Firewall configuration: An application of multiagent metalevel argumentation

Andy Applebaum[1], Zimi Li[2], Karl Levitt[3], Simon Parsons[4], Jeff Rowe[3], and Elizabeth I. Sklar[4]

[1]The MITRE Corporation, McLean, VA 22102 USA
[2]Department of Computer Science, The Graduate Center, City University of New York, New York, NY 10016 USA
[3]Department of Computer Science, University of California, Davis, CA 95616 USA
[4]Department of Informatics, King's College London, Strand, London WC2R 2LS UK

November 7, 2016

Abstract

Firewalls are an important tool in the assurance of network security. Packet filtering firewalls are configured by providing a set of rules that identify how to handle individual data packets that arrive at the firewall. In large firewall configurations, conflicts may arise between these rules. Argumentation provides a way of handling conflicts such that their origin is illuminated, and hence can help a system administrator understand the effects of a given configuration. To show how argumentation might help in this domain we examine the use of a system of metalevel argumentation for firewall configuration, showing how it makes conflicts and their origins clear, and showing how different instantiations of a metalevel argumentation system provide alternative ways to resolve conflicts.

1 Introduction

Assuring network security is a major problem today. It is a problem that is considered both in academic computer science, which aims to come up with new techniques for securing networks, and in the practical world of information technology, where system administrators struggle to prevent unauthorised users from breaking into the networks that they manage. Firewalls, first introduced in 1987 [25], are one of the core components of a network security implementation. A *firewall* is a combination of hardware and software that isolates an organization's internal network from the Internet at large, allowing some packets to pass through and blocking others [27]. The decision about which packets to pass and which to block is made according to some *policy*, and the *configuration* of a firewall is the business of implementing this policy.

As we will discuss below, firewall policies are set by specifying a set of rules, and there are a number of well-recognised problems in doing this. These problems relate to rules conflicting, having domains that overlap, and include redundancy (where the effect of a policy differs from that intended because some rules can never have any effect). Such *anomalies* arise from the complexity of setting up firewall policies in complex environments such as large organisations, especially when different parts of the overall firewall policy are set by different individuals.

In this paper, we discuss how an argumentation-based framework can be used to analyse a firewall policy. In particular, we examine the use of the metalevel argumentation framework of [32], choosing this system as the basis of our investigation because we believe that the metalevel reasoning about the

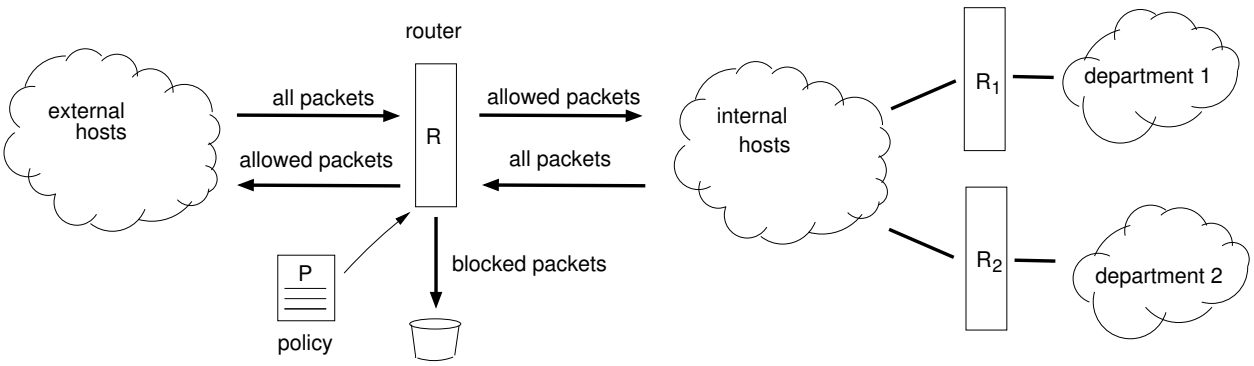

Figure 1: A packet-filtering firewall.

acceptability of arguments helps to make the reasons for conflict between policies especially clear, and makes it easy to understand how different strategies for resolving such conflicts work.

The remainder of this paper is organised as follows. Section 2 provides a brief introduction to firewalls and issues in their configuration. Section 3 introduces the specific metalevel argumentation system that we employ. Section 4 describes several ways in which this metalevel system can be used to represent firewall configurations and potentially resolve conflicts in the firewall rules. Section 5 discusses related work. Then Section 6 concludes.

2 Problems in firewall configuration

There are different types of firewall which function in different ways — packet-filtering firewalls, application/proxy firewalls, and network address translation. Packet-filtering firewalls operate at the network layer, not allowing packets to pass through the firewall unless they match the established policy rule set. Routers can provide a very common form of packet-filtering firewall. Packet-filtering usually makes decisions based on the following characteristics:

- Source and/or destination IP[1] addresses
- Source and/or destination port numbers
- Protocol types
- Other parameters within the IP header

A network administrator configures the firewall based on the policy, for example blocking and allowing packets based on what protocol they match and which IP address they have as their destination. A schematic of a packet-filtering firewall is given in Figure 1. Here the firewall is implemented by router *R*, which implements policy *P*. This inspects and filters incoming packets, aiming to remove any that attempt to attack internal hosts, and outgoing packets, both preventing users from accessing unauthorised resources outside the firewall and preventing attacks on external hosts from within the firewall.

Application firewalls, as indicated by the name, work at the application layer. These devices act as proxy machines for requested services. Requests are sent to a proxy machine, which then makes those requests to the Internet on behalf of the local client. A proxy machine acts as a buffer between "bad" remote users and the internal network client machines. Network address translation (NAT) also operates at the network layer, providing the capability to change the source and/or destination IP address. This is common when a private address space is used internally. The simplest type of NAT provides a one-to-one

[1]Internet Protocol

relationship between inside and outside IP addresses. In this type of NAT, only the IP header related to the IP address needs to be changed. The rest of the packet can be left unchanged. In the remainder of this paper, we concentrate on packet-filtering firewalls, but we believe that the techniques developed here could be applied to the other types we have listed above.

Table 1 lists some possible polices for a packet-filtering firewall and how they could be implemented. In particular, each row of Table 1 presents a rule for a single packet. A set of such rules makes up a firewall configuration. The first column in Table 1 numbers the rule for reference. The second column states the policy that the rule is intended to implement. The third, fourth, fifth and sixth columns break the policy implemented by a rule into the action that the firewall can perform (block or allow the packet) and the data that is available about each packet (the protocol it relates to, the source IP address, and the port the packet arrives at). The final column shows how a given rule would be implemented in the Linux `iptables` utility.

If a firewall were to be implemented with this set of rules, when a packet comes in it would be checked against rule 1. If rule 1 applied to that packet, then the action specified by rule 1 would be taken. Otherwise, the packet would be compared to rule 2. This process is repeated until one of the rules correctly specifies the packet; the first rule that does is the one that's applied, ignoring all rules after it. If no rule matches an incoming packet, some default rule (which might, for instance, be to let the packet pass since there is no specific rule to block it), would be applied. Imagine a packet arriving that uses TCP, coming from 55.55.55.55 on port 80. Rule 1 specifies blocking it, while rule 2 says to allow it. Since rule 1 is positioned before rule 2, the ultimate action is to block, but it is clear that there is some kind of a conflict occuring. As another example, a packet using UDP from 55.55.55.55 on port 53 would be blocked; while three of the rules say to block it, only the first one is actually "enforced".

The situations highlighted in both of these examples could be considered problematic, and both are what [2] calls an *anomaly* in a firewall policy. [2] defines four anomalies in terms of relations between rules:

Shadowing Rule a is said to shadow rule b if a has higher-priority than b, a and b specify different actions, and every packet that satisfies b also satisfies a.

In shadowing, the two rules are in conflict on *every* packet that the rules apply to.

Correlation Rule a and b are correlated if a and b specify different actions and some packets that satisfy a also satisfy b and vice versa.

In correlation, the rules conflict on *some* packets that the rules apply to.

Redundancy Redundancy occurs in two cases. In the first case, redundancy occurs if two rules a and b are such that all packets that satisfy a satisfy b, a and b specify the same action, and b is higher priority than a.

In the second case, redundancy occurs if all packets that satisfy a also satisfy b, a and b specify the same action, a is higher priority than b, and a is not involved in any correlation anomalies.

In both cases of redundancy, the lower priority rule will never be applied.

Generalization Rule a is said to generalize rule b if b has higher priority than a, a and b specify different actions, and every packet that satisfies b also satisfies a.

In generalization there is shadowing but the conflict is resolved by the priority.

In the example in Table 1, rule 4 shadows rule 5, since every instance of rule 5 is also an instance of rule 4, and they specify different actions. Rule 1 and rule 2 are correlated, since a TCP packet from 55.55.55.55 on port 80 is an instance of both rule 1 and rule 2, and they specify different actions. Rule 4 generalizes rule 2, since every instance of rule 2 is also an instance of rule 4, and they specify different actions. Rule 3 is redundant because every instance of rule 3 is also an instance of rule 4, and they specify the same action.

Now, in a small set of firewall rules such as these, it is easy enough to detect and fix anomalies. However, firewalls, especially in large organizations with many machines on a network, can include many hundreds of rules. In such a case, detection and correction of anomalies is much harder. The problem is even more complex when firewalls are composed of different components, each requiring some part of a policy that

Table 1: Examples of firewall polices and corresponding filtering rules. Each row lists an example of what might be required of a firewall, the corresponding firewall settings, and the implementation of these settings using the Linux `iptable` utility. The $*$ symbol indicates a wildcard, indicating any value is acceptable.

Rule	Policy	Firewall Setting				Linux implementation
		Action	Protocol	Source IP	Source port	
1	Block a malicious sender	block	$*$	55.55.55.55	$*$	iptables -A INPUT -p 0 -s 55.55.55.55 -j DROP
2	Allow Web services	allow	TCP	$*$	80	iptables -A INPUT -p tcp -dport 80 -j ACCEPT
3	Block DNS services	block	UDP	$*$	53	iptables -A INPUT -p udp -dport 53 -j DROP
4	Block all	block	$*$	$*$	$*$	iptables -A INPUT -p 0 -j DROP
5	Allow FTP services	allow	TCP	$*$	21	iptables -A INPUT -p tcp -dport 21 -j ACCEPT

is applied to a whole organisation, and as we shall discuss below, one can easily imagine scenarios in which decisions about whether to accept or reject specific packets requires complex reasons that need to combine information from a group of autonomous individuals. See [13] for a similar scenario in the domain of B2B applications. It is the need to deal with these complicated cases that is the reason we are using argumentation. In the remainder of the paper, we will describe how this can be done. In particular, we will use the *metalevel* approach presented in [32] since it provides a general approach to handling anomalies.

3 Metalevel argumentation

The metalevel argumentation framework of [32] is constructed on top of the standard Dung framework [20]. The idea is to make the conditions under which arguments are classified — for example as justified, rejected and defeated — and the definition of extensions — such as grounded, preferred and stable — expressible in a logical language. The advantage of doing this is that it becomes possible not just to have arguments about objects in some domain, but arguments (meta-arguments) about the status of those arguments. In this section, we introduce enough of this material to apply it to our firewall scenario. The description is an abbreviation of the presentation [32] with some minor modifications and additions (though naturally any faults in the interpretation of the original are ours).

3.1 Argumentation

The formal structure, taken from [32] is as follows. As [32] points out, the formalization is based not on Dung's classic presentation, but on the more recent labelling approach [15, 16, 45, 48] (nicely summarised in [10]). The basic notion is that of a *Dung argumentation framework*, a tuple $\langle \mathcal{A}, \mathcal{R} \rangle$ where $\mathcal{A}$ is a set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is a binary relation on $\mathcal{A}$ that identifies which arguments attack which other arguments.

In Dung's approach, as in [32], arguments are taken to be completely abstract entities with no internal structure, and the attack relation $\mathcal{R}$ is given. However, as a number of authors have pointed out — for example [5, 26, 39] — it is possible to construct *structured* arguments from some logic, and use the relationship between arguments to determine what attacks what. For example, given some logical language $\mathcal{L}$ and an inference relation $\vdash$, we might follow [5] by defining an argument as a pair (S, p) where S is a minimal set of formulae in $\mathcal{L}$ such that $S \vdash p$. p is the *conclusion* of such an argument and S is its *support*. In such a formulation, it is typical to say that one argument $S \vdash p$ attacks another $S' \vdash p'$ if either $p \equiv \neg p'$ or $\neg p \in S'$. That is (S, p) attacks another (S', p') if either the conclusions of the two arguments disagree, or the conclusion of the attacking argument disagrees with a member of the support of the attacked argument[2]. $\mathcal{A}$ is then the set of all arguments that can be constructed from some set of data Δ, and $\mathcal{R}$ is the set of all attacks between these arguments.

Given a set of arguments and attacks between them, the core of Dung's idea is that not all arguments are equal. It is possible to identify some arguments that we should consider *acceptable* — their conclusions are valid given what we know — and some that are not. The labelling approach gives us a simple way to determine whether an argument is acceptable or not. The approach can be described in terms of a *labelling function* L which maps from arguments to a set of labels $\{\texttt{IN}, \texttt{OUT}, \texttt{UNDEC}\}$. We then write $in(L)$ to indicate all arguments that are labelled `IN` by L, $out(L)$ to indicate all arguments that are labelled `OUT`, and $undec(L)$ to indicate all arguments that are labelled `UNDEC`.

Defined in this way, there is no relationship between a labelling and the attack relation over a set of arguments. The two are combined through the idea of legality. For a labelling L, an argumentation framework $\langle \mathcal{A}, \mathcal{R} \rangle$, and an argument $x \in \mathcal{A}$:

1. x is legally `IN` iff x is labelled `IN` and every $y \in \mathcal{A}$ that attacks x is labelled `OUT`.
2. x is legally `OUT` iff x is labelled `OUT` and there is at least one $y \in \mathcal{A}$ that attacks x and is labelled `OUT`.

[2] We discuss a specific $\mathcal{L}$ and $\vdash$ later in the paper.

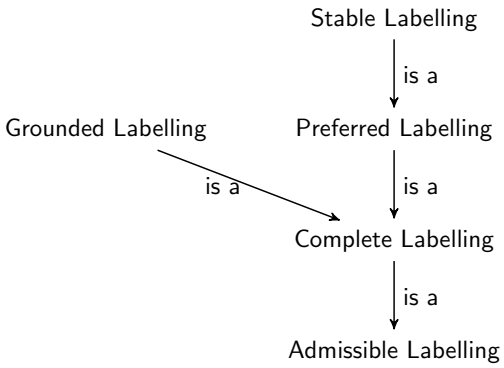

Figure 2: The relationship between labellings

3. x is legally UNDEC iff there is no $y \in \mathcal{A}$ that attacks x such that y is labelled IN, and there is at least one $y \in \mathcal{A}$ that attacks x such that y is labelled UNDEC.

Note that the UNDEC state occurs when x cannot be labelled IN (because it has at least one attacker that is not OUT), and cannot be labelled OUT (because it has no IN attacker). If an argument is not legally labelled, it is said to be illegally labelled. More precisely, an argument is illegally labelled l, where $l \in \{\texttt{IN}, \texttt{OUT}, \texttt{UNDEC}\}$ if it is not legally labelled l.

With the notion of legality tying labellings to attack relations, it is possible to recover Dung's idea that *extensions*, sets of arguments that are somehow coherent, can be identified within an argumentation framework. We do this through the notions of admissibility and completeness. An *admissible* labelling has no arguments that are illegally IN, and no arguments that are illegally OUT. A *complete* labelling is an admissible labelling that, in addition, has no arguments that are illegally UNDEC. Then, given a complete labelling L, we have that:

1. L is a *grounded* labelling iff there is no complete labelling with a smaller set of IN arguments.
2. L is a *preferred* labelling iff there is no complete labelling with a larger set of IN arguments.
3. L is a *stable* labelling if it contains no UNDEC arguments.

If L is a grounded labelling, then every argument x which is labelled IN in L is in Dung's grounded extension, if L is a preferred labelling then every x which is labelled IN in L is in the preferred extension, and if L is a stable labelling then every x which is labelled IN in L is in the stable extension. The relationship between labellings (and hence extensions) is shown in Figure 2, and note that other notions of acceptability, such as semi-stable [18], can also be captured using the same labelling approach.

Based on extension membership, we can then define the *status* of arguments. If x is in at least one extension, then x is *credulously* justified; if x is in all extensions, then it is *sceptically* justified; and if x is in no extensions, then it is *rejected*.

3.2 Metalevel argumentation

In [32], a *metalevel argumentation framework* is defined[3] as a tuple:

$$\langle \mathcal{A}, \mathcal{R}, \mathcal{A}_M, \mathcal{R}_M, \mathcal{C}, \mathcal{L}_\mathcal{C}, \mathcal{D} \rangle$$

where $\mathcal{A}$ is a set of arguments and $\mathcal{R}$ is an attack relation on object level arguments as in the previous section, and $\mathcal{A}_M$ and $\mathcal{R}_M$ are sets of arguments and attacks at the metalevel. $\mathcal{C}$ is a set of claims about the

[3]This is a less general subset of the system presented in [32], but sufficient for our purposes.

arguments in $\mathcal{A}_M$, that is a mapping from $\mathcal{A}$ to statements, $\mathcal{L}_\mathcal{C}$ is the language in which the claims are made, and $\mathcal{D}$ is a set of constraints on the attack relation $\mathcal{A}_M$ that are determined by the claims. As an example, [32] gives a metalevel argumentation framework that captures Dung's original argumentation system. In this system, $\mathcal{L}_\mathcal{C}$ includes a set of constants and a set of predicates. The set of constants C includes $\ulcorner x \urcorner$ for every $x \in \mathcal{A}$ (it is common practice to quote object level symbols in this way to make them constants at the metalevel). The set of predicates is:

$$\{\mathit{justified}, \mathit{defeat}, \mathit{rejected}\}$$

and $\mathcal{L}_\mathcal{C}$ has a set of well-formed formulae W defined by the following rules:

1. If $\ulcorner x \urcorner \in C$, then $\ulcorner x \urcorner \in W$
2. If $\ulcorner x \urcorner, \ulcorner y \urcorner \in W$, then $(\ulcorner x \urcorner, \ulcorner y \urcorner) \in W_\mathcal{R}$, $W_\mathcal{R} \subset W$
3. If $\ulcorner x \urcorner \in W$ and $\ulcorner x \urcorner \notin W_\mathcal{R}$, then $\mathit{justified}(\ulcorner x \urcorner) \in W$
4. If $\ulcorner x \urcorner \in W$ and $\ulcorner x \urcorner \notin W_\mathcal{R}$, then $\mathit{rejected}(\ulcorner x \urcorner) \in W$
5. If $\ulcorner x \urcorner, \ulcorner y \urcorner \in W$ and $\ulcorner x \urcorner, \ulcorner y \urcorner \notin W_\mathcal{R}$, then $\mathit{defeat}(\ulcorner x \urcorner, \ulcorner y \urcorner) \in W$

In other words, the language $\mathcal{L}_\mathcal{C}$ allows us to talk about any of the constants (which will represent arguments in $\mathcal{A}$), attacks between the arguments, whether arguments are justified or rejected, and whether one argument *defeats* another. The notion of defeat is necessary because exactly the kind of thing we want to capture is when there is an attack between two arguments, but there is something at the metalevel which overrides the attack. The labelling of arguments thus depends on defeats not on attacks.

We next need to define $\mathcal{A}_M$, which is the union of $\mathcal{A}_{M1}$, $\mathcal{A}_{M2}$ and $\mathcal{A}_{M3}$ where:

$$\begin{array}{l}\alpha \in \mathcal{A}_{M1}, \mathcal{C}(\alpha) = \mathit{justified}(\ulcorner x \urcorner) \text{ iff } x \in \mathcal{A} \\ \beta \in \mathcal{A}_{M2}, \mathcal{C}(\beta) = \mathit{rejected}(\ulcorner x \urcorner) \text{ iff } x \in \mathcal{A} \\ \gamma \in \mathcal{A}_{M3}, \mathcal{C}(\gamma) = \mathit{defeat}(\ulcorner x \urcorner, \ulcorner y \urcorner) \text{ iff } (x, y) \in \mathcal{R}\end{array}$$

so that arguments in $\mathcal{A}_M$ are statements about arguments in $\mathcal{A}$ being justified, rejected and defeating one another. Then the set of constraints on claims, $\mathcal{D}$ contains:

D1 if $\mathcal{C}(\alpha) = \mathit{defeat}(X, Y)$ and $\mathcal{C}(\beta) = \mathit{justified}(Y)$ then $(\alpha, \beta) \in \mathcal{R}_M$.

D2 if $\mathcal{C}(\alpha) = \mathit{defeat}(X, Y)$ and $\mathcal{C}(\beta) = \mathit{rejected}(X)$ then $(\beta, \alpha) \in \mathcal{R}_M$.

D3 if $\mathcal{C}(\alpha) = \mathit{justified}(X)$ and $\mathcal{C}(\beta) = \mathit{rejected}(X)$ then $(\alpha, \beta) \in \mathcal{R}_M$.

which together define the contents of $\mathcal{R}_M$. For example, the first of these says that a claim that X defeats Y is an attack on the claim that Y is justified. As [32] shows, computing the justified arguments in $\mathcal{A}_M$ will identify the justified arguments in $\mathcal{A}$ consistently across the different definitions of extensions.

As presented so far, and as described in [32], this metalevel argumentation system, just like Dung's system [20], has an asbtract notion of an argument. The members of $\mathcal{A}_M$ have no internal structure. However, one can (and we will below) construct the members of $\mathcal{A}_M$ from a set of statements Δ_M in some language $\mathcal{L}_M$ using an inference mechanism $\vdash_M$. When this is done, $\mathcal{L}_M$, like $\mathcal{L}_\mathcal{C}$, will contain constants $\ulcorner x \urcorner$ for every $x \in \mathcal{A}$ since $\mathcal{L}_M$ will be statements about these arguments. For example, a sentence in $\mathcal{L}_M$ might describe how one argument is preferred to another, and an argument in $\mathcal{A}_M$ that is constructed from such statements might describe how an attack from $\mathcal{R}$ is not a defeat because of this preference. The attacks between these arguments then populate $\mathcal{R}_M$.

4 Arguing about firewall policies

Having introduced some of the issues in firewall configuration, and the metalevel argumentation approach of [32], in this section we discuss how the latter can be used to model some aspects of firewall configuration in order to illuminate anomalies and potentially provide a means to support system administrators in solving them.

4.1 Scenario

We consider a simple scenario in which an organization operates a hierarchical network of routers (see Figure 1 again). The root node, R, is the master router which ultimately implements the firewall policy for the organization. The child nodes, R_1 and R_2, are gateways to different departments within the organization which have different requirements. R_1 and R_2 are stakeholders in the implemented policy and send their preferred policies to R. R then combines these policies to create the overall policy P for the organization. If the policies put forward by R_1 and R_2 conflict, R must resolve the conflict in order to create this overall policy.

Currently this merging of policies would be done by hand. In the simplest case, this is done by just concatenating the firewall rules. It is not hard, though, to imagine the process being automated with the routers being under the control of software agents. A_R is the agent controlling R and A_{R_1} and A_{R_2} are the agents controlling R_1 and R_2 respectively. Indeed, a process that implements a software-based packet filtering firewall would meet the description of a basic reactive agent [50], receiving a sequence of percepts in the form of data about incoming packets, and making a sequence of "accept"/"block" decisions. The policies for R_1 and R_2 are set by system administrators in the relevant departments, A_{R_1} and A_{R_2}, and these advocate for their policies with A_R. This agent (A_R) merges the policies and the combined policy is set by the system administrator with overall responsibility for the whole organization, based on information provided by the agent controlling R.

The model of structured argumentation that we use in the following examples is that of [28], which is a defeasible subset of ASPIC+ [33]. In terms of the element introduced above, $\mathcal{L}$ contains a set of propositions p_i, their negations $\neg p_i$, and a set of defeasible rules $p_1, \ldots, p_n \Rightarrow p_{n+1}$. Inference in this system is achieved by the application of:

$$\frac{p_1, \ldots, p_n \qquad p_1, \ldots, p_n \Rightarrow p_{n+1}}{p_{n+1}}$$

If p follows from a set of formulae S using this inference rule alone, we denote this by $S \vdash p$, and, if S is minimal, (S, p) is an argument. Note that all we are doing here is to make it possible to use the metalevel argumentation model from [32] with structured arguments. Our intention is not to propose a new model of argumentation; we are just using an existing model of metalevel argumentation with a subset of the well-known ASPIC+ system.

4.2 A simple metalevel model of a firewall

Now suppose that R_1 has a policy to deny all DNS traffic in order to enhance system security, while R_2 has a policy to allow HTTP traffic in order to support web services. We can model R_1's policy as:

$$\begin{array}{rcl} \mathit{secure\_system} & & \\ \mathit{secure\_system} & \Rightarrow & \neg \mathit{allow\_DNS} \\ \neg \mathit{allow\_DNS} & \Rightarrow & \neg \mathit{allow\_UDP} \end{array}$$

In addition to the justification of R_1's policy, this captures the fact that one approach to disallowing DNS traffic is to block all UDP traffic since DNS runs over UDP. (Such an approach would obviously block other protocols and applications that use UDP, and so this might be considered rather heavy-handed.) From this set of policy information, it is possible for A_{R_1} to construct the argument:

$$(\{\mathit{secure\_system}, \mathit{secure\_system} \Rightarrow \neg \mathit{allow\_DNS}, \neg \mathit{allow\_DNS} \Rightarrow \neg \mathit{allow\_UDP}\}, \neg \mathit{allow\_UDP}) \qquad (1)$$

which has the conclusion to block UDP traffic. We will call this argument n since it concerns name resolution. Similarly, R_2's policy can be modelled as:

$$\begin{array}{rcl} \mathit{allow\_WS} & & \\ \mathit{allow\_WS} & \Rightarrow & \mathit{allow\_TCP} \end{array}$$

giving A_{R_2} the argument:

$$(\{allow\_WS, allow\_WS \Rightarrow allow\_TCP\}, allow\_TCP)$$

with the conclusion that TCP traffic should be allowed. We will call this argument t.

We can imagine that A_R engages both A_{R_1} and A_{R_2} in an inquiry dialogue [49] to discover their requirements (for example following the protocol in [38]), a process that results in both n and t (including all the information on which they are based) being passed to A_R. In addition, A_R knows that:

$$allow\_WS \Rightarrow allow\_DNS$$
$$allow\_DNS \Rightarrow allow\_UDP$$

since web services require name resolution and hence require UDP. In addition to n and t, A_R can thus construct an argument w (an argument about the requirements of web services):

$$(\{allow\_WS, allow\_WS \Rightarrow allow\_DNS, allow\_DNS \Rightarrow allow\_UDP\}, allow\_UDP)$$

Clearly w and n attack one another[4]. In the formulation given above, we then have:

$$\begin{aligned} \mathcal{A} &= \{w, n, t\}, \\ \mathcal{R} &= \{(w, n), (n, w)\} \end{aligned}$$

which has a single grounded extension $\{t\}$ which does not specify what to do about UDP traffic. Before we consider how we can use argumentation to represent different solutions to this scenario, let's work through the full metalevel formulation. The previous section gave a metalevel formulation of a standard Dung framework. In this, the set of metalevel arguments includes statements about the justification and defeat of every argument in $\mathcal{A}$, and statements about defeat for every attack in $\mathcal{R}$. Thus we have:

$$\begin{aligned} \mathcal{A}_M = \{&defeat(\ulcorner w \urcorner, \ulcorner n \urcorner), defeat(\ulcorner n \urcorner, \ulcorner w \urcorner), \\ &justified(\ulcorner w \urcorner), justified(\ulcorner n \urcorner), justified(\ulcorner t \urcorner), \\ &rejected(\ulcorner w \urcorner), rejected(\ulcorner n \urcorner), rejected(\ulcorner t \urcorner)\} \end{aligned}$$

The constraints on claims then result in the following set of attacks:

$$\begin{aligned} \mathcal{R}_M = \{&(defeat(\ulcorner n \urcorner, \ulcorner w \urcorner), justified(\ulcorner w \urcorner)), (justified(\ulcorner w \urcorner), rejected(\ulcorner w \urcorner)), \\ &(rejected(\ulcorner w \urcorner), defeat(\ulcorner w \urcorner, \ulcorner n \urcorner)), (defeat(\ulcorner w \urcorner, \ulcorner n \urcorner), justified(\ulcorner n \urcorner)), \\ &(justified(\ulcorner n \urcorner), rejected(\ulcorner n \urcorner)), (rejected(\ulcorner n \urcorner), defeat(\ulcorner n \urcorner, \ulcorner w \urcorner)), \\ &(justified(\ulcorner t \urcorner), rejected(\ulcorner t \urcorner))\} \end{aligned}$$

The first six of these attack relations form a cycle in the argument graph[5] shown in Figure 3, which has no consistent labelling. The last pair of arguments listed in $\mathcal{R}_M$ can be labelled consistently so that t is justified. The result, then, is the single stable extension:

$$E = \{justified(\ulcorner t \urcorner)\}$$

and the corresponding single stable extension of the object level (as we already identified just considering the object level system) is $\{t\}$.

An understandable question to ask at this point is what the metalevel framework brings over a formulation of the problem using a standard Dung framework, or a structured argumentation system like ASPIC+, which builds a Dung framework from a knowledge base. In terms of what can be represented, and what

[4]The form of attack here is a *rebut*, an attack between the conclusions of arguments. While rebuts can be problematic in some argumentation systems [17], they do not cause problems when arguments are, as here, chains of defeasible rules [28].

[5]This is the weather example from [32, page 19] without the preference argument.

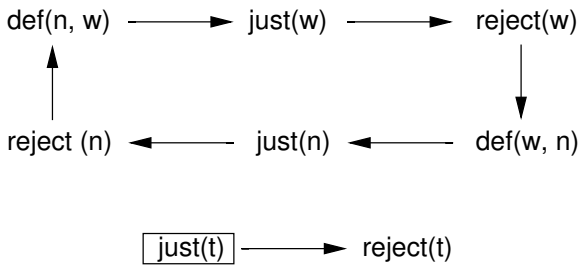

Figure 3: The metalevel argument graph for the basic firewall example. Each argument corresponds to an argument or an attack at the object level. A box is drawn around arguments that can legally be labelled IN.

can be formally concluded, the metalevel approach brings nothing. As [32] clearly shows, the metalevel approach it suggests, the one we have used here, captures precisely what a Dung framework can capture, and draws exactly the same conclusions. We are, of course, using a structured argumentation framework to generate the arguments, but these could be generated by ASPIC+. Indeed, given that we are using a structured argumentation system which is a subset of ASPIC+, one might argue that we *are* using ASPIC+. However, to focus on what can be represented and what conclusions are drawn is to miss the point. The point of using the metalevel approach lies in what is computed along the way.

In a standard Dung framework, the output is a set of labellings, telling us which arguments are justified (IN) and which are not. Frequently, in implementations, this is presented to the user in the form of an argument graph, colour-coded to indicate which arguments are IN, OUT, and UNDEC. Such an output is clear about *what* the conclusions are, but not *why*. Even in systems such as ArgTrust [37], which present views of structured arguments, a user who wants to understand *why* has to reconstruct the inference process for themselves. The metalevel approach, both in the way it is presented in [32] and the way we are using it here, is one way to address the explanation of *why*. In the same way that a Dung argument graph makes clear the relationship between arguments, the metalevel argument graph exposes the relationship between the status of arguments and thus the reasoning process that leads to the labelling attached to each argument. We believe that this means that the metalevel approach can be of service in situations, like network security, where the users of argumentation technology are unlikely to also be experts in argumentation. In this particular case, we see that the advantage of A_R using a metalevel framework, rather than a standard Dung framework (which would enable it to reach the same conclusion), is that A_R can use the metalevel framework to explain the resulting policy to the adminstrator with overall responsibility for the organization. For example, the argument graph that results in this case is given in Figure 3. This makes it clear that the fact there is no justified argument for w or n is the symmetry between them. Each attacks the other, and there is no reason to privilege one attack over the other.

4.3 A metalevel model using preferences

While the resolution of the conflicting policies achieved above is correct from the perspective of argumentation theory, it is not very satisfying from an application point of view — the resulting policy is unhelpful since it provides no decision on UDP traffic. A natural way to improve the situation is to express some kind of preference between web services and security (in our case) to resolve the conflict between w and n one way or the other. This is not a new idea, having been introduced at the object level in argumentation systems such as [4, 40].

As discussed in [32], preferences can easily be introduced into a metalevel argumentation framework. If we follow the approach described in [32], we consider that stating a preference w $>_P$ n — that the argument in favor of allowing UDP to support web services is strictly preferred to the argument in favor of blocking UDP to enhance security — is equivalent to stating a metalevel argument that n does not defeat w.

The formal description of the metalevel system is that of the previous section, with the same set of

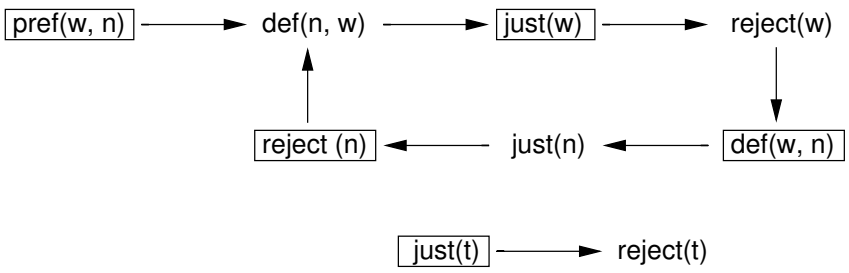

Figure 4: The metalevel argument graph for the firewall example with preferences. Each argument corresponds to an argument or an attack at the object level. A box is drawn around arguments that can legally be labelled IN.

arguments and attacks at the object level:

$$\begin{aligned} \mathcal{A} &= \{w, n, t\}, \\ \mathcal{R} &= \{(w, n), (n, w)\} \end{aligned}$$

but with an additional argument — the argument that w is preferred to n at the metalevel (we assume that the claim language $\mathcal{L}_C$ contains an additional predicate $preferred(\cdot, \cdot)$ to express this preference):

$$\begin{aligned} \mathcal{A}_M = \{&defeat(\ulcorner w \urcorner, \ulcorner n \urcorner), defeat(\ulcorner n \urcorner, \ulcorner w \urcorner), \\ &justified(\ulcorner w \urcorner), justified(\ulcorner n \urcorner), justified(\ulcorner t \urcorner), \\ &rejected(\ulcorner w \urcorner), rejected(\ulcorner n \urcorner), rejected(\ulcorner t \urcorner), \\ &preferred(\ulcorner w \urcorner, \ulcorner n \urcorner)\}, \end{aligned}$$

The set of metalevel attacks also has an additional member, the attack of $preferred(\ulcorner w \urcorner, \ulcorner n \urcorner)$ on $defeat(\ulcorner n \urcorner, \ulcorner w \urcorner)$:

$$\begin{aligned} \mathcal{R}_M = \{&(defeat(\ulcorner n \urcorner, \ulcorner w \urcorner), justified(\ulcorner w \urcorner)), (justified(\ulcorner w \urcorner), rejected(\ulcorner w \urcorner)), \\ &(rejected(\ulcorner w \urcorner), defeat(\ulcorner w \urcorner, \ulcorner n \urcorner)), (defeat(\ulcorner w \urcorner, \ulcorner n \urcorner), justified(\ulcorner n \urcorner)), \\ &(justified(\ulcorner n \urcorner), rejected(\ulcorner n \urcorner)), (rejected(\ulcorner n \urcorner), (defeat(\ulcorner n \urcorner, \ulcorner w \urcorner)), \\ &(preferred(\ulcorner w \urcorner, \ulcorner n \urcorner), defeat(\ulcorner n \urcorner, \ulcorner w \urcorner)), \\ &(justified(\ulcorner t \urcorner), rejected(\ulcorner t \urcorner))\} \end{aligned}$$

As Figure 4 shows, this additional attack now breaks the cycle[6] and we have a single stable extension at the metalevel:

$$E = \{justified(\ulcorner w \urcorner), defeat(\ulcorner w \urcorner, \ulcorner n \urcorner), rejected(\ulcorner n \urcorner), (preferred(\ulcorner w \urcorner, \ulcorner n \urcorner), justified(\ulcorner t \urcorner)\}$$

with the corresponding object level extension $\{w, t\}$, with a policy that allows TCP and UDP.

Again, we believe that this metalevel structure provides a means to explain the outcome to the administrator. Comparing Figure 4 and Figure 3, it is clear that the preference for w over n "fixes" the cycle of arguments so that $defeat(\ulcorner n \urcorner, \ulcorner w \urcorner)$ does not hold, resulting in w being justified.

4.4 A metalevel model using values

Metalevel argumentation can capture more than just the application of preferences. In this section, we briefly consider how we can use the same metalevel argumentation framework to apply ideas from value-based argumentation [12] to capture a situation in which different parties have different views about which policy to adopt.

[6]This section of the argument graph is now exactly the weather example from [32, page 19].

In a *Value-based Argumentation Framework* (VAF), we are concerned with values promoted by each attack and their relative strengths. This is necessarily subjective and audience dependent [12, 32] and provides a way for R to reason about implementing policies specific to R_1 and R_2. To see how this might work, consider a variation on the example discussed above, where the network is that of a research university[7]. R_1 belongs to a research computing facility (IT) which uses BitTorrent (BT) to provide updates to the host machines. This is deemed to be mission-critical by IT. R_2 belongs to the Chancellor's office (CO), which has deemed BitTorrent to be a legal liability and thus seeks to deny any BT traffic.

Let c denote the policy "Deny BitTorrent" and a denote the policy "Allow BitTorrent". As in the examples in Section 4, we have a situation in which there are mutual attacks between the arguments. In a value-based framework, we can annotate the policies to introduce values associated with the arguments by different parties. For example, "Deny BitTorrent" may be associated with the "Preventing Piracy" (p) value, while "Allow BitTorrent" may be associated with the "Allowing Mission Critical Services" (m) value. R may use reasoning specific to audiences pertinent to R_1 and R_2. If R_1 serves research-related computing facilities, the relevant audience is the advocate for the corresponding policy, IT (r_1). And if R_2 is controlling access to student dormitories, then the relevant audience is the advocate for the corresponding policy, CO (r_2). Further, each audience indicates which values are more important: r_1 prefers p to m, while r_2 prefers m to p.

We can formulate this as a value-based argumentation framework in the language of metalevel argumentation as (following [32]):

$$(\mathcal{A} = \{c, a\}, \mathcal{R} = \{(c, a), (a, c)\}, \mathcal{V} = \{p, m\}, \{val(c) = p, val(a) = m\}, \mathcal{P} = \{r_1 = \{(m, p)\}, r_2 = \{(p, m)\}\})$$

which leads to two audience-specific VAFs for r_1 and r_2. Our set of claims then includes the values, $val(c) = p$ and $val(a) = m$, and the preferences over values, $preferred_{r_1}(m, p)$ and $preferred_{r_2}(p, m)$. We can then formulate the audience-specific metalevel argumentation framework for r_1 as follows.

$$\begin{aligned}
\mathcal{A} &= \{\, c, a\}, \\
\mathcal{R} &= \{\, (c, a), (c, a)\}, \\
\mathcal{A}_M &= \{\, defeat(\ulcorner c \urcorner, \ulcorner a \urcorner), defeat(\ulcorner c \urcorner, \ulcorner a \urcorner), justified(\ulcorner c \urcorner), \\
&\quad\; justified(\ulcorner a \urcorner), rejected(\ulcorner c \urcorner), rejected(\ulcorner a \urcorner), \\
&\quad\; preferred_{r_1}(\ulcorner m \urcorner, \ulcorner p \urcorner) \\
\mathcal{R}_M &= \{\, (preferred_{r_1}(\ulcorner m \urcorner, \ulcorner p \urcorner), defeat(\ulcorner c \urcorner, \ulcorner a \urcorner)), \\
&\quad\; (defeat(\ulcorner c \urcorner, \ulcorner a \urcorner), justified(\ulcorner a \urcorner)), \\
&\quad\; (justified(\ulcorner a \urcorner), rejected(\ulcorner a \urcorner)), \\
&\quad\; (rejected(\ulcorner a \urcorner), defeat(\ulcorner a \urcorner, \ulcorner c \urcorner)), \\
&\quad\; (defeat(\ulcorner a \urcorner, \ulcorner c \urcorner), justified(\ulcorner c \urcorner)), \\
&\quad\; (justified(\ulcorner c \urcorner), rejected(\ulcorner c \urcorner)), \\
&\quad\; (rejected(\ulcorner c \urcorner), (defeat(\ulcorner c \urcorner, \ulcorner a \urcorner))\}.
\end{aligned}$$

A corresponding metalevel argumentation framework can be formulated for r_2. This leads to two audience specific frameworks for r_1 and r_2. The preferred extensions, for each audience, are:

$$E_{r_1} = \{preferred_{r_1}(m, p), justified(a), defeat(a, c), rejected(c)\}$$

and

$$E_{r_2} = \{preferred_{r_2}(p, m), justified(c), defeat(c, a), rejected(a)\}.$$

With the audience-specific extensions, the system administrator may reason that the CO prefers values promoted by preventing piracy over allowing mission critical services, while the IT prefers values promoted by allowing mission critical services over preventing piracy.

[7]This is, of course, a fictional university and bears no ressemblance to any institution at which any of the authors work or may have worked.

4.5 Structured reasoning at the metalevel

As noted above, the idea of using preferences to resolve conflicting arguments is not new, and the same is true of values. The new aspect that the metalevel approach brings is the ability to clearly see what the the preferences and values are doing, that is *how* they resolve the conflict. Examining the metalevel arguments and attacks in the example of Section 4.3, it is clear that the preferences are behind the resolution of the conflict by defeating $defeat(n, w)$, and in turn preventing that argument from making w unjustified. Similarly, looking at the metalevel arguments and attacks in the example of Section 4.4, it is clear how the different preferred extensions relate to the two audiences R_1 and R_2. In applications such as ours, where the justification for using argumentation is to be able to explain to users the structure of the problem and how to reason about it, this ability to use the metalevel system to explain how arguments are resolved at the object level is a powerful feature.

However, the models that we have discussed so far do not fully exploit the power of the metalevel framework. The astute reader will have spotted that though we have described how structured arguments can be used at the object level to connect firewall rules to arguments — as in the argument labelled (1) for example — we have yet to explore the construction of arguments at the metalevel. Allowing reasoning at this level allows us to construct arbitrary arguments at the metalevel that can resolve conflicts at the object level.

Consider, as an example, this variation on the use of preferences, where A_R has the following information in its Δ_M:

$$\begin{aligned} prefer(promote\_WS, be\_secure) & \\ achieves(\ulcorner w \urcorner, promote\_WS) & \\ achieves(\ulcorner n \urcorner, be\_secure) & \\ prefer(X, Y), achieves(Z, X), achieves(W, Y) & \Rightarrow preferred(Z, W) \\ preferred(Z, W), (W, Z) \in \mathcal{R} & \Rightarrow (preferred(Z, W), defeat(W, Z)) \end{aligned}$$

where X, Y, Z and W are variables and *achieves* is a predicate that captures the relationship between an object level argument and a metalevel proposition. This is metalevel information about a preference for arguments about firewall policies that promote web services over security, about the specific policies supported by the object level arguments w and n, about how to combine preferences and information about arguments in general (if you prefer one policy to another, then you prefer the argument that supports it), and about how preference relates to defeat.

From this information, it is clear that A_R can construct an argument for

$$(preferred(\ulcorner w \urcorner, \ulcorner n \urcorner), defeat(\ulcorner n \urcorner, \ulcorner w \urcorner))$$

which of course is the crucial piece in the use of preferences to resolve the circle of attacks in the metalevel representation of the conflict between w and n (see Figure 4 again). Abstracting from this, we have a general mechanism by which we can program A_R to figure out how to resolve conflicts in object level arguments — we provide it with knowledge in Δ_M from which it can construct metalevel arguments about which attacks are themselves defeated. The same mechanism could be used in our values example, to side with a particular audience and therefore rule out one of the preferred extensions.

One kind of reasoning that one might perform at this level that is of particular interest to us is reasoning about trust. A_R might wish to resolve the conflict between w and n based on what it knows about the trustworthiness of A_{R_1} and A_{R_2}. In [36], we described an argumentation system that could be used to infer the degree of trust between agents, and how this derived information could be combined with beliefs from those agents. Such reasoning could be employed in the metalevel argumentation framework to identify attacks on $defeat(\ulcorner n \urcorner, \ulcorner w \urcorner)$ or $defeat(\ulcorner w \urcorner, \ulcorner n \urcorner)$. To give a very simple example:

$$\begin{aligned} trust(self, A_{R_1}) \\ \neg trust(self, A_{R_2}) \end{aligned}$$

$$\begin{aligned} trust(self, X), \neg trust(self, Y) &\Rightarrow more\_trustworthy(X, Y) \\ source(\ulcorner n \urcorner, A_{R_1}) & \\ source(\ulcorner w \urcorner, A_{R_2}) & \\ source(X, Y), source(W, Z), more\_trustworthy(Y, Z) &\Rightarrow preferred(X, W) \end{aligned}$$

from which A_R could, using elements from the previous example, construct an argument for:

$$(preferred(\ulcorner n \urcorner, \ulcorner w \urcorner), defeat(\ulcorner w \urcorner, \ulcorner n \urcorner))$$

which would provide an alternative way to resolve the object level conflict between w and n — the fact that n is provided by a more trustworthy source than w is an argument against w defeating n.

4.6 Discussion

The examples in the previous section have demonstrated how metalevel argumentation can be used to represent firewall policies in such a way that a conflict in the rules is exposed at the metalevel. They have further shown how metalevel argumentation can capture the way that existing approaches, such as preference-based and value-based reasoning, can be used to resolve such conflicts. The key advantage that we see to using metalevel argumentation is that conflicts and their resolution are exposed at the metalevel, rather than being buried in the computation of justified arguments. The metalevel approach does not provide any reasoning that is not provided by existing approaches, but by handling arguments at the metalevel as objects that can be themselves the subject of arguments, reasoning about those arguments, and the conflicts between them, becomes explicit. In firewall configuration, the final decision about how to set the firewall up is going to be taken by a human system administrator — someone who is probably not an expert in argumentation — and our hypothesis is that any additional clarity will make their job much easier. We already have some evidence that providing arguments for and against decision options helps users to become more confident in their decisions [43], albeit in the domain of intelligence analysis rather than firewall configuration. More recent work [1], provides some support for the use of argumentation in maintaining secure services. In particular, this latter work showed that using argumentation to support the definition of firewall policies provided more complete and correct policies at the cost of some additional effort. However, we do not yet have any empirical evidence concerning the use of metalevel argumentation in handling firewall policies.

We also note that we started by discussing four kinds of anomaly in firewall rules — shadowing, correlation, redundancy and generalization. Our examples are all instances of shadowing. Before we can claim that metalevel argumentation can, in general, help with detecting and resolving firewall anomalies we must first also demonstrate that the approach can deal with the other forms of anomaly.

Finally, we point out a further extension of the approach we have been detailing in this paper. In [35], we identified a number of argumentation schemes for reasoning about trust. In particular, we identified two broad classes of scheme: schemes for deriving trust and schemes for propagating trust. The schemes for deriving trust are exemplified by the scheme for direct experience: if A has previous successful interactions with B, then that is the basis of an argument that A should trust B in future interations. The schemes for propagating trust are exemplified by transitivity: if A trusts B and B trusts C, then that forms the basis for an argument that A should trust C. Of course, both these schemes can be flawed, and [35] captures these flaws as critical questions. These identify situations in which the argument generated by the scheme is not sound. For example, direct experience can be questioned if B is not known to be the same individual with whom A has interacted in the past. Similarly, transitivity can be questioned if the context in which A trusts B is not the same as the context in which B trusts C.

These trust schemes and the relevant critical questions could be implemented in the same metalevel argumentation framework that we have been discussing. At the object level we encode the schemes as defeasible rules that allow the relevant arguments to be constructed, for example capturing transitivity with the rule:

$$trust(X, Y), trust(Y, Z) \Rightarrow trust(X, Z)$$

Then we capture the critical questions at the metalevel. For example, the critical question about context can be captured in a very similar way to that in which we captured trustworthiness in the previous section. In this case, we need metalevel knowledge about the context in which trust exists between agents, and a rule to identify if transitivity was used in the construction of an argument at the object level, and to ensure that the object-level argument is defeated if contexts do not match:

$$
\begin{aligned}
& trust\_context(self, teacher, childcare) \\
& trust\_context(teacher, mechanic, car\_repair) \\
& \ulcorner trust(X,Y), trust(Y,Z) \Rightarrow trust(X,Z) \urcorner \in Support(\ulcorner a \urcorner), \\
& trust\_context(X,Y,C), trust\_context(Y,Z,D), C \neq D \quad \Rightarrow \quad defeat(CQ_{context}, \ulcorner a \urcorner)
\end{aligned}
$$

where the symbol $CQ_{context}$ identifies the critical question about context. This is close to the approach adopted by [34], though the latter paper captures critical questions though the use of undercutting [39] rules, which (in our example) would deny the applicability of the transitivity rule rather than (as we do) ensuring that the argument generated using the rule is defeated. Of course, undercutting attacks are implicitly a metalevel notion.

Now, one important difference between the example above, and that in Section 4.5, is that in the latter example, reasoning about trust was at the metalevel, and in our example here it is at the object level. There need be no conflict between the two approaches. As [51] points out, it is possible to have many levels of metalevel reasoning. We can easily construct a system in which arguments about policy happen at level 0, then level 1 contains rules, facts and arguments making inferences about trust, just as in Section 4.5. These level 1 arguments treat arguments at level 0 as object level arguments. Level 2 then contains fact rules and arguments about the applicability of arguments about trust (capturing critical questions about the schemes that generated level 1 arguments), and treating level 1 arguments as object level elements to be manipulated.

5 Related Work

Given the importance of security, the central role that firewalls play in ensuring security, and the complexity of configuring firewalls, there has been considerable work on approaches to support firewall configuration. [2] implemented a set of algorithms in "Firewall Policy Advisor", a user-friendly tool. These algorithms use a firewall policy tree to deal with centralized and distributed firewalls [3]. [52] introduced FIREMAN, a static analysis toolkit to check anomalies in individual firewalls as well as among distributed firewalls.

Similar work has been carried out based on the use of formal logic. [22] proposed a formal logic for reasoning about the meaning of the firewall rules. Logic may be used to prove properties of a rule set and to detect a number of anomalies within a rule set. [23, 24] used ordered binary decision diagrams (BDD's) to represent rule sets as a boolean expressions, again proving a means to analyze the rule sets. However, the system does not allow the definition of rules using a logic programming syntax. In contrast, [21] presented a tool based on constraint logic programming (CLP) for analyzing firewall rules. This tool was implemented using Eclipse CLP, which makes it easy to express and extend the knowledge base of the system. [11] used a related approach, modelling firewall policies using a spatio-temporal logic that makes it possible to use model checking to find anomalies.

Argumentation-based firewall configuration has been studied by several authors. [8] described a technique based on Argumentation for Logic Programming with Priorities, allowing administrators to use high-level abstractions in specifying their network security requirements. In [9], this work was extended to automatically generate firewall policies from higher-level requirements; and in previous work, we [6, 42] have discussed how to use argumentation to handle firewall anomalies and, more generally, to address cyber-security. [41] applied Defeasible Logic Programming (DeLP) for validation and reconfiguration of a firewall, while [14] used and argument-based logic programming engine to enforce security requirements dynamically. Similarly, [29] showed how to analyze security systems using abstract argumentation framework.

More work associated with firewall configuration has been published that goes beyind the use of logic. [19] developed models for policy conflict analysis, taking into account semantic information about policy specifications. [53] proposed a policy algebra framework for security policy enforcement in hybrid firewalls, making use of the basic algebra used in rule sets, such as addition, conjunction, subtraction and summation. [30] presented a firewall analysis engine named Fang, based on a combination of a graph algorithms and a rule-base simulator.

Our use of metalevel argumentation sets our work apart from all of the work cited. All approaches to reasoning about firewall rules yet published have concentrated on object level reasoning. Our work is the first we are aware of to look at reasoning about firewall rules at the metalevel. Indeed, there is very little that has been written on metalevel argumentation. As already mentioned, metalevel argumentation was formally introduced in [32], and the idea of metalevel argumentation has been used [31] to provide an abstract integration of accrual and dialectical argumentation and to integrate argumentation-based reasoning about preferences with the object level arguments. [44] presented an argumentation-based model of social interaction integrating both object-level and metalevel argumentation, and, as already mentioned, [34] integrated structured argumentation and metalevel argumentation to express argumentation schemes. Both these latter papers are a close fit with what we discuss here, though neither addresses our firewall domain.

Finally, where our work touches on the idea of bringing in reasoning about trust at the metalevel, we are clearly beginning to overlap with the work of Villata *et al.* [46, 47] who have written quite extensively about how to represent and resolve arguments that attack arguments about the trustworthiness of agents. If this kind of reasoning were to be incorporated into our framework, it would require a second level of metareasoning — the first metalevel would be used to make statements about the trustworthiness of arguments and the effect of such statements on object-level defeats, and the second metalevel would make statements attacking these statements of trustworthiness. Resolving arguments at the second level would then inform which statements hold at the first level, and hence what arguments were preferred at the object level.

6 Conclusions

This paper has discussed the application of metalevel argumentation to the problem of modelling firewall configurations. In particular, we have shown how to use the metalevel argumentation system of [32] such that object level arguments are concerned with which packets to accept and block in a firewall, and the metalevel arguments identify — and potentially resolve — conflicts between these object level arguments due to shadowing anomalies in the firewall rules. We have argued that since a human system administrator will ultimately have to set the firewall policy, the use of a metalevel formalism — which makes explicit the (metalevel) arguments that explain *why* conflicts in rules arise and *how* they may be resolved — is appropriate. Since the initial version of this paper was published [7], we have conducted human subject experiments [1, 43] which show that using argumentation can be advantageous in complex tasks, giving some initial support for this contention. However, further experiments will be required to confirm this advantage in the firewall configuration domain.

Acknowledgements

Research was partially funded by the National Science Foundation, under grant CNS 1117761 and Army Research Laboratory and Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the National Science Foundation, or the U.S. Government.

Thanks to Sanjay Modgil and the anonymous reviewers for their helpful comments on the paper.

References

[1] N. Ajmeri, C.-W. Hang, S. Parsons, and M. Singh. Aragorn: Eliciting and maintaining secure service policies. (in submission), 2016.

[2] E. Al-Shaer and H. Hamed. Firewall policy advisor for anomaly discovery and rule editing. In *Integrated Network Management, 2003. IFIP/IEEE Eighth International Symposium on*, pages 17–30. IEEE, 2003.

[3] E. Al-Shaer and H. Hamed. Discovery of policy anomalies in distributed firewalls. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 4, pages 2605–2616. IEEE, 2004.

[4] L. Amgoud and C. Cayrol. On the acceptability of arguments in preference-based argumentation framework. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 1–7, 1998.

[5] L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artifical Intelligence*, 34(3):197–215, 2002.

[6] A. Applebaum, K. N. Levitt, J. Rowe, and S. Parsons. Arguing about firewall policy. In *Proceedings of the 4th International Conference on Computational Models of Argument*, pages 91–102, Vienna, Austria, 2012. IOS Press.

[7] A. Applebaum, Z. Li, A. R. Syed, K. Levitt, S. Parsons, J. Rowe, and E. I. Sklar. Firewall configuration: An application of multiagent metalevel argumentation. In *Proceedings of the Workshop on Argumentation in Multiagent Systems (ArgMAS) at Autonomous Agents and MultiAgent Systems (AAMAS)*, Valencia, Spain, June 2012.

[8] A. Bandara, A. Kakas, E. Lupu, and A. Russo. Using argumentation logic for firewall policy specification and analysis. *Large Scale Management of Distributed Systems*, pages 185–196, 2006.

[9] A. Bandara, A. Kakas, E. Lupu, and A. Russo. Using argumentation logic for firewall configuration management. In *Integrated Network Management, 2009. IM'09. IFIP/IEEE International Symposium on*, pages 180–187. IEEE, 2009.

[10] P. Baroni, M. Caminada, and M. Giacomin. An introduction to argumentation semantics. *The Knowledge Engineering Review*, 2011.

[11] N. Basumatary and S. M. Hazarika. Model checking a firewall for anomalies. In *Emerging Trends and Applications in Computer Science (ICETACS), 2013 1st International Conference on*, pages 92–96. IEEE, 2013.

[12] T. J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–428, 2003.

[13] J. Bentahar, R. Alam, Z. Maamar, and N. C. Narendra. Using argumentation to model and deploy agent-based B2B applications. *Knowledge-Based Systems*, 23(7):677–692, 2010.

[14] T. Bouyahia, M. S. Idrees, N. Cuppens-Boulahia, F. Cuppens, and F. Autrel. Metric for security activities assisted by argumentative logic. In *Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance*, pages 183–197. Springer, 2015.

[15] M. W. A. Caminada. On the issue of reinstatement in argumentation. In *Proceedings of the 10th European Conference on Logic in Artificial Intelligence*, pages 111–123, Liverpool, UK, 2006.

[16] M. W. A. Caminada. An algorithm for computing semi-stable semantics. In *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 222–234, Verona, Italy, 2007.

[17] M. W. A. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5–6):286–310, 2007.

[18] M. W. A. Caminada, W. A. Carnielli, and P. E. Dunne. Semi-stable semantics. *Journal of Logic and Computation*, 22:1207–1254, 2012.

[19] S. Davy and B. Jennings. Harnessing models for policy conflict analysis. *Inter-Domain Management*, pages 176–179, 2007.

[20] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. *Artificial Intelligence*, 77:321–357, 1995.

[21] P. Eronen and J. Zitting. An expert system for analyzing firewall rules. In *Proceedings of the 6th Nordic Workshop on Secure IT Systems (NordSec 2001)*, pages 100–107, 2001.

[22] J. Govaerts, A. Bandara, and K. Curran. A formal logic approach to firewall packet filtering analysis and generation. *Artificial Intelligence Review*, 29(3):223–248, 2008.

[23] S. Hazelhurst. Algorithms for analysing firewall and router access lists. Technical report, Department of Computer Science, University of the Witwatersrand, 2000.

[24] S. Hazelhurst, A. Fatti, and A. Henwood. Binary decision diagram representations of firewall and router access lists. Technical report, Department of Computer Science, University of the Witwatersrand, 1998.

[25] K. Ingham and S. Forrest. Network firewalls. *Enhancing computer security with smart technology*, pages 9–40, 2006.

[26] P. Krause, S. Ambler, M. Elvang-Gørannson, and J. Fox. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11 (1):113–131, 1995.

[27] J. Kurose and K. Ross. *Computer networking: a top-down approach*. Addison-Wesley, 2010.

[28] Z. Li and S. Parsons. On argumentation with purely defeasible rules. In *9th International Conference on Scaleable Uncertainty Management*, Quebec City, 2015.

[29] F. Martinelli, F. Santini, and A. Yautsiukhin. Network security supported by arguments. In *Privacy, Security and Trust (PST), 2015 13th Annual Conference on*, pages 165–172. IEEE, 2015.

[30] A. Mayer, A. Wool, and E. Ziskind. Fang: A firewall analysis engine. In *Security and Privacy, 2000. S&P 2000. Proceedings. 2000 IEEE Symposium on*, pages 177–187. IEEE, 2000.

[31] S. Modgil and T. J. Bench-Capon. Integrating dialectical and accrual modes of argumentation. *COMMA*, 216:335–346, 2010.

[32] S. Modgil and T. J. M. Bench-Capon. Metalevel argumentation. *Journal of Logic and Computation*, 6(21):959–1003, 2011.

[33] S. Modgil and H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.

[34] J. Müller, A. Hunter, and P. Taylor. Meta-level argumentation with argument schemes. In *Scalable Uncertainty Management*, pages 92–105. Springer, 2013.

[35] S. Parsons, K. Atkinson, Z. Li, P. McBurney, E. I. Sklar, M. Singh, K. Haigh, K. Levitt, and J. Rowe. Argument schemes for reasoning about trust. *Argumentation & Computation*, 5(2–3):160–190, 2014.

[36] S. Parsons, E. I. Sklar, and P. McBurney. Using argumentation to reason with and about trust. In *Proceedings of the 8th International Workshop on Argumentation in Multiagent Systems*, Taipei, Taiwan, 2011.

[37] S. Parsons, E. I. Sklar, J. Salvit, H. Wall, and Z. Li. ArgTrust: Decision making with information from sources of varying trustworthiness (Demonstration). In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, St Paul, MN, 2013.

[38] S. Parsons, M. Wooldridge, and L. Amgoud. Properties and complexity of formal inter-agent dialogues. *Journal of Logic and Computation*, 13(3):347–376, 2003.

[39] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1:93–124, 2010.

[40] H. Prakken and G. Sartor. Argument-based logic programming with defeasible priorities. *Journal of Applied Non-classical Logics*, 1997.

[41] P. Rajkhowa, S. M. Hazarika, and G. R. Simari. An application of defeasible logic programming for firewall verification and reconfiguration. In *Quality, Reliability, Security and Robustness in Heterogeneous Networks*, pages 527–542. Springer, 2013.

[42] J. Rowe, K. Levitt, S. Parsons, E. I. Sklar, A. Applebaum, and S. Jalal. Argumentation logic to assist in security administration. In *Proceedings of the 2012 Workshop on New Security Paradigms*, pages 43–52. ACM, 2012.

[43] E. I. Sklar, S. Parsons, Z. Li, J. Salvit, S. Perumal, H. Wall, and J. Mangels. Evaluation of a trust-modulated argumentation-based interactive decision-making tool. *Journal of Autonomous Agents and Multi-Agent Systems*, 30(1):136–173, 2016.

[44] E. I. Sklar, S. Parsons, and M. P. Singh. Towards an argumentation-based model of social interaction. In *Proceedings of the Workshop on Argumentation in Multiagent Systems (ArgMAS) at the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). St. Paul, MN*, 2013.

[45] B. Verheij. A labeling approach to the computation of credulous acceptance in argumentation. In *Proceedings of the 20th International Joint Conference on Aritificial Intelligence*, pages 623–628, Hyderabad, India, 2007.

[46] S. Villata, G. Boella, D. M. Gabbay, and L. van der Torre. Arguing about trust in multiagent systems. In *Proceedings of the 11th Symposium on Artificial Intelligence of the Italian Association for Artificial Intelligence*, Brescia, Italy, 2010.

[47] S. Villata, G. Boella, D. M. Gabbay, and L. van der Torre. Arguing about the trustworthiness of the information sources. In *Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Belfast, UK, 2011.

[48] G. Vreeswijk. An algorithm to compute minimally grounded and admissible defence sets in argument systems. In *Proceedings of the First International Conference on Computational Models of Argument*, pages 109–120, Liverpool, UK, 2006.

[49] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, Albany, NY, USA, 1995.

[50] M. Wooldridge. *An Introduction to Multiagent Systems*. Wiley, 2nd edition, 2009.

[51] M. J. Wooldridge, S. Parsons, and P. McBurney. The meta-logic of arguments. In *Proceedings of the 4th International Conference on Autonomous Agents and Multi-Agent Systems*, Utrecht, The Netherlands, July 2005.

[52] L. Yuan, H. Chen, J. Mai, C. Chuah, Z. Su, and P. Mohapatra. Fireman: A toolkit for firewall modeling and analysis. In *Security and Privacy, 2006 IEEE Symposium on*, pages 15–pp. IEEE, 2006.

[53] H. Zhao and S. Bellovin. Policy algebras for hybrid firewalls. In *Annual Conference of ITA (ACITA)*, volume 2007, 2007.