Semantics of Trust

Tim Muller

University of Luxembourg
tim.muller@uni.lu

Abstract. The meaning assigned to the word 'trust' is diverse. We present a formalism that allows various interpretations of trust.
To this end, we introduce terms that specify the observations of agents, called connections. Then we apply epistemic semantics to reason about the knowledge of agents. We allow specifications of interpretations of trust in terms of facts, and analyze whether agents know the relevant facts. If agents know that a target is trustworthy under an interpretation, that agent trusts the target.
We illustrate the formalism on three specific existing interpretations.

1 Introduction

When one asks an ethicist for a definition of trust, one might expect a response such as: "Trust (...) is letting other persons (natural or artificial, such as firms, nations, etc.) take care of something the trustor cares about, (...)" [1]. An economist might say: "Trust is the willingness to permit the decisions of others to influence your welfare" [2]. An interesting sociological definition is: "(Trust is the) undertaking of a risky course of action on the confident expectation that all persons involved in the action will act competently and dutifully." [3]. A game theoretical view on the issue is formulated as: "(Trust) is the mutual confidence that one's vulnerability will not be exploited in an exchange." [4]. [1]

In this paper, we argue that trust is something more abstract, and we refer to more specific notions, such as the aforementioned, as specific *meanings* of trust. First we define the general notions regarding trust:

Definition 1. *An* observation *is any contingent fact, that is witnessed to be true. A* trust assessment *is a boolean expectation based on own observations and possibly observations of others.* Trust *is a positive trust assessment, and* distrust *a negative. An* interpretation *of trust defines under which condition to trust or distrust an agent. Interpretations are formulated as predicates on agents. Given an interpretation, if an agent with perfect information trusts another agent, then the latter is* trustworthy*.*

The meaning of trust assessments differs according to the context. The meaning depends on what an expectation denotes, as illustrated by the aforementioned examples. An interpretation of trust determines which condition yields a positive expectation. A trust system contains zero or more interpretations and meanings of trust at the same time. To illustrate this with an example:

[1] Thanks to Harvey S. James, Jr. for collecting trust related quotes on his webpage.

Example 1. A user visits the Microsoft website, to get a software update. The software has a certificate, showing that the update is from Microsoft. After seeing the certificate, the user trusts the software and installs it.

One *meaning* of trust here, is that the user has the *expectation* that the software is not malware. An *interpretation* of trust is whether the software is certified. As the software is certified, the user has a positive *trust assessment*, hence the user *trusts* that the software. The usage of the word "trust" in the previous sentence implicitly uses the aforementioned interpretation and meaning. As common practice dictates, the meaning and interpretation of trust are implicit in situations where they are obvious.

In the example, a secondhand observation is the reason that the user gives a positive assessment. Namely, the user observed that a particular certifying authority has issued a certificate to Microsoft. Apparently there is some type of trust in the issuer of the certificate. Let's look at the situation with a finer grain. We introduce a new meaning of trust in the system, which is used alongside the aforementioned meaning of trust. The end-user trusts that the certifying authority is honest, and he trusts the ability of the certifying authority to assess the intentions of particular software developers. The new interpretation of trust is being a particular certifying authority.

In this example, different types of trust are dependant. The first interpretation could be formulated as: "A truthful certifying authority signed a software certificate". Which depends on a meaning of trust, such as: "We expect certifying authorities are truthful".

We observe that trust systems generally solve two particular problems, possibly at the same time: How to *collect* observations, and which *interpretation(s)* of trust to use. An example of a system designed to optimize the collection of data is [5], where agents not only delegate jobs to other agents that they need done, but also send out challenges to other agents. An agent will know the solution of a challenge, and if the challenged agent fails to return the right result, he will be trusted less. This only works in specialized settings, as the relevance of this solution depends on the indistinguishability between regular jobs and challenges. Whereas a paper such as [6] investigates ways to interpret trust. It focusses on different properties of interpretations, and formal denotations of these properties. How these interpretations relate to actual models is not defined.

One of the goals of this paper is to enable us to easily recombine the good aspects of existing, specialized solutions. A solution for both problems is needed for recommender systems. Recommender systems have several aspects that are studied, as shown in [7]. Those aspects represent different subproblems. Filtering, for example, is the process of turning a real item into a subject in the recommender system. Filtering is purely about data collection. Matching is the way a recommender system calculates which opinions agents should follow. What type of matching to use is an interpretation of trust. There are aspects, such as how feedback is given and how sensitive it is to changes over time, that depend on both data collection and interpretation.

Interpretations of trust are usually defined in terms of facts. When the software has a certificate from a particular certifying authority, the software is trustworthy on that basis. If an agent knows the right facts, the agent knows that another agent is trustworthy, and hence trust the agent. Given an interpretation, the problem of trusting can be reduced to a problem of having the right knowledge.

In most concrete contexts, the relation between observations and knowledge is obvious. As a consequence, the transformation of observations into knowledge is an underexposed issue. We, however, are interested in a more general context. Hence, we need to be explicit in our dealings with observations. To represent observations in this generalized setting, we introduce a formalism. The formalism is, although fundamentally different, derived from subjective logic [8], in the sense that it uses dilution and fusion. The fundamental differences lie in the fact that terms in subjective logic represent opinions in the form of probabilities and estimates, while terms in our formalism represent observations in the form of facts and statements. There is an axiomatization presented in [9], which abstracts from subjective logic. Due to this abstraction, some properties that are axiomatized will still hold in our context. Having set a formalism to express observations, we need to be able to transform these observations into knowledge in a general way. We should not, generally, treat secondhand observations as knowledge, nor should we disregard them totally. Hence, we should evaluate the trust in the source that claimed to have made the observation. "Are sources of information capable and honest" is a possible meaning of trust. This type of trust will be used to evaluate third-party observations, and is therefore assigned a special status, namely *reliability*. In other words, reliability is an instance of trustworthiness, where trust has the meaning "being capable and honest".

2 Connections and Networks

There are two well-studied aspects of trust systems. We recall data collection and interpretation, mentioned in the introduction. One of our goals is to glue these together in a generic way. As mentioned before, we need to transform observations into knowledge in an abstract way. To achieve this, we need a formal way to state the observations of agents. The formal representation of the observations of an agent is called the *connection* of that agent.

Every user in a trust system has a (possibly empty) set of observations regarding other users. If a has an observation about b, then b does not necessarily have an observation about a. If a has a secondhand observation regarding b, it means that another user, say c, claims to have a particular observation regarding b. This is considered as a's observation regarding b, albeit secondhand. However, to determine the value of c's claim, a can use its observations about c to assess the reliability of c. Recall that reliability is an instance of trustworthiness, for a meaning of trust such as "c speaks the truth". Since we want to keep our formalism general, we do not assume any interpretation of trust. As we need

an interpretation to assess trust, it is generally not possible to assess whether c speaks the truth. In Section 3, we address this problem.

In the preceding paragraph we hinted at the shape of connections. A connection that represents lack of observations is denoted as ε. We are able to denote a connection representing precisely one observation, by just giving the predicate representing the observation. Since an agent can have more than one observation, we should be able to combine connections, we call this a fusion of connections, denoted $\_ + \_$. Furthermore, we should be able to express that another agent, say a, makes a claim. This is called dilution, and is written $a.\_$. The syntax is defined over agents, who may state claims, and observations regarding agents, which we will denote as a predicate on said agents. The syntax of connections is defined over finite sets of agents $\mathcal{A}$ and possible observations $\mathcal{P}$ regarding agents:

$$\varphi ::= \varepsilon \mid P(a) \mid a.\varphi \mid \varphi + \varphi$$

for agent $a \in \mathcal{A}$ and observation $P \in \mathcal{P}$ regarding an agent. We treat observations as logical predicates, in other words $P \in \mathcal{P} : \mathcal{A} \to \{\top, \bot\}$. Logical relationships may exist between predicates.

For example, let $O(d)$ denote that door d is open, and $C(d)$ denote that d is closed. Obviously, in situations where a door d is open, d cannot be closed, denoted $O(d) \vDash^S \neg C(d)$. A set of those semantic rules S is called a signature. A signature defines the logical relationships between predicates. For this paper, it suffices to realize that, given a signature, not all combinations of predicates can be satisfied. The door cannot be both open and closed. We will refer to this property as *consistency*:

Definition 2. *A set of predicates Γ is consistent when there is no predicate $\varphi \in \Gamma$ such that $\Gamma \vDash^S \neg\varphi$.*

Using the example, $\Gamma = \{O(d), C(d)\}$ is, as we expect, not consistent, since $\{O(d), C(d)\} \vDash^S \neg C(d)$.

In [9], it was shown that fusion of opinions is associative and commutative, and that total uncertainty is the identity element. On that basis, we expect the fusion of connections also to be associative and commutative. Section 3 trivially shows the expectation is correct. And since an empty connection yields no information, we expect the empty connection to be the identity element. This allows us to treat every connection as a summation. If one of the summands is a particular observation, we can say that the connection contains that observation.

Let $\mathcal{C}$ be the set of all connections allowed by our syntax. We note that, as a consequence of the definition, $\mathcal{C}$ is countable. Using the notion of connections, we define a network. A network defines which agent has what connection. A network is represented as a function on agents $N : \mathcal{A} \to \mathcal{C}$. Rather than using the function type notation, we define a countable set of networks as $\mathcal{N}$. Note that the range of a network is a subset of $\mathcal{C}$. If a connection is in the range of a network, we simply say that the connection exists in the network.

Example 2. Figure 1 depicts an example of a network. To formally describe this network, we need to define a set of agents. We pick $\mathcal{A} = \{a, b, c\}$. We

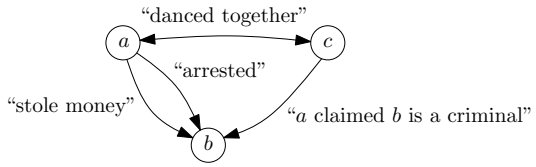

Fig. 1. An example of a network N of three agents

furthermore take a set of predicates: Let $D_\alpha(\beta)$ denote "β danced with α". Let $C(\alpha)$ denote "α is a criminal". Then $N(a) = C(b) + C(b) + D_a(c)$, $N(b) = \varepsilon$ and $N(c) = D_c(a) + a.C(b)$.

Let $P(a)$ and $Q(b)$ be inconsistent predicates. Agents only have observations that correspond to the real world. If an agent has observation $P(a)$, no other agent can have observation $Q(b)$. Since $P(a)$ has been observed, it must hold, and since it contradicts $Q(b)$, $Q(b)$ cannot hold, and hence not be observed. E.g. if a car is red, no agent can make the observation that the car is blue. Mind that it is, of course, possible that an agent claims that he observed that the red car is blue.

3 Generic Semantics

In order to successfully transform observations into knowledge, we have introduced a formalism to unambiguously represent observations in Section 2. We will now present a method that takes a connection and assigns a semantics to it, in the form of knowledge. The association of knowledge of agents with a particular model is a well-studied subject in logics, called epistemic logic. There is an introduction to semantics of logics and modal logic in [10]. An in-depth discussion of modal logics and their semantics can be found in [11]. Since this section will deal with epistemic modal logic and knowledge in general, we present [12] as a source of information about epistemic logic in particular.

Given a network, there should be a unique method to evaluate specific interpretations of trust. Formulated negatively; if, given a network, two methods of evaluating trust yield different trust assessments, then the two methods do not use the same interpretations. A network is seen as a model that has epistemic semantics. Paragraph 3.1 shows how this is done. In Paragraph 3.2, we formulate interpretations in epistemic logic. Together, this automatically yields a general method to evaluate interpretations in a model.

3.1 Epistemic Semantics

We assert that there is a real world, and in this world all facts are true or false. No agent, in general, will be able to distinguish the real world from other possible worlds, as he has not observed all facts. Within our model we distinguish three types of facts:

Whether a property holds.
Whether a claim has been made.
Whether a particular agent would lie about something.

The real world can now be defined using valuations, assignments of truth values, over these facts. In the real world, a property either holds or it does not, a claim has been made or it has not, and an agent is willing to lie or he is not.

We call the mapping from predicates $\mathcal{P} \times \mathcal{A}$, to truth values, a predicate valuation $\theta_{\mathcal{P}}$. We call a mapping from pairs of agents and connections $\mathcal{A} \times \mathcal{A} \times \mathcal{C}$, to whether a claim has been made, a statement valuation $\theta_{\mathcal{S}}$. We call a mapping from pairs of agents and connections $\mathcal{A} \times \mathcal{A} \times \mathcal{C}$, to truthfulness of claims, a reliability valuation $\theta_{\mathcal{R}}$. This mapping reasons about hypothetical situations. Namely, how reliable an agent would be, when he would state a particular claim to another agent. For example, for a perfectly honest agent a, $\theta_{\mathcal{R}}(a, b, x)$ will be true for all b and x. Whereas, for a compulsive liar it will be always be false. For a rational agent a, it will be true in those situations when it serves a's best interest to speak the truth. The particular reliability valuation defines the reliability of agents.

Let $W_{\mathcal{P}}$ be the set of all valuations $\theta_{\mathcal{P}}$.
Let $W_{\mathcal{S}}$ be the set of all valuations $\theta_{\mathcal{S}}$.
Let $W_{\mathcal{N}}$ be the set of possible networks.
Let $W_{\mathcal{R}}$ be the set of all valuations $\theta_{\mathcal{R}}$.

We can see that $x \in W_{\mathcal{P}} \times W_{\mathcal{S}}$ denotes all objective facts in a world. The set of possible worlds W is a subset of $W_{\mathcal{P}} \times W_{\mathcal{S}} \times W_{\mathcal{N}} \times W_{\mathcal{R}}$. Note, again, that if it is observed that $P(a)$ holds, then $P(a)$ must hold. Hence some combinations of valuations and networks cannot exist. The possibility relation defines which worlds can exist, and which are contradictory. The possibility relation, $\mathit{\Pi}$, enforces three conditions:

The consistency of the valuation over predicates, as defined in definition 2.
The connections of all agents in the network must adhere to reality; which is defined by the conformance function.
Furthermore, when a claims is reliable and has been made, then it must also conform to reality.

The conformance function, π, enforces that the observations x of an agent a are in line with reality:

If an agent has no observations, they cannot contradict reality.
If an agent observed a fact, then it must be true.
If an agent received a claim, then the claim must have been made.
If an agent has several connections, all must be in line with reality.

Let the set of possible worlds W be defined as $\{(w_{\mathcal{P}}, w_{\mathcal{S}}, w_{\mathcal{N}}, w_{\mathcal{R}}) \mid w_{\mathcal{P}} \in W_{\mathcal{P}} \wedge w_{\mathcal{S}} \in W_{\mathcal{S}} \wedge w_{\mathcal{N}} \in W_{\mathcal{N}} \wedge w_{\mathcal{R}} \in W_{\mathcal{R}} \wedge \mathit{\Pi}(w_{\mathcal{P}}, w_{\mathcal{S}}, w_{\mathcal{N}}, w_{\mathcal{R}})\}$, where $\mathit{\Pi}(p, c, n, r)$ is

the possibility relation on valuations, defining whether a world $w = (p, c, n, r)$ is possible:

$$\begin{aligned} \Pi(w_{\mathcal{P}}, w_{\mathcal{S}}, w_{\mathcal{N}}, w_{\mathcal{R}}) = \; & consistent(\{w_{\mathcal{P}}(P, a) | P \in \mathcal{P}, a \in \mathcal{A}\}) \wedge \\ & \forall_{a \in \mathcal{A}}(\pi(w_{\mathcal{P}}, w_{\mathcal{S}}, a, w_{\mathcal{N}}(a))) \wedge \\ & \forall_{a,b \in \mathcal{A}, x \in \mathcal{C}}(w_{\mathcal{R}}(a, b, x) \Rightarrow w_{\mathcal{S}}(a, b, x) \Rightarrow \pi(w_{\mathcal{P}}, w_{\mathcal{S}}, b, x)) \end{aligned}$$

and $\pi(p, c, a, x)$ is the conformance function defining if a specific connection x of an agent a is possible when the objective facts are defined by (p, c).

$$\pi(w_{\mathcal{P}}, w_{\mathcal{S}}, a, x) = \begin{cases} x = \varepsilon \vee \\ x = P(b) \wedge w_{\mathcal{P}}(P, b) \vee \\ x = c.x \wedge w_{\mathcal{S}}(a, c, x) \vee \\ x = x + y \wedge \pi(w_{\mathcal{P}}, w_{\mathcal{S}}, a, x) \wedge \pi(w_{\mathcal{P}}, w_{\mathcal{S}}, a, y) \end{cases}$$

In Kripke semantics, worlds may be related to each other.In epistemic semantics, two worlds are related to each other, if there is a particular agent cannot tell the difference between the two worlds. Such a relation is called an accessability relation. In our context, two worlds are related if an agent has the same connection in both worlds. For each agent $a \in \mathcal{A}$, let $\mathbf{R}^a$ be the relation over worlds, such that $(w_{\mathcal{P}}, w_{\mathcal{S}}, w_{\mathcal{N}}, w_{\mathcal{R}})\mathbf{R}^a(w'_{\mathcal{P}}, w'_{\mathcal{S}}, w'_{\mathcal{N}}, w'_{\mathcal{R}})$ iff $w_{\mathcal{N}}(a) = w'_{\mathcal{N}}(a)$.

Furthermore, we have functions $\Theta_{\mathcal{P}} : W \to \mathcal{P} \times \mathcal{A} \to \{\top, \bot\}$ defined as $\Theta_{\mathcal{P}}(w_{\mathcal{P}}, w_{\mathcal{S}}, w_{\mathcal{N}}, w_{\mathcal{R}}) = w_{\mathcal{P}}$, $\Theta_{\mathcal{S}} : W \to \mathcal{A} \times \mathcal{A} \times \mathcal{C} \to \{\top, \bot\}$ defined as $\Theta_{\mathcal{S}}(w_{\mathcal{P}}, w_{\mathcal{S}}, w_{\mathcal{N}}, w_{\mathcal{R}}) = w_{\mathcal{S}}$ and $\Theta_{\mathcal{R}} : W \to \mathcal{A} \times \mathcal{A} \times \mathcal{C} \to \{\top, \bot\}$ defined as $\Theta_{\mathcal{R}}(w_{\mathcal{P}}, w_{\mathcal{S}}, w_{\mathcal{N}}, w_{\mathcal{R}}) = w_{\mathcal{R}}$. These are functions are projections of worlds onto their respective valuations. In particular, $\Theta_{\mathcal{P}}$ determines if in a particular world, a particular observable predicate holds, i.e. a predicate from $\mathcal{P}$. Whereas $\Theta_{\mathcal{S}}$ determines if in a particular world, a particular claim x has been made by an agent a towards another agent b. We introduce the following notation for such a predicate: $\boldsymbol{C}^a_b(x)$. Finally, $\Theta_{\mathcal{R}}$ determines if in a particular world, a particular agent a is reliable when claiming x towards b. We introduce the following notation for such a predicate: $\boldsymbol{R}^a_b(x)$.

We find the following relation between observational models and epistemic logic in a straightforward manner:

Definition 3. *The observational model is a Kripke model, defined by* $\mathcal{K} = \langle W, \mathbf{R}^a, \mathbf{R}^b, \ldots, \Theta_{\mathcal{P}}, \Theta_{\mathcal{S}}, \Theta_{\mathcal{R}} \rangle$.

$$\begin{aligned} \mathcal{K}, w \models & \top \\ \mathcal{K}, w \models & \varphi \wedge \psi && \text{iff } \mathcal{K}, w \models \varphi \text{ and } \mathcal{K}, w \models \psi \\ \mathcal{K}, w \models & \neg\varphi && \text{iff } \mathcal{K}, w \not\models \varphi \\ \mathcal{K}, w \models & P(b) && \text{iff } \Theta_{\mathcal{P}}(w)(P, b) \\ \mathcal{K}, w \models & \boldsymbol{C}^a_b(x) && \text{iff } \Theta_{\mathcal{S}}(w)(a, b, x) \\ \mathcal{K}, w \models & \boldsymbol{R}^a_b(x) && \text{iff } \Theta_{\mathcal{R}}(w)(a, b, x) \\ \mathcal{K}, w \models & K_a(\varphi) && \text{iff for all } w' \text{ such that } w\mathbf{R}^a w' \text{: } \mathcal{K}, w' \models \varphi \end{aligned}$$

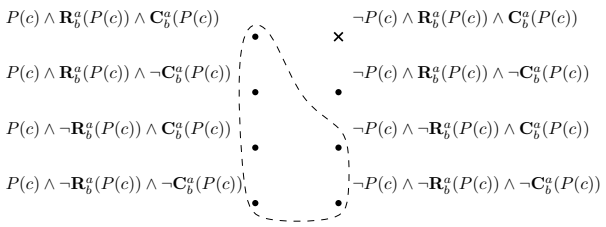

Fig. 2. A set of 7 possible worlds.

Example 3. Figure 2 depicts some possible worlds as dots annotated with the truth values of some predicates. The world depicted as a cross is not a possible world using the possibility relation. If $\boldsymbol{R}_b^a(P(c))$ and $\boldsymbol{C}_b^a(P(c))$ then $\pi(w_\mathcal{P}, w_\mathcal{R}, b, P(c))$. However, $w_\mathcal{P}(\neg P(c))$, thus $\neg\pi(w_\mathcal{P}, w_\mathcal{R}, b, P(c))$.

The points in the area marked by the dashed line represent states where the following statement holds: "If b is reliable to a when claiming that $P(c)$, then $P(c)$ holds." This type of collection of worlds will turn out to be relevant when dealing with claims.

3.2 Interpretations of Trustworthiness

Suppose we introduce an interpretation of trust into the model. An interpretation of trust tells us what predicate needs to be true in order for an agent to be trustworthy. Hence, we may view an interpretation as a predicate on agents. Let I be such an interpretation. Let the set of predicates in the system be $\{P\}$, and $I(a) = P(a)$. We interpret trust, such that an agent a is trustworthy when $P(a)$ holds. We formally define interpretations of trustworthiness:

Definition 4. *Let I be an* interpretation *and a an agent. We introduce a generic symbol for* trustworthiness*, $T^I(a)$. The semantics associated with this type of trust is:*

$$\mathcal{K}, w \models T^I(a) \text{ iff } \mathcal{K}, w \models I(a)$$

We define trust using the following notion; knowing an agent is trustworthy means being able to trust that agent. We can apply the definition of knowledge in Definition 3. We see that, only defining when an agent is trustworthy, is sufficient to define when it is trusted.

Proposition 1.

$$\mathcal{K}, w \models K_a(T^I(b)) \text{ iff } \mathcal{K}, w \models K_a(I(b))$$

Before we proceed, notice that it might be possible that in some world, an agent might be trustworthy to one agent, but not to another. A solution to the aforementioned problem, would be to index interpretations, in order to represent to whom trustworthiness applies. If there are N agents, then there are N different

interpretations of trust, one for each agent. We write that agent b is trustworthy to agent a under interpretation I as $T^{I_a}(b)$. An agent a trusts b under I, is then written $K_a(T^{I_a}(b))$. In social systems, it is common to have a situation where a knows that c is trustworthy towards b. In general, however, this has no consequence for agent a. Our formalism is flexible enough to allow a model where it does affect a.

Earlier, we discussed that some interpretations of trust are, in a sense, self-referential. Namely, when we trust that certain statements from agents to agents are reliable. We define a finite set ζ of claims that are asserted to be reliable. When an agent is trying to figure out whether another agent is trustworthy, it will assume that the claims they have heard are true, when the are in ζ. In other words, reliability of all the claims in ζ is a condition under which to determine interpretations of trustworthiness. Trustworthiness under ζ is more formally defined in Definition 5.

A world where a claim from ζ is falsely stated, is considered to be a possible world. However, agents assess their trust in a model under the assumption that such a world is not relevant. For example, a friend tells you that his car broke down. This does, generally, not imply that you know his car broke down. He might be lying, exaggerating or mistaken. It does, however, mean that you will assess trust in his car dealer using the assumption that your friends car broke down.

As a sidetrack, it is interesting to think about the approach that does interpret trust in truthfulness as a source of knowledge. In other words, a world where a claim from ζ is falsely stated is not considered possible. This approach is worked out into fine detail in [13]. What we call claims are called announcements there. An announcement updates the knowledge. However, such an approach is inherently dynamic. Hence, in [13] the emphasis must be on mixing epistemic logic with dynamic logics. This is a difficult task, since determining a false announcement to be reliable will lead to logical contradiction. For this reason, we prefer a static model. Our approach allows us to have a static model in a natural way.

Using the notion of reliable claims, we extend the definition of trustworthiness to:

Definition 5. *Let $T^I_\zeta(a)$ denote that, given that the claims in ζ are reliable, a is* trustworthy *under interpretation I.*

$$\mathcal{K}, w \models T^I_\zeta(a) \text{ iff } \mathcal{K}, w \models \bigwedge_{(b,c,x)\in\zeta} (\boldsymbol{R}^b_c(x)) \Rightarrow T^I(a)$$

Only defining when an agent is trustworthy is sufficient to define when it is trusted.

Proposition 2.

$$\mathcal{K}, w \models K_a(T^I_\zeta(b)) \text{ iff } \mathcal{K}, w \models K_a((\bigwedge_{(c,d,x)\in\zeta} \boldsymbol{R}^c_d(x)) \Rightarrow I(b))$$

We remark that $T^I(b)$ is, by definition, equivalent to $T^I_\emptyset(b)$.

We will see that the set builder for the set ζ can depend on conditional trust itself. When chaining trust, for example, it is possible that a delegates trust to b who further delegates that trust to c. Hence, we trust c, if the claim of delegation D of b is reliable, in other words, if $D \in \zeta$. However, D is in ζ, only when we somehow trust b, which is the case only when a was reliable. The next section will provide examples of interpretations of trust, and in particular chained interpretations of trust.

4 Existing Interpretations

The main idea of the generalized semantics that we present, is that they are easy to translate to. We will naively fill in the relevant observations and the interpretation of trust. If our approach works, then we are able to prove the given naive translations correct. Correctness of a translate means that, if we can derive trust in the original model, then we can derive trust in the translation, and vice versa.

4.1 Flat Public Key Infrastructure

A flat public key infrastructure is the most basic type of public key infrastructure. There are users and certifying authorities. A certifying authority signs certificates of users. A user u trusts another user u' when the certificate of u' has been signed by some trusted certifying authority. There is a special type of certifying authority, called a *root certifier*. Every user trusts precisely those certifying authorities that are root certifiers. Note that in Paragraph 4.2, root certifiers can delegate their trust to other certifying authorities.

In order to do an analysis, we present a more formal way to describe what a flat public key infrastructure is: Let U be a set of users. Let C be a disjoint set of certifying authorities, and C_{root} be the subset representing root certifiers. Let $S(c, u)$ denote that certifying authority c has signed a certificate of user u. We define that $observes_u(\varphi)$ means that u observes φ. Then user u trusts correctness of the public key of u' iff $\exists_{c \in C}(observes_u(\text{cisarootcertifier}) \wedge observes_u(S(c, u')))$.

The set of agents $\mathcal{A}$ is equivalent to the union of the users and the certifiers. The set of predicates $\mathcal{P}$ is $\{P\} \cup \{S_a \mid a \in C\}$. The predicate $P(a)$ is true when a is a root certifier, $a \in C_{root}$. The predicate $S_a(b)$ is true when $S(a, b)$, thus when a signed certificate of b.

Then we need to define the network. The connection of a user a contains precisely the observations of a. In other words, a has a connection $\sum_{b \in \mathcal{A}, observes_a(P(b))} P(b) + \sum_{b,c \in \mathcal{A}, observes_a(S(b,c))} b.S_b(c)$.

Lastly, we formally interpret trust. An agent a trusts an agent b, if there is a root certifying authority that claims to have signed b's certificate. We call this the basic interpretation of trust. We define such a basic interpretation of trust as follows:

$$T^{I^b}_\zeta(a)$$

where

$$I^b(a) = \bigvee_{b \in \mathcal{A}} (P(b) \wedge S_b(a))$$

and

$$\zeta = \{(a, b, S_b(c)) \mid a, b, c \in \mathcal{A} \wedge P(b)\}$$

Note that ζ represents the claims regarding signed certificates made by the root certifiers. In other words, for the basic trust interpretation, we assert that root certifiers are reliable in claims about certificates.

We prove correctness of the interpretation:

Lemma 1. *User u trusts correctness of the public key of u', iff $\mathcal{K}, w \models K_u T_\zeta^{I^b}(u')$.*

Proof. We prove the implication in both ways.

$\Rightarrow$ Agent u observes there is a root certifier $c \in C_{root}$, such that u observes that c claims $S(c, u')$. As a consequence, the connection of u contains a summand $P(c)$ and a summand $c.S_c(u')$. Since u has a connection $P(c) + y$, $w_{\mathcal{N}}(u) = P(c) + y$, $\Pi(w_{\mathcal{P}}, w_{\mathcal{S}}, w_{\mathcal{N}}, w_{\mathcal{R}}) \Rightarrow \pi(w_{\mathcal{P}}, w_{\mathcal{S}}, u, P(c) + y) \Rightarrow w_{\mathcal{P}}(P, c)$. Similarly, for $c.S_c(u')$ we see that $\boldsymbol{C}_c^u(S_c(u'))$. In all accessible worlds for u, $w_{\mathcal{N}}(u)$. Hence, $w_{\mathcal{P}}(P, c)$ holds in every world w', such that $w\mathbf{R}^u w'$, and thus $K_u P(c)$ and $K_u \boldsymbol{C}_c^u(S_c(u'))$.
As a consequence of $P(c)$, $(u, c, S_c(u')) \in \zeta$. We see, using the possibility relation and $\boldsymbol{C}_c^u(S_c(u'))$, that in all worlds where $\boldsymbol{R}_c^u(S_c(u'))$, $\pi(w'_{\mathcal{P}}, w'_{\mathcal{S}}, c, S_c u')$ also holds, and thus $w'_{\mathcal{P}}(S_c, u')$. Since $w'_{\mathcal{P}}(S_c, u')$ holds in every possible world for u, $\mathcal{K}, w' \models K_a(S_c(u))$. Hence, we can apply Proposition 2, to see that $\mathcal{K}, w \models K_u T_\zeta^{I^b}(u')$.

$\Leftarrow$ By definition, if $\mathcal{K}, w \models K_u T_\zeta^{I^b}(u')$ then $\mathcal{K}, w \models K_u((\bigwedge_{(c,d,x)\in\zeta} \boldsymbol{R}_d^c(x)) \Rightarrow I^b(u'))$. We get $\mathcal{K}, w \models K_u((\bigwedge_{c,d\in\mathcal{A}\wedge x\in\mathcal{C}\wedge P(d)} \boldsymbol{R}_d^c(x)) \Rightarrow \exists_{b\in\mathcal{A}}(P(b)\wedge S_b(u')))$ by applying the definition of ζ and I^b. Hence, for all w' such that $w\mathbf{R}^u w'$, $\mathcal{K}, w' \models (\bigwedge_{c,d\in\mathcal{A}\wedge x\in\mathcal{C}\wedge P(d)} \boldsymbol{R}_d^c(x)) \Rightarrow \exists_{b\in\mathcal{A}}(P(b) \wedge S_b(u'))$. We need to push the knowledge into the model. Let $w' = (w'_{\mathcal{P}}, w'_{\mathcal{S}}, w'_{\mathcal{N}}, w'_{\mathcal{R}})\mathbf{R}^u w$, we push the knowledge into the model by analyzing the shape of $w'_{\mathcal{N}}(u)$. We recall that every connection can be seen as a summation. We need only analyze what summands must exist in $w'_{\mathcal{N}}(u)$.
If there is no agent a such that there is a summand $P(a)$, then there is a world w', such that for no agent a, $P(a)$. In such a world $\zeta = \emptyset$, and clearly $T_\emptyset^{I^b}(u')$ does not hold, since there is no a with $P(a)$.
Hence, we can assume that there is an agent a, such that there is a summand $P(a)$. By applying the possibility relation, we see that for all worlds w', $w'_{\mathcal{P}}(P, a)$. Therefore $\zeta = \{(b, a, S_a(c)) \mid a, b, c \in \mathcal{A} \wedge P(a)\}$.
We furthermore know that agent a exist where $P(a)$ and $a.S_a(u')$ are summands of u's connection. Assume that all agents a for which there is a summand $P(a)$, there is no summand $a.S_a(u')$ in u's connection, then there is a world w', such that for no agent a, $P(a) \wedge S_a(u')$. In such a world $\zeta = \{b, a, S_a(c)) \mid a, b, c \in \mathcal{A} \wedge c \neq u' \wedge P(a)\}$. Therefore, $\mathcal{K}, w' \not\models K_u T_\zeta^{I^b}(u')$.

Therefore, there is an agent a, such that there is a summand $P(a)$ and a summand $a.S_a(u')$. Hence, $observes_u(P(a))$ and $observes_u(\boldsymbol{C}_a^u(S_a(u'))$, and u trusts correctness of the public key of u'.

4.2 Hierarchical Public Key Infrastructure

An N hierarchical public key key infrastructure is a generalization of the flat public key infrastructure. There are users and certifying authorities. A certifying authority can sign certificates of users and extend trusts. A user u trusts another user u' when the certificate of u' has been signed by some trusted certifying authority. There is a special type of certifying authority, called a *root certifier*. All root certifiers are trusted by all agents. If there is a chain, with length at most N, of trust extensions from a root certifier to another certifying authority c, then every user trusts c. We note that a 0 hierarchical public key infrastructure is, in fact, equivalent to the flat public key infrastructure. When we want to describe a hierarchical public key infrastructure, where we do not care about the length of a chain, we simply pick N greater than the number of certifying authorities.

Let U be a set of users. Let C be a disjunct set of certifying authorities, and C_{root} be the subset representing root certifiers. Root certifiers are trusted by every user in the system in their role as certifying authority. Let $H(c, c')$ denote that certifying authority c transfers part of its authority to certifier c'. For notational simplicity, we assume that $H(c, c)$. Let $S(c, u)$ denote that certifying authority c has signed a certificate of user u. Again, $observes_u(\varphi)$ means that u observes φ. We define that for $n > 0$, $D_n(u, c)$ holds for user u and certifying authority c, if there is a certifying authority c', such that $D_{n-1}(u, c')$ and $observes_u(H(c, c'))$. And for $n = 0$, we define that $D_0(u, c)$ holds when $observes_u(c \in C_{root})$. As a consequence of reflexivity of H, if $D_{n-1}(u, c)$ then $D_n(u, c)$. Again, u trusts correctness of the public key of u' iff $\exists_{c \in C}(D_N(c, u'))$.

The set of agents $\mathcal{A}$ is equivalent to the union of the users and the certifiers. The set of predicates $\mathcal{P}$ is $\{P\} \cup \{S_a \mid a \in C\} \cup \{H_a \mid a \in C\}$. The predicate $P(a)$ is true when a is a root certifier, $a \in C_{root}$. The predicate $H_a(b)$ is true when $H(a, b)$, thus when a extends his trust to b. The predicate $S_a(b)$ is true when $S(a, b)$, thus when a signed certificate of b. Then we need to define the network. The connection of a user a contains precisely the observations of a. In other words, a has a connection $\sum_{b \in \mathcal{A}, observes_a(P(b))} P(b) + \sum_{b,c \in \mathcal{A}, observes_a(S(b,c))} b.S_b(c) + \sum_{b,c \in \mathcal{A}, observes_a(H(b,c))} b.H_b(c)$.

An agent a trusts an agent b, if there is a trusted certifying authority that claims to have signed b's certificate. The first instance of "trust", namely trusting a certifying authority, differs from the second, namely trusting a user. We refer, in this context, to the first instance as intermediate trust, and the second instance as final trust. There are multiple types of intermediate trust, depending on their chain length. We define intermediate trust with a chain length only pertaining direct connections.

$$T_{\emptyset}^{I_0^i}(a)$$

where:

$$I_0^i(a) = P(a)$$

and:

$$\zeta_0 = \emptyset$$

We define intermediate trust with longer chain lengths $1 \leq n \leq N$:

$$T^{I^i_n}_{\zeta_n}(a)$$

where:

$$I^i_n(a) = \bigvee_{b \in \mathcal{A}} (H_b(a) \wedge T^{I^i_{n-1}}_{\zeta_{n-1}}(b))$$

and:

$$\zeta_n = \{(a, b, H_b(c)) \mid a, b, c \in \mathcal{A} \wedge T^{I^i_{n-1}}_{\zeta_{n-1}}(b)\}$$

Using intermediate trust, final trust can easily be defined.

$$T^{I^f}_{\zeta'}(a)$$

where:

$$I^f(a) = \bigvee_{b \in \mathcal{A}} (S_b(a) \wedge T^{I^i_N}_{\zeta_N}(b))$$

and:

$$\zeta' = \{(a, b, S_b(c)) \mid a, b, c \in \mathcal{A} \wedge T^{I^i_N}_{\zeta_N}(b)\}$$

Lemma 2. *User u trusts correctness of the public key of u', iff $\mathcal{K}, w \models K_u T^{I^f}_{\zeta'}(u')$.*

Proof. This proof is highly similar to that of Lemma 1. The forward side, soundness, is a simple extension of the flat case. Hence we only mention the completeness case, where we apply modus tollens as in Lemma 1.

If there is no agent a such that there is a summand $P(a)$, then there is a world w', such that for no agent a, $P(a)$. In such a world $\zeta_0 = \emptyset$, and clearly $T^{I^i_0}_{\emptyset}(u')$ does not hold, since there is no a with $P(a)$. As a consequence $T^{I^i_N}_{\zeta_N}(u')$ also does not hold, and neither does $T^{I^f}_{\zeta'}(u')$. Hence, we can assume that there is an agent a, such that there is a summand $P(a)$. By applying the possibility relation, we see that for all worlds w', $w'_{\mathcal{P}}(P, a)$. Therefore $\zeta_1 = \{(b, a, S_a(c)) \mid a, b, c \in \mathcal{A} \wedge P(a)\}$.

For all $1 \leq i \leq N$. We know that agents a exist such that $T^{I^i_{i-1}}_{\zeta_{i-1}}(a)$. If there is no agent a such that $T^{I^i_{i-1}}_{\zeta_{i-1}}(a)$ and there is no summand $a.H_a(b)$, then there is a world w', such that for no agent a, $H_a(b) \wedge T^{I^i_{i-1}}_{\zeta_{i-1}}(a)$. Such a world, however, cannot exist. Therefore, there must be an agent a with $T^{I^i_{i-1}}_{\zeta_{i-1}}(a)$, such that there is a summand $a.H_a(b)$.

Lastly, we know that agents a exist such that $T^{I^i_N}_{\zeta_N}(a)$. If there is no agent a such that $T^{I^i_N}_{\zeta_N}(a)$ and there is no summand $a.S_a(b)$, then there is a world w', such that for no agent a, $S_a(b) \wedge T^{I^i_N}_{\zeta_N}(a)$. Again, such a world must exist. Therefore, there must be an agent a with $T^{I^i_N}_{\zeta_N}(a)$, such that there is a summand $a.S_a(b)$.

We have derived in the last three paragraphs that there must exist a set of agents L, such that (per paragraph): There is an element $L_0 \in L$, where $w_{\mathcal{N}}(u)$ has a summand $P(L_0)$. There are intermediate agents L_i in L, such that $w_{\mathcal{N}}(u)$ has $L_i.H_{L_i}(L_{i+1})$. There is an element $L_N \in L$, where $w_{\mathcal{N}}(u)$ has a summand $S_{L_N}(u')$.

We can apply the definition of D to see that u trusts correctness of the public key of u'.

4.3 Web of Trust

An N Web of Trust is a variation of the hierarchical public key infrastructure. There are only users. Users can sign certificates, and by doing that, extend trust. A user u trusts another user u' when there is a chain of certificates from u to u', with length at most N. We note that a 0 web of trust simply means that users only trust certificates that they signed themselves. When we want to describe a web of trust with unlimited chaining, we simply pick N greater than the number of users.

We define the following model of a web of trust more formally: Let U be a set of users. Let N be an integer, representing the chain length threshold. Let $E(a, b)$ denote that agent a extends his trust to b. Let $D_n(a, b)$ denote that there is a chain with length at most n of trust-extensions from a to b. More formally: $D_0(a, b) = E(a, b)$, and $D_n(a, b) = (E(a, b) \vee \exists_{c \in \mathcal{A}}(E(a, c) \wedge D_{n-1}(c, b)))$. Agent a trusts agent b if and only if it observes $D_n(a, b)$.

The set of agents $\mathcal{A}$ is equivalent to the set of users U. The set of predicates $\mathcal{P}$ is $\{E_a \mid a \in \mathcal{A}\}$. The predicate $E_a(b)$ is true when $E(a, b)$, thus when a extends his trust to b. We assume that $E_a(a)$, so that we can lengthen a chain without breaking it. Then we need to define the network. The connection of a user a will contain precisely the observations of a. In other words, a has a connection $\sum_{b \in \mathcal{A}, observes_a(E(a,b))} E_a(b) + \sum_{b,c \in \mathcal{A}, observes_a(E(b,c))} b.E_b(c)$.

An agent a trusts another agent b, if there is a list L of users of size $n \leq N$, such that $a = L_0$, $b = L_n$ and for all $0 \leq i < n$, $E(L_i, L_{i+1})$. For all $0 \leq n \leq N$:

$$T_{\zeta_n}^{I_n^t}(a)$$

where:

$$I_n^t(a) = \bigvee_{b \in \mathcal{A}} (E_b(a) \wedge T_{\zeta_{n-1}}^{I_{n-1}^t}(b))$$

and:

$$\zeta_n = \{(a, b, E_b(c)) \mid a, b, c \in \mathcal{A} \wedge T_{\zeta_{n-1}}^{I_{n-1}^t}(b)\}$$

Lemma 3. *Let u and u' be users. Let there be a transitive interpretation, as defined above. User u trusts correctness of the public key of u', iff $\mathcal{K}, w \models K_u T_{\zeta_N}^{I_N^t}(u')$.*

Proof. Since I^t is essentially the same as I^i, this is a straightforward adaptation from Lemma 2.

5 Conclusion

A generalized semantics is given, requiring observations, in the form of connections, and interpretations, in the form of objective predicates. We formalized connections, by providing a clear syntax and semantics. We showed how interpretations of trust can act transitively, using examples. We asserted that if an agent knows that a target is trustworthy, he will trust the target.

We studied three particular examples, namely flat and hierarchical public key infrastructures, and PGP's web of trust. In all of them, we have instantiated the network in a way that we found natural. We gave the interpretations of trust as a straightforward translation, from natural language, to predicate logic. Then we proceeded to prove equivalence between our model and the existing model. We have hence successfully applied our approach to these three examples.

There is an important aspect that has not been discussed in this paper. The question whether the generalized semantics can be completely axiomatized. Due to the work on algebraization of predicate logic in [14], we expect that this is possible. Therefore, we will proceed to research such axiomatizations. They should axiomatize the core of all trust systems. Furthermore, it should be researched whether there is a way to generate axiomatizations given interpretations. If a standard way exists to axiomatize interpretations, then the axiomatizing existing systems is reduced to formalizing existing models in our formalism.

Another interesting exercise would be to formalize more complex models, that do not have a simple and concise logical description. If there is a translation into our formalism, proving equivalence to the original might not be feasible for such systems. We are interested in formalizing simple recommender systems. Given that the number of possible ratings is finite, we should be able to describe the real world of recommender systems in our language. Our intuition is therefore that formalizing recommender systems should be possible in our formalism.

An interesting existing formalism is subjective logic, as defined by Jøsang in [8]. In subjective logic, reasoning is done over opinions. It is assumed that these opinions are independent. This leads to the lack of idempotency of fusion in subjective logic. In subjective logic, fusing two identical opinions means fusing two different, independent opinions that happen to have the same value. Hence, we suspect that our methodology can be combined with subjective logic in a sequential way. Assume there is a way to transform arbitrary connection expressions to a form with independent fusion. Then we can input any network, reformulate it in an equivalent network, with only independent fusion, and apply subjective logic.

Finally, it is worth studying probability instead of possibility. Every world will have a certain probability, depending on its valuation. Interpretations will no longer be black and white issues, but assign opinions to situations. We expect that this approach will lead to subjective logic, when all information is independent. If it is equivalent to subjective logic, that would strongly support both the probability approach and subjective logic. Furthermore, the probability approach should also be able to fuse opinions with overlapping information properly, hence extending subjective logic.

Acknowledgements

Thanks to Sjouke Mauw and Baptiste Alcalde, for their input and steering. Thanks to Wojtek Jamroga for his advice regarding semantics of our model.

References

1. Baier, A.: Trust and antitrust. Ethics **96**(2) (1986) 231
2. Sobel, J.: Can we trust social capital? Journal of Economic Literature **40**(1) (March 2002) 139–154
3. Lewis, J.D., Weigert, A.: Trust as a Social Reality. Social Forces **63**(4) (1985) 967–985
4. Barney, J., Hanson, M.: Trustworthiness as a source of competitive advantage. Long Range Planning **28** (August 1995) 127–127(1)
5. Staab, E., Fusenig, V., Engel, T.: Towards trust-based acquisition of unverifiable information. In: Cooperative Information Agents XII. Volume 5180 of LNCS., Springer Verlag (2008) 41–54
6. Demolombe, R.: Reasoning about trust: A formal logical framework. In: Trust Management. Volume 2995 of Lecture Notes in Computer Science. Springer (2004) 291–303
7. Montaner, M., Lpez, B., de la Rosa, J.L.: A taxonomy of recommender agents on the internet. Artificial Intelligence Review **19** (2003) 285–330
8. Jøsang, A.: A logic for uncertain probabilities. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **9**(3) (2001) 279–311
9. Alcalde, B., Mauw, S.: An algebra for trust dilution and trust fusion. In: Formal Aspects in Security and Trust. Volume 5983 of Lecture Notes in Computer Science. Springer (2010) 4–20
10. Huth, M.R.A., Ryan, M.: Logic in computer science: modelling and reasoning about systems. Cambridge University Press, New York, NY, USA (2000)
11. Blackburn, P., Van Benthem, J., Wolter, F.: Handbook of Modal Logic. Springer (2006)
12. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: Reasoning About Knowledge. MIT Press, Cambridge, MA, USA (2003)
13. Baltag, A., Moss, L.S., Solecki, S.: The logic of public announcements, common knowledge, and private suspicions. Technical report, Amsterdam, The Netherlands, The Netherlands (1999)
14. Langel, J., Kohlas, J.: Algebraic structure of semantic information and questions. predicate logic: an information algebra. Technical Report 08-02, Department of Informatics, University of Fribourg (2008)