VĚDECKÉ SPISY VYSOKÉHO UČENÍ TECHNICKÉHO V BRNĚ

Edice Habilitační a inaugurační spisy, sv. 668
ISSN 1213-418X

Karel Slavíček

MATEMATICKÉ METODY DETEKCE ÚNIKU DAT MEDICÍNSKÉHO KOMUNIKAČNÍHO SYSTÉMU REDIMED

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta elektrotechniky a komunikačních technologií
Ústav telekomunikací

Mgr. Karel Slavíček, Ph.D.

MATEMATICKÉ METODY DETEKCE ÚNIKU DAT MEDICÍNSKÉHO KOMUNIKAČNÍHO SYSTÉMU REDIMED

MATHEMATICAL METHODS OF INFORMATION LEAKAGE DETECTION IN MEDICINE COMMUNICATION SYSTEM REDIMED

ZKRÁCENÁ VERZE HABILITAČNÍ PRÁCE
OBOR: TELEINFORMATIKA

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
NAKLADATELSTVÍ VUTIUM

BRNO 2020

KLÍČOVÁ SLOVA
medicínské obrazové informace, detekce anomálií, bezpečnost, matematické modely

KEYWORDS
medical picture data, anomaly detection, security, mathematical modelling

MÍSTO ULOŽENÍ:
Originál habilitační práce je uložen na vědeckém oddělení VUT FEKT v Brně

ISBN 978-80-214-5898-7
ISSN 1213-418X

Obsah

1 Úvod **5**
1.1 Metropolitní PACS systém MeDiMed 5
1.2 Radiologické komunikační centrum Redimed 6

2 Bezpečnostní aspekty zpracování medicínských obrazových dat **7**

3 Použité matematické nástroje **11**
3.1 Popisná a inferenční statistika 11
3.1.1 Statistická závislost dvou náhodných veličin 14
3.2 Entropické modely 15

4 Netechnické aspekty detekce úniku dat **16**
4.1 Právní aspekty 16
4.2 Psychologické aspekty 17

5 Analýza logů systému Redimed **18**
5.1 Analýza provozu malých uživatelů 19
5.2 Analýza provozu velkých uživatelů 20

6 Závěr **27**

Literatura **29**

Představení autora

Mgr. Karel Slavíček, Ph.D.
Ústav telekomunikací
Fakulta elektrotecniky a komunikačních technologií
Vysoké učení technické v Brně
Technická 12
616 00 Brno
e-mail:slavicekkarel@feec.vutbr.cz

Mgr. Karel Slavíček, Ph.D., narozen 25.9.1969 v Chrudimi.

Karel Slavíček úspěšně absolvoval magisterský studijní obor informatika na Přírodovědecké fakultě MU v roce 1993 a poté doktorský studijní program matematické inženýrství na Fakultě strojního inženýrství VUT. Od roku 1993 zaměstnán na Ústavu výpočetní techniky MU postupně na pozici systémového analytika a posléze výzkumného a vývojového pracovníka. Od roku 2004 pravidelně přednáší na Ústavu telekomunikací FEKT VUT v Brně.

Podílel se na budování páteřní datové sítě brněnských vysokých škol a na budování sítě CESNET a komunikační infrastruktury systému zpracování medicínských obrazových informací MeDiMed. V rámci pedagogické činnosti garantuje a realizuje výuku v oblasti architektury sítí včetně vedení diplomových a bakalářských prací. Řada vyžádaných přednášek pro ČTU i několik komerčních subjektů, pravidelně recenzuje příspěvky v odborném periodiku a hodnotí výzkumné projekty pro agenturu TAČR.

Vědeckovýzkumné aktivity uchazeče dříve orientovány převážně na vysokorychlostní a optické sítě, aktuálně zaměřeny na aplikovaný výzkum v oblasti IoT, zejména senzorů instalovatelných na osobě, senzoriky obecně a komunikační infrastruktury pro IoT. V této oblasti úspěšně řídil a i v současnosti vede několik výzkumných grantových projektů, především mezinárodních projektů EUREKA, kde zpravidla pracuje jako národní koordinátor projektu. Aktuálně je hlavním řešitelem projektu MVČR na výzkum senzorů na bázi PM vláken, na kterém spolupracuje VUT, MU a UNOB.

1 Úvod

Oblast budování a rozvoje vysokorychlostních datových sítí prodělala v uplynulých dvou dekádách bouřlivý rozvoj. Ruku v ruce se změnami kvantitativními, kdy přenosová rychlost vzrostla o několik řádů, dochází i ke změnám kvalitativním, zejména z pohledu spolehlivosti a dostupnosti služeb.

Rychlost vývoje je možné dokumentovat na příkladech brněnské akademické počítačové sítě a sítě národního výzkumu CESNET na jejichž rozvoji a provozu jsem měl tu čest se podílet. Vývoj brněnské akademické počítačové sítě je dokumentován v řadě odborných i popularizačních článků [12], [26], [25], [11], [20], [10].

Optokabelová síť brněnských univerzit je unikátní minimálně v evropském měřítku. Existence privátní optické sítě umožnila vznik důležitých aplikací, zejména v oblasti zdravotní péče. Zdravotní péče je jednou z posledních oblastí, kde se začaly informační a komunikační technologie využívat. Je to dáno nejen přirozeně vysokými nároky lékařů na výkon a spolehlivost ICT systému, ale do jisté míry i zdravě konzervativním přístupem lékařské veřejnosti. Ve chvíli, kdy výpočetní a komunikační systémy dosáhly parametrů potřebných pro jejich nasazení v oblasti medicíny se situace změnila a dnes si již například radiologii nedokážeme bez ICT systémů představit. K rozvoji využití ICT v lékařství významně přispěl i projekt MeDiMed.

1.1 Metropolitní PACS systém MeDiMed

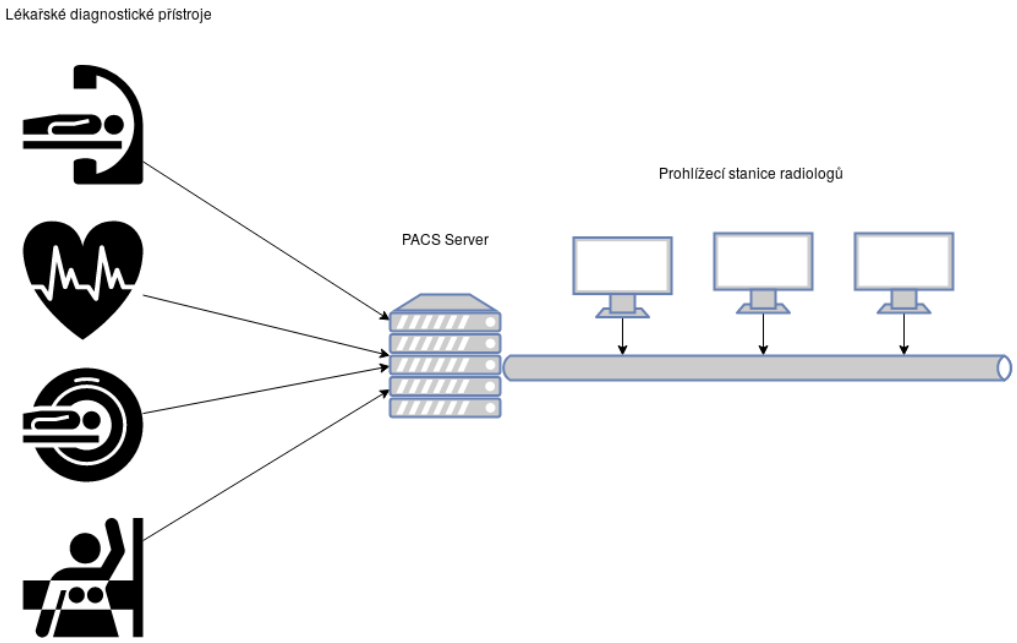
Na konci devadesátých let minulého století dosáhl pokrok v ICT technologiích takové úrovně, že je bylo možné začít využívat i pro přenos medicínských obrazových informací z lékařských diagnostických přístrojů, jako např. počítačový tomograf, magnetická rezonance, pozitronový emisní tomograf a další. Přechod na digitální formu zpracování obrazové dokumentace proběhl relativně rychle a dnes se již původní filmový materiál prakticky nepoužívá.

Používání digitálních zobrazovacích systémů a metod a využívání technologie PACS (Picture Archiving and Communication System) má řadu technických i ekonomických výhod. Umožňuje zvýšit přesnost diagnózy, umožňuje rychlejší přístup k obrazovým datům pacienta a nižší potřebu opakovaných vyšetření). PACS systémy jsou zpravidla používány paralelně s nemocničním informačním systémem (NIS), který slouží pro běžnou evidenci pacientů, jejich diagnóz a průběhu léčby.

Pro systémy NIS je nejběžnějším standardem HL7 [1], pro systémy PACS je dnes jediným používaným standardem DICOM [2]. Existence standardu DICOM významně přispěla k rozvoji digitalizace medicínských obrazových dat. Průkopníkem tohoto řešení v brněnském regionu byl Metropolitní PACS systém MeDiMed.

Vznik systému MeDiMed byl podmíněn existencí kvalitní optokabelové sítě brněnských vysokých škol, neboť běžné internetové přípojky v době vzniku tohoto řešení neposkytovaly

dostatečnou přenosovou kapacitu pro přenos velkého objemu dat, které generují lékařské diagnostické přístroje. Pro připojení brněnských nemocnic do tohoto systému proto byly využity vyhrazené optické vlákna a vznikla tak dedikovaná síť pro potřeby lékařské diagnostiky. Vybudování regionálního centra pro podporu zpracování medicínských obrazových informací MeDiMed bylo podpořeno řadou projektů, na jejich řešení jsem měl možnost se podílet. Pokroky budování systému MeDiMed byly popsány v řadě publikací [15], [13], [16], [16], [14], [18], [48],[17], [31], [47], [49], [19], [30], [50], [51], aktuální informace bývají zveřejňovány na webových stránkách projektu [5].



Obr. 1.1: Struktura systému PACS.

Obecná struktura PACS systému je znázorněna na obrázku 1.1. Jednotlivé diagnostické přístroje, tzv. modality ukládají obrazová data do PACS serveru, odkud tyto data stahují prohlížecí diagnostické stanice radiologů.

Systém MeDiMed je využíván řadou regionálních nemocnic. Každý zdravotnický subjekt, který toto řešení využívá, má svůj vlastní server i diskový prostor, kde se nachází jeho obrazová data.

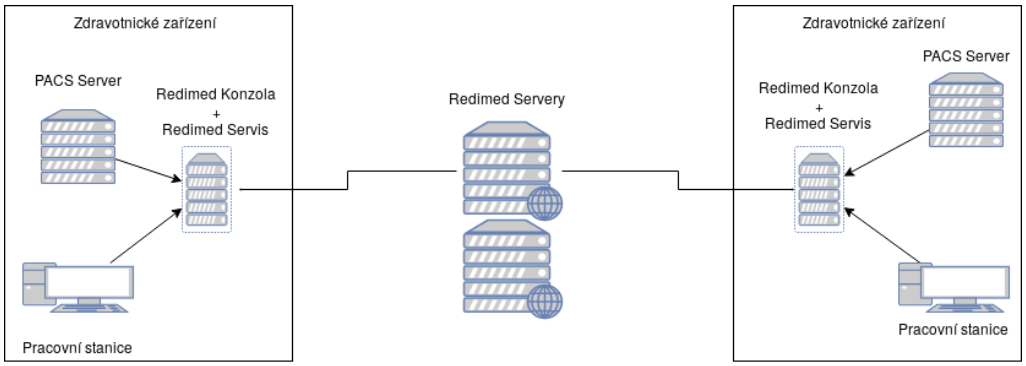
Z důvodů zajištění vysoké odolnosti a dostupnosti systému jsou data uloženy ve dvou samostatných a geograficky oddělených systémech. Bezpečnost přenosu je zajištěna vyhrazenými vlákny, případně využitím kryptograficky zabezpečených tunelů.

1.2 Radiologické komunikační centrum Redimed

S rozvojem digitalizace zpracování medicínských obrazových dat rostla i potřeba výměny snímků mezi zdravotnickými institucemi a to nejen v případě velkých nemocnic s rozsáhlým

přístrojovým vybavením, tak i v případě malých organizací a soukromých lékařských praxí. S nárůstem počtu uživatelů již nebylo možné všechny účastníky připojit vyhrazeným optickým vláknem a v případě institucí s menší potřebou komunikace by to ani nebylo účelné. Z toho důvodu vzniklo v rámci systému MeDiMed speciální radiologické komunikační centrum Redimed.

Komunikační systém Redimed je určený pro elektronickou výměnu medicínské obrazové dokumentace dle standardu DICOM případně dalších souborů mezi zdravotnickými institucemi navzájem. Zdravotnickou institucí zde může být kromě nemocnic a poliklinik i domácí pracoviště radiologů, menší privátní centra a praktičtí lékaři, případně akademická pracoviště lékařských fakult a to nejen v České republice. Jedná se o čistě softwarové řešení, které je použitelné jak pro přímou spolupráci dvou nemocnic případně nemocnice a soukromého radiologa, tak i pro spolupráci v rámci rozsáhlých sítí zdravotnických profesionálů. Struktura systému Redimed je znázorněna na obrázku 1.2.



Obr. 1.2: Struktura komunikačního systému Redimed.

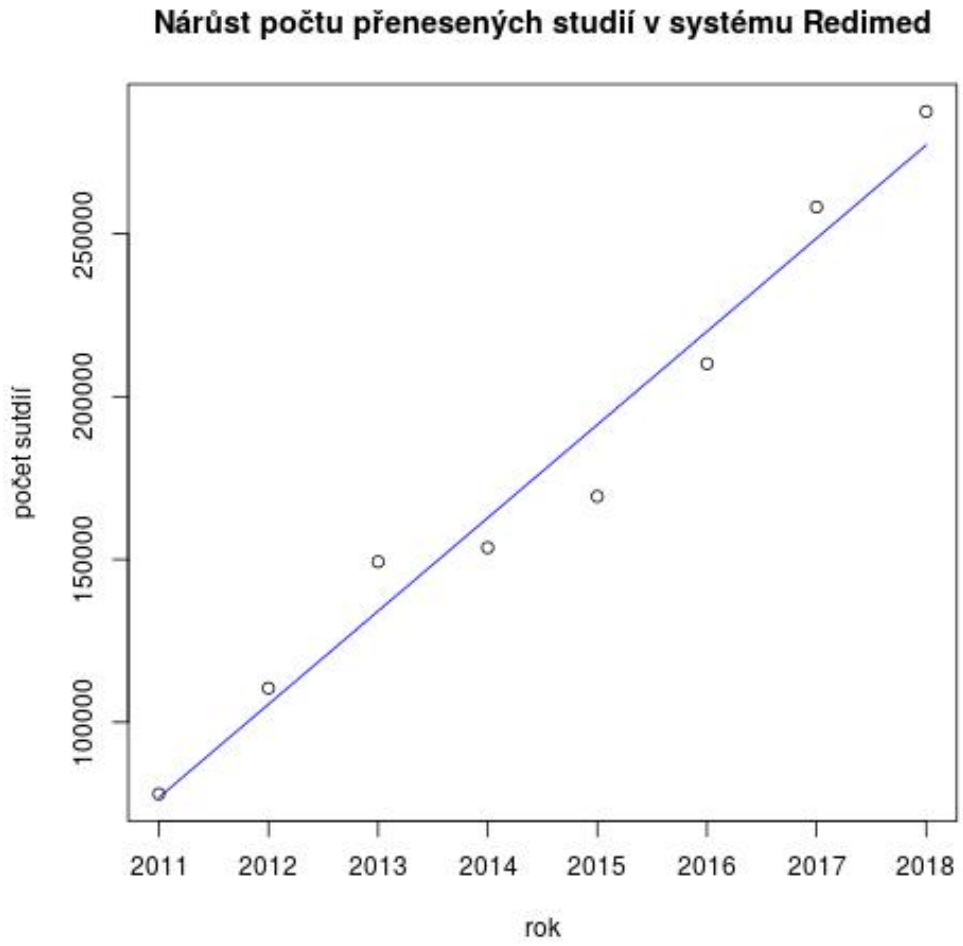
Systém Redimed si velmi rychle získal oblibu mezi zdravotnickými zařízeními všech velikostí, o čemž svědčí růst počtu přenášených studií i objemu dat. Pro lepší představu jsou tyto kvantitativní parametry uvedeny v grafech 1.3 a 1.4. Z grafu je patrný přibližně lineární nárůst objemu provozu, který aktuálně dosahuje stovek tisíc studií ročně.

Systém Redimed aktuálně využívá více než 570 institucí a počet uživatelů stále roste.

Přenos principielně citlivých medicínských informací mezi nezávislými zdravotnickými zařízeními a univerzálnost řešení systému Redimed, která je potřeba pro snadné zapojování jednotlivých účastníků do projektu, otevírá i nové možnosti úniku dat o pacientech.

2 Bezpečnostní aspekty zpracování medicínských obrazových dat

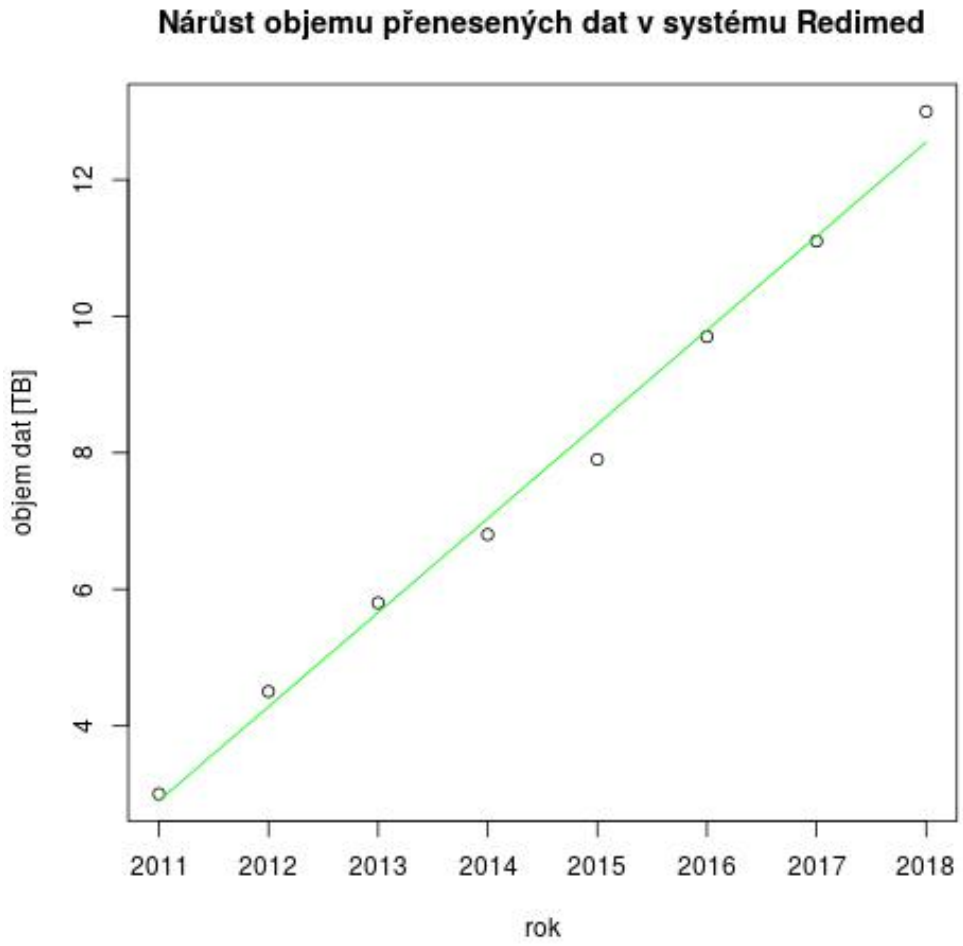
Současně s rozvojem komunikační a výpočetní technologie narůstá její sepětí s každodenním životem společnosti a tím se bohužel do digitálního prostředí přesouvá i nežádoucí činnost.



Obr. 1.3: Počet studií přenesených systémem Redimed.

Prakticky neustále dochází k nejrůznějším útokům na počítačové systémy, obsahující cenná a důležitá data. Různé nežádoucí či přímo podvodné aktivity využívají relativní anonymity elektronického prostředí. Počítače připojené k celosvětové síti Internet jsou prakticky neustále bombardovány pokusy o prolomení svého zabezpečení ve snaze získat přístup k datům v nich uloženým a tyto počítače ovládnout a využít k dalším útokům. Některé studie uvádí, že pokud k Internetu připojíme nový počítač, aniž by byl zabezpečen alespoň instalací nejnovější bezpečnostních aktualizací, bude útočníkem ovládnut v průměru za 4 minuty.

Medicínská data představují zajímavý cíl potencionálních útoků. Mohou obsahovat velmi citlivé údaje o pacientech. Tato data mohou být předmětem zájmu například pojišťovacích společností, kdy znalost zdravotního stavu člověka může vést ke změně chování pojišťovny při uzavírání pojistných smluv. V případě nemocí psychologického nebo venerologického typu může únik informací ze zdravotní dokumentace pacienta způsobit tomuto pacientovi řady nepříjemností a jistý typ vyloučení ze společnosti. Rovněž informace o existenci např. implantátů používaných v plastické chirurgii můžou v případě veřejně známých osobností působit značné nepříjemnosti. Vedle těchto snadno představitelných problémů způsobených únikem dat o kon-



Obr. 1.4: Objem dat přenesený systémem Redimed.

krétním pacientovi může být problémem i únik např. dlouhodobě shromažďovaných dat pro vědecký výzkum.

Zatím nedošlo k útoku na systémy zpracování medicínských obrazových dat, což můžeme přičítat dílem odpovědnému chování uživatelů a dílem relativně menší zajímavosti těchto dat ve srovnání s daty vládních institucí či bank. Přesto je nutné se na takovou možnost připravit a mít k dispozici automatizované nástroje, které na případnou nežádoucí aktivitu upozorní.

Navíc s nárůstem počtu uživatelů systému Redimed lze očekávat, že poroste i nebezpečí případného útoku nebo pokusu o únik informací z tohoto systému. Rovněž nástup využívání miniaturní elektroniky moderně označované jako IoT jistě v dohledné době zasáhne i oblast lékařství a přibude tak více dat, která jsou zajímavá pro potencionální útočníky.

Obecně lze možné útoky na ICT systém s cílem získat neoprávněný přístup k datům rozdělit na dvě kategorie:

- Útok na ICT infrastrukturu. Do této kategorie spadá odposlech na fyzickém přenosovém médiu, napadení aktivních síťových prvků, které nativně poskytují možnost kopírování přenášených dat za účelem diagnostiky případných problémů infrastruktury, dále na-

padení operačního systému případně programového vybavení serverů uchovávajících a zpracovávajících data.

- Útok na uživatelské úrovni. Touto kategorií rozumíme neoprávněné kopírování dat s využitím přístupových údajů uživatele, který má k těmto datům přístup. V tomto případě není podstatné, zda tak činí uživatel sám, či zda došlo k prozrazení nebo zcizení jeho přihlašovacích údajů.

Útoky na ICT infrastrukturu jsou předmětem celé řady výzkumných projektů a prací. Většina typů útoků je dobře prozkoumána a řady odborníků se zabývají možnostmi prevence, detekce a obrany proti těmto útokům. Dosud málo probádanou oblastí jsou možnosti automatické detekce úniku dat inicializovaných samotnými uživateli systému.

Možnosti detekce takto inicializovaného úniku dat jsou omezené. Pokud např. lékař, který má přístup do nemocničního systému PACS odešle jednu konkrétní studii na svůj soukromý účet systému ReDiMed (nebo účet spřátelené zdravotnické instituce či lékaře), není algoritmicky rozhodnutelné, zda tak učinil oprávněně či nikoli, aniž by to musela posuzovat nějaká další autorita. Pokud by však takto odesílal velké soubory studií, neměla by taková skutečnost uniknout pozornosti vhodných automatizovaných nástrojů pro zpracování logů událostí systému.

Nebezpečí útoku na uživatelské úrovni vzrůstá úměrně rostoucímu počtu uživatelů systému Redimed i nárůstu počtu přenášených studií. S nárůstem objemu přenášených dat roste přitažlivost systému pro případného útočníka. Jakýkoli pokus o prolomení zabezpečení systému stojí jisté úsilí a je spojeno s určitou mírou rizika odhalení a případných trestněprávních následků. Proto je nepravděpodobné, že by se vyskytlo příliš mnoho útoků na systém, který neobsahuje dostatečné množství citlivých nebo jinak hodnotných dat.

Zároveň s nárůstem počtu uživatelů roste i riziko ztráty přihlašovacích údajů nebo úmyslného zneužití systému samotným uživatelem. V mnoha případech je prakticky nemožné rozlišit, zda daný datový přenos inicioval oprávněný uživatel systému, nebo zda došlo k úniku jeho přihlašovacích údajů a přenos inicioval neznámý útočník.

V zásadě existují jen dvě možnosti detekce úniku dat na této úrovni:

- Podrobný audit jednotlivých datových přenosů
- Matematická analýza datových toků

Audit jednotlivých přenosů není příliš představitelný vzhledem k množství přenášených obrazových studií. Navíc detailní kontrola přenosů by způsobovala administrativní zátěž pro uživatele a brzdila by další rozvoj využívání systému Redimed a spolupráce zdravotnických institucí.

Jedinou reálnou možností obrany proti úniku dat iniciovanému na uživatelské úrovni tedy zůstává matematická analýza datových toků a vyhledávání neobvyklých situací. Odchylka od obvyklého stavu může a nemusí znamenat nežádoucí únik dat. Počet přenášených studií za jednotku času přirozeným způsobem kolísá, čímž se snižuje spolehlivost určení toho, co je či není obvyklý provoz.

Matematickým zpracováním logů událostí dokážeme identifikovat situaci, kdy by došlo k

významnému objemu nežádoucí komunikace, tj. situaci, kdy někdo kopíruje větší množství obrazových studií. Tímto postupem není možné zabránit úniku několika jednotlivých studií. Pokud máme za úkol ochránit jednotky velmi citlivých studií, není jiná možnost, než striktní evidence přístupu k nim.

Jiný problém, kterým jsme se v rámci projektu MeDiMed zabývali, je ochrana anonymizovaných studií užitých k výzkumným a výukovým účelům. V tomto případě se uživatelé obávají neautorizovaného užití jimi publikovaných výsledků. Pro takový případ je nutné zajistit publikované obrazové studie dodatečnou informací, např. vodoznakem [43], [42], [41].

3 Použité matematické nástroje

V této kapitole připomeneme matematické nástroje a postupy, které se dají použít pro analýzu logů událostí systému Redimed. Kvantitativní analýza logů událostí může pomoci odhalit nežádoucí přenosy dat a přitom zachovat nezbytnou míru anonymity uživatelů.

Jako překvapivě účinné se ukázaly základní nástroje popisné statistiky, jejichž přehled je uveden v následující podkapitole. Vedle těchto nejjednodušších nástrojů byly zkoumány i možnosti využití metod analýzy časových řad a spektrální analýzy pro vyhledávání periodických vzorů provozu.

3.1 Popisná a inferenční statistika

Slovo statistika má v kontextu kontextu zpracování dat minimálně dva různé významy. Jednak označuje vědní disciplínu, ale používá se též k označení některých vlastností sledované veličiny, např. aritmetický průměr je statistikou v tomto smyslu. Statistiku jakožto vědní disciplínu můžeme dále dělit na statistiku popisnou, která se zabývá numerickým popisem získaných dat, a statistiku induktivní, která se zabývá hledáním zákonitostí v získaných datech. V této kapitole připomeneme nejdůležitější poznatky z teorie pravděpodobnosti a matematické statistiky, které budeme používat pro analýzu dat přenosového systému Redimed. Podrobnější informace je možno nalézt v klasických učebnicích [7], [8] [62], [54], vysokoškolských učebních textů [40], [24]. Velmi pěkný on-line přehled užité matematiky včetně elegantně zpracovaných kapitol o teorii pravděpodobnosti a matematické statistiky nabízí Ústav matematiky fakulty strojního inženýrství VUT [3].

Popisná statistika se zabývá empiricky zjištěnými hodnotami a má svůj protějšek v teorii pravděpodobnosti, která pracuje s teoretickými matematickými modely. Základním pojmem teorie pravděpodobnosti je náhodný jev. Náhodný jev je výsledek nějakého pokusu nebo děje, který může či nemusí nastat. Může být popsán slovně, např. “při hodu kostkou padne šestka” (koneckonců teorie pravděpodobnosti vznikala na popud hazardních her), nebo může mít číselný charakter, např. počet lidí ve frontě na zmrzlinu je vyšší než 10, nebo počet přenesených CT snímků za poslední hodinu je nižší než 5. Číselně kvantifikovatelný stav náhodného děje pak nazýváme náhodnou veličinou. (Např. počet přenesených medicínských studií za den).

Tato náhodná veličina má pak řady praktických realizací, tj. v našem případě zjištěných počtů přenesených snímků, které průběžně měříme každý den. Tím vznikne statistický soubor pozorovaných počtů přenesených snímků. V tomto místě se nám potkává rovina teoretická - náhodná veličina - s rovinou empirickou – statistický soubor praktických realizací této náhodné veličiny.

Pravděpodobnostní chování náhodné veličiny X popisujeme pomocí její distribuční funkce

$$F(x) = P(X < x). \tag{3.1}$$

Distribuční funkce $F(x)$ vyjadřuje pravděpodobnost, že náhodná veličina X nabývá hodnoty menší než x.

Pro popis statistického souboru používáme dva základní typy charakteristik:

- míry polohy a
- míry variability

Jako míra polohy se nejčastěji používá aritmetický průměr, případně u některých náhodných veličin medián a u kategorických dat modus. Aritmetický průměr statistického souboru $x_1, x_2, ...x_n$ rozsahu n je definován jako

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.2}$$

Pro analýzu dat, které mohou mírně kolísat, např. počet přenesených obrazových studií za hodinu, může být výhodné číselnou řadu naměřených hodnot $x_1, x_2, ...x_n$ tzv. vyhladit, tj. nahradit posloupnost $x_1, x_2, ...x_n$ posloupností aritmetických průměrů několika (optimálně lichého počtu) sousedních hodnot, čímž získáme řadu $\hat{x}_2, \hat{x}_3, ...\hat{x}_{n-1}$, kde

$$\hat{x_j} = \frac{1}{3} \sum_{i=j-1}^{j+1} x_i. \tag{3.3}$$

V tomto případě mluvíme o tzv. klouzavém průměru. Klouzavý průměr jakkoli je ve své podstatě jednoduchý, je velmi vhodný nástroj pro detekci významných odchylek od obvyklého stavu a kromě detekce potenciálně nežádoucích přenosů medicínských obrazových dat jsem jej s úspěchem využívali i pro analýzu logů událostí aktivních prvků datových sítí a podobné aplikace [52], [45], [46], [53].

U klouzavého průměru někdy využíváme i vážený aritmetický průměr

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{3.4}$$

kde $w_i \geq 0$ jsou váhy, které vyjadřují relativní význam jednotlivých hodnot x_i.

Aritmetickému průměru, jakožto charakteristice empirických dat, odpovídá v teoretické rovině střední hodnota, která je mírou polohy odpovídající náhodné veličiny, jakožto teoretického modelu. Střední hodnota je definována jako

$$E(X) = \sum_{x} xp(x), \tag{3.5}$$

kde x probíhá celý definiční obor náhodné veličiny X v případě diskrétní náhodné veličiny a

$$E(X) = \int_{-\infty}^{\infty} f(x)dx \tag{3.6}$$

v případě náhodné veličiny se spojitým rozdělením pravděpodobnosti.

Nevýhodou aritmetického průměru i střední hodnoty je jejich citlivost na byť i malý počet velmi odlehlých měření. Bude-li např. Dané zdravotnické zařízení přenášet každý den práve 10 obrazových studií po dobu devíti dnů a jeden den jich přenese 30, vyjde nám průměrná hodnota 12. Proto v některých případech používáme jinou charakteristiku polohy a sice medián. Medián je takový prvek x_k statistického souboru $x_1, x_2, ...x_n$, pro který platí, že počet prvků x_i takových, že $x_i < x_k$ je stejný, jako počet prvků x_j takových, že $x_j > x_k$. Tj. x_k leží přesně "uprostřed" hodnot $x_1, x_2, ...x_n$. V praxi je však výpočet mediánu výrazně pracnější, než výpočet střední hodnoty, resp. aritmetického průměru, proto jej používáme jen tam, kde aritmetický průměr není vhodný.

Pro popis variability měřených nebo pozorovaných dat používáme více charakteristik:

- Variační rozpětí
- Kvantily
- Směrodatná odchylka
- Variační koeficient

Variačním rozpětím R statistického souboru $x_1, x_2, ...x_n$ rozumíme rozdíl největší a nejmenší hodnoty tohoto souboru:

$$R = x_{max} - x_{min} \tag{3.7}$$

Kvantily jsou takové hodnoty, pro které platí, že příslušný počet prvků statistického souboru $x_1, x_2, ...x_n$ má hodnotu vyšší, resp. nižší, než daný kvantil. Hovoříme zpravidla o dolním či horním kvartilu jako o takové hodnotě, že 75% prvků statistického souboru $x_1, x_2, ...x_n$ má hodnotu vyšší, resp. nižší, než tento kvantil. Obdobně mluvíme o decilech v případě, že tuto vlastnost požadujeme pro 90% hodnot, případně o dalších percentilech. Obdobně jako variační rozpětí můžeme definovat i percentilové rozpětí. V praxi se však příliš nepoužívá. Horní percentil, tj. číslo "nad kterým leží" jen 1% pozorovaných nebo očekávaných hodnot často používáme jako prahovou hodnotu pro určení toho, kdy již stav systému považujeme za neobvyklý a je vhodné vyvolat manuální intervenci.

Nejčastěji používanou charakteristikou míry variability statistického souboru se používá výběrová směrodatná odchylka. Přívlastek výběrová se používá pro odlišení empirické charakteristiky a teoretické charakteristiky odpovídající náhodné veličiny. Výběrovou směrodatnou odchylku statistického souboru $x_1, x_2, ...x_n$ definujeme jako

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{3.8}$$

U teoretickým modelů se jako míra variability náhodné veličiny X používá zpravidla rozptyl

$$D(X) = E([X - E(X)]^2). \tag{3.9}$$

Protože rozptyl nemá stejnou jednotku, jako hodnoty náhodné veličiny X, resp. odpovídajícího statistického souboru $x_1, x_2, ...x_n$, definujeme i pro náhodnou veličinu X směrodatnou odchylku jako

$$\sigma(X) = \sqrt{D(X)}. \tag{3.10}$$

V praxi nás však často zajímá relativní velikost výběrové směrodatné odchylky vzhledem k hodnotě aritmetického průměru. Proto zavádíme ještě další charakteristiku a tou je variační koeficient

$$v_x = \frac{s_x}{\overline{x}}. \tag{3.11}$$

3.1.1 Statistická závislost dvou náhodných veličin

Pro analýzu datových přenosů budeme využívat i statistickou závislost náhodných veličin. Statistická závislost nemusí znamenat a v mnoha případech ani neznamená kauzalitu. Můžeme ji s výhodou použít v situacích, kdy potřebujeme zjistit, jestli se náš zkoumaný jev chová obdobným způsobem jako jiné jevy podobného charakteru. Např. rozdíl v objemu přenesených dat v den státního svátku a v den následující u nemocnice A bude korespondovat s rozdílem v objemu přenesených dat v těchto dnech u nemocnice B. Přitom mezi těmito jevy není příčinná souvislost, ale protože oba jevy souvisí se stejným kalendářem, je zde jistá statistická závislost.

Pro popis statistické závislosti dvou jevů používáme obvykle Pearsonův korelační koeficient. Existují i další možnosti, např. Spearmanův korelační koeficient, ale v našem případě je Pearsonův korelační koeficient (dále jen korelační koeficient) zcela vyhovující.

Korelační koeficient dvou statistických souborů $X = x_1, x_2, ...x_n$ a $Y = y_1, y_2, ...y_n$ je definován vztahem

$$r = \frac{\sum_{i=1}^{n}(x_i-)(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}. \tag{3.12}$$

Korelační koeficient vyjadřuje míru závislosti odchylky od průměrné hodnoty u dvou statistických souborů. Vztah 3.12 je možné upravit do tvaru jednoduššího pro výpočet

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{(n-1)s_x s_y}. \tag{3.13}$$

Korelační koeficient nabývá hodnot od -1 do 1, přičemž hodnoty blízké 1, resp. -1 znamenají silnou lineární závislost (přímou, resp. nepřímou) statistických souborů, zatímco hodnoty blízké 0 signalizují, že mezi sledovanými statistickými soubory lineární závislost není.

Koeficient korelace mezi dvěma statistickými soubory má samozřejmě i svůj protějšek v teorii pravděpodobnosti v podobě koeficientu korelace dvou náhodných veličin. Ten je pro náhodné jevy X a Y definován jako

$$\rho_{X,Y} = \frac{E((X - \mu_X)(Y - \mu_Y}{\sigma_X \sigma_Y}. \tag{3.14}$$

Využití korelace mezi objemem přenášených dat v jednotlivých nemocnicích nám umožní přesněji identifikovat očekávaný průběh přenosu dat.

3.2 Entropické modely

Existuje celá řada prací, které se pro detekci anomálního chování sítě, zejména pro detekci bezpečnostních útoků, snaží využít míru neuspořádanosti, nebo složitosti datových toků. Jednou z velmi zajímavých prací na toto téma je [59], kde autoři využili entropii k detekci tehdy aktuálních internetových útoků, tzv. červů (Nachi worm, Welchia worm, Blaster worm a další). Studiu chování tohoto typu škodlivého software se v té době věnovala řada prací, např. [34], a analýza síťového provozu a jeho anomálií byla přirozeným vyústěním [35], [55], [61], [21]. Následovala celá řada pokračovatelů, kteří se snažili optimalizovat vzorkování datových toků pro přesněší a zejména rychlejší detekci, jako např. [39], [23], [9], [29]. Současně se zkoumaly i další možnosti využití statistických metod pro analýzu datových toků [60], [33], včetně distribuce pravděpodobnosti [22], [36]. Velmi zajímavá je i práce, která zkoumá možnosti využití klasických partií matematické statistiky - testování hypotéz [32].

Nutno dodat, že v těchto pracech byla entropie použita jako odhad jiné míry složitosti či strukturovanosti zachycených vzorků dat, která by z teoretického pohledu více odpovídala situaci. Touto mírou je Kolmogorovská složitost (Kolmogorov Complexity) [27], [37].

Na rozdíl od entropie, která popisuje průměrný očekávaný informační obsah zprávy nebo symbolu, který je vybrán jistým nahodilým postupem z dané množiny zpráv, nebo symbolů, popisuje Kolmogorovská složitost informační obsah dané zprávy nebo symbolu. Kolmogorovskou složitost daného objektu můžeme formálně definovat jako minimální velikost popisu (slovního či algoritmického) tohoto objektu, tj. např. jako minimální velikost počítačového programu, kterým je možné daný objekt vygenerovat. Pro praxi je však přímé použití Kolmogorovké složitosti nepříliš vhodné, proto bývá k jejímu odhadu využívána právě entropie.

V mnoha praktických aplikacích, včetně [59] se pro odhad velikosti entropie datového vzorku využívají standardní kompresní algoritmy jako je např. Lempel–Ziv– Oberhumer, který používá všeobecně známý komprimační nástroj ZIP. Jako odhad entropie dat využijeme poměr velikosti původních a komprimovaných dat, který většina implementací tohoto algoritmu poskytuje.

Entropie byla původně využívána v oblasti termodynamiky pro popis rozložení energie v systému. Tímto modelem se v průběhu dalších let inspirovala informatika a využila entropii k popisu množství přenášené informace. Používané entropické modely a možnosti jejich použití pro detekci nežádoucích aktivit v rámci systému přenosu medicínských obrazových dat budou podrobněji diskutovány v samostatných podkapitolách.

Entropie tak, jak ji chápeme v informatice se zpravidla odkazuje na původní práci Clauda Shannona [44] a tato disciplína byla dále rozvíjena v polovině minulého století [38]. Teorie entropie se samozřejmě rozvíjela i čistě matematickým směrem, velmi zajímavá je například práce [58], to už je ale mimo možnosti přímého využití pro detekci nežádoucích přenosů dat v medicínském prostředí. Zajímavý algebraický přístup k definici entropie je v článku [6]. Entropie exponenciálního typu, kam patří např. Tsallisova entropie, jsou diskutovány v [57]. To je jeden z důležitých modelů entropie, který je možné použít pro analýzu spektra příjemců snímků přenášených systémem Redimed. Entropie tohoto typu jsou parametrizovatelné, proto

je možné je přizpůsobovat konkrétním aplikacím [56].

Entropie není vhodná pro analýzu datových toků ve smyslu počtu přenesených studií a objemu přenesených dat, poskytuje ale velmi zajímavé výsledky v oblasti analýzy složení komunikujících partnerů. Nejdůležitější entropické modely, které jsem pro tuto analýzu využil, jsou popsány v následujících kapitolách.

V informatice je nejvíce využívaným modelem entropie je entropie nazvaná po Claude Elwoodovi Shannonovi. Vedle Shannonovy entropie existuje ještě řada dalších modelů. Z těch známějších jmenujme alespoň Tsallisovu entropii a Rényiho entropii.

Rényiho entropii $H_\alpha(S)$ pro systém S s konečnou množinou stavů $s_1, s_2, ...s_n$ s pravděpodobnostmi výskytu těchto stavů $P(s_i)$ definujeme vztahem

$$H_\alpha(S) = \frac{1}{1-\alpha} log_2(\sum_{i=1}^{n} p_i^\alpha), \tag{3.15}$$

kde $\alpha \geq 0$ a $\alpha \neq 1$ Pro $\alpha- > 1$ Rényiho entropie konverguje k Shannonově entropii. Při $\alpha- > \infty$ tento model entropie zdůrazňuje stavy s nejvyšší pravděpodobností výskytu a pro $\alpha- > 0$ naopak význam častěji se vyskytujících stavů klesá.

Tsallisovu entropii $H_q(S)$ pro systém S s konečnou množinou stavů $s_1, s_2, ...s_n$ s pravděpodobnostmi výskytu těchto stavů $P(s_i)$ definujeme vztahem

$$H_q(S) = \frac{1}{1-q}(\sum_{i=1}^{n} p_i^q - 1), \tag{3.16}$$

kde $q \neq 1$. Zde podobně jako v případě Rényiho entropie parametrem q určujeme výslednou podobu modelu.

Po důkladném zkoumání vlastností jednotlivých modelů entropie byl jako nejvhodnější model pro detekci anomálií v objemu přenosu medicínských obrazových informací zvolen Rényiho model entropie.

4 Netechnické aspekty detekce úniku dat

Pokud potřebujeme skutečně funkční systém pro detekci anomálních přenosů dat, musíme kromě sofistikovaných technických či matematických postupů vzít v úvahu i některé aspekty netechnické. Jedná se zejména o právní problematiku, neboť shromažďování a zpracování dat podléhá určitým právním úpravám a v neposlední řadě i o záležitosti psychologické, protože výstupy detekčního systému budou následně zpracovávány lidskou obsluhou a je proto třeba brát v úvahu i reakci lidí na stereotypní výstupy a pod.

4.1 Právní aspekty

Úkol zabezpečit výpočetní a komunikační systém proti neoprávněné manipulaci s daty úzce souvisí i s právní problematikou. Hranice mezi tím, co reálně potřebujeme pro zajištění bezpečnosti provozu a tím, co vyžadují nejrůznější zákonné předpisy a normy je mnohdy velmi tenká a snadno se můžeme dostat za hranu zákona.

Existuje celá řada právních norem, které upravují chování uživatelů v kyberprostoru a řeší počítačovou kriminalitu a to jak na národní tak i na evropské či mezinárodní úrovni. Právní předpisy definují pojem "Počítačová kriminalita"jako trestnou činnost, které se odehrává v kyberprostoru, tj. má souvislost s informačními a komunikačními technologiemi. Výraz kyberprostor (cyberspace) byl poprvé použit spisovatelem Williamem Gibsonem v románu Neuromancer [28] z roku 1984. Román pojem kyberprostor zpopularizoval natolik, že se začal používat jako odborný termín pro „prostor“ počítačových systémů a sítí v němž probíhá on-line komunikace.

Porušení tajemství je jakékoli neoprávněné narušení posílané zprávy nebo neveřejného přenosu počítačových dat se snahou zjistit jejich obsah. Podmínkou trestnosti není, aby tento obsah musel být někomu dalšímu sdělen. Složitější na posouzení je, kdy je porušení tajemství oprávněné: k určité míře sledování může být oprávněn například zaměstnavatel při kontrole činnosti zaměstnance.

Další otázkou je, jak široce se má chápat neveřejný přenos počítačových dat. Je chráněno pouze tajemství vlastního obsahu zprávy, kvůli kterému přenos probíhá, nebo i doprovodná technická data a údaje o probíhajícím datovém provozu? V tomto bodě se právní otázky potkávají s největším právnickým fenoménem posledních let - GDPR (General Data Protection Regulation) [4].

GDPR představuje nový právní rámec ochrany osobních údajů v evropském prostoru. Cílem této právní úpravy je hájit práva občanů EU proti neoprávněnému zacházení s jejich daty, zejména osobními údaji. GDPR se týká nejen firem a institucí, ale i jednotlivců a online služeb, které zpracovávají data uživatelů. GDPR zavedlo astronomické pokuty za porušování pravidel a v mnoha případech vyvolává nejistotu ohledně toho, jaké údaje ještě můžeme zpracovávat.

Je např. IP adresa osobní údaj ve smyslu zákona? Podle způsobu přiřazení IP adresy konkrétnímu zařízení, způsobu využívání tohoto zařízení a způsobu a době uchovávání záznamů o přidělení IP adresy může ale také nemusí být z IP adresy zjistitelná (alepoň po nějakou dobu) identita jejího uživatele.

4.2 Psychologické aspekty

Jakýkoli systém detekce anomálií nemůže algoritmicky rozhodnout, zda detekovaná anomálie představuje např. bezpečnostní hrozbu, či poruchu technologie, nebo se jedná jen o odchylku od běžného stavu, která má racionální vysvětlení a nepředstavuje skutečný problém. Systém potřebuje lidskou obsluhu, která provede příslušné vyhodnocení a rozhodnutí a případně spustí odpovídající reakci. Pro správné fungování detekčních systémů je zapotřebí vzít v úvahu i určité známé vzorce lidského chování. Například, pokud se bude příliš často opakovat falešný poplach, lidská obsluha se v relativně krátkém čase na takovou situaci adaptuje tím, že poplachové zprávy bude apriori považovat za falešný poplach a bude je prostě ignorovat.

Mám dlouholetou zkušenost s provozem rozsáhlé datové sítě, s dohledovým centrem pracujícím v režimu nepřetržitého provozu. Není jednoduché přesvědčit obsluhu dohledového cen-

tra, aby každému hlášení dohledového systému věnovala dostatečnou pozornost. Hlavními problémy, se kterými jsem se v praxi setkal jsou:

- Často se opakující falešný poplach. Pokud se falešný poplach nebo hlášení o chybě častěji opakuje, obsluha na něj přestává reagovat. Typickým příkladem je plánovaná údržba rozložená do více etap, pokud není v předstihu nahlášena dohledovému centru. V praxi jsem se opakovaně setkal s následující situací: systém nahlásil výpadek spojení k zákazníkovi, obsluha po telefonickém rozhovoru zjistila, že zákazník provádí údržbu svého zařízení a jedná se tak o plánovaný, lež nenahlášený výpadek. Po pár dnech se situace opakovala. Při třetím, maximálně čtvrtém opakování situace již obsluha dohledového centra zákazníka nekontaktovala a prohlásila, že "zákazník určitě zase provádí neohlášenou údržbu"aniž by zjišťovala skutečný stav věci.
- Často se opakující problémy, které se nakonec "vyřeší samy". Typickým příkladem jsou opakující se krátkodobé výpadky napájení. I v tomto případě obsluha dohledového centra velmi rychle dospěje do stavu, že v případě signalizace výpadku uzlu, který byl tímto problémem postižen, prostě prohlásí "To bude určitě zase výpadek napájení, počkáme, jestli se to za hodinu nespraví samo.".
- Příliš velká úroveň vnoření dohledovaných prvků. V situaci, kdy na přehledové mapě stavu sítě máme pod jednou ikonkou, která barevně signalizuje stav sítě, schovaný celý kampus, bude výpadek jednoho (třeba i nepodstatného) síťového prvku signalizován změnou barvy ikony pro celý kampus. Při změně barvy obsluha dohledá příčinu signalizace poruchy, dále však již stav zbytku kampusu nekontroluje s tím, že "červenou barvu této ikony způsobuje nefunkční dohledový modul UPS v rozvodné skříni na půdě".

Při konstrukci jakéhokoliv systému pro detekci anomálních stavů přenosu dat musíme mít na zřeteli všechny tyto právní a psychologické apekty.

5 Analýza logů systému Redimed

Jak již bylo zmíněno v úvodu práce, ke dni odevzdání této práce měl medicínský komunikační systém Redimed něco přes 570 uživatelů. Ne všichni uživatelé používají Redimed stejným způsobem a stejnou měrou. Řada uživatelů je pouze pasivními příjemci dat. Jedná se zejména o privátní praxe radiologů, kteří se věnují vyhodnocování snímků zaslaných z jiných zdravotnických institucí. Dalšími typickými uživateli tohoto typu jsou praktičtí lékaři, kteří tak mají k dispozici obrazovou dokumentaci pacienta, kterého např. odeslali do jiného zdravotnického zařízení a mohou mu následně podrobněji vysvětlit způsob a průběh léčby v nemocnici apod.

Počet uživatelů, kteří aktivně odesílají data je mnohem menší, než celkový počet uživatelů systému. Vývoj počtu aktivních odesilatelů obrazových informací je zachycen v tabulce 5.1. Počet aktivních odesilatelů průběžně roste po celou dobu existence systému Redimed. Dá se říct, že v roce 2015 systém překonal počáteční fázi, kdy se uživatelé teprve seznamovali s jeho možnostmi a hledali vhodný způsob využití odpovídající právě jejich potřebám a režimu práce. Pro analýzu a predikci toho, jak by se měl systém chovat a co už je odchylka od očekávaného

stavu, na kterou by bylo vhodné upozornit správce, proto použijeme data z let 2015 - 2018.

Tab. 5.1: Počet aktivních odesilatelů medicínských obrazových informací v systému Redimed.

Rok	2011	2012	2013	2014	2015	2016	2017	2018
Počet aktivních odesilatelů	83	105	105	127	145	172	183	198

I uživatelé, kteří aktivně odesílají snímky pomocí systému Redimed, využívají tento systém různým způsobem. Je zde řada uživatelů, kteří si systém Redimed chtěli jen vyzkoušet, nebo kteří mají jen velmi malé potřeby odesílat vlastní snímky spolupracujícím institucím. Uživatelé této kategorie odesílají nejvýše desítky až stovky snímků ročně. Pak jsou tady uživatelé střední velikosti, tj. uživatelé, kteří odesílají průměrně alespoň 5 snímků každý pracovní den, tj. zhruba 1000 a více snímků ročně. Systém Redimed má i určitý počet velkých uživatelů, kteří odesílají průměrně desítky snímků denně a některé speciální uživatele. Speciálními uživateli jsou například stanice určené k přeposílání nikoli celých snímků, ale pouze jejich hlaviček do systému pro výpočet radiační zátěže pacientů.

5.1 Analýza provozu malých uživatelů

U malých uživatelů je prakticky nemožné provádět nějaké statistické vyhodnocení jejich provozu způsobem, který by umožňoval predikovat očekávaný profil provozu a upozornit na neobvyklý stav. Důvodem je relativně velký rozptyl vyhodnocovaných dat, který by vedl buďto k situaci, že bude relativně málo citlivý na změnu počtu přenášených studií a tím i relativně benevolentní k případnému útočníkovi, nebo naopak příliš striktní a v tom případě by generoval příliš mnoho falešných poplachů. Pro ilustraci se podívejme na profil provozu polikliniky z jednoho menšího města (pod 10 tis. obyvatel), která odesílá zhruba 200 snímků ročně.

Protože se snažíme najít především situace, kdy uživatel (naše příkladová poliklinika) odesílá více dat, než je obvyklé, můžeme pro analýzu provozu použít upravený vstupní soubor, v němž využijeme pouze data těch měsíců, ve kterých bylo odesláno alespoň 10 studií. Základní popisné statistiky ročního profilu provozu této polikliniky ve smyslu počtu přenesených studií v původním i upraveném souboru jsou shrnuty v tabulce 5.2.

Z tabulky 5.2 lze snadno vidět, že počty odesílaných studií v jednotlivých měsících jsou velmi proměnlivé a i v případě, kdy neuvažujeme měsíce, ve kterých není žádný nebo jen minimální provoz, kolísá počet odeslaných studií o téměř 40%. K této skutečnosti je třeba ještě přičíst fakt, že uživatel může kdykoli odeslat jednotky studií navíc, např. z důvodu testování spojení. V případě této polikliniky např. 5 studií navíc znamená o 10% vyšší počet studií, než je aktuální dosažené měsíční maximum.

Z uvedeného přehledu je zřejmé, že v případě malých uživatelů radiologického komunikačního systému Redimed, sofistikovaná matematická analýza datových toků nedává příliš smysl.

Tab. 5.2: Analýza měsíčního počtu odesílaných studií příkladové polikliniky.

Statistika	**Původní soubor**	**Upravený soubor**
Rozsah souboru	12	8
Variační rozpětí	47	33
Průměr	17,33	25,88
Směrodatná odchylka	11,83	9,92
Variační koeficient	0,68	0,38

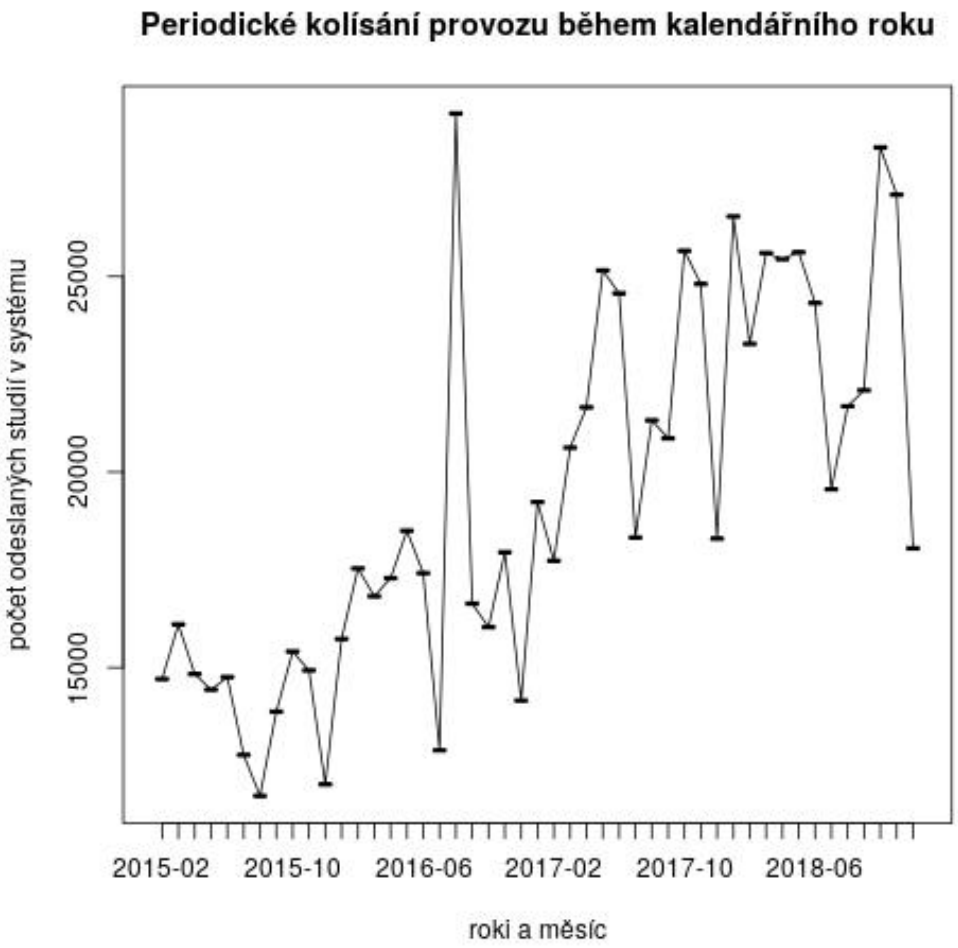
U uživatelů této velikosti je nejlépe použitelnou metodou prosté sledování počtu odeslaných studií v každém měsíci. Pokud by počet počet odeslaných studií překročil vhodně stanovený násobek maximální hodnoty uplynulého roku je vhodné uživatele upozornit. Jako "bezpečná hodnota", tj. stav, kdy nebudeme generovat příliš mnoho planých poplachů a dokážeme zachytit nástup případného útoku, se jeví dvojnásobek maxima odeslaných studií za měsíc (měřeno v předchozím kalendářním roce). Takovéto nastavení kontroly dokáže vstřebat i průběžný mírný nárůst komunikace uživatele.

5.2 Analýza provozu velkých uživatelů

U velkých uživatelů, kteří odesílají dostatečné množství snímků, bývá provoz ustáleněší a relativně méně variabilní (ve smyslu velikosti variačního koeficientu), proto můžeme použít více matematických nástrojů. V případě velkých uživatelů množství odesílaných studií více koresponduje s běžným kalendářem a můžeme proto s výhodou použít některé metody analýzy časových řad a vzájemnou korelaci statistik z různých zdravotnických institucí pro zpřesnění odhadu očekávaného počtu odesílaných studií.

U dostatečně velkých uživatelů, tj. takových kteří odesílají v průměru alespoň 5 snímků denně (stačí v pracovní dny) už se projevují periodické znaky chování uživatelů svázané s kalendářem. Na průběhu grafu počtu odeslaných snímků je zřetelně vidět roční periodický průběh, kde se projevuje vliv letních prázdnin a vánočních svátků. Graf celkového počtu studií odesílaných systémem Redimed v průběhu posledních čtyř let je na obrázku 5.1. Na grafu je zřetelně vidět 8 lokálních minim provozu v době letních prázdnin a vánočních svátků.

Pro demonstraci možností matematické analýzy profilu datových toků jsem jsem vybral jednu z větších nemocnic v krajském městě. Nemocnice odesílá ročně zhruba 5000 snímků. Graf měsíčních úhrnů počtu snímků odeslaných z této nemocnice je na obrázku 5.2. Graf zřetelně kopíruje charakteristické chování celého systému. Tuto vlastnost můžeme s výhodou využít pro modelování periodického kolísání počtu přenesených snímků během roku.

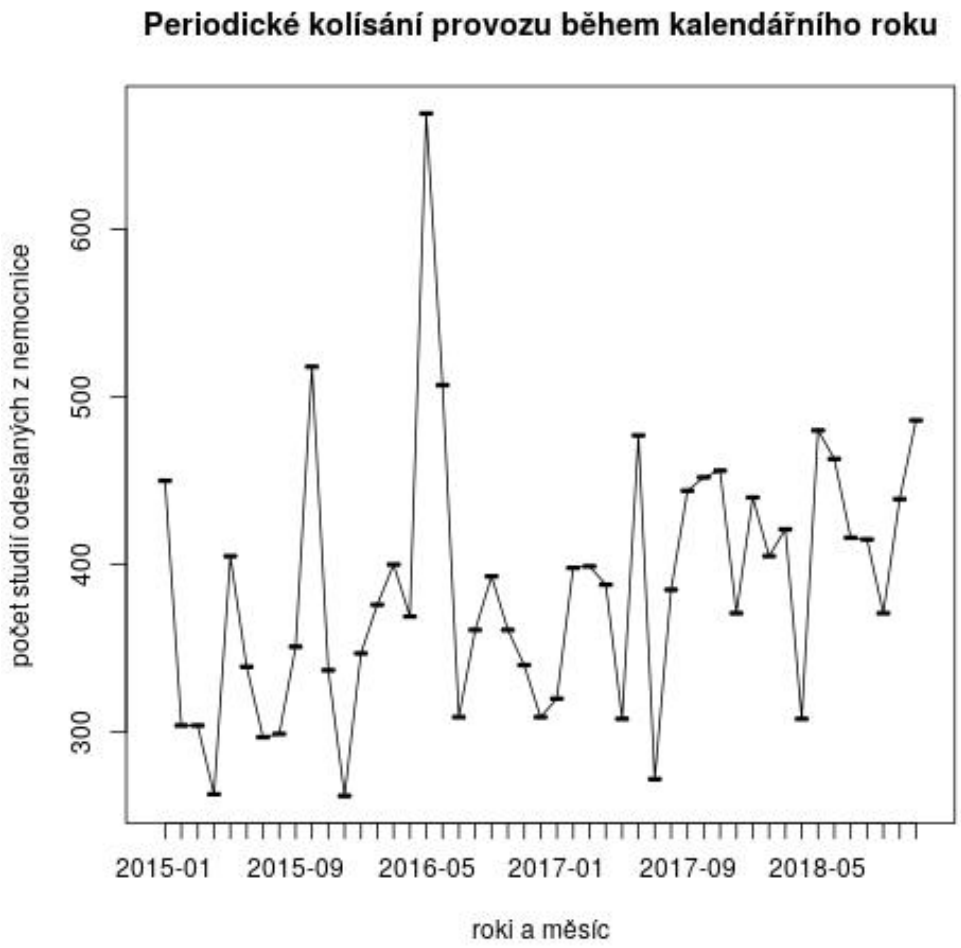


Obr. 5.1: Měsíční úhrny počtu přenesených zpráv v systému Redimed.

Pro srovnání s malou zdravotnickou organizací se podívejme, jak vypadal počet odesílaných studií během roku 2018. Statistická analýza těchto dat je uvedena v tabulce 5.3. Variační koeficient je v tomto případě výrazně menší, než u polikliniky z předchozího případu. Můžeme tedy zdánlivě přesněji předpovídat budoucí chování systému a citlivěji nastavit meze, při jejichž překročení bude systém kontaktovat správce dané nemocnice. Při podrobnějším pohledu ale zjistíme, že variační rozpětí počtu snímků odesílaných z této nemocnice je stále relativně velké. Pokud nastavíme prahovou hodnotu pro generování upozornění správců systému dostatečně vysoko (tak abysme minimalizovali vznik falešných poplachů), zůstane mám ještě pčíliš široký prostor pro případného útočníka. Pokud by se počet odesílaných snímků za měsíc zvýšil jen o málo desítek, detekční systém by na takovou situaci nereagoval.

Proto potřebujeme ještě další podpůrné analýzy, aby bylo možné zpřesnit odhad toho, jestli je objem přenášených dat v rámci obvyklého stavu.

V průběhu měsíce odesílá tato nemocnice více snímků, než poliklinika v menším městě za celý rok. To můžeme s výhodou využít pro včasnější zachycení nežádoucích toků dat, neboť na analýzu provozu této nemocnice použijeme stejnou metodu, jako na analýzu provozu menší



Obr. 5.2: Měsíční úhrny počtu zpráv odeslaných z vybrané nemocnice.

polikliniky za delší časové období.

Na počtu studií přenesených v jednotlivých dnech je dobře patrná týdenní perioda. Zároveň je ale patrné i to, že tato perioda neodpovídá kalendáři zcela přesně. Ještě lépe je to viditelné z grafu na obrázku 5.3. Tyto nepravidelnosti jsou dány nepravidelnostmi v pracovním kalendáři. Prakticky neexistuje měsíc (snad s výjimkou srpna, který je ale ovlivněn prázdninovým provozem a čerpáním dovolené u řady lidí), ve kterém by se nevyskytoval alespoň jeden státní svátek, nebo krátkodobé školní prázdniny. Z toho důvodu je poměrně komplikované využít periodické chování uživatelů pro zpřesnění odhadu počtu snímků, které mají být v daný den přeneseny. Metody analýzy časových řad, kterým jsem analýzu počtu odesílaných studií taktéž podrobil, nám sice nabízí řešení pro vyrovnání odchylek v pracovním kalendáři, bohužel ale za cenu nižší spolehlivosti odhadu.

Pro řešení tohoto problému se osvědčilo využití korelační analýzy. Vlivy nepravidelnosti kalendáře se projevují u všech nemocnic stejně. Pokud u některé nemocnice nastane výraznější pokles počtu odeslaných studií vlivem nepravidelnosti v pracovním kalendáři, dá se očekávat, že stejně budou reagovat i další nemocnice. Takový pokles nebo naopak nárůst počtu ode-

Tab. 5.3: Analýza měsíčního počtu odesílaných studií příkladové větší nemocnice.

Rozsah souboru	12
Variační rozpětí	178
Průměr	414,58
Směrodatná odchylka	55,06
Variační koeficient	0,13

slaných snímků se neprojeví úplně stejně ve všech nemocnicích. Je to dáno jednak drobnými rozdíly v organizaci práce, samozřejmě rozdíly ve velikosti a dopravní dostupnosti, ale potřebu komunikovat ovlivňují i urgentní případy, které zpravidla v době státních svátků neošetřují všechny nemocnice.

Korelace nárůstu a poklesu počtu odeslaných snímků z vybraných nemocnic je z tabulky patrná na první pohled. Pro potřeby deteke anomálií v datových tocích ji však potřebujeme uchopit vhodným matematickým nástrojem. Korelační koeficienty pro ověření dostatečné statistické vazby mezi počtem odeslaných snímků tří příkladových nemocnic jsou zde:

$$r_{A,B} = 0,69 \tag{5.1}$$

$$r_{A,C} = 0,69 \tag{5.2}$$

V obou případech vychází dostatečně vysoký korelační koeficient a proto můžeme datový provoz těchto nemocnic použít jako vzájemnou referenci.

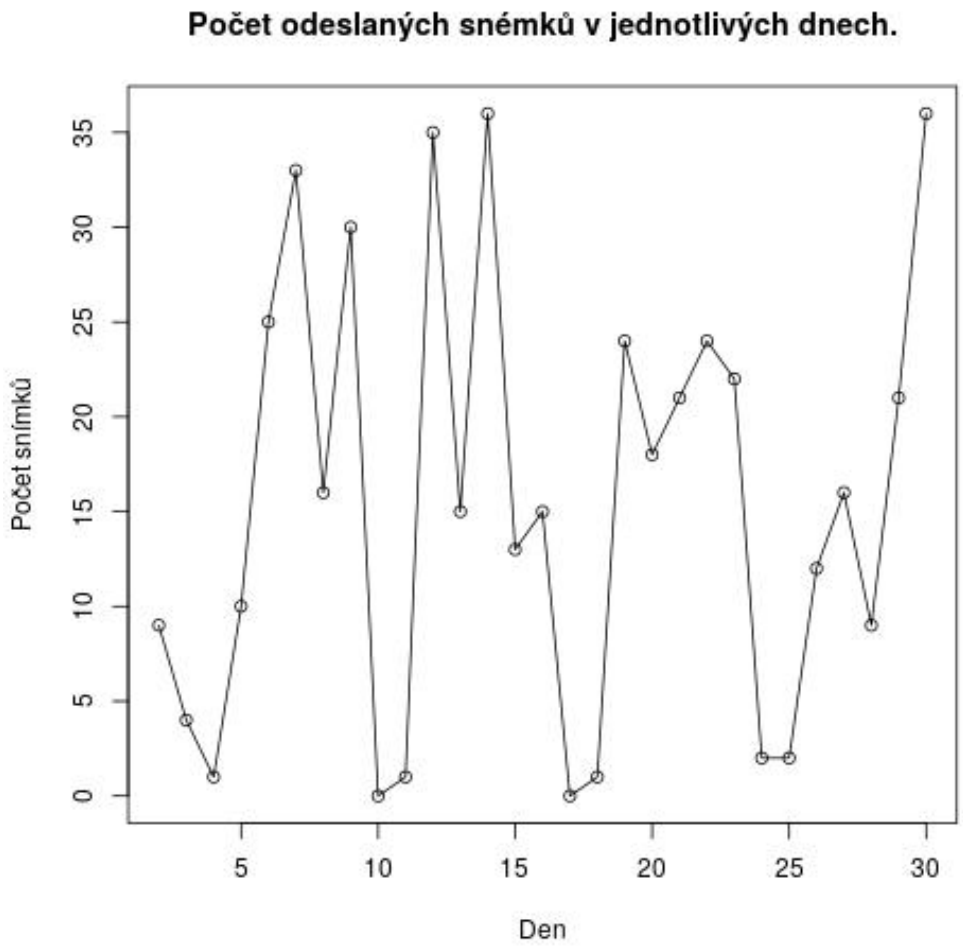
Pro odhad očekávaného počtu přenesených studií používáme relativní přírůstek nebo úbytek počtu přenesených studií mezi dvěma po sobě jdoucími dny vypočtený pomocí předpisu 5.3.

$$\delta(x_i) = \frac{x_i - x_{i-1}}{avg7(x_i)}, \tag{5.3}$$

kde $avg7(x_i)$ je klouzavý průměr počtu přenesených studií za poslední týden, tj.

$$avg7(x_i) = 1/7 \sum_{j=i-7}^{i-1} x_j. \tag{5.4}$$

Jako referenční hodnotu, ke které vztahujeme změny používáme týdenní průměr z několika důvodů: průměrná hodnota na přiměřeně dlouhé časové období do určité míry potlačuje vliv odlehlých měření a současně by se mohlo stát, že daná nemocnice např. pře víkend žádné snímky neodesílá a bylo by nutné ošetřit stavy, které by při výpočtu mohly vést k pokusu o dělení nulou. Použití relativních procentuálních přírůstků nám pomáhá vyrovnat sezónní kolísání počtu odesílaných snímků během roku a zjednodušuje srovnání podobných nemocnic, kterým řešíme nepravidelnosti v kalendáři způsobené státními svátky.



Obr. 5.3: Graf počtu odeslaných studií příkladové nemocnice během měsíce listopadu roku 2018.

U velkých zdravotnických zařízení je zapotřebí sledovat i očekávaný pokles počtu odeslaných studií v době slabého provozu. Celkový objem přenášených dat je zde natolik významný, že případný útočník by mohl snadno využít provozního sedla a v době slabého provozu odeslat data, která potřebuje, aniž by si toho někdo všimnul.

V případě velkých nemocnic, které odesílají desítky snímků denně, má smysl se zabývat i rozložením provozu v průběhu dne. V tabulce je příklad rozložení odesílání snímků během dne u nemocnice, která byla diskutována v předchozím příkladu. V tomto případě jsem použil data ze srpna letošního roku, protože historické záznamy se s touto přesností neuchovávají.

U nemocnice této velikosti nemáme dostatečné množství dat k tomu, aby bylo možné konstruovat spolehlivé statistické modely rozložení datového provozu během dne. Přesto však máme k dispozici některé základní údaje, které nám mohou pomoci detekovat neobvyklé stavy. U analýzy provozu během dne se však přibližujeme hranicím toho, kde analýza dat začne narážet na právní překážky, zejména v podobě GDPR.

Z analýzy rozložení provozu během dne můžeme snadno zjisti odchylky od běžného stavu

způsobené např. tím, že pracovník, který je za odesílání snímků zodpovědný, začal pracovat se zpožděním, měl delší obědovou pauzu a podobně. To je stav, který by mohl vést k řadě nepříjemností a zhoršování vztahů s uživateli systému. Účelem systému pro detekci neobvyklého provozu není zkoumat pracovní morálku zaměstnanců připojených institucí. Přesto má smysl se u velkých nemocnic zabývat i problémem rozložení provozu během dne. Při vyhodnocování analýzy je však třeba velké opatrnosti, protože zde lze očekávat určité množství anomálií, které mají přirozené vysvětlení a nepředstavují bezpečnostní hrozbu. Pokud např. laborant zodpovědný za odesílání snímků ráno zaspí, je pravděpodobné, že snímky, které během dne vznikly bude odesílat později, než obvykle a podle technických možností může být hustší provoz během dne, provoz i během obědové pauzy, nebo mohou být snímky odeslány později odpoledne nebo večer.

U velkých nemocnic je zajímavé sledovat i spektrum příjemců snímků, které daná nemocnice odesílá. Pro analýzu spektra příjemců odesílaných snímků se jako nejvhodnější jeví použití entropie. Postupně jsem testoval několik matematických modelů postavených jak na tradiční Shannonově entropii, tak i modely postavené na Rényiho nebo Tsallisově entropii s různými hodnotami parametru α, resp. q.

Jednotlivé nemocnice, které odesíllají obrazová data pomocí systému Redimed, používají různé způsoby práce: některé nemocnice mají stále spektrum partnerů, kterým posílají snímky, jiné komunikují s širším spektrem partnerů, přičemž ale jednotlivým partnerům posílají jen malé množství snímků. V obou případech je možné použít pro analýzu spektra komunikujících partnerů Shannonovu entropii, případně i tradiční statistické metody.

Nejsložitější situace nastává u nemocnic, které mají širší spektrum komunikujících partnerů, přičemž jednomu či dvěma posílají obvykle větší množství dat, než ostatním. Občasné výkyvy v množství dat posílaných "větším"partnerů způsobují, že použití statistickým metod selhává z důvodu velkého rozptylu dat. Podobně je tomu i v případě Shannonovy entropie.

V tomto případě potřebujeme parametrizovatelné mdely entropie, které je možné přizpůsobit profilu provozu dané nemocnice. Po delším testování Renyiho a Tsallisovy entropie jsem dospěl k závěru, vhodnější je použití Rényiho entropie. Parametr α je třeba přizpůsobit jak profilu provozu nemocnice, tak i počtu přenášených snímků.

Uvažujme množinu partnerů, kterým nemocnice odeslala snímky za jednotku času, v tomto případě jeden den. Pravděpodobnost jednotlivých stavů, tj. odeslání snímku danému příjemci můžeme popsat pomocí relativních četností. Z analýzy záměrně vynechávám víkendový provoz, protože během víkendu je provo minimální, někdy úplně nulový.

Při analýze struktury příjemců je třeba vzít v úvahu některé základní vlastnosti entropie:

- S rostoucím počtem komunikujících partnerů entropie obecně klesá.
- Entropie roste s rovnoměrností rozložení provozu mezi komunikující partnery.

Tyto vlastnosti je třeba v rámci konstrukce rozhodovacího kritéria vhodným způsobem vyrovnávat. Po mnoha experimentech, z nichž většina vedla k nepříliš uspokojivým výsledkům se podařilo optimalizovat kompenzace vlivu počtu málo frekventovaných příjemců následující

modifikací Rényiho entropie:

$$H_{C\alpha}(S) = \frac{1}{1-\alpha} log_2(\sum_{i=1}^{n} p_i^{\alpha}) \frac{1}{log_2^{1,3}(n)}, \tag{5.5}$$

kde pro danou příkladovou nemocnici je experimentálně zjištěná optimální hodnota parametru $\alpha = 0.8$. Korekční činitel

$$\frac{1}{log_2^{1,3}(n)} \tag{5.6}$$

kompenzuje vliv počtu pozorování, tj. v tomto případě počtu komunikujících partnerů.

Vliv počtu komunikujících partnerů na hodnotu entropie můžeme snadno demonstrovat na jednoduchém příkladu:

Nemocnice posílá 8 snímků hlavnímu partnerovi a po jednom snímku jednomu až třem dalším partnerům. Hodnoty Rényiho entropie vypočtené pro hodnotu parametru $\alpha = 0.8$ v jednoduché a modifikované verzi s kompenzací popsanou vztahem 5.6 jsou v tabulce 5.4.

Tab. 5.4: Příklad vlivu počtu komunikujících partnerů na hodnotu entropie.

Počet partnerů	H_α	$H_{C\alpha}$
8+1	0,572	0,572
8+1+1	1,030	0,566
8+1+1+1	1,409	0,572

Vypočtenou hodnotu entropie porovnáváme s horní a dolní prahovou hodnotou. Pro případ naší příkladové nemocnice je horní prahová hodnota 0,30 a dolní prahová hodnota hodnota 0,01. Překročení prahových hodnot ještě samo o sobě neznamená, že budem informovat uživatele o podezření na něco nekalého. Informaci o případném překročení prahových hodnot ještě kombinujeme s celkovým množstvím odeslaných snímků a množstvím komunikujících partnerů v daném časovém úseku. Mezní stavy, kdy např. v daný den odesíláme snímky jen jednomu z mála hlavních partnerů prostě jen z důvodu slabého provozu, není účelné se pokoušet zachytit matematickým výpočtem. Např. v případě entropických modelů bude v tomto případě entropie maximální možná. (Máme jen jednoprvkovou množinu stavů, tudíž pravděpodobnost výskytu daného jednoho stavu je 1.) Podobně je tomu v případě, že z nějakého důvodu v daný den neposíláme snímky většímu odběrateli. K tomu může být racionální důvod, např. externí radiolog, který pro danou nemocnici popisuje snímky, má dovolenou. V entropickém modelu se taková situace projeví maximalizací entropie.

Entropické modely nepoužíváme jen pro krásu této kapitoly matematiky. Účelem je odhalit neobvyklé rozložení struktury partnerů, kterým daná zdravotnická instituce odesílá snímky. Z neobvyklých stavů jsou významné jan takové, kdy by mohlo dojít k nežádoucímu úniku dat mimo zdravotnické zařízení. Tj. stav, kdy se nám objeví nový partner, kterému odesíláme větší množství snímků, nebo několik partnerů, kterým odesíláme středně velké množství snímků.

Ostatní situace, které mohou vyvolat změnu entropie, nejsou z pohledu úniku dat relevantní a je třeba je detekovat jiným způsobem, abychom potlačili vznik falešných poplachů.

6 Závěr

S růstem popularity radiologického komunikačního systému Redimed a počtem jeho uživatelů roste i nebezpečí, že se v řadách uživatelů (resp. v případě větších nemocnic jejich zaměstnanců) najde někdo, kdo bude chtít tento systém zneužít pro neoprávněné kopírování lékařské dokumentace pacientů. Proto je zapotřebí systém Redimed vybavit automatizovanými nástroji pro odhalování nežádoucí komunikace dříve, než nastane reálný pokus o zneužití tohoto systému.

Existuje široká škála matematických nástrojů, které jsou vhodné k analýze množství přenášených obrazových studií a odhalování neobvyklých datových toků. Žádný z těchto nástrojů však nedokáže pokrýt celou šíři způsobů, jakými uživatelé systém Redimed využívají. Teprve kombinací několika metod a postupů je možné vytvořit systém, který by přiměřeně citlivě reagoval na neobvyklé situace a přitom nevyvolával více než malé množství falešných poplachů. Falešné poplachy otupují pozornost lidské obsluhy, která jediná dokáže signály generované automatickým systémem analýzy dat posoudit a rozlišit situace, které jsou skutečně problematické od neobvyklých, ale přitom legitimních stavů.

Jakýkoli automatický detekční systém (pokud nemá vyvolávat enormní množství falešných poplachů) má určitou minimální hladinu citlivosti a není možné jej použít pro detekci úniku dat v množství menším, než je tato hranice. Automatický detekční systém není možné použít pro odhalení neoprávněného odesílání jednotek medicínských obrazových studií, přesto však je účelné takové systémy vyvíjet a v praxi používat, protože mohou odhalit stavy, kdy by docházelo k masivnímu úniku dat a to jsou právě ty situace, které by měl být schopen odhalit provozovatel komunikačního systému.

Matematická analýza datových toků je jen jednou ze složek zabezpečení komunikačního systému. Řeší jen jednu třídu možných útoků a neměl by odvést pozornost od zabezpečení dalších prvků celého komunikačního řetězce. Na druhou stranu matematická analýza datových toků byť nezajišťuje přímo filtrování provozu, ale slouží jako signalizace neobvyklých stavů, patří k těm nejzajímavějším metodám detekce útoků na komunikační systém. Umožňuje reagovat nejen na dosud známé typu útoků, ale upozorní i na nové typy útoků, pokud při nich dochází k úniku většího objemu dat. To jsou právě ty útoky, které jsou v případě práce s medicínskými informacemi ty nejkritičtější.

Matematické metody detekce neobvyklých datových toků je třeba neustále vyvíjet a přizpůsobovat přirozeným změnám chování uživatelů komunikačního systému, růstu objemu přenášených dat, zapojování dalších uživatelů do systému a vzniku a nasazování nových aplikací. Na druhou stranu je zde i dostatečný prostor pro další vývoj. Stále je možné zpřesňovat hranici toho, kdy je už vhodné datový tok považovat ze neobvyklý a informovat lidskou obsluhu.

Vyhledávání neobvyklých datových toků, které by mohly znamenat napadení systému vnějším či vnitřním nepřítelem je nikdy nekončící proces. V okamžiku, kdy útočník s dostatečným telekomunikačním vzděláním, zjistí, jak přesně detekční systém funguje, může upravit model útoku tak, aby zabránil nebo alespoň výrazně ztížil detekci útoku. Na druhou stranu zdokonalování metod detekce vede ke zmenšování objemu dat, které dokáže útočník získat aniž by byl odhalen.

Literatura

[1] URL <http://www.hl7.org/implement/standards/index.cfm?ref=nav>

[2] URL <https://www.dicomstandard.org/current/>

[3] URL <https://mathonline.fme.vutbr.cz/default.aspx>

[4] The EU General Data Protection Regulation (GDPR).
URL <https://eugdpr.org/>

[5] VÝUKA A VÝZKUM V OBLASTI MEDICÍNSKÝCH OBRAZOVÝCH INFORMACÍ. 2019.
URL <https://www.medimed.cz/>

[6] Amblard, P.-O.; Vignat, C.: A note on bounded entropies. *Physica A: Statistical Mechanics and its Applications*, roèník 365, è. 1, 2006: str. 50–56, doi:10.1016/j.physa.2006.01.002.

[7] Andel, J.: Matematická statistika. *SNTL/Alfa, Praha*, roèník 346, 1978.

[8] Anděl, J.: *Statisticke metody.* Matfyzpress, 2007.

[9] Choi, B.-Y.; Park, J.; Zhang, Z.-L.: Adaptive random sampling for traffic load measurement. In *IEEE International Conference on Communications, 2003. ICC'03.*, roèník 3, IEEE, 2003, s. 1552–1556.

[10] Dostál, O.: Metropolitní archív medicínských obrazových informací. *Zpravodaj ÚVT MU*, roèník XII, è. 5, 2002.

[11] Dostál, O.; Filka, M.; Slavíček, K.: Brněnská akademická ATM síť. In *Sborník mez. konference COFAX*, Bratislava: DT Bratislava, 1998, ISBN 80-233-0405-4, s. 79–82.

[12] Dostál, O.; Filka, M.; Šárek, M.: Optická síť VŠ. In *Sborník referátů konference Optické komunikace - OK 94*, Praha, 1994.

[13] Dostál, O.; Javorník, M.; Slavíček, K.: Management of Interhospital Processing of Medical Multimedia Data. In *Contemporary Trends in Top Management Education: How to Accomodate Demand and Suplply*, Brno: B.I.B.S.,a.s., 2004, ISBN 80-86575-74-8, s. 70–70.

[14] Dostál, O.; Javorník, M.; Slavíček, K.: Opportunity of Current ICT in the Processing of Medical Image Information. In *In Proceedings of the International Conference International Association of Science and Technology for Development. Mexico: EASTED*, Mexico: The International Association of Science and Technology for Development, 2006, ISBN 0-88986-545-0, s. 193–195.

[15] Dostál, O.; Javorník, M.; Slavíček, K.; aj.: MEDIMED-Regional Centre for Archiving and Interhospital Exchange of Medicine Multimedia Data. In *Proceedings of the Second IASTED International Conference on Communications, Internet, and Information Technology*, Scottsdale, Arizona, USA: International Association of Science and Technology for Development- IASTED, 2003, ISBN 0-88986-398-9, s. 609–614.

[16] Dostál, O.; Javorník, M.; Slavíček, K.; aj.: Developement of Regional Centre for Medical Multimedia Data Processing. In *Communications, Internet, and Information Technology*, St. Thomas (USA): ACTA Press, 2004, ISBN 0-88986-445-4, s. 632–636.

[17] Dostál, O.; Javorník, M.; Slavíček, K.; aj.: Integration of Telemedicine Activities in the Czech Republic. In *4th International Conference on Innovations in Information Technology, Innovations '07. IEEE*, Dubai, United Arab Emirates: UAE University, 2007, ISBN 978-1-4244-1840-4, s. 532–536.

[18] Dostál, O.; Slavíček, K.: Wireless Technology in Medicine Applications. In *Personal Wireless Communications*, Praha: Springer Verlag, 2007, ISBN 978-0-387-74158-1, s. 316–324.

[19] Dostál, O.; Slavíček, K.; Javorník, M.: System for Effective Collaboration in the Area of Medical Imaging. In *International Conference on Advanced Computer Science and Information Systems*, Bali: Faculty of Computer Science, Universitas Indonesia, Depok, 16424, 2010, s. 207 – 212.

[20] Dostál, O.; Slavíček, K.; Javorník, M.; aj.: *ICT Systems Monitoring*. Saarbrucken: LAMBERT Academic Publishing, 2012, ISBN 978-3-8473-7231-8.

[21] Duffield, N.; Lund, C.; Thorup, M.: Properties and prediction of flow statistics from sampled packet streams. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, ACM, 2002, s. 159–171.

[22] Duffield, N.; Lund, C.; Thorup, M.; aj.: Estimating flow distributions from sampled flow statistics. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, ACM, 2003, s. 325–336.

[23] Duffield, N.; aj.: Sampling for passive internet measurement: A review. *Statistical Science*, roèník 19, è. 3, 2004: s. 472–498.

[24] Dupac, V.; Huskova, M.: *Pravdepodobnost a matematicka statistika*. Karolinum, 2013.

[25] Filka, M.: *Optoelektronika pro telekomunikace a informatiku*. Prof. Ing. Miloslav Filka, Csc. a kol., 2017.

[26] Filka, M.; Dostál, O.; Slavíček, K.: ATM síť Brněnských vysokých škol. In *Sborník přednášek celostátní konference s mezinárodní účastí TELEKOMUNIKACE 98*, Brno: VUT Brno, 1998, ISBN 80-214-1140-6, s. 35–37.

[27] Fortnow, L.: Kolmogorov complexity. *Aspects of Complexity*, 2001, doi:10.1515/9783110889178.73.

[28] Haven, T. D.; Gibson, W.; Gibson, W.; aj.: *Neuromancer*. Alpha-Comic Verlag, 1990.

[29] Hohn, N.; Veitch, D.: Inverting sampled traffic. *IEEE/ACM Transactions on Networking*, roèník 14, è. 1, 2006: s. 68–80.

[30] Javorník, M.; Dostál, O.; Slavíček, K.: Regional Medical Imaging System. *World Academy of Science, Engineering and Technology*, roèník 7, 2011, ISSN 2010-376X.

[31] Javorník, M.; Dostál, O.; Slavíček, K.; aj.: Knowledge Management and Cost - Effectiveness in the Area of Medical Image Data. In *The 38th International Conference on Computers Industrial Engineering*, Beijing, China: University, Beijing,China, 2008, ISBN 978-7-121-07437-0, s. 883–887.

[32] Jung, J.; Paxson, V.; Berger, A. W.; aj.: Fast portscan detection using sequential hypothesis testing. In *IEEE Symposium on Security and Privacy, 2004. Proceedings. 2004*, IEEE, 2004, s. 211–225.

[33] Keiner, L. E.; Yan, X.-H.: A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery. *Remote sensing of environment*, roèník 66, è. 2, 1998: s. 153–165.

[34] Kim, J.; Radhakrishnan, S.; Dhall, S. K.: Measurement and analysis of worm propagation on Internet network topology. In *Proceedings. 13th International Conference on Computer Communications and Networks (IEEE Cat. No. 04EX969)*, IEEE, 2004, s. 495–500.

[35] Lakhina, A.; Crovella, M.; Diot, C.: Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM computer communication review*, roèník 34, ACM, 2004, s. 219–230.

[36] Lakhina, A.; Crovella, M.; Diot, C.: Mining anomalies using traffic feature distributions. In *ACM SIGCOMM computer communication review*, roèník 35, ACM, 2005, s. 217–228.

[37] Li, M.; Vitányi, P.; aj.: *An introduction to Kolmogorov complexity and its applications*, roèník 3. Springer, 2008.

[38] Lindley, D. V.: Information Theory and Statistics. Solomon Kullback. New York: John Wiley and Sons, Inc.; London: Chapman and Hall, Ltd.; 1959. Pp. xvii, 395. $12.50. *Journal of the American Statistical Association*, roèník 54, è. 288, 1959: s. 825–827, doi:10.1080/01621459.1959.11691207, <https://doi.org/10.1080/01621459.1959.11691207>.
URL <https://doi.org/10.1080/01621459.1959.11691207>

[39] Mai, J.; Sridharan, A.; Chuah, C.-N.; aj.: Impact of Packet Sampling on Portscan Detection. *IEEE J.Sel. A. Commun.*, roèník 24, è. 12, Prosinec 2006: s. 2285–2298, ISSN 0733-8716, doi:10.1109/JSAC.2006.884027.
URL <https://doi.org/10.1109/JSAC.2006.884027>

[40] Michalek, J.: *Matematicka statistika pro informatiky: urceno pro posl. fak. prirodoved.* SPN, 1987.

[41] Roček, A.; Javorník, M.; Slavíček, K.; aj.: Reversible Watermarking in Medical Imaging with Zero Distortion in ROI. In *Proceedings of 24th IEEE International Conference on Electronics, Circuits and Systems (ICECS 2017)*, New York: IEEE, 2017, ISBN 978-1-5386-1911-7, s. 356–359, doi:http://dx.doi.org/10.1109/ICECS.2017.8292071.
URL <https://ieeexplore.ieee.org/document/8292071/>

[42] Roček, A.; Slavíček, K.; Dostál, O.; aj.: A new approach to fully-reversible watermarking in medical imaging with breakthrough visibility parameters. *Biomedical Signal Processing and Control*, roèník 29, 2016, ISSN 1746-8094, doi:http://dx.doi.org/10.1016/j.bspc.2016.05.005.
URL <https://doi.org/10.1016/j.bspc.2016.05.005>

[43] Roček, A.; Slavíček, K.; Javorník, M.: RONI Size and another Attributes of Representative Sample of Medical Images in Common Hospital Operation, Related to Securing by Watermarking Methods. In *International Conference on Image Processing, Production and Computer Science (ICIPCS'16)*, editace P. O. M. M. Ahamed, London: URENG, 2016, ISBN 978-93-84422-62-2, s. 44–51.

[44] Shannon, C. E.: A Mathematical Theory of Communication. *Bell System Technical Journal*, roèník 27, è. 3, 1948: s. 379–423, doi:10.1002/j.1538-7305.1948.tb01338.x, <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x>.
URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x>

[45] Slavíček, K.; Dostál, O.; Javorník, M.: Mathematical Processing of Event Logs. In *New Information and Multimedia Technologies NIMT - 2010*, Brno: VUT Brno, 2010, ISBN 978-80-214-4126-2, s. 58–61.

[46] Slavíček, K.; Dostál, O.; Javorník, M.: Mathematical processing of event logs from nerwork devices. In *2010 International Conference on Intelligent Network and Computing*, Chengdu: IEEE, 2010, ISBN 978-1-4244-8271-9.

[47] Slavíček, K.; Dostál, O.; Javorník, M.; aj.: MEDIMED - Regional Centre for Medicine Image Data Processing. In *Knowledge Discovery and Data Mining*, USA: IEEE Computer Society, 2010, ISBN 978-0-7695-3923-2, s. 310 – 313.

[48] Slavíček, K.; Javorník, M.; Dostál, O.: Technology backround of international collaboration on medicine multimedia knowledge base establishment. In *Proceedings of the 2nd WSEAS International Conference on COMPUTER ENGINEERING and APPLICATIONS(CEA'08)*, Acapulco, Mexico, January 25-27, 2008: Published by WSEAS Press, 2008, ISBN 978-960-6766-33-6, s. 137–142.

[49] Slavíček, K.; Javorník, M.; Dostál, O.: Redundancy in Processing of Medical Image Data. In *Fourth International Conference on Computer Sciences and Convergence Information Technology*, Seoul, Korea: IEEE Computer Society Conference Publishing Services, 2009, ISBN 978-1-4244-5244-6, s. 519–523.

[50] Slavíček, K.; Javorník, M.; Dostál, O.: Extension of the Shared Regional PACS CenterMeDiMed to Smaller Healthcare Institutions. In *The Eleventh International Conference on Networks*, editace P. L. T. G. I. Pozniak-Koszalka, Saint Gilles, Reunion Island: IARIA, 2012, ISBN 978-1-61208-183-0, s. 83–87.

[51] Slavíček, K.; Javorník, M.; Dostál, O.: *MEDIMED Shared Regional PACS Center*. Croatia: InTech, první vydání, 2013, ISBN 978-953-51-1102-3, s. 43–62.
URL `<http://dx.doi.org/10.5772/55904>`

[52] Slavíček, K.; Ledvinka, J.; Javorník, M.; aj.: Mathematical Processing of Syslog Messages from Routers an Switches. In *Information and Automation for Sustainability*, Colombo: IEEE Catalog Number CFP0809B, 2008, ISBN 978-1-4244-2900-4, s. 463–468.

[53] Slavíček, K.; Schindler, V.; Dostál, O.; aj.: Kalman filter improvement for gyroscopic mouse movement smoothing. In *IIE Int'l Proceedings of International Conference on Research in Science, Engineering and Technology*, editace P. S. Z. Thaweesak, Kuala Lumpur: International Institute of Engineers, 2013, ISBN 978-93-82242-47-5, s. 43–48.

[54] Sikulova, M.; Karpisek, Z.: *Pravdepodobnost a matematicka statistika.* PC-DIR, 1997.

[55] Theriault, K.; Vukelich, D.; Farell, W.; aj.: Network traffic analysis unsing behaviour-based clustering. Whitepaper, BBN Technologies.

[56] Vignat, C.; Plastino, A.: Density operators that extremize Tsallis entropy and thermal stability effects. *Physica A: Statistical Mechanics and its Applications*, roèník 361, è. 1, 2006: str. 139–160, doi:10.1016/j.physa.2005.07.013.

[57] Vignat, C.; Plastino, A.; Plastino, A.: Correlated Gaussian systems exhibiting additive power-law entropies. *Physics Letters A*, roèník 354, è. 1-2, 2006: str. 27–30, doi:10.1016/j.physleta.2006.01.041.

[58] Voigt, J.: Stochastic operators, information, and entropy. *Comm. Math. Phys.*, roèník 81, è. 1, 1981: s. 31–38.
URL `<https://projecteuclid.org:443/euclid.cmp/1103920158>`

[59] Wagner, A.; Plattner, B.: Entropy based worm and anomaly detection in fast IP networks. In *14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise (WETICE'05)*, June 2005, s. 172–177, doi:10.1109/WETICE.2005.35.

[60] Xu, K.; Zhang, Z.-L.; Bhattacharyya, S.: Profiling internet backbone traffic: behavior models and applications. In *ACM SIGCOMM Computer Communication Review*, roèník 35, ACM, 2005, s. 169–180.

[61] Yegneswaran, V.; Barford, P.; Ullrich, J.: Internet intrusions: Global characteristics and prevalence. *ACM SIGMETRICS Performance Evaluation Review*, roèník 31, è. 1, 2003: s. 138–147.

[62] Zvara, K.; Stepan, J.: *Pravdepodobnost a matematicka statistika*. Matfyzpress, 2012.

Abstrakt

Tato habilitační práce se zabývá matematickou analýzou datových toků v systému výměny medicínských obrazových informací ReDiMed pro účely detekce anomálních stavů, zejména bezpečnostních incidentů případně poruch komunikačního systému. Na příkladu reálných dat komunikačního systému ReDiMed jsou demonstrovány vlastnosti a omezení klíčových metod pro co nejpřesnější detekci anomálií datových toků. Komunikační systém Redimed je určený pro elektronickou výměnu medicínské obrazové dokumentace pacientů mezi různými zdravotnickými zařízeními. Počet uživatelů systému Redimed v posledních letech významným způsobem narůstá a s nárůstem počtu uživatelů roste i riziko zneužití tohoto systému. Systém Redimed je založen na aktivním odesílání medicínských obrazových studií zdravotními institucemi. Proto je detekce anomálních datových toků významnou součástí zabezpečení systému Redimed.

Abstract

This habilitation thesis deals with mathematical analysis of data flows in system of medicine picture data exchange system Redimed. The reason of this analysis is detection of security incidents and communication system failures. Properties and limitations of key methods form point of view of anomaly detection accuracy are demonstrated on real data examples. Communication system Redimed is designed for electronic exchange of multimedia patients' documentation between various healthcare institutions. The number of Redimed users is growing rapidly during past few years. Together with growing number of users grows the risk of Redimed system abuse. The Redimed system is based on active dispatching of medicine picture data by the healthcare institutions. The traffic anomaly detection is for this reason meaningful part of Redimed system security.