VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

APLIKACE PRO STATISTICKOU ANALÝZU ICS KOMUNIKACE

APPLICATION FOR STATISTICAL ANALYSIS OF ICS COMMUNICATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **ANDREA CHIMENTI**
AUTHOR

VEDOUCÍ PRÁCE **Ing. IVANA BURGETOVÁ, Ph.D.**
SUPERVISOR

BRNO 2022

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav informačních systémů (UIFS) Akademický rok 2021/2022

Zadání bakalářské práce

Student: **Chimenti Andrea**
Program: Informační technologie
Název: **Aplikace pro statistickou analýzu ICS komunikace**
Application for Statistical Analysis of ICS Communication
Kategorie: Data mining

Zadání:

1. Seznamte se s protokoly pro ICS komunikaci a po dohodě s vedoucí vyberte protokol (protokoly), na který se zaměříte.
2. Seznamte se s dostupnými datovými sadami obsahujícími komunikaci ve zvoleném protokolu.
3. Seznamte se se základními způsoby vizualizace dat a se základy statistického popisu dat.
4. Po dohodě s vedoucí navrhněte aplikaci pro statistickou analýzu a vizualizaci dat průmyslové komunikace využívající zvolený protokol.
5. Navrženou aplikaci implementujte a otestujte na dostupných datových sadách.
6. Zhodnoťte dosažené výsledky

Literatura:

- Matoušek, P.: Description and analysis of IEC 104 Protocol. FIT-TR-2017-12, Brno: Fakulta informačních technologií VUT v Brně, 2017.
- Skiena, S.S.: The Data Science Design Manual. Springer, 2017, 445 s. ISBN 978-3-319-55443-3

Pro udělení zápočtu za první semestr je požadováno:

- Body 1 až 4.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce: **Burgetová Ivana, Ing., Ph.D.**
Vedoucí ústavu: Kolář Dušan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2021
Datum odevzdání: 11. května 2022
Datum schválení: 11. října 2021

Abstrakt

Cílem této práce je návrh a implementace aplikace, sloužící ke statistické analýze síťového provozu v ICS (Industrial Control Systems) komunikaci. Práce se nejprve věnuje představení řídících systémů průmyslové komunikace a jejich nejrozšířenějších protokolů. Podrobně rozebraný je protokol IEC 104. Následuje představení základních metod popisné statistiky, pomocí kterých lze průmyslovou komunikaci analyzovat. V práci je využito několika datových sad ve formátu CSV, které zachycují záznamy průmyslové komunikace. Na těchto datových sadách je ukázáno využití statistických metod. Dále práce obsahuje návrh a popis implementace aplikace, díky které lze datové sady analyzovat a získat textový i grafický popis dat. Hlavním cílem aplikace je usnadnit uživateli hledání stabilních charakteristik, kterých lze využít k detekci anomálií a útoků. V závěru je použití aplikace demonstrováno na datových sadách obsahujících různé typy útoků.

Abstract

This work aims to present the design and implementation of an application for statistical analysis of network traffic in ICS (Industrial Control Systems) communication. In the first place, the work presents Industrial Control Systems and some of their most common protocols. The protocol IEC 104 is described in more detail. This is followed by an introduction to the basic methods of descriptive statistics, that can be used to analyze industrial communication. For this purpose, several CSV datasets, that capture fragments of industrial communication, have been used. These datasets are used to show how some of the previously described statistical methods can be used. The work then describes the implementation of an application, which allows to analyze the datasets and obtain various statistics and a visual representation of the data. The main objective of the application is to make it easier for the user to find stable characteristics that can be used for anomaly and attack detection. Finally, the benefits that the application brings are demonstrated on a set of datasets containing different types of attacks.

Klíčová slova

průmyslová komunikace, řídící systémy, SCADA systémy, IEC 60870-5-104, statistická analýza, míry polohy, míry variability, spojnicový graf, vektorizace, Python

Keywords

industrial communication, industrial control systems, SCADA systems, IEC 60870-5-104, statistical analysis, measures of position, measures of variability, line graph, vectorization, Python

Citace

CHIMENTI, Andrea. *Aplikace pro statistickou analýzu ICS komunikace.* Brno, 2022. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Ivana Burgetová, Ph.D.

Aplikace pro statistickou analýzu ICS komunikace

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením paní Ing. Ivany Burgetové, Ph.D. Uvedl jsem všechny literární prameny, publikace a další zdroje, ze kterých jsem čerpal.

.....
Andrea Chimenti
10. května 2022

Poděkování

Rád bych poděkoval své vedoucí práce paní Ing. Ivaně Burgetové, Ph.D. za odborné vedení práce, za mnohé hodiny strávené konzultacemi a za cenné informace, které mi byla vždy ochotna poskytnout. Dále bych chtěl poděkovat své rodině a svým nejbližším přátelům za veškerou podporu, kterou mi poskytovali po dobu celého studia.

Obsah

Kapitola 1

Úvod

„Data hýbou světem“ – tuhle větu už jistě ve svém životě slyšel každý. Pravdou je, že tzv. „data science“ (neboli česky „datová věda“) je v posledních letech rychle se rozvíjejícím oborem. Díky moderním technologiím a možnostem které nabízí, lze analyzovat čím dál větší množství dat. Díky tomu lze pracovat s vhledy a informacemi o datech, které by se v minulosti jen těžko získávaly.

Jednou z oblastí, ve kterých vzniká potřeba analýzy velkého množství dat, jsou průmyslové řídící systémy, jejichž popis je uveden v kapitole 2. Tyto systémy se většinou skládají z množství stanic, které mezí sebou různě komunikují. Je tedy v zájmu operátorů systémů mít přehled o tom, co se v komunikaci děje. Prohlížení surových záznamů ve formě `csv` nebo `pcap` souborů však nepřináší požadované výsledky. Z tohoto důvodu představuje práce řešení, díky kterému lze jednoduše získat množství statistik a vizualizovat zkoumaná data.

Jedním se základních kamenů datové analýzy je popisná statistika. Kapitola 3 se proto zabývá jejími základními koncepty. V kapitole jsou uvedeny základní charakteristiky polohy, charakteristiky variability a některé metody pro vizualizaci dat.

Pro účely této práce byly využity datové sady, obsahující záznamy průmyslové komunikace, které vznikly na *Fakultě elektrotechniky a komunikačních technologií* a *Fakultě informačních technologií* spadající pod *Vysoké učení technické v Brně*. Popis datových sad a navržených metod k jejich analýze je uveden v kapitole 4.

V kapitole 5 je uveden návrh aplikace, která umožňuje uživateli (operátorovi) provést analýzu průmyslové komunikace bez nutnosti hlubokých znalostí programování apod. Aplikace poskytuje uživateli popis komunikačního profilu zkoumané datové sady a umožňuje mu nalézt v komunikaci stabilní hodnoty. Získané informace mohou být využity k hledání anomálií nebo útoků. Popis implementace aplikace, její architektura a uživatelské pohledy, které nabízí jsou uvedeny v kapitole 6.

Práce je zakončena experimentální částí v kapitole 7, kde je demonstrováno vyhledání stabilních hodnot v komunikaci a jejich následné použití pro detekci anomálií.

Kapitola 2

Industrial Control Systems

Termín *Industrial control systems* (dále jen „ICS“) souhrnně označuje různé typy průmyslových řídících systémů a prostředky pro jejich realizaci. ICS se skládají z řídicích prvků (např. elektrických, mechanických, hydraulických, pneumatických atd.), které jsou navzájem propojeny a řízeny tak, aby umožnily dosažení průmyslových cílů. Řízení může být plně automatizované, nebo může být ovládáno operátorem. ICS se obvykle používají v energetice, vodárenství, chemickém průmyslu, dopravě, a dalších průmyslových odvětvích. Může se jednat např. o systémy pro správu průtoku kapalin potrubím v ropné rafinerii, pro monitorování stavu elektrické rozvodové sítě apod. Jednotlivé druhy systémů se liší architekturou a možnostmi škálovatelnosti. Finální podoba systémů závisí na cílové průmyslové oblasti a na způsobu využití daného systému. Mezi nejpoužívanější systémy patří SCADA (Supervisory Control And Data Acquisition), které budou podrobněji představeny v podkapitole 2.1, a DCS (Distributed Control Systems).

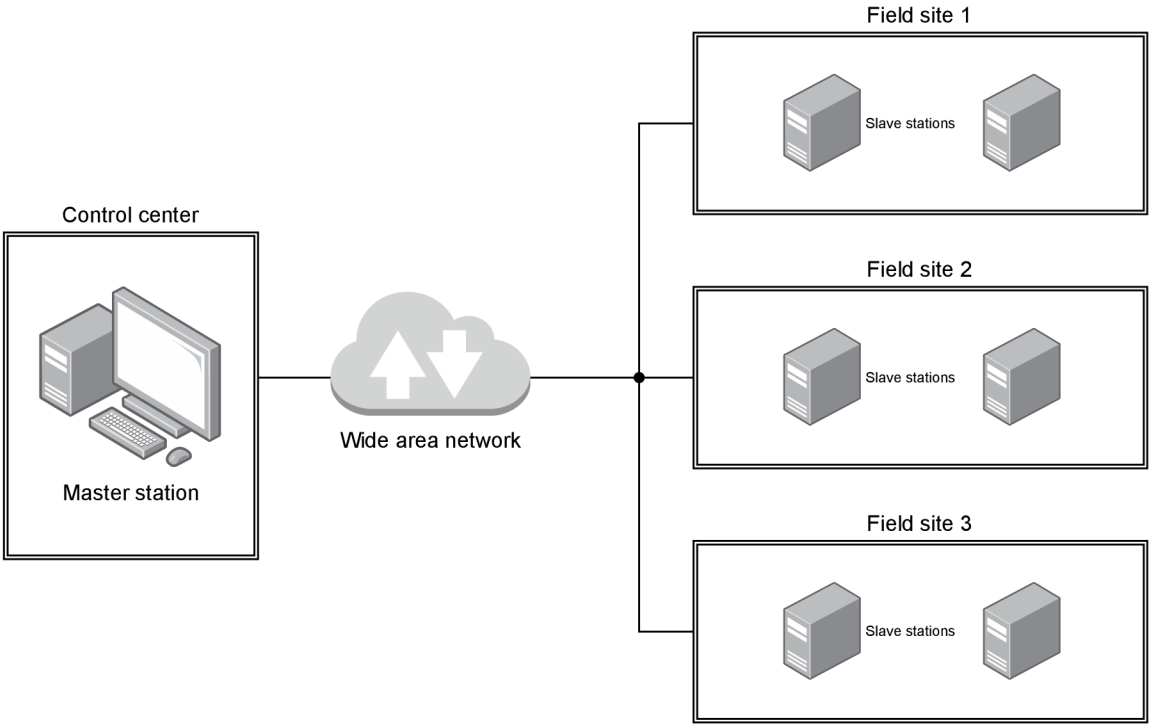
Aby bylo možné systémy běžně využívat v praxi, musí být jasně definován jejich komunikační profil. Za tímto účelem vzniklo několik standardů jako např. IEC 60870-5 nebo DNP3. V podkapitole 2.2 bude představen protokol IEC 60870-5-104, definující komunikační profil SCADA systémů, který k přenosu dat využívá transportní vrstvu TCP/IP.

Pro průmyslové řídící systémy je typické, že vznikly jako softwarová nadstavba nad fyzickými systémy, umožňující monitorování a ovládání fyzických prvků. Díky zvyšující se míře zakomponování informačních technologií do průmyslových odvětví, došlo ke vzniku mnoha „chytrých“ systémů (např. chytré budovy, chytré továrny, ...) využívajících ICS. Kromě benefitů, ve formě vyšší efektivity, konektivity atd., došlo zároveň k nárůstu počtu rizik, které s sebou použití řídících systémů obnáší. Možnosti zneužití některých druhů zranitelnosti jsou rozebrány v podkapitole 2.3 [14].

2.1 SCADA systémy

Supervisory Control And Data Acquisition (dále jen „SCADA“) systémy jsou určeny k centralizovanému sběru a monitorování dat ze zařízení, která se nacházejí v různých geografických oblastech. Monitorování a ovladání zařízení se provádí pomocí centrální řídící stanice. Agregovaná data se v reálném čase zobrazují textově nebo graficky operátorovi, který může celý systém monitorovat a vzdáleně reagovat na případné události. Dle nastavení systému se některé změny mohou provádět i automaticky. Své uplatnění nacházejí SCADA systémy např. v systémech pro distribuci plynu, sběr odpadních vod, koordinaci městské hromadné dopravy atd.

Typická architektura SCADA systému zahrnuje jednu řídící stanici a několik koncových zařízení. Řídící stanice poskytuje výpočetní zdroje pro příjem a zpracování dat a poskytuje operátorovi potřebné informace a ovládací prvky. Koncová zařízení sbírají svá lokální data a komunikují s řídící stanicí. Řídící stanice si může data od koncových zařízení vyžádat a zároveň může odesílat kontrolní příkazy. Komunikace řídící stanice a koncových zařízení může být založena na různých komunikačních kanálech, u nově vzniklých systémů se většinou využívá TCP/IP protokolu [14].



Obrázek 2.1: Obvykle používáná architektura systémů SCADA. Řídící stanice je umístěna v kontrolním centru a podřízené stanice jsou rozprostřeny do několika fyzicky i logicky oddělených oblastí.

2.1.1 Funkce SCADA systémů

SCADA systémy nabízejí velké množství monitorovacích a kontrolních funkcí. V následujícím výčtu jsou uvedeny ty nejdůležitější z nich. Převzato z [10].

- Zobrazení stavu technologických procesů v reálném čase.
- Vzdálené řízení technologických procesů pomocí řídící stanice.
- Možnost přímých zásahů do technologických procesů operátory.
- Automatické řízení technologických procesů za účelem zvýšení efektivity.
- Pravidelné generování provozních zpráv.
- Grafické zobrazení údajů o technologickém procesu za účelem vypracování efektivních provozních strategií.

2.1.2 Komunikační standardy SCADA systémů

Koncem 90. let došlo ke vzniku dvou otevřených komunikačních standardů, známým jako DNP3 a IEC 60870-5, které definují podobu komunikace mezi stanicemi v rámci SCADA systémů. Ta zahrnuje získávání informací a odesílání kontrolních příkazů mezi fyzicky vzdálenými zařízeními. Standardy se vyznačují možností spolehlivého přenosu relativně malých paketů v pevně daném pořadí. V tomto ohledu se liší od více obecně založených protokolů jako je FTP nebo TCP/IP, které nejsou příliš vhodné pro SCADA systémy. Vzhledem ke společnému základu protokolů je funkcionalita na nižších vrstvách podobná. Ve vyšších vrstvách však pracují odlišně [2].

IEC 60870-5 byl vytvořen Mezinárodní elektrotechnickou komisí (dále jen „IEC“, z anglické zkratky pro *International Electrotechnical Commission*). Standard je primárně využíván v systémech pro správu elektrické přenosové soustavy a to hlavně v evropských zemích [2]. V kapitole 2.2 bude podrobněji rozebrán protokol IEC 60870-5-104, patřící do této rodiny.

DNP3, celým názvem *Distributed Network Protocol 3*, byl navržen společností *Harris* a určen primárně pro použití v energetickém průmyslu. Na rozdíl od protokolu IEC 60870-5, je však využíván i v jiných odvětvích jako např. plynárenství, zpracování odpadních vod apod. Protokol má velkou podporu v Severní a Jižní Americe, Asii, Jižní Africe a Austrálii [2]. Práce se dále tímto standardem zabývat nebude.

2.2 Protokol IEC 60870-5-104

Protokol IEC 60870-5-104 (dále jen „IEC 104“) poskytuje komunikační profil pro zasílání základních telekomunikačních zpráv mezi řídící stanicí a podřízenými stanicemi u SCADA systémů. Vznikl jako nástupce protokolu IEC 60870-5-101 (dále jen „IEC 101“), oproti kterému se liší ve způsobu přenosu informací na nižších vrstvách. Kombinuje aplikační vrstvu protokolu IEC 101 s transportní vrstvou TCP/IP architektury [2]. Následující obsah kapitoly se zabývá vlastnostmi aplikační vrstvy a informace v ní obsažené jsou platné pro oba protokoly [8].

2.2.1 Komunikační profil

Základním předpokladem je rozdělení stanic do hierarchie, kde každá stanice má jasně danou pozici a práva.

Typy stanic

Typ stanice určuje pozici stanice v hierarchii systému.

- **Řídící stanice:** stanice, ze které jsou posílány příkazy podřízeným stanicím. Typicky se jedná o osobní počítač se SCADA systémem, se kterým přímo interaguje operátor systému. Řídící stanice obvykle komunikuje na portu 2404. Někdy je také označována jako *master* nebo *centrální stanice*.
- **Podřízená stanice:** stanice, která je ovládána řídící stanicí. Může se jednat o senzor, hydraulický lis, turbínu apod. Někdy je také označována jako *slave* nebo *koncová stanice*.

Režimy přenosu

Režim přenosu udává, které stanice mají právo iniciovat komunikaci.

- **Nevyvážený přenos:** řídící stanice reguluje datový provoz podáváním dotazů podřízeným stanicím. Iniciuje všechny přenosy zpráv, zatímco podřízené stanice na tyto zprávy pouze odpovídají.
- **Vyvážený přenos:** jakákoliv stanice má právo iniciovat přenos zprávy. Stanice mohou současně vystupovat jako řídicí a podřízené.

Směry komunikace

Směr komunikace udává, odkud kam se přenáší data.

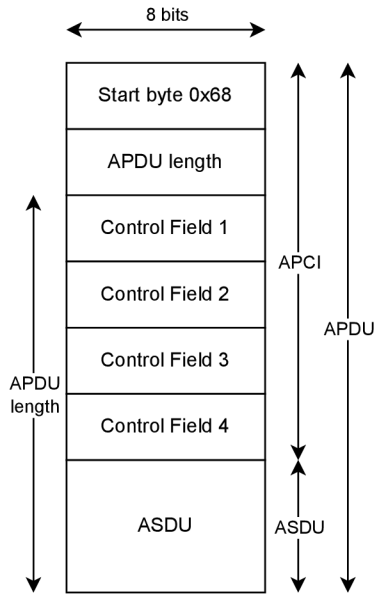
- **Monitorovací směr:** od podřízené stanice k řídící.
- **Kontrolní směr:** od řídící stanice k podřízené.
- **Obrácený směr:** stav, kdy podřízená stanice posílá dotazy a řídící stanice odpovídá daty.

2.2.2 Struktura aplikačních dat

Aplikační data protokolu IEC 104 tvoří APDU (Application Protocol Data Unit) jednotka, dělící se dále na APCI (Application Protocol Control Information) a ASDU (Application Service Data Unit), viz ilustrace 2.2.

APCI jednotka začíná 8 bity s fixní hodnotou `0x68`, za kterými následuje 8 bitů, které udávají délku APDU. Následují čtyři 8 bitové kontrolní pole (*control fields*).

ASDU jednotka obsahuje 48 bitů pro identifikaci dat a až 127 datových objektů, sloužící k uchování užitečných dat. Některé položky, které slouží k identifikaci jsou dále popsány v této kapitole [8].



Obrázek 2.2: APDU jednotka [8].

Formát APCI jednotky

Protokol IEC 104 definuje celkově 3 formáty, které udávají podobu kontrolních polí. I, S a U formát. S a I formáty ukládají sekvenční čísla odeslaných zpráv. Pokud je tento čítač neplatný, pak je spojení ukončeno. Možnost zneužití tohoto chování bude popsána v podkapitole 2.3. ASDU je přítomna pouze v případě, že je použit I-formát.

- **I-formát:** z angl. výrazu *Information transfer format*. Slouží k číslovanému přenosu informací mezi řídicí a podřízenou stanicí.
- **S-formát:** z angl. výrazu *numbered supervisory functions*. Používá se k provádění očíslovaných kontrolních funkcí.
- **U-formát:** z angl. výrazu *unnumbered control functions*. Slouží k provádění nečíslovaných řídicích funkcí.

Položky ASDU

Poznámka: V následujícím výčtu jsou uvedeny pouze položky relevantní pro tuhle práci.

- **Typ:** Udává typ přenášených dat.
- **COT:** Důvod přenosu, z angl. zkratky pro *cause of transmission*. Používá se při interpretaci dat cílovou stanicí. Každý typ ASDU má definovanou podmnožinu platných COT kódů, které jsou pro něj smysluplné.
- **COA:** Společná ASDU adresa, z angl. zkratky pro *common ASDU address*. Je asociovaná se všemi datovými objekty v ASDU bloku. Většinou je používána pro identifikaci celé stanice. Alternativní možností je identifikace sektoru stanice, při které lze stanici rozdělit na několik logických částí.
- **IOA:** Adresa informačního objektu, z angl. zkratky pro *Information object address*. Slouží k identifikaci konkrétních dat konkrétní stanice.

2.3 Útoky na SCADA systémy

Přestože je protokol IEC 104 široce využíván, zabezpečení nebylo při vzniku prioritou. Protokol postrádá důležité bezpečnostní prvky jako je šifrování, ochrana integrity nebo autentizace. V této podkapitole budou představeny některé kybernetické útoky, vůči kterým nemá protokol vestavěnou ochranu. Následující výčet útoků čerpá z [5].

Neautentizovaný přístup

Protokol nedisponuje mechanismem ověření identity. Útočník se může připojit ke koncové stanici a bez nutnosti autentizace začít odesílat kontrolní příkazy. Útočník může například odeslat dotazovací příkaz a zjistit stanicí používané hodnoty IOA (viz 2.2.2) . Typický průběh útoku vypadá následovně:

1. Zjištění IP adresy a portu koncové stanice.
2. Vydávání se za řídící stanici a odeslání požadavku na připojení koncové stanici.

3. Kvůli chybějící autentizaci, dojde ze strany koncové stanice k přijetí požadavku.
4. Odeslání libovolných kontrolních příkazů koncové stanici.

Manipulace sekvenčních čísel APCI

Pokud stanice obdrží paket s neočekávaným sekvenčním číslem v jednotce APCI, dojde k ukončení spojení. Útočník může zneužít této vlastnosti a následujícími kroky zapříčinit výpadek systému (tzv. *Denial of Service*):

1. Vložení se do komunikace mezi stanicemi. Z útočníka se stane tzv. *Man in the middle*.
2. Odchycení příchozích paketů.
3. Modifikace paketů. Konkrétně změna pořadového čísla APCI.
4. Výpočet nového *TCP checksum*.
5. Navrácení paketů do původní komunikace.

Manipulace TCP toku

Komunikace mezi řídící a koncovou stanicí probíhá v jednom TCP proudu, který je aktivně udržován otevřený. Útočník, který vystupuje jako *Man in the middle*, může využít zranitelnosti TCP a způsobit ukončení komunikace, vedoucí k nesprávné funkci nebo dokonce výpadku systému. Tento typ útoku se liší od 2.3 pouze ve třetím bodě. V tomto případě mohou být k modifikaci paketů použity následující metody:

- Sekvenční číslo TCP toku se změní na neočekávanou hodnotu.
- Synchronizační příznak se nastaví na hodnotu `FIN`[1].

Záplava pakety s příznakem SYN

Jedná se o běžný kybernetický útok, v anglickém jazyce je znám pod pojmem *SYN Flood*. Během útoku se útočník snaží zahltit cílový systém velkým množstvím paketů se synchronizačním příznakem nastaveným na hodnotu `SYN`[2]. Cílové stanice jsou zahlceny množstvím nových falešných žádostí o spojení a nestíhají odbavovat legitimní požadavky. Dochází k znatelnému zpomalení systému [12].

Vkládání podvržených paketů

Jedná se o pokročilou formu útoku, která umožňuje útočníkovi v pozici *Man in the middle* vkládat do komunikace pakety s libovolnými hodnotami ASDU. Útočník tak může převzít ovládání nad celou energetickou sítí. K úspěšnému provedení útoku musí útočník správně upravit sekvenční čísla (APCI i TCP) všech paketů. V případě, že by tak neučinil, došlo by k ukončení spojení. Detekce takového útoku může být velmi obtížná, jelikož nedochází k přerušení spojení a navíc může útočník podvrhnout data, která se budou zobrazovat operátorovi. Typický průběh útoku:

[1] Příznak `FIN` se používá u TCP komunikace k označení jejího konce.
[2] Příznak `SYN` se používá u TCP komunikace k vyžádání nového spojení.

1. Vložení se do komunikace mezi stanicemi. Z útočníka se stane tzv. Man in the middle.
2. Odchycení příchozích paketů a učení se APCI a TCP sekvencí.
3. Vložení podvrženého paketu s libovolnou hodnotou ASDU a vypozorovanými sekvenčními čísly do komunikace.
4. Sledování a udržování vnitřního stavu sekvenčních čísel.
5. Zachycení všech zbylých paketů a opravení sekvenčních čísel tak, aby spojení nebylo přerušeno.

Kapitola 3

Statistický popis dat

Statistickým popisem dat se zabývá tzv. *deskriptivní statistika*. Jedná se o disciplínu, která kvantitativně popisuje hlavní vlastnosti souboru dat a snaží se numerickým nebo grafickým popisem vystihnout podstatné informace o daných datech. Umožňuje uživateli nalézt rozložení hodnot atributů a tím pádem lépe pochopit zkoumaná data. Předmětem zkoumání deskriptivní statistiky je statistický soubor dat, který může reprezentovat buď celou populaci, nebo pouze její část (tzv. *výběrový soubor*).

Numerický popis dat se člení na míry polohy (viz 3.1) a míry variability (viz 3.2). Mezi míry polohy patří např. modus, medián nebo střední hodnota. Mezi míry variability patří např. rozptyl, směrodatná odchylka, minimum, maximum apod. Grafický popis dat nabízí velké množství metod zobrazení, z nichž nejčastěji používané jsou např. spojnicové grafy, histogramy, koláčové grafy apod. Vybrané metody budou popsány v podkapitole 3.3.

3.1 Míry polohy

Míry polohy vyjadřují „kde“ se data nacházejí. Charakterizují „střed“ datového souboru, kolem kterého se hodnoty pohybují. Mají důležitou výpovědní hodnotu o tom, jak data vypadají a jak se chovají. Mohou nabývat i takových hodnot, které se v datovém souboru přímo nenacházejí, například průměrná hodnota v souboru přirozených čísel může být racionálním číslem, které není přirozené. Pro charakteristiky polohy výběrového souboru platí, že pouze odhadují skutečné hodnoty celé populace [18].

Střední hodnota

Střední hodnota datového souboru je nejzákladnější charakteristikou polohy. Označuje se řeckým písmenem μ a je totožná s aritmetickým průměrem všech prvků souboru. Význam střední hodnoty je největší u datových souborů se symetrickým rozdělením dat a s nízkým výskytem odlehlých hodnot. V takovém případě by měla rozdělovat soubor na dva přibližně stejně velké celky. Tahle vlastnost je negativně ovlivňována výskytem odlehlých hodnot, které mají schopnost výrazně vychýlit střední hodnotu. Střední hodnota se počítá jako součet všech hodnot vydělených jejich počtem [13]:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{3.1}$$

kde:

N je celkový počet prvků v datovém souboru.

x_i je i-tý prvek datového souboru.

Důležitou vlastností průměru (a tím pádem i střední hodnoty) je „nevychýlenost“, která zajišťuje, že průměr všech možných výběrových průměrů, počítaný přes všechny výběrové soubory dané velikosti, je roven průměru celé populace. Díky této vlastnosti je zajištěno, že průměr výběrového souboru dobře aproximuje průměr celé populace a tudíž je jeho výpovědní hodnota relevantní i přes neúplnost dat [16].

Medián

Medián je přesným středem datové sady. Rozděluje sadu na dvě stejně velké části, kde počet prvků s větší hodnotou než medián je stejný jako počet prvků s hodnotou menší. V případě lichého počtu prvků v datové sadě je mediánem „prostřední“ prvek seřazené datové sady 3.2. Pro sady se sudým počtem prvků je mediánem průměr dvou „prostředních“ prvků seřazené datové sady 3.3. Čím více je distribuce dat symetrická, tím leží medián blíže střední hodnotě a přidaná výpovědní hodnota není příliš vysoká. U datových sad s asymetrickým rozložením může však vzdálenost a vzájemná poloha mediánu a střední hodnoty napovědět hodně o tom, jak vypadá distribuce a jak moc je asymetrická. Výrazně se lišící hodnoty navíc naznačují, že se v datové sadě nachází mnoho odlehlých hodnot, které ovlivňují střední hodnotu [13].

$$Med = x_{\frac{N+1}{2}} \qquad \text{pro lichá } N \tag{3.2}$$

$$Med = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} \qquad \text{pro sudá } N \tag{3.3}$$

kde:

N je celkový počet prvků v datovém souboru.

x_n je n-tý prvek seřazeného datového souboru.

Modus

Modus je nejčastější hodnota, která se vyskytuje v datové sadě. Při jeho výpočtu se většinou datová sada rozdělí do intervalů, ve kterých se nalezne ten nejvíce zastoupený a jako modus se použije jeho střed krajních hodnot. Často je využíván k detekci chyb nebo anomálií (v datové sadě může být například nejčastější hodnotou výchozí nebo chybová hodnota). Datové sady, pro které modus nabývá více hodnot, označujeme jako *multimodální*. Multimodalita může naznačovat, že v datový soubor vznikl smícháním dvou nebo více unimodálních populací [16].

Q-Kvantily

q-kvantily rozdělují seřazený soubor dat na q přibližně stejně velkých částí. k-tý q-kvantil je hodnota Q_k náhodné veličiny X, pro kterou za předpokladu že:

$$0 < k < q \wedge k, q \in \mathbb{N}$$

platí:

$$P(X \leq Q_k) \geq \frac{k}{q} \wedge P(X \geq Q_k) \geq 1 - \frac{k}{q} \tag{3.4}$$

Pro následující hodnoty q se q-kvantily označují jako:

- $q = 2$: medián
- $q = 3$: tercil
- $q = 4$: kvartil
- $q = 5$: kvintil
- $q = 10$: decil
- $q = 100$: percentil

3.2 Míry variability

Míry variability zkoumají, jak jsou prvky v datovém souboru vzájemně blízké či vzdálené. Hodnotí rozptýlenost hodnot statistického souboru kolem nějaké střední hodnoty. Pro charakteristiky variability výběrového souboru opět platí, že pouze odhadují skutečné hodnoty celé populace [19].

Směrodatná odchylka a rozptyl

Nejběžnějším měřítkem variability je rozptyl, který se označuje symbolem σ^2. Udává, jak moc jsou data soustředěna kolem střední hodnoty. Nízký rozptyl naznačuje, že datové body mají tendenci být těsně seskupeny kolem středu. Vysoký rozptyl naopak znamená, že mají tendenci se od středu vzdalovat. Kromě rozptylu se často používá i směrodatná odchylka, která je druhou odmocninou rozptylu a označuje se symbolem σ. Pro výpočet rozptylu se používá součet čtverců odchylek jednotlivých prvků od střední hodnoty [4]:

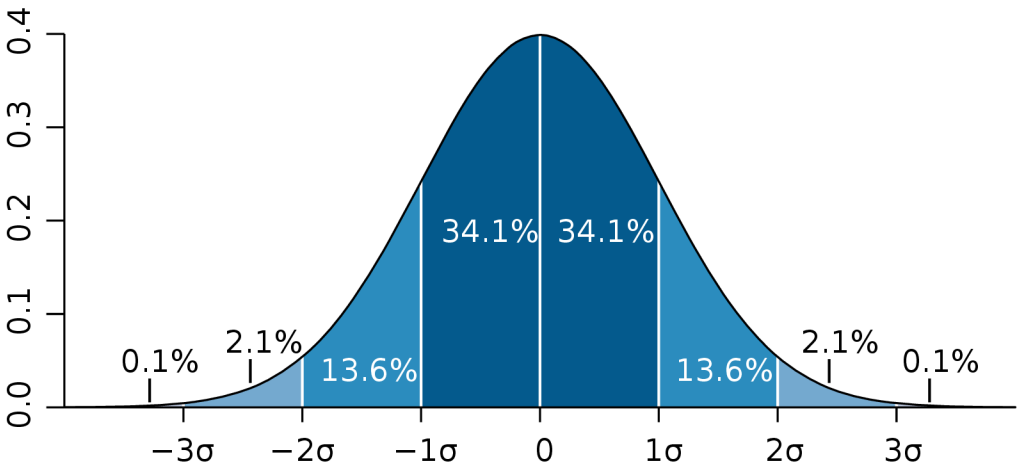
$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{3.5}$$

Pro výpočet směrodatné odchylky poté použijeme druhou odmocninu rozptylu:

$$\sigma = \sqrt{\sigma^2} \tag{3.6}$$

Pro datový soubor s distribucí, která se alespoň přibližně řídí normálním rozdělením, lze pomocí směrodatné odchylky odhadnout podíly hodnot, které spadají do určitých intervalů. Přibližné rozložení dat je ukázáno na obrázku 3.1 a platí pro něj platí následující vlastnosti [4]:

- V intervalu $\langle \mu - \sigma, \mu + \sigma \rangle$ leží přibližně 68 % hodnot.
- V intervalu $\langle \mu - 2\sigma, \mu + 2\sigma \rangle$ leží přibližně 95 % hodnot.
- V intervalu $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ leží přibližně 99,7 % hodnot.



Obrázek 3.1: Ukázka rozložení dat u normálního rozdělení [15].

Variační šíře

Udává rozdíl mezi největší a nejmenší hodnotou v datovém souboru. Je přímo ovlivněna extrémními hodnotami. Nejedná se o interval, ale o číslo udávající šíři intervalu. Výpočet vypadá následovně [19]:

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

kde:

x_n je n-tý prvek seřazeného datového souboru.

Mezikvartilové rozpětí

Mezikvartilové rozpětí (angl. *Interquartile range*, zkratka IQR) udává šířku oblasti, ve které leží středních 50 % hodnot. Nabízí uživateli robustní alternativu ke klasickému rozptylu, která není ovlivňována odlehlými hodnotami. Počítá se jako rozdíl mezi třetím a prvním kvartilem (tedy mezi 75. a 25. percentilem) [4]:

$$IQR = q_3 - q_1 = p_{75} - p_{25} \tag{3.8}$$

kde:

q_n je n-tý kvantil.

p_n je n-tý percentil.

Pomocí mezikvartilového rozpětí lze provést detekci odlehlých hodnot. Jelikož jimi není ovlivňováno, snižuje se oproti jiným metodám pravděpodobnost selhání. Metoda navíc nepředpokládá normální rozdělení dat. K detekci se používají tzv. brány, které jsou definovány následujícím způsobem [3]:

$$lower\_gate = q_1 - 1.5 \cdot IQR \tag{3.9}$$

$$upper\_gate = q_3 + 1.5 \cdot IQR \tag{3.10}$$

Pro odlehlé hodnoty x potom platí:

$$x < lower\_gate \vee x > upper\_gate \tag{3.11}$$

3.3 Vizualizační metody popisu dat

Vizualizační metody jsou důležitou součástí popisné statistiky, která efektivně ukazuje kvantitativní vlastnosti datových souborů jako je rozložení hodnot jejich atributů. Může se jednat o přehledové tabulky, různé druhy grafů, histogramy apod. Mezi nejdůležitější aspekty grafického popisu dat patří následující [13]:

- **Explorační funkce:** usnadňuje uživateli orientaci v datech. Pro uživatele je snazší pochopit chování a distribuci dat.
- **Detekce chyb:** usnadňuje uživateli nalezení chyb a anomálii v datovém souboru. Může se jednat o špatně naměřené hodnoty, nesprávně načtená data atd.
- **Komunikační funkce:** umožňuje snadno data prezentovat ostatním uživatelům, kteří nemají hluboký vhled do aplikační domény a neznají kontext vzniku datového souboru.

Histogram

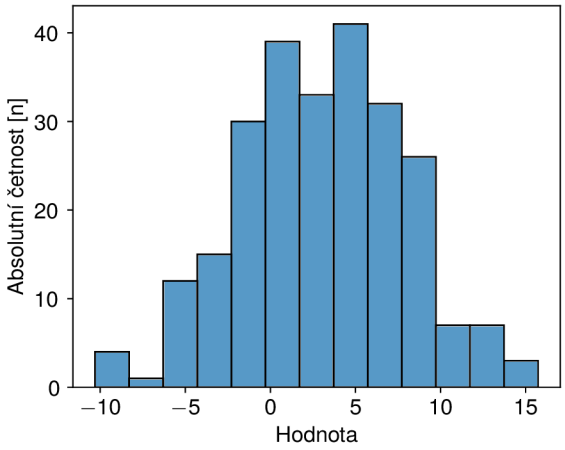
Histogram je sloupcový graf, který znázorňuje absolutní nebo relativní četnosti hodnot nějakého atributu datové sady [16]. Příklad histogramu je uveden na obrázku 3.2. Na vodorovnou osu jsou vyneseny hodnoty atributu a na svislou osu (relativní) četnosti. Někdy je histogram doplněn o křivku znázorňující odhad funkce hustoty pravděpodobnosti (viz obrázek 3.2b). Další častá úprava histogramu spočívá v tom, že namísto vynesení samostatného sloupce pro každou hodnotu atributu, jsou hodnoty atributů rozděleny do intervalů (tříd) a sloupce jsou vyneseny až pro tyhle intervaly. V takovém případě je důležité zvolit správnou velikost intervalu. Příliš malé intervaly mohou způsobit, že hodnot které do nich spadají, bude příliš málo, nebo že šum bude rušit čitelnost a přehlednost histogramu. Naopak příliš velké intervaly mohou způsobit ztrátu informace o rozdělení, podle kterého se data řídí [13].

Liniový graf

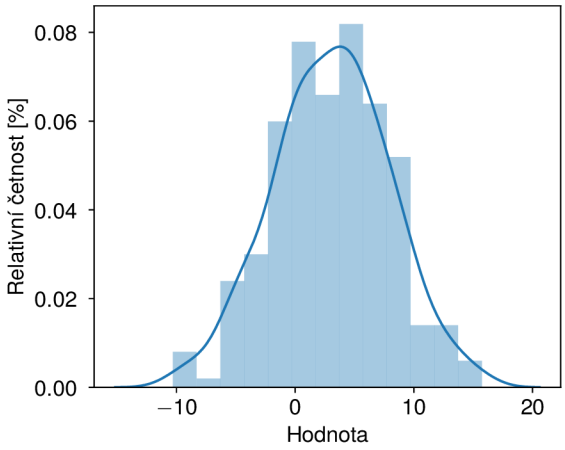
Liniový (nebo také spojnicový) graf je jedním z nejlepších způsobů zobrazení vývoje zkoumaného atributu v čase. Používá se pro časové řady atributů, u kterých hodnota atributu v jednom časovém okamžiku přímo souvisí s hodnotou atributu v následujícím časovém okamžiku [6]. Při tvorbě spojnicového grafu je potřeba věnovat zvýšenou pozornost volbě měřítka svislé osy. Většina nástrojů pro tvorbu spojnicových grafů automaticky volí jako interval svislé osy extrémní hodnoty zobrazovaných dat. V některých případech může docházet k zanedbání kontextu datového souboru a špatné interpretaci dat uživatelem [13]. Ukázka použití liniového grafu je popsána v podkapitole 4.4.1.

Tabulka

Tabulka je nejpřesnějším způsobem zobrazení dat. Na rozdíl od jiných grafických metod nedochází u tabulek ke ztrátě informací nepřesným vykreslením, zaokrouhlením apod. a umož-



(a) Histogram ukazující absolutní četnost hodnot.



(b) Histogram ukazující relativní četnost hodnot a odhad funkce hustoty pravděpodobnosti.

Obrázek 3.2: Příklad histogramů ukazující absolutní a relativní četnost hodnot datové sady. Datová sada obsahuje 500 vzorků, které byly vygenerovány z normálního rozdělení s parametry $\mu = 3$, $\sigma = 5$.

ňují tak v dalším zpracování využít přesné hodnoty dat. Dále vynikají v zobrazení vysoce multidimenzionálních dat, u kterých klasické vizualizační metody selhávají [13].

Koláčový graf

Koláčový graf je ideálním nástrojem pro zobrazení poměru kategorických atributů. Je založen na principu rozdělení kruhu do oblastí, které svou velikostí poměrově odpovídají poměru zastoupení jednotlivých kategorií daného atributu v sadě dat. Vždy zobrazuje pouze hodnoty jednoho atributu a je vhodný pro porovnávání. Většinou lze plnohodnotně nahradit sloupcovým grafem [13]. Příklad koláčového grafu je uveden na obrázku 3.3.

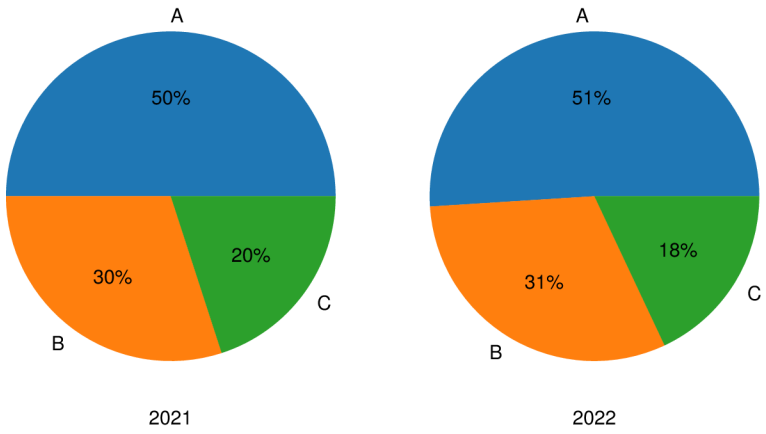
Krabicový graf

Krabicový graf, je graf ve tvaru obdélníku doplněný o tzv. *vousy*, který zobrazuje významné kvantily daného atributu (viz obrázek 3.4). Uvnitř obdélníkového tvaru je čarou naznačena pozice mediánu, samotný obdélník značí polohu prvního a třetího kvartilu. Délka obdélníku tedy odpovídá hodnotě mezikvartilového rozpětí a ohraničuje 50 % hodnot. Vousy dosahující za hranice obdélníkového tvaru pak většinou signalizují polohu minima a maxima [11].

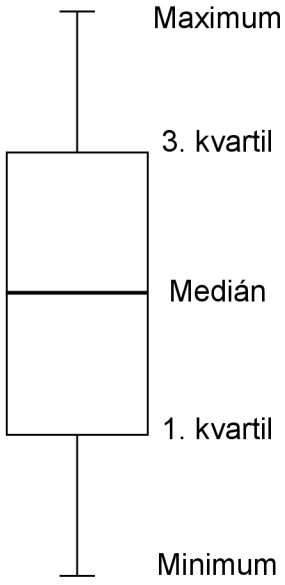
Houslový graf

Houslový graf zobrazuje podobně jako krabicový graf významné kvantily daného atributu. Navíc ale přidává i zobrazení odhadu hustoty pravděpodobnosti. Je vhodný zejména pro data, která se řídí multimodální distribucí[1], protože uživatel získá oproti krabicovému grafu výrazně lepší představu o opravdovém rozložení dat [7]. Porovnání krabicového a houslového grafu je vyobrazeno na obrázku 3.5.

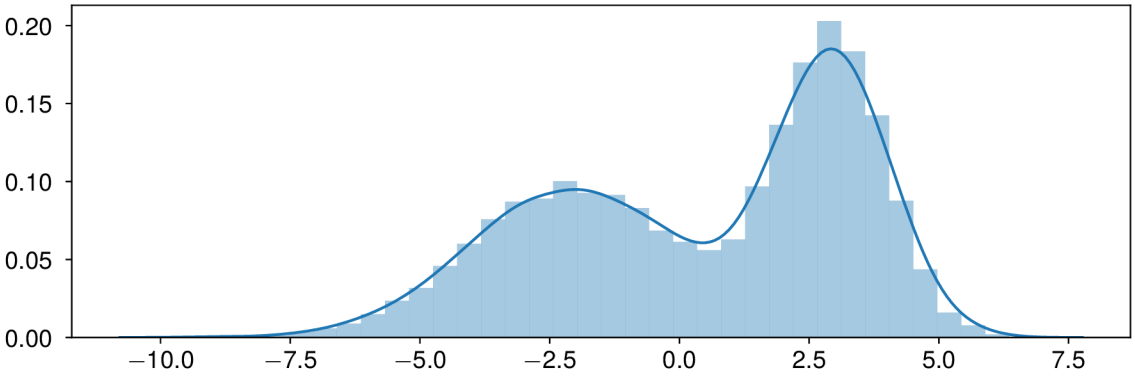
[1] Distribuce, která má dva nebo více lokálních maxim.



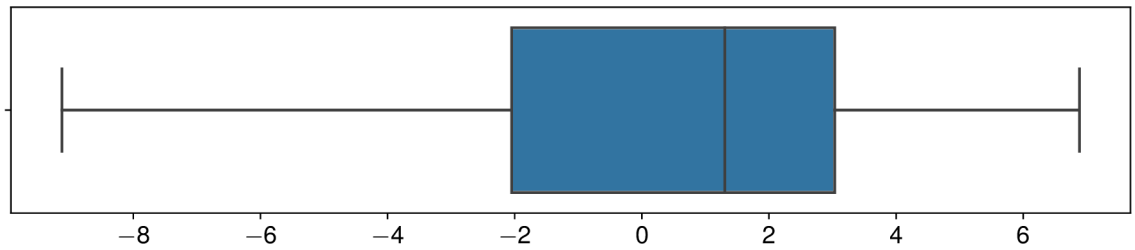
Obrázek 3.3: Ukázka dvou koláčových grafů, které znázorňují různé poměry zastoupení kategorického atributu v datech mezi lety 2021 a 2022. Kategorický atribut nabývá hodnot *A*, *B* a *C*. Obrázek znázorňuje, že porovnání poměrů je pro uživatele velmi snadné i přesto, že rozdíl mezi poměry je minimální.



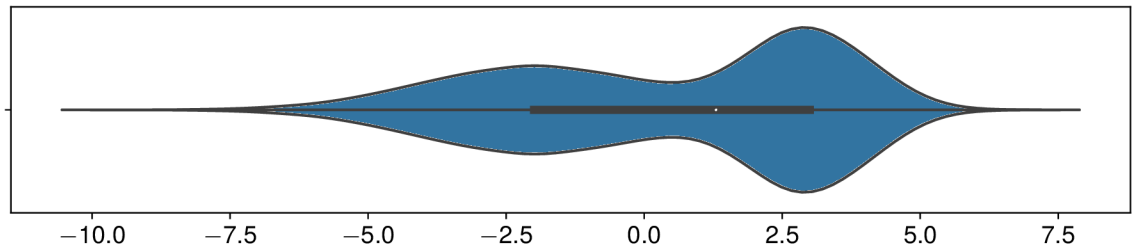
Obrázek 3.4: Důležité kvantily, které lze pomocí krabicového grafu zkoumat.



(a) Histogram dat řídící se bimodální distribucí. Distribuce vznikla spojením dvou normálních rozdělení s parametry $\mu = -2$, $\sigma = 2$ a $\mu = 3$, $\sigma = 1$. Křivka znázorňuje odhad hustoty pravděpodobnosti.



(b) Z krabicového grafu nelze poznat, jakým rozdělením se data opravdu řídí. Uživatel se může mylně domnívat, že se data řídí unimodálním normálním rozdělením.



(c) Houslový graf ukazuje i odhad hustoty pravděpodobnosti zkoumaných dat. Uživatel tak získává lepší představu o podobě rozdělení, kterým se data řídí.

Obrázek 3.5: Porovnání krabicového a houslového grafu na datové sadě řídící se bimodální distribucí. Je patrné, že houslový graf značně lépe vystihuje opravdové chování dat. Zdrojový kód, ze kterého čerpá obrázek inspiraci, je dostupný z [7].

Kapitola 4

Analýza datových sad

Práce se bude zabývat datovými sadami, které vznikly na *Fakultě elektrotechniky a komunikačních technologií* a *Fakultě informačních technologií* spadající pod *Vysoké učení technické v Brně*. Sady jsou veřejně dostupné na GitHubu v repositáři [9]. Celkově se jedná o 10 datových sad, z nichž 4 zachycují běžný provoz (viz 4.1) a zbylých 6 zachycuje provoz, ve kterém se vyskytuje útok (viz 4.2). Sady jsou ve formátu CSV a obsahují záznamy komunikací ve SCADA systémech využívajících protokolu IEC 104 (viz 2.2). Každý řádek (záznam) v CSV souborech reprezentuje jeden zachycený paket a každý sloupec jeden atribut komunikace. Atributy jsou podrobněji popsány v podkapitole 4.3. V poslední podkapitole 4.4 jsou pak popsány metody pro analýzu sad.

4.1 Datové sady s běžným provozem

Datové sady s běžným provozem obsahují záznamy komunikace, která probíhala standardním způsobem a nebyla ovlivněna žádnými útoky nebo anomáliemi. Základní vlastnosti sad jsou popsány v tabulce 4.1. Pro zjednodušení orientace v textu, bude před představením datových sad zavedeno jejich značení písmeny z latinské abecedy. V práci budou dále užívána pouze značení definována ve výčtu:

- **A:** `mega104-14-12-18-ioa.csv`
- **B:** `mega104-17-12-18-ioa.csv`
- **C:** `10122018-104Mega-ioa.csv`
- **D:** `13122018-mega104-ioa.csv`

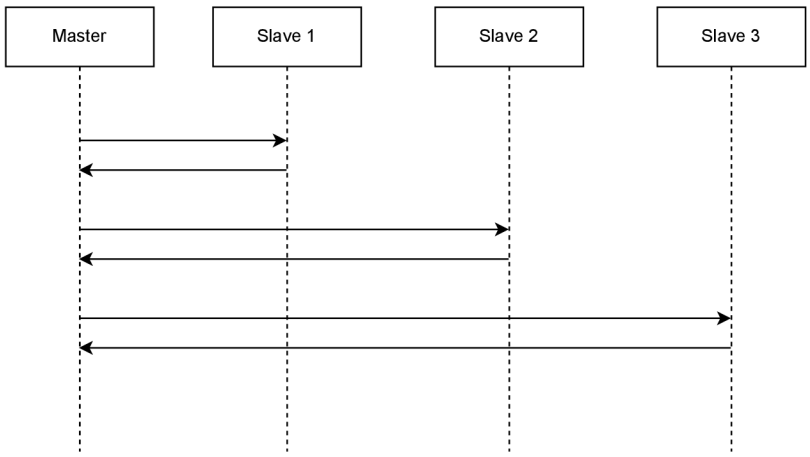
Datové sady A a B zachycují komunikaci v tzv. *peer-to-peer* profilu, který reprezentuje komunikaci mezi dvěma uzly [1]. U datové sady A se jedná o komunikaci mezi řídící stanici `192.168.11.248:2404` a podřízenou stanicí `192.168.11.111:56693`. U datové sady B vystupuje řídící stanice pod stejnou adresou, ale komunikuje s podřízenou stanicí s adresou `192.168.11.111:61254`.

Datové sady C a D zachycují komunikaci v tzv. *master-oriented* profilu, který reprezentuje komunikaci, kde vystupuje jedna řídící stanice a několik podřízených stanic. Ukázka průběhu *master-oriented* komunikace je na obrázku 4.1. Jak v datové sadě C, tak v datové sadě D je adresa řídící stanice `192.168.11.248:2404`. Podřízené stanice v obou datových sadách vystupují pod IP adresou `192.168.11.111` a liší se pouze v hodnotě portu. V datové

	A	B	C	D
Celkový počet paketů	14597	58930	104533	1460829
Časový interval	15:38:01.17	67:55:00.59	04:53:21.02	71:17:34.79
Počet zařízení	2	2	4	14
Počet kom. dvojic	1	1	3	13
Master → Slave	10151	40330	80243	1118552
	69.54%	68.44%	76.76%	76.57%
Slave → Master	4446	18600	24290	342277
	30.46%	31.56%	23.24%	23.43%

Tabulka 4.1: Přehled základních statistik datových sad. Pod pojmem *Master* → *Slave* se rozumí počet paketů ve směru od řídící stanice k libovolné podřízené stanici. Pod pojmem *Slave* → *Master* se rozumí počet paketů ve směru od libovolné podřízené stanice k nadřízené stanici.

sadě C se objevují stanice s porty `49784`, `49830` a `49849`. V datové sadě D jsou hodnoty portů `49849`, `50874`, `51197`, `51696`, `52142`, `52351`, `53015`, `53174`, `54224`, `55223`, `55382`, `55675` a `56693`.



Obrázek 4.1: Schéma komunikace *master-oriented* profilu. Řídící stanice periodicky komunikuje s množinou podřízených stanic a obsluhuje vždy pouze jednu podřízenou stanici. Z toho vyplývá, že komunikace probíhá v jednom okamžiku pouze mezi dvěma zařízeními [1].

4.2 Datové sady s útoky

Datové sady s útoky vznikly v rámci projektu *Bezpečnostní monitorování řídící komunikace ICS v energetických sítích (BONNET)*[1] na *Fakultě informačních technologií Vysokého učení technického v Brně*. Sady obsahují záznamy komunikace, ve které byl uměle nasimulován útok nebo anomálie. Jedná se o upravené verze jedné ze základních sad představených v podkapitole 4.1, konkrétně sady B. Všechny útoky byly generovány na aplikační vrstvě. Názvy sad a jejich popis je uveden výčtem:

[1] https://www.fit.vut.cz/research/project/1303/.en

- `connection-loss.cvs:` v rámci komunikace dochází dvakrát ke ztrátě spojení. Konkrétně v časech:

  1. od 16:27:57.68 do 16:37:48.63 (10 minut, 146 chybějících paketů)
  2. od 08:08:01.20 do 09:08:25.95 (1 hodina, 921 chybějících paketů)

- `switching-attack.csv:` cílem útoku je zapnutí/vypnutí zařízení. Celkově je odesláno 24 sérií paketů s parametry $asduType = 46$ (Double cmd), $numix = 1$, $cot = 6$ (Act), $cot = 7$ (ActCon), $cot = 10$ (ActTerm), $oa = 0$, $addr = 655535$, $ioa = 2$. Útok začíná v čase 06:27:55:00, trvá 10 minut a do komunikace zanáší 72 nových paketů.

- `scanning-attack.csv:` horizontální a vertikální skenování. Útok tohoto typu je možné provést díky zranitelnosti IEC 104 protokolu popsané v sekci *Neautentizovaný přístup* v podkapitole 2.3.

  1. Cílem horizontálního skenování je nalezení IP adresy řídící stanice. Útok začíná v čase 10:32:07 a končí v 10:49:10. Z podvržené adresy `192.168.11.102:45280` jsou odesílány IEC 104 U-příkazy *TestFrame Act* (s atributy $fmt = 0x03$, $uType = 0x10$) na port 2404 používaný řídící stanicí. Pokud se řídící stanice na dané IP adrese nachází, odpoví paketem *TestFrame Conf* (s atributy $fmt = 0x03$, $uType = 0x20$).
  2. Vertikální skenování prozkoumává informační objekty zařízení. V tomto případě je útok směřován na zařízení s IP adresou `192.168.11.111`. Za účelem skrytí své identity, používá útočník podvrženou IP adresu `192.168.11.248`, která patří řídící stanici. Útočník odesílá dotaz s atributy $asduType = 100$ (General Interrogation) a $cot = 6$ (Activation). Pokud informační objekt existuje, odesílá napadené zařízení odpověď s atributy $asduType = 100$ (General Interrogation) a $cot = 7$ (Activation Conf), jinak odesílá odpoveď s atributem $cot = 47$ (Unknown object address). K útoku se využívají výchozí hodnoty atributů $addr = 65535$ a $oa = 0$. Přestože je ioa adresa dlouhá 16 bitů (a může tedy nabývat 2^{16} hodnot), útok prohledává pouze hodnoty adres od 0 do 127. Útok začíná v čase 01:02:18 a končí v 01:23:19.

- `dos-attack.csv:` útočník se snaží zahltit systém stovkami legitimních požadavků. K tomu používá podvrženou IP adresu `192.168.11.248`, ze které posílá pakety s atributy $asduType = 36$ (Measured value, short floating point, with time tag) a $cot = 3$ (Spontaneous event). Útok začíná v 23:50:02 a končí v 01:18:29. Obsahuje celkově 1049 podvržených požadavků. Další útok je opakován v 02:30:05 a trvá do 04:01:54.

- `rogue-devices.csv:` do systému je zaneseno tzv. *rogue*[2] zařízení, které se vydává za legitimní IEC 104 zařízení s IP adresou `192.168.11.246`. Útočník používá ke komunikaci sekvenci paketů s atributy $asduType = 36$ (Measured value, short floating point with time tag) a $cot = 3$ (Spontaneous event). Útok začíná v 15:19:00 a končí v 15:46:03.

- `injection-attack.cvs:` Útočník odesílá neobvyklé požadavky s atributy $asduType = 45$ (Single command) a $cot = 6$ (Activation) na informační objekt s $ioa \subseteq \{2, 31, 32\}$.

[2]Zařízení, které je neoprávněně zavedeno do sítě, za účelem poškození provozovatele sítě. Může např. odposlouchávat komunikaci v síti, odesílat podvržené kontrolní příkazy jiným zařízením nebo provádět jiné aktivity, které představují bezpečnostní hrozbu pro provozovatele nebo uživatele sítě.

Zacílená stanice odesílá odpověď s parametrem *cot* = 7 (Activation Conf). Útok začíná v 19:35:19, končí v 19:41:06 a skládá se z 83 paketů.

Další útok typu *injection attack* se objevuje v 21:05:32, kdy útočník zahájí přenos souboru do napadené stanice s IP adresou `192.168.11.111`. Útočník odesílá pakety s atributy *asduType* 122 (Call directory, select file), 120 (File ready), 121 (Section ready), 123 (Last section), 124 (Ack file) a 125 (Segment). Útočník přistupuje k informačnímu objektu s *ioa* = 65537, který není typicky přístupný. Útok končí v 21:21:14 a skládá se z 221 paketů.

4.3 Atributy datových sad

Všechny popsané sady obsahují těchto 16 atributů:

- **TimeStamp:** Časové razítko paketu.
- **Relative Time:** Relativní čas zachycení paketu v rámci datové sady v sekundách.
- **srcIP:** IP adresa zdrojové stanice.
- **dstIP:** IP adresa cílové stanice.
- **srcPort:** Port zdrojové stanice.
- **dstPort:** Port cílové stanice.
- **ipLen:** Délka IP paketu.
- **len:** Délka aplikačních dat.
- **fmt:** Formát APCI jednotky (viz 2.2.2). Konkrétní hodnoty vyskytující se v dostupných datových sadách:
  - `0x00`: I-format
  - `0x01`: S-format
  - `0x03`: U-format
- **uType:** Funkce rámce, pouze pro pakety v U-formátu.
- **asduType:** Typ ASDU jednotky. (viz A.1)
- **numix:** Počet hodnot v IOA poli.
- **cot:** Důvod přenosu (viz 2.2.2).
- **oa:** Adresa původce. Z angl. zkratky pro *originator address*.
- **addr:** Linková adresa zařízení, může být dlouhá 8 nebo 16 bitů. Hodnoty `#FF` a `#FFFF` reprezentují broadcastové adresy.
- **ioa:** Pole adres informačních objektů (viz 2.2.2).

4.4 Metody popisu ICS komunikace

V této podkapitole budou představeny metody, které mohou být použity k analýze datových sad popsaných v podkapitolách 4.1 a 4.2. Vhledy, které se analýzou získají, poskytují uživateli celkový přehled o odchycené komunikaci a umožňují mu odhalit případné neočekávané události, útoky a anomálie.

4.4.1 Počet paketů v čase

Jendou z nejzákladnějších vlastností, kterou lze u průmyslové komunikace sledovat, je jednoduché měření počtu paketů v čase. Měření se většinou provádí pro jednu komunikační dvojici, ale lze provést agregovaně i pro více stanic současně. V první řadě je potřeba zvolit velikost časového okna, ve kterém se bude počet paketů měřit. Úskalí volby velikosti časového okna jsou popsány dále v této podkapitole v sekci *Volba velikosti časového okna*. Výsledky lze vizualizovat pomocí spojnicového grafu. Graf přehledně zobrazuje počty paketů v jednotlivých časových oknech a jejich vývoj v čase. Uživatel by měl být schopen z grafu vyčíst průběh sledované komunikace, její stabilitu, nebo případně nalézt její výpadky.

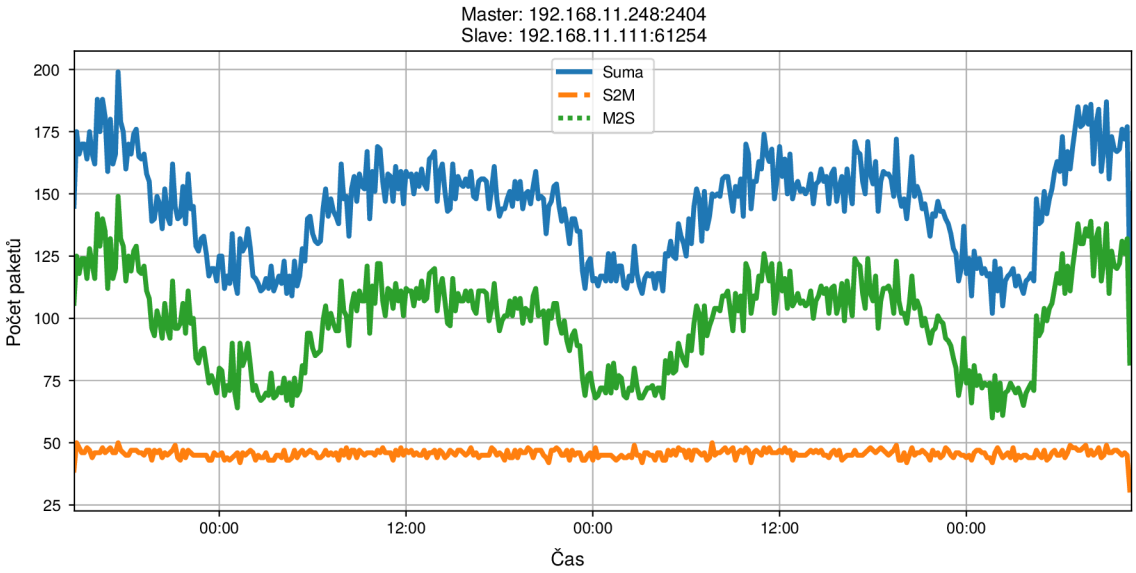
Jedním z možných rozšíření grafů, které vyobrazují komunikaci mezi dvěma stanicemi, je vynesení celkově tří křivek. Jednu křivku pro počet paketů ve směru od řídící stanici k podřízené (dále jen *M2S*), další pro počet paketů ve směru od podřízené stanice k řídící (dále jen *S2M*) a poslední pro součet obou směrů. Ukázka rozšíření je znázorněna na obrázku 4.2.

Volba velikosti časového okna

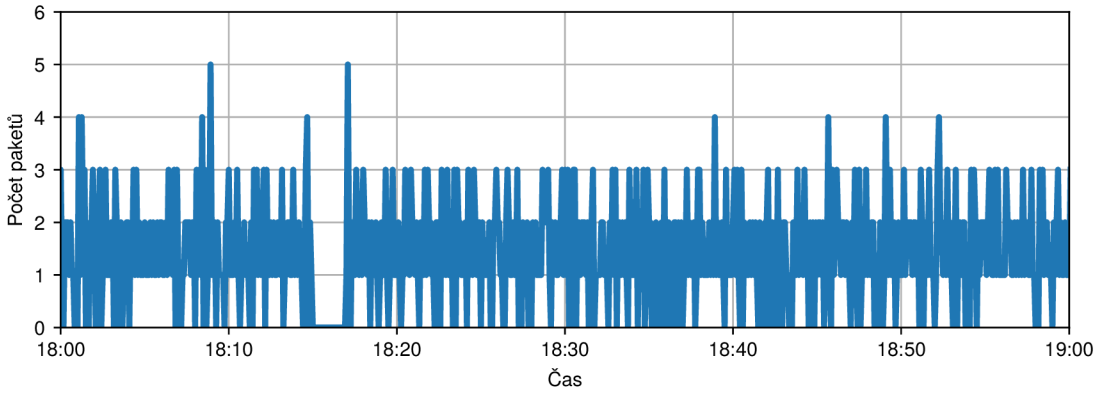
Volba velikosti časového okna je důležitým faktorem, který může představovat rozdíl mezi kvalitní a nekvalitní analýzou komunikace. Závisí na vlastnostech datové sady a na míře detailu, kterou uživatel vyžaduje. Ukázka dopadu velikosti časových oken na podobu spojnicových grafů je na obrázku 4.3.

V případě volby příliš malých časových oken, může dojít k vynulování některých oken a uživatel může nesprávně nabýt dojmu, že na některých místech došlo k výpadku komunikace. Ve skutečnosti je pouze velikost zvoleného časového okna menší než přirozená prodleva paketů v komunikaci. Příliš malé hodnoty také zanášejí do grafů šum, který ztěžuje jejich čitelnost (viz 4.3a).

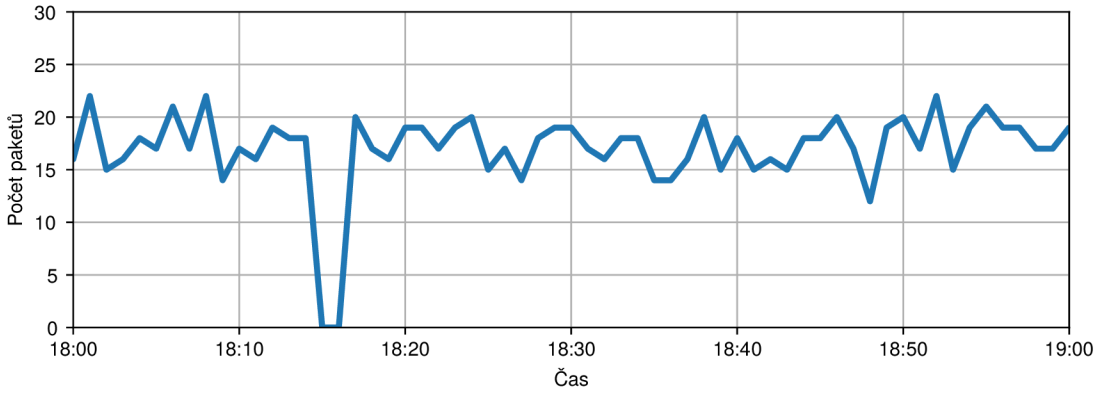
V případě volby příliš velkých časových oken, můžou naopak některé výpadky být uživateli zcela skryty. Uživatel může např. zvolit 60 minutové okno a zkoumat komunikaci kde došlo k výpadku na 5 minut. V takovém případě bude vliv výpadku na hodnotu v časovém okně minimální a z grafu nebude možné anomálii vyčíst (viz 4.3c).



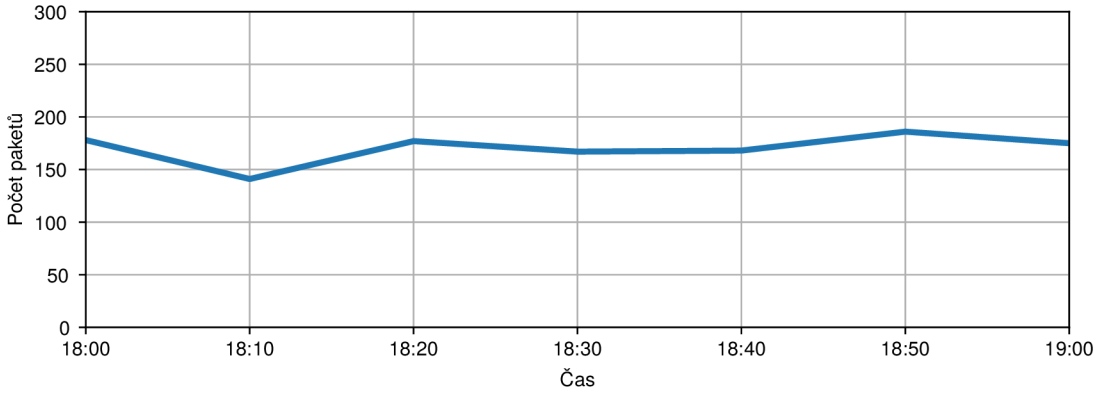
Obrázek 4.2: Spojnicový graf znázorňující počet zachycených paketů na záznamu komunikace z datové sady B a to konkrétně mezi stanicemi `192.168.11.248:2404` (master) a `192.168.11.111:61254` (slave). Velikost časových oken je 10 minut. Z grafu lze vidět, že tok paketů ve směru od podřízené stanice k řídící je v průběhu celého dne stabilní. V opačném směru je situace odlišná a jeví se, že během noci dochází k útlumu toku. Rozdělení komunikace na jednotlivé směry přináší nové hodnotné informace, které by jinak nebyly vidět.



(a) Velikost časového okna: 5 s. Graf je příliš jemný a obsahuje hodně šumu.



(b) Velikost časového okna: 1 min. Ideální volba časového okna, ve které lze výpadek jednoduše nalézt.

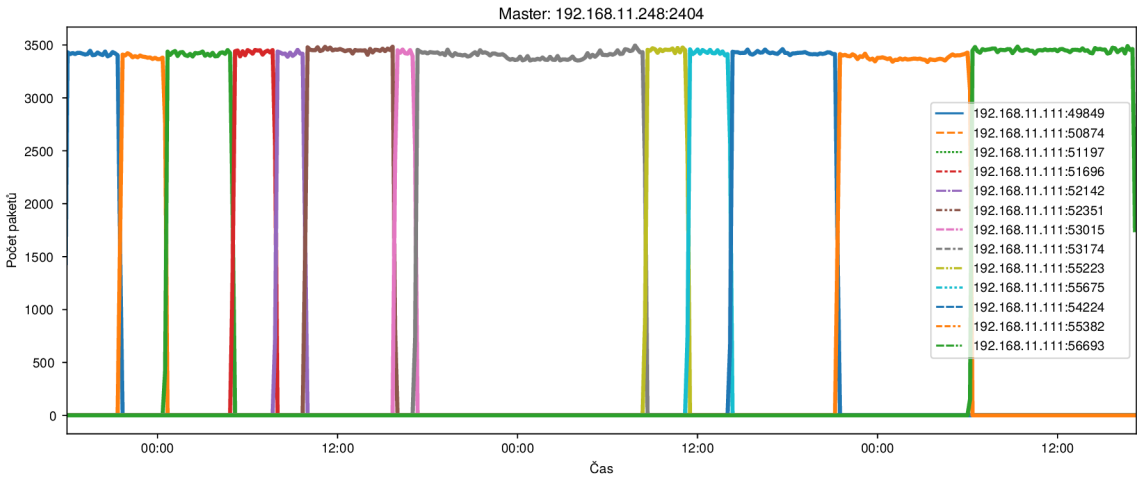


(c) Velikost časového okna: 10 min. Graf je příliš hrubý a informace o výpadku v něm nelze vyčíst.

Obrázek 4.3: Ukázka dopadu volby velikosti časového okna na čitelnost a informační hodnotu grafů. Do výstřižku komunikace z datové sady A (mezi stanicemi `192.168.11.248:2404` a `192.168.11.111:56693`) byl uměle zaveden výpadek v čase od 18:15 do 18:17. Z uživatelského hlediska je v tomto případě vhodná pouze volba časového okna o velikosti 60 *s*. Zbylé dva grafy mají pro uživatele nízkou informační hodnotu.

4.4.2 Komunikující dvojice

V *master-oriented* komunikačním profilu, který je znázorněn na obrázku 4.1, komunikuje řídící stanice v jeden okamžik pouze s jednou podřízenou stanicí. Lze tedy sestavit graf, který zobrazuje kdy probíhá komunikace s jednotlivými podřízenými stanicemi. Ukázka takového grafu je na obrázku 4.4. Díky tomuto grafu je možné určit v jakých časech došlo k přepnutí komunikačních dvojic a jak dlouho řídící stanice komunikovala s jednotlivými podřízenými stanicemi.



Obrázek 4.4: Ukázka grafu pro analázu *master-oriented* komunikace v datové sadě D. Řídící stanice střídavě komunikuje celkově se 13 podřízenými stanicemi. Velikost časového okna je 10 minut.

4.4.3 Inter-arrival time

Hodnota *Inter-arrival time* (Δt) reprezentuje uběhnutý čas mezi zachycením dvou po sobě jdoucích paketů v komunikaci. Hodnotu Δt je možné měřit zvlášť pro jeden směr, nebo pro oba směry zároveň. V případě IEC 104 protokolu a *master-oriented* komunikačního profilu se jeví jako výhodnější použití obousměrného měření [1].

Inter-arrival time může být využit k výpočtu jeho popisných charakteristik míry a variability nad celou datovou sadou. V takovém případě lze sledovat např. následující vlastnosti:

- **Minimální hodnota:** pomáhá uživateli s výběrem velikosti časového okna. Velikost časového okna by měla být vždy výrazně větší než minimální hodnota inter-arrival time. V opačném případě dochází ke ztrátě informace ve spojnicovém grafu.
- **Maximální hodnota:** v případě že je maximální hodnota výrazně vychýlená od střední hodnoty a třetího kvartilu, může to značit, že se v komunikaci nachází delší doba, kdy komunikace neprobíhala. Tzn. v komunikaci může být výpadek nebo útok.
- **Střední hodnota:** napovídá jaká asi byla „obvyklá“ pozorovaná hodnota. V kombinaci s mediánem pomáhá odhalit případnou asymetričnost rozložení hodnot inter-arrival time.

4.4.4 Stabilita atributů

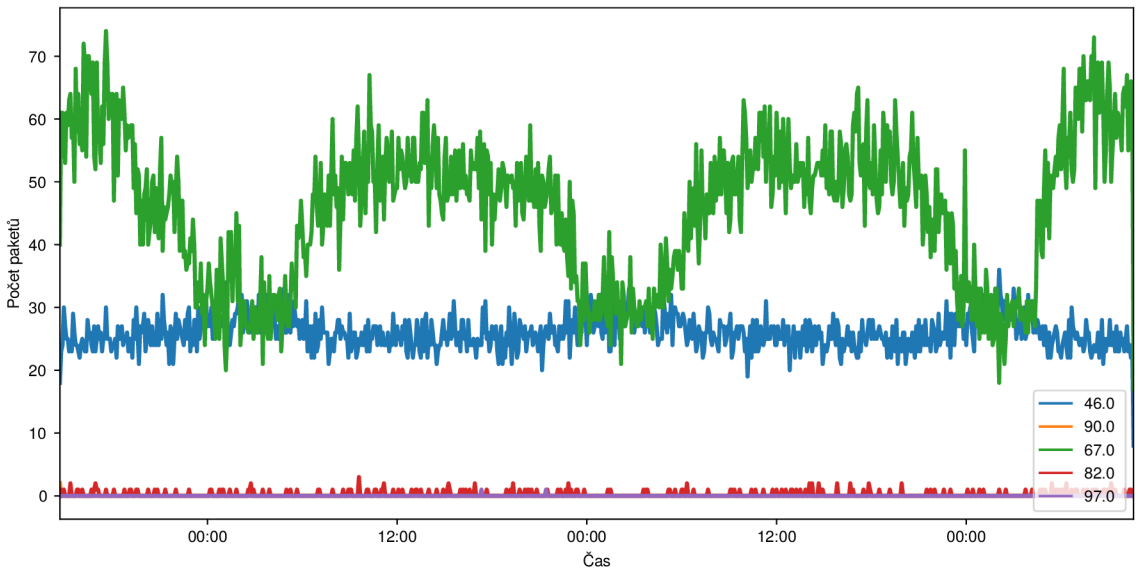
Důležitým ukazatelem, který umožňuje uživateli detekovat anomálie v komunikaci, je stabilita hodnot atributů IEC 104 komunikace. Datová sada je opět rozdělena do časových oken, pro která se počítá počet odchycených paketů s danou hodnotou atributu. Pro některé hodnoty některých atributů je pak typické, že jsou jejich toky stabilní, případně opisují nějakou periodickou křivku. Kromě sestrojení spojnicového grafu je vhodné využít i popisných charakteristik míry a variability.

Ukázka použití metody na datové sadě B

Za účelem ukázky metody byla vybrána datová sada B na celém jejím časovém intervalu. Zkoumaným atributem je *ipLen* a všechny jeho hodnoty, kterých v datové sadě nabývá (celkově 5 hodnot). Velikost časového okna je 5 minut. Předzpracování dat do časových oken je naznačeno v tabulce 4.2. Vykreslením hodnot tabulky vzniká spojnicový graf, který je vyobrazen na obrázku 4.5. Základní charakteristiky míry a variablity jsou popsány v tabulce 4.3.

	46	90	67	82	97
14:45	23	0	61	1	0
14:50	24	0	61	0	0
...					
10:25	22	0	66	1	0
10:30	24	0	64	0	0

Tabulka 4.2: Tabulka časových oken pro jednotlivé hodnoty atributu *ipLen*. Hodnoty buněk značí, kolik paketů s danou hodnotou atributu *ipLen* bylo zachyceno v daném časovém okně. V tabulce jsou uvedeny začátky časových oken. První a poslední časová okna nejsou uvedena, jelikož jejich délka je kratší než 5 minut a zkreslovala by statistiky.



Obrázek 4.5: Spojnicový graf znázorňující průběhy toků paketů s různými hodnotami atributu *ipLen*. Graf vychází z tabulky 4.2.

	μ	σ	$\mu - 3\sigma$	$\mu + 3\sigma$	Outliers
46	26,09	2,58	18,33	33,87	2
67	45,95	11,51	11,39	80,51	0
82	0,24	0,48	-1,22	1,70	20
97	0,00	0,05	-0,15	0,51	2

Tabulka 4.3: Základní popisné charakteristiky jednotlivých hodnot atributu *ipLen*. Hodnota „90“ není uvedena, jelikož se nachází pouze v prvním časovém okně, které nebylo do statistik započítáno. Sloupec „Outliers“ udává, kolik hodnot spadá mimo interval $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$.

Z údajů, které poskytují graf 4.5 a tabulka 4.3 lze vyvodit, které atributy se chovají stabilně a předvídatelně. Cílem analýzy ICS komunikace je nalezení takových atributů, které mohou pomoci s detekcí útoků a anomálií. Ideálně by měl atribut mít malou velikost intervalu $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ a spodní hranice by měla být větší jak 0, tak aby bylo možné detekovat i výpadky. Ve výše uvedeném příkladě se jeví jako ideální hodnota 46 která splňuje oba požadavky. Pokud je interval příliš velký, je těžké zachytit odlehlé hodnoty. To lze pozorovat u hodnoty atributu 67, která má velký rozptyl a nebyla detekována žádná hodnota mimo interval 3σ. Hodnoty atributů, které se v grafu pohybují kolem nuly, nemá příliš smysl uvažovat.

Kapitola 5

Návrh aplikace ICS Analyzer

Tato kapitola se zabývá návrhem nového nástroje, který slouží pro statistickou analýzu a vizualizaci dat průmyslové komunikace využívající protokolu IEC 104. Za tímto účelem byla navržena *desktopová* aplikace *ICS Analyzer*. Před samotným představením aplikace jsou nejprve v podkapitole 5.1 popsány nedostatky v aktuálně dostupných možnostech analýzy. Následuje popis požadavků, které by mělo navrhované řešení splňovat (viz 5.2). Významnou součástí návrhu je způsob, kterým nástroj přistupuje ke zpracování dat, jenž je uveden v podkapitole 5.3. Poslední podkapitola 5.4 se zabývá popisem navrženého uživatelského rozhraní.

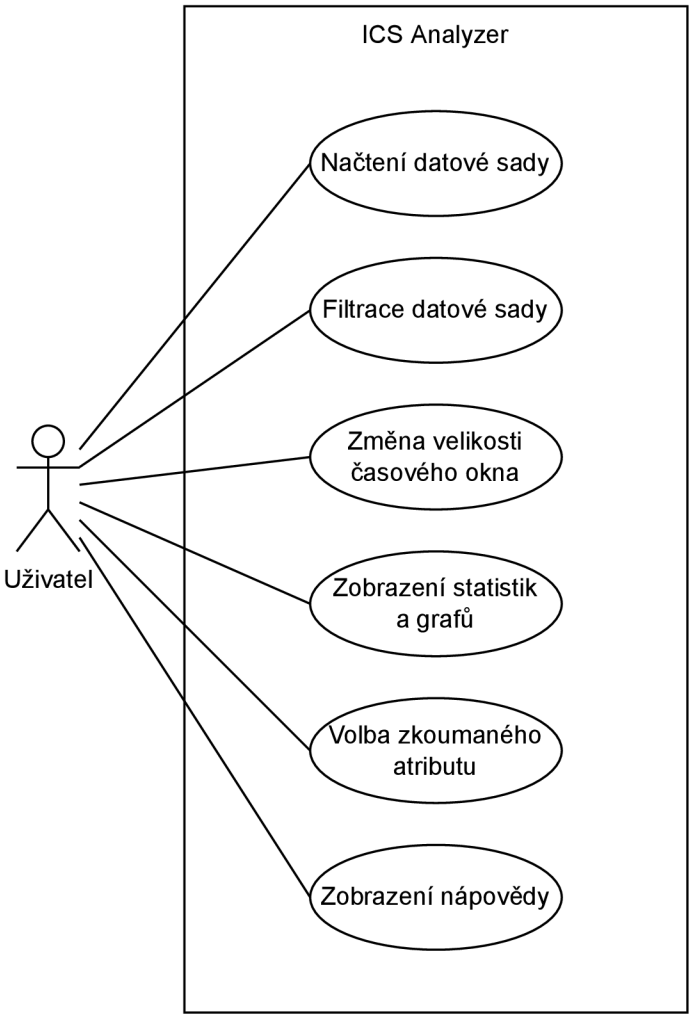
5.1 Stávající řešení

Aktuálně neexistuje volně dostupný nástroj, který by uživateli umožňoval jednoduše zpracovat a analyzovat datové sady průmyslové komunikace ve formátu `csv`. Uživatel je odkázán na „ruční“ zpracování dat. Za tímto účelem lze použít např. tabulkové procesory jako je *Microsoft Excel* nebo *LibreOffice Calc*, které ovšem nejsou stavěny pro zpracování rozsáhlých datových sad (např. nástroj *LibreOffice Calc* nedokáže načíst celou datovou sadu D) a navíc vyžadují hlubokou znalost použitého nástroje. Další možností je použití programovacích jazyků *Python* (s využitím knihovny *Pandas*) nebo *R*, které nabízejí uživateli nástroje pro zpracování obecných tabulkových dat. Mimo požadavku na znalost těchto jazyků je problémem vysoká míra úsilí, které musí uživatel vynaložit, aby z dat získal relevantní vhledy. Uživatel musí vždy nejprve provést několik časové náročných kroků (jako např. načtení dat, vytvoření filtrů, vykreslení grafů atd.), které odvádí jeho pozornost od samotné analýzy. Aplikace *ICS Analyzer* bude tyto kroky provádět automaticky a nabídne tak uživateli možnost se plně soustředit na analýzu datové sady.

5.2 Analýza požadavků

Primárním cílem navrhované aplikace je poskytnutí rychlého a jednoduchého způsobu analýzy datových sad průmyslové komunikace. Uživatel by neměl být zatěžován nepodstatnými úkony a měl by mít možnost se soustředit pouze na samotnou analýzu komunikace. Aplikace by měla uživateli poskytovat relevantní vhledy o zkoumané komunikaci a usnadňovat tak nalezení stabilních charakteristik nebo případných anomálií v komunikaci. Konkrétní možnosti interakce uživatele s aplikací jsou znázorněny diagramem užití na obrázku 5.1. Celkové požadavky lze rozdělit do několika kategorií:

- **Načítání datové sady:** aplikace by měla umožňovat přímočaře načíst datovou sadu ze souboru ve formátu `csv`. Přestože bude primárním využitím aplikace analýza datových sad popsaných v podkapitole 4.1, měla by být koncipována tak, aby byla co nejobecnější a mohla být využita i pro jiné datové sady. Podrobnější popis průběhu načítání a podmínky, které musí datová sada splňovat, jsou blíže popsány v podkapitole 6.4.
- **Přehledné zobrazení sady:** soubory ve formátu `csv` bývají pro uživatele nepřehledné a obtížně čitelné. Z tohoto důvodu by měla aplikace nabízet možnost zobrazení načtených dat v přehledné tabulce.
- **Zobrazení základních statistik datových sad:** před zahájením podrobnějšího zkoumání datové sady je vhodné, aby uživatel získal základní představu o její podobě. Z tohoto důvodu by měla aplikace uživateli poskytovat základní popisné statistiky o načtených datech.
- **Filtrace datové sady:** v některých případech může uživatel chtít analyzovat pouze část načtené datové sady, která splňuje jím zvolené podmínky. Aplikace by měla uživateli nabízet možnost filtrace datové sady na základě různých kritérií. Návrh systému filtrace je blíže popsán v sekci 5.3.1.
- **Nastavení velikosti časových oken:** klíčovou součástí analýzy je zobrazení počtu zachycených paketů (různých typů) v jednotlivých časových oknech. Velikost časového okna je proto důležitým parametrem, který by měl být nastavitelný.
- **Analýza komunikace:** aplikace by měla disponovat sadou pohledů, které umožní uživateli nalézt stabilní charakteristiky nebo anomálie v komunikaci. Vhledy by měly vycházet z metod popsaných v kapitole 4.4. Popis implementovaných vhledů je uveden v podkapitole 6.5.



Obrázek 5.1: Diagram případu užití (angl. *use case*) demonstrující různé způsoby, kterými uživatel může interagovat s aplikací.

5.3 Způsob zpracování dat

Důležitou součástí návrhu je způsob zpracování datové sady a to od jejího načtení po generování statistik a grafů. Aplikace bude pro manipulaci s daty na pozadí využívat knihovny *Pandas* a její datové struktury zvané „datový rámec“ (angl. *dataframe*, viz konec podkapitoly 6.1). Navržená sekvence zpracování je vyjádřena blokovým schématem na obrázku 5.2. Její jednotlivé kroky jsou:

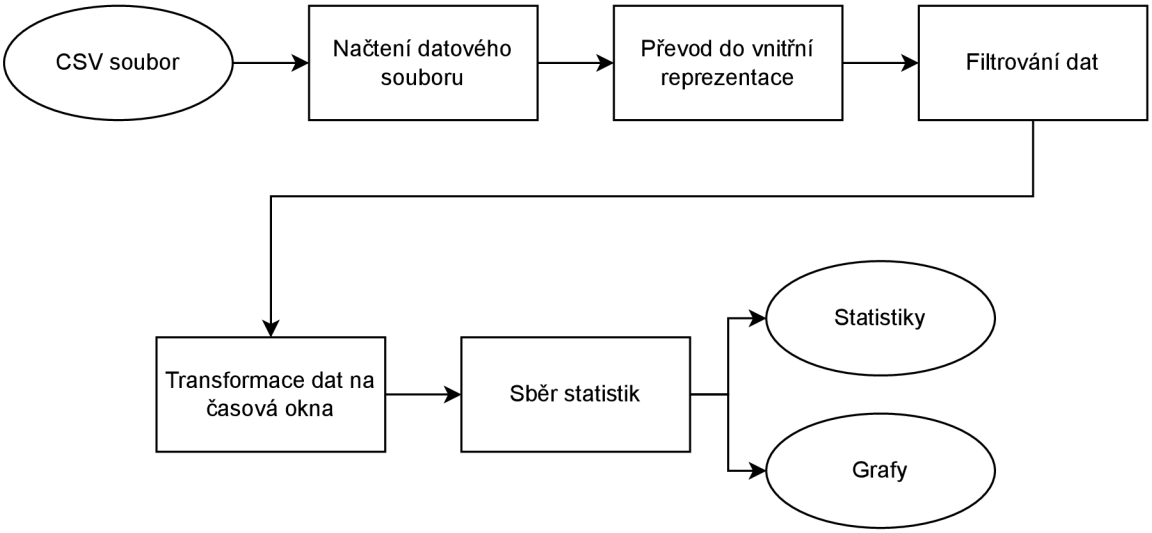
1. **Načtení datového souboru:** v první řadě je potřeba datový soubor korektně načíst. Z dat obsažených v souboru v *csv* formátu se vytvoří datový rámec, který jednotlivým atributům přiřadí datové typy. V tomto kroku je důležité, aby byly datové typy správně zvoleny, jinak načtení selže. Aplikace by měla provádět automatickou detekci datových typů a umožňovat uživateli jejich případnou změnu.

2. **Převod do vnitřní reprezentace:** po načtení dat se do datového rámce přidají nové sloupce, které budou uživateli skryty a usnadní (programátorovi) další zpracování. Jedná se např. o sloupce pro označení stanic vnitřním identifikačním číslem apod.

3. **Filtrování dat:** filtrace dat bude probíhat na základě pěti kritérií, které bude moct uživatel měnit. Datový rámec bude vyfiltrován tak, aby splňoval všechna nastavená kritéria. Výčet filtrů je uveden v podkapitole 5.3.1.

4. **Transformace dat na časová okna:** před zobrazením grafů bude potřeba data přetransformovat do podoby, která agreguje data do časových oken, jejichž velikost je volena uživatelem. Tato transformace je blíže popsána v sekci 5.3.2.

5. **Sběr statistik:** ve chvíli kdy budou data vhodně předzpracována, bude možné začít generovat statistiky a grafy. Zatímco generování grafů očekává data již transformovaná na časová okna, některé statistiky tento krok vyžadovat nebudou.



Obrázek 5.2: Blokové schéma (angl. *block diagram*) ukazující průběh zpracování datové sady.

5.3.1 Filtrace datové sady

Aplikace bude nabízet celkově 5 kriterií, pomocí kterých bude možné filtrovat datovou sadu:

- Adresa řídící stanice
- Adresy podřízených stanic
- Směr komunikace
- Začátek a konec komunikace (interval)
- Hodnoty zvoleného atributu

5.3.2 Transformace dat na časová okna

Pomocí transformace dat na časová okna lze z dat lépe získávat informaci o počtu zaznamenaných paketů s danou hodnotou atributu v daném intervale. V tabulce původních dat se nejprve vybere atribut, pro který se provede transformace. Aby měla transformace smysl, měl by atribut nabývat kategorických hodnot. Nemá smysl provádět transformaci

pro spojité atributy (např. relativní čas), protože by se v nové tabulce objevovaly pouze záznamy (řádky) s jedinou nenulovou hodnotou. Následně se vytvoří nová tabulka, ve které jsou z jednotlivých hodnot vybraného atributu vytvořeny nové sloupce, které reprezentují absolutní četnost jejich zastoupení v časových oknech. Agregační funkcí je v tomto případě funkce *počet*. Grafické znázornění transformace je ukázáno na obrázku 5.3.

Časové razítko	**Hodnota atributu**
17:52:18	a
17:52:24	b
17:53:50	b
17:57:43	a
18:01:17	a
18:02:28	a
18:02:45	a
18:06:33	b
18:06:42	c

↓

Začátek časového okna	**a**	**b**	**c**
17:50	1	2	0
17:55	1	0	0
18:00	3	0	0
18:05	0	1	1

Obrázek 5.3: Ukázka transformace dat na časová okna na náhodně vygenerovaných datech. Hodnoty atributu (a, b, c) byly vybrány pouze pro ilustrační účely. Velikost časového okna je 5 minut a agregační funkcí je funkce *počet*.

Poznámka: Kromě funkce počet existují další široce užívané agregačních funkce. V případě, že atribut je číselného typu, lze použít např. funkce pro výpočet sumy, výpočet střední hodnoty, nalezení minima nebo maxima apod. Existuje i možnost vytvoření vlastní agregační funkce. V této práci je však využita pouze agregační funkce počet.

5.4 Uživatelské rozhraní

Uživatelské rozhraní aplikace se bude skládat z menu, informačního panelu a šesti karet. Společně tyto prvky budou tvořit přímočarý nástroj pro analýzu průmyslové komunikace.

Menu

Menu bude hlavním prostředkem ovládání aplikace a bude nabízet 3 skupiny příkazů. Hierarchie příkazů a jejich funkcionalita je popsána výčtem:

1. **File:** přikazy pro základní ovládání aplikace.
    - **Load csv:** otevření dialogu pro načtení `csv` souboru. Načítání datové sady bude blíže popsáno v podkapitole 6.4.
    - **Exit:** ukončení aplikace.

2. **Filter:** příkazy pro filtraci datové sady.
    - **Select master station:** otevření dialogu pro výběr řídící stanice.
    - **Select slaves:** otevření dialogu pro výběr podmnožiny podřízených stanic. Výběr bude omezen na stanice, které komunikují s již zvolenou řídící stanicí.
    - **Select direction:** otevření dialogu pro filtraci komunikace podle směru. Celkově budou k dispozici 3 druhy filtrace: obousměrná, M2S a S2M.
    - **Change start and end time:** otevření dialogu pro omezení zkoumaného intervalu datové sady. Uživatel bude moct volit jeho začátek a konec.
    - **Change time window size:** otevření dialogu pro změnu velikosti časového okna.
    - **Select attribute:** otevření dialogu pro výběr atributu.
    - **Select attribute values:** otevření dialogu pro výběr konkrétních hodnot již zvoleného atributu.
3. **Help:** příkazy nápovědy.
    - **Show help:** zobrazení nápovědy.
    - **About:** zobrazení základních informací o aplikaci.

Informační panel

Informační panel by měl být vždy viditelným prvkem v horní části okna, který bude zobrazovat uživateli aktuálně aplikované nastavení a filtry datové sady. Konkrétně by měl zobrazovat nastavení filtrů pro adresu řídící stanice, adresy podřízených stanic, směr a časový interval. Panel by navíc měl ukazovat i nastavenou velikost časového okna a název zvoleného atributu.

Panel karet

Stěžejním prvkem uživatelského rozhraní bude panel karet, který by měl nabízet uživateli celkově 6 různých pohledů na načtená data. Popis a význam jednotlivých pohledů je blíže rozebrán v podkapitole 6.5. Jejich finální podoba je pak ukázána na obrázcích v příloze D.

Kapitola 6

Implementace aplikace ICS Analyzer

Následující kapitola se zabývá implementačními detaily aplikace *ICS Analyzer*. K jejímu vývoji byl vybrán jazyk *Python 3* a některé jeho široce užívané knihovny, jenž jsou představeny v podkapitole 6.1. Architektura aplikace a její rozdělení do logických celků je popsáno v podkapitole 6.2.

Během vývoje bylo potřeba vyřešit značné množství překážek. Jednou z nejvýznamnějších byla vysoká časová náročnost funkcí pro zpracování datových sad. Problém byl vyřešen použitím tzv. *vektorizace*, která je popsána v podkapitole 6.3.

Aplikace nabízí uživateli přímočarý způsob načítání datové sady. Jeho průběh je popsán v podkapitole 6.4. Následuje představení jednotlivých pohledů (karet). Každý pohled nabízí unikátní náhled na datovou sadu a poskytuje uživateli jinou informaci o komunikaci. Jejich podrobný popis je uveden v podkapitole 6.5.

6.1 Použité nástroje a knihovny

Implementace aplikace je založena na jazyce *Python*[1] 3.10.4, který nabízí jak prostředky pro zpracování dat, tak pro tvorbu *desktopových* aplikací a jejich uživatelského rozhraní. V implementaci jsou využity nové konstrukce verze 3.10 (např. *structural pattern matching* nebo nová syntaxe pro *type hinting*) a tudíž není aplikace zpětně kompatibilní se staršími verzemi.

K tvorbě grafického uživatelského rozhraní byla použita knihovna *PyQt6*[2]. Jedná se o robustní a populární nástroj pro tvorbu multiplatformních uživatelských rozhraní, který vychází z frameworku *Qt* pro *C++*. *PyQt6* poskytuje uživateli možnost tvorby rozhraní jak „programově“, tak pomocí nástroje *Qt Designer*. Pro účely této práce byla zvolena první varianta.

K načtení, zpracování a analýze datových sad byly využity knihovny *NumPy*[3] a *Pandas*[4]. Ke grafickému zobrazení dat pak byly použity knihovny *Matplotlib*[5] a *Seaborn*[6]. Všechny

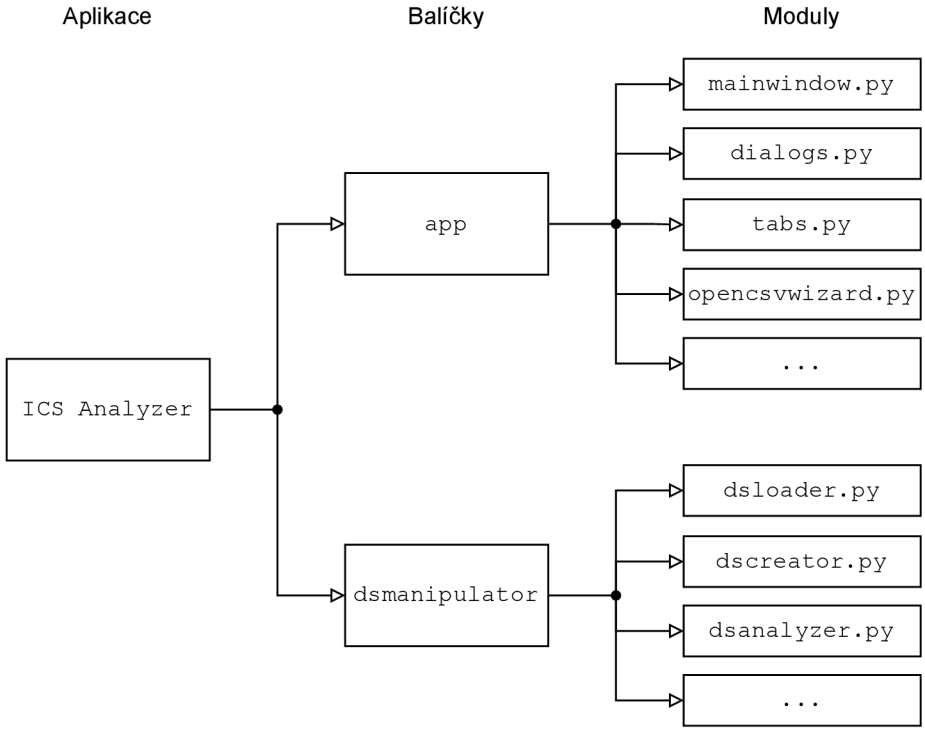
[1] https://www.python.org/downloads/release/python-3104/
[2] https://pypi.org/project/PyQt6/6.2.3/
[3] https://numpy.org/
[4] https://pandas.pydata.org/
[5] https://matplotlib.org/
[6] https://seaborn.pydata.org/

zmíněné knihovny jsou široce využívány komunitou a patří mezi standardní nástroje pro zpracování a analýzu dat.

Za speciální zmínku stojí datová struktura *datového rámce* (angl. *dataframe*), která je obecně nejpoužívanějším objektem knihovny *Pandas* [17]. Jedná se o dvourozměrnou datovou strukturu, kde řádky reprezentují jednotlivé záznamy a sloupce reprezentují atributy datové sady. Sloupce jsou určeny názvem a datovým typem. Povinným sloupcem je tzv. „index“, který slouží k adresaci záznamů. V aplikaci se datové rámce využívají k uchovávání načtených dat a manipulaci s nimi.

6.2 Architektura

Zdrojový kód aplikace je rozdělen na dva logické celky, které jsou implementovány ve formě *Python* balíčků. Prvním z nich je balíček `app`, jenž implementuje hlavní části aplikace jako je uživatelské rozhraní apod. Druhým je balíček `dsmanipulator`, který poskytuje nástroje ke zpracování datových rámců obsahující data z průmyslové komunikace. Architektura aplikace je naznačená blokovým schématem na obrázku 6.1.



Obrázek 6.1: Blokové schéma architektury aplikace *ICS Analyzer*. Schéma obsahuje pouze nejdůležitější moduly.

Balíček `app`

Balíček obsahuje implementaci grafického uživatelského rozhraní a funkcionality aplikace. Využívá metod nabízených balíčkem `dsmanipulator`. Nejdůležitějšími moduly balíčku jsou:

- `mainwindow`: implementace hlavního okna aplikace.

- `dialogs`: implementace dialogů aplikace (např. dialog pro změnu řídící stanice, intervalu, ...).
- `tabs`: implementace pohledů (karet) aplikace. Jednotlivé pohledy jsou blíže popsány v podkapitole 6.5.
- `opencsvwizard`: implementace dialogu pro načítání `csv` souborů, jehož podrobnější popis je uveden v podkapitole 6.4.

Balíček `dsmanipulator`

Balíček nabízí funkcionalitu pro načítání a zpracování datových rámců, které obsahují záznamy průmyslové komunikace. Je koncipován tak, aby byl použitelný i jako samostatná jednotka (např. pro „ruční“ analýzu datových sad v nástroji *Jupyter Notebook*[7]). Stěžejními moduly balíčku `dsmanipulator` jsou:

- `dsloader`: funkce pro detekci vlastností `csv` souboru (např. oddělovače sloupců) a jeho načtení.
- `dscreator`: funkce pro vnitřní zpracování datového rámce. Jedná se např. o funkce pro převod indexového sloupce na časovou řadu nebo přidání sloupců pro reprezentaci různých vlastností (vnitřní identifikátor stanice, komunikující dvojice, *inter-arrival time* ...).
- `dsanalyzer`: funkce pro analýzu datového rámce, generování statistik a grafů.

6.3 Vektorizace

Jedním z problémů, které vznikly během vývoje aplikace (konkrétně modulu `dscreator`), byla nízká efektivita funkcí pro zpracování datových rámců, což znemožňovalo použití aplikace pro větší datové sady. Z tohoto důvodu je v implementaci aplikace využito tzv. *vektorizace*, díky které lze významně urychlit výpočet některých operací nad datovými rámci. Metody vektorizace jsou nabízeny převážně knihovnou *NumPy* a v aplikaci jsou primárně využity v modulu `dscreator`. Příklad užití vektorizace v aplikaci je ukázán na obrázku 6.2.

[7] https://jupyter.org/

```python
def add_station_id(
    df: pd.DataFrame,
    station_ids: dict[Station, int],
) ->pd.DataFrame:

    df["SRC station id"] =df.apply(
        lambda row: station_ids[Station(row["srcIP"], row["srcPort"])],
        axis=1,
    )

    return df
```

(a) Původní „naivní“ implementace, která nevyužívá vektorizace. Při zpracování datové sady D byl čas potřebný k vykonání funkce 18.6 s.

```python
def add_station_id_vectorized(
    df: pd.DataFrame,
    station_ids: dict[Station, int],
) ->pd.DataFrame:

    def get_station_id(ip, port):
        return station_ids[Station(ip, port)]

    get_station_id_vectorized =np.vectorize(get_station_id)

    # prevod sloupcu datoveho ramce do numpy poli
    srcIPs =df["srcIP"].values
    srcPorts =df["srcPort"].values

    df["SRC station id"] =get_station_id_vectorized(srcIPs, srcPorts)

    return df
```

(b) Implementace využívající vektorizace. Při zpracování datové sady D byl čas potřebný k vykonání funkce 1.72 s.

Obrázek 6.2: Funkce `add_station_id` přidává do datového rámce nový sloupec s vnitřním identifikátorem zdrojové stanice. Pro každý záznam v datové sadě musí funkce nahlédnout do slovníku `station_ids`. Díky vektorizaci, lze tuto operaci provádět paralelně a snížit potřebný čas pro vykonání funkce. Při zpracování datové sady D bylo dosaženo téměř 11násobného zrychlení (testováno v prostředí *Jupyter Notebook*, systém: Zorin OS 15.3, CPU: Intel i7-8750H, RAM: 16 GB). *Poznámka: Pro účely ukázky byla funkce oproti implementaci v aplikaci zjednodušena.*

6.4 Načítání datových sad

Aplikace podporuje načítání datových sad ze souborů ve formátu `csv`. Podmínkou je, aby soubor obsahoval kompletní záznamy některých povinných atributů. Konkrétně se jedná o tyto atributy: Časové razítko, IP adresa zdrojové stanice, IP adresa cílové stanice. Pro optimální funkcionalitu aplikace by se v datové sadě měly nacházet i nepovinné atributy značící relativní čas, port zdrojové stanice a port cílové stanice. V případě, že některý z těchto atributů chybí, budou některé části aplikace omezeny. Bez specifikace portů, může

automatická detekce řídící stanice selhat. Při chybějícím relativním čase, nebudou k dispozici popisné statistiky *inter-arrival time*.

Po kliknutí na tlačítko pro načtení nové datové sady se uživateli zobrazí obrazovka pro výběr oddělovače dat, která je zobrazena na obrázku C.1. Aplikace se nejprve pokusí oddělovač detekovat sama a zobrazí uživateli názvy sloupců, které by vznikly použitím daného oddělovače. V případě, že detekce selže (např. když se v názvech sloupců nachází jiné znaky běžně používané jako oddělovače), může uživatel volbu změnit a aplikace mu zobrazí nové názvy sloupců.

Na další obrazovce (viz C.2) volí uživatel datové typy sloupců a sloupce, které náleží speciálním atributům jako je časové razítko, IP adresa atd. Mezi podporované datové typy patří: časové razítko, textový řetězec a číselná hodnota. I v tomto případě se aplikace snaží automaticky detekovat datové typy atributů a to na základě prvních 10 000 záznamů. Je však žádoucí, aby uživatel detekované datové typy překontroloval a upravil, jelikož špatně zvolený datový typ vede k selhání načtení datové sady. Jelikož byla aplikace stavěna primárně pro zpracování datových sad představených v kapitole 4, je detekce pro tyhle datové sady optimalizována a nevyžaduje zásah uživatele. Dále uživatel volí, které sloupce reprezentují speciální atributy. Celkově aplikace podporuje 6 speciálních atributů, z nichž 3 jsou povinné (časové razítko, IP adresa zdrojové stanice, IP adresa cílové stanice) a 3 nepovinné (relativní čas, port zdrojové stanice, port cílové stanice). Při jakékoliv změně konfigurace, provede aplikace její kontrolu. Nejprve dojde k ověření zda-li jsou zvolené datové typy sloupců kompatibilní s reprezentací atributů. Poté se aplikace pokusí na pozadí načíst 15 000 záznamů a ověřit, že nedošlo k problémům při načítání. Pokud aplikace nalezne problém, vypíše se upozornění a uživateli je znemožněno postupovat dále.

6.5 Pohledy

Pohledy nabízí způsob, jakým efektivně zkoumat načtená data. Celkově aplikace podporuje 6 pohledů, které se liší informační hodnotou a mění se na základě zvoleného nastavení a filtrů. U každého pohledu je uvedeno, které nastavení jej ovlivňuje. Nadpisy sekcí v této podkapitole reflektují pojmenování pohledů v aplikaci.

Dataset table

Prvním pohledem je jednoduché promítnutí načteného datového souboru do tabulky. Tabulka přehledně zobrazuje data načtená z `csv` a usnadňuje jejich čtení a prohlížení. Pro uživatele má největší význam ve chvíli, kdy už se v komunikaci orientuje a potřebuje zkontrolovat např. hodnoty atributů jednotlivých paketů, jejich přesné pořadí apod. Pohled reaguje pouze na změny filtrů a jeho ukázka je na obrázku D.1.

General statistics

Druhý pohled souhrnně zobrazuje základní statistiky datového souboru. Dělí se na dva sloupce. Levý sloupec zobrazuje statistiky načtených dat bez aplikovaných filtrů a mění se pouze při načtení nové datové sady. Pravý sloupec zobrazuje statistiky filtrovaných dat a mění se při jakékoliv změně filtrů. Uživateli dává pohled základní představu o vlastnostech datové sady kterou zkoumá. Ve statistikách jsou obsažený informace jako např. délka časového intervalu dat, popisné míry inter-arrival time (jehož význam je popsán v podkapitole 4.4.3), seznam unikátních hodnot atributů atd. Ukázka pohledu je na obrázku D.2.

All pairs

Pohled obsahuje jeden graf pro každou komunikační dvojici v datové sadě, které jsou vždy zobrazeny všechny (nezávisle na filtrech). Grafy ukazují vývoj jak celkového počtu paketů v čase, tak pro každý směr zvlášť. Pohled pomáhá uživateli určit, která stanice je řídící a ukazuje v jakých časech mezi sebou stanice komunikují. Metoda analýzy komunikace pomocí počtu paketů je popsána v podkapitole 4.4.1. Pohled reaguje pouze na změnu velikosti časového okna a na změnu intervalu, ostatní filtry pohled neovlivňují. Ukázka pohledu je k nalezení v příloze D.3.

Selected slaves

Čtvrtý pohled poskytuje uživateli náhled na průběh komunikace řídící stanice s jednotlivými podřízenými stanicemi.V horní části pohledu se nachází spojnicový graf, který vychází z metody popsané v podkapitole 4.4.2. Pod grafem se nachází tabulka, která ukazuje přesné hodnoty hranic (viz 6.1). Ukázka pohledu je na obrázku D.4.

Adresa podřízené stanice	**Čas prvního paketu**	**Čas posledního paketu**	**Délka trvání**	**Počet paketů**
192.168.11.11:49784	13:03:10.31	13:20:03.94	00:16:53.63	9905
192.168.11.11:49830	13:20:04.97	13:29:26.00	00:09:21.03	3011
192.168.11.11:49849	13:29:27.03	17:56:31.33	04:27:04.30	91617

Tabulka 6.1: Ukázka tabulky na komunikaci z datové sady C. Řádky reprezentují vlastnosti komunikace řídící stanice `192.168.11.248:2404` s jednotlivými podřízenými stanicemi.

Attribute table

Pokud je uživatelem vybrán atribut, zobrazí se v tomto pohledu tabulka časových oken, jejíž podoba je naznačena v tabulce 6.2. Pohled reaguje na všechna možná nastavení a jeho ukázka je v příloze D.5.

	46	**67**	**82**	⋯
17:15	75	46	0	
17:20	25	60	2	
17:25	24	73	0	
⋮				

Tabulka 6.2: Ukázka tabulky časových oken na datech z datové sady A. Na vertikální „ose“ jsou vyznačeny začátky časových oken, na horizontální „ose“ jsou vyznačeny vybrané hodnoty atributu *ipLen*. Vnitřní buňky tabulky značí počet odchycených paketů v daném intervalu s danou hodnotou atributu.

Attribute statistics

Poslední pohled nabízí bližší pohled na vybraný zkoumaný atribut. Přímo vychází z metody popsané v podkapitole 4.4.4. V horní části pohledu je vygenerován spojnicový graf, který

ukazuje počty paketů v časových oknech pro každou ze zvolených hodnot zkoumaného atributu. V dolní části pohledu se nachází přehledová tabulka, která ukazuje pro každou hodnotu atributu některé jeho míry polohy a variability. Popisné charakteristiky jsou počítány z počtu zachycených paketů v jednotlivých časových oknech. Hlavička přehledové tabulky je demonstrována tabulkou 6.3. Pohled reaguje na všechna možná nastavení. Ukázka viz D.6.

Attr. value	μ	σ	σ^2	q_1	q_2	q_3	IQR	$\mu - 3\sigma$	$\mu + 3\sigma$	Interval size	Outliers
...	...	...	...	...	...	...	...	...	...	...	...

Tabulka 6.3: Ukázka hlavičky přehledové tabulky. Sloupec *Interval size* reprezentuje velikost rozdílu $(\mu + 3\sigma) - (\mu - 3\sigma)$ a sloupec *Outliers* počet hodnot, které spadají mimo interval 3σ.

Kapitola 7

Experimenty

Cílem experimentů je ukázat, že lze aplikaci *ICS Analyzer* využít k nalezení stabilních charakteristik a anomálií v komunikaci. Pro experimenty bude využita datová sada B, obsahující běžnou komunikaci (viz 4.1), a 6 datových sad s útoky, jejichž popis je uveden v podkapitole 4.2. V první řadě je potřeba v běžné komunikaci najít takové typy paketů, jejichž počet zachycení v čase vykazuje známky stability. Typy paketů budou rozlišovány pomocí různých hodnot jednotlivých atributů. Průběh výběru relevantních typů je popsán v podkapitole 7.1. Dalším krokem je použití aplikace k detekci anomálií v datových sadách s útoky, ke které by ideálně měla být využita znalost stabilních charakteristik. Výsledky analýzy datových sad s útoky jsou uvedeny v podkapitole 7.2.

7.1 Hledání stabilních atributů

Stabilním atributem se v této práci rozumí konkrétní hodnota atributu, jejíž počty se v jednotlivých časových oknech výrazně nemění. Pro stabilní atributy je typické, že hodnota jejich rozptylu není příliš velká. Je však problematické určit hraniční hodnotu rozptylu tak, aby se o atributu dalo říct, že je dostatečně stabilní. Tuhle hranici je většinou potřeba určit experimentálně. Obecně lze požadavky na hledané atributy shrnout těmito body:

1. **Nízký rozptyl:** atribut by neměl mít příliš vysoký rozptyl, jelikož detekce odlehlých hodnot je pak příliš benevolentní a nezachytí vše co by měla. Ideálně by interval $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ (dále jen „interval 3σ“) měl být co nejmenší, tak aby byl detekovatelný i malý výkyv hodnot.

2. **Příslušnost k intervalu 3σ:** v ideálním případě by do intervalu 3σ měly padnout všechny hodnoty z normální komunikace. Většinou je však nutné připustit i několik málo hodnot, které padnou mimo interval. V případě, že by už v běžné komunikaci padalo mnoho hodnot mimo interval 3σ, vznikaly by v analýze falešné pozitivní hodnoty (tj. normální hodnoty označené jako anomálie).

3. **Kladná hodnota spodní hranice intervalu 3σ:** obecně není příliš vhodné volit atributy, které se pohybují příliš „nízko“ tzn. jejich střední hodnota se blíží nule. U takových atributů je obtížné detekovat výpadky. Proto minimálním požadavkem je aby byla spodní hranice intervalu 3σ větší než 0.

Dobrou praktikou je neomezovat se pouze na „nejstabilnější“ atribut v komunikaci, ale uvažovat i ty méně stabilní. Různé druhy útoků a anomálií mohou ovlivňovat různé atributy.

Nalezené stabilní atributy

Pomocí aplikace *ICS Analyzer* byly v datové sadě B nalezeny atributy, které splňují výše popsané požadavky. Tyto atributy, jejich střední hodnota, rozptyl a interval 3σ jsou uvedeny v tabulce 7.1. Mezi atributy *ipLen* a *len* byla nalezena vysoká korelace, proto jsou jejich hodnoty v tabulce uvedeny společně. Dále budou tyto atributy označovány pojmem „nalezené stabilní atributy“.

Atribut	Hodnota Atributu	Směr komunikace	μ	σ	$\mu - 3\sigma$	$\mu + 3\sigma$	Velikost intervalu 3σ
fmt	0x01	obousměrná S2M	19.54	2.43	12.27	26.82	14.55
ipLen	46	obousměrná	26.10	2.59	18.33	33.87	15.53
len	4						
ipLen	46	S2M	22.82	1.14	19.41	26.23	6.82
len	4						

Tabulka 7.1: Nalezené atributy vykazující stabilitu a jejich základní popisné charakteristiky.

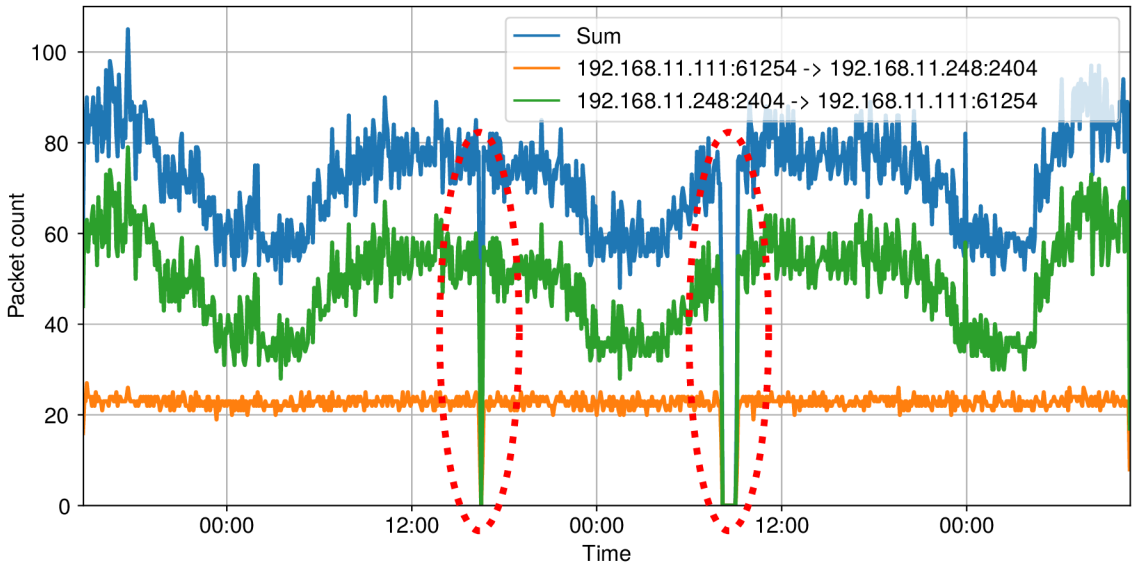
7.2 Detekce útoků na dostupných sadách

Tato podkapitola se zabývá použitím aplikace *ICS Analyzer* k detekci útoků v datových sadách z podkapitoly 4.2. K nalezeným stabilním atributům z podkapitoly 7.1 je v experimentech věnována zvýšená pozornost. Není však pravidlem, že by díky těmto atributům bylo vždy možné jednoznačně detekovat útok. Všechny grafy v této podkapitole jsou vygenerovány aplikací *ICS Analyzer* s velikostí časového okna 5 minut a pro obousměrnou komunikaci.

Datová sada `connection-loss.csv`

Výpadek komunikace je velmi snadno odhalitelný již z grafu ukazující počet paketů v čase. Takový graf je dostupný v pohledu *All pairs* (je jen jeden, protože se v komunikaci nachází pouze jedna komunikační dvojice) a jeho podoba je ukázána na obrázku 7.1.

Dalším ukazatelem, který napovídá tomu, že v komunikaci došlo k výpadku, je vysoká maximální hodnota *inter-arrival time*, která je dle statistik aplikace 3637,29 s (údaj je dostupný v pohledu *General statistics*).

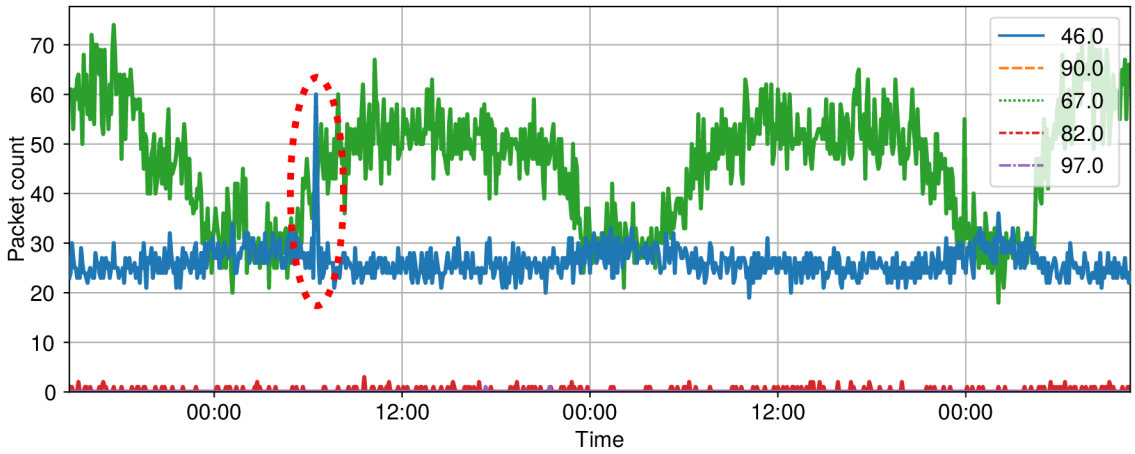


Obrázek 7.1: Graf počtu paketů podle směrů. Z grafu lze jasně vidět, že v komunikaci došlo ke dvěma výpadkům. V obou případech došlo k výpadku v obou směrech komunikace.

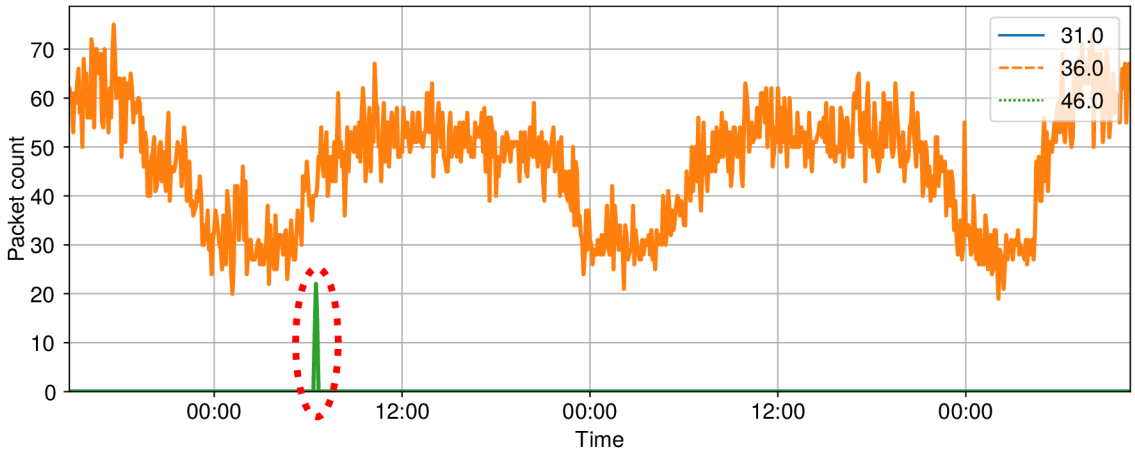
Datová sada `switching-attack-mod.csv`

Poznámka: Originální datová sada musela být pro účely použití v aplikaci mírně upravena. Aplikace předpokládá časovou souslednost časových razítek, která však v původní datové sadě `switching-attack.csv` *není dodržena. Z tohoto důvodu je následující útok popsán na upravené datové sadě* `switching-attack-mod.csv`*, ve které došlo k mírné úpravě některých časových razítek.*

Všechny vybrané stabilní atributy vykazují známky kladného vychýlení v době útoku. Nejvýraznějšího vychýlení dosahuje atribut *ipLen* s hodnotou 46 (viz obrázek 7.2a). Útok lze však zpozorovat i díky jiným ukazatelům. V komunikaci se objevují nové hodnoty atributů, které se v běžné komunikaci (v sadě B) vůbec neobjevují. Konkrétně se jedná o *asduType* s hodnotou 46 a *cot* s hodnotami 6 a 7. Na obrázku 7.2b je ukázáno neočekávané zachycení paketů s hodnotou atributu *asduType* 46. Tento atribut je typickým případem atributu, který by nesplnil podmínku č. 3 uvedenou v podkapitole 7.1 a mohl by být zavržen ještě před počátkem experimentu. Experiment však ukazuje, že k detekci útoku lze použit i takovéto atributy a to díky jejich možnému kladnému vychýlení.



(a) Graf vývoje toku jednotlivých hodnot atributu *ipLen*. Během útoku došlo k výraznému vychýlení u hodnoty atributu *ipLen* 46 (modrá křivka).



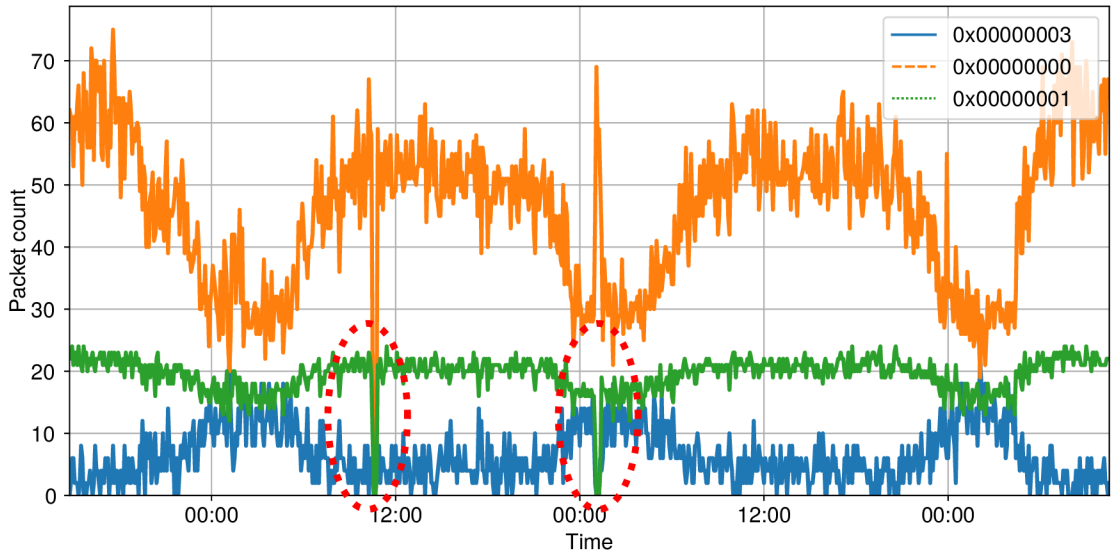
(b) Graf vývoje toku jednotlivých hodnot atributu *asduType*. Během útoku došlo k zachycení podezřelých paketů s hodnotou atributu *asduType* 46.

Obrázek 7.2: Grafy ukazující vliv útoku na tok hodnot atributů *ipLen* a *asduType*.

Datová sada `scanning-attack.csv`

Poznámka: Před zahájením analýzy bylo potřeba nastavit v aplikaci adresu řídící stanice. Automatická detekce v tomhle případě chybně zvolila jednu z podvržených adres.

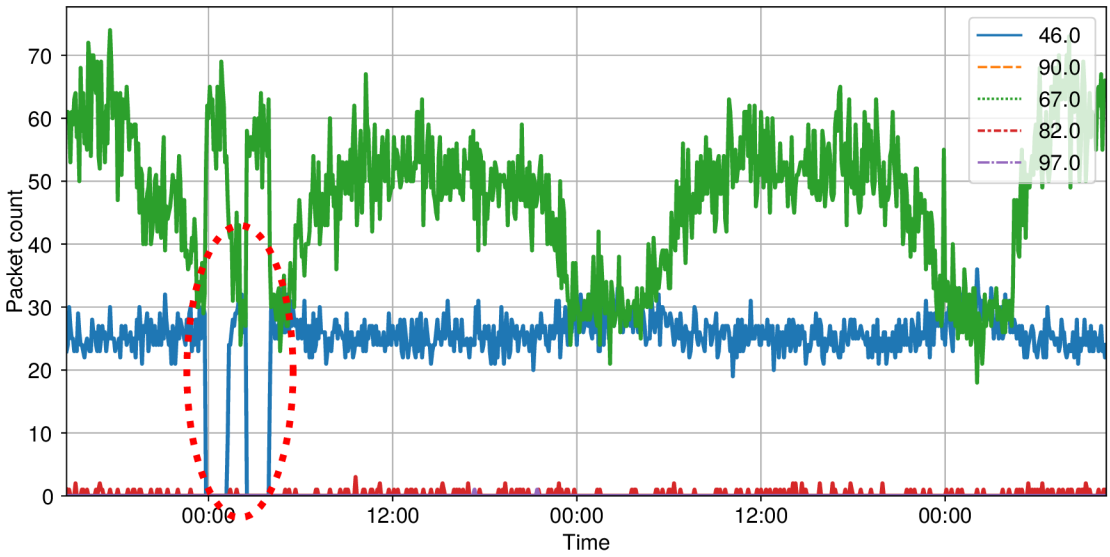
Oba dva typy útoků (horizontální a vertikální skenování) lze jednoduše detekovat pomocí všech vybraných stabilních atributů. Během útoků se objevují časová okna, kde počet zachycených paketů s danými hodnotami atributů je nulový. Na obrázku 7.3 je ukázka propadu pro hodnotu atributu *fmt* `0x01`.



Obrázek 7.3: Graf vývoje toku jednotlivých hodnot atributu *fmt*. Během obou útoků došlo k výpadku paketů s hodnotou atributu `0x01`. Anomálie lze vypozorovat i u paketů s hodnotou atributu `0x00`, kdy u horizontálního útoku došlo k propadu, kdežto u vertikálního útoku došlou k prudkému nárůstu. Zatímco u horizontálního útoku lze pomocí aplikace nalézt výpadek mimo interval 3σ, u vertikálního útoku tomu již tak není. Zde se ukazuje proč atribut s velkým rozptylem není vhodným kandidátem pro detekci anomálií.

Datová sada `dos-attack.csv`

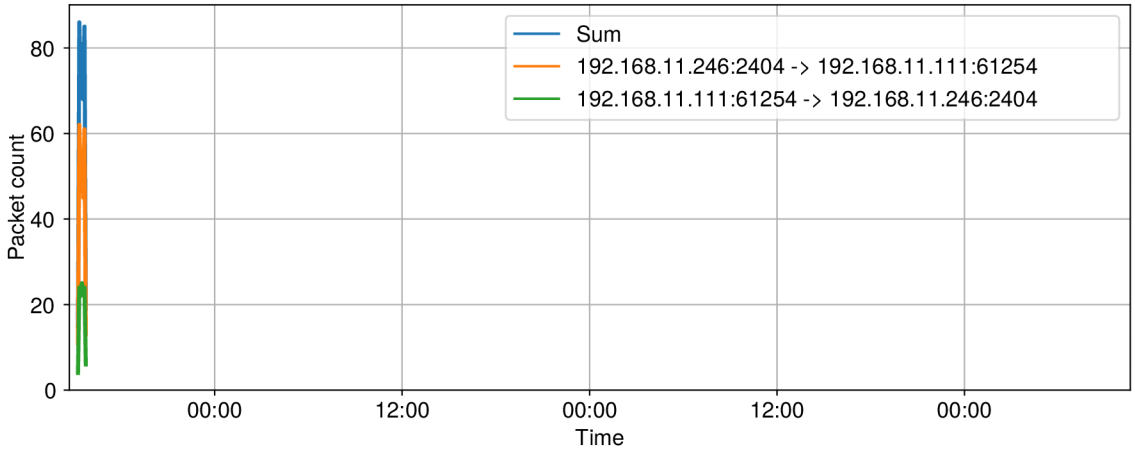
Útok v této datové sadě je snadno detekovatelný pomocí všech vybraných stabilních atributů. Pro demonstraci je vybrán atribut *ipLen* s hodnotou 46. Aplikace hlásí výpadek hodnoty mimo interval 3σ celkem v 35 časových oknech (což odpovídá 2 hodinám a 55 minutám). Z grafu na obrázku 7.4 je pak útok jasně viditelný.



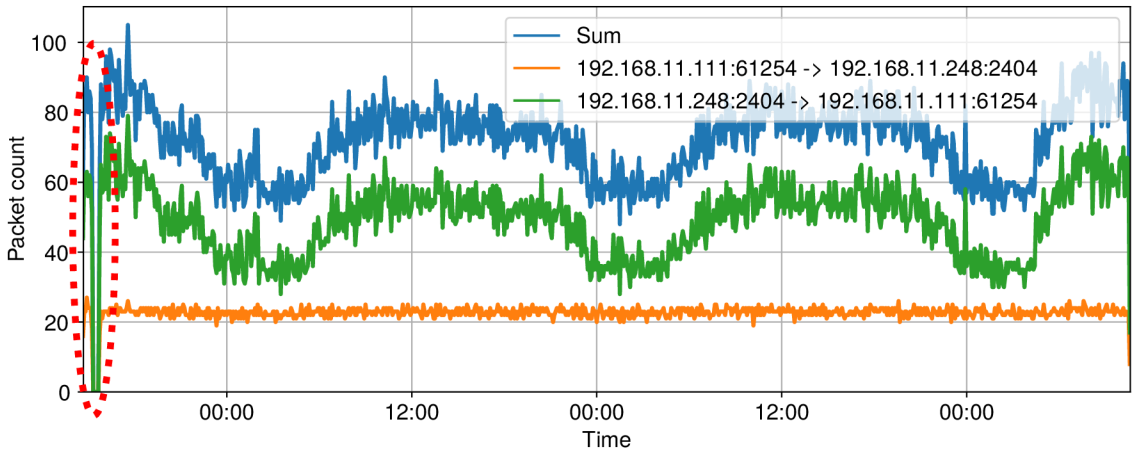
Obrázek 7.4: Graf vývoje toku jednotlivých hodnot atributu *ipLen*. Během útoku nebyly zachyceny žádné pakety s hodnotou *ipLen* 46. Naopak paketů s hodnotu 67 bylo zachyceno více než napovídá periodický průběh zelené křivky v grafu. Aplikace však pro hodnotu 67 nehlásí žádné výpadky mimo interval 3σ a odhalení je tak možné pouze na základě vizualizace dat grafem.

Datová sada `rogue-devices.csv`

Útok v této datové sadě lze v aplikaci vypozorovat z grafů pro počty zachycených paketů v jednotlivých komunikačních dvojicích (dostupných v pohledu *All pairs*). Je velmi podezřelé, že se v komunikaci nachází 2 řídící stanice (označené portem 2404). Z grafů na obrázku 7.5 lze navíc vidět, že stanice s adresou `198.168.11.246:2404` komunikuje jen velmi krátce.



(a) Komunikace podvržené řídící stanice.



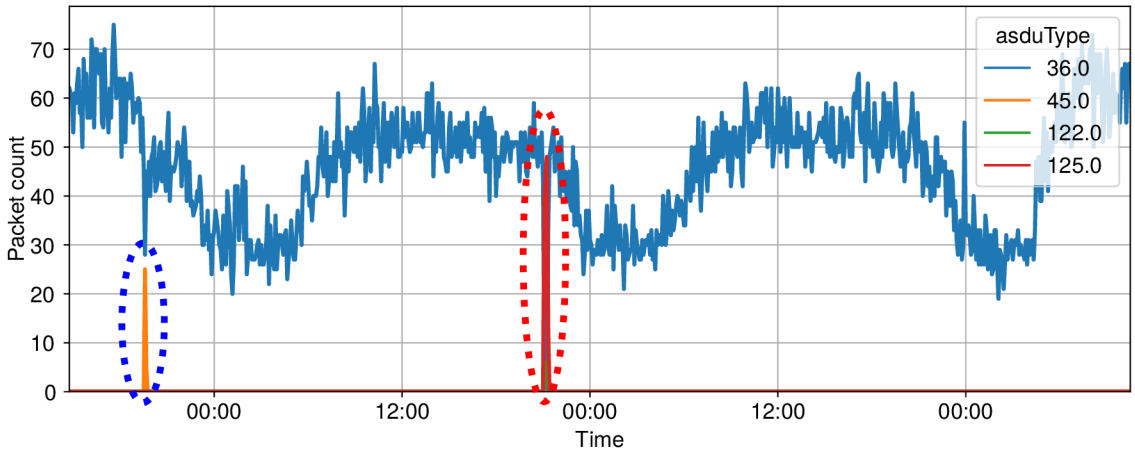
(b) V době útoku dochází k výpadku komunikace mezi opravdovou řídící stanicí a podřízenou stanicí.

Obrázek 7.5: Grafy počtů paketů všech komunikačních dvojic v datové sadě.

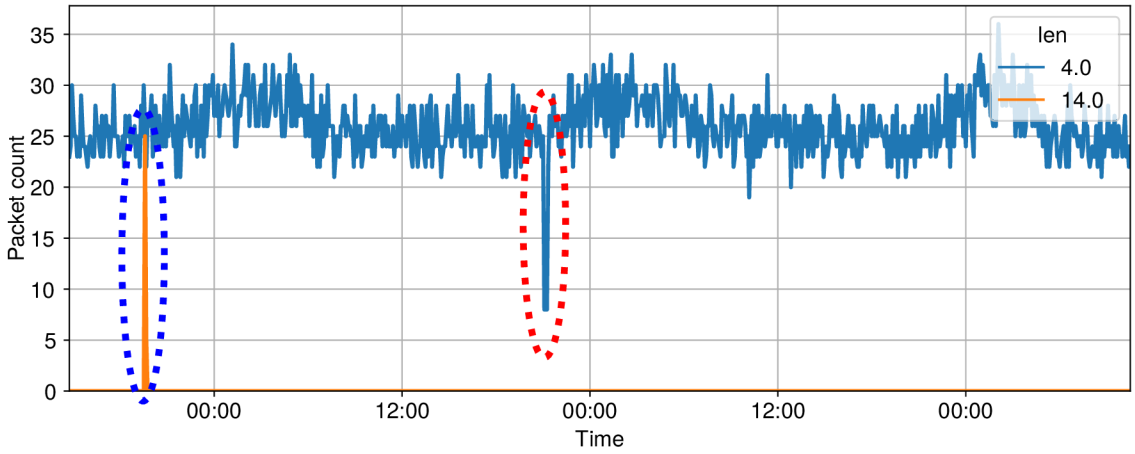
Datová sada `injection-attack.csv`

V komunikaci se objevují 2 útoky. První z nich není detekovatelný vybranými stabilními atributy. V komunikaci se však nacházejí jiné ukazatele, pomocí kterých lze podezřelou aktivitu odhalit. U hodnot atributu *asduType* 45 a *len* 14 dochází k prudkému nárůstu počtu zachycených paketů s danými vlastnostmi. Nárůsty jsou na obrázku 7.6 zakroužkovány modrou barvou.

Druhý útok již je detekovatelný pomocí atributu *len* s hodnotou 4. Podobně jako u jiných typů útoků, dochází opět k propadu počtu detekovaných paketů. V komunikaci lze navíc pozorovat podezřele zvýšenou aktivitu pro hodnoty atributu *asduType* 122 a 125. Zmíněné anomálie jsou na obrázku 7.6 zakroužkovány červenou barvou.



(a) Graf vývoje toku vybraných hodnot atributu *asduType*. V prvním útoku dochází k nečekanému nárůstu paketů s hodnotou 45. V druhém útoku pak dochází k ještě prudšímu nárůstu paketů s hodnotami 122 a 125. Zároveň dochází k poklesu paketů s hodnotou 36, což je ovšem v tomto grafu překryto jinými křivkami (v aplikaci by se použila filtrace hodnot atributů).



(b) Graf vývoje toku vybraných hodnot atributu *len*. V prvním útoku dochází k prudkému nárůstu paketů s hodnotou 14. U druhého útoku naopak dochází k poklesu paketů s hodnotou 4.

Obrázek 7.6: Grafy ukazující vliv útoků na tok hodnot atributu *asduType* a *len*. Pro zvýšení přehlednosti grafů jsou zobrazeny pouze hodnoty atributů relevantní pro detekci útoků.

Kapitola 8

Závěr

Cílem práce bylo navrhnout a naimplementovat aplikaci pro statistickou analýzu průmyslové komunikace využívající protokolu IEC 104. Před započetím vývoje aplikace bylo potřeba nastudovat celkově tři velké tématické okruhy. Prvním z nich byla tématika ICS komunikace a protokolu IEC 104, která byla důležitá pro porozumění obsahu datových sad. Dalším okruhem byly metody pro statistický popis, analýzu a zpracování dat. A to jak z teoretického hlediska (charakteristiky míry a variability), tak i z praktického (knihovny *Pandas*, *Seaborn*, ...). Poslendím okruhem byla problematika návrhu a vývoje grafického uživatelského rozhraní pomocí framworku *PyQt6* pro jazyk *Python*.

Navržená aplikace *ICS Analyzer* umožňuje uživateli načíst datovou sadu se záznamem průmyslové komunikace ve formátu `csv`. Uživatel může před provedením analýzy datovou sadu vyfiltrovat na základě pěti kritérií. Aplikace poskytuje mimo obecných statistik datové sady také vhledy, díky kterým může uživatel nalézt stabilní prvky v komunikaci. Zejména potom stabilita v počtu zachycených paketů určitého typu v čase se jeví jako dobrý ukazatel pro detekci různých druhů anomálií.

Prostoru pro budoucí rozšíření aplikace je poměrně hodně. V první řadě by bylo dobré do aplikace přidat novou kartu, která by umožňovala zkoumat korelaci mezi různými atributy datové sady. Dále by aplikace mohla nabízet nové filtrovací kritérium, které by bylo založené na hodnotě *inter-arrival time*. Mezi menší možné úpravy patří přidání možnosti volby počátečního data načítané datové sady a dopočet relativního času z časového razítka.

Na závěr bych chtěl dodat, že si z práce odnáším mnoho cenných znalostí, a to zejména na poli zpracování a analýzy dat, kterému bych se chtěl v buoucím profesním životě blíže věnovat.

Literatura

[1] Burgetová, I., Matoušek, P. a Ryšavý, O. Anomaly Detection of ICS Communication Using Statistical Models. In: *2021 17th International Conference on Network and Service Management (CNSM)*. 2021, s. 166–172. DOI: 10.23919/CNSM52442.2021.9615510.

[2] Clarke, G., Reynders, D. a Wright, E. *Practical Modern SCADA Protocols : DNP3, 60870. 5 and Related Systems*. 1. vyd. Saint Louis, MO: Elsevier Science & Technology, 2004. 170 s. ISBN 075067995.

[3] Frost, J. *Interquartile Range (IQR): How to Find and Use It* [online]. [cit. 2022-04-28]. Dostupné z: https://statisticsbyjim.com/basics/interquartile-range/.

[4] Frost, J. *Measures of Variability: Range, Interquartile Range, Variance, and Standard Deviation* [online]. [cit. 2022-04-28]. Dostupné z: https://statisticsbyjim.com/basics/variability-range-interquartile-variance-standard-deviation/.

[5] György, P. a Holczer, T. Attacking IEC 60870-5-104 Protocol. *CEUR WORKSHOP PROCEEDINGS* [online]. 2021, sv. 2874, s. 140–150, [cit. 2022-04-10]. Dostupné z: https://m2.mtmt.hu/api/publication/32062985.

[6] Košťáková, T. *O složitém jednoduše, aneb, Nebojte se statistiky, nekouše*. 1. vyd. Praha, CZ: Český statistický úřad, 2019. ISBN 978-80-250-2908-4.

[7] Lewinson, E. *Violin plots explained* [online]. říjen 2019 [cit. 2022-04-30]. Dostupné z: https://towardsdatascience.com/violin-plots-explained-fb1d115e023d.

[8] Matoušek, P. *Description and analysis of IEC 104 Protocol*. 2017. 38 s. Dostupné z: https://www.fit.vut.cz/research/publication/11570.

[9] Matoušek, P. *Datasets* [https://github.com/matousp/datasets]. GitHub, 2020 [cit. 2022-04-30].

[10] Nicola, M., Nicola, C.-I., Duță, M. et al. SCADA Systems Architecture Based on OPC and Web Servers and Integration of Applications for Industrial Process Control. *International Journal of Control Science and Engineering* [online]. 2018, sv. 8, č. 1, s. 13–21, [cit. 2022-04-10]. DOI: 10.5923/j.control.20180801.02. Dostupné z: http://article.sapub.org/10.5923.j.control.20180801.02.html.

[11] Pavlík, T. a Dušek, L. *Biostatistika* [online]. Multimediální podpora výuky klinických a zdravotnických oborů :: Portál Lékařské fakulty Masarykovy univerzity, 2012 [cit. 2022-04-30]. ISSN 1801-6103. Dostupné z: https://portal.med.muni.cz/clanek-590-biostatistika.html.

[12] Radoglou Grammatikis, P., Sarigiannidis, P., Giannoulakis, I. et al. Attacking IEC-60870-5-104 SCADA Systems. In: *2019 IEEE World Congress on Services (SERVICES)*. 2019, 2642-939X, s. 41–46. DOI: 10.1109/SERVICES.2019.00022.

[13] Skiena, S. S. *The Data Science Design Manual*. 1. vyd. New York, NY: Springer, 2017. ISBN 978-3-319-55443-3.

[14] Stouffer, K., Lightman, S., Pillitteri, V. et al. Guide to Industrial Control Systems (ICS) Security. [online]. Revision 2. květen 2015, [cit. 2022-04-07]. DOI: 10.6028/NIST.SP.800-82r2. Dostupné z: https://csrc.nist.gov/publications/detail/sp/800-82/rev-2/final.

[15] Toews, M. W. *Standard deviation diagram*. 2007 [cit. 2022-04-26]. File: `Standard deviation diagram.svg`. Dostupné z: https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg.

[16] Záhora, J. *Učebnice statistiky*. Univerzita Karlova v Praze, Lékařská fakulta v Hradci Králové, 2015. ISBN 978-80-88176-00-8.

[17] *Intro to data structures* [online]. [cit. 2022-04-28]. Dostupné z: https://pandas.pydata.org/docs/user_guide/dsintro.html#dataframe.

[18] *Míry polohy* [online]. 2022 [cit. 2022-04-26]. ISSN 1804-6517. Dostupné z: https://www.wikiskripta.eu/index.php?title=M%C3%ADry_polohy&oldid=452356.

[19] *Míry variability* [online]. 2022 [cit. 2022-04-28]. ISSN 1804-6517. Dostupné z: https://www.wikiskripta.eu/w/M%C3%ADry_variability.

Příloha A

Typy ASDU

Code	Description	Valid COTs
31	Double point information with time tag	3,5,11,12
36	Measured value, short floating point value with time tag	2,3,5,11,12,20,20+G
45	Single command	6,7,8,9,10,44,45,46,47
46	Double command	6,7,8,9,10,44,45,46,47
100	(General-) Interrogation command	6,7,8,9,10,44,45,46,47
120	File ready	13
121	Section ready	13
122	Call directory, select file, call file, call section	5,13
123	Last section, last segment	13
124	Ack file, Ack section	13
125	Segment	13

Tabulka A.1: Vybrané kódy určující typ ASDU jednotky [8].

Příloha B

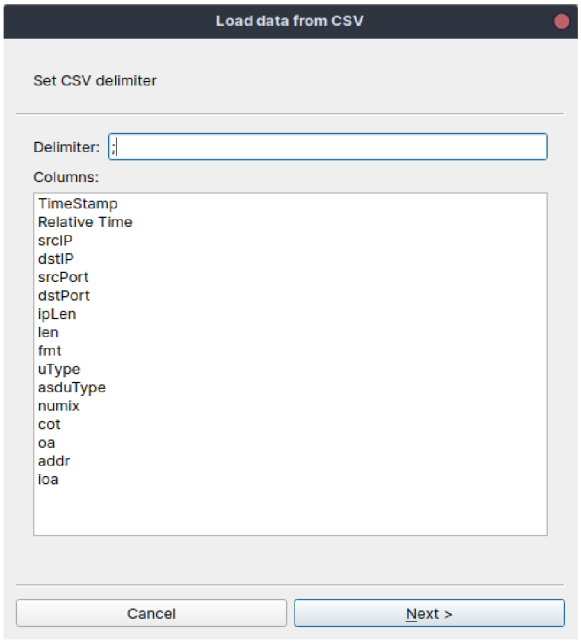
Kódy COT

Code	Cause of Transmission	Abbreviation
1	periodic, cyclic	per/cyc
2	background interrogation	back
3	spontaneous	spont
4	initialized	init
5	interrogation or interrogated	req
6	activation	act
7	confirmation activation	actcon
8	deactivation	deact
9	confirmation deactivation	deactcon
10	termination activation	actterm
11	feedback, caused by distant command	retrem
12	feedback, caused by local command	retloc
13	data transmission	file
14–19	reserved for further compatible definitions	
20	interrogated by general interrogation	inrogen
21	interrogated by interrogation group 1	inro1
22	interrogated by interrogation group 2	inro2
	...	
36	interrogated by interrogation group 16	inro16
37	interrogated by counter general interrogation	reqcogen
38	interrogated by interrogation counter group 1	reqco1
39	interrogated by interrogation counter group 2	reqco2
	...	
44	type-Identification unknown	uknown_type
45	cause unknown	uknown_cause
46	ASDU address unknown	unknown_asdu_address
47	Information object address unknown	unknown_object_address

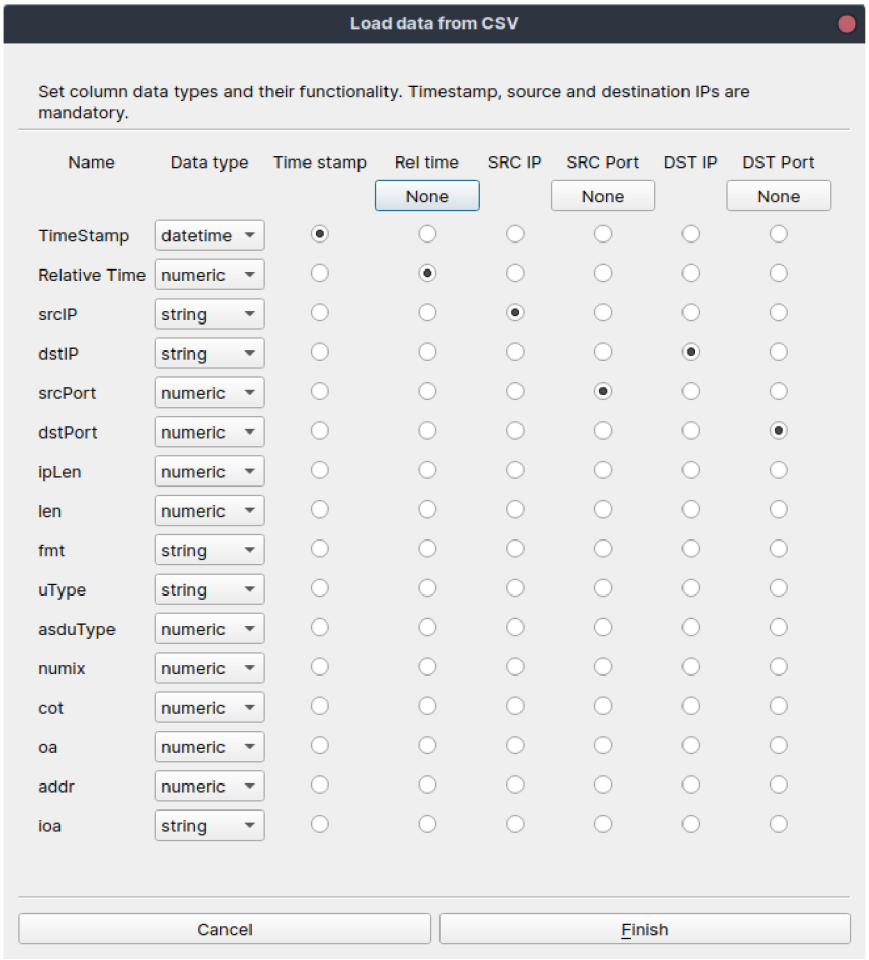
Tabulka B.1: Kódy COT a jejich význam [8].

Příloha C

Dialog pro načtení CSV souboru



Obrázek C.1: Výběr oddělovače a zobrazení názvů sloupců, které by vznikly použitím zvoleného oddělovače.

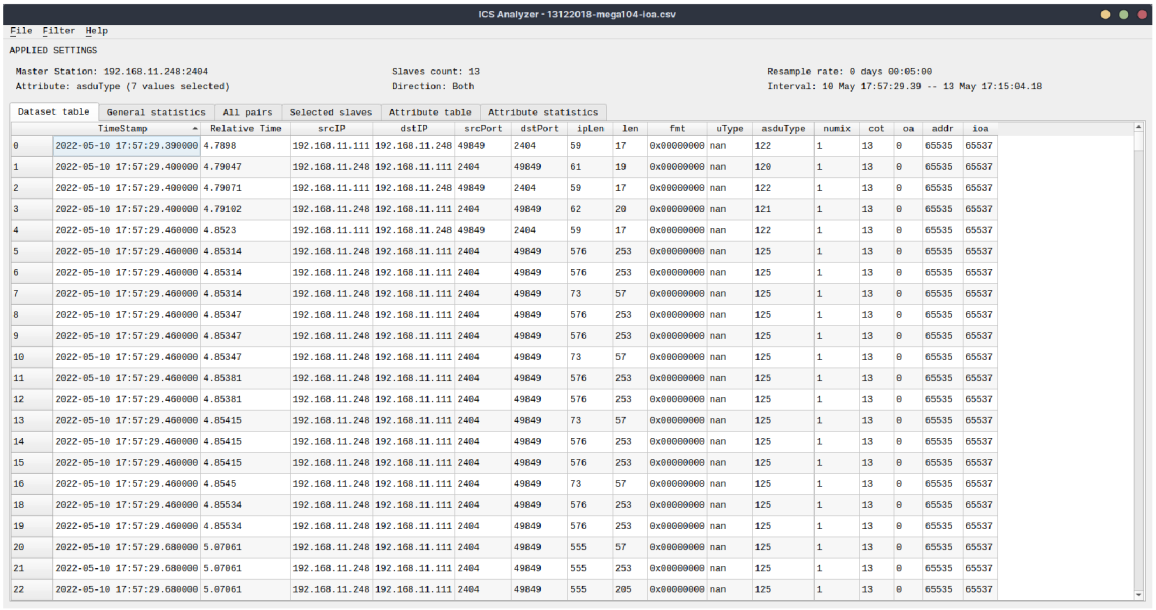


Obrázek C.2: Výběr datových typů sloupců a sloupců reprezentujících speciální atributy.

Příloha D

Pohledy aplikace

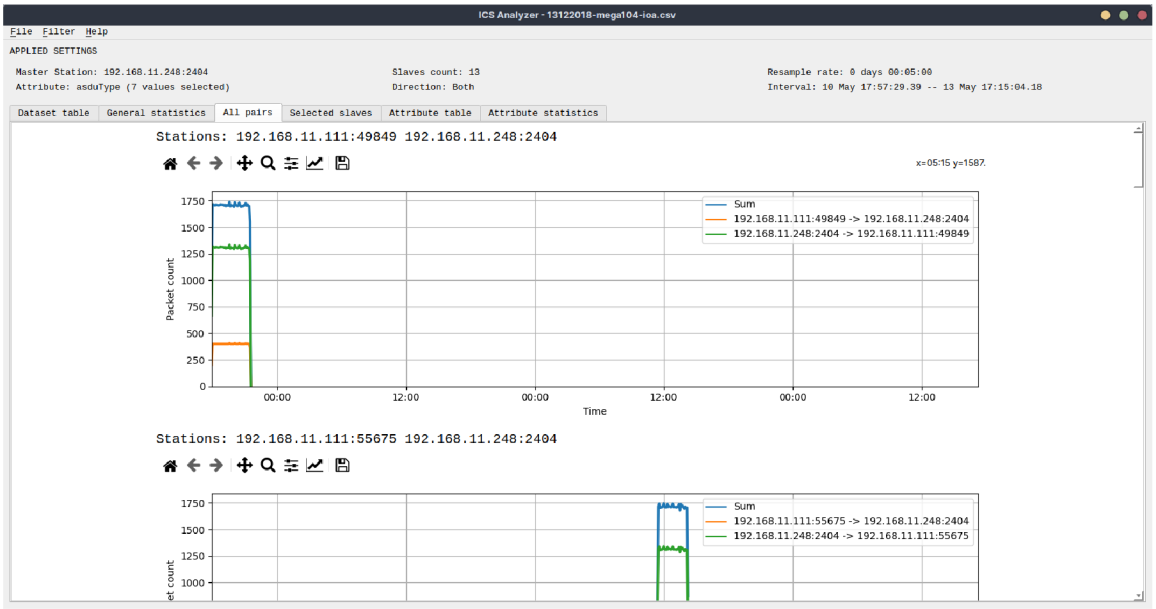
Ukázka všech pohledů aplikace s načtenou datovou sadou D.



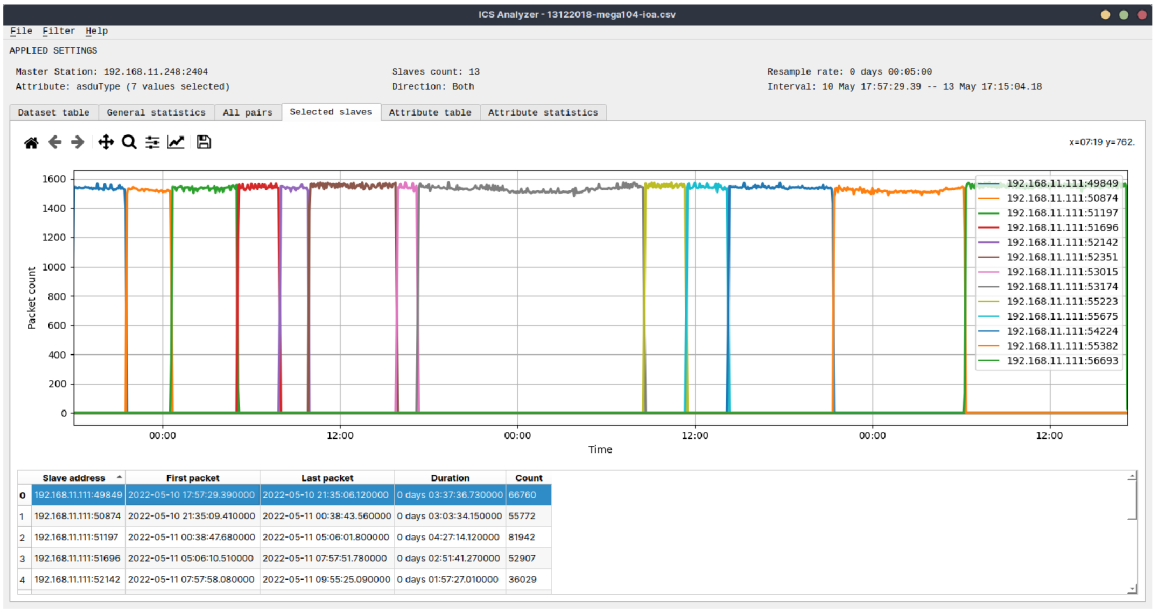
	TimeStamp	Relative Time	srcIP	dstIP	srcPort	dstPort	ipLen	len	fmt	uType	asduType	numix	cot	oa	addr	ioa
0	2022-05-10 17:57:29.390000	4.7898	192.168.11.111	192.168.11.248	49849	2404	59	17	0x00000000	nan	122	1	13	0	65535	65537
1	2022-05-10 17:57:29.400000	4.79047	192.168.11.248	192.168.11.111	2404	49849	61	19	0x00000000	nan	120	1	13	0	65535	65537
2	2022-05-10 17:57:29.400000	4.79071	192.168.11.111	192.168.11.248	49849	2404	59	17	0x00000000	nan	122	1	13	0	65535	65537
3	2022-05-10 17:57:29.400000	4.79102	192.168.11.248	192.168.11.111	2404	49849	62	20	0x00000000	nan	121	1	13	0	65535	65537
4	2022-05-10 17:57:29.460000	4.8523	192.168.11.111	192.168.11.248	49849	2404	59	17	0x00000000	nan	122	1	13	0	65535	65537
5	2022-05-10 17:57:29.460000	4.85314	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
6	2022-05-10 17:57:29.460000	4.85314	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
7	2022-05-10 17:57:29.460000	4.85314	192.168.11.248	192.168.11.111	2404	49849	73	57	0x00000000	nan	125	1	13	0	65535	65537
8	2022-05-10 17:57:29.460000	4.85347	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
9	2022-05-10 17:57:29.460000	4.85347	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
10	2022-05-10 17:57:29.460000	4.85347	192.168.11.248	192.168.11.111	2404	49849	73	57	0x00000000	nan	125	1	13	0	65535	65537
11	2022-05-10 17:57:29.460000	4.85381	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
12	2022-05-10 17:57:29.460000	4.85381	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
13	2022-05-10 17:57:29.460000	4.85415	192.168.11.248	192.168.11.111	2404	49849	73	57	0x00000000	nan	125	1	13	0	65535	65537
14	2022-05-10 17:57:29.460000	4.85415	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
15	2022-05-10 17:57:29.460000	4.85415	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
16	2022-05-10 17:57:29.460000	4.8545	192.168.11.248	192.168.11.111	2404	49849	73	57	0x00000000	nan	125	1	13	0	65535	65537
18	2022-05-10 17:57:29.460000	4.85534	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
19	2022-05-10 17:57:29.460000	4.85534	192.168.11.248	192.168.11.111	2404	49849	576	253	0x00000000	nan	125	1	13	0	65535	65537
20	2022-05-10 17:57:29.680000	5.07061	192.168.11.248	192.168.11.111	2404	49849	555	57	0x00000000	nan	125	1	13	0	65535	65537
21	2022-05-10 17:57:29.680000	5.07061	192.168.11.248	192.168.11.111	2404	49849	555	253	0x00000000	nan	125	1	13	0	65535	65537
22	2022-05-10 17:57:29.680000	5.07061	192.168.11.248	192.168.11.111	2404	49849	555	205	0x00000000	nan	125	1	13	0	65535	65537

Obrázek D.1: Dataset table.

Obrázek D.2: General statistics

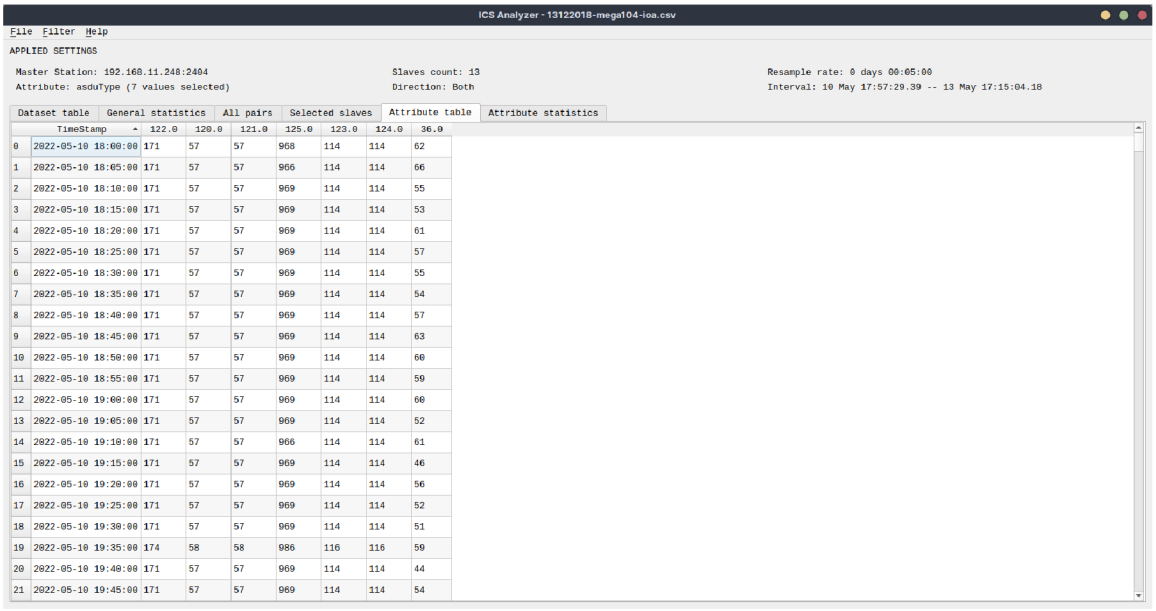


Obrázek D.3: All pairs



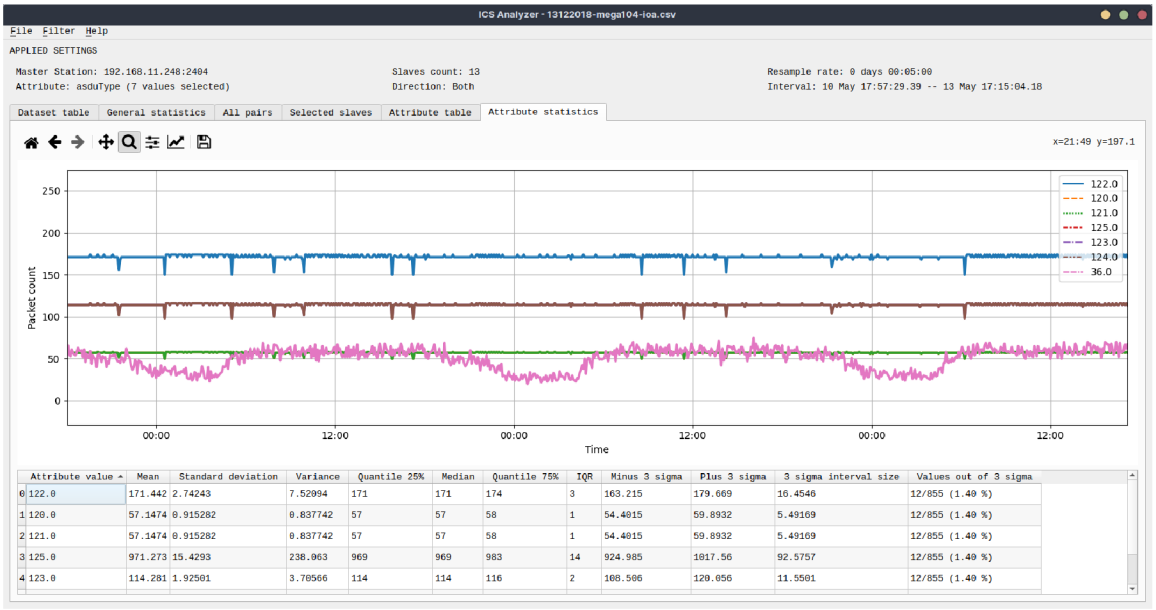
	Slave address	First packet	Last packet	Duration	Count
0	192.168.11.111:49849	2022-05-10 17:57:29.390000	2022-05-10 21:35:06.120000	0 days 03:37:36.730000	66760
1	192.168.11.111:50874	2022-05-10 21:35:09.410000	2022-05-11 00:38:43.560000	0 days 03:03:34.150000	55772
2	192.168.11.111:51197	2022-05-11 00:38:47.680000	2022-05-11 05:06:01.800000	0 days 04:27:14.120000	81942
3	192.168.11.111:51696	2022-05-11 05:06:10.510000	2022-05-11 07:57:51.780000	0 days 02:51:41.270000	52907
4	192.168.11.111:52142	2022-05-11 07:57:58.080000	2022-05-11 09:55:25.090000	0 days 01:57:27.010000	36029

Obrázek D.4: Selected slaves



	TimeStamp	122.0	120.0	121.0	125.0	123.0	124.0	36.0
0	2022-05-10 18:00:00	171	57	57	968	114	114	62
1	2022-05-10 18:05:00	171	57	57	966	114	114	66
2	2022-05-10 18:10:00	171	57	57	969	114	114	55
3	2022-05-10 18:15:00	171	57	57	969	114	114	53
4	2022-05-10 18:20:00	171	57	57	969	114	114	61
5	2022-05-10 18:25:00	171	57	57	969	114	114	57
6	2022-05-10 18:30:00	171	57	57	969	114	114	55
7	2022-05-10 18:35:00	171	57	57	969	114	114	54
8	2022-05-10 18:40:00	171	57	57	969	114	114	57
9	2022-05-10 18:45:00	171	57	57	969	114	114	63
10	2022-05-10 18:50:00	171	57	57	969	114	114	60
11	2022-05-10 18:55:00	171	57	57	969	114	114	59
12	2022-05-10 19:00:00	171	57	57	969	114	114	60
13	2022-05-10 19:05:00	171	57	57	969	114	114	52
14	2022-05-10 19:10:00	171	57	57	966	114	114	61
15	2022-05-10 19:15:00	171	57	57	969	114	114	46
16	2022-05-10 19:20:00	171	57	57	969	114	114	56
17	2022-05-10 19:25:00	171	57	57	969	114	114	52
18	2022-05-10 19:30:00	171	57	57	969	114	114	51
19	2022-05-10 19:35:00	174	58	58	986	116	116	59
20	2022-05-10 19:40:00	171	57	57	969	114	114	44
21	2022-05-10 19:45:00	171	57	57	969	114	114	54

Obrázek D.5: Attribute table



	Attribute value	Mean	Standard deviation	Variance	Quantile 25%	Median	Quantile 75%	IQR	Minus 3 sigma	Plus 3 sigma	3 sigma interval size	Values out of 3 sigma
0	122.0	171.442	2.74243	7.52094	171	171	174	3	163.215	179.669	16.4546	12/855 (1.40 %)
1	120.0	57.1474	0.915282	0.837742	57	57	58	1	54.4015	59.8932	5.49169	12/855 (1.40 %)
2	121.0	57.1474	0.915282	0.837742	57	57	58	1	54.4015	59.8932	5.49169	12/855 (1.40 %)
3	125.0	971.273	15.4293	238.063	969	969	983	14	924.985	1017.56	92.5757	12/855 (1.40 %)
4	123.0	114.281	1.92501	3.70566	114	114	116	2	108.506	120.056	11.5501	12/855 (1.40 %)

Obrázek D.6: Attribute statistics