On Trust Confusional, Trust Ignorant, and Trust Transitions

Yoshinobu Kawabe, Yuki Koizumi, Tetsushi Ohki, Masakatsu Nishigaki, Toru Hasegawa, Tetsuhisa Oda

On Trust Confusional, Trust Ignorant, and Trust Transitions

Yoshinobu Kawabe[1], Yuki Koizumi[2], Tetsushi Ohki[3], Masakatsu Nishigaki[3], Toru Hasegawa[2], and Tetsuhisa Oda[1]

[1] Graduate School of Business Administration and Computer Science, Aichi Institute of Technology
Yachigusa 1247, Yakusa-cho, Toyota, Aichi 470-0392, Japan
{ kawabe, oda }@aitech.ac.jp
[2] Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
{ ykoizumi, t-hasegawa }@ist.osaka-u.ac.jp
[3] Graduate School of Science and Technology, Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8011, Japan
{ ohki, nishigaki }@inf.shizuoka.ac.jp

Abstract. This paper introduces a two-dimensional representation for trust values that uses two metrics: "trust" and "distrust." With this representation, we can deal with such contradictory arguments as "The message is basically trustworthy but simultaneously not trustworthy." Such situations can be caused when a message is consistent with other messages, but the message is sent from an unknown sender. We also explore how to analyze the transitions of two-dimensional trust values with a theory of distributed algorithms and compare our trust representation with Jøsang's subjective logic.

Keywords: Two-dimensional trust representation · fuzzy logic · I/O-automaton theory · safety/liveness properties · subjective logic

1 Introduction

During recent large-scale disasters, social media have been actively used to exchange various information about victims. Although such social media messages are helpful during disasters, some might be unreliable. For example, when a huge earthquake struck northern Osaka on the morning of June 18, 2018, many fake messages were distributed on Twitter and rapidly retweeted all over Japan, causing many problems.

Even if a message's content is true at one specific moment, the message may "become untrue" as time passes. For example, even if the following message, "A person is seriously injured but still alive," is true immediately at the beginning of a disaster, it might be false an hour later; the person might be dead. In this sense, some messages may not be reliable. If one receives a message from an unknown sender, one might also suspect that it is unreliable. This might happen even if the message's content is relatively consistent.

To deal with such situations, we must properly evaluate the trust of messages and senders. Marsh and Dibben introduced a trust value, which ranges from -1 to 1, and classified trust notions into *trust*, *distrust*, *untrust*, and *mistrust* [13]. Their classification is one-dimensional; i.e., trust and distrust are at both extremities. However, for the notions of trust and distrust, Lewicki et al. [10] suggested that they are located at entirely separate dimensions. Cases exist where a one-dimensional expression is not sufficient for trust values.

Trust is a property that is closely related to human impressions. We believe that a technique for impression formation based on mathematical psychology should be applied for trust values. Oda [5][17][18][19] developed a Fuzzy-set Concurrent Rating (FCR) method with fuzzy logic that enables us to measure and analyze human impressions. Since the FCR method allows two or more dimensions for representing a truth value, trust and distrust notions can be described two-dimensionally by applying them to a trust representation. This enables us to describe situations in (i) confusional trust (e.g. "Although he can basically be trusted, in some cases he is not trustworthy") and (ii) ignorant trust (e.g. "Since I have never met him, I have no impression of him."). In this paper, we introduce a FCR-based, two-dimensional trust representation and show how it corresponds to the conventional trust representation of Marsh and Dibben.

We also deal with transitions of trust. If we regard a two-dimensional trust value as a state of an automaton, we can discuss properties defined with a series of state transitions. With results from the theory of distributed algorithms, we discuss safety-related trust properties (e.g. "A user never reaches a state of distrust" and "If a user exits the distrust region, she never returns to it") and liveness-related trust properties (e.g., "A user can finally reach the trust region."). We also discuss an efficient proof method for trust-related safety properties based on I/O-automaton theory.

This paper is organized as follows. After showing some notions and notations in Section 2, we introduce a two-dimensional trust representation in Section 3. In Section 4, we model and analyze trust transitions. Finally, in Section 5, we compare our trust representation with Jøsang's subjective logic.

2 Preliminaries

2.1 The FCR Method

A rating scale method (Fig. 1) is often used for questionnaires, where such adjectives as "poor," "fair," "average," "good," and "excellent" are given from which respondents choose. One problem with this method is that they tend to choose the middle item in the scale. This problem presents two cases. The first case is that the respondent has multiple answer candidates that are located at both extremities. The respondent usually chooses one of them, but if it is difficult for the respondent to choose one, a middle item may be chosen instead. The chosen middle item is not the true answer; the middle item is usually "average" or "neutral," which complicates analysis. In the second case, since the respondent lacks sufficient knowledge/interest to answer, she chooses the middle value.

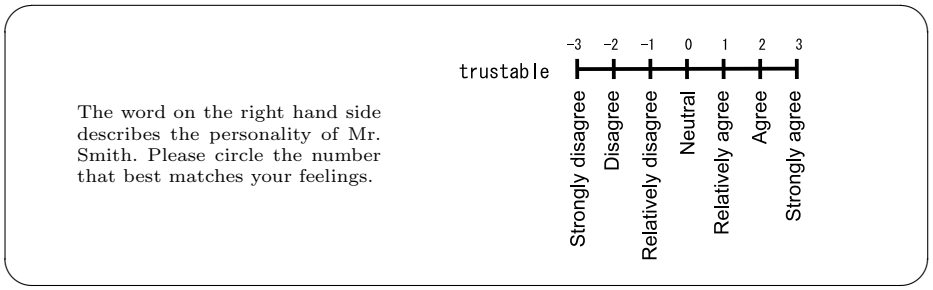

Fig. 1. Conventional questionnaire

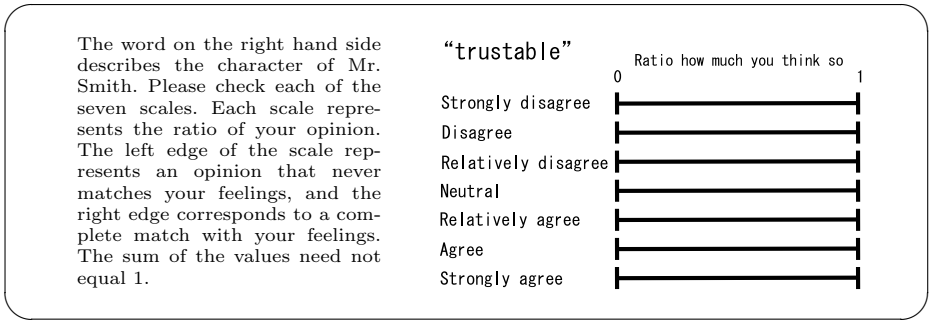

Fig. 2. Rating with FCR method

To overcome this problem of choosing the middle item, in the FCR method, respondents are requested to describe their confidence in each item (Fig. 2); in other words, the respondents answer how much they believe the truthiness in each item. Then by applying fuzzy inference, we calculate the true answers of the respondents. From a theoretical viewpoint we have no restrictions on the dimensions (i.e., the number of items), but for simplicity we just employ two dimensions in the rest of this paper.

Hyper Logic Space Model The FCR method employs the Hyper Logic Space model (HLS) as a logic space for multiple-dimensional multiple-valued logic. Figure 3 shows a two-dimensional space based on $true$ and $false$. For any $t, f \in [0,1]$, pair (t,f) is called an observation. t and f are independent; we do not assume such conditions as $t+f=1$. We call $\{(t,f) \,|\, t,f \in [0,1] \wedge t+f>1\}$ the region of contradiction. $\{(t,f) \,|\, t,f \in [0,1] \wedge t+f<1\}$ is called the region of ignorance, or the region of irrelevance. Finally, $\{(t,f) \,|\, t,f \in [0,1] \wedge t+f=1\}$ is the consistent region.

Integration Value Given observation (t,f), we need to calculate an actual truth value, which is called an integration value. Integration values can be cal-

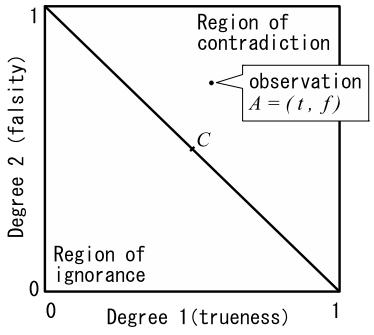

Fig. 3. Two-dimensional HLS

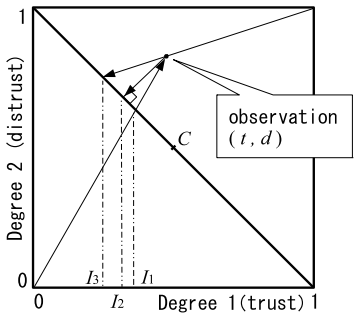

Fig. 4. Graphical calculation for integration values

culated in several ways, and we employ the reverse-item averaging method, where integration value I_2 is defined with $I_2(t, f) = \dfrac{t + (1 - f)}{2}$. The integration value is the average of the degree of the positive elements and the complementary degree of the negative elements. $I_2(t, f)$ is calculated in a graphical manner (Fig. 4). The result of calculation is the value of "degree 1" after drawing a perpendicular line from (t, f) to Fig. 4's diagonal line.

Degree of Contradiction Another important value in the FCR method is the degree of contradiction [5][17] or the contradiction-irrelevance degree. In the field of personality psychology, some situations are allowed, including "I like it, but I don't like it" or "I don't care for it at all." The degree of such confusion/irrelevance is formulated with the degree of contradiction.

For observation (t, f), degree of contradiction $C(t, f)$ should satisfy $C(t, f) = 1$ for complete confusion, $C(t, f) = -1$ for complete ignorance, and $C(t, f) = 0$ for a consistent situation. $C(t, f) = t + f - 1$ is usually employed where $C(t, f)$ represents the distance between (t, f) and the consistent region.

2.2 Trust Classification by Marsh and Dibben

A conventional trust value is a real number in $[-1, 1)$. Readers interested in the details of calculating trust values can find them here [13], but in this paper we omit them since they are beyond the scope of this paper and directly handle the calculated trust values. Marsh and Dibben introduced the following four notions of trust:

– *Trust*: This notion represents a case where a trust value is positive and exceeds a predefined value called a cooperation threshold. In this case, a trustee should be trusted, and the trust value is regarded as a measure of how much an agent believes the trustee.

- *Distrust*: Here the trust value is negative, and an agent believes that a trustee will actively work against her in a given situation.
- *Untrust*: Although the trust value is positive, it is not high enough to produce cooperation. An agent cannot determine if a trustee is actually trustworthy.
- *Mistrust*: Initial trust has been betrayed. More precicely, mistrust represents a situation either a former trust was destroyed or a former distrust was healed.

The mistrust notion is a time-related trust property discussed in Section 4. We address trust, distrust, and untrust notions in the following section. For these properties, see studies by Primiero [20] (on distrust and mistrust) and [21] (on trust and distrust).

3 FCR-Based Two-Dimensional Trust Representation

Suppose that you received a message, and you calculated its trust value. If the trust value is 0.9 and the cooperation threshold is 0.85, then from the definition of the trust notion, the message should be trusted. However, can you say that you have absolutely no distrust about this message? Since the maximum trust value is 1, a deficit of 0.1 exists. In this sense, the message might not be trusted enough.

We believe that this situation is caused by the limitations of the power of one-dimensional expressions. Hence, in this study we employ the degrees of trust $Trust$ and distrust $DisTrust$ defined with $Trust = DisTrust = \{\, v \mid 0 \leq v \leq 1\}$ and define a two-dimensional trust value as an element of $Trust \times DisTrust$. Following the FCR method, a two-dimensional trust value is also called an observation in this paper.

3.1 Understanding Two-Dimensional Trust Values

We semantically understand two-dimensional trust values by observing some of them.

Observation $(1, 0) \in Trust \times DisTrust$ has a high degree of trust (1) and a low degree of distrust (0). $(1, 0)$ represents a case where a trustee is completely trusted; this observation corresponds to (conventional) trust value 1. Observation $(0, 1)$ represents a case of complete distrust and corresponds to trust value -1. Observation $(0.5, 0.5)$, which falls exactly between $(1, 0)$ and $(0, 1)$, corresponds to 0 in conventional trust values.

To define such trust notions as trust, distrust, and untrust in our two-dimensional trust model, we employ the following transformation:

$$\left[\begin{pmatrix} \cos\frac{\pi}{4} & -\sin\frac{\pi}{4} \\ \sin\frac{\pi}{4} & \cos\frac{\pi}{4} \end{pmatrix} \left\{ \begin{pmatrix} t \\ d \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} + \begin{pmatrix} \frac{\sqrt{2}}{2} \\ 0 \end{pmatrix}\right] \times \frac{1}{\frac{\sqrt{2}}{2}} = \begin{pmatrix} t-d \\ t+d-1 \end{pmatrix}.$$

Figure 5 shows the transformation and observations $(1, 0)$, $(0, 1)$, and $(0.5, 0.5)$ are respectively mapped to $(1, 0)$, $(-1, 0)$, and $(0, 0)$. Below, the resulting point

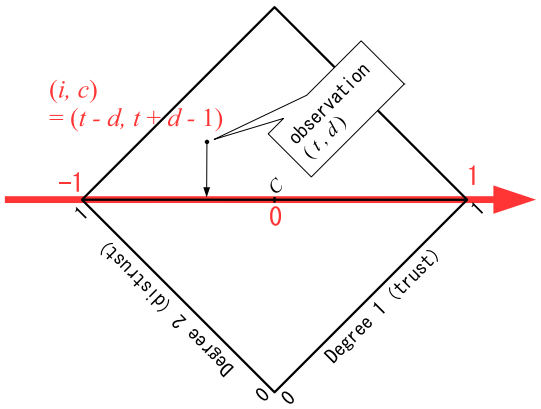

Fig. 5. Graphically understanding calculation of (i, c)

of the transformation is called (i, c). First element $i = t - d$ can be calculated with the reverse-item averaging method in Section 2.1. Actually, the value of i is calculated by normalizing $I_2(t, d)$ to be a value in region $[-1, 1]$; note that the range of integration value $I_2(t, d)$ was originally $[0, 1]$.

The value of i was regarded as a conventional trust value given by Marsh and Dibben. From the definition of $i = t - d$, a net trust value is calculated by subtracting the degree of trust from the degree of distrust, which matches our intuition. From Fig. 5, the consistent region, which is the line between $(1, 0)$ and $(0, 1)$ before the transformation, corresponds to the set of conventional trust values. Observation (t, d) in the consistent region satisfies $t + d = 1$ and is regarded as an assumption on the trust and distrust degrees. The theory of conventional trust values implicitly introduces this assumption.

Trust notions are defined with the value of i[1]. Let CT be a cooperation threshold. If we have $i = t - d \geq CT$, then it is a case of trust; if i is negative then it is case of distrust; if we have $0 \leq i < CT$, then it is a case of untrust. Note that for the case of distrust, condition $i < 0$ is equivalent to $t < d$; i.e., a trustee is distrusted if the degree of distrust exceeds the degree of trust.

3.2 New Classification on Untrust

As shown in Fig. 4, integration values can be graphically calculated. Observing the graphical calculation, the two-dimensional trust values in the same perpen-

[1] In this paper, we only define trust notions with the value of $i = t - d$ without the value of c. Our paper's trust notions are formalized with "linear" functions. For example, the trust and distrust notions are defined with restrictions $d \leq -t + CT$ and $d > t$. This is just for simplicity, and we believe it is possible to provide a finer definition for trust notions with both i and c; that is, we believe a "non-linear" definition is possible. This is future work.

dicular line have identical integration values. For example, observation $A = (t, d)$ and its nearest point on diagonal line $A' = (\frac{t+(1-d)}{2}, 1 - \frac{t+(1-d)}{2})$ have the same integration value. However, for observations A and A', the distance from the diagonal line is different. The distance between observation (t, d) and the diagonal line in Fig. 4 is given by $|t + d - 1|$, which is the absolute value of second element $c = t + d - 1$ of point (i, c) defined in the previous section. The formula of c is equivalent to the degree of contradiction-irrelevance $C(t, d)$ of the FCR method.

If $C(t, d)$ is positive and high, then it is a state of confusion; both trust and distrust degrees are high. If $C(t, d)$ is negative and low (i.e., the absolute value of c is high), then it is regarded as a state of irrelevance; in this case, both the trust and distrust degrees are low. In the field of fuzzy logic, a state of confusion is caused by information overload, and a state of irrelevance is caused by a lack of information [5][17]. For information overload, there is too much evidence about a trustee, some of which may increase the trust value on the trustee, but others may increase the distrust value. This causes confusion, which leads to a situation where you cannot determine whether the trustee is trustworthy. If you lack sufficient evidence, i.e., if you ignore the trustee, you cannot discuss whether she is trustworthy.

This discussion demonstrates that two cases exist where one cannot determine whether the trustee is trustworthy. Therefore, we introduce two types of new untrust notions:

- *Untrust confusional*: the trustee is both trusted and distrusted. Formally, this is a case with $0 < i < CT$ and $c \geq 0$.
- *Untrust ignorant*: the trustee is ignored; in other words, the trustee is neither trusted nor distrusted. Formally, this is a case with $0 < i < CT$ and $c < 0$.

The original untrust notion [13] corresponds to the notion of untrust ignorant, and in this paper, we introduce a new kind of untrust notion from the viewpoint of confusion.

3.3 Example

In three countries, an opinion poll was conducted about the approval ratings of each country's governments. We used the following items to answer this question: "*Do you trust your government*?"

1. *I have no idea*;
2. *Yes, I do*;
3. *No, I do not*;
4. *Sometimes yes, sometimes no.*

For country c, the number of answers for each item is $a_1^c, \ldots, a_4^c$; also, we have $s^c = a_1^c + a_2^c + a_3^c + a_4^c$. In this example, we calculate the degrees of trust t_c and distrust d_c of the government with $t_c = \frac{a_2^c + a_4^c}{s^c}$ and $d_c = \frac{a_3^c + a_4^c}{s^c}$.

A survey was conducted with 100 residents each in the countries of X, Y, and Z, and the following are the results:

$$\begin{aligned}(a_1^X, a_2^X, a_3^X, a_4^X) &= (10, 20, 30, 40),\\ (a_1^Y, a_2^Y, a_3^Y, a_4^Y) &= (50, 30, 10, 10), \text{ and}\\ (a_1^Z, a_2^Z, a_3^Z, a_4^Z) &= (20, 25, 5, 50).\end{aligned}$$

For each country, the following are the degrees of trust t_c and distrust d_c:

$$(t_X, d_X) = (0.6, 0.7),\ (t_Y, d_Y) = (0.4, 0.2) \text{ and } (t_Z, d_Z) = (0.75, 0.55).$$

For each country we can also calculate the values of i and c:

$$(i_X, c_X) = (-0.1, 0.3),\ (i_Y, c_Y) = (0.2, -0.4) \text{ and } (i_Z, c_Z) = (0.2, 0.3).$$

From this result, the following analysis is possible. For country X, there is some degree of distrust of the government, and citizens in country X are somewhat confused since the degree of contradiction is positive. For country Y, the degree of trust exceeds the degree of distrust, but the degree of contradiction is negative, which suggests that the people have little interest in their government. For countries Y and Z, although their integration values are the same, the degree of contradiction is positive for country Z. Note that we can compare countries Y and Z, even though the conventional trust model cannot since the degree of contradiction is not addressed.

4 Transitions of Two-Dimensional Trust Values

Mistrust is a property with regard to misplaced trust. If the first estimation for an observation is in the region of trust (i.e. $i = t - d \geq CT$), but the next estimation is changed to the region of distrust (i.e. $i < 0$), then the first estimation was incorrect. With this understanding, mistrust can be modeled as a property for the changes or transitions of a trust value.

Transition-related trust properties must be analyzed, including mistrust or swift trust [14][23], especially during disasters [3][9][15]. In this section, we regard an observation as a state and analyze the transitions of trust values.

4.1 Dealing with Trust Values as States

I/O-automaton [11][12] is a formal system for distributed algorithms, where a distributed system is modeled as a state machine and its properties are formalized with observable actions. Some actions, such as keyboard input, display output, and open communication in the Internet are observable, and others are unobservable, such as internal processing and secret communication.

Formally, automaton X has set of actions $sig(X)$, set of states $states(X)$, set of initial states $start(X) \subset states(X)$, and set of transitions $trans(X) \subset states(X) \times sig(X) \times states(X)$. Transition $(s, a, s') \in trans(X)$ is written as $s \xrightarrow{a}_X s'$. In this paper, a state is a tuple of values. Each element of the tuple has

a corresponding distinct variable name. A variable's name is used as an access function to its value. Such modeling is standard in I/O-automaton theory and its extensions, such as [8]. In this paper, we use variables tr and dis for trust and distrust values. The trust and distrust degrees in state $s \in states(X)$ are called $s.tr$ and $s.dis$.

4.2 Formalizing Trace-Based Trust Properties

Let $\alpha \equiv s_0 \xrightarrow{a_1}_X s_1 \xrightarrow{a_2}_X \cdots \xrightarrow{a_n}_X s_n$ be a transition sequence of automaton X. We define $tr(\alpha)$ as a sequence of all the external (i.e., observable) actions in $a_1 a_2 \cdots a_n$, and write $s_0 \overset{tr(\alpha)}{\Longrightarrow}_X s_n$. If s_0 is an initial state, $tr(\alpha)$ is called a trace of α. A trace is a sequence of observable actions in an execution from an initial state.

In I/O-automaton theory, various properties of distributed systems are defined with traces (Section 8.5.3 of [12]). Well-known characteristics are safety and liveness properties. Informally, a safety property means that nothing bad ever happens in a system. For example, the following are safety properties: "no division by zero error occurs" and "after reaching special state s, the system never reaches an error state." A liveness property means that something good might happen. "A program can terminate" and "from any state, the system can reach an initial state" are typical liveness properties.

If we regard Section 3's observations as states, we can define safety/liveness properties related to trust transitions. "An observation never reaches the region of distrust" and "after reaching the regions of trust or untrust, an observation never reaches the region of distrust" are trust safety properties. "An observation can reach the region of trust" is a trust liveness property.

Formalizing Trust Safety Properties Let CT be a cooperation threshold. We define the regions of trust $T(CT)$, distrust D, and untrust $U(CT)$:

$$\begin{cases} T(CT) = \{\, (t,d) \,|\, t \in Trust \ \wedge\ d \in DisTrust \wedge\ t - d \geq CT \,\} \\ D \quad\ \ = \{\, (t,d) \,|\, t \in Trust \ \wedge\ d \in DisTrust \wedge\ t < d \,\} \\ U(CT) = Trust \times DisTrust \setminus (T(CT) \cup D), \end{cases}$$

where $0 < CT \leq 1$ holds. Sets $T(CT)$, D, and $U(CT)$ correspond to Section 3.1's definitions for trust notions, where "linear" functions are employed; exploring a "non-linear" setting is a future work.

We introduce a predicate for the reachability from state s to state s':

$$\begin{aligned} &reachable(s, s') \\ &\quad \Longleftrightarrow s = s' \vee \exists s'' \in states(X)\, \exists a \in sig(X) [\, s \xrightarrow{a}_X s'' \wedge reachable(s'', s') \,], \end{aligned}$$

and we define predicate $nonDistr(s)$:

$$nonDistr(s) \Longleftrightarrow \forall s' \in states(X)\, [\, reachable(s, s') \implies (s'.tr,\, s'.dis) \notin D \,].$$

With these predicates, we can formalize a trust safety property, "an observation never reaches the region of distrust," with $\forall s \in start(X)\,[\,nonDistr(s)\,]$. Another safety property,"after reaching the region of trust or the region of untrust, an observation never reaches the region of distrust," is formalized:

$$\forall s \in states(X)\,\forall s' \in states(X) \\ [\,(reachable(s,s') \wedge (s'.tr, s'.dis) \notin D) \implies nonDistr(s')\,].$$

Formalizing Trust Liveness Properties We can also formalize trust liveness properties. Let $n \in \mathcal{N}$ be a natural number. We define $reach^n(s,s')$ to represent that state s' is reachable from state s with n-steps as follows:

$reach^0(s,s') \Longleftrightarrow s = s'$, and
$reach^{k+1}(s,s') \Longleftrightarrow \exists s'' \in states(X)\,\exists a \in sig(X)[\,reach^k(s,s'') \wedge s'' \xrightarrow{a}_X s'\,]$.

With $reach^n(s,s')$, "an observation can reach the region of trust" is defined:

$$\forall s \in states(X)\,\exists s' \in states(X)\,\exists n \in \mathcal{N}\,[\,reach^n(s,s') \wedge (s'.tr, s'.dis) \in T(CT)\,].$$

4.3 Efficient Proof Method for Trace-Based Trust Properties

Although we can directly prove the logic formulae in the previous section, this is inefficient. By employing a result in I/O-automaton theory, a more efficient proof is possible.

Figure 6 shows the specification of automaton `testerSafety`, which describes the transitions of a two-dimensional trust value. It is written in an I/O-automaton-based specification language called IOA [2]. This specification has three variables. Variables `tr` and `dis` are for the degrees of trust and distrust. Variable `stateOfAgent` is used for a trustee's internal state. Automaton `testerSafety` has three actions: `move`, `inDistr`, and `notInDistr`. Each action is described in a precondition-effect style, where the `pre`-part has a condition to fire the action and the `eff`-part has the action's body. Action `move` shows that an observation moves from (`pt`, `pd`) to (`pt` + `dt`, `pd` + `dd`) when event `ev` occurs. Actions `inDistr` and `notInDistr` are special observable qualities that denote whether the current observation is in the region of distrust. Action `inDistr` is enabled if $(\texttt{tr}, \texttt{dis}) \in D$ holds, which is equivalent to $\texttt{tr} < \texttt{dis}$, and action `notInDistr` is enabled if $(\texttt{tr}, \texttt{dis}) \notin D$ holds.

To specify automaton `testerSafety`, we need a concrete definition for predicate `condition` in the `pre`-part of action `move`. If we define this predicate with

$$\texttt{condition(stateOfAgent, ev, pt, pd, dt, dd)} \\ \Longleftrightarrow (\texttt{pt}, \texttt{pd}) \notin D \wedge (\texttt{pt} + \texttt{dt}, \texttt{pd} + \texttt{dd}) \notin D$$

then action `inDistr` cannot be enabled. None of `testerSafety`'s traces have occurrences of action `inDistr`. This creates a set of traces $traces(\texttt{testerSafety})$,

```
automaton testerSafety
  signature
    internal move(ev:Event, pt: VL, pd: VL, dt: VL, dd: VL)
    output inDistr(t:VL, d:VL)
    output notInDistr(t:VL, d:VL)

  states
    tr: VL := 0,    % VL ranges over [-1, 1]
    dis: VL := 0,   % but we assume 0 <= tr, dis <= 1 at any state
    stateOfAgent: agtState := InitState

  transitions
    internal move(ev, pt, pd, dt, dd)
      pre    pt = tr
          /\ pd = dis
          /\ (0 <= (pt + dt) /\ (pt + dt) <= 1)
          /\ (0 <= (pd + dd) /\ (pd + dd) <= 1)
          /\ condition(stateOfAgent, ev, pt, pd, dt, dd)
      eff tr := tr + dt;
          dis := dis + dd;
          stateOfAgent := change(stateOfAgent, ev)

    output inDistr(t, d)
      pre tr < dis /\ t = tr /\ d = dis
      eff do nothing

    output notInDistr(t, d)
      pre ~(tr < dis) /\ t = tr /\ d = dis
      eff do nothing
```

Fig. 6. Automaton `testerSafety` written in IOA language

where $traces(A)$ is used for the set of all the traces of automaton A and specifies the trust safety property "an observation never reaches the region of distrust in automaton `testerSafety`." Automaton `testerSafety` obviously satisfies $\forall s \in start(\texttt{testerSafety})\,[\,nonDistr(s)\,]$.

Automaton `traceSafety` is the specification automaton for a safety property, but we need to deal with a safety property of a concrete system. Let A be an automaton and let $traces(A)$ be a corresponding trace set. If trace inclusion $traces(A) \subseteq traces(\texttt{testerSafety})$ holds, then automaton A satisfies the safety property defined with automaton `testerSafety`'s traces. Therefore, to show $\forall s \in start(A)\,[\,nonDistr(s)\,]$, it suffices to show the trace inclusion.

I/O-automaton theory provides techniques that prove a trace inclusion of (possibly infinite-state) systems, which can be applied with a theorem-proving tool [2][22]. Finding a forward simulation between automata is one of the techniques. Forward simulation f from I/O-automaton $Conc$ to I/O-automaton $Abst$ is a binary relation over states satisfying the following conditions:

Initial state correspondence: For any initial state $a \in start(Conc)$, there is initial state $b \in start(Abst)$ and $f(a, b)$ holds.

Step correspondence: For any reachable states $a_1, a_2 \in states(Conc)$, $b_1 \in states(Abst)$ and any action $\pi_{Conc} \in sig(Conc)$, if $f(a_1, b_1)$ and $a_1 \xrightarrow{\pi}_{Conc} a_2$ hold, then there is a state $b_2 \in states(Abst)$ satisfying $f(a_2, b_2)$ and $b_1 \stackrel{\beta}{\Longrightarrow}_{Abst} b_2$ with $\beta = tr(a_1 \xrightarrow{\pi}_{Conc} a_2)$.

From Theorem 3.10 of [11], if there is a forward simulation from *Conc* to *Abst*, then we have $traces(Conc) \subseteq traces(Abst)$. Therefore, to show trace inclusion $traces(A) \subseteq traces(\texttt{testerSafety})$, it suffices to find a forward simulation from A to `testerSafety`. This leads to a safety property: "an observation never reaches the distrust region in A."

4.4 Example

Figure 7 shows an I/O-automaton `bbdSystem`, which is a specification of a communication system that sends a user's message to an online bulletin board after evaluating an observation. Specifically, by action `get_mes`, the system receives a message from a user, and an observation is evaluated with actions `discard_mes` and `approve_mes`. If pair (`tr+evalTr(i, m)`, `dis+evalDis(i, m)`) of the next state's observation falls in the distrust region, the message is discarded by action `discard_mes`; otherwise, it is sent by actions `approve_mes` and `say`.

If we hide observable actions `get_mes` and `say` in `bbdSystem`, that is, if we deal with these observable actions as internal ones, we can find a forward simulation from automaton `bbdSystem`$\backslash\{$`get_mes`, `say`$\}$ to automaton `testerSafety`. Consequently, we have $traces($`bbdSystem`$\backslash\{$`get_mes`, `say`$\}) \subseteq traces($`testerSafety`$)$ that provides the safety property defined with automaton `testerSafety`. A complete computer-assisted proof is found in [24].

5 Discussion

In this section we compare our two-dimensional trust representation with a similar approach found in Jøsang's subjective logic [7].

5.1 Two-Dimensional Representation in Subjective Logic

In probabilistic logic [16], the truth values of propositions are probabilities and are given based on the frequency of events. The confidence on a truth value is high if enough attempts can be made; for example, we can confirm the truthiness of proposition "the probability of heads when flipping a coin is 0.5" if we can toss the coin many times. In subjective logic, truth values are defined from an epistemic viewpoint. The confidence of a truth value is high if we know how a situation happens. For example, the confidence of the proposition, "the probability that Lee Harvey Oswald killed John F. Kennedy is 0.5" is high if the dynamics of the case are well-known; however, many aspects of this case remain unknown, so the confidence is not actually high.

```
automaton bbdSystem
  signature
    input get_mes(i:ID, m:MES)
    internal discard_mes(i:ID, m:MES)
    internal approve_mes(i:ID, m:MES)
    output say(i:ID, m:MES)
    output inDistr(t:VL, d:VL)
    output notInDistr(t:VL, d:VL)

  states
    tr: VL := 0,    % VL ranges over [-1, 1]
    dis: VL := 0,   % but we assume 0 <= tr, dis <= 1 at any state
    flg: Bool := false,
    mesQ: Seq[MES] := empty

  transitions             % Note: input actions does not have the
    input get_mes(i, m)   % "pre"-part since they are always enabled.
      eff mesQ := mesQ || (packet(i, m) -| empty)

    internal discard_mes(i, m)
      pre ~flg /\ mesQ ~= empty
               /\ packet(i, m) = head(mesQ)
               /\ ((tr + evalTr(tr, m))-(dis + evalDis(dis, m))) < 0
      eff mesQ := tail(mesQ)

    internal approve_mes(i, m)
      pre ~flg /\ mesQ ~= empty
               /\ packet(i, m) = head(mesQ)
               /\ ((tr + evalTr(tr, m))-(dis + evalDis(dis, m))) >= 0
      eff flg := true

    output say(i, m)
      pre flg /\ mesQ ~= empty
              /\ packet(i, m) = head(mesQ)
      eff tr := tr + evalTr(tr, m);
          dis := dis + evalDis(dis, m);
          mesQ := tail(mesQ);
          flg := false

    Outputs "inDistr" and "notInDistr" are defined as in the case of
    automaton "testerSafety."
```

Fig. 7. System never sends a message if a user might be distrusted

Subjective logic uses a domain, which is a set of distinct opinions. If a domain consists of opinions x and $\bar{x}$, it is called a binary domain. In this study, we deal with a binary domain where one opinion x corresponds to the trust notion and its contrary opinion $\bar{x}$ represents the distrust notion. A binomial opinion in subjective logic is defined with the following quadruple $\omega_x^A = (b_x^A, d_x^A, u_x^A, a_x^A)$:

- b_x^A: the amount of observer A's belief in x;
- d_x^A: the amount of observer A's disbelief in x;
- u_x^A: the amount of observer A's uncertainty about x;
- a_x^A: the prior probability in the absence of belief or disbelief.

We assume $0 \leq b_x^A, d_x^A, u_x^A, a_x^A \leq 1$ and $b_x^A + d_x^A + u_x^A = 1$ hold for any b_x^A, d_x^A, u_x^A and a_x^A. Values b_x^A, d_x^A, and u_x^A are depicted with a triangle in Fig. 8. The

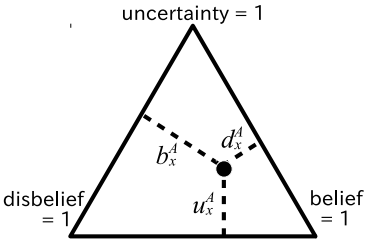

Fig. 8. Triangular Representation of Binomial Opinion in Subjective Logic

right bottom is the case where $b_x^A = 1$, the left bottom is where $d_x^A = 1$, and the top vertex is where $u_x^A = 1$.

A two-dimensional trust representation by b_x^A and d_x^A is possible in subjective logic, where b_x^A and d_x^A represent the degrees of trust and distrust.

5.2 Comparing Two Types of Trust Representations

Pair (b_x^A, d_x^A) in subjective logic corresponds to (t, d) of the FCR method in Section 3. We compare the two types of trust representations.

Considering Three Kinds of Opinions We consider three kinds of opinions below. The first one is an opinion where $b_x^A = 1$ and is shown at the bottom right of the Fig. 8's triangle. This opinion's correspondence in the FCR-based model is in $\{ (t, d) \mid t = 1 \wedge d \in DisTrust \}$ where the degree of trust is 1. However, as discussed later, since subjective logic does not deal with the region of confusion, thus the point where $b_x^A = 1$ exactly corresponds to $(1, 0)$ in the HLS model. Actually, in subjective logic the state where $b_x^A = 1$ is for an absolute opinion on x, and from a trust viewpoint trustee x is completely trusted by agent A.

The second type of opinion is where $u_x^A = 0$. In this case, we have $b_x^A + d_x^A = 1$; that is, the opinion is dogmatic and completely consistent. For this situation, the degree of contradiction is 0 in the FCR-based model. From a discussion in Section 3.1, such opinions are regarded as a conventional trust value by Marsh and Dibben.

Finally, we consider the case where $u_x^A = 1$. Here $b_x^A = d_x^A = 0$ holds since we have $0 \leq b_x^A, d_x^A \leq 1$ and $b_x^A + d_x^A + u_x^A = 1$. In subjective logic, this state is called vacuous or undefined. Observation $(0, 0)$ of the FCR-based model corresponds to this opinion.

Difference of Two Representations Observation $(0, 0)$ represents the total uncertainty. However, in the FCR-based model, this is not the only observation of it. We have another observation, $(1, 1)$, where the trustee is highly trusted but simultaneously highly distrusted. In this situation, you cannot determine

whether the trustee is trustworthy. From the following discussion, the subjective-logic-based approach cannot deal with such uncertainty.

From condition $b_x^A + d_x^A + u_x^A = 1$, we have $-u_x^A = b_x^A + d_x^A - 1$. The right hand side of this formula is equivalent to the degree of contradiction since pair (b_x^A, d_x^A) corresponds to (t, d) of the FCR method. $-u_x^A$ is the degree of contradiction. Moreover, we have $0 \leq u_x^A \leq 1$, which leads to $-1 \leq b_x^A + d_x^A - 1 \leq 0$. Hence, binomial opinions in subjective logic are either in the region of ignorance or in the consistent region. Therefore, we conclude that subjective logic does not deal with the region of contradiction. A similar logic space model without the region of contradiction is found in the A-IFS model [1].

As described in Section 3.2, a state of confusion is caused by information overload, and a state of irrelevance is caused by a lack of information. This observation suggests the following difference between subjective-logic- and FCR-based approaches:

- Let A be an observer and let x be a trustee. At the beginning of the computation, observer A is ignorant of trustee x, and no evidence exists upon which to judge whether x is trustworthy. Thus, we have $b_x^A = d_x^A = 0$ and $u_x^A = 1$ in subjective logic. In this study we assume that trust and/or distrust degrees increase if the observer collects evidence on x. If this is the case, as time passes, the values of b_x^A and d_x^A increase, and the value of u_x^A decreases. Finally, after collecting enough evidence, the value of u_x^A becomes 0. Since confusing situations are ignored in subjective logic, in the subjective-logic-based approach, there is an implicit assumption that an observer can finally calculate a trustee's trustworthiness.
- On the other hand, in the FCR-based trust representation, we have $t = d = 0$ at the beginning, as in the case of the subjective-logic-based approach. Thus, we have $C(0, 0) = -1$ for the degree of contradiction. Note that the absolute value of $C(t, d)$ can be seen as the degree of uncertainty, which is maximum in the beginning. Hereafter, if the observer collects evidence about the trustee, the values of t and d increase and the value of $|C(t, d)|$ decreases. If the two-dimensional trust value (t, d) is near the consistent region, then $C(t, d)$ is almost 0, and in this situation the observer can calculate a trustee's trustworthiness. However, in the FCR-based model, we have the region of contradiction. If more evidence is collected, the values of t and d further increase, and the value of $|C(t, d)|$ also increases. Finally, the value of $C(t, d)$ becomes nearly 1, which is a contradiction. If an observer has too much evidence, she may not accurately evaluate the trustee. This is an assumption in the FCR-based approach.

The setting in the FCR-based approach is reasonable, but the assumption in the subjective-logic-based approach is considered too strong. Actually, in the example of Section 3.3, the cases for countries X and Z cannot be dealt with in the subjective-logic-based approach since the degree of contradiction is positive.

If we modify t_c and d_c with $t_c = \dfrac{a_2^c + 0.5 \times a_4^c}{s^c}$ and $d_c = \dfrac{a_3^c + 0.5 \times a_4^c}{s^c}$ in the example of Section 3.3, then $t_c + d_c - 1 \leq 0$ is always satisfied. The weight

of 0.5 is introduced for variable a_4^c, and this enables us to handle the cases for countries X and Z in the subjective-logic-based approach. We can see that, in the modified example, the trust and distrust values of a respondent are 0.5 if the respondent chooses the fourth item:

4. *Sometimes yes, sometimes no.*

Note that the sum of the values equals 1. Hence, when we employ the weight of 0.5 in estimating t_c and d_c, we implicitly assume that a respondent choosing the fourth item can consistently evaluate the trustworthiness to her government. However, we do not use such an assumption in the example of Section 3.3, since we address the suggestion [10] that trust and distrust notions should be located at completely separate dimensions. Hence, in order to deal with trust and distrust degrees independently, we need to handle not only the case of $t_c + d_c \leq 1$ but also the case of $t_c + d_c > 1$. Therefore, the region of contradiction $\{(t, d) \mid t + d > 1\}$ is required. Actually, in the modeling of Section 3.3's example, the trust and distrust degrees of the fourth item's respondent are 1, which means that the sum of the trust and distrust degrees equals 2. Some readers may consider this modeling is coarse, but we believe that a more accurate evaluation is possible if we define $t_c = \dfrac{a_2^c + a_{4pos}^c}{s^c}$ and $d_c = \dfrac{a_3^c + a_{4neg}^c}{s^c}$ with:

$$a_{4pos}^c = \sum_{i \in S_{4th}} \text{trust degree of respondent } i, \text{ and}$$
$$a_{4neg}^c = \sum_{i \in S_{4th}} \text{distrust degree of respondent } i,$$

where S_{4th} is the set of respondents choosing the fourth item.

6 Conclusion

This paper proposed a two-dimensional trust representation based on fuzzy logic. An observation was given as a pair of trust and distrust degrees, and we discussed the validity of its representation by showing a mapping to conventional one-dimensional trust representation. We also introduced a new classification of untrust. Additionally, this paper discussed such trace-based trust properties as safety properties and liveness properties. We showed how a simulation-based proof method for trace inclusion can be applied for trust safety properties. Finally, we compared our two-dimensional trust representation with a trust representation based on subjective logic.

It is important to ensure the applicability of this paper's modeling of trust properties and the proof technique to actual systems. This study is a part of a research project on disaster communication systems, and future work will prove the trust properties of a real communication system with social media, such as a communication system for disaster management [4][6]. In real systems, an analyst may receive conflicting evidence from different sources, which means that

some source of information provides wrong evidence. We believe that the degree of contradiction is applicable to handle this situation. If there are many wrong information sources, then the degree of contradiction becomes high. Hence, the analyst can use the degree of contradiction to judge whether she should discard and re-collect evidence.

Although this paper has discussed safety properties such as "if a user exits the distrust region, she never returns to it," this sort of assertion is considered too strong in the real world. To use this paper's techniques for real systems, we need proper sufficient conditions. Finding such conditions is an important future work.

Finally, we must introduce a "non-linear" definition for trust notions (see the footnote in Section 3.1), which employ both i and c.

Acknowledgments This work was supported by the National Institute of Information and Communications Technology in Japan (Contract No. 193).

References

1. Atanassov, K.T.: Intuitionistic Fuzzy Sets: Theory and Applications. Physica-Verlag GmbH, Heidelberg, Germany, Germany, 1st edn. (2010)
2. Bogdanov, A.: Formal verification of simulations between I/O-automata. Master's thesis, Massachusetts Institute of Technology (2000)
3. Busa, M. G., Musacchio, M.T., Finan, S., Fennell, C.: Trust-building through social media communications in disaster management. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1179–1184. WWW '15 Companion, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2740908.2741724
4. Chen, J., Arumaithurai, M., Fu, X., Ramakrishnan, K. K.: CNS: Content-oriented notification service for managing disasters. In: Proceedings of ACM Conference on Information-Centric Networking. pp. 122–131. ACM (2016)
5. Deng, J., Oda, T., Umano, M.: Fuzzy logical operations in the two-dimensional hyper logic space concerning the fuzzy-set concurrent rating method. Journal of Japan Association for Management Systems **17**(2), 33–42 (2001)
6. Jahanian, M., Xing, Y., Chen, J., Ramakrishnan, K.K., Seferoglu, H., Yuksel, M.: The evolving nature of disaster management in the internet and social media era. In: 2018 IEEE International Symposium on Local and Metropolitan Area Networks, LANMAN 2018, Washington, DC, USA, June 25-27, 2018. pp. 79–84 (2018). https://doi.org/10.1109/LANMAN.2018.8475116
7. Jøsang, A.: Subjective Logic: A Formalism for Reasoning Under Uncertainty. Springer Publishing Company, Incorporated, 1st edn. (2016)
8. Kaynar, D., Lynch, N., Segala, R., Vaandrager, F.: The Theory of Timed I/O Automata, Second Edition. Morgan & Claypool Publishers, 2nd edn. (2010)
9. Lemieux, F.: The impact of a natural disaster on altruistic behaviour and crime. Disasters **38**(3), 483–499 (Jul 2014)
10. Lewicki, R.J., McAllister, D.J.B., Bies, R.J.: Trust and distrust: New relationships and realities. Academy of Management Review **23**, 438–458 (1998)
11. Lynch, N., Vaandrager, F.: Forward and backward simulations — part I: Untimed systems. Information and Computation **121**(2), 214–233 (Sep 1995). https://doi.org/10.1006/inco.1995.1134

12. Lynch, N.A.: Distributed Algorithms. Morgan Kaufmann Publishers (1996)
13. Marsh, S., Dibben, M.R.: Trust, untrust, distrust and mistrust – an exploration of the dark(er) side. In: Proceedings of the Third International Conference on Trust Management. pp. 17–33. iTrust'05, Springer-Verlag, Berlin, Heidelberg (2005). https://doi.org/10.1007/11429760_2
14. Meyerson, D., Weick, K.E., Kramer, R.M.: Swift Trust and Temporary Groups in Trust in Organizations: Frontiers of Theory and Research. SAGE (1995)
15. Murayama, Y.: Issues in disaster communications. Journal of Information Processing **22**(4), 558–565 (2014). https://doi.org/10.2197/ipsjjip.22.558
16. Nilsson, N.J.: Probabilistic logic. Artif. Intell. **28**(1), 71–88 (Feb 1986). https://doi.org/10.1016/0004-3702(86)90031-7
17. Oda, T.: Fundamental characterestics of fuzzy-set concurrent rating method. Journal of Japan Association for Management Systems **12**(1), 23–32 (1995), In Japanese
18. Oda, T.: Fuzzy set theoretical approach for improving the rating scale method : Proposing and introducing the FCR-method and the IR-method as novel rating methods. Japanese Psychological Review **56**(1), 67–83 (2013), In Japanese
19. Oda, T.: Measurement technique for ergonomics, section 3: Psychological measurements and analyses (3) measurements and analyses by kansei evaluation. The Japanese Journal of Ergonomics **51**(5), 293–303 (2015), In Japanese
20. Primiero, G.: A calculus for distrust and mistrust. In: Habib, S.M., Vassileva, J., Mauw, S., Mühlhäuser, M. (eds.) Trust Management X. pp. 183–190. Springer International Publishing, Cham (2016)
21. Primiero, G., Raimondi, F., Bottone, M., Tagliabue, J.: Trust and distrust in contradictory information transmission. Applied Network Science **2**, 12 (2017). https://doi.org/10.1007/s41109-017-0029-0
22. Soegaard-Andersen, J.F., Garland, S.J., Guttag, J.V., Lynch, N.A., Pogosyants, A.: Computer-assisted simulation proofs. In: CAV '93. Lecture Notes in Computer Science, vol. 697, pp. 305–319. Springer-Verlag (1993)
23. Wildman, J., Shuffler, M., Lazzara, E., Fiore, S., Burke, S.: Trust development in swift starting action teams: A multilevel framework. Group & organization management **37**(2), 137–170 (2012)
24. `https://aitech.ac.jp/kwb/proof4testerSafety/`