Noname manuscript No.
(will be inserted by the editor)

Liar liar, pants on fire; or how to use subjective logic and argumentation to evaluate information from untrustworthy sources

Andrew Koster · Ana L. C. Bazzan · Marcelo de Souza

the date of receipt and acceptance should be inserted later

Abstract This paper presents a non-prioritized belief change operator, designed specifically for incorporating new information from many heterogeneous sources in an uncertain environment. We take into account that sources may be untrustworthy and provide a principled method for dealing with the reception of contradictory information. We specify a novel Data-Oriented Belief Revision Operator, that uses a trust model, subjective logic, and a preference-based argumentation framework to evaluate novel information and change the agent's belief set accordingly. We apply this belief change operator in a collaborative traffic scenario, where we show that (1) some form of trust-based non-prioritized belief change operator is necessary, and (2) in a direct comparison between our operator and a previous proposition, our operator performs at least as well in all scenarios, and significantly better in some.

1 Introduction

One of the principal challenges in distributed information systems is to make sense of conflicting information. Whether this is a crowdsourcing system where many contributors send conflicting information, or an application where different sensors send different information; all such systems have to decide what information is truthful, based on a limited model for the provenance of this information. We can see this in the perspective of an intelligent agent having to revise its beliefs upon receiving new information.

Most work on belief revision falls within the framework of the AGM postulates (Alchourrón et al. 1985), which defines desiderata for a *prioritized belief change*

Andrew Koster
E-mail: andrew@andrewkoster.net

Ana L. C. Bazzan
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
E-mail: bazzan@inf.ufrgs.br

Marcelo de Souza
Santa Catarina State University, Ibirama, SC, Brazil
E-mail: marcelo.desouza@udesc.br

operator (in other words, new information is always accepted). However, in practice, new information may be rejected for any number of reasons, one of which is that it conflicts with more credible information in the belief set. Operators that take this into account are called *non-prioritized belief change* operators (Hansson 1999a). Such belief change operators attempt to explicitly deal with uncertainty and untrustworthy information.

In this work we focus on the problem of belief revision in an open multi-agent system, where agents must rely on communication to support their decision-making. We propose a novel belief change operator that is designed specifically to deal with conflicting information from a variety of more or less trustworthy sources. We describe its theoretical foundations, and empirically show some of its advantages in a collaborative traffic scenario, such as Waze.

These scenarios rely on many different drivers to report salient situations — e.g. speed traps, dangerous situations and intense traffic. They combine these reports with both data that is automatically collected by the same application, and information provided by road authorities, to maintain a map with up-to-date information about travel time and important incidents. In such a collaborative traffic application, reported incidents are often wrong, due to the dynamicity of the environment, simple error, or a malicious act (Ben Sinai et al. 2014), thus demonstrating the need for new models of belief revision that take the trustworthiness of information into account. In fact, the work of Ben Sinai et al. makes the use of trust in belief revision all the more urgent: they demonstrate a fairly simple and practical sybil attack on a collaborative traffic application; computational trust has shown to be useful in combating this type of attacks (Adler et al. 2008). Examples of computational trust models designed specifically for use in vehicular communication can be found in (Huang et al. 2010; Raya et al. 2008; Zhang 2011). Farah et al. (2013) propose a method for updating information about the world state using belief functions, and discounting information over time. In spirit, this is a similar approach to the one we propose, however it is very specifically engineered for dealing with temporal and locational events, whereas our approach is a more general approach to dealing with untrustworthy information in a broader sense.

2 Background and related work

Non-prioritized belief change operators can be seen as operators that allow an agent a choice in what beliefs to maintain (although they generally still require other axioms, such as maintaining consistency). As such, over the years a number of different methods have been proposed for choosing what beliefs should be held after a revision operation (Falappa et al. 2011; Hansson 1999b). A recent focus for this is to use argumentation for choosing these beliefs (Krümpelmann et al. 2012; Pereira et al. 2011; Tang et al. 2012). Nevertheless, such methods assume that a certain ordering, generally based on a confidence in the predicate's truthfulness, is established. In both Tang et al. and Pereira et al.'s work this is explicitly based on trust, just as we propose to do in this work. Tang et al. work on a problem that is orthogonal to ours: their work establishes an abstract framework in which the trustworthiness of agents and of information is simultaneously evaluated. In order to do this, they make some assumptions about the nature of trust (in particular that it is transitive) that we do not require. Furthermore, they assume that every

proposition has a single source, and do not provide a mechanism for dealing with the situation where multiple agents communicate the same proposition (or its negation), and the credibility of a single piece of information must be incorporated into this framework. Pereira et al.'s model is more similar in its aims to our own, and proposes to use possibilistic logic to evaluate the credibility of propositions. Nevertheless, this method has a problem when multiple sources give conflicting information: for instance, if three sources with a trustworthiness of 0.8 state $\neg a$, and one source with a trustworthiness of 0.9 states a, their process will end with a being believed, albeit with a low confidence (0.55). A more intuitive outcome of this situation would be to believe $\neg a$: three trustworthy sources state that $\neg a$ is true, as opposed to one, slightly more trustworthy source, stating a. In order to achieve this, a more principled approach to aggregating information from different sources is required.

Thus the related works in this area are lacking a correct treatment of aggregating the credibility of similar information from different sources, based on the trustworthiness of the source. This problem has been studied in the area of information fusion, and a large number of approaches have been proposed to take uncertainty into account explicitly (Dubois and Prade 1988; Jøsang 2002; Shafer 1976; Smarandache and Dezert 2005) when aggregating information from various different, possibly conflicting, sources. Of these, subjective logic (Jøsang 2002) stands out to us, because of its stochastic foundations (it is at its core a Bayesian update function), and because of its elegant treatment of the various problematic corner cases often used to illustrate faults in these theories (Jøsang 2012). Nevertheless, Information Fusion methods do not deal with logical inconsistency of different statements, and thus cannot be considered as full-fledged belief change operators.

We propose to use subjective logic in a novel belief change operator. In this, we follow the philosophy of Data-Oriented Belief Revision (Paglieri and Castelfranchi 2005), with two separate informational categories. An agent receives (large quantities of) conflicting data, and a belief change operator is used to "persuade the agent to believe" some set of this data over the rest. Our operator incorporates subjective logic into an argumentation framework, so as to ensure the resulting beliefs are consistent. The reason for using an argumentation framework is that it is a logical framework that is designed to deal with contentious information: it allows for inconsistent input, models the conflicts, and provides a method for resolving them. The number of abstract argumentation frameworks has exploded since the idea was first proposed (Dung 1995). We choose to base our framework on the abstract argumentation framework of Amgoud and Vesic (2014). We choose this because the framework is concise and simple, yet unlike Dung's original framework (Dung 1995), it allows for the explicit treatment of a preference ordering over arguments, which allows us to represent the credibility of arguments. There are other abstract argumentation frameworks that take preferences into account, such as those proposed by Amgoud and Cayrol (1998) and Bench-Capon (2003), however Amgoud and Vesic point out a number of flaws in how these frameworks deal with preferences; for our case in particular, it is important that if an argument A attacks argument B, but B is preferred over A that not only is B not defeated, but the attack relation is reversed, a property of the framework we chose, but not of any other preference-based argumentation framework. This is important, because other frameworks can result in sets of arguments that are not conflict-free being

admissible. Due to the way we construct arguments, conflicting arguments being admitted will result in a logically inconsistent belief set; whereas it is trivial to prove that if the admissible set of arguments is conflict-free, the set is logically consistent. Because beliefs being logically consistent is a strongly desirable property of a belief revision operator, it is important to use the right argumentation framework.

3 Preliminaries

Before we present our belief change operator, we summarize the theories that we use for its definition: subjective logic, and Vesic and Amgoud's argumentation framework.

3.1 Subjective logic

Subjective logic is a type of probabilistic logic that explicitly takes uncertainty and belief ownership into account (Jøsang 2002, 2012). Subjective logic is defined both for the binomial and the multinomial case. In our case the binomial case is sufficient, but the extension to a multinomial belief function is trivial.

Definition 1 (Binomial opinion)
Let φ be a proposition. Agent A's **binomial opinion** about the truth of φ is the tuple $\omega_\varphi^A = (b, d, u, a)$, where b is the degree of belief that φ is true, d the degree of belief that φ is false, u the uncertainty about the probability of φ, and a the prior probability of φ in the absence of information. These components satisfy the conditions that $b, d, u, a \in [0, 1]$ and $b + d + u = 1$. Furthermore, the **expected value**, or agent A's degree of certainty in the truth of φ is: $p(\omega_\varphi^A) = b + a \cdot u$.

Jøsang et al. provide two fusion operators for binomial opinions of this kind: cumulative and averaging fusion. The latter is intended for situations where the observations are dependent. The former is useful when different observations (or opinions) are independent, and thus it is what is needed in our work.

Definition 2 (Cumulative fusion)
Let φ be a proposition and $\omega_\varphi^A = (b^A, d^A, u^A, a^A)$ and $\omega_\varphi^B = (b^B, d^B, u^B, a^B)$ be two different opinions about φ. The cumulative opinion

$$\omega_\varphi^{A\oplus B} = (b^{A\oplus B}, d^{A\oplus B}, u^{A\oplus B}, a^{A\oplus B})$$

is computed as:

If $u^A \neq 0$ or $u^B \neq 0$:

$$b^{A\oplus B} = \frac{b^A u^B + b^B u^A}{u^A + u^B - u^A u^B} \qquad d^{A\oplus B} = \frac{d^A u^B + d^B u^A}{u^A + u^B - u^A u^B}$$

$$u^{A\oplus B} = \frac{u^A u^B}{u^A + u^B - u^A u^B} \qquad a^{A\oplus B} = \frac{a^A u^B + a^B u^A - (a^A + a^B) u^A u^B}{u^A + u^B - 2u^A u^B}$$

Elsewise, in which case both $u^A = 0$ and $u^B = 0$, a different formula is used that deals with crisp beliefs (those without any uncertainty). For the details we

refer to (Jøsang 2012). The cumulative fusion operator is commutative, idempotent and associative.

Finally, we need to define how to aggregate opinions over a conjunction of different propositions; this will also be needed in our belief change operator.

Definition 3 (Conjunction of opinions)
Let φ, ψ be proposition and $\omega_\varphi = (b_\varphi, d_\varphi, u_\varphi, a_\varphi)$ and $\omega_\psi = (b_\psi, d_\psi, u_\psi, a_\psi)$ be an agent's opinion regarding them. The conjunctive opinion

$$\omega_\varphi \owedge \omega_\psi = \omega_{\varphi\wedge\psi} = (b_{\varphi\wedge\psi}, d_{\varphi\wedge\psi}, u_{\varphi\wedge\psi}, a_{\varphi\wedge\psi})$$

is computed as follows:

$$\begin{aligned}
b_{\varphi\wedge\psi} &= b_\varphi b_\psi + \frac{(1-a_\varphi)a_\psi b_\varphi u_\psi + a_\varphi(1-a_\psi)b_\psi u_\varphi}{1-a_\varphi a_\psi} \\
d_{\varphi\wedge\psi} &= d_\varphi + d_\psi - d_\varphi d_\psi \\
u_{\varphi\wedge\psi} &= u_\varphi u_\psi + \frac{(1-a_\psi)b_\varphi u_\psi + (1-a_\varphi)b_\psi u_\varphi}{1-a_\varphi a_\psi} \\
a_{\varphi\wedge\psi} &= a_\varphi a_\psi
\end{aligned}$$

For the justification and properties of these operators we refer to the original works (Jøsang 2002, 2012).

3.2 Rich preference-based argumentation framework

We now give a very brief overview of Amgoud and Vesic's rich preference-based argumentation. We repeat definitions from the original works, but assume the reader has a basic knowledge of argumentation theory.

Definition 4 (Preference-based argumentation framework)
A preference-based argumentation framework is a triplet $\langle \mathcal{A}, \mathcal{R}, \succeq \rangle$ where $\mathcal{A}$ is a set of arguments, $\mathcal{R}$ is a binary relation between arguments $\mathcal{A}$, such that $A\mathcal{R}B$ means that A attacks B, and $\succeq$ is a pre-order of arguments $\mathcal{A}$ such that $A \succeq B$ means that A is preferred over B.

Given a preference-based argumentation framework, a rich argumentation framework can be defined as follows:

Definition 5 (Rich preference-based argumentation framework)
Given a preference-based argumentation framework $\langle \mathcal{A}, \mathcal{R}, \succeq \rangle$, the rich preference-based argumentation framework is defined as $\langle \mathcal{A}, \mathcal{R}_r \rangle$ where $\mathcal{R}_r = \{(a,b)|\ (a,b) \in \mathcal{R} \wedge a \succeq b\} \cup \{(b,a)|\ (a,b) \in \mathcal{R} \wedge b \succeq a\}$.

In other words, if an argument a attacks argument b, but b is strictly preferred over a, then the relation is inverted in the rich framework. In Amgoud and Vesic's work, the rich preference-based argumentation framework is more complex than this: they also show how the preference relation gives rise to a refinement relation between extensions of the argumentation framework. However, this is not necessary for our work. We choose to adopt skeptic semantics with regards to the

acceptability of arguments, but note that in most cases, there will only be a single stable extension, as proved in property 5 of (Amgoud and Vesic 2014). We elaborate on this, and on how exactly the preference ordering is computed in the next section.

Definition 6 (Stable extensions)
Let $\langle \mathcal{A}, \mathcal{R}_r \rangle$ be a rich preference-based argumentation framework. $\mathcal{E} \subseteq \mathcal{A}$ is a stable extension iff it is conflict free ($\nexists a, b \in \mathcal{E} : a \mathcal{R}_r b$) and attacks any argument in $\mathcal{A} - \mathcal{E}$ ($\forall b \in \mathcal{A} - \mathcal{E} : \exists a \in \mathcal{E} : a \mathcal{R}_r b$).

4 Proposed Belief Revision Operator

Our proposed approach for dealing with untrustworthy information is to update an agent's beliefs using Data-Oriented Belief Revision Operator, that uses trust, subjective logic and argumentation to compute an agent's beliefs based upon received information. First we give the formal specification of this operator, and describe the intuitions behind it, and second we provide details on how the operator can be implemented.

4.1 Specification

In order to decide what to believe, all non-prioritized belief change operators perform three steps:

1. Collect information, potentially from multiple sources, with associated trustworthiness.
2. Aggregate the information in some form.
3. Decide (select) what information to believe.

Mostly, these steps are considered individually, with work in the fields of knowledge acquisition and computational trust providing ways of performing step (1), whereas information fusion operators provide methods for step (2). Work on belief revision operators focuses mostly on step (3). In contrast, we propose a method that integrates steps (1) and (2) into the belief revision process. For this, we argue that a trust model should be used to assess the trustworthiness of a source, and use this to assess the credibility of the information received. Our model is agnostic to the specifics of the trust model, and we henceforth assume that it has been applied, and the trustworthiness of each individual source is known. Following a common assumption in the literature, we assume that the credibility of a communicated message is equal to the trustworthiness of the sender. In theory, our model is mostly agnostic to the underlying logic: we require strong negation to compute *opinions* in subjective logic (see below), and we require the logic to be *sound* and *complete*. For instance, a first-order logic, or some description logics could be used instead of the proposition logic that we use in the remainder of this work. However, in practice, the complexity of using an argumentation framework with a richer logic may lead to computational intractability.

Step (1) thus results in a multiset $\mathcal{I}$, defined such that for any proposition φ, communicated by a sender with trustworthiness v, $(\varphi, v) \in \mathcal{I}$. Note that for a

single proposition φ we may have multiple entries in $\mathcal{I}$, and even exact duplicates. Furthermore, we may have entries for its negation, $\neg\varphi$ as well. Moreover, we define the set of propositions $\mathcal{P} = \{\varphi | (\psi, v) \in \mathcal{I} \wedge \varphi \equiv \psi\}$. In other words, $\mathcal{P}$ contains a single representation for a proposition and all equivalent propositions that have been communicated.

In step (2) we compute the trustworthiness of each proposition $\varphi \in \mathcal{P}$, by aggregating together the information from different sources. For any proposition $\varphi \in \mathcal{P}$, let $O^+_\varphi = \{(v, 0, (1-v), a_\varphi) | \forall v : (\psi, v) \in \mathcal{I} \wedge \varphi \equiv \psi\}$ and $O^-_\varphi = \{(0, v, (1-v), a_\varphi) | \forall v : (\psi, v) \in \mathcal{I} \wedge \varphi \equiv \neg\psi\}$, where the prior probability a_φ is specified by the designer; e.g. if unknown or unimportant a value of 0.5 can be chosen, but if known it can be specified appropriately. We specify

$$\omega_\varphi^{O^+ \cup O^-} = \omega_1 \oplus \omega_2 \oplus \cdots \oplus \omega_n$$

with $\omega_1, \ldots \omega_n \in O^+_\varphi \cup O^-_\varphi$, and $\oplus$ the cumulative fusion operator as defined above. Note that if (b, d, u, a) is the fused opinion of φ, then (d, b, u, a) is the fused opinion of $\neg\varphi$. We use the cumulative fusion operator because we are fusing differing opinions of a single propositions together.

As an example of this process, let us flesh out our example from the related work section, with four different sources (here, cars). Car A, with trustworthiness 0.9, reports that the road is clear, while cars B, C and D, all with trustworthiness 0.8, report that the road is not clear (congested). In this case, the fused opinion $\omega_{clear}^{A \oplus B \oplus C \oplus D} = (0.406, 0.549, 0.045, 0.5)$, assuming the prior knowledge about the road is 0.5. Similarly, we could compute the fused opinion for $\neg clear$ and obtain $\omega_{\neg clear}^{A \oplus B \oplus C \oplus D} = (0.549, 0.406, 0.045, 0.5)$.

Now in step (3) we map the arguments.

Definition 7 (Arguments)
Let $\mathcal{P}$ be the propositions from step (1), and $\models$ the entailment relationship of propositional logic. The set $\mathcal{A}$ is the set of all arguments that can be generated from $\mathcal{P}$ such that $(H, h) \in \mathcal{A}$ iff $H \subseteq \mathcal{P}$ and $H \models h$, and additionally $\forall H' \subset H : H' \not\models h$.

Note that this is the theoretical set of arguments. In practice, a far more pragmatic approach will be required to keep the problem tractable. Thus instead of starting from $\mathcal{P}$ and computing all possible arguments, a practical solution will start from the set of interesting information (for instance, information necessary to achieve an agent's desires) and construct all arguments from $\mathcal{P}$ with respect to this information. This will result in a belief base, instead of a belief set, as would be the case if all possible arguments based on $\mathcal{P}$ were to be taken into account.

Definition 8 (Undercut and rebuttal)
An argument $(H, h) \in \mathcal{A}$ attacks $(H', h') \in \mathcal{A}$ iff either:

- $\exists \Phi \subseteq H' : \Phi \cup \{h\} \models \bot$, known as an undercut attack.
- $\{h, h'\} \models \bot$, known as a rebuttal attack.

If an argument A undercuts or rebuts A' then $A \mathcal{R} A'$.

Continuing our example, we construct two arguments about the state of the road: $\alpha = (\{clear\}, clear)$ and $\beta = (\{\neg clear\}, \neg clear)$. These two arguments attack each other: they both rebut the other argument, and thus we have $\mathcal{R} = \{(\alpha, \beta), (\beta, \alpha)\}$.

So far, the arguments do not take into account the trust in the underlying communication, or the binomial opinions based on this trust as computed in step (2). The trustworthiness of information is represented in the preference ordering over the arguments. Because we are fusing opinions of different propositions together, we use the conjunctive fusion operator. We thus define the preference order of an argument as follows:

Definition 9 (Preference ordering)
For an argument $A \in \mathcal{A}$ with support H, we define its level as the expected value of the conjunctive opinion over all propositions in H.

Let $H = \{\varphi_1, \varphi_2, \ldots, \varphi_n\}$, then we define

$$\omega_{\wedge H} = (b_{\wedge H}, d_{\wedge H}, u_{\wedge H}, a_{\wedge H}) = \omega_{\varphi_1} \mathbin{⍓} \omega_{\varphi_2} \mathbin{⍓} \cdots \mathbin{⍓} \omega_{\varphi_n}$$

The level of argument A is thus $level(A) = p(\omega_{\wedge H}) = b_{\wedge H} + u_{\wedge H} a_{\wedge H}$.

We define the preference relation as follows: let $A, B \in \mathcal{A}$, then $A \succeq B$ iff $level(A) \geq level(B)$.

An argument framework for our belief change operator is thus a tuple $\langle\ \mathcal{A},\ \mathcal{R},\ \succeq\ \rangle$ with $\mathcal{A}, \mathcal{R}$ and $\succeq$ as previously defined. Given such an argument framework, the rich argument framework can be constructed. It is easy to prove that there will only be more than one stable extension if multiple conflicting arguments have the exact same credibility. In practice, this should hardly ever occur, but if it does, we argue that a skeptic approach should be taken: only accept those arguments that are in *all* stable extensions, which in general will lead to rejecting all of the mutually conflicting arguments. Intuitively this makes sense, because there is no way for an agent to consistently decide between equally credible conflicting arguments.

This final step results in a set $\mathcal{B}$, the beliefs of the agent, defined as the union of the conclusions of all the skeptically accepted arguments in the stable extensions of $\langle\ \mathcal{A},\ \mathcal{R},\ \succeq\ \rangle$.

Demonstrating the argumentation in our example, we see that $level(\alpha) = 0.406 + 0.045 * 0.5 = 0.43$. Similarly, $level(\beta) = 0.57$, and thus $\beta \succeq \alpha$. This results in the repaired attack relation $\mathcal{R}_r = \{(\beta, \alpha)\}$, and the framework has a single stable extension: $\{\beta\}$. $\neg clear$ is the conclusion of argument β, and thus we have the resulting belief base: $\{\neg clear\}$. The agent thus believes the road is congested. Contrast this with the original resolution of the same scenario described in Section 2, using Pereira et al.'s operator, which resulted in the agent believing a, or in our traffic example, that the road is clear.

As discussed in further depth by Falappa et al. (2009), argumentation-based revision is not easily considered in any of the non-prioritized postulate systems for belief revision. It is easy to show that our framework does not fit well with either of the main approaches *expansion + consolidation* or *decision + revision*. It is our belief that our proposed operator is best described by what Hansson (1999a) calls an *integrated choice* operator. However, it is ongoing work to define a set of postulates for this class of operators. There clearly is a set of governing principles, both for our operator, and for other data-based revision methods (Paglieri and Castelfranchi 2005). For starters, we maintain consistency of the agent's beliefs after revision.

4.2 Implementation

Let $\mathcal{X}$ be a set of agents. Any $y \in \mathcal{X}$ can send a proposition φ to other agents in $\mathcal{X}$. When x receives proposition φ from agent y, x updates its information base $\mathcal{I}' = \mathcal{I} \cup \{\langle \varphi, y, t \rangle\}$, where t is the time at which the information was received. With this, x can compute the trustworthiness v of this individual message, using its computational trust model to compute the trustworthiness of source y, as we proposed. Whenever the agent's reasoning system requires a piece of information, the belief revision process is run, as described in the previous section. This is implemented as in Algorithm 1. The complexity of this algorithm is largely dependent on the choice of what arguments are constructed from an information base, using the **constructargs** method. The choice for this method is domain dependent. We will see in the next section that we construct 8 arguments from 8 propositions, however in a more general case this may be more complex. This is especially the case if the underlying logic is a predicate logic instead of a propositional one. In the most general case, the number of arguments is combinatorial (or worse) in the size of $\mathcal{I}$. However, in practice, this does not seem to occur. If we start from a given set of arguments $\mathcal{A}$ of size m, and an information base $\mathcal{I}$ of size n, the complexity of the Algorithm 1 is $O((n+m)^2)$. Nevertheless, this also does not take into account the computation of the stable extensions of an argumentation framework on line 24, for which we refer to the literature on argumentation. One popular method is to use reinstatement labelling (Caminada 2006). Nevertheless, this problem has been shown to be *co-NP-complete* (Dunne and Wooldridge 2009) in the worst case. This is expected, as belief revision, in general, is either *NP-complete* or *co-NP-complete*, depending on the exact problem specification.

The method **trust** for computing the trustworthiness of an agent in a given trust model, is domain dependent (just as **constructargs** is). For the experimental scenario we describe these methods in the next section. For ways of computing the stable extensions of an argumentation framework (**stable**), we refer to the literature on argumentation.

The proof that Algorithm 1 is a correct implementation of the belief revision operator described in the previous section is trivial, when realizing that line 11 applies the cumulative fusion operator of Definition 2 to every proposition that is stored in the agent's information base $\mathcal{I}$, and subsequently lines 12–17 apply the conjunctive operator of Definition 3 to attain the confidence in an argument, needed to construct the preference ordering of Definition 9. Subsequently, the algorithm iterates over all arguments in $\mathcal{A}$ and constructs the attack relations (on lines 22 and 23) as defined in Definitions 5 and 8. Finally, on line 24 the skeptically accepted arguments of stable extensions of the argument framework are computed and returned as the agent's belief base.

In addition to our own operator, we implemented two other belief revision operators. The first does not use trust; the second is an implementation of Pereira et al.'s operator:

Without trust The strength of a proposition is simply equal to the number of sources communicating it. An argument's level is equal to the minimum strength of the propositions in the support. For instance, if ten sources say a and six sources say b, then $level((\{a, b\}, a \wedge b)) = 6$. Other than that, it works the same as the full belief change operator above.

Algorithm 1: Belief revision algorithm

Data: trust (a computational trust model), $\mathcal{I}$ (agent's information base)
Result: $\mathcal{B}$: the agent's beliefs

$\mathcal{A} :=$ **constructargs**$(\mathcal{I})$
$\Phi := \{\varphi \mid \exists y, t' : \langle \varphi, y, t' \rangle \in \mathcal{I} \vee \langle \neg\varphi, y, t' \rangle \in \mathcal{I}\}$
for $\varphi \in \Phi$ **do**
    BinTrust := []
    for $\langle \psi, y, t' \rangle \in \mathcal{I}$ **do**
        $v :=$ **trust**(y)
        case $\psi = \varphi$ **do**
            BinTrust.append((v, 0, 1 - v, a_φ))
        case $\psi = \neg\varphi$ **do**
            BinTrust.append((0, v, 1 - v, 1 - a_φ))
    $(b, d, u, a) := \bigoplus$ BinTrust
    for $(A, \alpha) \in \mathcal{A}$ *such that* $\varphi \in A \vee \neg\varphi \in A$ **do**
        let $c =$ ***confidence****(A)* **in**
            if $c = None$ **then**
                confidence(A) $:= (b, d, u, a)$
            else
                confidence(A) $:= c \oslash (b, d, u, a)$
$\mathcal{R} := \emptyset$
for $(A, \alpha) \in \mathcal{A}$ **do**
    for $(B, \beta) \in \mathcal{A} - \{(A, \alpha)\}$ **do**
        let $(b_A, d_A, u_A, a_A) =$ ***confidence****(A)* ***and*** $(b_B, d_B, u_B, a_B) =$ ***confidence****(B)* **in**
            if $((\{\alpha, \beta\} \models \bot) \vee (\{\alpha\} \cup B \models \bot))$
                $\wedge\ b_A + u_A a_A \geq b_B + u_B a_B$ **then**
                $\mathcal{R} := \mathcal{R} \cup ((A, \alpha), (B, \beta))$
$\mathcal{B} := \{\alpha | (A, \alpha) \in \bigcap$ **stable**$(\mathcal{A}, \mathcal{R})\}$

Pereira et al.'s operator Briefly speaking Pereira et al.'s operator relies on a fuzzy reinstatement labeling algorithm to compute the confidence of the final arguments. This is implemented in an iterative process computing the following equation (Eq. 7 in the original work): $\alpha_{t+1}(A) = \frac{1}{2}\alpha_t(A) + \frac{1}{2} min\{\mathcal{L}(A), 1 - max_{B:B\mathcal{R}A}\alpha_t(B)\}$, where $\alpha_t(A)$ is the confidence in argument A in iteration t and $\mathcal{L}(A)$ is the initial confidence value assigned to the argument at $t = 0$. Their framework differs from ours in that they assume sources communicate arguments, rather than statements, and $\mathcal{L}(A)$ is simply the trustworthiness of the most trustworthy source that communicated argument A. This equation is computed for all the arguments in $\mathcal{A}$, and they prove it converges. Their method for subsequently choosing what arguments to believe relies on a possibilistic interpretation of this confidence value, but in our experiment scenario this is the same as using the confidence α in an argument as its level and proceeding with the argumentation framework introduced before.

5 Experimental scenario and results

To compare the performance of the three methods, experiments were performed, in which we simulate communication in a real road scenario. We use Netlogo (Wilenski 1999) to model and simulate this scenario, and we describe the model and simulation below.

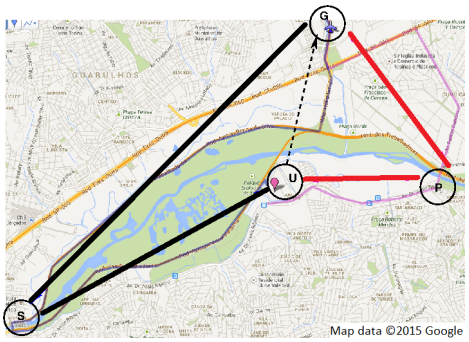


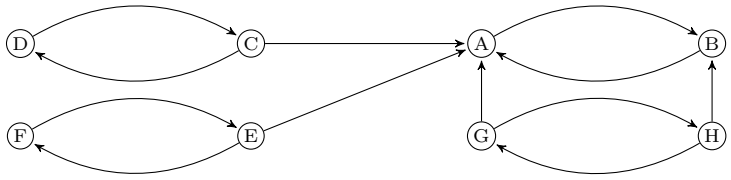
Fig. 1 Map of city C, with alternatives from G to U.

5.1 Agent-based Traffic Model

The model consists of a road network, as seen in Figure 1, with two alternative paths, GPU and GSU, to connect location G to U. The travel time along each path is dependent on the traffic volume along these roads, and computed according to the fundamental diagram of traffic flow (Greenshields 1935). The set of driver agents commuting from G to U are denoted with $\mathcal{X}$. These agents represent cars that can decide for themselves which of the two edges to pick to go from G to U. There are many other drivers that make up the environment, but since these are not going from G to U, they are treated as background traffic. However, they do influence the travel time along the paths, and thus the travel time of the agents in $\mathcal{X}$. There are three different classes of agents: *private drivers*, *professional drivers* and *authority drivers*, and every agent belongs to exactly one class. These different classes are used to initialize the simulation with heterogeneous behaviours, and they play a part in the trust model. However, other than that, all agents function in the same manner.

The simulation runs for 100 iterations. In each iteration, an agent decides, based on its beliefs about the state of the roads, to drive along either GSU or GPU. Both paths present problems. GSU has more traffic, and is generally congested. GPU is a peripheral path, and although the distance is shorter, it is an area that the majority of users assume to be poor and dangerous. Although this is not true, most people are not comfortable taking this path and thus GSU is congested while GPU is clear. In case a collaborative traffic system is used, our belief operator can make agents in $\mathcal{X}$ decide to travel along GPU. This is a challenge, however, when an agent receives conflicting information, especially about the dangerousness of P.

It is reasonable to assume that not all agents communicate and that not all have access to the same information. Furthermore, in traffic scenarios, storing trust values for each individual agent does not make sense because drivers seldom meet again. However, agents can be grouped into *classes* of sources, and a trustworthiness can then be associated with an entire class (this is sometimes called stereotyping (Burnett et al. 2010)). This is not far from what happens in reality: as decision-makers in traffic, we often associate different levels of trustworthiness to different classes of actors, such as taxi drivers, trucks or traffic police. This influences the confidence we have on information coming from these sources.



A: $(\{cS, \neg cP, \neg dP, cS \wedge \neg cP \wedge \neg dP \rightarrow \neg bS\}, \neg bS)$
B: $(\{dP, dP \rightarrow bS\}, bS)$ C: $(\{cP\}, cP)$ D: $(\{\neg cP\}, \neg cP)$
E: $(\{\neg cS\}, \neg cS)$ F: $(\{cS\}, cS)$ G: $(\{dP\}, dP)$ H: $(\{\neg dP\}, \neg dP)$

Fig. 2 Arguments (nodes) generated by the driver agents and the attack relations (arrows) between them.

These classes are used to refine the trust model that we adapt from (Koster et al. 2013), a simple trust model that is specifically designed for communication in traffic. The trust in a class of sources, and thus for any individual member of that class, is simply the expected likelihood that communication from members of that class is truthful, conditioned upon the past messages received from all members of that group. This is, to summarize very briefly the statistical model behind it, computed by the formula: $Trust(C) = \frac{|Truthful(C)|+1}{|Total(C)|+2}$, where C is a class of agents, $Truthful(C)$ the set of truthful messages received from members of class C, and $Total(C)$ the set of all messages, both truthful and untruthful.

The agents base their decision of which road to take on their beliefs regarding three propositions: cP, cS and dP, representing the congested state of the paths through P and S, and whether P is dangerous or not. The initial binomial opinions in these propositions are (0, 0, 1, 0.5) for the first two, because we assume agents have no a priori knowledge about the traffic situation on either path, and (0, 0, 1, a_d) for the last one. The value a_d is the level of discomfort a driver has with taking the peripheral path P, and we model this as the a priori confidence that he has in the belief that the road is dangerous. For each agent, the value of a_d is chosen from a normal distribution with tunable parameters.

These propositions are combined with a number of rules, to decide which of the two paths are best. In particular, the following strict rules are used. A strict rule is simply a predicate with binomial opinion (1, 0, 0, 1). The set of strict rules are: $dP \rightarrow bS$ and $\neg cP \wedge cS \wedge \neg dP \rightarrow \neg bS$. bS represents that S is the best path. The driver will pick S if bS, P if $\neg bS$ and will pick at random if the reasoning is inconclusive about what route is best.

In Figure 2 we graph the resulting argumentation framework. Here we see both types of attack, undercut and rebuttal. Argument A rebuts argument B, and vice versa, because $bS \wedge \neg bS \models \bot$. This results in two attack relations being added and thus two arrows in the graph; one each way. However, for undercut attacks there is only one arrow. Argument C undercuts argument A, because $\neg cP$ is in the support of argument A, and $\{cP, \neg cP\} \models \bot$ means that C undercuts A. However, there is no attack the other way, because the conclusion of argument A does not contradict anything in argument C. Nevertheless, the preferences can cause these attacks (and thus arrows) to flip directions, when the rich preference-based argument framework is computed.

We initialize our NetLogo model with 13,000 non-agent cars driving along GSU, and 1,000 cars driving along GPU, which causes GSU to be congested, and

GPU to be clear. We run the simulation with 300 agents, of which 200 are private drivers, 70 are professional drivers, and 30 are authority drivers. Moreover, we are interested in a scenario where untrustworthy agents negatively affect the system. Thus, we simulate what happens if liar agents misuse the communication channel to disseminate false status of traffic conditions and/or false information about the dangerousness of P. Liar agents always communicate the opposite of the truth. We tuned the percentage of liar agents (per class) to ensure that a solution without trust will fail to give correct results: in other words, an agent must necessarily decide whether newer information is true. In those situations where we do not need to discard information from untrustworthy sources, a prioritized belief revision operator could be used. This is achieved with 20% of private, 30% of professional and 5% of authority drivers lying. Moreover, the agents can use a sybil attack to propagate their lies, as in Ben Sinai et al.'s work 2014. We emphasize that these values are chosen to tune the scenario to test those situations where trust and belief revision systems can make a system can make a difference. This means that enough agents must lie, and these lies must have a meaningful impact on the system. Our experiments investigate how different choices for belief revision methods impact agents' functioning in such systems.

Summarizing, the simulation is initialized and then runs for 100 iterations. In each iteration the agents do the following:

1. Revise their beliefs according to the model used, and decide whether to drive along GPU or GSU.
2. Receive the truthful situation of the path used (congestion and danger of that path).
3. Compute trustworthiness of the three different agent classes based on the communication in the previous iteration and the information from step 2.
4. If a communicative agent, broadcast a message (truthful or a lie, depending on configuration). Selected agents receive messages. This is randomized at each iteration.

For further details on the model, we have described it fully following the ODD protocol (Grimm et al. 2006)[1] in the work by Souza et al. (2015).

5.2 Results

In the simulation, agents repeatedly travel between G and U. Given the particular parameters of the scenario, the best choice for every agent is to drive along P.

For the sake of comparing the role of different components of the model, several cases were devised, as outlined in Table 1. Specifically, we see that cases *0–2* disregard discomfort with the periphery. In cases *0* and *3*, all agents have perfect information about traffic conditions on both routes. While this assumption is not realistic, these cases serve as benchmark for comparing the belief revision operators to.

As the performance measure of the experiments, we use the percentage of agents in $\mathcal{X}$ that travel along path P, because since it is clear while S is congested,

[1] The ODD protocol was designed to standardize the method for describing Agent-Based Models in order to facilitate understanding and duplication, and thus solves a number of the problems existing in simply sharing a code base.

Case	(a)	(b)	(c)	(d)
0	√	-	-	-
1	-	-	-	-
2	-	-	√	-
3	√	√	-	-
4	-	√	-	-
5	-	√	√	-
6	-	√	√	√

Table 1 Simulation cases — (a) agents have complete information on congestion; (b) discomfort is considered; (c) liar agents; (d) dangerousness of P is communicated.

the traffic from G to U is better off using P. Each experiment was repeated 100 times; Table 2 shows the mean of these runs. The variance in some of the results is due to a subset of agents choosing their route randomly.

In *case 0*, all agents have perfect information and choose P. In *case 1*, with only true information being communicated, all methods obtain the optimal output.

Case 2 is the first interesting case, with false information causing the simple majority rule to fail: because of the way we chose the parameters, a very slight majority of the information received will be false, and thus the system breaks down. This case also uncovers an interesting shortcoming of Pereira et al.'s model, which is that if multiple equally trustworthy agents communicate conflicting information, then they believe neither.

This causes the model to choose randomly between the routes. Our model, on the other hand, combines the majority rule and the trust in the subjective logic operator and reaches the correct conclusion: P is clear.

When we add a discomfort level a_d (an a priori confidence in the belief that the periphery is dangerous), we see that things are worse still. In *case 3* agents have perfect knowledge about the congestion on both routes, but not about the danger level. Only 15% of the agents believe P is not dangerous. Case *3* is a baseline for cases *4, 5* and *6*. In *case 4* as in *case 1*, with truthful communication about congestion, all methods converge to the baseline value. Also as expected, *case 5* follows a similar pattern as *case 2*.

In *case 6* we add communication, including liars, about the danger in P. Note that, in addition to communicating about traffic, liar agents will (falsely) assert that the peripheral route is dangerous. We firstly see that our approach manages to correctly filter out the false information and eventually all agents learn that P is not dangerous, and 100% of the agents choose P. Interestingly, the inverse happens with Pereira et al.'s model: agents that initially believed P was not dangerous (the 15% of *case 3* and *4*) end up not knowing what to believe or even believing it is dangerous.

6 Discussion

As we described at the start of the previous section, the experimental scenario was tuned to demonstrate the interesting properties of our model. Nevertheless we should mention that if we perform a grid search over the parameters there are no situations that our model performs worse than either of the other methods; belief revision ignoring trust and using Pereira et al.'s model. This result follows our intuition: the use of the cumulative fusion operator has been shown to work well

Case	Operator	% of agents	Std dev
0	-	100	0
1	-	100	0
1	Pereira et al.	100	0
1	Our approach	100	0
2	-	0	0
2	Pereira et al.	50.1	2.81
2	Our approach	100	0
3	-	15.56	2.21
4	-	15.81	1.86
4	Pereira et al.	15.26	1.9
4	Our approach	15.72	1.9
5	-	0	0
5	Pereira et al.	7.75	1.36
5	Our approach	15.67	1.9
6	-	0	0
6	Pereira et al.	1.56	2.61
6	Our approach	100	0

Table 2 Percentage of agents in peripheral path P

in other domains with conflicting information: it takes both the numbers of reports into account as well as the confidence in each of those reports. By only taking the numbers (model without trust) or choosing the single data point with the highest confidence (Pereira et al's model), the other models discard information, however that is not always wrong. There may be other application domains where lots of low confidence data should be treated as noise and ignored (in which case picking the maximum might work better than the cumulative fusion), or domains where the confidence assigned to data points is meaningless noise, and ignoring trust values is the best option. We believe such approaches are nevertheless suboptimal: meaningless noise should ideally be filtered out before the belief revision phase, where we only revise our beliefs based on meaningful information, in which case the number of reports and the confidence in each of them should be aggregated in a principled manner.

Another discussion point is whether operators like these should still be considered belief *revision* operators. Strictly speaking what we do, and what Pereira et al. and other Data-oriented Belief Revision operators do, is to build a new belief set from the information base every time we run the revision algorithm. This is fundamentally different from traditional operators that follow the AGM postulates do: the *inclusion* postulate states that when revising a set K with p, the result will be a subset of $K \cup \{p\}$: in other words, no sentences need to be added to K other than p. Our method, and in fact any DBR operator, can break even this postulate: let our information base be $\mathcal{I} = \{(b, 0.1), (b \rightarrow a, 1), (c, 0.5), (c \rightarrow \neg b, 1)\}$. Following our algorithm, this results in a belief set: $\mathcal{B} = \{c, c \rightarrow \neg b, \neg b\}$. However, if we revise with the new information $(\neg c, 0.9)$, we find that the resulting belief set $\mathcal{B}' = \{\neg c, b, b \rightarrow a, a\}$ is not a subset of $\mathcal{B}$, we have suddenly adopted beliefs a. Worse still, we can give similar examples that break *vacuity*. This seems to indicate that the operator we are proposing is not rightly *revision* at all, but rather something entirely different (we were questioned whether this should not be called merging instead). Nevertheless, we believe that *revision* is the right term, although it is definitely a different type of revision than has traditionally been considered, and we add our voices to those of Paglieri and Castelfranchi in calling

this type of operation Data-Oriented Belief Revision to separate it from traditional approaches. We stick with revision, because while the mechanism and properties of the operator are different, the operator still fulfills the goal of *revising* a single agent's beliefs based on newly received information. Belief merging, in contrast, implies that there are two or more different sets of beliefs that need to be merged, and its use in the literature refers mostly to aggregating the beliefs of a group of agents (Pigozzi 2015); a different problem altogether.

7 Conclusion

In this work, we proposed a new belief change operator and evaluated it with a comparison with the most similar operator from the literature. Here we need to make two remarks. Firstly, in the direct comparison between Pereira et al. and our approach, we uncover an unexpected property of their model. Due to choosing the message with maximum trust, they rely heavily on the accuracy of the trust model, and we suspect that this is one of the reasons for the bad performance in our experiments. In contrast, by using techniques from the information fusion domain, and pairing this with a more robust argumentation framework, our operator infers the truth even with a rough group-based trust model that is only approximately correct.

A second remark is that we tuned the scenario to display some specific characteristics, mostly regarding the share of agents communicating falsehoods. We wish to emphasize that in reality we can expect a larger majority of communication to be truthful, but in a large open environment with potentially untrustworthy agents, one must nevertheless be robust to this possibility and this rules out the use of a naive majority-rule approach to choosing whom to trust.

In general, we show that our operator performs at least as well as the alternatives and is more robust. Finally, when a small number of agents report the truth, we show that our operator can allow other agents to quickly assimilate this knowledge. We believe this is an important property that can help overcome another problem in belief change in a social environment: false institutional memory (Eck and Soh 2010). Exploring this is future work.

References

T. B. Adler, K. Chatterjee, L. De Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to wikipedia content. In *Proceedings of the 4th International Symposium on Wikis*, page 26, Porto, Portugal, 2008. ACM.

C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functios. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

L. Amgoud and C. Cayrol. On the acceptability of arguments in preference-based argumentation. In *UAI'98*, pages 1–7, Madison, USA, 1998. Morgan Kauffman Publishers.

L. Amgoud and S. Vesic. Rich preference-based argumentation frameworks. *International Journal of Approximate Reasoning*, 55(2):585–606, 2014.

M. Ben Sinai, N. Partush, S. Yadid, and E. Yahav. Exploting social navigation. *arXiv preprint arXiv:1410.0151*, 2014.

T. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

C. Burnett, T. J. Norman, and K. Sycara. Bootstrapping trust evaluations through stereotypes. In *Proceedings of AAMAS'10*, pages 241–248, Toronto, Canada, 2010. IFAAMAS.

M. Caminada. On the issue of reinstatement in argumentation. In *10th European Conference on Logics in Artificial Intelligence (JELIA'06)*, pages 111–123, Liverpool, UK, 2006. Springer.

D. Dubois and H. Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4(3):244–264, 1988.

P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonoic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

P. E. Dunne and M. Wooldridge. Complexity of abstract argumentation. In I. Rahwan and G. R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 85–104. Springer, 2009.

A. Eck and L.-K. Soh. Dynamic facts in large team information sharing. In *Proceedings of AAMAS'13*, pages 1217–1218, Saint Paul, USA, 2010. IFAAMAS.

M. A. Falappa, G. Kern-Isbender, and G. R. Simari. Belief revision and argumentation theory. In I. Rahwan and G. R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 341–360. Springer, 2009.

M. A. Falappa, A. J. García, G. Kern-Isbender, and G. R. Simari. On the evolving relation between belief revision and argumentation. *The Knowledge Engineering Review*, 26(1): 35–43, 2011.

M. B. Farah, D. Mercier, É. Lefèvre, and F. Delmotte. A high-level application using belief functions for exchanging and managing uncertain events on the road in vehicular ad hoc networks. *Annals of telecommunications - annales des télécommunications*, 69(3):185–199, 2013.

B. D. Greenshields. A study of traffic capacity. In *Proceedings of the 14th Annual Meeting of the Highway Research Board*, pages 448–481, 1935.

V. Grimm, U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, G. Huse, et al. A standard protocol for describing individual-based and agent-based models. *Ecological modelling*, 198(1):115–126, 2006.

S. O. Hansson. *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Springer, 1999a.

S. O. Hansson. A survey of non-prioritized belief revision. *Erkenntnis*, 50(2–3):413–427, 1999b.

D. Huang, X. Hong, and M. Gerla. Situation-aware trust architecture for vehicular networks. *Communications Magazine, IEEE*, 48(11):128–135, 2010.

A. Jøsang. The consensus operator for combining beliefs. *Artificial Intelligence*, 141(1):157–170, 2002.

A. Jøsang. Interpretation of fusion and hyper opinions in subjective logic. In *15th International Conference on Information Fusion*, pages 1225–1232, Singapore, 2012. IEEE.

A. Koster, A. G. B. Tettamanzi, A. L. C. Bazzan, and C. d. C. Pereira. Using trust and possibilistic reasoning to deal with untrustworthy communication in vanets. In *16th IEEE Annual Conference on Intelligent Transportation Systems*, pages 2355–2360, The Hague, The Netherlands, 2013. IEEE.

P. Krümpelmann, M. Thimm, M. A. Falappa, A. J. García, G. Kern-Isbender, and G. R. Simari. Selective revision by deductive argumentation. In *Theory and Applications of Formal Argumentation (TAFA'11)*, pages 147–162, Barcelona, Spain, 2012. Springer.

F. Paglieri and C. Castelfranchi. Revising beliefs through arguments: Bridging the gap between argumentation and belief revision in MAS. In *Argumentation in Multi-Agent Systems*, pages 78–94. Springer, 2005.

C. d. C. Pereira, A. G. B. Tettamanzi, and S. Villata. Changing one's mind: Erase or rewind? possibilistic belief revision with fuzzy argumentation based on trust. In *IJCAI'11*, pages 164–171, Barcelona, Spain, 2011. AAAI Press.

G. Pigozzi. Belief merging and judgment aggregation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2015 edition, 2015. `http://plato.stanford.edu/archives/fall2015/entries/belief-merging/`.

M. Raya, P. Papadimitratos, V. D. Gligor, and J.-P. Hubaux. On data-centric trust establishment in ephemeral ad hoc networks. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*. IEEE, 2008.

G. Shafer. *A Mathematical Theory of Evidence*, volume 1. Princeton University Press, Princeton, USA, 1976.

F. Smarandache and J. Dezert. Information fusion based on new proportional conflict redistribution rules. In *8th International Conference on Information Fusion*, volume 2,

Philadelphia, USA, 2005. IEEE.

M. d. Souza, A. Koster, and A. L. C. Bazzan. Technical description of an agent-based model for testing the effect of com-munication, trust and belief revision methods in a collaborative traffic scenario, 2015. Available at `http://goo.gl/1gx1YA`.

Y. Tang, K. Cai, P. McBurney, E. Sklar, and S. Parsons. Using argumentation to reason about trust and belief. *Journal of Logic and Computation*, 22(5):979–1018, 2012.

U. Wilenski. Netlogo, 1999. `http://ccl.northwestern.edu/netlogo` Center for Connected Learning and Computer-based Modeling, Northwestern University, Evanston, IL, USA.

J. Zhang. A survey on trust management for vanets. In *Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on*, pages 105–112. IEEE, 2011.