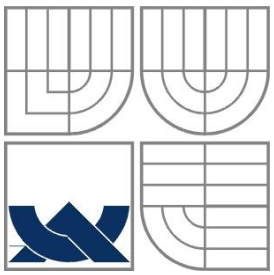
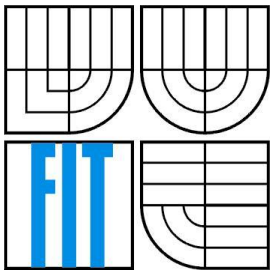


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

DETEKCE ŠKODLIVÝCH DOMÉN POMOCÍ ANALÝZY DNS PROVOZU

MALICIOUS DOMAINS DETECTION USING ANALYSIS OF DNS TRAFFIC

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

EVA AMBRUŠOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR

ING. MICHAL KOVÁČIK

BRNO 2015

Abstrakt

Předmětem této práce je detekce škodlivých domén založena na pasivní analýze DNS provozu. Představuje návrh a implementaci systému, který realizuje detekci DNS anomálií na základě skladby doménového jména. Využívá při tom entropii a frekvenční charakteristiku n-gramů. Systém byl testován na DNS datech získaných z reálného provozu a jeho testováním a analýzou výsledků byla ověřena funkčnost implementovaných detektorů.

Abstract

The aim of this thesis is the detection of malicious domains based on passive analysis of DNS traffic. It represents the design and implementation of a system which proceeds DNS anomaly detection based on a structure of the domain name by using the entropy and a frequency characteristics of n-grams. The system was tested on DNS data obtained from the real traffic and the functionality of implemented detectors was verified by testing and analysis of results.

Klíčová slova

DNS, škodlivá doména, detekce, analýza pasivního DNS provozu

Keywords

DNS, malicious domain, detection, passive DNS analysis

Citace

Ambrušová Eva: Detekce škodlivých domén pomocí analýzy DNS, bakalářská práce, Brno, FIT VUT v Brně, 2015

Detekce škodlivých domén pomocí analýzy DNS provozu

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením pana Ing. Michala Kováčika. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

……
Eva Ambrušová
18. května 2014

Poděkování

Rada by som poďakovala vedúcemu mojej bakalárskej práce Ing. Michalovi Kováčikovi za cenné rady a pripomienky počas tvorby práce a taktiež za poskytnuté materiály.

Obsah

1 Úvod

V súčasnej dobe, kedy internet nadobúda obroského rozmachu a je neoddliteľnou súčasťou života, je hrozba počítačových útokov veľmi aktuálna. Útočníci sa snažia nájsť nedostatky internetu a nelegálnym spôsobom tak získavajú informácie či služby, ktoré by im bežne neboli prístupné.

Veľké množstvo týchto útokov je zamerané na službu DNS a jej servery, pričom počet a sila útokov stále viac a viac narastá. Služba DNS sa používa na preklad IP adresy každého počítača v sieti na jemu prislúchajúce doménové meno, čo je pri používaní internetu kľúčové, keďže pre človeka je jednoduchšie zapamätať si slovo než číselnú kombináciu. Servery DNS sú obeťou útokov najmä z toho dôvodu, že takmer každá internetová komunikácia je založená na princípe dotaz a odpoveď. V praxi to znamená, že klient pošle dotaz na server DNS a ten sa mu snaží poskytnúť najlepšiu možnú odpoveď.

Existuje široké spektrum útokov na službu DNS a takisto aj možností ich odhalenia. Možnými útokomi sú napríklad podvrhnutie záznamov DNS nasadením podvodného serveru, alebo zneužitie protokolu za účelom posielania odlišného typu dát.

Útokmi, ktorých detekciou sa zaoberá táto práca, sú založené na algoritmickom generovaní doménových mien. Princípom je opätovné posielanie dotazov na DNS server, pričom algoritmické generovanie mena zabezpečí jeho častú zmenu a tým sa útočník vyhne blokovaniu prístupu. Škodlivé domény tohto typu, nazývané tiež DGA (*Domain Generation Algorithm*) domény, sa svojou skladbou odlišujú od legitímnych domén, ktorých účelom je jednoduchá zapamätateľnosť a preto sa skladajú z bežne používaných slov. Práve túto vlastnosť je možné využiť na detekciu domén DGA pôvodu.

Táto práca ponúka informácie o princípe a fungovaní služby DNS, poskytuje náhľad na rozdelenie anomálií v DNS a zaoberá sa tiež možnosťami monitorovania sieťovej prevádzky. Jej zameraním je detekcia škodlivých domén na základe skladby doménového mena.

Druhá kapitola popisuje princíp fungovania služby DNS, jeho architektúru a zdroje dát, ktoré sú použiteľné na potrebnú analýzu. Tretia kapitola ponúka náhľad na aktuálne anomálie objavujúce sa v DNS a zameriava sa tiež na škodlivé domény ako také a spôsoby ich detekcie. Štvrtá a piata kapitola predstavuje návrh a implementáciu systému realizujúceho detekciu škodlivých domén. Šiesta kapitola pojednáva o testovaní implementovaného systému a analýze výsledkov. V závere práce sú predostreté návrhy na vylepšenie implementovaného detektoru a možné pokračovanie práce.

2 Systém DNS

Každé sieťové rozhranie v sieti musí disponovať jedinečným identifikátorom, ktorý predstavuje IP adresa. Pre človeka je ľahšie si zapamätať slovo, a nie číslo, a práve tento fakt bol primárnym dôvodom pre vznik služby *Domain Name System* (ďalej DNS), ktorá zabezpečuje preklad IP adries na doménové mená a späť.

Služba DNS obsahuje celosvetovú databázu doménových mien a IP adries, ktoré k nim prislúchajú. Táto databáza je z dôvodu jej rozsahu distribuovaná na viacero počítačov, na ktorých bežia menné servery, nazývané tiež nameservery. IP adresu doménového mena zisťujeme dotazom na server DNS pomocou resolverov, ktoré sa dotazujú na informácie z menných serverov.

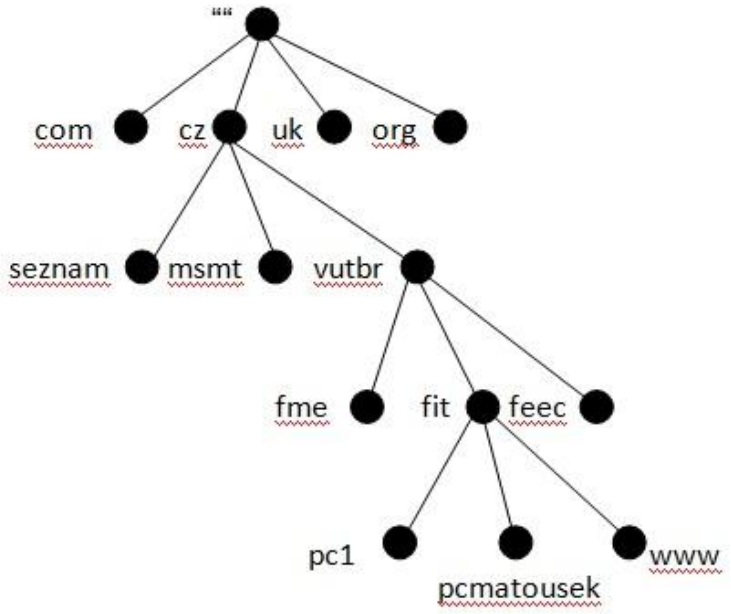
Táto kapitola slúži ako úvod do problematiky služby DNS. Prvá časť popisuje jej architektúru, druhá časť sa zaoberá možnosťami monitorovania sieťovej prevádzky.

2.1 Architektúra DNS

Systém DNS tvoria tri hlavné komponenty – priestor doménových mien, servery DNS a resolver. Nasledujúce podkapitoly obsahujú podrobnejší popis jednotlivých komponent. Uvedené informácie vychádzajú prevažne z [1].

2.1.1 Priestor doménových mien

Z dôvodu efektívneho vyhľadávania sa logický priestor všetkých doménových mien ukladá v systéme DNS pomocou hierarchického usporiadania záznamov do stromu. Z pohľadu algebry sa jedná o acyklický graf, viď Obrázok 1.



Obrázok 1: Hierarchické usporiadanie doménových mien v DNS [1]

Najvyššou doménou je root doména označovaná bodkou. Koreň stromu DNS, root, je pomenovaný textovým reťazcom nulovej dĺžky. Názvy ostatných uzlov sú tvorené textovým reťazcom (bez bodky) a majú dĺžku stanovenú na maximálne 63 znakov. Tieto názvy sú však iba súčasťou doménových mien, a nie samotnými doménami, pretože doména je podstromom v grafe doménových mien. Cesta od listu ku koreňu slúži na uloženie, prípadne vyhľadávanie doménových adries, a z toho vyplýva, že doménové meno je cesta od uzlu ku koreňu stromu.

Doména, ktorá má vrchol v uzle vo vzdialenosti 1 od koreňu grafu, sa nazýva doména prvej úrovne (anglicky *Top Level Domain – TLD*). Doména vo vzdialenosti 2 od koreňu stromu DNS je doménou druhej úrovne, atď. Listy stromu označujú konkrétne sieťové zariadenia.

Úplné (absolútne) doménové meno určitého uzlu (anglicky *Fully Qualified Domain Name – FQDN*) je postupnosť mien uzlov tvoriacich úplnú cestu od listu až ku koreňu, ktoré sú oddelené bodkami (napr. „`www.fit.vutbr.cz.`“). Ukončujúca bodka slúži ako oddeľovač, za ktorým nasleduje meno koreňu DNS stromu tvoreného reťazcom nulovej dĺžky. Pri práci s Internetom obvykle túto bodku pri preklade dopĺňa použitá aplikácia.

DNS je decentralizovaný systém, a teda jednotlivé jeho časti sú uložené na lokálnych serveroch DNS tvoriacich systém DNS. Fyzické časti priestoru DNS pod jednotnou správou sa nazývajú zóny. Zatiaľčo doména je časť priestoru adries so spoločným suffixom, zónu tvoria časti priestoru uložené na konkrétnom serveri.

Na spätné dohľadanie doménovej adresy k IP adrese sa využíva reverzné (spätné) mapovanie. V datovom priestore DNS bola pre tieto účely vytvorená špeciálna doména `in-addr.arpa.`, ktorej uzly sú pomenované číslicami reprezentujúcimi IP adresu v štvorbitovom dekadickom formáte oddelenom bodkami, takteiž zapísané v reverznom tvare (napr. 1.1.168.192). Táto technika je užitočná napríklad v prípade autorizácie počítača, kedy sa kontroluje výskyt záznamu IP adresy stanice v systéme DNS. Systém to môže vyhodnotiť ako použitie podvrhnutej IP adresy v prípade chýbajúceho doménového mena a komunikáciu odmietnuť.

2.1.2 Server DNS

Servery DNS uchovávajú dáta z priestoru doménových mien a ich úlohou je odpovedať na dotazy smerujúce na databázu DNS. Dáta uchovávané vo forme množiny záznamov sú uložené v lokálnom súbore, alebo si ich server načíta z iného DNS serveru pomocou prenosu zón.

Rozlišujeme nasledujúce typy DNS serverov:

- **primárny** – obsahuje úplné záznamy o spravovaných doménach, ktoré sú uložené lokálne v súbore. Pre každú doménu existuje práve jeden primárny server poskytujúci autoritatívne odpovede pre tieto domény.
- **sekundárny** – získava dáta od primárneho serveru pomocou prenosu zónových súborov obsahujúcich databázu konkrétnej domény. Sekundárny server je tiež autoritatívny server pre danú doménu a zaisťuje pravidelný prenos zónových dát a aktualizáciu dát.
- **záložný** – pracuje ako sprostredkovateľ medzi klientom a cieľovým DNS serverom. Umožňuje zrýchliť proces rezolúcie doménového mena, pretože si uchováva odpovede na svoje dotazy a opätovne ich používa. Poskytuje však neautoritatívne odpovede.

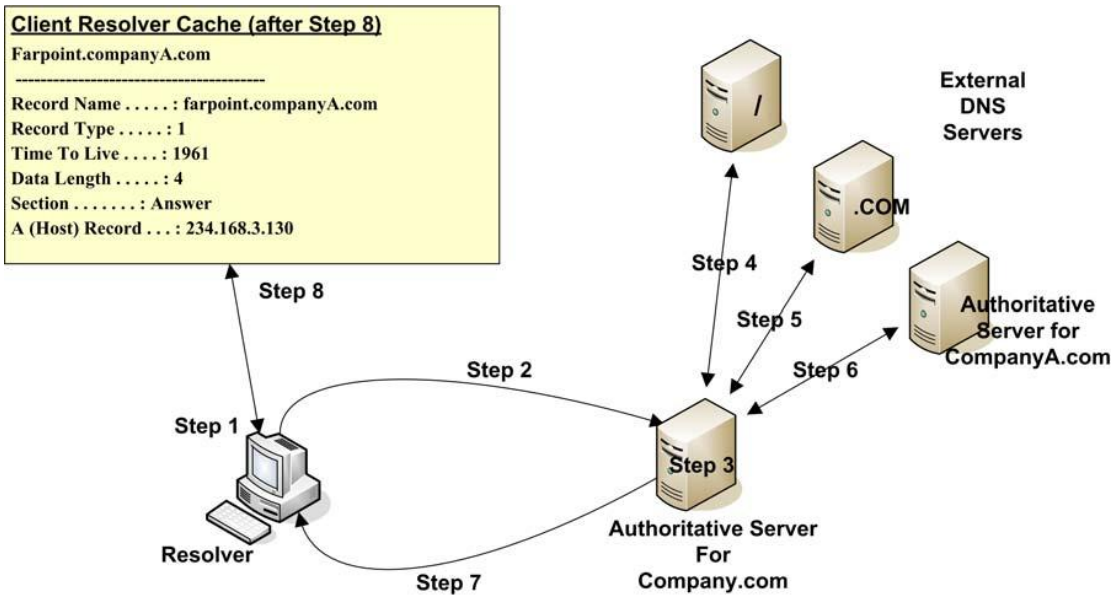
2.1.3 Resolver

Pomocou klientského programu resolver pristupujú užívateľské programy k dátam uloženým v systéme DNS. Základnou úlohou resolveru je zaistiť preklad mena na IP adresu, a to posielaním dotazov serverom DNS, interpretovaním odpovede od nich a poskytovaním týchto dát užívateľskému

programu, ktorý o ne žiadal. Ak dotazovaný DNS server v danom okamihu nepozná odpoveď na žiadosť, vráti odkaz na ďalší server, kde je hľadaná informácia uložená.

Proces hľadania odpovede v systéme DNS, rezolúcia, je realizovaný nasledovne. Server DNS sa spýta koreňového serveru DNS, ktorého adresa mu je známa, na najvyššiu doménu a následne postupuje po stromovej štruktúre od koreňu až k uzlu obsahujúcemu hľadanú informáciu. Koreňový server [5] je autoritatívny server pre všetky domény TLD, ktorý ku každej doméne TLD pozná adresu autoritatívneho serveru. Koreňový server DNS pri dotaze na akékoľvek doménové meno odpovedá adresou serveru DNS, ktorý informáciu obsahuje. Servery DNS pre domény každej úrovne obsahujú informácie o príslušných subdoménach úrovne nižšej o 1. Existuje 13 koreňových serverov rozmiestnených po celom svete.

Servery DNS rozlišujú dva typy dotazov, podľa ktorých sa líši ďalší postup rezolúcie. Prvý typ dotazu, *rekurzívny*, spočíva v tom, že pokiaľ server, ktorý dotaz prijal, nie je autoritatívnym serverom pre danú doménu, musí sa rekurzívne dotazovať ďalších serverov. Ak odpoveď nenájde, vracia chybu. Druhý typ dotazu, *iteratívny*, spočíva v odpovedaní serveru na dotaz najlepšou odpoveďou, ktorú môže poskytnúť. Server sa pozrie do svojej lokálnej databázy a pokiaľ nenájde odpoveď, vráti adresy serverov, ktoré sú najbližšie k hľadanej adrese.



Obrázok 2: Príklad rezolúcie rekurzívneho dotazu [8]

Kvôli redukcii počtu dotazov na systém DNS používa väčšina serverov DNS lokálnu vyrovnávaciu pamäť cache, do ktorej priebežne ukladá odpovede od serverov pre budúce vyhľadávanie. V niektorých prípadoch servery DNS disponujú tzv. negatívnou cache, v ktorej si ukladajú odpovede od autoritatívnych serverov o tom, že hľadaný záznam v danej doméne neexistuje. Existencia pamäti cache zrýchľuje proces vyhľadávania odpovede, odpovede získané z cache však môžu byť neautoritatívne.

2.2 Zdroje sieťových dát

V prípade využitia DNS ako nástroja na detekciu škodlivých domén je dôležité sa zamerať na zdroje dát, ktoré je možné na tento účel použiť. Táto kapitola sa zameriava na základné vlastnosti

protokolov, pomocou ktorých je možné monitorovať sieťovú prevádzku a na základe analýzy získaných dát posudzovať korektnosť navrhovaných metód. Úvod čerpá prevažne z [2].

Prvým spôsobom monitoringu siete bolo použitie protokolu SNMP (*Simple Network Management Protocol*). Na základe štatistík zo základných stavebných prvkov siete je pomocou neho možné zistiť rôzne informácie, napríklad o počte prenesených paketov, bajtov, chýb, a pod. V dnešnej dobe však aj napriek podpore mnohými zariadeniami nemá veľké využitie, pretože poskytuje minimum dát použiteľných pre bezpečnostnú analýzu.

Ďalším, v dnešnej dobe veľmi populárnym riešením, je použitie tzv. flow dát. Pod týmto pojmom sa rozumie združenie paketov zo siete do sieťových tokov, ktoré sú identifikované zdrojovou a cieľovou IP adresou, zdrojovým a cieľovým portom a typom protokolu. S hlavičkou toku sú ukladané najčastejšie informácie o počte prenesených dát, trvaní toku, či príznakoch v prípade TCP komunikácie. Výhodou tohto riešenia je jeho rýchlosť, keďže sú uchovávané iba určité údaje o tokoch.

Metodiky SNMP a flow dát sú však nedostatočné, pretože pomocou nich nie je možné zachytávať tie položky jednotlivých paketov alebo hlavičiek DNS, ktoré sú dôležité pre aplikačný protokol DNS. Ideálnym riešením by bolo zaznamenávať úplne všetky informácie o paketoch, to je však pamäťovo a výpočtovo náročné. Kompromisom medzi monitorovaním tokov a kompletných paketov je použitie protokolov NetFlow alebo IPFIX. Jednou z možností ukladania takto zachytených dát sú súbory typu PCAP.

2.2.1 NetFlow

NetFlow [3] je protokol spoločnosti Cisco vytvorený za účelom monitorovania sieťovej prevádzky, ktorý umožňuje administrátorom sledovať dáta zo siete v reálnom čase. V súčasnosti je to veľmi rozšírený protokol pre monitorovanie sietí.

Jeho prvou verziou bola masovo používaná verzia 5. Verzia 6 bola oproti nej rozšírená o podporu tunelovanej prevádzky a verzia 7 naviac disponuje informáciami o switchoch. Dnes je veľmi využívaná verzia 9, ktorá zaviedla väčšiu flexibilitu pomocou šablón a okrem položiek z verzie 5 dovoľuje prenášať napríklad aj IPv6 adresy s portami. Všetky uvedené verzie protokolu NetFlow sú Cisco proprietárne a nikdy neboli štandardom.

Sieťový tok protokolu NetFlow verzie 5 je sekvencia paketov identifikovaná nasledovnými údajmi:

- zdrojová a cieľová adresa
- zdrojový a cieľový port
- protokol za stanovený interval času
- rozhranie, na ktorom bol tok zachytený
- typ služby (Type of Service)

Pri sieťových tokoch je možné rozlišovať smer komunikácie, preto pre jedno spojenie typu klient-server existujú dva toky. Okrem uvedených položiek sú pre tok zaznamenávané aj počty prenesených paketov a bajtov v toku, časové značky začiatku a konca toku a v prípade TCP protokolu aj nastavené príznaky.

Sledovaniu tokov je nutné prispôsobiť aj vnútornú štruktúru siete. Tento účel plnia dva hlavné prvky technológie NetFlow, a to kolektor a jeden alebo viac exportérov, ktoré sú nasadené do siete.

Exportér je zodpovedný za samotné vytváranie záznamov o tokoch z monitorovaných IP paketov, prípadne za ich aktualizáciu v NetFlow cache, ktoré ďalej posiela kolektoru. Kolektor zbiera a ukladá zaznamenané dáta, ktoré je následne možné analyzovať. Prácu exportérov buď vykonávajú routre ako svoju druhotnú funkciu, alebo sú na ňu využívané špeciálne monitorovacie zariadenia nazývané sondy.

2.2.2 IPFIX

Na rozdiel od NetFlow je IPFIX (*Internet Protocol Flow Information eXport*) [4] štandardizovaný IETF (*Internet Engineering Task Force*) protokol, ktorý vychádza z protokou NetFlow v9. Je značne flexibilný, keďže nemá pevne danú štruktúru prenášaných dát. Dovoľuje definovať nové položky pre prenos, čím umožňuje prenášať potrebné informácie, a tiež podporuje možnosť prenosu premenných dĺžok polí (napr. URL, doménové meno).

Informačné prvky informačného modelu IPFIX sú rozdelené do 12 skupín podľa ich sémantiky a použiteľnosti. Na mapovanie paketov do sieťového toku protokolu IPFIX môžu slúžiť informácie z týchto prvkov:

- IP hlavička
- transportná hlavička
- sub-IP hlavička
- ďalšie vlastnosti odvodzované z paketov

Pokiaľ tieto informácie neslúžia ako identifikátory toku, môžu sa ich hodnoty v jednotlivých paketoch vrámci toku líšiť. Pre také časti paketov platí, že ich hodnota je určená prvým zachyteným paketom korešpondujúceho toku, s výnimkou situácie, kedy popis tohto informačného prvku explicitne nešpecifikuje inú sémantiku. Toto pravidlo umožňuje zápis všetkých údajov o toku týkajúcich sa hlavičkových polí paketu len jeden krát a z nasledujúcich paketov sú aktualizované iba informácie odvodzované z viac než jedného paketu (napr. minimá/maximá toku, časové značky toku, a pod.).

2.2.3 PCAP

PCAP (*packet capture*) [15] predstavuje rozhranie na zachytávanie paketov počas monitorovania sieťovej prevádzky. Súbory typu PCAP uchovávajú kompletné informácie o každom pakete a sú veľmi rozšírené pri realizovaní pasívnej analýzy zachyteného sieťového toku.

V systémoch typu Unix je s využitím knižnice `libpcap` možné implementovať aplikácie schopné zachytiť a analyzovať sieťovú prevádzku, alebo prečítať už zachytenú prevádzku a analyzovať ju. Dostupným nástrojom na čítanie a analyzovanie záznamov typu PCAP je Tcpdump pre operačný systém Linux alebo Wireshark pre operačný systém Windows.

Typickou príponou súborov tohto typu je `.pcap`, `.cap` či `.dmp`.

3 Anomálie v DNS

Pod pojmom anomálie v DNS rozumieme situáciu, kedy je DNS hlavným činiteľom či cieľom útoku, prípadne je zneužitá niektorá z jeho vlastností, čo robí útok niekoľkonásobne silnejším.

Potreba riešiť bezpečnosť internetových služieb prišla až s rozmachom počítačových sietí a so zvyšujúcim sa počtom užívateľov. Služba DNS sa stáva, rovnako ako ostatné protokoly a služby Internetu, záujmom hackerov, ktorých cieľom je využíť každú jej nedokonalosť vo svoj prospech. Možnosť zneužitia DNS sa stupňuje tým, že ide o otvorený protokol, ktorý nepoužíva žiadne mechanizmy zabezpečenia pre overenie pravosti dát zo serveru. Systém DNS dôveruje DNS odpovedi na základe zdrojovej adresy, portu a transakčného ID, a tým môže jednoducho dôjsť k doručeniu nepravej odpovede.

Táto kapitola popisuje anomálie v DNS rozdelené na dve skupiny podľa [6]. Najskôr bude uvedený prehľad o útokoch, ktoých priamym cieľom nie je DNS a následne budú popísané útoky, ktoré sú zamerané priamo na DNS servery. Posledná podkapitola obsahuje prehľad o problematike škodlivých domén a ich detekcii.

3.1 Útoky necielené priamo na DNS servery

Útoky zneužívajúce DNS servery sa snažia získať a zneužiť administratívne informácie, prípadne nepozornosť užívateľov, a neberú si za cieľ samotné DNS servery.

Jedným zo spôsobov, ako takýto útok realizovať, je tzv. okupovanie doménového priestoru so zámerom jeho zneužitia na nekalé obchodné praktiky. Útočník si úmyselne registruje také doménové meno, ktorého názov sa veľmi podobá známej doméne. U užívateľa je v takomto prípade veľká pravdepodobnosť, že adresy omylom zamení. Cieľom tejto praktiky je buď získanie väčšieho počtu klientov (ak sa jedná napr. o e-shop), alebo profit útočníka na predaji takto zaregistrovanej domény vlastníkovi značky za oveľa vyššiu cenu.

Ďalším zo spôsobov je napríklad registrovanie domény, ktorej názov je podobizňou existujúcej domény, prípadne názvom so zámernou typografickou chybou. Na alternatívnu doménu sa užívateľ obvykle dostane vďaka svojej nepozornosti. Tú má útočník registrovanú za účelom reklamy, distribúcie škodlivého softvéru, či za účelom realizovať phishing[1]. Táto praktika sa najčastejšie realizuje zneužitím pravopisných chýb v doméne (napr. www.gogle.com), typografických chýb (napr. www.butbr.cz) alebo chyby v TLD (napr. www.fit.vutbr.com).

Existuje niekoľko predstaviteľov útokov tohto typu, dvomi najvýznamnejšími sú:

- **DNS Reflection** – jedná sa o útok založený na DoS (*Denial-of-Service*), čo značí odopretie služby. Pri tomto útoku je využitých viacero počítačov vo forme prostredníkov pre útok, s ktorých narastajúcim počtom sa znižuje pravdepodobnosť odhalenia útočníka, keďže pri útoku sa strieda veľké množstvo IP adries. Útočník používa podvrhnutú zdrojovú IP adresu tak, aby boli vygenerované niekoľkonásobne väčšie odpovede, ktoré sú následne doručené na adresu obeti.
- **DNS Tunneling** – jedná sa o zneužitie služby DNS, ktorého podstatou je zapúzdrenie dát do klasickej DNS prevádzky, ktorá väčšinou nebýva nijakým spôsobom obmedzovaná. DNS pakety môžu byť použité na vytvorenie skrytého komunikačného kanálu nazývaného DNS tunel. Ten je možné použiť pre prenos dát bez toho, aby reagoval firewall.

[1] phishing – činnosť, pri ktorej sa podvodník snaží vylákať od používateľov rôzne citlivé údaje, typicky heslá

3.2 Útoky cielené priamo na DNS servery

Cieľom útokov popisovaných v tejto časti sú DNS servery. Ide o poškodenie DNS serveru úpravou jeho záznamov, čo môže útočník zneužiť bez vedomia obete, alebo o samotné odoprenie prístupu k serveru.

Známymi útokmi tohto typu sú:

- **Denial-of-Service útok** – najzákladnejší útok na DNS server. Útok tohto typu však vo veľkej miere neovplyvní správanie DNS serveru a často je len sprievodným útokom alebo súčasťou zložitejších útokov. Útočník sa pomocou nich snaží zneprístupniť, alebo aspoň veľmi spomaliť prístup k určitej on-line službe. Pokiaľ sa na útoku podieľa väčšia skupina počítačov, jedná sa o DDoS útok (*Distributed Denial of Service*). Tieto počítače sú väčšinou napadnuté malwarom a stávajú sa súčasťou botnetu[2].
- **Rekurzívny útok** – špeciálny prípad DDoS útoku. Hlavnou myšlienkou je zasielanie dotazov na neexistujúce doménové mená. Tieto doménové mená sú často náhodne generované a keďže sa nemôžu nachádzať v cache pamäti lokálneho serveru, musí dotaz zakaždým spracovať autoritatívny server. To spôsobí zaslanie množstva nevyžiadaných dotazov na server, ktorý sa stal obeťou útoku.
- **Cache poisoning** – ide o zmenu obsahu cache pamäti serveru za účelom zmeny pôvodne korektného mapovania doménových mien na IP adresy. Dôvodom je absencia akéhokoľvek zabezpečenia DNS záznamov a overovania zdroja odpovede, čo umožňuje útočníkovi presmerovať dotazy na server, ktorý je v plnej moci útočníka. Ďalším dôvodom úspešnosti útoku cache poisoning je fakt, že neexistuje spoľahlivý spôsob určenia legitímnosti odpovede. Získané informácie útočník využíva na nelegálne účely (napr. phishing).
- **Buffer overflow** - ide o útok využívajúci existenciu chýb v DNS softvéri. Jeho realizácia je možná v prípade, že softvér nesprávne zaobchádza s väčším obsahom pamäte než je možné, a tak dôjde k pretečeniu dát mimo alokovanú pamäť. Útočník môže vykonať tento typ útoku dvomi spôsobmi, a to zasielaním dotazov na DNS server alebo použitím autoritatívneho DNS serveru.
- **Útok skenovaním portov** – je typ útoku, ktorý okrem DNS serverov cieli často aj na iné služby. V prípade, že nevyvolá žiadnu reakciu, je veľmi jednoduché ho odhaliť, avšak v opačnom prípade je náročné identifikovať, že sa jedná o tento typ útoku. Útok spočíva v posielaní správ na rôzne porty a podľa ich otvorenia využíva útočník známe exploity[3] pre prístup do inak nedostupnej služby.

3.3 Škodlivé domény v DNS

V počítačových sieťach sa často vyskytujú aj anomálie založené na zneužívaní samotných doménových mien. Tieto anomálie zahŕňajú napríklad použitie domény ako webovej stránky sťahujúcej škodlivý kód do počítača, použitie domény na phishing, a pod.

Táto kapitola ponúka popis techniky fast-flux, spôsobujúcej ťažšie blokovanie škodlivých domén v sieti, a náhľad na možnosti detekcie škodlivých domén vychádzajúci z [2].

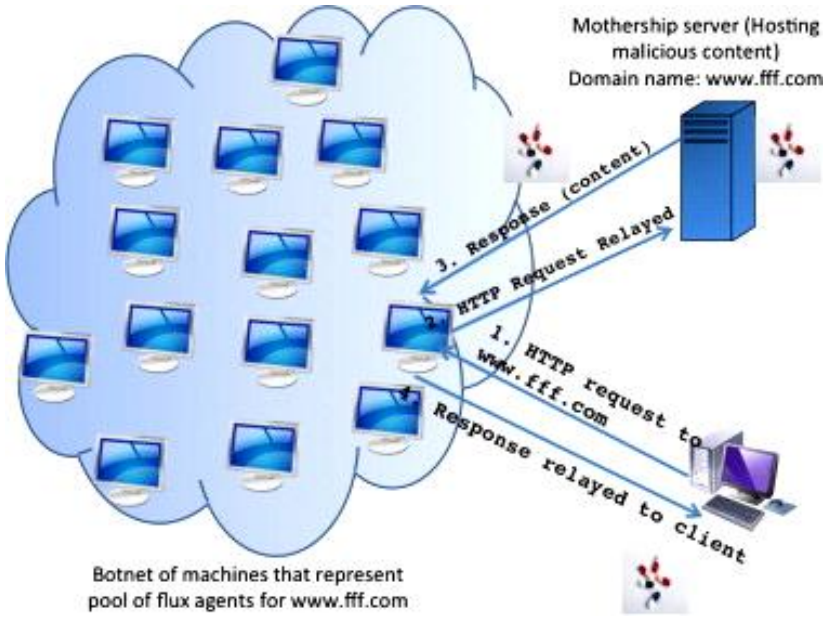
[2] botnet – sieť počítačov infikovaných malware-om riadená z jedného centra hackerom
[3] exploit – program (príp. sekvencia príkazov) využívajúci programátorskú chybu, ktorá spôsobí pôvodne nezamýšľanú činnosť software

3.3.1 Technika fast-flux

Fast-flux je technika využívaná botnetmi zameraná na sťaženie blokovania škodlivých domén. Je to technika využívaná botnetmi na maskovanie phishingu a doručovania malware. Spočíva v priradení niekoľkých IP adries k jednému FQND domény, ktoré zvyšuje náročnosť lokalizovania škodlivých domén a ich následného blokovania.

Rozlišujeme dve varianty techniky fast-flux. Princíp *single fast-flux* spočíva v neustálom menení DNS A[4] záznamov pre jedno doménové meno. Zahŕňa to výskyt DNS záznamov s veľmi nízkou hodnotou TTL, po ktorú je IP adresa priradená určitej doméne a uchovávaná v cache menného serveru. Obvyklá hodnota je niekoľko hodín, prípadne dní, pričom upravená hodnota je nastavená len na niekoľko minút. Tým je vytvorený neustále sa meniaci zoznam cieľových IP adries pre konkrétne doménové meno, ktorý môže obsahovať stovky až tisícky záznamov.

Vylepšením je technika *double fast-flux*, ktorej sofistikovanosť spočíva v tom, že na jej odstavenie nestačí len odstránenie menného serveru škodlivej domény, ako je tomu u single fast-flux. Povoľuje totiž dynamické zmeny vo vstupnej zóne menného serveru, čo prináša možnosť pre viacero rôznych IP adries stať sa autoritatívnym serverom. To predstavuje ďalšiu vrstvu zabezpečujúcu odolnosť škodlivých serverov.



Obrázok 3: Základná myšlienka techniky fast-flux [7]

3.3.2 Spôsoby detekcie

Existuje niekoľko spôsobov detekcie škodlivých domén. Pri detekcii na základe analýzy NetFlow dát bolo zistené, že NetFlow dáta obsahujú veľmi limitované množstvo informácií, na rozdiel od paketov s plným obsahom, a pre detekciu sú nepostačujúce. Je však možné pre podobný účel použiť IPFIX, ktorý obsahuje kľúčové položky ako informácie o TTL, dotazovaných doménových menách, a pod.

[4] DNS A záznam (address record) – záznam obsahujúci IPv4 adresu priradenú danému menu

Ďalším spôsobom je sledovanie domén, na ktoré chodí abnormálny alebo koncentrovaný počet dotazov, prípadne sledovanie dotazov posielaných na neexistujúce doménové mená, ktoré by mohli počítačom v botnete napomáhať ku kontaktovaniu svojho C&C[5] serveru.

Detekciu je tiež možné realizovať pasívnou analýzou zachytených dát so zameraním na dotazované doménové meno, typ záznamu, dáta odpovede, TTL a časové razítko prvého objavenia sa záznamu. Na týchto dátach je následne realizovaný manuálny rozbor, na základe ktorého je možné odhaliť domény využívajúce typografickú chybu vo svojom názve za účelom presmerovania klienta na škodlivý server.

Domény, ktoré sa podieľajú na akejkoľvek škodlivej činnosti, je možné detekovať tiež na základe sledovania 15 príznakov extrahovaných z DNS prevádzky. Tie sú rozdelené do 4 rôznych skupín nasledovne:

1) Časové príznaky:
   a) krátky život
   b) denná podobnosť
   c) opakované vzory
   d) pomer prístupov
2) Príznaky DNS odpovedí:
   a) počet odlišných IP adries
   b) počet odlišných krajín
   c) počet domén s rovnakou IP adresou
   d) výsledok reverzného DNS dotazu
3) Príznaky hodnoty TTL:
   a) priemerná hodnota TTL
   b) štandardná odchýlka od TTL
   c) počet odlišných hodnôt TTL
   d) počet zmien hodnoty TTL
   e) percentuálne zastúpenie hodnôt TTL
4) Príznaky doménového mena:
   a) percentuálne zastúpenie numerických znakov
   b) percentuálna dĺžka LMS[6]

Na základe sledovaných časových príznakov je možné realizovať detekciu v dvoch fázach, a to globálne, kedy sa určia domény s krátkou dobou života, a lokálne, kedy sa analyzuje správanie domén počas ich života. Príznaky súvisiace s obsahom DNS odpovedí sú sledované hlavne z dôvodu priradenia niekoľkých IP adries jednej doméne. To sa deje kvôli rozdeleniu záťaže na väčší počet strojov, no môže to byť zneužité škodlivými doménami na zvýšenie ich životnosti. Príznaky týkajúce sa hodnoty TTL sú zaznamenávané primárne z dôvodu výrazne nižšej hodnoty TTL, pokiaľ sa jedná o systém zameriavajúci sa na svoju vysokú dostupnosť. Škodlivé domény využívajú túto techniku na zníženie šance na ich výskyt v DNS blackliste[7]. Posledná skupina príznakov týkajúca sa vlastností doménového mena sa zaoberá doménou druhej úrovne. Na základe vyvodenia záveru z percentuálneho zastúpenia číslic v doménovom mene bolo vyvolaných veľa falošných poplachov o jej škodlivosti, preto sa táto technika spresňuje vyhľadávaním najdlhšieho zmysluplného reťazca v doménovom mene.

[5] Command & Control server – centralizovaný počítač posielajúci príkazy botnetu a prijímajúci odpovede od napadnutých počítačov
[6] LMS – najdlhší zmysluplný reťazec (Longest meaningful string)
[7] DNS blacklist – zoznam potenciálne nebezpečných DNS domén

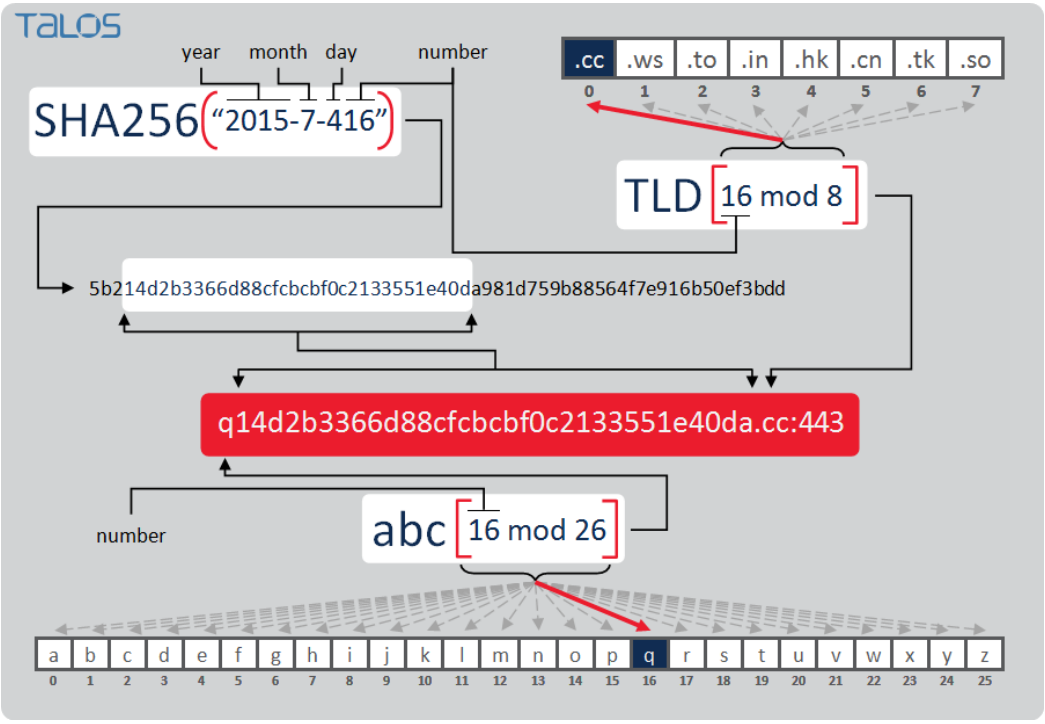
4 Návrh aplikácie

Táto kapitola popisuje návrh systému realizujúceho detekciu škodlivých domén zloženého z niekoľkých detektorov. Na začiatku poskytuje prehľad o príčine vzniku škodlivých domén a ich podobe. Ďalej znázorňuje štruktúru systému spracovávajúceho zachytenú DNS prevádzku, ktorá je následne podrobená analýze jednotlivými detektormi. Detektory môžu pracovať spolu alebo oddelene, podľa parametrov zadaných užívateľom. V prípade detekcie škodlivej domény je jej meno zaznamenané do výstupného textového súboru, spolu s dôvodom, prečo je považovaná za škodlivú. Výstupom sú tiež grafy znázorňujúce vlastnosti analyzovaných domén.

4.1 Detekcia škodlivých domén

V súčasnej dobe sú doménové mená vo veľkej miere zneužívané na škodlivé účely. Jednou z možných techník zneužívania je technika fast-flux, ktorá je používaná botnetmi tak, že každá časť botnetu algoritmicky generuje obrovské množstvo doménových mien a dotazov, až kým aspoň jeden z nich nie je rozpoznaný. Následne je kontaktovaná príslušná IP adresa, ktorá je ďalej použitá ako C&C server. Detailnejší rozbor tejto problematiky ponúka kapitola 3.3.1.

Takéto doménové mená sa nazývajú tiež DGA (*Domain Generation Algorithm*) domény. Zámerom algoritmu DGA je znemožniť detekciu škodlivosti vytvorenej domény na základe jej výskytu v určitom blackliste, keďže je unikátna a jej skladba sa mení často. Obrázok 4 predstavuje možnú podobu tvorby DGA domény, ktorá je schopná denne vytvoriť až 333 rôznych doménových mien. Vstupom tohto algoritmu je aktuálny dátum a číslo, na základe ktorého sa vyberie jedna z ôsmych TLD.



Obrázok 4: Príklad tvorby DGA domény [14]

Keďže sa jedná o algoritmicky generované doménové mená, je zrejmé, že z hľadiska svojej znakovej skladby majú škodlivé domény iné vlastnosti než legitímne, ktoré sú prevažne tvorené slovami bežného jazyka. Na detekciu škodlivosti takýchto domén je preto možné využiť techniku skúmajúcu znakovú skladbu doménových mien.

Táto práca sa zaoberá dvomi podobnými technikami. Jedna z nich je založená na entropii merajúcej neusporiadanosť hlások v doménovom mene, ďalšia využíva frekvenčnú analýzu n-gramov, ktorá skúma prítomnosť reťazcov typických pre konkrétny jazyk v doménovom mene. Nasledujúci príklad zobrazuje možnú podobu DGA domény.

Príklad: Doména DGA pôvodu.

`ripo3psiaz3tvfm3hyyuzzg4ek2wv65nsvvyun.4dbr5vphjdhdsrxbqaaaaaaa.aaaad4mkm4tocfnlkaaaaabcnjngczer3p7soyozczishiavs3qxviwyu3unnl`

4.2 Druhy sledovaných anomálií

Detekcia škodlivých domén je v tejto práci realizovaná predovšetkým analýzou entropie a n-gramov doménového mena. Systém je zložený z troch hlavných detektorov, ktorými analyzované doménové meno prechádza sekvenčne. Pozitívny či negatívny výsledok jedného detektoru nemá vplyv na realizáciu analýzy v nasledujúcom detektore.

Možnosťou redukcie analyzovaných dát je vylúčiť z analýzy tie domény, ktoré škodlivé nie sú. Zoznam takýchto domén, tzv. whitelist, je možné získať pomocou služby Alexa Top 500 Global Sites [11], ktorá eviduje zoznam najpopulárnejších domén. Na implementáciu a testovanie detektorov je použitý zoznam obsahujúci 1 000 000 celosvetovo najpopulárnejších doménových mien.

Keďže v bežnej prevádzke sa vyskytujú doménové mená zložené z niekoľkých úrovní, pre správne fungovanie aplikácie je nutné doménové meno rozdeliť na doménové úrovne (viď Tabuľka 1) a s nimi pracovať zvlášť. Subdomény zložené z troch a menej znakov nenesú dostatočnú informáciu, na základe ktorej by bolo možné určenie jej legálneho alebo DGA pôvodu, preto sú z analýzy vyradené.

Ak je čo len v jednej zo subdomén detekovaná určitá anomália, je celé doménové meno označené za škodlivé.

Príklad: Rozdelenie doménového mena zachyteného z reálnej prevádzky na jednotlivé subdomény a príznak zaradenia do analýzy na základe dĺžky.

Doménové meno: `mlocate.spotlife.net.intern.fahrschule-moritz.at`

Subdoména	Zaradená do analýzy
mlocate	áno
spotlife	áno
net	nie
intern	áno
fahrschule-moritz	áno
at	nie

Tabuľka 1: Rozdelenie doménového mena na analýzu

4.2.1 Detekcia na základe entropie

Jedným zo spôsobov, ktorým je v tejto práci odhaľovaná škodlivosť domén v dátovom toku, je hodnota entropie doménového mena. Entropia je veličina popisujúca stupeň neusporiadanosti v množine dát, v tomto prípade znakov v doménovom mene. Väčšia neusporiadanosť značí vyššiu hodnotu entropie.

Legitímne doménové mená sú väčšinou zložené zo zmysluplných slov, keďže účelom ich existencie je jednoduchšia zapamätateľnosť pre človeka. Doménové mená generované strojom majú vyššiu hodnotu entropie a skladajú sa z neusporiadanejšej množiny znakov a preto podobné prípady môžu indikovať DNS anomáliu. [9]

Táto metóda pravdaže nemôže s úplnou istotou určiť škodlivosť domény. Existujú legitímne doménové mená, ktoré disponujú vysokou hodnotou entropie z dôvodu, že sú tvorené viacerými zmysluplnými slovami oddelenými pomlčkou alebo neoddelenými ničím, viď Tabuľka 2. Obsahujú teda veľký počet rozličných znakov a tým hodnota ich entropie narastá. Keďže hodnota entropie rastie aj s narastajúcou dĺžkou doménového mena, nedá sa táto metóda považovať za úplne spoľahlivú ani v prípade doménových mien s nadpriemernou dĺžkou. V podobných situáciách dochádza k `false positive` detekcii a tento problém je možné eliminovať napríklad pridaním daného doménového mena do whitelistu. Detailnejší rozbor podobných prípadov je predostretý v kapitole 6.

Doména	Entropia
kinotipfilmykeshlednutizdarma	4.03039
elmundialdefutbolbrasil2014	4.01415
sportmarketing2015	3.9477
pujcky-pro-pravnicke-osoby	3.84411

Tabuľka 2: Legitímne domény s vysokou hodnotou entropie

4.2.2 Detekcia na základe zhody n-gramov

N-gramová frekvenčná charakteristika je podľa [10] taktiež použiteľná na detekciu DNS anomálií, napr. DNS tunelov. Tak, ako v prirodzenom jazyku, aj domény a subdomény zachytené v DNS dotazoch a odpovediach obsahujú prirodzenejšie sekvencie hlások, než doménové mená vygenerované útočníkom. Škodlivé domény majú okrem značne vyššej hodnoty entropie taktiež obrovskú nesúmernosť frekvencie n-gramov v porovnaní s typickou doménou.

N-gram je definovaný ako sled n po sebe nasledujúcich položiek z určitej sekvencie. V tejto práci sa za n-gram považuje postupnosť 3 alebo 4 po sebe nasledujúcich hlások z doménového mena.

Detekcia spočíva v zaznamenávaní výskytu všetkých unikátnych n-gramov nájdených v predloženom zozname legitímnej prevádzky a výpočte jeho frekvenčnej charakteristiky, podľa čoho je tento n-gram ohodnotený. Analyzovanému doménovému menu sa následne priradí atribút predstavujúci súčet frekvenčných charakteristík všetkých n-gramov z daného doménového mena v legitímnej prevádzke, na základe ktorého je možné určiť, či sa jedná o legitímnu alebo škodlivú doménu. S nižšou zhodou n-gramov s n-gramami z databázy legitímnych doménových mien rastie pravdepodobnosť, že skúmaná doména je škodlivá.

Frekvenčná charakteristika n-gramu je počítaná nasledovne:

$$p_{n-gram} = \frac{x}{l} \cdot 100$$

Kde:

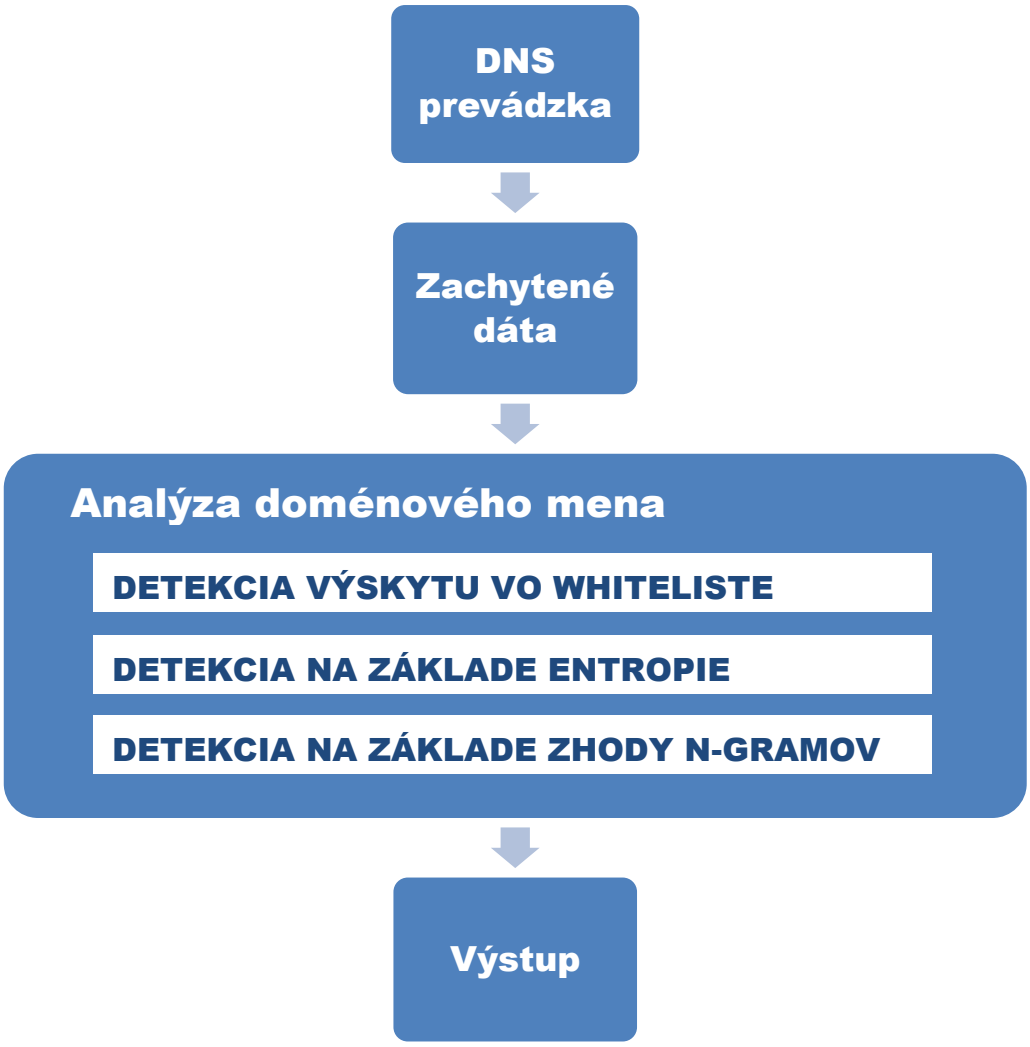
p_{n-gram} je frekvenčná charakteristika n-gramu predstavujúca pravdepodobnosť výskytu n-gramu v zozname

x je počet výskytov daného n-gramu v zozname

l je počet subdomén dlhších ako 3 z predloženého zoznamu

4.3 Architektúra systému

Obrázok 4 znázorňuje architektúru systému vyvinutého na detekciu. Vstupom sú doménové mená zachytené z DNS prevádzky. Jednotlivé detektory na základe hranice určenej experimentovaním označia analyzované doménové meno za legitímne alebo škodlivé. Výstupom je následne textový súbor a grafy, ktoré poskytujú informácie o zachytených škodlivých doménových menách.



Obrázok 5: Architektúra systému

4.4 Získavanie zdrojových dát

Ako je uvedené v kapitole 2.2, protokol IPFIX ponúka možnosť efektívneho monitorovania DNS prevádzky. Jeho výhodou je, že umožňuje analyzovať len vybrané položky aplikačných protokolov a teda nie je nutné, aby zaznamenával celé pakety. Tým sa znižujú jeho pamäťové nároky a rovnako aj nároky na výpočtový výkon.

Zdrojom vstupných dát do tejto práce sú IPFIX dáta z DNS pluginu pre Flowmon Exportér od spoločnosti INVEA [12], ktorý poskytuje možnosť spracovávať DNS prevádzku analyzovaním a následným extrahovaním niektorých častí DNS paketu z jeho aplikačnej vrstvy. Ide o časti reprezentujúce informácie podstatné pre sieťovú analýzu prevádzky, ako napr. transakčné ID, celkový počet DNS odpovedí, návratový kód, typ dotazu, či dotazované doménové meno. Práve posledná zmienená položka bola zo zachytených dát používaná ako vstup pre vyvíjaný detektor škodlivých domén.

Doménové mená z DNS pluginu sú z pamäťových dôvodov ukladané len po 128 znakov, preto je aj v implementácii tohto detektoru realizované rovnaké obmedzenie. Predpokladom však je, že počet paketov, ktoré obsahujú dlhšie doménové meno, je z hľadiska legitímnej prevádzky mizivý a preto toto obmedzenie nemá viditeľný dopad na efektívnosť detekcie.

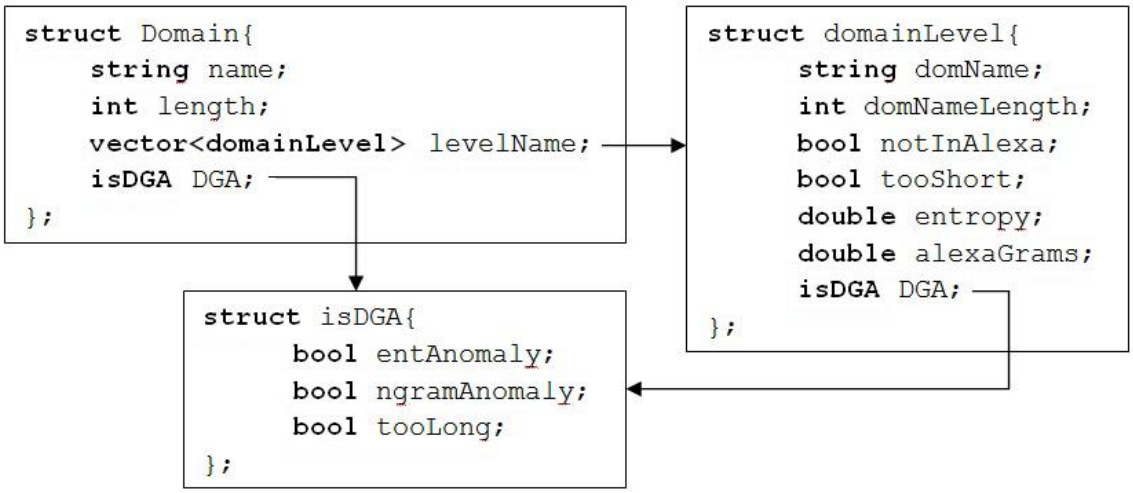
5 Implementácia

Implementácia systému vychádza z návrhu uvedenom v predchádzajúcej kapitole. Aplikácia bola implementovaná v programovacom jazyku C++. Najprv bolo nutné spracovať vstupné dáta vhodným spôsobom a následne implementovať navrhnuté detektory. Detailnejší rozbor týchto krokov popisujú nasledujúce podkapitoly.

5.1 Vstupné dáta

Vstupom aplikácie je textový súbor obsahujúci na každom riadku jedno doménové meno ohraničené úvodzovkami (napr. "`www.youtube.com`"). Vstupné dáta použité na testovanie aplikácie predstavujú zoznam zachytených doménových mien zo siete CESNETu a boli poskytnuté vedúcim bakalárskej práce.

Na ukladanie vlastností analyzovaného doménového mena boli vytvorené tri spolu súvisiace dátové štruktúry, ktoré znázorňuje Obrázok 6.



Obrázok 6: Štruktúry určené na ukladanie informácií o doménovom mene

Štruktúra `Domain` ukladá nasledovné informácie:

- `name` – reťazec predstavujúci celé doménové meno zbavené ohraničujúcich úvodzoviek
- `length` – dĺžka reťazca `name`
- `levelName` – subdomény uložené v dátovom kontajneri `vector`
- `DGA` – príznak škodlivosti

Štruktúra `domainLevel` ukladá tieto detailné informácie o každej subdoméne:

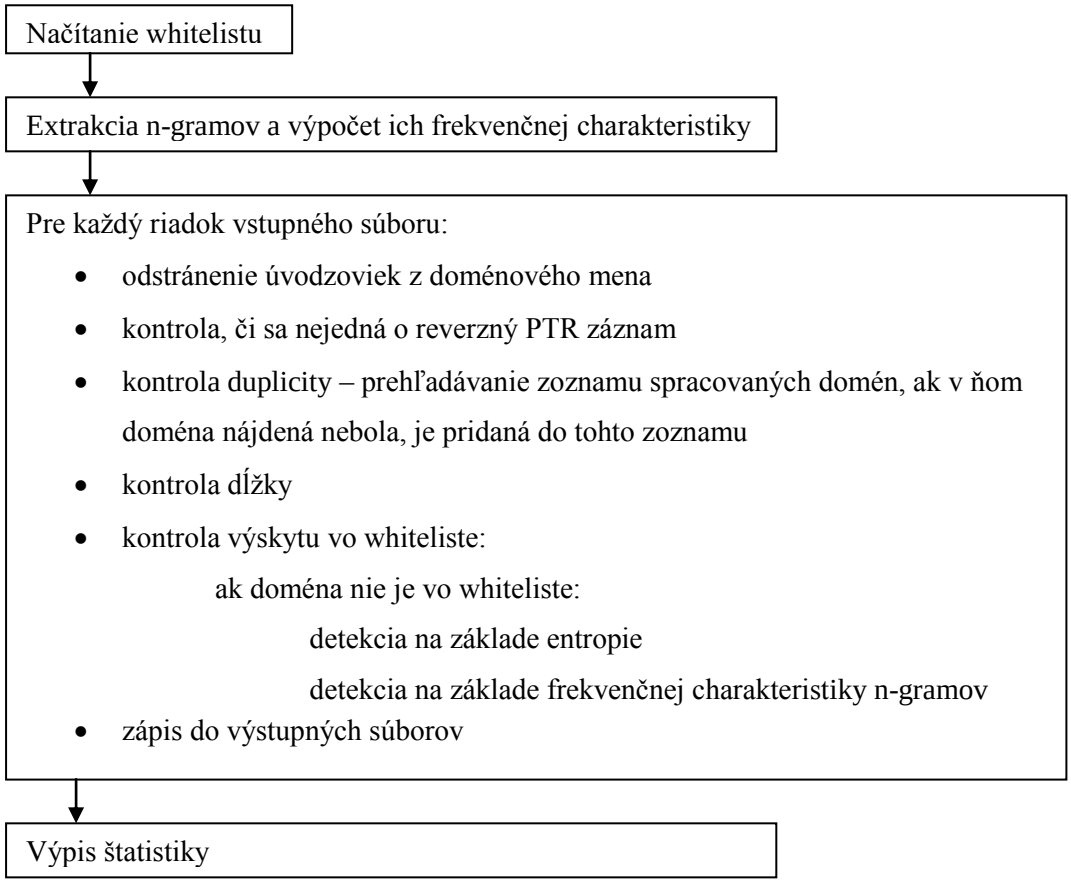
- `domName` – meno subdomény
- `domNameLength` – dĺžka reťazca `domName`
- `notInAlexa` – príznak výskytu vo whiteliste, nastavený na `true` ak sa doména vo whiteliste nenachádza
- `tooShort` – príznak, ktorý je nastavený na true, ak je dĺžka domény menej ako 3
- `entropy` – hodnota entropie subdomény

- `alexaGrams` – súčet frekvenčnej charakteristiky všetkých n-gramov nájdených v subdoméne
- `DGA` - príznak škodlivosti

Štruktúra `isDGA` bola vytvorená na uloženie prípadného dôvodu označenia doménového mena za škodlivé. Ak je jeden z týchto príznakov nastavený na true aspoň u jednej zo subdomén, je tento príznak nastavený na `true` aj u celého doménového mena. Obsahuje tri položky:

- `entAnomaly` – `true`, ak bola zistená anomália na základe hodnoty entropie
- `ngramAnomaly` – `true`, ak bola zistená anomália na základe frekvenčnej charakteristiky n-gramov
- `tooLong` – `true`, ak dĺžka doménového mena prekračuje 128 znakov

Spracovanie vstupných dát predstavuje funkcia `analyse` a je realizovaná nasledujúcim algoritmom, ktorý zobrazuje základné jadro programu pri spustení so všetkými prípustnými parametrami:



Obrázok 7: Spracovanie vstupných dát

Počas načítavania vstupných dát sa z každého spracovávaného reťazca na vstupe odstránia ohraničujúce úvodzovky. Následne sa inicializuje štruktúra `Domain` predstavujúca jedno analyzované doménové meno hodnotami získanými zo vstupného reťazca.

Keďže vstupné dáta neobsahujú informáciu o type DNS záznamu, v doménovom mene sa môžu okrem obyčajných prekladov A a AAAA záznamov objavovať aj iné položky. Príkladom je dotaz na `in-addr.arpa` doménu, čo je reverzný PTR záznam zaisťujúci preklad IP adresy na doménové meno. Podobné dáta boli teda z analýzy vylúčené porovnaním doménového mena najvyššej úrovne na zhodu s reťazcom „`arpa`" a doménového mena druhej úrovne na zhodu s reťazcom „`in-addr`".

Následne môže prebehnúť kontrola, či je dané doménové meno v rámci vstupných dát unikátne a nebolo už analyzované. Kontrola je nepovinná a jej realizácia môže byť parametricky doplnená užívateľom.

Potom je každé doménové meno podrobené analýze jednotlivými detektormi, ktorej výsledky sú uložené do výstupných súborov.

5.2 Implementované detektory

Táto kapitola poskytuje detailnejší náhľad na implementáciu jednotlivých detektorov.

5.2.1 Detektor výskytu vo whiteliste

Prvým krokom na realizáciu tohto detektoru je načítanie zoznamu legitímnych domén z predloženého súboru, ktoré je realizované pred samotným spracovávaním doménových mien. Predpokladá sa, že tento zoznam obsahuje na každom riadku práve jedno legitímne doménové meno. Tento krok je naimplementovaný vo funkcii `loadAlexa`, ktorá reťazec na každom riadku rozdelí podľa bodiek, čím extrahuje reťazce predstavujúce subdomény doménového mena a každý z týchto reťazcov dlhší ako 3 uloží do dátového kontajneru `vector` typu `string`.

Detekcia výskytu subdomén v zozname doménových mien Alexy je implementovaná vo funkcii `isInAlexa`, ktorá v prípade výskytu domény vo whiteliste vracia `true`, v opačnom prípade `false`. Algoritmus funkcie znázorňuje Obrázok 8. Doménové mená nájdené vo whiteliste sú z ďalšej analýzy vyradené, pokiaľ bol užívateľom zadaný príslušný parameter.

Algoritmus 1: Funkcia `isInAlexa`

```
/* pre každú subdoménu spracovávanej domény                    */
foreach level of Domain do
   /* pre každé meno z whitelistu                               */
   foreach alexaName of alexaDomains do
   /* meno z whitelistu sa zhoduje s menom subdomény            */
   if alexaName == level.domName then
      level.notInAlexa = false;
   else
      level.notInAlexa = true;
   end if
end foreach
/* subdoména sa nenašla vo whiteliste                           */
if level.notInAlexa then
   return false;
end if
return true;
end foreach
```

Obrázok 8: Algoritmus funkcie loadAlexa

Dátový kontajner `vector` bol na uloženie použitý z toho dôvodu, že vopred nie je jasné, koľko legitímnych domén bude treba načítať a s jeho použitím je k dispozícii neobmedzené množstvo pamäti bez nutnosti opätovnej alokácie.

Z uvedených informácií je zrejmé, že zoznam legitímnych doménových mien musí byť editovateľný užívateľom. Kvôli zvýšeniu flexibility programu je teda očakávané, že tento zoznam je uložený v súbore, ktorého názov je zadaný parametrom pri spúšťaní. Ďalej sa predpokladá, že zoznam obsahuje na každom riadku práve jedno legitímne doménové meno.

5.2.2 Detektor kontroly názvu domény na základe entropie

Táto detekcia je implementovaná funkciou `checkEntropy`, ktorá na vstupe očakáva celé doménové meno ako položku štruktúry `Domain`. Funkcia využíva základné matematické operácie na výpočet entropie podľa vzorca uvedeného v kapitole 5.2.2.1 a tento výpočet realizuje pre každú subdoménu domény uvedenej na vstupe. Ak hodnota entropie priradená jednej jeho subdoméne prekročí maximálnu hranicu pre legitímnu doménu, príznak DGA doménového mena je nastavený na pozitívny.

5.2.2.1 Výpočet entropie

Hodnota entropie jedného slova sa vypočíta nasledovne:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Kde:
n je počet písmen v slove
p_i je pravdepodobnosť výskytu i-tého písmena v slove

Príklad: Výpočet entropie *H* pre slovo `google`.

$$H = -\left(\left(\frac{2}{6}\log_2\left(\frac{2}{6}\right)\right)_g + \left(\frac{2}{6}\log_2\left(\frac{2}{6}\right)\right)_o + \left(\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right)_l + \left(\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right)_e\right) = 1{,}91830$$

Slovu `google` odpovedá hodnota entropie 1,91830. Keď túto hodnotu porovnáme s priemernou hodnotou všetkých predložených legitímnych domén, ktorá vychádzala 2,80398, je patrné, že hodnota entropie u slova `google` je badateľne nižšia. Toto slovo teda z hľadiska entropie spadá pod stanovenú medzu ohraničujúcu legitímne domény a potvrdzuje to aj fakt, že sa jedná o celosvetovo známu legitímnu doménu.

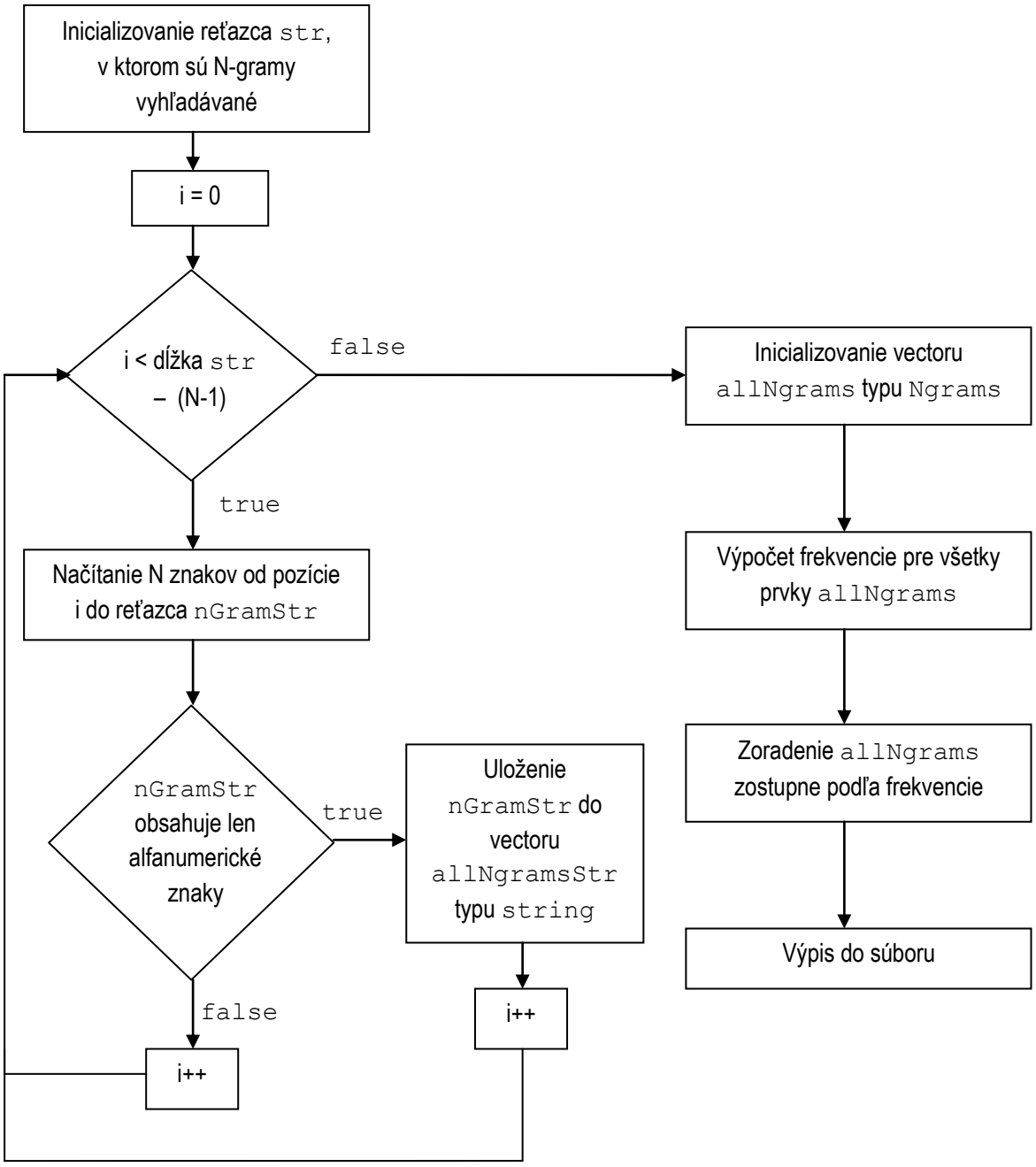
5.2.3 Detektor kontroly názvu domény na základe n-gramov

Na ukladanie n-gramov a ich frekvencií bola vytvorená štruktúra `Ngrams`, ktorá obsahuje nasledovné položky:

- **`string`** `ngramName` – reťazec tvoriaci n-gram
- **`double`** `freq` – frekvenčná charakteristika n-gramu

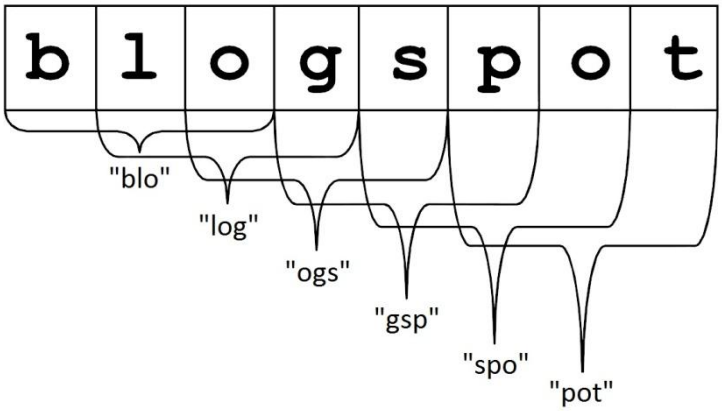
Aplikácia na vstupe očakáva súbor, v ktorom sa budú vyhľadávať n-gramy a následne počítať pravdepodobnosti ich výskytu. Typicky sa jedná o whitelist, užívateľ však môže predložiť napríklad aj množinu najfrekventovanejších slov určitého jazyka. Operáciu vykonáva funkcia `findNgrams`, ktorá v tomto súbore vyhľadá všetky n-gramy, vypočíta k nim frekvenčnú charakteristiku a kvôli znovupoužiteľnosti ich uloží do súboru v zložke `output` v tvare:

```
ing 3.66656
log 2.75471
ine 2.66273
```



Obrázok 9: Algoritmus funkcie `findNgrams`

Obrázok 9 znázorňuje princíp algoritmu funkcie `findNgrams`. Krok výpočtu frekvencie pre prvky vectoru `allNgrams` spočíva v tom, že pre každý prvok `allNgramsStr` sa vykoná prehľadávanie prvkov vectoru `allNgrams`, ktoré doposiaľ nadobudol. Ak je v položke `ngramName` prvku vectoru `allNgrams` reťazec zhodný s reťazcom `allNgramsStr`, položka `freq` tohoto prvku sa inkrementuje. Inak sa do vectoru `allNgrams` pridá nový prvok, ktorého položka `ngramName` je inicializovaná na aktuálny reťazec `allNgramsStr` a položka `freq` je rovná 1. Po tomto prehľadaní položky `freq` vectoru `allNgrams` obsahujú počet výskytov jednotlivých n-gramov, preto je nutné pre všetky jeho prvky túto hodnotu podeliť počtom slov zo zoznamu, z ktorého boli n-gramy počítané, pričom sa zanedbajú n-gramy s počtom výskytov menším ako 100. Táto hodnota je následne vynásobená číslom 100 kvôli získaniu percentuálneho ohodnotenia.



Obrázok 10: 3-gramy extrahované zo slova `blogspot`

Princíp činnosti funkcie `getNgramsOfDN`, ktorá realizuje extrakciu n-gramov z doménového mena, vyplýva z obrázku 10. Funkcia vracia takto extrahované reťazce n-gramov vo forme `vectoru` typu `string`, ktorý je súčasťou vstupu funkcie `countNgrams`.

Vstupom funkcie `countNgrams` je taktiež súbor obsahujúci zoznam n-gramov a ich frekvencií. Zoznam je funkciou prechádzaný po riadkoch, pričom každé spracovanie riadku zahŕňa inicializovanie novej položky vectoru `alexaGrams`, ktorý tvoria prvky štruktúry `Ngrams`, viď Obrázok 11.

vector<Ngrams> `alexaGrams`

0	1	2	...
ngramName == "ing" freq == 3.66656	ngramName == "log" freq == 2.75471	ngramName == "ine" freq == 2.66273	...

Obrázok 11: Inicializovaný vector `alexaGrams`

Príklad: Počítanie pravdepodobnosti 3-gramov pre doménu `blogspot`.

Prvým krokom je extrakcia všetkých 3-gramov z daného slova a ich uloženie do dátového kontajnera `vector`. 3-gramy extrahované z tohto slova znázorňuje Obrázok 10.

Následne sa inicializuje `vector` obsahujúci prvky štruktúry `Ngrams`, ktoré predstavujú dvojice *3-gram, frekvenčná charakteristika n-gramu* načítané z predloženého (prípadne vypočítaného) zoznamu. Každý 3-gram zo slova `blogspot` sa potom vyhľadá v tomto `vectore` a položka subdomény `alexaGrams`, predstavujúca súčet frekvenčných charakteristík 3-gramov, sa inkrementuje o pravdepodobnosť prislúchajúcu k danému 3-gramu:

$$alexaGrams = 0{,}73_{blo} + 1{,}07_{log} + 0{,}08_{ogs} + 0{,}03_{gsp} + 0{,}63_{spo} + 0{,}17_{pot} = 2{,}71$$

Počítanie n-gramov je možné doplniť parametricky. Zvolením výpočtu n-gramov sa však značne predĺži výpočtová doba, keďže predložený whitelist kvôli relevantnosti analýzy obsahuje veľké množstvo doménových mien, ktorý je potom nutné niekoľkokrát prehľadávať. Optimalizácia je navrhnutá v kapitole 6.2.1.

Jednou z možností optimalizácie rýchlosti je tiež spustenie analýzy s vopred vyhľadanými n-gramami a ich pravdepodobnosťami. Preto ak je užívateľom vybraná voľba uvedená v predošlom

odseku, n-gramy a ich pravdepodobnosti sa uložia do súboru `3grams`, prípadne `4grams` v zložke `src_data` zoradené zostupne podľa hodnoty pravdepodobnosti. Tento súbor je potom možné znovu použiť na analýzu iného zoznamu doménových mien bez nutnosti opätovného vyhľadávania n-gramov vo whiteliste, čím sa skráti výpočtová doba programu.

5.3 Výstup aplikácie

Implementovaný detektor škodlivých domén je konzolová aplikácia, ktorej výstupom je textový súbor obsahujúci analýzu predložených doménových mien. Ďalšou formou výstupných dát sú grafy znázorňujúce výsledky analýzy. Po skončení behu aplikácie sa na štandardnom výstupe zobrazí stručná štatistika o analyzovaných doménových menách a počte odhalených anomálií.

Výstupný textový súbor `analysis.txt` obsahuje detailný prehľad analyzovaných doménových mien. Obsahuje hodnoty entropie alebo frekvenčnej charakteristiky n-gramov pre každú subdoménu, spolu s dôvodom prípadného označenia doménového mena za škodlivé. Po skončení behu aplikácie sa na štandardnom výstupe zobrazí stručná štatistika reprezentujúca celkový počet analyzovaných doménových mien a počet zistených anomálií. Pri týchto výpisoch sa zobrazuje iba analýza týkajúca sa spustených detektorov.

Príklad: Ukážka zo súboru `analysis.txt` pri spustení detektoru entropie. Pre každé doménové meno je vypísané jeho celé znenie a ku každej jeho subdoméne je vypísaná hodnota entropie alebo frekvenčnej charakteristiky n-gramov, podľa zvoleného detektoru. Pokiaľ je subdoména príliš krátka a nebola analyzovaná, výpis hodnoty je nahradený reťazcom "---". Nasleduje informácia o detekcii anomálie a v prípade jej výskytu aj dôvod označenia doménového mena za škodlivé.

```
DOMENOVE MENO : registracia.azet.sk

sk -- entropia: ---
azet -- entropia: 2
registracia -- entropia: 2.91398

Vyskyt anomalie: nie
-------------------------
DOMENOVE MENO : www.sslgamesloadbalancer-714831773.eu-west-
1.windowsupdate.amazonaws.com

com -- entropia: ---
amazonaws -- entropia: 2.6416
windowsupdate -- entropia: 3.39275
eu-west-1 -- entropia: 2.72548
sslgamesloadbalancer-714831773 -- entropia: 3.96474

Vyskyt anomalie: ANO
Dovod: VYSOKA HODNOTA ENTROPIE
```

```
--------------------------
DOMENOVE MENO : photos-h.ak.instagram.com

com -- entropia: ---
instagram -- entropia: 2.9477
ak -- entropia: ---
photos-h -- entropia: 2.5

Vyskyt anomalie: nie
```

Výstup aplikácie tvoria taktiež súbory obsahujúce dáta určené na generovanie grafov, ktoré sa ukladajú do zložky `output`. Po skončení behu aplikácie je možné v zložke `src` príkazom `./plot_all.sh` vytvoriť grafy zobrazujúce výsledky analýzy. Na generovanie grafov je potrebné mať nainštalovaný program Gnuplot.

Príklad: Ukážka vstupného súboru na tvorbu grafu `entropy_length.jpeg` pomocou nástroja Gnuplot.

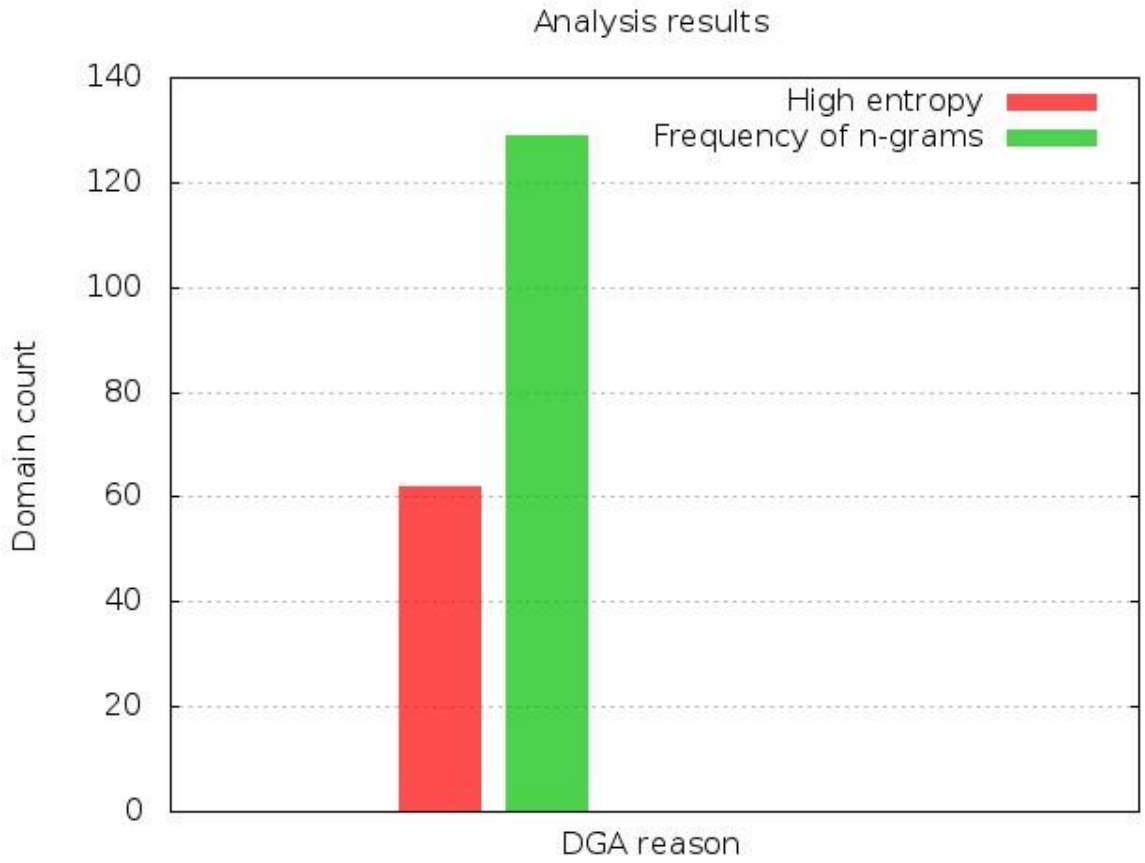
```
9 2.32193 1
9 2.32193 0
8 1.92193 0
2 -1 0
9 2.32193 1
```

Hodnota v prvom stĺpci predstavuje dĺžku domény, hodnota v druhom stĺpci jej entropiu. Ak je analyzovaná doména príliš krátka, je na tomto mieste vypísaná konštanta -1 a doména v grafe nie je zobrazená. Hodnota v treťom stĺpci slúži na určenie prítomnosti domény vo whiteliste a jej následné farebné zobrazenie. Ak sa doména nachádza v predloženom whiteliste, v treťom stĺpci je vypísaná hodnota 1, inak 0.

Výstupom je jeden štatistický graf a tri grafy zobrazujúce vlastnosti subdomén analyzovaných doménových mien, pričom tie, ktoré sú analýzou označené za škodlivé, sú zobrazené červenou farbou. Obdoba týchto grafov bola použitá počas testovania aplikácie a analýzy výsledkov, v tomto prípade sa však dbalo na rozlišovanie domén podľa ich výskytu vo whiteliste. Niektoré z foriem týchto grafov sú uvedené v kapitole 6. Jedná sa o grafy:

- `ngram_len.jpeg` – zobrazuje frekvenčnú charakteristiku n-gramov v závislosti na dĺžke domény
- `ngram_ent.jpeg` – zobrazuje frekvenčnú charakteristiku n-gramov v závislosti na entropii domény. V prípade, že analýza na základe entropie, alebo analýza na základe frekvenčnej charakteristiky nebola spustená, sa tento graf vytvorí ako prázdny graf.
- `entropy_len.jpeg` – zobrazuje entropiu domén v závislosti na jej dĺžke
- `statistics.jpeg` – zobrazuje počet domén označených ako DGA na základe entropie a počet domén označených ako DGA na základe frekvenčnej charakteristiky n-gramov. Ukážku poskytuje Obrázok 12.

V prípade, že aplikácia bola spustená len s parametrami `-i` a `-w`, a teda je realizovaná iba detekcia výskytu domény vo whiteliste, výstupné grafy sú prázdne.



Obrázok 12: Výstupný graf `statistics.jpeg` zobrazujúci výsledky analýzy

6 Analýza výsledkov detekcie

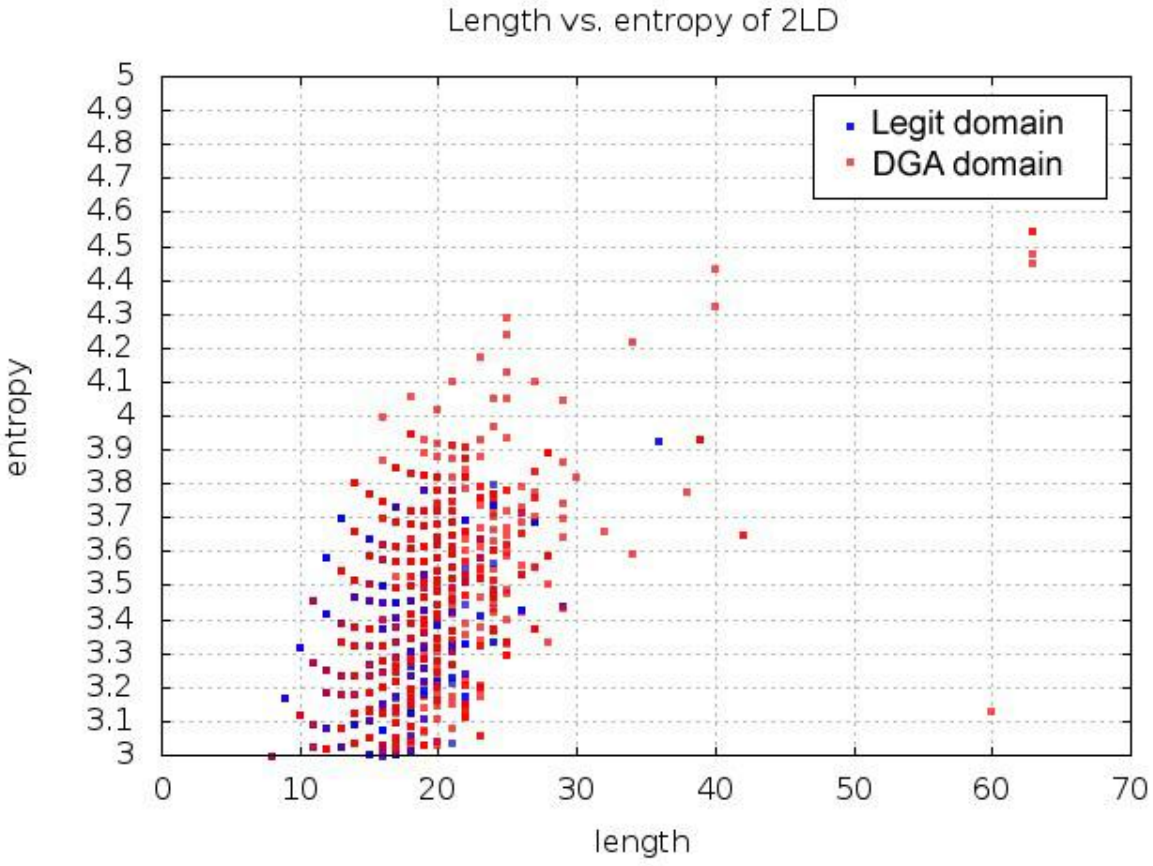
Táto kapitola sa zaoberá analýzou výsledkov detekcie a testovaním systému, ktoré bolo realizované priebežne počas vývoja. Testovanie bolo spúšťané na dátach obsahujúcich zmes legitímnych a DGA doménových mien zachytených exportérmi v sieti CESNETu.

Kľúčovým krokom detekcie škodlivých domén bolo stanoviť hranice oddeľujúce DGA prevádzku od legitímnej. Z toho dôvodu boli po výpočte entropie a frekvenčnej charakteristiky n-gramov vytvorené grafy zobrazujúce tieto ich vlastnosti spolu s ich výskytom vo whiteliste. Dáta, s ktorými bola táto analýza spustená, obsahovali doménové mená z reálnej prevádzky.

6.1 Detektor entropie

Priemerná hodnota entropie predložených legitímnych domén dosahuje 2,80398. Niektoré doménové mená túto hodnotu prekračujú o viac než tretinu, preto je možné už na základe tejto informácie odhaliť potenciálne škodlivú doménu.

Na základe výsledkov detekcie spustenej na dátach z reálnej prevádzky bol zostavený graf, viď Obrázok 13. Graf zobrazuje doménové mená druhej úrovne s hodnotou entropie vyššou ako 3, pričom rozlišuje doménové mená vyskytujúce sa vo whiteliste od tých, ktoré v ňom nájdené neboli. Z grafu vidno, že väčšina domén s hodnotou entropie vyššou než 3,8 sa nenachádza vo whiteliste, a preto je

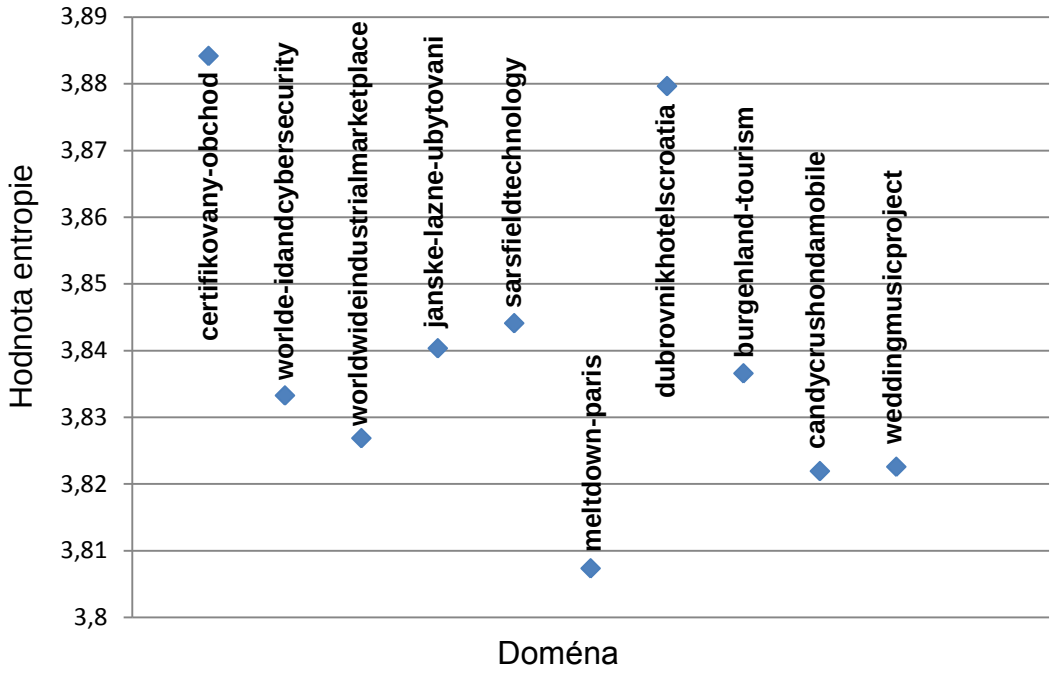


Obrázok 13: Hodnota entropie v porovnaní s dĺžkou domény

možné ich považovať za škodlivé. Z grafu je tiež zrejmé, že hodnota entropie pre doménové meno rastie s jeho narastajúcou dĺžkou, čo však nemá veľkú výpovednú hodnotu pri stanovovaní hľadanej medze.

Na základe uvedeného grafu bola ako maximálna hodnota entropie pre legitímnu doménu stanovená hodnota 3,8 a následne realizované testovanie jej správnosti. Počas testovania detektoru na vzorke dát obsahujúcej 181356 doménových mien bolo identifikovaných 6013 domén, ktoré túto hranicu prekračujú.

Analyzovaním domén označených za škodlivé a ich hodnôt entropie sa zistilo, že približne každá pätnásta doména identifikovaná ako škodlivá predstavuje `false positive` výsledok. Na nasledujúcom grafe (Obrázok 14) sú uvedené príklady takýchto doménových mien.

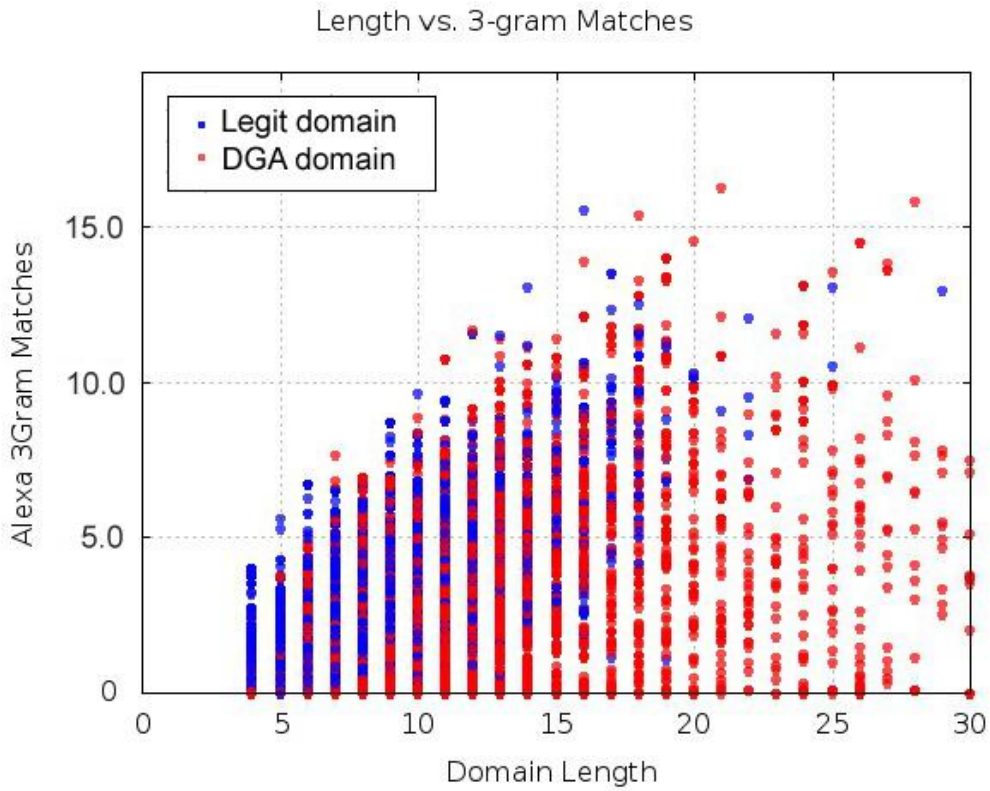


Obrázok 14: Legitímne doménové mená s vysokou hodnotou entropie

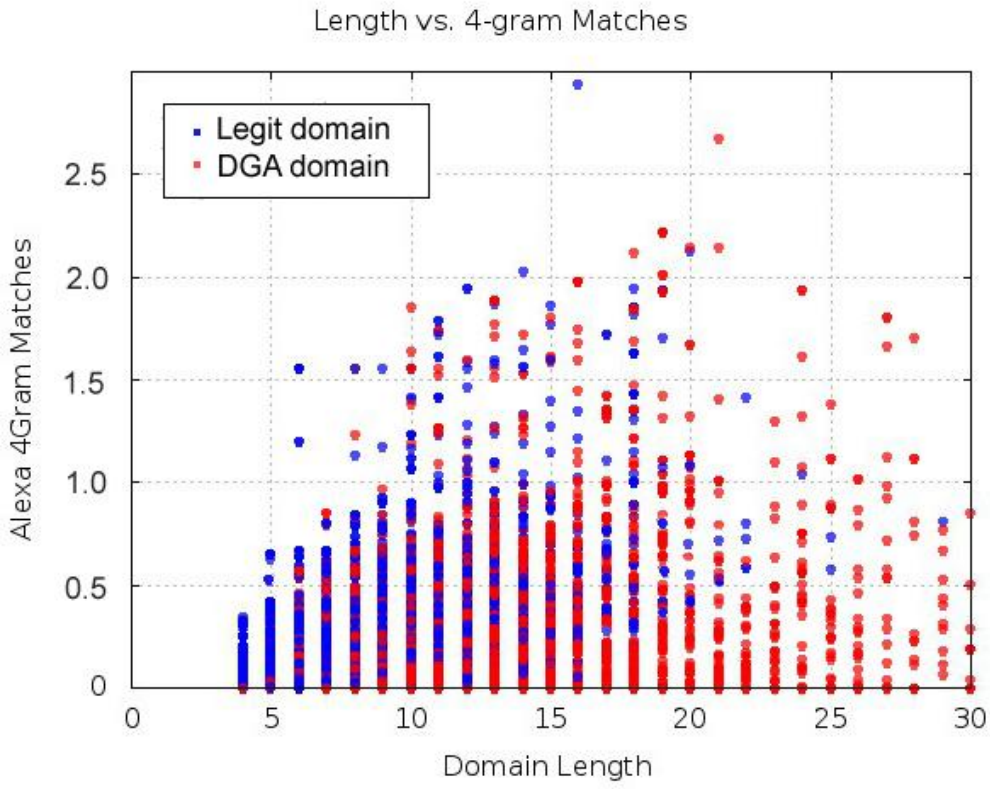
Uvedený problém by bolo možné riešiť posunutím stanovenej medze. Interval hodnoty entropie, do ktorého spadal najväčší počet `false positive` výsledkov, tvorili hodnoty 3,8 až 3,9. Do tohto intervalu však spadala aj hodnota entropie správne identifikovaných škodlivých domén, ktoré tvorili takmer až tretinu zo všetkých týchto domén. Posunutie hranice by teda spôsobilo okrem redukcie `false positive` výsledkov aj redukciu množstva detekovaných DGA domén, preto je tento problém vhodné riešiť iným spôsobom, napríklad pridaním doménového mena do whitelistu.

6.2 Detektor n-gramov

Tak, ako v prípade entropie, aj z hodnôt frekvenčnej charakteristiky n-gramov v doménovom mene boli určené medze, do ktorých spadajú legitímne domény. Z nasledujúcej dvojice grafov (Obrázok 15 a Obrázok 16) je zrejmé, že frekvencia n-gramov v legitímnych doménach narastá s ich dĺžkou. Z tejto informácie bol vyvodený záver, že dlhšie domény disponujúce nízkou frekvenčnou charakteristikou n-gramov budú pravdepodobne DGA pôvodu. Stanovené medze zobrazuje Tabuľka **3**.



Obrázok 15: Frekvenčná charakteristika 3-gramov v závislosti na dĺžke domény



Obrázok 16: Frekvenčná charakteristika 4-gramov v závislosti na dĺžke domény

3-gramy		4-gramy	
dĺžka domény	frekvenčná charakteristika n-gramu	dĺžka domény	frekvenčná charakteristika n-gramu
13	0,5	13	0,05
14	1	14	0,05
15	1,5	15	0,05
16	2	16	0,1
17	2,5	17	0,15
18	3	18	0,2
19	3,5	19	0,3
20	4	20	0,4
21	4,5	21	0,5

Tabuľka 3: Medze stanovené na detekciu DGA domén s využitím frekvenčnej charakteristiky

Ak je dĺžka reťazca tvoriaceho doménu väčšia ako stanovená hranica a súčet frekvenčných charakteristík n-gramov vyskytujúcich sa v nej je nižší ako stanovená hranica, doména je označená za škodlivú.

Testovaním tohto detektoru dochádzalo k `false positive` výsledkom najmä vtedy, keď sa jednalo o analýzu doménových mien, ktorých názov pozostáva z iných než anglických slov.

Príklad: Legitímne doménové mená s nízkou hodnotou frekvenčnej charakteristiky n-gramov.

```
velkoobchodni-ceny
zdravaakrasnazahrada
janske-lazne-ubytovani
holbein-gymnasium
```

Keďže implementovaný detektor využíva na počítanie frekvenčnej charakteristiky n-gramov whitelist zložený z celosvetovo najpopulárnejších doménových mien, je v ňom vysoká koncentrácia slov anglického jazyka a domény iných jazykov sú ohodnotené nižšou hodnotou. Implementovaný detektor by teda bolo možné rozšíriť o analyzovanie frekvenčnej charakteristiky n-gramov konkrétneho jazyka.

6.2.1 Optimalizácia detektoru

Bolo by zbytočné zahrnúť do analýzy n-gramy, ktoré sa síce v zozname legitímnych domén nachádzajú, no počet ich výskytov je nízky. Preto bolo z hľadiska rýchlosti programu výhodné stanoviť hranicu určujúcu hodnotu počtu výskytov, pod ktorú je výskyt daného n-gramu v analyzovanej doméne irelevantný. Táto hranica bola stanovená na základe nasledovnej dvojice grafov, viď Obrázok 17 a 18.

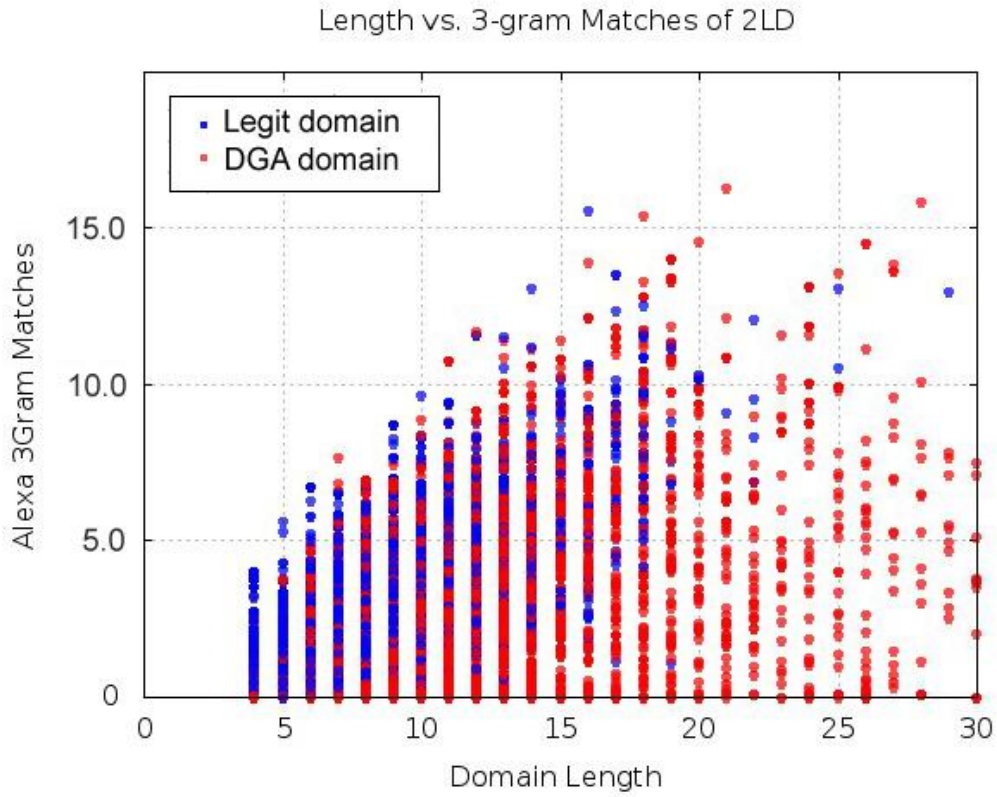
Obrázok 17: Graf počtu výskytov 3-gramov v predloženom whiteliste

Obrázok 18: Graf počtu výskytov 4-gramov v predloženom whiteliste

Obrázok 17 predstavuje graf počtu výskytov jednotlivých 3-gramov v predloženom whiteliste. Najvyšší počet výskytov z tohto zoznamu mal 3-gram `ing`, u ktorého táto hodnota dosahovala 33429. Pre zvýšenie prehľadnosti a presnejšie určenie hranice je zobrazená iba časť grafov relevantná pre určenie žiadanej hodnoty.

Pravdepodobnosť výskytu n-gramu, ktorý sa v predloženom zozname nachádza 100 krát, je približne 0,0001, čo je zanedbateľná hodnota. Na základe tohto poznatku a grafu (Obrázok 17) bola hranica určujúca zanedbateľnosť výskytu 3-gramu stanovená na 100. Rovnaká hranica zanedbania bola určená aj pri detekcii na základe 4-gramov, viď Obrázok 18.

Výpočet výskytov n-gramov bol na rovnakej vzorke dát spustený pred aj po stanovení hranice. Nasledujúci graf (Obrázok 19) potvrdzuje fakt, že 3-gramy vyskytujúce sa v predloženom whiteliste menej ako 100 krát nemajú viditeľný vplyv na výsledok a nie je ich teda nutné do analýzy zahrnúť. Graf predstavuje zvyšujúce sa hodnoty výskytov 3-gramov vzhľadom k narastajúcej dĺžke domény. Potvrdzujúce výsledky vychádzali aj po zopakovaní analýzy so stanovenou hranicou v prípade 4-gramov.



Obrázok 19: Frekvenčná analýza 3-gramov v závislosti na dĺžke domény po stanovení hranice zanedbateľných výskytov

Súčasťou aplikácie je kvôli optimalizácii aj súbor obsahujúci 3-gramy z whitelistu, ktorý bol súčasťou vývoja, a k nim prislúchajúce hodnoty frekvenčnej charakteristiky. Je umiestnený v zložke `src_data` s názvom `3grams`. Výhodou spustenia detektoru s týmito dátami na vstupe je skrátenie výpočtovej doby z desiatok minút na jednotky sekúnd.

7 Záver

Táto práca sa zaoberá detekciou škodlivých domén na základe pasívnej analýzy DNS. Obsahuje teoretickú časť, ktorá je zameraná na základný popis služby DNS, možnosti monitorovania sieťovej prevádzky a prehľad anomálií v DNS. Na základe naštudovaných metód detekcie bol navrhnutý systém zložený z niekoľkých detektorov anomálií, ktoré môžu pracovať spolu alebo oddelene.

Pre implementáciu bol zvolený jazyk C++ a nástroj Gnuplot, pomocou ktorého sú tvorené grafy zobrazujúce výsledky analýzy. Funkčnosť implementovaného systému bola testovaná na dátach z reálnej DNS prevádzky. Výsledkom je detektor použiteľný prevažne v oblasti správy a zabezpečenia sietí a to svojou schopnosťou odhaliť prípadný DNS útok na základe atypickej skladby doménového mena.

Napriek tomu, že výsledný detektor účinne detekuje domény DGA pôvodu, stále je tu priestor na vylepšovanie aplikácie a pridávanie rôznych funkcionalít. Ako je uvedené v kapitole 6, počas analýzy na základe frekvenčnej charakteristiky dochádzalo k najväčšiemu počtu false positive výsledkov vtedy, keď sa jednalo o doménové meno skladajúce sa z iných než anglických slov. Dôvodom je fakt, že n-gramy a ich frekvenčná charakteristika boli čerpané zo zoznamu najpopulárnejších svetových domén. Spresnenie výsledkov by preto mohlo byť docielené implementovaním takého detektoru, ktorý by realizoval analýzu s možnosťou voľby konkrétneho jazyka. K redukcii false positive výsledkov by prispelo aj pridanie legitímnych domén prekračujúcich stanovené hranice do whitelistu, a to na základe dlhodobého testovania.

Tiež by bolo možné implementovať ďalšie detektory analyzujúce skladbu doménového mena, ktoré by pracovali na princípe analýzy výskytu číslic v doménovom mene či frekvenčnej charakteristiky znakov určitého jazyka. Potenciálnym rozšírením je tiež predloženie blacklistu a tým identifikovanie domény za škodlivú a jej možné vyradenie z analýzy.

Literatúra

[1] Matoušek, P.: *Síťové aplikace a jejich architektura*. Akademické nakladatelství, VUTIUM, Brno, 2014. ISBN 978-80-214-3766-1

[2] Kováčik, M., Detekce síťových anomálií a bezpečnostních incidentů s využitím DNS dat, pojednání k tématu disertační práce, Brno: FIT VUT v Brně, 2014.

[3] Claise, B.: *Cisco Systems NetFlow Services Export Version 9*, RFC 3954, 2004.

[4] Quittek, J.; Bryant, S.; Claise, B.;, Aitken, P.; Meyer, J.: *Information Model for IP Flow Information Export*, RFC 5102, 2008.

[5] Bush, R.; Karrenberg, D.; Kosters, M.; Plzak, R.: *Root Name Server Operational Requirements*, RFC 2870, 2000.

[6] Roolvink, S.: Detecting attacks involving DNS servers : A netflow data based approach. 2008. [online]. [cit. 2014-12-29]. Dostupné z: http://essay.utwente.nl/58497/

[7] Basheer, N.: Fast Flux Watch: A mechanism for online detection of fast flux networks. 2014. [online]. [cit. 2015-01-02]. Dostupné z: http://www.sciencedirect.com/science/article/pii/S2090123214000034

[8] Olzak, T.: DNS resource record integrity is still a big, big problem. 2008. [online]. [cit. 2015-01-02]. Dostupné z: http://www.techrepublic.com/blog/it-security/dns-resource-record-integrity-is-still-a-big-big-problem/

[9] Farnham, G.: Detecting DNS Tunneling. 2013. [online]. [cit. 2015-01-02]. Dostupné z: http://www.sans.org/reading-room/whitepapers/dns/detecting-dns-tunneling-34152

[10] Born, K.; Gustafson, D.: NgViz: detecting DNS tunnels through n-gram visualization and quantitative analysis. In *Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research*, CSIIRW '10, 2010, ISBN 978-1-4503-0017-9.

[11] Alexa: The top 500 sites on the web. [online]. [cit. cit. 2015-01-02]. Dostupné z: http://www.alexa.com/topsites

[12] Kováčik, M.: Liberouter: DNS plugin [online]. [cit. 2015-05-13]. Dostupné z: https://www.liberouter.org/technologies/dns-plugin/

[13] YADAV, Sandeep, Ashwath Kumar Krishna REDDY, A.L. Narasimha REDDY, Supranamaya RANJAN, Ying ZHANG, Yongzheng ZHANG a Jun XIAO. 2010. Detecting algorithmically generated malicious domain names. *Proceedings of the 10th annual conference on Internet measurement - IMC '10* [online]. [cit. 2015-05-17]. Dostupné z: http://kodu.ut.ee/~koit/KT/imc104-yadav.pdf

[14] Threat Spotlight: Dyre/Dyreza: An Analysis to Discover the DGA. [online]. [cit. 2015-05-17]. Dostupné z: http://blogs.cisco.com/security/talos/threat-spotlight-dyre

[15] Pcap - Packet Capture library. [online]. [cit. 2015-05-18]. Dostupné z: http://www.tcpdump.org/manpages/pcap.3pcap.html

Príloha A

Parametre pre spustenie aplikácie

`-h` vypíše nápovedu

`-i <názov súboru>` určenie vstupného súboru obsahujúceho na každom riadku jedno doménové meno ohraničené úvodzovkami

`-o <názov súboru>` určenie výstupného súboru. Pokiaľ tento parameter nie je zadaný, výstup je uložený do súboru `analysis.txt`

`-w <názov súboru>` určenie súboru obsahujúceho zoznam legitímnych doménových mien, ktoré je možné vylúčiť z analýzy

`-e` spustí detekciu na základe entropie

`-n <N>` spustí detekciu na základe n-gramov, dĺžka n-gramov N je zvolená ako číslo 3 alebo 4. Tento parameter je použiteľný iba v kombinácii s parametrom `-f`, ktorý špecifikuje názov súboru obsahujúceho vopred vyhľadané n-gramy a k nim prislúchajúce pravdepodobnosti.

`-c <N>` spustí detekciu na základe n-gramov, dĺžka n-gramov N je zvolená ako číslo 3 alebo 4. Tento parameter je použiteľný iba v kombinácii s parametrom `-f`, ktorý špecifikuje názov súboru obsahujúceho zoznam slov, v ktorom sa majú vyhľadať n-gramy a dopočítať ich pravdepodobnosti výskytu

`-f <názov súboru>` pri použití v kombinácii s parametrom `-n` špecifikuje súbor obsahujúci n-gramy a k nim prislúchajúce pravdepodobnosti. Pri použití v kombinácii s parametrom `-c` špecifikuje súbor, z ktorého sa majú vypočítať pravdepodobnosti výskytu v ňom nájdených n-gramov.

`-d` zamedzí výskyt duplicitných doménových mien vo výstupnom súbore

`-a` zaistí výpis legitímnych doménových mien do výstupného súboru. Ak nie je zvolený, vypíše sa iba analýza škodlivých mien

Parameter `-i` je povinný. Parametre `-n` a `-c` sú nekombinovateľné. Pri voľbe parametru `-n` alebo `-c` je prítomnosť parametru `-f` povinná.

Príloha B

Obsah CD

Priložené CD obsahuje zdrojové kódy aplikácie a dáta potrebné pre jej beh uložené v nasledujúcej adresárovej štruktúre:

- src – adresár zdrojových súborov
- src_data – adresár obsahujúci dáta potrebné na beh aplikácie
- output – prázdny adresár, do ktorého sa ukladajú výstupné súbory pre generovanie grafov
- BP – bakalárska práca vo formáte pdf
- README – pokyny k použitiu aplikácie