PROJEKT

Multitenantné operačné centrum kybernetickej bezpečnosti riešené
ako otvorená cloudová služba s prvkami strojového učenia

Analýza dostupných dataset-ov pre detekciu útokov v sieťovej prevádzke

Príloha k výstupu V7

Previazané na pracovný balík KPB3

Obsah

V rámci kybernetickej bezpečnosti existuje viacero druhov útokov a v ďalšom texte budú jednotlivé dostupné datasety a nástroje uvádzané podľa príslušnosti k danému typu útoku, ktoré jednotlivý dataset pokrýva. Napriek tomu, že SPAM a phishing predstavujú rôzne vektory útoku, vzhľadom k nedostatku datasetov pre tieto typy útokov sú v ďalšom texte zhrnuté v jednej sekcii.

Typy útokov

Narušenie siete

KDD'99 a NSL-KDD.

Dataset KDD'99 je štatisticky vopred spracovaný súbor údajov, ktorý je k dispozícii od roku 1999 od DARPA, jedná sa o aktualizovanú verziu DARPA98. Jedná sa o najčastejšie používaný dataset. Dataset má tri komponenty, základné a obsahové príznaky a príznaky súvisiace s prevádzkou, čo predstavuje celkovo 41 príznakov/parametrov pre bežnú prevádzku a prevádzku so simulovanými útokmi. Dataset NSL-KDD je verziou datasetu KDD'99, v ktorej sú odstránené redundantné záznamy, čo napomáha klasifikačným systémom produkovať nezaujaté výsledky (unbiased). Datasety obsahujú rôzne typy útokov, ako napríklad Neptune-DoS, pod-DoS, Smurf-DoS a buffer-overflow.

https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Moore set

Tento dataset pripravili v roku 2005 výskumníci z Intel Research. Zahŕňa trasovanie reálnej prevádzky zhromaždené monitorovaním vysoko-rýchlostnej siete. Každý objekt v Moore datasete predstavuje jeden tok TCP paketov medzi klientom a serverom, ktorý pozostáva z 248 príznakov. Informácie v príznakoch sú získané z hlavičiek paketov, zatiaľ čo klasifikačné triedy boli odvodené pomocou analýzy založenej na obsahu.

https://www.cl.cam.ac.uk/research/srg/netos/projects/archive/nprobe/data/papers/sigmetrics/index.html

LBNL2005

V Národnom laboratóriu Lawrence Berkeley (LBNL) v roku 2005 zozbierali prevádzku v LBNL/ICSI v rámci projektu Enterprise Tracing Project počas troch mesiacov v rokoch 2004 a 2005 na dvoch smerovačoch. Obsahuje úplné hlavičky sieťovej

prevádzky zaznamenané v stredne veľkom podniku pokrývajúcom 22 podsietí a zahŕňa sledovacie údaje pre širokú škálu prevádzky vrátane webu, e-mailu, zálohovania a streamovaných médií. Keďže sledovanie prevádzky nie je úplne anonymizované, všetky pakety neobsahujú užitočné dáta (payload).

https://www.icir.org/enterprise-tracing/download.html

CAIDA datasety

Centrum pre aplikovanú analýzu internetových údajov (CAIDA) so sídlom v superpočítačovom centre Kalifornskej univerzity v San Diegu zhromažďuje rôzne údaje z geograficky a topologicky odlišných lokalít a sprístupňuje ich výskumnej komunite v čo najväčšej možnej miere, pričom rešpektuje súkromie jednotlivcov a organizácií, ktoré poskytnú údaje alebo prístup k sieti. CAIDA-DDoS Dataset obsahuje približne jednu hodinu anonymizovanej prevádzky z DDoS útoku. Tento typ útoku ("odmietnutie služby") sa pokúša zabrániť prístupu k cieľovému serveru využitím všetkého výpočtového výkonu servera a celej šírky pásma v sieti, ktorá spája server s internetom. Trasovanie zahŕňa iba prevádzku útoku na obeť a reakcie obete na útok. Neútočná prevádzka bola v najväčšej možnej miere odstránená.

https://www.caida.org/catalog/datasets/ddos-20070804_dataset/

Kyoto2006+

Kyoto2006+ je verejne dostupný referenčný dataset pozostávajúci z 24 štatistických príznakov, ktorý je postavený na monitorovaní troch rokoch sieťovej prevádzky v rokoch 2006 - 2009. Zahŕňa bežné servery aj honeypot-y rozmiestnené na Kjótskej univerzite v Japonsku označené ako normálny (žiadny útok), útok (známy útok) a neznámy útok. Zahŕňa rôzne útoky vykonávané proti honeypotom, ako je shellcode, exploit-y, DoS, skenovanie portov, spätný rozptyl a malvér. Aktualizovaná verzia datasetu obsahuje ďalšie údaje zozbierané v rokoch 2006 - 2015.

http://www.takakura.com/Kyoto_data/

UNIBS2009

Dataset zostavený Univerzitou v Brescii v roku 2009. Pozostáva zo sledovania prevádzky zhromaždenej prostredníctvom nástroja Tcpdump na okrajovom smerovači univerzitnej kampusovej siete počas troch po sebe nasledujúcich pracovných dní (2009.9.30, 2009.10.1 a 2009.10.02), pričom sieť bola pripojená k internetu prostredníctvom 100 Mbps uplink. Dataset poskytuje skutočné label-y (označenia typu

prevádzky) a sledovanie prevádzky zahŕňa web (HTTP a HTTPS), mail (POP3, IMAP4, SMTP a ich varianty Secure Sockets Layer), Skype, P2P (BitTorrent, Edonkey), SSH (Secure Shell), FTP (File Transfer Protocol) a MSN.

https://fkie-cad.github.io/COMIDDS/content/datasets/unibs/

UNB ISCX-2012

Dataset Installation Support Center of Expertise (ISCX)-2012 bol pripravený na ISCX na University of New Brunswick. Je postavený na 7-dňovej sieťovej prevádzke a pozostáva z viac ako dvoch miliónov prevádzkových paketov charakterizovaných 20 príznakmi, ktoré majú nominálne, celočíselné alebo desatinné hodnoty. Dataset obsahuje úplný payload paketov vo formáte pcap.

https://www.unb.ca/cic/datasets/ids.html

CTU-13

Dataset CTU-13 zostavilo České vysoké učení technické. Pozostáva z prevádzky botnetov zachytených na univerzite v roku 2011. Dataset obsahuje trinásť scenárov, ktoré pokrývajú rôzne útoky botnetov, ktoré používajú rôzne protokoly a vykonávajú rôzne akcie zmiešané s bežnou prevádzkou a prevádzkou na pozadí. Dataset je k dispozícii vo forme jednosmerného toku, obojsmerného toku a zachytávania paketov.

https://www.stratosphereips.org/datasets-ctu13/

UNSW-NB15

Dataset UNSW-NB15 bol zostavený v roku 2015 Univerzitou Nového Južného Walesu v Canberre na School of Engineering and IT, UNSW Canberra na ADFA pomocou malej, emulovanej siete počas 31 hodín získavaním normálnych a škodlivých nespracovaných sieťových paketov. Pozostáva z deviatich typov útokov: analysis, backdoors, DoS, exploits, generic, fuzzers, reconnaissance, shell code and worms. Pozostáva z viac ako dvoch miliónov záznamov, z ktorých každý sa vyznačuje 49 príznakmi s nominálnymi, celočíselnými alebo desatinnými hodnotami.

https://research.unsw.edu.au/projects/unsw-nb15-dataset

AWID 2015

Aegean Wi-Fi Intrusion Dataset (AWID), publikovaný v roku 2015, obsahuje najväčšie množstvo údajov o sieti Wi-Fi (normálna prevádzka a prevádzka obsahujúca útok) zozbieraných v skutočných sieťových prostrediach. 16 typov útokov možno zoskupiť do 3 kategórií: flooding, impersonation, and injection. Dataset obsahuje viac ako 5 miliónov vzoriek, z ktorých každá sa vyznačuje 154 príznakmi, ktoré predstavujú polia rámca WLAN spolu s metadátami fyzickej vrstvy.

https://icsdweb.aegean.gr/awid/download-dataset

ISCXVPN2016

ISCXVPN2016, ktorý zverejnila UNB v roku 2016, zahŕňa prevádzku zachytenú pomocou nástrojov Wireshark a tcpdump, pričom generuje celkové množstvo 28 GB údajov. Pre VPN bol použitý externý poskytovateľ služieb VPN pripojený pomocou OpenVPN (režim UDP). Na generovanie SFTP a FTPS prevádzky bol použitý externý poskytovateľ služieb a Filezilla ako klient.

https://www.unb.ca/cic/datasets/vpn.html

CIDDS

Coburg Intrusion Detection Datasets (CIDDS), pripravené na Coburg University of Applied Sciences (Hochschule Coburg), pozostávajú z niekoľkých datasetov založených na sieťových tokoch vytvorených vo virtuálnych prostrediach pomocou OpenStack. Najpoužívanejší súbor údajov databázy CIDDS, CIDDS-001, vydaný v roku 2017, pokrýva štyri týždne jednosmerných prevádzkových tokov, z ktorých každý sa vyznačuje 19 príznakmi s nominálnymi, celočíselnými alebo desatinnými hodnotami. Dataset zahŕňa útoky ako DoS, skenovanie portov a brute-force SSH.

https://github.com/markusring/CIDDS

CIC-IDS2017

Dataset bol vytvorený v emulovanom sieťovom prostredí v CIC. Je postavený na 5 dňoch (3. júla až 7. júla 2017) sieťovej prevádzky a zahŕňa celý rad najbežnejších typov útokov vrátane FTP patator, SSH patator, DoS slowloris, DoS Slowhttptest, DoS Hulk, DoS GoldenEye, Heartbleed, brute-force, XSS, SQL Injection, Infiltration, Bot, DDoS a Port Scan, z ktorých každý sa vyznačuje 80 príznakmi extrahovanými pomocou nástroja CICFlowMeter. Dataset obsahuje aj úplný payload paketov vo formáte pcap.

https://www.unb.ca/cic/datasets/ids-2017.html

CIC-DDoS2019

Verejne dostupný dataset pre detekciu narušenia s útokmi DoS aplikačnej vrstvy od Kanadského inštitútu pre kybernetickú bezpečnosť a obsahuje najnovšie DDoS útoky, ktoré sú podobné údajom z reálneho sveta. Zahŕňa výsledky analýzy sieťových technológií pomocou nástroja CICFLOWMeter-V3, ktorá obsahuje tok tokenov založený na zdroji časovej pečiatky a cieľovom zdroji IPS a portových protokoloch a útokoch. Autori vykonali osem DoS útokov na aplikačnú vrstvu. Normálne správanie používateľov bolo generované kombináciou výsledných stôp s prevádzkou bez útokov datasetu ISCX 2012. Dataset je k dispozícii vo formáte založenom na paketoch a obsahuje 24 hodín sieťovej prevádzky.

https://www.unb.ca/cic/datasets/ddos-2019.html

NDSec-1

Dataset obsahuje súbory s prevádzkou a logovacie súbory sieťových útokov syntetizované výskumníkmi zo sieťových zariadení. Je verejne dostupný a v roku 2016 bol zachytený vo formáte založenom na balíkoch. Obsahuje ďalšie informácie o syslogu a denníku udalostí systému Windows. Kompozície útokov zahŕňajú botnet, brute-force (proti FTP, HTTP a SSH), DoS (HTTP, SYN a UDP flooding), exploits, skenovanie portov, spoofing a injektovanie XSS/SQL.

https://www2.hs-fulda.de/NDSec/NDSec-1/Files/

BoT-IoT

Dataset obsahuje viac ako 72 miliónov záznamov vrátane útokov typu DDoS, DoS, OS, skenovania služieb, keyloging a exfiltrácie údajov. Nástroj Node-red bol použitý na simuláciu správania siete zariadení IoT. MQTT protokol vytvára komunikáciu medzi strojmi (M2M). Testovacie scenáre IoT sú meteorologická stanica, inteligentná chladnička, pohybom aktivované svetlá, diaľkovo aktivovaná garážová brána a inteligentný termostat.

https://www.kaggle.com/datasets/vigneshvenkateswaran/bot-iot

IoT-23

Dataset pozostáva z 23 sieťových scenárov prevádzky IoT, vrátane 20 PCAP súborov z infikovaných zariadení IoT a troch skutočných sieťových prenosov IoT. Pre každý škodlivý scenár bol vytvorený špecifický malvér, ktorý bol spustený na Raspberry Pi pomocou niekoľkých protokolov a rôznych akcií. Zachytávanie sieťovej prevádzky pre neškodné scenáre bolo získané zo sieťovej prevádzky troch skutočných zariadení IoT: inteligentnej LED lampy Philips HUE, domáceho inteligentného osobného asistenta Amazon Echo a inteligentného zámku dverí Somfy. Škodlivé aj neškodné scenáre boli spustené v kontrolovanom sieťovom prostredí s neobmedzeným internetovým pripojením, ako každé skutočné zariadenie IoT.

https://www.stratosphereips.org/datasets-iot23

UGR'16

Dataset UGR'16, ktorý v roku 2018 navrhli Maciá-Fernández a kol. a obsahuje sledovanie siete NetFlow zozbierané zo skutočnej siete ISP úrovne 3, ktorá sa skladá z virtualizovaných a hostovaných služieb niekoľkých organizácií a klientov vrátane WordPress, Joomla, e-mailu, FTP atď. Senzory NetFlow boli nainštalované v hraničných smerovačoch siete, aby zachytili všetku prichádzajúcu a odchádzajúcu prevádzku od poskytovateľa internetových služieb. Poskytované sú dva datasety: jeden pre trénovacie modely (kalibračný súbor) a druhý pre testovanie výstupov modelov (testovací súbor).

https://nesg.ugr.es/nesg-ugr16/

Kitsune2019

Dataset o sieťových útokoch Kitsune, Kitsune2019, bol pripravený na Ben-Gurionovej univerzite v Negeve v Izraeli a bol vydaný v máji 2018. Dataset sa skladá z 9 súborov pokrývajúcich 9 rôznych situácií útokov na komerčný video monitorovací systém založený na IP a sieť IoT: OS (Operating System) Scan, Fuzzing, Video Injection, ARP Man in the Middle, Active Wiretap, SSDP Flood, SYN DoS, Secure Sockets Layer Renegotiation a Mirai Botnet. Obsahuje 27 170 754 vzoriek, z ktorých každá sa vyznačuje 115 skutočnými príznakmi.

https://www.kaggle.com/datasets/ymirsky/network-attack-dataset-kitsune

NETRESEC datasety

NetreSec je softvérová spoločnosť, ktorá sa špecializuje na monitorovanie sieťovej bezpečnosti a forenznú analýzu. Rovnako udržiavajú pcap repozitáre

zhromaždené z rôznych internetových zdrojov. Jedná sa o zoznam voľne dostupných verejných úložísk so zachytenou verejnou internetovou paketovou prevádzkou. Väčšina webových stránok uvedených na ich webových stránkach poskytuje súbory Full Packet Capture (FPC), iné však poskytujú iba skrátený obsah.

https://www.netresec.com/?page=PcapFiles

MAWI

Archív MAWI pozostáva z pokračujúcej zbierky trasovania dennej internetovej prevádzky zachytenej v rámci chrbticovej siete WIDE na niekoľkých vzorkovacích miestach. Nástroj Tcpdump sa používa na načítanie trasovacích údajov a IP (Internet Protocol) adresy v trasách sú šifrované pomocou upravenej verzie nástroja Tcpdpriv (MAWI Working Group Traffic Archive (http://www.wide.ad.jp)). Vzorkovací bod-F pozostáva z denných záznamov trasovania na tranzitnom spojení WIDE k dodávateľskému poskytovateľovi internetových služieb a je v prevádzke od 01/07/2006.

https://mawi.wide.ad.jp/mawi/

ADFA

Dataset bol vyvinutý na University of New South Wales na hodnotenie sysémov HIDS v OS Linux a Windows. Tento dataset bol zamýšľaný ako zástupca datasetu zahŕňajúceho štruktúru a metodológiu moderných bezpečnostných útokov a ako náhrada starších datasetov ako sú KDD alebo UNM. Dataset obsahuje záznamy z operačných systémov Linux aj Windows; sú vytvorené z vyhodnocovania systémových volaní HIDS. Ubuntu Linux verzie 11.04 bol použitý ako hostiteľský operačný systém na zostavenie ADFA-LD (Creech & Hu, 2014b). Niektoré z prípadov útokov v ADFA-LD boli odvodené od nového zero-day malvéru, vďaka čomu je tento súbor údajov vhodný na zdôraznenie rozdielov medzi prístupmi SIDS a AIDS k detekcii narušenia siete. Zahŕňa tri odlišné kategórie údajov, pričom každá skupina údajov obsahuje surové dáta s trasovaním systémových volaní. Každý tréningový dataset bol zhromaždený od hostiteľa pre bežné aktivity, pričom správanie používateľov siahalo od prehliadania webu až po prípravu dokumentu LATEXX.

https://research.unsw.edu.au/projects/adfa-ids-datasets

NGIDS-DS

Dataset bol navrhnutý na University of New South Wales na hodnotenie linuxových HIDS. Sieťové pakety medzi útočiacim systémom a systémom obete boli

zachytené v jednom súbore pcap. Pozostáva z obrovského množstva normálnych a abnormálnych vektorov generovaných pomocou nástroja IXIA perfect-storm uloženého v niekoľkých súboroch CSV. Systémové volania a časy ich vykonávania boli zachytené z operačného systému Linux obete ako sady príznakov, ktoré sa použijú na vyhodnotenie efektívnosti nových HIDS.

https://fkie-cad.github.io/COMIDDS/content/datasets/ngids_dataset/

SPAM a phishing

SPAM Email

Dataset o e-mailoch SPAM obsahuje celkovo 4601 e-mailov vrátane 1813 spamových e-mailov a 2788 legitímnych e-mailov, z ktorých každý sa vyznačuje 58 atribútmi. V roku 1999 ho Hewlett Packard daroval UCI Machine Learning Repository.

http://archive.ics.uci.edu/dataset/94/spambase

Ling-Spam

Dataset Ling-Spam, ktorý navrhol Androutsopoulos a kol. v roku 2000, obsahuje spam aj legitímne e-maily získané z e-mailového distribučného zoznamu, lingvistického zoznamu, so zameraním na jazykové záujmy týkajúce sa výskumných príležitostí, pracovných ponúk a diskusií o softvéri. Dataset obsahuje 2 893 rôznych e-mailov, z ktorých 2 412 sú skutočné e-maily zozbierané zo zoznamu a 481 sú spamové e-maily získané od jedného z autorov korpusu.

https://metatext.io/datasets/ling-spam-dataset

WEBSPAM-UK2006

Dataset WEBSPAM-UK2006 bol vytvorený ako množina webových stránok (s doménou UK) stiahnutých Laboratóriom webových algoritmov Univerzity v Miláne (Università degli Studi di Milano) a manuálne vyhodnotený skupinou dobrovoľníkov v roku 2006. Dataset pozostáva z jednotlivých anotácií, URL a hypertextových odkazov a obsahu HTML stránok 77 741 046 webových stránok.

https://chato.cl/webspam/datasets/uk2006/

SpamAssassin

Apache SpamAssassin je antispamová platforma s otvoreným zdrojovým kódom, ktorá poskytuje filter pre klasifikáciu e-mailov a blokovanie spamu. SpamAssassin Public mail korpus je výber 6 047 e-mailov, ktoré pripravil SpamAssassin v roku 2006. Z celkového počtu je 1 897 spamových správ a 4 150 legitímnych e-mailov.

https://www.kaggle.com/datasets/beatoa/spamassassin-public-corpus

TREC2007 Public corpus

Verejný korpus TREC 2007 obsahuje všetky e-mailové správy doručené na konkrétny server. Server obsahoval niekoľko účtov, ktoré sa prestali používať a niekoľko účtov typov "honeypot" zverejnených na webe, ktoré boli použité na registráciu do niekoľkých služieb, niektoré legitímne a niektoré nie. Dataset TREC obsahuje 75 419 správ, z ktorých 25 220 sú legitímne e-maily a 50 199 nevyžiadané správy, pričom sú jednotlivé správy rozdelené do troch sub-korpusov.

https://plg.uwaterloo.ca/~gvcormac/treccorpus07/

SMS Spam Collection

Dataset o zbere spamu SMS je verejne dostupný korpus, ktorý vytvoril Almeida a spol. v roku 2011. Ide o anotovaný dataset so 5574 SMS správami zozbieraných z mobilných telefónov, pričom 747 správ je typu spam a 4827 typu ham.

https://archive.ics.uci.edu/dataset/228/sms+spam+collection

"Gold Standard" opinion spam

Dataset navrhol Ott et al. v roku 2011. Korpus obsahuje 1 600 recenzných textov, 800 klamlivých a 800 pravých, o 20 hoteloch v oblasti Chicaga. Skutočné recenzie boli získané z recenzovaných webových stránok ako TripAdvisor, Expedia a Yelp a tie klamlivé boli vytvorené pomocou Amazon Mechanical Turk (AMT). V datasete obsahuje 400 písaných recenzií s negatívnou citovou polaritou a 400 s pozitívnou citovou polaritou.

https://aclanthology.org/P11-1032/

Spear phishing email dataset (2011) & Benign email dataset (2013)

Tieto datasety boli pripravené službou skenovania podnikovej pošty spoločnosti Symantec. Dataset *Spear phishing email* obsahuje 1 467 e-mailov z 8 kampaní a *Bening*

email obsahuje 14 043 e-mailov. E-maily boli odoslané v rokoch 2011 až 2013 a obsahujú prílohy, anonymné informácie o zákazníkoch a osobné údaje.

https://www.kaggle.com/c/19fall-spear-phishing-detection/data

EPVME-Dataset

Dataset zahŕňa 6 nových typov scenárov škodlivých e-mailových útokov, ktoré využívajú zraniteľnosti v mechanizmoch autentifikácie zabezpečenia e-mailov (SPF a DMARC) a vykresľovaní používateľského rozhrania e-mailov. Obsahuje 37 283 škodlivých e-mailov a všetky vzorky e-mailov majú informácie o hlavičke a tele e-mailu. Vo vytvorených škodlivých ukážkach sa s výnimkou niektorých špecifických polí, do ktorých sa vkladajú užitočné dáta (payload), náhodne vyberie iný obsah vrátane položiek Od, Do, Predmet a hlavný text.

https://github.com/sunknighteric/EPVME-Dataset

MovieLens dataset

GroupLens Research zhromaždil a sprístupnil súbory údajov o hodnotení z webovej stránky MovieLens (https:// movie lens. org). Dataset sa zbieral v rôznych časových obdobiach v závislosti od veľkosti súboru. MovieLens 20 M obsahuje 20 miliónov hodnotení a 465 000 tagov aplikovaných na 27 000 filmov od 138 000 používateľov zhromaždených od januára 1995 do marca 2015 [120].

https://www.kaggle.com/datasets/grouplens/movielens-20m-dataset

Netflix

Dataset Netflix pozostáva zo zoznamov všetkých filmov a televíznych relácií dostupných na Netflixe spolu s podrobnosťami, ako sú - obsadenie, režiséri, hodnotenia, rok vydania, trvanie atď.

https://www.kaggle.com/datasets/shivamb/netflix-shows

PhishTank, DeltaPhish, Phish-Labls a Anti-Phishing Working Group (APWG9)

sú antiphishingové zdroje, ktoré verejne hlásia phishingové webové stránky v snahe znížiť podvody a krádeže identity spôsobené phishingom a súvisiacimi incidentmi.

https://www.phishtank.com/

https://www.phishlabs.com/

https://apwg.org/

Malware

Malrec Dataset

Malvérový sandbox systém Malrec, ktorý využíva deterministický záznam a prehrávanie celého systému PANDA na zachytenie vysoko verných a celo-systémových stôp spustenia malvéru s nízkou časovou a priestorovou réžiou. Dataset predstavuje nový súbor 66 301 záznamov malvéru zozbieraných za obdobie dvoch rokov. Systém Malrec a jeho dataset môžu pomôcť poskytnúť štandardizovaný benchmark na hodnotenie výkonnosti budúcich dynamických analýz.

https://giantpanda.gtisc.gatech.edu/malrec/dataset/

The Malimg dataset

Tento dataset navrhla v roku 2011 Kalifornská univerzita v Santa Barbare a obsahuje 9458 obrázkov malvéru z 25 rodín.

https://paperswithcode.com/dataset/malimg

The Malware Genome Project

Dataset navrhli výskumníci zo Štátnej univerzity v Severnej Karolíne v roku 2011, obsahuje 1260 vzoriek malvéru pre Android patriacich do 49 rôznych rodín malvéru zozbieraných od augusta 2010 do októbra 2011.

http://www.malgenomeproject.org/

Malheur

Dataset navrhnutý v roku 2011, je nástroj pre automatickú analýzu správania malvéru v prostredí sandbox.

https://github.com/rieck/malheur

Malicia dataset

Dataset publikovaný v roku 2013, obsahuje 11 688 binárnych súborov malvéru zhromaždených z 500 serverov na sťahovanie počas 11 mesiacov vo formáte Windows Portable Executable. Cieľom práce bolo identifikovať hostiteľov, ktorí šíria malvér vprostredí a zbierať vzorky malvéru. Na zber vzoriek malvéru zriadili honeypot a klienti v tomto honeypote sa odvolávali na databázu URL malvéru na stiahnutie webovej stránky rozlíšením IP adresy.

https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset

CTU-Malware

Dataset CTU-Malware, ktorý tiež zostavilo České vysoké učení technické, pozostáva zo stoviek záznamov (nazývaných scenáre) rôznych vzoriek malvérovej komunikácie. Korpus zahŕňa malvér aj bežné vzorky.

https://www.stratosphereips.org/datasets-ctu13/

Microsoft Malware Classification Challenge

V roku 2015 spoločnosť Microsoft spustila Microsoft Malware Classification Challenge spolu s vydaním súboru údajov pozostávajúceho z viac ako 20 000 vzoriek malvéru patriacich do deviatich rodín. Každý súbor malvéru obsahuje identifikátor, čo je 20-znaková hodnota hash, ktorá jednoznačne identifikuje súbor a označenie triedy, čo je celé číslo, ktoré predstavuje jednu z deviatich rodín, do ktorých môže malvér patriť.

https://www.kaggle.com/c/malware-classification

USTC-TFC2016

Dataset USTC-TFC2016, publikovaný v roku 2017, pozostáva z desiatich typov malvérovej prevádzky z verejných webových stránok, ktoré boli zhromaždené zo skutočného sieťového prostredia v rokoch 2011 až 2015. Spolu s takouto škodlivou prevádzkou obsahuje neškodná časť desať typov bežnej prevádzky, ktoré boli zhromaždené pomocou IXIA BPS, profesionálneho zariadenia na simuláciu sieťovej prevádzky. Veľkosť datasetu je 3,71 GB vo formáte pcap.

https://www.kaggle.com/datasets/randasrour/ustctfc2016

CICAndMal2017

Dataset malvéru pre Android CICAndMal2017, ktorý v roku 2018 zverejnil CIC, pozostáva zo štyroch kategórií malvéru a to adware, ransomware, scareware a SMS Malware a 80 príznakov sieťovej prevádzky extrahovaných pomocou nástroja CICFlowMeter. Dataset obsahuje 5 065 neškodných aplikácií z Google Play publikovaných v rokoch 2015, 2016 a 2017 a 426 vzoriek malvéru patriacich do 42 jedinečných rodín malvéru. Dataset je plne anotovaný a obsahuje sieťovú prevádzku, protokoly, hovory API/SYS, štatistiky telefónov a výpisy pamäte rodín malvéru.

https://www.unb.ca/cic/datasets/andmal2017.html

CICMalDroid2020

Dataset je komplexnou množinou vzoriek malvéru pre Android určených na výskum a analýzu v oblasti mobilnej bezpečnosti. Obsahuje viac ako 17 000 aplikácií pre Android (APK) vrátane škodlivých aj neškodných vzoriek. Škodlivé vzorky sú rozdelené do rôznych skupín malvéru, ako je adware, ransomware, scareware a SMS malware. Každá vzorka je označená ako neškodná alebo škodlivá s ďalšími metaúdajmi o rodine malvéru. Dataset bol zozbieraný pomocou kombinácie reálnych aplikácií z obchodu Google Play a známych vzoriek malvéru z rôznych zdrojov. Na extrahovanie príznakov zo súborov APK sa použili techniky dynamickej a statickej analýzy vrátane volaní API, povolení, sieťovej prevádzky a systémových volaní. Tieto príznaky sú užitočné na trénovanie modelov strojového učenia na detekciu a klasifikáciu malvéru.

https://www.unb.ca/cic/datasets/maldroid-2020.html

VxHeavens

je webová stránka venovaná poskytovaniu informácií o škodlivom softvéri. Archív obsahuje viac ako 17 000 programov patriacich do 585 rodín malvéru (trójske kone, vírusy, červy).

https://github.com/opsxcq/mirror-vxheaven.org

Doplnkové datasety a zdroje

IEEE Test Feeders

Už takmer dve desaťročia podvýbor pre analýzu distribučných systémov (DSA) pracovná skupina pre testovacie zásobníky (test feeders) (TFWG) konštruuje verejne dostupné distribučné testovacie zásobníky pre použitie akademickými pracovníkmi.

Cieľom týchto testovacích zásobníkov je vytvoriť modely distribučných systémov, ktoré odrážajú širokú škálu možností návrhu a analytických problémov. 13-zbernicové a 123-zbernicové zásobníky sú súčasťou systémov Test Feeder vytvorených v roku 1992 na vyhodnotenie a porovnávanie algoritmov pri riešení nevyvážených trojfázových radiálnych systémov. Podvýbor DSA ich schválil počas letného stretnutia Power and Energy Society (PES) v roku 2000. Schneider a kol. zhŕňajú úsilie TFWG a zamýšľané použitie testovacích zásobníkov.

https://cmte.ieee.org/pes-testfeeders/resources/

XSSed

Projekt XSSed bol vytvorený vo februári 2007. Je to archív zraniteľných webových stránok s cross-site scriptingom (XSS) a poskytuje informácie o veciach súvisiacich so zraniteľnosťami XSS.

http://www.xssed.com/archive

NeCTAR (National eResearch Collaboration Tools and Resources)

Cloudová platforma, ktorú v roku 2012 spustila austrálska Research Data Commons, poskytuje austrálskej výskumnej komunite rýchly, interaktívny a samoobslužný prístup k rozsiahlej výpočtovej infraštruktúre, softvéru a údajom.

https://ardc.edu.au/services/ardc-nectar-research-cloud/

The Mobile-Sandbox

Navrhnutý Univerzitou v Erlangene-Norimberi v Nemecku v roku 2014. Jedná sa o statický a dynamický systém pre analýzu určený ako podpora pre analytikov pri odhaľovaní škodlivého správania malvéru.

https://github.com/floe/mobile-sandbox

Credit Card Fraud

Dataset bol zozbieraný a analyzovaný počas výskumnej spolupráce spoločností Worldline a Machine Learning Group (http://mlg.ulb.ac.be) ULB (Université Libre de Bruxelles) v oblasti získavania veľkých dát a odhaľovania podvodov. Dataset obsahuje transakcie uskutočnené európskymi držiteľmi kariet kreditnými kartami v septembri 2013. Tento súbor údajov predstavuje transakcie, ktoré sa uskutočnili za dva dni, pričom je k dispozícii 492 podvodov z celkového počtu 284 807 transakcií. Dataset je veľmi

nevyvážený, trieda s pozitívnou anotáciou (označujúca prítomnosť podvodu) predstavuje 0,172 % všetkých transakcií. Obsahuje iba číselné vstupné premenné, ktoré sú výsledkom transformácie PCA. Bohužiaľ, pôvodné príznaky a ďalšie základné informácie o údajoch nie sú poskytované z dôvodu problémov s dôvernosťou údajov. Jediné parametre, ktoré sa netransformujú pomocou PCA, sú "čas" a "suma". Príznak "čas" obsahuje sekundy, ktoré uplynuli medzi každou transakciou a prvou transakciou v datasete. Príznak "suma" je suma vykonanej transakcie.

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Twitter ISIS Dataset

Dataset ISIS na Twitteri, publikovaný v roku 2018, pozostáva z tweetov/retweetov súvisiacich s ISIS v arabčine zhromaždených od februára 2016 do mája 2016. Dataset obsahuje tweety a súvisiace informácie, ako je ID používateľa, ID opätovného tweetu, hashtagy, počet sledovateľov, počet sledovateľov, obsah, dátum a čas. Približne 53 miliónov tweetov sa zozbieralo na základe 290 hashtagov ako napr. Štát islamského kalifátu a Islamský štát.

https://www.kaggle.com/datasets/fifthtribe/how-isis-uses-twitter

Italian Retweets Timeseries

Dataset Italian Retweets Timeseries, zverejnený v roku 2019, obsahuje časové údaje o približne 5 121 132 retweetoch od 47 947 používateľov prevzatých z talianskej Twittersphere zverejnených medzi 18. 6. 2018 a 1. 7. 2018.

https://zenodo.org/records/2653138

5GAD-2022 5G attack detection dataset

Dataset s údajmi o sieťovej prevádzke 5G pre použitie s nástrojmi strojového učenia na porovnávanie možností detekcie útokov pre viacero rôznych modelov. Dataset obsahuje simulovanú prevádzku 5G siete s normálnymi podmienkami a s rôznymi druhmi sieťových útokov, celkovo 10 útokov.

https://github.com/IdahoLabResearch/5GAD

Normálna prevádzka

Normálne-1UE

Tento súbor údajov pozostáva z bežnej internetovej prevádzky zaznamenanej na sieťových rozhraniach. Konkrétne obsahuje sieťovú prevádzku simulovanú na jednom UE

s rôznymi automatizovanými úlohami vrátane streamovania videí z YouTube, prístupu k 500 populárnym webovým stránkam, sťahovania súborov cez FTP, pripojenia zdieľania SAMBA a sťahovania súborov z neho a konferenčného hovoru cez Microsoft Teams.

Normálne-2UE

Pre Normal-2UE boli použité rovnaké typy prevádzky ako v Normal-1UE, až na to, že tentoraz bolo generovanie prevádzky rozdelené medzi 2 UE.

Prevádzka s útokmi

Táto časť datasetu obsahuje prevádzku, v ktorej sú prítomné rôzne typy útokov (celkovo 10), z ktorých väčšina sa spolieha na volania REST API do rôznych častí jadra.

Typy útokov:

Prieskumné útoky (Reconnaissance Attacks):

- *AMFLookingForUDM* - Tento útok sa vykonáva vyžiadaním informácií o sieťovej funkcii zjednotenej správy údajov (UDM) a zároveň zosobnením funkcie správy prístupu a mobility (AMF). Interne sa tento útok javí ako neškodná systémová požiadavka a využíva skutočnosť, že funkcia sieťového úložiska (NRF) nekontroluje, či je zdrojom požiadavky skutočne AMF.
- *GetAllNFs* - Tento útok sa vykonáva rovnako ako AMFLookingForUDM s tým rozdielom, že nie je špecifikovaný typ target-nf. To má za následok, že NRF vráti žiadateľovi všetky sieťové funkcie (NF).
- *GetUSerData* - Tento útok požaduje od UDM informácie týkajúce sa používateľa so subscriberID=0000000003.
- *randomDataDump* - Tento útok využíva nedostatok overenia vstupu vo free5GC a nastavuje requester-nf-type na náhodný reťazec pri zadávaní požiadavky nf-instances na NRF. NRF bude stále odpovedať všetkými NF.
- *automatedRedirectWithTimer* - Tento útok pochádza zo správy spoločnosti Positive Technologies 5G Samostatný základný bezpečnostný výskum, časť 4.3. V podstate počúva sieťovú prevádzku, zatiaľ čo sa UE pripája. Kód útoku načúva požiadavke na vytvorenie relácie protokolu PFCP (Packet Forwarding Control Protocol) a potom skontroluje, či je adresa UE v zozname adries obetí. Ak je UE obeťou, útok zaznamenáva informácie o relácii, ktoré používa na presmerovanie prevádzky z UE. To sa dosiahne odoslaním žiadosti o úpravu relácie PFCP do funkcie používateľskej roviny (UPF) s zisteným ID relácie a ID pravidla akcie presmerovania (FARID). Útok odošle dve takéto žiadosti o úpravu, počká 5 sekúnd, odošle ďalšie dve žiadosti o úpravu, aby sa UE vrátil na normálnu cestu, počká ďalších 5 sekúnd a potom zopakuje.

Re-konfigurácia siete:

- *FakeAMFInsert* - Tento útok zaregistruje falošný AMF v NRF. To sa dosiahne spustením príkazu curl na VLOŽENIE objektu JSON do NRF. V prostredí, kde bol

tento útok spustený, neexistuje žiadna kontrola oprávnenia, ktorá by zabránila útočníkovi zaregistrovať falošný AMF. Následne sa spustí útok FakeAMFDelete na odstránenie falošného AMF. ID inštancie musí byť univerzálne jedinečný identifikátor (UUID) verzie 4, avšak free5GC nekontroluje ID inštancie ani iné podrobnosti o AMF pred jeho pridaním do jadra. To zahŕňa aj to, či je inštancia-id správne naformátované UUID, ktoré pozostáva z hexadecimálnych hodnôt v reťazci. Niekoľko znakov "1" v reťazci UUID útoku bolo nahradených znakom "I" pri písaní útočného kódu, takže reťazec nebol správne naformátovaný ako hexadecimálny. ID inštancie bolo preto neplatné UUID, ale bez ohľadu na to bolo akceptované free5GC ID inštancie.

- *randomAMFInsert* - Tento útok je rovnaký ako "FakeAMFInsert" s tým rozdielom, že ID inštancie je náhodne vygenerovaný UUID.

DOS útoky:

- *CrashNRF* - Tento útok sa spolieha na zneužitie vo free5GC, kde chybná požiadavka na funkciu sieťového úložiska (NRF) spôsobí jeho zlyhanie. Od free5GC v3.1.1 sa zdá, že toto zneužitie bolo opravené, pretože táto požiadavka HTTP GET už nebude mať za následok zlyhanie jadra.
- *FakeAMFDelete* - Tento útok je prevádzkovaný v spojení s FakeAMFInsert, kde fakeAMF je hexadecimálne ID relácie falošného AMF vloženého do systému. Tento útok spolu s GetAllNF na nájdenie ďalších AMF by mohol byť zneužitý na odstránenie legitímnych AMF zo siete, čím by sa narušila funkčnosť siete.
- *automatedDropWithTimer* - Tento útok je podobný útoku "automatedRedirectWithTimer", ale strieda presmerovanie používateľskej prevádzky a prerušenie používateľskej prevádzky, čím efektívne odpojí používateľa od dátovej siete (DN).

The Numenta Anomaly Benchmark (NAB)

Korpus NAB s 58 dátovými súbormi časových radov je navrhnutý tak, aby poskytoval údaje pre výskum detekcie anomálií v streamovanom dátovom toku. Pozostáva z údajov získaných z reálnej prevádzky ako aj z umelých časových radov a obsahuje vyznačené úseky s anomálnym správaním. Väčšina dát pochádza z reálneho sveta z rôznych zdrojov, ako sú metriky servera AWS, Twitter, metriky klikania na reklamy, údaje o návštevnosti a ďalšie.

https://github.com/numenta/