VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ
DEPARTMENT OF COMPUTER SYSTEMS

REPUTACE ZDROJŮ ŠKODLIVÉHO PROVOZU

REPUTATION OF MALICIOUS TRAFFIC SOURCES

TEZE DISERTAČNÍ PRÁCE
DISSERTATION THESIS STATEMENT

AUTOR PRÁCE **Ing. VÁCLAV BARTOŠ**
AUTHOR

ŠKOLITEL **doc. Ing. JAN KOŘENEK, Ph.D.**
SUPERVISOR

BRNO 2018

Abstrakt

Při zajišťování bezpečnosti počítačových sítí je mimo jiné nezbytné získávat a zpracovávat informace o existujících hrozbách, ať už odvozené z hlášení vlastních detekčních nástrojů či pocházející od třetích stran. Mezi takové informace patří i seznamy síťových entit (IP adres, doménových jmen, URL apod.), které byly identifikovány jako škodlivé. V mnoha případech však prostá binární informace, zda je daná entita škodlivá či nikoliv, nestačí. Je vhodné mít ke každé entitě i další data popisující jí prováděné škodlivé aktivity a také shrnující skóre, které její reputaci vyjádří číselně. To umožní jednak rychlé zhodnocení míry hrozby, kterou určitá entita představuje, a zároveň umožní entity porovnávat a řadit. Tato práce se zabývá návrhem právě takového reputačního skóre. Navržené skóre, nazvané *Future Maliciousness Probability* (FMP skóre), je hodnota mezi 0 a 1 přiřazená každé podezřelé síťové entitě a vyjadřující pravděpodobnost, že bude daná entita v nejbližší době (znovu) provádět určitou škodlivou činnost. Výpočet tohoto skóre je tedy založen na předpovědi budoucích útoků. Tato předpověď vychází z historie přijatých hlášení o bezpečnostních událostech a z dalších relevantních dat týkajících se dané entity a je založena na pokročilých metodách strojového učení. Metoda výpočtu skóre je v práci nejprve popsána obecně, pro libovolný typ entity a vstupní data, a poté je přizpůsobena pro konkrétní případ – hodnocení IPv4 adres na základě hlášení ze systému pro sdílení bezpečnostních událostí a doplňujících dat z reputační databáze. Tato metoda pak byla vyhodnocena na reálných datech. Kvůli potřebě získat dostatečně velkou a kvalitní datovou sadu pro toto vyhodnocení se část práce věnuje i oblasti detekce bezpečnostních událostí (framework pro analýzu dat o síťových tocích NEMEA a návrh několika nových detekčních metod) a vývoji otevřené reputační databáze NERD, která slouží k udržování profilů nahlášených IP adres. Data z těchto systémů pak byla využita jak pro vyhodnocení přesnosti predikce, tak pro vyhodnocení vybraných případů použití výsledného FMP skóre.

Klíčová slova

síťová bezpečnost, reputace, reputační skóre, reputační databáze, predikce útoků, strojové učení, analýza síťového provozu

Citace

BARTOŠ, Václav. *Reputace zdrojů škodlivého provozu.* Brno, 2018. Teze disertační práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Školitel doc. Ing. Jan Kořenek, Ph.D.

Abstract

An important part of maintaining network security is collecting and processing information about about cyber threats, both from network operator's own detection tools and from third parties. A commonly used type of such information are lists of network entities (IP addresses, domains, URLs, etc.) which were identified as malicious. In many cases, the simple binary distinction between malicious and non-malicious entities is not enough. It is beneficial to keep other supplementary information for each entity, which describes its malicious activities, and also a summarizing score, which evaluates its reputation numerically. Such a score allows for quick comprehension of the level of threat the entity poses and allows to compare and sort entities. The goal of this work is to design a method for such summarization. The resulting score, called *Future Maliciousness Probability* (FMP score), is a value between 0 and 1, assigned to each suspicious network entity, expressing the probability that the entity will do some kind of malicious activity in a near future. Therefore, the scoring is based of prediction of future attacks. Advanced machine learning methods are used to perform the prediction. It is based on previously received alerts about security events and on other relevant data related to the entity. The method of computing the score is first described in a general way, usable for any kind of entity and input data. Then a more concrete version is presented for scoring IPv4 address by utilizing alerts from an alert sharing system and supplementary data from a reputation database. This method is then evaluated on a real world dataset. In order to get enough amount and quality of data for this dataset, a part of the work is also dedicated to the area of security analysis of network data. A framework for analysis of flow data, NEMEA, and several new detection methods are designed and implemented. An open reputation database, NERD, is also implemented and described in this work. Data from these systems are then used to evaluate precision of the predictor as well as to evaluate selected use cases of the scoring method.

Keywords

network security, reputation, reputation score, reputation database, attack prediction, machine learning, network traffic analysis

Obsah

1 Úvod

Počítačové sítě a jejich prostřednictvím poskytované služby jsou dnes zcela běžnou součástí života většiny z nás. Je proto důležité zajistit nejen maximální spolehlivost těchto sítí a služeb, ale také jejich zabezpečení. Na všechny počítačové systémy připojené k internetu totiž prakticky neustále míří nejrůznější kybernetické hrozby [1, 2]. Těmi mohou být například různé podvodné emaily a phishing, útoky typu odepření služby (DDoS), šíření malware či pokusy o ovládnutí cizích systémů útočníky za různými účely, jako je změna obsahu webových stránek, odposlech citlivých dat, průmyslová i státní špionáž či sabotáž, často jsou také napadené systémy jen využity pro vykonávání dalších útoků. Stále častější jsou také masivní úniky osobních dat. V posledních letech se navíc objevují nové hrozby, jako například ransomware, tedy malware, jenž nějakým způsobem (obvykle zašifrováním dat) znepřístupní počítačový systém a vyžaduje zaplacení výkupného, nebo tzv. cryptojacking, tedy neoprávněné využívaní cizích výpočetních prostředků pro těžbu kryptoměn.

Proti kybernetickým hrozbám lze bojovat řadou způsobů. Důležité je například monitorování síťového provozu a detekce škodlivých činností pomocí různých typů systémů typu IDS (systém pro detekci průniků). Kromě toho je však vhodné získávat i informace o potenciálních hrozbách od třetích stran. Těmi mohou být například seznamy různých indikátorů hrozeb (blacklisty), informace o objevených zranitelnostech, existujícím malware, technikách používaných známými skupinami hackerů apod. Někdy jsou tato data vytvářena a spravována konkrétním subjektem a poskytována buď zdarma či v rámci placené služby, existují ale i různé platformy pro sdílení takových informací přímo mezi zapojenými organizacemi, které pak data zároveň přijímají i vytvářejí. Sdílena jsou někdy i přímo strojová hlášení z detekčních systémů.

I přes rozvoj platforem pro sdílení informací na vyšší úrovni abstrakce jsou podle nedávného průzkumu [3] stále nejčastěji používaným zdrojem kyberbezpečnostních informací běžné blacklisty (také nazývané seznamy indikátorů kompromitace, *Indicators of Compromise, IoC*), především seznamy škodlivých IP adres a URL.

Tyto blacklisty lze chápat jako jednoduchou formu hodnocení reputace z hlediska bezpečnostních hrozeb. Toto hodnocení je však jen binární – určitá entita (IP adresa, URL či jiný identifikátor) na seznamu buď je, nebo není, je tedy označena za škodlivou, nebo za neškodnou. Přitom skutečnost bývá složitější, různé adresy mohou představovat různou míru a různé druhy rizika a zdroje dat, ze kterých se vychází, mohou být různě spolehlivé. Žádná míra škodlivosti, spolehlivosti ani jiné doplňující informace (např. čas přidání na seznam, počet detekovaných pokusů o útok) však v blacklistech obvykle uváděny nejsou.

Navíc nemá uživatel žádný vliv na pravidla a meze, na základě nichž je rozhodováno o zařazení entity na blacklist, a kvůli absenci doplňujících dat ani nelze blacklist později filtrovat. Nelze si tak například zvolit, zda mají být zahrnuty

i adresy jen mírně podezřelé, či jen ty, u nichž je velká jistota, že jsou škodlivé. Podobně v případě technických omezení na maximální délku aplikovaného blacklistu neexistuje způsob, jak z poskytnutého blacklistu vybrat pouze určitý počet adres tak, aby to byly ty, které představují největší hrozbu.

Jako řešení výše uvedených nedostatků současného stavu autor této práce navrhuje shromažďovat data o detekovaných bezpečnostních událostech v otevřené reputační databázi, a to tak, aby mohly být dále poskytovány nejen hotové seznamy entit, které jsou dle určitých pravidel vyhodnoceny jako škodlivé, ale i všechna další dostupná data o těchto entitách. Ta by uživatelům poskytovala detailní informace o potenciálních hrozbách a umožňovala každému vytvořit si seznam škodlivých entit dle vlastních kritérií. Velmi důležitou součástí systému poskytujícího tato data by měla být i možnost ohodnotit jednotlivé entity pomocí skóre, které by číselně vyjádřilo jejich reputaci či míru hrozby, kterou představují. To by jednak usnadnilo rychlé zhodnocení podezřelé entity člověkem analytikem, zároveň by umožnilo entity porovnávat či řadit. Takové seřazení pak v důsledku umožní i vytváření optimálních blacklistů uživatelem definované velikosti pouhým výběrem prvních n záznamů s nejhorší reputací.

Několik prací se již v minulosti takovému číselnému hodnocení škodlivosti či reputace síťových entit věnovalo [4, 5, 6], vždy však šlo jen o hodnocení celých skupin IP adres – síťových prefixů či čísel autonomních systémů. Takové hodnocení je založeno na pozorování, že adresy ve stejné síti se často chovají podobně, pokud je tedy v jedné síti detekováno několik škodlivých adres, je u ostatních adres v této síti vyšší pravděpodobnost, že jsou nebo brzy budou také škodlivé. Využití těchto korelací je užitečné, hodnocení celých sítí je však problematické v tom, že všechny adresy v dané síti jsou hodnoceny stejně, ačkoliv některé mohou představovat větši hrozbu než jiné – přinejmenším v tom, že některé již byly skutečně detekovány jako škodlivé (a to potenciálně s různou intenzitou), u jiných se pouze předpokládá tato možnost na základě jejich příslušnosti do stejné sítě.

Hlavním cílem této disertační práce je tedy návrh metody výpočtu reputačního skóre, které by bylo přiřazováno každé jednotlivé podezřelé entitě (např. IP adrese). Mělo by shrnovat informace jak o předchozím chování této entity, tak i o chování blízkých či jinak podobných entit (např. ostatních adres ve stejné síti). Tím se spojí výhody obou výše uvedených přístupů k hodnocení reputace – vysoká granularita blacklistů (které rozlišují jednotlivé entity, avšak bez skóre) a využití korelací mezi blízkými entitami, používané dosud pouze při hodnocení celých sítí. Kromě toho by měla metoda výpočtu reputačního skóre umožňovat zahrnout i případná další dostupná data relevantní pro vyhodnocení míry hrozby, kterou entita představuje.

Dále by skóre mělo být založeno na explicitní predikci budoucího chování hodnocené entity, neboť právě odhad budoucího chování je důležitý pro obranu a prevenci útoků. Pouhé shrnutí informací o předchozím chování, jež je základem většiny existujících metod, nemusí být vhodné. Přestože budoucí chování zpravidla vychází z chování minulého, může být ovlivněno i řadou dalších faktorů.

Jediné práce, které v tomto kontextu pracují s předvídáním, které IP adresy pravděpodobně v blízké budoucnosti zaútočí, jsou práce na téma tzv. prediktivního blacklistování [7, 8]. V těchto pracech je však cílem vždy pouze vytvoření blacklistu, nikoliv číselné ohodnocení adres, navíc jsou na blacklist uváděny jen celé /24 prefixy, ne samostatné adresy.

Reputační skóre splňující výše uvedená kritéria lze využít řadou způsobů. Jak již bylo zmíněno, může sloužit pro rychlé zhodnocení škodlivosti entity člověkem či pro vytváření blacklistů s uživatelem volenými parametry, jako je limit na počet záznamů či mezní hodnota míry hrozby, kterou musí entita překročit, aby byla na seznam přidána. Skóre může být dále využito i jako vstup jiných algoritmů, například jako jedno z kritérií pro prioritizaci incidentů v SIEM systémech, v rámci algoritmů rozlišujících škodlivý provoz od legitimního, např. při ochraně proti spamu nebo DDoS útokům, nebo pro řízení míry detailu monitorování provozu jednotlivých entit.

Kromě návrhu metody hodnocení reputace síťových entit se část práce věnuje také oblasti detekce bezpečnostních hrozeb. Cílem je především získání dostatečného množství dat pro hlavní část této práce, neboť hlášení o detekovaných škodlivých aktivitách jsou hlavním vstupem pro určení reputačního skóre, nově vyvinuté detekční metody a framework, v němž byly implementovány, jsou však zároveň samy o sobě významným přínosem ve své oblasti.

1.1 Přínosy práce

Jádrem a hlavním přínosem disertační práce je:

- Návrh obecné metody hodnocení reputace síťových entit na základě předpovědi jejich budoucího chování (kap. 2).

- Ověření navržené metody nad reálnými daty o škodlivých IP adresách (kap. 6 a 7).

Aby mohlo být dosaženo dobrých výsledků v tomto hlavním tématu práce, bylo nutné získat velké množství kvalitních dat a analyzovat je. To vedlo k několika vedlejším přínosům této práce:

- Návrh systému pro analýzu síťového provozu (NEMEA) a několika nových metod pro detekci škodlivého síťového provozu na základě analýzy dat o síťových tocích (kap. 3).

- Návrh a implementace reputační databáze síťových entit (kap. 4).

- Analýza charakteristik bezpečnostních hlášení a zdrojů škodlivého provozu (kap. 5).

1.2 Struktura tezí

V kapitole 2 je popsána hlavní myšlenka práce, metoda výpočtu reputačního skóre v obecné podobě, tj. bez ohledu na typ entity, a je zde uvedena její formální definice. Další části práce se pak zabývají aplikací této metody v konkrétním případě a souvisejícími implementačními činnostmi. Nejprve je v kapitole 3 popsán přínos v oblasti analýzy síťového provozu a detekce útoků, především je kapitola věnována systému NEMEA. V kapitole 4 je pak přestaven systém reputační databáze NERD. Následuje popis a analýza dat použitých v této práci, a to v kapitole 5. Na základě kontextu a analýzy dat z předchozích kapitol je pak v kapitole 6 obecná metoda výpočtu FMP skóre konkretizována pro případ IPv4 adres a dat ze systémů Warden a NERD. Vyhodnocení vlastností této metody podle různých kritérií je provedeno v kapitole 7. Celou práci pak shrnuje a uzavírá kapitola 8.

2 Obecná metoda vyhodnocování reputace síťových entit

Z motivace v úvodu této práce vyplývá potřeba číselného hodnocení reputace jednotlivých IP adres, případně i jiných identifikátorů, které by vyjadřovalo míru hrozby asociovanou s danou entitou, využívalo všech dostupných informací a navíc bylo prediktivní, tedy vyjadřovalo očekávanou míru hrozby v nejbližší budoucnosti. Ačkoliv určité metody hodnocení repoutace síťových entit již existují, žádná nemá všechny potřebné vlastnosti. Tato kapitola se zabývá návrhem nové metody hodnocení reputace, která tyto požadavky splňuje.

V této kapitole je metoda představena v obecné formě, aplikovatelná na různé typy síťových entit. V kapitole 6 je pak tato obecná metoda konkretizována pro použití s IPv4 adresami a bezpečnostními hlášeními ze sdílecího systému.

2.1 Základní koncept

Hlavní myšlenkou navržené metody je ohodnotit každou entitu číslem, které vyjádří míru škodlivosti či hrozby, kterou tato entita představuje. Konkrétně by toto číslo mělo odpovídat *pravděpodobnosti, že bude daná entita vykazovat škodlivé chování během určitého časového intervalu v blízké budoucnosti (predikční okno)*. Tuto pravděpodobnost, a tedy výsledek navrhované metody pro hodnocení reputace síťových entit, označujeme jako *Future Misbehavior Probability score* (zkráceně *FMP skóre*).

Definice skóre pomocí pravděpodobnosti budoucího škodlivého chování je zvolena proto, že právě predikce budoucích útoků je důležitá pro prevenci a obranu (minulým útokům již nezabráníme), protože je ale predikce vždy založená jen na informacích z minulosti, funguje zároveň tato pravděpodobnost dobře i jako shrnutí předchozích aktivit dané entity.

Teoreticky je možné do skóre zahrnout kromě pravděpodobnosti budoucích útoků i očekávanou míru jejich závažnosti. To je však velmi problematické. Neexistuje totiž žádný obecně použitelný způsob hodnocení míry závažnosti kyberbezpečnostních událostí, ta vždy závisí na konkrétních vlastnostech cílové sítě, v ní aplikovaných pravidlech a mnoha dalších faktorech, a jde tedy o věc značně subjektivní. Tato práce se tedy závažností predikovaných útoků přímo nezabývá. Částečně je ale problematika závažnosti pokryta tím, že jsou predikovány pravděpodobnosti různých typů útoků zvlášť (viz dále).

2.1.1 Výpočet FMP skóre

Hodnocení entit pomocí FMP skóre je tedy založeno na předpovědi jejich budoucího škodlivého chování. Tato předpověď by měla být založena na všech dostupných datech o dané entitě – především na informacích o jejím předchozím škodlivém chování, ale i na dalších relevantních datech, která lze k entitě získat.

Způsob odvození pravděpodobnosti budoucích útoků na základě takových dat však nemusí být přímočarý a navrhnout příslušný prediktor ručně by bylo velmi obtížné. Obvykle však není problém získat velké množství dat o předchozích škodlivých činnostech entit z historických záznamů, predikční model je tedy možné vytvořit pomocí metod strojového učení s učitelem. Tento přístup je použit v této práci.

Ideální prediktor, tedy takový, který dokáže vždy přesně předpovědět budoucí chování dané entity, by přiřazoval FMP skóre pouze s hodnotami 1.0 nebo 0.0, podle toho, jestli se entita bude nebo nebude během predikčního okna chovat škodlivě. Takový prediktor je však v praxi nedosažitelný. Jakýkoliv reálný prediktor může pouze určit pravděpodobnost škodlivého chování na základě jemu dostupných informací v čase predikce. Cílem je tedy navrhnout co nejpřesnější prediktor ve smyslu co nejlepšího odhadu pravděpodobnosti pro všechny entity.

Dále je vhodné poznamenat, že v praxi je nemožné vědět o veškerém škodlivém chování dané entity, jsou známy jen ty události, které byly detekovány a byla přijata příslušná hlášení. Část útoků mohla zůstat nepovšimnuta, buď kvůli nedokonalosti detektorů, nebo jednoduše proto, že zdroj i cíl útoku leží mimo monitorovanou síť. Ve výsledku je tedy možné predikovat pouze budoucí *hlášení* související s danou entitou, nikoliv skutečně provedené útoky.

2.1.2 Varianty

FMP skóre může být obecné, předpovídající jakýkoliv typ škodlivé aktivity, resp. jakoukoliv kategorii hlášení, nebo specifické jen pro konkrétní typ. Například můžeme určit FMP skóre v kontextu DDoS útoků a zvlášť FMP skóre v kontextu skenování portů, každé odpovídající pravděpodobnosti budoucích útoků daného typu. Podobně je možné mít různá FMP skóre pro specifické cíle, určující pravděpodobnost útoků například pro konkrétní podsítě či typy služeb. Pokud je

v textu potřeba taková FMP skóre rozlišit, je možné použít dolní index, např. FMP_{scan}. Ve zbytku této kapitoly však mezi těmito variantami nebudeme rozlišovat, protože jediný rozdíl je v tom, co konkrétně je považováno za škodlivé aktivity, které mají být predikovány.

Důležitým parametrem FMP skóre je také délka predikčního okna. Tu je třeba vždy zvolit s ohledem na očekávané použití. V této práci je uvažována délka predikčního okna 24 hodin, což autor považuje za hodnotu vhodnou pro většinu aplikací. V praxi by tak mělo být FMP skóre pro každou entitu aktualizováno alespoň jednou denně, vždy na další predikční období.

2.2 Formální definice

Hlavním vstupem pro výpočet FMP skóre jsou hlášení o škodlivých aktivitách prováděných konkrétními entitami. Tato hlášení mohou mít různé formáty a obsahovat různé informace, pro účely navrhované metody však musí obsahovat minimálně: (*i*) čas detekce, t, a (*ii*) identifikátor entity (např. IP adresu), která je hlášena jako zdroj dané aktivity, e. Dále je vhodné, pokud hlášení obsahují: (*iii*) typ či kategorii hlášené události, c, (*iv*) objem či intenzitu události, v (přesný význam závisí na typu události, může to být např. počet pokusů o navázání spojení), a (*v*) identifikátor detektoru, d. V následujícím textu předpokládáme, že hlášení obsahují všech pět těchto atributů, ale metoda může být s určitými omezeními aplikována i pokud jsou dostupné jen první dva.

Hlášení (angl. *alert*) tedy můžeme definovat jako pětici $a = (t, e, c, v, d)$. Množinu všech dostupných hlášení označme A. Čas, ve kterém je prováděna predikce (*aktuální čas* či *čas predikce*), je označován jako t_0. *Predikční okno*, T_p, je definováno jako časový interval délky w_p bezprostředně následující po t_0, tedy $T_p = (t_0, t_0 + w_p)$. Prediktor využívá jako vstup informace o hlášeních přijatých v určitém období v minulosti, v *historickém okně*, $T_h = (t_0 - w_h, t_0)$, kde w_h je délka tohoto okna.

Jeden *vzorek* dat (*sample*) je definován jako souhrn vlastností entity e v konkrétní čas predikce t_0. Každý vzorek je reprezentován tzv. *feature vectorem*[1] $\mathbf{x}_{e,t_0} = (x_1, x_2, ..., x_k)_{e,t_0}$. Tento vektor se skládá z různých atributů získaných jak z dat o hlášeních přijatých během historického časového okna tak z dalších doplňujících dat známých o dané entitě v čase t_0 (více o volbě a výpočtu atributů je uvedeno v kap. 2.3).

Výstup, který má být předpovězen, y_{e,t_0}, je binární hodnota označující, zda k dané entitě existuje nějaké hlášení v příslušném predikčním okně,

$$y_{e,t_0} = \begin{cases} 1 & \text{pokud } \exists a \in A : a = (t, e, \cdot, \cdot, \cdot), t \in T_p \\ 0 & \text{jinak.} \end{cases} \tag{1}$$

[1] Česky např. *vektor rysů* či *vektor atributů*, žádný český ekvivalent však v oboru ustálený není a proto je dále v tomto textu používán původní anglický výraz. Jednotlivé prvky vektoru pak budou nazývány *atributy*.

V případě, že má být vypočítáno FMP skóre specifické pro určitý kontext, může být výše uvedená podmínka více omezující, např. kategorie hlášení musí mít konkrétní hodnotu. Vzorky, pro něž platí $y_{e,t_0} = 1$, náleží do tzv. *pozitivní třídy*, ostatní tvoří tzv. *negativní třídu*.

Úlohou pro strojové učení je vytvořit model, který dokáže pro zadaný feature vector $\mathbf{x}_{e,t_0}$ co nejpřesněji odhadnout pravděpodobnost, že $y_{e,t_0} = 1$, tedy že daná entita bude v predikčním okně nahlášena jako škodlivá. Tato úloha je v oblasti strojového učení označována jako *odhad pravděpodobnosti binárních tříd* (angl. *binary class probability estimation problem*). Jde vlastně o běžnou klasifikaci do dvou tříd, kdy nás však nezajímá přiřazení jedné konkrétní třídy každému vzorku, ale spíš pravděpodobnost příslušnosti vzorků do jednotlivých tříd.

Výstup predikčního modelu, označený $\hat{y}_{e,t_0}$, je odhadem pravděpodobnosti pozitivní třídy pro příslušný feature vector,

$$\hat{y}_{e,t_0} \approx p(y_{e,t_0} = 1 | \mathbf{x}_{e,t_0}). \tag{2}$$

Pro vytvoření prediktoru je použit standardní proces strojového učení. Nejprve je potřeba připravit datovou sadu pro trénování modelu. To v tomto případě znamená, že je vybráno několik časových okamžiků v rozmezí, ze kterého jsou dostupná data. Označme je jako vzorkovací časy, $T_s = t_1, ..., t_m$. Pro každý takový časový okamžik, $t_0 \in T_s$, je pak pro každou entitu $e \in E$ vypočítán feature vector a příslušná třída, $(\mathbf{x}_{e,t_0}, y_{e,t_0})$. Tím vznikne datová sada o počtu $|E \times T_s|$ vzorků. Dále bude pro označení vzorku pro jednoduchost používán pouze index i, tedy např. $\mathbf{x}_i$ a y_i.

Takto vytvořená datová sada je pak náhodně rozdělena na dvě části, trénovací a testovací sadu, z nichž první je použita pro natrénování modelu, druhá pro jeho vyhodnocení.

V této obecné části návrhu metody není doporučen žádný konkrétní model strojového učení. Pro různé typy dat mohou být vhodné různé modely a jejich konfigurace, obvykle je proto nutné provést experimenty s různými modely a vybrat ten s nejlepšími výsledky.

Pro vyhodnocení kvality predikce lze použít metriku zvanou *Brierovo skóre* (BS). V případě binárního problému a za předpokladu označení tříd čísly 0 a 1 je BS definováno jako:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2, \tag{3}$$

kde N je počet vzorků použitých k vyhodnocení. Brierovo skóre nabývá hodnot mezi 0 a 1, nižší hodnoty znamenají lepší predikci, tedy přesnější odhad pravděpodobnosti. Cílem při trénování modelu je tedy minimalizace BS.

Jakmile je predikční model natrénován a dosahuje uspokojivých výsledků, může být použit pro přiřazování FMP skóre novým vzorkům síťových entit. Pro

každou novou entitu je tedy vypočítán odpovídající feature vector, popisující předchozí s ní související hlášení a další dostupné informace, a ten je použit jako vstup natrénovaného modelu. Výstup modelu, $\hat{y}$, je pak přímo použit jako FMP skóre dané entity v daný predikční čas,

$$FMP(e, t_0) = \hat{y}_{e,t_0} = f(\mathbf{x}_{e,t_0}), \tag{4}$$

kde funkce f reprezentuje natrénovaný model.

Protože charakteristiky chování škodlivých entit, na nichž je založena predikce, se mohou v čase měnit, je v praktickém nasazení nutné v pravidelných intervalech přetrénovávat model nad aktuálními daty.

Pokud je vyžadován výpočet různých FMP skóre pro specifický kontext (např. předvídajících zvlášť jednotlivé typy útoků), je nutné pro každé takové skóre natrénovat samostatný model. Vzorky použité pro trénování takových modelů jsou označeny jako pozitivní ($y_i = 1$) pouze tehdy, pokud v predikčním okně existuje hlášení splňující příslušná kritéria, např. ohlašuje konkrétní typ útoku, ostatní hlášení jsou ignorována (vzorky mají $y_i = 0$).

2.3 Návrh feature vectoru

Jak již bylo zmíněno, feature vector použitý jako vstup pro predikci budoucích hlášení o škodlivém chování síťové entity obsahuje dva základní typy informací: (*i*) atributy odvozené z informací o předchozích hlášeních vztahujících se k této nebo k podobným entitám (např. k sousedním IP adresám) a (*ii*) atributy odvozené z jiných informací než z hlášení (např. zda je entita na nějakém veřejném blacklistu).

Konkrétní množinu atributů je nutné vždy navrhnout specificky pro daný typ entity, tedy např. zda jde o IP adresy či doménová jména, a s ohledem na to, jaká data jsou k dispozici. Zejména v případě atributů založených na hlášeních je však možné navrhnout doporučenou skupinu atributů, které by měly být použitelné ve většině případů.

Jako základní informace, které by měly být vždy obsažené ve feature vectoru, tedy navrhujeme použít:

- Počet hlášení
- Celkový objem či intenzita nahlášených událostí
- Počet detektorů, které příslušná hlášení vygenerovaly
- Čas od posledního hlášení
- Průměr a medián intervalů mezi jednotlivými hlášeními v historickém okně.

První tři hodnoty mohou být vypočítány přes různě dlouhá časová období, např. počty za poslední den a za celé historické okno. Pro každý interval tak dostaneme samostatný atribut. Případně je možné se na tyto hodnoty dívat jako

na časovou řadu, spočítat tedy např. počet hlášení za každý den v historickém okně, a pak za atribut použít *exponenciálně vážený plovoucí průměr* (*exponentially weighted moving average*, EWMA) této časové řady.

Při výpočtu všech výše zmíněných atributů by měla být brána v úvahu ta hlášení, která hlásí jako škodlivou právě tu entitu, pro kterou je tvořen feature vector. Tak však zachytíme pouze korelace mezi hlášeními o stejné entitě v čase. Pokud pro daný typ entity dává smysl zabývat se i korelacemi v prostoru, tedy že existují korelace mezi chováním blízkých či obecně nějak podobných entit, je možné jako další část feature vectoru použít stejné atributy, avšak beroucí v úvahu i hlášení o všech ostatních entitách, které jsou dle nějakého kritéria dostatečně podobné té entitě, pro kterou je vektor počítán (v případě IP adres to mohou být např. všechny IP adresy ve stejném /24 prefixu nebo ve stejném autonomním systému).

Mnohé z atributů mohou nabývat velmi vysokých hodnot (např. počet hlášení či jejich celkový objem) což může některým metodám strojového učení činit problémy. Všechny takové hodnoty proto doporučujeme před zpracováním transformovat funkcí $log(x + 1)$. V případě časových intervalů mezi hlášeními (které bývají nastaveny na nekonečno, pokud byla přijata méně než dvě hlášení) je pak vhodnější použít spíše funkci $exp(-x)$.

3 Inovace v oblasti detekce nežádoucího provozu

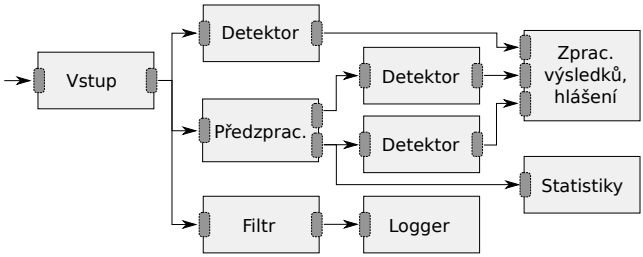
Obecný princip ohodnocování reputace prostřednictvím FMP skóre, navržený v předchozí kapitole, je dále v této práci konkretizován pro hodnocení reputace škodlivých IPv4 adres na základě hlášení z různých detekčních systémů a dalších doplňujících dat z reputační databáze. Tato kapitola se věnuje získávání hlášení, tedy oblasti detekce bezpečnostních událostí.

Hlavním výsledkem práce autora v této oblasti je nový framework pro snadnou implementaci detekčních nástrojů využívajících data o IP tocích a řada v něm implementovaných detektorů, včetně několika zcela nových detekčních metod. Ty výrazně pomohly v získání vhodné a dostatečně velké datové sady pro další části této práce, samozřejmě jsou však i samy o sobě významným přínosem. Následující podkapitoly popisují tento framework a vybrané případy jeho využití.

3.1 Systém pro proudové zpracování dat o síťových tocích (NEMEA)

Framework NEMEA (*Network Measurements Analysis*) [9]. byl navržen jako platforma pro snadné a efektivní zpracování dat o IP tocích, především za účelem detekce různých bezpečnostních událostí. Mezi jeho hlavní vlastnosti patří:

- Modulární flexibilní architektura, snadná rozšiřitelnost



Obrázek 1: Příklad několika modulů NEMEA a jejich propojení

- Proudové zpracování flow dat v reálném čase
- Vysoká propustnost
- Podpora flow dat rozšířených o položky z aplikační vrstvy

NEMEA je navržena jako heterogenní modulární systém. Každý modul implementuje nějaký konkrétní úkol, například předzpracování dat, filtrování, detekci konkrétního typu útoků či anomálií, hlášení výsledků apod. Moduly běží jako nezávislé procesy a předávají si mezi sebou data pomocí jednosměrných komunikačních rozhraní. Data jsou předávána jako potenciálně nekonečný proud jednotlivých zpráv – flow záznamů, hlášení o detekovaných událostech apod. Příklad jednoduché instance systému NEMEA je znázorněn na obrázku 1.

Každá instance systému NEMEA je složena z různých modulů, které mohou být různě propojeny. Obvykle bývají moduly propojeny do stromové struktury nebo acyklického orientovaného grafu s jedním modulem sloužícím jako hlavní vstup dat do systému. Tento modul sbírá, případně generuje, flow data a posílá je ke zpracování dalším modulům. Na druhé straně systému jsou pak obvykle moduly pro záznam výsledků do souborů, databáze, či pro odesílání hlášení emailem či do jiných systémů.

Systém je založen na principu proudového zpracování dat. Standardně jsou tedy všechna data mezi moduly předávána přímo v operační paměti, bez nutnosti meziukládání na disk nebo do databáze. To umožňuje zpracovávat v reálném čase data i z velkých sítí na jediném výkonném serveru. V případě potřeby je však možné celý systém i distribuovat – jednotlivé moduly mohou běžet na různých serverech a komunikovat přes síť.

V typickém nasazení pro velké sítě je jako vstupní modul použit plugin pro kolektor IPFIXcol, který získává NetFlow nebo IPFIX data ze sond rozmístěných na různých místech v síti a posílá je dalším modulům. Při monitorování malých sítí či pro testovací účely je možné jako hlavní vstup použít i modul *flow meter*, který monitoruje přímo pakety na lokálním síťovém rozhraní (případně čte

pcap soubor), vytváří z nich flow záznamy a odesílá je přímo ostatním NEMEA modulům.

Komunikace mezi moduly je implementována vlastní knihovnou TRAP (*Traffic Analysis Platform*). Ta slouží jako sjednocující vrstva pro různé způsoby meziprocesové komunikace, jako jsou UNIX domain sockety, TCP sockety, nebo např. i ukládání a čtení do a ze souboru, a poskytuje další obecné funkce používané NEMEA moduly. Jako datový formát je většinou modulů používán vlastní, vysoce efektivní, ale zároveň flexibilní, binární formát UniRec. Lze však posílat i JSON či obecná nestrukturovaná data.

Celý systém je volně dostupný jako *open-source*[2]. Součástí základní distribuce systému je množství modulů pro běžné úlohy zpracování dat i několik modulů pro detekci nejběžnějších typů škodlivého provozu. Zamýšleným hlavním přínosem je však poskytnutí jednotné platformy pro analýzu flow dat, nad níž může výzkumná komunita implementovat stávající i zcela nové metody analýzy a tyto implementace sdílet, čímž systém podporuje výzkum v této oblasti.

3.2 Vybrané možnosti použití systému NEMEA

NEMEA byla navržena především jako framework pro snadnou implementaci metod pro detekci škodlivého provozu, už základní systém však obsahuje množství modulů, včetně některých jednoduchých detektorů. Tato kapitola stručně shrnuje možnosti, jak lze systém NEMEA v praxi použít.

Mezi nejčastější nežádoucí aktivity na internetu patří skenování sítě. NEMEA poskytuje moduly pro detekci jak horizontálního, tak vertikálního skenování portů. Podstatně škodlivějším typem útoku detekovatelným na síťové vrstvě jsou volumetrické útoky odepření služby – *Denial of Service* (*DoS*) a *Distributed DoS* (*DDoS*). Systém NEMEA poskytuje pro detekci DoS útoků několik možností, jednoduché útoky dokáže detekovat modul HostStats, pracující na principu výpočtu profilů provozu jednotlivých IP adres a následném porovnání těchto profilů s definovanými pravidly, jiný modul umožňuje detekovat překročení nastavených limitů množství provozu pro jednotlivé podsítě, a v době psaní tohoto textu je ve vývoji detektor, jehož metoda je inspirována nástrojem FastNetMon. Dále je k dispozici modul specializující se na detekci amplifikačních útoků. S pomocí tradičních flow dat obsahujících informace po transportní vrstvu lze detekovat i některé útoky na konkrétní aplikace. V systému NEMEA byl například implementován detektor slovníkových útoků na autentizované služby, jako např. SSH.

Pokud jsou k dispozici flow data rozšířená o informace z hlaviček vybraných aplikačních protokolů, lze systém NEMEA využít pro detekci řady dalších útoků. Typickým příkladem je např. protokol DNS, analýzou jehož provozu lze odhalit řadu škodlivých činností – například tunelování dat běžného provozu

[2] https://github.com/CESNET/Nemea

pomocí DNS dotazů a odpovědí. Metoda pro detekci takových tunelů byla také implementována jako modul pro systém NEMEA [10].

Dalším příkladem protokolu, na nějž bývá směřována řada útoků, je protokol *Session Initiation Protocol* (SIP). Byly navrženy NEMEA moduly pro detekci skenování uživatelských jmen, slovníkové hádání hesel [11] a pokusy o zneužití SIP ústředen hádáním vytáčecího schématu [12].

Pro detekci škodlivého provozu však není vždy nutné navrhovat speciální metodu a implementovat nový modul, často stačí základní moduly, např. obecný filtr. Tímto způsobem byly například v roce 2014 detekovány pokusy o zneužití známé zranitelnosti – *Shellshock*. Stačilo pomocí filtrovacího modulu vyhledávat jistý regulární výraz v hlavičkách protokolu HTTP (které již v té době byly v datech o síťových tocích v síti CESNET k dipozici). Bylo tak odhaleno velké množství skenů testujících zranitelnost i mnoho skutečných pokusů o zneužití (snaha o spuštění kódu, který na serveru stáhne a spustí malware). Dalším příkladem ad-hoc detekce s pomocí filtrovacího modulu je zaznamenávání přístupů na IP adresy a URL *command & control* serverů získané při předchozí analýze malware, což umožňuje odhalení zařízení v monitorované síti, která jsou tímto malware napadena.

3.3 Zpracování hlášení o detekovaných událostech

Detekční moduly zpravidla generují jako výstup jednoduché záznamy s informacemi popisujícími detekované bezpečnostní události. Další zpracování těchto záznamů, jako je např. logování či hlášení operátorům, probíhá v systému NEMEA unifikovaným způsobem prostřednictvím speciální sady modulů. Ty převádí výstupy jednotlivých detektorů do jednotného formátu a výsledné zprávy mohou ukládat do souboru či do databáze, odeslat je emailem, nebo do systému pro sdílení hlášení Warden.

3.4 Shrnutí přínosu

Do termínu odevzdání této práce (prosinec 2018) přispěl systém NEMEA ke vzniku 9 recenzovaných publikací na vědeckých konferencích a 7 dalších odborných publikací. Na 5 z nich se přímo podílel i autor této práce. Systém NEMEA byl také využit v 20 bakalářských a 15 diplomových pracech (z nichž 8, resp. 2, vedl autor této práce, u několika dalších byl konzultantem), jejichž náplní byl zpravidla vývoj nějakého NEMEA modulu, příp. rozšíření frameworku.

Většina z vyvinutých detektorů je v současnosti nasazena pro monitorování provozu v české akademické síti CESNET. Začíná se používat i v několika jiných sítích, a to nejen v ČR. Hlášení o škodlivém provozu generovaná detektory v síti CESNET jsou denně používána v praxi pro zajišťování bezpečnosti sítě a také tvoří významnou část dat použitých pro výzkum popisovaný v následujících kapitolách.

4 Reputační databáze síťových entit

Reputační skóre navrhované v této práci je přiřazováno jednotlivým entitám, jako jsou IP adresy, sítě, domény apod., a mělo by být založeno na všech dostupných informacích o těchto entitách. Je tedy třeba udržovat profily nahlášených zdrojů škodlivého chování, v nichž budou tyto informace udržovány. Nejde přitom jen o agregaci hlášení, ale i o získávání dalších relevantních dat z jiných zdrojů.

Systém udržující takové profily můžeme nazývat reputační databází. Ačkoli vytváření profilů škodlivých IP adres a jiných entit na základě dostupných informací samozřejmě není nová myšlenka, žádná veřejně dostupná databáze tohoto typu, ani žádná otevřená implementace, kterou by bylo možné nasadit pro zpracování vlastních dat, nebyla k dispozici.

Autor této práce proto se proto mimo jiné věnoval i návrhu a implementaci právě takové otevřené reputační databáze. Jejím účelem je jednak poskytnutí dat pro tuto disertační práci, ale také praktické využití bezpečnostními týmy při jejich každodenním boji proti kybernetickým hrozbám.

Původní myšlenka takového systému, stejně jako ohodnocování reputace IP adres, byla publikována v [13], plánované vlastnosti systému byly také představeny v [14]. Tato kapitola stručně popisuje vlastnosti a architekturu této reputační databáze.

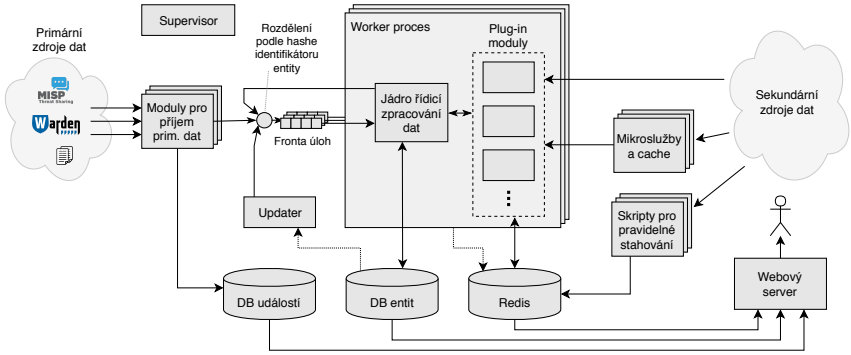
4.1 Systém NERD

Z uživatelského hlediska je systém NERD (*Network Entity Reputation Database*) webový portál, ve kterém může kdokoliv vyhledat jakoukoliv IP adresu, doménové jméno, či jiný síťový identifikátor (obecně *entitu*) a získá všechny dostupné informace týkající se dané entity a související nějak s bezpečností – například seznam všech souvisejících bezpečnostních hlášení informace o přítomnosti entity na nějakém blacklistu, informace z DNS, databází whois, geolokační data apod. Systém také umožňuje vyhledat všechny entity splňující zadaná kritéria a seřadit je podle různých atributů.

Za tímto portálem je komplexní modulární systém, který slouží k získávání dat z nejrůznějších zdrojů, jejich zpracování a uložení do databáze a k pravidelnému obnovování. Různé aspekty tohoto systémy jsou popsány v následujících podkapitolách.

4.2 Typy ukládaných dat

Základní datovou jednotkou je *záznam entity* – strukturovaný záznam (formát JSON) uchovávající všechny informace o konkrétní entitě. Entitou může být např. IP adresa, síťový prefix, číslo autonomního systému (ASN), doménové jméno apod. Každý záznam entity se skládá z množství *atributů*.



Obrázek 2: Architektura systému NERD

Zdroje dat ukládaných do záznamů lze rozdělit do dvou tříd – *primární* a *sekundární*. Primární datové zdroje jsou ty, které označují určité síťové entity jako škodlivé a na základě nichž jsou vytvářeny záznamy entit v databázi. Tyo jsou především hlášení z detekčních systémů.

Sekundární zdroje dat jsou ty, z nichž se získávají dodatečné informace o již známých entitách. Jde například o DNS dotazy pro zjištění doménového jména navázaného k IP adrese či naopak, geolokaci nebo dotazy do blacklistů a jiných databází.

Některé atributy nepochází přímo z primárních ani sekundárních zdrojů, ale jsou odvozeny lokálně z ostatních atributů. Příkladem může být přiřazování štítků podle převažujícího typu škodlivé aktivity nebo FMP skóre odvozené z ostatních dat pomocí metody prezentované v kapitole 6. Tyto atributy se nazývají *odvozené atributy*.

U některých atributů, jejichž hodnota se může často měnit, jako je například přítomnost entity na různých blacklistech, je udržována nejen aktuální hodnota, ale i historie všech předchozích hodnot za určité období.

4.3 Architektura

Kvůli potřebě vysoké flexibility a snadné škálovatelnosti je systém NERD založen na principu modulární architektury. Skládá se z množství vzájemně spolupracujících komponent, zpravidla pracujících jako samostatné procesy. Architektura systému je znázorněna na obrázku 2.

Součástí systému je několik databází. V té hlavní, *databázi entit*, jsou uloženy všechny záznamy entit, jak byly popsány v kapitole 4.2. Dále je zde samostatná databáze pro ukládání originálních dat z primárních zdrojů, tedy přijatých hlá-

šení. Nakonec je tu rychlá *in-memory key-value* databáze (Redis), pro ukládání různých, obvykle krátkodobých či rychle se měnících, pomocných dat.

Hlavní vstup systému je reprezentován množinou modulů pro příjem primárních dat. Ty přijímají zprávy, hlášení, či seznamy entit z externích zdrojů a do globální fronty vkládají *úlohy* požadující vytvoření nebo úpravu záznamů entit (tzv. *update requests*).

Tyto úlohy jsou zpracovávány jádrem systému – množinou pracovních procesů (*workers*). Ty aplikují požadované změny nad záznamy entit. Dále také obohacují záznamy o data ze sekundárních zdrojů a vypočítávají odvozené atributy. To je prováděno pomocí množiny zásuvných modulů, díky čemuž je snadné přidat, změnit či odstranit sekundární datové zdroje či pravidla pro odvozování atributů jen změnou těchto modulů.

Tyto worker procesy mohou běžet paralelně v libovolném množství, díky čemuž je výkon systému snadno škálovatelný.

Většina zásuvných modulů slouží k získávání dat z externích zdrojů, a to zpravidla buď přímými dotazy do rozhraní těchto zdrojů, protřednictvím speciální mikroslužby či cahce, nebo jsou data zvláštním skriptem pravidelně stahována, předzpracována a uložena lokálně.

Pravidelné aktualizace dat v záznamech entit jsou řízeny komponentou *updater*. Ta kontroluje časové známky poslední aktualizace v záznamech a vydává úlohy (update requests) pro aktualizace daných atributů, pokud jsou starší než určitá doba.

Poslední komponentou systému je webový server. Jeho prostřednictvím jsou všechna data poskytována uživatelům přes webové grafické rozhraní nebo jiným systémům přes REST API.

5 Použitá data a jejich charakteristiky

Pro ověření obecného principu určování reputace síťových entit, představeného v kapitole 2, je v následujících kapitolách navržena a ověřena konkrétní varianta výpočtu FMP skóre pro ohodnocování IP adres. Pro návrh takové varianty je třeba nejprve znát charakteristiky chování škodlivých IP adres a typické vlastnosti souvisejících hlášení.

Cílem této kapitoly tedy je: (*i*) uvést, jaká konkrétní datová sada byla použita dále v této práci, a (*ii*) prostřednictvím analýzy této datové sady zjistit základní charakteristiky chování zdrojů škodlivého provozu.

Tato kapitola částečně vychází z analýzy dat ze systému Warden provedené v roce 2015 a popsané v technické zprávě [15]. Zde je však tato analýza zopakována nad novějšími daty, konkrétně z druhé poloviny roku 2017. Stejná data jsou pak použita i v kapitole 7 pro vyhodnocení metody určování reputace IP adres.

Tabulka 3: Počet hlášení, nahlášených IP adres a počet detektorů, které hlášení vygenerovaly, podle kategorie.

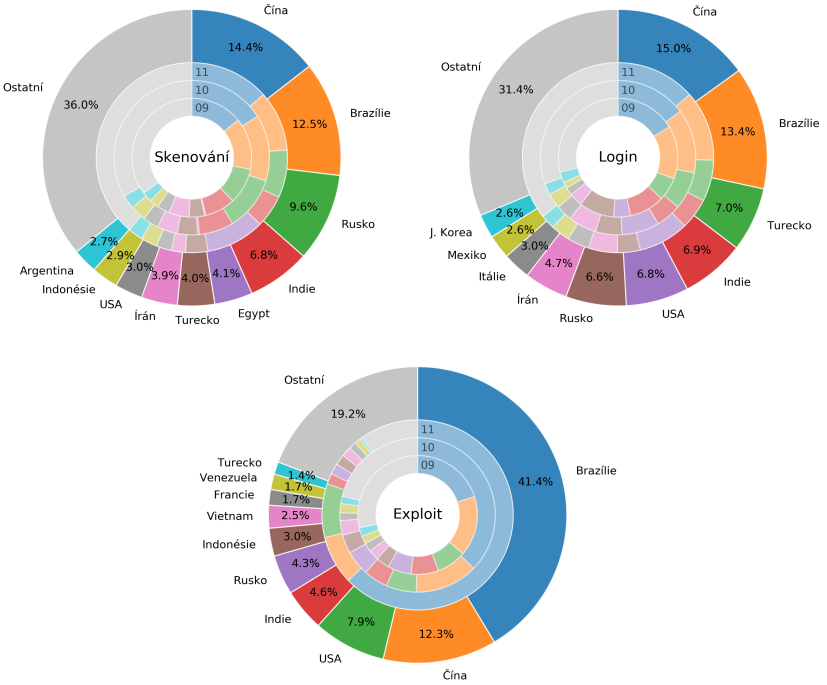
	skenování	login	exploit
září			
– počet hlášení	40 158 463	2 864 168	411 223
– počet adres	1 758 801	103 402	20 664
– počet detektorů	8	4	4
říjen			
– počet hlášení	51 326 622	2 566 588	130 221
– počet adres	1 801 762	92 512	20 946
– počet detektorů	10	6	5
listopad			
– počet hlášení	54 519 766	2 396 084	56 447
– počet adres	2 324 460	80 132	24 609
– počet detektorů	7	5	3

5.1 Datová sada

Základem pro metodu hodnocení reputace jsou data typu hlášení o bezpečnostních událostech. Data tohoto typu použitá v této práci pochází ze systému pro sdílení hlášení Warden, provozovaného sdružením CESNET. Do tohoto systému jsou svedena hlášení z více než 30 detekčních nástrojů různých druhů. Zdroje dat jsou umístěny jak přímo v síti CESNET a v kampusových sítích připojených univerzit, tak i u několika komerčních a zahraničních partnerů. Systém Warden celkově zpracuje kolem 1,7 milionu hlášení za den (cca 20 za sekundu). Přibližně 1/4 hlášení pochází z detektorů vyvinutých v rámci frameworku NEMEA (viz kap. 3). Jsou to přitom často unikátní typy dat, které nejsou detekovány jinými systémy (např. pokusy o neoprávněný přístup a DDoS útoky jsou z většiny hlášeny detektory v NEMEA).

Pro tuto práci byla použita data ze tří měsíců – od 1. 9. 2017 do 30. 11. 2017. Každý měsíc je zde analyzován samostatně. Porovnáním výsledků z různých časových období lze určit, jak jsou zjištěné charakteristiky stabilní v čase.

Ačkoliv jsou do systému Warden hlášeny události mnoha různých typů, pouze tři kategorie hlášení jsou dostupné v dostatečném počtu a kvalitě, aby mohly být použity v této práci. Jde o kategorie *skenování*, *login*, *exploit*, označující po řadě jakékoliv pokusy o skenování portů či sítě, neoprávněné pokusy o přihlášení k autentizovaným službám a pokusy o zneužití nějaké zranitelnosti. V tabulce 3 jsou uvedeny počty hlášení, počty nahlášených zdrojových IP adres, a počty detektorů pro jednotlivé kategorie hlášení.



Obrázek 4: Deset zemí s největším počtem škodlivých IP adres pro různé typy útoků.

Pro zachování jednoduchosti a stručnosti, pokud se dále v této práci zmiňují *škodlivé aktivity* či *útoky*, myslí se tím události jakékoliv kategorie, včetně skenování.

5.2 Geografické rozložení zdrojů škodlivého provozu

Pro každou IP adresu v datové sadě byla zjištěna její pravděpodobná geografická pozice na úrovni země, a to pomocí volně dostupné databáze GeoLite2 od společnosti MaxMind[3]. V grafech na obrázku 4 je znázorněno 10 zemí s nejvyšším počtem nahlášených adres pro jednotlivé typy útoků. Tři vnitřní kruhy vždy znázorňují poměr nejčastějších zemí v jednotlivých měsících (ve směru zevnitř ven: září, říjen, listopad), vnější kruh pak průměr přes všechny tři měsíce.

Při porovnání jednotlivých grafů si můžeme všimnout, že jsou zde značné podobnosti, například Čína a Brazílie jsou vždy na prvních dvou příčkách. Na

[3] https://dev.maxmind.com/geoip/geoip2/geolite2/

dalších pak vždy najdeme v první desítce Indii, Rusko, USA a Turecko, jejich pořadí a hlavně relativní zastoupení se už však výrazně liší. Obecně tedy můžeme říci, že geografické rozložení zdrojů síťových útoků se liší podle typu útoku.

Porovnání dat z jednotlivých měsíců odhalí, že sice v čase dochází k jistým změnám v geografickém rozložení škodlivých IP adres, až na několik výjimek jsou to však změny jen malé a celkově je rozložení poměrně stabilní.

Výše uvedené statistiky vycházejí z absolutního počtu škodlivých IP adres v dané zemi, logicky proto v grafech převažují poměrně velké země s velkým množstvím aktivních IP adres. Dále bylo proto zkoumáno i relativní zastoupení škodlivých adres vzhledem k celkovému počtu adres v jednotlivých zemích (grafy zde však z důvodu nedostatku místa nejsou uvedeny). I v těchto relativních počtech jsou mezi jednotlivými zeměmi velmi výrazné rozdíly. Některé země, velké i malé, se ukazují jako výrazně více prostoupené škodlivými adresami, než jiné, a to i o několik řádů (některé země mají až 2 % adres nahlášených jako zdroje skenování, zatímco u jiných jsou to jen tisíciny procenta). Zároveň platí, že toto relativní zastoupení se v jednotlivých měsících minimálně u některých zemí výrazně liší, způsob výpočtu statistik pro predikci by tedy měl vycházet jen z nedávných dat.

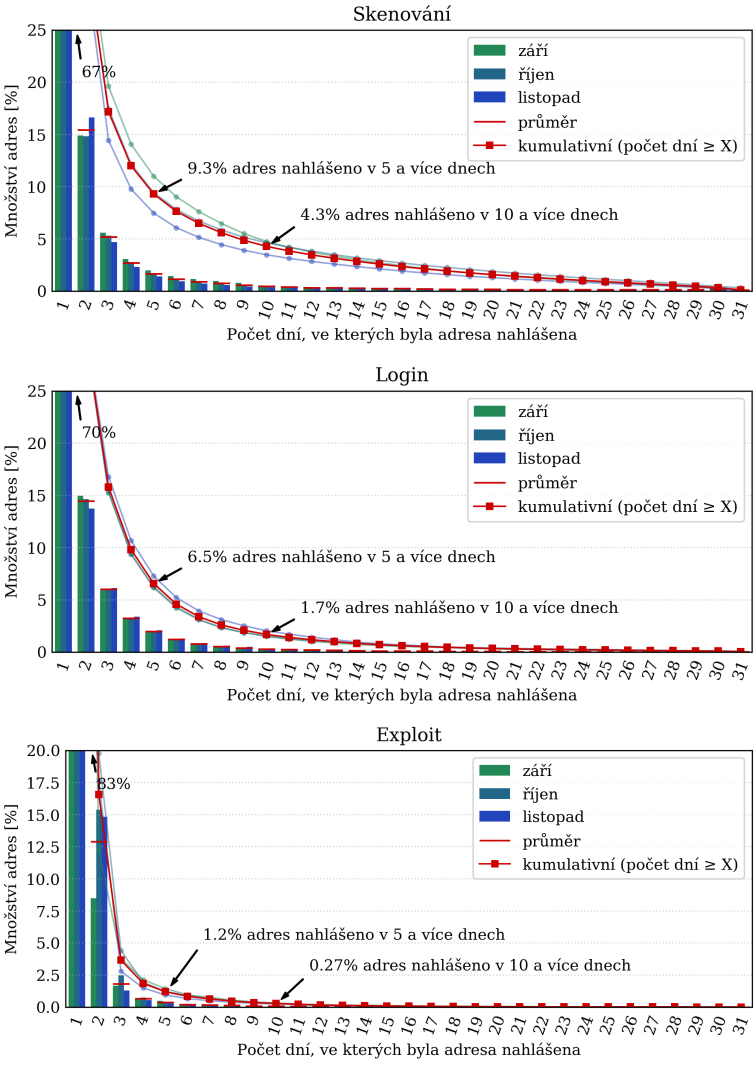
5.3 Korelace hlášení v čase

Pravděpodobně nejdůležitějšími informacemi pro vyhodnocení reputace síťové entity a pro predikci případných budoucích škodlivých aktivit jsou data o jejím předchozím chování. Intuitivně lze předpokládat, že entity škodící v nedávné minulosti budou v budoucnu zdrojem dalších útoků pravděpodobněji, než ostatní entity. Byla tedy provedena základní analýza korelací hlášení v čase, zaměřená především na to, zda, jak často a jak dlouho se v hlášeních opakují stejné zdroje útoků.

Pro studium těchto charakteristik byla hlášení v datové sadě rozdělena podle dne detekce, tzn. pro každý den v uvedených třech měsících byl vytvořen seznam IP adres, které byly v tento den nahlášeny jako zdroj daného typu útoku.

Pro zjištění, zda je běžné, že je stejná adresa hlášena opakovaně po delší dobu, bylo pro každou adresu spočítáno, v kolika dnech v určitém měsíci byla nahlášena jako škodlivá. Tyto počty jsou uvedeny na obrázku 5. Na vodorovné ose je uveden počet dnů v měsíci, výška každého sloupce ukazuje, kolik adres se v daném měsíci vyskytlo *právě* v tolika dnech. Čára označená jako „kumulativní“ hodnota pak ukazuje, jaké procento adres bylo nahlášeno v *alespoň* tolika dnech (jde tedy o součet výšky sloupce na dané pozici a všech sloupců napravo).

Ve všech případech je většina adres nahlášena jen v jediném dni, viz první sloupec v každém grafu (67 %, 80 %, resp. 83 % adres pro hlášení typu skenování, login a exploit). Navíc další analýzou lze zjistit, že velká část adres je nahlášena jen jedním jediným hlášením, konkrétně 44 %, 20 %, 57 % adres pro jednotlivé typy hlášení. Útoky z většiny adres tedy netrvají dlouho – buď proto, že jsou tyto



Obrázek 5: Rozložení množství adres podle počtu dní v měsíci, v nich byly nahlášeny jako škodlivé.

adresy skutečně škodlivé jen krátce (buď je příslušné nakažené zařízení rychle opraveno nebo, a to je pravděpodobnější, jde např. o mobilní zařízení, které se pohybuje po různých IP adresách), nebo může jít o útoky nízké intenzity, navíc rozptýlené přes velké množství cílových sítí, takže sítě pokryté detektory přispívajícími do systému Warden nevidí útok vícekrát než jednou za měsíc.

Na pravé straně grafů jsou pak adresy, které jsou hlášeny velmi často. Například v případě skenování je v průměru 9,3 % adres nahlášeno v pěti či více dnech v měsíci. Takové adresy již lze považovat za dlouhodobě aktivní útočníky a lze u nich očekávat podstatně vyšší pravděpodobnost dalších hlášení.

Podrobnější analýza dat navíc odhalí, že těchto 9,3 % (180 tis.) skenujících adres je zodpovědných za 65 % všech hlášení typu skenování. Podobně zatímco jen k 6,5 % (5 800) adres byl nahlášen pokus o přihlášení (login) v pěti a více dnech, tyto adresy jsou zodpovědné za 60 % příslušných hlášení. V případě hlášení typu exploit je pak pouhých 1,2 % (265) adres zodpovědných za rovných 50 % hlášení (vždy jde o průměr za všechny tři měsíce).

Porovnání jednotlivých grafů na obrázku 5 odhalí, že míra opakovaných hlášení o stejné adrese se liší podle typu útoku. Zatímco hlášení typu skenování a login vykazují podobný podíl adres nahlášených jen v jednom dni, čára znázorňující kumulativní hodnoty leží u skenování podstatně výš. To znamená, že pokud je nějaká skenující adresa hlášena opakovaně, pak její aktivita trvá v průměru déle, než u hlášení typu login. Hlášení typu exploit pak vykazují zdaleka nejmenší míru opakování – jen velmi málo adres je nahlášeno ve více než několika málo dnech.

6 Predikce škodlivého chování IP adres

Tato kapitola popisuje, jak je obecný princip hodnocení reputace síťových entit pomocí FMP skóre, navržený v kapitole 2, aplikovaný v konkrétní situaci – hodnocení IPv4 adres na základě hlášení ze systému Warden a některých dalších dat dostupných v systému NERD. Jde tedy o tzv. *proof-of-concept* navržené metody, ukázku a vyhodnocení jejího použití v praxi.

6.1 Zdroj dat a nastavení parametrů

Jako data pro experimenty byla použita hlášení ze systému Warden, podrobně popsaná v kapitole 5. Ta byla dále doplněna o data o nahlášených IP adresách ze systému NERD, např. přiřazená doménová jména či přítomnost adres na blacklistech.

Délka historického okna w_h je 7 dní, délka predikčního okna w_p je 1 den. Cílem je tedy určit pravděpodobnost, že k dané IP adrese bude přijato hlášení během následujících 24 hodin, a to na základě informací o hlášeních z předchozího týdne.

6.2 Příprava datové sady

Originální datová sada, tedy všechna hlášení za uvedené období, obsahuje 155 milionů hlášení z 23 různých zdrojů, Celkem je v ní nahlášeno 5,3 milionu různých IP adres. Jak bylo uvedeno v kapitole 5, naprostá většina hlášení (téměř 95 %) hlásí různé typy skenování sítě (kategorie *Recon.Scanning*). Dále jsou výrazně zastoupeny kategorie *Attempt.Login* a *Attempt.Exploit*. Ty jsou pro účely vyhodnocení sloučeny do jedné, pokusy o neoprávněný přístup.

Pro každou IP adresu tedy budou určovány dva typy FMP skóre, $FMP_{\text{skenování}}$, predikující hlášení kategorie *Recon.Scanning*, a $FMP_{\text{přístup}}$, predikující hlášení libovolné z kategorií *Attempt.**.

Připomeňme, že jeden *vzorek* pro strojové učení je množina vlastností IP adresy v konkrétní predikční čas. Pro přípravu datové sady tedy bylo zvoleno 24 časových okamžiků v rámci tří měsíců, ze kterých pochází použitá data, a každý byl použit jako jeden predikční časPro každý takový čas byl vytvořen seznam všech IP adres, které byly alespoň jednou nahlášeny daným typem hlášení v příslušném historickém okně (tedy v předchozích 7 dnech), a pro každou z nich byl vypočítán feature vector $\mathbf{x}_i$ a určena třída y_i.

Takto bylo získáno 12,3 milionu vzorků pro predikci hlášení typu *skenování* a 765 000 vzorků pro predikci hlášení typu *přístup*. Z obou těchto datových sad byla náhodně vybrána podmnožina vzorků, které jsou použity jako testovací sada (600 000 pro *skenování*, 100 000 pro *přístup*). Zbývající vzorky jsou použity pro trénování.

6.3 Feature vector

Návrh atributů feature vectoru odvozených z předchozích hlášení vychází z obecných doporučení z kapitoly 2.3. Pro každou kategorii hlášení (*skenování* a *přístup*) je pro každou IP adresu vypočítána následující sada atributů, přičemž jsou brána v úvahu hlášení obsahující danou adresu:

1. Počet hlášení v posledním dni
2. Celkový počet pokusů o navázání spojení (objem útoku) za poslední den
3. Počet detektorů, které tuto adresu nahlásily během posledního dne
4. Počet hlášení v posledním týdnu
5. Celkový počet pokusů o navázání spojení (objem útoku) za poslední týden
6. Počet detektorů, které tuto adresu nahlásily během posledního týdne
7. EWMA počtu hlášení za den (z dat za poslední týden)
8. EWMA celkového počtu pokusů o spojení za den (z dat za poslední týden)
9. EWMA binární posloupnosti označující přítomnost hlášení (0 nebo 1) v každém dni (z dat za poslední týden)

10. Čas od posledního hlášení (v počtu dní)
11. Průměrný interval mezi hlášeními v posledním týdnu (v počtu dní, nekonečno v případě méně než dvou hlášení)
12. Medián intervalů mezi hlášeními v posledním týdnu (v počtu dní, nekonečno v případě méně než dvou hlášení)

Tato sada atributů je doplněna ještě jednou podobnou sadou, pro kterou jsou však brána v úvahu všechna hlášení obsahující jakoukoliv IP adresu ze stejného /24 prefixu, jako má vyhodnocovaná adresa (tato délka prefixu byla zvolena jako nejvhodnější pro určování podobně se chovajících adres, stejná byla použita pro agregaci zdrojů či cílů útoku v řadě dřívějších prací, např. [7, 8, 6, 16]). Tato tzv. *prefixová* sada atributů obsahuje atributy 1–9 z předchozího seznamu a navíc tyto dva následující:

- Počet různých IP adres v daném prefixu nahlášených za poslední den
- Počet různých IP adres v daném prefixu nahlášených za poslední týden

Protože existují nezanedbatelné korelace mezi událostmi typu skenování a pokusy o přístup [17, 18], vždy jsou jako vstup použity atributy vypočítané z obou kategorií hlášení, bez ohledu na to, která kategorie má být predikována.

Další dva atributy využívají informací o geolokaci IP adres a jejich příslušnosti pod určitý autonomní systém (AS). Pro každou zemi a AS byla tedy vypočítána její tzv. *škodlivost*, která vyjadřuje poměr počtu škodlivých IP adres (dle hlášení přijatých v posledním týdnu) z dané země či AS vůči celkovému počtu adres v této zemi či AS. Jako vstupní atributy jsou pak u každé adresy použity tyto poměry pro zemi a AS, do nichž daná adresa náleží. Celkem tedy feature vector obsahuje 48 atributů odvozených z přijatých hlášení.

Další část vektoru tvoří několik atributů vycházejících z jiných zdrojů dat. Tyto jsou všechny binární, nabývají hodnotu 1, pokud je určitá vlastnost splněna, jinak 0. Zaprvé jde o přítomnost dané IP adresy na 5 veřejných blacklistech a na seznamu dynamicky přidělovaných adresních rozsahů.

Dále je pomocí DNS dotazů ke každé IP adrese zjištěno odpovídající doménové jméno a na něj je aplikována sada ručně navržených pravidel. Například jsou vyhledávána klíčová slova jako „static“, „dynamic“, „dsl“ nebo různými způsoby vložená IP adresa. Výsledkem jsou další 4 atributy.

Celkem se tedy feature vector skládá z 58 atributů, určených vždy pro konkrétní IP adresu a okamžik v čase.

6.4 Předzpracování dat, trénování a způsob použití prediktoru

Data jsou pomocí podvzorkování vyvážena (aby pozitivní a negativní třída měly stejný počet vzorků) a hodnoty některých atributů jsou nelineární transformací

Tabulka 6: Brierovo skóre různých modelů vypočítané na testovací části datových sad *skenování* a *přístup*

	skenování	přístup
NN, 2 vrstvy	0,06462	0,05486
NN, 3 vrstvy	0,06459	0,05424
GBDT(100, 3)	0,06713	0,05287
GBDT(200, 7)	0,06284	0,05065

normalizovány, jak bylo uvedeno v kap. 2.3. Pak jsou data využita pro natrénování daného modelu strojového učení, přičemž cílem trénování je minimalizace Brierova skóre vypočítaného přes všechny vzorky trénovací datové sady. Pro data typu *skenování* a *přístup* je vždy vytvořen samostatný model.

Při vyhodnocování jsou pak natrénovanému modelu předloženy vzorky testovací datové sady.

Při praktickém nasazení by se pak postupovalo stejně – ke každé IP adrese by byl vypočítán feature vector, ten by byl předložen předem natrénovanému modelu (či modelům, pro různé typy predikovaných útoků) a výsledek by pak byl použit jako FMP skóre dané adresy.

7 Vyhodnocení

V této kapitole jsou popsány experimenty vyhodnocující metodu určování FMP skóre IP adres na základě předchozích hlášení a dalších informací, jak byla definována v kapitole 6.

7.1 Vyhodnocení kvality predikčních modelů

Pro experimenty byly vybrány dvě třídy modelů strojového učení, *neuronové sítě* (NN) a tzv. *gradient boosted decision trees* (GBDT, také známé jako *xgBoost*), což je model založený na skupině (*ensemble*) mnoha rozhodovacích stromů. Bylo experimentováno s různými konfiguracemi těchto modelů. Tabulka 6 ukazuje Brierovo skóre několika vybraných modelů pro obě datové sady.

Neuronové sítě měly 2, resp. 3, skryté vrstvy, všechny sestávající z 56 uzlů (stejně jako je počet atributů feature vectoru) s aktivační funkcí typu ReLU. Výstupní vrstva má jeden uzel a aktivační funkci typu sigmoid. Modely typu GBDT sestávají ze 100 rozhodovacích stromů o maximální hloubce 3, resp. 200 stromů o maximální hloubce 7.

Model $GBDT(200, 7)$ dosahuje na obou datových sadách nejlepšího Brierova skóre, hodnoty jsou však u všech modelů velmi podobné a blízké nule, což znamená, že všechny uvedené modely dokáží kvalitně predikovat budoucí hlášení.

Pomocí kalibračních křivek pravděpodobnosti (zde z důvodu nedostatku místa neprezentovaných) bylo ověřeno, že předpovězená hodnota skutečně velmi dobře aproximuje skutečnou pravděpodobnost, že daná adresa bude během predikčního okna nahlášena jako škodlivá. Pokud tedy například pro skupinu vzorků model predikuje hodnoty blízké 0,6, skutečně přibližně 60 % z těchto vzorků nahlášených je, 40 % není. Výsledky pro datovou sadu *skenování* jsou v tomto ohledu velmi přesné, u datové sady přístup je v rozmezí predikované pravděpodobnosti od 0,4 do 0,9 přesnost nižší, kvůli malému počtu takových vzorků, celkově jsou však výsledky stále dostatečně dobré.

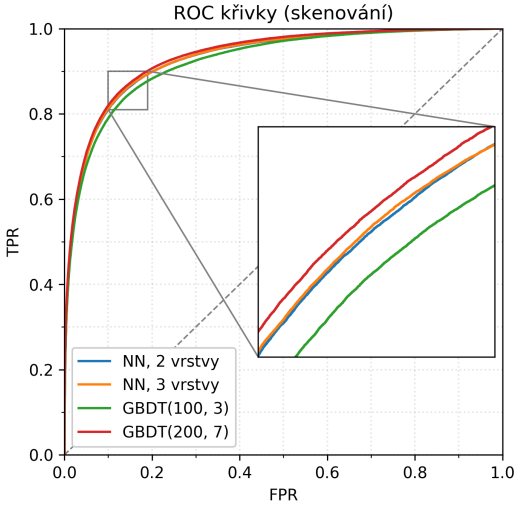
Na obrázcích 7 a 8 jsou uvedeny ROC křivky jednotlivých modelů, které ukazují vztah mezi poměrem pravdivě pozitivních (*true positives*, TP) a falešně pozitivních (*false positives*, FP) výsledků v případě, že je na skóre aplikována určitá mez, pomocí níž jsou vzorky rozděleny na škodlivé a neškodné. Každý bod křivky odpovídá určité hodnotě této meze. Čím více se křivka odchyluje od hlavní diagonály odchyluje a blíží se levému hornímu rohu, tím přesnější predikce je.

Všechny křivky jsou poměrně hladké a vzájemně velmi podobné. Jediný výraznější rozdíl je mezi datovými sadami, kdy hlášení typu *skenování* se zdají být snáze predikovatelné než hlášení typu *přístup*. Konkrétně lze z ROC křivek vyčíst například to, že pokud je v případě skenování mez nastavena tak, aby míra falešně pozitivních výsledků byla 10 %, je možné zachytit (a zablokovat v případě použití jako blacklistu) až 80 % všech opakujících se zdrojů skenováníJe důležité poznamenat, že falešně pozitivní výsledek zde nutně neznamená zablokování legitimní IP adresy, příslušná adresa může být stále škodlivá, pouze během predikčního okna neprovedla žádný útok vůči sledované síti. V takovém případě jde pouze o zbytečně zabrané místo na blacklistu. To umožňuje posunout mez i do oblasti s poměrně vysokým počtem falešně pozitivních výsledků a zablokovat tak naprostou většinu opakujících se zdrojů škodlivého provozu. Jedinou cenou za to je větší velikost blacklistů.

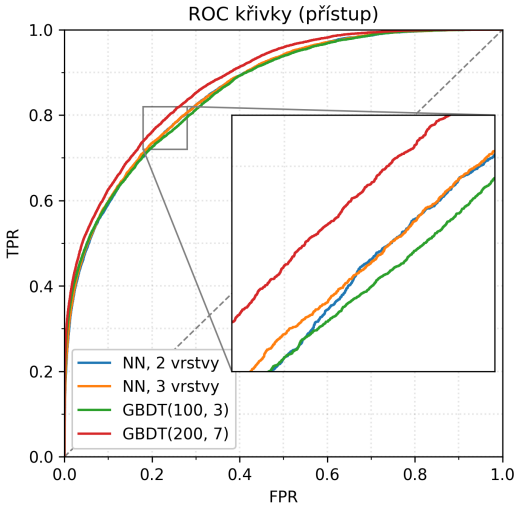
Dle všech kritérií dosahuje nejlepších výsledků model $GBDT(200, 7)$. Ve zbytku práce je tedy nadále používán pouze tento model.

Dále byl vyhodnocen i vliv jednotlivých vstupních atributů. Atributy byly rozděleny do několika skupin – atributy odvozené pouze z hlášení o dané adrese a o predikované kategorii, z hlášení o druhé z kategorií, z hlášení o ostatních adresách ve stejném prefixu, doplňující příznaky (přítomnost na blacklistech a příznaky odvozené z hostname) a škodlivost příslušné země a AS. Pak byla vyhodnocována kvalita prediktoru při použití různých kombinací těchto skupin.

Ukázalo se, že poměrně dobrých výsledků lze dosáhnout i se základními daty, pouze hlášeními o dané adrese a dané kategorii hlášení, každá z ostatních skupin však výsledky dokáže nezanedbatelně vylepšit. ZároveŇ však žádná z dodatečných skupin nevylepší výsledky tolik, jako kombinace všech. Z toho lze tedy odvodit, že žádná ze skupin atributů feature vektoru není zbytečná.



Obrázek 7: ROC křivky pro 4 různé modely a testovací datovou sadu *skenování*



Obrázek 8: ROC křivky pro 4 různé modely a testovací datovou sadu *přístup*

Tabulka 9: Porovnání blacklistů různých typů a velikostí pomocí metrik hit-count, hit-rate a procenta zablokovaných útočníků

N	blacklist	T	hit-count	hit-rate	% útočníků
100	FMP	0,99	100	100 %	2,3 %
	$GWOL_1$	–	83	83 %	1,9 %
	$GWOL_7$	–	71	71 %	1,6 %
500	FMP	0,65	443	89 %	10,1 %
	$GWOL_1$	–	236	47 %	5,4 %
	$GWOL_7$	–	233	47 %	5,3 %
2000	FMP	0,18	862	43 %	19,7 %
	$GWOL_1$	–	650	33 %	14,9 %
	$GWOL_7$	–	579	29 %	13,2 %
388444	uceprotect	–	463	0,12 %	10,6 %
8063	bl.de-ssh	–	336	4,2 %	7,2 %
1503	bfb	–	70	4,7 %	1,6 %

7.2 Využití FMP skóre pro vytváření prediktivních blacklistů

Dále bylo vyhodnoceno jednoho z možných použití FMP skóre – generování blacklistů volitelné velikosti. Při tomto použití je na konci každého dne sestaven seznam IP adres s největším FMP skóre (blacklist) a ten je pak používán pro blokování provozu[4] během následujícího dne.

Velikost či restriktivnost blacklistu je definována uživatelem (administrátorem sítě) – buď je použit pevně daný počet IP adres s nejvyšším skóre, nebo jsou použity všechny adresy se skóre vyšším než určitá mez.

Pro vyhodnocení efektivity blacklistů je zde použita metrika *hit-count*. Ta je definovaná (v souladu s [7, 8]) jako počet IP adres na blacklistu, které jsou správně předpovězeny jako škodlivé, tzn. skutečně je během predikčního okna (tj. následujícího dne) detekován útok z dané adresy. V případě použití blacklistu pro blokování provozu je to tedy počet úspěšně zablokovaných útočníků. Zde navíc definujeme metriku *hit-rate*, což je hodnota *hit-count* vydělená velikostí blacklistu. Určuje tedy, jaké procento záznamů v blacklistu se ukázalo jako užitečné.

Pro hodnocení jsou zde využita jen data typu *přístup*, protože pokusy o neoprávněný přístup jsou jistě závažnější události než skenování a dává tedy větší smysl zdroje takových aktivit blokovat.

Pro vyhodnocení těchto byly na základě FMP skóre vytvořeny blacklisty několika různých velikostí. Pro porovnání byly na základě stejných dat (tj. hlášení ze systému Warden) vytvořeny také blacklisty založené na základní metodě

[4] Nebo aplikaci rate-limitingu či jiných restriktivních opatření dle potřeb uživatele.

GWOL (*global worst offender list*, termín převzat z [7, 8]). V případě této metody se blacklist skládá z těch IP adres, ke kterým bylo za určité předchozí období přijato nejvíce hlášení. Konkrétně jsou použity období 1 den ($GWOL_1$) a 7 dní ($GWOL_7$). Stejně jako v případě blacklistů založených na FMP skóre lze i v případě GWOL generovat blacklisty různé délky, vždy tedy vzájemně porovnáváme blacklisty se stejným počtem záznamů.

Dále jsou pro porovnání vyhodnoceny tři reálné blacklisty poskytované třetími stranami, konkrétně UCEPROTECT[5], blocklist.de-SSH[6] (*bl.de-ssh*) a BruteForceBlocker[7] (*bfb*). Tyto blacklisty jsou založené na jiných datech. Jejich velikost je pevně daná.

V tabulce 9 jsou uvedeny různé metriky pro všechny testované blacklisty (blacklisty jsou vytvořeny pro tři různé dny, hodnoty v tabulce jsou průměrem z těchto dní). FMP a GWOL blacklisty jsou generovány ve velikostech 100, 500 a 2000 záznamů. Sloupec označený T ukazuje hodnotu meze FMP skóre odpovídající dané velikosti blacklistu. Jinými slovy, FMP blacklisty obsahují vždy ty adresy, které splňují $FMP_{\text{pristup}} \geq T$. Sloupec *hit-count* ukazuje počet adres, které ve dni, pro který byl blacklist připraven, skutečně zaútočily a byly by pomocí blacklistu zablokovány. *Hit-rate* je hodnota *hit-count* vydělená N, tedy procento záznamů v blacklistu, které úspěšně zablokovaly nějaký útok. Všechny hodnoty v tabulce jsou průměrem ze tří testovaných dnů.

Z tabulky je zřejmé, že obecně mají menší blacklisty vyšší hodnoty hit-rate. To je očekávané, protože tyto obsahují jen adresy s největší pravděpodobností budoucích útoků (resp. nejaktivnější v předchozích dnech v případě GWOL). Obzvlášť efektivní je FMP blacklist o délce 100 záznamů, u nějž skutečně zaútočilo všech 100 uvedených adres. Celkově jsou ve všech případech FMP blacklisty výrazně efektivnější než ty vytvořené metodou GWOL.

V průměru bylo v každém dni nahlášeno 4376 různých útočících IP adres. Poslední sloupec v tabulce 9 ukazuje, kolik z těchto adres by bylo kterým blacklistem zablokováno. Hodnoty se nezdají být nijak vysoké, je však nutné poznamenat, že přibližně 60 % útočníků v každém dni je „nových“, tzn. nebyli v předchozím týdnu ani jednou detekováni a jejich útoky je tak téměř nemožné předvídat. Maximální dosažitelné procento zablokovaných útočníků je tedy kolem 40 %.

Blacklisty třetích stran se ukázaly být z pohledu *hit-rate* velmi neefektivní, neboť pouze velmi malé procento adres uvedených na blacklistu bylo skutečně detekováno jako zdroj nějakého útoku. To je dáno tím, že tyto blacklisty jsou vytvářeny na základě zcela jiných zdrojů dat a můžou tak uvádět i útočníky, kteří necílí na žádné sítě či protokoly sledované detektory přispívajícími do systému Warden.

[5] http://www.uceprotect.net/en/
[6] https://www.blocklist.de/en/
[7] http://danger.rulez.sk/index.php/bruteforceblocker/

Nicméně další analýza ukázala, že pokud není problémem přílišná velikost blacklistů, je výhodné zkombinovat FMP blacklist s těmito blacklisty třetích stran. Seznam sjednocující FMP blacklist s mezí 0,5 (681 záznamů) se všemi třemi blacklisty třetích stran (celkem 397 241 záznamů) dokáže zablokovat 24,1 % útočících IP adres. Je však třeba myslet také na to, že příliš velký blacklist může zvýšit pravděpodobnost zablokování legitimního provozu.

Vyhodnocení lze shrnout tak, že blacklisty vytvořené na základě FMP skóre jsou velmi efektivní. Při stejné velikosti dokáží zablokovat výrazně více útočníků než blacklisty typu GWOL vytvořené na základě stejných dat. I v porovnání s různými blacklisty třetích stran dokáží FMP blacklisty zablokovat srovnatelný či větší počet útočníků, ovšem při mnohem menší velikosti blacklistu a tedy s nižšími nároky na výkon a také s nižší pravděpodobností zablokování legitimního provozu.

7.3 Využití FMP skóre pro efektivnější obranu proti DDoS útokům

Další možné využití FMP skóre je jako jedno z kritérií pro rozlišení škodlivého a legitimního provozu při obraně proti DDoS útokům. Implementace a vyhodnocení této možnosti použití bylo provedeno v rámci diplomové práce T. Jánského [19], jejímž byl autor této disertační práce konzultantem. Výsledky byly také publikované v konferenčním příspěvku [20].

Práce se zabývá obranou proti tzv. objemovým DDoS útokům, tj. takovým, při kterých je cílový server nebo jeho síťové připojení zaplaveno obrovským množstvím požadavků či obecně jakýmkoliv síťovým provozem. Obrana proti takovým útokům obecně spočívá v rozpoznání a zahození škodlivého provozu, tedy toho, který je vygenerován útočníkem a nepředstavuje reálné požadavky uživatelů. Klíčovou úlohou je schopnost spolehlivě rozpoznat škodlivý provoz od toho legitimního.

Práce vylepšuje stávající algoritmus v konkrétním zařízení *DDoS Mitigation Device (DMD)*. Ten je založen pouze na pozorování aktuálního objemu provozu z jendotlivých IP adres a jsou blokovány vždy ty nejaktivnější. Ty však nemusí být nutně škodlivé. Ve vylepšené verzi je proto vždy automaticky zjištěno i FMP skóre všech IP adres, které jsou zdrojem nějakého provozu, a primárně jsou blokovány ty, jež mají vysoké skóre (jsou tedy pravděpodobně škodlivé), bez ohledu na objem provozu. Až když zablokování takového provozu nestačí, začnou se blokovat i ostatní adresy podle jejich příspěvku k celkovému provozu, tak jako v původním algoritmu.

Expreimenty provedené v [19, 20] ukázaly, že tento přístup dokáže výrazně snížit množství nesprávně zablokovaného legitimního provozu i při velkých DDoS útocích.

8 Závěr

V této práci byla představena metoda pro číselné vyjádření reputace síťových entit (především IP adres) z hlediska bezpečnostních hrozeb. Toto číslo, nazvané *Future Misbehavior Probability score* či *FMP skóre*, slouží jako shrnutí všech dostupných bezpečnostně relevantních informací o dané entitě a zároveň jako předpověď jejího budoucího chování. Konkrétně FMP skóre vyjadřuje pravděpodobnost, že daná entita bude v příštích 24 hodinách detekována jako zdroj určitého nežádoucího chování, přičemž určení této pravděpodobnosti je prováděno pomocí strojového učení s využitím všech dostupných dat o dané entitě i ostatních „blízkých“ či podobných entitách. Tento obecný princip byl dále konkretizován pro případ ohodnocování škodlivých IPv4 adres a ověřen na reálných datech.

Bylo ukázáno, že všechny testované modely strojového učení dokáží dostatečně přesně odhadovat skutečnou pravděpodobnost budoucích hlášení a mohou být tedy využity pro výpočet FMP skóre.

Skóre může být využito hned několika způsoby. Tím základním je shrnutí dostupných informací o dané entitě do snadno uchopitelného čísla, prezentovaného uživateli, které člověku umožňuje rychle vyhodnotit a porovnat škodlivost jednotlivých entit. Podobným způsobem lze takové číselné ohodnocení využít i strojově, jak bylo ukázáno na příkladu filtrování DDoS útoků v kapitole 7.3.

Při kombinaci s dalšími údaji, jako je například vyhodnocení důležitosti cíle útoku, může být FMP skóre využito i pro prioritizaci incidentů, procesu, jehož cílem je napovědět bezpečnostnímu operátorovi, kterými událostmi se zabývat přednostně.

Další možností využití je vytváření blacklistů s adresami s nejvyšším FMP skóre, tedy s největší pravděpodobností budoucích útoků, které pak lze použít například k blokování provozu těchto adres. Výhodou oproti tradičním blacklistům je možnost zvolit si libovolně velikost takového blacklistu, resp. mezní hodnotu FMP skóre. Experimenty provedené v kapitole 7.2 ukázaly, že blacklisty vytvořené podle FMP skóre dokáží být velmi efektivní z hlediska počtu zablokovaných útočníků vzhledem k velikosti blacklistu.

V porovnání s předchozími pracemi zabývajícími se hodnocením škodlivosti síťových entit je navržená metoda unikátní v tom, že kombinuje výhody všech předchozích přístupů – jde o číselné hodnocení umožňující adresy porovnávat a řadit, pracuje se s jednotlivými adresami namísto celých sítí, přičemž jsou ale zároveň zahrnuty i informace o dalších adresách ve stejné síti, a hodnocení je založeno na explicitní predikci budoucího chování, namísto pouhého shrnutí předchozích aktivit. Navíc je význam FMP skóre jednoduše interpretovatelný, což usnadňuje jeho používání, případně nastavovaní mezí. Metoda je unikátní i tím, že umožňuje využít prakticky jakákoliv dostupná data, například odhad, zda je adresa dynamicky přidělovaná, či geolokační informace. Předchozí práce

vždy vycházely jen z analýzy blacklistů nebo z přijatých hlášení o detekovaných bezpečnostních událostech.

V práci byl také popsán přínos autora v oblasti detekce škodlivého provozu – návrh několika metod detekce škodlivého provozu, významný podíl na návrhu frameworku pro analýzu síťových dat (NEMEA), jeho implementaci, včetně implementace navržených detekčních metod, a nasazení celého systému v reálné síti. Tato činnost pomohla získat velké množství dat o síťových útocích, jejichž analýza pomohla odhalit či ověřit vlastnosti zdrojů škodlivého chování, na jejichž znalosti staví navržená metoda hodnocení reputace. Kromě toho, že byla data z detekčního systému NEMEA využita pro tuto práci, je systém již několik let úspěšně používán v praxi a přispěl i ke vzniku řady akademických prací.

V neposlední řadě byl popsán i systém pokročilé reputační databáze NERD, jenž byl také vytvořen v souvislosti s touto prací. Jeho účelem je jednak získávat dodatečné informace ke škodlivým IP adresám, zároveň slouží jako platforma, pro niž je metoda výpočtu FMP skóre primárně určena.

Výpočet FMP skóre do systému NERD již implementován. Skóre je zde počítané pro všechny adresy, k nimž tento systém udržuje nějaký záznam, je pravidelně aktualizované a volně dostupné, kdokoliv tak může tato data využít v praxi. Samozřejmostí je i možnost vytvářet blacklisty dle zadaných kritérií a také možnost dotazovat se na skóre IP adres nejen přes webové GUI, ale i programové API, což umožňuje snadnou integraci do jiných systémů.

Na využití tohoto skóre je již připravován systém obrany proti DDoS útokům provozovaný organizací CESNET. Za podobným účelem je plánováno využití FMP skóre i v rámci evropského projektu GN4[8]. Jednou z aktivit tohoto projektu je vývoj nové verze systému Firewall on Demand[9] (FoD), provozovaného organizací GÉANT, v němž by měla být v případě DDoS útoku automaticky navrhována pravidla pro blokování škodlivého provozu. FMP skóre ze systému NERD by mělo být jedním z klíčových kritérií určujících, které adresy či sítě budou při útoku navrženy k blokování.

Dále je FMP skóre poskytované systémem NERD využito v evropském projektu PROTECTIVE[10], jehož cílem je vývoj nástroje pro sdílení a pokročilou analýzu kyberbezpečnostních informací. FMP skóre je zde využito především jako jedno z kritérií pro prioritizaci hlášení, ale také je ve webovém rozhraní zobrazováno uživateli jako jedna ze základních informací o IP adresách.

[8] https://www.geant.org/Projects/GEANT_Project_GN4
[9] https://www.geant.org/Networks/Network_Operations/Pages/Firewall-on-Demand.aspx
[10] https://protective-h2020.eu/

Reference

[1] ENISA: ENISA Threat Landscape Report 2017. Leden 2018, doi:10.2824/967192.
URL https://www.enisa.europa.eu/publications/enisa-threat-landscape-report-2017

[2] Symantec: 2018 Internet Security Threat Report. březen 2018.
URL https://www.symantec.com/security-center/threat-report

[3] Ponemon Institute: Third Annual Study on Exchanging Cyber Threat Intelligence: There Has to Be a Better Way. Research Report, leden 2018.

[4] Collins, M. P.; Shimeall, T. J.; Faber, S.; aj.: Using Uncleanliness to Predict Future Botnet Addresses. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, New York, NY, USA: ACM, 2007, s. 93–104, doi:10.1145/1298306.1298319.

[5] Shue, C. A.; Kalafut, A. J.; Gupta, M.: Abnormally Malicious Autonomous Systems and Their Internet Connectivity. *IEEE/ACM Transactions on Networking*, ročník 20, č. 1, únor 2012: s. 220–230, ISSN 1063-6692, doi:10.1109/TNET.2011.2157699.

[6] van Wanrooij, W.; Pras, A.: Filtering spam from bad neighborhoods. *International Journal of Network Management*, ročník 20, č. 6, 2010: s. 433–444, doi:10.1002/nem.753.

[7] Zhang, J.; Porras, P.; Ullrich, J.: Highly Predictive Blacklisting. In *Proceedings of the 17th Conference on Security Symposium*, SS'08, Berkeley, CA, USA: USENIX Association, 2008, s. 107–122.

[8] Soldo, F.; Le, A.; Markopoulou, A.: Blacklisting Recommendation System: Using Spatio-Temporal Patterns to Predict Future Attacks. *IEEE Journal on Selected Areas in Communications*, ročník 29, 08 2011: s. 1423–1437.

[9] Cejka, T.; Bartos, V.; Svepes, M.; aj.: NEMEA: A Framework for Network Traffic Analysis. In *12th International Conference on Network and Service Management (CNSM 2016)*, IEEE, 2016, s. 195–201, doi:10.1109/CNSM.2016.7818417.

[10] Čejka, T.; Rosa, Z.; Kubátová, H.: Stream-wise Detection of Surreptitious Traffic over DNS. In *Proc. of 19th IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, IEEE, 2014, doi:10.1109/CAMAD.2014.7033254.

[11] Jansky, T.; Cejka, T.; Bartos, V.: Hunting SIP Authentication Attacks Efficiently. In *Security of Networks and Services in an All-Connected World (AIMS 2017)*, LNCS 10356, Springer, 2017, doi:10.1007/978-3-319-60774-0_9.

[12] Cejka, T.; Bartos, V.; Truxa, L.; aj.: Using Application-Aware Flow Monitoring for SIP Fraud Detection. In *Intelligent Mechanisms for Network Configuration and Security (AIMS 2015)*, LNCS 9122, Springer, 2015, s. 87–99, doi:10.1007/978-3-319-20034-7_10.

[13] Bartoš, V.; Kořenek, J.: Evaluating Reputation of Internet Entities. In *IFIP International Conference on Autonomous Infrastructure, Management and Security (AIMS'16)*, LNCS 9701, Springer, 2016, s. 132–136, doi:10.1007/978-3-319-39814-3_13.

[14] Bartoš, V.: Creating a Network Reputation Database. Poster, TNC'16 conference, 2016.

[15] Bartoš, V.: Analysis of alerts reported to Warden. Technická zpráva 1/2016, CESNET, únor 2016.

[16] Moura, G. C. M.; Sperotto, A.; Sadre, R.; aj.: Evaluating Third-Party Bad Neighborhood Blacklists for Spam Detection. In *Proceedings of IFIP/IEEE International Symposium on Integrated Network Management 2013*, IEEE, květen 2013, ISBN 978-1-4673-5229-1, s. 252–259.

[17] Bartoš, V.; Žádník, M.: An Analysis of Correlations of Intrusion Alerts in an NREN. In *19th International Workshop on Computer-Aided Modeling Analysis and Design of Communication Links and Networks (CAMAD)*, IEEE, prosinec 2014, s. 305–309.

[18] Hofstede, R.; Hendriks, L.; Sperotto, A.; aj.: SSH Compromise Detection Using NetFlow/IPFIX. *SIGCOMM Computer Communication Review*, ročník 44, č. 5, říjen 2014: s. 20–26, ISSN 0146-4833, doi:10.1145/2677046.2677050.

[19] Jánský, T.: Informovaná mitigace DDoS útoků na základě reputace. Diplomová práce, ČVUT, 2018.

[20] Jánský, T.; Čejka, T.; Žádník, M.; aj.: Augmented DDoS Mitigation with Reputation Scores. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*, ARES 2018, New York, NY, USA: ACM, 2018, doi:10.1145/3230833.3233279.

Životopis

Osobní údaje

Jméno a příjmení:	Václav Bartoš
Národnost:	česká
Datum narození:	24. 5. 1987
Email:	xbarto11 <at> stud.fit.vutbr.cz
	vaclavbartos <at> centrum.cz

Vzdělání

od 2011 Fakulta informačních technologií, VUT v Brně.
Doktorský studijní program Výpočetní technika a informatika.

2009–2011 Fakulta informačních technologií, VUT v Brně.
Magisterský studijní program Informační technologie, obor počítačové a vestavěné systémy.
Diplomová práce: Detekce anomálií v síťovém provozu.
Prospěl s vyznamenáním (tzv. červený diplom).

2006–2009 Fakulta informačních technologií, VUT v Brně.
Bakalářský studijní program Informační technologie.
Bakalářská práce: Verifikace generického propojovacího systému na FPGA.
Prospěl s vyznamenáním (tzv. červený diplom).

2002–2006 SPŠ a SOU Lanškroun.
Obor slaboproudá elektrotechnika.

Akademické zkušenosti

2011–2016 Vedení více než 20 bakalářských a diplomových prací, obvykle na témata související s analýzou síťového provozu pro bezpečnostní účely.

Pracovní zkušenosti

od 2010 CESNET, výzkum a vývoj v oblasti monitorování sítě a analýzy provozu pro bezpečnosntí účely, účast na řadě národních i evropských projektů, od r. 2015 člen týmu CESNET-CERTS.

2007–2010 CESNET, vývojář firmware pro FGPA

Účast na projektech

2017–2021 CTI (Vybudování a ověřovací provoz systému Cyber Threat Intelligence), projekt MV ČR.
Cílem projektu vytvoření komplexního systému pro sběr a analýzu hlášení a jiných kyberbezpečnostních informací pro použití státními úřady (shrnuje výsledky SABU a několika dalších projektů do jednoho komplexního systému).

2016–2019 PROTECTIVE, evropský projekt výzvy H2020.
Cílem je vývoj platformy pro sdílení a zpracování kyberbezpečnosntích informací, včetně jejich korelace, kontextualizace a prioritizace.
https://protective-h2020.eu/

2016–2019 SABU (Sdílení a analýza bezpečnostních událostí), projekt MV ČR.
Cílem je vývoj platformy pro sdílení a zpracování kyberbezpečnosntích informací a rozšíření sdílecí komunity v rámci ČR.
https://sabu.cesnet.cz/

2016–2019 GÉANT GN4-2, Activity JRA2, Task 6 (Security services for the infrastructures).
Vývoj služeb Reputation shield a Firewall on Demand a jejich vzájemného propojení.

2015–2016 GÉANT GN4-1, Activity SA3, Task 1 (Security Service Development).
Návrh a vývoj služby Reputation shield, reputační databáze síťových entit (také známé jako NERD).

2012–2013 GÉANT GN3, Activity JRA2, Task 4 (Multidomain Security).
Vývoj nástroje nfQuery sloužícího detekci bezpečnostních událostí zasahujících více sítí.