VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

DETEKCE ANOMÁLIÍ V SÍŤOVÉ KOMUNIKACI
ANOMALY DETECTION OF ICS TRAFFIC

PROJEKT - PŘENOS DAT, POČÍTAČOVÉ SÍTĚ A PROTOKOLY
PROJECT - DATA COMMUNICATIONS, COMPUTER NETWORKS AND PROTOCOLS

AUTOR PRÁCE **Bc. RADEK DUCHOŇ**
AUTHOR

BRNO 2022

Abstrakt

Cílem této práce je analyzování síťové komunikace nad protokolem IEC104. Po této analýze se bude věnovat vytvoření a natrénování statistického modelu pro detekci anomálií a na závěr se bude věnovat experimentům s modelem.

Abstract

The aim of this work is to analyze ICS traffic on protocol IEC104. Statistical model for detection of anomalies should be created and trained after that analysis and a conclusion will be experiments with that model.

Klíčová slova

IEC-104, detekce anomálií, statistické modely

Keywords

IEC-104, anomaly detection, statistical models

Citace

DUCHOŇ, Radek. *Detekce anomálií v síťové komunikaci.* Brno, 2022. Projekt - Přenos dat, počítačové sítě a protokoly. Vysoké učení technické v Brně, Fakulta informačních technologií.

Obsah

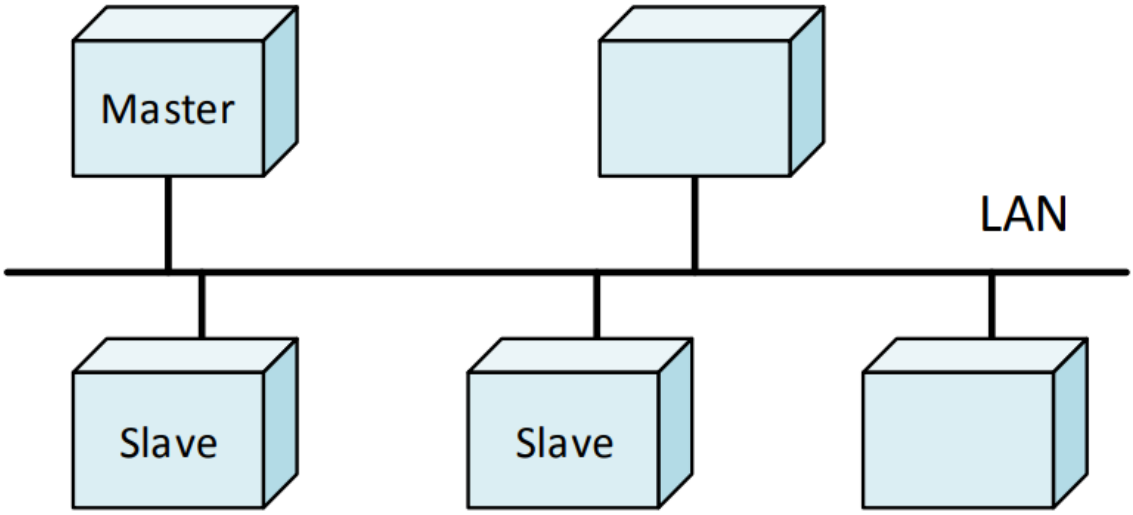
Kapitola 1

Úvod

Cílem této analýza síťové komunikace pro systémy SCADA a následné vytvoření a natrénování statistického modelu pro detekci anomálií.

Konkrétní komunikace, které se tato práce věnuje, je založena na přenosovém protokolu IEC 60870-5-104, který je rozšířením síťového protokolu IEC 60870-5-101. Protokol pracuje nad síťovým protokolem TCP/IP a topologie sítě je typu Master/Slave, viz obrázek 1.1 níže. Řídící stanice (Master) musí v tomto protokolu začínat komunikaci s kontrolovanou stanicí (Slave) a konkrétní přenášená data mají pevně definovanou strukturu.

Struktuře dat v paketech se budu věnovat v sekci 2, kde se budu také věnovat filtrovaní paketů z poskytnutých datasetů pro následnou analýzu. Na závěr se budu věnovat vytvořenému statistickému modelu pro analýzu anomálií v komunikaci.



Obrázek 1.1: Diagram komunikace [2]

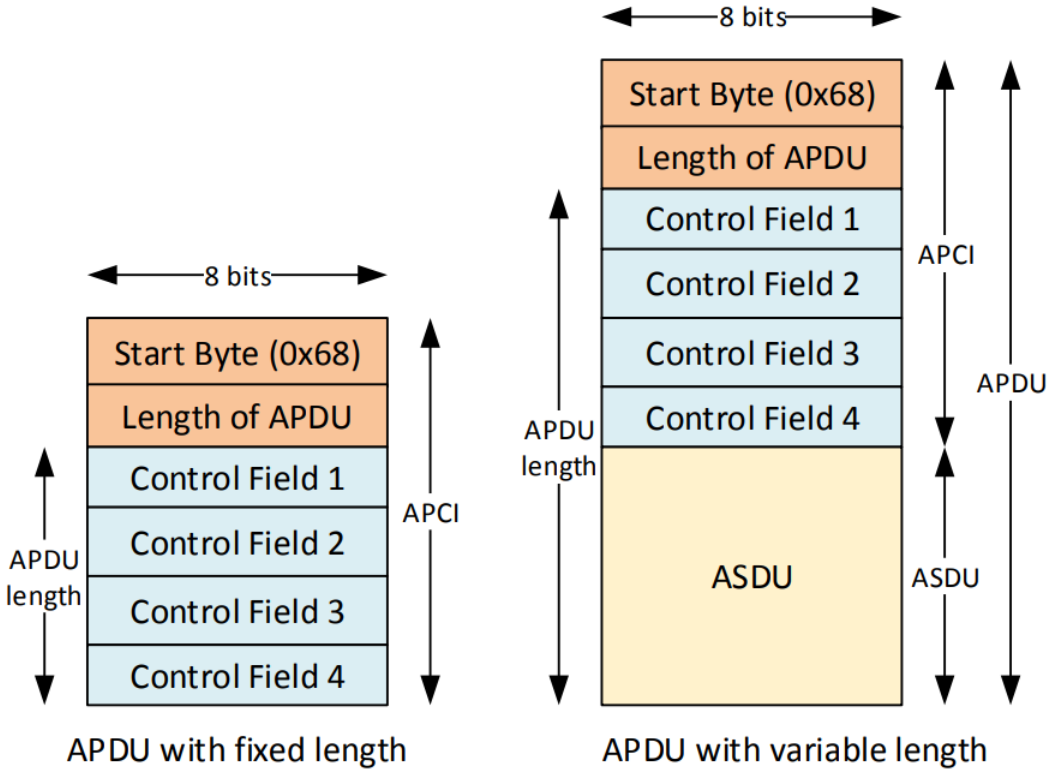
Kapitola 2

Analýza a filtrování packetů

Jelikož datasety, které byly využity v rámci tohoto projektu obsahovaly velké množství paketů, které nebyly vhodné pro následnou analýzu, bylo nutné započít jejich filtrováním, k tomu bylo nutné pochopit strukturu těchto dat a odlišit je od ostatních (např. běžných ACK nebo ICMP) paketů.

2.1 Struktura dat

Jak je vidět na obrázku 2.1, data zde mají jasně definovanou strukturu, kterou lze využít k filtrování paketů, které spadají do IEC-104 komunikace.



Obrázek 2.1: Struktura dat [2]

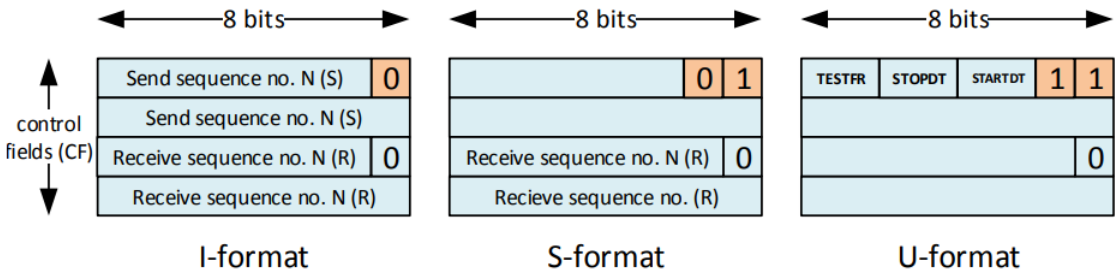
První oktet dat zde začíná vždy pevně definovanou hodnotou `0x68`.

Následuje oktet, v němž se nachází délka APDU, což je jednotka dat aplikačního protokolu (z angl. Application Protocol Data Unit.

Na závěr se zde nachází řídící pole o délce čtyř oktetů, které mohou mít tři různé varianty, tzv. I-format, S-format a U-format, viz obrázek 2.2.

Toto je minimální pevně dná struktura, volitelně může pokračovat dále ASDU, neboli jednotka dat aplikační služby (z anglického Application Service Data Unit).

Do délky APDU zapsané ve druhém oktetu jsou započítány přinejmenším 4 byty řídícího pole, nepočítají se do ní však první dva oktety se start bytem a délkou APDU, dále se do něj počítá ASDU. Minimální hodnota, která zde může být zapsaná je tak 4, maximální hodnota, které může oktet legální nabývat, je 253. Data tak nemohou být delší než 255 bytů celkem.



Obrázek 2.2: Struktura dat [2]

2.2 filtrace komunikace a extrakce dat

Pro filtrování dat z datasetu ve formátu `pcapng` jsem využil knihovnu `scapy` pro jazyk Python 3, která nabízí třídu `PcapNgReader` pro čtení těchto dat, přičemž jsem pro filtraci aplikoval znalosti popsané v minulé sekci 2.1.

Prvním důležitým využitým údajem je start byte v hodnotě 0x68, dle testů nad poskytnutými datasety již pouze tato informace stačila k odfiltrování naprosté většiny pro model nevalidních dat. Další využitou informací je minimální délka dat, která musí být alespoň 6 bytů, poslední položkou, kterou jsem využil je povinný nulový bit na nejméně významné pozici (LSB) ve třetím oktetu řídícího pole. Např. z datasetu `mega104-17-12-18.pcapng` tak bylo vyfiltrováno 58929 paketů zájmu.

Ze získaných paketů zájmu byla následně exportována užitečná data do formátu `csv`, konkrétně se využívá zdrojová a cílová adresa z IP hlavičky paketů a časová stopa paketů, měl jsem také snahu o využití např. délky paketů, avšak nesetkal jsem se v tomto případě s úspěchem.

Kapitola 3

Statistický model

Pro detekci anomálií byl vytvořen statistický model, při jehož konstrukci jsem vycházel z [1]. Tento model je založený na mezipaketových intervalech, směru komunikace a pravidla tří sigma, podle kterého při normálním rozdělení bude 99.7 % dat ve vzdálenosti tří sigma od průměru dat.

3.1 Konstrukce statistického modelu

Model byl zkonstruován dle následujícího algoritmu:

1) Spočtení mezipaketových intervalů jako rozdíl časových stop dvou po sobě jdoucích paketů jako $\Delta t_{i+1} = t_{i+1}$ - t_i.

2) Rozřazení paketů do skupin dle směru komunikace, tedy na základě zdroje a cíle. K tomu byly využity IP adresy zařízení.

3) Pro každý směr se nalezne nejlepší bod rozdělení z kandidátů, detailně viz 3.2.

4) Pro dané nejlepší body rozdělení spočítáme intervaly, které považujeme za normální, dle pravidla sigma tedy průměr s možnou odchylkou ± tří sigma.

3.2 Nalezení nejlepšího kandidáta bodu rozdělení

Pro nalezení nejlepšího kandidáta lze postupovat následovně:

1) Zvolit kandidáty, jako kandidáty jsem zvolil kvartily neboli 25 percentil, 50 percentil a 75 percentil mezipaketových intervalů pro daný směr. Kromě kvartilů jsem jako kandidáta spočítal také průměr a oproti [1] jsem přidal také kvadratický průměr. Kvadratický průměr bohužel neměl na nic vliv na dodaných datasetech, jelikož byly ve všech případech vybrání jiní kandidáti na bod rozdělení jako lepší.

2) Pakety rozdělíme do časových oken, např. po 300 sekundách.

3) Pro každého kandidáta na bod rozdělení spočítáme potřebné charakteristiky, zde průměr a směrodatnou odchylku počtu paketů v časových oknech s větším mezipaketovým intervalem, než je zvolený kandidát a s menším mezipaketovým intervalem, než je zvolený kandidát.

4) Jako nejlepšího kandidáta vybereme bod, pro který jsme nalezli nejmenší směrodatnou odchylku a pro který zároveň platí, že průměr méně 3 sigma je větší než nula, jelikož by některé chyby nešlo detekovat, kdyby spodní hranice počtu paketů v okně byla nula či méně.

3.3 Detekce anomálií

Implementovaná detekce anomálií vypisuje dva typy hlášení, varování (která by byla brána jako chyby, kdyby se vycházelo pouze z jedné hodnoty) a chyby.

V případě varování jde o znamení, že v rámci jednoho intervalu došlo k překročení vymezených hodnot, tím to způsobem by ovšem vznikalo relativně vysoké množství falešně pozitivních hlášení.

Z výše zmíněného důvodu se vypisuje také druhá kategorie hlášení, která vychází z [1], kde bylo navrženo tříhodnotové testování, to nicméně dále rozšiřuji do trochu komplexnější podoby. V citovaném textu dochází pouze k uchování informace, zda byla chyba v posledních třech oknech. Chyba se následně hlásí pouze, pokud je chyba v alespoň dvou ze tří posledních oken. Tato metoda je dle mého testování však poměrně náchylná na falešně negativní výsledky při krátkodobých útocích s mezerami, či krátkodobých výpadcích. Proto jsem vytvořil další pravidla, která měla za cíl zmírnit falešně negativní výsledky s cílem nezvýšit razantně falešně pozitivní výsledky.

Pravidlo, které jsem vytvořil navrch obnáší uchování nejen informace, jestli byla v posledních třech oknech anomálie, ale také konkrétní počet paketů v daných třech oknech. Chyba je potom hlášena navíc v případě, že průměr z posledních třech oken spadne mimo meze normálních dat. Díky tomu se při testování podařilo odhalit krátkodobé výpadky, které postihly pouze jedno okno, nebo simulované útoky, kdy byl v takovém jednom okně počet paketů velmi razantně větší, než spadá do normálních dat.

Další drobnou změnu, kterou jsem zavedl je schopnost pamatovat si výskyt abnormálních hodnot v posledních devíti oknech a generovat chybu, pokud se objevily abnormální data alespoň ve třech z těchto devíti oken. Při testech této detekce opět nebylo pozorováno zvýšené množství falešně pozitivních hlášení.

Pro testování detekce anomálií byl dle návrhu natrénován statistický model za pomoci dvou třetin datasetu, jeho úspšnost byla poté testována na zbývající třetině datasetu, která byla případně upravována doplněním nadměrného množství testovacích dat do některých oken, či případně odstraněním některých dat.

Kapitola 4

Závěr

Podařilo se vytvořit statistický model založený na [1], který dle mého názoru obsahoval slabiny v možnostech detekce určitých typů krátkodobých anomálií. Tuto vlastnost se mi podařilo dle testů s generovanými anomáliemi vylepšit primárně pomocí metody průměrování posledních tří časových oken, kterou jsem popsal v sekci 3.3.

Dále jsem se pokoušel aplikovat zmíněný typ detekce anomálií za využití jiné charakteristické vlastnosti. Konkrétně jsem zkoušel využít velikost paketů namísto mezipaketových intervalů, jelikož dle mých pozorování jsou velikosti paketů poměrně stabilních velikostí. Tato metoda se však nesetkala s úspěchem a generovala velmi vysoké množství falešně pozitivních hlášení, pravděpodobně jelikož v tomto případě mají velikosti buď velmi malé nebo velmi velké velikosti, ale nikoliv něco uprostřed.

Dále bych chtěl zmínit způsob, jak dále vylepšit detekci bez nepatřičně velkého zvýšení falešně negativních zpráv, který jsem však nestihl aplikovat a otestovat. Rád bych ho zde v závěru alespoň navrhl, jedná se o možnost analyzovat na základě překrývajících se časových oken. Další okno by tak mohlo začít již například v polovině předešlého, tímto způsobem by mohlo dojít ke zvýšení schopnosti detekce krátkých útoků a výpadků na hranici dvou časových oken.

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.

[2] Ing. Petr Matoušek Ph.D., M.A.. *Description and analysis of IEC 104 Protocol* [online]. [cit. 2022-04-22]. Dostupné z: https://www.fit.vut.cz/research/publication-file/11570/TR-IEC104.pdf.