A Dynamic Model of Trust in Dialogues

Gideon Ogunniye, Nir Oren, and Timothy J. Norman

Department of Computing Science, University of Aberdeen, AB24 3UE, Scotland, UK
{g.ogunniye,n.oren,t.j.norman}@abdn.ac.uk

Abstract. ...

Keywords:

1 Introduction

Within a dialogue, participants exchange advance arguments aimed at reaching some conclusion. Typically, these participants have partial information and individual preferences and goals, and the aim of the dialogue is for the parties to reach some outcome based on these individual contexts. Importantly, some dialogue participants may be *malicious* or *incompetent*, and the inputs from these parties should be discounted, based on the lack of trust ascribed to them.

While previous work [3] has considered how trust and reputation of participants should be updated following the justified conclusions of a dialogue, we observe that in long-lasting human discussions, trust can change during the dialogue itself. In turn, such changes in trust may require untrusted agents to present more evidence for their arguments to be believed, while the burden of proof reduces on highly trusted agents. Thus, there appears to be a feedback cycle which we would like to capture within more formal dialogue.

We therefore seek to address the following questions. 1) How can should trust affect the justified conclusions obtained from a dialogue? 2) How should trust change during the course of a dialogue based on the utterances made by the dialogue participants?

2 Background

2.1 Trust Graph

We use the basic concept of graph theory to represent the set of agents that are involved in a dialogue and the trust relation over them. A graph G is a pair of disjoint sets (V, E). V is the set of *vertices* (or nodes). E is the set of *edges* (or arcs) and is a set of subsets of V. In this work, we are interested in showing how participating agents of a dialogue can be represented on a trust graph and how the trust relation over them can be updated based on their behaviours in the dialogue.

We start with the definition of trust relation over a set of participating agents Ags in a dialogue.

Definition 1. *Trust relation tr over a set of participating agents in a dialogue is a preference ordering over them represented by the relationship* $\succeq$.

$$tr : Ags \times Ags \rightarrow \succeq$$

For simplicity, we assume the preference ordering $\succeq$ is a preorder (a reflexive and transitive relation) on Ags. $tr(Ag_i, Ag_j)$ denotes trust relation over agents Ag_i and Ag_j. We represent this trust relation on a directed trust graph.

Definition 2. *A trust graph for a set of agents Ags is a pair*

$$\mathcal{T} : \langle Ags, \{tr\}\rangle$$

where $\langle tr\rangle$ *is a trust relation over the agents in Ags such that if* $tr(Ag_i, Ag_j)$ *exist, then* $\{Ag_i, Ag_j\}$ *is a directed arc in* $\mathcal{T}$.

In the trust graph as shown in Figure 1, the set of participating agents is the set of vertices, and the trust relation over them is the set of arc. A directed path between agents on the trust graph indicates that we can obtain trust relation over the set of agents in Ags that are connected through the path.

Definition 3. *Let* $\mathcal{T} : \langle Ags, \{tr\}\rangle$ *be a trust graph and* $Ag_i, Ag_n \in Ags$. *A directed path P from* Ag_i *to* Ag_n *is a sequence of vertices* $\langle Ag_i, Ag_{i+1}, ..., Ag_n\rangle$ *from agent* Ag_i *to* Ag_n *such that* Ag_i *is the initial vertex of P and* Ag_n *is the terminal vertex and* $tr(Ag_i, Ag_n)$ *is as follows:*

$$tr(Ag_i, Ag_n) = tr(Ag_i, Ag_{i+1}) \uplus tr(Ag_{i+1}, Ag_{i+2}, ..., \uplus tr(Ag_{n-1}, Ag_n)$$

where function $\uplus$ *is a function for transitive operation on trust relation over the set of agents connected to agents* AG_i *and* AG_N *through the path P.* $\uplus$ *combines trust relation along a path P*

There can be several possible paths between two agents. For instance, if the two paths between Ag_i and Ag_n are

$$P' = \langle Ag_i, Ag'_{i+1}, ..., Ag_n\rangle \text{ and } P'' = \langle Ag_i, Ag''_{i+1}, ..., Ag_n\rangle$$

and

$$tr(Ag_i, Ag_n)' = tr(Ag_i, Ag'_{i+1}) \uplus tr(Ag'_{i+1}, Ag'_{i+2}, ..., \uplus tr(Ag'_{n-1}, Ag_n)$$
$$tr(Ag_i, Ag_n)'' = tr(Ag_i, Ag''_{i+1}) \uplus tr(Ag''_{i+1}, Ag''_{i+2}, ..., \uplus tr(Ag''_{n-1}, Ag_n)$$

The trust relation over Ag_i and Ag_n is then given as:

$$tr(Ag_i, Ag_n) = tr(Ag_i, Ag_n)' \oplus tr(Ag_i, Ag_n)''$$

where $\oplus$ is a function for antisymmetric operation on trust relation over the set of agents connected to Ag_i and Ag_n through the paths P' and P''. $\oplus$ combines trust relations along two paths.

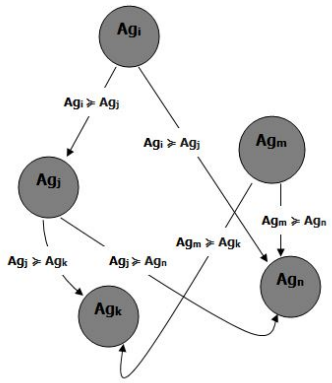

Fig. 1. Example trust graph

2.2 Trust Graph

Given this graph theory, we can now represent a set of arguments exchanged in a dialogue and the set of attacks between them as well as the trust ratings of their sources.

3 The System

We consider a system where dialogue participant i is modelled through a *commitment store* $CS_i \subseteq A$, containing a set of arguments. At any point in time, a participant may *add* or *retract* arguments from their commitment store. An argument may be added to a commitment store if it is not already present within it (and was not previously present), and may be retracted only if it was already present. We also consider the *universal commitment store* $UCS = \bigcup_i CS_i$. The dialogue then consists of a sequence of *add* and *retract* moves, where each move references both an argument and a dialogue participant (e.g., $add(\alpha, a)$ denotes that a adds an argument α to their commitment store.

Each dialogue participant also has an associated trust rating, encoded through a preference ordering over all dialogue participants (this preference ordering is represented by the relationship $\succeq$). Given a universal commitment store, a set of *attacks* between arguments[1], and a preference ordering over dialogue participants, we can instantiate an abstract argumentation framework by transforming attacks into defeats as done in [2]: argument a defeats an argument b iff a attacks b and there are some dialogue participants α, β such that $a \in CS_\alpha, b \in CS_\beta$ and $\alpha \succeq \beta$.

Our approach is based on the following observations.

– A dialogue participant whose arguments are self-contradicting should be less trusted than a consistent participant.

[1] Like defeats, an attack is a binary relation over arguments.

– A dialogue participant who is unable to justify their arguments should be less trusted than one who can.
– A dialogue participant who regularly retracts arguments should be less trusted than one who does not.

We seek to formalise each of these observations within our framework.

3.1 Self Contradicting Arguments

A dialogue participant i is self contradicting iff there are two arguments $a, b \in CS_i$ such that a attacks b or vice-versa.

We write SC_i to denote the number of self contradicting arguments dialogue participant i has.

3.2 Lack of Justification

There are several ways to formalise a lack of justification. For example, one could consider partial arguments, though these should be distinct from enthymemes. Given the abstract nature of our system, we consider unjustified arguments as those that are defeated, but not reinstated (i.e., those that do not appear within the extension according to the semantics under which the dialogue operates).

Formally, an argument a lacks justification for a dialogue participant i iff $a \in CS_i$ and $a \notin \mathcal{E}(UCS, D)$. Here, $\mathcal{E}$ represents the extension(s) obtained on the argumentation framework $\langle UCS, D \rangle$.

We note again that additional definitions of the lack of justification are possible. For example, one could require that the dialogue participant in question be the one to advance the reinstating argument.

We write LJ_i to denote the number of arguments associated with dialogue participant i that lack justification.

3.3 Argument Retraction

The number arguments retracted by an dialogue participant i is denoted AR_i.

3.4 Other Properties

Increase in trust is equally important in dialogues as an incentive for consistent and trustworthy behaviour. In this regard, we consider some factors that can allow for increase in trust assigned to a dialogue participant in the course of a dialogue. The two factors considered for increase in trust during dialogues include:

– Void Precedence
– Defence Precedence

3.5 Void Precedence

Trust in a participant who has non-attacked arguments in its commitment store should increase in a dialogue. The degree of such increase can be based on the number of non-attacked arguments the participant has in its commitment store denoted as VP_i.

3.6 Defence Precedence

Trust in a participant should increase for each of the defended arguments in its commitment store. Similarly, the degree of such increase can be based on the number of defended arguments in the participant's commitment store denoted as DP_i.

3.7 Dynamic Trust

At any point in the dialogue, we may update the associated trust graph. The trust relation over the dialogue participants value provides us with a total order — over the participants. In turn, this total order is used to compute defeats at that point in the dialogue.

With dynamic trust rating of dialogue participants, and the use of such ratings for resolving attacks in a dialogue, we attempt to address a particular limitation of the use abstract argumentation formalism in dialogues. The formalism is restrictive for dialogues in that there is no deviation from the turn-taking procedure of attack and counterattack until one party or the other cannot make a further move. In general, it seems reasonable to hold that a dialogue may not reach any conclusion in a reasonable time when all the dialogue participants are allowed to continue to advance arguments and counterarguments as long as they want.

By applying trust computation, we specify that at a particular stage of a dialogue, the less trusted participant cannot advance arguments that will defeat the arguments of a more trusted participant. In view of this, burden of proof is shifted to the less trusted participant to put forward evidences in support of its other claims in a dialogue otherwise its claims may be discounted for the dialogue to reach a conclusion. A participant main claim in a dialogue may be discounted if it fails to defend other sub-claims that defend the claim. This condition is similar to the pattern of argumentation called elenchos or elenchus where a person is examined with regard to his first statement by answering questions and making further statements in the hope that the meaning and the truth-value of his first statement will be determined [4].

4 Evidential Relevance

In our model, dialogue participants and most importantly, the less trusted ones are expected to provide evidence as a vital tool to win an argument. Intuitively,

the set of claims that are adequately backed up with relevant evidences should lead to the justified conclusion of a dialogue. To determine what evidence counts for a claim, we formulate argument schemes for reasoning about the relevance of evidences. Here by *evidence*, we are indicating a factual claim that is acceptable (or almost acceptable) to all the parties in a dialogue. Therefore, the proposed argument schemes are not intended to discount an evidence per se, but to determine when an evidence counts for evaluating the truth of a claim. This requirement is important to discount evidences that are misleading in a dialogue irrespective of whether they are true or not.

We set two criteria namely *credible* and *relevance* for ranking evidences in a dialogue. First, an evidence is credible if it is accepted (or almost acceptable) to be true by all the parties in a dialogue. Second, an evidence that satisfy the first condition must make the claim(s) it supports probable enough based on the set of argument schemes that are used to evaluate its relevance.

5 Argument Schemes for Reasoning about Evidences

Argument schemes represent templates for making presumptive inference, formed by premises supporting a conclusion and critical questions (CQs) that can be put forward against an argument. In general, argument schemes are used to evaluate arguments on two grounds. First, to determine whether an argument is valid or invalid. Second, to determine if an argument is relevant or irrelevant. Our focus is primarily on the latter use of argument schemes for ranking the relevance of evidences.

For each use of an argument scheme by the proponent of an argument in our model, there is a matching set of critical questions appropriate for the scheme. Specifically, an opponent use the critical questions to challenge the proponent's argument and the proponent must reply accordingly. How strong or weak an evidence is evaluated to be depends solely on this process. An evidence may be discounted when its relevance to the truth of the claim it supports is not sufficiently substantiated. In addition, a claim with more relevant evidence will defeat a claim with less relevant evidence and the trust ratings of respective proponents and opponents of such claims will be updated accordingly.

In this work we use the following argument schemes to reason about the relevance of evidences:

- Argument from Self-Contradiction
- Argument fro Retraction
- Argument from Context
- Argument from Time

5.1 Argument from Self-Contradiction

Argument from self-contradiction is adapted from Walton's argument from inconsistent commitment [6] defined as follows:

Initial Commitment Premise: α has claimed or indicated that he is committed to proposition a
Opposed Commitment Premise: Other evidence in this particular case shows that α is not really committed to a.
Conclusion: α's commitments are inconsistent.

And the critical questions are:
CQ1: What is the evidence supposedly showing that α is committed to a?
CQ2: What further evidence in the case is alleged to show that α is not committed to a ?
CQ3: How does the evidence from premise 1 and premise 2 prove that there is a conflict of commitments?
In this model, we adapt this scheme to state that an evidence does not count in a dialogue if the evidence leads to inconsistency in the commitment store of the participant that provide it.

5.2 Example 1:

Let F = ⟨ A, D ⟩ be an AF with A = { a, b, c, d, e, f } and D = { (b, c), (b, a), (c, e), (e, d), (d, a), (f,b) } as shown in figure 2. Let the commitment stores of α and β in the dialogue be CS_α = { b, c, d } and CS_β = { a, e, f } respectively.

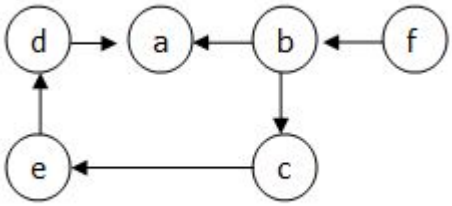

Fig. 2. An argumentation framework

From figure 2, α contradicts its commitments by advancing argument b to attack argument a as b is equally attacking one of its previous argument c. So, β can argue that α cannot provide an evidence to support b in the dialogue using argument from self-contradiction.

5.3 Argument from Retraction

We use *argument from retraction* to evaluate how much latitude a dialogue participant should have in retracting its prior commitments in a dialogue. For

example, if a participant asserts an argument a at a particular stage of a dialogue and went further to advance some sub-arguments, b and c to support a. If the participant retracts a at subsequent stage of the dialogue, it will still need to retract the sub-arguments b and c or other arguments that are closely related to a.

According to [5], the *external stability adjustment* in dialogues requires that once a particular proposition is retracted, some premises or warrant leading into that proposition will also have to be retracted. The method of external stability adjustment therefore add another kind of penalty to retraction of commitments in our model. The first penalty entails that a participant looses trust rating for retraction of commitments and the second penalty in the sense that the participant would have to retract other related arguments to the argument retracted. However, this second penalty may not be applicable at a particular stage of a dialogues if other arguments in the participant's commitment stores are not dependent on the argument it retracted.

The scheme is as follows:
α retracts argument a from its commitment store
Arguments b and c are sub-arguments of a or are closely related to a .
Therefore, α must retracts b and c .

The critical questions to the scheme are:
CQ1: Did α actually retract a from its commitment store?
CQ2: What is the evidence supposedly showing that b and c are closely related to a?
CQ3: Is there a room for questioning whether b or c can be retained even tough a has been retracted?

5.4 Argument from Context

Argument from Context like rhetorical theory of argumentation examines evidences provided in a dialogue based on the specific knowledge base of the participants in the dialogue and the context in which an evidence may be regarded as supporting/attacking a claim. In particular, this scheme helps to capture the rhetorical context of evidences by contextualising evidences to a particular audience and a particular topic of discussion in such a way that what counts as strong evidence in one dialogue may be regarded as weak or irrelevant evidence in another.

The scheme is defined as follows:
Evidence e sufficiently support claim a in context C
Evidence e is advanced to support claim a in context C
Therefore, evidence e is relevant to support claim a in context C.
Critical questions include:
CQ1: Is evidence e supporting claim a in context C?
CQ2: Is there a room to question the relevance of evidence e to claim a even though e is advanced in context C?

For example, the argument *That movie was a bomb, so we shouldnt show it again* expresses quite different arguments in England versus the United States, since in England calling a movie a bomb is to say it was very good, whereas in the United States saying something was bomb is to say it was awful.

5.5 Argument from Time

This argument scheme imposes a time-constraint on when evidence can be considered as relevant to support a claim. For instance, suppose an evidence that *Lionel Messi won the Ballon d' or in 2015* was accepted as a strong support for the claim that *Messi is the best footballer in the world in 2015*. Can the evidence still be relevant for the same claim in 2016? The notion of time in argumentation can as well be likened to rhetorical theory of argumentation where what count as strong evidence in one dialogue may be weak or irrelevant in another.

The scheme is as follows:
Generally, if e occurs in time T, then a will (might) be true.
In this case, e occurs (might occur) in time T.
Therefore, in this case, s will (might be) true.
Critical Questions include:
CQ1: Did evidence e advanced occur in time T?
CQ2: Are there other factors that would or will make evidence e not relevant to support a even though e occur in time T?

6 Discussion

7 Conclusions

References

1. Bollobás, B.: Modern graph theory. Springer Science & Business Media. 184 (2013)
2. Modgil, S., Prakken, H.: A General Account of Argumentation and Preferences. Artificial Intelligence Journal. 361-397 (2012)
3. Paglieri, F., Castelfranchi, C., da Costa P., Falcone, R., Tettamanzi, A., Villata, S.: Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation. Computational and Mathematical Organization Theory. Springer. 20(2), 176-194 (2014)
4. Robinson, R.: Plato's earlier dialectic (1953)
5. Walton, D., Krabbe, E. C.: Commitment in dialogue: Basic concepts of interpersonal reasoning. SUNY press. (1995)
6. Walton, D.: Methods of argumentation Cambridge University Press. (2013)