# Deep Analysis and Detection of Firewall Anomalies Using Knowledge Graph

Abdelrahman Osman Elfaki[a] and Amer Aljaedi[b]
*College of Computing and Information Technology, University of Tabuk, Tabuk 71491, Saudi Arabia*

Keywords: Firewall Anomalies, Firewall Rules, Network Policies, Knowledge Graph.

Abstract: Implementing firewall policy with defining firewall rules is a cumulative process that could take place in different periods and depend on the network conditions, which makes it prone to errors and difficult to validate without effective tools. Such tools should be carefully designed to capture and spot firewall configuration errors and anomalies. The solution in this paper consists of four steps, which are: formalizing the firewall rules by using FOL, defining the general form of the anomaly, collecting all active destinations' IP addresses and port numbers in updated lists, and applying the proposed FOL rules for detecting firewall anomalies. The general form has been represented by using knowledge graph for supporting visualization aiming to detect firewall anomalies by extracting knowledge from the knowledge graph and its formalization rules. The proposed method is efficient and capable of discovering all types of firewall anomalies.

## 1 INTRODUCTION

Nowadays, many daily activities depend on the internet services, such as online shopping, learning, or entertainment. Some of these activities are sensitive and require the deployment of security measures. Firewall is a cornerstone in internet security as it is can be defined as a network security system that controls and monitors incoming and outgoing network traffic based on a predetermined security set of rules (Boudriga, 2017). This set of rules has been commonly declared as firewall rules or filtering rules in the literature. In this paper, we will use these two interchangeable names to refer to the set of firewall rules.

According to the research work in (Chao, 2018), filtering rules in firewalls could reach hundreds or even thousands of rules. Hence, it is impossible to ensure the correctness and efficiency of these rules without automated tools. Such automated tools should validate the set of filtering rules and assist in the optimization process. According to (Voronkov et al., 2018) setting up a firewall is a complicated process and an error-prone task. As facts, the legacy firewall rules might be designed and implemented by different administrators (Hu et al., 2012). In addition, the firewall policies might be maintained in the network by more than one administrator. These facts explicate the complicated nature of managing firewall policies and their implementation. In practice, the network may face security issues due to firewall configuration errors. In literature, these configuration errors in firewall rules have been collectively described as firewall anomalies. The prominent firewall anomalies that have been discussed in literature are *shadowing*, *redundancy*, *correlation*, *generalization*, and *irrelevant* (Abbes et al., 2016; As-Suhbani & Khamitkar, 2017; Ahmed & Askari, 2018; Kim et al., 2021). In the following, we discuss each anomaly case in the firewall rule context.

*Shadowing* anomaly rule can be identified as a preceding rule (i.e., superior rule by its order in the set) that matches all the networking traffic packets which also could match other subsequent rules in the set. Hence, such the subsequent rules will never be counted since they are shadowed by the preceding rule that its matching fields cover a larger range of addresses. In the following, *fr* denotes a firewall rule. A rule $fr_2$ is shadowed by rule $fr_1$ if $fr_2$ follows $fr_1$, and the $fr_2$ matching fields are subsets of the corresponding fields in $fr_1$ while the actions of $fr_1$ and $fr_2$ are different. Table 1 illustrates an instance of the

[a] https://orcid.org/0000-0002-8881-0504
[b] https://orcid.org/0000-0003-4099-5025

Elfaki, A. and Aljaedi, A.
Deep Analysis and Detection of Firewall Anomalies Using Knowledge Graph.
DOI: 10.5220/0011704900003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 411-417
ISBN: 978-989-758-626-2; ISSN: 2184-4313

shadowing anomaly. On the other hand, if the actions of $fr_1$ and $fr_2$ are same, then this case represents *redundancy* in the firewall rule set.

Table 1: An instance of shadowing anomaly.

| Index | Proto. | Src-IP | Src-Port | Dst-IP | Dst-Port | Action |
|---|---|---|---|---|---|---|
| $I_1$ | TCP | 10.0.0.0/28 | * | 20.0.0.0/28 | 22 | Deny |
| $I_2$ | TCP | 10.0.0.5 | * | 20.0.0.3 | 22 | Allow |

According to Karafili et al. (2020), *correlation* anomaly is defined as two firewall rules with opposite action conditions, but they can match some packets of each other. Table 2 illustrates an instance of *correlation* anomaly. In case the packet fields are `(TCP,10.0.0.5,*,20.0.0.3, 22)`, the action of $fr_1$ will drop the network packet, as it is matched first with $fr_1$ before it can reach $fr_2$, which forwards it. Therefore, the result of *correlation* anomaly could cause conflicting actions on the network packets.

Table 2: An instance of correlation anomaly.

| Index | Proto. | Src-IP | Src-Port | Dst-IP | Dst-Port | Action |
|---|---|---|---|---|---|---|
| $I_1$ | TCP | 10.0.0.0/28 | * | 20.0.0.3 | 22 | Deny |
| $I_2$ | TCP | 10.0.0.5 | * | 20.0.0.0/28 | 22 | Allow |

According to (Ahmed & Askari, 2018), the *generalization* anomaly occurs when there are two rules having contradictory decisions, and all the packets matched by one of the rules are a subset of the matched packets of the second rule. Table 3 shows an instance of *generalization* anomaly. In case the packet `(TCP,10.0.0.5, *,20.0.0.3, 22)`, the action of $fr_1$ will drop the packet, whereas the action of $fr_2$ would forward it. Note that the second rule is generalization of the first rule, and if the order of the rules is reversed, it will result in shadowing anomaly. The *generalization* is considered merely warning, especially when updating the firewall policy, as it is used to make exceptions for the general rules.

Table 3: An instance of generalization anomaly.

| Index | Proto. | Src-IP | Src-Port | Dst-IP | Dst-Port | Action |
|---|---|---|---|---|---|---|
| $I_1$ | TCP | 10.0.0.5 | * | 20.0.0.3 | 22 | Deny |
| $I_2$ | TCP | 10.0.0.0/28 | * | 20.0.0.0/28 | 22 | Allow |

The last firewall anomaly is *irrelevant* anomaly. According to (Ahmed & Askari, 2018), an *irrelevant* firewall rule is a rule that would not match any packets that traverse the firewall system. To the best of our knowledge, we can summarize the *irrelevant* firewall rules in the following two common cases: 1) An erroneous entry of the IP addresses within the firewall rule such as identical IPs for the source and destination addresses, and 2) The firewall rule contains irrelevant addresses according to the network addressing scheme of the 5-tuple traffic flow (i.e., non-existent destinations or services). For instance, in the first rule of Table 4, the IP addresses of the source and destination are identical. Hence, the firewall rule $fr_1$ will not match any received packet.

Table 4: Instances of irrelevant anomaly.

| Index | Proto. | Src-IP | Src-Port | Dst-IP | Dst-Port | Action |
|---|---|---|---|---|---|---|
| $I_1$ | TCP | 10.0.0.5 | * | 10.0.0.5 | 22 | allow |
| $I_2$ | UDP | 10.0.0.5 | * | 50.0.0.3 | 70 | allow |

The second rule $fr_2$ of Table 4 presents an instance of *irrelevant* anomaly (case 2). The flow tuple `(UDP, 50.0.0.3, 70)` would represent a non-existent destination if there was no machine in the network with that IP address `(50.0.0.3)`, or a non-existent service if there was no application that is listening on port 70, or both. Hence, the *irrelevant* rule will not match any received packet since the destination, or the service does not exist or has been removed from the network.

Knowledge graph has been utilized in many domains for knowledge representation and reasoning (Elfaki et al., 2019). In this paper, the firewall anomalies have been described by using knowledge graph and predicate calculus. The knowledge graph can assist in providing pictorial visualization for network administrators while predicate calculus aims to provide solution formalization.

## 2 RELATED WORK

Hu et al. (2012) have suggested a firewall policy anomaly management framework, named as Firewall Anomaly Management Environment (FAME). This framework has been used for anomaly detection and resolution. For anomaly detection, a rule-based segmentation technique has been adopted, and for conflict resolution, a risk level has been utilized as a strategy. The risk level is defined based on Common Vulnerability Scoring System (CVSS). The implementation of FAME has been done based on JavaBDD which has scalability issues.

Abbes et al. (2016) proposed Firewall Anomaly Tree (FAT) as a representation technique for firewall rules. The essential concept of their technique is the drawing paths of the filtering rules in a tree, and hence the common paths that reflect anomalies could be detected. While it is obvious the paths that are taken by individual rules, it is not clear how aggregate rules could be illustrated in the graph. The work in (Von et al., 2020) used Knowledge Graph to handle the difficulties of extracting knowledge from multi-source security logs. Also, the work in (Wang et al., 2021) used Knowledge Graph application for extracting knowledge from 15 social attack incidents and scenarios. The results show how Knowledge Graph is capable to define social engineering threat elements and potential threats to victims. Wang et al. (2021) developed Knowledge Graph for predicting the possible attack path. They considered CVSS's quantitative indicators for a single vulnerability and combine the network security evaluation method to calculate the possible paths.

# 3 METHODOLOGY

In this section, we discuss the methodology that has been adopted to develop our proposed method. According to (Chao, 2018), the form of standard firewall rule is `<order><protocol><src_ip><src_port><dst_ip><dst_port><action>`. In our proposed method, the firewall rule has been represented by using First Order Logic (FOL). A predicate named "rule" with eight parameters. In the following, the syntax and semantic of this predicate have been explained in detail.

**Syntax:** *fr* (`index, proto, scr_IP, src_port, dst_IP, dst_port, action`).

**Semantic:** *fr* denotes firewall rule, `index` denotes the identifier for the rule (i.e., it can be used to prioritize rules), `proto` denotes the used networking protocol, `scr_IP` denotes the source IP address, `src_port` denotes source port of the machine, `dst_IP` denotes the destination IP address, `dst_port` denotes the destination port in the other end of the communication, `action` denotes the instruction that will be taken when the network packet matches the rule. Table 5 shows an instance of standard firewall rule, which will be represented in the proposed method as: *fr*(`I,tcp,10.0.0.5, *,20.0.0.3, 22, allow`).

## 3.1 Assumptions

The proposed method has two assumptions, which are: a rule represents a single packet, and it uses IP addresses and ports that are defined in updated lists. In the following, the methodology has been discussed in steps.

1) Symbolizing the firewall rules by using FOL;
2) Defining the general form of the anomaly;
3) Collecting all active IPs addresses and ports in updated lists;

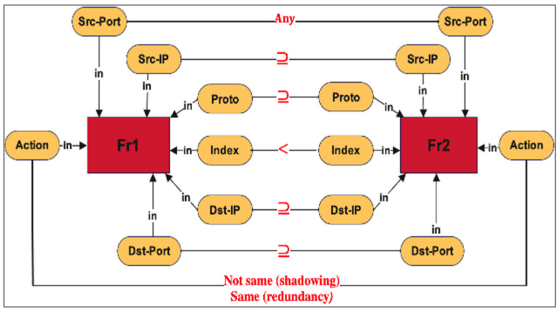

Figure 1: Knowledge graph depicts shadowing and redundancy anomalies.

4) Appling the proposed FOL rules for detecting firewall anomalies.

## 3.2 Validation Rules

Regarding aforementioned discussion, the firewall anomalies could be summarized into four categories as follows:

$$\texttt{Shadowing} \Longrightarrow \left(\forall l, \left(fr_a^l \supseteq fr_b^l\right)\right) \wedge \left(fr_a^{action} \neq fr_b^{action}\right) \quad (1)$$

$$\texttt{Redundancy} \Longrightarrow \left(\forall l, \left(fr_a^l \supseteq fr_b^l\right)\right) \wedge \left(fr_a^{action} = fr_b^{action}\right) \quad (2)$$

$$\begin{aligned}\texttt{Correlation} \Longrightarrow & \left(\exists l, \left(fr_a^l \supseteq fr_b^l\right)\right) \wedge \left(\exists l', \left(fr_a^{l'} \subseteq fr_b^{l'}\right)\right) \\ & \wedge \left(fr_a^{action} \neq fr_b^{action}\right)\end{aligned} \quad (3)$$

$$\begin{aligned}\texttt{Generalization} \Longrightarrow & \left(\forall l, \left(fr_a^l \subseteq fr_b^l\right)\right) \\ & \wedge \left(fr_a^{action} \neq fr_b^{action}\right)\end{aligned} \quad (4)$$

Where fr is a firewall rule, *a* and *b* are orders of the firewall rules such as a > b, while *action* is the instruction of the firewall rule ∈ (allow, deny), which is basically means forward the network packet or drop it. The *l* is the individual matching field of the firewall rule such as source IP (`scr_IP`), destination IP (`dst_IP`), source port (`src_port`), destination port (`dst_port`), and the communication protocol (`proto`).

# 4 FIREWALL ANOMALIES DETECTION

In this section, we present and discuss the first order calculus predicates that are used to detect firewall rules anomalies.

Table 5: An instance of standard firewall rule.

| Index | Proto. | Src-IP | Src-Port | Dst-IP | Dst-Port | Action |
|---|---|---|---|---|---|---|
| I₁ | TCP | 10.0.0.5 | * | 20.0.0.3 | 22 | allow |

## 4.1 Shadow Anomaly

The general form of shadow anomaly as below:

$$\forall\ fr, pk:\ (I_1<I_2)\wedge(pk_1 \in fr_1)\wedge(pk_2 \in fr_1)\wedge(pk_2 \in fr_2)\wedge(pk_1 \backslash == pk_2)\Rightarrow fr_2 \text{ is shadowed by } fr_1 \tag{5}$$

The equation (5) denotes that rule *fr*$_2$ comes after *fr*$_1$ in the firewall rules ordering, i.e., *fr$_1$* will execute first. The *fr$_2$* can match two types of packets ($pk_1$ and $pk_2$, and *fr$_2$* has only type $pk_2$. Hence, equation (5) shows the general shadow anomaly, where all packet matching fields in *fr$_2$* are belonging to *fr$_1$* (i.e., subsets of *fr$_1$*). Thus, *fr$_2$* is shadowed by *fr$_1$*. This case represents *shadowing* anomaly, and if the actions are equal, then it is considered as *redundancy*. Figure 1 shows knowledge graph depicts *shadow* anomaly.

In the following, the *shadowing* anomaly has been classified into four cases as following:

$$\forall: I, proto, \text{src\_IP}, \text{src\_port}, \text{dst\_IP}, \text{dst\_port}, action:\ fr_1(I_1, proto_1, \text{src\_IP}_1, \text{src\_port}_1, \text{dst\_IP}_1, \text{dst\_port}_1, action_1)\wedge fr_2(I_2, proto_2, \text{src\_IP}_2, \text{src\_port}_2, \text{dst\_IP}_2, \text{dst\_port}_2, action_2)\wedge(I_1<I_2)\wedge(proto_1 = proto_2)\wedge(\text{src\_IP}_1 = \text{src\_IP}_2)\wedge(\text{dst\_IP}_1 = \text{dst\_IP}_2)\wedge(\text{dst\_port}_1 = \text{dst\_port}_2)\wedge(action_1 \sim= action_2)\Rightarrow \text{Shadowing case1} \tag{6}$$

Equation (6) presents the first case of a shadowing rule, where *fr$_1$* and *fr$_2$* are denoted by predicates *fr$_1$* and *fr$_2$* respectively. The *fr$_1$* and *fr$_2$* have the same protocols, source IPs, destination IPs, destination ports, and different actions. The *fr$_1$* is preceding *fr$_2$*. We have called it first case of *shadowing*. In the following, the rest of other possible cases of *shadowing* rules are presented.

$$\forall: I, proto, \text{src\_IP}, \text{src\_port}, \text{dst\_IP}, \text{dst\_port}, action:\ fr_1(I_1, proto_1, \text{src\_IP}_1, \text{src\_port}_1, \text{dst\_IP}_1, \text{dst\_port}_1, action_1)\wedge fr_2(I_2, proto_2, \text{src\_IP}_2, \text{src\_port}_2, \text{dst\_IP}_2, \text{dst\_port}_2, action_2)\wedge(I_1<I_2)\wedge((proto_1 = any)\vee(proto_2 \in proto_1))\wedge(\text{src\_IP}_2 = \text{src\_IP}_1)\wedge(\text{dst\_IP}_2 = \text{dst\_IP}_1)\wedge(\text{dst\_port}_2 = \text{dst\_post}_1)\wedge(action_2 \sim= action_1)\Rightarrow \text{Shadowing case2} \tag{7}$$

Equation (7) denotes that *fr*$_2$ is shadowed by *fr*$_1$. In *fr$_1$*, the protocol is equivalent to “any”, i.e., all types of protocols, TCP or UDP, which means the protocol of *fr*$_2$ is part of the protocols in *fr*$_1$ regardless of the protocol type in *fr*$_2$. The other case, $proto_2 \in proto_1$, means that the protocol in *fr*$_2$ is belonging to the protocol in *fr*$_1$, i.e., $proto_1 = \{TCP, UDP\}$, hence $proto_2 \in proto_1$. The rest of packet matching fields are equal in *fr*$_1$ and *fr*$_2$ but with different actions.

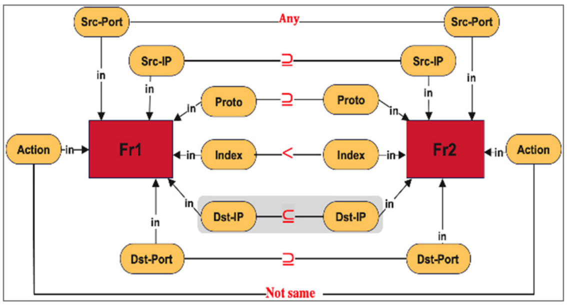

Figure 2: Knowledge graph depicts a case of correlation anomaly.

$$\forall: I, proto, \text{src\_IP}, \text{src\_port}, \text{dst\_IP}, \text{dst\_port}, action:fr_1(I_1, proto_1, \text{src\_IP}_1, \text{src\_port}_1, \text{dst\_IP}_1, \text{dst\_port}_1, action_1)\wedge fr_2(I_2, proto_2, \text{src\_IP}_2, \text{src\_port}_2, \text{dst\_IP}_2, \text{dst\_port}_2, action_2)\wedge(I_1<I_2)\wedge(proto_2 = proto_1)\wedge((\text{src\_IP}_1 = any)\vee(\text{src\_IP}_2 \in \text{src\_IP}_1))\wedge(\text{dst\_IP}_2 = \text{dst\_IP}_1)\wedge(\text{dst\_port}_2 = \text{dst\_post}_1)\wedge(action_2 \sim= action_1)\Rightarrow \text{Shadowing case3} \tag{8}$$

Equation (8) denotes that *fr$_2$* is shadowed by *fr$_1$*. In this equation, there are two scenarios of the source IP of *fr$_1$*, either equal to “any”, or it is an aggregate IP. An aggregate source IP means that it covers a range of source IP addresses in single rule. The source IP of *fr$_2$* is a single address or it covers a smaller range of addresses that belong to the source IP of *fr$_1$*. For instance, suppose source IP of *fr$_1$* = {10.0.0.1-10.0.0.10} and the source IP of *fr$_2$* is {10.0.0.5}, hence source IP of *fr$_2$* is belonging to *fr$_1$*. In equation 8, the source IP of *fr$_2$* is belonging to *fr$_1$* and the rest of packet matching fields are equivalent, but with different actions. Therefore, *fr$_1$* is *shadowing fr$_2$*.

$$\forall: I, proto, \text{src\_IP}, \text{src\_port}, \text{dst\_IP}, \text{dst\_port}, action:\ fr_1(I_1, proto_1, \text{src\_IP}_1, \text{src\_port}_1, \text{dst\_IP}_1, \text{dst\_port}_1, action_1)\wedge fr_2(I_2, proto_2, \text{src\_IP}_2, \text{src\_port}_2, \text{dst\_IP}_2, \text{dst\_port}_2, action_2)\wedge(I_1<I_2)\wedge(\text{src\_IP}_2 = \text{src\_IP}_1)\wedge((\text{dst\_IP}_1 = any)\vee(\text{dst\_IP}_2 \in \text{dst\_IP}_1))\wedge(action_2 \sim= action_1)\Rightarrow \text{Shadowing case4} \tag{9}$$

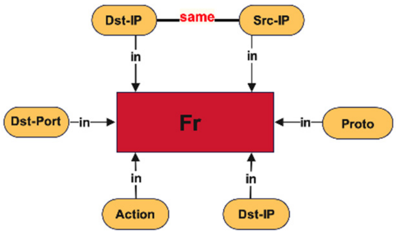

Figure 4: Knowledge graph depicts a case of irrelevant anomaly.

Equation (9) denotes that *fr$_2$* is shadowed by *fr$_1$*. In this equation, there are two scenarios of the destination IP address of rule 1 as either equal to "any", or it is an aggregate IP address. An aggregate destination IP means that it covers a range of destination IP addresses in single rule, *$fr_1$*. The destination IP of *$fr_2$* is a single address or it covers a smaller range of addresses that belong to the destination IP of *$fr_1$*. For instance, suppose destination IP of *$fr_1$* = {10.0.0.10-10.0.0.20} and destination IP of *$fr_2$* is {10.0.0.11}, hence destination IP of *$fr_2$* is belonging to *$fr_1$*. In equation (9), the destination IP of *$fr_2$* is belonging to *$fr_1$* and the rest of packet matching fields are equivalent, but with different actions. Therefore, *$fr_2$* is shadowed by *$fr_1$* .

$$\forall:\text{I,proto,src\_IP,src\_port,dst\_IP,dst\_port,action}:\text{fr}_1(\text{I}_1,\text{proto}_1,\text{src\_IP}_1,\text{src\_port}_1,\text{dst\_IP}_1,\text{dst\_port}_1,\ \text{action}_1)\wedge\text{fr}_2(\text{I}_2,\text{proto}_2,\text{src\_IP}_2,\text{src\_port}_2,\text{dst\_IP}_2,\ \text{dst\_port}_2,\ \text{action}_2)\wedge(\text{I}_1<\text{I}_2)\wedge(\text{src\_IP}_2=\text{src\_IP}_1)\wedge(\text{dst\_IP}_2=\text{dst\_IP}_1)\wedge((\text{dst\_port}_1=\text{any})\vee(\text{dst\_port}_2\in\text{dst\_port}_1))\wedge(\text{action}_2{\sim}{=}\text{action}_1)\Longrightarrow\text{Shadowing case5}\tag{10}$$

Equation (10) denotes that *$fr_2$* is shadowed by *$fr_1$*. In this equation, there are two scenarios of the destination port of *$fr_1$*, either equal to "any", or it is a range of port numbers, which aggregate a number of destination ports in single rule; *$fr_1$*. The destination port of *$fr_2$* is a single port or it can be a smaller range of port numbers that belong to the destination port numbers in *$fr_1$*. For instance, suppose destination port of *$fr_1$* = {2069-3068} and the destination port of *$fr_2$* is {3035}, hence the destination port of *$fr_2$* is belonging to *$fr_1$*. In equation (10), destination port of *$fr_2$* is belonging to *$fr_1$* and the rest of packet matching fields are equivalent, but with different actions. Therefore, *$fr_2$* is shadowed by *$fr_1$*.

## 4.2 Correlation Anomaly

The difference between *shadowing* and *correlation* anomalies is that in *shadowing* all packets of *$fr_2$* are covered by *$fr_1$* with different actions, whereas in *correlation* some packets of *$fr_2$* are covered in *$fr_1$*, and also some packets of *$fr_1$* can be matched by *$fr_2$*, which has different action. Table 2 shows an instance of *correlation* anomaly. Figure 2 shows knowledge graph represents *correlation* anomaly. The general form of *correlation* anomaly as follows:

$$\exists\,\text{pk},\forall\,\text{fr}:(\text{I}_1<\text{I}_2)\wedge(\text{pk}_1\in\ \text{fr}_1)\wedge(\text{pk}_2\in\text{fr}_2)\wedge((\text{pk}_2\in\text{fr}_1)\vee(\text{pk}_1\in\ \text{fr}_2))\Longrightarrow\ \text{fr}_2\text{ is correlated with }\text{fr}_1\tag{11}$$

Equation (11) denotes the general form of *correlation* anomaly, where pk denotes packets and *fr* denotes firewall rule. In equation (11), packet $\text{pk}_1$ belongs to *$fr_1$*, packet $\text{pk}_2$ belongs to *$fr_2$*, and *$fr_1$* is preceding *$fr_2$*. Some of the packets that are belonging to *$fr_2$* are belonging as well to *$fr_1$*, or some of packets that are belonging to *$fr_1$* are also belonging to *$fr_2$*. Hence, *$fr_2$* is correlated with *$fr_1$*. By using distributive theory, equation (11) could be seen as two parts, part one is: $(\text{I}_1<\text{I}_2)\wedge\ (\text{pk}_1\in\text{fr}_1)\ \wedge(\text{pk}_2\in\text{fr}_1)$, the second part is: $(\text{I}_1<\text{I}_2)\wedge(\text{pk}_1\in\ \text{fr}_1)\wedge(\text{pk}_2\in\ \text{fr}_2)\wedge\ (\text{pk}_1\in\text{fr}_2)$ . In regard to part one, if the packets belong to *$fr_2$* and in the same time can match *$fr_1$*, *$fr_2$* is considered shadowed by *$fr_1$*. For instance, $(\text{pk}_1\in\text{fr}_1)\wedge(\text{pk}_2\in\text{fr}_2)\wedge(\text{pk}_2\in\ \text{fr}_1)\Longrightarrow\text{Shadowing}$. Hence, the analysis of *shadowing* could be applied to the first part of *correlation*. The second part of general *correlation* equation: $(\text{I}_1<\text{I}_2)\wedge(\text{pk}_1\in\ \text{fr}_1)\wedge(\text{pk}_2\in\ \text{fr}_2)\wedge(\text{pk}_1\in\text{fr}_2)$ is discussed below:

$$\forall:\text{I,proto,src\_IP,src\_port,dst\_IP,dst\_port,action}:\ \text{fr}_1(\text{I}_1,\text{proto}_1,\text{src\_IP}_1,\text{src\_port}_1,\text{dst\_IP}_1,\text{dst\_port}_1,\text{action}_1)\wedge\ \text{fr}_2(\text{I}_2,\text{proto}_2,\text{src\_IP}_2,\text{src\_port}_2,\text{dst\_IP}_2,\text{dst\_port}_2,\text{action}_2)\wedge(\text{I}_1<\text{I}_2)\wedge((\text{proto}_2=\text{any})\vee(\text{proto}_1\in\text{proto}_2))\wedge(\text{src\_IP}_2=\text{src\_IP}_1)\wedge(\text{dst\_IP}_2=\text{dst\_IP}_1)\wedge(\text{dst\_port}_2=\text{dst\_post}_1)\ \wedge(\text{action}_2{\sim}{=}\text{action}_1)\Longrightarrow\text{Correlation1}\tag{12}$$

Equation (12) denotes that *$fr_2$* is correlated with *$fr_1$* as the rule with index $\text{I}_1$ is part of the rule with index $\text{I}_2$. In *$fr_2$*, the protocol type is equivalent to "any", i.e., all types of protocols, which means the protocol of *$fr_1$* will be part of the protocol in *$fr_2$* regardless of the protocol type in *$fr_1$*. The other case, $\text{proto}_2\in\text{proto}_1$, means that the protocol in *$fr_2$* is belonging to the protocol in *$fr_1$*, i.e., $\text{proto}_1=\{\text{TCP},\ \text{UDP}\}$, hence $\text{proto}_2\in\text{proto}_1$. The rest of packet matching fields are equal in *$fr_1$* and *$fr_2$*, but with different actions.

$$\forall:\text{I,proto,src\_IP,src\_port,dst\_IP,dst\_port,action}:\text{fr}_1(\text{I}_1,\text{proto}_1,\text{src\_IP}_1,\text{src\_port}_1,\text{dst\_IP}_1,\text{dst\_port}_1,\text{action}_1)\wedge\ \text{fr}_2(\text{I}_2,\text{proto}_2,\text{src\_IP}_2,\text{src\_port}_2,\text{dst\_IP}_2,\text{dst\_port}_2\ ,\text{action}_2)\wedge(\text{I}_1<\text{I}_2)\wedge(\text{proto}_2=\text{proto}_1)\wedge((\text{src\_IP}_2=\text{any})\vee(\text{src\_IP}_1\in\text{src\_IP}_2))\ \wedge\ (\text{dst\_IP}_2=\text{dst\_IP}_1)\wedge(\text{dst\_port}_2=\text{dst\_post}_1)\wedge(\text{action}_2{\sim}{=}\text{action}_1)\Longrightarrow\text{Correlation2}\tag{13}$$

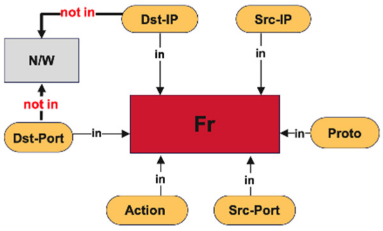

Figure 5: Knowledge graph depicts irrelevant anomalies of non-existent destination.

Equation (13) denotes that $fr_2$ is correlated with $fr_1$. In this equation, there are two scenarios of the source IP of $fr_2$, either equal to "any", or it is an aggregate source IP address. The source IP address of $fr_1$ is a single address or might be a smaller range of source IP addresses that belong to the source IP range of $fr_2$.

$$\forall{:}\,\text{I,proto,src\_IP,src\_port,dst\_IP,dst\_port,action}{:}\ \text{fr}_1(\text{I}_1,\text{proto}_1,\text{src\_IP}_1,\text{src\_port}_1,\text{dst\_IP}_1,\text{dst\_port}_1,\text{action}_1) \land \text{fr}_2(\text{I}_2,\text{proto}_2,\text{src\_IP}_2,\text{src\_port}_2,\text{dst\_IP}_2,\text{dst\_port}_2,\text{action}_2) \land (\text{I}_1<\text{I}_2) \land (\text{src\_IP}_2 = \text{src\_IP}_1) \land ((\text{dst\_IP}_2 = \text{any}) \lor (\text{dst\_IP}_1 \in \text{dst\_IP}_2)) \land (\text{action}_2 \sim= \text{action}_1) \Longrightarrow \text{Correlation3} \tag{14}$$

The equation (14) denotes that $fr_2$ is correlated with $fr_1$. In this equation, there are two scenarios of the destination IP of $fr_2$, either equal to "any", or it is an aggregate IP for destination addresses in single rule, $fr_2$. The destination IP of $fr_1$ is a single address or it is a smaller range of destination IP addresses that belong to the destination IP range in $fr_2$. Hence, the destination IP of $fr_1$ is belonging to $fr_2$. In equation (14), destination IP of $fr_1$ is belonging to $fr_2$ and the rest of packet matching fields are equivalent in $fr_1$ and $fr_2$, but with different actions.

$$\forall{:}\,\text{I,proto,src\_IP,src\_port,dst\_IP,dst\_port,action}{:}\ \text{fr}_1(\text{I}_1,\text{proto}_1,\text{src\_IP}_1,\text{src\_port}_1,\text{dst\_IP}_1,\text{dst\_port}_1,\text{action}_1) \land \text{fr}_2(\text{I}_2,\text{proto}_2,\text{src\_IP}_2,\text{src\_port}_2,\text{dst\_IP}_2,\text{dst\_port}_2\ ,\text{action}_2) \land (\text{I}_1<\text{I}_2) \land (\text{src\_IP}_2 = \text{src\_IP}_1) \land (\text{dst\_IP}_1 = \text{dst\_IP}_1) \land ((\text{dst\_port}_2 = \text{any}) \lor (\text{dst\_port}_1 \in \text{dst\_port}_2)) \land (\text{action}_2 \sim= \text{action}_1) \Longrightarrow \text{Correlation3} \tag{15}$$

Equation (15) denotes that $fr_2$ is correlated with $fr_1$. In this equation, there are two scenarios of the destination port number(s) of $fr_2$, either equal to "any", or it is aggregate port numbers. Aggregate destination port numbers means there are multiple destination ports in a single rule, $fr_2$. The destination port of $fr_1$ is a single port or can be a smaller range of destination port numbers that belong to the destination port of $fr_2$. In equation (15), the destination port of $fr_1$ is belonging to $fr_2$ and the rest of packet matching fields are equivalent in $fr_1$ and $fr_2$, but with different actions. Therefore, $fr_1$ is correlated with $fr_2$.

### 4.3 Generalization Anomaly

The generalization anomaly occurs when there are two rules having contradictory decisions and all the packets matched by one of the rules are a subset of the packets that match the second rule. The difference between shadowing and generalization anomalies could be represented by the following equation:

*Shadowing*: $fr_2 \subseteq fr_1$:- Rule2 is subset of rule1
*Generalization*: $fr_1 \subseteq fr_2$:- Rule1 is subset of rule2

Table 3 shows an instance of *generalization* anomaly. The general form of the *generalization* anomaly as follows:

$$\forall\ \text{fr, pk}{:}\ (\text{fr}_1<\text{fr}_2) \land (\text{pk}_1 \in \text{fr}_1) \land (\text{pk}_2 \in \text{fr}_2) \land (\text{pk}_1 \in \text{fr}_2) \Longrightarrow \text{fr}_2 \text{ is generalization of } \text{fr}_1 \tag{16}$$

This general form of *generalization* anomaly is similar to the second part of *correlation* anomaly. Hence, same equations that are applied to satisfy second part of the *correlation* anomaly are suitable to satisfy the *generalization* anomaly which are equations 12-15.

### 4.4 Irrelevant Anomaly

As it mentioned in Section 1, the *irrelevant* firewall rule is a rule that would not match any transmitted packets in the network. Such anomaly could occur when the firewall rules are not updated along with the changes in network topology or configurations (i.e., added/ removed devices or addressing scheme) or due to rules misconfigurations. Equation (17) denotes *irrelevant* anomaly (rule 1 in Table 4). In equation (17), the source and destination IP addresses are same, which is misconfiguration:

$$\forall{:}\,\text{I,proto,src\_IP,src\_port,dst\_IP,dst\_port,action}{:}\text{fr}(\text{I,proto,src\_IP,src\_port,dst\_IP, dst\_port, action}) \land (\text{src\_IP} = \text{dst\_IP}) \Longrightarrow \text{Irrelevant1} \tag{17}$$

$$\forall{:}\,\text{I,proto,src\_IP,src\_port,dst\_IP,dst\_port,action}{:}\text{fr}(\text{I, proto, src\_IP, src\_port, dst\_IP, dst\_port, action}) \land ((\text{dst\_IP} \lor \text{dst\_port}) \notin \text{network}) \Longrightarrow \text{Irrelevant2} \tag{18}$$

Equation (18) denotes that the *irrelevant* case could happened if destination IP address or the destination port number is currently not existing in the network. Figures 4, and 5 shows the knowledge graphs depicting *irrelevant* cases in equation (17) and equation (18) respectively.

## 5 DISCUSSION AND CONCLUSION

In this paper, the knowledge graph-based approach has been introduced for providing deep analysis of networking firewall anomalies which are: *shadowing*, *generalization*, *correlation*, *irrelevant*, and *correlation*. Knowledge graph provides visualization that can assist networking administrators in deploying their firewall rules safely with the capability to detect anomalies in the rule set. In our work, knowledge graph is formulated as predicate calculus rules to provide solid mathematical representation for the existing anomalies, which in turn bridges the gap in developing automation tools and effective solutions. As a methodology, a general pattern for each anomaly has been formally defined, and then the cases satisfied each general pattern have been presented and analysed to ensure that our method covers all possible scenarios of well-know firewall rules' anomalies.

As a future work, we will implement a handy solution for our approach, including a solver for automatic knowledge extraction from the predicate calculus rules, which will be connected to KgBase, a knowledge graph builder free tool. On the other hand, the work in (Nguyen & Sakama, 2019) will be used to prove the generalization of our approach.

## REFERENCES

Boudriga, N. (2017). Security of Mobile Communications. *In Proceedings of the IEEE International Conference on Signal Processing and Communications*. IEEE.

Chao, C. (2018). A Feasible Anomaly Diagnosis Mechanism for Stateful Firewall Rules. *In Proceedings of the 27th International Conference on Computer Communication and Networks (ICCCN)*. IEEE.

Voronkov, A., Iwaya, L., Martucci, L., Lindskog, S. (2018). Systematic Literature Review on Usability of Firewall Configuration. *IN ACM Computing Surveys*, issue 6, ISSN 0360-0300. ACM.

Hu, H., Ahn, G., Kulkarni, K. (2012). Detecting and Resolving Firewall Policy Anomalies. *In IEEE Transactions on Dependable and Secure Computing*, issue 3, pp. 318-331. IEEE.

Abbes, T., Bouhoula, A., Rusinowitch, M. (2016). Detection of Firewall Configuration Errors with Updatable Tree. In International Journal of Information Security, issue 15, pp. 301-317. SpringerLink.

As-Suhbani H., Khamitkar, S. (2017). Using Data Mining for Discovering Anomalies from Firewall Logs: a Comprehensive Review. *In International Research Journal of Engineering and Technology (IRJET)*, vo. 4, issue 11, pp. 419-423.

Ahmed, Z., and Askari, S. (2018). Firewall Rule Anomaly Detection: A Survey. *In International Journal of Computational Intelligence & IoT*, vo. 2, pp. 722-727. SSRN.

Karafili, E., Valenza, F., Chen, Y., Lupu, E. (2020). Towards a Framework for Automatic Firewalls Configuration via Argumentation Reasoning. *In Proceedings of the IEEE/IFIP Network Operations and Management Symposium (NOMS)*. IEEE.

Elfaki, A., Aljaedi, A., Duan, Y. (2019). Mapping ERD to knowledge graph. *In Proceedings of the IEEE World Congress on Services (SERVICES)*. IEEE.

Von, V., Cao, S., Di, X., Gong, Y., Ren, W., Zhang, X. (2020). Knowledge Extraction and Knowledge Graph Construction Based on Campus Security Logss. *In Proceedings of the 6th International Conference on Artificial Intelligence and Security*. SpringerLink.

Wang, Z., Zhu, H., Liu, P., Sun, L. (2021). Social Engineering in Cybersecurity: a Domain Ontology and Knowledge Graph Application Examples. *In Cybersecurity*, no. 31, pp. 1-21. https://doi.org/ 10.1186/ s42400-021-00094-6

Wang, Y., Sun, Z., Han, Y. (2021). Network Attack Path Prediction Based on Vulnerability Data and Knowledge Graph. *In International Journal of Innovative Computing, Information and Control*, vo. 17, no. 5. http://www.ijicic.org/ ijicic-170518.pdf

Nguyen, H., Sakama, C. (2019). A New Algorithm for Computing Least Generalization of a Set of Atoms. *In Proceedings of the International Conference on Inductive Logic Programming*. SpringerLink.

Kim, T., Kwon, T., Lee, J., Song, J. (2021). F/Wvis: Hierarchical Visual Approach for Effective Optimization of Firewall Policy. *In IEEE Access*, vo. 9, pp. 105989 – 106004. IEEE.