UNIVERZITA PAVLA JOZEFA ŠAFÁRIKA V KOŠICIACH
PRÍRODOVEDECKÁ FAKULTA

OBOHACOVANIE INDIKÁTOROV KOMPROMITÁCIE V PODVODNÝCH E-MAILOVÝCH SPRÁVACH

2023 **Monika RAPAVÁ**

UNIVERZITA PAVLA JOZEFA ŠAFÁRIKA V KOŠICIACH
PRÍRODOVEDECKÁ FAKULTA

OBOHACOVANIE INDIKÁTOROV KOMPROMITÁCIE V PODVODNÝCH E-MAILOVÝCH SPRÁVACH

BAKALÁRSKA PRÁCA

Študijný program:	Informatika
Pracovisko (katedra/ústav):	Ústav informatiky
Vedúci bakalárskej práce:	Mgr. Eva Marková

Košice 2023 **Monika RAPAVÁ**

Univerzita P. J. Šafárika v Košiciach
Prírodovedecká fakulta

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Monika Rapavá
Študijný program: informatika (jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 18. - informatika
Typ záverečnej práce: Bakalárska práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Obohacovanie indikátorov kompromitácie v podvodných e-mailových správach

Cieľ:
1. Analyzovať indikátory kompromitácie podvodných e-mailových správ
2. Porovnať prístupy na obohatenie indikátorov kompromitácie v podvodných e-mailových správach prostredníctvom threat intelligence
3. Navrhnúť, implementovať a vyhodnotiť nástroj na analýzu podvodných e-mailových správ pomocou threat intelligence a obohacovania indikátorov kompromitácie

Literatúra:
1. LEGG, Phil; BLACKMAN, Tim. Tools and techniques for improving cyber situational awareness of targeted phishing attacks. In: 2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA). IEEE, 2019. p. 1-4.
2. CONTI, Mauro; DARGAHI, Tooska; DEHGHANTANHA, Ali. Cyber threat intelligence: challenges and opportunities. In: Cyber Threat Intelligence. Springer, Cham, 2018. p. 1-6.
3. COHEN, Aviad; NISSIM, Nir; ELOVICI, Yuval. Novel set of general descriptive features for enhanced detection of malicious emails using machine learning methods. Expert Systems with Applications, 2018, 110: 143-169.
4. TOUNSI, Wiem; RAIS, Helmi. A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers & security, 2018, 72: 212-233.

Vedúci: Mgr. Eva Marková
Oponent: RNDr. Zoltán Szoplák
Ústav: ÚINF - Ústav informatiky
Riaditeľ ústavu: doc. RNDr. Ondrej Krídlo, PhD.

Dátum zadania: 11.05.2021

Dátum schválenia: 15.05.2023 doc. RNDr. Ondrej Krídlo, PhD.
riaditeľ ústavu

Poďakovanie

Týmto sa chcem poďakovať vedúcej svojej práce Mgr. Eve Markovej za odborné vedenie, cenné rady a veľkú pomoc pri tvorbe tejto bakalárskej práce.

Abstrakt v štátnom jazyku

Indikátory kompromitácie (IOCs) sú časti údajov, ktoré identifikujú potencionálne škodlivú aktivitu v sieti alebo v systéme. Threat intelligence (TI) tvorí množina takýchto nazbieraných dát, ktoré sú posúdené a použité v súvislosti s bezpečnostnými hrozbami a zraniteľnosťami. Preto zohrávajú dôležitú úlohu v oblasti kybernetickej bezpečnosti. V tejto práci analyzujeme indikátory kompromitácie v škodlivých e-mailoch a následne pomocou TI obohacujeme tieto nazbierané dáta. Výstupom tejto práce je súhrn štatistík z extrahovaných atribútov pred a po obohacovaní, a taktiež hľadanie vzťahov medzi indikátormi kompromitácie z rôznych e-mailov pochádzajúcich z rôznych dostupných zdrojov. Týmto spôsobom je možné zjednodušiť prvotnú analýzu škodlivých e-mailov, čo slúži ako pomoc pri včasnom reagovaní na útoky rôzneho druhu.

Kľúčové slová: indikátory kompromitácie, sociálne inžinierstvo, threat intelligence

Abstrakt v cudzom jazyku

Indicators of compromise (IOCs) are segments of data which identify potentially malicious activity on a network or system. Threat intelligence (TI) is a collection of such gathered data that is assessed and used in the context of security threats and vulnerabilities. Therefore, they have a significant role to play in cybersecurity. In this paper, we analyze the indicators of compromise in malicious emails and enrich these collected data using TI. The output of this work is a summary of statistics from the extracted attributes before and after enrichment and also finding relationships between the indicators of compromise from different emails from a variety of available sources. This can simplify the initial analysis of malicious emails, which is helpful for early response to attacks of various kinds.

Keywords: indicators of compromise, social engineering, threat intelligence

Obsah

Zoznam ilustrácií

Zoznam tabuliek

Zoznam skratiek a značiek

TI	Threat intelligence
DNS	Domain Name System, systém názvov domén
RFC	Request for Comments, žiadosť o komentáre
SIEM	Security Information and Event Management, manažment bezpečnostných informácií a udalostí
API	Application Programming Interface, rozhranie pre programovanie aplikácií
IoC	Indicator of Compromise, indikátor kompromitácie
SPF	Sender Policy Framework
DKIM	DomainKeys Identified Mail
DMARC	Domain-based Message Authentification, Reporting and Conformance

Úvod

Vo svete neustále rozvíjajúcich sa kybernetických útokov sa pravidelne stretávame s veľkým množstvom rôznorodých podvodných e-mailov. Keďže práve e-mailová komunikácia je jednou z pravidelne využívaných, útočníci si túto cestu uvedomujú s cieľom jednoducho doručiť tieto podvodné e-maily do e-mailových schránok bežných používateľov. Pokiaľ sa jedná o cielené kybernetické útoky, tie sa často zameriavajú na organizácie, ktoré zdieľajú nejaký spoločný znak. Preto je podľa niektorých vhodné zdieľať informácie v reálnom čase na zamedzenie opakovaných útokov, alebo zmiernenie dopadov pri poškodení. To nás vedie k pojmu threat intelligence (TI), ktorý by bolo možné definovať ako funkčná obrana na zníženie rozdielu medzi pokročilejšími útokmi a prostriedkami na obranu organizácie. Zahŕňa to proces analýzy, založený na identifikácii, zbere a obohatení o relevantné informácie. Nie vždy je totiž potrebné uchovávať veľké množstvo dát pri tvorení modelu, keďže zostanú nevyužité [1].

Hlavným cieľom tejto práce je zjednodušenie prvotnej analýzy podvodných e-mailových správ, vďaka čomu dokážeme znížiť čas pri reakcii na bezpečnostný incident od prvého prijatia e-mailu. Výstupom je súhrn štatistík z extrahovaných atribútov s využitím metódy obohacovania pomocou TI.

Prvým cieľom práce je identifikovať relevantné indikátory kompromitácie podvodných e-mailových správ a ich následná analýza. V druhom cieli je potrebné porovnať prístupy na obohacovanie indikátorov kompromitácie prostredníctvom threat intelligence. Ako posledným cieľom je návrh a implementácia nástroja na analýzu podvodných e-mailových správ s vyhodnotením štatistík z extrahovaných atribútov pred a po obohacovaní pomocou threat intelligence.

Práca je rozdelená do štyroch základných kapitol. Prvá kapitola je venovaná základnej terminológii e-mailov. Jej súčasťou je taktiež rozbor samotného e-mailu, kde sa podrobnejšie venujeme analýze hlavičkových atribútov a pozrieme sa taktiež na podobné práce súvisiace s touto problematikou. V druhej kapitole sa zameriavame na základnú definíciu threat intelligence a kategóriám jej zdrojov. V tejto časti taktiež porovnávame existujúce threat intelligence služby podľa rôznych kategórií a popisujeme ich odlišnosti. V tretej kapitole sa nachádza samotný návrh celkového priebehu procesu, pričom každej časti sa venujeme podrobnejšie v samostatných podkapitolách. Štvrtá, a zároveň posledná kapitola predstavuje celkový súhrn výsledkov

a štatistík pred a po obohacovaní pomocou TI, pričom vybrané služby sú rozdelené podľa obohacovaných atribútov. Pomocou vyhodnotených štatistík v poslednej časti tejto kapitoly uvádzame príklady odporúčaní, ktoré možno vykonať pri podvodných e-mailových správach.

1 Základná terminológia e-mailov

V rámci e-mailovej komunikácie medzi odosielateľom a príjemcom je prítomných viacero strán.

S-MUA, R-MUA (S – Sender's/R – Receiver's Mail User Agent) – Zložka e-mailového systému, ktorá umožňuje používateľom prijímanie a odosielanie e-mailov. Môže sa jednať či už o desktopovú, alebo webovú aplikáciu. Existujú desiatky rôznych MUA, pričom používateľ si môže vybrať z množstva rôznych funkcionalít podľa svojich preferencií [2].

S-MSA, H-MSA (S – Sender's/H - MHS's Message Submission Agent) –Pripravuje e-maily na odosielanie do MHS, pričom využíva protokoly SMTP a TCP. S-MSA pridáva polia hlavičky ako sú atribúty „Message-ID" a „Date", pričom mení časti e-mailu podľa miestneho internetového prostredia. H-MSA zodpovedá za to, aby správa vyhovovala tomuto prostrediu [3].

MTA (Message Transfer Agent) – Funguje podobne ako router, aby sa e-mail preniesol k príjemcovi. Preposielanie sa uskutočňuje pomocou série MTA, kým sa e-mail nedostane k cieľovému MDA. Keďže v porovnaní s paketovým prenosom majú správy väčšiu veľkosť a čas prenosu je taktiež vyšší, MTA vykonáva funkciu MTA odosielateľa aj príjemcu. Tento agent taktiež pridáva informáciu o sledovaní (v prípade zlyhania pri prenose je možné prenos opakovať) [3]. Pred odoslaním sú e-mailové správy ukladané do úložiska, teda pri prípadnom oneskorení na strane odosielateľa ho MTA pravidelne kontroluje (zvyčajne 10 až 30 minút), či je možné poštu odoslať. Tá sa odošle len vtedy, ak je príjemca na doručenie pripravený a IP adresa servera bola získaná prostredníctvom DNS. Ak sa správu nepodarí doručiť v časovom období približne 3 až 5 dní, pošta sa vráti odosielateľovi [4].

H-MDA, R-MDA (MHS's/R - Receiver's Message Delivery Agent) – Proces doručenia e-mailu z H-MDA na R-MDA. Cieľom H-MDA je presmerovanie e-mailu na uvedenú adresu, ak je tak uvedené v preferenciách adresáta. Prenos z tohto agenta do úložiska e-mailových správ sa uskutočňuje pomocou protokolu POP, alebo IMAP [3].

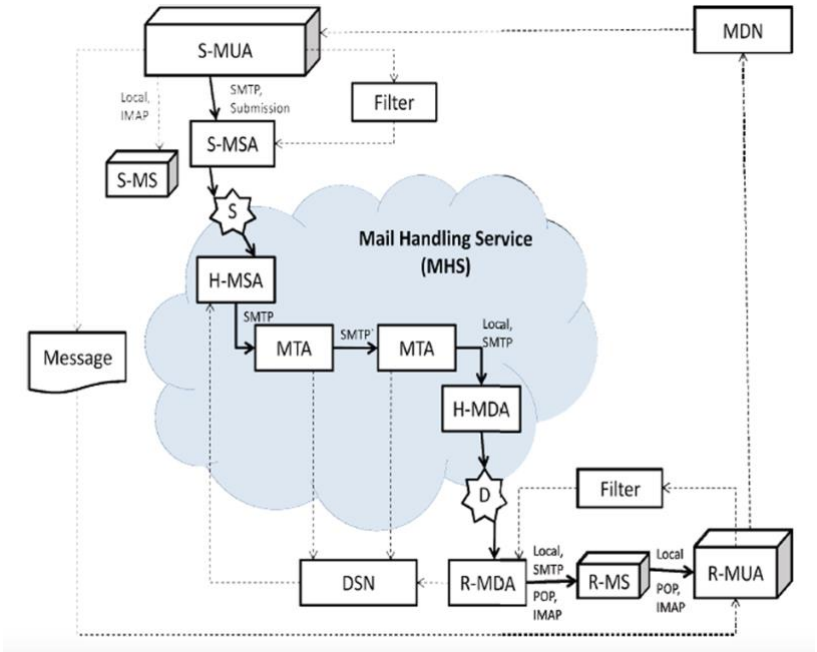
MHS (Mail Handling Service) – Servis, ktorý uskutočňuje samotný prenos e-mailových správ medzi odosielateľom a príjemcom. MHS prijíma e-mail od jedného odosielateľa a doručuje ho ostatným. Skladá za z agentov MTA [3].

MS (S – Sender's/R – Receiver's Message store) – Úložisko e-mailových správ, či už na serveri, alebo na lokálnom zariadení. Toto úložisko využíva MUA [3].

SMTP (Simple Mail Transfer Protocol) - Protokol, ktorý sa využíva na prenos elektronickej pošty medzi jednotlivými uzlami. SMTP transakcia pozostáva z obálky a správy, ktorá sa následne skladá z hlavičky a tela, pričom obálka sa posiela oddelene od samostatnej správy pomocou príkazov „MAIL FROM" a „RCPT TO" [4].

IMAP/POP – Protokoly, ktoré využíva MUA na získavanie e-mailov z e-mailovej schránky servera. Preferovanejší je IMAP protokol, ktorý udržiava všetky e-maily na serveri, pričom POP ich maže zo servera po tom, čo boli stiahnuté [2].

Na obrázku č.1 môžeme vidieť celkový proces prenosu e-mailovej správy z popísaných komponentov. Môžeme si všimnúť, že práve MUA je úvodnou aj záverečnou zložkou od odoslania k prijatiu e-mailu.

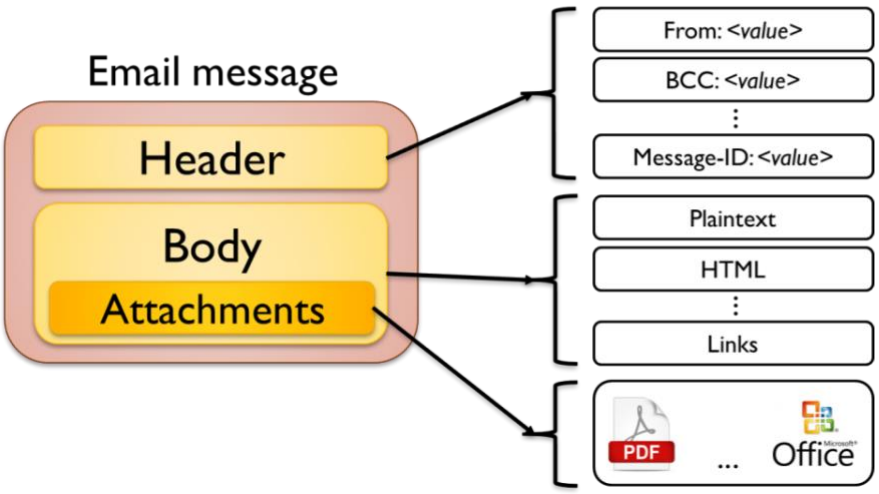


Obr. 1 E-mailová komunikácia [3]

1.1 Štruktúra e-mailovej správy

E-mailová správa sa skladá z dvoch hlavných častí: hlavička a telo správy. Túto štruktúru e-mailu môžeme vidieť zakreslenú na obrázku č. 2. Pokiaľ by sme sa pozreli

na e-mailovú správu v textovom formáte, hlavička je uvedená na začiatku e-mailu a telo nasleduje bezprostredne po nej, pričom ich oddeľuje prázdny riadok. Telo e-mailu obsahuje text, ktorý je bežne viditeľný pre používateľa, pričom prítomné môžu byť taktiež prílohy zakódované pomocou štandardných podporovaných kódovaní. Po formálnej stránke telo môže obsahovať jednoduchý text (plaintext), HTML (Hypertext Markup Language), alebo kombináciu oboch (multipart email). Pokiaľ je e-mail formátovaný ako HTML, jeho súčasťou môžu byť vložené súbory či obrázky. Prítomné ochrany na e-mailových serveroch však môžu obmedziť veľkosť e-mailovej správy (vrátane príloh), či blokovať vybrané prípony [5]. Pokiaľ hovoríme o samotnom obsahu tela podvodných e-mailov, často sa v ňom nachádzajú kľúčové slová či slovné termíny, pomocou ktorých dokážeme rozlíšiť podvodné e-maily od tých legitímnych. Cieľom útočníkov je totiž využiť psychologickú slabosť či nepozornosť ľudí na zvýšenie pravdepodobnosti vykonania požadovanej aktivity. Príkladom môžu byť nezrovnalosti v platobných údajoch, deaktivácia účtu, či doručenie fiktívnej zásielky [6]. Túto aktivitu môže používateľ vykonať prostredníctvom vloženého URL odkazu, na ktorý klikne a ten ho presmeruje na podvrhnutú webovú stránku. Po zadaní údajov ako sú používateľské mená, heslá, či čísla kreditných kariet sa následne odošlú útočníkovi, ktorý k nim získa prístup. Ďalšou možnosťou po kliknutí na URL odkaz je stiahnutie si škodlivého kódu (malvéru) do zariadenia, pomocou ktorého získajú útočníci priamo informácie o online účte obete [7]. Bližšej analýze hlavičky sa venujeme v nasledujúcej podkapitole.



Obr. 2 Štruktúra e-mailovej správy [5]

1.2 Hlavičkové atribúty

Každý e-mail má práve jednu hlavičku, v ktorej sú uložené dodatočné informácie o správe. Vo všeobecnosti platí, že každý atribút je zapísaný v inom riadku a pozostáva z názvu a priradenej hodnoty. Počet týchto atribútov môže byť rôzny v závislosti od servera, ktorý e-mail odoslal. Vždy prítomné sú však povinné atribúty, ktoré musí obsahovať každý e-mail. Sú nimi napríklad atribúty „Date“ či „From“. Hodnoty takmer akýchkoľvek hlavičkových atribútov je však možné sfalšovať, no atribút „Received“ je generovaný e-mailovým severom príjemcu, a teda sú hodnoty v tomto atribúte považované za dôveryhodné [5].

Hlavička môže obsahovať viac ako jedno pole „Received”, ktoré je potrebné čítať zdola nahor. Spodná časť tohto atribútu predstavuje prvý e-mailový server, ktorý sa podieľal na prenose a posledný sa nachádza na úplnom vrchu [2]. Postupne sa teda jednotlivé polia pridávajú na vrch pomyselného zásobníka. Informácie, ktoré sú obsiahnuté v hlavičke môžu pomôcť napr. pri zistení odosielateľa e-mailu (útok na identitu odosielateľa e-mailu), či ďalších základných informácií, ktoré sú pri prvotnom prehliadaní e-mailu „neviditeľné“. Samotná analýza by mala zahŕňať najmä preskúmanie IP adresy odosielateľa, formát správy, SPF (Sender Policy Framework), DKIM (DomainKeys Identified Mail) a DMARC (Domain-based Message Authentification, Reporting and Conformance) obsiahnuté v atribúte „Authentication-Results“ ako aj protokol, ktorý bol použitý na iniciáciu (HTTP alebo SMTP) [8].

Keďže súčasťou tejto práce je extrakcia indikátorov kompromitácie, je nevyhnutné si vysvetliť jednotlivé hlavičkové atribúty pre správnu interpretáciu dát a ďalšiu prácu s nimi.

Názov atribútu	**Popis**	**Zdroj**	**Kategória (podľa RFC 5322) [9]**	**Typ**
From	adresa odosielateľa (autora), ktorý je zodpovedný za napísanie správy	hlavička		String

Reply-To	indikuje adresu, na ktorú majú byť odosielané prípadné odpovede	hlavička	Originator Fields	String
Sender	určuje poštovú schránku agenta zodpovedného za skutočný prenos správy	hlavička		String
Subject	predmet správy	hlavička	Informational Fields	String
Date	dátum a čas odoslania správy	hlavička	The Origination Date Field	RFC 822, 1036, 1123, 2822
Content-type	formát správy	hlavička, telo		String
Message-Id	unikátny identifikátor správy	hlavička	Identification Fields	String
Authentication-Results **A) SPF**	TXT DNS záznam, v ktorom je uvedené, ktorý poštový server je oprávnený odosielať emaily pod určitou doménou	hlavička		Categorical
B) DKIM	nástroj na podpisovanie hlavičky odchádzajúcich emailov privátnym kľúčom	hlavička		Categorical
C) DMARC	overenie odosielateľa správy pomocou DKIM a SPF, a zároveň ako sa má príjemca vysporiadať s neúspešným overením	hlavička		Categorical
D) IP Address		hlavička		String
Return-Path	emailová adresa, kam má	hlavička	Trace Fields	String

	byť zaslané hlásenie v prípade chybových hlásení pri posielaní e-mailu			
X-Forefront-Antispam-Report **A) CIP**	obsahuje informácie o správe a o spôsobe jej spracovania	hlavička	Optional Fields	String
B) LANG		hlavička		Categorical
C) SFV		hlavička		Categorical
D) SCL		hlavička		Numerical
E) CAT		hlavička		Categorical
X-Microsoft-Antispam - BCL	obsahuje ďalšie informácie o hromadnej pošte a phishingu	hlavička	Optional Fields	Numerical

Tab. 1 Polia e-mailovej hlavičky

Tabuľka č. 1 predstavuje zoznam niektorých hlavičkových atribútov. Každý atribút obsahuje stručný popis ako aj zdroj, kde je možné ho v samotnej štruktúre e-mailovej správy nájsť. Ďalšou časťou tabuľky je rozdelenie podľa kategórií určených v RFC 5322 [9]. Posledný stĺpec obsahuje informáciu o type samotného atribútu, napr. či sa jedná o numerickú, kategorickú hodnotu atď. V nasledujúcej časti si bližšie rozoberieme niektoré hlavičkové atribúty podľa kategórií z tabuľky č.1.

- The Origination Date Field
    - **Date -** Formát tohto hlavičkového atribútu ukážeme na nasledujúcom príklade: Mon, 01 Jan 2022 12:01:00 +0100. Prvá časť obsahuje informáciu o dni v týždni, nasleduje deň v roku, mesiac, rok a čas. Posledná časť označuje časový posun oproti koordinovanému svetovému času UTC (Coordinated Universal Time). Znamienka + a – interpretujú rozdiely v čase na východnej (+) a západnej (-) pologuli oproti UTC.

Prvé dve číslice po znamienku reprezentujú rozdiel v počte hodín a ďalšie dve v počte dodatočných minút.

- Originator Fields
    - **Reply-To –** Jedná sa o nepovinný parameter, ktorý indikuje adresu(y), na ktorú sa autor správy odkazuje s prípadnými odpoveďami na daný e-mail. Ak tak nie je uvedené, odpoveď je odoslaná na e-mailovú adresu z časti „From" [9].
- Identification Fields
    - **Message-ID -** Unikátny identifikátor správy, ktorý pri prenose e-mailovej správy generuje prvý Mail Submission Agent (MSA) na ceste od odosielateľa k príjemcovi. Tento identifikátor sa týka práve jednej verzie e-mailovej správy, pričom pri akejkoľvek zmene e-mailu sa automaticky generuje nový identifikátor. Príklad formátu Message-ID: <20221025173552.7EB12C6423AASHC8@google.com> [9].

Oba nasledujúce atribúty vyššej úrovne sa pridávajú do hlavičky nadštandardne (označuje sa prvým znakom X), keďže sa nejedná o povinné polia a na ich poradí nezáleží. Jedná sa o tzv. „antispamové" atribúty, ktoré obsahujú informácie o samotnom skenovaní e-mailu na spam, malware a ostatné.

- Optional Fields
    - **X-Forefront-Antispam-Report** [10] – Obsahuje informácie o e-mailovej správe a o tom, ako bola spracovaná. Samotné informácie sú v tvare názov poľa a hodnota, ktoré sú oddelené bodkočiarkou. Niektoré z vybraných atribútov: CAT, CIP, CTRY, SCL, SFV atď.
        - CAT (Category) popisuje kategóriu pomocou filtrovacej služby dostupnej v službe Office365 s názvom Exchange Online Protection (EOP). Príklady jednotlivých kategórií: SPM (Spam), MALW (Malware), PHSH (Phishing), BULK, HPHSH alebo HPHISH (High confidence phishing) atď.
        - CTRY (Country) označuje krajinu určenú pripájacou IP adresou, ktorá však nemusí byť rovnaká ako krajina získaná z IP adresy odosielateľa.

        - LANG (Language) predstavuje jazyk, v ktorom je e-mailová správa písaná. Nadobudnuté hodnoty sú zapísané ako kód krajiny, ktorý reprezentuje daný jazyk napr. Slovakia ako SK.
        - SCL (Spam Confidence Level) skóre, ktoré hovorí o tom, nakoľko sa jedná o vyhodnotenie ako spam z filtrovania pomocou EOP. Výsledkom filtrovania je jedna číselná hodnota. Vo všeobecnosti platí, čím vyššie číslo „spam skóre“, tým je vyššia pravdepodobnosť, že sa jedná o spam. Všetky e-mailové správy obsahujúce SCL:0, alebo SCL:-1 hovoria o tom, že sa o spam podľa filtrovania nejedná. Hodnoty SCL:8, alebo SCL:9 označujú e-maily ako „High confidence spam“, teda sa s vysokou pravdepodobnosťou o spam jedná.
    - **X-Microsoft-Antispam** [10] – Obsahuje dodatočné informácie o spame a phishingu. Nadobúda práve jeden atribút so skratkou BCL.
        - BCL (Bulk Confidence Level) skóre generujúce sa na základe označenia od príjemcu, či sa jedná o hromadne rozposielaný e-mail. Pri hromadne rozposielaných e-mailoch sa odosielatelia líšia v spôsoboch odoslania e-mailu, napr. ako je tvorený samotný obsah, alebo ako získavajú zoznam príjemcov. Ak sa jedná o „dobrých“ hromadných odosielateľov, tí posielajú relevantné e-maily, ktoré generujú malé množstvo sťažností od príjemcov. Sú tu však takí, ktorí posielajú nesúvisiace e-maily, kedy sa generuje viac sťažností. Podľa tohto postupu sa na záver vyhodnotia všetky sťažnosti, a každému e-mailu sa pridelí jedna číselná hodnota.

Posledné z vybraných atribútov na bližšiu analýzu sa nachádzajú v hlavičkovom atribúte s názvom „Authentication-Results“. Jedná sa taktiež o „antispamový“ atribút, ktorý obsahuje informácie o výsledkoch SPF, DKIM a DMARC.

- **SPF (Sender Policy Framework)** [11] – DNS záznam, v ktorom je uvedené, ktorý poštový server je oprávnený odosielať e-maily pod určitou doménou. Pokiaľ je SPF záznam zverejnený v DNS, tak obsahuje zoznam IP adries, ktoré

môžu odosielať e-maily pod danou doménou. Pri odosielaní e-mailovej správy je úlohou prijímajúceho servera, aby porovnal IP adresu odosielateľa s autorizovanými IP adresami v SPF zázname. Existuje však slabé miesto pri tomto overovaní, čím je preposielanie e-mailov.

Syntax SPF: spf=<pass (IP address)|fail (IP address)|softfail (reason)|neutral|none|temperror|permerror> smtp.mailfrom=<domain>.

- **DKIM (DomainKeys Identified Mail)** [11] – predstavuje pomyselný vodoznak pre e-mailové správy. Jedná sa o akúsi formu nástroja, ktorý umožňuje odosielateľovi elektronicky podpísať hlavičku odosielaného e-mailu privátnym kľúčom. Na základe tejto dodatočnej informácie dokáže príjemca overiť, že jemu doručený e-mail pochádza z domény, o ktorej tvrdí, že pochádza. Príjemca si taktiež dokáže potvrdiť, že s daným e-mailom nebolo manipulované na ceste medzi odosielateľom a príjemcom. Úlohou prijímajúceho servera je overiť správnosť tohto podpisu. Overenie podpisu sa robí pomocou verejného kľúča uvedeného v DNS zázname pre danú doménu.
- **DMARC (Domain-based Message Authentication, Reporting and Conformance)** [12] – je kombináciou vyššie spomínaných SPF a DKIM. Ak má dôjsť k úspešnej DMARC hodnote, je potrebné, aby došlo k doménovej zhode v atribútoch „From", „Return-Path" a doménach uvedených v SPF a DKIM. S týmito informáciami dokáže server rozhodnúť o tom, či e-mailovú správu prijať, odmietnuť, alebo inak označiť podľa prednastavených pravidiel.

1.2.1 Podobné práce

V tejto podkapitole sa budeme venovať porovnaniu podobných prác, ktoré súvisia s extrahovaním hlavičkových atribútov. Táto časť slúži ako pomoc pri výbere našej výstupnej množiny a atribútov, ktoré sa následne rozhodneme analyzovať.

V práci EmailProfiler [13] autori vyvinuli automatizovaný prístup, ktorý slúži ako ochrana používateľov pred spear-phishingovými útokmi. Jedná sa o vyššiu formu cielenia podvodných e-mailových správ, pričom sú jednotliví príjemcovia vyberaní podľa určitých kritérií. Pri tejto práci boli vybrané nasledujúce hlavičkové atribúty - „Return-Path", „x-mailman-version", „x-originating-hostname", „x-originating-ip", „x-spam-flag", „x-virus-scanned", „carbon copy" v kombinácii s informáciami uvedenými v texte e-mailu.

Ďalšia práca [14] porovnáva existujúce techniky klasifikácie spamu pomocou hlavičkových atribútov. Cieľom bolo zistiť, či je možné len pomocou hlavičkových atribútov klasifikovať spamové e-maily pomocou množiny atribútov „CC“ (carbon copy), „Date“, „Delivered To“, „DKIM Signature“, „From“, „Message-ID“, „Reply-To“, „Return-Path“, „Received“, „Subject“, „To“. Pre ohodnotenie daného e-mailu ako spam boli vytvorené špecifické pravidlá pre jednotlivé hodnoty atribútov.

Autori tretej vybranej práce [15] predstavili návrh modelu na detekciu phishingových e-mailov, pričom najperspektívnejšie riešenie pri výbere atribútov predstavovalo kombináciu hlavičkových atribútov s atribútmi obsiahnutými v tele správy pomocou techník dolovania údajov. Okrem rôznych získaných údajov napr. o URL odkazoch z tela e-mailu boli vybrané aj nasledujúce atribúty z hlavičky – „From“, „Message-ID“ a „Content-Type”.

Pasupatheeswaran [16] skúmal podrobne len jeden atribút s názvom „Message-ID“. Jeho cieľom bolo z tohto atribútu získať užitočné informácie, ktoré by dokázali napomôcť pri forenznej analýze. Zameral sa taktiež na pochopenie fungovania agenta MTA (Mail Transfer Agent) verzie 8.14.

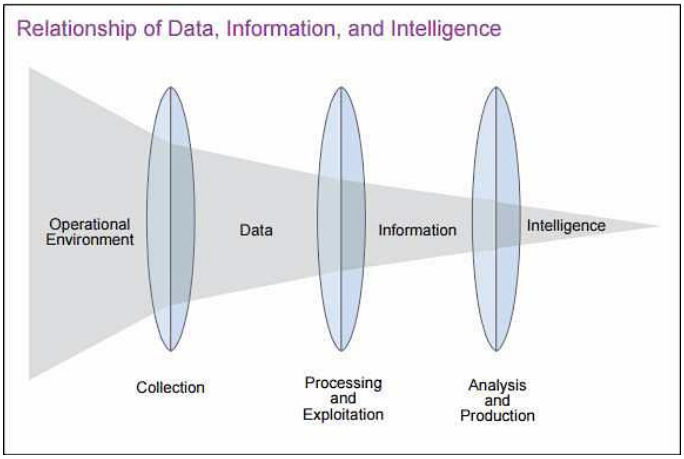
Na účely forenznej analýzy je taktiež zameraná práca [2], ktorá analyzuje mechanizmus konštrukcie kľúčových slov, ktoré sú bežne používané v hlavičke e-mailovej správy. Vybraná bola nasledujúca množina atribútov – „From”, „To”, „Date”, „Bcc” (blind carbon copy), „Subject”, „Return-Path”, „Message-ID” a „Received”, pričom sa autori zamerali najmä na atribúty „Received” a „Message-ID“, pomocou ktorých je možné lepšie identifikovať integritu a autenticitu e-mailu.

Našim záverom vyplývajúcim z predchádzajúcich prác sú nasledujúce vybrané atribúty – „From”, „Reply-To”, „Sender”, „Message-ID”, „Date”, „Content-Type”, „Authentication-Results”, „Return-Path”, „X-Forefront-Antispam-Report”, “X-Microsoft-Antispam” a URL odkazy. Celkový počet atribútov je 11, pričom 10 z nich pochádza z hlavičky a jeden z tela e-mailovej správy. Špecifickými sú 2 hlavičkové atribúty označené počiatočným písmenom „X“. Tie boli vybrané po bližšej analýze antispamových atribútov, ktoré sú prítomné pri odosielaní e-mailovej správy a vkladané ako dodatočné parametre v rámci služby Office365.

2 Threat intelligence

Ďalšou časťou našej práce je obohacovanie získaných atribútov pomocou threat intelligence a porovnanie jednotlivých prístupov. Z toho dôvodu sa v tejto kapitole venujeme vysvetleniu tohto pojmu, ako aj problematike zdieľania dát, či základnému členeniu zdrojov TI pre lepšie pochopenie pri samotnej práci.

Pojem cyber threat intelligence (CTI), resp. threat intelligence (TI) je aktuálnou témou pre mnohé organizácie, ktoré čelia nárastu kybernetických útokov. Problémom často býva identifikácia informácií, ktoré možno považovať za súčasť, no ešte väčším problémom je pochopenie samotného pojmu. Preto je potrebné si bližšie identifikovať dva pojmy, a to intelligence a threat intelligence. Kombinácia dát, informácií a inteligencie totiž tvorí životný cyklus a pri pochopení týchto odlišností možno vyťažiť maximum pri zdokonaľovaní a obohacovaní kybernetickej bezpečnosti. Tento vzájomný vzťah medzi popísanými pojmami môžeme vidieť na obrázku č. 3. Hlavným účelom inteligencie je operatívna činnosť zameraná na detekciu, prevenciu a reakciu. Zozbierané dáta sú spracované na poskytnutie informácií. Tieto informácie sa ďalej analyzujú a transformujú do formátu, ktorý možno implementovať do inteligencie [17]. Threat intelligence tvoria akékoľvek poznatky o hrozbách založené na dôkazoch, ktoré môžu pomôcť pri včasnom reagovaní na útok, alebo jeho zabráneniu. Zahŕňa technické indikátory, mechanizmy a využiteľné rady o už existujúcej, alebo vznikajúcej hrozbe [1]. Je dôležité poznamenať, že automatizácia a nástroje vyvinuté od rôznych spoločností síce zvyšujú efektivitu, ale dôležitým článkom naďalej zostáva ľudský faktor.



Obr. 3 Vzťah medzi pojmami dáta, informácia a inteligencia [17]

Poznáme 3 kategórie zdrojov TI, interné, externé zdroje a zdroje získané v rámci komunít. Za interné zdroje možno považovať údaje zozbierané v rámci organizácie z internej siete v podobe záznamov (logov), upozornení (alertov), reportov, či SIEM systému zavedeného v organizácii. Externé zdroje môžu pochádzať z verejne dostupných miest. Spomínaný problém identifikácie informácií spočíva najmä vo filtrácii dát pochádzajúcich z rôznych zdrojov. Z toho dôvodu je potrebné rozlišovať medzi relevanciou a kvalitou dát pochádzajúcich najmä z takto získaných zdrojov. Za dáta z poslednej kategórie pochádzajúce z komunít sú považované akékoľvek zdieľané dáta medzi členmi jednej komunity [17]. Ďalšou otázkou taktiež zostáva problém zdieľania dát, pri ktorej sa názory organizácii líšia. Hlavnou myšlienkou zdieľania dát je zlepšovanie situačného povedomia medzi zainteresovanými stranami, kde práve prostredníctvom zdieľania informácií o aktuálnych hrozbách či zraniteľnostiach možno promptne reagovať [1]. Pojem CTI vznikol s cieľom zefektívniť reakcie na útoky, kedy dokážu bezpečnostní analytici získavať informácie a včasne z nich rozpoznať relevantné indikátory kybernetických útokov [18].

Rozlišujeme dva rôzne prístupy k zdieľaniu CTI: manuálny a automatizovaný. Príkladmi nevýhod manuálneho prístupu môžu byť časové obmedzenie pri zdieľaní (pomalé), alebo miera ľudskej chybovosti pri spracovaní. Tieto nedostatky však vieme pomocou automatizovaného prístupu eliminovať. Napriek tomu je však manuálne zdieľanie často používaným, konkrétne v prípadoch, ak medzi jednotlivými stranami už existuje nejaký vzájomný vzťah. Zdieľanie CTI však prebieha aj v celosvetovom meradle, no každá krajina má stanovené vlastné pravidlá a predpisy, čo je možné zdieľať s ostatnými a čo už je potrebné anonymizovať [19].

2.1 Dostupné služby

V tejto podkapitole sa venujeme porovnaniu voľne dostupných služieb, pomocou ktorých je možné z extrahovaných atribútov obohatiť dáta získaním prídavných informácií. Na porovnanie sme vybrali 12 služieb poskytujúcich API, ktorými sú Shodan [20], AbuseIPDB [21], ip-api [22], ipapi [23], Verifalia [24] a EmailHippo [25], Email Dossier [26], Pulsedive [27], URLScan [28], Virustotal [29], CheckPhish [30], WhatIsMyIp [31].

	Vstupný atribút	Maximálny počet dopytov	Dostupná dokumentácia	Potrebná registrácia	Potrebný API kľúč/Session kľúč
Shodan	IP adresa	10 000 výsledkov za mesiac (ak členom)	áno	áno	áno
AbuseIPDB	IP adresa, doména	1000 dopytov za deň	áno	áno	áno
ip-api	IP adresa, doména	45 požiadaviek za minútu	áno	nie	nie
ipapi	IP adresa	1000 dopytov za mesiac	áno	áno	áno
Verifalia	E-mailová adresa	15 HTTPS požiadaviek za sekundu a 25 e-mailových adries (dopytov) za deň	áno	áno	nie
EmailHippo	E-mailová adresa	Prvých 100 dopytov zdarma, 220 požiadaviek za sekundu (platené konto)	áno	áno	áno
Email Dossier	E-mailová adresa	Podľa zakúpeného počtu	áno	áno	áno
Pulsedive	IP adresa, doména, URL odkaz	30 požiadaviek za minútu, 1000 za deň	áno	áno	áno
URLScan	URL odkaz	4 požiadavky za minútu a 500 za deň	áno	áno	áno
Virustotal	IP adresa, doména URL odkaz, hash	60 verejných skenov za minútu a 5000 za deň	áno	áno	áno
CheckPhish	IP adresa, doména, URL odkaz	25 dopytov za deň a 250 za mesiac	nie	áno	áno
WhatIsMyIp	hlavička e-mailu (mimo API), IP adresa	24 dopytov za deň	áno	áno	áno

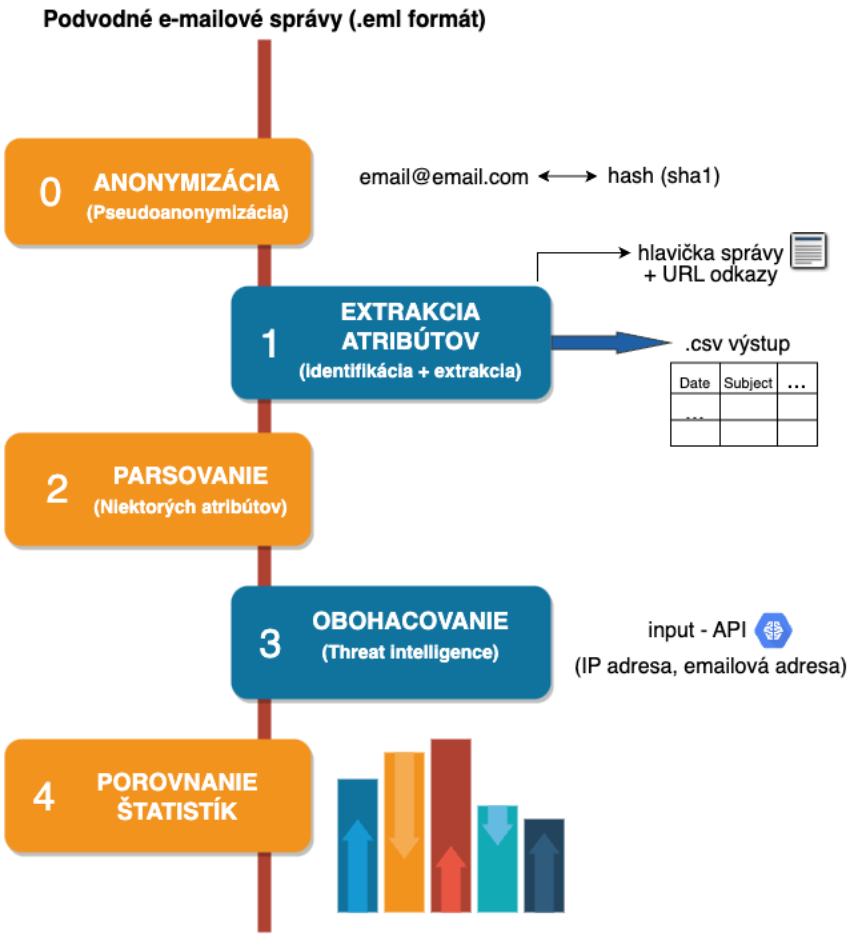
Tab. 2 Porovnanie služieb

V tabuľke č. 2 porovnávame jednotlivé služby podľa rôznych kategórií. Prvou kategóriou bolo porovnanie vstupných atribútov. Z výsledkov uvedených v tabuľke je

možné vidieť, že veľké množstvo služieb berie ako vstupný parameter najmä IP adresy, pričom služby Verifalia a EmailHippo sú zamerané na overenie e-mailovej adresy. Obe tieto služby sú však limitované, pri prvej z nich je možný len malý počet požiadaviek za deň, pričom služba EmailHippo je platená pri prekročení dopytovania nad 100 e-mailových adries (po registrácii). Služba WhatIsMyIp ponúka oproti ostatným vybraným službám overenie hlavičky e-mailovej správy. Táto možnosť je však aktuálne dostupná len na webovej stránke pri manuálnom zadávaní vstupu. Najmenej obmedzení na počet dopytov pri dlhodobejšom dopytovaní sa, má služba Shodan a AbuseIPDB s bezplatnou registráciou. Dokumentácia na oficiálnej webovej stránke bola dostupná na každej z vybraných okrem služby CheckPhish, pričom povinnosť registrácie pre efektívnejšie využívanie API nebola potrebná len pri službe ipapi. Na autentifikačné účely bolo vyžadované použitie API kľúča vo väčšine nami vybraných služieb. Pri službe Email Dossier bol využívaný tzv. session kľúč, pomocou ktorého nie je potrebné odosielať prihlasovacie meno a heslo pri každej požiadavke. Pre väčší obmedzujúci počet dopytov za deň sme pri obohacovaní nami vybraných atribútov vylúčili služby EmailHippo, CheckPhish a WhatIsMyIp. Taktiež sme vylúčili službu Email Dossier, ktorá ponúka využívanie API len v platenej verzii.

3 Návrh modelu

V tejto kapitole sa venujeme návrhu samotného modelu. Návrh pozostáva z niekoľkých bodov práce vysvetlených v jednotlivých podkapitolách, ktorých výsledným cieľom je automatizácia incident response (IR) pre podvodné e-maily. Na obrázku č. 4 môžeme vidieť graficky znázornené jednotlivé fázy popisovaných činností.



Obr. 4 Model

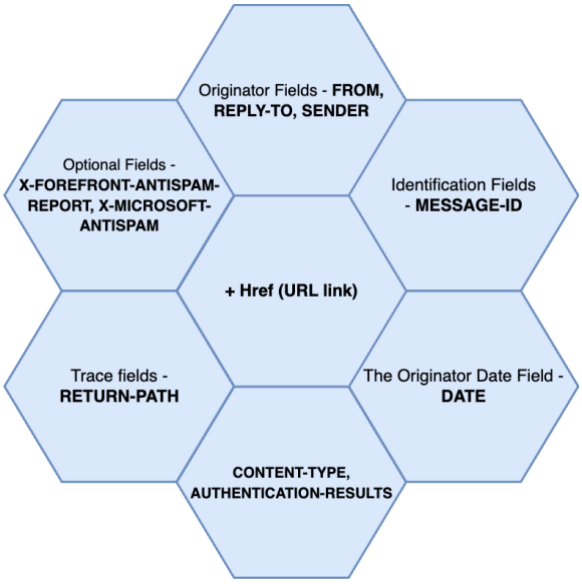
Vstupným parametrom tejto práce sú podvodné e-maily pochádzajúce z rôznych dostupných zdrojov. Dataset pozostáva z 212 podvodných e-mailov zbieraných medzi rokmi 2021-2023 vo formáte .eml.

3.1 Anonymizácia

V predprípravnej fáze bolo hlavnou úlohou anonymizovať e-mailové adresy príjemcov z jednotlivých e-mailov. Dôvodom je ochrana osobných údajov pri ďalšej možnej práci s datasetom a jeho zverejnení. Údaje uvedené pred zavináčom boli nahradené SHA1 hashom. Špecifickejšie sa jedná o pseudoanonymizáciu, kde je možné sa spätne dopátrať k e-mailovej adrese príjemcu. Vytvorená bola preto tabuľka s dvojicami hodnôt e-mailová adresa-hash.

3.2 Extrakcia atribútov

Pri ďalšej fáze bolo potrebné sa bližšie pozrieť na samotnú štruktúru e-mailu. Po bližšom preskúmaní atribútov z e-mailovej správy sme sa v tejto práci rozhodli extrahovať nasledujúcich 10 hlavičkových atribútov a 1 atribút z tela zobrazených na obrázku č. 5. Tieto hlavičkové atribúty sme následne uložili do .csv formátu pre ich bližšiu analýzu.



Obr. 5 Extrahované atribúty z e-mailovej správy

Z tela e-mailovej správy boli získané URL odkazy, ktoré sme po extrakcii uložili do samostatného .csv formátu. Súčasťou tohto dokumentu sú taktiež priradené názvy e-mailov k jednotlivým odkazom. URL odkazy sme získali pomocou regulárneho výrazu, pričom extrahovaná bola prvá časť odkazu po doménu (po prvú lomku uvedenú

za doménou). V celkovom počte sa jedná o 603 odkazov, či už sa jedná o samotné phishingové odkazy, alebo odkazy na priložené obrázky. Keďže však nie všetky extrahované odkazy boli v správnom formáte, bolo potrebné ich redukovať (odstránené boli taktiež linky, ktoré odkazovali na službu Google, keďže legitimitu nebolo potrebné posudzovať). Konečný zoznam preto obsahuje 133 unikátnych odkazov.

3.3 Parsovanie atribútov

Výstup s extrahovanými atribútmi bolo potrebné rozparsovať na získanie dodatočných informácií, keďže niektoré atribúty obsahovali viacero kategórií, popísaných v podkapitole 1.2 Hlavičkové atribúty. Jedná sa napríklad o atribút „Date“, ktorý obsahuje viacero údajov reprezentujúcich čas, či dátum. Oddeľovačom medzi jednotlivými údajmi bola medzera, pomocou ktorej sme ich dokázali rozdeliť do samostatných stĺpcov a následne pracovať len s vybranými. Ďalším špecifickým príkladom je získanie hodnôt SPF, DKIM a DMARC, ktoré sú uvedené v atribúte „Authentication-Results“. Keďže sa jedná o komplexnejší výstup a samotné oddelenie pomocou jedného vybraného znaku by nebolo efektívne, v tomto prípade bolo potrebné pomocou lambda funkcie zistiť počiatočné a konečné znaky. Podobne boli získané IP adresy z rovnakého hlavičkového atribútu. Postupným parsovaním sme dokázali získavať väčšie množstvo jednotlivých hodnôt vybraných atribútov.

3.4 Obohacovanie pomocou threat intelligence

Ďalšou časťou tohto modelu bolo obohacovanie získaných dát pomocou threat intelligence. Obohacované boli IP adresy, e-mailové adresy odosielateľov a URL odkazy. Z porovnávaných nástrojov v podkapitole 2.1 bolo vybraných 8 služieb, ktoré poskytujú bezplatné API s voľne dostupným kódom na implementáciu. Tieto služby sme rozdelili do troch kategórií podľa vstupných atribútov:

- služby na obohacovanie IP adries – Shodan, AbuseIPDB, ip-api, ipapi, Pulsedive,
- služba na obohacovanie e-mailových adries – Verifalia,
- služby na obohacovanie URL odkazov – URLScan, Pulsedive, Virustotal.

Pre rôzne služby bolo potrebné mierne upraviť vstupné dáta. Pokiaľ by sme sa pozreli na IP adresy, z každej e-mailovej správy bola extrahovaná jedna IP adresa z hlavičkového atribútu „Authentication-Results“. Po postupnom prechádzaní každého e-mailu boli tieto IP adresy vkladané do zoznamu, alebo boli uložené do súboru v .csv formáte. Takto extrahované IP adresy boli následne zadané na vstup do vybraných služieb poskytujúcich bezplatné API. Ak sa napríklad pozrieme na odlišnosti vstupov pri službách Shodan a AbuseIPDB, pri službe Shodan bolo potrebné jednotlivé IP adresy oddeliť čiarkou, pričom služba AbuseIPDB si vyžadovala každú IP adresu v samostatnom riadku.

Samotný proces získania obohatených údajov o URL odkaze bol oproti obohacovaniu IP adries mierne komplikovanejší. Pri celkovom procese od zadávania URL odkazov do vybraných služieb, až po ich samotné vyhodnotenie bolo niekoľko špecifických krokov pre dané služby. Prvý krok bol však pre trojicu vybraných služieb rovnaký. Ten zahŕňal extrakciu samotných odkazov z tela e-mailovej správy, ktorú sme popísali vyššie v tejto kapitole. Ostatné kroky sa však pre jednotlivé služby odlišovali, z toho dôvodu budú popísané pre každú z vybraných samostatne. Spoločným znakom pre všetky tri služby je však komplexnosť výstupných parametrov a ich samotná štruktúra. Preto sme vyhodnotené štatistické údaje neuvádzali a popísané budú len jednotlivé parametre.

3.5 Porovnanie štatistík

Spoločným znakom pre túto časť bolo načítanie extrahovaných údajov do dátového rámca. Porovnanie štatistík bolo však vykonané na dvoch miestach. Ako prvé sa je porovnanie atribútov a ich hodnôt po parsovaní, teda získané údaje z nášho datasetu a jeho extrakcie. Súčasťou je základná štatistika nad dátami, ktorá sa týkala najmä získania počtov jednotlivých hodnôt a vysvetlenie ich výsledkov v podvodných správach. Takto sme dokázali získať napr. vzájomné porovnanie medzi dňami v týždni vzhľadom na počet odoslaných e-mailov v konkrétny deň, či porovnanie jednotlivých krajín určených pomocou pripájajúcej IP adresy. Druhé porovnanie sa nachádza v úplnom závere po obohacovaní pomocou threat intelligence. Venovali sme sa porovnaniu výstupov z jednotlivých služieb, ako aj celkovému porovnaniu, nakoľko sme dokázali získať nové obohatené dáta.

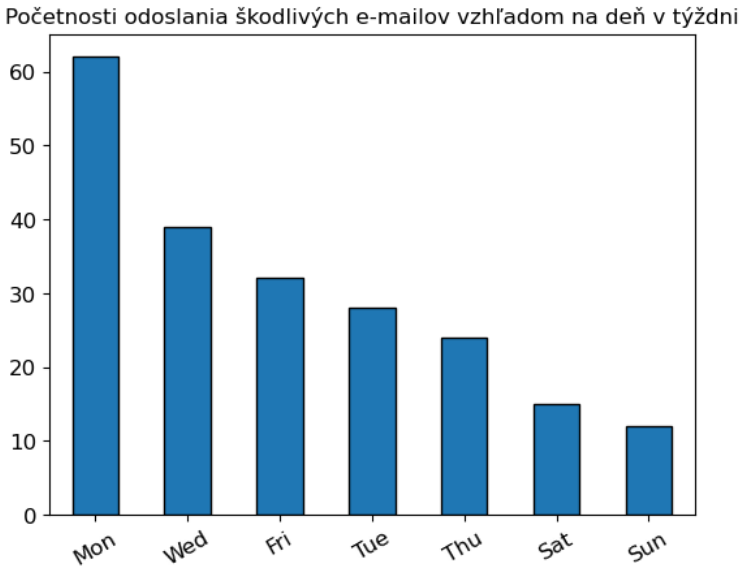
4 Vyhodnotenie

Pri implementácií jednotlivých častí práce sme pracovali s programovacím jazykom Python verzie 3, ktorý poskytuje množstvo rôznorodých knižníc. Táto kapitola je členená na tri podkapitoly podľa fázy obohacovania vyhodnotených štatistík, pričom posledná podkapitola je venovaná samotnému zhrnutiu všetkých výsledkov.

4.1 Základné štatistiky z extrahovaných atribútov

Nasledujúce vyhodnotenia pochádzajú z extrahovaných a rozparsovaných hlavičkových atribútov, ktoré sa nám podarilo získať. Jedná sa o základné štatistické údaje jednotlivých atribútov. Bližšie sme sa zamerali na atribút „Date“ teda dátum odoslania e-mailovej správy, kde sme zisťovali mieru početnosti odoslania škodlivých e-mailov vzhľadom na deň v týždni. Týmto atribútom sme sa zaoberali taktiež z dôvodu porovnania so štatistikou vytvorenou v roku 2021 [32], ktorá hovorí o najfrekventovanejších dňoch odosielania podvodných e-mailových správ v týždni so zameraním na služby Facebook a Microsoft. Výslednými dňami s najvyššou frekvenciou odosielania boli najmä štvrtok a piatok. Dôvodom práve týchto dní by mohla byť únava a nepozornosť používateľov, ktorí môžu byť v závere pracovného týždňa náchylnejší na vykonanie požadovanej akcie (kliknutie na URL odkaz, stiahnutie prílohy, ...) v podvrhnutej e-mailovej správe.

Zo štatistiky zobrazenej na obrázku č. 6, ktorá bola vytvorená z nášho datasetu je najfrekventovanejším dňom odosielania podvodných e-mailov pondelok, pričom najmenej frekventované sú dni sobota a nedeľa. Dni sú usporiadané od najvyššej frekvencie odosielania po najnižšiu. Na obrázku č. 7 môžeme vidieť presné číselné hodnoty prislúchajúce daným dňom. Z tohto porovnania usudzujeme, že útočníci si už nevyberajú špecifické dni na realizáciu útokov a podvodné e-mailové správy sú odosielané v akýkoľvek deň v týždni.



Obr. 6 Stĺpcový graf odoslaných e-mailov vzhľadom na deň v týždni

	Count
Mon	62
Wed	39
Fri	32
Tue	28
Thu	24
Sat	15
Sun	12

Obr. 7 Počet odoslaných e-mailov vzhľadom na deň v týždni

Pozreli sme sa aj na početnosti vzhľadom na rok zobrazené na obrázku č. 8. Keďže bol dataset podvodných e-mailových správ zbieraný najmä v roku 2022, jeho počet je najväčší vzhľadom na ostatné roky.

	Count
2022	159
2021	34
2023	19

Obr. 8 Počet odoslaných e-mailov vzhľadom na rok

Zbierané boli taktiež e-mailové adresy odosielateľov („From") a e-mailové adresy pri odpovedaní („Return-Path"). Na obrázku č. 9 môžeme vidieť nezhodu týchto atribútov až 150-krát (hodnota „False"), čo môže naznačovať podvodné e-mailové správy. Vo väčšine prípadov legitímnych e-mailov sa totiž tieto dva atribúty zhodujú, keďže ako odosielatelia legitímneho e-mailu sa chceme dozvedieť o prípadných problémoch s doručením na rovnakej e-mailovej adrese. Podvodné správy bývajú často „jednosmerné", teda je zmenená spiatočná e-mailová adresa a často jediným účelom je vykonanie akcie od používateľa a prítomnosť chybových hlásení nie je v tomto prípade dôležitá.

	True	False
Count	62	150

Obr. 9 Počet nezhôd v e-mailovej adrese medzi atribútmi From a Return-Path

Po extrahovaní hlavičkového atribútu „Authentication-Results" a jeho rozparsovaní sme získali hodnoty SPF, DKIM a DMARC. Ich nadobudnuté hodnoty sme porovnali v spoločnej tabuľke. Obrázok č. 10 zobrazuje počet jednotlivých nadobudnutých hodnôt. Pokiaľ sa pozrieme na stĺpce jednotlivo, najpočetnejšími hodnotami SPF boli „pass" (83) a „none" (81) v tomto poradí. Pri hodnote DKIM je hodnota „none" s najväčším počtom (126), čo taktiež platí pri DMARC hodnote s počtom 106. Jediná úspešná hodnota „pass" pri SPF, DKIM a DMARC je v nízkom počte, čo môže byť taktiež jedným z indikátorov podvodných e-mailov. Označenie „pass" pri každom z trojice týchto atribútov však nemusí automaticky naznačovať legitímny e-mail. Príkladom môže byť elektronicky podpísaný e-mail, pričom túto možnosť nemajú všetky organizácie, či jednotlivci, a teda k úspešnej hodnote v tomto prípade nedôjde. Jedná sa však o jeden zo stupňov overenia, ktorý sa pri e-mailoch často sleduje. Z toho dôvodu je potrebné pri rozhodovaní sa, či sa jedná o legitímny, alebo podvodný e-mail sa zamerať na viaceré aspekty týchto atribútov a rozhodovať sa na základe viacerých atribútov.

	SPF	DKIM	DMARC
pass	83	76	31
none	81	126	106
softfail	21	-	-
fail	14	9	26
temperror	7	-	7
permerror	4	-	6
neutral	2	-	-
timeout	-	1	-
bestguesspass	-	-	36

Obr. 10 Počet nadobudnutých hodnôt SPF, DKIM a DMARC

Zaujímavými atribútmi boli taktiež získané počty v súvislosti s krajinou určenou pomocou pripájajúcej IP adresy. Ten bol extrahovaný z antispamového atribútu „X-Forefront-Antispam-Report“ z časti „CTRY“. Podľa obrázku č. 11 môžeme vidieť, že na prvom mieste je umiestnená krajina Spojené Štáty (kód krajiny – „US“) s počtom 79. Ďalšími krajinami s počtami hodnôt 23 a 14 sú Slovensko (kód krajiny – „SK“) a Nemecko (kód krajiny – „DE“) v tomto poradí.

	Count
US	79
SK	23
DE	14
NL	11
TR	9
JP	7
ID	5
...	...
BD	1
GE	1
AL	1
CZ	1
CL	1
IL	1
TW	1

Obr. 11 Porovnanie krajín s ich počtom

4.2 Štatistiky po obohacovaní pomocou TI

Druhé vyhodnotenie dát bolo vykonané v poslednom bode tejto práce, konkrétne sa jedná o štatistiky po obohacovaní pomocou TI. Zo získaných atribútov sme obohacovali už spomínané IP adresy**,** e-mailové adresy odosielateľov a URL odkazy. Podľa týchto vstupných parametrov sú členené nasledujúce podkapitoly.

4.2.1 Služby na obohacovanie IP adries

Shodan

Služba Shodan slúži ako vyhľadávač, ktorý zhromažďuje informácie o zariadeniach pripojených k internetu. Vstupným parametrom pre túto službu bola množina IP adries uložená v súbore vo formáte .csv, pričom tieto IP adresy boli oddelené čiarkami. Množinu výstupných údajov sme si stanovili na 8 atribútov k zadanej IP adrese zo vstupu. Jedná sa o nasledujúce atribúty – „Organization", „Country Code", „Country Name", „City", „Latitude", „Longtitude", „Host Names", „Ports".

Z celkového počtu 212 IP adries nebolo možné zistiť dodatočné informácie o 84 IP adresách z dôvodu nedostupnosti dát. Pokiaľ by sme sa pozreli na porovnanie informácií pred obohacovaním, pričom sme z antispamového atribútu („X-Forefront-Antispam-Report") získali atribút „Country Code", tak sme pomocou tejto služby dokázali obohatiť IP adresu o ďalších 7 vybraných atribútov.

Obrázok č. 12 hovorí o počte otvorených portov získaných pomocou atribútu „Ports" zoradených zostupne. Vybraných bolo prvých 15 najpočetnejších, pričom otvorené porty s najvyšším počtom sú 80, 22 a 443.

Ports	Count
80	76
22	70
443	69
53	63
993	60
25	44
995	41
465	38
21	37
587	36
110	36
143	34
3306	29
2087	26
2082	24

Obr. 12 Porovnanie počtu otvorených portov

AbuseIPDB

Služba AbuseIPDB slúži na kontrolu IP adries, kde je možné nahlasovať dané IP adresy, ktoré sú zapojené do škodlivých aktivít a pridávať komentáre s krátkym popisom. Túto možnosť nahlasovania IP adries má každý používateľ po registrácii. Keďže pomocou služby Shodan nebolo možné získať informácie o všetkých IP adresách, pri obohacovaní sme pokračovali danou službou. Rovnako ako pri predchádzajúcej sme na vstup zadali .csv formát s IP adresami, pričom v tomto prípade bola každá IP adresa uvedená na samostatnom riadku. AbuseIPDB poskytuje taktiež iné a jedinečné výstupné atribúty oproti predchádzajúcej službe, preto boli v tomto prípade vybrané nasledujúce atribúty – „ipAddress", „isPublic" , „hostnames", „isWhitelisted", „usageType", „totalReports", „numDistinctUsers", „lastReportedAt", „countryCode", „isp", „domain", „abuseConfidenceScore".

Výstup z tejto služby nám poskytol dodatočné informácie o každej z 212 IP adries, ktoré boli na vstupe zadané. Na obrázku č. 13 môžeme vidieť zoradených 15 IP adries podľa posledného atribútu – „abuseConfidenceScore", ktorý sme zadávali ako výstupnú množinu od najvyššej hodnoty po najnižšiu. Jedná sa o atribút, ktorý predstavuje percento z počtu nahlásenia danej IP adresy. Prvá IP adresa je nahlásená so

skóre až 100% s celkovým počtom nahlásení 68. IP adresy na ďalších troch priečkach sa pohybujú so skóre medzi 70-90% s pomerne veľkým množstvom počtu nahlásení daných IP adries.

IP adresa	Abuse Score	Country Code	Total
188.225.21.131	100	RU	
167.89.16.17	88	US	
198.21.0.135	83	US	
167.89.10.181	75	US	
209.17.115.113	24	US	
203.146.237.187	21	TH	
165.22.51.52	18	SG	
74.208.4.197	18	US	
209.17.115.59	16	US	
209.17.115.59	16	US	
198.61.254.16	14	US	
18.156.147.178	14	DE	
103.230.126.139	12	MY	
211.1.227.5	10	JP	
185.236.231.107	10	PT	

Obr. 13 Porovnanie IP adries podľa abuse skóre

Taktiež sme sa pozreli na získané hodnoty z atribútu „usageType“. Tento atribút so sebou nesie informáciu o type použitia konkrétnych IP adries, ktoré sú rozdelené do niekoľkých kategórií. Na obrázku č. 14 môžeme vidieť jednotlivé kategórie priradené k počtu výskytu. Na prvom mieste s veľkým počtom 158 sa nachádzajú IP adresy priradené kategórii ako „Data Center/Web Hosting/Transit“. Na nasledujúcich priečkach môžeme vidieť ďalšie kategórie s nižším výskytom, či už sa jedná o rozsah IP adries, ktoré spadajú do akademického prostredia, alebo ďalších.

	Usage Type
Data Center/Web Hosting/Transit	158
University/College/School	24
Fixed Line ISP	18
Commercial	3
Government	2
Mobile ISP	1

Obr. 14 Počet klasifikácií typov použitia

ip-api

V poradí už treťou službou na obohacovanie IP adries sme vybrali službu s názvom ip-api. Jedná sa o jedinú službu, pri ktorej nebolo potrebné sa registrovať a teda ani generovanie akéhokoľvek API kľúča na dopytovanie. Obohacovanie bolo úspešné pri počte 207 IP adries zo vstupu, pričom do výstupnej množiny sme si vybrali všetky dostupné atribúty, ktorými sú „status“, „continent“, „continentCode“, „country“, „countryCode“, „region“, „regionName“, „city“, „district“, „zip“, „lat“, „lon“, „timezone“, „offset“, „currency“, „isp“, „org“, „as“, „asname“, „reverse“, „mobile“, „proxy“, „hosting“, „query“.

Môžeme si všimnúť podobnosť výstupných atribútov z jednotlivých služieb, no každá služba ponúka svoje špecifické atribúty, ktoré nie sú implementované v inej. V tejto službe sú týmito atribútmi napr. „timezone“, kde je uvedené časové pásmo a „offset“, pričom ten predstavuje posun oproti časovému pásmu UTC DST (Coordinated Universal Time - Daylight saving time) v sekundách. Za týmto atribútom môžeme taktiež vidieť uvedenú menu v danej krajine („currency“). Na chvoste výstupnej množiny sa nachádzajú boolean hodnoty, či sa jedná o mobilné pripojenie („mobile“), alebo či sa jedná o Proxy, VPN, alebo Tor adresu („proxy“).

Práve tieto dva vybrané atribúty môžeme vidieť na nasledujúcom obrázku č. 15. Porovnávali sme uvedené boolean hodnoty v oboch atribútoch, Konkrétnejšie, vyhodnotenie ako pravdivé bolo len v troch prípadoch atribútu „mobile“ a v ôsmych atribútu „proxy“.

	Mobile	Proxy
false	204	199
true	3	8

Obr. 15 Porovnanie nadobudnutých hodnôt atribútov mobile a proxy

ipapi

Nasleduje rovnomenná služba s názvom ipapi, pomocou ktorej sa nám podarilo obohatiť všetky extrahované IP adresy. Tie boli vo forme listu zadané na vstup do API tejto služby. Vybrané parametre do výstupnej množiny boli všetky dostupné z bezplatného balíčka po registrácii. Sú nimi „ip", „type", „continent_code", „continent_name", „country_code", „country_name", „region_code", „region_name", „city", „zip", „latitude", „longtitude", „location", „capital", „languages", „name", „native", „country_flag", „country_flag_emoji", „country_flag_emoji_unicode", „calling_code", „is_eu".

Jedná sa zväčša o základné atribúty, ktoré sú dostupné aj v iných službách. Z toho dôvodu sme nad samotnými atribútmi nerobili žiadne štatistické údaje. Avšak nachádzajú sa tu taktiež špecifické atribúty, ktoré stoja za zmienku. Pomocou atribútu „capital" sa vieme dopátrať k hlavnému mestu príslušnej krajiny, „native" nám prezradí úradný jazyk a atribút „calling_code" zas telefónnu predvoľbu. Ako zaujímavosť vo výstupnej množine je uvedený taktiež atribút „country_flag_emoji", ktorý predstavuje emotikon vlajky príslušnej krajiny a jeho unicode zápis v jeho susednom atribúte („country_flag_emoji_unicode").

Pulsedive

Posledná zo služieb, ktoré sme si vybrali pre obohacovanie IP adries má názov Pulsedive. Ide o threat intelligence platformu, pričom je možné skenovať rôzne druhy IoC. Pulsedive v sebe integruje trojicu nástrojov Shodan, Virustotal a AbuseIPDB, výsledky z ktorých sú však dostupné až v platenej verzii. V našej práci sa nám však podarilo pozrieť sa samostatne na všetky zo spomínaných.

Ako vstupné parametre do tejto služby sme si zvolili IP adresy a URL odkazy (uvedené nižšie). Výsledkom obohatenia bolo však len 47 IP adries s ich dodatočnými, no užitočnými a rozsiahlymi informáciami. Keďže sa jedná o komplexnejší výstup,

v tomto prípade neuvádzame konkrétnu výstupnú množinu atribútov a ani vykonané štatistické údaje. Dôvodom bol veľký počet parametrov, pričom každý výstup o jednej IP adrese sa líšil, a preto sa nám nepodarilo zjednotiť jednotlivé výstupy. Na obrázku č. 16 však uvádzame ukážku obohatených atribútov pre jednu IP adresu.


```
"qid": null,
"iid": 30513877,
"indicator": "212.193.30.247",
"type": "ip",
"risk": "medium",
"risk_recommended": "medium",
"manualrisk": 0,
"retired": "No recent activity",
"stamp_added": "2021-12-18 22:11:35",
"stamp_updated": "2022-04-01 18:48:16",
"stamp_seen": "2021-12-18 22:18:40",
"stamp_probed": "2021-12-18 22:18:40",
"stamp_retired": "2022-04-01 18:48:16",
"recent": 0,
"submissions": 0,
"umbrella_rank": null,
"umbrella_domain": null,
"riskfactors": [
    {
        "rfid": 60,
        "description": "found in threat feeds",
        "risk": "medium"

],
"redirects": {
    "from": [],
    "to": []
},
"threats": [],
"feeds": [
    {
        "fid": 45,
        "name": "Blocklist.de Blocklist",
        "category": "abuse",
        "organization": "Blocklist.de",
        "pricing": "free",
        "stamp_linked": "2021-12-18 22:11:35"

],
"comments": [],
"attributes": {
    "port": [
        "3389"
    ],
    "protocol": [
        "RDP"
    ]
},
"properties": {
    "geo": {
        "countrycode": "CZ",
        "country": "Czechia",
        "lat": "50.0853",
        "long": "14.411",
        "isp": "Delis LLC",
        "asn": "AS211252",
        "org": "Delis LLC"
```


Obr. 16 Ukážka výstupu Pulsedive

Napriek tomu sme vybrali niekoľko atribútov, ktoré uvádzame aj s vysvetlením. Sú nimi napríklad atribúty „risk“ – ohodnotenie IP adresy podľa získaných výstupov, „riskfactors“ – rozvetvenejší atribút, ktorý obsahuje popisy jednotlivých faktorov s ich samostatne ohodnoteným rizikom, či „attributes“, ktorý obsahuje ďalšie tri podatribúty – „port“ (podobne ako pri službe Shodan označuje tento atribút otvorené porty dostupné z internetu), „protocol“ a „technology“. V atribúte „feeds“ môžeme vidieť výsledky dopytov na iné služby, pričom každý dopyt je označený svojim identifikátorom, názvom, priradenou kategóriou a ďalšími. Ďalším užitočným je atribút „properties“, ktorý v sebe zahŕňa základné informácie, či abuse kontakt o danej IP adrese s názvom

koreňového atribútu „whois“. Jedná sa skutočne o rozsiahlu službu, ktorá v sebe zahŕňa množstvo užitočných informácií.

4.2.2 Služba na obohacovanie e-mailových adries

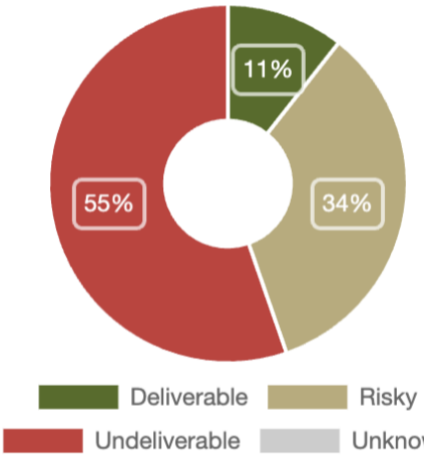
Pri obohacovaní e-mailových adries sme z porovnávaných služieb v podkapitole 2.1 vybrali len službu Verifalia, ktorá ponúka bezplatnú verziu po registrácii.

Verifalia

Služba Verifalia ponúka real-time overovanie e-mailových adries, pričom overuje validitu, či rizikovosť jednotlivých e-mailových adries. Jedná sa však o najviac limitované API služby z množiny porovnávaných služieb. Maximálny počet e-mailových adries na overenie je stanovený na 25 dopytov za deň pri registrácii v rámci základného balíčka. Z dôvodu tohto obmedzenia prebehlo obohacovanie len na vzorke 124 e-mailových adries. Výstupná množina bola v tomto prípade stanovaná na nasledujúcich 21 atribútov – „InputData“, „Classification“, „Status“, „EmailAddress“, „EmailAddressLocalPart“, „EmailAddressDomainPart“, „AsciiEmailAddressDomainPart“, „HasInternationalDomainName“, „HasInternationalMailboxName“, „IsCatchAllFailure“, „IsDisposableEmailAddress“, „IsRoleAccount“, „IsDnsFailure“, „IsMailboxFailure“, „IsNetworkFailure“, „IsSmtpFailure“, „IsSuccess“, „IsSyntaxFailure“, „IsTimeoutFailure“, „SyntaxFailureIndex“, „CompletedOn“.

O každej e-mailovej adrese zo vstupného .csv formátu boli zistené dodatočné informácie, ako je napríklad kategorizácia podľa klasifikácie či statusu. Posledné tri atribúty z našej vybranej množiny hovoria o tom, v ktorej fáze overovania validácie nastalo zlyhanie.

Obrázok č. 17 predstavuje percentuálne zastúpenie porovnania jednotlivých klasifikácií e-mailových adries z atribútu „Classification“. Môžeme vidieť, že až 55% tvoria e-mailové adresy, ktoré boli označená ako „Undeliverable“ (Nedoručiteľné), 34% ako „Risky“ (Riskantné) a len zvyšných 11% ako „Deliverable“ (Doručiteľné).



Obr. 17 Porovnanie klasifikácií e-mailových adries

Pozreli sme sa taktiež na výsledky atribútu „Status“. Obrázok č. 18 predstavuje porovnanie statusov s ich počtom výskytu. Väčšina týchto statusov označuje konkrétnejšiu chybu pri overovaní, či už „MailboxValidationTimeout“ – vypršanie časového limitu (25), „MailboxDoesNotExist“ - neexistencia e-mailovej schránky (25), alebo „DomainDoesNotExist“ – neexistencia domény (18). S úspešným statusom „Success“ bolo vyhodnotených len 14 e-mailových adries. Z týchto výsledkov môžeme usúdiť, že väčšina e-mailových adries zo vstupu neboli validné, keďže nastal problém v určitej fáze ich overovania. Útočníci tak mohli vytvoriť náhodné e-mailové adresy, pomocou ktorých rozposielali podvodné e-mailové správy.

	Status
MailboxValidationTimeout	25
MailboxDoesNotExist	25
DomainDoesNotExist	18
ServerIsCatchAll	15
Success	14
DomainIsMisconfigured	12
SmtpConnectionTimeout	7
UnhandledException	2
DomainPartCompliancyFailure	2
MailboxHasInsufficientStorage	1
DomainHasNullMx	1
LocalEndPointRejected	1
CatchAllValidationTimeout	1

Obr. 18 Porovnanie statusov s ich počtom výskytu

4.2.3 Služby na obohacovanie URL odkazov

V nasledujúcej časti uvádzame len prvé dve služby z vybraných, keďže po analýze výsledkov zo služby Pulsedive sme nedostali dodatočné atribúty, ktoré by sme mohli popísať v porovnaní s výslednými atribútmi z obohacovania IP adries.

URLScan

Pri analýze škodlivých odkazov je ako prvou voľbou často práve služba s názvom URLScan. Táto služba slúži na skenovanie a analýzu vložených odkazov, pričom počas skenovania zaznamenáva aktivitu, kam nás jednotlivé odkazy môžu zaviesť. Z toho dôvodu sa vo všeobecnosti jedná o nesmierne užitočný nástroj, pokiaľ si nie sme istí legitimitou priloženého odkazu.

Pri samotnom procese získania dodatočných informácií o URL odkazoch v našej práci po úvodnom prvom kroku extrakcie odkazov nasledovala časť dopytovania sa s nasledujúcimi výstupnými atribútmi – „message", „uuid", „result", „api", „visibility", „options", „url". Prvý atribút „message" predstavuje krátky popis výsledku dopytu. Zo získaného počtu 133 URL odkazov bolo správou „Submission successful" označených 99 z nich. Tento výstup bol uložený vo formáte .csv, pričom ostatné neúspešné hlásenia sme odfiltrovali.

Pri ďalšom kroku získania obohatených údajov o URL odkazoch bolo potrebné extrahovať a rozparsovať atribút „result". Formát tohto atribútu je v nasledujúcom tvare: https://urlscan.io/result/uuid. Pre každý takýto odkaz bolo potrebné získať poslednú časť s označením „uuid", ktorá hovorí o unikátnom vygenerovanom identifikátore pre každý sken. Tieto identifikátory sme si uložili do samostatného listu a požiadavku na získanie obohatených údajov sme opakovali. Keďže samotné generovanie bolo pri komplexnejšom výstupe časovo náročné, náš záverečný výstup z tejto služby obohacuje 52 URL odkazov. Tento výstup je vo formáte JSON, ktorý obsahuje nasledujúce atribúty najvyššej úrovne – „task", „page", „lists", „data", „meta", „stats", „verdicts". Na obrázku č. 19 môžeme vidieť ukážku výstupu obohatenia jedného URL odkazu, pričom si môžeme všimnúť tri atribúty najvyššej úrovne – „task", „verdicts" a „stats".

```
"task": {
  "uuid": "901126c1-5c66-4140-ab84-a5ac0289a675",
  "time": "2023-04-20T11:23:12.597Z",
  "url": "https://chitterchatter.co.uk/",
  "visibility": "public",
  "method": "api",
  "source": "6c31ba32",
  "tags": [],
  "reportURL": "https://urlscan.io/result/901126c1-5c66-4140-ab84-a5ac0289a675/",
  "screenshotURL": "https://urlscan.io/screenshots/901126c1-5c66-4140-ab84-a5ac0289a675.png",
  "domURL": "https://urlscan.io/dom/901126c1-5c66-4140-ab84-a5ac0289a675/"
},
"verdicts": {
  "overall": {
    "score": 100,
    "categories": [
      "phishing"
    ],
    "brands": [
      "genericcloudflare"
    ],
    "tags": [
      "phishing"
    ],
    "malicious": true,
    "hasVerdicts": true
  },
  "urlscan": {
    "score": 100,
    "categories": [
      "phishing"
    ],
    "brands": [
      {
        "key": "genericcloudflare",
        "name": "Generic Cloudflare",
        "country": [
          "us"
        ],
        "vertical": [
          "Online"
        ]
      }
    ],
    "tags": [
      "phishing"
    ],
    "malicious": true,
    "hasVerdicts": true
  },
"stats": {
  "resourceStats": [
    {
      "count": 2,
      "size": 11925,
      "encodedSize": 2202,
      "latency": 0,
      "countries": [
        "US"
      ],
      "ips": [
        "2a06:98c1:3121::3"
      ],
      "type": "Document",
      "compression": "5.4",
      "percentage": 40
    },
```


Obr. 19 Ukážka výstupu URLScan

Prvý atribút „task" zahŕňa základné informácie o požiadavke, ako je časový údaj začiatku skenovania daného URL odkazu, metóda zadávania (v našom prípade hodnota „api") a taktiež URL odkazy na snímky jednotlivých stránok. Súčasťou atribútu „page" sú atribúty týkajúce sa samotnej stránky, ktorými sú napríklad IP adresa kontaktovaná pre primárnu požiadavku, s ktorou súvisí aj informácia o krajine, meste, či hlavičkový atribút HTTP „Server". Nasleduje zoznam kontaktovaných IP adries (a krajín s nimi spojenými), domén, URL odkazov, TLS certifikátov a SHA256 hashov a ďalších obsiahnutých v atribúte „lists". Predposledný atribút „stats" obsahuje vypočítané štatistiky podľa typu, protokolu, IP adries a iné. Vyhodnotenie verdiktov ohľadom škodlivého obsahu je uvedené v poslednom atribúte „verdicts" najvyššej úrovne. Jedná sa o verdikty zo samotnej služby, či komunity, súčasťou ktorých je aj vyhodnotené skóre na základe získaných informácií (hodnoty sa pohybujú od -100 až 100, pričom skóre 100 naznačuje škodlivý obsah).

Virustotal

Ďalšou bežne používanou službou, a zároveň poslednou v našom zozname na obohacovanie URL odkazov je služba Virustotal. Špecifikum tejto služby je analýza vstupného parametru, v našom prípade URL odkazu, pomocou viac ako 70-tich antivírusových skenerov a ich vyhodnotení.

Po prvotnom získaní URL odkazov bolo potrebné pre každý z nich vygenerovať URL identifikátor. Tento reťazec znakov sme vytvorili pomocou base64 kódovania, pričom výsledné identifikátory nie sú vyplnené znakom „=“, ako to zvyčajne pri tomto kódovaní býva. Po tomto kroku sme už dopyty smerovali v upravenom tvare URL odkazu uvedeného pre túto službu. Podobne ako pri službe URLScan sa jednalo o zložitejší a rozvetvený výstup rozdelený na viacero úrovní a atribúty jednotlivých odkazov sa odlišovali. Zároveň bolo generovanie obmedzené na 4 dopyty za minútu, z toho dôvodu bolo obohacovanie vykonané na vzorke 46 URL odkazov. Na obrázku č. 20 uvádzame ukážku výstupu z danej služby.


```
"data": {
    "attributes": {
        "last_modification_date": 1682716017,
        "times_submitted": 7078,
        "total_votes": {
            "harmless": 5,
            "malicious": 25
        },
        "threat_names": [],
        "redirection_chain": [
            "http://www.w3.org/"
        ],
        "last_submission_date": 1682716005,
        "last_http_response_content_length": 29311,
        "last_http_response_headers": {
            "Content-Length": "7857",
            "CF-Cache-Status": "HIT",
            "Age": "0",
            "expires": "Fri, 28 Apr 2023 21:16:12 GMT",
            "vary": "Accept-Encoding",
            "Server": "cloudflare",
            "last-modified": "Thu, 27 Apr 2023 09:00:11 GMT",
            "Connection": "keep-alive",
            "x-backend": "www-mirrors",
            "x-request-id": "7be5ea235e43410d",
            "content-location": "Home.html",
            "content-security-policy": "upgrade-insecure-requests",
            "Content-Encoding": "gzip",
            "strict-transport-security": "max-age=15552000; includeSubdomains; preload",
            "etag": "\"727f-5fa4d957d5cc0-gzip\"",
            "Cache-Control": "max-age=600",
            "Date": "Fri, 28 Apr 2023 21:06:47 GMT",
            "CF-RAY": "7bf245684f4fc50c-ORD",
            "alt-svc": "h3=\":443\"; ma=86400, h3-29=\":443\"; ma=86400",
            "Content-Type": "text/html; charset=utf-8",
            "Accept-Ranges": "bytes"
        },
        "reputation": -86,
        "tags": [],
        "last_analysis_date": 1682716005,
        "first_submission_date": 1284591183,
        "categories": {
            "Forcepoint ThreatSeeker": "information technology",
            "Sophos": "information technology",
            "Xcitium Verdict Cloud": "mobile communications",
            "BitDefender": "computersandsoftware"
        },
        "last_http_response_content_sha256": "61a5b712900488d8035ec3517e6f3096d40637cbc9a917893170eba18e1eed62",
        "last_http_response_code": 200,
        "last_final_url": "https://www.w3.org/",
        "url": "http://www.w3.org/",
        "title": "World Wide Web Consortium (W3C)",
        "last_analysis_stats": {
            "harmless": 73,
            "malicious": 0,
            "suspicious": 1,
            "undetected": 15,
            "timeout": 0
        },
```


Obr. 20 Ukážka výstupu Virustotal

Pokiaľ sa pozrieme na vygenerovaný výstup a niektoré jeho výstupné atribúty, príkladmi atribútov venujúcich sa štatistikám sú napríklad „times_submitted“, „total_votes“, či „reputation“. Prvý z uvedených atribútov hovorí o počte skenovaní jednotlivého odkazu. V našom prípade bola najvyššia hodnota s počtom 16 847. V atribúte „total_votes“ sme sa mohli dozvedieť o počte negatívnych („malicious“) a pozitívnych („harmless“) hodnotení získaných od Virustotal komunity. So spomínaným atribútom súvisí taktiež nasledujúci atribút „reputation“, pričom tento obsahuje jednu číselnú hodnotu všetkých hlasov komunity. Posledným atribútom, ktorý by sme chceli spomenúť v rámci štatistických údajov z výsledkov je atribút „last_analysis_stats“, súčasťou ktorého sú kategórie s ich pridelenými číselnými hodnotami. Jedná sa o zosumarizovanie výsledkov z rôznych URL skenerov, pričom sú zaradené do piatich kategórií – harmless, malicious, suspicious, timeout a undetected. Samostatné vyhodnotenie z jednotlivých skenerov sa nachádza v atribúte s názvom „last_analysis_results“.

V atribúte „category“ si vieme pozrieť ďalšie kategórie a ich službu (URL skener), ktorá danému URL odkazu pridelila kategóriu napr. „Sophos: information technology“. Ďalší zaujímavý atribút s názvom „trackers“ obsahuje všetky nájdené trackery v danom URL odkaze (skripty na webových stránkach, ktoré získavajú dáta o preferenciách a aktivite používateľa). Pomocou tohto atribútu sme v našom výslednom reporte zaznamenali 8 rôznych trackerov.

Pokiaľ by sme sa pozreli na atribúty „last_final_url“, „outgoing_links“ a „self“ môžeme si všimnúť, že ich výslednými hodnotami je odkaz, alebo zoznam odkazov. Pomocou atribútu „last_final_url“ si vieme pozrieť stránku, kam by nás pôvodný URL odkaz presmeroval. Atribút „outgoing_links“ obsahuje zoznam odkazov na rôzne iné domény spojené s našim skenovaným odkazom a posledný atribút „self“ v sebe zahŕňa odkaz na stránku Virustotal-u, kde sú uvedené kompletné výsledky pre daný URL odkaz.

4.3 Zhrnutie

Poslednou časťou tejto práce je zhrnutie výsledkov pred a po obohacovaní pomocou TI. Počet získaných atribútov najvyššej úrovne z e-mailovej správy bol 11 pred obohacovaním (10 z hlavičky a jeden z tela). Následným parsovaním sa počet

navýšil o 17, konkrétne atribút „Authentication-Results“ v sebe zahŕňa 4 hodnoty - SPF, DKIM, DMARC a IP adresu, atribút „X-Forefront-Antispam-Report“ nám ponúkol 12 dodatočných podatribútov (hodnoty niektorých však neboli pri každom vyplnené) a „X-Microsoft-Antispam“ 1 podatribút. Celkovo sa nám teda podarilo pred obohacovaním získať z jednej e-mailovej správy 25 atribútov (pričom atribúty „Sender“ a „Reply-To“ neobsahovali vždy hodnotu).

Súčet všetkých atribútov získaných po obohacovaní pomocou TI bol komplikovanejší, keďže výstupy z niektorých služieb boli rozvetvené a nebolo možné stanoviť presný počet. Z toho dôvodu sme sa rozhodli uviesť len približné počty. Pri procese spočítavania jednotlivých atribútov sme odstránili duplicitné atribúty, ktoré sa v rôznych službách opakovali. Do úvahy sme v službách – Pulsedive, URLScan a Virustotal počítali atribúty, ktoré sa opakovali pri viacerých výstupoch, a teda sa nejednalo o jedinečné atribúty danej IP adresy, či URL odkazu. Pri službe Pulsedive sme tento počet stanovili na 11, pri službe URLScan na 18 a službe Virustotal na 9 atribútov, pričom každý z nich po rozparsovaní obsahoval dodatočné informácie, ktoré sme však nezapočítavali. Počet atribútov získaných z vybraných služieb na obohacovanie IP adries sme stanovili na celkový počet 53, pri obohacovaní adries odosielateľov sa jedná o celkový počet 18 a URL odkazov o 38 atribútov. Po súčte jednotlivých obohacovaní sa dostávame k záverečnému približnému počtu získaných atribútov pre jednu e-mailovú správu pomocou obohacovania TI, ktorým je číslo 109. Takéto získané údaje bolo možné získať pomocou threat intelligence platforiem, ktoré poskytujú zdieľané informácie.

Pokiaľ by sme sa pozreli na počet získaných atribútov pri tejto práci v rámci analýzy podvodných e-mailov zobrazených v tabuľke č.3, celkovo sa nám podarilo získať približne 134 atribútov od základnej identifikácie atribútov dostupných v e-mailovej správe po ich obohacovanie pomocou TI .

	(približný) počet získaných atribútov
pred obohacovaním	25
po obohacovaní	109
SPOLU	**134**

Tab. 3 Počet získaných atribútov pri analýze podvodného e-mailu

Po analýze podvodných e-mailov a vytvorených štatistikách je možné stanoviť sadu odporúčaní, ktoré možno využiť pri nahlasovaní e-mailu resp. niektorých jeho atribútov. Pomocou získaných dodatočných informácií z obohacovania prostredníctvom TI vieme získať širší obraz o daných atribútoch. Práve vybrané obohacované atribúty – IP adresy, e-mailové adresy odosielateľov a URL odkazy sú tými, ktoré vieme istým spôsobom korigovať resp. nahlasovať. Príkladom môže byť e-mailová adresa, ktorá bola po overení validity vyhodnotená ako riskantná, či neexistujúca. Pri IP adrese sa vieme pozrieť na počet nahlásení resp. jej reputáciu z rôznych služieb a aké aktivity boli prostredníctvom nej vykonávané. Taktiež si vieme porovnať krajinu, ktorá je udávaná pri danej IP adrese s textom, v akom je daný e-mail písaný a čo je jeho zámerom. Prostredníctvom služieb na overenie URL odkazov sa vieme pozrieť na snímku stránky a taktiež ju porovnať s obsahom danej e-mailovej správy. Uvedené sú taktiež ďalšie informácie o odkaze, ktoré nám naznačujú, či sa jedná o legitímny obsah, alebo sa jedná o napodobnenie stránky už známeho poskytovateľa. Z toho dôvodu sú jednotlivé získané hodnoty vhodnými ukazovateľmi, či je potrebné dané IP adresy, e-mailové adresy či URL odkazy nahlásiť. Nahlasovanie je vo všeobecnosti možné rôznymi spôsobmi, či už priamo v e-mailovom klientovi, pričom označíme celý e-mail napríklad ako odosielateľa hromadnej pošty, alebo je možné nahlásiť jednotlivé atribúty. V tom prípade je jednou z možností nahlasovací formulár v rôznych TI službách. Pokiaľ však pri kontrole IP adresy zistíme, že patrí pod rozsah konkrétnej spoločnosti, ktorá poskytuje nahlasovací formulár, v tom prípade ju vieme nahlásiť priamo tam.

Záver

E-mailová komunikácia je jednou z bežne využívaných, ktorá nás sprevádza na každom kroku. Pri rôznych útokoch či už na organizácie, alebo jednotlivcov sú často počiatočným prienikom do systému práve podvodné e-mailové správy. Ak teda dokážeme skrátiť čas prvotnej analýzy od prijatia podvodného e-mailu, môžeme na útoky reagovať včas.

Hlavným cieľom tejto bakalárskej práce bola automatizácia samotného procesu analýzy podvodných e-mailových správ, súčasťou ktorého uvádzame záverečný súhrn štatistík zo získaných atribútov. Pomocou vyvinutého nástroja dokážeme promptne reagovať na podvodné e-mailové správy zadané na vstupe.

Prvým cieľom tejto práce bolo identifikovať, extrahovať a analyzovať indikátory kompromitácie z podvodných e-mailových správ. Z toho dôvodu sa v prvej kapitole venujeme základnej terminológii e-mailových správ, pričom dôležitú časť tvorí pochopenie ich samotnej štruktúry. Keďže sme sa rozhodli extrahovať atribúty najmä z hlavičky e-mailových správ, bližšiu pozornosť im venujeme v samostatnej podkapitole. Súčasťou tejto kapitoly je taktiež porovnanie podobných prác, ktoré súvisia s touto problematikou, pomocou ktorých sme dokázali stanoviť výsledné extrahované atribúty.

V druhom cieli sme porovnali prístupy na obohacovanie indikátorov kompromitácie prostredníctvom threat intelligence. V úvodnej časti druhej kapitoly bol pojem threat intelligence predstavený, pričom ďalšia časť bola venovaná porovnaniu vybraných služieb, ktoré ponúkajú možnosť obohatiť vstupné dáta. Vybrané služby sme rozdelili do troch základných kategórií podľa obohacovaných atribútov.

Posledným cieľom bolo navrhnúť a implementovať nástroj na analýzu podvodných e-mailových správ. Súčasťou neho ponúkame štatistické údaje z extrahovaných atribútov. Tomuto cieľu sme sa venovali v posledných dvoch kapitolách. V tretej kapitole sme predstavili samotný návrh modelu, pričom sme bližšie popísali každú vykonávanú časť. Posledná kapitola bola venovaná vyhodnoteniu získaných údajov. Tú sme rozdelili na dve ďalšie podkapitoly podľa rozlišovania skúmaných atribútov pred a po obohacovaní pomocou TI, pričom na záver tejto kapitoly uvádzame výsledné zhrnutie. Počet získaných atribútov z každej e-mailovej správy pred obohacovaním sa pohybuje okolo 25 atribútov, pričom po obohacovaní sa nám podarilo získať približne

109 atribútov. Celkovo tak približne hovoríme o konečnom počte 134 získaných atribútov z jednej e-mailovej správy.

Na záver taktiež uvádzame krátku sadu odporúčaní pre incident response. Jedná sa o príklady všeobecných odporúčaní, ktoré možno vykonať pomocou získaných hodnôt z vykonaných štatistík.

V budúcnosti je možné túto prácu doplniť o väčšie množstvo extrahovaných atribútov, ktoré by sa obohacovali v ďalších rôznych službách poskytujúcich TI. Keďže sme v tejto práci analyzovali zväčša hlavičkové atribúty, zaujímavé by mohli byť taktiež atribúty z tela e-mailovej správy, pričom by sme sa mohli pozrieť aj na samotný text.

Zoznam použitej literatúry

1. TOUNSI, Wiem; RAIS, Helmi. A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers & security, 2018, 72: 212-233.
2. GUO, Hong; JIN, Bo; QIAN, Wei. Analysis of email header for forensics purpose. In: 2013 International Conference on Communication Systems and Network Technologies. IEEE, 2013. p. 340-344.
3. KUMARI, Anuradha; AGRAWAL, Nitin; LILHORE, Umesh. Attack over email system. International Journal of Scientific Research & Engineering Trends, 2017, 3.5: 200-206.
4. RIABOV, Vladimir V. SMTP (Simple Mail Transfer Protocol). River College, 2005.
5. COHEN, Aviad; NISSIM, Nir; ELOVICI, Yuval. Novel set of general descriptive features for enhanced detection of malicious emails using machine learning methods. Expert Systems with Applications, 2018, 110: 143-169.
6. FANG, Yong, et al. Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism. IEEE Access, 2019, 7: 56329-56340.
7. ALMOMANI, Ammar, et al. A survey of phishing email filtering techniques. IEEE communications surveys & tutorials, 2013, 15.4: 2070-2090.
8. CHARALAMBOU, E., et al. Email forensic tools: A roadmap to email header analysis through a cybercrime use case. Journal of Polish Safety and Reliability Association, 2016, 7.1.
9. RFC 5322 [online]. [cit. 2023-04-15]. Dostupné z: https://www.rfc-editor.org/rfc/rfc5322#section-3.3
10. Anti-spam message headers [online]. [cit. 2023-04-15]. Dostupné z: https://learn.microsoft.com/en-us/microsoft-365/security/office-365-security/message-headers-eop-mdo?view=o365-worldwide
11. How DKIM SPF & DMARC work [online]. [cit. 2023-04-15]. Dostupné z: https://www.socinvestigation.com/what-are-spf-dkim-and-dmarc-protection-against-spoofing-and-phishing/

12. Dmarc Alignment [online]. [cit. 2023-04-15]. Dostupné z: https://dmarcian.com/alignment/
13. DUMAN, Sevtap, et al. Emailprofiler: Spearphishing filtering with header and stylometric features of emails. In: 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC). IEEE, 2016. p. 408-416.
14. KULKARNI, Priti; ACHARYA, Haridas. Comparative analysis of classifiers for header based emails classification using supervised learning. International Research Journal of Engineering and Technology, 2016, 3.03: 1939-1944.
15. SMADI, Sami, et al. Detection of phishing emails using data mining algorithms. In: 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA). IEEE, 2015. p. 1-8.
16. PASUPATHEESWARAN, Satheesaan. Email'Message-IDs' helpful for forensic analysis?. 2008.
17. ABU, Md Sahrom, et al. Cyber threat intelligence–issue and challenges. Indonesian Journal of Electrical Engineering and Computer Science, 2018, 10.1: 371-379.
18. CONTI, Mauro; DARGAHI, Tooska; DEHGHANTANHA, Ali. Cyber threat intelligence: challenges and opportunities. Springer International Publishing, 2018.
19. WAGNER, Thomas D., et al. Cyber threat intelligence sharing: Survey and research directions. Computers & Security, 2019, 87: 101589.
20. Shodan [online]. [cit. 2023-04-15]. Dostupné z: https://www.shodan.io
21. AbuseIPDB [online]. [cit. 2023-04-15]. Dostupné z: https://www.abuseipdb.com/
22. Ip-api [online]. [cit. 2023-04-27]. Dostupné z: https://ip-api.com/
23. Ipapi [online]. [cit. 2023-04-27]. Dostupné z: https://ipapi.com/
24. Verifalia [online]. [cit. 2023-04-30]. Dostupné z: https://verifalia.com/developers
25. EmailHippo [online]. [cit. 2023-04-30]. Dostupné z: https://tools.emailhippo.com/
26. Email Dossier [online]. [cit. 2023-05-06]. Dostupné z: https://centralops.net/co/EmailDossier.aspx

27. Pulsedive [online]. [cit. 2023-05-01]. Dostupné z: https://pulsedive.com/
28. URLScan [online]. [cit. 2023-05-03]. Dostupné z: https://urlscan.io/
29. Virustotal [online]. [cit. 2023-05-03]. Dostupné z: https://www.virustotal.com/gui/home/upload
30. Checkphish [online]. [cit. 2023-05-06]. Dostupné z: https://checkphish.ai/
31. WhatIsMyIp [online]. [cit. 2023-05-06]. Dostupné z: https://www.whatismyip.com/
32. Statistics of phishing attacks in 2021 [online]. [cit. 2023-05-10]. Dostupné z: https://www.bleepingcomputer.com/news/security/social-media-phishing-attacks-are-at-an-all-time-high/

Prílohy

Príloha A: Bakalárska práca v elektronickej podobe, zdrojové kódy k anonymizovaniu, extrahovaniu atribútov z e-mailových správ a obohacovaniu pomocou TI v elektronickej podobe