OSTBAYERISCHE
TECHNISCHE HOCHSCHULE
REGENSBURG

Studienarbeit zum Thema Visualisierung hochdimensionaler Daten

Im Studienfach IT-Security hochgradig vernetzter Anwendungen und Systeme

WiSe 24/25

Vorgelegt von: Lukas Wolf
Matrikelnummer: 3434260

Betreuer: Prof. Dr. Rudolf Hackenberg

Abgabedatum: 04.02.2025

Kurzfassung

In dieser Arbeit werden verschiedene Methoden und Diagrammtypen zur Visualisierung hochdimensionaler Daten analysiert. Durchgeführt werden die Analysen anhand von Daten, aufgezeichnet von einem Network Intrusion Detection System, zum Netzwerkverkehr im Normalbetrieb und im Denial of Service-Angriffsfall. Zum Vergleich absoluter Kennzahlen stellen sich besonders Linien- und Flächendiagramme als geeignet heraus. Die Größenordnung im Unterschied der mittleren Anzahl von durch Hosts verarbeiteter Datenpakete um den Faktor 42,75 wird in einem solchen Diagramm deutlich ersichtlich. Mithilfe der Hauptkomponentenanalyse werden hochdimensionale Daten auf typischerweise ein bis drei Dimensionen reduziert. Werden diese Dimensionen in Streudiagrammen gegenüber gestellt, so sammeln sich ähnliche Daten in der visuellen Darstellung als Cluster. Bezogen auf die zugrundeliegenden Daten konnten starke Indize festgestellt werden, die auf einen TCP-Syn Flooding-Angriff und einen gezielten UDP Flooding-Angriff hinweisen.

Inhaltsverzeichnis

1 Einleitung

Im Bereich der Data Science ist es heutzutage üblich mit umfangreichen, hochdimensionalen Datenmengen zu arbeiten. Diese Daten sind somit allerdings nur schwer greifbar. Um daraus neue Erkenntnisse zu erlangen, ist ein gezieltes, strukturiertes Vorgehen nötig. Im Umfang dieser Arbeit werden hierzu verschiedene Verfahren zur Visualisierung hochdimensionaler Daten analysiert und eingeordnet.

Die betrachteten Daten liegen dabei im Kontext der Netzwerküberwachung. Zugrunde liegen mehrere Datensätze, die jeweils den Netzwerkverkehr über einen bestimmten Zeitraum abbilden, welche mit dem Network Intrusion Detection System (NIDS) Suricata[1] aufgezeichnet wurden. Es wird dabei unterschieden zwischen Netzwerkverkehr im Normalbetrieb und im Angriffsfall. Unterschiede dieser beiden Arten sollen visuell ausgearbeitet werden, wobei diese Erkenntnisse dabei helfen sollen, Angriffe in einem Netzwerk frühzeitig zu erkennen. Die betrachteten Datensätze umfassen bis zu 70 Dimensionen. Betrachte Angriffstypen sind Denial of Service (DoS) [SKK*97] und Ransomware [Bre16], wobei der Fokus auf DoS-Angriffen liegt.

Die Daten werden in Kapitel 2 zunächst in geeigneten Datenstrukturen aufbereitet. Diese umfassen eine tabellarische Darstellung und eine Darstellung als multidimensionaler Data Cube [GBL*96]. Bei der explorativen Analyse in Kapitel 3 werden zunächst Daten über einen zeitlichen Verlauf dargestellt. Weiter können aus dem Cube mit Slicing-Operationen gezielt Datensegmente entnommen werden, um einzelne Features im Detail zu analysieren. In Kapitel 4 werden die Daten mithilfe der Hauptkomponentenanalyse [WEG87] auf einen kleinen Dimensionsraum reduziert. Auch im Bezug auf Netzwerke [Bul18] erwies sich das Verfahren als geeignete Methode zur Kategorisierung des Datenverkehrs.

[1] Open Information Security Foundation: `www.openinfosecfoundation.org`, letzter Zugriff: 04.02.2025

2 Preprocessing

Über einen beliebigen Zeitraum werden vom NIDS *Ereignisse* aufgezeichnet, wobei ein Ereignis bestimmte Kennzahlen (Features) im Netzwerk über eine Zeit von 30 Sekunden festhält. Erfasste Features sind beispielsweise die Anzahl ankommender Pakete im gesamten Netzwerk, an einem bestimmten Host oder eines bestimmten Protokolls oder auch die Anzahl aktiver Verbindungen. Der aufgezeichnete Netzwerkverkehr umfasst ein Subnetz mit Maske 255.255.255.0. Bevor die Daten analysiert werden, werden diese vorher in geeigneten Datenstrukturen aufbereitet.

2.1 Aufbau des Datensatzes

Für jede Aufzeichnung liegt ein Datensatz im JSON-Format [Bra14] vor. Eine Aufzeichnung bildet isoliert entweder den normalen Netzwerkverkehr oder ein bestimmtes Angriffsszenario ab. Der Datensatz besteht aus einer Liste von Ereignissen, in welchen die Features in sogenannten *Paket-Queues* gegliedert sind. Durch diese Gliederung können Features über den gesamten Netzwerkverkehr, speziell für einen Host oder für ein bestimmtes Protokoll ausgewertet werden.

2.1.1 Arten von Datensätzen

Der Datensatz während des *Normalbetriebs* umfasst einen Zeitraum von etwa drei Stunden. Von *DoS-Angriffen* liegen fünf Aufzeichnungen über einen Zeitraum von jeweils 20 Minuten bis drei Stunden vor. Die beiden *Ransomware*-Datensätze bilden jeweils einen Zeitraum von etwa einer Stunde ab.

2.1.2 Paket-Queues

Das NIDS Suricata überwacht den Netzwerkverkehr mithilfe von Paket-Queues. Über den Zeitraum eines Ereignisses, also 30 Sekunden, werden ankommende Datenpakete in diese Queues eingereiht. Bei Abschluss des Ereignisses werden Kennzahlen anhand der eingereihten Pakete erfasst und die Queues vor Beginn des nächsten Ereignisses geleert. Wie in Abbildung 2.1 zu sehen ist, sind Queues in die drei Kategorien Universal Queue, Host Queues und Protocol Queues aufgeteilt. Diese werden gesammelt auch als *Selected Queues* bezeichnet. Die *General Queue* beinhaltet Auswertungen über alle Selected Queues. Die für jede Queue ermittelten Kennzahlen

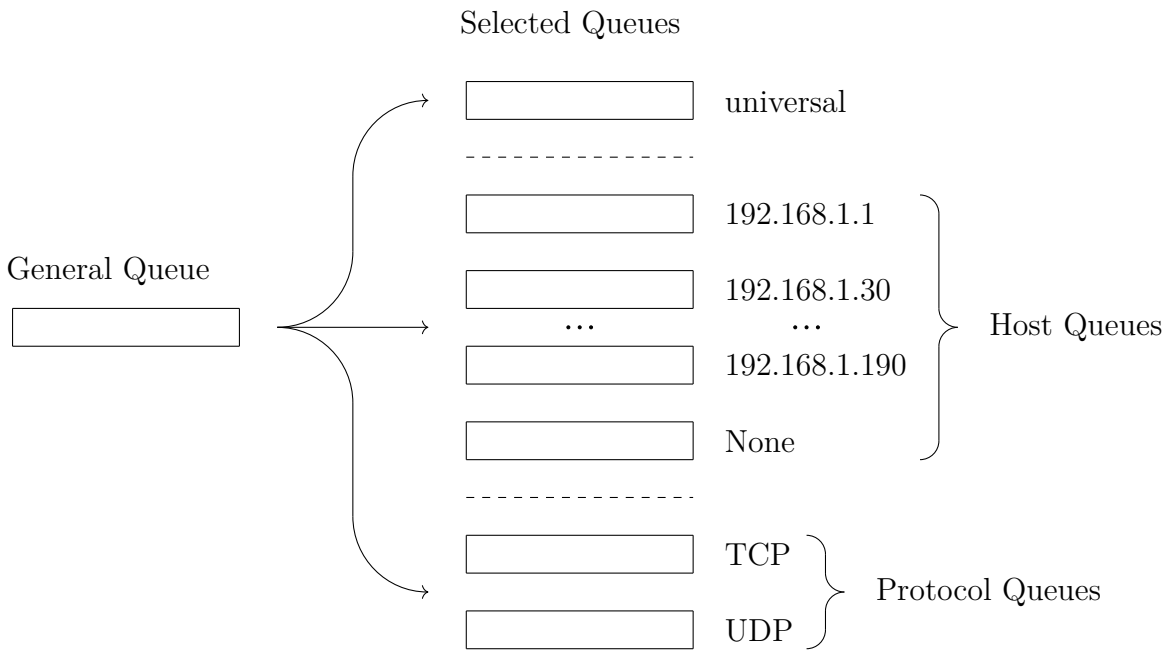

Abbildung 2.1: In der Abbildung werden die im NIDS Suricata verwalteten Paket-Queue-Typen dargestellt. Die Selected Queues sind in die drei Kategorien Universal, Host und Protocol eingeteilt, wobei ein Datenpaket jeder Kategorie jeweils einmal zugeordnet wird. In der General Queue werden Kennzahlen aller Selected Queues aggregiert bereitgestellt.

sind im Datensatz dem Ereignis als Feature zugeordnet und gegliedert für jede Selected Queue und die General Queue abrufbar.

Selected Queues Ein im Netzwerk ankommendes Datenpaket wird jeder der drei Kategorien einmal zugeordnet und dort jeweils in einer konkreten Selected Queue eingereiht. Die *Universal Queue* bildet den gesamten Netzwerkverkehr ab und enthält somit jedes ankommende Datenpaket. Bei den *Protocol Queues* wird nach bestimmten Protokollen der Datenpakete unterschieden. Betrachtet werden hier die Transportprotokolle TCP und UDP. Jedem Host im Netzwerk ist eine *Host Queue* zugeordnet, welche die an diesem Host ankommenden Datenpakete umfasst. Host Queues umfassen darüber hinaus untergeordnete *Connection Queues*, welche konkrete TCP- bzw. UDP-Verbindungen abbilden. Datenpakete einer Host Queue werden somit zusätzlich einer Verbindung zugeordnet, wobei eine Verbindung identifiziert wird durch das Quadrupel aus Sender- und Empfänger-IP und Sender- und Empfänger-Port. Die Selected Queues umfassen im Datensatz etwa 20 Features.

General Queue In der *General Queue* werden die Ergebnisse der einzelnen Queues kombiniert und aggregiert. Dabei werden Kennzahlen wie das arithmetische Mittel, der Median, Summen oder das Maximum für bestimmte Kennzahlen der Selected Queues gebildet. Die General Queue umfasst im Datensatz etwa 50 Features.

General Queue Data Layout

Event ID	General Queue Connection Features	General Queue Features
0	features like:	features like:
1	connection queues	mean queue length
2	length sum	
3		longest queue
4	connection count	
5	total	host count
...	...	...

Selected Queues Data Layout

Event ID	Queue	Selected Queue Connection Features	Selected Queue Features
0	universal	features like:	features like:
0	host 1	connection count	queue length
0	host 2		
0	tcp	connection queues	relative queue
1	universal	length median	length
1	host 1		
...	...	...	...

Tabelle 2.1: Die Tabelle zeigt die Aufteilung der Daten nach General Queue links und nach Selected Queues rechts. Ein Ereignis (Event) wird in einer Zeile dargestellt, eine Spalte stellt ein Feature dar. Allgemeine Features und Connection Queue-Features werden separat aufgeführt. Der Datenbereich ist in den Tabellen nicht gefüllt, stattdessen sind einige Beispiele für mögliche Features aufgeführt.

2.1.3 Typen von Features

Im Datensatz wird zwischen allgemeinen Features und Connection Queue-Features unterschieden. Allgemeine Features beziehen sich auf den Zustand der Selected Queues, Connection Queue-Features beschreiben dahingegen den Zustand der vorher beschriebenen Connection Queues. Da Verbindungen ausschließlich Hosts zugeordnet sind, werden diese Features nur an den Host-Selected Queues und zusammengefasst in der General Queue erfasst.

2.2 Datenaufbereitung

Um die Daten zu analysieren, werden die Datensätze im JSON-Format in Datenstrukturen geladen und aufbereitet. Es werden hierfür eine tabellarische Struktur und ein multidimensionaler Data Cube [GBL*96] gewählt, welche umfangreiche Operationen bieten, auf bestimmte Daten und Aggregationen dieser zuzugreifen. Konkret wurden hierfür Pandas Dataframes [McK11] eingesetzt.

2.2.1 Tabellarische Struktur

Der schematische Aufbau der Struktur ist in Tabelle 2.1 zu sehen. Die Daten eines Datensatzes werden in zwei Tabellen aufgeteilt. Eine bildet die General Queue ab, die andere gesammelt alle Selected Queues. Im General Queue Layout stellt eine Zeile ein Ereignis (Event) dar, also einen 30-sekündigen Abschnitt des Netzwerkverkehrs. Beim Selected Queue Layout wird jedes Event zusätzlich nach Selected Queue unterschieden. Eine Zeile wird somit eindeutig identifiziert aus der Kombination der Event ID und der jeweiligen Queue. Auf den Spalten sind die Features angeordnet, wobei die Unterscheidung von allgemeinen Features und Connection Queue-Features berücksichtigt wird.

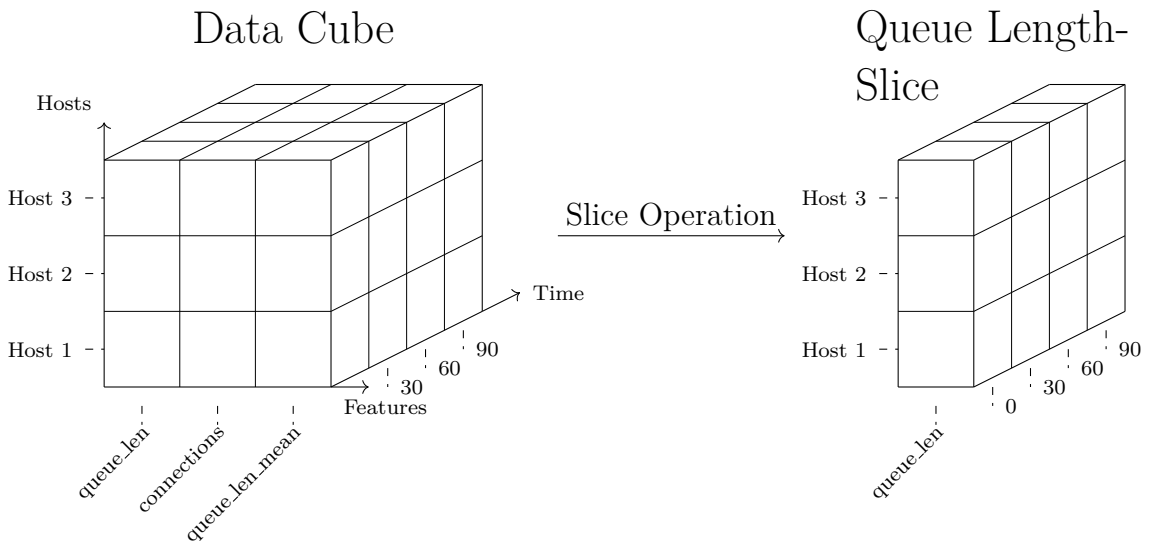

Abbildung 2.2: Die Abbildung stellt einen Data Cube dar, welcher in die drei Dimensionen Features, Zeit und Hosts gegliedert ist. Für gezielte Analysen können beliebige Datensegmente durch eine Slice-Operation entnommen werden.

2.2.2 Multidimensionaler Data Cube

Für eine detaillierte Analyse der Host-Selected Queues werden die Daten des Normalbetriebs und während eines DoS-Angriffs jeweils in einer vereinfachten Form eines Data Cubes organisiert. Zusätzlich zu den beiden Dimensionen Feature und Zeit in der tabellarischen Struktur werden die Daten hier zusätzlich nach einer dritten Dimension, nach den einzelnen Hosts, unterschieden, wie in Abbildung 2.2 dargestellt. Durch die Slice-Operation können einzelne Datensegmente aus dem Cube entnommen werden und einzeln betrachtet werden. Es kann beispielsweise ein bestimmtes Feature selektiert werden, welches dann nach Zeit und Hosts aufgegliedert ist. Durch Aggregation eines Slices können Zusammenhänge zu Features der General Queue hergestellt werden. Realisiert wurde der Cube mittels eines Multi-Index-Dataframes.

3 Explorative Datenanalyse

Bei der explorativen Datenanalyse werden gezielt einzelne Features des Normal- und das DoS-Angriffsszenarios über den zeitlichen Verlauf der Ereignisse analysiert. Die General Queue und die Selected Queues werden dabei separat untersucht und dabei Zusammenhänge hergestellt.

3.1 Evaluation der General Queue

Zunächst werden die Connection Queue Features "Summe der Längen aller Connection Queues", "Anzahl aller Verbindungen", Mittel und Median der Connection Queue-Längen über den zeitlichen Verlauf analysiert und auf Zusammenhänge eingegangen. Die Features stehen in folgendem Zusammenhang:

$$\text{Mittel der Längen der Connection Queues} = \frac{\text{Summe der Längen der Connection Queues}}{\text{Anzahl Verbindungen}} \tag{3.1}$$

Länge der Connection Queues In Abbildung 3.1 wird die Summe der Längen aller Connection Queues dargestellt. Anders formuliert beschreibt das Feature die Anzahl aller Datenpakete, die von Hosts insgesamt verarbeitet werden. Im Normalbetrieb liegt die Kurve relativ konstant beim Mittel von 1090,70 mit einer Standardabweichung von 39,32, wohingegen sie sich während des Angriffs in einem Bereich von 34495 bis 52490 um das Mittel 46631,60 bei einer

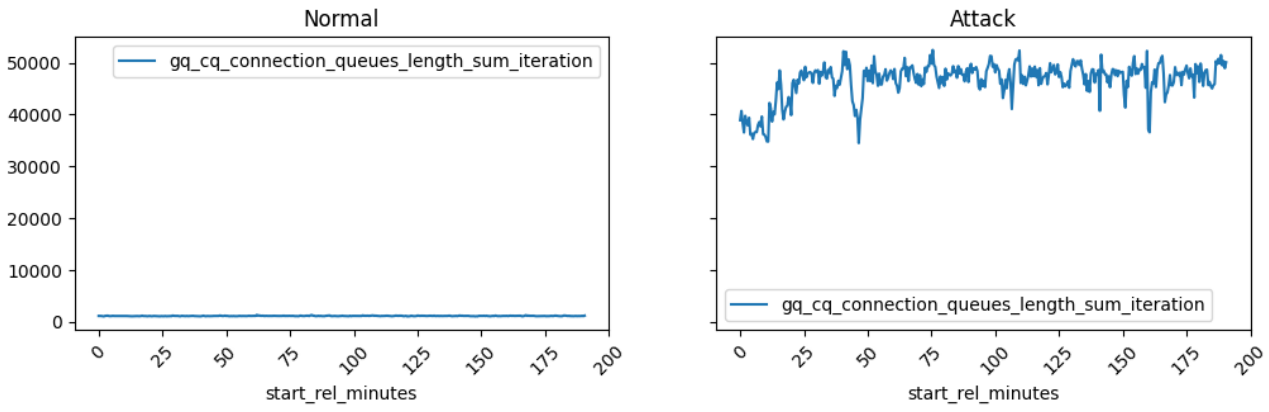

Abbildung 3.1: Im Diagramm ist die Summe der Längen aller Connection Queues im Normalbetrieb und im DoS-Angriffsfall über einen Zeitraum von drei Stunden abgebildet.

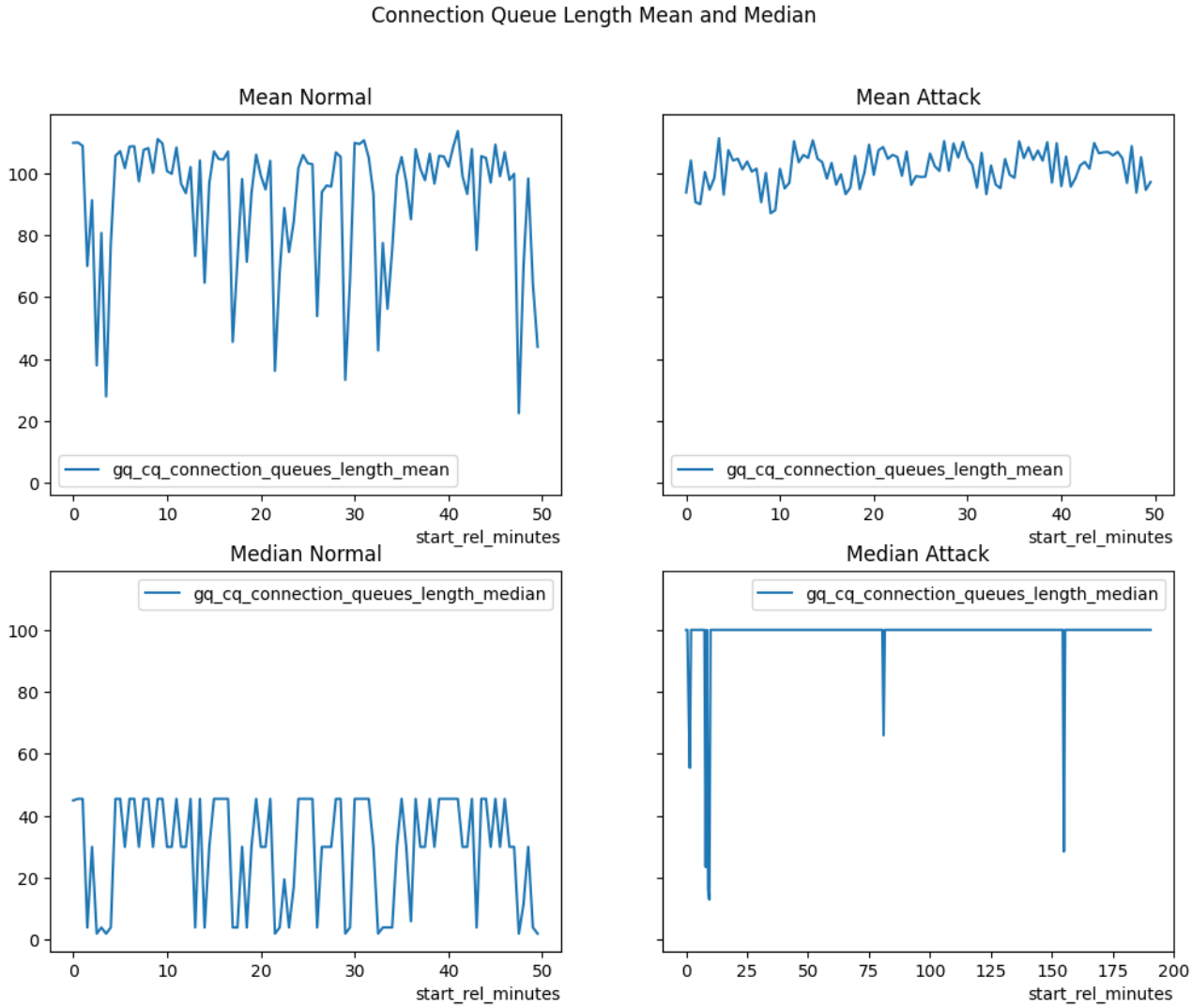

Abbildung 3.2: Abgebildet sind Diagramme, die das Mittel und den Median im Normal- und im DoS-Angriffsfall zeigen. Der Median im Angriffsfall wird über einen Zeitraum von drei Stunden dargestellt. Bei den anderen Diagrammen handelt es sich um einen Ausschnitt von 50 Minuten zur klareren Darstellung der Fluktuation.

Standardabweichung von 3516,43 bewegt. Das Mittel der beiden Kurven unterscheidet sich um den Faktor 42,75. Die große Diskrepanz der Kurven in diesen Faktoren ist ein starkes Indiz für einen DoS-Angriff.

Im Mittel gibt es im Normalbetrieb 3,05 aktive Hosts pro Ereignis, während es beim Angriff 10,02 sind. Insgesamt werden über den Zeitraum normal 6 beziehungsweise beim Angriff 14 verschiedene Hosts erreicht. Es fällt auf, dass beim DoS-Angriff mehr Hosts erreicht werden, wobei selbst 14 Hosts nur einen Bruchteil der potentiell 255 Hosts im Subnetz darstellen. Es ist allerdings nicht bekannt, ob über die erreichten Hosts hinaus im Netzwerk weitere verfügbar sind.

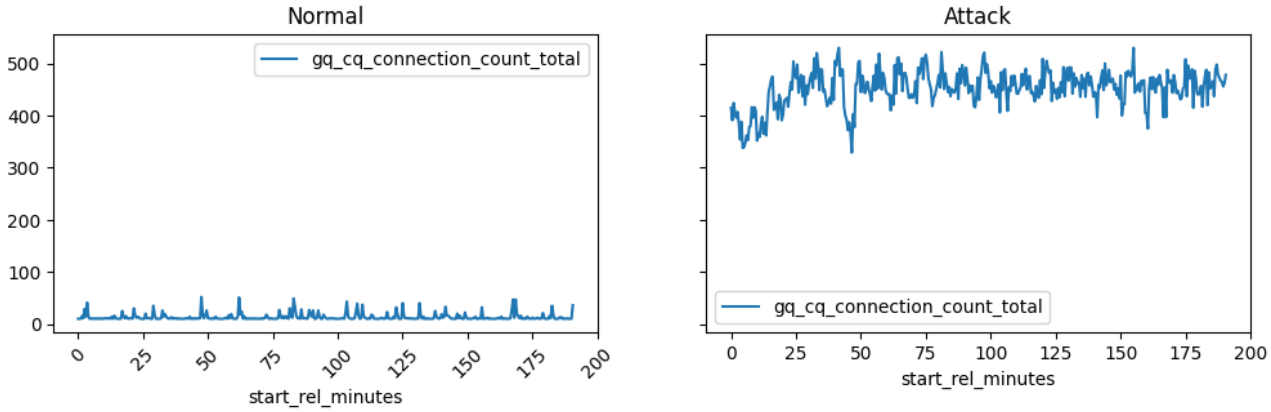

Abbildung 3.3: Die Diagramme zeigen die insgesamte Anzahl an Verbindungen im Normal- und DoS-Angriffsszenario über einen Zeitraum von drei Stunden.

Mittel und Median Bei Betrachtung des Mittels und des Medians im Normalbetrieb fällt auf, dass die Werte über den zeitlichen Verlauf stark fluktuieren, was in Abbildung 3.2 ersichtlich wird. Dies weist darauf hin, dass der Netzwerkverkehr, anders als beim Angriff, regelmäßig entlastet wird. Auf der Angriffsseite verlaufen die Kurven vergleichsweise sehr hoch und relativ konstant. Besonders der Median ist beinahe vollständig konstant bei einem Wert von 100 und somit um einen Faktor von 2,20 höher als der Maximalwert im Normalbetrieb. Das bedeutet, dass mindestens die Hälfte aller Connection Queues eine Länge von mindestens 100 besitzt. Bei Betrachtung einzelner Hosts konnten allerdings auch höhere Connection Queue-Längen als 100 festgestellt werden. Es wird vermutet, dass beim Angriff an einen großen Teil der Verbindungen pro 30 Sekunden exakt 100 Pakete geschickt werden.

Die Werte von Mittel und Median verlaufen in den beiden Szenarien in einer ähnlichen Größenordnung. Besonders das Mittel des Mittels liegt sehr nah beieinander mit einem Wert von 91,89 auf der Normalseite und 103,22 auf der Angriffsseite. Unter Betrachtung dieser Werte und Gleichung 3.1 kann die große Diskrepanz in Abbildung 3.1 nicht auf die individuelle Connection Queue-Längen zurückgeführt werden.

Anzahl von Verbindungen Der dritte Faktor in Gleichung 3.1 ist die gesamte Anzahl der Verbindungen, welcher in Abbildung 3.3 dargestellt wird. Hier zeichnet sich ein verhältnismäßig ähnliches Bild wie in Abbildung 3.1, wobei die Verbindungszahl im Angriffsfall sich im Mittel um den Faktor 33,59 unterscheidet verglichen zum Normalbetrieb. Daraus lässt sich ableiten, dass bei Betrachtung der vorliegenden Daten die Summe aller Connection Queue-Längen, also die Summe aller Pakete, die an Hosts anliegen, hauptsächlich durch die Anzahl an Verbindungen bestimmt wird und die individuelle Queue-Länge einen untergeordneten Einfluss hat.

Über den gesamten Zeitraum liegt der Anteil an TCP-Verbindungen während des DoS-Angriffs bei 98,44% bis 99,46%, was die Vermutung nahe legt, dass es sich dabei um einen Syn Flooding-Angriff [SKK*97] handelt.

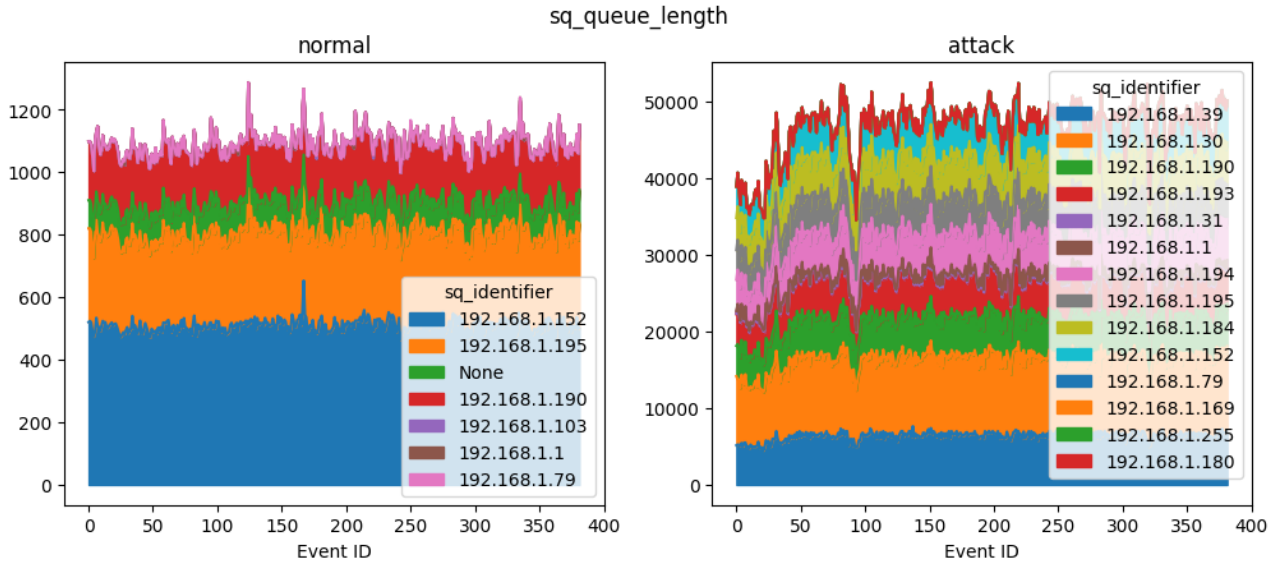

Abbildung 3.4: In den Diagrammen werden gestapelt die Längen aller Host-Queues im Normal- und DoS-Angriffsfall über einen Zeitraum von drei Stunden abgebildet. Werden die einzelnen Queue-Längen aufsummiert, so entspricht das dem General Queue-Feature der Anzahl aller von Hosts verarbeiteten Datenpaketen.

3.2 Evaluation der Selected Queues

Zur Auswertung der Features der Selected Queues werden die vorbereiteten Data Cubes eingesetzt. Die Selected Queue-Features werden mittels einer Slicing-Operation aus dem Cube selektiert und in Flächendiagrammen über die zugrundeliegenden Hosts dargestellt. Bei den Abbildungen in diesem Abschnitt ist die Event ID auf der x-Achse abgebildet. Diese kann auch als zeitlicher Verlauf gesehen werden, da die Ereignisse chronologisch geordnet sind. Ein Ereignis umfasst eine Dauer von 30 Sekunden, somit kann die ID halbiert werden, um die Startzeit des Events in Minuten zu erhalten.

Queue-Länge Abbildung 3.4 zeigt die Längen der einzelnen Host-Queues über einen zeitlichen Verlauf von drei Stunden. Es fällt auf, dass die Längen der einzelnen Host-Queues sowohl im Normalbetrieb als auch im Angriffsfall relativ konstant verlaufen, wobei beim Angriff in den ersten 60 Ereignissen und etwa um das Ereignis 100 eine stärkere Abweichung davon sichtbar ist. Die entstehenden Kurven zeigen sich auch in Abbildung 3.1. Bei weiteren Analysen fällt auf, dass die Summe der Host-Queue-Längen anders als erwartet nicht gleich der Länge der Universal-Queue beziehungsweise der Summe der Längen der Protokoll-Queues ist sondern deutlich höher. Der Grund für die Diskrepanz ist zum noch nicht geklärt. Es konnten auch kaum Broadcast-Pakete festgestellt werden, welche möglicherweise dupliziert werden könnten.

Verbindungsanzahl Die Verbindungsanzahl wird in Abbildung 3.5 dargestellt. Dabei fallen auf der Normalseite starke Spitzen auf, welche größtenteils auf die Hosts 192.168.1.79,

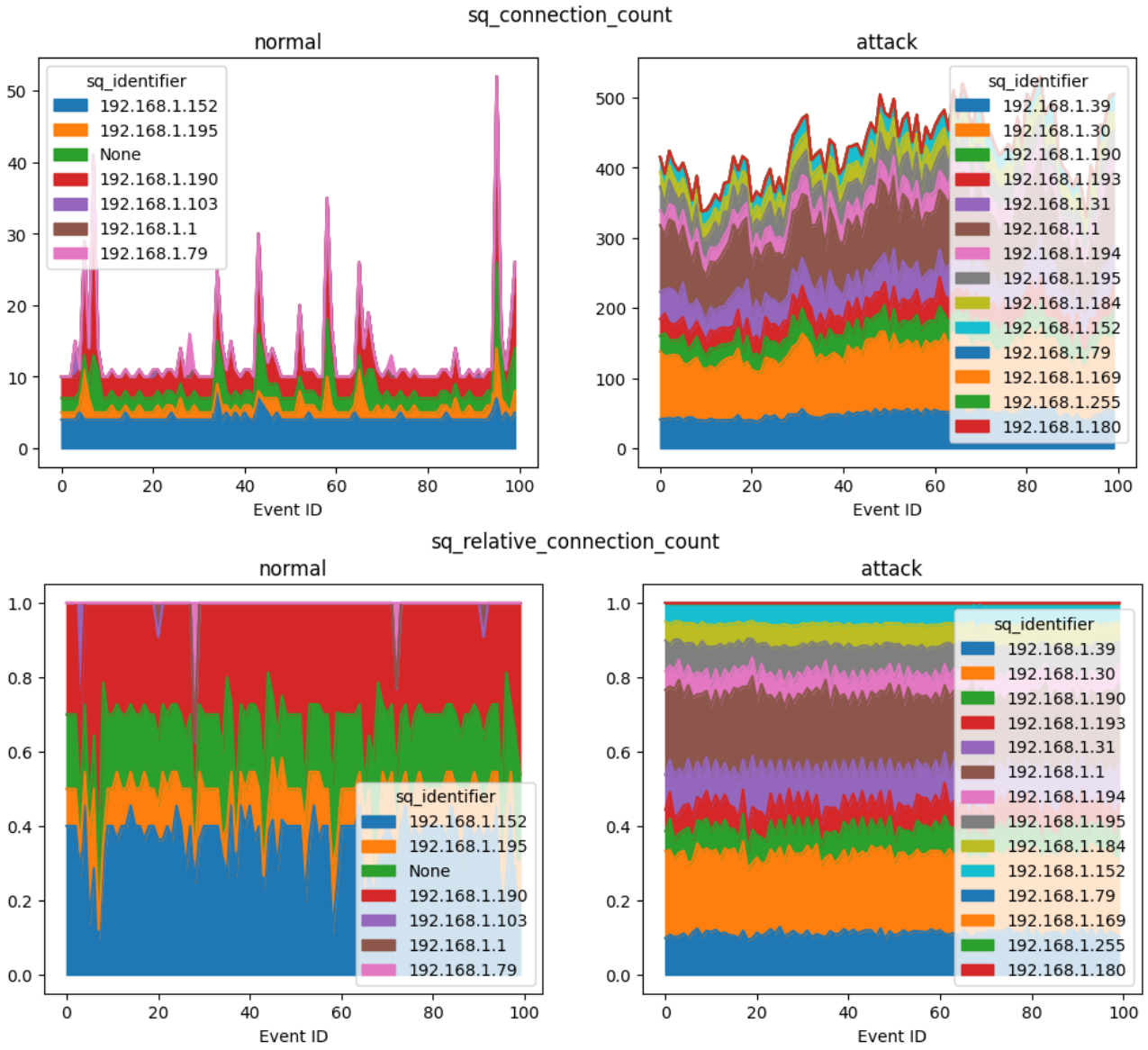

Abbildung 3.5: Die Diagramme stellen die Anzahl an Verbindungen je Host im Normal- und im DoS-Angriffsszenario gestapelt über einen Zeitraum von 50 Minuten dar. In der oberen Zeile ist die absolute Anzahl an Verbindungen gestapelt abgebildet, welche in Summe dem General Queue-Feature aller Verbindungen der Host-Queues entspricht. Dagegen zeigt die untere Zeile den relativen Anteil an Verbindungen je Host gestapelt, was in Summe 1,0 beziehungsweise 100% ergibt. Die Host-Queue "None" beinhaltet alle Datenpakete, die keinem aktiven Host zugeordnet sind.

192.168.1.190 und die Host-Queue "None", welche nicht zuordenbare Pakete enthält, zurückzuführen sind. Insgesamt weisen die einzelnen Hosts ansonsten jeweils eine relativ konstante Anzahl an Verbindungen auf. Die Hosts mit der verhältnismäßig größten Anzahl an Verbindungen sind 192.168.1.30, 192.168.1.1 und 192.168.1.39 und machen bereits etwa die Hälfte aller Verbindungen aus. Dies zeigt sich noch klarer in den unteren beiden Diagrammen von Abbildung 3.5 bei Betrachtung des Anteils eines Hosts an die gesamte Verbindungsanzahl. Des Weiteren spiegeln sich die entstehenden Kurven auch in Abbildung 3.3 im Zeitraum 0 bis 50 Minuten wieder.

4 Hauptkomponentenanalyse

Die Hauptkomponentenanalyse, im Englischen Principal Component Analysis (PCA), ist ein Verfahren, das eine beliebige Anzahl an Eingangsdimensionen auf eine bestimmte Anzahl an Ausgangsdimensionen reduziert, wobei die Varianz der Ausgangsdaten möglichst beibehalten werden soll. Diese Ausgangsdimensionen werden Hauptkomponenten, im Englischen Principal Components (PCs), genannt. Typischerweise wird eine Reduktion auf zwei bis drei Dimensionen vorgenommen, um beispielsweise eine möglichst ganzheitliche Darstellung der Daten in einem Diagramm abzubilden.

Ähnliche Features, also Features, die eine gewisse lineare Abhängigkeit aufweisen, werden durch die Analyse in einer Hauptkomponente gruppiert und repräsentieren den höchsten Anteil der durch die Hauptkomponente dargestellte Größe. In einer Analyse werden also besonders Features gebündelt, die eine hohe Korrelation aufweisen. Bei der Analyse kann die sogenannte *Ladung* [WEG87, S. 41] von Elementen einer Hauptkomponente ermittelt werden, mit welcher bestimmt werden kann, zu welchem Anteil eine bestimmte Hauptkomponente durch eine Eingangsdimension beeinflusst wird, wie in Gleichung 4.1 zu sehen ist. Eine solche Verteilung ist in Tabelle 4.1 zu sehen. Die einzelnen Hauptkomponenten stehen orthogonal [WEG87] zueinander, sind also linear unabhängig und bilden jeweils einen Teil der Gesamtvarianz der Ausgangsdaten ab.

$$\text{Anteil von Feature } i \text{ an der Hauptkomponente } j = \frac{(\text{Ladung}_{ij})^2}{\sum_k (\text{Ladung}_{kj})^2} \tag{4.1}$$

In den nachfolgenden Hauptkomponentenanalysen werden die Ausgangsdaten jeweils auf drei Hauptkomponenten abgebildet. Analysen zur General Queue und zu den Selected Queues werden separat durchgeführt.

4.1 Analyse der General Queue

Für die Hauptkomponentenanalyse der General Queue werden nicht numerische Features ausgeschlossen, da diese in der Berechnung nicht einbezogen werden können. Darüber hinaus werden Features bezüglich der Queue-Priorität ausgeschlossen. Die Analyse wird zunächst separat für den Normal- und den DoS-Angriffsdatensatz durchgeführt, wobei jeweils insgesamt 38 Features einbezogen werden. Die Verteilung auf die ersten beiden Hauptkomponenten ist in Abbildung 4.1 dargestellt.

Anteil an PC1		Anteil an PC2	
gq_active_router_count	6,24%	gq_median_host_queue_length	11,83%
gq_active_router_in_cache_count	6,24%	gq_mean_host_queue_length	11,73%
protocol_queue_count_layer_4_diff	6,02%	gq_cq_connection_queues_length_sum_iteration	10,38%
gq_protocol_queue_count_layer_4_below	6,02%	gq_popped_pkts_iteration	10,11%
gq_protocol_queue_count	6,02%	gq_mean_growth_rate	6,60%
gq_protocol_queue_length_entropy	5,98%	gq_cq_connection_count_total	6,24%
gq_protocol_queue_length_variance	5,77%	gq_mean_queue_length	5,91%
gq_median_protocol_queue_length	5,77%	gq_median_growth_rate	5,44%
gq_mean_protocol_queue_length	5,77%	gq_host_count	4,12%
...		...	

Tabelle 4.1: In der Tabelle ist ein Beispiel, wie die prozentualen Anteile der Features auf die Hauptkomponenten 1 und 2 verteilt sind, wobei nur die ersten neun Features mit dem größten Beitrag abgebildet sind.

4.1.1 Getrennte Analyse der Szenarien

Für die beiden Szenarien werden die ersten drei Hauptkomponenten analysiert, welche auf der Normalseite 46,83%, 18,29% beziehungsweise 10,53% und in Summe 75,66% der Gesamtvarianz der Daten repräsentieren. Auf der Angriffsseite sind das 42,53%, 17,97% beziehungsweise 13,48% und in Summe 73,98%. Ein Ausschnitt der genauen Verteilungen von PC1 und PC2 der PCA des Normalbetriebs ist in Tabelle 4.1 abgebildet.

PC1 des Normalbetriebs ergibt sich größtenteils aus Features, die sich auf die Protokoll-Queues beziehen. Es wird angenommen, dass die hohe Varianz dieser Komponente darauf zurückzuführen ist, dass es eine Gruppe von Ereignissen gibt, die keine UDP-Verbindungen aufweisen. Diese Gruppe spiegelt sich auch im linken Cluster auf der Normalseite von Abbildung 4.1. Auf diesen Umstand wird in der nachfolgenden kombinierten PCA genauer eingegangen. In PC2 sind am stärksten die Features vertreten, die sich auf Queue-Länge beziehungsweise Anzahl verarbeiteter Pakete beziehen. Weitere Vertreter sind darüber hinaus beispielsweise die Verbindungszahl und die Anzahl aktiver Hosts. PC3 repräsentiert hauptsächlich Features, die sich auf

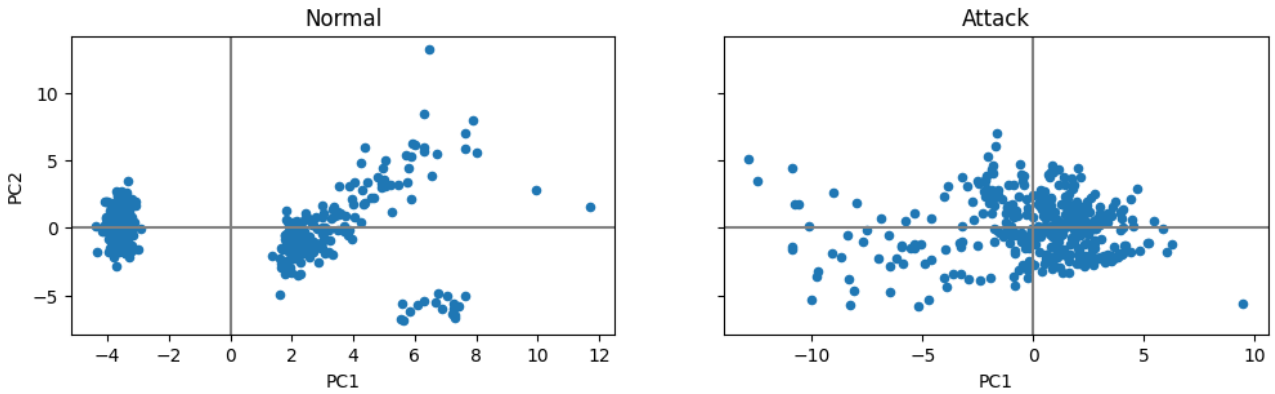

Abbildung 4.1: Die Abbildung visualisiert die ersten beiden Komponenten der Hauptkomponentenanalyse von Normalbetrieb und DoS-Angriff. Auf der Normalseite ist die Bildung verschiedener Cluster sichtbar.

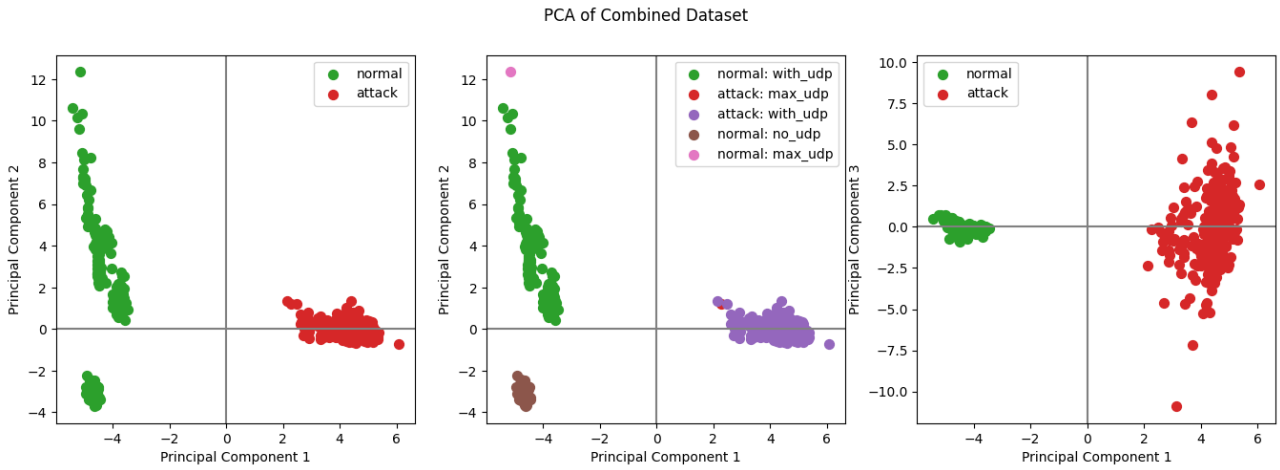

Abbildung 4.2: Die Diagramme stellen die Ergebnisse der kombinierten Hauptkomponentenanalyse von Normal- und Angriffsdatensatz dar. In ersten beiden Grafiken stellt das Verhältnis von PC1 und PC2 dar, wobei zweitere nähere Details der Cluster und spezielle Ereignisse zeigt. Die dritte Grafik bildet bildet PC3 im Verhältnis zu PC1 ab.

Host-Queues beziehen, wie zum Beispiel die Anzahl aktiver Hosts Entropie und Differenz der Host-Queue-Länge zum Vorgänger.

Beim Angriffsdatensatz spiegeln sich in PC1 Features bezüglich der Queue-Längen ähnlich zu PC2 auf der Normalseite. PC2 ist vergleichbar mit PC1 des Normalbetriebs mit Features, die sich auf Protokoll-Queues beziehen. In PC3 finden sich besonders eine stark abgegrenzte Gruppe von vier Features, die die Anzahl an aktiven Hosts beziehungsweise die Gesamtzahl an Queues beschreiben. Die Komponente wird durch diese Features bereits zu 92,77% bestimmt. In den vorangegangenen Komponenten ist die Zusammensetzung der Komponente deutlich stärker auf eine höhere Anzahl an Features verteilt. Die ersten drei Features von PC3 belaufen sich auf den exakt selben Einfluss, was die Vermutung nahelegt, dass diese zueinander linear abhängig sein könnten und somit dieselbe Information abbilden würden. Dies wurde im Umfang dieser Arbeit allerdings nicht weiter analysiert.

4.1.2 Kombinierte Analyse

Bei der kombinierten Analyse werden der Normal- und der Angriffsdatensatz zusammengeführt und die Hauptkomponentenanalyse auf diese Zusammenstellung angewandt. Einzelne Datensätze werden mit Labeln versehen, um Gruppen zusammengehörender Daten visuell unterscheidbar zu machen, wie in Abbildung 4.2 zu sehen ist. Nach Durchführung der Analyse repräsentieren die ersten drei Hauptkomponenten 62,00%, 17,97% und 6,34% der Gesamtvarianz der Daten, was in Summe 86,31% ausmacht. Auffällig ist, dass hier somit ein um etwa 10% höherer Abbildungsgrad erreicht wird als bei den beiden vorherigen Analysen.

Die Aufteilung der ersten beiden Hauptkomponenten ist vergleichbar mit der Angriffseite bei der getrennten Analyse, wobei PC1 hauptsächlich durch Features bezüglich der Queue-Länge

und bestimmt wird und PC2 durch Features bezüglich der Protokoll-Queues. PC3 stellt sich besonders aus Features zusammen, die die Wachstumsraten der Selected Queue-Längen darstellen.

In den Diagrammen von Abbildung 4.2 ist deutlich zu erkennen, dass sich Daten bezüglich PC1 stark in Normal auf der linken Seite und Angriff auf der rechten Seite aufspalten. Dies ist dadurch zu begründen, dass beim DoS-Angriff eine um ein Vielfaches höherer Verkehr an Paketen vorliegt und somit auch Queues deutlich stärker gefüllt sind. Die Normalseite ist bezogen auf PC2 sichtbar stärker gestreut als die Angriffsseite und bildet zudem mehrere Cluster. Weitere besondere Eigenschaften dieser Komponente werden im mittleren der Diagramme ersichtlich. Dort fällt auf, dass Cluster unten links vollständig aus Ereignissen besteht, die ausschließlich TCP-Verbindungen und keine UDP-Verbindungen aufweisen. Das Ereignis am oberen Ende der PC2-Skala ist auch das Ereignis mit dem größten gemessenen Anteil an UDP-Verbindungen von 15,28%. Der starke Ausschlag an dieser Stelle kommt daher, dass das alleinstehende Maximum bezogen auf die Vergleichsdaten tatsächlich eine Ausnahme ist. Das Mittel des Anteils an UDP-Verbindungen im Normalbetrieb liegt im Vergleich dazu bei nur 1,62%. Auf der Angriffsseite verläuft dieser Anteil in einem Bereich von 0,54% und 1,56%. Insgesamt betrachtet lässt die Komponente allerdings nur bedingt Rückschlüsse auf das tatsächliche Verhältnis von TCP- und UDP-Verbindungen zu. Die hier beschriebenen Erkenntnisse spiegeln sich auch in PC1 der Normalseite von Abbildung 4.1.

Weiter wird das Verhältnis zwischen PC1 und PC3 betrachtet, wie in Abbildung 4.2 rechts dargestellt. Dort fällt auf, dass sich, anders als in PC2, ein enges Cluster auf der Normalseite bildet und die Angriffsseite eine deutlich stärkere Streuung bezüglich PC3 aufweist. Es wird vermutet, dass dies auf den Unterschied in der Größenordnung des Wachstums der Queues in den beiden Szenarien zurückzuführen ist. Das mittlere Wachstum der im Normalbetrieb beläuft sich betragsmäßig auf 0,17 bis 75,57 Pakete mit einer Varianz von $2,4 \cdot 10^2$ Im Angriffsfall liegt dies deutlich höher zwischen 0,69 und 3114 Paketen mit einer Varianz von $1,5 \cdot 10^5$.

4.1.3 Verteilung verschiedener Angriffstypen

In diese Auswertung werden der Normaldatensatz, die fünf DoS-Angriffdatensätze und zusätzlich die beiden Ransomware-Angriffdatensätze einbezogen. Wie bei den vorangegangenen PCAs fließen die 38 ausgewählten Features in die Berechnung ein. Es ergibt sich eine Verteilung der Varianz auf die ersten drei Hauptkomponenten von 56,34%, 11,24% und 8,85%, womit insgesamt 76,44% abgedeckt werden. PC1 beinhaltet wieder größtenteils Features bezüglich der Queue-Länge. Der Fokus bei PC2 liegt auf der Anzahl der Protokoll-Queues, wohingegen er bei PC3 hauptsächlich auf der Verteilung zwischen UDP- und TCP-Verbindungen liegt. Die Verteilung der Daten wird in Abbildung 4.3 visualisert.

Zunächst ist zu beachten, dass Ransomware-Angriffe sich bezüglich der Menge des Datenverkehrs, dargestellt durch PC1, ein ähnliches Verhalten aufweisen, wie im Normalbetrieb. Die DoS-Angriffe bilden gesammelt ein Cluster und weisen bezüglich PC1 deutlich höhere Werte

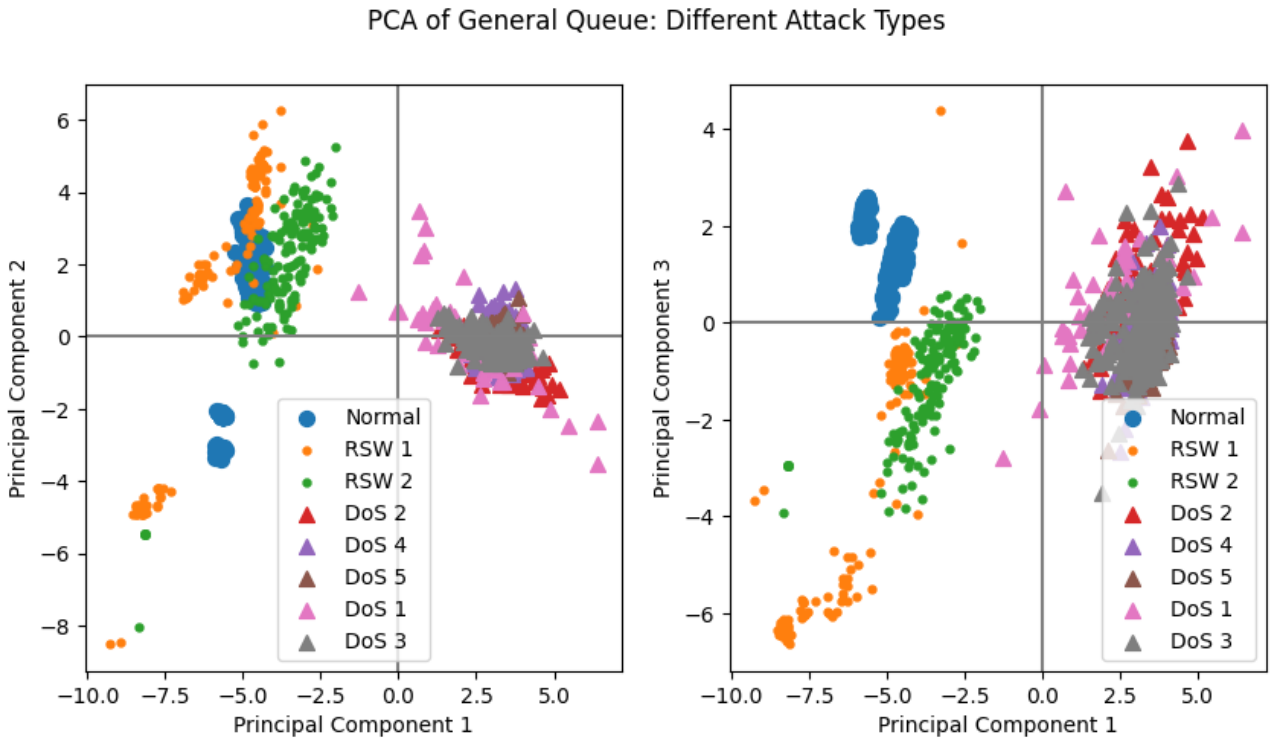

Abbildung 4.3: Es wird die Hauptkomponentenanalyse des kombinierten Datensatzes aus Normalbetrieb, DoS- und Ransomware-Angriffen dargestellt. Die Hauptkomponenten 2 und 3 werden jeweils der ersten Hauptkomponente gegenüber gestellt.

auf als die anderen Szenarien. Die stärkere Streuung von Datenpunkten in PC2 beim Normal- und Ransomware-Szenario ist dadurch begründet, dass dort in einigen Ereignissen nur eine Protokoll-Queue, die TCP-Queue, vorhanden ist und bei den Ransomware-Angriffen auch Ereignisse komplett ohne Protokoll-Queue vorkommen.

PC3 weist mit der Verteilung von UDP- und TCP-Verbindungen einen ähnlichen Sachverhalt auf, bietet aber durch die Sicht auf die Anzahl an Verbindungen eine detailliertere Aufschlüsselung dessen. Auffällig ist, dass die Datenpunkte der Ransomware-Angriffe unterhalb des Normal-Clusters sammeln, was darauf zurückzuführen ist, dass dort ein verhältnismäßig hoher Anteil an UDP-Verbindungen vorliegt von im Mittel 26,60%. Das Mittel der anderen beiden Szenarien liegt dagegen jeweils unter 1,62%. Bei weiteren Analysen des Normal- und des Ransomware-Szenarios wird ein Unterschied in der Anzahl an Verbindungen und in der mittleren Connection Queue-Länge deutlich. Die mittlere Anzahl an Verbindungen pro Ereignis im Normalbetrieb beträgt 13,48 beziehungsweise 90,33 im Angriffsfall. Auf der anderen Seite liegt das Mittel der mittleren Connection Queue-Länge bei 91,79 beziehungsweise 16,95. Bei den Ransomware-Angriffen sind also verhältnismäßig viele Verbindungen vorhanden, mit einer im Mittel niedrigen Länge der zugehörigen Connection Queues. Bisher konnte allerdings kein kausaler Zusammenhang diesbezüglich hergestellt werden.

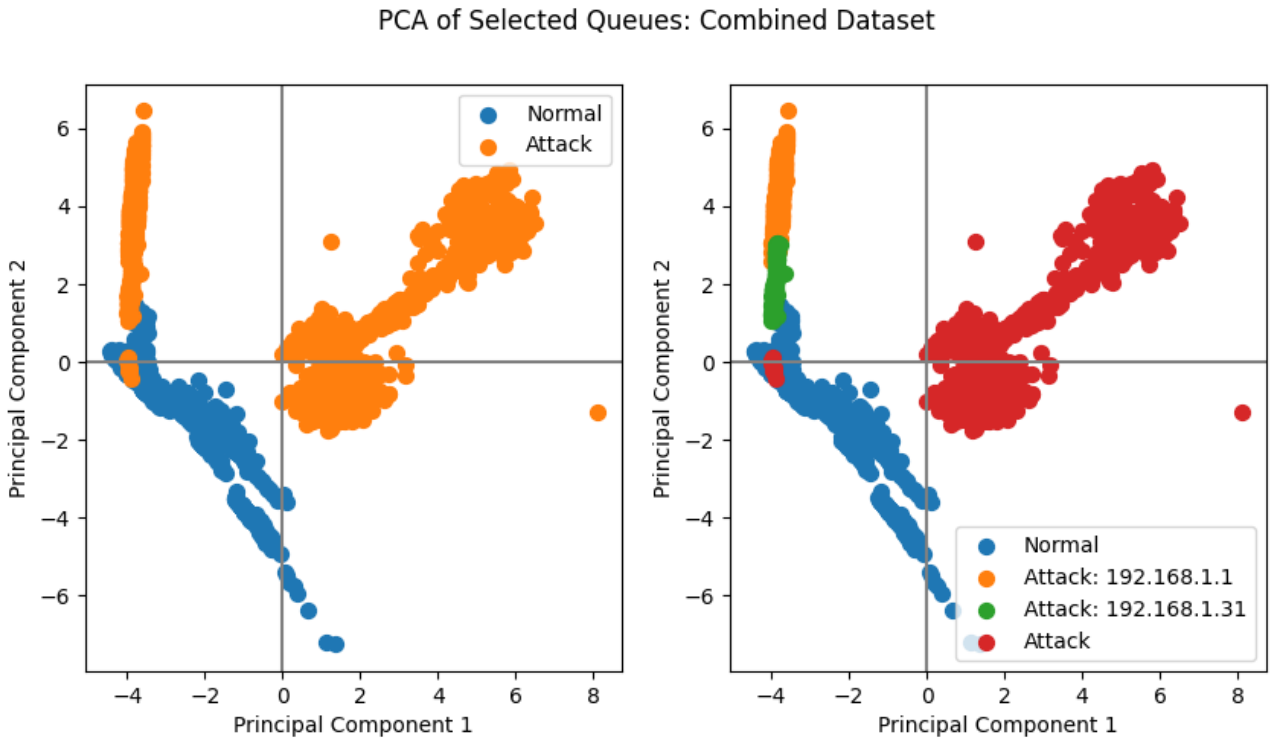

Abbildung 4.4: In der Abbildung wird die Hauptkomponentenanalyse der Host-Selected Queues des Normalbetriebs und des Angriffsfalls visualisiert. Es werden die Hauptkomponenten 1 und 2 analysiert, wobei in der rechten Grafik spezielle Cluster ausgezeichnet sind.

4.2 Analyse der Selected Queues

Bei der Hauptkomponentenanalyse der Selected Queues werden speziell die Host-Selected Queues einbezogen. Dies ermöglicht die Zuordnung von Ereignissen zu den zugehörigen Hosts. In der Analyse enthalten ist der Datensatz des Normalbetriebs und ein Datensatz eines DoS-Angriffs, wobei insgesamt 16 Features berücksichtigt werden. Features bezüglich der Queue-Priorität werden ausgeschlossen.

Die Hauptkomponenten 1 bis 3 bilden 47,96%, 24,92% und 14,63% der Gesamtvarianz der Daten ab, was in Summe 87,51% ergibt. In Abbildung 4.4 wird das Verhältnis zwischen PC1 und PC2 aufgezeigt. PC1 bezieht sich gleichermaßen auf Features, die mit der Länge von Connection Queues als auch der individuellen Länge der Host-Queues in Verbindung stehen. Dagegen werden unter PC2 vor allem Features zusammengeführt, die sich auf die Anzahl der Verbindungen pro Host beziehen. In PC3 werden Features zusammengefasst, die Bezug auf zu den anderen Hosts relative Größen haben, wie die relative Verbindungsanzahl oder die relative Queue-Länge. Auf diese Komponente wird hier nicht näher eingegangen.

Bei Betrachtung des Normalbetriebs in der Abbildung fällt auf, dass die Ereignisse bezüglich PC1 und PC2 eine negative Korrelation aufweisen. In den Extremen gibt es Hosts mit wenigen Verbindungen, die jeweils stark ausgelastet sind und auf der anderen Seite Hosts mit viele Verbindungen, die jeweils nur wenig beansprucht werden. Dies legt die Vermutung nahe, dass

die einzelnen Hosts im Normalbetrieb eine annähernd gleichbleibende *gesamte Auslastung* balancieren, wenn die gesamte Auslastung als Produkt aus Anzahl an Verbindungen und Länge der Host-Queues verstanden wird.

Im Angriffsfall bilden sich zwei Cluster. Das rechte davon weist eine positive Korrelation bezogen auf PC1 und PC2 auf. Das Verhältnis zwischen der Anzahl an Verbindungen und der Länge der Host-Queues bleibt also annähernd gleich. Es wird vermutet, dass beim zugrundeliegenden DoS-Angriff eine Vielzahl an Verbindungen aufgebaut wird und diese direkt voll ausgelastet werden.

Das linke Angriffs-Cluster wird besonders durch die Hosts 192.168.1.1 und 192.168.1.31 repräsentiert, wobei angenommen wird, dass 192.168.1.1 das Standard-Gateway im Subnetz ist. Bezüglich PC1 verläuft das Cluster auf einem vergleichsweise konstanten niedrigen Wert. Bei genauerer Analyse wird festgestellt, dass die Verbindungen auf diesen beiden Hosts beinahe ausschließlich UDP-Verbindungen sind. Die Verbindungsanzahl ist ähnlich verteilt wie im rechten Cluster, wobei das linke Cluster im Mittel eine um den Faktor 1,58 erhöhte Anzahl aufweist. Es wird vermutet, dass neben dem bislang angenommenen Syn Flooding-Angriff auch ein UDP Flooding-Angriff [SRP20, S. 5332] auf diese beiden Hosts stattfindet. Dabei wird eine Vielzahl an nicht zustellbaren UDP-Paketen, typischerweise mit zufälligem Zielport, an einen Empfänger gesendet. Dieser sendet je Paket eine Kontrollnachricht an die angegebene Sender-Adresse, um über die Nichtzustellbarkeit zu informieren. Bei einer Flut von solchen UDP-Paketen wird der Host dadurch blockiert. Die hohe Anzahl unterschiedlicher Zielports spiegelt sich in der hohen Anzahl an UDP-Verbindungen wieder.

5 Fazit

Im Umfang dieser Arbeit werden verschiedene visuelle Darstellungen auf die Eignung zur Analyse hochdimensionaler Daten untersucht. Zur Aufbereitung der Daten eignet sich neben einer tabellarischen Darstellung besonders die Datenstruktur eines Data Cubes. Dimensionen können hierbei dynamisch und flexibel angeordnet und gegenüber gestellt werden.

Durch Gegnüberstellung bestimmter Features und der zeitlichen Dimension in Form von Liniendiagrammen werden Inkonsistenzen in der Kontinuität ersichtlich. Flächendiagramme eignen sich besonders, um Zusammenhänge verschiedener Ebenen von Daten zu untersuchen, wie beispielsweise die gesamte Anzahl aller Host-Datenpakete mit der individuellen Anzahl von Paketen der einzelnen Hosts in Abbildung 3.4. Mit diesen Diagrammtypen können verschiedene Szenarien in absoluten Zahlen miteinander verglichen werden. Beispielsweise wird beim Vergleich des normalen Netzwerkbetriebs mit einem DoS-Angriff die starke Diskrepanz in der Länge der Paket-Queues deutlich, wie in Abbildung 3.1 ersichtlich.

Auf den Einsatz von Diagrammen mit drei Koordinatenachsen wurde verzichtet. Stattdessen wird der Zusammenhang von drei Dimensionen durch die Gegenüberstellung zweier separater 2D-Diagramme, wie beispielsweise in Abbildung 4.2, untersucht. Weitere Dimensionen können gegebenenfalls durch die Verwendung unterschiedlicher Farben und Marker eingebracht werden, was in Abbildung 4.3 zum Einsatz kommt.

Die Hauptkomponentenanalyse stellt sich als besonders geeignetes Mittel heraus, um wesentliche Unterschiede in hochdimensionalen Daten zu erkennen. Selbst bei hochdimensionalen Daten bis zu 50 Dimensionen wie bei der Hauptkomponentenanalyse der General Queue bei einer Kombination verschiedener Datensätze konnte mit zwei Hauptkomponenten eine Abdeckung der Gesamtvarianz der Daten von mindestens 67,58% erreicht werden. Sie bietet eine visuelle Analyse der Unterschiede, die durch die Bildung von Clustern deutlich werden. In der Analyse des Netzwerkverkehrs bilden besonders die Ereignisse der Typen “Normalbetrieb” und “DoS-Angriff” stark voneinander abgrenzbare Cluster. Zudem wird ersichtlich, dass in den analysierten Szenarien, die wesentlichen Unterschiede bereits der ersten Hauptkomponente entnommen werden können. Ein Ransomware-Angriff unterscheidet sich bezüglich des Umfangs des Netzwerkverkehrs weniger stark vom Normalbetrieb. Bezogen auf andere Hauptkomponenten wird dennoch eine Abgrenzung ersichtlich, wobei allerdings noch kein kausaler Zusammenhang festgestellt wurde. Eine Überlegung für weitere Analysen ist, die zeitliche Dimension den Hauptkomponenten gegenüber zu stellen. Es ist zudem denkbar, die Hauptkomponentenanalyse in einem NIDS einzusetzen, um automatisiert bestimmte Angriffsszenarien zu erkennen.

Literaturverzeichnis

[Bra14] Tim Bray. *The JavaScript Object Notation (JSON) Data Interchange Format*. RFC 7159. März 2014. DOI: 10.17487/RFC7159. URL: https://www.rfc-editor.org/info/rfc7159.

[Bre16] Ross Brewer. "Ransomware attacks: detection, prevention and cure". In: *Network Security* 2016.9 (2016), S. 5–9. ISSN: 1353-4858. DOI: 10.1016/S1353-4858(16)30086-1.

[Bul18] Viktoras Bulavas. "Investigation of network intrusion detection using data visualization methods". In: *2018 59th International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*. 2018, S. 1–6. DOI: 10.1109/ITMS.2018.8552977.

[GBL*96] Jim Gray, Adam Bosworth, Andrew Layman und Hamid Pirahesh. "Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total". In: *Proceedings of the Twelfth International Conference on Data Engineering*. ICDE '96. USA: IEEE Computer Society, 1996, S. 152–159. ISBN: 0818672404.

[McK11] Wes McKinney. "pandas: a Foundational Python Library for Data Analysis and Statistics". In: *Python High Performance Science Computer* (Jan. 2011).

[SKK*97] C.L. Schuba, I.V. Krsul, M.G. Kuhn, E.H. Spafford, A. Sundaram und D. Zamboni. "Analysis of a denial of service attack on TCP". In: *Proceedings. 1997 IEEE Symposium on Security and Privacy (Cat. No.97CB36097)*. 1997, S. 208–223. DOI: 10.1109/SECPRI.1997.601338.

[SRP20] Mikail Mohammed Salim, Shailendra Rathore und Jong Hyuk Park. "Distributed denial of service attacks and its defenses in IoT: a survey". In: *The Journal of Supercomputing* 76 (2020), S. 5320–5363. DOI: 10.1007/s11227-019-02945-z.

[WEG87] Svante Wold, Kim Esbensen und Paul Geladi. "Principal component analysis". In: *Chemometrics and Intelligent Laboratory Systems* 2.1 (1987). Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, S. 37–52. ISSN: 0169-7439. DOI: 10.1016/0169-7439(87)80084-9.