A Framework for Using Trust to Assess Risk in Information Sharing

Chatschik Bisdikian[1] ⋆, Yuqing Tang[2], Federico Cerutti[3], and Nir Oren[3]

[1] IBM Research Division,Thomas J. Watson Research Center,P.O. Box 704,Yorktown Heights, NY 10598, USA
bisdik@us.ibm.com

[2] Carnegie Mellon University, Robotics Institute, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
yuqing.tang@cs.cmu.edu

[3] University of Aberdeen, School of Natural and Computing Science, King's College, AB24 3UE, Aberdeen, UK
f.cerutti@abdn.ac.uk
n.oren@abdn.ac.uk

Abstract. In this paper we describe a decision process framework allowing an agent to decide what information it should reveal to its neighbours within a communication graph in order to maximise its utility. We assume that these neighbours can pass information onto others within the graph, and that the communicating agent gains and loses utility based on the information which can be inferred by specific agents following the original communicative act. To this end, we construct an initial model of information propagation and describe an optimal decision procedure for the agent.

1 Introduction

Information lies at the heart of successful decision making, and is therefore clearly valuable. When acting as part of a society, one must often divulge information to others when pursuing some goal. In doing so, the costs and benefits that such a divulgence will bring must be weighed up against each other. One of the most critical factors in this calculation is the trust placed in the entity to which one is providing the information — an untrusted individual might pass private information onto others, or may act upon the information in a manner harmful to the information provider.

In this paper we seek to provide a trust based decision mechanism for assessing the risks, and through these the costs and benefits, associated with the release of information. Using our mechanism, an agent can decide how much information to provide in order to maximise its own utility. We situate our work within the context of a multi-agent system. Here, an agent must assess the risk of divulging information to a set of other agents, who in turn may further propagate the information. The problem the agent faces is to identify the set of information that must be revealed to its neighbours (who will potentially propagate the information further) in order to maximise its own utility.

⋆ This paper is dedicated to the memory of Chatschik Bisdikian who recently passed away.

In the context of a multi-agent system, the ability of an agent to assess the risk of information sharing is critical when agents have to reach agreement, for example when coordinating, negotiating or delegating activities. In many contexts, agents have conflicting goals, and inter-agent interactions must take the risk of a hidden agenda into account. Thus, a theory of risk assessment for determining the right level of disclosure to apply to shared information is vital in order to avoid undesirable impacts on an information producer.

As a concrete example, consider the work described in [1], wherein information from accelerometer data attached to a person can be used to make either *white-listed* inferences, which are ones that the person desires others to infer, or *black-listed* inferences, which are those the person would rather not reveal. Using such accelerometer data, the person may wish a doctor to be able to determine how many calories they burn in a day, but might not want others to be able to infer what their state is (e.g. sitting, running or asleep). The person must thus identify which parts of the accelerometer data should be shared in order to enable or prevent their white- or black-listed inferences. While [1] examined how inferences can be made (e.g. that the sharing of the entropy of FFT coefficients provides a high probability of detecting activity level and low probability of detecting activity type), this work did not consider the *impacts* of sharing such information when it is passed on to others.

In this paper we consider an alternative situation, more applicable to the multi-agent systems domain. Here, a governmental espionage agency has successfully placed spies within some hostile country. It must communicate with these spies through a series of handlers, some of which may turn out to be double-agents. It must therefore choose what information to reveal to these handlers in order to maximise the benefits that spying can bring to it, while minimising the damage they can do. It is clear that the choices made by the agency depend on several factors. First, it must consider the amount of trust it places in the individual agents. Second, it must also take into account the amount of harm these agents can do with the information. Finally, it must consider the benefits that can accrue from providing the agents with the information. The combination of the first and second factors together provide a measure of the risk of information sharing. Now when considering the second factor, an additional detail must be taken into account, namely that the agents may already have some information available to themselves together with the information provided by the agency. Therefore, the final risk to itself does not only depend on the information provided, but rather on the level of harm that can occur due to the inferences made by the hostile agents.

The remainder of this paper is structured as follows. Section 2 describes the decision making process that an agent should follow, and examines its properties. Then Sect. 3 illustrates our proposal with a numeric example, and we contrast our approach with existing work in Sect. 4, where we also describe intended paths for future work. Section 5 concludes the paper.

2 The Model of Risk in Information Sharing

We consider a situation where an information producer shares information with one or more information consumers. These consumers can, in turn, forward the information

to other consumers, who may also forward it on, repeating the cycle. Furthermore, a consumer may or may not use the information provided as expected by the provider (e.g. based on some usage level agreement), and, hence, the producer needs to assess the risk that it will incur if the information is misused. In other terms, the producer has the goal to communicate part of the message to the desired consumers unless this will give rise to an unacceptable risk. This risk represents the potential harm that the producer will receive if some (undisclosed) part of the message is derived by some of the consumers, either desired or undesired. In fact, once the producer has communicated some pieces of information with the desired consumers, it cannot forbid the consumers to share what they know with other agents.

In the following subsection we will describe a model for this case. In what follows, we will use upper-case letters, e.g. X, to represent random variables (r.v.'s), lower-case letters, e.g. x, to represent realisation instances of them, and $F_X(\cdot)$ and $f_X(\cdot)$ to represent the probability distribution and density of the r.v. X, respectively.

2.1 The Model

We consider a set of agents able to interact with their neighbours through a set of communication links, as embodied by a communication graph or network. We assume that each agent knows the topology of this network. We introduce the concept of a *Framework for Risk Assessment* FRA that considers the set of agents, the messages that can be exchanged, the communication links of each agents, a producer that is willing to share some information, and the recipients of the information, which are directly connected to the producer within the communication graph.

Definition 1. *A* Framework for Risk Assessment *(FRA) is a 6-ple:*

$$\langle \mathcal{A}, \mathcal{C}, M, ag, m, Tg \rangle$$

where:

- $\mathcal{A}$ *is a set of agents;*
- $\mathcal{C} \subseteq \mathcal{A} \times \mathcal{A}$ *is the set of communication links among agents;*
- M *is the set of all the messages that can be exchanged;*
- $ag \in \mathcal{A}$ *is the* producer*, viz. the agent that shares information;*
- $m \in M$ *is a message to be assessed;*
- $\mathcal{A} \setminus \{ag\}$ *is the set of* consumers*, and in particular:*
  - $Tg \subseteq \mathcal{A} \setminus \{ag\}$ *are the* desired consumers*, and* $\forall ag_X \in Tg, \langle ag, ag_X \rangle \in \mathcal{C}$*;*
  - $\mathcal{A} \setminus (\{ag\} \cup Tg)$*, are the* undesired consumers*.*

Given a framework FRA, ag will make use of the risk assessment procedure described in this paper to determine how to share information. Then given a message, and a *degree of disclosure*, there is a function that returns a new message which is a *derived* version, according to the degree of disclosure, of the original message. Intuitively, a derived message conveys less information than the original, with the degree of disclosure specifying how much less information is provided. For instance, if we know that "the country A is going to invade the country B" (original information), then the information

"country B is going to be invaded" without specifying that the invader will be the country A is derived from the original given a degree of disclosure less than 1. It is beyond the scope of this paper to introduce a metric for evaluating how information should be derived, and it is left for future work.

Definition 2. *Given $\mathcal{A}$ a set of agents, a message $m \in M$, $ag_1, ag_2 \in \mathcal{A}$, $x_{ag_1}^{ag_2}(m) \in [0,1]$ is the* degree of disclosure of message m *used between the agent ag_1 and the agent ag_2, where $x_{ag_1}^{ag_2}(m) = 0$ implies no sharing and $x_{ag_1}^{ag_2}(m) = 1$ implies full disclosure between the two agents.*

We define the disclosure function *as follows:*

$$d : M \times [0,1] \mapsto M$$

$d(\cdot,\cdot)$ accepts as input a message and a degree of disclosure of the same message, and returns the disclosed part of the message as a new message.

In the following, when evident from the context, we will omit the indication of the message to which a degree of disclosure is related.

Since we assume that an information producer will only share information if doing so provides it with some benefit, the provision of information is equivalent to the producer "selling" the information for some profit.

Definition 3. *Given a FRA $\langle \mathcal{A}, \mathcal{C}, M, ag, m, Tg \rangle$, let $ag_X \in Tg$, $P(x_{ag}^{agx})$ be a r.v., described by $F_P(\cdot; x_{ag}^{agx})$ and $f_P(\cdot; x_{ag}^{agx})$, which represents the* benefit *agent ag receives when sharing the message m with a degree of disclosure x_{ag}^{agx} with agent ag_X.*

Given a FRA, the decision whether or not to share the information with the recipient must take into account several factors, and in particular:

1. the probability that an agent in possession of the message will forward it onward;
2. the levels of disclosure of messages exchanged between two agents;
3. the ability of each agent to infer knowledge from the received (disclosed) message;
4. the impacts that the inferred knowledge has on the information producer.

Each of these factors is subjective, and computed with respect to an agent's beliefs. With regards to agent ag's beliefs, we formalise these factors as follows (leaving ag implicit in the notation). Moreover, hereafter for the sake of generality we will silently drop the distinction between desired and undesired consumers.

Definition 4. *Given a FRA $\langle \mathcal{A}, \mathcal{C}, M, ag, m, Tg \rangle$, $\forall ag_1, ag_2 \in \mathcal{A} \setminus \{ag\}$:*

- *$s_{ag_1}^{ag_2} \in [0,1]$ is the probability that ag_1 will propagate to ag_2 the disclosed part of m that it receives;*
- *$x_{ag_1}^{ag_2} \in [0,1]$ is the assumed disclosure degree of communications between the two agents.*

We need now to introduce two operators for combining disclosure degrees when information is shared across multiple agents.

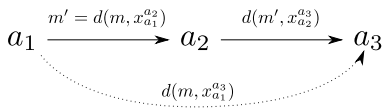

Fig. 1. Combining disclosure degrees: single communication path (Def. 5).

Figure 1 depicts the simplest case where the first operator can be applied. Let us suppose that there are three agents ag_1, ag_2, ag_3 and that ag_1 plays the role of the producer of the information m, that is shared with ag_2 with a degree of disclosure $x_{ag_1}^{ag_2}$ (the message sent to ag_2 is $m' = d(m, x_{ag_1}^{ag_2})$). Then, ag_2 shares what it received (m') with ag_3 sending it a message ($d(m', x_{ag_2}^{ag_3})$) derived from a new degree of disclosure $x_{ag_2}^{ag_3}$ which clearly applies to m' instead of m.

The operator we introduce in Def. 5 is aimed at computing the equivalent disclosure degree ($x_{ag_1}^{ag_3}$) to use for deriving a new message from m that ag_1 can send to ag_3 and such that $d(m', x_{ag_2}^{ag_3}) = d(m, x_{ag_1}^{ag_3})$. We require the introduced operator to be (i) transitive, and (ii) that the returned value $x_{ag_1}^{ag_3}$ should not be greater than $x_{ag_1}^{ag_2}$. This "monotonicity" requirement founds an intuitive explanation in the fact that ag_2 does not know the original information m, just its derived version $d(m, x_{ag_1}^{ag_2})$, and we assume that ag_2 does not make any kind of inference before sharing its knowledge. For instance, if ag_1 shares with ag_2 the 70% of the message m (whatever this means), and ag_2 shares with ag_3 the 50% of the information it received, what ag_3 receives is the 35% of m. The monotonicity requirement is strictly related to the assumption that an agent cannot share what it derived: this requirement will be relaxed in future developments of the FRA framework.

Definition 5. *Given a FRA* $\langle \mathcal{A}, \mathcal{C}, M, ag, m, Tg \rangle$, $\forall ag_1, ag_2, ag_3 \in \mathcal{A}$, $\langle ag_1, ag_2 \rangle$, $\langle ag_2, ag_3 \rangle \in \mathcal{C}$, *let* $m' = d(m, x_{ag_1}^{ag_2})$ *be the message sent by* ag_1 *to* ag_2 *(see Fig. 1). Then the message sent by* ag_2 *to* ag_3 *is:*

$$d(m', x_{ag_2}^{ag_3}) = d(m, x_{ag_1}^{ag_3})$$

where

- $x_{ag_1}^{ag_3} = \langle s_{ag_1}^{ag_2}, x_{ag_1}^{ag_2} \rangle \odot \langle s_{ag_2}^{ag_3}, x_{ag_2}^{ag_3} \rangle$;
- $\odot$ *is a transitive function such that*

$$\odot : ([0,1] \times [0,1]) \times ([0,1] \times [0,1]) \mapsto [0,1]$$

- $x_{ag_1}^{ag_3} \leq x_{ag_1}^{ag_2}$.

More interesting is the case where there are multiple path for reaching another consumer (whether desired or not). This is the case depicted in Fig. 2, where ag_1 shares (with different degree of disclosure $x_{ag_1}^{ag_2}$ and $x_{ag_1}^{ag_3}$) the same information to ag_2 and ag_3, and then both ag_2 and ag_3 share the information with ag_4.

Again, the transitive operator described in Def. 6, in collaboration with the one defined in Def. 5, should compute the degree of disclosure ($x_{ag_1}^{ag_4}$) of an equivalent message directly sent from ag_1 towards ag_4. The monotonicity requirement is still mandatory,

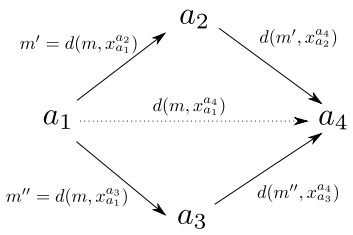

Fig. 2. Combining disclosure degrees: multiple communication path (Def. 6).

and in this case the derived degree of disclosure $x_{ag_1}^{ag_4}$ cannot be greater than the minimum of the disclosure degrees used for sharing the information with ag_2 ($x_{ag_1}^{ag_2}$) and with ag_3 ($x_{ag_1}^{ag_3}$). As before, this is mandatory since we assume that each agent does not make any inference before sharing the information.

Definition 6. *Given a FRA* $\langle \mathcal{A}, \mathcal{C}, M, ag, m, Tg \rangle$, $\forall ag_1, ag_2, ag_3, ag_4 \in \mathcal{A}$, $\langle ag_1, ag_2 \rangle$, $\langle ag_1, ag_3 \rangle$, $\langle ag_2, ag_4 \rangle$, $\langle ag_3, ag_4 \rangle \in \mathcal{C}$, *let* $m' = d(m, x_{ag_1}^{ag_2})$ *be the message sent by* ag_1 *to* ag_2, *and let* $m'' = d(m, x_{ag_1}^{ag_3})$ *be the message sent by* ag_1 *to* ag_3 *(see Fig. 2). Then there is a* merge *function that merges the message sent by* ag_2 *to* ag_4, *with the message sent by* ag_3 *to* ag_4 *as follows:*

$$merge(d(m', x_{ag_2}^{ag_4}), d(m'', x_{ag_3}^{ag_4})) = d(m, x_{ag_1}^{ag_4})$$

where

- $x_{ag_1}^{ag_4} = \left(\langle s_{ag_1}^{ag_2}, x_{ag_1}^{ag_2} \rangle \odot \langle s_{ag_2}^{ag_4}, x_{ag_2}^{ag_4} \rangle\right) \oplus \left(\langle s_{ag_1}^{ag_3}, x_{ag_1}^{ag_3} \rangle \odot \langle s_{ag_3}^{ag_4}, x_{ag_3}^{ag_4} \rangle\right)$;
- $\oplus$ *is a transitive function s.t.*

$$\oplus : [0,1] \times [0,1] \mapsto [0,1]$$

- $x_{ag_1}^{ag_4} \leq \min\{x_{ag_1}^{ag_2}, x_{ag_1}^{ag_3}\}$.

We do not introduce specific $\odot$ and $\oplus$ operators, leaving their definition to future work. However, we note that these definitions allow us to treat the transmission of information between any two agents in the network as a transmission between directly connected agents, subject to changes in the total level of disclosure. Computing this level of disclosure requires ag to make some assumptions, which can be computed from specific $\odot$ and $\oplus$ instantiations. Given this, we no longer distinguish between "known" disclosure degrees, as specified by the producer ag when transmitting the original message, and the "derived" disclosure degree — the one that ag derives using $\odot$ and $\oplus$.

We now turn our attention to the core of the decision process for assessing risk, which is based on the following definitions of *inferred knowledge* and of the *impact* that inferred knowledge has on ag.

Definition 7. *Given a FRA* $\langle \mathcal{A}, \mathcal{C}, M, ag, m, Tg \rangle$, $\forall ag_1, ag_2 \in \mathcal{A} \setminus \{ag\}$, $y_{ag_{2|x_{ag_1}^{ag_2}}} \in [0,1]$ *is the assumed (by* ag*)* amount of knowledge *of* m *that* ag_2 *can infer given*

$x_{ag_1}^{ag_2}$ *according to the r.v.* $I_{ag_2}(x_{ag_1}^{ag_2})$ *with distribution and density* $F_{I_{ag_2}}(\cdot; x_{ag_1}^{ag_2})$ *and* $f_{I_{ag_2}}(\cdot; x_{ag_1}^{ag_2})$ *respectively.*

In particular, $y_{ag_{2_{|x_{ag_1}^{ag_2}}}} = 0$ *will imply no knowledge and* $y_{ag_{2_{|x_{ag_1}^{ag_2}}}} = 1$ *inferred knowledge of equal extend as the one* ag_1 *has*[4].

We write $\mathcal{I}_{ag_{2_{|x_{ag_1}^{ag_2}}}}$ *to represent the family of r.v.'s* $I_{ag_2}(x_{ag_1}^{ag_2})$.

Previously, we suggested that the provision of information yields a positive utility to an agent. However, the provision of information to undesirable agents can also result in some sort of *impact*. In the following we will consider an impact equal to 0 when the producer will experience no harm at all by a consumer, while an impact equal to 1 means that the producer will receive the highest damage possible (i.e. total destruction) by a specific consumer.

Definition 8. *Given a FRA* $\langle \mathcal{A}, \mathcal{C}, M, ag, m, Tg \rangle$, $\forall ag_1, ag_2 \in \mathcal{A} \setminus \{ag\}$, $y_{ag_{2_{|x_{ag_1}^{ag_2}}}} \in [0,1]$, *the assumed (by* ag*)* impact *(normalized in the interval* $[0,1]$*) that an information producer* ag *incurs when an information consumer* ag_1 *makes use of the information inferred* $y_{ag_{|ag_1}}$ *from a message* m *disclosed according to a degree* $x_{ag}^{ag_1}$ *is represented by* $\mathcal{B} = \{B(y_{ag_{|ag_1}}) \in [0,1], y_{ag_{|ag_1}} \in [0,1]\}$, *a collection of r.v.'s, indexed by* $y_{ag_{|ag_1}}$, *with distribution* $F_B(\cdot; y_{ag_{|ag_1}})$ *and density* $f_B(\cdot; y_{ag_{|ag_1}})$, *reflecting* ag*'s beliefs regarding the level that* ag_1 *will impact* ag *when* ag_1 *knows (or has inferred)* $y_{ag_{|ag_1}}$.

In particular, $B(y_{ag_{|ag_1}}) = 0$ *should be interpreted as no impact, while* $B(y_{ag_{|ag_1}}) = 1$ *should be interpreted as complete impact (e.g. total destruction) in a situation (or context) of interest.*

We will refer to the r.v. $B(y_{ag_{|ag_1}})$ *as the* behavioral $y_{ag_{|ag_1}}$-trust *(or, simply,* $y_{ag_{|ag_1}}$*-trust) about the consumer* ag_1 *held by the producer* ag.

The r.v. B reflects how much ag can trust ag_1, if the former completely discloses the information to the latter[5]. Hereafter, for the sake of clarity, we will drop the sub- and super-scripts within our notation when the context is clear.

Figure 3 provides a graphical interpretation of *inference* (Def. 7) and *impact* (Def. 8) when a producer ag shares a message m with a degree x with a consumer.

We need to merge the above definition in order to compute the probabilities of inference and trust after the producer ag shares information with agents in Tg which can then propagate across the network.

Proposition 1. *Given a FRA* $\langle \mathcal{A}, \mathcal{C}, M, ag, m, Tg \rangle$, *an agent* $ag_Y \in \mathcal{A}$ *that has received a message* $d(m, x)$, *with* $x = x_{ag}^{ag_Y}$. *Let* y *be the inferred (by* ag_Y*) information according to the r.v.* $I(x)$ *(with probability* $\approx f_I(y; x)\, dy$*). Then, assuming that the impact* z *is independent of the degree of disclosure* x *given the inferred information* y, ag *expects a level of risk* z *described by the r.v.* $R(x)$ *with density:*

[4] For simplicity, we currently assume that agents share communicated information, but do not share any inferences they may make.

[5] Note that in dynamically evolving systems, it is expected that $B(y_{ag_{|ag_1}})$ may also evolve as the producer may accumulate experience transacting with the consumer. In this case, $F_B(\cdot; y_{ag_{|ag_1}})$ could represent a prior distribution, or a steady state one (if one can be reached). In our current discussion, we will silently assume a system that has reached steady state.

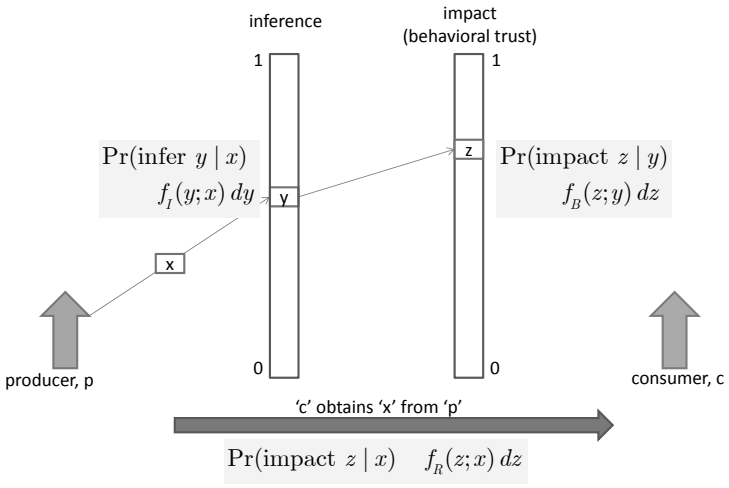

Fig. 3. The probabilities of inference and trust following sharing.

$$f_R(z;x) = \int_0^1 f_B(z;y)\, f_I(y;x)\, dy.$$

Proof.

$$\begin{aligned} F_R(z;x) = \Pr\{B \le z|x\} &= \int_0^1 \Pr\{B \le z, I = y|x\}\, dy \\ &= \int_0^1 \Pr\{B \le z|I = y, x\} f_I(y;x)\, dy = \int_0^1 F_B(z;y)\, f_I(y;x)\, dy, \end{aligned}$$

The density function is easily derived from the distribution $F_R(z;x)$ *since* $f_R(z;x) = \frac{d}{dz} F_R(z;x)$. □

We can now define the net benefit of sharing information, where the *benefit* is the metric dual to the impact metric, viz. it represents the advantages that the producer will have by sharing an information with a specific consumer.

Definition 9. *Given a FRA* $\langle \mathcal{A}, \mathcal{C}, M, ag, m, Tg \rangle$*, let be* $P(x_{ag}^{ag_Y})$ *the r.v. representing the* benefit *that the producer expects by sharing* m *with* $ag_Y \in \mathcal{A}$ *in function of the disclosure degree.*

Moreover, for each $ag_Y \in \mathcal{A}$*, the* net benefit *for the producer to share information with* ag_Y *is described by:* $C = P - R$*, with an average, or expected benefit,* $\mathbb{E}\{C(x_{ag}^{ag_Y})\} = \mathbb{E}\{P(x_{ag}^{ag_Y})\} - \mathbb{E}\{R(x_{ag}^{ag_Y})\}$.

Note that the producer may *value* its information x according to the benefit it receives. For example, for each x, it may value the information at level $R(x)$ (and above) in order to protect itself from any expected impacts.

Finally, any time we need a single value instead of a distribution, we can exploit the same idea of descriptors of a random variable, by introducing descriptors for trust and risk.

Definition 10. *Let $h(\cdot)$ be a function defined on $[0,1]$, and $x \in [0,1]$ be a level of disclosure. We define*

$$t_h^B(x) = \int_0^1 h(w)\, f_B(w;x)\, dw, \tag{1}$$

*to be the x-*trust descriptor *induced by $h(\cdot)$.*

We similarly defined a risk descriptor as follows:

Definition 11. *Let $h(\cdot)$ be a function defined on $[0,1]$, and $x \in [0,1]$ be a level of disclosure.We define*

$$t_h^R(x) = \int_0^1 h(w)\, f_R(w;x)\, dw, \tag{2}$$

*to be the x-*risk descriptor *induced by $h(\cdot)$.*

Typical $h(\cdot)$ will include the moment generating functions, such as $h(k) = k, k^2$, etc., and entropy $h(k) = -ln\big(f_K(k)\big)$ for the density of some r.v. K. In the following we use the expectation as the risk descriptor, leaving other possible functions for future work.

2.2 Properties of the Model

In this section we illustrate two notable properties of our model. The first one is with regards to the case where a consumer can derive the full original message, which, unsurprisingly, leads to the worst case impact.

Proposition 2. *When a consumer is capable of gaining maximum knowledge, then $f_I(y;x) = \delta(y-1)$, where $\delta(\cdot)$ is the Dirac delta function, and $F_R(z;x) = F_B(z) \triangleq F_B(z;1)$, i.e., the risk coincides with the 1-trust (Def. 8).*

Proof. By the definition of the inference r.v. $I(x)$, when ag_X is believed to gain maximum knowledge then the density $f_I(y;x)$ carries all its weight at the point $y = 1$ for all x. Hence, $f_I(y;x) = \delta(y-1)$ and it follows from the definition of the Dirac delta function, see also Prop. 1

$$F_R(z;x) = \int_0^1 F_B(z;y)\, f_I(y;x)\, dy = \int_0^1 F_B(z;y)\, \delta(y-1)\, dy = F_B(z;1). \tag{3}$$

□

The second property pertains to the case where agent ag shares information with more than one consumer. Assuming a non-homogeneous situation where the trust and impact levels for ag against each consumer varies, it is clearly beneficial to identify conditions where these impacts balance (and, hence, indicate crossover thresholds) across the multiple agents.

For two agents ag_1, ag_2 having corresponding inference and behavioural trust distributions $F_{I_j}(y;x)$ and $F_{B_j}(z;y)$, $j \in \{1,2\}$, then for the shared information to have similar impact, x_1 and x_2 should be selected, such that:

$$\begin{gathered} F_{R_1}(z;x_1) = F_{R_2}(z;x_2) \Leftrightarrow \\ \int_0^1 F_{B_1}(z;y)\, f_{I_1}(y;x_1)\, dy = \int_0^1 F_{B_2}(z;y)\, f_{I_2}(y;x_2)\, dy. \end{gathered} \tag{4}$$

Note that the above relationship implies the r.v.s R_1 and R_2 are equal in distribution, which is expected to be too stringent a condition to satisfy, and quite possibly too hard to attain. In general, one may want to consider equalities on the average, such as, finding x_1 and x_2 satisfying:

$$\mathbb{E}\{g(R_1(x_1))\} = \mathbb{E}\{g(R_2(x_2))\}, \tag{5}$$

for appropriate functions $g(\cdot)$.

Proposition 3. *If $g(z) = z$ then to attain the same level of impact when ag shares information with ag_1, ag_2, the information disclosure levels x_1 and x_2 must satisfy*

$$\mathbb{E}_{I_1}\big\{\mathbb{E}\{R_1(x_1)|I_1\}\big\} = \mathbb{E}_{I_2}\big\{\mathbb{E}\{R_2(x_2)|I_2\}\big\}. \tag{6}$$

Proof. The case where $g(z) = z$ corresponds to the regular averaging operator, and (5) *becomes:*

$$\begin{gathered} \int_0^1 \int_0^1 z f_{B_1}(z;y)\, f_{I_1}(y;x_1)\, dy\, dz = \int_0^1 \int_0^1 z f_{B_2}(z;y)\, f_{I_2}(y;x_2)\, dy\, dz \\ \Leftrightarrow \int_0^1 f_{I_1}(y;x_1) \left[\int_0^1 z f_{B_1}(z;y)\, dz\right] dy = \int_0^1 f_{I_2}(y;x_2) \left[\int_0^1 z f_{B_2}(z;y)\, dz\right] dy \\ \Leftrightarrow \mathbb{E}_{I_1}\big\{\mathbb{E}\{R_1(x_1)|I_1\}\big\} = \mathbb{E}_{I_2}\big\{\mathbb{E}\{R_2(x_2)|I_2\}\big\}. \end{gathered} \tag{7}$$

□

3 An Example

To illustrate our proposal, let us suppose that British Intelligence sent two spies, James and Alec, to France. James is a clever agent, very loyal to Britian, while Alec is less brilliant and his trustworthiness is highly questionable. After some months, London informed her men that in three weeks France will be invaded by an European country: it hopes that James and Alec can recruit new agents in France thanks to this information. However, the intelligence agency does not specify how this invasion will take place, although they already know it is very likely to come from the East. However, both James and Alec may infer the following additional piece of information, namely that Spain, Belgium and Italy have no interest in invading France, while Germany does. London does not want to share the information that the invasion will be started by Germany, because they are the only ones aware of these plans, and a leak would result in a loss

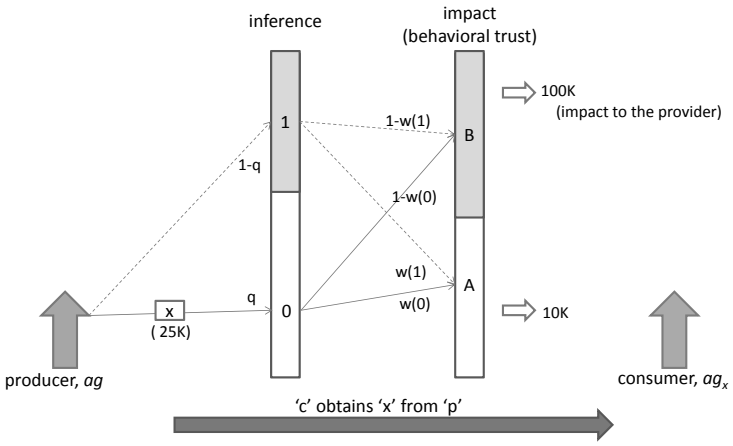

Fig. 4. A simple binary inference/impact case.

of credibility for the UK government. Therefore, British Intelligence has to assess the risk in order to determine whether or not it is acceptable to inform its agents that France will be invaded by an European country.

In order to formalise this situation, let us consider discrete random variables for ease of presentation, where probability mass functions are used instead of densities. As per Fig. 4, inferences can be of type "0" or "1"; ag believes that each can be made with probabilities q and $1-q$ respectively, by an information consumer. In our scenario, inference of type "0" means that the information "Germany will be the invader" is not inferred, while it is inferred by the inference of type "1".

For each of the possible inferences there are two levels of impact, noted with A and B with impact costs of $10K$ of and $100K$ respectively, as shown in the figure, while the utility cost of having new agent recruited is of $25K$. With the probabilities shown in the figure, an information producing agent is characterised by the triple $\alpha = \langle q, w(0), w(1)\rangle$.

Therefore, James' behaviour can be characterised by the triple $\alpha_1 = \langle 0.1, 0.9, 0.9\rangle$, while Alec is characterised by the triple $\alpha_2 = \langle 0.6, 0.6, 0.4\rangle$, which shows that if Alec infers that the invader will be the Germany, it is more likely that he will defect, leading to the worst impact case for the UK.

When agent ag shares message m at disclosure level x with a particular consumer agent ag_X, the average impact $\mathbb{E}\{h\}$ anticipated by ag is given by

$$\begin{aligned}\mathbb{E}\{h\} &= q\big\{10w(0) + 100[1-w(0)]\big\} + (1-q)\big\{10w(1) + 100[1-w(1)]\big\}\\ &= q\big\{100 - 90w(0)\big\} + (1-q)\big[100 - 90w(1)\big]\\ &= 100 - 90\big\{q[w(0) - w(1)] + w(1)\big\}.\end{aligned} \tag{8}$$

Unsurprisingly, in the case of James the average impact ($\mathbb{E}\{h_1\} = 10K$) is sensibly lower that the average impact in the case of Alec ($\mathbb{E}\{h_2\} = 53.2K$).

Moreover, let us consider ag having the option to share information with one of two agents, ag_1 (James) or ag_2 (Alec) each characterised by the triples α_1 and α_2. In this case, the average impacts anticipated by ag from each of agents will be equal when:

$$\begin{aligned}
\mathbb{E}\{h_1\} &= \mathbb{E}\{h_2\} \Leftrightarrow \\
100 - 90\big\{q_1[w_1(0) - w_1(1)] + w_1(1)\big\} &= 100 - 90\big\{q_2[w_2(0) - w_2(1)] + w_2(1)\big\} \Leftrightarrow \\
q_1\{w_1(0) - w_1(1)\} + w_1(1) &= q_2\{w_2(0) - w_2(1)\} + w_2(1) \Leftrightarrow \\
q_2 &= q_1 \frac{w_1(0) - w_1(1)}{w_2(0) - w_2(1)} + \frac{w_1(1) - w_2(1)}{w_2(0) - w_2(1)}.
\end{aligned} \tag{9}$$

It is interesting to note that in this case, the levels of impact (e.g. $10K$ and $100K$) have no effect in the equality of the impacts, which only depends on the relationships between the triplets α_1 and α_2.

In the ideal case that $q_1 = 1$, i.e., only "0" can be inferred, then

$$q_2 = \frac{w_1(0) - w_1(1) + w_1(1) - w_2(1)}{w_2(0) - w_2(1)} = \frac{w_1(0) - w_2(1)}{w_2(0) - w_2(1)}, \tag{10}$$

which because $0 \leq q_2 \leq 1$ implies that (10) could be valid only if $w_1(0)$ and $w_2(0)$ are simultaneously larger or smaller than $w_2(1)$. Then, for example, if $w_1(0)$ and $w_2(0)$ are both larger than $w_2(1)$, then $w_1(0) \leq w_2(0)$ will also be required.

If we were to assume that $w_1(1) = w_2(1)$, i.e., agents ag_1 and ag_2 were to behave similarly when they infer "1", then it follows from (9):

$$\frac{q_2}{q_1} = \frac{w_1(0) - w_1(1)}{w_2(0) - w_2(1)}. \tag{11}$$

Thus, to attain similar average impact the q's must be inversely proportional to the difference $w(0) - w(1)$. Furthermore, given that the q's are non-negative, the differences $w_i(0) - w_i(1)$, $i = 1, 2$, must have the same sign otherwise no q_2 can be found that will result to equal impact.

Now let us suppose that agent ag benefits at level $P(x)$ when sharing a message with disclosure level x, in the figure $\bar{P}(x) = \mathbb{E}\{P(x)\} = 25K$. Then, the expected net benefit $\bar{C}(x) = \mathbb{E}\{C(x)\}$ for the agent will be:

$$\bar{C}(x) = \bar{P}(x) - 100 + 90\big\{q[w(0) - w(1)] + w(1)\big\}. \tag{12}$$

Since $\bar{C}(x) \geq 0$ is desired, we must necessarily have:

$$\begin{aligned}
100 - \bar{P}(x) &\leq 90\big\{q[w(0) - w(1)] + w(1)\big\} \Leftrightarrow \\
\frac{100 - \bar{P}(x)}{90} &\leq q[w(0) - w(1)] + w(1) \Leftrightarrow \\
\frac{100 - \bar{P}(x)}{90} &\leq qw(0) + (1 - q)w(1) \leq 1.
\end{aligned} \tag{13}$$

The right hand side of the expression above represents the probability of experiencing impact at level A, see Fig. 4, which immediately necessitates that $\bar{P}(x) \geq 10K$. In

other words, the minimum valuation of the information should be at least as large as the minimum impact expected to occur, cf. Def. 9.

To conclude our example, from Eq. 13, which assesses the risk and the trust model (the triples in this discrete case), we can see that Britian can share the information that France is going to be invaded with James ($\frac{75}{90} \leq 0.9 \leq 1$), but not with Alec ($\frac{75}{90} \not\leq 0.52 \leq 1$).

4 Discussion and Future Work

The work of this paper makes use of an unspecified trust model as a core input into the reasoning process. Our probabilistic underpinnings are intended to be sufficiently general to enable it to be instantiated with arbitrary models, such as [2, 3]. Unlike these models, our work is not intended to compute a specific trust value based on some set of interactions, but rather to decide how to use the trust value output by the models.

The use of trust within a decision making system is now a prominent research topic, see [4, 5] for an overview. However, the most part of the works in this area generally assumes that agents will interact with some most trusted party, as determined by the trust model. This assumption reflects the basis of trust models on action and task delegation rather than information sharing. [6] is an exception to this trend; while still considering tasks, Burnett explicitly takes into account the fact that dealing with a trusted party may be more expensive, and thus lead to a lower utility to an agent when delegating a low priority task. Burnett's model therefore considers both risk and reward when selecting agents for interaction.

Another body of work relevant to this paper revolves around information leakage. Work such as [7] considers what information should be revealed to an agent given that this agent should not be able to make specific inferences. Unlike this paper, such work does not consider the potential benefits associated with revealing information.

Finally, there is a broad research area devoted at assessing risk in different contexts. As summarised in [8], which compares seven definitions of trust[6], the notion of risk is the result of some combination of uncertainty about some outcome, and a (negative) payoff for an intelligent agent and his goals. While this definition is widely accepted (with minor distinctions), different authors have different point of view when it comes to formally define what is meant by *uncertainty*. In [9], instead of providing a formal definition of risk, the authors introduced a scenario-based risk analysis, considering (i) the *scenario*, (ii) its *likelihood*, and (iii) the *consequences* of that scenario. They also introduce the notion of *uncertainty* in the definition of likelihood and of consequences. Doing so allows them to address the core problem of such models, viz. that complete information of all possible scenarios is required. The connection between risk and trust has been the subject of several studies, e.g. [10] shows a formal model based on epistemic logic for dealing with trust in electronic commerce where the risk evaluation is one of the components that contribute to the overall trust evaluation, [11] proposes a conceptual framework showing the strict correspondence between risk and some definition of trust, [4] discusses the connection between risk and trust in delegation. However,

[6] Although not considered in [8], the definition provided in [4] follows the others.

to our knowledge our work is the first attempt to consider risk assessment in trust-based decision making about information sharing.

There are several potential avenues for future work. First, we have assumed that trust acts as an input to our decision process, and have therefore not considered the interplay between risk and trust. We therefore seek to investigate how both these quantities evolve over time. Another aspect of work we intend to examine is how the trust process affects disclosure decisions by intermediate agents with regards to the information they receive. More informally, agents might not propagate information from an untrusted source onwards, as they might not believe it. Such work, together with a more fine grained representation of the agents' internal beliefs could lead to interesting behaviours such as agents lying to each other [12]. Other scenarios of interest can be easily envisaged, and they will be investigated in future work. For instance, a slightly modified version of the framework proposed in this paper can be used for determining the degree of disclosure in order to be reasonably sure that a desired part of the message will actually reach a specific agent with which we do not know how to communicate. This is the situation when an organisation tries to reach an undercover agent by sharing some information with the enemy, hoping that somehow some pieces of information will reach eventually the agent. Other interesting cases of study will be addressed in future work since the ultimate goal of our research is to identify utility maximising utterances, given a knowledge rich (but potentially incomplete or uncertain) representation of the multi-agent system.

5 Conclusions

In this paper we described a framework enabling an agent to determine how much information should disclose to others in order to maximise its utility. This framework assumes that any disclosure could be propagated onwards by the receiving agents, and that certain agents should not be allowed to infer some information, while it is desirable that others do make inferences from the propagated information. We showed that our framework respects certain intuitions with regards to the level of disclosure used by an agent, and also identified how much an information provider should disclose in order to achieve some form of equilibrium with regards to its utility. Potential applications can be envisaged in strategic contexts, where pieces of information are shared across several partners which can have hidden agenda. Therefore it is of primary importance to determine the level of disclosure of the information in order to received some benefit or to contribute to the coalition, without be harmed.

To our knowledge, this work is the first to take trust and risk into account when reasoning about information sharing, and we are pursuing several exciting avenues of future work in order to make the framework more applicable to a larger class of situations.

Acknowledgements. The authors thank the anonymous reviewers for their helpful comments.

Research was sponsored by US Army Research laboratory and the UK Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The

views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory, the U.S. Government, the UK Ministry of Defense, or the UK Government. The US and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

1. Chakraborty, S., Raghavan, K.R., Srivastava, M.B., Bisdikian, C., Kaplan, L.M.: Balancing value and risk in information sharing through obfuscation. In: Proceedings of the 15th Int'l Conf. on Information Fusion (FUSION '12). (2012)
2. Jøsang, A., Ismail, R.: The beta reputation system. In: Proceedings of the 15th Bled Electronic Commerce Conference. (2002)
3. Teacy, W.T.L., Patel, J., Jennings, N.R., Luck, M.: Travos: Trust and reputation in the context of inaccurate information sources. Autonomous Agents and Multi-Agent Systems **12**(2) (2006) 183–198
4. Castelfranchi, C., Falcone, R.: Trust theory: A socio-cognitive and computational model. Wiley Series in Agent Technology (2010)
5. Urbano, J., Rocha, A., Oliveira, E.: A socio-cognitive perspective of trust. In Ossowski, S., ed.: Agreement Technologies. Volume 8 of Law, Governance and Technology Series. Springer Netherlands (2013) 419–429
6. Burnett, C., Norman, T.J., Sycara, K.: Trust decision-making in multi-agent systems. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume One. IJCAI'11, AAAI Press (2011) 115–120
7. Mardziel, P., Magill, S., Hicks, M., Srivatsa, M.: Dynamic enforcement of knowledge-based security policies. In: Proceedings of the 24th IEEE Computer Security Foundations Symposium. (2011) 114–128
8. Wang, X., Williams, M.A.: Risk, uncertainty and possible worlds. In: Privacy, security, risk and trust (passat), IEEE Third International Conference on Social Computing (SOCIALCOM). (2011) 1278–1283
9. Kaplan, S., Garrick, B.J.: On the quantitative definition of risk. Risk Analysis **1**(1) (1981) 11–27
10. Tan, Y.H., Thoen, W.: Formal aspects of a generic model of trust for electronic commerce. Decision Support Systems **33**(3) (July 2002) 233–246
11. Das, T.K., Teng, B.S.: The Risk-Based View of Trust: A Conceptual Framework. Journal of Business and Psychology **19**(1) (2004) 85–116
12. Caminada, M.W.: Truth, lies and bullshit; distinguishing classes of dishonesty. In: Social Simulation workshop (SS@IJCAI). (2009) 39–50