A Difficulty in Trust Modelling

No Author Given

No Institute Given

Abstract. There are various trust models, and no pair are the same. Why is there no consensus on what the best trust model is yet? In this paper, we prove some impossibility results, that suggests that no perfect trust model exists. We provide a general meta-model of trust models. A new concept, belief-consistence, is introduced. If a trust model is not belief-consistent, then it cannot model anyone's beliefs. We show that trust models without uncertainty can only be belief-consistent is an extremely simple case. Moreover, full trust models cannot be belief-consistent, even with uncertainty. Finally, we show how any discrepancies in belief can be blown up into arbitrary errors.

1 Introduction

Trust is an important tool for decision making. Although trust is a complicated concept, people have good (but imperfect) intuitions about trust. Computational trust concerns capturing the idea of trust in (numerical) values. There is no consensus on a general trust model for computational trust. In this paper, we provide several impossibility results. Notably, no trust model in which learning may occur can be consistent with a user's beliefs about other users' behaviour. Moreover, the gap between the model and the beliefs will be arbitrarily large.

The various impossibility results are quite technical in nature. We define the concepts relating to trust models in as general terms as possible (Section 2). In Sections 3-5, we prove all the technical results. These sections are important to verify the claims, as well as to understand the precise technical meaning. However, a reader who is only interested in the meaning and implications of our impossibility results can skip ahead to Section 6.

There are various trust models, with various assumptions. In our paper, we are interested in model with both *fusion* (learning from new evidence) and *chaining* (rating/reporting/transitive trust). A leading trust model that supports various operations, including fusion and chaining, is Subjective Logic [6, 7]. There are various progressively more powerful flavours of SL, but each idea can be expressed in our meta-model. There are models attempting to model the inner state of agents, typically with HMMs, such as [14, 3]. Bayesian models (e.g. [13, 1]) quite naturally capture the idea of learning from evidence. Models could be based on machine learning [4]. Finally, the Beta model [5] is an example of a trust model that relies on taking a distribution over a parameter as a trust opinion. Our meta-model must be able to capture these various trust models.

A core technique to obtaining our results is the introduction of the notion of *belief-consistency.* Users (agents) have some unknown behaviour – deterministic, stochastic, strategic, Markovian, or otherwise. If one were to know the behaviour of a user, then they would know the probabilities that the user would act in a specific way, in every context. Although the user does not know the behaviour of the other user, they do have some beliefs about the likelihoods of the behaviours. If the other user acts in a way that is highly improbable for a behaviour, then the relative likelihood of that behaviour should go down in the observer's belief, proportionally. Similarly, for acts that are probable in a behaviour, the relative likelihood of that behaviour must increase proportionally. The trust values of the corresponding trust opinions must reflect this change in beliefs when fusing opinions.

The structure of the paper is as follows: Section 2 introduces the formalism and the concepts that allow us to reason about trust models in general. Section 3 applies the new techniques on a partial trust model with only fusion, where there is no notion of uncertainty, to prove the first impossibility result. Then we introduce uncertainty in Section 4, and prove the more general impossibility result for full trust models. In Section 5, we show that minute errors in a trust model can have arbitrarily bad consequences, by providing a network that can exacerbate any differences. Everything is finally put into perspective in Section 6. Section 6 can be read without the technical details necessary for the proofs.

2 Trust Model

In this section, we provide a meta-model for trust. We try to keep the technicalities to a minimum. For that reason, we define some mathematical notion as a shortcut. The parameters involved may be discrete or continuous, so the notionation is chosen to cover both types of variables. $\varDelta(x)$ denotes a distribution of x – this could be probability mass or probability density. Similarly, we define $\mathbb{E}_f(x)$ as the expected value of $f(x)$ – which may be a sum or integral. We can write $f(\mathbf{x})$ to mean $\sum_{x \in \mathbf{x}} f(x)$ or $\int_{\mathbf{x}} f(x)dx$, depending on whether the set $\mathbf{x}$ is discrete or continuous. Finally, if $f : Y \to (X \to Z)$, then we write $f(x; y)$ instead of $f(y)(x)$.

2.1 Trust Modelling

There are two major types of trust models: Descriptive trust models, intended to faithfully model people's level of trust, and prescriptive trust models, intended to model what the level of trust should be. The difference becomes apparent when people's (dis)trust is misguided. The former type of model is interesting if we want to understand human behaviour, and an application could be trustworthy AI. The latter type of model is interesting for optimising decision-making based on trust, with computer security being an important application. Throughout the paper, we consider the latter type.

In this section, we attempt to capture prescriptive trust models in the most general sense. In order to derive our results, we must still make some assertions. Trust involves agents; people or devices with some unknown behaviour. The behaviour can be strategic, stochastic, Markovian or other. It assigns probability to actions in given contexts. Past actions are a form of evidence that can be used to establish trust. Another form of evidence is second-hand, where another agent reports their opinion. The trust model uses trust opinions to model the trust arising from the evidence. The trust model supports two operations, fusion (adding evidence) and chaining (reporting opinions). We will more formally define these concepts in the remainder of this section.

In the remainder of the paper, we refer to agents as users, although our results remain true for non-human agents as well. We define the set of users $\mathcal{U} = \{u, u', u_0, \dots\}$. The users have the following set of actions available: $\mathcal{A} = \{a, a', a_0, \dots\}$. Their behaviour dictates their actions stochastically, depending on context. The context $c \in \mathcal{C}$ could encode their intentions, goals, or state, as well as external information like the topic or time. Hence, their behaviour can be defined as a function $b : \mathcal{C} \to \Delta(\mathcal{A})$, and we define $\mathcal{B} = \{b, b', b_0, \dots\}$ as the set of possible behaviours. The actual behaviour of a user is not known to the other users.

A trust opinion is abstractly defined as a set $\mathcal{O} = \{o, o', o_0, \dots\}$, with two operations $\_ + \_$ and $\_ \cdot \_$ as fusion and chaining, respectively. Within the trust model, the trust opinions are simply a collection of values, supporting fusion and chaining. However, the trust opinions are supposed to model an actual state of affairs: A trust opinion o is about an user, $\mathbf{u}(o) \in \mathcal{U}$, acting positively; $\mathbf{a}^+ \subseteq \mathcal{A}$ is the set of actions considered positive. The value $\mathbf{t}_c(o)$ represents the trust value of a trust opinion $o \in \mathcal{O}$ in a given context $c \in \mathcal{C}$ – how probable the holder of o thinks it is that user $\mathbf{u}(o)$ performs an action in $\mathbf{a}^+$, given context c.

A trust network is a collection of trust opinions that can be described using fusion and chaining on these trust opinions (more information on trust networks can be found in [7]). In particular, a trust network is a series-parallel graph, where serial composition is represented by chaining, and parallel composition by fusion. The depicted graph in Figure 1, for example, has the corresponding formula $o_1 \cdot (o_2 + o_3 + o_4 \cdot o_5) + o_6$. Observe that we assume $\_ \cdot \_$ binds stronger than $\_ + \_$. Further, fusion can be proven to be associative, so no parenthesis are needed for larger sums, but the same is not true for chaining [10].

The source of the network represents the user forming a trust opinion, and all outgoing arrows are their direct observations. This is a type of evidence $e \in \mathcal{E}$, consisting of a user u' (to whom the arrow is pointing) performing a specific action a in a context c. The function $\mathbf{e}(o) \in \mathcal{E}$ provides the evidence that resulted in o. All incoming arrows to a user have the same context and definition of good actions. A claim is a different type of evidence, and is represented by a pair of opinions.

The importance of the opinion on the intermediate user in a trust chain, is that the opinion tells us whether the claim is factual. There is a context c (namely the context of them making a claim), such that $\mathbf{t}_c(o_0) = \mathbf{t}_c(o_1) \implies$

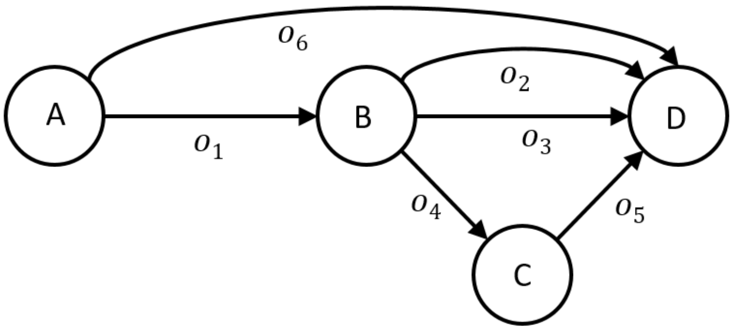

Fig. 1. An example trust network. A ultimately forms an opinion on D.

$o_0 \cdot o' = o_1 \cdot o'$. Furthermore, as $\mathbf{t}_c(o)$ goes to 1, $o \cdot o'$ goes to o', as a perfectly trustworthy user's claim can be treated as evidence.

There is a special trust opinion, μ, which represents the opinion resulting from no evidence. The opinion μ is a unit element with respect to fusion, as $o + \mu$ would represent having some opinion o, and updating it with no additional evidence. Although the order in which actions happen may be important (this would be represented in the context of the actions), the order in which one learns about the evidence is irrelevant. Therefore, fusion is associative, and furthermore it is commutative.

2.2 Belief-Consistency

While users' actual behaviour is unknown, users still have beliefs about what the other users' behaviour may be. The belief assigns a likelihood to each behaviour, so a belief is a function $\beta : \mathcal{B} \to \mathbb{R}^{\geq 0}$. Let $\mathscr{B}$ be the set of all potential beliefs. In this paper, we never assume what users' beliefs actually are. Rather we use the beliefs as a tool to show that no trust model can accurately reflect any beliefs the user may have.

The link between trust opinions and beliefs is through an interpretation. An interpretation is a function $\iota : \mathcal{O} \to \mathscr{B}$. This allows us to define *belief-consistency*. A trust model is belief-consistent iff exists an interpretation ι, such that:

- For all $o \in \mathcal{O}$, with $\beta = \iota(o)$, $\mathbf{t}_c(o) = \mathbb{E}_{\beta(b)}(b(\mathbf{a}^+); c))$.
- If $\mathbf{e}(o) = (u, a, c)$, $b_0(a; c) = x \cdot b_1(a; c)$, $\beta = \iota(o')$, and $\beta(b_0) = y \cdot \beta(b_1)$, then $\beta' = \iota(o + o')$ must satisfy $\beta'(b_0) = x \cdot y \cdot \beta'(b_1)$.
- With $\beta = \iota o + o'$, the value $\min_{b \in \mathcal{B}} \beta(b; c)$ is decreasingly monotonic with $\mathbf{t}_{c'}(o)$.

In other words, there is an interpretation such that 1) the expected probabilities of actions given the beliefs equal the trust value of o, 2) the belief that a user has a certain behaviour changes proportionally to the probability of them performing an observed action, and 3) an increased likelihood of all possible behaviours if the claim is decreasingly trustworthy. The second point allows to implicitly merge beliefs. One can verify that $\beta_0(b; c)$ and $\beta_1(b; c)$ can be merged by taking $\beta(b; c) = \frac{\beta_0(b;c)\cdot\beta_1(b;c)}{\beta_\mu(b;c)}$, where $\beta_\mu = \iota\mu$.

Finally, we will be referring to *non-trivial* trust models. Observe that it is perfectly allowed to have $|\mathcal{B}| = 1$ (i.e. $\mathcal{B} = \{b\}$), making $\iota(o) = \iota(\mu)$ for all $o \in \mathcal{O}$, since all $\beta \in \mathscr{B}$ must be $\beta(b) = x$, for a constant x. No learning occurs if there is only one behaviour that we consider, since regardless of the evidence, the behaviour must be the one that is considered. While such a model is trivially belief-consistent, the trust model is meaningless, as it does not reflect users' ability to learn to trust.

3 Models without Uncertainty

In this section, we investigate trust models that do not have a notion of uncertainty. In other words, the only value(s) that we have is/are the trust value(s). Recall that the trust value represents the trust opinion holder's belief that the relevant action will be positive.

For this section, the only operation we study is fusion. This means that the results obtained in this section are actually true for any (belief-consistent) model of information fusion, not just trust models. Note that an important example here is Dempster-Shafer theory [2, 12], with Dempster's rule of combination filling the role of fusion. Our impossibility results will demonstrate that a measure of uncertainty is necessary – this is mass not assigned to a single outcome. Dempster-Shafer theory and the link between Subjectiv Logic and Beta models both show that with uncertainty, there are belief-consistent models of just fusion.

Uncertainty values may be implicitly represented in some trust models (e.g. [11]). For example, take a trust system based on ratings, using the average rating as the trust value (e.g. Google reviews, see [8]). Such a system must keep track of the number of ratings to compute the average, and this value is used for fusion. A trust opinion could be a pair (r, n), and fusion would then be: $(r_0, n_0) + (r_1, n_1) = (\frac{r_0\cdot n_0 + r_1\cdot n_1}{n_0+n_1}, n_0 + n_1)$. The second value, n, is a measure of certainty, making its inverse a measure of uncertainty. Such a model is *not* an example of a model without uncertainty.

Specifically, by the model not having uncertainty, we mean that the fusion of two trust opinions must only depend on the trust values. A naive formalisation of this notion is that for a specific context c, **WNU** if $\mathbf{t}_c(o_0) = \mathbf{t}_c(o_1)$, then $\mathbf{t}_c(o_0 + o') = \mathbf{t}_c(o_1 + o')$. The intuition behind **WNU** is that there should not be a distinction when updating opinions, when they have the same trust value, as this is the only relevant value for that context. Below, we demonstrate that this formalisation is too weak, and provide an improved formalisation of the idea.

It is possible to construct a non-trivial belief-consistent trust model that has the property that, for arbitrary $c \in \mathcal{C}$, $\mathbf{t}_c(o_0) = \mathbf{t}_c(o_1) \implies \iota(o_0) = \iota(o_1)$. Formally:

Proposition 1. *If $|\mathcal{B}|$ is finite, then there is an interpretation ι, such that $\mathbf{t}_c(o_0) = \mathbf{t}_c(o_1) \implies \iota(o_0) = \iota(o_1)$.*

Proof. Let $B \subset \mathcal{B}$ be a finite set behaviours, such that for each behaviour $b \in B$, $b(\mathbf{a}^+) \in \mathbb{Q}$. Further, let the initial belief $\iota(\mu) = \beta$ be distributed so that, for all $b, b' \in B$, there is no $x \in \mathbb{Q}$, so that $\beta(b) \cdot x = \beta(b')$. Since every action has a rational probability of occurring in every behaviour, the likelihood of each behaviour can be expressed as a rational number, multiplied by the initial value. The expected value contains all the rationally independent values with a rational weight. Two expected values can only be equal, if all weights are equal, thus having the same relative likelihood. □

An example of rationally independent values could be: $\beta(b_0) = \sqrt{2}$, $\beta(b_1) = \sqrt{3}$, $\beta(b_2) = \sqrt{5}$, etc. So, for example, for any $o \in \mathcal{O}$ and $\beta' = \iota(o)$, $\beta'(b_2) = \frac{n \cdot \sqrt{5}}{m}$. Clearly, any finite $|\mathcal{B}|$ allows all $b \in \mathcal{B}$ to be given their own prime square root.

The corollary of Proposition 1 is that there exist non-trivial belief-consistent trust models following the rule that if $\mathbf{t}_c(o_0) = \mathbf{t}_c(o_1)$, then $\mathbf{t}_c(o_0 + o') = \mathbf{t}_c(o_1 + o')$. The reason is that we just showed that $\mathbf{t}_c(o_0) = \mathbf{t}_c(o_1)$ implies $\iota(o_0) = \iota(o_1)$, thus $\iota(o_0 + o') = \iota(o_1 + o')$, which implies $\mathbf{t}_c(o_0 + o') = \mathbf{t}_c(o_1 + o')$. We did not explicitly provide a model of o, but observe that, via Proposition 1, $\mathbf{t}_c(o)$ has a one-to-one relation with $\iota(o)$. Hence, $\mathbf{t}_c(o_0 + o_1)$ can be computed by translate $\mathbf{t}_c(o_0)$ and $\mathbf{t}_c(0_1)$ to $\iota(o_0)$ and $\iota(0_1)$, then taking $\iota(b; o_0 + o_1) = \frac{\iota(b;o_0) \cdot \iota(b;o_1)}{\iota(b;\mu)}$, and translating that back to $\mathbf{t}_c(o_0 + o_1)$. From this definition, it is possible for two values o_0 and o_1 to have roughly similar trust values, but for $o_0 + o'$ and $o_1 + o'$ to have completely different trust values. We are effectively encoding uncertainty in the insignificant digits.

A stronger, more appropriate formalisation of not having uncertainty, is to say that trust opinions cannot 'overtake' each other when fusing with the same opinion. In other words, **SNU** if $\mathbf{t}_c(o_0) \leq \mathbf{t}_c(o_1)$, then $\mathbf{t}_c(o_0 + o') \leq \mathbf{t}_c(o_1 + o')$. **SNU** is a natural axiom, if there is no notion of uncertainty. The tricky construction from before does not work for the stronger formalisation.

It turns out that if the set $\mathcal{B} \geq 2$, then the axiom must be broken. Intuitively, performing about as many good actions as bad actions, increases the belief in a balanced behaviour. However, a strong belief in balanced behaviour means that observing a single action is not swaying beliefs very much. Formally:

Theorem 1. *If $|\mathcal{B}| > 2$, then there is no belief-consistent trust model that satisfies **SNU**.*

Proof. By way of contradiction, let ι be a belief-consistent trust model. Partition $\mathcal{B}$, into B_{low}, B_{high} and B_{mid}, such that for all $b_0 \in B_{low}$, $b_1 \in B_{mid}$, $b_2 \in B_{high}$, $b_0(\mathbf{a}^+) < b_1(\mathbf{a}^+) < b_2(\mathbf{a}^+)$. We say $\iota(B; o)$ is negligible if $\iota(B; o) < \epsilon \cdot \iota(\mathcal{B}; o)$ – less than a fraction ϵ of total likelihood. We pick ϵ sufficiently small that $\iota(B_{low}; \mu), \iota(B_{mid}; \mu), \iota(B_{high}; \mu)$ are all non-negligible. There exists a trust opinion o_{mid}, such that $\iota(B_{low}; o_{mid}) + \iota(B_{high}; o_{mid}) < \epsilon^2 \cdot \iota(\mathcal{B}; o_{mid})$. The evidence is a sufficiently large collection of evidence containing both positive actions ($a \in \mathbf{a}^+$) and negative ones ($a \notin \mathbf{a}^+$). $\iota(\mu + o_|)$ must satisfy that

$\iota(B_{high}; \mu + o_{mid}) = x^n \cdot y^m \cdot \iota(B_{high}; \mu)$, where x is close to 0, y close to 1 and n, m are very large. This term is smaller than ϵ^2, for sufficiently large n. Similarly for $\iota(B_{low}; \mu + o_|)$, for sufficiently large m.

If $\mathbf{t}_c(\mu) \leq \mathbf{t}_c(o_{mid})$, then consider the opinion o_{high}, such that $\iota(B_{low}; o_{high}) + \iota(B_{mid}; o_{high}) < \epsilon \cdot \iota(\mathcal{B}; o_{high})$ and $\iota(B_{mid}; o_{high}) > \epsilon^2 \cdot \iota(\mathcal{B}; o_{high})$. The opinion o_{high} follows from sufficiently many $a \in \mathbf{a}^+$. Now $\mathbf{t}_c(\mu + o_{high})$ is a value in the high range, since all behaviours is the low and mid range have negligible belief. Whereas $\mathbf{t}_c(o_{mid} + o_{high})$ is a value in the mid range, since ϵ^2 is negligible compared to ϵ. But this implies $\mathbf{t}_c(\mu) \leq \mathbf{t}_c(o_{mid})$, yet $\mathbf{t}_c(\mu + o_{high}) > \mathbf{t}_c(o_{mid} + o_{high})$.

If $\mathbf{t}_c(\mu) < \mathbf{t}_c(o_{mid})$, then we can apply the same argument, but with high/low swapped, making $\mathbf{t}_c(o_{mid}) \leq \mathbf{t}_c(\mu)$, yet $\mathbf{t}_c(o_{mid} + o_{low}) > \mathbf{t}_c(\mu + o_{low})$. □

The intuition behind the proof is quite simple. If we have a large amount of evidence, then a new piece of evidence will not sway the opinion much. The opinion that sways the most, is total uncertainty (μ). While this intuition is clear, the proof is quite technical. The advantage is that the proof is extremely general, and works for all sorts of trust models. However, the theorem assumed $\mathcal{B} > 2$. The reason is that we needed to partition $\mathcal{B}$ into 3 non-empty sets. This technicality is not-at-all obvious from the intuition, but turns out to be crucially important!

The converse of Theorem 1 turns out to be true as well. If $|\mathcal{B}| \leq 2$, then it is possible to have a belief-consistent trust model. Of course, for $|\mathcal{B}| = 1$, this is trivially true. However, for $|\mathcal{B}| = 2$, there is a non-trivial belief-consistent trust model for fusion, without uncertainty.

Theorem 2. *There is a non-trivial belief-consistent trust model that satisfies* ***SNU****, but only if* $|\mathcal{B}| = 2$

Proof. Let there be two behaviours, $\{b_0, b_1\} = \mathcal{B}$ such that $0 < b_0(a; c) \neq b_1(a; c) < 1$ for all $a \in \mathcal{A}, c \in \mathcal{C}$. Take $r = b_0(\mathbf{a}^+; c)$ and $s = b_1(\mathbf{a}^+; c)$, with $r < s$. Then $\mathbf{t}_c(o) \propto r \cdot \iota(b_0; o) + s \cdot \iota(b_1; o)$. If $\mathbf{t}_c(o_0) \leq \mathbf{t}_c(o_1)$, then $\frac{\iota(b_1;o_0)}{\iota(b_0,o_0)} \leq \frac{\iota(b_1;o_1)}{\iota(b_0,o_1)}$, and $\mathbf{t}_c(o_0 + o') = x \cdot \frac{\iota(b_1;o_0)}{\iota(b_0,o_0)} \leq x \cdot \frac{\iota(b_1;o_1)}{\iota(b_0,o_1)} = \mathbf{t}_c(o_1 + o')$, where x is the constant relative belief in behaviour b_1 over b_0 in $\iota(o')$. □

The intuition is that, if there are only two possible behaviours, then the trust value represents the relative belief in one behaviour over the other. There is a one-to-one correspondence between belief in one behaviour and the trust value. Fusing with the same opinion leads to an identical relative shift in belief. The corollary of Theorem 2 is that it is possible to do information fusion without any notion of uncertainty, if the underlying belief is sufficiently simple. An example of such simple underlying belief would be: agents are either good, and act positively 90% of the time, or they are malicious and act negatively 90% of the time.

4 Models with Uncertainty

Trust models almost always have a measure of uncertainty. This can be explicit, like in Subjective Logic [7] or here [9]. But often this is implicit, when taking average scores or ratings, the inverse of the total number of scores is a measure of uncertainty.

In the previous section, we have seen that without uncertainty, even just fusion is not a possible operation. This strongly suggests that a full trust model should at least have a notion of uncertainty. Unfortunately, we prove in this section that even uncertainty is not sufficient to have a full, non-trivial, belief-consistent trust model.

As before, we need to agree on the basic notion of uncertainty first. First, the opinion based on no evidence has maximum uncertainty (**MU**): $o \neq \mu \implies \mathbf{u}_c(\mu) > \mathbf{u}_c(o)$. Secondly, uncertainty goes down, when fusing opinions (**FU**): $\mathbf{u}_c(o_0 + o_1) \leq \mathbf{u}_c(o_1)$. Thirdly, chaining introduces some uncertainty (**CUB**): $\mathbf{u}_c(\mu) \geq \mathbf{u}_c(o_0 \cdot o_1) > \mathbf{u}_c(o_1)$. Fourthly, the more trust in the user making the claim (w.r.t. the context c' of them making the claim), the less uncertainty (**CUC**): $\mathbf{t}_{c'}(o_0) \leq \mathbf{t}_{c'}(o_1) \implies \mathbf{u}_c(o_0 \cdot o') \geq \mathbf{u}_c(o_1 \cdot o')$. And finally, the uncertainty of a chain scales with the claim (**CUS**): $\mathbf{u}_c(o_0) \leq \mathbf{u}_c(o_1) \implies \mathbf{u}_c(o' \cdot o_0) \leq \mathbf{u}_c(o' \cdot o_1)$.

To put the axioms we impose on uncertainty in perspective, we prove that there exist beliefs, such that when merged together, they form the beliefs of μ – total uncertainty.

Proposition 2. *For any β_μ, there exist β_0 and β_1, such that $\beta_\mu(b;c) \propto \frac{\beta_0(b;c) \cdot \beta_1(b;c)}{\beta_\mu(b;c)}$*

Proof. Partition $\mathcal{B}$ into B_0 and B_1. Set $\beta_0(b;c) = \beta_\mu(b;c)^2$, if $b \in B_0$ and $\beta_0(b;c) = 1$, if $b \notin B_0$. Similarly, set $\beta_1(b;c) = \beta_\mu(b;c)^2$, if $b \in B_1$ and $\beta_1(b;c) = 1$, if $b \notin B_1$. Then $\frac{\beta_0(b;c) \cdot \beta_1(b;c)}{\beta_\mu(b;c)} = \frac{\beta_\mu(b;c)^2 \cdot 1}{\beta_\mu(b;c)} = \beta_\mu(b;c)$. □

The choice of β_0 and β_1 was fairly arbitrary, and there are many pairs of beliefs that have the property that they merge to the belief corresponding to total uncertainty. If there exists o_0 and o_1 such that $\iota(o_0) = \beta_0$ and $\iota(o_1) = \beta_1$, then either **MU** or **FU** must not hold, since **MU** states that $\mathbf{u}_c(\mu) > \mathbf{u}_c(o_0)$, but then **FU** states that $\mathbf{u}_c(\mu + o_1) \geq \mathbf{u}_c(o_0 + o_1)$, but then $\mathbf{u}_c(o_1) = \mathbf{u}_c(\mu + o_1) \geq \mathbf{u}_c(o_0 + o_1) = \mathbf{u}_c(\mu)$ which contradicts **MU**, which states $\mathbf{u}_c(\mu) > \mathbf{u}_c(o_1)$. Note that this proof did not refer to chaining – only to fusion. We know that there are trust models with fusion in which the axioms **FU** and **MU** are both true. This is because no opinions o_0 and o_1 can ever be reached via fusion alone.

Unfortunately, we can adopt the sort of argument presented in Proposition 2 to work for opinions that can be reached via chaining. Rather than relying on two defined beliefs, the argument relies on a limit of a growing fusion of chains. Nevertheless, the conclusion is that The formal statement:

Theorem 3. *There is no full, non-trivial, belief-consistent trust model, that where uncertainty follows **MU**, **FU**, **CUC** and **CUS**.*

Proof (sketch). The proof relies on a limit argument. Let σ_n be the set of all sequences of actions of length n (so $|\sigma_n| = |\mathcal{A}|^n$). For each $0 \leq i < |\mathcal{A}|^n$, the opinion o_i is the opinion whose evidence is the sequence of actions $(a_0, \ldots, a_{n-1})$ in a context c. The opinions $o'_0, o'_1, \ldots, o'_{|\mathcal{A}|^n - 1}$ are chosen opinions with high

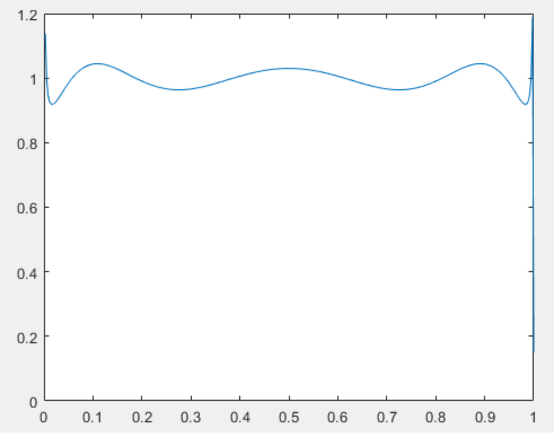

Fig. 2. An illustration of a curve that approaches μ (1) as $n \to \infty$, using $n = 9$

$\mathbf{t}_{c'}(o')$ (where c' is the context of providing a good claim). We can then take the limit of $o'_0 \cdot o_0 + o'_1 \cdot o_1 + \cdots + o'_{|\mathcal{A}|^n-1} \cdot o_{|\mathcal{A}|^n-1}$, as $n \to \infty$. With the right choices for o'_i, the limit of this expression is μ.

If there are finitely many behaviours, then, as n goes to infinity, each sequence of actions has only one state with non-negligible likelihood. By changing o'_i, we can increase or decrease the likelihood of any specific behaviour in the final belief, thus allowing us to have the final belief match μ. Since each likelihood function $\iota(o'_i \cdot o_i)$ must exceed some value ϵ, there is no behaviour that cannot be pumped up by increasing the appropriate o'_i.

The argument for infinite sets of behaviours $\mathcal{B}$ is essentially the same. For each n, we can only adjust finitely many points. It is important to recall that all parameters are continuous or discrete. Therefore the function we obtain in each step is also continuous. Further, any behaviour close to a behaviour corresponding exactly to a sequence of actions will actually still have a very high relative likelihood of generating those actions. That means that the function is smooth, and that every point converges to the desired distribution.

A fusion of lots of opinions, each of which has less uncertainty than μ, results in μ. We reach a similar contradiction to Proposition 2. □

We underspecified trust chaining, and demonstrated that for no implementation of trust chaining, we can avoid the impossibility result. As a result, unfortunately, the proof of Theorem 3 is not constructive. It will not be possible to explicitly provide a mechanism that computes the o'_i, without having more specification on the chaining operation. To illustrate the proof, however, Figure 2 shows how the trick from the proof works for some specific implementation of chaining. In particular, we use the extended Beta model [5, 10] as the trust model. The implementation of chaining used is $\iota(\beta; o \cdot o') = \mathbf{t}_c(o) \cdot \iota(\beta; o') + (1 - \mathbf{t}_c(o)) \cdot \iota(\beta; \mu)$. Furthermore, $\mu = 1$ – the uniform distribution. We see in Figure 2 that the line

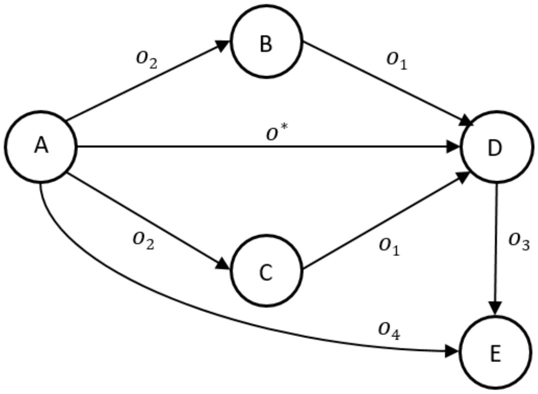

Fig. 3. Diagram of the trust network construction from Theorem 4

is tending to a straight line on 1. The figure uses $n = 9$ – the curve increasingly looks like $f(x) = 1$ as n increases.

As with Theorem 1, Theorem 3 has a straightforward intuition: If everyone is saying something different, then we are back to square one, and we have total uncertainty. But here again, the technical details do matter! We relied on the axioms of uncertainty for the contradiction, as a combination of operations that should lower uncertainty resulted in maximal uncertainty. Here, we used properties 2) and 3) of belief-consistency. The proof works for any number of behaviours higher than 1.

5 Approximate Trust Models

From Section 3, we conclude that any reasonable trust model must have some notion of uncertainty, otherwise even information fusion is not possible. From Section 4, we then conclude that no full non-trivial belief-consistent trust model exists even with uncertainty. However, practical examples of trust models with uncertainty exists. Subjective Logic [7] is a particularly expressive trust model, with several different variations and extensions. Although we have shown that all incarnations of Subjective Logic must not be belief-consistent, it may still mimic our beliefs closely.

In this section, we provide a construction in the form of a trust network, where substituting a trust opinion o for a highly similar trust opinion o' results in a radically different final trust opinion. The implication of this result is that the error of any trust model is unbounded, with respect to any interpretation. In other words, not only is there no non-trivial belief-consistent trust model, any non-trivial trust model deviates arbitrarily from the beliefs.

Subtly different interpretations of opinions can lead to radically different opinions in a trust network. Formally:

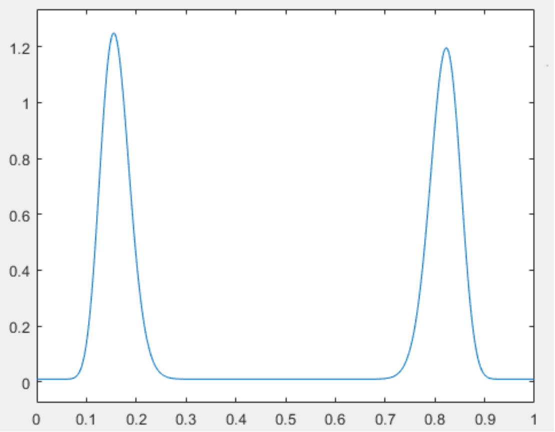

Fig. 4. A Beta model-based belief distribution of $o_2 \cdot o_0 + o_2 \cdot o_1$

Theorem 4. *There exists a trust network, such that changing the interpretation of one single opinion from β to β' results in a totally different opinion.*

Proof. Take $\iota(o^*) = \beta$ and $\iota(o^*) = \beta'$ as (subtly) different interpretations of o^*. Say $\beta(\mathbf{b}; c) > \beta'(\mathbf{b}; c)$ for the interval $\mathbf{b}$, but $\beta(\mathbf{b}'; c)' < \beta'(\mathbf{b}'; c)$ for the interval $\mathbf{b}$. Let o_0 be the opinion based on a large sequence of actions generated by $b \in \mathbf{b}$ and o_1 based on $b' \in \mathbf{b}'$. Let o_2 be an opinion so that $\mathbf{t}_c(o_2)$ is close to 1 for the context of making a claim. Let o_3 be an overwhelmingly positive opinion, and o_4 very negative. Take $((o_2 \cdot o_0 + o_2 \cdot o_1) + o^*) \cdot o_3 + 0_4$. Only behaviours in the crucial part of o^* are non-negligible in $(o_2 \cdot o_0 + o_2 \cdot o_1)$ by design. Fusion with o^* exaggerates the different, as the resulting graph will only have the behaviours where they differ as non-negligible. The value $\mathbf{t}_c((o_2 \cdot o_0 + o_2 \cdot o_1) + o^*))$ is higher for β than for β', so with balancing, in the former case, our opinion will be close to o_3 and close to o_4 in the latter. □

The construction is shown in Figure 3. The ABCD diamond blows up the difference between the two interpretations of o^*. To illustrate how the diamond blows up the difference, an example model in the shape of a Beta model is provided. If o^* differs around $0.1 - 0.2$ and $0.77 - 0.87$, then multiplying o^* with the diagram in Figure 4 will exaggerate the differences.

6 Discussion

In this section, the results are discussed on a non-technical level. We will reference the relevant theorem by name, so that the reader can draw their own conclusions as well. As our results were formulated as generally as possible, the proofs were extremely technical, even if the intuition was obvious. However, we argue that "obvious" intuitions may not be true, so a technical proof is required.

In Section 2, we introduce the trust model and the notion of of belief-consistency. A trust model is a computational model, in which trust opinions are represented as values, with which operations can be done. In particular, the two operations are fusion and chaining. Fusion is the idea that we can learn more evidence, and update our opinions. Chaining is the idea that users can provide reports/recommendations/ratings to one another.

Our goal was to have a prescriptive trust model. So the fusion or chaining of two opinions must be logically coherent. However, there are lots of unknowns, so we cannot tell if an opinion is wrong, Users have beliefs, and there should be a mapping between the trust opinions in the model and the beliefs of the users. We do not know what the user's beliefs are, but we do know that these beliefs must be internally consistent (in a prescriptive model).

If a trust model is not belief-consistent, then some trust opinions are guaranteed to be wrong. Various trust opinions in the trust model cannot all be true, as they model different realities. So while we do not model the actual beliefs of any user, we have proven that trust models cannot capture their beliefs consistently.

The first impossibility result (Theorem 1) is that any belief-consistent trust model without uncertainty can have at most one of two behaviours. Basically, the trust value represents the relative likelihood of one behaviour over the other. These are extremely simple models, where users are (for example) just good or bad, and both types of users have a fixed stochastic behaviour, Such a model is theoretically interesting, but fails to model real-life agents appropriately. The probability for a "good" user to act positively must not be 1, since then, observing negative actions leads to a contradiction. However, if the probability for a good user to act positively is some value $v < 1$, then our trust value can never exceed (or even reach) v, regardless of the number of positive actions we observe. It is commonly accepted that uncertainty is a useful tool for trust modelling, and Section 3 supports this notion.

An issue encountered in Section 3, was that uncertainty could be encoded into the precise trust values. Proposition 1 shows the mechanism to accomplish this. This is a more general problem, in the sense that multiple (up to countably infinitely many) values can be encoded in a single real number between 0 and 1, for example, by interleaving their digits. There are rules that we need to specify what counts as a trust value, to prevent "cheating" like in Proposition 1. The rule for uncertainty-free trust values is that fusion with a constant must be monotonic.

The second impossibility result (Theorem 3) is our main result. It shows that even with uncertainty values, there is no belief-consistent trust model (except the trivial one). The intuition behind the proof is that if lots of people all say different things, then we know nothing.

The set of trust opinions that can be obtained via fusion is actually extremely limited compared to the potential set of beliefs. The desirable rule that more evidence leads to less uncertainty is not generally true outside this limited set (Proposition 2). Trust chaining is an operation that typically moves outside of this limited set. From this, we conclude that the combination of fusion and chaining does, unfortunately, not lend itself to being captured in a trust model. This is the titular difficulty in trust modelling.

A possible way to rescue the existence of belief-consistent trust models, is to reject one of the axioms of uncertainty. The candidate axiom that lends itself most to being questioned is axiom **FU**: For any opinions o_0 and o_1, o_0+o_1 has less uncertainty than o_1. For example, imagine o_0 and o_1 being polar opposites, both with low uncertainty, this feels as though there should be uncertainty arising from this seeming contradiction. However, while uncertainty about the correctness of the data or the model, we specifically mean agent uncertainty – uncertainty arising from not knowing a user's behaviour. Typically, trust models consider the agent uncertainty to still have gone down – this is what the mathematics typically tells us. After all, we can see uncertainty as the inverse of the number of observations, and the number of observations undeniably went up. Axiom **FU** is part of our definition of agent uncertainty, and we argue that this is sensible, as it relates to the inverse of the number of observations.

Finally, in Section 5, the question of impact is addressed. While it may be impossible to have a belief-consistent trust model, perhaps there is a trust model that is good enough. Arguably, may state-of-the-art trust models are good enough for purpose. However, we show, in Theorem 4, that any deviation can be blown up to cause an arbitrarily large error. Fortunately, the construction is unlikely to occur in practice, and we conjecture that good trust models for practical purposes do exist. This does mean that there is no perfect trust model, so models have to be chosen pragmatically.

Based on our impossibility results, we conjecture that there is no trust model in which trust opinions have a finite representation. Specifically, our conjecture is that even with a large number of trust values, uncertainty values and auxiliary values, no general belief-consistent trust model exists. The current tools are not sufficient to address this conjecture, so some additional formalism is required.

7 Conclusion

We formalise trust models as mathematical objects in which operations occur on trust opinions. The trust operations are the two most ubiquitous ones in the field: Fusion, allowing users to learn from new evidence, and chaining, modelling reports/ratings/recommendations/transitive trust.

The core question in the paper was "what are the limitations of trust models?" We introduce the notion of belief-consistence to help us answer this question. If evidence changes our beliefs in a specific way, then the trust model's job is to reflect this change. It turns out that this is too much to ask.

We consider trust models without uncertainty, and prove that the only trust model that works without uncertainty, has an overly simplified underlying belief. It can only capture the belief that agents have one of two possible behaviours. This is not a realistic model for most applications. Models with uncertainty can capture fusion and be belief-consistent.

Full trust models with uncertainty were then studied. Scenarios with both fusion and chaining are too complex to be modelled. No such trust model can be belief-consistent. Finally, we show that even approximating the beliefs cannot be enough in all cases. There is a trust network that can blow up any tiny discrepancy.

There are still questions to be answered, including the big question whether finite belief-consistent trust models can exist at all.

References

1. Che, S., Feng, R., Liang, X., Wang, X.: A lightweight trust management based on bayesian and entropy for wireless sensor networks. Security and Communication Networks **8**(2), 168–175 (2015)
2. Dempster, A.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat **38**(4), 325–339 (1967)
3. ElSalamouny, E., Sassone, V., Nielsen, M.: Hmm-based trust model. In: International Workshop on Formal Aspects in Security and Trust. pp. 21–35. Springer (2009)
4. Eziama, E., Jaimes, L.M., James, A., Nwizege, K.S., Balador, A., Tepe, K.: Machine learning-based recommendation trust model for machine-to-machine communication. In: 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). pp. 1–6 (2018). https://doi.org/10.1109/ISSPIT.2018.8705147
5. Ismail, R., Jøsang, A.: The beta reputation system. BLED 2002 proceedings p. 41 (2002)
6. Jøsang, A.: Artificial reasoning with subjective logic. In: Proceedings of the second Australian workshop on commonsense reasoning. vol. 48, p. 34. Citeseer (1997)
7. Jøsang, A.: Subjective logic. Springer (2016)
8. Levy, S.E., Duan, W., Boo, S.: An analysis of one-star online reviews and responses in the washington, dc, lodging market. Cornell Hospitality Quarterly **54**(1), 49–63 (2013)
9. Muller, T.: An unforeseen equivalence between uncertainty and entropy. In: Trust Management XIII: 13th IFIP WG 11.11 International Conference, IFIPTM 2019, Copenhagen, Denmark, July 17-19, 2019, Proceedings 13. pp. 57–72. Springer (2019)
10. Muller, T., Schweitzer, P.: On beta models with trust chains. In: IFIP International Conference on Trust Management. pp. 49–65. Springer (2013)
11. Ou, W., Luo, E., Tan, Z., Xiang, L., Yi, Q., Tian, C.: A multi-attributes-based trust model of internet of vehicle. In: Network and System Security: 13th International Conference, NSS 2019, Sapporo, Japan, December 15–18, 2019, Proceedings 13. pp. 706–713. Springer (2019)
12. Shafer, G.: A mathematical theory of evidence, vol. 42. Princeton university press (1976)
13. Teacy, W.L., Luck, M., Rogers, A., Jennings, N.R.: An efficient and versatile approach to trust and reputation using hierarchical bayesian modelling. Artificial Intelligence **193**, 149–185 (2012)
14. Xiao, S., Dong, M.: Hidden semi-markov model-based reputation management system for online to offline (o2o) e-commerce markets. Decision Support Systems **77**, 87–99 (2015)