ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Ανάλυση μεθόδων κυβερνοεπίθεσης που στηρίζονται σε τεχνητή νοημοσύνη

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ψυχάρη Παναγιώτη - Νικόλαου

Επιβλέπων : Κοκολάκης Σπύρος

Μέλη εξεταστικής επιτροπής: Κρητικός Κυριάκος, Στεργιόπουλος Γιώργος

Σάμος, Ιανουάριος 2021

Η σελίδα αυτή είναι σκόπιμα λευκή

του

Ψυχάρης Παναγιώτης - Νικόλαος

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

Λίστα Σχημάτων

Ακρωνύμια

AI	Artificial intelligence
IT	Information technology
ML	Machine learning
DL	Deep Learning
RNN	Recurrent neural network

Περίληψη

Η τεχνητή νοημοσύνη αφορά την μελέτη που ασχολείται κατά κύριο λόγο με τα αντικείμενα της επιστήμης των υπολογιστών και των μαθηματικών με σκοπό την σχεδίαση και δημιουργία ευφυών υπολογιστικών συστημάτων. Με το στοιχείο της ευφυΐας, αποβλέπει στην ανάπτυξη τεχνικών σε συστημάτα τα οποία θα διαθέτουν χαρακτηριστικά που προσομοιώνουν την ανθρώπινη νοητική ικανότητα και συμπεριφορά. Παρά το γεγονός ότι τα πρώτα βήματα στον κλάδο της τεχνητής νοημοσύνης πραγματοποίηθηκαν πριν από περίπου εξήντα χρόνια, η συνεχής και έντονη ερευνητική δραστηριότητα που ακολούθησε, τουλάχιστον σε θεωρητικό επίπεδο, έθεσε τα θεμέλια για την δημιουργία ενός κλάδου που με την έλευση ισχυρών υπολογιστών θα κατάφερνε να παρουσιάσει τις πρωτοφανείς ικανότητές του. Τα τελευταία χρόνια έχει αρχίσει να χρησιμοποιείται σε διάφορους τομείς και τα υπολογιστικά συστήματα που την χρησιμοποιούν γίνονται πιο διαδεδομένα και χρήσιμα. Αυτές οι νέες δυνατότητες διευκολύνουν πολλές διαδικασίες και εφαρμόζονται για να κάνουν το διαδικτυακό περιβάλλον ασφαλέστερο και πιο προσιτό. Αν και οι εφαρμογές αυτής της τεχνολογίας προσφέρουν σημαντικά οφέλη, μπορούν να χρησιμοποιηθούν και με κακόβουλο σκοπό. Παρά τις πολλές μελέτες για την τεχνητή νοημοσύνη στην ασφάλεια πληροφοριακών συστημάτων, η ερευνητική κοινότητα δεν έχει ασχοληθεί αρκετά με τις επιθέσεις που μπορούν να πραγματοποιηθούν στον κυβερνοχώρο αξιοποιώντας την. Οι στόχοι αυτής της έρευνας είναι η ανάλυση επιθέσεων που στηρίζονται σε τεχνητή νοημοσύνη με ποικίλους τρόπους ξεκινώντας με την παρουσίαση τους, τον λόγο που θεωρούνται σημαντικές, την παρουσίαση παραδειγμάτων για να κατανοηθεί ο τρόπος λειτουργίας τους, ο βαθμός επικινδυνότητάς τους και τέλος, αναφέρονται οι μέθοδοι που μπορούν να χρησιμοποιηθούν για την αντιμετώπισής τους. Μέσα από την μελέτη των επιθέσεων αποφασίστηκαν να οριστούν οι συγκεκριμένοι στόχοι επειδή η επίτευξη τους θα είχε ως αποτέλεσμα να γίνουν κατανοητές οι πολύπλευρες ιδιότητες που φέρουν οι εφαρμογές της τεχνητής νοημοσύνης. Άλλα οφέλη από τους στόχους αυτούς είναι ότι θα γνωστοποιηθεί η ευελιξία αυτής της τεχνολογίας που επιτρέπει στους ειδικούς να την χρησιμοποιήσουν με διάφορους τρόπους για την ολοκλήρωση των επιθυμητών ενεργειών ακόμα κι αν πρόκειται για κακόβουλη δραστηριότητα και επίσης θα μπορεί ο αναγνώστης να προσδιορίσει μέχρι ποιο σημείο φτάνουν οι δυνατότητες της τεχνητής νοημοσύνης στον χώρο της ασφάλειας και αν αυτές οι δυνατότητες οδηγούν σε αποτελέσματα που είναι σπουδαιότερα από αυτά που πετυχαίνουν οι υπάρχουσες συμβατικές τεχνικές κυβερνοεπιθέσεων. Όσον αφορά την ανάλυση, έγινε μια εισαγωγή στα στάδια που περιλαμβάνει μια κυβερνοεπίθεση ξεκινώντας από την αναζήτηση στόχου, συνεχίζοντας στην συλλογή πληροφοριών, έπειτα στην δημιουργία προφίλ στόχου, την ανίχνευση ευπάθειας και τέλος τις ίδιες τις μεθόδους επίθεσης. Στην συνέχεια, πραγματοποιήθηκε λεπτομερής περιγραφή 6 μεθόδων κυβερνοεπίθεσης που στηρίζονται σε τεχνητή νοημοσύνη οι οποίες ταξινομήθηκαν σε 3 διαφορετικές κατηγορίες με βάση τον τρόπο προσέγγισης που ακολουθούν για να αποκτήσουν πρόσβαση και να διεισδύσουν σε ενα σύστημα. Η πρώτη κατηγορία αφορά τις επιθέσεις που χρησιμοποιούν κακόβουλα bots που έχουν σχεδιαστεί για να λειτουργούν με έναν αυτοματοποιημένο τρόπο και στηρίζονται σε τεχνητή νοημοσύνη για να πραγματοποιούν συγκεκριμένες διαδικασίες με μεγαλύτερη ακρίβεια. Η

δεύτερη κατηγορία εστιάζει σε επιθέσεις που αποσκοπούν στην απόκτηση μη εξουσιοδοτημένης πρόσβασης σε συστήματα με πιο αποτελεσματικούς και ευφυείς τρόπους. Τέλος, η τρίτη κατηγορία περιέχει τις μεθόδους που χρησιμοποιούν adversarial training για να αποφύγουν τα συστήματα ανίχνευσης και να μαθαίνουν να διεξάγουν αυτοματοποιημένες επιθέσεις. Κάθε μια από τις παραπάνω κατηγορίες διαθέτει λεπτομερή ανάλυση δύο μεθόδων επίθεσης όπου μέσα από την κατανόηση τους προσφέρεται μια διαφορετική οπτική για την ασφάλεια που μπορεί να βοηθήσει στον σχεδιασμό πιο αποτελεσματικών μεθόδων αντιμετώπισης τους. Συμπερασματικά, από τη ανάλυση που πραγματοποιήθηκε αποδεικνύεται ότι η τεχνητή νοημοσύνη αποτελεί σημαντικό εργαλείο στη διαδικασία σχεδιασμού και πραγματοποίησης των κυβερνοεπιθέσεων καθώς οι ιδιότητες που την διακρίνουν από τις παραδοσιακές προσεγγίσεις είναι τόσο εξελιγμένες που πολλές φορές ο μόνος αποτελεσματικός τρόπος ενίσχυσης της άμυνας των συστημάτων για αυτές τις επιθέσεις είναι να χρησιμοποιούνται επίσης ευφυείς αλγορίθμοι.

Λέξεις Κλειδιά: *Τεχνητή νοημοσύνη, Κυβερνοασφάλεια, Πληροφοριακά συστήματα*

Abstract

Artificial intelligence is the study that deals primarily with the subjects of computer science and mathematics in order to design and create intelligent computer systems. With the element of intelligence, it aims to develop techniques in systems that will have features that simulate human mental ability and behavior. Although the first steps in the field of artificial intelligence were made about sixty years ago, the continuous and intense research activity that followed, at least in theory, laid the foundations for the creation of a field that with the advent of powerful computers would be able to show its unprecedented abilities. In recent years it has started to be used in various fields and the computer systems that use it are becoming more widespread and useful. These new features facilitate many processes and are implemented to make the online environment safer and more accessible. Although applications of this technology offer significant benefits, they can also be used with malicious intent. Despite the many studies on artificial intelligence being utilized in information systems security, the research community has not dealt enough with the attacks in cyberspace that are powered by it. The objectives of this research are to analyze attacks based on artificial intelligence in a variety of ways, starting with their presentation, the reason they are considered important, the presentation of examples to understand how they work, their degree of danger and finally, the methods of dealing with them.

Through the study of the attacks, it was decided to set these specific goals because achieving them would result in understanding the multifaceted properties of artificial intelligence applications. Another benefit of these goals is improved understanding of the flexibility of this technology which allows experts to use it in various ways to complete the desired actions even if it is a malicious activity and also the reader will be able to determine how far the capabilities of artificial intelligence reach in the field of security and whether these capabilities lead to results that are more important than those achieved by existing conventional cyber-attack techniques. Regarding the analysis, an introduction was made on the stages of a cyber-attack starting from the search for a target, continuing to gathering information, then creating a target profile, detecting vulnerabilities and finally the attack methods themselves. Then, a detailed description of 6 cyber-attack methods that are based on artificial intelligence was performed which were classified into 3 different categories based on the approach they follow to gain access and penetrate a system. The first category concerns attacks that use malicious bots designed to operate in an automated manner and rely on artificial intelligence to perform specific processes with greater accuracy. The second category focuses on attacks aimed at gaining unauthorized access to systems in more efficient and intelligent ways. Finally, the third category contains methods that use adversarial training to avoid detection systems and learn to carry out automated attacks. Each of the above categories has a detailed analysis of two methods of attacks where through their understanding, a different perspective on security is offered that can help design more effective methods of dealing with them. In conclusion, the analysis that was carried out shows that artificial intelligence is an important tool in the process of planning and

performing cyber-attacks as the properties that distinguish it from traditional approaches are so sophisticated that often the only effective way to strengthen the defense of the systems for these attacks are to also implement intelligent algorithms.

Keywords: *Artificial intelligence, Cybersecurity, Information systems*

1

Εισαγωγή

1.1 Κυβερνοασφάλεια και τεχνητή νοημοσύνη

Τα τελευταία χρόνια, έχει σημειωθεί σημαντική αύξηση στη χρήση ηλεκτρονικών συσκευών, συστημάτων και πόρων που δημιουργούν, αποθηκεύουν ή επεξεργάζονται δεδομένα. Τεχνολογίες όπως το Internet of Things (IoT), τις έξυπνες συσκευές, τους αισθητήρες, το cloud computing, τα μεγάλα δεδομένα, το Διαδίκτυο, τα κινητά τηλέφωνα και τις ασύρματες τεχνολογίες, χρησιμοποιούνται στον τομέα της εκπαίδευσης, της υγείας, των επικοινωνιών, τις τραπεζικές υπηρεσίες, τις κυβερνήσεις και τις επιχειρήσεις σε όλο τον κόσμο. Ως εκ τούτου, η ασφάλεια της κοινωνίας εξαρτάται σε μεγάλο βαθμό από την ύπαρξη ασφαλών υποδομών. Ένα τμήμα αυτών των υποδομών λειτουργεί στον κυβερνοχώρο ο οποίος μπορεί να υποβληθεί σε σοβαρές επιθέσεις. Πιο συγκεκριμένα, μια έρευνα αποκάλυψε ότι το 90% των εταιρειών στις Ηνωμένες Πολιτείες ήταν στόχος μιας επίθεσης στον κυβερνοχώρο, με το 80% να υποφέρει σημαντικές οικονομικές απώλειες (Hinde, 2003).

Οι χρήσεις των ψηφιακών τεχνολογιών προκαλούν μεγάλες προκλήσεις στο επίπεδο ασφάλειας, προστασίας δεδομένων και κανονισμών που ακολουθούνται από οργανισμούς για την αντιμετώπιση απειλών στον κυβερνοχώρο. Η παροχή όμως και η διατήρηση της ασφάλειας είναι μια πολύ περίπλοκη εργασία. Προκειμένου να αντιμετωπίσει αυτές τις προκλήσεις, η βιομηχανία ασφάλειας έχει αναπτύξει ειδικές υποδομές, αλγορίθμους, αρχιτεκτονικές και εφαρμογές λογισμικού για την προστασία υπολογιστικών πόρων που περιλαμβάνουν λογισμικό (software), υλικό (hardware), ηλεκτρονικά δεδομένα καθώς και το ίδιο το δίκτυο (network) από μη εξουσιοδοτημένη πρόσβαση ή επιθέσεις ευπάθειας (vulnerability attacks) που προορίζονται για εκμετάλλευση.

Παράλληλα, η ανάπτυξη εργαλείων αυξανόμενης πολυπλοκότητας παρατηρείται συνεχώς από τους ανθρώπους για την εξυπηρέτηση των αναγκών τους. Οι ηλεκτρονικοί υπολογιστές είναι από πολλές απόψεις, ένα ακόμα εργαλείο. Έχουν την δυνατότητα να εκτελέσουν ένα είδος αριθμητικών διαδικασιών με μεγάλη ταχύτητα και αξιοπιστία. Παρ' όλα αυτά, ιδιαίτερο ενδιαφέρον είχε η ιδέα της δημιουργίας ενός υπολογιστικού συστήματος που να μπορεί να σκεφτεί. Για αυτόν τον λόγο, τα τελευταία χρόνια έχει σημειωθεί εκτεταμένη έρευνα στον τομέα της τεχνητής νοημοσύνης η οποία αποτελεί μια από τις αναδυόμενες τεχνολογίες που προσπαθεί να προσομοιώσει την ανθρώπινη λογική. Οι εφαρμογές αυτής της τεχνολογίας έχουν πραγματοποιηθεί σε ένα μεγάλο εύρος τομέων με εντυπωσιακά αποτελέσματα.

1.2 Αντικείμενο διπλωματικής

1.2.1 Σκοπός της διπλωματικής

Ο σκοπός αυτής της εργασίας είναι να πραγματοποιηθεί ανάλυση των μηχανισμών κυβερνοασφάλειας που χρησιμοποιούν αλγορίθμους τεχνητής νοημοσύνης καθώς και οι εφαρμογές τους. Η σημασία αυτής της ανάλυσης πηγάζει από την ολοένα και αυξανόμενη ανάγκη για προστασία ενός μεγάλου εύρους προβλημάτων με θεματολογία που καλύπτει από νομικά έως οικονομικά ζητήματα.

Η ανάπτυξη της τεχνολογίας και συγκεκριμένα η διαθεσιμότητα της υπολογιστικής ισχύος σε προσιτό κόστος έχει επιτρέψει να δημιουργηθούν εκλεπτυσμένες μέθοδοι τεχνητής νοημοσύνης που έχουν γνωρίσει άνθιση και στον τομέα της ασφάλειας πληροφοριακών συστημάτων. Ο μεγάλος αριθμός από πεδία εφαρμογών που έχουν επωφεληθεί από την τεχνητή νοημοσύνη έχει ως επόμενο η χρήση της να αξιοποιείται τόσο σε μεθόδους με κακόβουλο σκοπό σε κυβερνοεπιθέσεις αλλά και ως μέσο ενίσχυσης της ασφάλειας ενάντια αυτών των επιθέσεων.

Στην συνέχεια θα ακολουθήσει ανάλυση της κυβερνοασφάλειας, της τεχνητής νοημοσύνης καθώς και τι μπορεί να επιφέρει ένας συνδυασμός αυτών των δύο.

1.2.2 Στόχοι της διπλωματικής

Πιο συγκεκριμένα, οι στόχοι της διπλωματικής είναι να εντοπιστούν και να αναλυθούν τα ακόλουθα ερωτήματα:

- Ποιές είναι οι μέθοδοι τεχνητής νοημοσύνης που χρησιμοποιούνται σε κυβερνοεπιθέσεις;
- Με ποιο τρόπο αμύνονται οι επιχειρήσεις/οργανισμοί σήμερα αξιοποιώντας μεθόδους τεχνητής νοημοσύνης;
- Ποιά είναι η απόδοση αυτών των μέτρων προστασίας;

1.3 Δομή της διπλωματικής

Μετά την εισαγωγή, στο κεφάλαιο 2 παρουσιάζεται το αντικείμενο της κυβερνοασφάλειας και της τεχνητής νοημοσύνης. Η κατανόηση των βασικών στοιχείων τους θα είναι χρήσιμη για να είναι πιο ευανάγνωστα τα τμήματα που αναφέρονται στον τρόπο διεξαγωγής των κυβερνοεπιθέσεων στην συνέχεια. Στο κεφάλαιο 3 περιγράφεται η μεθοδολογία καθώς και η ανάλυση των τεχνικών κυβερνοεπιθέσεων και οι τρόποι αντιμετώπισής τους. Συνεχίζοντας, στο κεφάλαιο 4 αναφέρονται τα συμπεράσματα από την ανάλυση και τέλος στο κεφάλαιο 5 είναι οι αναφορές που χρησιμοποιήθηκαν.

2

Θεωρητικό υπόβαθρο

2.1 Κυβερνοασφάλεια

Η ασφάλεια στον κυβερνοχώρο αποτελείται από το σύνολο των μεθόδων που επιδιώκουν να παρέχουν προστασία για να μειώσουν τον κίνδυνο από επιθέσεις καθώς και να διατηρήσουν την ακεραιότητα των δικτύων. Το υποπεδίο της ασφάλειας πληροφορικής προβλέπεται να αυξηθεί γρήγορα στο άμμεσο μέλλον με εκτιμήσεις ειδικών να υπολογίζουν ότι θα ξεπεράσει το 10% του συνολικού μεριδίου αγοράς πληροφορικής έως το έτος 2025, το οποίο σήμερα ανέρχεται σε 4 τρισεκατομμύρια δολάρια (Shandilya, 2020).

Η ασφάλεια στον κυβερνοχώρο έχει γίνει πρωταρχικό μέλημα τόσο για οργανισμούς όσο και για παγκόσμιες κυβερνήσεις. Πιο αναλυτικά, η ασφάλεια δεδομένων παρέχει μια άλλη οπτική σε όλες τις πρακτικές ασφάλειας είτε αφορά προστασία από επιτιθέμενους με σκοπό την καταστροφή ή την κλοπή δεδομένων, είτε προστασία από εταιρίες που διαχειρίζονται δεδομένα χρηστών. Στην πρώτη περίπτωση απαιτείται η ανάλυση απειλών που αφορά τη διαδικασία αξιολόγησης ύποπτων ενεργειών στον κυβερνοχώρο (Solanki, 2020). Μια απειλή για την ασφάλεια στον κυβερνοχώρο μπορεί να είναι καταστροφική επειδή συνήθως αφορά μη εξουσιοδοτημένη πρόσβασση σε δεδομένα. Στην δεύτερη όμως περίπτωση που ασχολείται με την προστασία δεδομένων χρηστών από εταιρίες, η ασφάλεια επιτυγχάνεται με πολιτικές όπως ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR). Μέσα από αυτόν τον κανονισμό, καθίσταται απαραίτητο για τους προμηθευτές λογισμικού να τηρούν τις ανάγκες των σχετικών νόμων και να διασφαλίζουν την ποιότητα σε σχέση με την ασφάλεια.

Η ασφάλεια στον κυβερνοχώρο είναι σημαντική διότι οποιοσδήποτε μπορεί να είναι πιθανός στόχος επίθεσης. Επιχειρηματικοί οργανισμοί, κυβερνητικοί φορείς, ιδρύματα, ακόμη και άτομα με ασήμαντη ταυτότητα βρίσκονται σε κίνδυνο με αποτέλεσμα η εφαρμογή προληπτικών

μέτρων να κρίνεται απαραίτητη. Οι κυβερνοεπιθέσεις μπορούν να περιλαμβάνουν απειλές όπως άρνηση υπηρεσίας (DoS attack), κακόβουλα λογισμικά, ιούς λογισμικού και μηνύματα ηλεκτρονικού "ψαρέματος". Αυτές οι απειλές στοχεύουν οντότητες με διαδικτυακή παρουσία και όταν αφορούν υπηρεσίες όπως τραπεζικά συστήματα μπορούν να επηρεάσουν μεγάλο μέρος του πληθυσμού που τις χρησιμοποιεί.

Κατά τον σχεδιασμό ενός πλάνου ασφάλειας είναι σημαντικό να δωθεί σημασία σε κάποιους συγκεκριμένους τομείς. Οι κύριοι τομείς που καλύπτονται από την ασφάλεια στον κυβερνοχώρο είναι η ασφάλεια εφαρμογών, η ασφάλεια καταστροφών, η ασφάλεια πληροφοριών και η ασφάλεια δικτύου. (Deepa, 2014). Συγκεκριμένα, η ασφάλεια εφαρμογών εστιάζει στην ανάπτυξη μέτρων ασφαλείας για την προστασία από απειλές που εντοπίζονται στο σχεδιασμό, την ανάπτυξη εφαρμογών, την εγκατάσταση και τη συντήρηση της εφαρμογής. Στη συνέχεια, η ασφάλεια καταστροφών ασχολείται με την ανάπτυξη διαδικασιών ασφάλειας στον κυβερνοχώρο που περιλαμβάνουν την ανάπτυξη σχεδίων εκτίμησης κινδύνου, τον καθορισμό προτεραιοτήτων και τον καθορισμό στρατηγικών αποκατάστασης σε περίπτωση καταστροφής. Ακολούθως, η ασφάλεια πληροφοριών επικεντρώνεται στην προστασία των πληροφοριών από μη εξουσιοδοτημένη πρόσβαση. Τέλος, η ασφάλεια του δικτύου έχει ως κύριο στόχο την εξασφάλιση προστασίας, ακεραιότητας, χρηστικότητας και αξιοπιστίας στη συνδεσιμότητα του δικτύου. Τα στοιχεία της ασφάλειας του δικτύου περιλαμβάνουν λογισμικό προστασίας από ιούς (anti-virus), τείχη προστασίας (firewalls), εικονικό ιδιωτικό δίκτυο (vpn) και σύστημα πρόληψης εισβολών για τον εντοπισμό απειλών που εξαπλώνονται ταχέως στα δίκτυα υπολογιστών (Padmavathi, 2019).

*2.1.1 **Οι τρεις πτυχές της κυβερνοασφάλειας***

Η ασφάλεια στον κυβερνοχώρο στηρίζεται στην εμπιστευτικότητα, την ακεραιότητα και την διαθεσιμότητα. Η σημασία τους γίνεται αντιληπτή και από το γεγονός οτι οι επαγγελματίες στον τομέα της ασφάλειας αξιολογούν τις απειλές με βάση τον πιθανό αντίκτυπο που έχουν στην εμπιστευτικότητα, την ακεραιότητα και τη διαθεσιμότητα των δεδομένων, των εφαρμογών και των συστημάτων ενός οργανισμού.

Σήμερα, στην εποχή της διαδικτύου, η πρόσβαση σε πληροφορίες είναι πιο σημαντική από ποτέ. Γι’ αυτό τον λόγο, στο πλαίσιο της ασφάλειας πληροφοριών, η εμπιστευτικότητα εξασφαλίζει ότι οι πληροφορίες που πρέπει να παραμείνουν μυστικές, παραμένουν μυστικές και ότι μόνο τα άτομα που έχουν εξουσιοδότηση να έχουν πρόσβαση μπορούν να έχουν πρόσβαση. Η μη εξουσιοδοτημένη πρόσβαση σε εμπιστευτικές πληροφορίες μπορεί να έχει καταστροφικές συνέπειες, όχι μόνο στις εφαρμογές εθνικής ασφάλειας, αλλά και στο εμπόριο και τη βιομηχανία. Οι κύριοι μηχανισμοί προστασίας της εμπιστευτικότητας στα συστήματα πληροφοριών είναι η κρυπτογραφία και οι έλεγχοι πρόσβασης. Παραδείγματα απειλών για την εμπιστευτικότητα είναι κακόβουλο λογισμικό, εισβολείς, κοινωνική μηχανική, μη ασφαλή δίκτυα και συστήματα με κακή διαχείριση.

Σχετικά με την δεύτερη πτυχή που αφορά την ακεραιότητα, πραγματοποιείται έλεγχος της αξιοπιστίας, της προέλευσης, της πληρότητας και της ορθότητας των πληροφοριών καθώς και της πρόληψης της ακατάλληλης ή μη εξουσιοδοτημένης τροποποίησης των πληροφοριών. Η ακεραιότητα στο πλαίσιο ασφάλειας πληροφοριών είναι σημαντικό να εξασφαλίζεται όχι μόνο

στην ίδια την πληροφορία αλλά και στην προέλευση της - δηλαδή στην ακεραιότητα της πηγής όπου προέρχεται.

Οι μηχανισμοί προστασίας της ακεραιότητας ομαδοποιούνται σε δύο ευρείς τύπους: προληπτικούς μηχανισμούς, όπως ελέγχους πρόσβασης που αποτρέπουν τη μη εξουσιοδοτημένη τροποποίηση πληροφοριών και μηχανισμούς ντετέκτιβ, οι οποίοι αποσκοπούν στην ανίχνευση μη εξουσιοδοτημένων τροποποιήσεων όταν οι προληπτικοί μηχανισμοί έχουν πλέον αποτύχει.

Οταν ο έλεγχος της ακεραιότητας αφορά δεδομένα, επικεντρώνεται στο αν έχουν τροποποιηθεί με μη εξουσιοδοτημένο τρόπο. Γενικότερα, η ακεραιότητα των δεδομένων καλύπτει τα στάδια της αποθήκευσης τους, της επεξεργασίας τους, όπως επίσης και της μεταφοράς τους. Στην περίπτωση που ο έλεγχος της ακεραιότητας αφορά συστήματα, αυτό που ελέγχει είναι η ποιότητα με την οποία ένα σύστημα είναι σε θέση να εκτελεί την επιδιωκόμενη λειτουργία του απαλλαγμένο από μη εξουσιοδοτημένο χειρισμό.

Τέλος, η τρίτη πτυχή που αφορά την διαθεσιμότητα, εξασφαλίζει την έγκαιρη και αξιόπιστη πρόσβαση και χρήση πληροφοριών. Η διαθεσιμότητα πληροφοριών, αν και συνήθως αναφέρεται τελευταία, δεν παύει να έχει υψηλή σημασία στην ασφάλεια τους. Αυτό γίνεται κατανοητό από το γεγονός ότι δεν υπάρχει μεγάλη σημασία στην εμπιστευτικότητα και ακεραιότητα εάν οι εξουσιοδοτημένοι χρήστες πληροφοριών δεν μπορούν να έχουν πρόσβαση σε αυτές και να τις χρησιμοποιήσουν. Το ίδιο συμβαίνει και στην περίπτωση όπου υπάρχουν μέθοδοι ασφάλειας όπως εξελιγμένης κρυπτογράφησης και ελέγχοι πρόσβασης αλλά την ίδια ώρα αυτές οι πληροφορίες που προστατεύονται δεν είναι προσβάσιμες από εξουσιοδοτημένους χρήστες όταν τις χρειάζονται. Επομένως, παρά το ότι αναφέρεται τελευταία στην τριάδα, η διαθεσιμότητα είναι εξίσου σημαντική και απαραίτητη για την ασφάλεια των πληροφοριών, όπως η εμπιστευτικότητα και η ακεραιότητα. Οι επιθέσεις που μπορούν να σημειωθούν κατά της διαθεσιμότητας είναι γνωστές ως επιθέσεις άρνησης υπηρεσίας (DoS) και συζητούνται στο Κεφάλαιο 3.

Οι τρεις πτυχές που αναφέρθηκαν παραπάνω μπορούν να επηρεαστούν από πολλούς παράγοντες. Για παράδειγμα, οι φυσικές και ανθρωπογενείς καταστροφές προφανώς μπορούν να επηρεάσουν τη διαθεσιμότητα, καθώς και την εμπιστευτικότητα και την ακεραιότητα των πληροφοριών. Αν και η συχνότητα και η σοβαρότητά τους διαφέρουν σε μεγάλο βαθμό - οι φυσικές καταστροφές είναι σπάνιες αλλά σοβαρές, ενώ τα ανθρώπινα λάθη είναι συχνά αλλά συνήθως όχι τόσο σοβαρά όσο οι φυσικές καταστροφές. Και στις δύο περιπτώσεις, ο σχεδιασμός για την εξασφάλιση της συνεχής λειτουργίας των επιχειρήσεων όπως και της αποκατάστασης τους όταν συμβαίνουν καταστροφές, περιλαμβάνει τη χρήση τακτικών και αξιόπιστων αντιγράφων ασφαλείας (back ups) που αποσκοπούν στην ελαχιστοποίηση των απωλειών.

*2.1.2 **Υπηρεσίες ασφάλειας***

Για να επιτευχθεί η εμπιστευτικότητα (confidentiality), η ακεραιότητα (integrity) και η διαθεσιμότητα (availability), απαιτείται η εφαρμογή ορισμένων υπηρεσιών ασφάλειας. Οι υπηρεσίες αυτές περιλαμβάνουν σημαντικές διαδικασίες που αφορούν τις μεθόδους ταυτοποίησης (identification), αυθεντικοποίησης (authentication) και εξουσιοδότησης (authorization), όπου είναι οι κύριοι ελέγχοι που στοχεύουν στην προστασία της τριάδας που αναφέρθηκε προηγουμένως.

2.1.2.1 Ταυτοποίηση

Η ταυτοποίηση αποτελεί το πρώτο βήμα στην ακολουθία ταυτοποίησης - αυθεντικοποίησης - εξουσιοδότησης που εκτελείται καθημερινά πολλαπλές φορές από

ανθρώπους και υπολογιστές, όταν απαιτείται πρόσβαση σε πληροφορίες ή πόρους επεξεργασίας πληροφοριών. Ενώ τα στοιχεία των συστημάτων ταυτοποίησης διαφέρουν ανάλογα με το ποιος ή τι απαιτείται, ορισμένες ιδιότητες αναγνώρισης συνεχίζουν να ισχύουν ανεξάρτητα από αυτά τα στοιχεία. Μερικές από αυτές τις ιδιότητες είναι το πεδίο εφαρμογής (scope), η τοποθεσία (locality) και η μοναδικότητα των αναγνωριστικών (uniqueness of ID).

2.1.2.2 Αυθεντικοποίηση

Όσον αφορά την αυθεντικοποίηση, η οποία συμβαίνει αμέσως μετά την ταυτόποίηση και πριν από την έγκριση, αποτελεί την διαδικασία όπου επαληθεύει την αυθεντικότητα της ταυτότητας που έχει δηλωθεί στο προηγούμενο στάδιο. Με άλλα λόγια, είναι στο στάδιο αυθεντικοποίησης που αποδεικνύεται ότι κάποιος είναι πράγματι το άτομο ή το σύστημα που ισχυρίζεται ότι είναι.

Υπάρχουν τρεις μέθοδοι αυθεντικοποίησης που στηρίζονται είτε σε ερωτήματα τα οποία αφορούν κάποια γνώση που μπορεί να έχει ο χρήστης, είτε κάποιο φυσικό αντικείμενο που διαθέτει ή μέσω βιομετρικού ελέγχου. Είναι σημαντικό να σημειωθεί ότι ‘αρκετή’ διασφάλιση μπορεί να σημαίνει διαφορετικούς βαθμούς διασφάλισης ανάλογα με το εν λόγω περιβάλλον και εφαρμογή, και επομένως μπορεί να απαιτεί διαφορετικές προσεγγίσεις στον έλεγχο ταυτότητας. Για παράδειγμα, οι απαιτήσεις αθυεντικοποίησης ενός κρίσιμου εθνικού συστήματος ασφαλείας είναι προφανές ότι διαφέρουν από αυτές μιας μικρής εταιρείας (Mohammed, 2017). Επειδή οι διαφορετικές μέθοδοι έχουν διαφορετικό κόστος και ιδιότητες, καθώς και διαφορετικές αποδόσεις επένδυσης, η επιλογή της καταλληλότερης μεθόδου για ένα συγκεκριμένο σύστημα ή οργανισμό θα πρέπει να γίνει αφού εξεταστούν προσεκτικά αυτοί οι παράγοντες.

Από τις μεθόδους αυθεντικοποίησης που αναφέρθηκαν παραπάνω, ο ελέγχος ταυτότητας που στηρίζεται σε κάποια γνώση που μπορεί να έχει ο χρήστης είναι οι κωδικοί πρόσβασης, οι φράσεις πρόσβασης, οι μυστικοί κωδικοί και οι προσωπικοί αριθμοί αναγνώρισης (PIN). Σε αυτές τις μεθόδους ελέγχου ταυτότητας, υπονοείται ότι αν κάποιος γνωρίζει κάτι που υποτίθεται ότι είναι γνωστό μόνο από τον Χ, τότε συμπεραίνεται ότι θα πρέπει να είναι όντως ο Χ (αν και στην πραγματική ζωή αυτό δεν συμβαίνει πάντα). Αυτός ο έλεγχος είναι η πιο συχνά χρησιμοποιούμενη μέθοδος αυθεντικοποίησης χάρη στο χαμηλό κόστος και την εύκολη εφαρμογή σε συστήματα πληροφοριών. Ωστόσο, η χρήση μόνο αυτού του ελέγχου ενδέχεται να μην παρέχει πλήρη προστασία και να μην είναι κατάλληλος για συστήματα που απαιτούν υψηλή ασφάλεια.

Η δεύτερη μέθοδος ελέγχου ταυτότητας είναι επίσης ευρέως χρησιμοποιούμενη και στηρίζεται σε κάποιο φυσικό αντικείμενο που διαθέτει ένας χρήστης. Παράδειγμα αυτής της μεθόδου αποτελούν τα κλειδιά (keys). Όπως χρησιμοποιούνται για να κλειδώσουν και να ξεκλειδώσουν πόρτες, αυτού του τύπου η πιστοποίηση στα συστήματα πληροφοριών συνεπάγεται ότι εάν κάποιος διαθέτετει κάποιο είδος διακριτικού, όπως μια έξυπνη κάρτα ή ένα διακριτικό (token) USB, τότε είναι το άτομο που ισχυρίζεται ότι είναι. Φυσικά, οι ίδιοι κίνδυνοι που ισχύουν για τα κλειδιά ισχύουν επίσης για έξυπνες κάρτες και USB tokens - μπορεί να κλαπεί, να χαθούν ή να καταστραφούν. Αυτή η μέθοδος ελέγχου ταυτότητας περιλαμβάνει ένα επιπλέον εγγενές κόστος ανά χρήστη όπου σε σύγκριση με χρήση μόνο κωδικών πρόσβασης, οι κωδικοί δεν έχουν κάποιο κόστος για την έκδοση νέο κωδικού πρόσβασης.

Η Τρίτη μέθοδος αυθεντικοποίησης αφορά τους βιομετρικούς ελέγχους. Ένας βιομετρικός έλεγχος στηρίζεται σε ένα φυσιολογικό ή χαρακτηριστικό συμπεριφοράς ενός ανθρώπου που μπορεί να διακρίνει ένα άτομο από το άλλο και το οποίο θεωρητικά μπορεί να χρησιμοποιηθεί για την ταυτοποίηση ή την επαλήθευση της ταυτότητας. Οι βιομετρικές μέθοδοι ελέγχου σε αυτή την

περίπτωση περιλαμβάνουν αναγνώριση δακτυλικών αποτυπωμάτων, ίριδας και αμφιβληστροειδούς, καθώς και αναγνώριση φωνής και υπογραφής μεταξύ άλλων. Αν και η λειτουργία τους είναι δυσκολότερο να κατανοηθεί από τις άλλες δύο μεθόδους, όταν χρησιμοποιείται σωστά, μπορεί να συμβάλλει σημαντικά στην ισχύ του ελέγχου. Παρ 'όλα αυτά, η βιομετρική είναι ένα πολύπλοκο θέμα και είναι πολύ πιο σύνθετο να αναπτυχθεί από τις άλλες δύο μεθόδους. Σε αντίθεση με τους άλλους ελέγχους αυθεντικοποίησης, ανεξάρτητα από το αν ένα κάποιος γνωρίζει ή όχι τον κωδικό πρόσβασης ή διαθέτει το απαραίτητο διακριτικό, τα βιομετρικά συστήματα ελέγχου ταυτότητας κατανοούν κατά πόσο κάποιος μοιάζει με το άτομο που ισχυρίζεται ότι είναι. Φυσικά αυτή η μέθοδος απαιτεί πολύ περισσότερη ρύθμιση και διαμόρφωση που εξαρτάται από την εγκατάσταση.

2.1.2.3 Εξουσιοδότηση

Μετά την ταυτοποίηση και την αυθεντικοποίηση στο στάδιο ελέγχου, παρέχεται στους χρήστες ένα σύνολο εξουσιοδοτήσεων που καθορίζουν τι δικαιώματα έχουν και τι μπορούν να κάνουν στο σύστημα. Αυτές οι εξουσιοδοτήσεις ορίζονται συνήθως από την πολιτική ασφάλειας του συστήματος και καθορίζονται από τον διαχειριστή ασφάλειας του συστήματος (Trnka, 2018). Τα προνόμια που παρέχονται σε αυτό το στάδιο μπορεί να κυμαίνονται από τα άκρα, που σημαίνει «να μην επιτρέπεται τίποτα» έως «να επιτρέπονται τα πάντα» και να περιλαμβάνεται οτιδήποτε στο μεταξύ. Όπως είναι κατανοητό, στην διαδικασία αναγνώρισης-πιστοποίησης-έγκρισης, το δεύτερο και το τρίτο στάδιο της εξαρτώνται από το πρώτο στάδιο και ο τελικός στόχος ολόκληρης της διαδικασίας είναι η επιβολή ελέγχου πρόσβασης και ευθύνης (accountability), η οποία περιγράφεται στη συνέχεια.

Μια άλλη σημαντική αρχή της ασφάλειας των πληροφοριών είναι η υπευθυνότητα. Συγκεκριμένα, η υπευθυνότητα αναφέρεται στη δυνατότητα εντοπισμού ενεργειών και συμβάντων που διεξήχθησαν στο παρελθόν, στους χρήστες, τα συστήματα ή τις διαδικασίες που τα πραγματοποίησαν, για να καθορίσουν την ευθύνη για ενέργειες ή παραλείψεις (Eggenschwiler, 2017). Ένα σύστημα μπορεί να μην θεωρείται ασφαλές εάν δεν παρέχει υπευθυνότητα, γιατί θα ήταν αδύνατο να εξακριβωθεί ποιος είναι υπεύθυνος και τι έκανε ή δεν έκανε στο σύστημα χωρίς αυτή τη διασφάλιση. Η λογοδοσία στο πλαίσιο των συστημάτων πληροφοριών παρέχεται κυρίως από αρχεία καταγραφής (logs) και από την παρακολούθηση κινήσεων (audit trail).

Απο τα παραπάνω, τα αρχεία καταγραφής συστήματος και εφαρμογών έχουν τη δυνατότητα να παρέχουν λεπτομερή πληροφορία ως ταξινομημένοι κατάλογοι συμβάντων και ενεργειών και αποτελούν τα κύρια μέσα για την απόδοση ευθύνης στα περισσότερα συστήματα. Ωστόσο, τα αρχεία καταγραφής μπορούν να θεωρηθούν αξιόπιστα μόνο εάν η ακεραιότητά τους είναι εξασφαλισμένη. Με άλλα λόγια, εάν κάποιος μπορεί να γράψει ή και να διαγράψει αρχεία καταγραφής ή το audit trail, δεν θα θεωρηθεί αρκετά αξιόπιστο για να χρησιμεύσει ως βάση για τη απόδοση ευθύνης. Επιπλέον, σε περίπτωση δικτύων ή συστημάτων επικοινωνίας, τα αρχεία καταγραφής πρέπει να έχουν σωστή χρονική σήμανση και ο χρόνος πρέπει να συγχρονίζεται σε ολόκληρο το δίκτυο, ώστε συμβάντα που επηρεάζουν περισσότερα από ένα συστήματα να μπορούν να συσχετιστούν και να αποδοθούν σωστά (Abbott, 2015).

Συνεχίζοντας**,** η μέθοδος του audit trail αν και είναι παρόμοια με τα αρχεία καταγραφής, έχει κάποιες σημαντικές διαφορές. Τα αρχεία καταγραφής εμφανίζουν συνήθως ενέργειες υψηλού επιπέδου, όπως ένα μήνυμα ηλεκτρονικού ταχυδρομείου που παραδίδεται ή μια ιστοσελίδα που προβάλλεται, ενώ το audit trail συνήθως παρουσιάζει πληροφορίες για λειτουργίες χαμηλότερου επιπέδου, όπως το άνοιγμα ενός αρχείου, την εγγραφή σε ένα αρχείο ή την αποστολή ενός πακέτου μέσω ενός δικτύου. Αν και το audit trail παρέχει πιο λεπτομερείς πληροφορίες σχετικά

με τις ενέργειες και τα συμβάντα που έλαβαν χώρα στο σύστημα, δεν είναι απαραίτητα πιο χρήσιμο, με την πρακτική έννοια της λέξης, από τα αρχεία καταγραφής, απλώς και μόνο επειδή η αφθονία λεπτομερειών στο audit trail το καθιστά περισσότερο απαιτητικό σε πόρους και χρονοβόρο για τη δημιουργία, την αποθήκευση και την ανάλυση (Asaro, 1999) . Μια άλλη πτυχή με την οποία διαφέρουν τα αρχεία καταγραφής και το audit trail είναι η πηγή τους: τα αρχεία καταγραφής δημιουργούνται συνήθως και ως επί το πλείστον από συγκεκριμένο λογισμικό ή εφαρμογές του συστήματος, ενώ ένα audit trail συνήθως διατηρείται από το λειτουργικό σύστημα ή τη μονάδα ελέγχου του.

Ένα άλλο στοιχείο της κυβερνοασφάλειας με ιδιαίτερη σημασία είναι η ιδιωτικότητα. Στο πλαίσιο της ασφάλειας των πληροφοριών αναφέρεται συνήθως στα δικαιώματα των ατόμων στο απόρρητο των προσωπικών τους πληροφοριών και στον επαρκή, ασφαλή χειρισμό αυτών των πληροφοριών από τους χρήστες του. Οι προσωπικές πληροφορίες συνήθως αναφέρονται σε πληροφορίες που προσδιορίζουν άμεσα αλλά και έμμεσα έναν άνθρωπο και σε πολλές χώρες, το απόρρητο των προσωπικών πληροφοριών προστατεύεται από νόμους που επιβάλλουν απαιτήσεις σε οργανισμούς που επεξεργάζονται προσωπικά δεδομένα και ορίζουν ποινές για μη συμμόρφωση. Ειδικότερα, η Ευρωπαϊκή Ένωση (ΕΕ) διαθέτει αυστηρή νομοθεσία για την προστασία των προσωπικών δεδομένων, η οποία περιορίζει τον τρόπο με τον οποίο οι οργανισμοί μπορούν να επεξεργάζονται προσωπικές πληροφορίες καθώς και τι μπορούν να κάνουν με αυτές. Το Σύνταγμα των ΗΠΑ εγγυάται επίσης ορισμένα δικαιώματα απορρήτου, αν και η προσέγγιση σε θέματα απορρήτου διαφέρει μεταξύ των Ηνωμένων Πολιτειών και της Ευρώπης.

Σε αυτό το σημείο αξίζει να αναφερθεί η σημασία της μη αποποίησης ευθύνης στο πλαίσιο της ασφάλειας πληροφοριών. Επειδή αποτελεί μία από τις ιδιότητες των κρυπτογραφικών ψηφιακών υπογραφών που προσφέρει τη δυνατότητα απόδειξης εάν ένα συγκεκριμένο μήνυμα έχει υπογραφεί ψηφιακά από τον κάτοχο του ιδιωτικού κλειδιού μιας συγκεκριμένης ψηφιακής υπογραφής (Moniz, 2011) την καθιστά ιδιαίτερα σημαντική. Η μη αποποίηση ευθύνης είναι ένα σχετικά αμφιλεγόμενο θέμα, εν μέρει επειδή είναι σημαντικό σε αυτή την εποχή και επειδή δεν παρέχει απόλυτη εγγύηση: ένας κάτοχος ψηφιακής υπογραφής, ο οποίος μπορεί να θέλει να αρνηθεί μια συναλλαγή κακόβουλα, πάντα να ισχυρίζεστε ότι το κλειδί ψηφιακής υπογραφής του είχε κλαπεί από κάποιον ο οποίος υπέγραψε πραγματικά την εν λόγω ψηφιακή συναλλαγή, αποκηρύσσοντας έτσι τη συναλλαγή.

Τέλος, ένα από τα πιο δύσκολα ζητήματα ασφάλειας πληροφοριών είναι το ζήτημα της λειτουργικότητας έναντι της διασφάλισης. Ο καλύτερος τρόπος για να γίνει αυτό κατανοητό είναι να ανατρέξει κάποιος στη δική του εμπειρία με υπολογιστές: πόσες φορές ένας υπολογιστής απέτυχε να κάνει κάτι που ήταν αναμενόμενο και πόσες φορές έκανε κάτι που δεν ήταν επιθυμητό? Είναι αυτή η διαφορά μεταξύ των προσδοκιών και του τι συμβαίνει στην πραγματικότητα που αναφέρεται ως λειτουργικότητα έναντι διασφάλισης. Ένα συγκεκριμένο σύστημα μπορεί να ισχυριστεί ότι εφαρμόζει έναν μεγάλο αριθμό από έξυπνα χαρακτηριστικά ασφαλείας, αλλά αυτό είναι πολύ διαφορετικό από το να υπάρχει υψηλός βαθμός εμπιστοσύνης ότι πράγματι τα εφαρμόζει σωστά και δεν θα συμπεριφέρεται με απροσδόκητο τρόπο (Shelupanov, 2019). Ένας άλλος τρόπος για να εξεταστεί το ζήτημα της λειτουργικότητας έναντι της διασφάλισης είναι ότι η λειτουργικότητα αφορά το τι μπορεί να κάνει ένα σύστημα και η διασφάλιση είναι αυτό που δεν θα κάνει ένα σύστημα.

2.1.3 Σημασία της ασφάλειας της πληροφορίας στους οργανισμούς

Η προσεκτική εφαρμογή των ελέγχων ασφάλειας πληροφοριών είναι ζωτικής σημασίας για την προστασία ενός οργανισμού καθώς και της νομικής θέσης, του προσωπικού και άλλων υλικών ή άυλων περιουσιακών στοιχείων. Η αδυναμία ενός οργανισμού να επιλέξει και να εφαρμόσει κατάλληλους κανόνες και διαδικασίες ασφάλειας ενδέχεται να έχει αρνητικό αντίκτυπο στην αποστολή του οργανισμού. Ειδικότερα στο σημερινό περιβάλλον που χαρακτηρίζεται από έντονη δραστηριότητα κακόβουλου κώδικα, παραβιάσεων συστημάτων και απειλών εσωτερικών πληροφοριών, τα δημοσιευμένα ζητήματα ασφάλειας ενός οργανισμού μπορεί να έχουν αρνητικές συνέπειες, ιδίως στην κερδοφορία και στη φήμη του.

Σε περιπτώσεις που οι πληροφορίες και τα συστήματα ενός οργανισμού συνδέονται με εξωτερικά συστήματα, οι ευθύνες εκτείνονται πέρα από τα όρια της οργάνωσης. Αυτό μπορεί να απαιτήσει από τη διοίκηση (1) να γνωρίζει ποιο γενικό επίπεδο ή τύπο ασφάλειας χρησιμοποιείται στα εξωτερικά συστήματα ή / και (2) να ζητήσει τη διασφάλιση ότι το εξωτερικό σύστημα παρέχει επαρκή ασφάλεια για τις πληροφορίες και το σύστημα του οργανισμού. Για παράδειγμα, οι πάροχοι υπηρεσιών cloud (cloud service providers) και οι συμμετέχοντες στην αλυσίδα εφοδιασμού cloud ενδέχεται να αναλάβουν τον διαχειριστικό ρόλο για την αποθήκευση, την επεξεργασία και τη μετάδοση πληροφοριών του οργανισμού (Nieles, 2017). Ωστόσο, είναι υποχρέωση του οργανισμού να διασφαλίσει ότι οι πάροχοι υπηρεσιών cloud και οι συμμετέχοντες στην αλυσίδα εφοδιασμού cloud παρέχουν ένα κατάλληλο επίπεδο ασφάλειας για τις πληροφορίες που αποθηκεύονται, υποβάλλονται σε επεξεργασία και μεταδίδονται.

Η παροχή αποτελεσματικής ασφάλειας πληροφοριών απαιτεί μια ολοκληρωμένη προσέγγιση που λαμβάνει υπόψη μια ποικιλία τομέων τόσο εντός όσο και εκτός του πεδίου ασφάλειας πληροφοριών. Αυτή η προσέγγιση εφαρμόζεται σε ολόκληρο τον κύκλο ζωής του συστήματος. Για παράδειγμα, η άμυνα σε βάθος (defense-in-depth) είναι μια αρχή ασφάλειας που χρησιμοποιείται για την προστασία πληροφοριών και συστημάτων από απειλές με την εφαρμογή πολυστρωματικών αντιμέτρων ασφάλειας. Αυτή η αρχή ασφάλειας χρησιμοποιεί διοικητικές άμυνες (π.χ. πολιτικές (policies) και τεχνολογίες ασφάλειας (π.χ. συστήματα εντοπισμού εισβολών, firewalls και λογισμικό προστασίας από ιούς) σε συνδυασμό με φυσικές άμυνες ασφάλειας (π.χ. φρουρούς) για την ελαχιστοποίηση της πιθανότητας μιας επιτυχημένης επίθεσης στο σύστημα. Αυτά τα μέτρα συμβάλλουν στη μείωση της πιθανότητας παραβίασης της ασφάλειας και στην αποφυγή άλλων επιζήμιων επιπτώσεων στην εμπιστευτικότητα, την ακεραιότητα ή τη διαθεσιμότητα. Επίσης έχουν την δυνατότητα να ενημερώνουν τον οργανισμό σχεδόν σε πραγματικό χρόνο μετά την έναρξη μιας επίθεσης (Katzke, 2010).

Τέλος, η ασφάλεια των πληροφοριών δεν είναι μια διαδικασία που θεωρείται στατική. Αντιθέτως, απαιτεί συνεχή παρακολούθηση και διαχείριση για την προστασία της εμπιστευτικότητας, της ακεραιότητας και της διαθεσιμότητας των πληροφοριών καθώς και για να διασφαλιστεί ότι οι νέες ευπάθειες και οι εξελισσόμενες απειλές εντοπίζονται σε σύντομο χρονικό διάστημα και υπάρχει η ανάλογη ανταπόκριση (Dempsey, 2011).

2.2 Τεχνητή νοημοσύνη

Η τεχνητή νοημοσύνη είναι η επιστήμη που συνδυάζει την πληροφορικής με τα μαθηματικά και επικεντρώνεται στη δημιουργία έξυπνων μηχανών και προγραμμάτων. Ο σκοπός της

τεχνητής νοημοσύνης είναι να προσπαθήσει να μιμηθεί την ανθρώπινη συνείδηση και να εκτελέσει καθήκοντα παρόμοια με αυτά που μπορούν τα ανθρώπινα όντα. Αν και η επιστήμη αυτή βρίσκεται σε ικανοποιητικό επίπεδο ώστε να μπορεί να βελτιώσει πολλές διαδικασίες της καθημερινής ζωής, οι προσδοκίες φαίνεται να ξεπερνούν την πραγματικότητα μιας και ακόμα μετά από πολλές δεκαετίες έρευνας, κανένας υπολογιστής δεν έχει καταφέρει να περάσει το Turing Test (ένα μοντέλο για τη μέτρηση της «νοημοσύνης») (McGuire, 2006).

Ένα από τα πιο σημαντικά υποπεδία της τεχνητής νοημοσύνης είναι η μηχανική μάθηση. Όπως ορίστηκε από τον Arthur Samuel το 1959, η μηχανική μάθηση είναι το πεδίο μελέτης που δίνει στους υπολογιστές τη δυνατότητα να μαθαίνουν χωρίς να είναι ρητά προγραμματισμένοι. Ο στόχος της μηχανικής μάθησης είναι η δημιουργία προγραμμάτων των οποίων η απόδοση βελτιώνεται αυτόματα με ορισμένες παραμέτρους εισόδου, όπως δεδομένα και κριτήρια απόδοσης μεταξύ άλλων. Τα προγράμματα αυτά βασίζονται σε δεδομένα για τη λήψη αποφάσεων ή προβλέψεων και πολλές φορές μπορεί να μην γίνεται αντιληπτό αλλά η μηχανική μάθηση χρησιμοποιείται ιδιαίτερα στην καθημερινή μας ζωή, από τη σύσταση προϊόντων σε διαδικτυακές πύλες μέχρι αυτοκίνητα που οδηγούν αυτόνομα (Manure, 2020). Αυτά τα χαρακτηριστικά την καθιστούν ικανή σε διάφορους τομείς και ως εκ τούτου επικεντρώνεται στο σχεδιασμό προγραμμάτων υπολογιστών και μηχανών ικανών να εκτελούν εργασίες που οι άνθρωποι είναι φυσικά καλοί, συμπεριλαμβανομένης της φυσικής κατανόησης της γλώσσας, της κατανόησης του λόγου και της αναγνώρισης της εικόνας.

Στα μέσα του εικοστού αιώνα, η μηχανική μάθηση προέκυψε ως υποσύνολο της τεχνητής νοημοσύνης, παρέχοντας μια νέα κατεύθυνση για το σχεδιασμό της τεχνητής νοημοσύνης αντλώντας έμπνευση από την αντιληπτική κατανόηση του τρόπου λειτουργίας του ανθρώπινου εγκεφάλου (Rosenblatt, 1958). Σήμερα, η μηχανική μάθηση παραμένει βαθιά συνδεδεμένη με την έρευνα για την τεχνητή νοημοσύνη. Ωστόσο, συχνά θεωρείται ευρύτερα ως επιστημονικό πεδίο που εστιάζει στο σχεδιασμό μοντέλων υπολογιστών και αλγορίθμων που μπορούν να εκτελέσουν συγκεκριμένες εργασίες, που συχνά περιλαμβάνουν αναγνώριση προτύπων, χωρίς να χρειάζεται να προγραμματιστούν ρητά.

Η μηχανική μάθηση κυρίως περιλαμβάνει τρεις τύπους:

1. Μηχανική μάθηση με επίβλεψη
2. Μηχανική μάθηση χωρίς επίβλεψη
3. Ενισχυτική μηχανική μάθηση

Οι αλγόριθμοι μάθησης με επίβλεψη και χωρίς επίβλεψη έχουν παρουσιάσει σπουδαίες δυνατότητες στην απόκτηση γνώσεων από μεγάλα σύνολα δεδομένων. Η μάθηση με επίβλεψη αντικατοπτρίζει την ικανότητα ενός αλγορίθμου να γενικεύει τη γνώση από τα διαθέσιμα δεδομένα εισόδου που έχουν ετικέτα (label), έτσι ώστε ο αλγόριθμος να μπορεί να χρησιμοποιηθεί για την πρόβλεψη νέων περιπτώσεων.

Η μάθηση χωρίς επίβλεψη αναφέρεται στη διαδικασία ομαδοποίησης δεδομένων χρησιμοποιώντας αυτοματοποιημένες μεθόδους ή αλγόριθμους σε δεδομένα που δεν έχουν ταξινομηθεί ή κατηγοριοποιηθεί. Σε αυτήν την περίπτωση, οι αλγόριθμοι πρέπει να «μάθουν» τις υποκείμενες σχέσεις ή χαρακτηριστικά από τα διαθέσιμα δεδομένα με βάση παρόμοια χαρακτηριστικά γνωρίσματα που έχουν (Berry, 2020).

2.2.1 Μηχανική μάθηση με επίβλεψη

Η μηχανική μάθηση με επίβλεψη είναι μια από τις πιο συχνές και επιτυχημένες μεθόδους μηχανικής μάθησης. Παρακάτω θα πραγματοποιηθεί αναλυτικότερη περιγραφή για αυτή τη μέθοδο καθώς και από τι απαρτίζεται.

Η μάθηση με επίβλεψη χρησιμοποιείται στην περίπτωση που επιδιώκεται να γίνει πρόβλεψη ενός συγκεκριμένου αποτελέσματος από δεδομένα εισόδου. Στη συνέχεια δημιουργείται ένα μοντέλο μηχανικής μάθησης, το οποίο θα χρησιμοποιηθεί για την εκπαίδευσή του μοντέλου (training dataset). Σκοπός αυτής της διαδικασίας είναι να επιτευχθεί πρόβλεψη για νέα δεδομένα τα οποία δεν είχαν συμπεριληφθεί στα δεδομένα που χρησιμοποιήθηκαν κατά την εκπαίδευση του μοντέλου. Ανάλογα με την κατάσταση των δεδομένων που χρησιμοποιούνται για την εκπαίδευση, μπορεί να κριθεί αναγκαίο να γίνει κάποιας μορφής καθαρισμός των δεδομένων ώστε να προσφέρουν όσο το δυνατόν καλύτερο αποτέλεσμα. Αυτές οι βελτιώσεις τις περισσότερες φορές απαιτούν ανθρώπινη προσπάθεια αλλά όταν ολοκληρωθούν, το αποτέλεσμα θα έχει πολύ μεγάλη χρησιμότητα.

2.2.2 Μηχανική μάθηση χωρίς επίβλεψη

Η δεύτερη οικογένεια αλγορίθμων μηχανικής μάθησης είναι χωρίς επίβλεψη. Σε αυτή την περίπτωση ο αλγόριθμος εκμάθησης λαμβάνει μόνο τα δεδομένα εισόδου και του ζητείται να εξάγει γνώση από αυτά τα δεδομένα.

Στη μηχανική μάθηση, το πρόβλημα της μη επιβλεπόμενης μάθησης είναι αυτό της προσπάθειας εύρεσης συσχέτισης σε δεδομένα χωρίς επισήμανση. Δεδομένου ότι τα παραδείγματα που δίνονται στον αλγόριθμο δεν φέρουν επισήμανση, δεν υπάρχει σφάλμα ή κάποια σχετική ένδειξη-ανταμοιβή (reward) για την ορθότητα ώστε να πραγματοποιηθεί αξιολόγηση της δυνατότητας εύρεσης μιας πιθανής λύσης. Αυτό αποτελεί το χαρακτηριστικό που διακρίνει τη μη επιβλεπόμενη από την επιβλεπόμενη μάθηση. Χωρίς επίβλεψη η εκμάθηση ορίζεται ως η εργασία που εκτελείται από αλγόριθμους που μαθαίνουν από ένα εκπαιδευτικό σύνολο που δεν φέρει σήμανση, χρησιμοποιώντας τα χαρακτηριστικά των εισόδων για κατηγοριοποίηση σύμφωνα με ορισμένα γεωμετρικά ή στατιστικά κριτήρια. Για αυτόν τον λόγο μια σημαντική πρόκληση στην μάθηση χωρίς επίβλεψη είναι η αξιολόγηση του κατά πόσο ο αλγόριθμος έμαθε κάτι χρήσιμο. Επομένως, είναι πολύ δύσκολο να εκτιμηθεί αν ένα μοντέλο είχε καλή απόδοση.

2.2.3 Ενισχυτική μηχανική μάθηση

Η ενισχυτική μηχανική μάθηση στηρίζεται στην ιδέα της μάθησης που επιτυγχάνεται αλληλεπιδρώντας με το περιβάλλον. Όπως και στην περίπτωση του ανθρώπου, όταν ένα βρέφος παίζει, κουνάει τα χέρια του ή κοιτάζει γύρω του, δεν έχει ρητό δάσκαλο, αλλά έχει άμεση αισθητική σχέση με το περιβάλλον του. Αυτή η διαδικασία παράγει πληθώρα πληροφοριών σχετικά με την αιτία και το αποτέλεσμα, για τις συνέπειες των ενεργειών και για το τι χρειάζεται για την επίτευξη στόχων. Για αυτό τον λόγο, η εκμάθηση από την αλληλεπίδραση είναι μια θεμελιώδης ιδέα που βασίζεται σχεδόν σε όλες τις θεωρίες της μάθησης και της νοημοσύνης.

Με άλλα λόγια, η ενισχυτική μηχανική μάθηση περιλαμβάνει την εκμάθηση της ικανότητας να πραγματοποιείται αντιστοίχιση καταστάσεων σε ενέργειες έτσι ώστε να μεγιστοποιηθεί ένα αριθμητικό σήμα ανταμοιβής. Ο εκπαιδευόμενος δεν ενημερώνεται για ποιες ενέργειες πρέπει να λάβει, όπως σε πολλές μορφές μηχανικής μάθησης, αλλά αντ 'αυτού πρέπει να ανακαλύψει ποιες ενέργειες αποδίδουν τη μεγαλύτερη ανταμοιβή κάνοντας δοκιμές. Στις πιο ενδιαφέρουσες περιπτώσεις, οι ενέργειες μπορεί να επηρεάσουν όχι μόνο την άμεση ανταμοιβή, αλλά και την επόμενη κατάσταση και μέσω αυτής, όλες τις επόμενες ανταμοιβές.

Μία από τις προκλήσεις που προκύπτουν στην ενισχυτική μηχανική μάθηση, και όχι σε άλλα είδη μάθησης, είναι η ανταλλαγή μεταξύ εξερεύνησης και εκμετάλλευσης. Για να επιτευχθεί υψηλή ανταμοιβή, ένας πράκτορας ενισχυτικής μάθησης (reinforcement learning agent) πρέπει να προτιμά ενέργειες που έχει δοκιμάσει στο παρελθόν και έχουν αποδόσει υψηλή ανταμοιβή. Για να ανακαλύψει τέτοιες ενέργειες, πρέπει να δοκιμάσει ενέργειες που δεν έχει επιλέξει προηγουμένως. Ο πράκτορας πρέπει να εκμεταλλευτεί αυτό που ήδη γνωρίζει για να λάβει ανταμοιβή, αλλά πρέπει επίσης να διερευνήσει για να κάνει καλύτερες επιλογές δράσης στο μέλλον. Το δίλημμα είναι ότι ούτε η εξερεύνηση ούτε η εκμετάλλευση μπορούν να επιδιωχθούν αποκλειστικά χωρίς να αποτύχουν στο έργο. Ο πράκτορας πρέπει να δοκιμάσει μια ποικιλία ενεργειών και σταδιακά να ευνοεί εκείνες που φαίνεται να είναι οι καλύτερες. Σε μια στοχαστική εργασία, κάθε δράση πρέπει να δοκιμάζεται πολλές φορές για να κερδίσει μια αξιόπιστη εκτίμηση της αναμενόμενης ανταμοιβής της. Αξίζει να αναφερθεί ότι το ζήτημα της εξισορρόπησης της εξερεύνησης και της εκμετάλλευσης δεν προκύπτει στην εποπτευόμενη και χωρίς επίβλεψη μάθηση, τουλάχιστον στις καθαρές μορφές τους.

Εξελίξεις σε αυτό το αντικείμενο έχουν επίσης επηρεαστεί έντονα από την ψυχολογία και τη νευροεπιστήμη, με σημαντικά οφέλη και προς τους δύο κλάδους. Από όλες τις μορφές της μηχανικής μάθησης, η ενισχυτική μηχανική μάθηση είναι η πλησιέστερη στο είδος της μάθησης που κάνουν οι άνθρωποι και άλλα ζώα και πολλοί από τους βασικούς αλγόριθμους που χρησιμοποιούνται αρχικά εμπνεύστηκαν από βιολογικά συστήματα μάθησης.

2.3 Συμπεράσματα

Η ασφάλεια στον κυβερνοχώρο περιλαμβάνει τεχνολογίες, αρχιτεκτονικές, υποδομές και εφαρμογές λογισμικού που έχουν σχεδιαστεί για την προστασία υπολογιστικών πόρων από επιθέσεις. Η ασφάλεια στον κυβερνοχώρο επικεντρώνεται σε τέσσερεις κύριους τομείς: την ασφάλεια εφαρμογών, την ασφάλεια καταστροφών, την ασφάλεια πληροφοριών και την ασφάλεια

δικτύου. Πολλοί μέθοδοι ασφάλειας έχουν σχεδιαστεί από ερευνητές για να εξασφαλιστεί προστασία στον κυβερνοχώρο από ανεπιθύμητους εισβολείς και ευαισθησίες. Παρ' όλα αυτά, η απόδοση των παραδοσιακών μεθόδων ασφάλειας δεν επαρκεί λόγω των συνεχώς εξελισσόμενων τύπων επιθέσεων που στοχεύουν υπολογιστές και δίκτυα.

Με την πρόοδο που έχει επιτευχθεί στον τομέα της τεχνητής νοημοσύνης, έχει ως αποτέλεσμα να έχουν αρχίσει να χρησιμοποιούνται οι ευφυείς αλγόριθμοι ως ισχυρό όπλο στις κυβερνοεπιθέσεις καθώς και στην βελτίωση της απόδοσης των συστημάτων ασφάλειας, την ανίχνευση απειλών, την ανίχνευση ευπάθειας σε IoT και ιστότοπους κοινωνικής δικτύωσης και γενικότερα εφαρμογές ασφάλειας στον κυβερνοχώρο. Σε αυτό το κεφάλαιο πραγματοποιήθηκε περιγραφή στις έννοιες της κυβερνοασφάλειας και της τεχνητής νοημοσύνης. Τα βασικά στοιχεία των δύο αυτών κλάδων που αναλύθηκαν σε αυτό το κεφάλαιο είναι χρήσιμα για την καλύτερη κατανόηση των μεθόδων που χρησιμοποιούνται στις κυβερνοεπιθέσεις και στις τεχνικές αντιμετώπισης τους που θα συζητηθούν λεπτομερώς στη συνέχεια.

3

Τεχνικές Τεχνητής Νοημοσύνης στις κυβερνοεπιθέσεις και την κυβερνοάμυνα

3.1 Μεθοδολογία

3.1.1 Ερευνητικό κίνητρο

Η τεχνητή νοημοσύνη έχει αποδείξει ότι μπορεί να να ωφελήσει με πολλούς τρόπους την κοινωνία και τον άνθρωπο. Η κακόβουλη χρήση αυτής της τεχνολογίας όμως μπορεί να προγραμματιστεί για να προκαλέσει βλάβες με έναν ιδιαίτερα ευφυή τρόπο που είναι δύσκολο να καταπολεμηθεί με συμβατικές προσεγγίσεις και πολλές φορές είναι δύσκολο να κατανοηθεί από τον ανθρώπινο εγκέφαλο.

Κίνητρο αυτής της διπλωματικής είναι η ανάλυση της κακόβουλης χρήσης της τεχνητής νοημοσύνης ως μέσο για την υποστήριξη κακόβουλων δραστηριοτήτων. Η μελέτη αυτή επικεντρώνεται στις μεθόδους που μπορεί να χρησιμοποιηθεί η τεχνητή νοημοσύνη ως όργανο στη διαδικασία επίθεσης η οποία πολλές φορές συναντάται να αποτελεί και βασικό εργαλείο στην κυβερνοάμυνα.

3.1.2 Μέθοδοι αναζήτησης πληροφορίας

Η ερευνητική διαδικασία περιελάμβανε τον καθορισμό ενός ερευνητικού ερωτήματος και τον σχεδιασμό μιας στρατηγικής για τη συλλογή σχετικών εργασιών, καθώς και την ανάλυση και αναφορά των ευρημάτων. Το κύριο ερευνητικό ερώτημα από το οποίο σχεδιάστηκαν τα ερωτήματα για την αναζήτηση πληροφορίας είναι: Ποιές είναι οι τεχνικές τεχνητής νοημοσύνης στις κυβερνοεπιθέσεις και πώς μπορούν να αντιμετωπιστούν;

Σε πρώτο στάδιο, για την συγκέντρωση υλικού πραγματοποιήθηκε αναζήτηση στο internet με βάση τις παρακάτω έννοιες: ‘ai based cyber attacks’, ‘malicious ai in cyber space’, ‘use of machine learning in cyber security’, ‘artificial intelligence assisted methods for cyber security’, ‘use of machine learning in cyber defense’. Η γλώσσα που χρησιμοποιήθηκε στις αναζητήσεις ήταν η Αγγλική ώστε να βρεθούν όσο το δυνατόν περισσότερα αποτελέσματα. Αρχικά, λόγω της ανάγκης εντοπισμού μεθόδων κυβερνοεπίθεσης που εμπεριέχουν τεχνητή νοημοσύνη το οποίο έχει ξεκινήσει να παρατηρείται λίγα χρόνια, κρίθηκε απαραίτητο να εξεταστούν λεπτομερώς όλα τα σχετικά ευρήματα από διάφορες πηγές που παρουσίαζαν ενδιαφέρον, ακόμα και από ιστοσελίδες για την επικαιρότητα στον χώρο της κυβερνοασφάλειας που δεν χαρακτηρίζονται από ιδιαίτερη αξιοπιστία ώστε να βρεθούν κάποιες στοιχειώδεις αναφορές. Αυτή η προσσέγγιση πραγματοποιήθηκε ώστε να ανακαλυφθούν πιο εύκολα οι κατάλληλες μέθοδοι και στη συνέχεια να χρησιμοποιηθούν ως φίλτρο για πιο στοχευμένη αναζήτηση σε πηγές με μεγαλύτερη αξιοπιστία.

Σε δεύτερο στάδιο, αφού εντοπίστηκαν τεχνικές επίθεσης που πληρούσαν την προυπόθεση να χρησιμοποιούν τεχνητή νοημοσύνη, επιλέχτηκαν αυτές που είναι πιο συχνές και πιο καταστροφικές. Στη συνέχεια πραγματοποιήθηκε αναζητήση των τεχνικών σε συνδιασμό με την προσθήκη λέξεων όπως ‘PDF’ ή ‘paper’. Αυτό εξασφάλιζε ότι η πληροφορία θα προέρχεται κατά κύριο λόγο από ερευνητικές εργασίες / journals οι οποίες συνήθως πραγματοποιούνται από άτομα που έχουν ασχοληθεί σοβαρά με το αντικείμενο.

Στο τρίτο και τελευταιο στάδιο πραγματοποιήθηκε έλεγχος της σχετικής πληροφορίας που εντοπίστηκε στις ερευνητικές εργασίες κυρίως μέσω ιστοτοπων όπως το research gate, emerald, κτλ, επιβεβαιώνοντας την πληροφορία με διασταύρωση από άλλες πηγές όπου αυτό ήταν δυνατόν και ύστερα αναζητήθηκε υλικό και σε βιβλία από εκδότες που έχουν φήμη για ποιοτικές εκδόσεις όπως Springer, MIT Press, O’Reilly κ.α. Πολλές φορές χρειάστηκε να χρησιμοποιηθεί πληροφορία που είχε προέλθει από διάφορες πηγές συμπεριλαμβανομένων της ακαδημίας, οργανισμών με αντικείμενο την κυβερνοασφάλεια όπως επίσης και εκδόσεις από τον Διεθνή Οργανισμό Τυποποίησης (ISO).

3.2 Ανάλυση

*3.2.1 **Τα στάδια μιας κυβερνοεπίθεσης***

Οι κυβερνοεπιθέσεις αφορούν την κακόβουλη χρήση κώδικα ή άλλων εργαλείων που επηρεάζουν αρνητικά την ασφάλεια ενός συστήματος και των χρηστών της. Σε αυτό το στάδιο της ανάλυσης, οι κυβερνοεπιθέσεις θα χωριστούν σε διάφορα στάδια για να γίνει πιο κατανοητή ως διαδικασία. Αυτά τα στάδια περιλαμβάνουν την αναζήτηση στόχου, την συλλογή πληροφοριών, την δημιουργία προφίλ στόχου, την ανίχνευση ευπάθειας και τέλος τις μεθόδους επίθεσης.

3.2.1.1 Αναζήτηση στόχου

Από τα πιο χρήσιμα χαρακτηριστικά που διαθέτει η τεχνητή νοημοσύνη είναι η ικανότητα να επεξεργάζεται να ερμηνεύει και να λαμβάνει πολύπλοκες αποφάσεις βάσει δεδομένων που αναλύει. Το γεγονός ότι το διαδίκτυο αποτελείται σε μεγάλο βαθμό από δεδομένα κειμένου, καθιστά την τεχνητή νοημοσύνη ιδιαίτερα σημαντική και έχει ως αποτέλεσμα την εύρεση και τον σχεδιασμό συνεχώς πιο αποδοτικών τρόπων εξόρυξης και ανάλυσης δεδομένων κειμένου (text mining) από πληθόρα επιστημόνων. Αυτές οι διαδικασίες εξυπηρετούν στην εξαγωγή χρήσιμων πληροφοριών που κρύβονται σε κειμένα και αυτό επιτυγχάνεται πλέον με μεγαλύτερη ευκολία μέσω της δυνατότητας εντοπισμού μοτίβων σε μεγάλες ποσότητες πληροφοριών.

Μέσω του διαδικτύου η αναζήτηση πληροφοριών μπορεί να πραγματοποηθεί εύκολα και γρήγορα. Για παράδειγμα, μέσω των ιστοτόπων που ασχολούνται με ειδήσεις, ένα άρθρο για την παραβίαση ασφαλείας μιας υπηρεσίας μπορεί να αυξήσει την πιθανότητα μιας μελλοντικής επίθεσης σε εταιρείες που χρησιμοποιούν την ίδια υπηρεσία (Perera, 2018). Αυτό έχει προκαλέσει την ανάπτυξη μεθόδων που εστιάζουν στην αποτελεσματική αναζήτηση και αναγνώριση του υποψήφιου στόχου μιας μελλοντικής κυβερνοεπίθεσης.

Τα στάδια για την συλλογή πληροφοριών χρησιμοποιώντας τεχνητή νοημοσύνη δεν διαφέρουν σε μεγάλο βαθμό από τις συμβατικές προσσεγγίσεις αλλά η αποδοτικότητα, ο τρόπος διεξαγωγής και οι αμυντικοί μηχανισμοί που απαιτούνται για την παρεμπόδιση τους έχουν αλλάξει εντυπωσιακά τα τελευταία χρόνια.

3.2.1.2 Συλλογή πληροφοριών

Το πρώτο στάδιο για την αναζήτηση στόχου είναι η συλλογή πληροφοριών και αποτελεί αναπόσπαστο τμήμα του σχεδιασμού της επίθεσης. Δραστηριότητες που μπορεί να περιλαμβάνει είναι η αξιολόγηση πιθανών τύπων στόχων, η αξιολόγηση εργαλείων, τεχνικών και διαδικασιών, η μελέτη γνωστών αντιμέτρων και η κατανόηση των επιπτώσεων και πιθανών αποκρίσεων των στόχων. Ο υποτομέας της τεχνητής νοημοσύνης που χρησιμοποιείται για αυτή την διαδικασία ονομάζεται Natural Language Processing (NLP) και έχει ως σκοπό την επεξεργασία και ανάλυση μεγάλων ποσοτήτων δεδομένων φυσικής γλώσσας.

Η συλλογή και επεξεργασία δεδομένων μπορεί να πραγματοποιηθεί από διάφορες πηγές που βασίζονται στον Ιστό. Έτσι, η χρήση μεθόδων NLP σε κατάλληλους ιστοτόπους όπως μέσα κοινωνικής δικτύωσης, ειδήσεις, κ.α μπορούν να εντοπίσουν και να συλλέξουν πληροφορίες σχετικές με κοινές πρακτικές ασφάλειας και μέτρα ασφαλείας (π.χ. εντός βιομηχανιών), γνωστές ευπάθειες, συμβάντα και ύποπτα γεγονότα, ονόματα οργανώσεων, επιχειρηματικές δραστηριότητες, τοποθεσίες και ώρες λειτουργίας, ονόματα υπαλλήλων, διευθύνσεις ηλεκτρονικού ταχυδρομείου, ονόματα χρηστών σε κοινωνικά μέσα, κ.α

3.2.1.3 Δημιουργία προφίλ στόχου

Η τεχνητή νοημοσύνη έχει ήδη ξεκινήσει να χρησιμοποιείται από τα μεγάλα κοινωνικά μέσα δικτύωσης και τις πλατφόρμες ηλεκτρονικού εμπορίου για τη δημιουργία προφίλ χρηστών (profiling) και στοχευμένη διαφήμιση που έχει σκοπό να επιστήσει την προσοχή και να ασκεί επιρροή για αγορές υπηρεσιών και προϊόντων. Μέσω αυτής της εξατομικευμένης διαφήμισης, οι διαφημιζόμενοι μπορούν να προσεγγίζουν τους χρήστες βάσει των ενδιαφερόντων και των δημογραφικών στοιχείων τους καθώς και άλλων πληροφοριών με μεγάλη επιτυχία που έχει ως αποτέλεσμα την απόκτηση νέων πελατών και την αύξηση των εσόδων τους (Brynjolfsson, 2017).

Οι μέθοδοι που χρησιμοποιούνται για την δημιουργία προφίλ μπορούν να εστιάζουν είτε σε άτομα (ή ομάδες ατόμων) είτε σε δίκτυα. Στην περίπτωση όπου η δημιουργία προφίλ εστιάζει σε άτομα, η ανάλυση των σελίδων που ένα άτομο διαθέτει στα μέσα κοινωνικής δικτύωσης μπορεί να προσφέρει λεπτομέρειες της προσωπικότητας του όπου μπορούν στη συνέχεια να χρησιμοποιηθούν για να προβλευθούν ακόμα και ψυχολογικές καταστάσεις όπως η κατάθλιψη.

Στην περίπτωση των διαφημίσεων, τα εξελιγμένα συστήματα τεχνητής νοημοσύνης έχουν την ικανότητα να αναγνωρίζουν ποιό είναι το κατάλληλο μήνυμα και να το παρουσιάζουν ακριβώς την κατάλληλη στιγμή για να μεγιστοποιήσουν την αποδοτικότητα μιας διαφήμισης. Με αυτόν τον τρόπο, η δημιουργία προφίλ στόχων χρησιμοποιείται για την αύξηση της πιθανότητας επιτυχίας και μία από τις πιο διαδεδομένες τεχνικές που χρησιμοποιείται για αυτή την διαδικάσια είναι τα *convolutional neural networks* (CNN) όπου είναι μια κατηγορία νευρωνικών δικτύων που έχουν αποδειχθεί πολύ αποτελεσματικά σε τομείς όπως η αναγνώριση εικόνας και η ταξινόμηση.

Από την άλλη πλευρά, στον τομέα της κυβερνοασφάλειας, όταν οι επιτιθέμενοι επιδιώκουν να δημιουργήσουν προφίλ για τον στόχο τους έχοντας συλλέξει τα ανεπεξέργαστα δεδομένα από διαφορετικές πηγές, αφιερώνουν χρόνο για να συσχετίσουν τα ανεπεξέργαστα δεδομένα προτού τα επεξεργαστούν. Το κίνητρο είναι να ανακαλύψουν συσχετισμούς μεταξύ των στόχων όπου αποτελεί ένα κρίσιμο στάδιο, επειδή μια επιτυχημένη επίθεση εξαρτάται από τις πληροφορίες που προέρχονται από αυτήν τη φάση. Η συσχέτιση δεδομένων πριν από την τελική επεξεργασία παρέχει πληροφορίες για το περιβάλλον και τη συμπεριφορά του στόχου και αυτές οι πληροφορίες στη συνέχεια χρησιμοποιούνται για τον σχεδιασμό των στοχευμένων επιθέσεων. Οι πιο χρήσιμες πληροφορίες που χρειάζεται ο επιτιθέμενος είναι διευθύνσεις ηλεκτρονικού ταχυδρομείου, λογαριασμοί σε μέσα κοινωνικής δικτύωσης, ονόματα εργαζομένων και ονόματα χρηστών, αναλόγως τον στόχο.

3.2.1.4 Ανίχνευση ευπάθειας

Η ευπάθειες που υπάρχουν σε λογισμικά παραμένουν ένα σοβαρό πρόβλημα που αντιμετωπίζουν οι εταιρείες λογισμικού και οι τελικοί χρήστες που τα χρησιμοποιούν. Τα τελευταία χρόνια, υπήρξαν αυξημένενες αναφορές σε θέματα ευπάθειας ασφαλείας με σοβαρές επιπτώσεις και το γεγονός οτι η κοινωνία μας είναι πιο τεχνολογικά εξαρτημένη από κάθε άλλη εποχή έχει ως αποτέλεσμα την αύξηση της ανάγκης για εργαλεία και μεθόδους ανίχνευσης ευπάθειας λογισμικού. Συγκεκριμένα, στην σημερινή εποχή τα προσωπικά δεδομένα που θα μπορούσαν να οδηγήσουν σε κλοπή ταυτότητας δημοσιεύονται πλέον στα μέσα κοινωνικής δικτύωσης και ευαίσθητες πληροφορίες όπως αριθμοί κοινωνικής ασφάλισης, πληροφορίες πιστωτικών καρτών και στοιχεία τραπεζικού λογαριασμού αποθηκεύονται πλέον σε υπηρεσίες αποθήκευσης cloud όπως το Dropbox ή το Google Drive.

Υπάρχουν ορισμένες παραδοσιακές μέθοδοι και εργαλεία που χρησιμοποιούν διάφορες προσεγγίσεις για τον εντοπισμό και την αναφορά σημείων που αποτελούν απειλή για την ασφάλεια των συστημάτων και των χρηστών. Οι πιο διαδεομένοι μέθοδοι παρουσιάζονται στη συνέχεια:

3.2.1.4.1 Fuzzing

Η μέθοδος Fuzzing αφορά την εισαγωγή μη έγκυρων δεδομένων εισόδου ως input σε μια εφαρμογή με σκοπό να προκαλέσει ως output μια συμπεριφορά που δεν αναμένεται. Αυτό μπορεί να επιτευχθεί με την εισαγωγή μεγάλων ποσοτήτων τυχαίων δεδομένων που ονομάζονται fuzz στο υπο μελέτη πρόγραμμα το οποίο δοκιμάζεται σε μια προσπάθεια να το κάνει να «καταρρεύσει» (system crash). Εάν βρεθεί μια ευπάθεια, ένα εργαλείο λογισμικού που ονομάζεται fuzzer μπορεί να χρησιμοποιηθεί για τον εντοπισμό πιθανών αιτιών.

Η δημιουργία αυτών των δεδομένων μπορεί να πραγματοποιηθεί με δύο τρόπους (Bekrar, 2011). Ο ένας τρόπος είναι να δημιουργηθούν τυχαία, τροποποιώντας σωστά δεδομένα χωρίς να απαιτείται γνώση των λεπτομερειών της εφαρμογής. Αυτή η μέθοδος είναι γνωστή ως Black Box Fuzzing και ήταν η πρώτη ιδέα fuzzing που είχε επινοηθεί. Από την άλλη πλευρά, η μέθοδος White Box Fuzzing συνίσταται στη δημιουργία δοκιμών με την προϋπόθεση ότι υπάρχει πλήρης γνώση του κώδικα της εφαρμογής και της συμπεριφοράς του συστήματος. Μία τρίτη μέθοδος είναι το Grey Box Fuzzing που βρίσκεται ανάμεσα στις δύο μεθόδους που αναφέρθηκαν, με σκοπό να επωφεληθεί και από τις δύο. Σε αυτή την περίπτωση απαιτείται μόνο ελάχιστη γνώση της συμπεριφοράς του στόχου.

Η μέθοδος Fuzzing λειτουργεί καλύτερα για την ανακάλυψη ευπαθειών που μπορούν να αξιοποιηθούν μέσω buffer overflow, DOS (denial of service), cross-site scripting και SQL injection. Αυτές οι επιθέσεις χρησιμοποιούνται συχνά για κακόβουλο σκοπό ώστε να επιτύχουν τη μεγαλύτερη δυνατή καταστροφή στον λιγότερο δυνατό χρόνο. Από τη φύση της αυτή η μέθοδος είναι λιγότερο αποτελεσματική για την αντιμετώπιση απειλών ασφαλείας που δεν προκαλούν σφάλματα προγράμματος, όπως spyware, ορισμένους ιούς, worms, Trojans και keyloggers.

Αν και θεωρείται απλή ως μέθοδος, προσφέρει υψηλή αναλογία οφέλους προς κόστος και συχνά μπορεί να αποκαλύψει σοβαρά ελαττώματα που παραβλέπονται όταν γράφεται ο κώδικας

για το λογισμικό. Παρ' όλα αυτά δεν παρέχει πλήρη εικόνα της συνολικής ασφάλειας, της ποιότητας ή της αποτελεσματικότητας ενός προγράμματος και για αυτό τον λόγο είναι πιο αποτελεσματική όταν χρησιμοποιείται σε συνδυασμό με άλλες μεθόδους.

3.2.1.4.2 Web Application Scanners

Από διαδικτυακά καταστήματα έως μεγάλες εταιρείες, οι διαδικτυακές υπηρεσίες (Web services) γίνονται όλο και περισσότερο ένα στρατηγικό μέσο διανομής περιεχομένου. Επιτρέπουν την επικοινωνία μεταξύ ενός παρόχου και ενός καταναλωτή και υποστηρίζονται από μια σύνθετη υποδομή λογισμικού, η οποία συνήθως περιλαμβάνει έναν διακομιστή εφαρμογών, το λειτουργικό σύστημα και ένα σύνολο συστημάτων (π.χ. βάσεις δεδομένων, πύλες πληρωμής κ.λπ.)

Οι Web application scanners είναι μια αυτοματοποιημένη λύση που εξετάζει εφαρμογές στον Ιστό για ευπάθειες στην ασφάλεια τους (Vieira, 2009). Υπάρχουν δύο βασικές προσεγγίσεις για τον έλεγχο εφαρμογών ιστού για ευπάθειες: Η πρώτη ονομάζεται White box testing όπου αποτελείται από την ανάλυση του πηγαίου κώδικα της διαδικτυακής εφαρμογής. Αυτό μπορεί να γίνει χειροκίνητα ή χρησιμοποιώντας εργαλεία ανάλυσης κώδικα όπως το Fortify Application Security, SonarQube ή Pixy. Το πρόβλημα είναι ότι η εξαντλητική ανάλυση του πηγαίου κώδικα μπορεί να είναι δύσκολη και δεν μπορεί να βρει όλα τα ελαττώματα ασφαλείας λόγω της πολυπλοκότητας του κώδικα. Η δεύτερη προσέγγιση ονομάζεται Black box testing και είναι επίσης γνωστή ως δοκιμή διείσδυσης (penetration testing). Σε αυτή τη μέθοδο ο σαρωτής (scanner) δεν γνωρίζει τα εσωτερικά της εφαρμογής και χρησιμοποιεί τεχνικές fuzzing όπως αναφέρθηκε προηγουμένως μέσω αιτημάτων HTTP (HTTP requests).

Οι Web application scanners αποτελούν συχνά έναν εύκολο τρόπο δοκιμής εφαρμογών για εύρεση ευπαθειών. Στην πραγματικότητα, παρέχουν έναν αυτόματο τρόπο αναζήτησης ευπαθειών, αποφεύγοντας την επαναλαμβανόμενη και κουραστική εργασία που απαιτεί ο έλεγχος εκατοντάδων ή και χιλιάδων δοκιμών με το χέρι για κάθε τύπο ευπάθειας. Οι περισσότεροι από αυτούς τους σαρωτές είναι ιδιόκτητα εργαλεία (π.χ Acunetix Web Vulnerability Scanner, IBM Rational AppScan, κ.α), αλλά υπάρχουν επίσης μερικοί δωρεάν σαρωτές εφαρμογών οι οποίοι είναι ανοικτού κώδικα (π.χ. Foundstone WSDigger).

3.2.1.4.3 Static vs dynamic analysis

Η στατική ανάλυση που χρησιμοποιεί την white box προσέγγιση θεωρείται η πιο αποδοτική από πλευράς κόστους με την ικανότητα εντοπισμού σφαλμάτων σε μια πρώιμη φάση του κύκλου ζωής ανάπτυξης λογισμικού. Η στατική ανάλυση μπορεί επίσης να ανακαλύψει σφάλματα που δεν θα εμφανιστούν σε μια δυναμική δοκιμή. Η δυναμική ανάλυση, από την άλλη πλευρά, είναι ικανή να εντοπίσει ένα δυσεύρετο ελάττωμα ή μια ιδιαίτερα περίπλοκη ευπάθεια που συνήθως δεν μπορεί να εντοπίσει η στατική ανάλυση. Μια δυναμική δοκιμή, ωστόσο, θα εντοπίσει ελαττώματα μόνο στο τμήμα του κώδικα που εκτελείται κατά την διάρκεια του ελέγχου. Η απόφαση μιας επιχείρησης για την καταλληλότερη μέθοδο σχετίζονται με τον τύπο της υπό μελέτη εφαρμογής, ο χρόνος και οι πόροι της εταιρείας μεταξύ άλλων.

*3.2.2 **Μέθοδοι Πρόσβασης και διείσδυσης***

Αυτή η ενότητα συνοψίζει τους τρόπους με τους οποίους οι τεχνολογίες τεχνητής νοημοσύνης ενισχύουν τον σχεδιασμό και την εκτέλεση κυβερνοεπιθέσεων με σκοπό την μη εξουσιοδοτημένη πρόσβαση σε συστήματα και δίκτυα. Οι παρακάτω μέθοδοι επιβεβαιώνουν ότι η τεχνητή νοημοσύνη έχει σημαντικό ρόλο στο συνεχώς μεταβαλλόμενο τοπίο του κυβερνοχώρου και η ικανότητα μάθησης που προσφέρει με εξελιγμένες τεχνικές όπως deep learning και reinforcement learning έχει ως αποτέλεσμα την διευκόλυνση των εγκληματικών ενεργειών με πιο αποτελεσματικό τρόπο. Ως εκ τούτου, η ενημέρωση των νέων τάσεων σχετικά με επιθέσεις στον κυβερνοχώρο έχει γίνει πολύ πιο σημαντική για την προώθηση κατάλληλων αμυντικών ενεργειών.

Οι επιθέσεις που θα αναλυθούν στη συνέχεια έχουν ομαδοποιηθεί σε αυτές που χρησιμοποιούν κακόβουλα bots, αυτές που εστιάζουν στην παράνομη απόκτηση πρόσβασης και σε αυτές που χρησιμοποιούν “Adversarial training” με σκοπό να ξεγελάσουν το σύστημα μέσα από εκπαίδευση μοντέλων μηχανικής μάθησης ώστε να πετύχουν τον σκοπό τους.

3.2.2.1 Επιθέσεις χρησιμοποιώντας κακόβουλα bots

Στις δύο ακόλουθες επιθέσεις γίνεται χρήση προγραμμάτων (bots) που έχουν σχεδιαστεί για να λειτουργούν με έναν αυτοματοποιημένο τρόπο και χρησιμοποιούν τεχνητή νοημοσύνη για να πραγματοποιούν συγκεκριμένες διαδικασίες με μεγαλύτερη ακρίβεια.

SNAP_R

Παρουσίαση της επίθεσης

Τα κοινωνικά δίκτυα και ειδικότερα το Twitter με την πρόσβαση σε πληθόρα προσωπικών δεδομένων, την παροχή ενός API που εξυπηρετεί την χρήση Bots και την επικράτηση των συντομευμένων συνδέσμων (shortened links) μέσα σε δημοσιεύσεις, είναι ιδανικοί χώροι για τη διάδοση περιεχομένου που δημιουργείται από μηχανές (bots).

Με βάση τα παραπάνω, έχει δημιουργηθεί πρόσφατα μια επίθεση η οποία χρησιμοποιώντας μεθόδους τεχνητής νοημοσύνης μπορεί να μετατρέψει τις πλατφόρμες κοινωνικής δικτύωσης σε όπλο και φέρει την ονομασία «Social Network Automated Phishing with Reconnaissance» (SNAP_R). Οι κυβερνοεπιθέσεις με αυτό το εργαλείο έχουν την δυνατότητα να πραγματοποιούν αυτοματοποιημένο spear phishing στο Twitter, αλλά παρόμοια εργαλεία θα μπορούσαν να προγραμματιστούν και για άλλα κοινωνικά δίκτυα.

Σήμερα, οι επιθέσεις ηλεκτρονικού ψαρέματος συμβαίνουν σε διάφορες πλατφόρμες και ιδιαίτερα στα μέσα κοινωνικής δικτύωσης τα οποία εκθέτουν αυτήν την ευπάθεια μέσω του συνεχούς αυξανόμενου αριθμού χρηστών με τάσεις να αποκαλύπτουν προσωπικά δεδομένα**.** Με αυτόν τον τρόπο οι επιτιθέμενοι μπορούν σταδιακά να συλλέξουν πληροφορίες που θα τους δώσουν πρόσβαση σε τραπεζικούς λογαριασμούς ή ακόμη και να δημιουργήσουν μια νέα ταυτότητα χρησιμοποιώντας τα στοιχεία του θύματός τους.

Στην περίπτωση του Twitter, υπάρχουν αρκετά χαρακτηριστικά που ωφελούν την χρήση τεχνητής νοημοσύνης και συγκεκριμένα το υποπεδίο με την ονομασία NLP (*Natural language processing)*. Για παράδειγμα, επειδή σε κάθε δημοσίευση επιτρέπεται περιορισμένος αριθμός λέξεων, οι δημοσιεύσεις που παράγονται από το μοντέλο έχουν μειωμένη πιθανότητα να έχουν γραμματικά λάθη. Ένα άλλο χαρακτηριστικό είναι η συνήθεια να περιλαμβάνονται στα μηνύματα shortened links τα οποία μπορούν να κρύψουν την πραγματική διεύθυνση και να αυξήσουν το ποσοστό των κλικ από τους χρήστες (Patil, 2020).

Το NLP που είναι το κύριο μέσο διεκπεραίωσης των παραπάνω, έχει χρησιμοποιηθεί με επιτυχία για πολλές εφαρμογές συμπεριλαμβανομένης της ανίχνευσης ανεπιθύμητων μηνυμάτων (spam) (Sahami, 1998). Κατά την προετοιμασία της επίθεσης phishing, το NLP μπορεί να διακρίνει τα επαναλαμβανόμενα μοτίβα κειμένου για τον εντοπισμό θεμάτων που ενδιαφέρει τους χρήστες/στόχους και να προβεί στη δημιουργία μηνυμάτων στα οποία είναι πιο πιθανό να ανταποκριθεί ο στόχος.

Εστιάζοντας πιο πολύ στο SNAP_R τώρα, ο τρόπος που λειτουργεί είναι να καταγράφει τα ονόματα των χρηστών του Twitter σε μια λίστα και στη συνέχεια προσπαθεί να εντοπίσει τους καταλληλότερους στόχους για την επίθεση με βάση την πιθανότητα επιτυχίας, η οποία καθορίζεται κατά κύριο λόγο από τα θέματα που συνηθίζουν να δημοσιεύουν. Όταν ορισμένοι χρήστες κριθούν ότι είναι ευάλωτοι, το SNAP_R τους επιλέγει ως στόχο και στη συνέχεια, δημοσιεύει αυτόματα περιεχόμενο που περιέχει και έναν σύνδεσμο (link) προκειμένου να τους κατευθύνει σε συγκεκριμένους ιστοτόπους για την απόκτηση ευαίσθητων ιδιωτικών στοιχείων και κωδικών.

Πιο αναλυτικά, μετά την επιλογή ενός στόχου, το SNAP_R εξάγει ένα θέμα και το ιστορικό δημοσιεύσεων για τα tweets του συγκεκριμένου χρήστη και ύστερα δημιουργεί το κατάλληλο μήνυμα, ώστε να προσελκύσει τους στόχους με μεγαλύτερη επιτυχία. Έχει παρατηρηθεί ότι τα πιο αποτελεσματικά tweets για το ηλεκτρονικό ψάρεμα είναι αυτά που διαθέτουν συχνά εμφανιζόμενες λέξεις τα οποία δημοσιεύονται σε ώρες που ιστορικά έχει καταγραφεί ότι ο στόχος συνηθίζει και ο ίδιος να κάνει δημοσιεύσεις (Seymour, 2016).

Για μια επιτυχημένη επίθεση, το SNAP_R εκτός της ικανότητας να εκτελεί σύνθετες και εξατομικευμένες επιθέσεις είναι σημαντικό να λαμβάνει υπόψη και τους κανόνες του Twitter οι οποίοι υπάρχουν στους όρους χρήσης του ιστοτόπου. Έτσι, έαν στέλνονταν μαζικά σύνδεσμοι ηλεκτρονικού ψαρέματος τυχαία, το πιθανό θα ήταν να γίνει αντιληπτό και να κλείσει ο συγκεκριμένος λογαριασμός. Για να αποφευχθεί αυτό, η μέθοδος που χρησιμοποιεί είναι να συγκεντρώνει πρώτα τους χρήστες σε ομάδες με βάση τα προφίλ τους, χρησιμοποιώντας δεδομένα όπως η ποσότητα προσωπικών πληροφοριών που αποκαλύπτουν στο προφίλ τους, οι αλληλεπιδράσεις που έχουν με τους ακολούθους τους (followers) και πόσο συχνά κάνουν δημοσιεύσεις. Επιπλέον, εάν από τις πληροφορίες τους δείχνουν ότι είναι στόχος υψηλής αξίας (όπως τίτλοι θέσεων εργασίας ή δημοτικότητα), αυτά τα δεδομένα λαμβάνονται υπόψη επίσης.

Η διαδικασία ομαδοποίησης των χρηστών επηρεάζεται από τη συχνότητα των δημοσιεύσεων και το αντικείμενο των δημοσιεύσεων στα οποία οι χρήστες τείνουν να ανταποκρίνονται. Άλλες χρήσιμες πληροφορίες είναι η τοποθεσία των χρηστών που μπορεί γνωστοποιηθεί μέσα από «geotagging», οι εκδηλώσεις στις οποίες έχουν παρευρεθεί ή πρόκειται να παρευρεθούν καθώς επίσης και πληροφορίες που είναι δυνατόν να εξαχθούν μέσα από την

ανάλυση του συναισθήματος των μηνυμάτων (sentiment analysis) που στέλνουν (Perveen, 2016). Όλα αυτά τα χαρακτηριστικά χρησιμοποιούνται ως παράμετροι στο SNAP_R για να δημιουργηθεί το κείμενο της δημοσίευσης και να καθοριστεί η ώρα που θα αναρτηθεί.

Μετά από τον καθορισμό των στόχων, το SNAP_R τους στέλνει ένα tweet που περιέχει έναν σύνδεσμο. Για τη δημιουργία των tweets, χρησιμοποιούνται μοντέλα Markov σε συνδυασμό με Long Short-Term Memory (LSTM) recurrent neural networks (Seymour, 2016). Τα μοντέλα Markov δημιουργούν κείμενο με τη λογική ότι μια μελλοντική κατάσταση εξαρτάται μόνο από την τρέχουσα κατάσταση και όχι από τα γεγονότα που συνέβησαν πριν από αυτήν. Αυτό έχει ως αποτέλεσμα το μοντέλο να μην λαμβάνει υπόψη το νόημα των λέξεων πριν από την τρέχουσα λέξη και το κείμενο που παράγεται να μη φαίνεται φυσιολογικό. Από την άλλη, τα LSTM λύνουν αυτό το πρόβλημα επειδή μπορούν να θυμούνται το νόημα μιας μεγάλης πρότασης όταν προσπαθούν να προβλέψουν ποια θα είναι η κατάλληλη λέξη κατά τη δημιουργία ενός κειμένου. Ωστόσο, η αύξηση της ακρίβειας των LSTM έχει μεγάλο υπολογιστικό κόστος και απαιτεί περισσότερο χρόνο και δεδομένα για την εκπαίδευση.

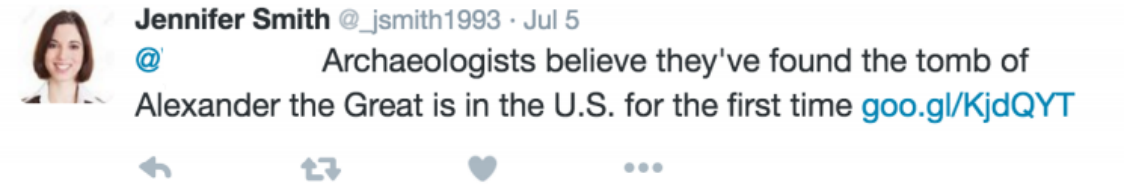
Για να αποφευχθεί το υπολογιστικό κόστος της επανεκπαίδευσης αυτών των μοντέλων για κάθε χρήστη, τα μοντέλα αυτά προ-εκπαιδεύονται σε δεδομένα από προηγούμενες επιθέσεις τύπου spear-phising και στη συνέχεια προστίθενται δεδομένα που αφορούν το θέμα συζήτησης που ενδιαφέρει τον κάθε χρήστη ξεχωριστά το οποίο εντοπίστηκε μέσα από την αυτοματοποιημένη δημιουργία προφίλ που είχε προηγηθεί.

Γιατί είναι σημαντική

Η τεχνική spear phising που χρησιμοποιεί το SNAP_R είναι εξέλιξη του phising (ηλεκτρονικό ψάρεμα) και στηρίζεται στην κοινωνική μηχανική (social engineering) όπου αποτελεί ένα από τα παλαιότερα αλλά και πιο αποτελεσματικά μέσα για εκμετάλλευση. Οι πρώτες προσπάθειες ηλεκτρονικού ψαρέματος ακολουθούσαν μια προσέγγιση όπου μόνο μερικοί στόχοι αν ανταποκρινόντουσαν ανάμεσα σε εκατομμύρια χρήστες ήταν αρκετό για να θεωρηθεί μια επίθεση επιτυχημένη. Αντιθέτως, οι Spear phishing επιθέσεις έχουν την διαφορά ότι είναι περισσότερο στοχευμένες διότι χρησιμοποιούν προσωπικές πληροφορίες σχετικά με τα προοριζόμενα θύματα στα μηνύματα/mail που στέλνουν, σε μια προσπάθεια να φαίνονται αυθεντικά. Η χρήση τεχνητής νοημοσύνης ευθύνεται σε μεγάλο βαθμό για την επιτυχή πραγματοποίηση όλων αυτών των σταδίων και έχει ως αποτέλεσμα να αυξάνεται η πιθανότητα ανταπόκρισης των στόχων στις επιθέσεις, γεγονός που δημιουργεί ανησυχίες στους διαδικτυακούς χρήστες όσον αφορά την ασφάλεια (Halevi, 2015).

Παράδειγμα

Κατά τη διάρκεια της συλλογής πληροφοριών εάν ένας χρήστης αναφέρει στις δημοσιεύσεις του θέματα που έχουν να κάνουν για παράδειγμα με αρχαιολογία, το SNAP_R μπορεί να κάνει μια δημοσίευση παρόμοια με της παρακάτω εικόνας που να προορίζεται για αυτόν τον χρήστη. Επειδή το κείμενο θα ενδιαφέρει τον υποψήφιο στόχο που το λαμβάνει, θα είναι πιο πιθανό να επισκεφτεί τον σύνδεσμο.



Εικόνα 1: Παράδειγμα αυτοματοποιημένης δημοσίευσης.
Πηγή: Blackhat[1]

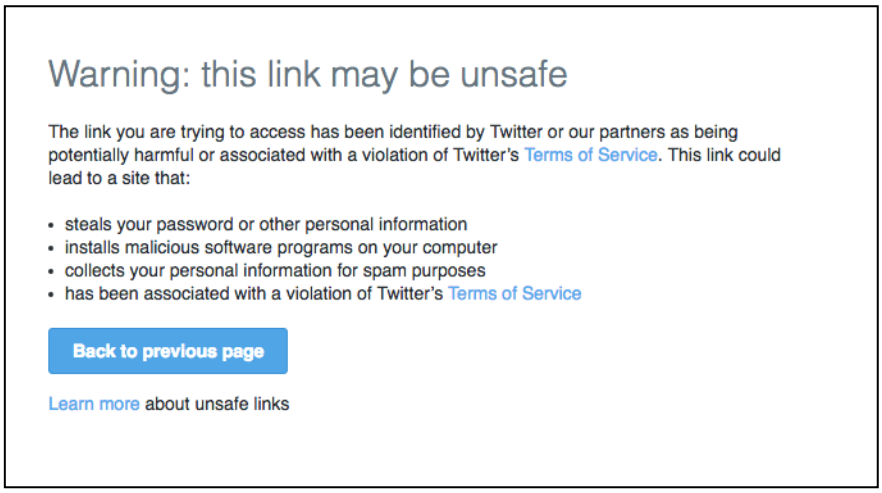
Εκτίμηση επικινδυνότητας

Αυτή η μέθοδος μπορεί να στοχεύσει μεγάλο αριθμό χρηστών και σε σύγκριση με παρόμοιες προσπάθειες μαζικών επιθέσεων ηλεκτρονικού ψαρέματος που είχαν περίπου 10% επιτυχία (Jakobsson, 2006), το SNAP_R πετυχαίνει περισσότερο από 45% (Seymour, 2016). Αυτό οφείλεται σε σπουδαίο βαθμό στην αξιοποίηση της επιστήμης των δεδομένων που ενισχύει την ικανότητα σχεδιασμού εξατομικευμένων μηνυμάτων και τον εντοπισμό ευάλωτων χρηστών με προσεγγίσεις που στηρίζονται στα ειδικά χαρακτηριστικά που έχουν τα μέσα κοινωνικής δικτύωσης και συγκεκριμένα το Twitter.

Μεθοδοι αντιμετώπισης

Σχετικά με τις μεθόδους άμυνας, το Twitter έχει καταφέρει σε μεγάλο βαθμό να μπορεί να ανακαλύψει το αυτοματοποιημένο ανεπιθύμητο περιεχόμενο με συνδέσμους ηλεκτρονικού ψαρέματος. Αυτό επιτυγχάνεται εμφανίζοντας ένα μήνυμα προειδοποίησης όταν οι χρήστες κάνουν κλικ σε έναν σύνδεσμο που οδηγεί σε ιστότοπο εκτός του Twitter (όπως παρουσιάζεται στην παρακάτω εικόνα) ή αποκλείοντας τον σύνδεσμο πριν την δημοσίευση από τον επιτιθέμενο (Twitter, 2020).

[1] https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf



Εικόνα 2: Προειδοποιητικό μήνυμα από το Twitter.
Πηγή: Twitter[2]

Ωστόσο, οι εξατομικευμένες spear phising επιθέσεις που πραγματοιούνται από το SNAP_R μπορούν να είναι επιτυχείς προτού γίνουν αντιληπτές από τα συστήματα ανίχνευσης του Twitter. Επομένως, αυτό χρειάζεται μεγαλύτερη ευαισθητοποίηση για την ανάπτυξη κατάλληλων αμυντικών μηχανισμών.

Fake reviews attack

Παρουσίαση της επίθεσης

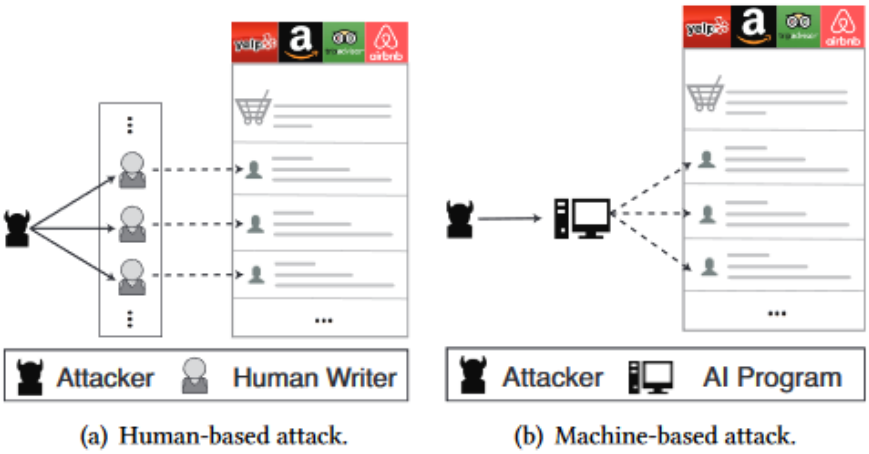
Τον τελευταίο καιρό υπάρχουν πολλές ανησυχίες σχετικά με την αξιοπιστία των διαδικτυακών πηγών πληροφοριών λόγω ψευδών κριτικών και παραπληροφόρησης. Σχετικά παραδείγματα είναι η διάδοση ψευδών κριτικών με σκοπό να βλάψουν ανταγωνιστές, να κερδίσουν πολιτικές εκστρατείες και να επηρεάσουν την κοινή γνώμη. Μέχρι τώρα συνηθιζόταν να γίνεται από ανθρώπους αλλά τελευταία έχει ξεκινήσει η αντικατάστασή τους χρησιμοποιώντας προγράμματα που στηρίζονται σε τεχνητή νοημοσύνη (Yao, 2017).

Ο στόχος αυτών των επιθέσεων είναι να δημιουργηθούν ψεύτικες κριτικές με τέτοιο τρόπο που να μην διαφέρουν από αυτές που γράφονται από ανθρώπους. Στην παραδοσιακή προσέγγιση σε αυτής της μορφής την επίθεση, οι επιτιθέμενοι πληρώνουν ανθρώπινους συγγραφείς για να κάνουν αυτές τις παράνομες ενέργειες στο διαδίκτυο. Δεδομένου ότι οι συγγραφείς είναι πραγματικοί άνθρωποι, μπορούν να κάνουν τις κριτικές να φαίνονται αληθινές και ως εκ τούτου να μην ανιχνεύονται από αυτοματοποιημένα εργαλεία ανίχνευσης. Ωστόσο, τέτοιες επιθέσεις απαιτούν σημαντικό κόστος, χρόνο και επίσης εάν κάποιος παράγει μαζικές κριτικές, θα ανιχνευθεί σύντομα με αποτέλεσμα να κλειστεί ο λογαριασμός του.

Ως εξέλιξη της παραδοσιακής μεθόδου που αναφέρθηκε παραπάνω, έχει δημιουργηθεί μια αυτοματοποιημένη επίθεση ψευδών κριτικών που χρησιμοποιεί ένα γλωσσικό μοντέλο που

[2] https://help.twitter.com/en/using-twitter/url-shortener

βασίζεται σε deep neural networks για την παραγωγή ρεαλιστικού περιεχομένου που προορίζεται για διαδικτυακές κριτικές και στοχεύουν να μην διακρίνονται από αυτές που δημιουργήθηκαν από πραγματικούς ανθρώπους.



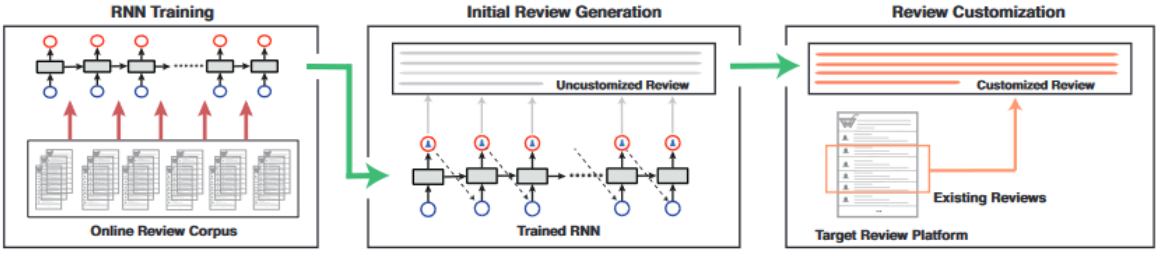
(a) Human-based attack. (b) Machine-based attack.

Εικόνα 3: Η αυτοματοποίηση με μεθόδους τεχνητής νοημοσύνης προσφέρει ταχύτερα αποτελέσματα.
Πηγή: Arxiv[3]

Σχετικά με τον τρόπο λειτουργίας αυτής της επίθεσης, σε πρώτη φάση γίνεται εκπαίδευση ενός Long short-term memory (LSTM) μοντέλου με τα δεδομένα εκπαίδευσης να αποτελούνται από ένα πραγματικό σύνολο με κριτικές (π.χ για εστιατόρια). Τα LSTM είναι ένας τύπος recurrent neural network και χρησιμοποιούνται σε σύνθετα προβλήματα που σχετίζονται με αυτόματη μετάφραση, αναγνώριση ομιλίας και άλλα. Στη συνέχεια, μετά τη διαδικασία εκπαίδευσης, δημιουργείται ένα σύνολο αρχικών κριτικών υπολογίζοντας πιθανότητες για να προβλέψει ποιοι χαρακτήρες είναι πιθανό να έρθουν στη συνέχεια με βάση τους προηγούμενους χαρακτήρες.

Στην δεύτερη και τελευταία φάση, πραγματοποιείται μια σειρά από διορθώσεις στις κριτικές που δημιούργησε ο αλγόριθμος στο προηγούμενο στάδιο με σκοπό να τις κάνει πιο συγκεκριμένες στο αντικείμενο που χρειάζεται ώστε να είναι και πιο αληθοφανείς. Αυτό επιτυγχάνεται με την αυτοματοποιημένη αντικατάσταση λέξεων με αντίστοιχες λέξεις κλειδιά που έχουν αναφερθεί ήδη σε παλαιότερες κριτικές με αντικείμενο που είναι παρόμοιο με αυτό που προορίζονται οι νέες ψευδείς κριτικές.

[3] https://arxiv.org/pdf/1708.08151.pdf



Εικόνα 4: Τα στάδια της επίθεσης.
Πηγή: Arxiv[4]

Γιατί είναι σημαντική

Αυτή η επίθεση θεωρείται σημαντική επειδή ένα μεγάλο τμήμα του πληθυσμού επιλέγει να συμβουλεύεται τις διαδικτυακές κριτικές προϊόντων και υπηρεσιών οι οποίες τελευταία ασκούν μεγάλη επιρροή στην λήψη αποφάσεων των πελατών. Έρευνες δείχνουν ότι το 91% του πληθυσμού ελέγχει τακτικά κριτικές και το 84% στηρίζεται σε αυτές πριν προχωρήσει σε αγορά (Inc.com, 2017). Καθώς οι καταναλωτές αλληλεπιδρούν όλο και περισσότερο μεταξύ τους μέσω του διαδικτύου, ανταλλάσσουν απόψεις και εμπειρίες με αποτέλεσμα οι επιχειρήσεις να αφιερώνουν όλο και περισσότερους πόρους στην παρακολούθηση, την ανάλυση και επιδιόρθωση της φήμης τους στο διαδίκτυο. Ως εκ τούτου, η επίδραση της συγκεκριμένης επίθεσης που επιδιώκει την μαζική παραποίηση των αξιολογίσεων έχει μεγάλο αντίκτυπο σε εταιρίες και καταναλωτές.

Παράδειγμα

Έχοντας ως στόχο την ιστοσελίδα Yelp που διαθέτει κριτικές εστιατορίων, η χρήση των εργαλείων που περιγράφονται σε αυτή την επίθεση με τεχνητή νοημοσύνη, κατάφερε να συνθέσει κριτικές που ήταν τόσο πειστικές που όχι μόνο δεν έγιναν αντιληπτές από τους πραγματικούς χρήστες και τα συστήματα ανίχνευσης αλλά δέχτηκαν και θετικές αντιδράσεις από τους αναγώστες της σελίδας που τις κατέστησαν ως χρήσιμες. (Yao, 2017)

Εκτίμηση επικινδυνότητας

Ο κίνδυνος που μπορεί να επιφέρει αυτή η επίθεση προέρχεται από την δυσκολία να γίνουν αντιληπτές οι κριτικές που δημιουργεί και να καταλήγει να παραπλανεί τους καταναλωτές σε αγορές που δεν θα είναι ικανοποιητικές. Αυτό μπορεί να επηρεάσει αρνητικά και τις επιχειρήσεις δεδομένου ότι μετά από μια αποτυχημένη αγορά οι πελάτες ίσως να μην ξαναπροτιμήσουν την ίδια εταιρία εξαιτίας έλλειψης εμπιστοσύνης.

[4] https://arxiv.org/pdf/1708.08151.pdf

Μέθοδοι αντιμετώπισης

Από άποψη άμυνας, αυτή η επίθεση είναι δύσκολο να αντιμετωπιστεί επειδή έχει την δυνατότητα να φτιάχνει τις κριτικές σε σύντομο χρονικό διάστημα και λόγω της μεθόδου που χρησιμοποιεί καταφέρνει να τις καθιστά πειστικές και να μην εντοπίζονται από λογισμικό ανίχνευσης. Παρ' όλα αυτά μία προσέγγιση που πολλές φορές καταφέρνει να διακρίνει τις ψευδείς κριτικές είναι ο ανιχνευτής λογοκλοπής, ο οποίος χρησιμοποιεί επίσης τεχνητή νοημοσύνη. Για παραδειγμα, μια απλή εκδοχή αυτής της επίθεσης στηρίζεται στην αντιγραφή ή μερική αντιγραφή πραγματικών κριτικών στις οποίες το μοντέλο είχε εκπαιδευτεί. Έτσι το πρόγραμμα ελέγχου λογοκλοπής σε περίπτωση που έχει πρόσβαση στη βάση δεδομένων με τις κριτικές που χρησιοποιήθηκαν για την εκπαίδευση του μοντέλου του επιτιθέμενου μπορεί κάνοντας συγκρίσεις να εντοπίσει ότι οι κριτικές δεν είναι πραγματικές.

3.2.2.2 Επιθέσεις παράνομης πρόσβασης

Οι δύο ακόλουθες επιθέσεις παρουσιάζουν τις δυνατότητες που προσφέρει η τεχνητή νοημοσύνη για την απόκτηση μη εξουσιοδοτημένης πρόσβασης σε συστήματα με πιο αποτελεσματικούς και ευφυείς τρόπους.

Intelligent password brute-force attack

Παρουσίαση της επίθεσης

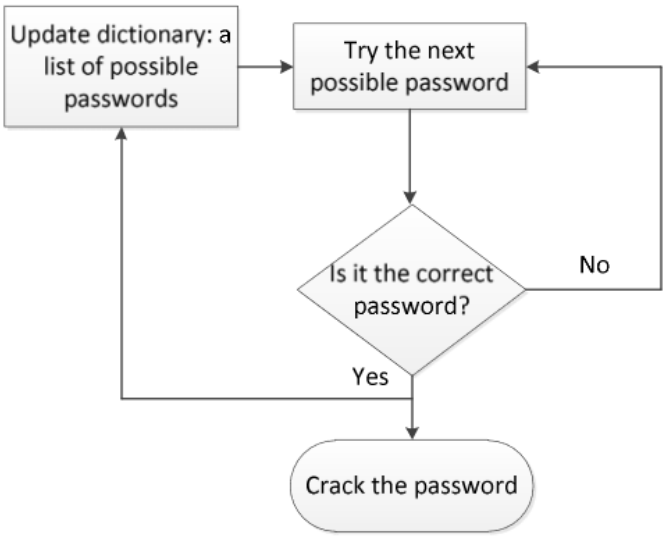
Οι κωδικοί πρόσβασης (Passwords), όταν χρησιμοποιούνται σωστά, είναι ένας εξαιρετικά απλός και αποτελεσματικός τρόπος για την προστασία δεδομένων και συστημάτων πληροφορικής από μη εξουσιοδοτημένη πρόσβαση. Ωστόσο, πολλά άτομα συνεχίζουν να χρησιμοποιούν κωδικούς πρόσβασης με τρόπο που τους εκθέτει σε κίνδυνο (Yildirim, 2019).

Από τις πιο διαδεδομένες επιθέσεις σε δίκτυα υπολογιστών που εστιάζουν στην παράνομη εύρεση του σωστού συνδιασμού κωδικών προσβασης είναι η επίθεση brute force. Σε μια τέτοια επίθεση στο πρωτόκολλο SSH (Secure shell), ο εισβολέας προσπαθεί να συνδεθεί στο λογαριασμό ενός χρήστη δοκιμάζοντας διαφορετικούς κωδικούς μέχρι να εντοπίσει τον σωστό κωδικό πρόσβασης. Λόγω του μεγάλου αριθμού πιθανών συνδυασμών οι εισβολείς χρησιμοποιούν αυτοματοποιημένο λογισμικό που ελέγχει δισεκατομμύρια συνδυασμούς γραμμάτων, αριθμών και συμβόλων επανειλημμένα.

Μία νέα παραλλαγή της brute force επίθεσης που χρησιμοποιεί τεχνητή νοημοσύνη, αποτελεί την νέα γενιά αυτής της κλασικής προσέγγισης στην απόκτηση κωδικών πρόσβασης. Από τις πιο σημαντικές διαφορές που διαθέτει σε σύγκριση με την απλή μέθοδο είναι η ικανότητα να μαθαίνει από παλαιότερα δεδομένα και να βελτιώνει αυτόματα την απόδοσή της. Συγκεκριμένα, μέσα από εκπαίδευση ενός recurrent neural network (RNN) μοντέλου αποκτά την ικανότητα αναγνώρισης μοτίβων σε παλιούς κωδικούς πρόσβασης και καταφέρνει να δημιουργεί αυτόματα νέους υποψήφιους κωδικούς πρόσβασης με μεγαλύτερη πιθανότητα επιτυχίας.

Γιατί είναι σημαντική

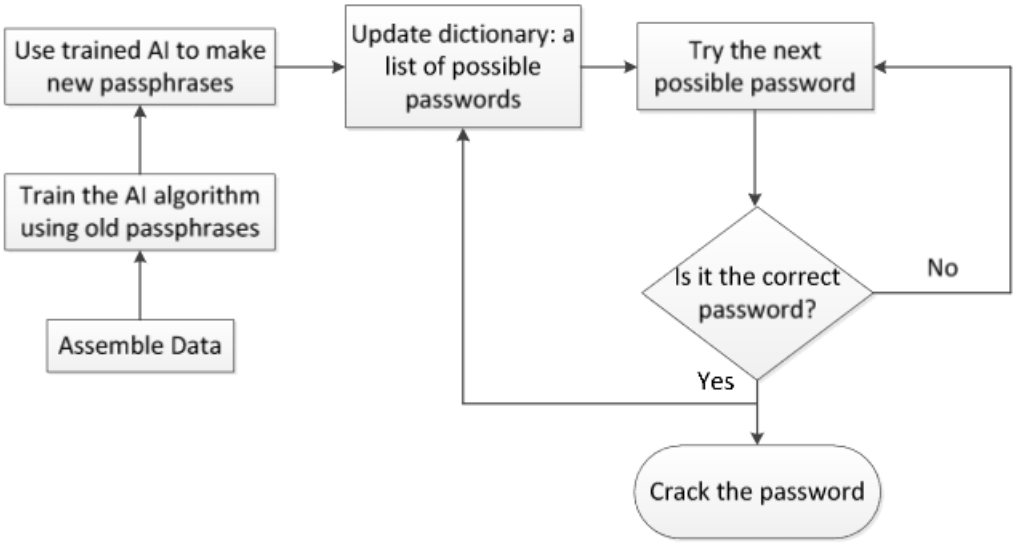
Αυτή η επίθεση είναι σημαντική διότι αποτελεί μια ενισχυμένη έκδοση μιας απειλής που είναι ήδη σε μεγάλο βαθμό ζημιογόνα. Αναλυτικότερα, ο παραδοσιακός τρόπος για την brute-force επιθέση βασίζεται σε ένα επεξεργασμένο λεξικό που περιέχει ένα σύνολο πιθανών κωδικών πρόσβασης το οποίο διαθέτει συνδυασμούς συνηθισμένων λέξεων και συμβόλων αλλά και κωδικούς που προέκυψαν από δημιοσιεύσεις κωδικών προηγούμενων παραβιάσεων. Πολλές φορές ακόμα και αυτή η μορφή της επίθεσης που δεν χρησιμοποιεί τεχνητή νοημοσύνη καταφέρνει να είναι αποτελεσματική, παρ' όλα αυτά οι μεγαλύτεροι κωδικοί πρόσβασης επειδή έχουν περισσότερες πιθανές τιμές απαιτούν ακόμη περισσότερους συνδυασμούς, καθιστώντας τους ιδιαίτερα δύσκολο να εντοπιστούν σε σύγκριση με τους μικρότερους. Αυτό έχει ως αποτέλεσμα, οι πόροι που απαιτούνται για αυτή την επίθεση να αυξάνονται εκθετικά με το αυξανόμενο μέγεθος του κωδικού πρόσβασης.



Εικόνα 4: Μέθοδος παραδοσικής brute-force επίθεσης που βασίζεται σε χρήση λεξικού (dictionary based). Πηγή: Aisel[5]

Το στοιχείο που καθιστά σημαντική την νέα αυτή παραλλαγή της επίθεσης brute force είναι το γεγονός ότι υποστηρίζεται από τεχνολογίες τεχνητής νοημοσύνης. Συγκεκριμένα, χρησιμοποιεί έναν αλγόριθμο μηχανικής μάθηση ανοιχτού κώδικα, που ονομάζεται Torch-rnn για την δημιουργία ενός γλωσσικού μοντέλου RNN με σκοπό την δημιουργία νέων υποψηφίων κωδικών ακολουθώντας ένα παρόμοιο μοτίβο που υπήρχε σε προηγούμενους κωδικούς πρόσβασης.

[5] https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1038&context=mwais2018



Εικόνα 5: Μέθοδος επίθεσης brute force με χρήση τεχνητής νοηοσύνης.
Πηγή: Aisel[6]

Παράδειγμα

Χαρακτηριστικό παράδειγμα αποτελεί ένας κωδικός σε ένα σύστημα που βασίζεται σε λέξεις που συνηθίζονται να χρησιμοποιούνται για αυτόν τον σκοπό. Η εφαρμογή της επίθεσης σε αδύναμους κωδικούς πρόσβασης έχουν υψηλότερο ποσοστό επιτυχίας και έτσι οι σωστοί συνδυασμοί χαρακτήρων εντοπίζονται σε σύντομο χροσνικό διάστημα.

Εκτίμηση επικινδυνότητας

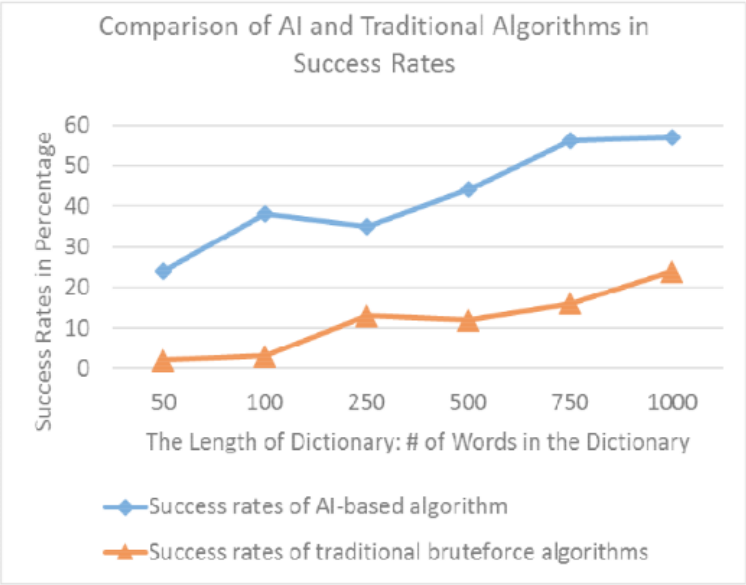
Η αποτελεσματικότητα αυτής της ανανεωμένης μεθόδου brute force την καθιστά ιδιαίτερα επικίνδυνη ειδικότερα σε περιπτώσεις όπου δεν δίνεται αρκετή σημασία στην επιλογή σύνθετου κωδικού πρόσβασης. Επίσης, η «ευφυία» που αποκτά μέσα από την εκπαίδευση του μοντέλου την κάνει σαφώς πιο αποδοτική από την απλή brute force προσέγγιση. Σύμφωνα με συγκρίσεις που έχουν γίνει μεταξύ εκπαιδευμένου μοντέλου με χρήση RNN και της παραδοσιακής brute force επίθεσης τα αποτελέσματα παρουσιάζουν πώς η χρήση τεχνητής νοημοσύνης βοηθά εντυπωσιακά στην εύρεση του σωστού κωδικού πρόσβασης. Ο παρακάτω πίνακας δείχνει την απόδοση των δύο μεθόδων για διαφορετικά μήκη του λεξικού (δηλαδή τον αριθμό των λέξεων στο λεξικό): 50, 100, 250, 500, 750 και 1000. Για κάθε δοκιμή, υπολογίζεται ο μέσος όρος 100 δοκιμών (Trieu, 2018).

[6] https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1038&context=mwais2018

# of Words in the Dictionary	Success Rates of AI-Based Algorithm	Success Rates of Traditional Algorithm
50	24%	2%
100	38%	3%
250	35%	13%
500	44%	12%
750	56%	16%
1000	57%	24%

Εικόνα 6: Σύγκριση των δύο μεθόδων με βάση την επιτυχή εύρεση του κωδικού πρόσβασης.

Πηγή: Aisel[7]



Εικόνα 7: Σύγκριση των δύο μεθόδων με βάση τα ποσοστά επιτυχίας εύρεσης κωδικού πρόσβασης.

Πηγή: Aisel[8]

[7] https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1038&context=mwais2018

[8] https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1038&context=mwais2018

Μέθοδοι αντιμετώπισης

Ορισμένες αμυντικές στρατηγικές για την αποφυγή εύρεσης του κωδικού πρόσβασης από αυτή την επίθεση μπορούν να περιλαμβάνουν την εφαρμογή ελέγχου ταυτότητας πολλών παραγόντων (Multi-factor authentication) και την επιλογή σύνθετων κωδικών πρόσβασης χρησιμοποιώντας τυχαίους συνδυασμούς που αποτελούνται από πολλούς χαρακτήρες. Αξίζει να αναφερθεί ότι ο ελέγχος ταυτότητας πολλών παραγόντων χρησιμοποιείται συχνά τον τελευταίο καιρό και παρόλο που δεν παρέχει πλήρη ασφάλεια, καθιστά αρκετά πιο δύσκολο να αποκτηθεί μη εξουσιοδοτημένη πρόσβαση από επιτιθέμενους επειδή απαιτεί μια δευτερεύουσα μορφή επαλήθευσης πριν από τη σύνδεση.

Επίθεση που στηρίζεται σε φωντητική σύνθεση

Παρουσίαση της επίθεσης

Μία από τις τεχνολογίες που έχει εξελιχθεί σε μεγάλο βαθμό με την υποστήριξη της τεχνητής νοημοσύνης είναι οι φωνητικοί βοηθοί που είναι γνωστοί και ως ευφυείς προσωπικοί βοηθοί. Οι φωνητικοί βοηθοί βρίσκονται πλέον ενσωματωμένοι σε διάφορα προϊόντα, από smartphone και συστήματα πλοήγησης αυτοκινήτου έως διάφορους άλλους τύπους οικιακών συσκευών. Τα τελευταία χρόνια ο αριθμός των συμβατών συσκευών έχει αυξηθεί εντυπωσιακά (Han, 2018) και οι μεγαλύτερες εταιρίες τεχνολογίας έχουν μια δική τους εκδοχή αυτής της τεχνολογίας με το "Google assistant" από την Google, το "Siri" από την Apple, την "Cortana" από τη Microsoft, την "Alexa" από την Amazon και το "Bixby" από τη Samsung να είναι οι πέντε δημοφιλέστερες που διατίθενται στην αγορά αυτή τη στιγμή.

Οι υπηρεσίες που παρέχονται από αυτούς τους φωνητικούς βοηθούς εξαρτώνται από το εύρος των εφαρμογών που έχουν αναπτυχθεί με σκοπό να τους χρησιμοποιούν και συνήθως καλύπτουν διάφορους τομείς όπως προσωπική οργάνωση, παροχή πληροφοριών, εκτέλεση παραγγελιών, ψυχαγωγία, οικιακή αυτοματοποίηση κ.α. Γενικότερα, η πρόσβαση σε αυτές τις υπηρεσίες αποκτάται μέσω των συμβατών συσκευών χρησιμοποιώντας φυσική γλώσσα, το οποίο απαιτεί την ενεργοποίηση του μικροφώνου και την συνεχή καταγραφή ήχων του περιβάλλοντος γύρω από τη συσκευή έτσι ώστε να διακρίνει τη λέξη αφύπνισης (π.χ. «Alexa» για το Amazon Alexa, «Ok Google» και «Hey Google» για τον Google Assistant κ.α. Όταν γίνεται αναφορά στην λέξη αφύπνισης, ενεργοποιείται στη συνέχεια η ηχογράφηση (Cheung, 2019) .

Τα παραπάνω χαρακτηριστικά στα οποία βασίζεται η λειτουργία των φωνητικών βοηθών μπορούν να χρησιμοποιηθούν κακόβουλα από επιτιθέμενους για να εισβάλλουν σε συμβατές συσκευές και να αποκτήσουν πρόσβαση σε προσωπικές πληροφορίες. Μία σχετική προσέγγιση που χρησιμοποιεί τεχνητή νοημοσύνη για αυτό το σκοπό, ξεκινάει με την κρυφή ηχογράφηση των λέξεων αφύπνησης και στην συνέχεια αποφασίζει ποια είναι η ιδανικότερη στιγμή που θα πραγματοποιηθεί η επίθεση.

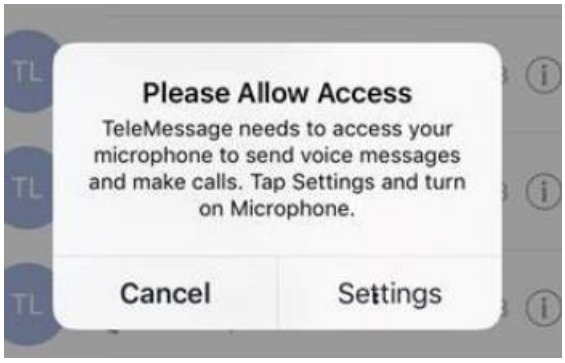
Γιατί είναι σημαντική

Αυτή η επίθεση είναι σημαντική επειδή ο επιτιθέμενος αποκτά πρόσβαση σε πολλές λειτουργίες της συσκευής στην οποία έχει εγκατασταθεί η εφαρμογή. Επιπροσθέτως, λόγω της συνήθειας να

παραχωρούνται διάφορα δικαιώματα σε εφαρμογές (apps) κατά την εγκατάστασή τους σε έξυπνα τηλέφωνα, είναι εύκολο να δωθούν σε κακόβουλη εφαρμογή χωρίς να είναι εμφανές ο σκοπός στον κάτοχο της συκευής.

Παράδειγμα

Ένα παράδειγμα που χρησιμοποιεί αυτή την προσέγγιση είναι η δημιουργία ενός παιχνιδιού που ελέγχεται από το μικρόφωνο της συσκευής ενός smartphone. Αρχικά, μέσω της εφαρμογής/παιχνιδιού ο επιτιθέμενος επιδιώκει να ξεγελάσει τον χρήστη ώστε να παραχωρήσει τα δικαιώματα (όπως πρόσβαση στο μικρόφωνο της συσκευής) που απαιτούνται για την εκτέλεση των κακόβουλων ενεργειών. Με αυτό τον τρόπο καταφέρνει να αποκτήσει πρόσβαση χωρίς να γίνεται αντιληπτός από τον χρήστη και τα προγράμματα προστασίας από ιούς.



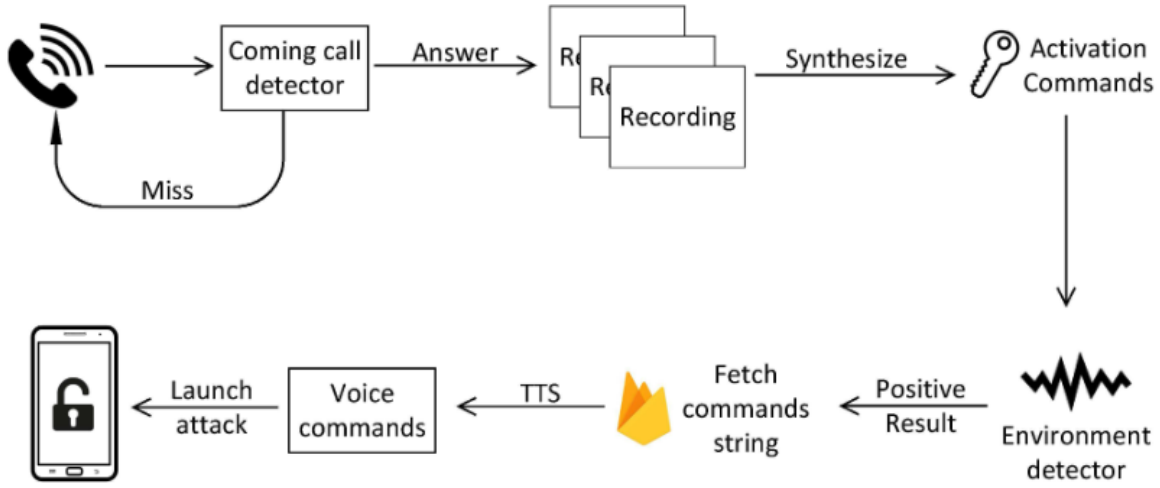
Εικόνα 8: Παράδειγμα αιτήματος αποδοχής για χρήση μικροφώνου από εφαρμογή.
Πηγή: Telemessage[9]

Μετά από την παραχώρηση των απαιτούμενων δικαιώματων, ο επιτιθέμενος έχει την δυνατότητα να εγγράφει κρυφά τη φωνή του χρήστη από εισερχόμενες και εξερχόμενες κλήσεις μέσω του μικροφώνου. Έχοντας συλλέξει αρκετά δεδομένα από τις ηχογραφήσες, ακολουθεί επεξεργασία στο φωνητικό υλικό εφαρμόζοντας Natural Language Processing (NLP) με σκοπό να συνθέσει τις λέξεις-κλειδιά που θα χρησιμοποιηθούν στην επίθεση σε επόμενο στάδιο.

Για τον καθορισμό της κάταλληλης ώρας που θα χρησιμοποιηθούν οι λέξεις κλειδιά κατά την επίθεση, λαμβάνονται υπόψη κάποιες πληροφορίες που υποδηλώνουν ότι ο χρήστης είναι λιγότερο πιθανό να αντιληφθεί την επίθεση. Αυτές οι πληροφορίες παρέχονται από την ίδια την συσκευή, η οποία στέλνει δεδομένα σχετικά με το περιβάλλον που συλλέγονται μέσω των ενσωματωμένων αισθητήρων (π.χ. μικρόφωνο για εγγραφή ήχου περιβάλλοντος, αισθητήρας φωτός περιβάλλοντος, μοτίβα κίνησης μέσω του επιταχυνσιόμετρου). Ύστερα τροφοδοτεί τα παραπάνω δεδομένα σε ένα πρόγραμμα που βασίζεται στον αλγόριθμο μηχανικής μάθησης με όνομα random forest για να αποφασίσει την κατάλληλη στιγμή που θα ξεκινήσει η επίθεση.

[9] https://www.telemessage.com/granting-enterprise-number-archiver-access-to-your-microphone

Όταν αποφασιστεί η κατάλληλη ώρα για την επίθεση, πραγματοποιείται ενεργοποίηση του φωνητικού βοηθού κάνοντας αναπαραγωγή των λέξεων αφύπνησης από το ηχείο της συσκευής με την φωνή του ίδιου του χρήστη που είχε ηχογραφηθεί σε προηγούμενο στάδιο. Μετά την ενεργοποίηση του φωνητικού βοηθού ένας μεγάλος αριθμός λειτουργιών μπορεί να πραγματοποιηθεί, όπως να ξεκλειδώσει η οθόνη, να αλλάξουν οι ρυθμίσεις του smartphone καθώς και πολλές άλλες που μπορεί να κάνει ένας άνθρωπος, όπως η αποστολή SMS / email και πραγματοποίηση τηλεφωνικών κλήσεων (Zhang, 2018).



Εικόνα 9: Τα στάδια της επίθεσης.
Πηγή: Arxiv[10]

Εκτίμηση επικινδυνότητας

Η διακριτικότητα που χαρακτηρίζει τον τρόπο που διεξάγεται αυτή η επίθεση, την καθιστά μεγάλη απειλή στους χρήστες έξυπνων τηλεφώνων. Αυτό σημαίνει ότι μπορεί να δρα για μεγάλο χρονικό διάστημα χωρίς να γίνεται αντιληπτό, έχοντας πρόσβαση σε ευαίσθητες πληροφορίες αλλά και την δυνατότητα να αλληλεπιδρά με την ίδια την συσκευή από απόσταση. Ένα άλλο στοιχείο που αυξάνει την επικινδυνότητα είναι το γεγονός ότι τα δικαιώματα για χρήση του μικροφώνου παραχωρούνται από τον ίδιο τον κάτοχο της συσκευής. Ο χρήστης βέβαια δεν γνωρίζει ότι η ηχογράφηση γίνεται και σε ώρες που δεν χρησιμοποιείται η εφαρμογή ούτε τι μπορεί να επιτευχθεί με αυτήν στην συνέχεια.

[10] https://arxiv.org/pdf/1805.06187.pdf

Μέθοδοι αντιμετώπισης

Σχετικά με μεθόδους αντιμετώπισης αυτής της επίθεσης, η εφαρμογή τεχνητής νοημοσύνης για προστασία θα μπορούσε να δώσει υποσχόμενα αποτελέσματα. Συγκεκριμένα, η εκπαίδευση ενός αλγορίθμου με ηχητικά δεδομένα ώστε να μάθει να αναγνωρίζει αν η πηγή των φωνητικών εντολών είναι μέσω του ενσωματωμένου ηχείου του smartphone (όπως συνέβη στην επίθεση που αναφέρθηκε παραπάνω) και να είναι σε θέση να μην ανταποκρίνεται σε αυτές τις εντολές, θα μπορούσε να παρέχει προστασία σε αυτού του είδους τις επιθέσεις.

3.2.2.3 Επιθέσεις χρησιμοποιώντας Adversarial training

Η τεχνητή νοημοσύνη έχει χρησιμοποιηθεί σε μεγάλο βαθμό για την ενίσχυση της άμυνας σε συστήματα όπως για παράδειγμα στην ανίχνευση κακόβουλου λογισμικού. Παρ' όλα αυτά, στις δύο παρακάτω μεθόδους παρουσιάζονται οι δυνατότητες που διαθέτει και στον σχεδιασμό κυβερνοεπιθέσεων χρησιμοποιώντας adversarial training για να αποφύγει τα συστήματα ανίχνευσης αλλά και να μαθαίνει να διεξάγει αυτοματοποιημένες επιθέσεις.

MalGAN

Παρουσίαση της επίθεσης

Ο αλγόριθμος MalGAN βασίζεται στα Generative Adversarial Networks (GANs) που αποτελούν μια προσέγγιση για generative modeling που χρησιμοποιεί deep learning μεθόδους. Το Generative modeling ανήκει στο unsupervised learning που περιλαμβάνει την αυτόματη ανακάλυψη και εκμάθηση των μοτίβων στα δεδομένα εισόδου (training data) με τέτοιο τρόπο ώστε το μοντέλο να μπορεί να χρησιμοποιηθεί για την παραγωγή νέων παραδειγμάτων που βασίζονται στο αρχικό σύνολο δεδομένων. Η βασική διαφορά με άλλες προσεγγίσεις είναι ότι οι δημιουργοί κακόβουλων λογισμικών δεν έχουν πρόσβαση στις λεπτομερείς παραμέτρους των μοντέλων μηχανικής εκμάθησης που χρησιμοποιούνται από συστήματα ανίχνευσης και ως εκ τούτου μπορούν να εκτελέσουν μόνο black box επιθέσεις. Στην περίπτωση όμως του MalGAN επιτυγχάνεται μεγάλο ποσοστό επιτυχίας χωρίς ανίχνευση από αυτά τα συστήματα επειδή διαθέτει επίσης έναν ανιχνευτή κακόβουλου λογισμικού όπου μπορεί μέσα από εκπαίδευση με Generative Adversarial Networks να εντοπίσει τι χρειάζεται για να μην γίνονται αντιληπτές οι επιθέσεις.

Γιατί είναι σημαντική

Η σημαντικότητα αυτής της επίθεσης προέρχεται από την πρωτοποριακή τεχνική που χρησιμοποιούν τα GANs. Τα GANs είναι ένας έξυπνος τρόπος εκπαίδευσης ενός generative model, όπου το πρόβλημα αντιμετωπίζεται ως περίπτωση «supervised learning» το οποίο έχει δύο υπο-μοντέλα: το generator model (G) που εκπαιδεύεται για να δημιουργεί νέα παραδείγματα και το discriminator model (D) που προσπαθεί να ταξινομήσει τα παραδείγματα ως αληθινά ή ψεύτικα (Goodfellow, 2014). Τα δύο μοντέλα εκπαιδεύονται μαζί ως αντίπαλοι σε ένα παιχνίδι μηδενικού αθροίσματος (όπου το κέρδος του ενός μοντέλου είναι ίσο με την ζημιά του άλλου),

έως ότου επιτευχθεί να ξεγελαστεί το discriminator model περίπου στο 50% των περιπτώσεων, που σημαίνει ότι το generator model δημιουργεί πειστικά παραδείγματα.

Παράδειγμα

Στην περίπτωση της επίθεσης με MalGAN που στηρίζεται στα παραπάνω, δημιουργούνται παραδείγματα κακόβουλου λογισμικού για να παρακάμψουν τα συστήματα ανίχνευσης τα οποία βασίζονται σε μηχανική μάθηση με black box προσέγγιση. Πιο αναλυτικά, στο πρώτο στάδιο πραγματοποιείται εκπαίδευση με δεδομένα από δείγματα που προέρχονται και από κακόβουλα προγραμμάτα και από μη κακόβουλα ώστε να μπορεί να γίνει ταξινόμηση στη συνέχεια. Ο τρόπος που γίνεται η εκπαίδευση βοηθά στην ελαχιστοποίηση της πιθανότητας ανίχνευσης των κακόβουλων προγραμμάτων με το να οδηγούν τον ταξινομητή να αναγνωρίζει εσφαλμένα κακόβουλα προγράμματα ως ασφαλή.

Εκτίμηση επικινδυνότητας

Η επίθεση με MalGAN είναι δύσκολο να αντιμετωπιστεί επειδή τα κακόβουλα προγράμματα έχουν υποστεί διαδικασίες ώστε να μην αναγνωρίζονται απο συστήματα ανίχνευσης και αυτό την καθιστά επικίνδυνη.

Μέθοδοι αντιμετώπισης

Η αξιολόγηση των αποτελεσμάτων (Hu, 2017) παρουσιάζει υψηλά ποσοστά επιτυχίας όπου τα συστήματα ανίχνευσης αδυνατούσαν να αναγνωρίσουν ότι κάποια προγράμματα που κρίνονταν ασφαλή ήταν στην πραγματικότητα κακόβουλα. Η ευελιξία που διαθέτει αυτή η επίθεση καθιστά την αντιμετώπισή της ιδιαίτερα δύσκολη, παρ' όλα αυτά, μια αμυντική προσέγγιση θα ήταν να επανεκπαιδευτεί ο ανιχνευτής κακόβουλου λογισμικού με βάση τα δημιουργημένα από το MalGAN παραδείγματα κακόβουλου λογισμικού. Ωστόσο, ακόμη και αν ο ανιχνευτής ενημερωθεί με τα παραδείγματα που δημιούργησε το MalGAN, τα νέα παραδείγματα κακόβουλου λογισμικού που θα δημιουργηθούν μελλοντικά από το MalGAN θα παραμείνουν μη ανιχνεύσιμα μέχρι την επόμενη ενημέρωση.

DeepHack

Παρουσίαση της επίθεσης

Το DeepHack είναι ένα εργαλείο ανοιχτού κώδικα που στηρίζεται σε τεχνητή νοημοσύνη και συγκεκριμένα στο Reinforcement Learning (RL), όπου είναι ένας τύπος μηχανικής μάθησης που επιτρέπει σε έναν πράκτορα (agent) να μαθαίνει μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Αυτό καθιστά το DeepHack ικανό να μπορεί να μαθαίνει να εισχωρεί στις βάσεις δεδομένων των εφαρμογών ιστού (web applications) χωρίς προηγούμενη γνώση του συστήματος. Σε σύγκριση με τις παραδοσιακές μεθόδους που ο εισβολέας απαιτείται να γράψει κώδικα για να εκτελέσει τις κακόβουλες ενέργειες, σε αυτή την περίπτωση δεν χρειάζεται.

Γιατί είναι σημαντική

Αυτή η μέθοδος είναι σημαντική επειδή ο επιτιθέμενος μπορεί να αποκτά πρόσβαση σε συστήματα με έναν δημιουργικό αλλά και αυτοματοποιημένο τρόπο χωρίς προηγούμενη γνώση της εφαρμογής και της βάσης δεδομένων.

Παράδειγμα

Για καλύτερη κατανόηση του τρόπου λειτουργίας αυτού του εργαλείου θα αναλυθεί μία επίθεση όπου κατάφερε να εκμεταλλευτεί ευπάθειες ενός τραπεζικού ιστοτόπου και να αποκτήσει πρόσβαση στη σχεσιακή βάση δεδομένων (relational database) και να πραγματοποιήσει SQL queries για απόκτηση πληροφοριών. Συγκεκριμένα, αφού εκπαιδευτεί ένα νευρωνικό δίκτυο με δεδομένα ώστε να μπορεί να κατανοήσει την σύνταξη ενός SQL ερωτήματος (SQL query), τότε προσπαθεί να προβλέψει τον επόμενο χαρακτήρα που ακολουθεί σε επόμενα SQL ερωτήματα με βάση τα δεδομένα που είχε συναντήσει κατά την εκπαίδευσή του. Σε κάθε προσπάθεια, ανταμείβεται με μια απάντηση που βασίζεται σε Boolean τιμή από τον απομακρυσμένο διακομιστή (remote server). Αυτό ενημερώνει το μοντέλο σχετικά με την εγκυρότητα του ερωτήματος που είχε συνταχθεί και έπειτα από πολλές προσπάθειες μαθαίνει ποιος χαρακτήρας πρέπει να τοποθετηθεί στη συνέχεια μέχρι η επιθυμητή πληροφορία να εξαχθεί επιτυχημένα από τη βάση δεδομένων (Petro, 2017).

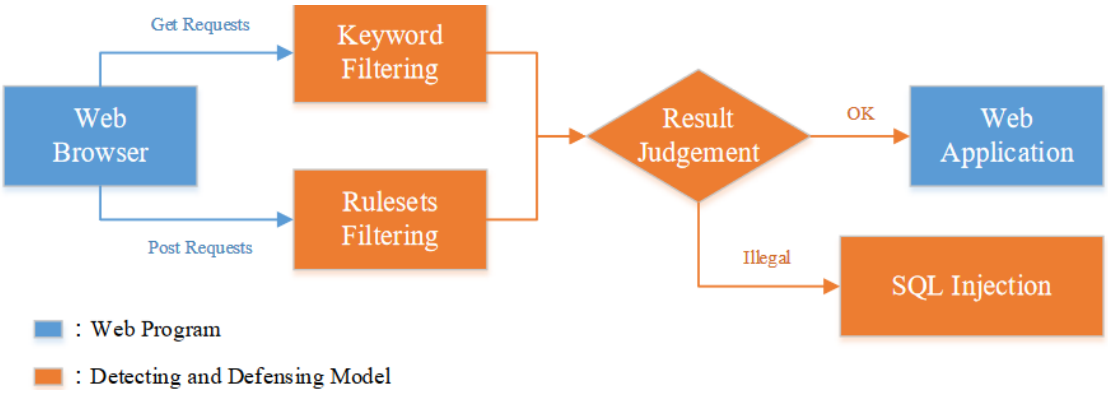
Εκτίμηση επικινδυνότητας

Η συγκεκριμένη επίθεση θεωρείται επικίνδυνη επειδή μπορεί να είναι πιο αποτελεσματική από την απλή SQL-Injection επίθεση λόγω της ικανότητας που αποκτά από την τεχνητή νοημοσύνη. Η διαφορά αυτής της μεθόδου με την παραδοσιακή επίθεση SQL-Injection είναι ότι μετά την εκπαίδευση του μοντέλου, η επίθεση πραγματοποιείται αυτόματα δοκιμάζοντας κάθε πιθανό χαρακτήρα για την σύνταξη του σωστού SQL query χρησιμοποιώντας brute force προσέγγιση που στηρίζεται επίσης και στα δεδομένα που είχε εκπαιδευτεί. Για παράδειγμα σε ένα input τύπου “SELECT * FRO”, το αμέσως επόμενο output θα είναι να προστεθεί το γράμμα “M”. Με αυτόν τον τρόπο καταφέρνει να δημιουργήσει τους κατάλληλους συνδυασμούς και να αποκτήσει πρόσβαση εάν υπάρχει η ευπάθεια που το επιτρέπει.

Μέθοδοι αντιμετώπισης

Αν και δεν υπάρχουν αμυντικοί μηχανισμοί για αυτόν τον τύπο επίθεσης σε περίπτωση που υπάρχουν ήδη ευπάθειες, η λήψη προληπτικών μέτρων μπορεί να μειώσει την πιθανότητα επιτυχίας της επίθεσης τόσο από την κλασική εκδοχή της όσο και από αυτήν που υποστηρίζεται από τεχνητή νοημοσύνη. Τα προληπτικά μέτρα αφορούν την εφαρμογή σωστών πρακτικών κατά της διαδικασία γραφής κώδικα και περιλαμβάνουν τον έλεγχο των στοιχείων που εισάγει ο χρήστης (Input type examination). Για παράδειγμα εάν ένα πεδίο ζητά αριθμητικό input, ο προγραμματιστής θα πρέπει να έχει σχεδιάσει το πρόγραμμα με τέτοιο τρόπο ώστε να απαγορεύει την είσοδο οποιουδήποτε άλλου τύπου δεδομένων (Mavromoustakos, 2016).

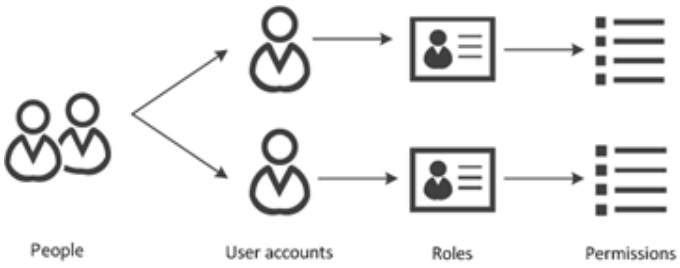
Τα επιτρεπόμενα πιθανά inputs δεν περιορίζονται μόνο στον τύπο δεδομένων αλλά μπορεί να εφαρμοστεί και φιλτράρισμα των HTTP requests που θα καθορίσει αν κάτι είναι επιτρεπτό ή όχι. Στην παρακάτω εικόνα παρουσιάζεται ένας τύπος φιλτραρίσματος που δεν πραγματοποιεί τον ίδιο έλεγχο σε Get και Post Requests όπως συνηθιζόταν παλαιότερα. Η εφαρμογή των κατάλληλων ελέγχων αναλόγως το request έχει σαν αποτέλεσμα την αποφυγή των ψευδώς θετικών (false positives) και ψευδώς αρνητικών αποτελεσμάτων (false negatives) (Song & Zhang, 2016).



Εικόνα 10: Μέθοδος φιλτραρίσματος σε input χρήστη.
Πηγή:ResearchGate[11]

Άλλο σημαντικό μέτρο είναι να χρησιμοποιούνται “Parameterized Queries” όπου κάθε στοιχείο εισόδου αποτελεί μια ξεχωριστή παράμετρο έτσι ώστε η βάση δεδομένων να τις αντιμετωπίζει πάντα ως δεδομένα και όχι ως μέρος μιας εντολής SQL (Song & Zhang, 2016). Τέλος, η βεβαίωση ότι κάθε εφαρμογή έχει τα δικά της διαπιστευτήρια βάσης δεδομένων και ότι αυτά τα διαπιστευτήρια έχουν τα ελάχιστα δικαιώματα που χρειάζεται η εφαρμογή είναι ιδιαίτερα σημαντικό μέτρο και πρέπει να τηρείται (Ma, Li & Lu, 2011).

11 https://www.researchgate.net/publication/314642498_A_Defense_Model_against_SQL_Injection_Based_on_Parameterized_Queries



Εικόνα 11: Σε κάθε ρόλο πρέπει να παρέχονται μόνο τα απαραίτητα δικαιώματα.
Πηγή: Cert[12]

3.3 Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάστηκαν τα στάδια διεξαγωγής μιας επίθεσης ξεκινώντας από την αναζήτηση στόχου, την συλλογή πληροφοριών, την δημιουργία προφίλ στόχου, την ανίχνευση ευπάθειας και στην συνέχεια πραγματοποιήθηκε παρουσίαση μεθόδων επιθέσεων με παραδείγματα που χρησιμοποιούν τεχνητή νοημοσύνη. Κατόπιν, εξηγήθηκαν οι τρόποι όπου αυτές οι μέθοδοι επιτυγχάνουν καλύτερα αποτελέσματα μέσα από αυτοματοποίησεις παραδοσιακών μεθόδων και εκπαίδευση μοντέλων με χρήση των κατάλληλων δεδομένων. Με αυτόν τον τρόπο οι επιθέσεις που διαξάγονται είναι ταχύτερες, έχουν μεγαλύτερη κλίμακα και λόγω της δυνατότητας εξατομίκευσης που παρέχουν, καταλήγουν να είναι και πιο αποτελεσματικές. Στη συνέχεια, για κάθε μία από τις παραπάνω επιθέσεις παρουσιάστηκαν τεχνικές αντιμετώπισης που σε μεγάλο βαθμό στηρίζονται και αυτές στην χρήση τεχνητής νοημοσύνης. Κλείνοντας, στον παρακάτω πίνακα παρουσιάζεται μια σύνοψη των μεθόδων κυβερνοεπίθεσης που αναλύθηκαν σε αυτό το κεφάλαιο. Ο πίνακας περιγράφει τις 6 μεθόδους ταξινομημένες σε 3 κύριες κατηγορίες και για κάθε μέθοδο γίνεται αναφορά στα χαρακτηριστικά τους, εστιάζοντας στις τεχνολογίες τεχνητής νοημοσύνης που χρησιμοποιούν για πετύχουν το σκοπό τους.

[12] https://www.cert.govt.nz

Κατηγορία επίθεσης	Επιθέσεις	Χαρακτηριστικά
Επιθέσεις χρησιμοποιώντας κακόβουλα bots	SNAP_R	Δυνατότητα αυτόματης δημιουργίας μηνυμάτων για εξατομικευμένη επίθεση τύπου Spear-phising. Χρησιμοποιεί: • Markov μοντέλα που έχουν εκπαιδευτεί με πρόσφατες δημοσιεύσεις του χρήστη-στόχου. → Εξατομικευμένο μήνυμα. • LSTM Neural Network που έχει εκπαιδευτεί σε πιο γενικά κείμενα. → φυσική χρήση γλώσσας. • Shortened links → Απόκρυψη πραγματικού URL.
	Fake reviews attack	Δημιουργία αυτοματοποιημένων αληθοφανών κριτικών για να επηρεάσουν την κοινή γνώμη. Χρησιμοποιεί: • LSTM Neural Network που έχει εκπαιδευτεί με πραγματικές κριτικές. → φυσική χρήση γλώσσας. • Αυτοματοποιημένη αντικατάσταση λέξεων με λέξεις κλειδιά → προσαρμογή γλώσσας στο υπο μελέτη αντικείμενο (π.χ εστιατόρια).
Επιθέσεις παράνομης πρόσβασης	Intelligent password brute-force attack	Brute force επίθεση που βασίζεται σε τεχνητή νοημοσύνη. Χρησιμοποιεί: • Εκπαίδευση ενός Recurrent Neural Network (RNN) μοντέλου → ικανότητα αναγνώρισης μοτίβων σε παλιούς κωδικούς πρόσβασης με αποτέλεσμα να δημιουργεί αυτόματα νέους υποψήφιους κωδικούς πρόσβασης με μεγαλύτερη πιθανότητα επιτυχίας.
	Επίθεση που στηρίζεται σε φωντητική σύνθεση	Απόκτηση πρόσβασης σε smartphone συσκευές μέσω φωνητικού βοηθού. Χρησιμοποιεί: • Μικρόφωνο της συσκευής → κρυφή ηχογράφηση του χρήστη. • Natural Language Processing (NLP) → επεξεργασία σε λέξεις - κλειδιά από την ηχογράφηση που θα χρησιμοποιηθούν στην επίθεση μέσω του φωνητικού βοηθού.

Επιθέσεις χρησιμοποιώντας Adversarial Training	MalGAN	Δυνατότητα μη ανιχνεύσιμων Black-Box επιθέσεων. Χρησιμοποιεί: • Αλγόριθμο που βασίζεται σε Generative Adversarial Network (GAN) → παραδείγματα κακόβουλου λογισμικού, τα οποία είναι σε θέση να παρακάμψουν μοντέλα ανίχνευσης που βασίζονται στη μηχανική μάθηση.
	DeepHack	Δυνατότητα εισχώρησης σε βάσεις δεδομένων εφαρμογών ιστού (web applications) χωρίς προηγούμενη γνώση του συστήματος. Χρησιμοποιεί: • Reinforcement Learning → Ικανότητα δημιουργίας αυτοματοποιημένων επιθέσεων τύπου SQL-Injection.

4

Συμπεράσματα

Στον χώρο της κυβερνοασφάλειας παρατηρούνται συνεχώς νέες προσεγγίσεις στο τρόπο που διεξάγονται κυβερνοεπιθέσεις. Τον τελευταίο καιρό οι προσπάθειες για βελτίωση των επιθέσεων έχουν ξεκινήσει να δίνουν έμφαση στην εφαρμογή τεχνικών που βασίζονται σε τεχνητή νοημοσύνη. Αυτή η μελέτη διερευνά τις επιθετικές δυνατότητες μέσω αυτοματοποίησης παραδοσιακών χειροκίνητων διαδικασιών, επιτρέποντας στους εισβολείς να διεξάγουν επιθέσεις ευρύτερου πεδίου, με μεγαλύτερη ταχύτητα και σε μεγαλύτερη κλίμακα. Σε αυτή την διπλωματική πραγματοποιήθηκε ανάλυση τεχνικών κυβερνοεπιθέσεων που χρησιμοπούν τεχνητή νοημοσύνη ξεκινώντας με την παρουσίαση τους, τον λόγο που θεωρούνται σημαντικές, παραδείγματα για να κατανοηθεί ο τρόπος λειτουργίας τους, ο βαθμός επικινδυνότητάς τους και τέλος, οι μέθοδοι αντιμετώπισής τους.

Μέσα από την ανάλυση αυτών των μεθόδων η διπλωματική εργασία συμβάλλει στην παροχή γνώσεων σχετικά με την κακόβουλη χρήση της τεχνητής νοημοσύνης στον κυβερνοχώρο, όπου είναι ένα θέμα που δεν έχει διερευνηθεί τόσο έντονα όσο οι μέθοδοι ανίχνευσης και άμυνας. Μέσα από αυτή την προσπάθεια επιδιώκεται να γνωστοποιηθούν οι κίνδυνοι που υπάρχουν όταν η τεχνητή νοημοσύνη χρησιμοποιείται για την βελτίωση των επιθέσεων. Η εστίαση και σε αυτήν την πλευρά θα έχει ως αποτέλεσμα οι ερευνητές να επινοούν πιθανές κακόβουλες τεχνικές πριν από τους παραβάτες και έτσι θα σχεδιάζονται προληπτικά ασφαλέστερα συστήματα.

Κατά τη ανάλυση που πραγματοποιήθηκε εντοπίστηκε ότι η τεχνητή νοημοσύνη μπορεί πράγματι να αποτελέσει σημαντικό εργαλείο στο σχεδιασμό και την πραγματοποίηση των κυβερνοεπιθέσεων καθώς οι ιδιότητες που την διακρίνουν από τις παραδοσιακές προσεγγίσεις είναι τόσο εξελιγμένες που πολλές φορές ο μόνος αποτελεσματικός τρόπος ενίσχυσης της

άμυνας των συστημάτων για αυτές τις επιθέσεις είναι να χρησιμοποιούνται επίσης ευφυείς αλγορίθμοι.

Κλείνοντας, ένα θέμα που αξίζει να μελετηθεί μελλοντικά είναι αυτό των δεδομένων (data sets) και η προστασία αυτών όταν αφορούν άτομα. Η σημασία των δεδομένων είναι ένας απολύτως κρίσιμος παράγοντας επειδή εάν δεν υπάρχουν τα σωστά δεδομένα, τότε ακόμη και οι πιο προηγμένες τεχνολογίες AI και Machine Learning χάνουν γρήγορα την ακρίβεια και τη λειτουργικότητά τους. Για τις επιχειρήσεις που συλλέγουν δεδομένα χρηστών είναι σίγουρα εύκολο να έχουν μεγάλη ποσότητα δεδομένων αλλά αυξάνονται ταυτόχρονα και οι περιορισμοί με τις νέες αλλαγές στην νομοθεσία περί προστασίας. Παρ' όλα αυτά η πρόσβαση σε δεδομένα θα καθορίσει την ύπαρξη μελλοντικής επιτυχίας και στον χώρο της τεχνητής νοημοσύνης και της κυβερνοασφάλειας.

5

Αναφορές

Abbott, R. G. (2015). Log Analysis of Cyber Security Training Exercises. *Procedia Manufacturing* .

Alhassan, M. M. (2017). Information Security in an Organization. *International Journal of Computer*.

Asaro, P. V. (1999). Effective Audit Trails A Taxonomy For Determination of Information Requirements. *PubMed*.

Bekrar. (2011). Finding software vulnerabilities by smart fuzzing. *Fourth IEEE International Conference on Software Testing* (σσ. 427-430). Berlin: IEEE.

Berry, M. W. (2020). *Supervised and Unsupervised Learning for Data Science.* New York: Springer.

Bloem, C. (2017, July 31). *84 Percent of People Trust Online Reviews As Much As Friends. Here's How to Manage What They See*. Ανάκτηση από Inc.: https://www.inc.com/craig-bloem/84-percent-of-people-trust-online-reviews-as-much-.html

Brauer, K. (2011, May 17). Authentication and security aspects in an international multi-user network. Turku, Finland.

Brynjolfsson. (2017). The Business of Artificial Intelligence. *Harvard Business Review 7*, 3-11.

Cheung, K. H. (2019). User Experience of Voice Assistants: Exploration of Anticipatory Assistance with Augmented Sensitivity toUsage Environments' Context. *Library and information sciences*, 5-6.

Deepa, T. P. (2014, January). Survey on need for cyber security in India. 1-5. Bangalore, Karnataka, India.

Dempsey, K. (2011). *Information Security Continuous Monitoring (ISCM) for Federal Information Systems and Organizations.* Gaithersburg: NIST.

Eggenschwiler, J. (2017). Accountability challenges confronting cyberspace governance. *Internet Policy Review*.

Feldman, R. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing.* Cambridge.

Goodfellow, I. J., Mirza, M., & Xu, B. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27* (σσ. 2672–2680). Montreal: MIT Press.

Halevi, T. (2015). Spear-Phishing in the Wild: A Real-World Study of Personality, Phishing Self-efficacy and Vulnerability to Spear-Phishing Attacks. *SSRN Electronic Journal*.

Han, S., & Yang, H. (2018). Understanding adoption of intelligent personal assistants: A parasocial relationship perspective. *Emerald Publishing Limited*, 618-636.

Hinde, S. (2003). The law, cybercrime, risk assessment and cyber protection. *Computers & Security*, 90-95.

Holeňa, M., & Kopp, M. (2020). *Classification Methods For Internet Applications.* New York: Springer.

Hu, W., & Ying, T. (2017). Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN. *arXiv e-prints*.

Igual, L. (2017). *Introduction to Data Science, A Python Approach to Concepts, Techniques and Applications.* New York: Springer.

Jakobsson, M., & Ratkiewicz, J. (2006). Designing ethical phishing experiments: a study of (ROT13) rOnl query features. *Proceedings of the 15th International Conference on World Wide Web* (σσ. 513–522). Edinburgh: Association for Computing Machinery.

Katzke, S. (2010). Guide for Applying the Risk Management Framework to Federal Information Systems. *NIST*.

Kowalski, G., & Maybury, M. (2000). *Information Storage and Retrieval Systems: Theory and Implementation.* New York: Springer.

LeCun, Y., & Boser, B. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 541-551.

Ma, X., Li, R., & Lu, Z. (2011). Specifying and enforcing the principle of least privilege in role-based access control. *Concurrency and Computation: Practice and Experience*, 1313-1331.

Manure, A. (2020). *Learn TensorFlow 2.0 Implement Machine Learning and Deep Learning Models with Python.* New York: Apress.

Mavromoustakos, S. (2016). Causes and Prevention of SQL Injection Attacks in Web. *Proceedings of the 4th International Conference on Information and Network Security* (σσ. 55-59). New York: Association for Computing Machinery.

McGuire, B. (2006, December). *The History of Artificial Intelligence.* Ανάκτηση από courses.cs.washington.edu: https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf

Mohammed, S. (2017). Password-based Authentication in Computer Security: Why is it still there? *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, 01-05.

Moniz, P. M. (2011, April). *Confidentiality, integrity and non-repudiation smartgrids.* Ανάκτηση από core.ac.uk: https://core.ac.uk/download/pdf/12428059.pdf

Müller, A. C. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists.* California: O'Reilly.

Nieles, M. (2017). *An Introduction to Information Security.* Gaithersburg: NIST.

Oren, J. C. (2016). *Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems.* Gaithersburg: NIST.

Padmavathi, G. (2019). *Handbook of Research on Machine and Deep Learning Applications for Cyber Security.* Pennsylvania : IGI Global.

Patil, S., Alabdullah, B., & Alrumayh, N. (2020). Detecting Phishing Websites Using Machine Learning. *IEEE*, 1-6.

Perera, I. (2018). Cyberattack Prediction Through Public TextAnalysis and Mini-Theories. *IEEE*, 3001-3010.

Perveen, N., Qaisar Rasool, & Akhtar, N. (2016). Sentiment Based Twitter Spam Detection. *International Journal of Advanced Computer Science and Applications*, 568-573.

Petro, D. (2017). *Weaponizing Machine Learning: Humanity was Overrated Anyway.".* Ανάκτηση από defcon: https://www.defcon.org/html/defcon-25/dc-25-speakers.html

Raghavan, P., Schütze, H., & Manning, C. (2009). *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 386–408.

Sahami, M., & Dumais, S. (1998). A Bayesian approach to filtering junk email. *AAAI*.

Seymour, J., & Tully, P. (2016). *Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter.* Ανάκτηση από Blackhat: https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf

Shandilya, S. K. (2020). *Advances in Cyber Security Analytics and Decision Systems.* New York: Springer.

Shanmugasundara, J., Shekita, E., & Kiernan, J. (2001). A general technique for querying XML documents using a relational database system. *SIGMOD*, 20–26.

Shelupanov, A., Konev, A., & Kostyuchenko, E. (2019). Information Security Methods. *Symmetry*.

Solanki, V. K. (2020). *Cyber Defense Mechanisms.* Florida: CRC Press.

Song, K., & Zhang, H. (2016). A Defense Model against SQL Injection Based on Parameterized Queries. *5th International Conference on Computer Sciences and Automation Engineering (ICCSAE 2015)* (σ. 516). Sanya: Atlnatis Press.

Stoneburner, G., Goguen, A., & Feringa, A. (2002). *Risk Management Guide for Information Technology Systems.* Gaithersburg: NIST.

Sutton, R. S. (2015). *Reinforcement Learning: An Introduction.* London: MIT Press.

Trieu, K., & Yang, Y. (2018). Artificial Intelligence-Based Password Brute Force Attacks. *AISEL*.

Trnka, M. (2018). Survey of Authentication and Authorization forthe Internet of Things. *Sec. and Commun. Netw.*

Twitter. (2020). *help.twitter.com*. Ανάκτηση από https://help.twitter.com/en/safety-and-security/phishing-spam-and-malware-links

Vieira, M., Antunes, N., & Madeira, H. (2009). Using Web Security Scanners to Detect Vulnerabilities in Web Services. *2009 IEEE/IFIP International Conference on Dependable Systems Networks* (σσ. 566-571). Lisbon: IEEE.

Yao, Y. (2017). Automated Crowdturfing Attacks and Defenses in Online Review Systems. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (σσ. 143–1158). New York: Association for Computing Machinery.

Yildirim, M. (2019). Encouraging users to improve password security and memorability. *International Journal of Information Security*, 741–759.

Zhang, R., Chen, X., & Wen, S. (2018). Using AI to Hack IA: A New Stealthy Spyware Against Voice Assistance Functions in Smart Phones. *IEEE*.

Zhao, L. (2018). Summary of vulnerability related technologies based on machine learning. *AIP Conference Proceedings* . Beijing: AIP.