Trust on Information Sources: A theoretical and computation approach

Alessandro Sapienza, Rino Falcone and Cristiano Castelfranchi,
Institute of Cognitive Science and Technologies,
ISTC-CNR, Rome, Italy,
{alessandro.sapienza, rino.falcone, cristiano.castelfranchi}@istc.cnr.it

Abstract— **We start from the claim that trust in information sources is just a kind of social trust. We are interested in the fact that the relevance and the trustworthiness of the information acquired by an agent X from a given number of sources strictly depends and derives from the X's trust on each of these sources with respect the kind of that information. In this paper, we analyze the different dimensions of trust in information sources and formalize the degree of subjective certainty or strength of the X's belief P, considering three main factors: the X's trust about P just depending from the X's judgment of the source's competence and reliability; the sources' degree of certainty about P; and the X's degree of trust that P derives from that given source. Finally we present a computational approach based on fuzzy sets.**

I. Dimensions of Trust in Information Sources

Which are the important specific dimensions of trust in information sources (TIS)? Many of these dimensions are quite sophisticated, given the importance of information for human activity and cooperation. We will simplify and put aside several of them.

First of all, we have to trust (more or less) the source (F) as competent and reliable in that domain, in the domain of the specific information content. Am I waiting for some advice on train schedule? On weather forecast? On the program for the examination? On a cooking recipe?

Is this F not only competent but also reliable (in general or specifically towards me)? Is F sincere and honest? Or leaning to lie and deceive? Will F do what has promised to do or "has" to do for his role? And so on.

These competence and reliability evaluations can derive from different reasons, basically:

a) Our previous *direct experience* with F (how F performed in the past interactions) on that specific information content , or better our "memory" about, and the adjustment that we have made of our evaluation of F in several interaction, and possible successes or failure relying on its information;

b) *Recommendations* (other individuals Z reporting their direct experience and evaluation about F) or *Reputation* (the shared general opinion of others about F) on that specific information content; [3; 4; 5; 12; 13];

c) *Categorization* of F (it is assumed that a source can be categorized and that it is known this category), exploiting inference and reasoning:

- inheritance from classes or groups were Z id belonging (as a good "exemplar");
- analogy: Z is (as for that) like Y, Y is good for, then Z too is good for;
- analogy on the task: Z is good/reliable for P he should be good also for P', since P and P' are very similar. (In any case: how much do I trust my reasoning ability?).

On this basis it is possible to establish the competence/reliability of F on the specific information content [2,6].

The two faces of F's trustworthiness (competence and reliability) are relatively independent[1]; we will treat them as such. Moreover, we will simplify these complex components in just one quantitative fuzzy parameter: F's estimated trustworthiness; by combining competence and reliability.
In particular we define the following fuzzy set: *terrible*, *poor*, *mediocre*, *good*, *excellent* (see figure 1) and apply it to each of the previous different dimensions (direct experience, recommendations and reputation, categorization).

These competence and reliability evaluations can derive from different reasons, basically:

Second, information sources have and give us a specific information that they know/believe; but believing something is not a yes/no status; we can be more or less convinced and sure (on the basis of our evidences, sources, reasoning). Thus a good source might inform us not only about P, but also about

[1] Actually they are not fully independent. For example, F might be tempted to lie to me if/when is not so competent or providing good products: he has more motives for fudging me.

its *degree of certainty about* P, its trust in the truth of P. For example: "It is absolutely sure that P", "Probably P", "It is frequent that P", "It might be that P", and so on.

Of course there are more sophisticated meta-trust dimensions like: how much am I sure, confident, in F's evaluation of the probability of the event or in his subjective certainty?[2] Is F not sincere? Or not so self-confident and good evaluator? For example, in drug leaflet they say that a given possible bad side effect is only in 1% of cases. Have I to believe that? Or they are not reliable since they want to sell that drug? For the moment, we put aside that dimension of how much meta-trust we have in the provided degree of credibility. We will just combine the provided certainty of P with the reliability of F as source. It in fact makes a difference if an excellent or a mediocre F says that the degree of certainty of P is 70% (**see §I.B**).

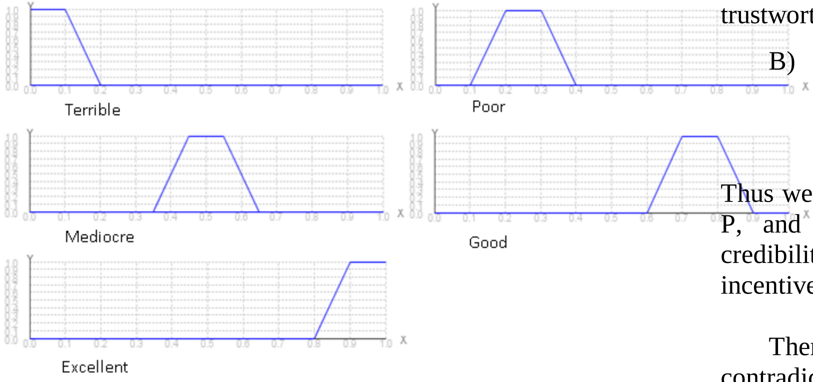

Figure 1: Representation of the five fuzzy sets

Third, especially for information sources it is very relevant the following form of trust: the trust we have that the information under analysis derives from that specific source, how much we are sure about that "transmission"; that is, that the communication has been correct and working (and complete); that there are no interferences and alterations, and I received and understood correctly; that the F is really that F (Identity).Otherwise I cannot apply the first factor: F's credibility.

Let's simplify also these dimensions, and formalize just the *degree of trust that F is F*; that the F of that information (I have to decide whether believe or not) is actually F. In the WEB this is an imperative problem: the problem of the real *identity* of the F, and of the reliability of the signs of that identity, and of the communication.

These dimensions of TIS are quite independent of each other (and we will treat them as such); we have just to combine them and provide the appropriate dynamics. For example, what happen if a given very reliable source F' says that "it is sure that P", but I'm not sure at all that the information really comes from F' and I cannot ascertain that?

A. *Additional problems and dimensions*

We believe in a given datum on the basis of its origin, its source: perception? communication? inference? And so on.

A) *The more reliable (trusted) the F the stronger the trust in P, the strength of the Belief that P.*

This is why it is very important to have a "memory" of the sources of our beliefs. However, there is another fundamental principle of the degree of credibility of a given Belief (its trustworthiness):

B) *The many the converging sources, the stronger our belief* (of course, if there are no correlations among the sources).

Thus we have the problem to combine different sources about P, and their subjective degrees of certainty, and their credibility, in order to weigh the credibility of P, and have an incentive due to a large convergence of sources.

There might be different heuristics for dealing with contradictory information and sources. One (prudent) agent might adopt as assumption the worst hypothesis, the weaker degree of P; another (optimistic) agent, might choose the best, more favorable estimation; another agent might choose the most reliable source. We will formalize only one strategy: the weighing up and combination of the different strengths of the different sources, avoiding however the psychologically incorrect result of probability values, where by combining different probabilities we always decrease the certainty, it never increases. On the contrary - as we said - convergent sources reinforce each other and make us more certain of that datum.

B. *Feedback on source credibility/TIS*

We have to store the sources of our beliefs because, since we believe on the basis of source credibility, we have to be in condition to adjust such credibility, our TIS, on the basis of the result. If I believe that P on the basis of source F1, and later I discover that P is false, that F1 was wrong or deceptive, I have to readjust my trust in F1, in order next time (or with similar sources) to be more prudent. And the same also in case of positive confirmation .

However, remember that it is well known [8] that the negative feedback (invalidation of TIS) is more effective and heavy than the positive one (confirmation). This asymmetry (the collapse of trust in case on negative experience versus the slow acquisition or increasing of trust) is not specific of trust and of TIS; it is -in our view- basically an effect of a general cognitive phenomenon. It is not an accident or weirdness if the disappointment of trust has much stronger (negative) impact than the (positive) impact of confirmation. It is just a sub-case of the general and fundamental asymmetry of negative vs. positive results, and more precisely of "losses" against "winnings": the well-known Prospect theory [9]. We do not evaluate in a symmetric way and on the basis of an "objective"

[2] In a sense it is a *transitivity principle* [7]: X trust Y, and Y trust Z; will X trust Z? Only if X trusts Y "as a good evaluator of Z and of that domain". Analogously here: will X trust Y because Y trusts Y? Only if X trust Y "as a good and reliable evaluator" of it-self.

value/quantity our progresses and acquisitions versus our failures and wastes, relatively to our "status quo". Losses (with the same "objective" value) are perceived and treated as much more severe: the curve of losses is convex and steep while that of winnings is concave. Analogously the urgency and pressure of the "avoidance" goals is greater than the impulse/strength of the achievement goals [10]. All this applies also to the slow increasing of trust and its fast decreasing; and to the subjective impact of trust disappointment (betrayal!) vs. trust confirmation. That's why usually we are prudent in deciding to trust somebody; in order do not expose us to disappointment and betrayals, and harms. However, also this is not always true; we have quite naive forms of trust just based on gregariousness and imitation, on sympathy and feelings, on the diffuse trust in that environment and group, etc. This also plays a crucial role in social networks on the web, in web recommendations, etc.

Moreover, in our theory [11] not always and automatically a bad result (or a good result) entails the revision of TIS. It depends on the "causal attribution": it has been a fault/defect of F or an interference on the environment? The result might be bad although F's performance was his best. Let us put aside here the feedback effect and revision on TIS.

C. *Plausibility: the integration with previous knowledge*

To believe something means not just to put it in a file in my mind; it means to *"integrate" it with my previous knowledge*. Knowledge must be at least non-contradictory, and possibly supported, justified: this explains that, and it is explained, supported, by these other facts/arguments. If there is *contradiction* I cannot believe P; either I have to reject P or I have to revise my previous beliefs in order to coherently introduce P. It depends on the strength of the new information (its credibility, due to its sources) and on the number and strength of the internal opposition: the value of the contradictory previous beliefs, and the extension and cost of the required revision. That is: *it is not enough that the candidate belief that P be well supported and highly credible*; is there an epistemic conflict? Is it "implausible" to me? Are there antagonistic beliefs? And which is their strength? The winner of the conflict will be the stronger "group" of beliefs. Even the information of a very credible source (like our own eyes) can be rejected!

II. FORMALIZING AND COMPUTING THE DEGREE OF CERTAINTY AS TRUST IN THE BELIEF

As we have said, there is a confidence, a trust in the beliefs we have and on which we rely.

Suppose X is a cognitive agent, an agent who has beliefs and goals. Given Bel_X, the set of the X's beliefs, then P is a belief of X if:

$$P \in Bel_X \qquad (1)$$

The degree of subjective certainty or strength of the X's belief P corresponds with the X's trust about P, and call it:

$$Trust_X(P) \qquad (2)$$

A. *Its origin/ground*

Concerning a single belief P, we have to consider n different sources asserting or denying P. The final value of $Trust_X(P)$ depends on X's trust towards every single source F of the information P (that could mean with respect the class of information to which P belongs):

$$Trust_X(F,P) \qquad (3)$$

In other words, we state that:

$$Trust_X(P) = f(Trust_X(F_1,P), \ldots, Trust_X(F_n,P)) \qquad (4)$$

Where n is the total number of sources.

Then to compute X's trust value, we have to compose the n sources' value in just one resulting factor.

Applying now the conceptual modeling previously described we have that $Trust_X(F,P)$ can be articulated in:

1. X's trust about P just depending from the X's judgment of the F's competence and reliability as derived from the composition of the three factors (direct experience, recommendation/reputation, and categorization), in practice the F's credibility about P on view of X:
$$Trust^1{}_X(F,P) \qquad (5)$$
2. F's degree of certainty about P: information sources give not only the information but also their certainty about this information; given that we are interested to this certainty, but we have to consider that through X's point of view, we introduce
$$Trust_X(Trust_F(P)) \qquad (6)$$
in particular, we consider that X completely trusts F, so that $Trust_X(Trust_F(P)) = Trust_F(P)$
3. the X's degree of trust that P derives from F: the trust we have that the information under analysis derives from that specific source:
$$Trust_X(Source(F,P)) \qquad (7)$$
4. the fact that F is supporting P or is opposing to it (not P):
$$Support_F(P) \qquad (8)$$

Resuming:

$$Trust_X(F,P) = f_3(Trust^1{}_X(F,P), Trust_X(Trust_F(P)), Trust_X(Source(F,P)), Support_F(P)) \qquad (9)$$

Here we could introduce a threshold for each of these 3 dimensions, allowing to reduce risk factors.

B. *A modality of computation*

1) *$Trust^1{}_X(F,P)$*

As specified in **§I** the value of $Trust^1{}_X(F,P)$ is a function of:

1. Past interactions;
2. The category of membership;
3. Reputation.

As previously said, each of these values is represented by a fuzzy set: terrible, poor, mediocre, good, excellent. We then compose them into a single fuzzy set, considering a weight for each of these three parameters. Those weights are defined in range [0;10], with 0 meaning that the element has no importance in the evaluation and 10 meaning that it has the maximal importance.

It is worth noting that the weight of experience has to be referred to a twofold meaning: it must take into account the numerosity of experiences (with their positive and negative values), but also the intrinsic value of experience for that subject.

However, the fuzzy set in and by itself is not very useful: what interests us in the end is to have a plausibility range, which is representative of the expected value of $Trust^1_X(F,P)$.

To get that, it is therefore necessary to apply a defuzzyfication method. Among the various possibilities (mean of maxima, mean of centers …) we have chosen to use the *centroid method*, as we believed it can provide a good representation of the fuzzy set. The centroid method exploits the following formula:

$$k = \left(\int_0^1 x\, f(x)\, dx\right) / \left(\int_0^1 f(x)\, dx\right) \qquad (10)$$

were f(x) is the fuzzy set function.

The value k, obtained in output, is equal to the abscissa of the gravity center of the fuzzy set.

This value is also associated with the variance, obtained by the formula:

$$\sigma^2 = \left(\int_0^1 (x-k)^2 f(x)\, dx\right) / \left(\int_0^1 f(x)\, dx\right) \qquad (11)$$

With these two values, we determine $Trust^1_X(F,P)$. as the interval $[k-\sigma;k+\sigma]$.

2) $Trust_X(F,P)$

Once we get $Trust^1_X(F,P)$., we can determine the value of $Trust_X(F,P)$. In particular, we determine a trust value followed by an interval, namely the uncertainty on $Trust_X(F,P)$.

For uncertainty calculation we use the formula:

Uncertainty $= 1 - (1 - \Delta Trust^1_X) * Trust_X(Trust_F(P)) * Trust_X(Source(F,P))$ *(12)*

$\Delta Trust^1_X = Max(Trust^1_X(F,P)) - Min(Trust^1_X(F,P))$

In other words, the uncertainty depended on the uncertainty interval of $Trust^1_X(F,P)$, properly modulated by $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$.

This formula implies that uncertainty:

- Increase / decrease linearly when $\Delta Trust^1_X$ increase / decrease;
- Increase / decrease linearly when $Trust_X(Trust_F(P))$ decrease / increase;
- Increase / decrease linearly when $Trust_X(Source(F,P))$ decrease / increase.

The inverse behavior of $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ is perfectly explained by the fact that when X is not so sure that P derives from F or F's degree of certainty about P is low, global uncertainty should increase.

The maximum uncertainty value is 1 (+-50%) meaning that X is absolutely not sure about its evaluation. On the contrary, the minimum value of uncertainty is 0, meaning that X is absolutely sure about its evaluation.

In a way similar to uncertainty, we used the following formula to compute a value of $Trust_X(F,P)$:

1) If $Support_F(P) = 1$, namely F is supporting P

$$Trust_X(F,P) = ½ + (Trust^1_X(F,P) - ½) * Trust_X(Trust_F(P)) * Trust_X(Source(F,P)) \qquad (13a)$$

2) If $Support_F(P) = 1$, namely F is opposing P

$$Trust_X(F,P) = ½ - (Trust^1_X(F,P) - ½) * Trust_X(Trust_F(P)) * Trust_X(Source(F,P)) \qquad (13b)$$

This formula has a particular trend, different from that of uncertainty. Here in fact the point of convergence is ½, value that does not give any information about how much X can trust F about P. Notice that, if F is supporting P:

- If $Trust^1_X(F,P)$ is less than ½, as $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ increase the value of trust will decrease going to the value of $Trust^1_X(F,P)$; as $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ decrease the value of trust will increase going to ½;
- If $Trust^1_X(F,P)$ is more than ½, as $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ increase the value of trust will increase going to the value of $Trust^1_X(F,P)$; as $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$decrease the value of trust will decrease going to ½;

Conversely, when F is opposing P:

- If $Trust^1_X(F,P)$ is less than ½, as $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ increase the value of trust will increase going to the value of $Trust^1_X(F,P)$; as $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ decrease the value of trust will decrease going to ½;
- If $Trust^1_X(F,P)$ is more than ½, as $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ increase the value of trust will decrease going to the value of $Trust^1_X(F,P)$; as $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ decrease the value of trust will increase going to ½;

3) Computing a final trust value: sources' aggregation

How to evaluate the contribution of different sources? In general, the average value is given by the average of individual sources' trust value.

This issue gets more complicated when you need to find an average uncertainty value: computing the average of uncertainties is not enough. For instance, suppose we have two sources, the former asserting 0 with uncertainty 0 and the latter asserting 1 with uncertainty 0. Intuitively, a trust value of 0.5 **is fine by me**, but it is implausible that uncertainty is equal to 0; on the contrary, it should take the maximum value.

Thus it is easy to note how global uncertainty depends on both the single values of uncertainty and the single trust values. Plus we state that **the greater the number of convergent**

sources towards a trust value, the lower the uncertainty I have. Then the formula to compute this global value should take into account these factors.

The domain of uncertainty [0,1] has been divided into 5 intervals of amplitude 0.2. Values falling in the same interval are considered convergent. Here is the used formula:

$$\mathrm{Unc} = \mathrm{Unc}_0 + \sum_j \sum_i^I \mathrm{Unc}_i \,/\, (I*N) \qquad (14)$$

where:

Unc_0 = minimum distance value between the computed medium trust value and each single trust value (of every single source);

j = intervals, $1<j<5$;

I = number of convergent sources in the **j-th** interval;

N= total sources' number ;

Unc_i = uncertainty on **i-th** source.

Thus it is worth noting that it is better to have two sources asserting the same thing, even if with a given value of uncertainty, than two sources asserting opposing information, even if with the utmost certainty.

III. Conclusion

In this work we have analyzed the nature of trust in information source also on the basis of our previous works [1; 14].

We identified which components influence this kind of trust and showed how them contribute to the creation of trust. We also showed how the degree of trust in an information P strictly depends and derives from the X's trust in the sources producing it with respect the kind of information.

Finally we provided a detailed framework and a computational model to deal with this kind of problem.

We consider necessary to specify that, although we described the model and the variable that influence it, we have not investigated some important parameters (such as the weights of past experience, category and reputation). In fact we think that these values are strongly linked to the context in which the model is applied and should emerge from it.

Acknowledgment

This work is partially supported by the Project PRISMA (PiattafoRme cloud Interoperabili per SMArt-government; Cod. PON04a2 A) funded by the Italian Program for Research and Innovation (Programma Operativo Nazionale Ricerca e Competitività 2007-2013).

References

[1] Castelfranchi, C., Falcone R., Pezzulo, (2003) Trust in Information Sources as a Source for Trust: A Fuzzy Approach, Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-03) Melburne (Australia), 14-18 July, ACM Press, pp.89-96.

[2] Falcone R., Piunti, M., Venanzi, M., Castelfranchi C., (2013), From Manifesta to Krypta: The Relevance of Categories for Trusting Others, in R. Falcone and M. Singh (Eds.) Trust in Multiagent Systems, ACM Transaction on Intelligent Systems and Technology, Volume 4 Issue 2, March 2013

[3] Yolum, P. and Singh, M. P. 2003. Emergent properties of referral systems. In Proceedings of the 2nd International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS'03).

[4] Conte R., and Paolucci M., 2002, Reputation in artificial societies. Social beliefs for social order. Boston: Kluwer Academic Publishers.

[5] Sabater-Mir, J. 2003. Trust and reputation for agent societies. Ph.D. thesis, Universitat Autonoma de Barcelona.

[6] Burnett, C., Norman, T., and Sycara, K. 2010. Bootstrapping trust evaluations through stereotypes. In Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10). 241248.

[7] Falcone R., Castelfranchi C., Trust and Transitivity: How trust-transfer works, 10th International Conference on Practical Applications of Agents and Multi-Agent Systems, University of Salamanca (Spain)28-30th March, 2012.

[8] Joana Urbano, Ana Paula Rocha, and Eugnio Oliveira, Computing Con_dence Values: Does Trust Dynamics Matter? In L. Sabra Lopes et al. (Eds.): EPIA 2009, LNAI 5816, pp. 520-531, 2009, Springer.

[9] Kahneman, Daniel, and Amos Tversky, "Prospect Theory: An Analysis of Decision Under Risk". Econometrica. XLVII (1979): 263-291.

[10] Higgins, E. T. (1997). Beyond pleasure and pain. American Psychologist, 52, 1280-1300.

[11] Falcone R., Castelfranchi, C. (2004), Trust Dynamics: How Trust is influenced by direct experiences and by Trust itself; Proceedings of the 3rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-04), New York, 19-23 July 2004, ACM-ISBN 1-58113-864-4, pages 740-747.

[12] Sabater-Mir J., Sierra C., (2001), Regret: a reputation model for gregarious societies. In 4th Workshop on Deception and Fraud in Agent Societies (pp. 61-70). Montreal, Canada.

[13] S. Jiang, J. Zhang, and Y.S. Ong. An evolutionary model for constructing robust trust networks. In Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2013.

[14] Castelfranchi C., Falcone R., Sapienza A., Information sources: Trust and meta-trust dimensions . CEUR Workshop Proceedings 2014 (In press)