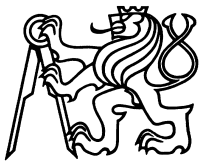
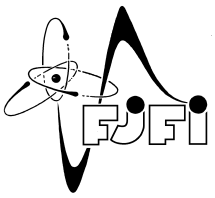


ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta jaderná a fyzikálně inženýrská



Detekce anomálií na službách v počítačové síti

Anomaly detection on computer network services

Diplomová práce

Autor: **Bc. Lukáš Kysilka**

Vedoucí práce: **Ing. Martin Kopp, Ph.D.**

Konzultant: **prof. RNDr. Ing. Martin Holeňa, CSc.**

Akademický rok: 2022/2023

České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská

Katedra: matematiky

Akademický rok 2022/2023

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student	Bc Lukáš Kysilka
Studijní program.	Aplikované matematicko-stochastické metody
Název práce (česky)	Detekce anomálií na službách v počítačové síti
Název práce (anglicky)	Anomaly detection on computer network services
Jazyk práce	čeština

Pokyny pro vypracování:

1 Seznamte se s tématem detekce anomálií v počítačových sítích.

2 Navrhněte metodu pro detekci objemových anomálií, tedy náhlých změn přeneseného objemu dat na sledovaných síťových službách

3 V návrhu metody zohledněte paměťovou náročnost, zejména pak potřebné množství historických dat pro správné fungování statistického modelu, a také výpočetní náročnost aktualizace parametrů zvoleného statistického modelu.

4 Pro zvolenou metodu navrhněte vhodné nastavení parametrů a postup jejich aktualizace

5 Experimentálně ověřte vlastnosti zvolené metody na datech z reálného síťového provozu

Doporučená literatura:

1. A. A. Cook, M. Göksel, F. Zhong, Anomaly detection for IoT time-series data: A survey. IEEE Internet of Things Journal, 2019, 6481-6494.
2. C. C. Aggarwal, An introduction to outlier analysis: In Outlier Analysis. Springer Publishing Company, Inc., 2017, 1-34.
3. M. Braei, S. Wagner, Anomaly detection in univariate time-series: A survey on the state-of-the-art. arXiv preprint, 2020.
4. P. Embrechts, C. Klüppelberg, T. Mikosch, Modelling Extremal Events. British actuarial journal, 1999, 5(2), 465-465.
5. M. Holeňa, P. Pulc, M. Kopp, Classification Methods for Internet Applications. Springer International Publishing, 2020.

Jméno a pracoviště vedoucí diplomové práce:

Ing. Martin Kopp, Ph.D.
Cisco, Karlovo náměstí 2097/10, 120 00 Praha Nové Město

Jméno a pracoviště konzultanta:

prof. RNDr. Ing. Martin Holeňa, CSc.
Ústav informatiky, Akademie věd České republiky
Pod Vodárenskou věží 271/2
182 07 Praha

Datum zadání diplomové práce: 31.10.2022

Datum odevzdání diplomové práce: 3.5.2023

Doba platnosti zadání je dva roky od data zadání.

V Praze dne 31.10.2022

garant oboru

vedoucí katedry

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta jaderná a fyzikálně inženýrská
-3-

děkan

Poděkování:
Chtěl bych zde poděkovat především svému školiteli Ing. Martin Kopp, Ph.D. za pečlivost, ochotu, vstřícnost a odbornost při vedení mé diplomové práci. Dále chci poděkovat společnosti Cisco, že mi poskytlo zázemí a příslušná data pro můj výzkum a jejím zaměstnancům, jmenovitě Jaroslavu Hlaváčovi a Janu Šterclovi za odborné rady při práci.

Čestné prohlášení:
Prohlašuji, že jsem tuto práci vypracoval samostatně a uvedl jsem všechnu použitou literaturu.

V Praze dne 3. května 2023

Bc. Lukáš Kysilka

Název práce:

Detekce anomálií na službách v počítačové síti

Autor: Bc. Lukáš Kysilka

Obor: Aplikované matematicko-stochastické metody

Druh práce: Diplomová práce

Vedoucí práce: Ing. Martin Kopp, Ph.D., Cisco

Konzultant: prof. RNDr. Ing. Martin Holeňa, CSc., Ústav informatiky, Akademie věd České republiky

Abstrakt:

Tato práce se zabývá detekcí anomálií v síťovém provozu, konkrétně na síťových službách aplikační vrstvy. Anomálie jsou v této práci definovány jako odlehlé hodnoty v množství síťové komunikace, například náhlý nárůst komunikace v krátkém časovém okně. Útočník může způsobit anomálii v síťovém provozu při neopatrném chování například při pokusu o krádež citlivých dat, prolomení hesla, nebo proniknutí do klíčových institucí. Tato práce se zaměřuje na nalezení vhodné statistické metody pro detekci volumetrických anomálií v síťové komunikaci a následnou implementaci detektoru. K nalezení vhodného statistického modelu byly použity QQ-plot, G-statistika a variační koeficient. Pro lepší adaptabilitu modelu na dynamické chování uživatelů na síti jsme do modelu přidali další parametry jako je vážení a postupné zapomínání dat. Celková funkčnost detektoru byla ověřena sérií experimentů na reálných datech z několika desítek firemních sítí.

Klíčová slova: anomálie, síťová služba, testování statistických hypotéz, Weibullovo rozdělení

Title:

Anomaly detection on computer network services

Author: Bc. Lukáš Kysilka

Abstract:

This thesis deals with anomaly detection in network traffic, specifically on network services of the application layer. In this thesis, anomalies are defined as outlying values in the amount of network traffic, such as a sudden increase in a short time window. An attacker can cause such an anomaly by his actions, for example, while attempting to steal sensitive data, cracking passwords, or penetrating key network locations. This work focuses on finding a suitable statistical method for detecting volumetric anomalies in network communications and implementing a detector based on that method. QQ-plot, G-statistic and variational coefficient were used to find such a statistical model. Additional parameters, such as weighting and gradual data forgetting, were added to make the model more adaptive to dynamic user behaviour. The detector's overall performance was verified by experiments on real data from several dozen corporate networks.

Key words: anomaly, network service, statistical hypothesis testing, Weibull distribution

Obsah

Úvod

Síťové služby jsou pro většinu společnosti zásadní a od jejich života v dnešní době neoddělitelná věc. Přestože je toto téma veřejnosti málo známé, síťové služby jsou využívány téměř neustále, v práci, doma, ve volném čase, zkrátka všude, kde se používá internetové spojení. Používají se například při posílání elektronické pošty, sdílení dat, stahování souborů nebo při hraní online her. Mezi známé síťové služby patří Hypertext Transfer Protocol (HTTP), který slouží pro přenos hypertextových dokumentů a jiných typů souboru.

Anomálie (z řeckého a-nomos) znamená nepravidelnost, výjimečnost, odchýlení se od obecného pravidla. V různých vědách se anomálie interpretuje jiným způsobem, například ve fyzice se můžeme setkat s anomálií vody. V našem odvětví, tedy v informatice a statistice, je anomálie chápána jako hodnota odlehlá od zbytku dat. V této práci data představují množství komunikací na síťových službách a odlehlá hodnota neobvykle velké množství těchto komunikací. Velký přenos dat může znamenat stahování velkého objemu dat, pokus o prolomení hesla a například u síťové služby Samba přítomnost malwaru.

Subjekt, který způsobí, že se napadené zařízení chová anomálně, můžeme nazvat útočníkem. Jeho cílem může být krádež osobních údajů, bankovních karet nebo prolomení hesla jakéhokoliv typu účtu. Dalším motivem útočníka může být proniknutí přes bezpečnostní systém do klíčových státních institucí, jako jsou exekutiva, parlament nebo také nemocnice, policie, tajné služby či dopravní nebo energetická infrastruktura a následná krádež citlivých dat.

Pro přiblížení tématu uvedu příklad. V březnu minulého roku Microsoft a platforma pro správu identit Okta detekovaly nelegální vniknutí do lokální sítě, za kterou stála nově vzniklá kriminální skupina LAPSUS$, která se specializuje na krádeže dat velkých společností a vyhrožuje jejich zveřejněním, pokud nebude zaplaceno výkupné. Skupina LAPSUS$ oznámila prostřednictvím svého kanálu na komunikační službě Telegram, že zveřejňuje zdrojový kód ukradený společnosti Microsoft. Cílem útoku této hackerské skupiny na Microsoft bylo ukrást řádově gigabyty zdrojových kódů. [1]. Stažení takového objemu dat by vyžadovalo velký přenos dat za krátký časový úsek, což je vzhledem k běžnému provozu a chování uživatelů anomální.

Protože lidí, kteří pracují na internetu a tedy využívají síťové služby, je hodně a bohužel existuje také mnoho útočníků, nemůže každou komunikaci kontrolovat člověk, ale je třeba, aby chování lidí hlídal a chránil počítačový program (detektor). Mým cílem bylo navrhnout vhodnou statistickou metodu pro takovýto detektor, implementovat ho, ověřit její funkčnost na základě historických dat, diskutovat výsledky a implementovat algoritmus do reálného provozu.

Metoda, kterou jsem zvolil, je založená na pravděpodobnosti a statistice. Z toho plyne, že pojem počet komunikací je chápán jako náhodná veličina, která se chová podle nějakého, nám neznámého rozdělení. Mým prvním úkolem bylo nalézt rozdělení, které nejlépe popisuje a predikuje počet komunikací uživatelů v počítačové síti. Podmínky pro náš distribuční model byly efektivita (rychlé vyhodnocení nové komunikace) a nízká výpočetní náročnost. Našim požadavkům nejlépe vyhovuje parametrický model, který je definován svými vnitřními parametry. Mezi parametrické modely se řadí Gaussovo, exponenciální, binomické rozdělení atd. Pro vysvětlení vnitřních parametrů uvedu příklad už zmíněného Gaussova

rozdělení. Jejími vnitřními parametry jsou střední hodnota a rozptyl. Z toho dále plyne otázka, jak nejefektivněji nalézt nejvhodnější parametrický model. Máme k dispozici velké množství historických dat a omezenou výpočetní kapacitu, proto musíme použít účinnou a úspornou statistickou metodu. Z analýzy historických dat nám vyšlo jako nejvhodnější Weibullovo rozdělení. Po nalezení nejlepšího parametrického modelu jsem definoval pravidlo, podle kterého se detektor rozhodne, jaký počet komunikací je a jaký není anomální.

Chování uživatelů na síti je velmi nepravidelné. Aktivita (počet komunikací na síťové službě) se liší v tom, zda je víkend nebo všední den, jestli je den či noc, mění se i během dne. Na tyto změny musí detektor vhodně reagovat, aby fungoval správně a detekoval relevantní hrozby. Proto druhým úkolem bylo ve zkušebním provozu optimalizovat vnější parametry detektoru. Z vnějších parametrů uveďme například čas potřebný k adaptaci modelu na data v závislosti na síťové službě a zákazníkovi. Dále pak velikost dat, která si detektor ukládá do paměti a na základě kterých aktualizuje své vnitřní parametry. Důsledkem nastavení vnějších parametrů je, s jakou flexibilitou detektor reaguje na aktuální aktivitu uživatelů. Vnějších parametrů je více a důkladně si je představíme v kapitole 5.

V následující kapitole si detailně probereme síťové služby, nad kterými má detektor pracovat. V druhé kapitole si zadefinujeme anomálii a budeme diskutovat jednotlivé metody detekce. V kapitole 3 zavedeme matematický fundament, následovaný detailním popisem nalezení správného modelu a nakonec diskusí vnějších parametrů modelu.

Kapitola 1

Síťové služby

1.1 Počítačová síť

Počítačová síť je soubor zařízení a softwaru, které umožňují komunikaci mezi počítači a dalšími zařízeními, jako jsou tiskárny, servery a úložiště. Tyto sítě jsou v dnešní době nezbytné pro mnoho podniků a organizací, protože umožňují sdílení dat a zdrojů a usnadňují spolupráci a komunikaci mezi zaměstnanci.

Počítačové sítě lze klasifikovat podle mnoha kritérií, jako je rozsah sítě, topologie sítě a druh použitého protokolu. Rozsah sítě se může lišit od malých lokálních sítí (LAN) až po globální sítě (WAN). Topologie sítě se týká fyzického uspořádání zařízení v síti a může být hvězdicová, kruhová, nebo sběrnicová síť.

Síťové služby, jako jsou e-mail, sdílení souborů a tisk, jsou důležitou součástí počítačových sítí. Tyto služby bývají součástí aplikační vrstvy sítě, která je zodpovědná za zprostředkování komunikace mezi aplikacemi na jednotlivých zařízeních. Aplikační vrstva obsahuje mnoho protokolů, jako je HTTP pro webové prohlížeče, SMTP pro e-mailové služby a FTP pro sdílení souborů.

V poslední době se objevují nové technologie, jako je Internet věcí (IoT) a 5G mobilní sítě, které umožňují větší množství zařízení, které mohou být propojené do sítě a rychlejší přenos dat. Tyto technologie přináší nové výzvy pro bezpečnost a správu sítě, ale také nabízí nové možnosti v oblasti automatizace a sledování v reálném čase.

V souhrnu lze říci, že počítačové sítě jsou důležitou součástí moderního světa a umožňují efektivní a spolehlivou komunikaci a sdílení zdrojů. Je důležité, aby byly správně navrženy a spravovány a aby byly bezpečné a spolehlivé pro své uživatele. [2]

1.2 Síťové služby

Počítačová síť a síťové služby jsou úzce propojeny a tvoří neoddělitelnou součást moderního informačního systému. Počítačová síť umožňuje přenos dat a komunikaci mezi různými zařízeními, zatímco síťové služby poskytují konkrétní funkcionality pro komunikaci a manipulaci s daty v rámci sítě. Tyto služby jsou obvykle implementovány pomocí "klient-server"architektury, kde server poskytuje službu a klient ji využívá. Klientské i serverové komponenty mohou běžet na různých počítačích v síti a komunikovat mezi sebou pomocí síťových protokolů.

Klient–server je síťová architektura, která odděluje klienta (často aplikaci s grafickým uživatelským rozhraním) a server, kteří mezi sebou komunikují přes počítačovou síť. Alternativou architektury klient–server je peer-to-peer. Klient–server popisuje vztah mezi dvěma počítačovými programy, v nichž

první program, klient, žádá o služby jiný program zvaný server. Příkladem je webový prohlížeč, tj. klientský program na uživatelském počítači, který může přistupovat k informacím na libovolném webovém serveru na světě. Na tomto modelu je založen například přístup na e-mail, web, přístup k databázi apod. A našim cílem bude najít klienta, který se na serveru chová anomálně ve smyslu velkého počtu požadavků na server.

Mezi nejpoužívanější protokoly v aplikační vrstvě patří FTP, HTTP, SMTP, POP3, IMAP, DNS a další. FTP (File Transfer Protocol) se používá k přenosu souborů mezi různými zařízeními. HTTP (Hypertext Transfer Protocol) je protokol pro přenos webových stránek a webových aplikací. SMTP (Simple Mail Transfer Protocol) je protokol pro přenos e-mailových zpráv mezi e-mailovými servery. POP3 (Post Office Protocol version 3) a IMAP (Internet Message Access Protocol) jsou protokoly pro přístup k e-mailovým schránkám na e-mailovém serveru. DNS (Domain Name System) je protokol pro překlad doménových jmen na IP adresy. Mezi další rozšířené protokoly na aplikační vrstvě patří například SSL/TLS (Secure Sockets Layer/Transport Layer Security) pro zabezpečené spojení, SSH (Secure Shell) pro vzdálený přístup k počítači a další.

Výhodou aplikační vrstvy je, že umožňuje uživatelům používat různé služby, jako jsou e-mailové služby, webové stránky nebo instant messaging, na jednom zařízení. Díky aplikační vrstvě je také možné využívat služby napříč různými platformami a operačními systémy. [3]

1.2.1 Příklady síťových služeb

V této podkapitole si uvedeme a detailněji popíšeme 4 síťové služby, ze kterých stahuji data, na kterých budu testovat navržený detektor. Zvolil jsem 2 služby, které uživatelé využívají častěji, a 2 služby, které se využívají méně.

Lightweight Directory Access Protocol (LDAP) je síťová služba, která slouží ke správě informací v adresářovém serveru. LDAP umožňuje ukládání, prohledávání a správu informací o uživatelích, zařízeních, aplikacích a dalších objektech v hierarchickém uspořádání, známém jako adresářová struktura. Tento protokol je široce používán v síťových aplikacích a operačních systémech pro autentizaci uživatelů, správu souborových systémů, řízení přístupu k síťovým zdrojům a mnoho dalších účelů. LDAP poskytuje standartizované rozhraní pro manipulaci s daty v adresáři a může být použit pro integraci s jinými síťovými službami, jako jsou například emailové servery a webové aplikace. [4]

Hypertext Transfer Protocol (HTTP) byl původně navržen k přenosu hypertextových dokumentů. V současnosti je tento protokol používaný k přenosu souborů v klient-server bázi. HTTP funguje na principu request-response mezi klientem (prohlížeč) a severem, který je postaven na tom, že klient požádá server o URL adresu, který ji následně doručí klientovi. V této komunikaci hraje nemalou roli protokol TCP, který je jakýmsi médiem mezi klientem a požadovaným serverem. Klientem může být například webový prohlížeč, zatímco serverem může být proces, pojmenovaný webový server běžící na počítači hostující jednu, nebo více webových stránek [5]. Anomálie v tomto protokolu pro mě znamená, že se uživatel abnormálně často snaží přihlásit na hypertextový odkaz, hypertextový dokument nebo na jiné zdroje.

Server Message Block (SMB) je síťový protokol, který umožňuje sdílení souborů, tiskáren a dalších prostředků mezi počítači v síti. Protokol SMB byl původně vyvinut pro operační systém Microsoft Windows, ale dnes je podporován i v jiných operačních systémech. S pomocí protokolu SMB mohou uživatelé přistupovat k souborům a tiskárnám umístěným na jiných počítačích v síti, jako by se nacházely přímo na jejich vlastním počítači. Protokol SMB funguje na aplikační vrstvě v síťovém modelu OSI a využívá transportní protokoly TCP nebo UDP pro přenos dat. Protokol SMB byl postupně vylepšován a rozšířen, včetně verzí SMB2 a SMB3, které přinášejí vylepšené funkce a zabezpečení dat. [6]

Secure Shell (SSH) je kryptografický síťový protokol, který umožňuje bezpečné vzdálené připojení k jinému počítači. SSH umožňuje šifrovat a ověřovat veškerou komunikaci mezi klientem a serverem,

čímž chrání citlivá data před odposlechy a útoky. SSH se používá pro správu a přenos dat mezi serverem a klientem, přičemž zajišťuje vysokou úroveň bezpečnosti a ochrany před zneužitím. Kromě vzdálené správy a přenosu dat lze pomocí SSH také vytvářet tunely pro bezpečné připojení k jiným službám a aplikacím. [7]

Kapitola 2

Metody detekce anomálií

2.1 Definice anomálie

V literatuře se můžeme setkat s různými definicemi anomálie. Hawkins v knize [8] definuje anomálii jako: "Pozorování, které se tak výrazně liší od ostatních, vzbuzuje podezření, že bylo generováno pomocí jiného mechanismu."

Alternativní definice anomálie od Barnetta a Lewise je: "Odlehlá hodnota je pozorování (nebo podmnožina pozorování), která se zdá být v rozporu se zbytkem tohoto souboru dat."[9]

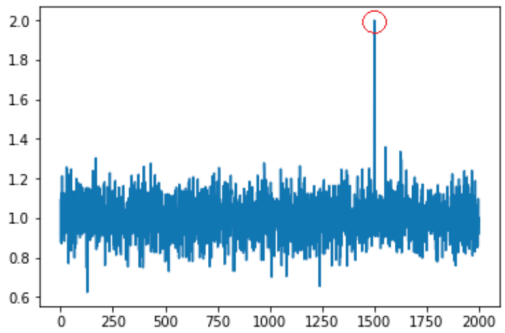
Definice anomálie podle [10] je: "Měřitelné důsledky neočekávané změny stavu systému, který je mimo jeho lokální, nebo globální normu."

Tyto definice obsahují důležité důsledky, které ve své práci používám:

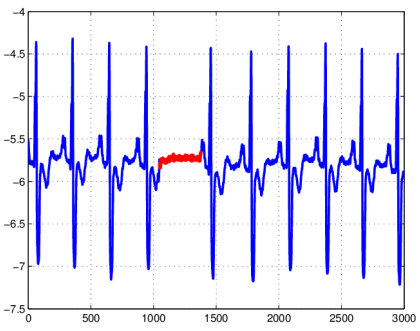
- Většina dat je neanomální (normální), jinými slovy anomálie je vzácný jev.
- Definice toho, co je normální chování, se v čase velmi rychle mění.

Uveďme si 3 základní typy anomálií.

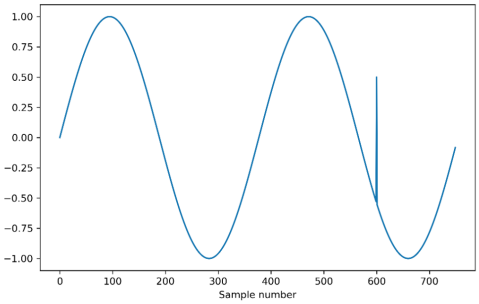
1. **Bodová anomálie:** "Pokud se pozorování výrazně odchyluje od zbytku dat, říkáme tomu bodová anomálie. Například velká bankovní transakce, která se liší od ostatních transakcí, je bodovou anomálií. Bod X_t je tedy považován za bodovou anomálii, pokud se jeho hodnota výrazně liší od všech bodů v intervalu $[X_{t-k}, X_{t+k}]$, kde $k \in \mathbb{R}$ je dostatečně velké,"obrázek 2.1.
2. **Kolektivní anomálie:** "Existují případy, kdy jednotlivé body nejsou anomálie, ale posloupnost bodů je označena jako anomálie. Například zákazník banky vybere ze svého bankovního účtu každý den v týdnu 500 USD. I když příležitostný výběr 500 USD je pro zákazníka normální, posloupnost výběrů je anomální chování,"obrázek 2.2.
3. **Kontextová anomálie:** "Některé body mohou být normální v určitém kontextu, zatímco v jiném kontextu mohou být detekovány jako anomálie: Denní teplota 30^o C u nás v létě je normální, zatímco stejná teplota v zimě je považována za anomálii,"obrázek 2.3. [11]



Obrázek 2.1: Bodová anomálie v náhodném gaussovském šumu [10]



Obrázek 2.2: Kolektivní anomálie v simulované časové řadě EKG [10]



Obrázek 2.3: Kontextové anomálie - hodnota v bodě 600 je stejná jako řada dalších pozorování $(\ldots 400, 500)$, avšak v kontextu tohoto pozorování je anomální [10]

Navržený detektor bude detekovat bodové anomálie a abychom ho zasadili do kontextu používaných metod, tak si v další sekci uvedeme pár příkladů detekce, popíšeme si pár typů detektorů, jejich podstatu, jak oni definují anomálii a na čem jsou založeny.

2.2 Úvod do metod detekce anomálií

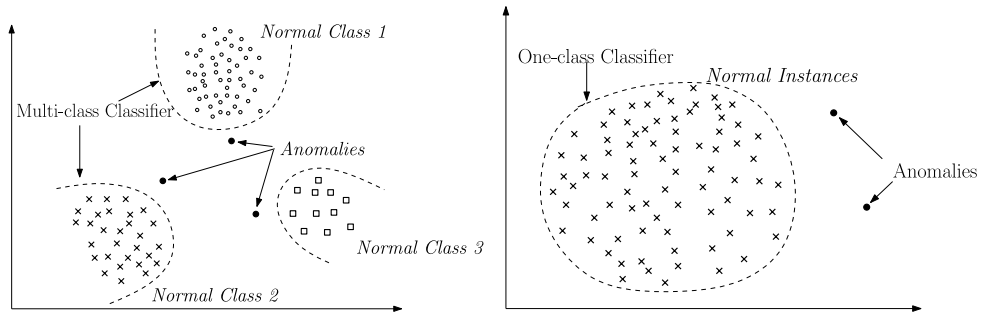
2.2.1 Detekce anomálií pomocí klasifikace

Detekce anomálií pomocí klasifikace [12] je metoda, která na základě historických dat rozhodne, jestli je dané pozorování anomální, nebo ne. Data, která se používají pro natrénování modelu, jsou označena (label) jako normální, nebo anomální. Této metodě se česky říká učení s učitelem, anglicky supervised learning.

Při detekci anomálií se používají klasifikační techniky, které pracují ve dvou fázích: v trénovací fázi se klasifikátor učí pomocí označených trénovacích dat a v testovací fázi se použije naučený klasifikátor k určení, zda je dané pozorování normální, nebo anomální.

Existují dvě kategorie klasifikačních metod pro detekci anomálií: metody více tříd a metody jedné třídy. V metodách klasifikace do více tříd učí klasifikátor rozlišovat každou normální třídu od ostatních a označuje testovací data jako anomální, pokud nebyly zařazeny ani do jedné ze tříd. Metody jedné třídy učí klasifikátor najít diskriminační hranici kolem normální množiny dat a označují testovací data jako anomální, pokud nespadají do této oblasti, viz obrázek 2.4.

Postup pro detekci anomálií pomocí klasifikace je následující:



Obrázek 2.4: Použití klasifikátoru k detekci anomálií. Na levém obrázku se klasifikuje do více tříd, na pravém obrázku se klasifikuje do jedné třídy. Převzato z [12].

1. Rozhodnutí, jak rozdělit trénovací data na skupiny.
2. Rozpoznání vlastností, které jsou vhodné pro klasifikaci.
3. Naučení modelu na základě trénovacích dat.
4. Naučený model použít ke klasifikaci neznámých dat.

Mezi klasifikátory, které se dají mimo jiné využít i pro detekci anomálií patří: různé speciální zypy neuronových sítí, jako auto-encodery [13], variační auto-encodery [14]. Další oblíbený klasifikátor pro detekci anomálií je tzv. one class support vector machines SVM [15]. Více informací o těchto metodách v [12].

Výhody a nevýhody detekce pomocí klasifikace

Výhody detekce pomocí klasifikace:

- Široce použitelné pro mnoho domén.
- Nejjednodušší přístup k hledání souvislostí v datech. [16]

Nevýhody detekce pomocí klasifikace:

- Pro natrénování modelu jsou potřeba označená data. Tyto data nejsou vždy k dispozici, nebo jich je malý počet.
- Trénování modelu je velmi časově náročné. [12]

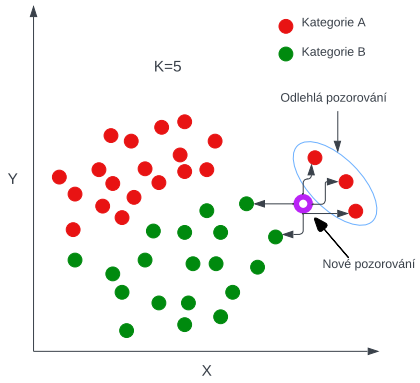
2.2.2 Detekce anomálií založené na vzdálenosti

Detekce anomálií založená na vzdálenosti je jedním z nejrozšířenějších přístupů k detekci anomálií v datech. Základem této detekční metody je měření vzdálenosti mezi jednotlivými pozorováními. Pokud je vzdálenost mezi daným pozorováním a ostatními daty v datasetu výrazně odlišná, je pravděpodobné, že se jedná o anomálii. V této detekční metodě existuje mnoho různých přístupů k výpočtu vzdálenosti a k identifikaci anomálií na základě těchto vzdáleností. Tyto metody mohou být použity na různých typech dat, včetně spojitých a kategorických dat.

Jednou z nejpoužívanějších technik detekce anomálií na základě vzdálenosti je detekce pomocí nejbližšího souseda (Nearest neighbor). Tato metoda vychází z předpokladu, že neanomální data se nachází

blízko sebe, zatímco anomálie jsou vzdáleny od svých nejbližších sousedů. [17] Tato technika vyžaduje určení vzdálenosti, případně podobnosti k nalezení nejbližších sousedů (například Eukleidovskou vzdálenost) nebo podobnosti mezi daty (například kosinová podobnost). Základním principem detekce anomálií pomocí nejbližšího souseda je: "Normální datové instance se vyskytují ve frekventovaných oblastech, zatímco anomálie se vyskytují daleko od jejich nejbližších sousedů."[12] Pojmem nejbližší soused pozorování A se myslí takové pozorování, které má ze všech dat nejmenší vzdálenost k pozorování A. Pro určení, zda je testovací pozorování anomální, se obvykle stanovuje prahová hodnota vzdálenosti.

Další metoda pro výpočet hodnoty anomálního skóre daného pozorování spočívá v počtu n nejbližších sousedů, kteří jsou vzdáleni maximálně d od daného pozorování. Tento postup může být interpretován jako odhad globální hustoty pravděpodobnosti pro každé pozorování, protože zahrnuje sousední data v hyperkouli o poloměru d. Například v 2D datové množině je hustota dat rovna $\frac{n}{\pi d^2}$. Inverzní hodnota hustoty pravděpodobnosti se používá jako hodnota anomálního skóre testovacích dat. [12] Některé techniky [18] místo výpočtu skutečné hustoty stanoví fixní poloměr d a vypočítají hodnotu anomálního skóre jako $\frac{1}{n}$, zatímco jiné techniky [18] stanoví fixní počet sousedů n a vypočítají hodnotu anomálního skóre jako $\frac{1}{d}$ viz obrázek 2.5. Více o detekci anomálií na základě nejbližšího souseda v článku [12].



Obrázek 2.5: Dvě třídy dat v příznakovém prostoru, kde fixní počet sousedů je nastaven na 5.

Výhody a nevýhody detekce založené na vzdálenosti

Výhody detekce založené na vzdálenosti:

- Není nutné znát rozdělení dat. [12]
- Užitečné pro vytváření modelů, které zahrnují nestandardní datové typy, například text. [19]

Nevýhody detekce založené na vzdálenosti:

- Není použitelné pro vysoce dimenzionální data.
- Velmi paměťově náročné. [19]

2.3 Statistická detekce anomálií

Základní princip detekce anomálií pomocí statistických metod je: Neanomální pozorování se vyskytují v oblastech s vysokou pravděpodobností stochastického modelu, zatímco anomální pozorování se vyskytují v oblastech s nízkou pravděpodobností stochastického modelu. [12].

Techniky detekce anomálií založené na statistických metodách předpokládají, že trénovací data jsou distribuovaná podle předpokládaného modelu. Tyto metody využívají statistický model, který popisuje neanomální chování dat, a následně používají statistický test k určení, zda nové pozorování patří do předpokládaného modelu. Pokud se nové pozorování vyskytuje v oblasti, která má podle natrénovaného modelu nízkou pravděpodobnost, je označena jako anomálie. Existují dva hlavní typy statistického modelu: parametrické a neparametrické. Parametrické modely předpokládají znalost systému hustot a odhadují parametry z předchozích dat. Na druhé straně neparametrické modely nevycházejí z předpokladu znalosti distribuce a statistický model je odhadován pouze na základě předchozích dat. V následujících podkapitolách budou popsány oba typy modelů.

2.3.1 Parametrické modely

Detekce anomálií pomocí parametrického modelu předpokládá, že normální data jsou generována parametrickým rozdělením s parametry θ a s hustotou pravděpodobnosti $f(x;\theta)$, kde x je pozorování. Anomální skóre pozorování x se rovná inverzní hodnotě hustoty pravděpodobnosti $f^{-1}(x;\theta)$. Parametry θ jsou odhadnuty z historických dat.

Pro určení anomálie může být použito testování statistických hypotéz. Nulová hypotéza H_0 pro takové testy je, že pozorování x bylo vygenerováno podle odhadnutého rozdělení $f(x,\theta)$. Pokud statistický test odmítne H_0, x je označeno za anomální pozorování. Testování hypotéz je spojeno s testovací statistikou, která může být později použita jako anomální skóre.

Parametrické techniky detekce anomálií lze kategorizovat podle předpokládaného typu distribuce takto:

2.3.1.1 Gaussovský model

Tato metoda detekce anomálií předpokládá, že data mají Gaussovo rozdělení. Parametry modelu jsou odhadnuty pomocí metody maximální věrohodnosti 3.6.2. Poté se použije prahová hodnota, aby se určilo, zda je pozorování anomální, nebo ne. Hodnota anomálního skóre je určena jako vzdálenost od odhadnutého průměru.

Jednoduchá technika detekce odlehlých hodnot, často používaná v oblasti procesního řízení kvality [20], prohlašuje všechna data, která jsou vzdálena více než 3σ od průměru distribuce μ, za anomální, kde σ je směrodatná odchylka. Hodnota pozorování, které má pravděpodobnost $0,997\%$ náleží intervalu $[\mu - 3\sigma, \mu + 3\sigma]$.

Pro detekci anomálií byly použity i sofistikovanější statistické metody, jak je popsáno v [9]. Dvě z nich si zde představíme.

Grubbův test, také známý jako maximální normovaný reziduální test, se používá pro jednorozměrná data, která mají Gaussovo rozdělení. Dané pozorování je prohlášeno za anomální pokud:

$$z > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}}, \tag{2.1}$$

kde n je počet pozorování, $t^2_{\alpha/(2n),n-2}$ je $\frac{\alpha}{2n}$-kvantil studentova rozdělení s $n-2$ stupni volnosti a z je anomální skóre, které se vypočítá jako:

$$z = \frac{|x - \bar{x}|}{\sigma}, \tag{2.2}$$

kde $\bar{x}$ je výběrový průměr a σ je směrodatná odchylka.

Ye a Chen [21] používají k detekci anomálií χ^2 statistiku. Trénovací fáze předpokládá, že data mají vícerozměrné normální rozdělení. Hodnota statistiky χ^2 je určena jako:

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - E_i)^2}{E_i}, \tag{2.3}$$

kde X_i je pozorovaná hodnota i-té proměnné a E_i je očekávaná hodnota i-té proměnné a n je počet proměnných. Velká hodnota značí, že vzorek dat obsahuje anomálie.

2.3.2 Neparametrické modely

Techniky detekce anomálií v této kategorii používají neparametrické statistické modely, kde statistický model není předem definován konkrétním systémem hustot, ale je stanoven z historických dat. Tyto techniky obvykle předpokládají méně vlastností dat, jako je hladkost hustoty, ve srovnání s parametrickými technikami. [12]

2.3.2.1 Histogram-based Outlier Score

V této podkapitole si představíme techniku detekce anomálií, která je založená na neparametrických statistických metodách. Histogram-based Outlier Score (HBOS) je metoda detekce anomálních hodnot, která pracuje s daty ve formě histogramu a počítá skóre pro každé pozorování na základě její pozice v histogramu.

Princip HBOS detektoru spočívá v tom, že odlehlé hodnoty se často nachází v řídce obydlených oblastech histogramu. Detektor HBOS proto pracuje tím, že rozdělí data do předem určených intervalů a zaznamená počty výskytů hodnot v každém z nich.

Výpočet anomálního skóre daného pozorování x se provede následovně: Nejprve se pro každou hodnotu určí, do kterého intervalu histogramu spadá. Poté se vypočítá pravděpodobnost každého intervalu, přičemž se předpokládá rovnoměrné rozložení hodnot v souboru dat.

$$HBOS(x) = \sum_{i=0}^{d} log\left(\frac{1}{hist_i(x)}\right), \tag{2.4}$$

kde d je počet znaků, $x \in \mathbb{R}^d$ je vektor znaků a $hist_i$ značí pravděpodobnost i-tého znaku. [22]. Znak v našem kontextu znamená počet komunikací na síťové službě. Důvod pro logaritmickou transformaci je ten, že je méně citlivý vůči extrémním hodnotám, které by způsobovaly velmi vysoké anomální skóre. [23]

Celkově lze říci, že HBOS je jednoduchý a rychlý algoritmus, který se dobře škáluje na velké množství dat. Výsledky HBOS jsou však závislé na velikosti histogramu a počtu intervalů, což může mít za následek snížení přesnosti detekce v některých případech. HBOS také není schopen detekovat odlehlé hodnoty, které jsou rozptýleny v rámci histogramu, což může být problém v datech s velkým rozptylem.

2.3.3 Výhody a nevýhody detekce pomocí statistických metod

Výhody statistických metod: [12]

- Pokud známe rozdělení dat, statistické metody poskytují řešení, které je statisticky podložené.
- Anomální skóre, které statistické metody poskytují, je spojeno s intervalem spolehlivosti, který lze použít jako doplňkovou informaci při rozhodování ohledně jakéhokoliv pozorování.
- Pokud je odhadnutý distribuční model robustní vůči anomáliím v datech, může fungovat, aniž by bylo potřeba mít označená trénovací data (*unsupervised model*).

Nevýhody statistických metod:

- Statistické metody spoléhají na předpoklad, že data pochází z určitého rozdělení, tento předpoklad často neplatí.
- Výběr statistiky pro testování hypotéz není jednoduchý úkol.
- Velký počet falešně pozitivních anomálií.

Kapitola 3

Statistická teorie

V této kapitole zavedu statistické pojmy, definice a tvrzení, na kterých vybuduji navržený detektor. Jedná se o odhadování distribučních funkcí, které se dělí na neparametrické a parametrické, dále pak metody detekce systému hustot: QQ-ploty a testy dobré shody, nakonec t-test a F-test. Pokud nebude psáno jinak, celou teorii jsem čerpal z [24].

Předpokladem pro statistickou teorii, se kterou budu pracovat, bude

- množina všech elementárních výsledků experimentu Ω, σ-algebra nad množinou Ω: $\mathcal{A}$.
- pravděpodobnostní míra na prostoru $(\Omega, \mathcal{A})$: $\mathbb{P} : \mathcal{A} \to \mathbb{R}$.
- náhodná veličina X a její konkrétní realizace x.
- výběr nezávislých a stejně rozdělených náhodných veličin *iid*.
- parametrický prostor $\Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$.
- hustota pravděpodobnosti f a její odhad $\widehat{f}$.
- kumulativní distribuční funkce F a její odhad $\widehat{F}$.
- systém hustot $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta \subset \mathbb{R}^k\}$.

3.1 Odhad kumulativní distribuční funkce

Empirická distribuční funkce je odhadem skutečné kumulativní distribuční funkce F, ze které pochází naše pozorování. Podle Glivenko-Cantelli teorému [24] konverguje s pravděpodobností 1 k tomuto skutečnému rozdělení F.

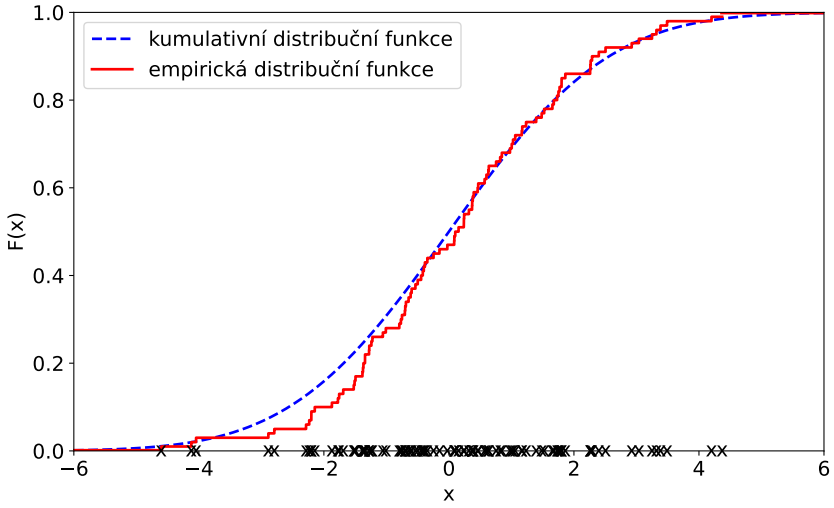
Definice 3.1.1 *(Empirická distribuční funkce). Nechť* $(X_i)_{i=1}^n \sim F$*, iid. Potom je empirická distribuční funkce definována předpisem:*

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_x(X_i), \quad x \in \mathbb{R}, \tag{3.1}$$

kde $\mathcal{I}_x(X)$ je indikátorová funkce a je definována předpisem

$$\mathcal{I}_x(X) = \begin{cases} 1, & \text{pokud } X \leq x \\ 0, & \text{pokud } X > x. \end{cases}$$

Na obrázku 3.1 vidíme odhad a skutečnou kumulativní distribuční funkci. S rostoucím počtem pozorování odhadnutá křivka konverguje ke skutečné (modré) kumulativní distribuční funkci.



Obrázek 3.1: Červená křivka reprezentuje empirickou distribuční funkci a modrá je skutečná kumulativní distribuce. Na ose x vidíme naměřená pozorování.

Definice 3.1.2 *(Jádrový odhad hustoty pravděpodobnosti). Nechť* $(X_i)_{i=1}^n \sim F$, *iid. Mějme jádro* $K(x)$ *takové, že* $K(x) \geq 0$, $\int_{\mathbb{R}} K(x)\,\mathrm{d}x = 1$, $\int_{\mathbb{R}} xK(x)\,\mathrm{d}x = 0$, $\int_{\mathbb{R}} x^2K(x)\,\mathrm{d}x > 0$. *Označme* $h \in \mathbb{R}^+$ *jako šířku okna, neboli vyhlazovací parametr. Pak definujeme jádrový odhad hustoty vztahem*

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad x \in \mathbb{R}.$$

Nejčastěji se používá uniformní, Epanechnikovo, normální, trojúhelníkové nebo Laplaceovo jádro. Bartlet-Epanechnikovo jádro má nejmenší integrated mean squared error (IMSE). Laplaceovo jádro má těžké chvosty.

3.2 PP/QQ-plot

Probability-probability plot (PP-plot) a quantile-quantile plot (QQ-plot) jsou grafické nástroje, které se používají pro správnou detekci sytému hustot. Před tím, než si zadefinujeme PP a QQ-plot, zadefinujeme si kvantil kumulativní distribuční funkce.

Definice 3.2.1 *(α-kvantil [25]). Nechť* $X \sim F$. *Definujeme* α*-kvantil rozdělení* F *jako*

$$x_\alpha = \inf\{x : F(x) \geq \alpha\}, \quad \forall \alpha \in [0, 1].$$

PP plot je graf, ve kterém vynášíme dvě distribuční funkce proti sobě a jeho definice je následující:

Definice 3.2.2 *(PP-plot). Nechť* $(X_i)_{i=1}^n \sim F$ *a iid. Nechť* $\widehat{F}_n(x)$ *je empirický odhad. Pak množinu uspořádaných dvojic*

$$\left\{F(x_{(k)}), \widehat{F}_n(x_{(k)})\right\}_{k=1}^n$$

nazveme PP-plot, kde F je náš volený statistický model. [26]

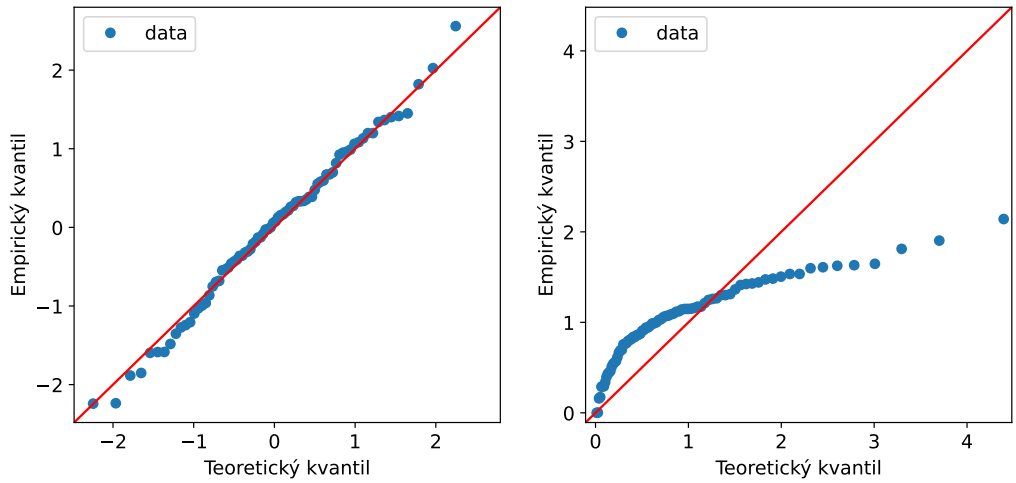
Kvantil-kvantil plot (QQ plot) 3.2 je podobný graf jako PP plot, funguje na stejném principu určení systému hustot, akorát místo distribucí se na osy vynáší jejich kvantily. Pro definici QQ-plotu potřebujeme zaručit existenci kvantilu. Ta plyne z věty 3.2.1.

Věta 3.2.1 *Nechť je kumulativní distribuční funkce spojité a prosté zobrazení, pak platí* $x_\alpha = F^{-1}(\alpha)$, *kde* $\alpha \in [0, 1]$. *[25]*

Definice 3.2.3 *(QQ-plot). Nechť* $(X_i)_{i=1}^n \sim F$ *a iid, kde F splňuje předpoklady věty 3.2.1. Nechť* $\widehat{F}_n(x)$ *je empirický odhad. Pak množinu uspořádaných dvojic*

$$\left\{F^{-1}(F(x_{(k)})), F^{-1}(\widehat{F}_n(x_{(k)}))\right\}_{k=1}^n$$

nazveme QQ-plot, kde F je náš volený statistický model. [26]



Obrázek 3.2: Na obrázcích můžeme vidět 2 QQ-ploty a s nimi vygenerovaná data, která mají normální rozdělení. Levý QQ-plot je nastavený na normální rozdělení, proto mají data lineární závislost. Pravý QQ-plot je nastavený na exponenciální rozdělení, proto na obrázku vidíme, že se empirický kvantil nerovná teoretickému kvantilu.

3.3 Statistické testy hypotéz

Testování statistických hypotéz je o tom, že na základě dat z experimentu ověříme platnost nějaké domněnky/hypotézy na úrovni celé populace či parametrů. Tato statistická disciplína nám dává poměrně mocný nástroj k tomu, abychom například rozhodli, jestli studijní výsledky žáka závisí na pohlaví, škole, rodičích, inteligenci atd.

Experiment provádíme na jednotlivých jedincích, přičemž často máme k dispozici tzv. pokusnou a kontrolní skupinu. Příklad těchto skupin může být dvojitě zaslepený experiment (double-blind experiment), pomocí kterého zkoumáme účinky daného léku. Mějme dvě skupiny pacientů, jedné podáme skutečný lék a druhé placebo, zároveň pacient neví, jestli dostal lék nebo jen placebo. Výsledky se pak zpracují právě pomocí testování statistických hypotéz. Statistické hypotézy nám říkají, jestli mezi zkoumanými skupinami je signifikantní rozdíl ve výsledcích nebo se výsledky neliší, jinými slovy, zda daný vstupní parametr má významný vliv na výsledek, nebo nemá.

Statistické hypotézy budeme používat k určení správného statistického modelu. Budeme počítat statistiky, které závisí na jednotlivých systémech hustot a tyto skupiny statistik budeme mezi sebou porovnávat. Nyní si pár statistických testů představíme.

3.3.1 Test dobré shody

Test dobré shody (Goodness of fit GoF) patří do skupiny asymptotických testů hypotéz. Tento test je jedním z velice užitečných nástrojů, jak identifikovat konkrétní typ statistického modelu z pozorovnání sledované veličiny X.

Definice 3.3.1 *(Test dobré shody). Mějme $(X_i)_{i=1}^n \sim F$, kde F je neznámá distribuce. Volme jednu konkrétní distribuci $F_0 \in \mathcal{F}$. Pak test hypotézy*

$$H_0 : F = F_0 \qquad vs. \qquad H_1 : F \neq F_0 \quad (resp. \quad F = F_1) \tag{3.2}$$

se nazývá test dobré shody (GoF) modelu na hladině významnosti $\alpha \in (0, 1)$ a testujeme H_0 na této signifikantní hranici pro chybu I. druhu.

Poznámka 3.3.1 *Hladina významnosti bude mít apriorní hodnotu $\alpha = 0,05$, kterou budu používat v této práci, pokud nebude řečeno jinak.*

Nejznámější test dobré shody se nazývá χ^2-test, který převádí celou úlohu na asymptotický test v multinomickém parametrickém modelu za cenu diskretizace oboru hodnot náhodného výběru $\mathbf{X}$. Vytvořme tedy dělení $\{A_1, \ldots, A_k\}$ oboru hodnot X na disjunktní množiny. Dále zavedeme

$$p_j = \mathbb{P}_F(X \in A_j), \qquad p_{0j} = \mathbb{P}_{F_0}(X \in A_j), \quad \forall j \in \{1, \ldots, k\}. \tag{3.3}$$

Mějme k dispozici množinu náhodného výběru pozorování $\mathbf{X} = \{X_i\}_{i=1}^n \sim F$ iid. a nechť

$$Y_j = |\mathbf{X} \cap A_j| \tag{3.4}$$

je počet těch pozorování $X_i \in \mathbf{X}$, která se vyskytují v j-tém intervalu A_j, $j \in \{1, \ldots, k\}$.

G-statistika

G-statistika je jedna ze 3 statistik, která se používá v testech dobré shody. Podle věty o ekvivalenci Pearsonovy, Neymanovy a G-statistiky všechny konvergují k χ^2 rozdělení. Pro naše účely si zadefinujeme poslední jmenovanou.

$$\Lambda_n(\mathbf{Y}) = -2 \sum_{j=1}^{k} Y_j \ln\left(\frac{np_{0j}}{Y_j}\right) \tag{3.5}$$

Poznámka 3.3.2 *Statistika 3.5 vyjadřuje Kullback–Leiblerovu divergenci mezi naměřenou a očekávanou četností v daném intervalu $D_{KL}(n\mathbf{p}\|\mathbf{Y})$, kde vektor $\mathbf{p}$ je n-tice p_j, které byly zadefinovány v rovnici* (3.3) *a $\mathbf{Y}$ je vektor naměřených četností Y_j ve všech množinách A_j zadefinovaný v rovnici* (3.4)*. Tato veličina je jenom přenásobena konstantou z důvodu konvergence k χ^2 rozdělení.*

3.3.2 t-test

Další z testů, který je velice užitečný a který budeme používat, je dvou-výběrový nepárový t-test, který porovnává střední hodnoty dvou gaussovských výběrů. Příkladem toho může být testování rozdílu střední hodnoty tlaku krve u kuřáků a nekuřáků. My budeme používat t-test pro testování střední hodnoty G-statistik a variačních koeficientů, které jsou závislé na systému hustot. Mějme tedy dva výběry:

$$X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$Y_1, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Pak podle centrálního limitního teorému platí:

$$\overline{X}_{n_1} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \tag{3.6}$$
$$\overline{Y}_{n_2} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \tag{3.7}$$

a také platí

$$\frac{(n_1 - 1)s_{n_1}^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$$
$$\frac{(n_2 - 1)s_{n_2}^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$

Budeme testovat hypotézu rovnosti středních hodnot obou náhodných výběrů dat

$$H_0 : \mu_1 = \mu_2 \qquad vs. \qquad H_1 : \mu_1 \neq \mu_2$$

na hladině α.
Rozlišujeme 3 případy:

1. Známe-li σ_1^2, σ_2^2.
2. Neznáme σ_1^2, σ_2^2 a víme: $\sigma_1^2 = \sigma_2^2$.
3. Neznáme σ_1^2, σ_2^2 a víme, že $\sigma_1^2 \neq \sigma_2^2$.

Druhá a třetí varianta se nám bude hodit. Zadefinujme si proto jejich statistiky a příslušné rozdělení:

Neznámé σ_1^2, σ_2^2, $\sigma_1^2 = \sigma_2^2$

Pokud σ_1^2, σ_2^2 neznáme, ale víme, že $\sigma_1^2 = \sigma_2^2$, pak volíme testovací statistiku jako

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2), \tag{3.8}$$

kde $s^2 = \frac{(n_1-1)s_{n_1}^2 + (n_2-1)s_{n_2}^2}{n_1+n_2-2}$ se nazývá pooled sample variance.

Neznámé $\sigma_1^2, \sigma_2^2, \sigma_1^2 \neq \sigma_2^2$

Pokud σ_1^2, σ_2^2 neznáme, ale víme, že $\sigma_1^2 \neq \sigma_2^2$, pak užíváme testovací statistiku

$$T_\nu(\mathbf{X}, \mathbf{Y}) = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2}}} \sim t(\nu), \tag{3.9}$$

kde

$$\nu = \frac{\left(\frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_{n_1}^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_{n_2}^2}{n_2}\right)^2}.$$

Pro neceločíselné ν interpolujeme $t(\nu)$ z hodnot sousedních $t(\lfloor\nu\rfloor)$ a $t(\lfloor\nu\rfloor + 1)$. Pro obě statistiky, jak pro T, tak pro T_ν, dostáváme kritický obor hodnot

$$W_\alpha = \left\{(\mathbf{x}, \mathbf{y}) : T(\mathbf{x}, \mathbf{y}) \geq t_{1-\frac{\alpha}{2}}(a)\right\},$$

kde $t_{1-\frac{\alpha}{2}}$ značí $1 - \frac{\alpha}{2}$ kvantil studentova rozdělení, a značí stupeň volnosti. Pro první případ je $a = n_1 + n_2 - 2$ a pro druhý případ $a = \nu$. $\mathbf{x}$ je vektor realizací náhodného výběru $\mathbf{X} = (X_i)_{i=1}^n$.

3.3.3 F-test

Za stejných předpokladů jako u t-testu testujeme hypotézu homogenity rozptylů dvou Gaussovských výběrů

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad vs. \qquad H_1 : \sigma_1^2 \neq \sigma_2^2 \tag{3.10}$$

na hladině α. Testovací statistiku volíme

$$F(\mathbf{X}, \mathbf{Y}) = \frac{s_{n_1}^2}{s_{n_1}^2} \sim F(n_1 - 1, n_2 - 1).$$

Kritický obor tohoto testu je

$$W_\alpha = \left\{(\mathbf{x}, \mathbf{y}) : F(\mathbf{x}, \mathbf{y}) \geq F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)\right\}.$$

3.4 Variační koeficient

V teorii pravděpodobnosti a statistice je variační koeficient (Coefficient of variation CV) také známý jako relativní směrodatná odchylka (RSD), znamená standardizovanou míru rozptylu rozdělení pravděpodobnosti. Často se vyjadřuje v procentech a je definován jako poměr rozptylu a střední hodnoty:

$$CV = \frac{\sqrt{D[X]}}{|E[X]|} \tag{3.11}$$

Tento koeficient se dá odhadnout pomocí statistiky:

$$\widehat{CV} = \frac{s_n}{|\overline{X_n}|}, \tag{3.12}$$

$\overline{X_n}$ je výběrový průměr a s_n^2 je výběrová směrodatná odchylka, později budeme značit empirical CV. [27]

3.5 Distribuce s těžkými chvosty

Na závěr této kapitoly si zadefinujeme distribuce s těžkými chvosty (heavy-tailed distribuce). Náhodná veličina, která má rozdělení s těžkými chvosty, je charakterizována velkým rozsahem svého oboru hodnot.

Existují 3 podtřídy distribucí s těžkými: fat-tailed, long-tailed a subexponential distribuce. Používají se dvě rigorózní definice heavy-tailed distribucí. Někteří autoři tento termín používají pro označení těch distribucí, které nemají všechny své momenty konečné, a někteří tento pojem používají pro distribuce, které nemají konečný rozptyl.

My si vyslovíme obecnou definici, která je nejpoužívanější a zahrnuje všechny naše později zmíněné distribuce.

Definice 3.5.1 *(Heavy-tailed distribuce [28]) Rozdělení náhodné veličiny $X \sim F$ nazveme heavy-tailed distribucí, jestliže momentová vytvářející funkce $M_X(t)$ má nekonečnou hodnotu $\forall t > 0$, neboli:*

$$\int_{-\infty}^{+\infty} e^{tx} \, \mathrm{d}F(x) = +\infty.$$

Mezi distribuce s těžkými chvosty, se kterými budeme dále pracovat, patří rozdělení Pareto, Log-normální distribuce, Lévyho distribuce, Weibullovo distribuce, Burrovo distribuce, Log-logistická distribuce, Log-gamma distribuce, Fréchetovo distribuce, Cauchyovo distribuce a Gumbelovo distribuce, které si zde představíme.

Weibullovo rozdělení

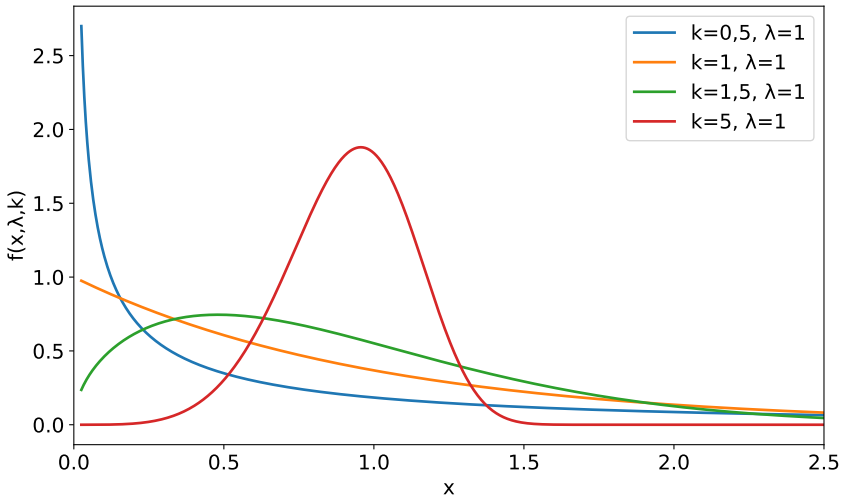
Weibullovo rozdělení je definované pro spojitou náhodnou veličinu X. Jmenuje se podle Waloddiho Weibulla, což byl švédský matematik. Funkce hustoty pravděpodobnosti Weibullova rozdělení je:

$$f(x; \lambda, k) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, & x \geq 0 \\ 0, & x < 0, \end{cases}$$

kde $k > 0$ je tvarovací parametr a $\lambda > 0$ je parametr měřítka. Její kumulativní distribuční funkce je definovaná jako:

$$F(x; \lambda, k) = \begin{cases} 1 - e^{-\left(\frac{x}{\lambda}\right)^k}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Weibullovo rozdělení souvisí s řadou dalších distribucí. Pro $k = 1$ máme exponenciální rozdělení a pro $k = 2$ a $\lambda = \sqrt{2}\sigma$ dostáváme Rayleighovo rozdělení. Na obrázku 3.3 můžeme vidět hustotu pravděpodobnosti pro konkrétní parametry. [29]



Obrázek 3.3: Hustota pravděpodobnosti Weibullova rozdělení pro jednotlivé parametry.

Vidíme, že Weibullovo rozdělení je velice flexibilní a pro správně naladěné parametry má opravdu těžké chvosty (modrá křivka). Toto rozdělení má široké využití. Používá se například pro analýzu přežití (očekávaná doba, než dojde k jedné události, jako je smrt biologického organismu), v teorii spolehlivosti, v hydrologii, v ekonomii, v sociologii, v předpovědi počasí, v rozdělení rychlosti větru atp. [30]

Exponenciální Weibullovo rozdělení

Exponenciální Weibullovo rozdělení je definované pro spojitou náhodnou veličinu X. V roce 1993 jej definovali Mudholkar a Srivastava v článku [31]. Exponenciální Weibullovo rozdělení rozšiřuje Weibullovo rozdělení o jeden tvarovací parametr. Funkce hustoty pravděpodobnosti exponenciálního Weibullova rozdělení je:

$$f(x;k,\lambda,\alpha) = \begin{cases} \alpha\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\left[1-e^{-(x/\lambda)^k}\right]^{\alpha-1}e^{-(x/\lambda)^k}, & x \geq 0 \\ 0, & x < 0, \end{cases}$$

kde $k > 0$ je první tvarovací parametr, $\alpha > 0$ je druhý tvarovací parametr a $\lambda > 0$ je škálovací parametr. Její kumulativní distribuční funkce je definovaná jako:

$$F(x;k,\lambda,\alpha) = \left[1-e^{-(x/\lambda)^k}\right]^{\alpha} \tag{3.13}$$

Speciální případy:

- $\alpha = 1$: Weibullovo rozdělení
- $k = 1$: exponenciálně-exponenciální rozdělení

Exponenciální Weibullovo rozdělení se používá pro odhad času do první poruchy, ve spolehlivosti systémů a v klinických experimentech. [32]

Zobecněné logistické rozdělení

Zobecněné logistické rozdělení, které je definované pro spojitou náhodnou veličinu X, se podle Johnsona [33] dělí na 4 různé systémy hustot. My budeme pracovat s typem I, protože tento typ je v knihovně *Statistical functions* v pythonu. Tento typ je také nazýván jako šikmá logistická distribuce. Funkce hustoty pravděpodobnosti zobecněného logistického rozdělení je:

$$f(x;\alpha) = \frac{\alpha e^{-x}}{(1+e^{-x})^{\alpha+1}}, \tag{3.14}$$

která je definovaná pro celé $\mathbb{R}$, kde $\alpha > 0$ je tvarovací parametr. Její kumulativní distribuční funkce je definovaná jako:

$$F(x;\alpha) = (1+e^{-x})^{-\alpha} \tag{3.15}$$

Zobecněné logistické rozdělení se používá v různých oblastech, jako jsou ekonometrie, těžba nerostných surovin nebo modelování růstu buněk. [34]

Logaritmicko-normální rozdělení

Zkráceně log-normální rozdělení je definováno pro spojitou náhodnou veličinu X, jejíž logaritmus má Gaussovo rozdělení: $X \sim Lognormal$, pak $ln(X) \sim \mathcal{N}$. Funkce hustoty pravděpodobnosti Log-normálního rozdělení je: [35]

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma x} exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right), \tag{3.16}$$

kde $\mu \in \mathbb{R}$ je střední hodnota a $\sigma^2 > 0$ je rozptyl. Její kumulativní distribuční funkce je definovaná jako:

$$F(x;\mu,\sigma^2) = \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right), \tag{3.17}$$

kde Φ je kumulativní distribuční funkce $\mathcal{N}(0,1)$.

Zobecněné Paretovo rozdělení

Zobecněné Paretovo rozdělení je definováno pro spojitou náhodnou veličinu X. Často se používá pro odhad chvostu jiné distribuce. Funkce hustoty pravděpodobnosti zobecněného Paretovo rozdělení je: [36]

$$f(x;\xi) = \begin{cases} (1+\xi x)^{-\frac{\xi+1}{\xi}}, & \xi \neq 0 \\ e^{-x}, & \xi = 0, \end{cases}$$

kde

- $x \geq 0$ pro $\xi \geq 0$.
- $0 \leq x \leq -1/\xi$ pro $\xi < 0$.

Proměnná x se dá nahradit $z = \frac{x-\mu}{\sigma}$, kde $\mu \in \mathbb{R}$ a $\sigma > 0$. Pro tvarovací parametr $\xi < 0$ se hustota pravděpodobnosti s touto transformací nedá zapsat v uzavřeném tvaru, ale je řešením diferenciální rovnice. Kumulativní distribuční funkce pro proměnnou x je definovaná jako:

$$F(x;\xi) = \begin{cases} 1 - (1+\xi x)^{-1\xi}, & \xi \neq 0 \\ 1 - e^{-x}, & \xi = 0, \end{cases}$$

Toto rozdělení se používá například v modelování srážek, v ekonomii a sociálních aktivitách. [37]

3.6 Metoda maximální věrohodnosti

Metoda maximální věrohodnosti (MLE) je ve statistice nejpoužívanější metoda, pomocí které hledáme bodové odhady parametrů. Tato metoda se snaží maximalizovat sdruženou hustotu pravděpodobnosti experimentu naměřených dat vzhledem k parametru, který je obsažen v navrženém rozdělení. Tímto způsobem získáme tzv. věrohodnostní funkci. Zde tedy předpokládáme znalost rozdělení.

Definice 3.6.1 *(Věrohodnostní funkce). Mějme $(X_i)_{i=1}^n$ s odpovídajícím systémem hustot $\mathcal{F}$. Pak definujeme věrohodnostní funkci vztahem*

$$L(\theta) = f(\mathbf{x}, \theta), \quad \forall\theta \in \Theta, \forall \mathbf{x} \in \mathbb{R}^n,$$

kde $\mathbf{x}$ je vektor realizací náhodného výběru $\mathbf{X} = (X_i)_{i=1}^n$. A logaritmickou věrohodnostní funkci definujeme vztahem

$$l(\theta) = \ln L(\theta).$$

Definice 3.6.2 *(Maximálně věrohodný odhad). Definujeme maximálně věrohodný odhad parametrů vztahem*

$$\widehat{\theta}_{ML}(\mathbf{X}) = \arg\sup_{\theta\in\Theta} L(\theta)$$

za předpokladu, že $\widehat{\theta}_{ML}$ je borelovsky měřitelná, jednoznačná a závisí na $\mathbf{X} = (X_i)_{i=1}^n$.

Kapitola 4

Navrhovaná metoda

Na své diplomové práci pracuji ve spolupráci se společností Cisco, která mi poskytla data reálných sítí. Jedno odvětví, kterému se Cisco věnuje, je zabezpečení sítí a komunikace. Vyvíjejí různé druhy detektorů anomálií, které se snaží detekovat negativní vnější vlivy (útoky) na síťových službách, pomocí kterých by se do firem například skrz zaměstnanecké počítače mohli nelegálně nabourat a získat citlivé informace, soukromé údaje lidí, obchodní strategie firem, jejich rozpočty atd. Vedlejším projevem některých typů útoků je zvýšená aktivita na konkrétní síťové službě a na detekci tohoto náhlého zvýšení se zaměřuje moje diplomová práce. Útočník se snaží prolomit heslo, rozšířít virus či stahovat větší množství dat. Budeme se snažit detekovat zvýšenou aktivitu na síti (burst) za nějaký časový úsek. Mluvíme tu o bodové anomálii.

Navržená metoda je založená na pravděpodobnosti a statistice, kde počet komunikací uživatele na dané službě za časový úsek je sledovaná náhodná veličina. První část mé diplomové práce je nalezení/odhadnutí modelu, podle kterého je náhodná veličina rozdělena. Pro nalezení správné distribuční funkce klademe důraz a omezení na to, aby náš detektor byl efektivní a úsporný z hlediska výpočetní a paměťové náročnosti. Detektor bude detekovat anomální komunikace po časových úsecích. Samotná detekce anomálních komunikací a aktualizace parametrů detektoru musí být velice rychlá, úsporná a v zásadě by neměla záviset na počtu/velikosti komunikací v časovém okně.

Model se dále dělí na tzv. uživatelský model a celosíťový model. Toto dělení je čistě technické. Uživatelský model znamená komunikaci konkrétního uživatele a celosíťový model značí komunikaci uživatelů v celé lokální síti. Uživatelský model hledá odchylku v chování toho konkrétního uživatele, oproti jeho dlouhodobému chování. Kdežto celosíťový model hledá uživatele, kteří se svým chováním vymykají od svých kolegů. Pro tyto modely budu dělat samostatné analýzy.

4.1 Odhad distribuční funkce

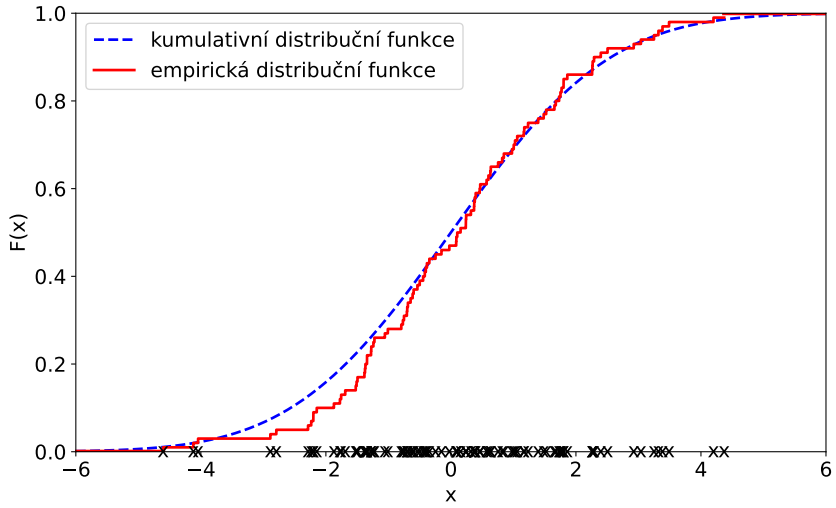
Jak jsem zmínil výše, v prvním kroku mé práce je nalézt správný distribuční model, který bude spolehlivě charakterizovat počty komunikací. Pro nalezení vhodného distribučního modelu jsem měl k dispozici desítky milionů komunikací napříč jednotlivými uživateli, celými sítěmi, službami (HTTP, HTTPS, LDAP, SMB, RDP a SSH) a délkami komunikace (den, týden, měsíc). Data, která mám k dispozici, jsou ve formě histogramu a s tímto formátem dat budu dále pracovat.

Existují dva základní typy odhadů distribučního modelu, parametrický a neparametrický, které si nyní představíme, popíšeme si jejich silné a slabé stránky a z toho pro nás bude plynout, jaký typ odhadu použijeme.

Empirický odhad kumulativní distribuční funkce

Základním odhadem neparametrického rozdělení je empirický odhad distribuční funkce. V pravděpodobnosti je empirická distribuční funkce $\widehat{F}$ spojená s empirickou mírou vzorku. Tato kumulativní distribuční funkce je skoková funkce, jejíž hodnota se v každém pozorování zvětší o $\frac{1}{n}$ viz. obrázek 4.1.

Nevýhoda této metody odhadu správného modelu v našem kontextu je, že tento odhad vůbec nepočítá s chvosty rozdělení. Po posledním pozorování je hodnota distribuční funkce 1, viz. rovnice (3.1) a obrázek 4.1, kde vidíme, že konec červené křivky se rovná hodnotě 1, což ale v důsledku znamená, že pravděpodobnost jakéhokoliv pozorování větší než maximum z našich pozorování je 0, což pro nás nemá smysl. Jak se později dozvíme, naše data mají těžké chvosty. Proto pro naši problematiku musíme volit jiný způsob odhadu distribuční funkce.

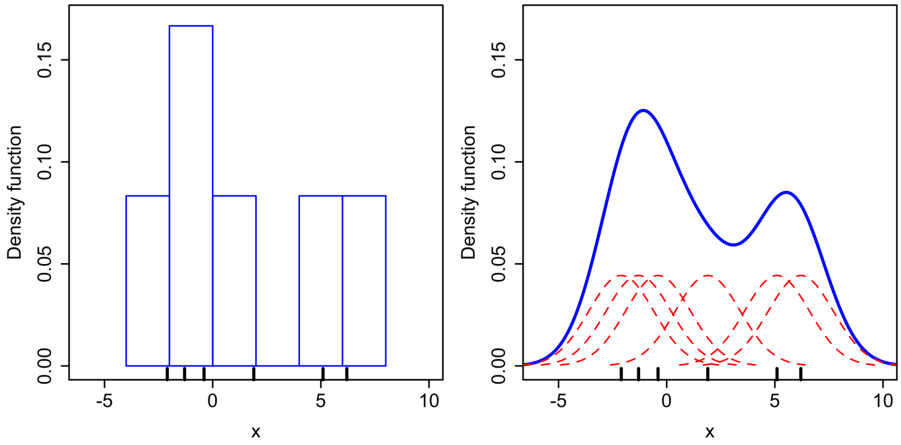


Obrázek 4.1: Červená křivka reprezentuje empirickou distribuční funkci a modrá je skutečná distribuce. Na ose x vidíme naměřená pozorování.

Jádrový odhad hustoty pravděpodobnosti

Dalším způsobem, jak odhadnout hustotu pravděpodobnosti, je jádrový odhad hustoty pravděpodobnosti 3.1.2. Tento odhad na rozdíl od empirického odhadu je hladký a počítá s chvosty dat viz. obrázek 4.2.

Tento způsob má jednu nevýhodu. Podle empirických počátečních dat se odhadne distribuční funkce a na základě ní bude detektor detekovat anomálie mezi aktuálními komunikacemi. To bohužel nepůjde, protože by si detektor musel do paměti ukládat celou distribuční funkci nebo všechna předchozí pozorování. To ale nesplňuje požadavek na paměťovou úspornost, protože detektor bude zpracovávat data po několikaminutových intervalech a komunikací k vyhodnocení bude za každý časový úsek veliké množství. Proto budeme muset zvolit druhou cestu a sice parametrický odhad distribuční funkce.



Obrázek 4.2: Vlevo vidíme histogram dat, vpravo (modrá křivka) jádrový odhad hustoty a jednotlivé komponenty pozorování (červená), převzato z [38].

Detekce systému hustot

Parametrický odhad distribuční funkce na základě dat znamená dvě věci, nalézt správný systém hustot $\mathcal{F}$ a odhadnout parametry tohoto systému. Parametry daného rozdělení se odhadnou pomocí metody maximální věrohodnosti 3.6.2.

Hlavní rozdíl proti neparametrickému odhadu hustoty pravděpodobnosti spočívá v tom, že si detektor pro průběžnou aktualizaci bude ukládat do paměti pouze vektor aktualizovaných parametrů daného rozdělení, tzn. n-tici čísel, která není závislá na počtu komunikací v časovém okně, nýbrž na zvoleném systému hustot.

Pojďme si představit metody, které se používají pro odhad správného systému hustot. Existují grafické metody a metody testování hypotéz.

Grafické metody

Známe veliké množství dat, které má normální rozdělení. Bohužel ne všechna data mají toto rozdělení, nebo jsou intuitivně pochopitelná. Pojďme si představit jednotlivé metody detekce sytému hustot.

Mezi grafické metody patří například probability-probability plot (PP plot) 3.2.2, quantile-quantile plot (QQ plot) 3.2.3, isothermal transformation diagrams (time-temperature-transformation (TTT) diagrams) a další. Představme si PP a QQ ploty, které použijeme.

PP-plot a QQ-plot jsou dva grafické nástroje používané v statistice k ověření shody mezi dvěma soubory dat nebo mezi daty a teoretickým modelem. PP-plot je graf, který vizualizuje podobnost mezi empirickým rozdělením dat a teoretickým kumulativním rozdělením dat. Pokud jsou data získána z teoretického rozdělení, tak by body na grafu měly ležet na diagonále, což naznačuje dokonalou shodu mezi daty a teoretickým modelem. QQ-plot je graf, který zobrazuje shodu mezi kvantily dat a kvantily teoretického rozdělení. Pokud jsou data získána z teoretického rozdělení, tak by body na grafu měly tvořit přibližně přímku, což naznačuje dobrou shodu mezi daty a teoretickým modelem.

Statistické hypotézy

Protože máme k dispozici veliké množství dat a jen těžko bychom prošli každý soubor, aplikovali na něj QQ-plot, vyzuálně vyhodnotili, jaká distribuce nejlépe fituje naše data, musíme použít automatický QQ-plot, který zanalyzuje všechny soubory dat a výsledkem bude skalár, pomocí kterého se vyhodnotí,

jestli daná distribuce věrohodně fituje na naše data, nebo ne. Pojmem **fitovat** budeme v této práci rozumět to, jak distribuce odhaduje zkoumaná data.

Pro tuto analýzu použijeme dvě statistiky, CV statistiku (3.11) a G-statistiku (3.5), které jsou výpočetně úsporné. Pro každý soubor dat spočítáme tyto statistiky v závislosti na systému hustot. Pro lepší představu viz. tabulka 4.1, kde počet řádků je počet souborů dat a každý sloupec je vypočítaná statistika pro každý systém hustot. Nakonec pomocí F-testu (3.10) a t-testu 3.3.2 vyhodnotíme data v tabulce 4.1, jaký systém hustot nejvěrohodněji fituje naše data. Postup práce analýzy souborů dat a následné vyhodnocení vysvětluji na začátku kapitoly 4.2.

Tabulka 4.1: Tabulka G-statistiky pro vybrané systémy hustot. V každém řádku je vypočítaná G-statistika pro konkrétní vzorek dat v závislosti na systému hustot.

$\Lambda_{Weibull}$	$\Lambda_{Expweibull}$	Λ_{Genlog}	$\Lambda_{Lognorm}$	$\Lambda_{Genpareto}$
4.6	8.6	17.2	16.1	17.7
3.6	92.7	53.6	33.2	213.6
996.6	1,046.2	226.9	203.3	261.3
72.9	112.7	296.2	411.0	425.0
5.5	9.2	10.8	280.3	690.9
3.4	39.8	51.9	52.1	1,379.5
53.8	70.3	92.4	334.8	68.6
990.8	967.9	6,400.8	1,504.6	5,788.8
383.4	400.2	385.8	1,367.4	313.8
5.3	3.3	7.7	13.2	19.4
1,216.8	1,655.6	512.7	2,063.9	13,447.1

⋮

4.2 Výsledky odhadu distribuční funkce

Po zadefinování statistických pojmů, metod a testů můžeme přejít na samotnou analýzu dat. Jak už bylo zmíněno výše, dat je k dispozici veliké množství. K tomuto množství dat bude zapotřebí nalézt "automatický QQ-plot". Není z časových důvodů možné, aby se pro každý soubor dat vygeneroval QQ-plot v závislosti na heavy-tailed distribuci 3.5.1 a vizuálně vyhodnocoval, jaké rozdělení nejlépe fituje data. Na druhou stranu nemohu pro jakékoliv množství dat počítat statistiky, které jsem zadefinoval v (3.5), protože je to velice výpočetně náročné. Postup analýzy dat bude následující:

1. Pro 300 náhodně vybraných souborů dat vygeneruji QQ-ploty pro všechny výše zmíněné heavy-tailed distribuce. Vizuálně vyhodnotím, jaké distribuce nejvěrohodněji fitují naše data.

2. Pro nejvhodnější kandidáty se spočítá ze všech souborů dat G-statistika (3.5), CV statistika (3.11) a z dat odhad CV statistiky (3.12).

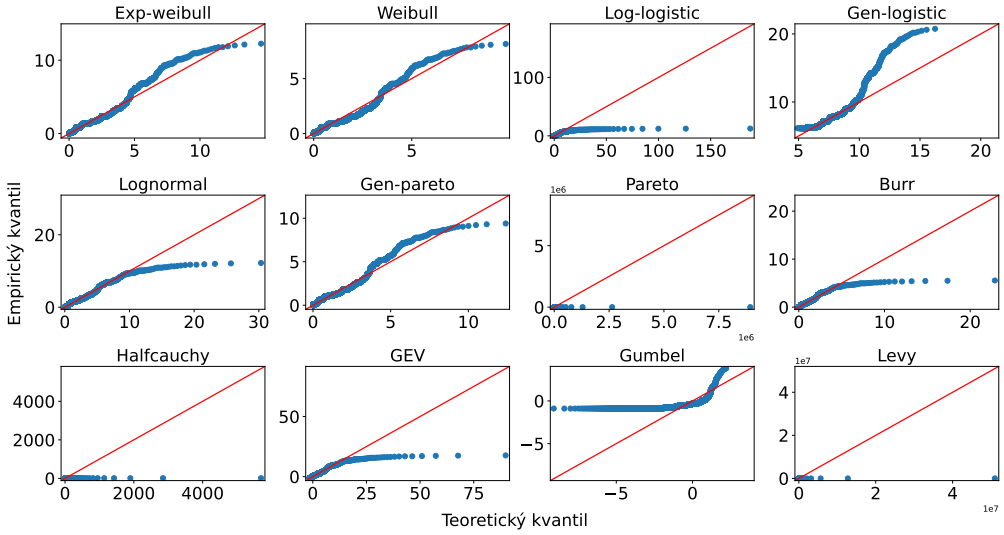
3. Z těchto statistik se provede analýza a podle t-testu a F-testu se dojde k nejvhodnějšímu modelu.

Poznámka 4.2.1 *Další možností bylo místo této G-statistiky, nebo CV statistiky, použít koeficient determinace R^2 z teorie regresní analýzy [39]. Hlavní myšlenka je taková, že si vygeneruji body QQ-plotu a ty se budu snažit proložit regresní přímkou, z toho bych vypočítal koeficient determinace R^2 a tuto statistiku bych dále analyzoval. Nevýhodou této metody je veliká výpočetní náročnost.*

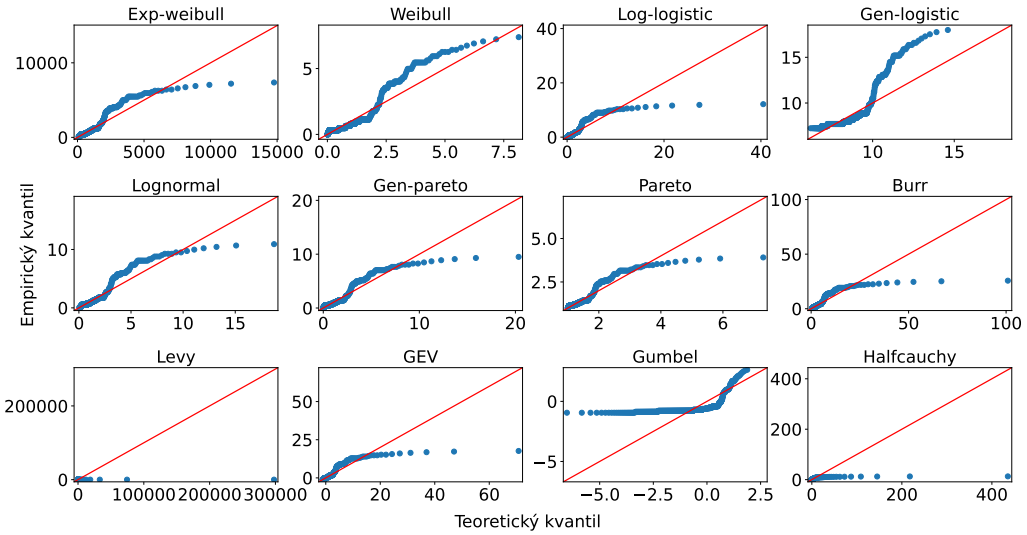
4.2.1 Analýza pomocí QQ-plotu

Analýza dat pro daný systém hustot v grafu QQ-plotu 3.2.3 (osy grafu jsou kvantily empirické distribuce a konkrétní distribuce) probíhá tak, že pokud daná distribuce věrohodně fituje data, jsou data v grafu vyobrazená v linii, která má funkční předpis $y = x$, tedy svírá s osou x úhel 45°. Naopak pokud daná distribuce data nefituje věrohodně, data jsou vyobrazená daleko od této linie, zejména jejich chvosty.

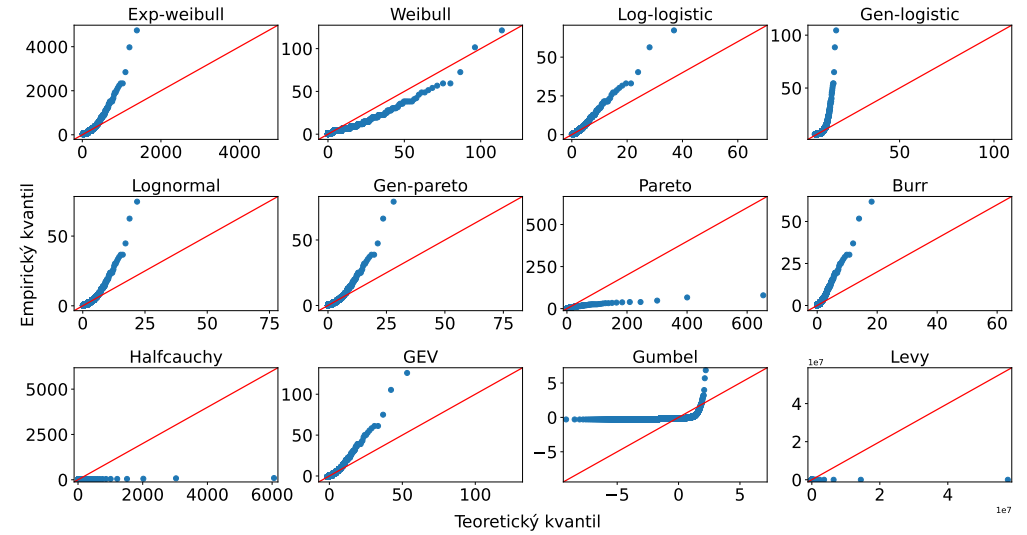
Pro analýzu jsem zvolil Exponenciální-weibullovo, Weibullovo, Log-logistické, Zobecněné logistické, Lognormální, Zobecněné Paretovo, Paretovo, Lévyho, Burrovo, Fisher-Tippettovo, Gumbelovo a Half-cauchyovo rozdělení. Na následujících obrázcích 4.3-4.5 vidíme několik vybraných.



Obrázek 4.3: QQ-plot počtu komunikací uživatelů podle jednotlivých distribucí celosíťového modelu za 24 hodin na síťové službě HTTPS



Obrázek 4.4: QQ-plot počtu komunikací uživatelů podle jednotlivých distribucí celosíťového modelu za 24 hodin na síťové službě SMB



Obrázek 4.5: QQ-plot počtu komunikací uživatelů podle jednotlivých distribucí celosíťového modelu za 24 hodin na síťové službě SSH

V QQ-plotech 4.3-4.4 vidíme, že data nejlépe popisuje Log-normální, Weibullovo, Exp-weibulovo, Gen-logistic a Gen-pareto, kde tyto systémy dobře fitují chvosty dat. U ostatních systémů vidíme, že distribuce špatně fitují převážně chvosty dat. V QQ-plotech na obrázku 4.5 dobře fituje data pouze Weibullovo rozdělení.

Ze všech dat jsou zde ilustrovány pouze 4 vzorky, tedy nemusí být zde jasně vidět, že právě těchto 5 distribucí data popisuje nejlépe. Zbylé systémy, hlavně Burr, GEV, Gumbel, Half-cauchy a Lévy, jak

vidíme, fitují naše data velice špatně. Vidíme, že pro většinu zbylých modelů platí, že mají proti datům lehké chvosty. Tedy pro výšší počty komunikací předpovídají skoro nulovou pravděpodobnost, což je pro náš model zásadní.

4.2.2 Analýza pomocí G-statistiky

Jak už bylo dříve řečeno, tuto statistiku použiji jako skalár, který znamená, jak správně daná distribuce fituje naše data. Kdyby nastala situace taková, že bychom měli napozorovaný jeden soubor dat (počet cestujících v metru každý den nebo objem průtoku řeky v daném místě atd.), tak na tato data implementuji PP/QQ-plot podle různých distribucí a vyhodnotím v jednom kroku, jaká distribuce nejlépe fituje data. Zde máme daleko zajímavější úlohu a sice více souborů dat, tudíž máme k dispozici daleko více informací o těchto datech, které se budeme snažit pomocí vhodných statistik a následného vyhodnocení získat.

V minulé kapitole nám vyšly kandidáti, kteří fungují dobře a ty budu používat pro další analýzu: Weibullovo, Exponenciální-weibullovo, zobecněné logistické, lognormální a zobecněné Paretovo rozdělení. Postup bude následující, vypočítám G-statistiku pro každý výběr dat a každou distribuci. Budu mít několik stovek tisíc skalárů a pomocí boxplotů, deskriptivní numerické statistiky a t-testu budu vyhodnocovat, jaký model bude nejvhodnější pro naše data.

K výpočtu jsem mezi ostatními statistikami zvolil G-statistiku (3.5), protože ostatní dvě (Pearsonova a Neymanova) mají ve jmenovateli buď naměřenou četnost, nebo očekávanou četnost. Kde tyto veličiny jsou v mnoha případech velice nízké a jenom kvůli jedné množině by mohl být výpočet statistiky nepřesný a právě tyto vysoké hodnoty se v logaritmu tolik nepromítnou.

Jak vidíme v kapitole 3.3.1, je třeba k výpočtu G-statistiky vhodně rozdělit obor hodnot naší náhodné veličiny. Pro ideální počet intervalů jsem zvolil Sturgesův vzorec [40], kde je počet definován $k = 1 + \lceil \log_2 n \rceil$, kde n je počet pozorování, podmínka: $n > 30$. Záměrně jsem použil tento vzorec z důvodu jednoduchosti na výpočet. Přesnější počet intervalů by nám dal Doaneův vzorec [41], který zahrnuje i šikmost, což je 3. centrální moment daného rozdělení, ale výpočet by byl pomalejší a nemá tak velký vliv na výsledek.

Deskriptivní numerická statistika

V následující tabulce vidíme základní numerické statistiky naší pozorované G-statistiky podle jednotlivých distribucí.

distribuce	průměr	směr. odchylka	medián	min.	max.	variační koef.
celosíťový model						
Weibull	23030	64467	765	0	384444	2,8
Exp-weibull	22118	88425	1008	2	919083	4,0
Gen-logistic	49654	172001	1195	4	1835005	3,5
Log-norm	67216	359693	1747	7	3961034	5,4
Gen-pareto	140833	510887	2965	11	5595623	3,6
uživatelský model						
Weibull	53	114	21	0	2743	2,2
Exp-weibull	64	108	34	0	3510	1,7
Gen-logistic	59	65	39	0	1175	1,1
Log-norm	234	457	100	0	14464	2,0
Gen-pareto	688	1305	146	0	30853	1,9

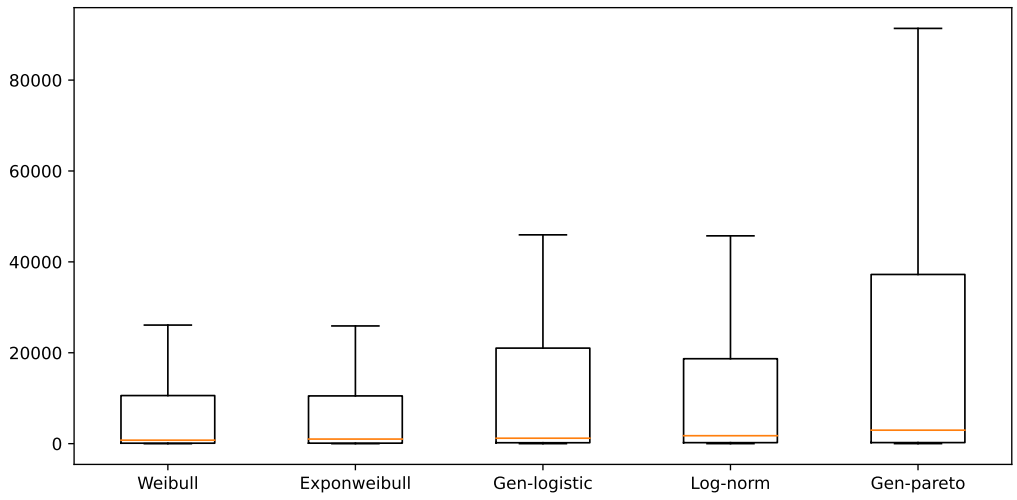
Tabulka 4.2: Tabulka deskriptivních statistik pro G-statistiku.

Ze statistik celosíťového modelu můžeme vidět, že Log-normální rozdělení mají o 2 krát větší průměr, medián a směrodatnou odchylku a Gen-pareto o řád než Weibullovo rozdělení. Z toho plyne, že většina dat není rozdělená podle Log-normální a Gen-pareto rozdělení.

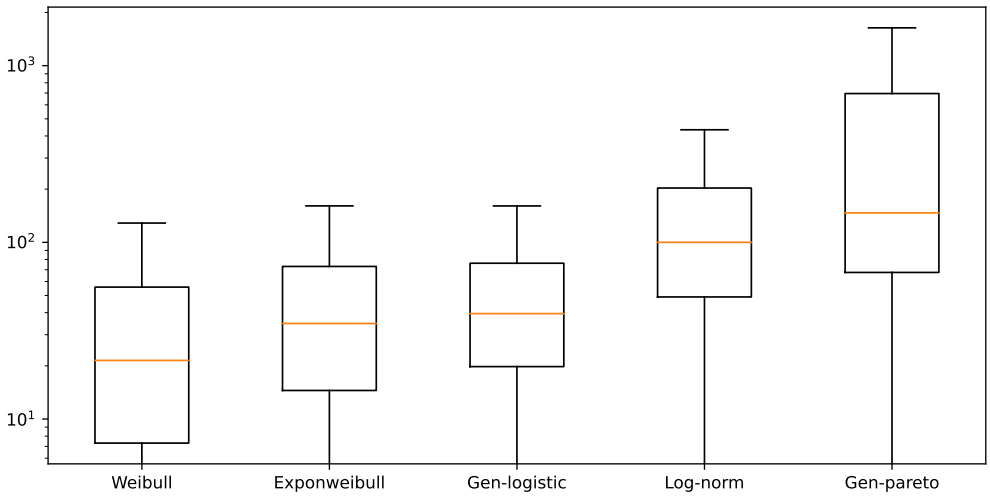
Pro data uživatelského modelu opět vidíme velký průměr, medián a směrodatnou odchylku pro rozdělení Log-normální a Gen-pareto, z toho opět vyplývá, že většina těchto dat takto rozdělená není.

Grafická analýza

Pro grafickou analýzu zkoumané G-statistiky jsem použil boxploty.



Obrázek 4.6: Boxploty G-statistiky vypočítané podle jednotlivých distribucí dat získaných z celosíťového modelu



Obrázek 4.7: Boxploty G-statistiky vypočítané podle jednotlivých distribucí dat získaných z uživatelského modelu

Z G-statistiky dat celosíťového modelu 4.6 vidíme, že Gen-pareto rozdělení má velké interkvartilové rozpětí a velice težké chvosty, tudíž není robustní na část pozorování. Naopak boxploty pro Weibullovo a Exponenciální-weibullovo rozdělení pro většinu dat si udržují nízké hodnoty. Medián dat se až na Log-normální a Gen-pareto tolik neliší.

Naopak rozdělení G-statistiky dat uživatelského modelu 4.7 se pro jednotlivé distribuce liší. Medián je nejmenší pro Weibullovo rozdělení a nejvyšší je pro Gen-pareto rozdělení. Weibullovo rozdělení má sice veliký variační koeficient, ale ten je způsoben nízkým průměrem statistiky. Nezapomeňme si na obrázku 4.6 všimnout logaritmického měřítka na ose y.

Test homogenity středních hodnot

Tento test slouží k tomu, aby rozhodl, zda se střední hodnoty dvou stejně rozdělených výběrů liší, nebo se rovnají na hladině významnosti α. Tento test nám rozhodne, zda se liší střední hodnota souboru G-statistik jednotlivých rozdělení, nebo ne. Jinými slovy t-test rozhodne o tom, jestli analyzované distribuce fitují data stejně kvalitně, nebo je mezi nimi významný rozdíl.

Pro tento test použijeme t-test, který jsme si definovali v 3.3.2. Pro aplikaci tohoto testu musíme nejdříve ověřit předpoklady, které jsou: normalita dat, nezávislost dat, znalost rozptylů výběrů dat (rovnají, nerovnají se).

První předpoklad je splněn podle CLT (3.6). Nezávislost dat je předpokladem celé práce. Třetí předpoklad/znalost toho, jestli jsou rozptyly obou výběrů stejné nebo ne, otestuji pomocí F-testu. V Pythonu jsem pro tento test použil Levene's test. Výsledek je v následující tabulce 4.3, kde srovnávám G-statistiky vypočítané podle jednotlivých rozdělení.

distribuce	celosíťový model	uživatelský model
Weibull, Exp-weibull	0,887	$\sim 10^{-16}$
Weibull, Gen-logistic	0,021	
Weibull, Log-norm	0,053	
Weibull, Gen-pareto	$\sim 10^{-17}$	
Exp-weibull, Gen-logistic	0,023	
Exp-weibull, Log-norm	0,051	
Exp-weibull, Gen-pareto	$\sim 10^{-20}$	
Gen-logistic, Log-norm	0,48	
Gen-logistic, Gen-pareto	0,007	
Log-norm, Gen-pareto	0,059	

Tabulka 4.3: Tabulka p-hodnot Leveneho testu, který testuje shodnost rozptylů dvou souborů dat, které jsou v prvním sloupci. Pokud je p-hodnota menší než 0,05, zamítáme hypotézu o shodnosti rozptylů.

V tabulce 4.3 vidíme, že žádná data (G-statistiky) uživatelského modelu nemají stejný rozptyl. Stejný rozptyl G-statistik dat z celosíťového modelu mají výběry z rozdělení Weibull vs. Exp-weibull, Weibull vs. Log-norm, Exp-weibull vs. Log-norm a Log-norm vs. Gen-pareto. Pro tyto dvojice nezamítáme hypotézu H_0 na hladině významnosti α. Proto tyto dvojice budeme testovat pomocí t-testu, konkrétně volíme variantu (3.8). Zbylé dvojice budeme testovat podle varianty (3.9). Výsledky vidíme v následující tabulce 4.4.

distribuce	celosíťový model	uživatelský model
Weibull, Exp-weibull	0,894	$\sim 10^{-20}$
Weibull, Gen-logistic	0,021	
Weibull, Log-norm	0,053	
Weibull, Gen-pareto	$\sim 10^{-20}$	
Exp-weibull, Gen-logistic	0,023	
Exp-weibull, Log-norm	0,051	
Exp-weibull, Gen-pareto	$\sim 10^{-20}$	
Gen-logistic, Log-norm	0,48	
Gen-logistic, Gen-pareto	0,007	
Log-norm, Gen-pareto	0,059	

Tabulka 4.4: Tabulka p-hodnot testu homogenity středních hodnot

Z této tabulky můžeme vyčíst, že střední hodnoty mají pro celosíťový model shodné Weibullovo a Exp-weibullovo rozdělení, dále Weibullovo a Log-normální, Exp-weibull a Log-normální a Log-normální a Gen-pareto na hladině α. Pro data uživatelského modelu se střední hodnoty všech distribucí na hladině α liší.

Nejlépe počty komunikací na síťových službách, podle t-testu a numerické statistiky, popisuje Weibullovo rozdělení pro celosíťový model. V uživatelském modelu nejlépe data popisuje Weibullovo a Exp-weibullovo rozdělení.

4.2.3 Analýza pomocí variačního koeficientu

Deskriptivní numerická statistika

V následující tabulce vidíme základní numerické statistiky pozorovaného koeficientu variace podle jednotlivých distribucí.

distribuce	průměr	směr. odchylka	medián	min.	max.	variační koef.
celosíťový model						
Empirical	1,2	0,46	1,1	0,3	2,6	0,38
Weibull	3,1	3,7	1,8	0,0	19	1,2
Exp-weibull	12	29	2,3	0,7	150	2,4
Gen-logistic	0,2	0	0,16	0,15	0,19	0
Log-norm	10^{34}	10^{35}	1000	0,3	10^{36}	15
uživatelský model						
Empirical	0,51	0,14	0,50	0,10	0,82	0,27
Weibull	7,0	19	1,5	0	160	2,7
Exp-weibull	1,8	2,0	1,3	0	16	1,1
Gen-logistic	0,2	0,14	0,16	0,14	1,17	0,7
Log-norm	10^{31}	10^{33}	1,7	0	10^{34}	14

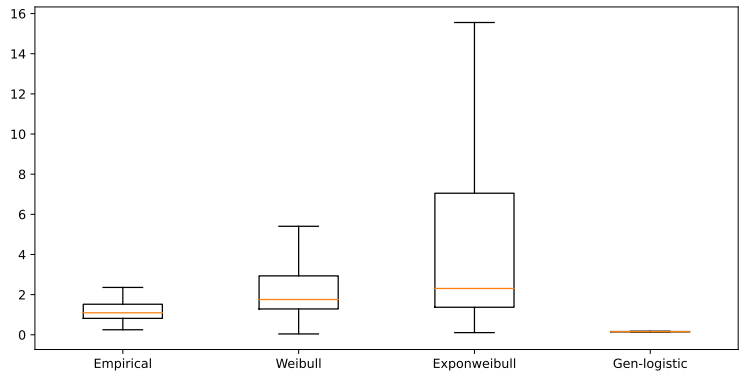
Tabulka 4.5: Tabulka deskriptivních statistik pro variační koeficient.

Ze statistik v tabulce 4.5 vidíme, že Log-normální rozdělení má pro oba typy dat obrovskou směrodatnou odchylku, z toho jasně plyne, že pro malou část dat je použitelné, ovšem pro většinu dat toto rozdělení není použitelné.

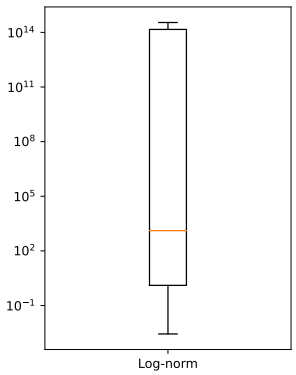
Poznámka 4.2.2 *V této části jsem neanalyzoval CV statistiku pro Gen-pareto distribuci, protože pro velkou část našich souborů dat výsledek neexistoval. Je to z toho důvodu, že pro konkrétní parametry rozdělení neexistuje druhý centrální moment-rozptyl.*

Grafická analýza

Pro grafickou analýzu zkoumaného variačního koeficientu opět použiji boxploty.

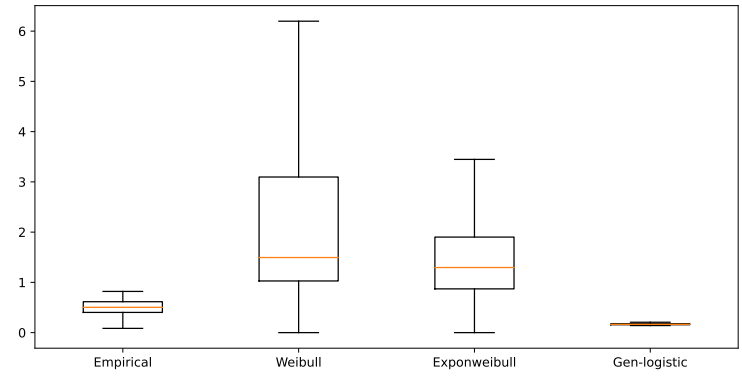


(a) Boxploty empirického variačního koeficientu a variačního koeficientu pro Weibullovo, Exp-weibullovo a Gen-logistické rozdělení.

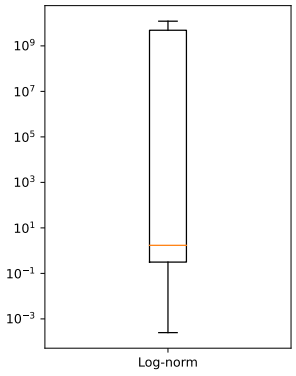


(b) Boxplot variačního koeficientu pro Log-normální rozdělení.

Obrázek 4.8: Boxploty variačního koeficientu vypočítaného podle jednotlivých distribucí dat získaných z celosíťového modelu.



(a) Boxploty empirického variačního koeficientu a variačního koeficientu pro Weibullovo, Exp-weibullovo a Gen-logistické rozdělení.



(b) Boxplot variačního koeficientu pro Log-normální rozdělení.

Obrázek 4.9: Boxploty variačního koeficientu vypočítaného podle jednotlivých distribucí dat získaných z uživatelského modelu.

Z boxplotů 4.8 vyplývá, že CV spočítaný pro Log-normální rozdělení se velice liší od empirical CV, má těžké chvosty. Pokud srovnáme empirical CV s ostatními, vidíme, že se mediány tolik neliší. U Exp-weibullova rozdělení zaznamenáváme relativně velké interkvartilové rozpětí, naopak Gen-logistic rozdělení má velice malé interkvartilové rozpětí kvůli nízké směrodatné odchylce. Ten samý závěr můžeme říci o boxplotech 4.9 až na interkvartilové rozpětí dat Weibullova a Exp-weibullova rozdělení, ta jsou v tomto případě opačná, Weibullovo rozdělení má větší interkvartilové rozpětí než Exp-weibullovo rozdělení.

Test homogenity středních hodnot

Opět před tím, než použijeme t-test pro rozhodnutí shodnosti středních hodnot, otestujeme rozptyl našich dat viz. tabulka 4.6

distribuce	celosíťový model	uživatelský model
empirical, Weibull	0,068	0,157
empirical, Exp-weibull	10^{-5}	10^{-18}
empirical, Gen-logistic	10^{-17}	10^{-11}
empirical, Log-norm	0,312	10^{-10}
Weibull, Exp-weibull	0,893	0,122
Weibull, Gen-logistic	0,062	0,162
Weibull, Log-norm	0,312	10^{-10}
Exp-weibull, Gen-logistik	10^{-5}	10^{-62}
Exp-weibull, Log-norm	0,331	10^{-10}
Gen-logistic, Log-norm	0,348	10^{-10}

Tabulka 4.6: Tabulka p-hodnot Leveneho testu

Důležité je pro nás srovnání s empirical CV. V tabulce 4.6 vidíme, že pro celosíťový model mají dvě rozdělení shodné rozptyly a ostatní ne. Pro uživatelský model je shodná jen jedna distribuce. Ostatní kombinace zamítaly hypotézu H_0 o shodnosti rozptylů. Podle této tabulky opět použijeme t-test.

distribuce	celosíťový model	uživatelský model
empirical, Weibull	0,063	0,157
empirical, Exp-weibull	10^{-20}	10^{-18}
empirical, Gen-logistic	10^{-17}	0,049
empirical, Log-norm	0,313	10^{-15}
Weibull, Exp-weibull	0,890	0,157
Weibull, Gen-logistic	0,051	0,163
Weibull, Log-norm	0,313	10^{-10}
Exp-weibull, Gen-logistik	10^{-10}	10^{-15}
Exp-weibull, Log-norm	0,351	10^{-12}
Gen-logistic, Log-norm	0,311	10^{-18}

Tabulka 4.7: Tabulka p-hodnot t-testu pro celosíťový model a uživatelský model

V tabulce 4.7 nezamítáme hypotézu o rovnosti středních hodnot na hladině α empirical CV s Weibullovým a Log-normálním rozdělením. Stejnou střední hodnotu na hladině α mají i tyto dvě rozdělení. Zde bude rozhodovat i grafická analýza.

Pro uživatelský model má s empirical CV stejnou střední hodnotu opět Weibullovo a Gen-logistic. I zde mají tyto dvě rozdělení stejné střední hodnoty na hladině α, tudíž i zde se rozhodne podle grafické a numerické analýzy.

4.2.4 Závěr k odhadu distribuční funkce

Závěrem k vybraným statistickým metodám je, že variační koeficient neurčí jednoznačně, jaký systém hustot nejlépe odhaduje data. Dále je z definice podmínka na sledovanou veličinu, aby $X \in \mathcal{L}_2$,

jinými slovy, aby existoval druhý centrální moment, což pro část vzorků dat při zobecněném Paretově rozdělení neplatilo. S kombinací deskriptivní numerické statistiky se pro kvalitní závěr použít dá. G-statistika, která se používají v testu dobré shody, je velice efektivní a jednoznačně určí správný systém hustot. Navíc její výpočet probíhá rychle. Pro nalezení správného systému hustot, by jistě šly použít i další statistiky, například koeficient determinace. Tato statistika je velice efektivní a byla by nejpřesnější, ale tato metoda je pro naše účely nepoužitelná z důvodu velké výpočetní náročnosti.

Závěrem výsledků je, že pozorovaná data, jak celosíťového modelu, tak uživatelského modelu nejlépe popisuje Weibullovo rozdělení. Z analýz G-statistiky 4.2.2 a variačního koeficientu 4.2.3 jasně vyplývá, že toto rozdělení nejlépe fituje data.

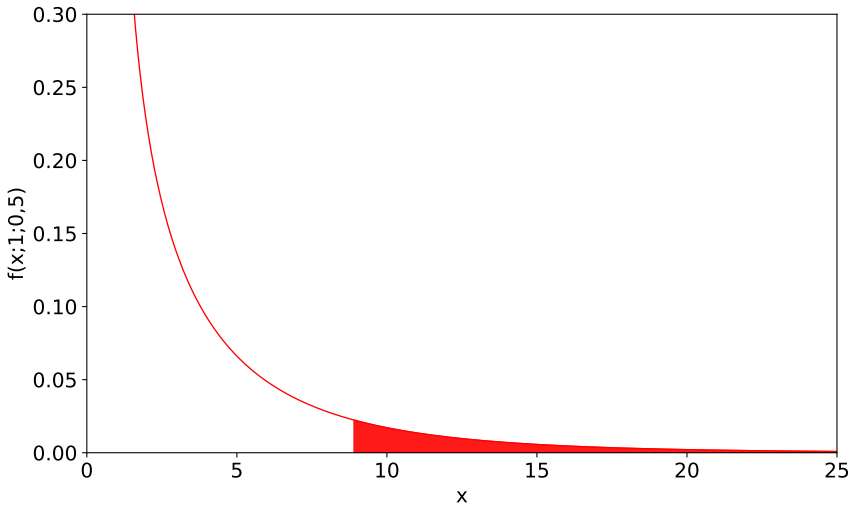
4.3 Kvantilové pravidlo

Jak bylo už dříve řečeno, navržený detektor funguje na klasických statistických metodách, tedy na základě správného modelu, který věrohodně fituje naše data. Navržený detektor poběží v reálném čase, to znamená, že musí být dostatečně úsporný a efektivní. Soubory dat, které tento detektor bude vyhodnocovat, budou přicházet po krátkých intervalech. Interval, po kterém dojde k detekci, se nazývá časové okno.

V tuto chvíli známe náš statistický model, podle kterého jsou počty komunikací rozděleny. A nyní přichází otázka, podle jakého pravidla se bude algoritmus rozhodovat, jestli daný počet komunikací je anomální, či není.

Za anomálii bude označen takový počet spojení, které nebude spadat do $0,95$-kvantilu Weibullova rozdělení. Tedy práh se bude rovnat $0,95$-kvantilu Weibullova rozdělení. Když počet komunikací bude menší než hodnota prahu, nebude označeno za anomální, když počet komunikací bude větší než hodnota prahu, bude označeno za anomální.

Jako příklad si uveďme nastavení aktuálních parametrů na hodnoty $(1; 0,5)$ (vypočítaný $0,95$-kvantil modelu má hodnotu 8,9), tak uživatel, který komunikoval v aktuálním časovém okně 9 krát, bude označen za anomálního, viz. obrázek 4.10.



Obrázek 4.10: Červněná křivka značí hustotu pravděpodobnosti Weibullova rozdělení a červená oblast značí anomální počet komunikací.

Kapitola 5

Experimenty

5.1 Úvod

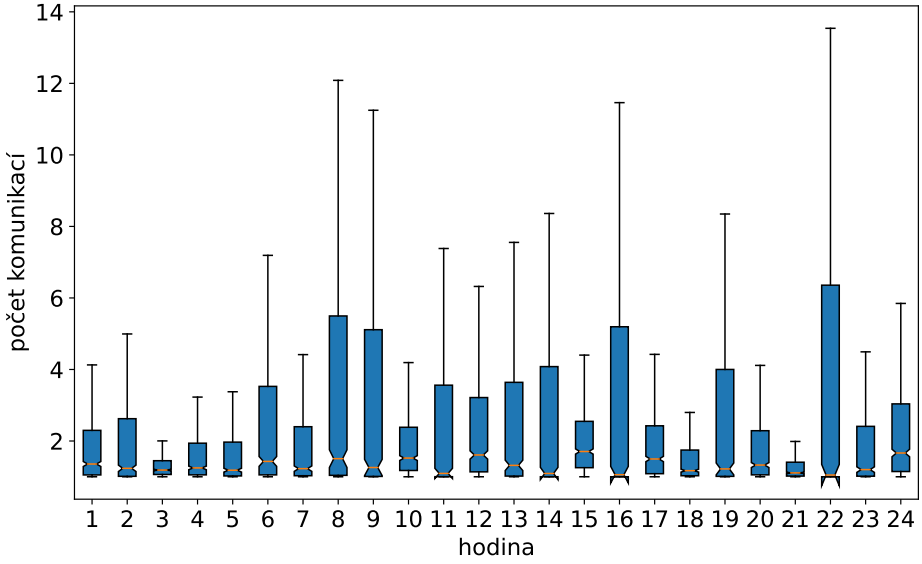
Nyní máme vybraný nejlepší parametrický model a zadefinované pravidlo, podle kterého bude algoritmus klasifikovat, jaký počet komunikací je anomální, jaký ne. V této kapitole důkladně popíšu postup druhé části práce. Jak již bylo zmíněno v úvodu, chování uživatelů na síti je dynamické a mění svůj charakter, proto je žádoucí do modelu přidat vnější parametry. Po jejich správném nastavení bude detektor lépe reflektovat aktuální aktivitu uživatelů na síti, a tím přesněji detekovat skutečné anomální počty komunikací. Navržené vnější parametry jsou:

- Způsob aktualizace vnitřních parametrů
- Warmup perioda: Jak dlouho bude detektor sbírat data, ze kterých odvodí iniciální nastavení parametrů.
- Forget perioda: Jak dlouho od přítomnosti do minulosti si bude model do paměti ukládat vnitřní parametry.

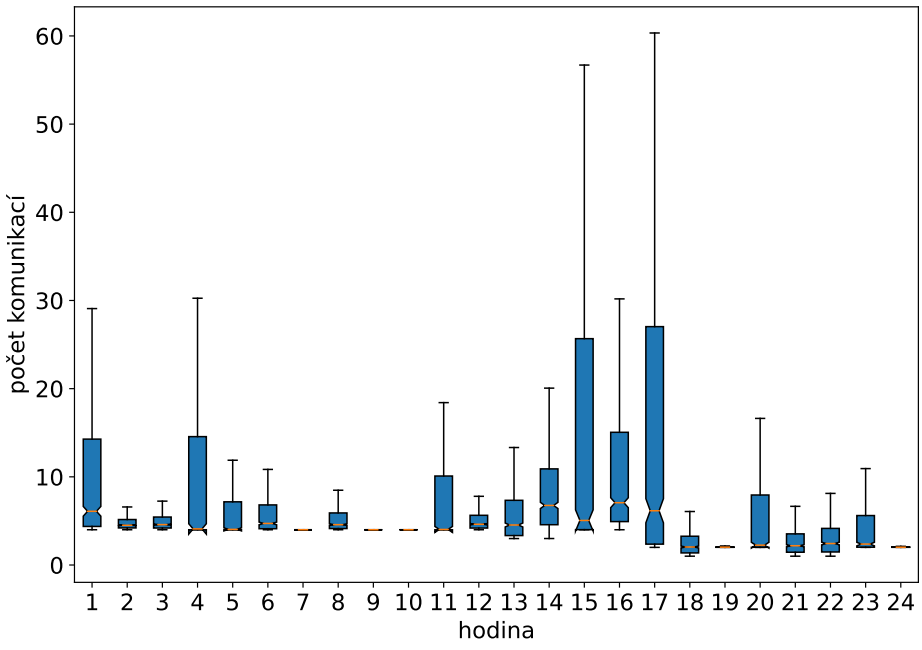
Dynamické chování uživatelů vidíme na obrázcích 5.1 pro konkrétní den a službu. Každý boxplot vyjadřuje rozdělení počtu komunikací za hodinu. U služby Lightweight Directory Access Protocol (LDAP) vidíme zvýšenou aktivitu mezi 8.-16. hodinou a u Samby (SMB) je zesílená komunikace mezi 15.-17. hodinou. Je žádoucí, aby větší aktivitu většiny uživatelů model reflektoval a pokládal toto chování za neanomální.

V následujících sekcích navrhnu a detailněji vysvětlím jednotlivé vnější parametry, budu diskutovat jejich motivaci, provedu experimenty a na jejich základě určím, jak mají být nastavené. Experimenty dělám pro časové období jednoho týdne, abych zachytil všechny změny chování uživatelů sítě. Experimenty provádím na detekčním systému (online režim). Základní parametry detekčního systému jsou:

- K dispozici máme data pro celosíťový model.
- Data jsou analyzována pro konkrétní den, službu a síť.
- Algoritmus vyhodnocuje data po 10 minutovém intervalu (časové okno).
- Při každém vyhodnocení dojde k aktualizaci parametrů.



(a) LDAP 3.7.2022



(b) SMB 5.7.2022

Obrázek 5.1: Boxploty značí rozdělení počtu komunikací pro každou hodinu na daných službách a dnech.

5.2 Způsob aktualizace vnitřních parametrů

Tento vnější parametr je klíčový pro přesnou klasifikaci detektoru. K aktualizaci vnitřních parametrů dochází pomocí váženého průměru z předchozích vnitřních parametrů, které má detektor uložené v paměti. Formální definice:

$$\theta_{new} = \frac{\sum_{k=1}^{n} \omega(k) \cdot \theta_k}{\sum_{k=1}^{n} \omega(k)}, \tag{5.1}$$

kde n je počet časových oken (například pro 4 hodinovou forget periodu a 10 minutové časové okno, bude $n = 4 * 6 = 24$), θ_k je k-tý parametr v pořadí od přítomnosti do minulosti a $\omega(k)$ je hodnota váhové funkce pro k-té časové okno. Z této definice plynou otázky:

- kdy aktualizovat vnitřní parametry modelu, zda před vyhodnocením, nebo po vyhodnocení nových dat.
- zda aktualizace závisí na počtu časových oken ve forget periodě, nebo na počtu dat v každém časovém okně.
- zda datům ve forget periodě přiřazovat váhu podle času vyhodnocení, tedy čím starší data, tím menší význam.
- zda bude aktualizace probíhat na základě všech dat v aktuálním časovém okně, nebo pouze těch, které by detektor nevyhodnotil jako anomální.

5.2.1 Moment aktualizace

V této kapitole zodpovím otázku, kdy aktualizovat parametry modelu, zda před, nebo po vyhodnocení aktuálních dat. Detektor, který vyhodnocuje aktuální data pouze na základě předchozích dat, nebude dostatečně reflektovat náhlé skokové nárůsty aktivity uživatelů. Bude o jedno časové okno zpožděný za detektorem, který detekuje anomálie až po aktualizaci parametrů. Důsledkem bude detekování vyššího množství anomálních komunikací, kde část z nich bude falešně pozitivní a přehlédnutí nižšího počtu falešně negativních anomálních komunikací. Detektor, který vyhodnocuje data až po aktualizaci parametrů, bude dostatečně ve větší míře reflektovat aktuální data. Důsledkem bude nižší množství anomálních komunikací. Jinými slovy bude větší množství falešně negativních anomálních komunikací a menší množství falešně pozitivních anomálních komunikací. Následným experimentem dostaneme odpověď na otázku, která možnost bude pro přesnější detekci skutečných anomálních komunikací přesnější.

Pro ilustraci pustím detektor napříč službami, který bude detekovat anomálie po aktualizaci vnitřních parametrů. Detekované anomální komunikace srovnám s detekovanými anomálními komunikacemi z detektoru, který bude detekovat anomálie před aktualizací vnitřních parametrů. Výsledkem bude vykreslení anomálních komunikací, které byly detekovány pouze jedním z detektorů. Ostatní vnější parametry jsou nastaveny:

- Forget a warmup periody jsou nastaveny na 4 hodiny.
- V aktualizaci vnitřních parametrů je zohledněn počet dat v časovém okně.
- V modelu je použita lineární váhová funkce. 5.4

Na obrázcích 5.2 vidíme týdenní aktivitu uživatelů na službách HTTP a SSH. Žlutá křivka reprezentuje $0,95$ kvantil počtu komunikací za každou hodinu. Červená křivka značí práh obou detektorů. Černě jsou vyobrazené ty anomální komunikace, které detektor, který aktualizuje vnitřní parametry před vyhodnocením, nedetekoval a detektor, který aktualizuje vnitřní parametry po vyhodnocení, detekoval. Zejména v obrázku 5.2a je vidět, že detekované anomální komunikace se nachází před, nebo při náhlém zvýšení aktivity uživatelů. Tyto anomálie, jak bylo řečeno, skutečně anomálie nejsou.

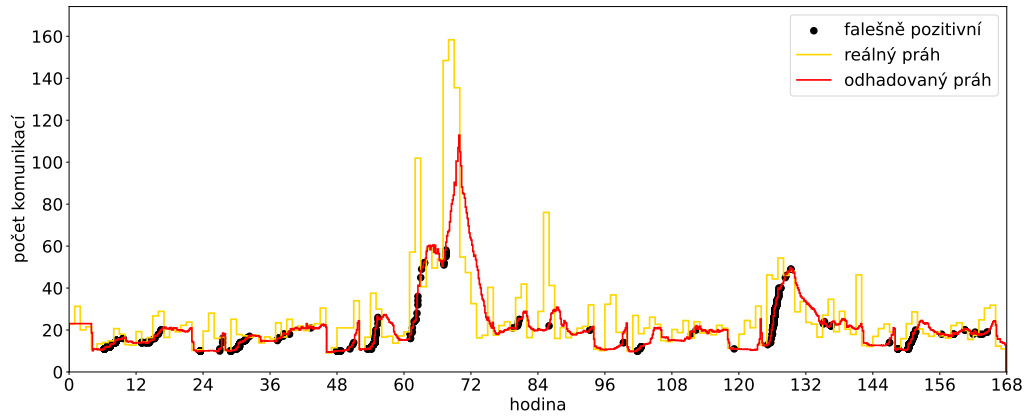
Detektor, který detekuje anomálie po aktualizaci vnitřních parametrů, má jednu výhodu. Neplatí hypotéza, že by neodhalil více skutečných anomálií na úkor nižší detekce falešně pozitivních anomálií. Detekuje méně falešně pozitivních anomálií a přitom neodhaluje žádné skutečné anomálie navíc.

Pro detektor, který klasifikuje počty komunikací před aktualizací vnitřních parametrů, vyplývá z obrázků 5.2 jedna nevýhoda. Neadaptuje se včas na náhlé zvýšení aktivity uživatelů, která je opodstatněná viz 8.-16. hodina na obrázku 5.1a, tedy bude detekovat více falešných anomálních komunikací.

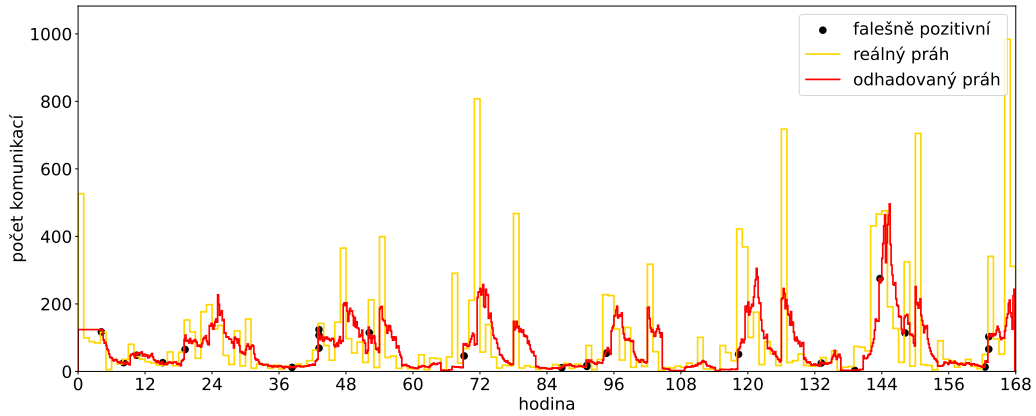
V tabulce 5.1 vidíme počet falešně pozitivních anomálií detekovaných detektorem, který detekuje anomálie před aktualizací vnitřních parametrů. Přesnost detekce skutečných anomálních komunikací detektoru, který vyhodnocuje data po aktualizaci vnitřních parametrů, se průměrně zvětšila o zhruba 15%. Z těchto důvodů bude detektor aktualizovat parametry před vyhodnocením aktuálních dat.

V důsledku takového nastavení momentu aktualizace dojde ke změně ve formální definici aktualizace vnitřních parametrů 5.1: $k \in \{0, \ldots, n-1\}$, kde $k = 0$ je hodnota proměnné pro aktuální časové okno:

$$\theta_{new} = \frac{\sum_{k=0}^{n-1} \omega(k) \cdot \theta_k}{\sum_{k=0}^{n-1} \omega(k)}, \tag{5.2}$$



(a) HTTP



(b) SSH

Obrázek 5.2: Reálná aktivita uživatelů (žlutá linie) a průběh prahu modelu (červená linie) s vyobrazenými falešně pozitivními anomálními komunikacemi (černé body) zachyceny detektorem, který detekuje anomálie před aktualizací vnitřních parametrů, 1.-7.3.2023.

Počet falešně pozitivních anomálií		
služby	relativní	absolutní
HTTP	11,8 %	1022
SSH	18,4 %	23

Tabulka 5.1: Tabulka počtu falešně pozitivních anomálií detekovaných detektorem, který detekuje anomálie, před aktualizací vnitřních parametrů ve dnech 1.-7.3.2023.

5.2.2 Počet uživatelů v časovém okně

Další vnější parametr, který je možno použít, je vážení vnitřních parametrů podle počtu uživatelů, kteří byli aktivní v rámci časového okna. Zohledňování počtu dat v časovém okně při aktualizaci vnitřních parametrů má významný efekt zejména u méně používaných služeb nebo sítí. Přidáním tohoto

parametru do aktualizace se předejde náhlým změnám hodnot prahu v modelu na základě neopodstatněné změny aktivity uživatelů, čímž je myšlena velká komunikace malého počtu uživatelů na rozdíl od uplynulých časových oken. Průběh prahu bude vyvážený vzhledem k počtu uživatelů. V případě více používaných služeb či sítí tento parametr nehraje významnou roli. Tuto motivaci uvedu konkrétněji. Bez zohlednění počtu uživatelů pro málo používanou službu bude mít vnitřní parametr, který byl odhadnutý na základě 5 pozorování, stejnou váhu jako parametr, který byl v dalším časovém okně odhadnut na základě 30 pozorování. U více používaných služeb rozdíl mezi zohledněním a bez zohledněním počtu uživatelů nebude tak velký, protože mezi 250 a 300 uživateli je daleko menší relativní rozdíl (0,83) než mezi 5 a 50 uživateli (0,1).

Experiment realizuji pro méně používanou službu SSH a pro častěji používanou službu HTTP. Detekce probíhá na týdenních datech s detektorem, který nezohledňuje a který zohledňuje počet uživatelů. Výsledkem bude vykreslení $0,95$ kvantilu počtu komunikací za hodinu a průběh prahů obou detektorů. Další vnější parametry detektoru jsou nastaveny:

- Forget a warmup periody jsou nastaveny na 4 hodiny.
- K aktualizaci vnitřních parametrů dochází před detekcí.
- Bez použití váhové funkce.

Na obrázcích 5.3 5.4 vidíme průběhy prahů detektorů pro služby SSH a HTTP a na obrázcích 5.3b vidíme zvětšené vybrané části grafu 5.3a mezi 3.-7. a 16.-20. hodinou. Nad hodnotou prahu jsou čísla, která značí počet uživatelů v časovém okně.

Efekt zohlednění počtu uživatelů v časovém okně můžeme vidět v levém obrázku 5.3b, kde v 6. hodině vzrostla hodnota prahu ze 100 na 250. Hodnota prahu detektoru, který zohledňuje počet uživatelů v časovém okně, je nižší, protože 3 uživatelé v prvním časovém okně jsou méně než průměr počtu uživatelů za poslední 4 hodiny. Naopak v pravém obrázku 5.3b se hodnota prahu detektoru, který zohledňuje počet uživatelů v časovém okně na začátku 19 hodiny, zvýšila více, protože počet uživatelů v tomto časovém okně je nadprůměrný oproti počtu v předchozích časových oknech.

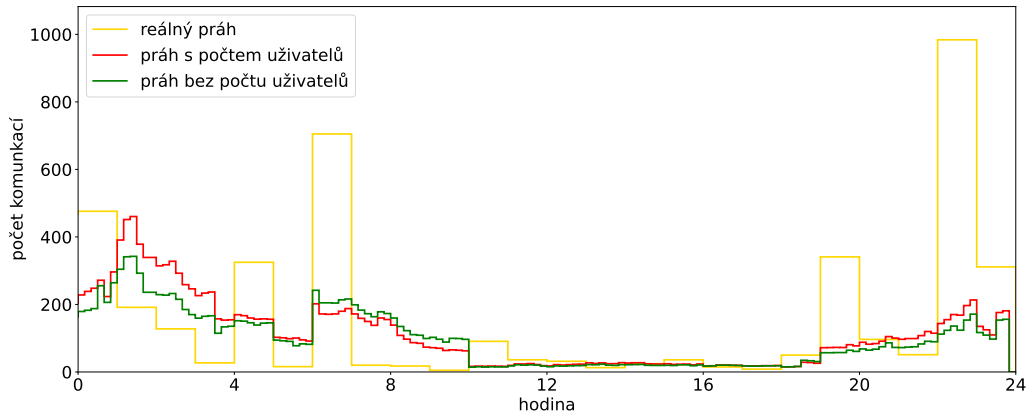
Detektor, který zohledňuje počet uživatelů v časovém okně, přesněji detekuje skutečné anomálie ze dvou důvodů. Pokud se náhle zvýší aktivita uživatelů a je opodstatněná, jinými slovy počet uživatelů, který zvýšil svoji aktivitu, je velký, například, když ráno přijdou lidé do práce nebo odpoledne z oběda, detektor se rychleji přizpůsobí aktivitě a neoznačí většinu komunikací za anomální. Pokud se náhle zvýší aktivita pouze u malé části uživatelů, detektor se náhle zvýšené aktivitě nepřizpůsobí a tuto malou část označí za anomální.

Na obrázku 5.4 vidíme průběhy prahů obou detektorů na službě HTTP. Můžeme si všimnout, že hodnoty prahů se významně neliší. Je to potvrzení domněnky, že pro více používané služby (225 uživatelů v každém časovém okně) zohledňování počtu uživatelů nemá významný vliv na průběh prahu.

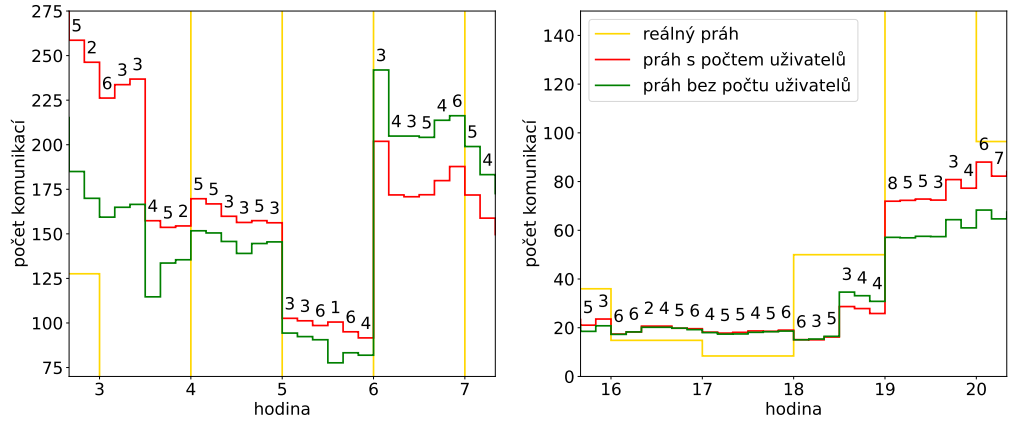
Z těchto důvodů bude aktualizace vnitřních parametrů záviset i na počtu uživatelů v časovém okně. Váhová funkce z formální definice aktualizace vnitřních parametrů (5.1) má tvar:

$$\omega(k) = l(k), \tag{5.3}$$

kde $l(k)$ je funkce počtu uživatelů v k-tém časovém okně.

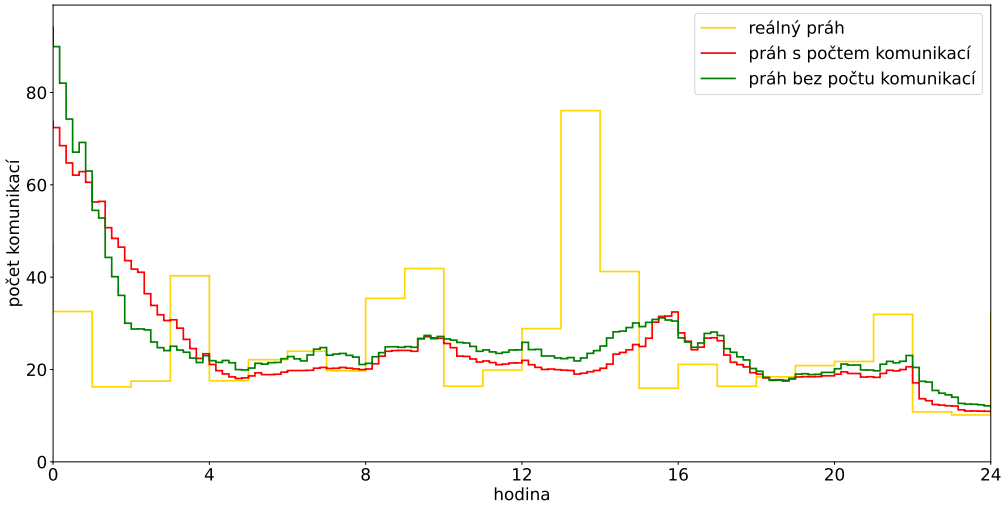


(a) Průběh prahu detektoru dne 7.3.2023.



(b) Zvětšená část grafu 5.3a mezi 3-7 hodinami (vlevo) a 16-20 hodinami (vpravo).

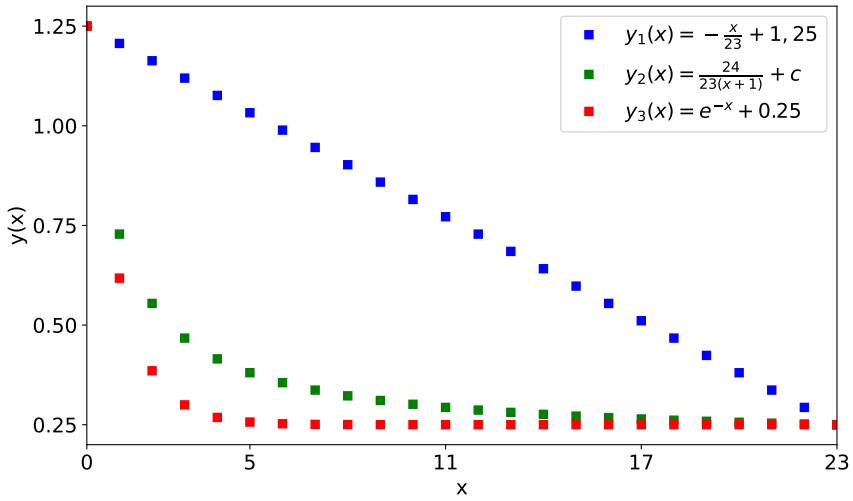
Obrázek 5.3: Reálná aktivita uživatelů (žlutá linie), průběh prahu detektoru (červená a zelená linie) na službě SSH 7.3.2023, kde čísla nad průběhem prahu značí počet uživatelů v časovém okně. Průměrný počet uživatelů v časovém okně je 5.



Obrázek 5.4: Reálná aktivita uživatelů (žlutá linie), průběh prahu detektoru (červená a zelená linie) na službě HTTP 4.3.2023, kde komunikovalo v časovém okně průměrně 225 uživatelů.

5.2.3 Váhová funkce

Váhovou funkci vnitřních parametrů, která se nachází na obrázku 5.5, zavádím proto, aby bez ohledu na délku forget periody měly aktuální data největší vliv na hodnotu vnitřních parametrů modelu. Čím jsou data starší, tím menší vliv by měla mít na hodnotu nových vnitřních parametrů. Motivace zavedení tohoto parametru je podobná jako v předchozí kapitole. Chceme, aby se detektor v dostatečné míře adaptoval na aktuální data a detekoval pouze ty počty komunikací, které skutečně anomální jsou. Dobrou ilustraci této motivace můžeme vidět na obrázku 5.1b, kde mezi 15.-17. hodinou vidíme významně vyšší aktivitu uživatelů než doposud a nechceme, aby detektor označil většinu těchto uživatelů za anomální. Chceme, aby se rychle přizpůsobil zvýšené aktivitě.



Obrázek 5.5: Váhové funkce y proměnné x, kde x představuje pořadí časového okna od přítomnosti do minulosti. Definiční obor funkce začíná od 0 a hodnota $y(0)$ značí váhu aktuálních dat. Funkce jsou nastaveny pro 24 časových oken, které odpovídají 4 hodinové forget periodě s 10 minutovým časovým oknem. Počáteční podmínky: aktuální data mají váhu 1.25 a nejstarší data mají váhu 0.25. Konstanta c je nastavena tak, aby odpovídala počátečním podmínkám: $c = 1,25 - \frac{24}{23}$.

Možností, jak má vypadat váhová funkce, je mnoho. Nabízí se exponenciální funkce, monomiální nebo lineární funkce viz obrázek 5.5. Nejvhodnější tvar váhové funkce je lineární funkce, protože zajistí rovnoměrné klesání váhy předchozích dat. U zbylých dvou funkcí vidíme rychlý pokles, kde od 5. časového okna mají data minimální váhu, která se po zbytek času moc nemění. Použití těchto váhových funkcí má ve výsledku podobný efekt jako nepoužití žádné váhové funkce a nastavení forget periody na 1 hodinu. V kapitole 5.4, kde se věnuji nastavení forget periody, uvidíme, že detektor s hodinovou forget periodou mění hodnotu svého prahu příliš prudce a není tedy použitelný. Definice váhové funkce:

$$y(x) = -\frac{x}{T} + 1,25 \tag{5.4}$$

$D_y = \{0, 1, 2, \ldots, T-1\}$, kde T je fixní parametr, který je roven počtu časových oken v nastavené forget periodě o jedna menší. O jedna menší je z důvodu započítávání aktuálních dat do váhové funkce.

Experimentem bude detekce na službách LDAP a HTTP. Na obě služby aplikuji detektor s lineární váhovou funkcí a bez váhové funkce. Abych zjistil, jak rozdílně detektor funguje, vykreslím anomální komunikace, které nebyly detekovány oběma detektory současně. Na základě výsledků se rozhodnu, jaký detektor detekuje více skutečných anomálních komunikací a méně neanomálních komunikací. Nastavení ostatních vnějších parametrů detektoru:

- Forget a warmup periody jsou nastaveny na 4 hodiny.
- V aktualizaci vnitřních parametrů je zohledněn počet dat v časovém okně.
- K aktualizaci vnitřních parametrů dochází před detekcí.

Na obrázcích 5.6 a 5.7 můžeme vidět žlutou křivku, která ilustruje 0,95 kvantil počtu komunikací za hodinu, červená a zelená křivka značí průběh prahů dektorů s/bez váhové funkce a černé a modré body zobrazují ty anomální komunikace, které byly detekovány pouze jedním z detektorů.

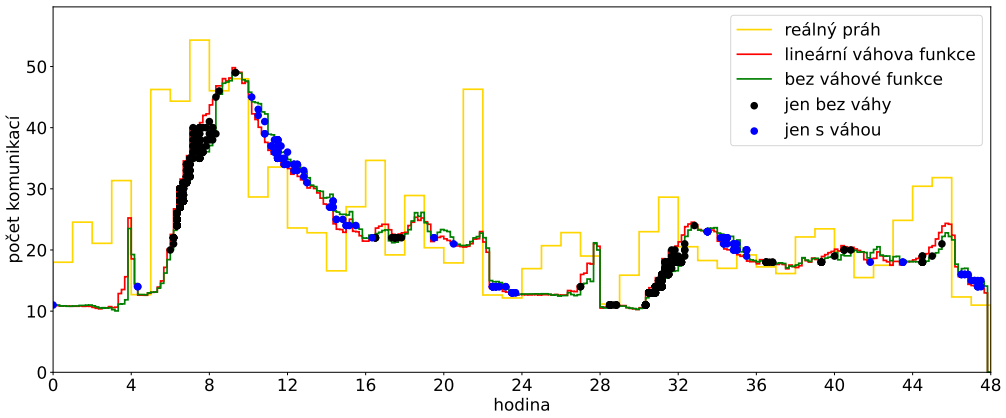
Rozdíl průběhu prahů obou detektorů není příliš velký zejména při malých změnách aktivity uživatelů. Naopak u skokových nárůstů/poklesů aktivity uživatelů se průběhy liší. Na obrázku 5.6b mezi 6.-8., nebo 30.-32. hodinou vidíme, že se detektor s váhovou funkcí rychleji přizpůsobil náhlému zvýšení aktivity a důsledkem je méně detekovaných anomálií, které skutečně anomální nejsou. Naopak při náhlém poklesu aktivity mezi 8.-16. hodinou vidíme u detektoru s váhovou funkcí rychlejší adaptaci a důsledkem je větší množství detekovaných komunikací, které vzhledem k nízké aktivitě skutečně anomální jsou.

V tabulce 5.2 vidíme množství anomálních komunikací, které byly detekovány pouze jedním z detektorů, a můžeme říci, že detekce modelem s váhovou funkcí je o 23% přesnější. Z těchto důvodů je váhová funkce do detektoru zahrnuta.

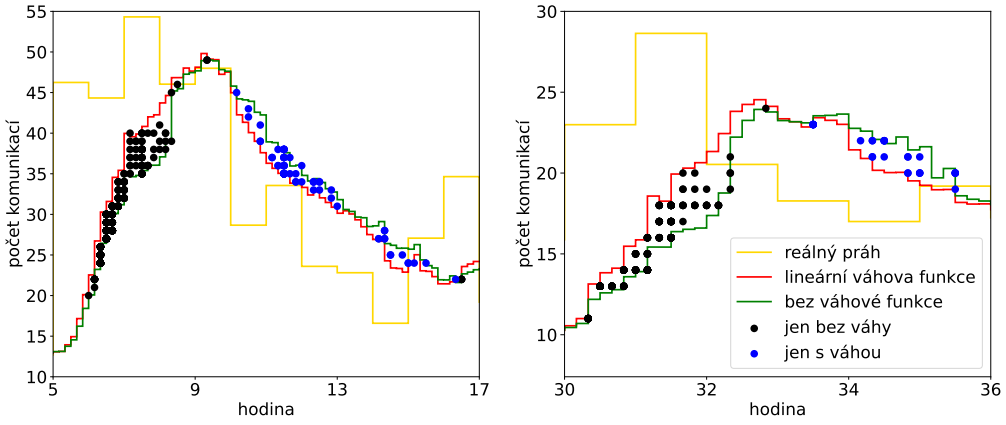
Váhová funkce z formální definice aktualizace vnitřních parametrů (5.1) má nový tvar:

$$\omega(k) = y(k) \cdot l(k), \tag{5.5}$$

kde $y(k)$ je hodnota funkce (5.4) v bodě k.

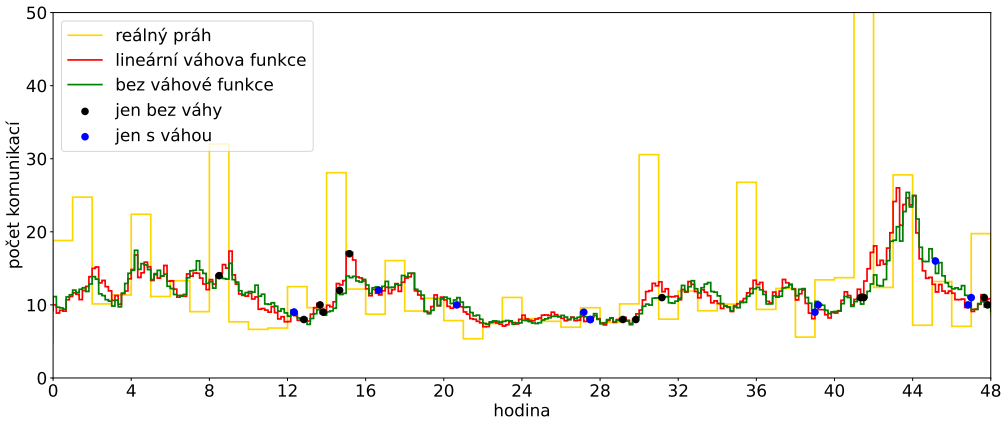


(a) Průběh pradu detektoru ve dnech 6.-7.3.2023.



(b) Zvětšená část grafu 5.6a mezi 5-17 hodinami (vlevo) a 30-36 hodinami (vpravo).

Obrázek 5.6: Reálná aktivita uživatelů (žlutá linie), průběh prahu detektoru (červená a zelená linie) s vyobrazenými anomálními komunikacemi (modré a černé body) zachyceny pouze jedním detektorem pro službu HTTP ve dnech 6.-7.3.2023.



Obrázek 5.7: Reálná aktivita uživatelů (žlutá linie), průběh prahu detektoru (červená a zelená linie) s vyobrazenými anomálními komunikacemi (modré a černé body) zachyceny pouze jedním detektorem pro službu LDAP ve dnech 2.-3.3.2023.

	bez váhové funkce	s váhovou funkcí
LDAP	10,4 % (17/164)	11,4 % (19/166)
HTTP	16,9 % (1548/9177)	7,29 % (600/8229)

Tabulka 5.2: Tabulka počtu anomálií, které byly detekovány pouze jedním detektorem ve dnech 1.-7.3.2023

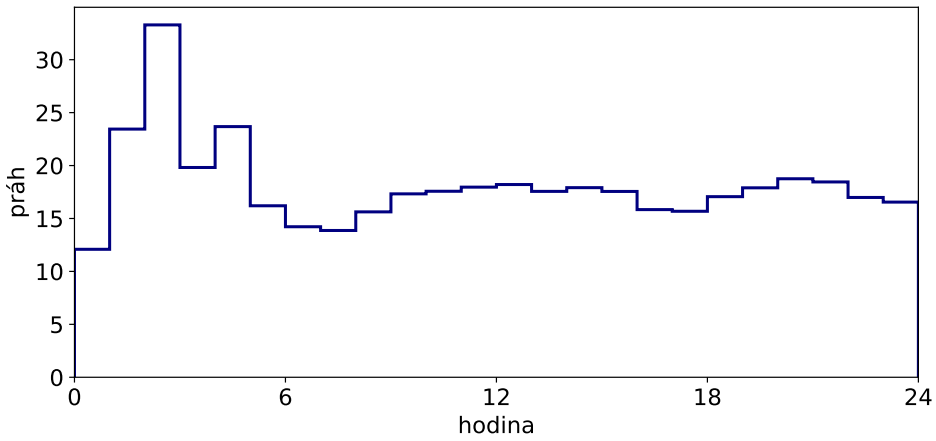
5.3 Warmup perioda

Warmup perioda je doba od spuštění detektoru $[0, t]$, po kterou bude pouze sbírat data pro počáteční inicializaci vnitřních parametrů a nebude detekovat anomální komunikace.

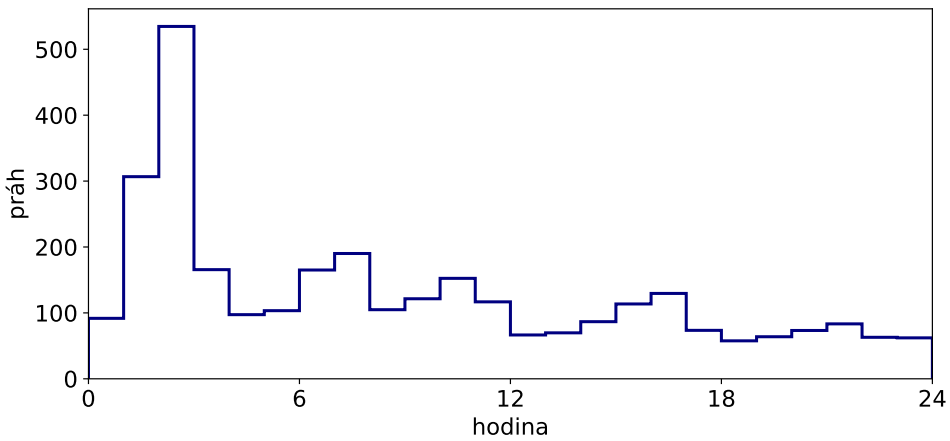
Detektor, který by neprovedl počáteční inicializaci parametrů, by byl z počátku velice nepřesný.

Pro stanovení správné délky warmup periody je nutné získat data z telemetrie, která jsou rozdělena po hodinách pro konkrétní síť a službu. Následně se odhadnou parametry distribuční funkce každou hodinu na základě všech předchozích dat a vypočítá se práh modelu. Tyto parametry budou použity k výpočtu prahu modelu pro každou hodinu. Pro určení warmup periody bude rozhodující průběh prahu a variability v čase.

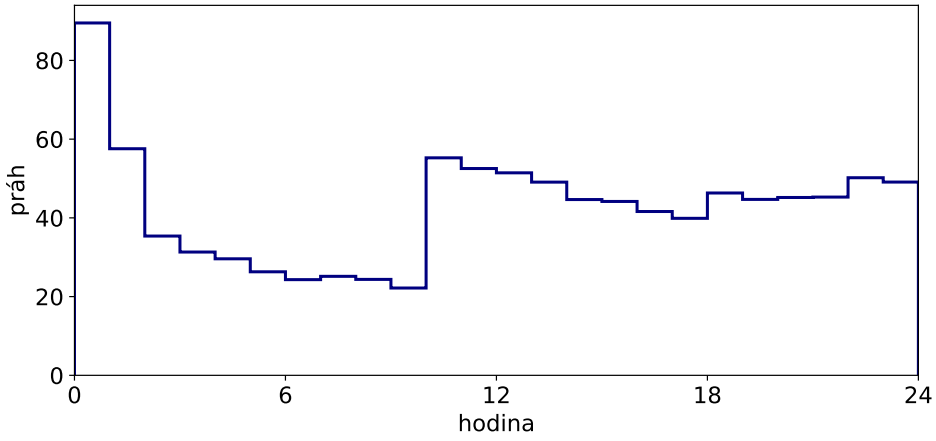
Stáhnu si z úložiště data napříč sítěmi na službách LDAP, HTTP a SSH pro 3 všední a 3 víkendové dny. Na obrázcích 5.8 můžeme v počátečních hodinách vidět významný rozptyl prahu. Ten je způsoben nedostatkem vstupních dat. Postupem času vidíme zejména na službách LDAP 5.8a a HTTP 5.8b konvergenci prahu. Větší rozptyl průběhu prahu u služby SSH 5.8c je způsoben malým počtem uživatelů, kterých jsou pouze desítky. Ze všech obrázků 5.8 můžeme usoudit, že k významnému snížení počátečního rozptylu dojde v řádu hodin. Hodnota warmup periody se proto bude pohybovat mezi 4-12 hodinami.



(a) LDAP 27.8.2022.



(b) HTTP 20.7.2022.



(c) SSH 27.8.2022.

Obrázek 5.8: Průběh prahu detektoru aktualizovaného novými daty na službách LDAP, HTTP a SSH.

5.4 Forget perioda

Forget perioda značí velikost časového intervalu x, který je myšlen od minulosti do přítomnosti $[t - x, t]$, na základě kterého si bude detektor do paměti ukládat všechny vnitřní parametry modelu, které proběhly v tomto intervalu. Délka forget periody je klíčová pro správnou detekci anomálního počtu komunikací. Její velikost stojí za flexibilitou detektoru, jinými slovy, jak rychle se budou parametry modelu měnit vzhledem k aktuálním datům.

Nyní stojíme před otázkou, zda udělat detektor, který si bude do paměti ukládat vnitřní parametry staré v řádu několika dní (dlouhodobý), nebo pouze v řádu hodin (krátkodobý). Rozdíl mezi těmito detektory je v rychlosti změny jejich prahu. Dlouhodobý detektor se nebude dostatečně adaptovat na aktuální aktivitu v síti. Pro příklad si uveďme průběh aktivity na obrázku 5.1b, kde vidíme nárůst aktivity mezi 15.-17. hodinou. Na tuto změnu by se dlouhodobý detektor neadaptoval a v tomto případě by většinu komunikací označil za anomální. Naopak krátkodobý detektor se snadno na tuto změnu adaptuje. Nevýhoda krátkodobého detektoru je, že krátká doba časového intervalu může způsobit, že detektor bude detekovat nová data na základě nepřesné informace o chování uživatelů na síti. My jsme se rozhodli pro krátkodobý detektor a ten bude následně analyzován.

Úkolem bude spustit detektor na detekčním systému s různými hodnotami forget periody a porovnat chování jednotlivých prahů, jejich dynamiku a flexibilitu. Pracuji s týdenními daty dne 21.-27.9.2022. Analýzu dělám pro služby LDAP, HTTP a SMB a pro 1, 4, 8, 12 a 24 hodinovou forget periodu. Ostatní vnější parametry jsou nastaveny následovně:

- Warmup perioda se rovná forget periodě.
- K aktualizaci vnitřních parametrů dochází před vyhodnocením dat.
- V aktualizaci vnitřních parametrů je zohledněn počet dat v časovém okně.
- V modelu je použita lineární váhová funkce (5.4).

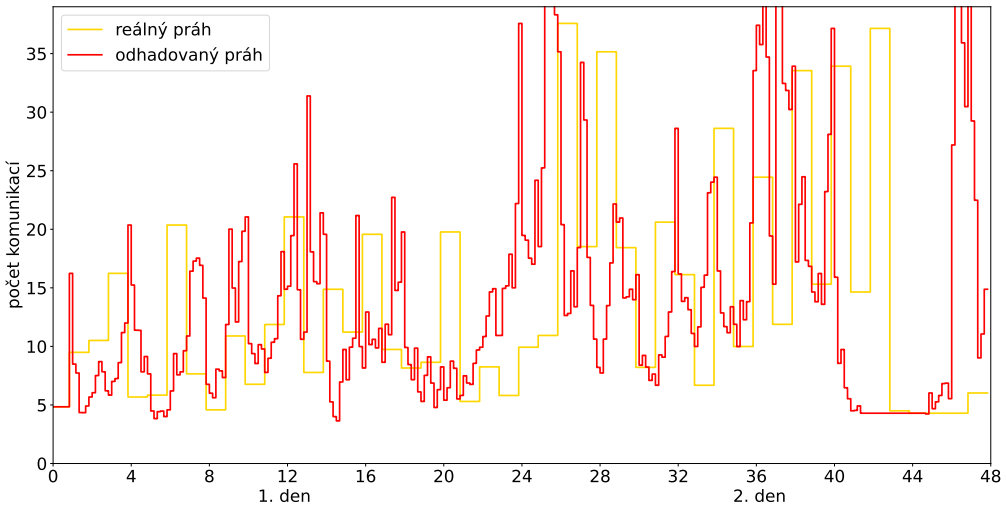
Na obrázcích 5.9 a 5.10 vidíme data pro dva vybrané dny ze zmiňovaného týdne. Žlutá barva prezentuje skutečnou aktivitu konkrétní sítě ve formě boxplotů a průběh $0,95$ kvantilu počtu komunikací pro každou hodinu. Modrá, resp. oranžová, resp. zelená, resp. červená křivka zobrazuje průběh prahu detektoru se 4, resp. 8, resp. 12, resp. 24 hodinovou forget periodou.

Na všech křivkách v obrázcích 5.9, 5.10a a 5.10b vidíme počáteční adaptaci na data. Počáteční hodnota prahu je vypočítaná na základě všech dat v intervalu warmup periody, proto se hodnoty prahu pro jednotlivé periody liší.

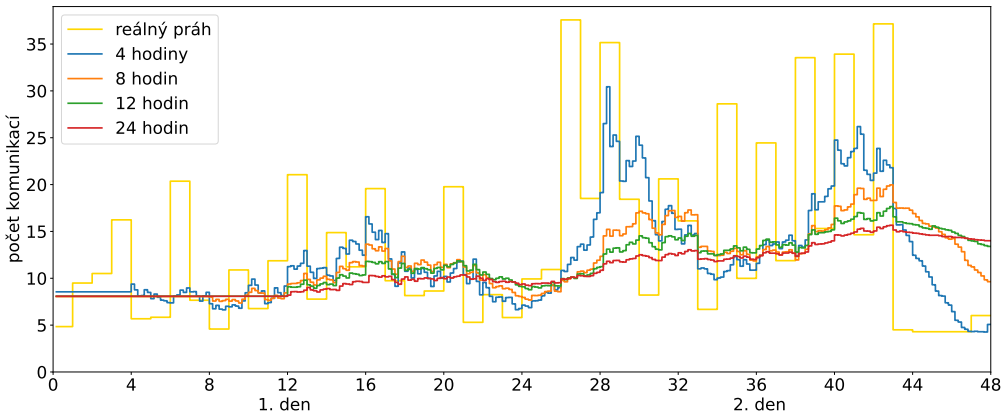
Na obrázku 5.9 vidíme průběh prahu v modelu, ve kterém je nastavena hodnota forget periody 1 hodina. Hodnota prahu se rychle mění, a to z důvodu nízkého počtu dat v paměti. Dynamické chování prahu je také způsobeno váhovou funkcí, protože nová data mají na aktualizaci parametrů významný vliv.

Jednotlivé průběhy prahů v závislosti na forget periodě projevují rozdílnou míru adaptability k aktuálním datům. Lze pozorovat, že model s forget periodou trvající 4 hodiny dokáže reagovat dostatečně rychle na aktuální data, zatímco model s forget periodou trvající 24 hodin působí konzervativně a zpožďuje svoji reakci na data. Modely s forget periodami trvajícími 4 a 8 hodin projevují klidné chování, ale zároveň dokáží adekvátně reflektovat dynamiku chování uživatelů. Při zvýšení aktivity (obrázek 5.10a 2 den, 5. hodina) nebo náhlé pasivitě uživatelů (obrázek 5.10c 1. den, 7. hodina) se rychleji přizpůsobí model s nižší forget periodou, zatímco model s vyšší forget periodou reaguje pomaleji a se zpožděním. Proto se budou v následující kapitole detailněji analyzovat modely se 4 a 8 hodinovou forget periodou, kde se na základě statistik vybere nejvhodnější model.

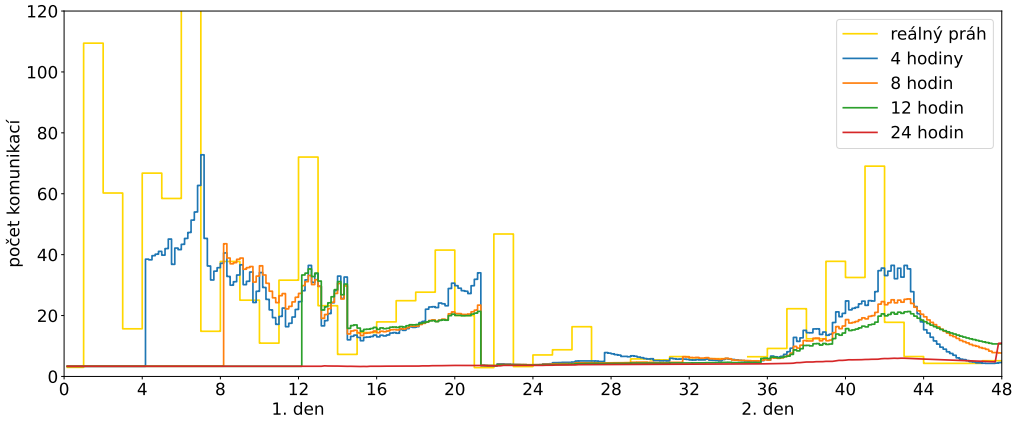
Na obrázku 5.10c ve 3. hodině 2. dne si můžeme všimnout neočekávaného nárůstu prahu modelu s 4 hodinovou forget periodou. Tento efekt je způsoben jevem, kdy skokovému nárůstu aktivity předchází pasivita. Model s nízkou hodnotou forget periody ztratí většinu dat z období pasivního chování a naopak má k dispozici data, která značí velkou aktivitu. Tento jev je navíc zesílen váhovou funkcí, která náhle zvýšenou aktivitu upřednostňuje.



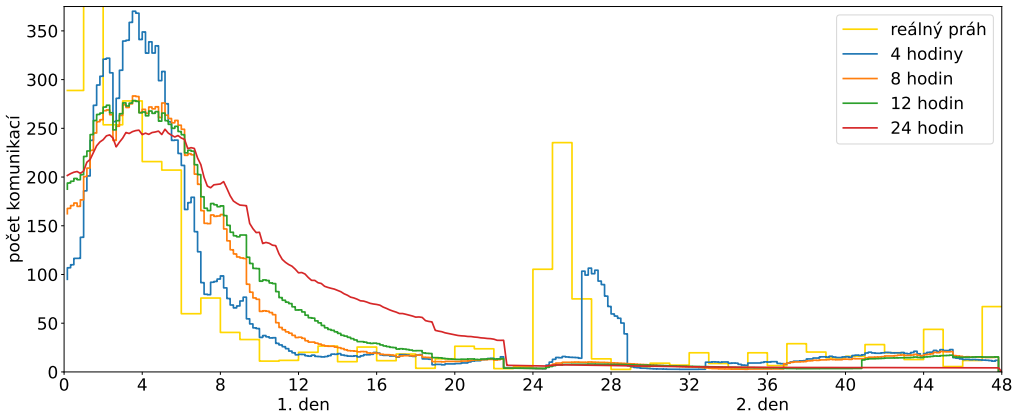
Obrázek 5.9: Reálná aktivita uživatelů (žlutá linie), průběh prahu detektoru (červená linie) na službě LDAP dne 21.-22.9.2022 s 1 hodinovou forget periodou.



(a) LDAP 21.-22.9.2022.



(b) SMB 21.-22.9.2022.



(c) HTTP 26.-27.9.2022.

Obrázek 5.10: Reálná aktivita uživatelů (žlutá linie), průběh prahu detektoru (modrá, oranžová, zelená a červená linie) v závislosti na délce forget periody na daných službách.

5.5 Finální vyhodnocení

V této kapitole provedu detailnější analýzu detektoru s forget periodami 4 a 8 hodin. Aby náš detektor detekoval anomálie co nejpřesněji, nebude stačit z obrázku posuzovat detektor pouze podle průběhu prahu, ale je třeba zavést vhodné veličiny, které budou měřit, jak se detektor chová a podle kterých se rozhodnu, jaký detektor je nejlepší.

Výběr veličin se bude opírat o kontext síťových služeb. V souvislosti s naším tématem víme, že množství anomálního počtu komunikací by mělo být menší než 5%, jinými slovy detektor by měl označovat největší počty komunikací vzhledem ke zbytku dat. Dále víme, že počet uživatelů, kteří se chovají anomálně vůči zbytku, je také malý. Pokud například dojde k vniknutí útočníka do lokální sítě, začne se chovat anomálně pouze ten počítač či série počítačů, které byly napadené. Posledním ukazatelem správné funkčnosti detektoru budou senzitivita [42] a specificita [43], které měří přesnost detektoru detekovat skutečné anomálie. Vybrané veličiny jsou následující:

1. **Množství anomálních komunikací**: Jakou část z celého množství komunikací označil detektor jako anomální.
2. **Množství anomálních uživatelů**: Jaké procento uživatelů z celé sítě detektor označil jako anomální.
3. **Senzitivita detektoru**: Vyjadřuje úspěšnost, s níž detektor zachytí přítomnost skutečné anomálie.

$$\text{senzitivita} = \frac{\#\text{ skutečných anomálií}}{\#\text{ skutečných anomálií} + \#\text{ falešně negativních anomálií}} \tag{5.6}$$

4. **Specificita detektoru**: Vyjadřuje schopnost detektoru přesně vybrat počty komunikací, které skutečně nejsou anomální.

$$\text{specifita} = \frac{\#\text{ skutečných anomálií}}{\#\text{ skutečných anomálií} + \#\text{ falešně pozitivních anomálií}} \tag{5.7}$$

Nastavení detektoru je stejné jako v 5.4:

- Warmup perioda se rovná forget periodě.
- K aktualizaci vnitřních parametrů dochází před vyhodnocením dat.
- V aktualizaci vnitřních parametrů je zohledněn počet dat v časovém okně.
- V modelu je použita lineární váhová funkce (5.4).

Pracuji s týdenními daty dne 1.-7.3.2023 a analýzu jsem provedl na službách LDAP, HTTP, SMB a SSH. Pro použití senzitivity a specificity je zapotřebí mít k dispozici seznam skutečných anomálií. Tento seznam byl vytvořen manuálním označením počtu komunikací ze seznamu všech komunikací, které se staly ve vymezeném časovém období.

Poznámka 5.5.1 *Je nutné si uvědomit, že seznam anomálních komunikací byl vytvořen subjektivním označením komunikací, proto není přesný, je pouze orientační. Část manuálně označených komunikací ve skutečnosti nemusí být anomální a část neoznačených komunikací může být reálně anomální. Proto do odhadnuté senzitivity a specificity přispívá negativně lidský faktor.*

Z tabulek 5.3 a 5.4 můžeme vidět, že detektor s 8 hodinovou forget periodou označuje méně komunikací a počtu uživatelů než detektor s 4 hodinovou forget periodou napříč všemi službami. Detektor s 8 hodinovou forget periodou vynechává víc skutečných anomálních komunikací, ale zároveň detekuje méně běžných komunikací. Dále vidíme, že na službách LDAP a SMB je detekováno zhruba 5 % uživatelů, což je chování, které je očekávané. Na službě HTTP je detekováno zhruba 10 % anomálních uživatelů. Na této službě bude žádoucí posunout práh výš. Naopak na službě SSH je detekováno méně než procento anomálních uživatelů, zde bude žádoucí posunout práh detektoru níž.

Na službách SMB a SSH vidíme očekávané chování v tom, že procento anomálních komunikací je podobné procentu anomálních uživatelů viz tabulka 5.4, naopak na službách LDAP a HTTP se procenta množství anomálních komunikací a anomálních uživatelů dramaticky liší. Tento rozdíl je způsoben v použití jednotlivých služeb.

V tabulkách 5.5 a 5.6 vidíme porovnání, jak přesné jsou detektory se 4 a 8 hodinovou forget periodou. Detektor s 8 hodinovou forget periodou má na všech službách větší senzitivitu. Na službách HTTP, SMB a SSH má specificitu podobnou jako detektor se 4 hodinovou forget periodou, ale na službě LDAP je u 8 hodinové forget periody specificita vyšší o 7 %.

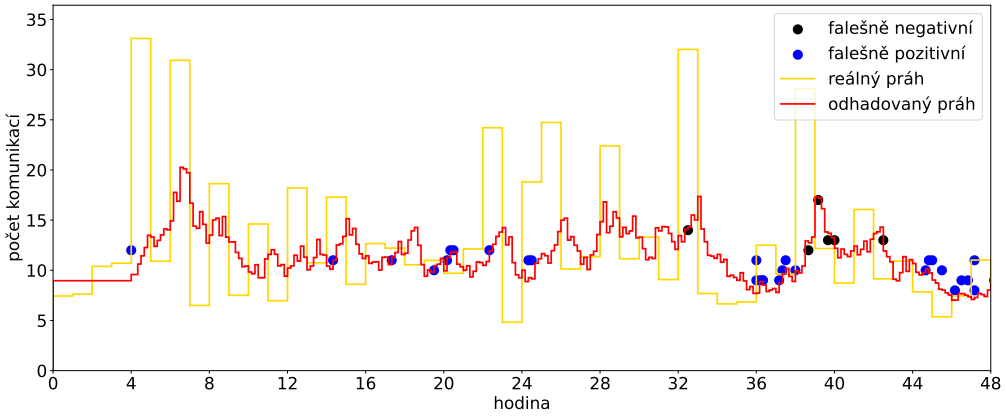
Na obrázcích 5.11, 5.12 a 5.13 vidíme reálný $0,95$ kvantil počtu komunikací seskupených po hodině, průběh odhadovaného prahu a zobrazené falešně pozitivní, nebo negativní komunikace uživatelů. Závěry z tabulky se shodují se závěry obrázků. V grafech 5.11a, 5.12a a 5.13a vidíme, že se průběh prahu chová dynamičtěji než v grafech 5.11b, 5.12b a 5.13b, proto při prudkém nárůstu, zapříčiněným náhlou aktivitou, pár anomálních komunikací neoznačí jako anomální.

	LDAP		HTTP	
forget perioda	komunikace (2 456 068)	uživatelé (39 788)	komunikace (3 026 031)	uživatelé (50 285)
4	0,29 %	5,34 %	0,61 %	10,1 %
8	0,27 %	4,69 %	0,53 %	9,83 %

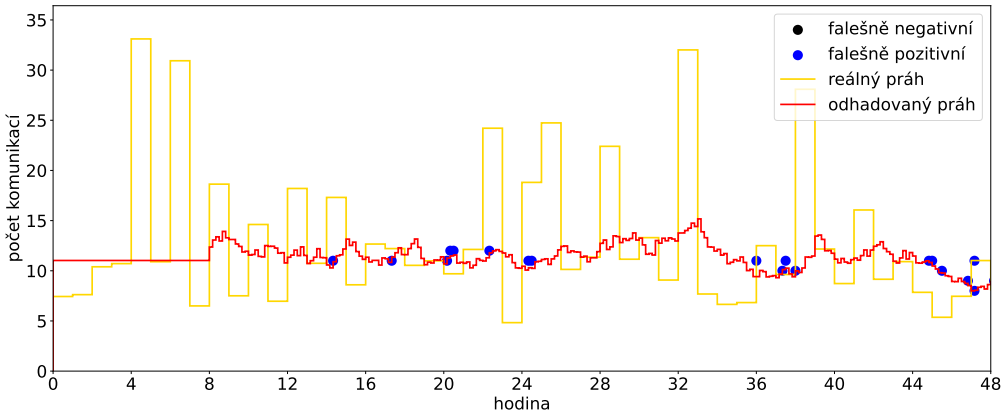
Tabulka 5.3: Tabulka veličin na službách LDAP a HTTP, pro data naměřena v týdnu 1.-7.3. napříč 6 sítěmi. Komunikace značí množství anomálního počtu komunikací. Uživatelé znamená množství anomálních uživatelů.

	SMB		SSH	
forget perioda	komunikace (22 759)	uživatelé (39 788)	komunikace (237 027)	uživatelé (50 285)
4	6,30 %	5,60 %	0,67 %	0,82 %
8	5,30 %	4,74 %	0,52 %	0,49 %

Tabulka 5.4: Tabulka veličin na službách SMB a SSH, pro data naměřena v týdnu 1.-7.3. napříč 6 sítěmi. Komunikace značí množství anomálního počtu komunikací. Uživatelé znamená množství anomálních uživatelů.

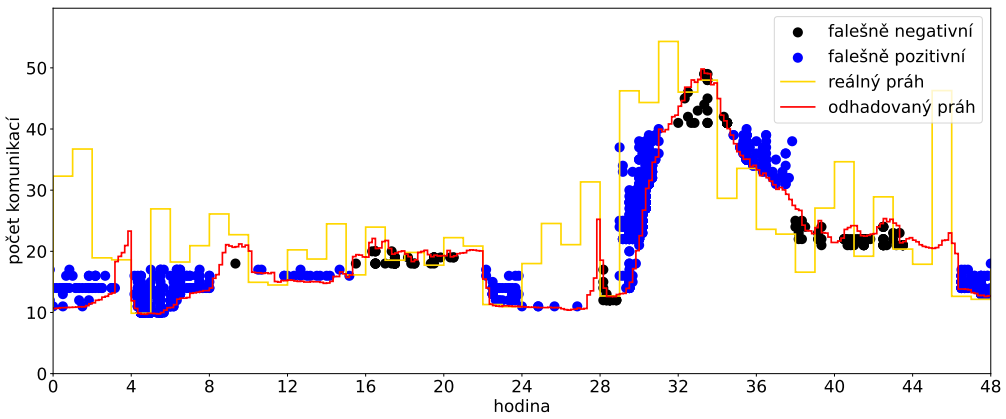


(a) Forget perioda 4 hodiny.

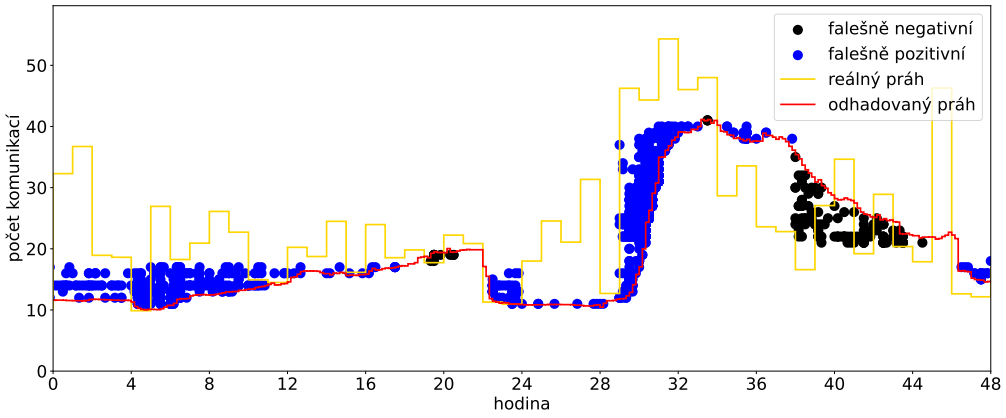


(b) Forget perioda 8 hodin.

Obrázek 5.11: Reálná aktivita uživatelů (žlutá linie) a průběh prahu detektoru (červená linie) s falešně pozitivními a negativními anomáliemi (modré a černé body) v závislosti na délce forget periody na službě LDAP 1.-2.3.2023.

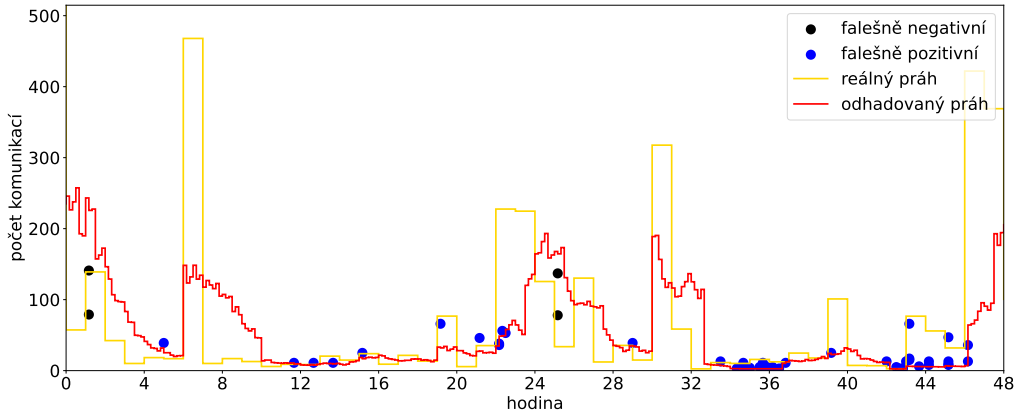


(a) Forget perioda 4 hodiny.

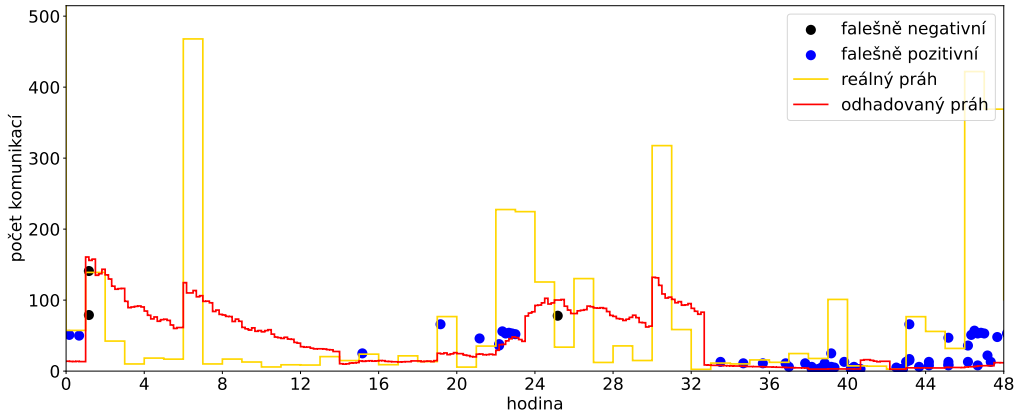


(b) Forget perioda 8 hodin.

Obrázek 5.12: Reálná aktivita uživatelů (žlutá linie) a průběh prahu detektoru (červená linie) s falešně pozitivními a negativními anomáliemi (modré a černé body) v závislosti na délce forget periody na službě HTTP 5.-6.3.2023



(a) Forget perioda 4 hodiny.



(b) Forget perioda 8 hodin.

Obrázek 5.13: Reálná aktivita uživatelů (žlutá linie) a průběh prahu detektoru (červená linie) s falešně pozitivními a negativními anomáliemi (modré a černé body) v závislosti na délce forget periody na službě SSH 4.-5.3.2023.

	LDAP		HTTP	
forget perioda	senzitivita	specificita	senzitivita	specificita
4	88,47 %	65,97 %	85,14 %	64,44 %
8	94,91 %	73,32 %	88,48 %	63,89 %

Tabulka 5.5: Tabulka senzitivity a specificity na službách LDAP a HTTP, pro data naměřena v týdnech 1.-7.3. a 12.-18.4. napříč 2 sítěmi.

	SMB		SSH	
forget perioda	senzitivita	specificita	senzitivita	specificita
4	76,70 %	65,25 %	75.38 %	44,98 %
8	94,06 %	66,87 %	83,27 %	42,77 %

Tabulka 5.6: Tabulka senzitivity a specificity na službách SMB a SSH, pro data naměřena v týdnech 1.-7.3. a 12.-18.4. napříč 2 sítěmi.

5.5.1 Shrnutí nastavení vnějších parametrů

V této podkapitole shrnu celkové nastavení detektoru, které jsme podle výsledků experimentu vybrali.

- **Statistický model:** Podle metod, které odhadují nejlepší systém hustot (QQ-plot, G-statistika a variační koeficient), bylo vybráno, jako statistický model, Weibullovo rozdělení.
- **Práh:** Detektor bude označovat za anomální ty počty komunikací, které jsou větší než je $0,95$-kvantil statistického modelu.
- **Způsob aktualizace vnitřních parametrů:** Nová hodnota vnitřních parametrů se vypočítá pomocí váženého průměru předchozích vnitřních parametrů.
- **Moment aktualizace:** Vnitřní parametry se budou aktualizovat před samotnou detekcí.
- **Počet uživatelů:** Při aktualizaci vnitřních parametrů je zohledněn počet uživatelů, kteří v daném časovém okně komunikovali.
- **Váhová funkce:** Do aktualizace vnitřních parametrů vstupuje lineární váhová funkce, která způsobí, že čím jsou vnitřní parametry starší, tím budou mít na aktuální hodnotu vnitřních parametrů menší vliv.
- **Warmup perioda:** Délka warmup periody musí být minimálně 4 hodiny dlouhá. Je ovšem navázána na délku forget periody.
- **Forget perioda:** Nejvhodnější délka forget periody je 8 hodin.

Detekce takto nastaveného detektoru funguje vcelku dobře. Senzitivita detektoru se pohybuje kolem 90 % a specificita kolem 60 %. Ovšem před nasazením do produkce bude nutné doladit nastavení prahu. Nejvíce chyb detektor dělá na hranici prahu.

Proto při další práci pro zdokonalení detektoru je třeba zavést tzv. *fuzzy threshold*, kde hranicí nebude diskrétní číslo, ale interval, který s sebou ponese informaci, s jakou pravděpodobností je daný počet komunikací anomální.

Závěr

Hlavním tématem této diplomové práce byla detekce anomálií, kde anomálie je chápána jako neobvykle vysoký počet spojení na danou síťovou službu. Konkrétněji se pak zaměřuje na nalezení vhodné metody, která bude dostatečně přesná, paměťově úsporná a schopná pracovat v reálném čase.

Cílem je detekovat bodové anomálie na síťových službách v aplikační vrstvě. Tyto anomálie mohou být způsobeny útočníkem při pokusu o prolomení hesla šifrované komunikace (protokol SSH), nebo například při jeho snaze nalézt v síti další zranitelná zařízení (protokol SMB). Pro tento typ se nejlépe hodí detekční metody založené na statistice.

Navrhovaná metoda tedy využívá statistický model pro spolehlivý popis pozorovaných dat. Pro fungování v reálném čase je nutná efektivita, rychlost a paměťová úspornost, což vedlo k volbě parametrického typu odhadu distribuční funkce. Pro zúžení kandidátů byl použit QQ-plot a také informace, že data mají těžké chvosty. Zvolenými kandidáty byly Weibullovo, exponenciální Weibullovo, zobecněné logistické, log-normální a zobecněné Paretovo rozdělení. Pro nalezení nejlepšího rozdělení z těchto kandidátů jsem na všechna data aplikoval, už zmíněnou, G-statistiku a variační koeficient. Z analýzy plyne, že pro celosíťový i uživatelský model, nejlépe data odhaduje Weibullovo rozdělení.

Druhým závěrem této práce je, že variační koeficient sám o sobě neumožňuje jednoznačně určit, který systém hustot nejlépe popisuje data, a že pro kvalitní závěry je třeba použít i další statistické metody. G-statistika jednoznačně určí nejvhodnější systém hustot a je rychlá a efektivní.

Protože chování uživatelů na síti je dynamické a mění svůj charakter, bylo třeba do modelu přidat tzv. vnější parametry.

Proto se druhá část práce zaměřuje na experimentální výzkum vnějších parametrů a jejich vlivu na adaptaci modelu na aktuální aktivitu uživatelů. Experimenty jsem prováděl na službách LDAP, HTTP, SMB a SSH vždy pro týden dat napříč sítěmi.

Nejprve bylo třeba definovat způsob aktualizace vnitřních parametrů. Zvolená metoda vypočítá nové vnitřní parametry Weibullova rozdělení pomocí váženého průměru historických vnitřních parametrů, které má detektor uložené v paměti.

Druhým vnějším parametrem byl moment aktualizace vnitřních parametrů. Otázkou zde bylo, zda aktualizovat vnitřní parametry před detekcí na nově příchozích datech, nebo až po té. Výsledky experimentu ukázaly, že se přesnost detektoru, který aktualizuje vnitřní parametry před detekcí, zvýšila o 15 %.

Třetím parametrem, který jsem diskutoval, bylo zohlednění počtu uživatelů v časovém okně. Přidáním tohoto parametru do modelu se předešlo náhlým změnám hodnot prahu detektoru a rozlišení zda je náhlá změna aktivity opodstatněná, či nikoliv. Tento parametr má významný efekt zejména u méně používaných služeb, jako jsou SSH, SMB.

Dalším vnějším parametrem byla váhová funkce. Váhovou funkci jsem zavedl proto, aby bez ohledu na délku forget periody měla aktuální data největší vliv na hodnotu vnitřních parametrů a čím jsou data starší, tím by měl být jejich vliv na nastaveni modelu menší. Proto jsem zvolil lineární váhovou funkci, kde vliv dat v čase rovnoměrně klesá. Detektor s váhovou funkcí detekuje anomálie o 23 % přesněji.

Posledním parametrem, který jsem zavedl, byla forget perioda, která znamená délku časového intervalu od minulosti do přítomnosti. Na základě délky forget periody si bude detektor do paměti ukládat vnitřní parametry komunikací, které proběhly v tomto časovém intervalu. Velikost tohoto parametru stojí za flexibilitou detektoru, jinými slovy, jak rychle se budou vnitřní parametry modelu měnit vzhledem k aktuálním datům. Zaměřil jsem se na krátkodobou variantu, což znamená délku forget periody v řádu hodin. Podle měření senzitivity a specificity modelu nejpřesněji detekuje anomální počty komunikací detektor s 8 hodinovou forget periodou.

Detekce anomálií s použitím detektoru s nastavením těchto vnějších parametrů funguje celkově dobře. Nicméně, pro nasazení detektoru do produkčního prostředí, bude nutné optimalizovat nastavení prahu. Nejčastěji detektor špatně detekuje počty komunikací na hranici prahu. Pro jeho zdokonalení bude proto vhodné pro další práci použít tzv. *fuzzy threshold*, který stanový interval namísto diskrétní hodnoty prahu. Tento interval bude obsahovat informaci o pravděpodobnosti, s jakou bude daný počet komunikací považován za anomální.

Reference

[1] KrebsonSecurity, *A Closer Look at the LAPSUS$ Data Extortion Group*, [Online; accessed 19-June-2022], 2022. URL: https://krebsonsecurity.com/2022/03/a-closer-look-at-the-lapsus-data-extortion-group/.

[2] J. Kurose a K. Ross, *Computer networking*, 7. vyd. Upper Saddle River, NJ: Pearson, dub. 2016.

[3] SDxCentral, *Network Service*, [Online; accessed 2-August-2022], 2015. URL: https://www.sdxcentral.com/resources/glossary/network-service/.

[4] A. S. Gillis, *What is LDAP (Lightweight Directory Access Protocol)?* Pros. 2022. URL: https://www.techtarget.com/searchmobilecomputing/definition/LDAP.

[5] Wikipedia contributors, *Hypertext Transfer Protocol — Wikipedia, The Free Encyclopedia*, [Online; accessed 2-August-2022], 2022. URL: https://en.wikipedia.org/w/index.php?title=Hypertext_Transfer_Protocol&oldid=1101281242.

[6] R. Sheldon a J. Scarpati, *What is the server message block (SMB) protocol? how does it work?* Srp. 2021. URL: https://www.techtarget.com/searchnetworking/definition/Server-Message-Block-Protocol.

[7] T. Ylonen, *What is SSH (secure shell)? | SSH Academy*, 2023. URL: https://www.google.com/amp/s/www.ssh.com/academy/ssh%5C%3fhs_amp=true.

[8] D. Hawkins, *Identification of outliers*, en. Springer Science & Business Media, dub. 2013.

[9] V. Barnett a T. Lewis, *Outliers in Statistical Data* (Wiley Series in Probability and Statistics), en, 3. vyd. Chichester, England: John Wiley & Sons, ún. 1994.

[10] A. A. Cook, G. Misirli a Z. Fan, "Anomaly detection for IoT time-series data: A survey," *IEEE Internet of Things Journal*, roč. 7, č. 7, s. 6481–6494, 2019.

[11] M. Braei a S. Wagner, "Anomaly detection in univariate time-series: A survey on the state-of-the-art," *arXiv preprint arXiv:2004.00433*, 2020.

[12] V. Chandola, A. Banerjee a V. Kumar, "Anomaly detection," *ACM Computing Surveys*, roč. 41, č. 3, s. 20–39, srp. 2007. DOI: 10.1145/1541880.1541882.

[13] Z. Chen, C. K. Yeo, B. S. Lee a C. T. Lau, "Autoencoder-based network anomaly detection," in *2018 Wireless telecommunications symposium (WTS)*, IEEE, 2018, s. 1–5.

[14] J. An a S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special lecture on IE*, roč. 2, č. 1, s. 1–18, 2015.

[15] K.-L. Li, H.-K. Huang, S.-F. Tian a W. Xu, "Improving one-class SVM for anomaly detection," in *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, IEEE, sv. 5, 2003, s. 3077–3081.

[16] M. Zamini a S. M. H. Hasheminejad, "A comprehensive survey of anomaly detection in banking, wireless sensor networks, social networks, and healthcare," *Intelligent Decision Technologies*, dub. 2019.

[17] J. Chen, J. Zhang, R. Qian, J. Yuan a Y. Ren, "An Anomaly Detection Method for Wireless Sensor Networks Based on the Improved Isolation Forest," *Applied Sciences*, roč. 13, s. 702, led. 2023. DOI: 10.3390/app13020702.

[18] E. M. Knorr, R. T. Ng a V. Tucakov, "Distance-based outliers: algorithms and applications," *VLDB J.*, roč. 8, č. 3-4, s. 237–253, ún. 2000.

[19] S. Omar, M. Ngadi, H. Jebur a S. Benqdara, "Machine Learning Techniques for Anomaly Detection: An Overview," *International Journal of Computer Applications*, roč. 79, říj. 2013. DOI: 10.5120/13715-1478.

[20] K. Chuntonov, A. Ivanov, B. Verbitsky a V. Kozhevnikov, "Gas Purification and Quality Control of the End Gas Product," roč. 5, s. 44–58, led. 2017. DOI: 10.4236/msce.2017.58005.

[21] N. Ye a Q. Chen, "An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems," *Quality and Reliability Engineering International*, roč. 17, s. 105–112, břez. 2001. DOI: 10.1002/qre.392.

[22] N. Paulauskas a A. Baskys, "Application of histogram-based outlier scores to detect computer network anomalies," *Electronics*, roč. 8, č. 11, s. 1251, 2019.

[23] A. Cavin, *Fast anomaly detection with python*, srp. 2022. URL: https://towardsdatascience.com/hbos-vs-iforest-on-macbook-pro-m1-c258d2b5fe6b.

[24] V. Kůs a M. Kovanda, *Matematická statistika*. Fakulta jaderná a fyzikálně inženýrská Praha, 2020, s. 36–46. URL: https://makovanda.eu/_data/01MAS.pdf.

[25] V. Kůs a M. Kovanda, *Míra a pravděpodobnost*. Fakulta jaderná a fyzikálně inženýrská Praha, 2020, s. 27. URL: https://makovanda.eu/_data/01MIP.pdf.

[26] J. Gibbons D a S. Chakraborti, *Nonparametric statistical inference* ("Statistics: A Series of Textbooks and Monographs"). "Boca Raton, FL": CRC Press, květ. 2003, s. 144–145.

[27] F. Pavelka a P. Klímek, *Aplikovaná statistika*. Vysoké učení technické, Fakulta managementu a ekonomiky ve Zlíně, 2000.

[28] T. Rolski, H. Schmidli, V. Schmidt a J. L. Teugels, *Stochastic processes for insurance and finance*. John Wiley & Sons, 2009.

[29] A. Kızılersü, M. Kreer a A. W. Thomas, "The weibull distribution," en, *Signif. (Oxf.)*, roč. 15, č. 2, s. 10–11, dub. 2018.

[30] C. Carrillo, J. Cidrás, E. Díaz-Dorado a A. F. Obando-Montaño, "An Approach to Determine the Weibull Parameters for Wind Energy Analysis: The Case of Galicia (Spain)," *Energies*, roč. 7, č. 4, s. 2676–2700, 2014, ISSN: 1996-1073. DOI: 10.3390/en7042676. URL: https://www.mdpi.com/1996-1073/7/4/2676.

[31] G. S. Mudholkar a D. K. Srivastava, "Exponentiated Weibull family for analyzing bathtub failure-rate data," *IEEE Trans. Reliab.*, roč. 42, č. 2, s. 299–302, čvn. 1993.

[32] M. Pal, M. Ali a J. Woo, "Exponentiated Weibull distribution," *Statistica*, roč. 66, s. 139–147, led. 2003. DOI: 10.6092/issn.1973-2201/493.

[33] N. L. Johnson, S. Kotz a N. Balakrishnan, *Continuous Univariate Distributions, Volume 2* (Wiley Series in Probability and Statistics), en, 2. vyd. Nashville, TN: John Wiley & Sons, dub. 1995.

[34] B. Lagos-Álvarez, N. Jerez-Lillo, J. P. Navarrete, J. Figueroa-Zúñiga a V. Leiva, "A Type I Generalized Logistic Distribution: Solving Its Estimation Problems with a Bayesian Approach and Numerical Applications Based on Simulated and Engineering Data," *Symmetry*, roč. 14, č. 4, 2022, ISSN: 2073-8994. DOI: 10.3390/sym14040655. URL: https://www.mdpi.com/2073-8994/14/4/655.

[35] N. L. Johnson, S. Kotz a N. Balakrishnan, *Continuous Univariate Distributions, Volume 1* (Wiley Series in Probability and Statistics), en, 2. vyd. Nashville, TN: John Wiley & Sons, říj. 1994.

[36] P. Embrechts, C. Kluppelberg a T. Mikosch, *Modelling extremal events* (Stochastic Modelling and Applied Probability), en, 1. vyd. Berlin, Germany: Springer, čvn. 1997.

[37] A. L. A. Martins, G. R. Liska, L. A. Beijo, F. S. d. Menezes a M. Â. Cirillo, "Generalized Pareto distribution applied to the analysis of maximum rainfall events in Uruguaiana, RS, Brazil," en, *SN Appl. Sci.*, roč. 2, č. 9, zář. 2020.

[38] Wikipedia contributors, *Kernel density estimation — Wikipedia, The Free Encyclopedia*, [Online; accessed 22-March-2023], 2023. URL: https://en.wikipedia.org/w/index.php?title=Kernel_density_estimation&oldid=1136297214.

[39] J. P. Barrett, "The coefficient of determination—some limitations," *The American Statistician*, roč. 28, č. 1, s. 19–20, 1974.

[40] H. A. Sturges, "The choice of a class interval," *Journal of the american statistical association*, roč. 21, č. 153, s. 65–66, 1926.

[41] D. P. Doane, "Aesthetic frequency classifications," *The American Statistician*, roč. 30, č. 4, s. 181–183, 1976.

[42] WikiSkripta, *Senzitivita testu* —, [Online; navštíveno 14. 03. 2023], 2022. URL: https://www.wikiskripta.eu/index.php?title=Senzitivita_testu&oldid=457004.

[43] WikiSkripta, *Specificita testu* —, [Online; navštíveno 14. 03. 2023], 2022. URL: https://www.wikiskripta.eu/index.php?title=Specificita_testu&oldid=457003.