πολύ συχνά οι επιστήμονες της πληροφορικής καλούνται να αντιμετωπίσουν μη δομημένα δεδομένα και για αυτό το σκοπό, έχουν επινοήσει κάποιες τεχνικές για την αντιμετώπισης των μη δομημένων δεδομένων. Τεχνικές όπως η εξόρυξη δεδομένων(data mining), η επεξεργασία φυσικής γλώσσας και η ανάλυση κειμένου έχουν επιστρατευτεί για την ανεύρεση μοτίβων σε μη δομημένες πληροφορίες ή την ερμηνεία αυτών. Μετά από την εφαρμογή κάποιας από αυτές τις τεχνικές, προσδίδονται στα δεδομένα ετικέτες με μεταδεδομένα προκειμένου να επιτευχθεί μία μορφή δομής και η βέλτιστη δυνατή κατανόησή τους.

Σε αυτή την εργασία η αντιμετώπιση των μη δομημένων δεδομένων γίνεται με την τεχνική web scraping. Σε αυτή την περίπτωση το ζητούμενο δεν είναι η κατανόηση των πληροφοριών που κρύβουν τα δεδομένα, αφού αυτή γίνεται εύκολα αν αναλογιστεί κανείς την εμπλοκή του ανθρώπινου παράγοντα. Το είδος των πληροφοριών αναγνωρίζεται από την ανάγνωσή τους στις δύο ιστοσελίδων των οποίων οι πληροφορίες αντλούνται από κώδικα HTML.

2.4.2 Structured Data – Δομημένα δεδομένα

Ο όρος **δομημένα δεδομένα** σε αντιδιαστολή με τον όρο μη δομημένα δεδομένα περιγράφει δεδομένα που δημιουργούνται χρησιμοποιώντας ένα προκαθορισμένο σταθερό σχήμα/μοντέλο και συνήθως οργανώνονται σε μορφή πίνακα.[17]. Κάθε κελί του πίνακα περιέχει μια διακριτή τιμή που αναφέρεται σε κάποιο συγκεκριμένο χαρακτηριστικό. Το σχήμα/μοντέλο επιβάλλει τους περιορισμούς που απαιτούνται για να καταστούν τα δεδομένα συνεπή. Η σχεσιακό μοντέλο βάσεων δεδομένων είναι ένα παράδειγμα δομημένων δεδομένων. Οι πίνακες συνδέονται χρησιμοποιώντας μοναδικά αναγνωριστικά (id's) και μια γλώσσα ερωτημάτων όπως η SQL χρησιμοποιείται για την αλληλεπίδραση με τα δεδομένα.

Τα δομημένα δεδομένα είναι σημαντικά για την καλύτερη κατανόηση των πληροφοριών και όχι μόνο[14]. Πολλοί ιστότοποι παράγονται από δεδομένα που αποθηκεύονται σε βάσεις δεδομένων. Αυτά τα δεδομένα διαμορφώνονται σε κώδικα HTML και αυτό δυσκολεύει τους ιχνηλάτες του ιστοχώρου να ερμηνεύσουν σωστά τις πληροφορίες. Τα δομημένα δεδομένα όμως, επιτρέπουν στις μηχανές αναζήτησης να μέσω των ιχνηλατών να κατανοήσουν καλύτερα τις τρέχουσες πληροφορίες μιας ιστοσελίδας και κατ' επέκταση η ιστοσελίδα και ο φορέας που αντιπροσωπεύει να έχουν καλύτερο πλασάρισμα στα αποτελέσματα αναζήτησης.
Κατά κάποιο τρόπο με τα δομημένα δεδομένα ο προγραμματιστής μπορεί να «μιλάει» στις μηχανές αναζήτησης προσδιορίζοντας τις πληροφορίες που εκθέτονται σε έναν ιστότοπο, διευκολύνοντας έτσι τη μηχανή αναζήτησης στο να συμπεριλάβει στα αποτελέσματα αναζήτησης τον συγκεκριμένο ιστότοπο, όταν γίνεται αναζήτηση για τις πληροφορίες που έχουν κατανοήσει οι ιχνηλάτες. Έτσι βελτιώνεται η προβολή του περιεχομένου ενός ιστότοπου.

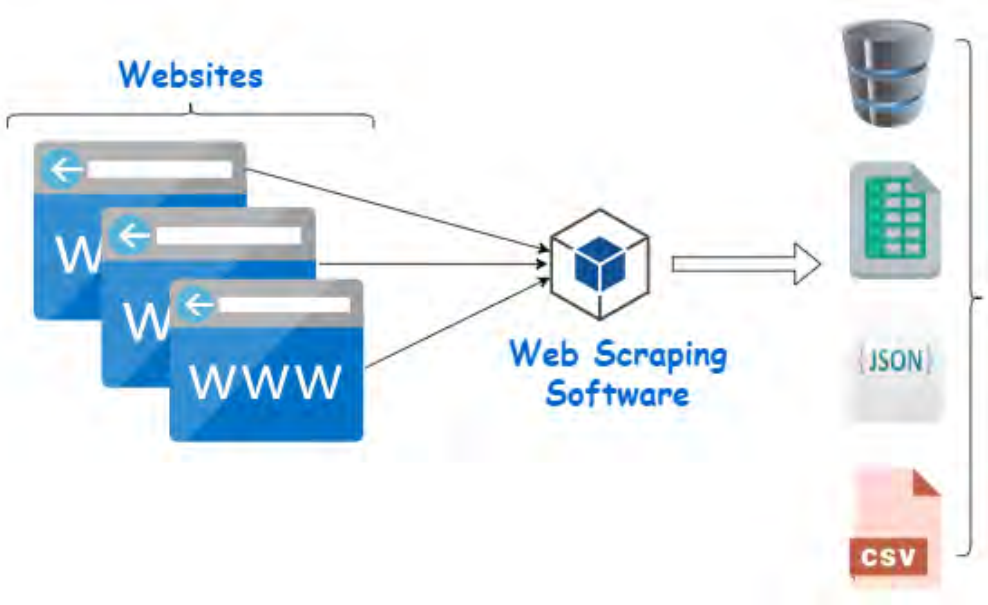
Στην εργασία αυτή εμπλέκονται δομημένα δεδομένα των οποίων οι άντληση γίνεται μέσω API που προσφέρονται από τις συγκεκριμένες ιστοσελίδες σε συνδυασμό με τον κώδικα σε python που πρέπει να γραφεί ώστε προσωρινά τα δεδομένα να αποθηκευτούν στη μνήμη του υπολογιστή πριν εισαχθούν στη βάση δεδομένων.

2.5 Web Scraping

Web scraping ή **web harvesting** ή **web data extraction** ή **screen scraping** είναι τεχνική που χρησιμοποιείται για την εξαγωγή μεγάλου όγκου δεδομένων από ιστοσελίδες[15]. Είναι μια μορφή αντιγραφής συγκεκριμένων δεδομένων που συλλέγονται από έναν ιστότοπο και αντιγράφονται σε μία βάση δεδομένων ή ένα υπολογιστικό φύλλο με σκοπό τη μελλοντική ανάκτηση ή χρήση.
Τα δεδομένα που εμφανίζονται σε ιστοσελίδες, μπορούν να προβληθούν μόνο μέσω ενός προγράμματος περιήγησης. Δεν προσφέρεται η δυνατότητα αποθήκευσης ενός αντιγράφου αυτών των δεδομένων για προσωπική χρήση. Η μόνη επιλογή που έχει ο ενδιαφερόμενος είναι να αντιγράψει με μη αυτόματο τρόπο δηλαδή να εκτελέσει μία πάρα πολύ κουραστική μακροσκελή εργασία που εν τέλει μπορεί να είναι αδύνατη αν ο όγκος των δεδομένων που τον ενδιαφέρουν είναι απαγορευτικός για οποιαδήποτε μη αυτόματη διαδικασία. Ή μπορεί να χρησιμοποιήσει την τεχνική web scraping η οποία αυτοματοποιεί τη διαδικασία της αντιγραφής από ιστοσελίδες.

Η αντιγραφή δεδομένων από μία ιστοσελίδα προϋποθέτει αρχικώς την ανίχνευση και την λήψη της ιστοσελίδας. Τότε το περιεχόμενο της μπορεί να αναλυθεί, να αναδιαμορφωθεί και τέλος να αντιγραφούν τα δεδομένα σε μια βάση δεδομένων, ένα υπολογιστικό φύλλο και άλλα.

Οι ιστοσελίδες δημιουργούνται σε γλώσσες HTML και XHTML και περιέχουν πληθώρα χρήσιμων δεδομένων σε μορφή κειμένου. Ωστόσο, οι περισσότερες από αυτές έχουν σχεδιαστεί για τελικούς χρήστες, για τον άνθρωπο δηλαδή και όχι για τη διευκόλυνση αυτοματοποιημένων χρήσεων. Γι΄ αυτό το λόγο έχουν αναπτυχθεί web scraping εργαλεία προκειμένου η εξαγωγή των δεδομένων να γίνεται πολύ γρήγορα και να μην απαιτείται ανθρώπινο χέρι.



Πηγή: https://www.webharvy.com/articles/what-is-web-scraping.html#:~:text=Web%20Scraping%20(also%20termed%20Screen,in%20table%20(spreadsheet)%20format.

Υπάρχουν ήδη ανεπτυγμένα εργαλεία τα οποία μπορούν να εγκατασταθούν τοπικά σε έναν υπολογιστή όπως τα Webharvy, Visual Web Ripper και το OutWit Hub και άλλα. Υπάρχουν και εργαλεία τα οποία είναι cloud-based όπως το import.io και το Mozenda.

Φυσικά, ό,τι λύσεις υπάρχουν υπό τη μορφή λογισμικού απαιτούν την ανάλυση και κατανόηση του HTML κώδικα των σελίδων και προϋποθέτουν την σύνταξη κατάλληλου κώδικα. Επομένως, το web scraping είναι εφικτό και με σύνταξη κώδικα από έναν προγραμματιστή ο οποίος θα λάβει υπόψιν του τις ιδιαίτερες ανάγκες που μπορεί να έχει αυτός που τον προσέλαβα και να επιτύχει το ίδιο αποτέλεσμα με ένα έτοιμο λογισμικό. Η διαδικασία που πρέπει να ακολουθήσει είναι να καταλάβει τη δομή του HTML κώδικα και στη συνέχεια να συντάξει τον δικό του κώδικα οποίος θα αντιγράφει αρχικά τα δεδομένα στη μνήμη του υπολογιστή μέσω κατάλληλης δομής δεδομένων που προσφέρεται από τη γλώσσα προγραμματισμού που χρησιμοποιεί και στη συνέχεια να γράψει τα δεδομένα σε μια βάση δεδομένων, ένα υπολογιστικό φύλλο και άλλα[16]. Πρέπει εδώ να τονιστεί ότι πριν γράψει σε όποιο μέσο του ζητηθεί, θα πρέπει να αναλύσει τα δεδομένα διότι αυτά ανακτώνται ως μη δομημένα δεδομένα. Επομένως θα πρέπει να σκεφτεί και ένα μοντέλο/σχήμα σύμφωνα με το οποίο θα αποθηκεύσει τα δεδομένα στη βάση δεδομένων ή οποιοδήποτε άλλο αποθηκευτικό μέσο του ζητηθεί.

Υπάρχει βέβαια και η δυνατότητα χρήσης API αν προσφέρεται από την εκάστοτε ιστοσελίδα. Όταν υπάρχει διαθέσιμο API τότε η δουλειά του προγραμματιστή είναι ευκολότερη, αφού δε χρειάζεται να ασχοληθεί καθόλου με τον HTML κώδικα. Η χρήση API επιστρέφει δομημένα δεδομένα σύμφωνα με το μοντέλο που χρησιμοποίησαν οι προγραμματιστές του οργανισμού που αντιπροσωπεύει μια ιστοσελίδα, γεγονός που αφαιρεί ένα επιπλέον καθήκον από τον προγραμματιστή.

API(application programming interface) καλείται μια διεπαφή υπολογιστών που καθορίζει τις αλληλεπιδράσεις μεταξύ πολλών ενδιάμεσων λογισμικών[18]. Ορίζει τα είδη των αιτημάτων και τον τρόπο που αυτά μπορούν να γίνουν καθώς και τις μορφές δεδομένων που πρέπει να χρησιμοποιηθούν, τις συμβάσεις που πρέπει να ακολουθηθούν και άλλα. Μπορεί να είναι, είτε σχεδιασμένο βάσει ενός βιομηχανικού προτύπου ώστε να εξασφαλίζεται η επιθυμητή λειτουργικότητα, είτε προσαρμοσμένο για ένα ειδικό στοιχείο.

Τα API επιτρέπουν το modular programming δηλαδή αυτή την τεχνική προγραμματισμού σύμφωνα με την οποία δίνεται έμφαση στο διαχωρισμό λειτουργικότητας ενός προγράμματος σε εναλλάξιμες λειτουργικές μονάδες έτσι ώστε για κάθε επιθυμητή λειτουργικότητα να εκτελείται μόνο ένα module και όχι ενδεχομένως ένα ολόκληρο πρόγραμμα[19]. Στην ανάπτυξη εφαρμογών ένα API απλοποιεί τον προγραμματισμό αφού αφαιρεί την υποκείμενη εφαρμογή και εκθέτει μόνο τα αντικείμενα, ενέργειες ή δεδομένα που χρειάζεται ο προγραμματιστής. Αν υποθέσουμε λοιπόν ότι έχουμε μια περίπλοκη εφαρμογή τηλεφωνικού καταλόγου δομημένων δεδομένων, όπου για κάθε επαφή υπάρχουν πολλές πληροφορίες όπως ονοματεπώνυμο, αριθμός τηλεφώνου, φυσική διεύθυνση κατοικίας, διεύθυνση ηλεκτρονικού ταχυδρομείου, ταχυδρομικός κώδικας κ.α., ένα κατάλληλα προσαρμοσμένο API επιτρέπει στον προγραμματιστή να εξάγει όποιες πληροφορίες χρειάζεται από τον τηλεφωνικό κατάλογο χωρίς να πρέπει να καταλάβει πώς λειτουργεί η εφαρμογή.

Για να γίνει πιο κατανοητός ο ρόλος ενός API στο διαδίκτυο γενικότερα δίνεται το εξής σενάριο:

Όταν ένας χρήστης θέλεις να μπει στη σελίδα του Facebook από ένα περιήγησης διαδικτύου(client-πελάτης), μία αίτηση φτάνει στον απομακρυσμένο server(εξυπηρετητή) της εταιρίας Facebook. Όταν ο περιηγητής λάβει απάντηση για την

αίτηση, ερμηνεύει τον κώδικα και προβάλει την αρχική σελίδα του Facebook. Ο απομακρυσμένος server του Facebook χρησιμοποιεί ένα API για να λάβει αιτήσεις και να αποστείλει απαντήσεις σε αυτές τις αιτήσεις. Δηλαδή δεδομένα. Γίνεται τώρα εύκολα αντιληπτό ότι κάθε φορά που επισκεπτόμαστε μία ιστοσελίδα ο περιηγητής μας «μιλάει» με ένα API το οποίο είναι μέρος ενός server.

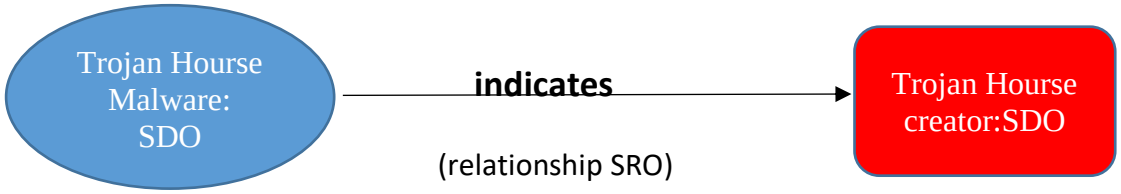
2.6 STIX - TAXII

2.6.1 STIX

Η βάση δεδομένων που δημιουργείται σε αυτή την εργασία έγινε με γνώμονα την ανάγκη για μία καθολική βάση δεδομένων που θα συμβάλει στην ασφάλεια στον κυβερνοχώρο διαμοιραζόμενη χωρίς περιορισμούς και χρηματικό κέρδος σε όλες τις ενδιαφερόμενες πλευρές. Η τεχνική επιτροπή πληροφοριών Cyber Threat Intelligence (CTI TC) του οργανισμού OASIS προτείνει τα πρότυπα STIX και TAXII για την περιγραφή απειλών και μέσον ανταλλαγής πληροφοριών αντίστοιχα. Τα δυο αυτά πρότυπα αναπτύχθηκαν υπό την εποπτεία του υπουργείου εσωτερικής ασφάλειας των Η.Π.Α. και πλέον η ανάπτυξή τους συνεχίζεται από τον οργανισμό OASIS.

Το STIX(Structured Treat Information eXpression) είναι ένα πρότυπο για την έκφραση πληροφοριών σχετικών με τις απειλές στο διαδίκτυο με δομημένο και καθορισμένο τρόπο[20]. Με βάση την μορφή αρχείων JSON επιτρέπει την αυτόματη ανταλλαγή πληροφοριών μεταξύ των πολλών εργαλείων που χρησιμοποιούνται για τη διασφάλιση της ασφάλειας ενός οργανισμού από απειλές στο διαδίκτυο.

Η τελευταία έκδοση STIX2.0 ορίζει δύο κατηγορίες αντικειμένων. Την STIX Domain Objects(SDO) και την STIX Relationship Objects(SRO). Τα SDO χάριν απλότητας μπορούν να θεωρηθούν ως κόμβοι ενός γραφήματος που συνδέονται με SRO όπως φαίνεται στον παρακάτω γράφο.



Το πρότυπο STIX2.0 ορίζει δώδεκα αντικείμενα τομέα STIX(SDO):

1. **attack-pattern**: Μια προσέγγιση που ακολουθεί ένας (ή περισσότεροι) κακόβουλοι δράστες για να θέσουν σε κίνδυνο έναν στόχο.

2. **campaign**: Μια συλλογή εχθρικών συμπεριφορών που περιγράφουν ένα σύνολο κακόβουλων δραστηριοτήτων ή επιθέσεων που στοχεύουν ένα συγκεκριμένο σύνολο θυμάτων για μια χρονική περίοδο.

3. **course-of-action**: Μια ενέργεια που έχει γίνει για την αποφυγή ή την αντιμετώπιση μιας επίθεσης

4. **identity**: Άτομα, οργανώσεις ή ομάδες, καθώς και τάξεις ατόμων, οργανισμών ή ομάδων.

5. **indicator**: Ενσωματώνει ένα μοτίβο STIX για τον εντοπισμό μιας κακόβουλης ή ύποπτης δραστηριότητας.

6. **intrusion-set**: Ένα σύνολο πόρων και εχθρικών συμπεριφορών που υπάρχει η υποψία ότι είναι χαρακτηριστικά ενός συγκεκριμένου threat-actor. Σε αντίθεση με το a campaign, δεν είναι συγκεκριμένο για ένα σύνολο στόχων ή για μια χρονική περίοδο.

7. **malware**: Κακόβουλο λογισμικό, επίσης γνωστό ως κακόβουλος κώδικας, που σκοπό έχει να θέσει σε κίνδυνο την εμπιστευτικότητα, την ακεραιότητα ή τη διαθεσιμότητα των δεδομένων συστημάτων του θύματος

8. **observed data**: Αντιπροσωπεύει πληροφορίες που παρατηρούνται σε ένα σύστημα ή δίκτυο (π.χ. διεύθυνση IP, αρχείο) ως ένα σύνολο STIX Cyber Observables.

9. **report**: συλλογή πληροφοριών απειλής με τη μορφή αντικειμένων STIX, SDO και SRO που αφορά ένα ή περισσότερα θέματα, όπως περιγραφή κακόβουλου δράστη, κακόβουλου λογισμικού ή τεχνικής εισβολής.

10. **threat-actor**: Άτομα, ομάδες ή οργανώσεις που είναι ύποπτοι για κακόβουλη συμπεριφορά.

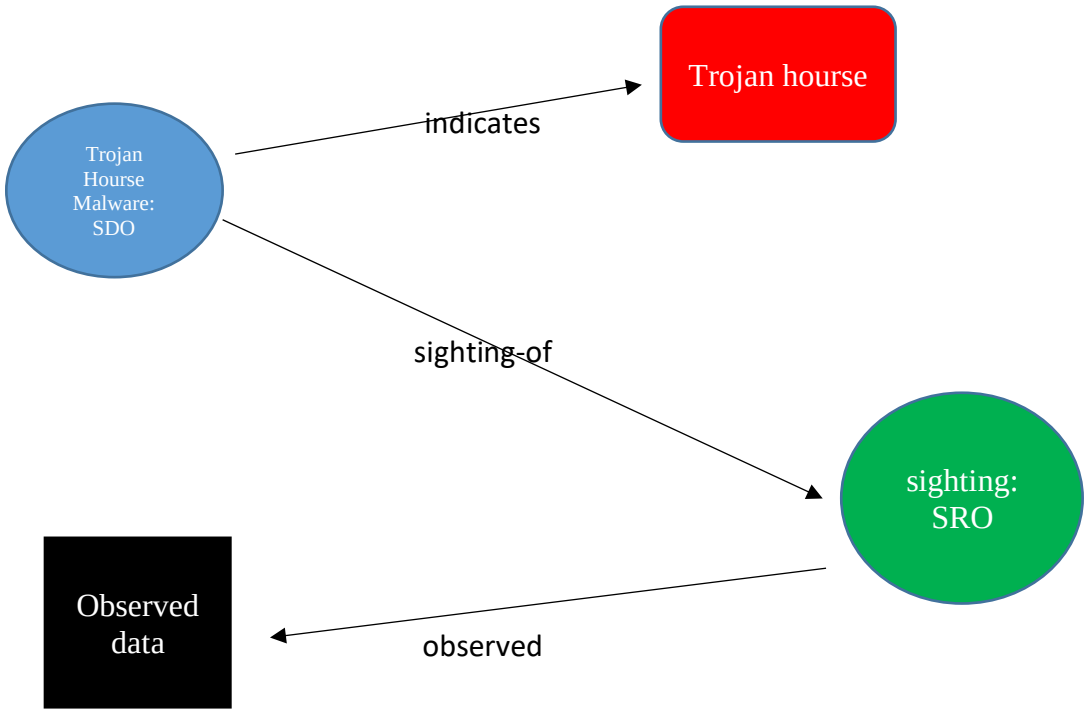
11. **tool**: Νόμιμο λογισμικό που μπορεί να έχει κακόβουλη χρήση

12. **vulnerability**: Ένα σφάλμα στο λογισμικό του οποίου η εκμετάλλευση μπορεί να επιτρέψει την παράνομη πρόσβαση σε ένα σύστημα ή δίκτυο.

Ορισμένα χαρακτηριστικά των SDO μπορεί να είναι μη δομημένα δεδομένα. Μπορεί όμως να παίρνουν τιμές που ορίζονται από τα λεξιλόγια οριζόμενα από το πρότυπο. Επί παραδείγματι, το επίπεδο ικανότητας ενός treat-actor μπορεί να παίρνει τιμές min, advanced, expert κ.α.. Η σημασία αυτών των τιμών ορίζεται από το STIX2.0.

Στα SRO αντικείμενα ορίζονται επίσης δύο αντικείμενα σχέσης STIX(SRO's):

1. Το **relationship** SRO το οποίο χρησιμοποιείται για τη σύνδεση δύο SDO αντικειμένων και περιγράφει τη σχέση τους μέσω της **relationship_type** ιδιότητας.
2. Το **sighting** SRO το οποίο εκφράζει την ανίχνευση ενός CTI(Cyber Threat Intelligence) όπως για παράδειγμα ενός malware.



παράδειγμα γράφου με αντικείμενα SDO, SRO

2.6.2 TAXII

Το TAXII(Trusted Automated Exchange of Intelligence Information) είναι ένα πρωτόκολλο ανταλλαγής πληροφοριών μέσω HTTPS[20]. Στο πρότυπο STIX ορίζεται ένα σύνολο απαιτήσεων για πελάτες/εξυπηρετητές και ένα REST API για την αλληλεπίδρασή τους με δύο τύπους υπηρεσιών όπως φαίνονται παρακάτω:

1. **Collection(συλλογή)**: Μία διεπαφή σε ένα αποθετήριο αντικειμένων που παρέχεται από έναν server ο οποίος επιτρέπει την εξυπηρέτηση πελατών από τον server αποκρινόμενος σε μια αίτηση/αίτημα**.**
2. **Channel(Κανάλι)**: επιτρέπει την ανταλλαγή πληροφοριών σύμφωνα με το μοντέλο publish-subscribe(δημοσίευση-εγγραφή)

Εδώ πρέπει να σημειωθεί ότι παρόλο που τα πρότυπα STIX και TAXII είναι αποτέλεσμα κοινής προσπάθειας και παρόλο που ένας διακομιστής TAXII πρέπει να μπορεί να χειριστεί το STIX, τα δύο αυτά πρότυπα είναι ανεξάρτητα. Δηλαδή είναι δυνατή η ανταλλαγή πληροφοριών STIX χωρίς τη χρήση του TAXII από τη μία, και από

την άλλη ένας διακομιστής TAXII μπορεί να διαχειριστεί και άλλες μορφές πληροφοριών εκτός του προτύπου STIX.

Το συμπέρασμα είναι ότι ο συνδυασμός των δύο αυτών προτύπων προσφέρουν τη δυνατότητα αξιόπιστης ανταλλαγής πληροφοριών σχετικά με απειλές στον κυβερνοχώρο και ότι η χρήση τους θα πρέπει να υιοθετηθεί από περισσότερους φορείς έρευνας σχετικά με την κυβερνοασφάλεια και αυτός είναι ο λόγος που χρησιμοποιούνται σε αυτή την εργασία.

2.6 Google Safe Browsing

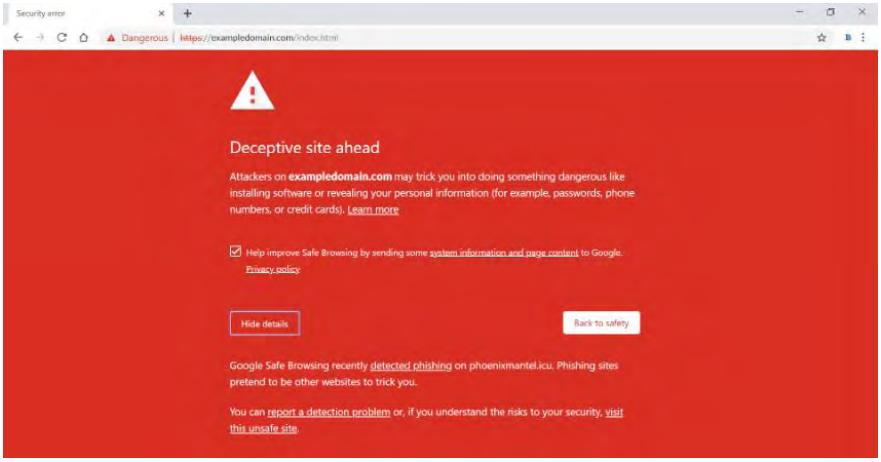
Το **google safe browsing** ξεκίνησε το 2007 για την προστασία των χρηστών σε όλο το διαδίκτυο από επίθεσης phishing(ψαρέματος) και έχει εξελιχθεί πλέον προσφέρονταw εργαλεία χρήσιμα για την προστασία από διάφορες απειλές στο διαδίκτυο όπως κακόβουλο λογισμικό[21].

Ο περιηγητής της Google και άλλοι περιηγητές χρησιμοποιούν το google safe browsing εμφανίζοντας ένα προειδοποιητικό μήνυμα όταν ο χρήστης έχει ζητήσει να επισκεπτεί έναν επικίνδυνο ιστότοπο ή προσπαθεί να κατεβάσει μια επιβλαβή εφαρμογή ή επιβλαβείς επεκτάσεις. Επιπλέον, οι ομάδες ασφαλείας της Google και του λογισμικού συστήματος για smartphones και tablets android για την ανάπτυξη υποδομής που σαρώνει όλες τις εφαρμογές που ανεβαίνουν στο google play store αλλά και τις αναβαθμίσεις αυτών ώστε να εξασφαλίζεται η προστασία των καταναλωτών.

Οι υπηρεσίες του google safe browsing είναι δωρεάν και διαθέσιμες προς χρήση από προγραμματιστές και εταιρίες που δραστηριοποιούνται στο διαδίκτυο. Η google παρέχει API προκειμένου οι προγραμματιστές να μπορούν εύκολα και γρήγορα να ενσωματώσουν το google safe browsing στον κώδικά τους και να παρέχουν προστασία στους πελάτες τους.

Στην παρακάτω εικόνα μπορείτε να δείτε ένα στιγμιότυπο από την προειδοποίηση που λαμβάνει ο χρήστης όταν προσπαθεί να επισκεφτεί έναν επικίνδυνο

ιστότοπο από τον περιηγητή Google Chrome.



Παρόμοιες ειδοποιήσεις λαμβάνουν οι χρήστες που χρησιμοποιούν τους περιηγητές Microsoft Edge και Mozilla Firefox μέσα από τα εργαλεία Smart Screen filter by Microsoft και Phishing protection by Mozilla Firefox αντίστοιχα που χρησιμοποιούν κάτω από το καπό το google safe browsing.

Στο πλαίσιο αυτής της εργασίας, το google safe browsing χρησιμοποιείται για διασταύρωση εγκυρότητας των απειλών που βρέθηκαν μέσω του web scraping σε προηγούμενα στάδια του κώδικα.

ΚΕΦΑΛΑΙΟ 3

ΥΛΟΠΟΙΗΣΗ - ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 ΒΑΣΙΚΑ ΕΡΓΑΛΕΙΑ

3.1.1 Python 2

Το βασικότερο συστατικό της υλοποίησης είναι η προγραμματιστική γλώσσα python. Η python είναι μια γλώσσα υψηλού επιπέδου που δημιουργήθηκε από τον Guido van Rossun και κυκλοφόρησε πρώτη φορά το 1991[22]. Με την python μπορεί κανείς να κάνει δομημένο προγραμματισμό, όπως γίνεται σε αυτή την υλοποίηση αλλά και αντικειμενοστραφή προγραμματισμό. Είναι διαθέσιμη και χρησιμοποιείται για ανάπτυξη λογισμικού σε πολλά λειτουργικά συστήματα.

Η Python 2 που χρησιμοποιείται στην παρούσα εργασία, κυκλοφόρησε το 2000[reference required unknow yet] και υποστηρίζεται μέχρι το 2020 με την τελευταία έκδοση 2.7.18. Από αυτό το έτος διακόπτεται η υποστήριξη και οι προγραμματιστές θα πρέπει να μεταναστεύσουν τον κώδικά τους από την python 2 στην python 3, η οποία κυκλοφόρησε το 2008 και θα συνεχίζει να υποστηρίζεται. Αυτό είναι αναγκαστικό να γίνει για λόγους ασφαλείας και συμβατότητας με μελλοντικά projects. Τα μεγάλα κλίμακας projects θα πρέπει να υποστούν κάποιες αλλαγές προσεκτικά, ώστε να διατηρηθεί η λειτουργικότητα τους στο ακέραιο.

Η εργασία αυτή ξεκίνησε να υλοποιείται και ολοκληρώθηκε με python 2, καθώς υπήρξαν κωλύματα από το σύστημα στο οποίο γράφτηκε ο κώδικας κατά την απόπειρα μετεγγραφής σε python 3.

3.1.2 Beautiful Soup

Η **Beautiful Soup** είναι μια βιβλιοθήκη της python και χρησιμοποιείται για την εξαγωγή δεδομένων από αρχεία HTML και XML[23]. Χρησιμοποιεί μεθόδους για

αναζήτηση, περιήγηση και μετατροπή ενός δέντρου προκειμένου να γίνει η εξαγωγή των επιθυμητών δεδομένων.
Μετατρέπει αυτόματα, έγγραφα που θέλει να εξερευνήσει ο προγραμματιστής σε μορφή Unicode και τα αντλούμενα δεδομένα σε UTF-8. Χρησιμοποιεί κάποιο αναλυτή κειμένου(parser) της python όπως ο **html5lib** ή ο **lxml** επιτρέποντας την επιλογή διαφορετικό στρατηγικών ανάλυσης κειμένου.

Μπορεί να αναλύσει και να περιηγηθεί σε οποιοδήποτε έγγραφο ζητηθεί. Κάποιες από τος πολύ σημαντικές λειτουργίες που εκτελεί είναι η αναζήτηση με βάση κριτηρίων που θέτε ο προγραμματιστής. Πρακτικά, κάποιες εντολές αναζήτησης θα μπορούσαν να ήταν η «εύρεση όλων των επικεφαλίδων με έντονο(bold) κείμενο» ή «εύρεση όλων των συνδέσμων», χωρίς να πρέπει να κάνει κάτι άλλο ο χρήστης ‘χειροκίνητα’ με επιπλέον κώδικα. Αυτό εξοικονομεί πολλές ώρες εργασίας στους προγραμματιστές.

Η εξαγωγή των επιθυμητών δεδομένων γίνεται σταδιακά. Ο προγραμματιστής πρέπει αρχικά να έχει μελετήσει αυτό που βλέπει ο τελικός χρήστης μια ιστοσελίδας στον περιηγητή του, και να επιλέξει ποια δεδομένα τον ενδιαφέρουν. Στη συνέχεια οφείλει να εξερευνήσει τον html κώδικα της ιστοσελίδας και να εντοπίσει ενδεχομένως διάφορα αναγνωριστικά αμέσως πριν και μετά από το κείμενο που θέλει να εξάγει, ώστε να δρομολογήσει την εξαγωγή τον δεδομένων με εντολές στον κώδικά του. Η παραπάνω διαδικασία διενεργείται πειραματικά, μέχρι το τελικό αποτέλεσμα να είναι αυτό που επιθυμεί ο προγραμματιστής και να μην του επιστρέφονται σκουπίδια δηλαδή άχρηστα ή λάθος δεδομένα που δεν έχουν νόημα ή δεν τα χρειάζεται. Αυτά όλα σαφώς πρέπει να εκτελεστούν ευλαβικά, όταν η εξαγωγή δεδομένων γίνεται από τον html κώδικα κάποιας ιστοσελίδας και όχι με άντληση δομημένων δεδομένων μέσω API.

Τέλος η βιβλιοθήκη αυτή για να χρησιμοποιηθεί, απαιτεί την εγκατάσταση απαραίτητων αρχείων στο σύστημα είτε ο προγραμματιστής επιλέγει να δουλεύει τον κώδικά τους σε scripts και να διαχειρίζεται τη μεταγλώττιση και την εκτέλεση του κώδικα με ένα terminal, είτε επιλέγει να δουλεύει με μία εφαρμογή που επιτρέπει τη δημιουργία και εκτέλεση κώδικα python όπως το **Jupyter Notebook** που τρέχει σε περιηγητή. Για λόγους ποικιλίας αλλά και κάποιων περιορισμών στις δυνατότητες του συστήματος οι οποίες επιφέρουν σφάλματα στη λειτουργία του Jupyter Notebook, για

διάφορα τμήματα του κώδικα χρησιμοποιήθηκαν αμφότερες και οι δύο παραπάνω επιλογές για τη συγγραφή και εκτέλεση του κώδικα.

3.1.3 urlib2

Το urlib2 είναι ένα module το οποίο προσφέρει ένα αναβαθμισμένο API για χρήση πόρων από το ίντερνετ που καθορίζονται από URL's[24]. Καθορίζει συναρτήσεις και κλάσεις προκειμένου να μπορούν να ανοίγουν τα URL's κατά βάση HTTP.

Αρχικά θέτουμε σε μία μεταβλητή το url που θέλουμε να ανοίξουμε και στη συνέχεια το ανοίγουμε με μία χρησιμοποιώντας την εντολή urlib2.urlopen(). Επιστρέφονται πληροφορίες τις οποίες κάνουμε ανάλυση με τη συνάρτηση BeautifulSoup η οποία παίρνει ως ορίσματα τη μεταβλητή με τις πληροφορίες που επιστράφηκαν και ένα feature το οποίο καθορίζει τον parser(αναλυτή κειμένου) που επιλέγουμε για να αναλύσουμε τις πληροφορίες.

3.2 Τα τέσσερα scripts του κώδικα

3.2.1 Το πρώτο script

Στο σημείο αυτό θα ξεκινήσουμε ην ανάλυση της πρακτικής εργασίας η οποία αποτελείται από τέσσερα scripts. Στο πρώτο script αρχικά δημιουργούμε και συνδεόμαστε μέσω ενός κέρσορα στη sqlite3 βάση δεδομένων στην οποία φτιάχνουμε ένα table με τα απαραίτητα πεδία που θα έχει κάθε εγγραφή. Τα πεδία αυτά είναι τα ip, domain_name, longitude, latitude, country, OS, και city. Με τα longitude και latitude μπορούμε να τοποθετήσουμε πάνω στον παγκόσμιο χάρτη την τοποθεσία του μηχανήματος που τρέχει η κάθε ιστοσελίδα ενώ με τα πεδία country και city προσφέρουμε πιο φιλικά την τοποθεσία του μηχανήματος. Με το πεδίο OS δηλώνεται το

λειτουργικό σύστημα που χρησιμοποιείται στο μηχάνημα και φυσικά με το πεδίο IP καθορίζεται η IP του μηχανήματος ενώ με το domain_name η ηλεκτρονική διεύθυνση της ιστοσελίδας που εμπεριέχει κίνδυνο. Σε επόμενο στάδιο ης ανάλυσης θα εξηγηθεί πώς ακριβώς εντοπίζονται όλα αυτά τα στοιχεία που συνθέτουν τις εγγραφές της βάσης δεδομένων.

Η sqlite είναι μία βιβλιοθήκη με την οποία μπορούμε να υλοποιήσουμε μια αυτοδύναμη, χωρίς διακομιστή βάση δεδομένων SQL[https://www.sqlite.org/about.html]. Είναι δωρεάν για χρήση για οποιοδήποτε σκοπό εμπορικό ή ιδιωτικό. Είναι η πιο διαδεδομένη βάση δεδομένων στον κόσμο και χρησιμοποιείται σε εφαρμογές υπολογιστών, εφαρμογές έξυπνων συσκευών και άλλα. Η ανάπτυξη της βιβλιοθήκης ξεκίνησε το 2009 και ο στόχος των προγραμματιστών είναι να υποστηριχθεί ως το 2050. Είναι ένα πολύ χρήσιμο δωρεάν εργαλείο που χρησιμοποιήθηκε για την κατασκευή της βάσης δεδομένων για αυτή την εργασία.

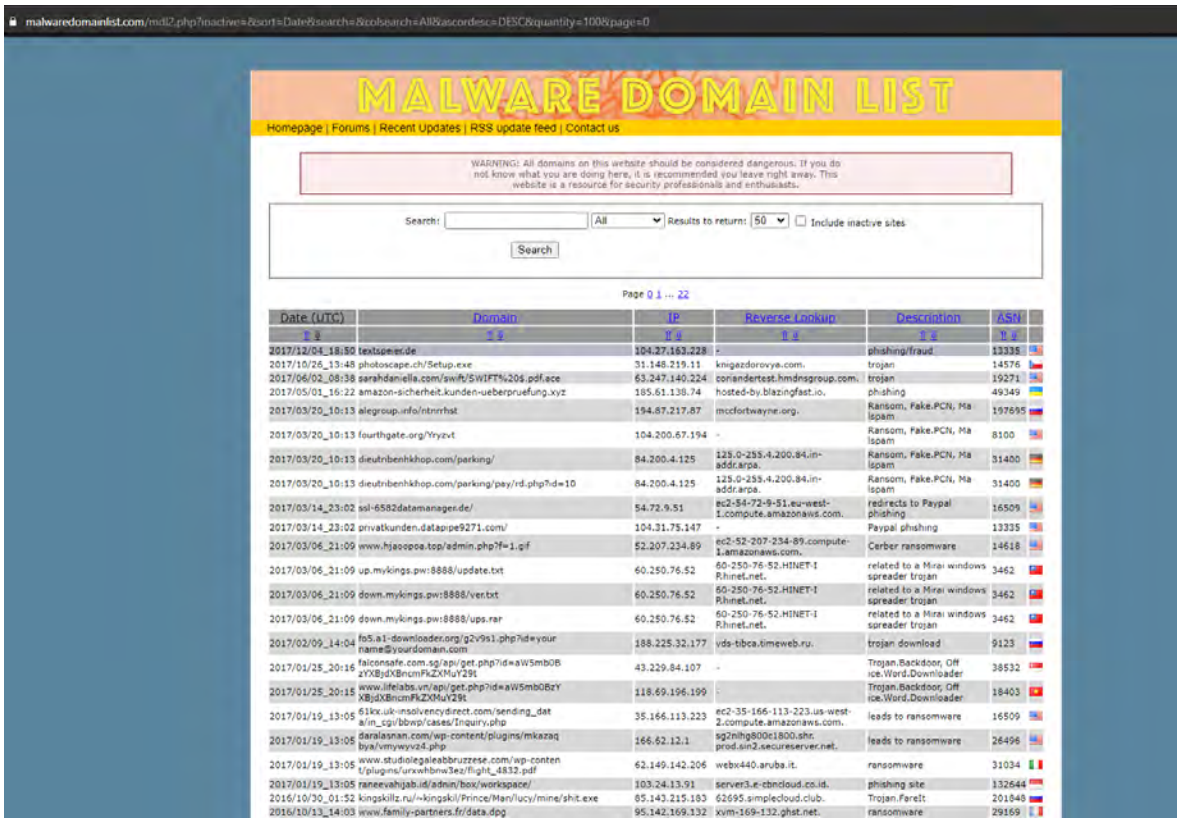
3.2.1.1 Εξόρυξη από τον HTML κώδικα της σελίδας

Αρχικά πριν γίνει οποιαδήποτε άλλη εργασία, δηλώνονται 3 λίστες στις οποίες θα αποθηκεύονται προσωρινά οι IP, τα domain names και περιγραφή της απειλής που κρύβεται και διανέμεται από την κάθε ιστοσελίδα κατ' αντιστοιχία.

Όπως αναφέρθηκε ήδη σε προηγούμενο κεφάλαιο, η εξόρυξη δεδομένων από το ίντερνετ γίνεται είτε με εξαγωγή μη δομημένων δεδομένων από τον HTML κώδικα μιας ιστοσελίδας, είτε με άντληση δομημένων δεδομένων μέσω ενός API που προσφέρει η εκάστοτε ιστοσελίδα. Σε αυτή την παρ άγραφο θα παρουσιαστεί η διαδικασία του screen scraping από δύο γνωστές ιστοσελίδες που διαθέτουν πληροφορίες που μας ενδιαφέρουν.

Η πρώτη ιστοσελίδα είναι η Malware Domain List την οποία μπορεί να επισκεφτεί κανείς αναζητώντας την διεύθυνση ηλεκτρονική διεύθυνση «https://www.malwaredomainlist.com/mdl2.php?inactive=&sort=Date&search=&colsearch=All&ascordesc=DESC&quantity=100&page=0». Στη διεύθυνση αναφέρονται IP διευθύνσεις και σχετικές πληροφορίες που μας ενδιαφέρουν πολύ. Το υλικό που χρειάζεται να πάρουμε από αυτή την ιστοσελίδα κατανέμεται σε 23 υποσελίδες οι

οποίες μπορούν να εμφανιστούν είτε επιλέγοντας με το ποντίκι την κατάλληλη υποσελίδα, είτε αλλάζοντας στο url ένα πολύ συγκεκριμένο σημείο.



Πηγή:

https://www.malwaredomainlist.com/mdl2.php?inactive=&sort=Date&search=&colsearch=All&ascordesc=DESC&quantity=100&page=0

Για να καταφέρουμε να πάρουμε όλο το υλικό αυτοματοποιημένα με κώδικα, δημιουργούμε αρχικώς μια επανάληψη της οποίας το σώμα εκτελείται όσες είναι και οι υποσελίδες που προσφέρουν επιθυμητό υλικό. Κάθε φορά που ξεκινάει μια καινούρια επανάληψη ανακτούμε το url της εκάστοτε υποσελίδας με τη βοήθεια της urlib2, ενώ αναλύουμε το κείμενο της κάθε υποσελίδας με την BeautifulSoup αναζητώντας παράλληλα με την εντολή soup.find_all κάποια αναγνωριστικά μέσα στον HTML κώδικα τα οποία μας βοηθούν να πλησιάσουμε τα επιθυμητά δεδομένα και που ανιχνεύτηκαν έπειτα από μελέτη του HTML κώδικα. Μετά από μερικές δεκάδες αποτυχημένων προσπαθειών, επιτυγχάνεται το αποτέλεσμα που πρέπει. Με τρεις διαδοχικές επαναλήψεις αποκτούμε τα δεδομένα IP, domain_name και περιγραφή της απειλής.

Στη συνέχεια μελετώνται τα δεδομένα που αντλήθηκαν, ώστε με κατάλληλο κώδικα να καθαριστούν από έξτρα χαρακτήρες, σύμβολα κτλ που τα συνόδευαν ώστε να

αποθηκευτούν στις τρεις λίστες καθαρά και έτοιμα να μπουν στη βάση δεδομένων.

Κάπως έτσι αποκτούμε υλικό για τις πρώτες 2300 εγγραφές που αργότερα θα καταχωρηθούν στη βάση δεδομένων.

Η δεύτερη ιστοσελίδα που χρησιμοποιείται για screen scraping είναι η VXvault της οποίας η ηλεκτρονική διεύθυνση είναι η «http://vxvault.net/ViriList.php?s=500&m=40».

Η ιστοσελίδα αυτή παρέχει πάνω από 1.5 εκατομμύρια πιστοποιημένα ενεργά ή ανενεργά domains με τις IP τους που είναι κακόβουλα.

Πηγή: http://vxvault.net/ViriList.php?s=500&m=40

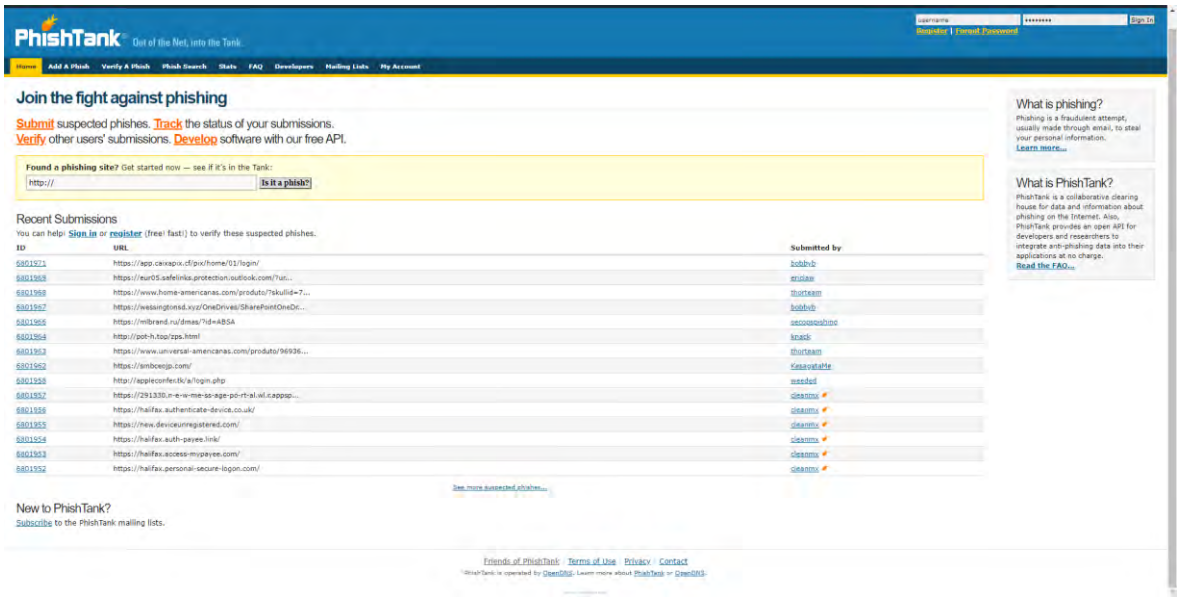
Η άντληση του υλικού από αυτή την ιστοσελίδα γίνεται με παρόμοια τακτική όπως αυτή που εφαρμόζεται για την Malware Domain List. Μελετήθηκε ο HTML κώδικας και πειραματικά επιτυγχάνεται η άντληση της τριπλέτας [IP,domain_name,περιγραφή_απειλής]. Για λόγους οικονομίας χρόνου, δεν αποκτήθηκε

όλο το υλικό. Δηλαδή στην επαναληπτική διαδικασία άντλησης των δεδομένων δεν συμπεριλήφθηκαν όλες οι υποσελίδες που προσφέρουν υλικό.

Μετά την άντληση των δεδομένων από αυτές τις δύο ιστοσελίδες ελέγχουμε αν οι λίστες IP, domain_names και description(της οποίας κάθε στοιχείο περιέχει την περιγραφή για μια απειλή) έχουν το ίδιο μήκος ώστε να υπάρχει η βεβαιότητα ότι υπάρχει αντιστοιχία μεταξύ IP, domain και περιγραφή απειλής των δεδομένων που έχουν αποκτηθεί. Στις δοκιμές που έγιναν, παρατηρήθηκαν διαφορετικά μήκη των λιστών γεγονός που αποδόθηκε στην αδυναμία εισαγωγής στοιχείων στη λίστα διότι δεν εγκρινόταν από τα πρότυπα ελέγχου για καθαρά δεδομένα. Επομένως χρειάστηκε επιπλέον κώδικας για να εντοπιστούν επαναλαμβανόμενα μοτίβα «σκουπιδιών» στα δεδομένα και ενίσχυση του κώδικα «καθαρισμού» ώστε να μπορούν όλα τα δεδομένα να εγκριθούν και να εισαχθούν στις αντίστοιχες λίστες.

3.2.1.2 Εξόρυξη με API

Σε αυτή την περίπτωση άντλησης δεδομένων η προσέγγιση είναι διαφορετική. Η επόμενη ιστοσελίδα που χρησιμοποιείται είναι η PhishTank της οποίας η ηλεκτρονική διεύθυνση είναι η «http://phishtank.org/». Η ιστοσελίδα αυτή ασχολείται κυρίως με περιπτώσεις απειλών οι οποίες εμφανίζονται μέσω ιστοσελίδων και έχουν να κάνουν με phishing απειλές, δηλαδή με περιπτώσεις όπου οι χρήστες του διαδικτύου παγιδεύονται από επιτήδειους και αποκαλύπτουν σημαντικά προσωπικά δεδομένα τους, όπως λογαριαμό email και password στοιχεία πιστωτικών καρτών κτλ.



Πηγή: http://phishtank.org/

Πολλοί μεγάλοι οργανισμοί που ασχολούνται με την κυβερνοασφάλεια χρησιμοποιούν τα δεδομένα που παρέχει η PhishTank. Στην παρακάτω εικόνα μπορείτε να δείτε τους σημαντικότερους από αυτούς τους οργανισμούς.

Πηγή: http://phishtank.org/friends.php

Η PhishTank παρέχει API με το οποίο μπορούμε να αντλήσουμε δομημένα δεδομένα για τη δική μας βάση δεδομένων. Αυτό που έπρεπε να γίνει είναι αρχικά να δημιουργηθεί ένας λογαριασμός στην ιστοσελίδα ώστε να μπορούμε να αποκτήσουμε ένα προσωπικό API key, και στη συνέχεια να προσθέσουμε κατάλληλο κώδικα σε αυτόν που παρέχεται από την ιστοσελίδα για το API ώστε να αρχίσει η άντληση δεδομένων. Τα δεδομένα που αποκτούνται έρχονται στη δομή dictionary της python. Εξερευνώντας τα δεδομένα επιλέγουμε να κρατήσουμε τα πεδία που μας ενδιαφέρουν τα οποία μέχρι

τώρα είναι τα IP, domain_name, και περιγραφή της απειλής, τα οποία αυτή τη φορά μπορούμε ευκολότερα να καταχωρήσουμε στη βάση αλλάζοντας μόνο την κωδικοποίησή τους. Σε αυτή την περίπτωση επειδή πολλοί οργανισμοί χρησιμοποιούν τη βάση δεδομένων της PhishTank, οι διαχειριστές έχουν βάλει όριο στη χρήση του API για λόγους καλής εξυπηρέτησης των πιο σημαντικών ενδιαφερόμενων πλευρών. Μία φορά ανά δώδεκα ώρες μπορεί κάποιος ιδιώτης να τραβήξει δεδομένα από το API, επομένως η ανάπτυξη του κώδικα καθυστέρησε αρκετά και κάθε φορά που μπορεί κανείς να χρησιμοποιήσει το API, πρέπει να έχει χτίσει πολύ προσεκτικά τον κώδικά του, ώστε τουλάχιστον να αποφεύγονται σφάλματα σύνταξης και να περιορίζεται στη διόρθωση λογικών σφαλμάτων.

Όταν όλα λειτούργησαν όπως πρέπει, έγινε εκ νέου έλεγχος για το μήκος των λιστών που αποθηκεύονται προσωρινά τα δεδομένα για τη διατήρηση της αντιστοιχίας μεταξύ των πληροφοριών.

Μέχρι αυτό το σημείου του script έχουμε γνώση για κάποια πράγματα που αφορούν τη βάση δεδομένων. Γνωρίζουμε πόσες θα είναι συνολικά οι εγγραφές του ενός από τα δύο tables έχει η βάση, αφού έχουμε δύο εκ των βασικών πεδίων. Το IP και το domain_name.

Το επόμενο τμήμα κώδικα του πρώτου script αφορά την απόκτηση των υπόλοιπων πεδίων για κάθε εγγραφή και τον έλεγχο διπλότυπων, αφού αντλούμε δεδομένα από 3 διαφορετικές πηγές για τα βασικά πεδία και όπως είναι φυσικό, μπορεί να συμβαίνει ένα ζεύγος [IP,domain_name] να είναι καταχωρημένο σε περισσότερες από μία πηγή. Πρέπει εδώ να τονιστεί ότι ο έλεγχος για διπλότυπες εγγραφές είναι ιδιαίτερα χρήσιμος και για την περίπτωση που ξανατρέξει το script ώστε να γίνει ενημέρωση της βάσης για νέες απειλές.

Πριν αναλυθεί το τελευταίο κομμάτι του κώδικα, πρέπει να αναφερθεί η επόμενη ιστοσελίδα από την οποία αντλούνται δεδομένα τα οποία αφορούν τα υπόλοιπα πεδία των εγγραφών. Πρόκειται για τη Shodan της οποίας η ηλεκτρονική διεύθυνση είναι η "https://shodan.io". Η Shodan χρησιμοποιείται παγκοσμίως από ερευνητές, επαγγελματίες στην ασφάλεια και γενικά μεγάλες επιχειρήσεις. Η σελίδα αυτή εντοπίστηκε στην κατά την έρευνα για απόκτηση πληροφοριών για μία IP διεύθυνση. Παρομοίως με τη PhishTank διαθέτει βάση δεδομένων με απειλές πολλών ειδών ενώ παρέχει επιπλέον πολλές πληροφορίες όσον αφορά την τοποθεσία των μηχανημάτων

αλλά και άλλα θέματα που μπορούν να ενδιαφέρουν έναν ερευνητή της κυβερνοασφάλειας. Προσφέρει πολύ καλό API το οποίο χρησιμοποιείται για αυτή την εργασία.

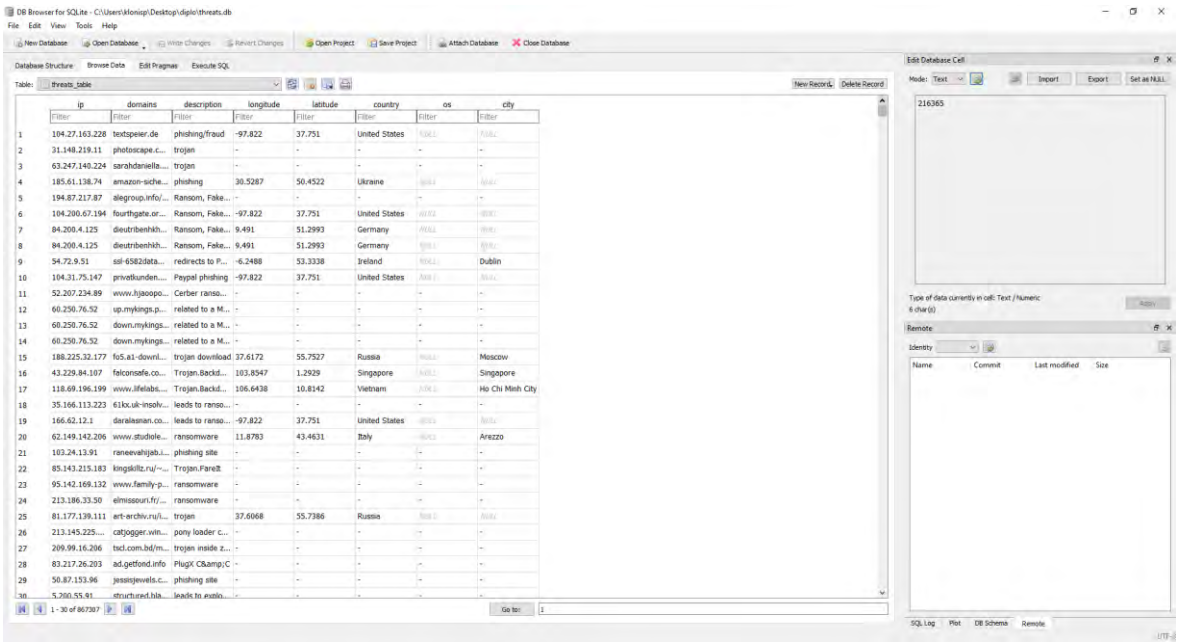
Αρχικά για να είναι εφικτό να έχουμε πρόσβαση στο API πρέπει να αποκτήσουμε ένα προσωπικό API key κατά παρόμοιο τρόπο με th PhishTank δημιουργώντας ένα λογαριασμό. Επίσης πρέπει να κάνουμε import τη βιβλιοθήκη της python shodan.
Στη συνέχεια μελετήθηκε ο python κώδικας για το API και προσαρμόστηκε στο script ώστε να παραχθεί η απαιτούμενη λειτουργικότητα. Πρέπει να σημειωθεί ότι παρόλο που έχει προχωρήσει η έρευνα για την ασφάλεια και παρόλο που τα εργαλεία και τα μέσα για την έρευνα έχουν αναβαθμιστεί πολύ, δεν είναι πάντα εφικτό να ανακτούμε όλες τις πληροφορίες που χρειάζονται για μία απειλή που πιστοποιηθεί. Γι΄ αυτό το λόγο στο τμήμα του κώδικα που αντλούνται δεδομένα από τη Shodan χρησιμοποιείται το σχήμα try-except ώστε να θέτουμε με NULL τα πεδία της βάσης που δεν είναι γνωστά.

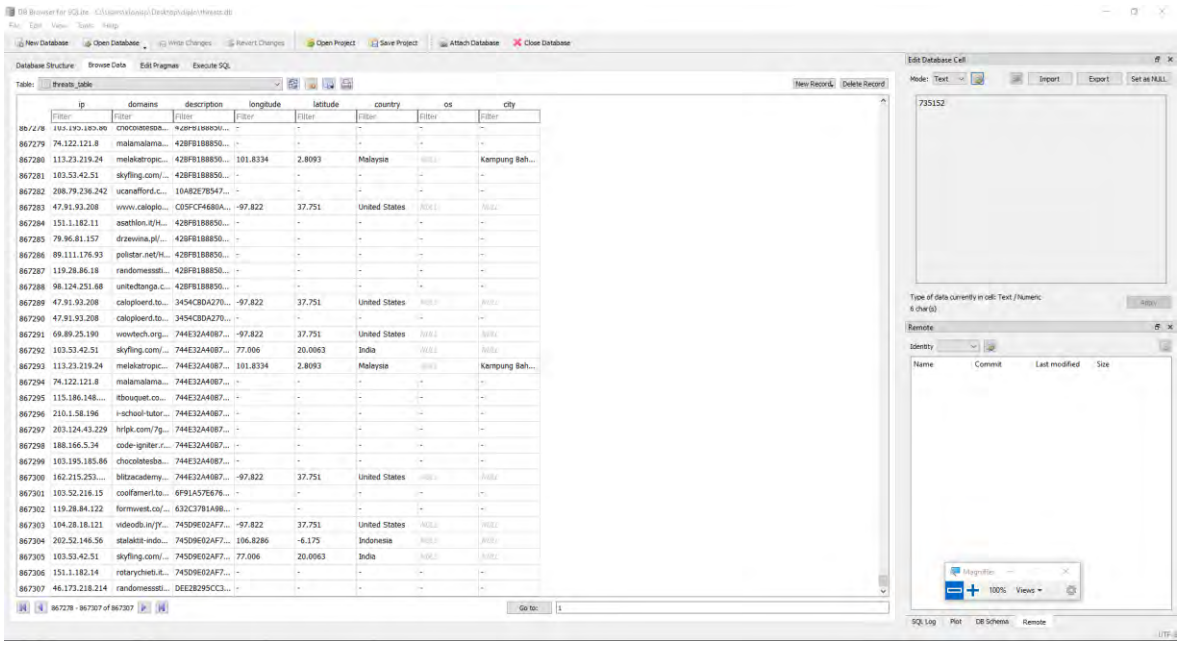
Επομένως, αρχικά γίνεται έλεγχος για διπλότυπα μέσα στη βάση. Ο έλεγχος αυτός λειτουργεί για τα τρεξίματα του script εκτός του πρώτου. Στη συνέχεια Τρέχει μία επανάληψη τόσες φορές όσες οι εγγραφές που θα περαστούν στη βάση δεδομένων. Μέσα στην επανάληψη υπάρχει ο κώδικας ο οποίος αντλεί δεδομένα από τη Shodan και ο κώδικας οποίος γράφει τις εγγραφές στη βάση. Όπως αναφέρθηκε και νωρίτερα τα πεδία που μπαίνουν στο table των εγγραφών είναι τα ip, domains, description, longitude, latitude, country, os, city. Εκτός από το πεδίο IP, τα υπόλοιπα πεδία έχουν κατά περιπτώσεις τιμές NULL αφού δεν υπάρχουν στοιχεία για όλες τις εγγραφές στη βάση του Shodan.

Το script ολοκληρώνεται κλείνοντας τη σύνδεση με τη βάση δεδομένων μας και εκτυπώνοντας στην οθόνη ένα φιλικό για το χρήστη μήνυμα τερματισμού. Για να τρέξει όλος αυτός ο κώδικας είναι προφανές ότι στην αρχή του script γίνονται κατάλληλα imports για τις βιβλιοθήκες της python που χρησιμοποιούνται. Ο συνολικός χρόνος που έτρεξε αυτό το script ήταν περίπου 13 ημέρες και δημιουργήθηκαν 867.307 εγγραφές.

3.2.1.3 Προβολή της βάσης δεδομένων σε γραφικό περιβάλλον

Για να μπορέσουμε να προβάλουμε γραφικά τη βάση δεδομένων που δημιουργήθηκε χρησιμοποιείται η εφαρμογή για λειτουργικό σύστημα Windows «DB Browser for SQLite». Παρακάτω μπορείτε να δείτε μερικά στιγμιότυπα της εφαρμογής όταν προβάλλεται το threats_table που δημιουργήθηκε στο πρώτο script. Συγκεκριμένα εμφανίζονται 2 στιγμιότυπα στα οποία προβάλλονται με αρίθμηση οι πρώτες 30 εγγραφές και οι τελευταίες 30 εγγραφές με την τελευταία εγγραφή να έχει αρίθμηση 867.307 γεγονός που πιστοποιεί τον αριθμό των εγγραφών που ισχυριζόμαστε ότι μπήκαν στη βάση.





	ip	domains	description	longitude	latitude	country	os	city
867279	74.122.121.8	malamalama...	42BFB188850...	-	-	-	-	-
867280	113.23.219.24	melakatropic...	42BFB188850...	101.8334	2.8093	Malaysia	-	Kampung Bah...
867281	103.53.42.51	skyfling.com/...	42BFB188850...	-	-	-	-	-
867282	208.79.236.242	ucanafford.c...	10A82E7B547...	-	-	-	-	-
867283	47.91.93.208	www.celoplo...	C05FCF4680A...	-97.822	37.751	United States	-	-
867284	151.1.182.11	asathlon.it/H...	42BFB188850...	-	-	-	-	-
867285	79.96.81.157	drzewina.pl/...	42BFB188850...	-	-	-	-	-
867286	89.111.176.93	polistar.net/H...	42BFB188850...	-	-	-	-	-
867287	119.28.86.18	randomesssti...	42BFB188850...	-	-	-	-	-
867288	98.124.251.68	unitedtanga.c...	42BFB188850...	-	-	-	-	-
867289	47.91.93.208	caloploerd.to...	3454C8DA270...	-97.822	37.751	United States	-	-
867290	47.91.93.208	caloploerd.to...	3454C8DA270...	-	-	-	-	-
867291	69.89.25.190	wowtech.org...	744E32A4087...	-97.822	37.751	United States	-	-
867292	103.53.42.51	skyfling.com/...	744E32A4087...	77.006	20.0063	India	-	-
867293	113.23.219.24	melakatropic...	744E32A4087...	101.8334	2.8093	Malaysia	-	Kampung Bah...
867294	74.122.121.8	malamalama...	744E32A4087...	-	-	-	-	-
867295	115.186.148....	itbouquet.co...	744E32A4087...	-	-	-	-	-
867296	210.1.58.196	i-school-tutor...	744E32A4087...	-	-	-	-	-
867297	203.124.43.229	hrlpk.com/7g...	744E32A4087...	-	-	-	-	-
867298	188.166.5.34	code-igniter.r...	744E32A4087...	-	-	-	-	-
867299	103.195.185.86	chocolatesba...	744E32A4087...	-	-	-	-	-
867300	162.215.253....	blitzacademy...	744E32A4087...	-97.822	37.751	United States	-	-
867301	103.52.216.15	coolfamerl.to...	6F91A57E676...	-	-	-	-	-
867302	119.28.84.122	formwest.co/...	632C3781A98...	-	-	-	-	-
867303	104.28.18.121	videodb.in/jY...	745D9E02AF7...	-97.822	37.751	United States	-	-
867304	202.52.146.56	stalaktit-indo...	745D9E02AF7...	106.8286	-6.175	Indonesia	-	-
867305	103.53.42.51	skyfling.com/...	745D9E02AF7...	77.006	20.0063	India	-	-
867306	151.1.182.14	rotarychieti.it...	745D9E02AF7...	-	-	-	-	-
867307	46.173.218.214	randomesssti...	DEE2B295CC3...	-	-	-	-	-

867278 - 867307 of 867307

3.2.2 Το δεύτερο script

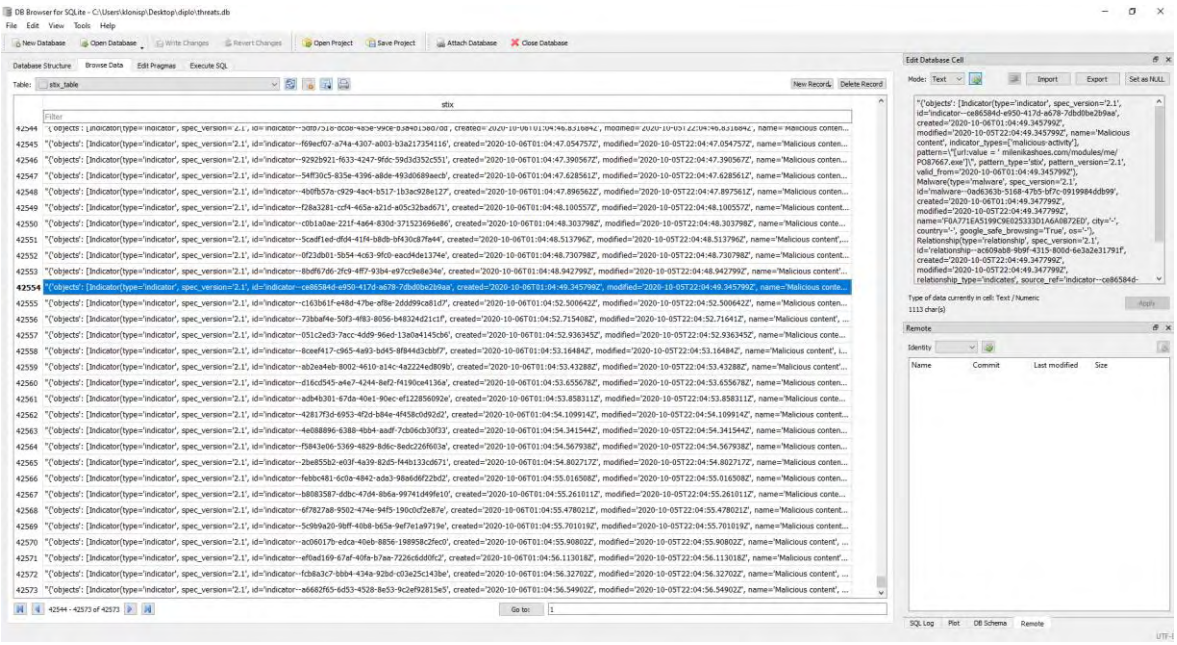
Στο δεύτερο script μετατρέπονται οι απειλές στο πρότυπο stix στο οποίο συμπεριλαμβάνεται και ο έλεγχος μέσω του google safe browsing. Ειδικά για να κάνουμε χρήση του google safe browsing πηγαίνουμε στη σελίδα του google safe browsing Lookup API και συνδεόμαστε με ένα google λογαριασμό ώστε να εξασφαλίσουμε ένα προσωπικό API key ώστε να μπορέσουμε να χρησιμοποιήσουμε το API της google για το google safe browsing.

Επιστρέφοντας στο script κάνουμε κατάλληλα imports για να μπορεί να τρέξει ο κώδικας χωρίς προβλήματα. Έπειτα, συνδεόμαστε στη βάση με κατάλληλο κώδικα, φτιάχνουμε ένα καινούριο table το stix_table και παίρνουμε όλες τις εγγραφές του threats_table από τη βάση μας και ελέγχουμε χρησιμοποιώντας το API του google safe browsing, αν οι απειλές είναι πιστοποιημένες και από τη google.

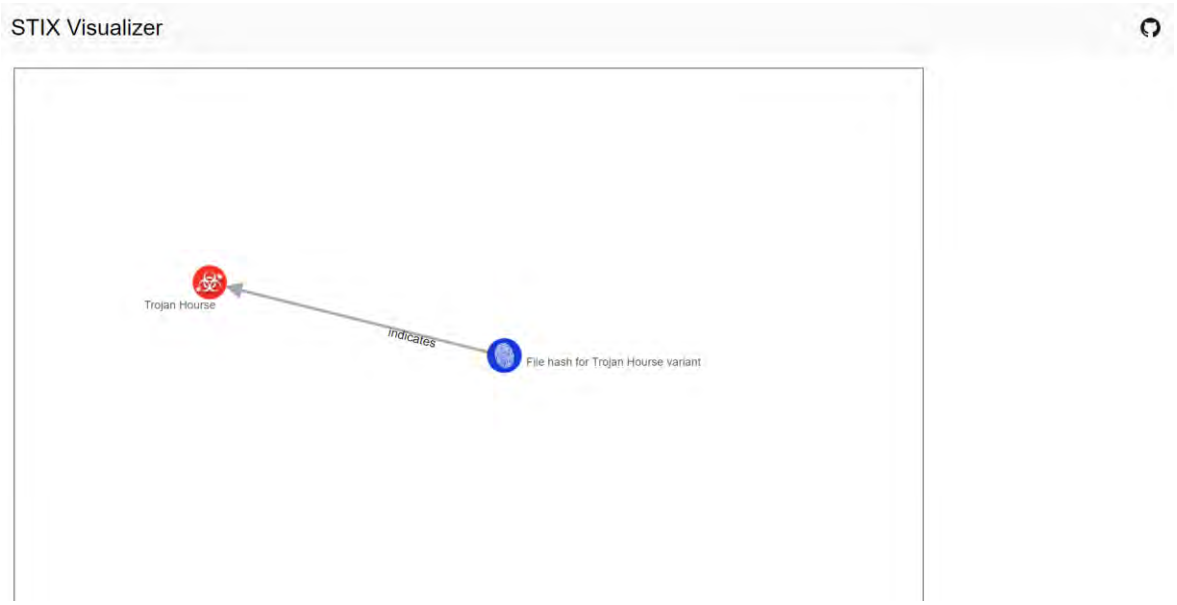
Στη συνέχεια προσαρμόζουμε τον ήδη διαθέσιμο κώδικα για το πρότυπο stix στις δικές μας απαιτήσεις αφαιρώντας επιπλέον μεταβλητές που αντιστοιχούν σε πληροφορίες που δε μας ενδιαφέρουν και εν τέλει προκύπτει πληροφορία που αντιστοιχεί στην αναπαράσταση μιας απειλής σε πρότυπο stix.

Τέλος, εισάγουμε στη βάση δεδομένων στο stix_table, τις απειλές σε μορφή stix. Όλος ο προηγούμενος κώδικα συμπεριλαμβάνεται σε μία επανάληψη οι οποία θα τρέξει τόσες φορές όσες είναι οι εγγραφές του threats_table της βάσης δεδομένων. Στην επανάληψη δε συμπεριλαμβάνεται σαφώς ο κώδικας σύνδεσης στη βάση, η δημιουργία του stix_table, η δήλωση του API key του google safe browsing και το url του google safe browsing το οποίο χρησιμοποιείται για να γίνει αίτηση(request) στη google. Στο τέλος του script αποσυνδεόμαστε από τη βάση με κατάλληλο κώδικα.

Για να δούμε το αποτέλεσμα που παράγεται με αυτό το script, πέρα από κατάλληλες εκτυπώσεις στην οθόνη μας, μπορούμε φυσικά να χρησιμοποιήσουμε εκ νέου την εφαρμογή «DB Browser for SQLite» όπου πλέον θα υπάρχει το νέο stix_table το οποίο θα περιέχει τις απειλές σε μορφή stix. Για εξοικονόμηση χρόνου το script αυτό έτρεξε περίπου 3 ώρες και στο stix_table εγγράφηκαν 42.573 εγγραφές, δλδ απειλές που μετατράπηκαν στο πρότυπο stix. Παρακάτω φαίνεται ένα στιγμιότυπο της εφαρμογής «DB Browser for SQLite» όπου προβάλλεται το stix_table.



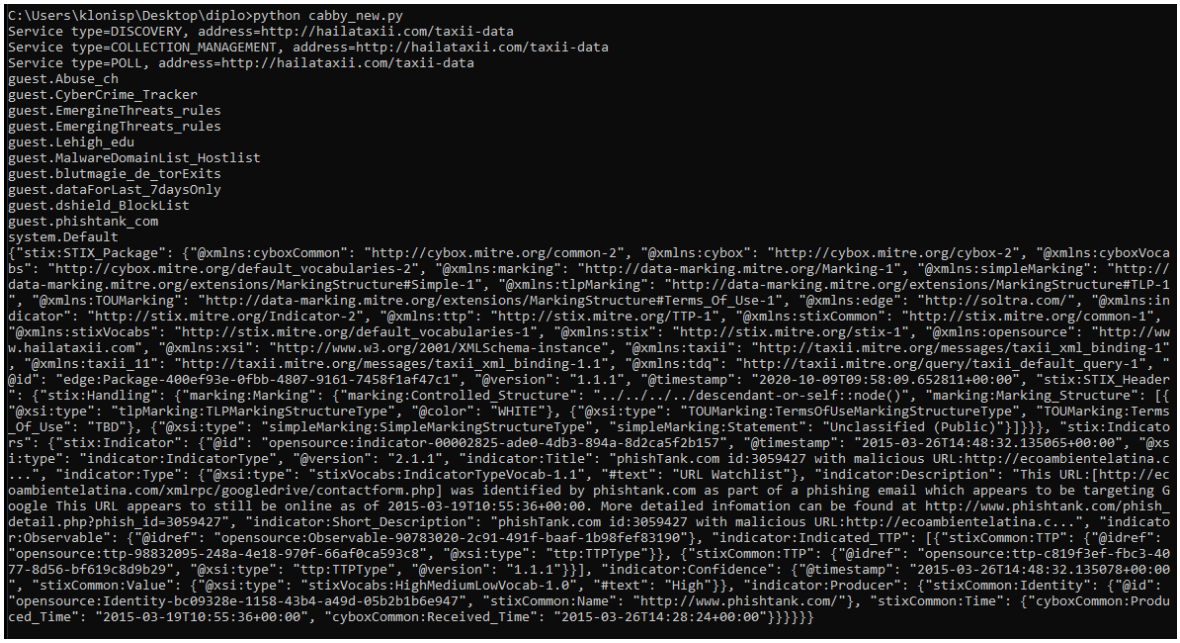
Μια απειλή σε μορφή stix μπορεί να αναπαρασταθεί ως γράφος. Αν ανατρέξουμε στην ιστοσελίδα της oasis για να χρησιμοποιήσουμε το stix visualizer και κάνουμε επικόλληση στο κατάλληλο παράθυρο μία εγγραφή stix από αυτές πού έχουμε στη βάση προκύπτει κάτι τέτοιο:



Η ιστοσελίδα που μπορεί κανείς να χρησιμοποιήσει το stix visualizer είναι η https://oasis-open.github.io/cti-stix-visualization/ .

3.2.3 Το τρίτο script

Στο τρίτο script χρησιμοποιούμε τη βιβλιοθήκη cabby της python για να αντλήσουμε δεδομένα σε stix από το repository της **HAIL A TAXII** που περιέχει απειλές σε stix format. Αφού εκτελέσουμε τον κατάλληλο κώδικα παίρνουμε για δοκιμή μία εγγραφή σε stix και η εικόνα του τερματικού που τρέχει το script είναι αυτή:

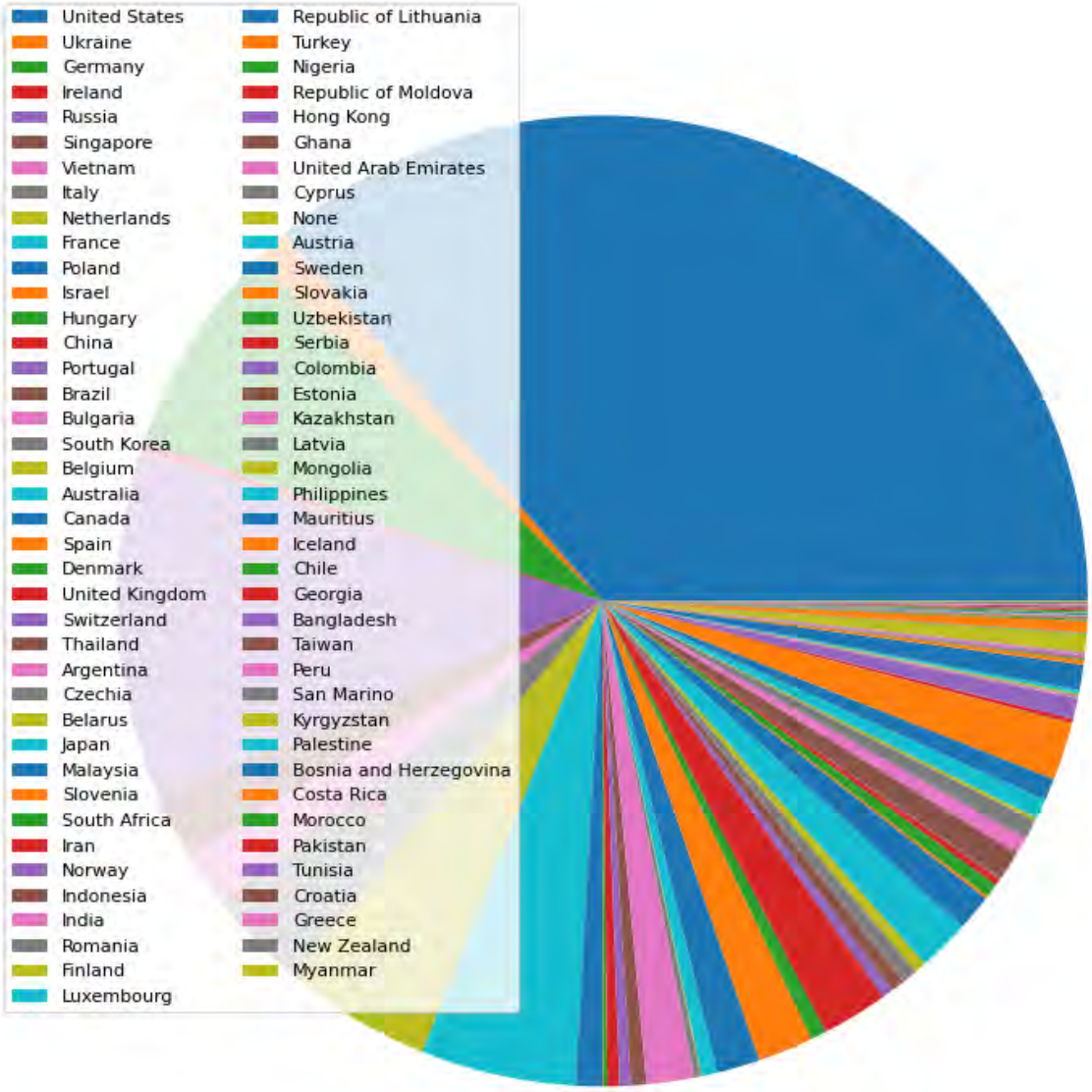


Στην αρχή παίρνουμε feedback για το service discovery, στη συνέχεια εμφανίζονται τα διαθέσιμα feeds και τέλος η εγγραφή σε μορφή stix.

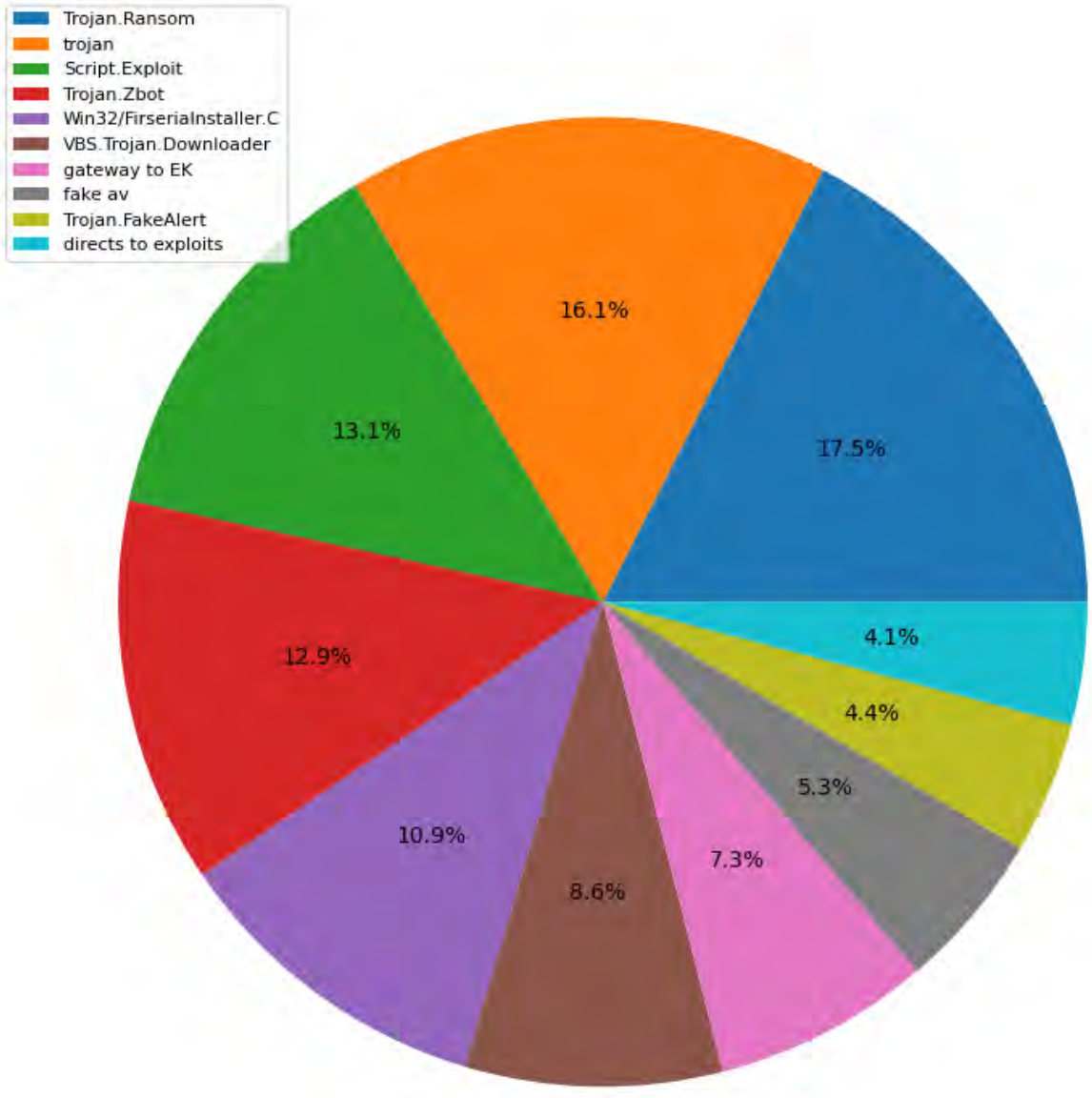
3.2.4 Το τέταρτο script – γραφήματα

Από το δεδομένα που υπάρχουν στη βάση μπορούν να βγουν κάποια ενδιαφέροντα στοιχεία στατιστικού ενδιαφέροντος. Το τέταρτο και τελευταίο script της εργασίας περιέχει κώδικα με τον οποίο κατασκευάζονται τρία γραφήματα.

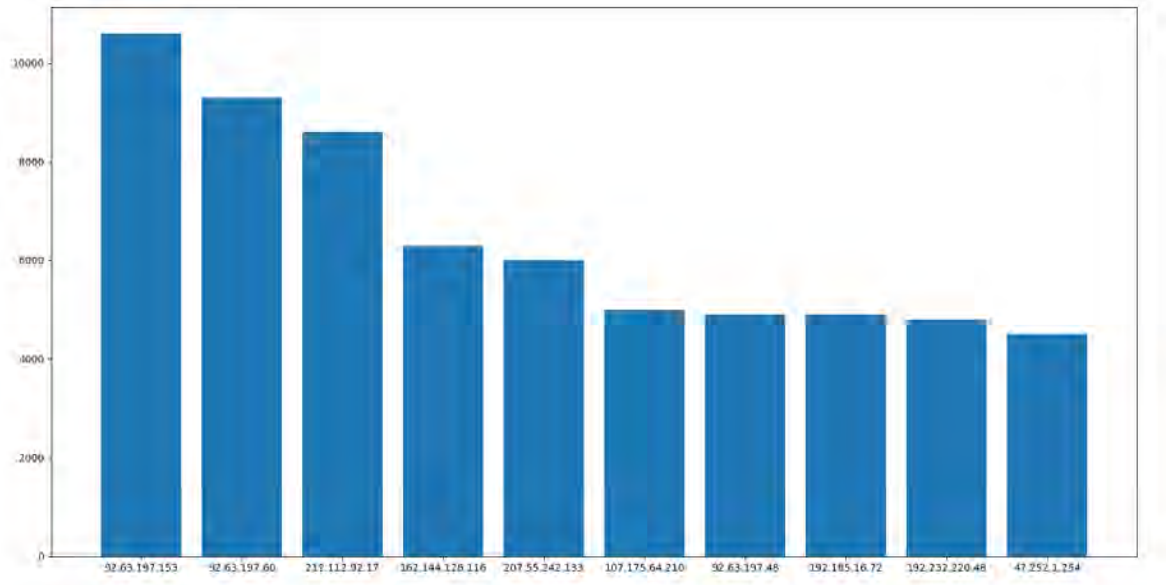
Το πρώτο γράφημα δείχνει τις 79 πρώτες χώρες όπου εντοπίζονται μαζικά μηχανήματα που φιλοξενούν server επικίνδυνων ιστοσελίδων. Η κατανομή των domain σε κάθε χώρα έγινε βάση των πληροφοριών που υπάρχουν στη βάση και αντλήθηκαν από τη Shodan. Το γράφημα είναι το παρακάτω:


United States
Ukraine
Germany
Ireland
Russia
Singapore
Vietnam
Italy
Netherlands
France
Poland
Israel
Hungary
China
Portugal
Brazil
Bulgaria
South Korea
Belgium
Australia
Canada
Spain
Denmark
United Kingdom
Switzerland
Thailand
Argentina
Czechia
Belarus
Japan
Malaysia
Slovenia
South Africa
Iran
Norway
Indonesia
India
Romania
Finland
Luxembourg
Republic of Lithuania
Turkey
Nigeria
Republic of Moldova
Hong Kong
Ghana
United Arab Emirates
Cyprus
None
Austria
Sweden
Slovakia
Uzbekistan
Serbia
Colombia
Estonia
Kazakhstan
Latvia
Mongolia
Philippines
Mauritius
Iceland
Chile
Georgia
Bangladesh
Taiwan
Peru
San Marino
Kyrgyzstan
Palestine
Bosnia and Herzegovina
Costa Rica
Morocco
Pakistan
Tunisia
Croatia
Greece
New Zealand
Myanmar

Το δεύτερο γράφημα που κατασκευάζεται αφορά τις 10 πρώτες απειλές σε συχνότητα που ανιχνεύονται στη βάση δεδομένων που κατασκευάστηκε. Παρακάτω βλέπετε το γράφημα.



Το τελευταίο γράφημα αφορά τις πρώτες δέκα IP διευθύνσεις σε συχνότητα εμφάνισης στη βάση. Ο λόγος που μια IP μπορεί να εμφανιστεί στη βάση περισσότερες από μία φορές είναι διότι μπορεί να φιλοξενείται στο ίδιο μηχάνημα παραπάνω από ένα κακόβουλο domain. Το γράφημα είναι το παρακάτω. Στον άξονα Υ παρουσιάζεται η συχνότητα εμφάνισης μια IP διεύθυνσης ενώ στον άξονα Χ οι δέκα συγκεκριμένες IP διευθύνσεις.



ΒΙΒΛΙΟΓΡΑΦΙΑ

[1]. *Computer History Museum*, 'Timeline of Computer History', [Online], Accessed: October 2020, available: https://www.computerhistory.org/timeline/computers/

[2]. Wikipedia, *'Template: Internet history timeline',* [Online], Accessed: October 2020, available: https://en.wikipedia.org/wiki/Template:Internet_history_timeline

[3]. Wikipedia, '*Cybercrime',* Accessed: October 9th 2020*,* available: *https://en.wikipedia.org/wiki/Cybercrime*

[4]. Monali S. Gaigole *et al*, 'The Study of Network Security with Its Penetrating Attacks and Possible Security Mechanisms', *International Journal of Computer Science and Mobile Computing*, Vol.4 Issue.5, May- 2015, pg. 728-735, available: www.ijcsmc.com

[5]. Malwarebytes, 'All about malware', [Online], Accessed: October 9th 2020, available: https://www.malwarebytes.com/malware/

[6]. Wikipedia, 'Internet protocol suite, [Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Internet_protocol_suite

[7]. Wikipedia, 'Link layer',[Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Link_layer

[8]. Wikipedia, 'Internet layer' ,[Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Internet_layer

[9]. Wikipedia, 'Transport layer', [Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Transport_layer

[10]. Wikipedia, 'Application layer', [Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Application_layer

[11]. R. Ramakrishnan and J. Gehrke,'An overview of database systems', *Database Management Systems, Thessaloniki,* Giola, 2012, pp 3-23

[12]. Wikipedia, 'Database', [Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Database

[13]. Wikipedia, 'Unstructured Data', [Online], Accessed:P October 9th 2020, available: https://en.wikipedia.org/wiki/Unstructured_data

[14]. D. Christodpulakis and A. Foka, Database II, [Online], available: http://www.dblab.upatras.gr/download/courses/db2/2007/dbII_05.pdf , Accessed: October 9th 2020

[15]. WebHarvy, 'What is Web Scraping', [Online], Accessed: October 9th 2020, available: https://www.webharvy.com/articles/what-is-web-scraping.html#:~:text=Web%20Scraping%20(also%20termed%20Screen,in%20table%20(spreadsheet)%20format.

[16]. Wikipedia, 'Web scraping', [Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Unstructured_data

[17]. WordLift, 'What is structured data from a technical standpoint', Accessed: October 9th 2020, available: https://wordlift.io/blog/en/entity/structured-data/

[18]. Wikipedia, 'API', [Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Modular_programming

[19]. Wikipedia, 'Modular programming', [Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/API

[20]. SEKOIA.IO Blog, 'Introduction to STIX and TAXII', [Online], Accessed: October 9th 2020, available: https://medium.com/sekoia-io-blog/stix-and-taxii-c1f596866384

[21]. Wikipedia, 'Google Safe Browsing', [Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Google_Safe_Browsing

[22]. Wikipedia, 'Python (programming language)', [Online], Accessed: October 9th 2020, available: https://en.wikipedia.org/wiki/Python_(programming_language)

[23]. Beautiful soup Documentation, [Online], Accessed: October 9th 2020, available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[24]. PyMOTW, 'urlib2 – Library for opening URLs', [Online], Accessed: October 9th 2020, available: https://pymotw.com/2/urllib2/

[25]. SQLite, 'About SQLite', [Online], Accessed: October 9th 2020, available: https://www.sqlite.org/about.html