On How Agents Make Friends: Mechanisms for Trust Acquisition

Babak Esfandiari[1], Sanjay Chandrasekharan[2]

[1] Department of Systems and Computer Engineering,
Carleton University, Ottawa, Ontario, Canada
babak@sce.carleton.ca

[2] Cognitive Science Ph.D. Program,
Carleton University, Ottawa, Ontario, Canada
schandr2@chat.carleton.ca

Abstract. We need models of trust to facilitate cooperation in multi-agent systems, where agents, human and artificial, do not know each other beforehand. This paper lists and proposes simple mechanisms for trust acquisition based on a very basic and general definition of trust, making no assumptions on the internal cognitive models of the involved agents. We also show how trust acquired one-on-one can be propagated in a social network of agents.

1 Introduction

There has been a recent burst of interest on the topic of trust, due in part to the importance it plays in e-commerce applications [22, 28]. For example, in e-commerce, it is important to know the credentials of the buyer or the seller before initiating a commercial transaction [28]. From an AI perspective, an evaluation of trust is needed to facilitate task allocation, and therefore cooperation, between agents in an open, multi-agent, system setting. In the context of collaborative interface agents [21], it is important for agents to be able to identify their true peers, i.e. the ones that can give them *relevant* advice, and only relevant advice, and information to increase productivity.

Different authors have given various definitions for the term Trust, as well as properties that trust must verify. Marsh [23] uses the definition by [14], which is commonly accepted in the literature:

"…trust, (or symmetrically, distrust) is a particular level of the subjective probability with which an agent will perform a particular action, both before he can monitor such action (or independently of his capacity to monitor it) and in a context in which it affects his own action."

Castelfranchi [5] approaches trust from a different perspective, and considers trust to be "the mental counterpart of delegation". It is also argued that "only an agent with goals and beliefs can trust". Both authors have formal models to arrive at a trust metric. The Marsh model is a simple mathematical one, while the Castelfranchi model is logic-based.

There are other models of trust developed for multi-agent systems, like [17, 25, 15, 1].

However most definitions we know of do not capture the temporal and dynamic aspects of trust, and therefore the notion of *trust acquisition* in multi-agent systems has been mostly unexplored. In this paper we list and propose different constructive and pragmatic mechanisms for trust acquisition, to allow multi-agent systems to use trust as a factor for:

- decision making: when should an agent delegate a task to another agent
- learning: whom should an agent learn from, and take advice
- obeying another agent

We also illustrate our mechanisms with examples from RoboCup (www.robocup.org) and Network Management.

2 Definition of Trust

The views on trust can be broadly classified into two: the cognitive view and the mathematical view. In the cognitive view, trust is made up of underlying beliefs, and trust is a function of the value of these beliefs. An example of this kind of model is [5]. The mathematical view ignores the role of underlying beliefs and uses a trust metric, based on variables like Perceived_competence, Perceived_risk, Utility of a situation for the agent involved, Importance of a situation etc. These models incorporate some aspects of game theory and the evolution of cooperation models. An example of this kind of model is the one put forward by Marsh [23].

Both views see trust as a variable with a threshold for action. When the value of the variable crosses the threshold, the agent executes an action. In the Marsh model, the action is cooperation, in the Castelfranchi model the action is delegation. The action is boolean in nature, the agent either delegates or not, or the agent either cooperates or not.

We do not see any major contradiction in these views. Further, we believe that a Boolean decision on trust could be based on a probabilistic evaluation, based on a threshold value, which would be determined a priori. Our simple definition will not make any assumptions on a particular mental model that leads to a trust decision:

Trust is a function T between any two agents of a set A of agents:

```
T: A x A -> [0,1]
```
ex: T(Jules, Jim) = 0.8, Jules trusts Jim 80%

We define *Trust Acquisition* as the process or mechanism that allows the calculation and update of T. In our definition, acquisition is not necessarily an 'increase' of T.

The following sections will explore different Trust Acquisition Mechanisms, by describing different ways to calculate and update T.

3 Trust Acquisition

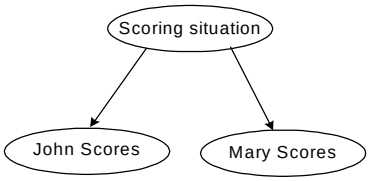
In this section we will first study mechanisms for one-on-one trust acquisition, and then we will focus on the propagation of that acquired trust in an acquaintance graph of agents.

3.1 Trust Acquisition by Observation

It is usually unlikely that an agent will have direct access to the mental states of another agent, which is needed to deterministically determine whether the agent can be trusted or not. Most trust decisions are based on observation, and it is an intuitive way to build a model. The observation model should capture the uncertainty that follows the (at least initial) unpredictability of the fellow agent. Given this, Bayesian networks are a logical choice of knowledge representation in this case.

Trust acquisition can then be performed by Bayesian learning [26]. In the simplest case of a known structure and a fully observable Bayesian network, the learning task is reduced to statistical considerations. Consider the following RoboCup-type of situation: Agents John and Mary are both in a potential goal scoring situation, John has the ball, and has to decide whether to pass it to Mary or go ahead and shoot the ball.

If John is able to observe past performances by Mary, consider the acquired statistics as a trust value and then ascribe them as a probability in the graph below, then by simply comparing them with his own performance, he will be able to make a decision. We are assuming here for the sake of simplicity that the fact of passing the ball will not have any effect on the situation, but indeed the success rate of passing could be another observed statistic to take into consideration.



S	P(J=T)	P(J=F)
T	60	40
F	10	90

S	P(M=T)	P(M=F)
T	70	30
F	10	90

$$T(John, Mary) = P(M=T|S)$$

Fig 1. John and Mary are both in a potential goal-scoring situation, John has the ball, and has to decide whether to pass it to Mary or go ahead and shoot the ball. John observes past performances by Mary, considers the acquired statistics as a trust value and then ascribe them as a probability in the graph. By simply comparing the probabilities with his own performance, John will be able to make a decision about passing the ball

More generally, for any two agents A and B, a delegation situation D, and an observed performance statistic S, we can evaluate observed trust T_{obs} as follows:

$$T_{obs}(A,B) = P(S|D)$$

Professional baseball players (and fans) are familiar with this type of reasoning all too well. One could also imagine reinforcement learning techniques: John passes the ball, Mary scores, so John receives positive feedback… For a more complex use of Bayesian networks to model agent relationships, see [2].

3.2 Trust Acquisition by Interaction

In [21], a simple set of protocols are proposed to allow learning interface agents to collaborate in order to learn from each other. It was important for such agents to determine which fellow agents to trust, since the users they were derived from possibly had very dissimilar behavior. The main protocols were:

- the exploratory protocol, where the agent asks the other agent questions for which it already knows the answer. The agent then increases the trust rating of the agents who give the expected answer.
- the query protocol, where the agent asks for advice from trusted agents.

Besides the email-filtering application described in [21], such protocols were also applied in the context of network management [11, 10]. Indeed, different network management operators can be ascribed different, and sometimes conflicting, tasks such as fault-monitoring, configuration or security management. Using the above protocols, their interface agents should then find their true peers with no prior knowledge about each other. The observed result of the ensuing collaboration was an acceleration in the learning process. However no new tasks were learned compared to the isolated case, which is not really surprising.

It is interesting to note that since the exploration and the query protocol only differ in the way the querying agent interprets the reply, such protocols can be used in contexts where not all agents are willing to collaborate: the queried agents cannot tell a test from an actual query.

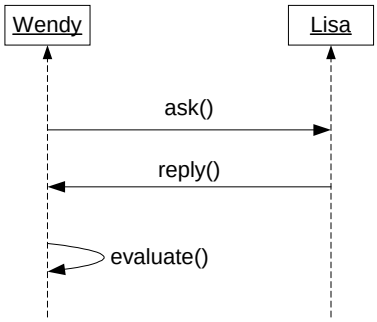


Fig 2. Trust(wendy, lisa) = wendy.evaluate(lisa, replies). The exploration and the query protocol only differ in the way the querying agent interprets the reply. Such protocols can therefore be used in trust contexts where not all agents are willing to collaborate: the queried agents cannot tell a test from an actual query

A simple and general way of evaluating interaction-based trust T_{inter} based on an agent B's replies to an agent A during the exploration phase is:

$T_{inter}(A,B)$ = number of correct replies / total number of replies

3.3 Trust Acquisition Using Institutions

In the human world, trust values are not arrived at using extensive computations. Humans "aggregate" trusting situations. These composite trust metrics are used in repeat situations to arrive at trust values. The aggregated trust values and situations are then shared. When these situations and values are shared by a community, these values become constants, and are then used to base a trust decision on. This is the notion of reputation. Reputation mechanisms are widely used to arrive at trust values in e-commerce settings [26]. Reputation is a metric that grows over time. However, reputations can also be created top-down, using institutions. Badges, uniforms etc., signify reputation based on institutions.

For instance, we will trust a person with our life if he is wearing a gun and a policeman's uniform. We will not trust him if he doesn't have that uniform. When the person doesn't have a uniform, we have to resort to more computations and queries to reach a trust value. When the person has a uniform, the trust situation becomes more easily computable, because the uniform is like a pattern, and all the agent has to do is pattern-matching.

However, note that this pattern is a *created* one. It is provided by an institution, to help in the trust computation. The uniform is an artificial pattern, created to make-up for the lack of exploitable structure in a trusting situation. In other words, we are changing the environment to aid in our trust computation.

Figure 3 [6], captures the relationship between traditional pattern-matching AI and Situated AI [4]. Instead of relying on a complex mental model of an unstructured world, we split the complexity, part to the agent, part to the environment. Then we provide structure to the environment for the agent to exploit, structure that gives the agent a "leg-up" in its computation. In the case of trust, institutional stamps (uniforms, degrees, seals etc) form the exploitable structure provided to the environment, so that agents can arrive at a trust value quickly.

Given an agent A and an agent B belonging to a given recognized institution, we can trivially assign institutionalized trust T_{inst} as follows:

$T_{inst}(A, B) = 1$

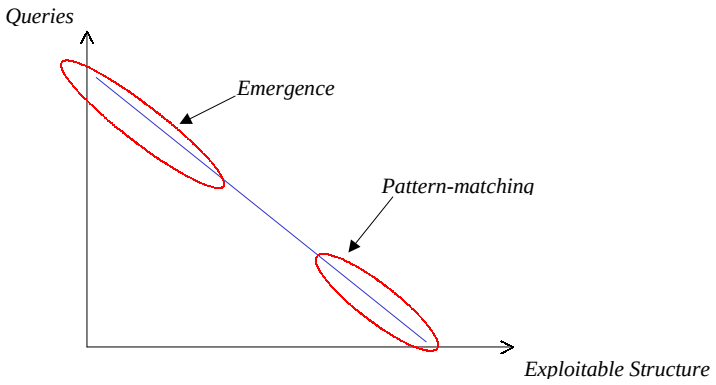


Fig 3. To reduce cognitive load, the agent will exploit structure in the environment to arrive at trust values. If there is more exploitable structure, the agent will resort to simple pattern-matching, if there is less structure the agent needs to query the environment more. Institutional signifiers like uniforms, badges, degrees etc. constitute exploitable structure for humans to arrive at trust values. Reputation mechanisms extend this result to multi-agent systems

Another aspect of institutionalization is the establishment of contracts and guarantees [3, 13]. A breach of contract should influence the trust value: hence the need for contract monitoring. Failure to fulfill a contract will then influence the further choice of contractees. In [12] a discounting factor is used in the evaluation of bidders in the contract-net protocol, and is based on past experience. It includes the possibility to unlearn past a given time window.

A badge or a uniform is just a signifier -- a pattern provided by institutions for people to use -- to minimize the cognitive load while doing trust computation. However, people also use other, informal, signifiers to reduce their cognitive load during trust computation. These signifiers range from skin and hair color to sex, nationality and dressing styles.

Since these are informal signifiers, they are not exactly institutional. They are considered the basis of prejudice, and computations based on these kinds of signifiers are usually considered prejudiced. However, these signifiers are, in a sense, institutionalized, because a large number of people use them.

Prejudice is acquired through induction, just as trust in institutions is acquired. And prejudice does decrease cognitive load. Given these similar properties, it is difficult to delineate between prejudice and institutional trust. From a theoretic machine learning view, prejudice formation can be seen as the result of underfitting the training data by following the Ockham's razor principle too closely.

4 Trust Propagation

Les amis de nos amis sont nos amis

Once one-on-one trust is established by agents with their peers using any of the above mechanisms, it becomes interesting to study the possibility of propagating it along a chain of acquaintances, to allow for trust acquisition based on second-hand (or more generally n^{th} hand) information. This is one form of reputation-based trust. We will take a Social Network Analysis approach [27] here, and apply it to the problem of trust.

Let us consider a graph representing a network of agents and their acquaintances. It consists in a directed, labeled graph, where an (a,b) edge represents the trust value that *a* has of *b*. Note that the choice of a directed graph highlights the fact that trust is not symmetric: T(a,b) is not necessarily equal to T(b,a). Also trust is not reflexive: T(a,a) is not necessarily equal to 1.

The figure below represents an example of a trust graph.

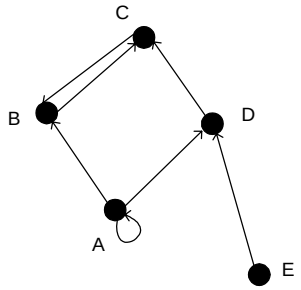


Fig 4. A directed graph for trust evaluation. In a multi-agent, distributed, setting, where the graph's edge values are not centrally known, the problem of calculation of the trust interval becomes equivalent to the problem of routing in a communication network

Edges are absent where the trust value is unknown. Our goal therefore will be to determine those values (i.e. create edges), as well as update the value of existing edges, based on trust propagation.

Trust is only weakly transitive [8], so our propagation model should take into account the decrease of trust along the chain. Here is a first possibility:

$T_{prop}(a,c) =? T(a,b_1) \times \ldots \times T(b_{1n},c)$ with (b_i) being the intermediate agents in a path from a to c

Problems are:

- different paths might give contradictory values
- cycles in a path can artificially decrease the trust value. In the above graph for example, one might decide to loop 3 times in A before reaching to a neighboring agent.

Our proposed solution to the above problems is to replace a strict trust calculation with a trust interval. The bounds of the interval are determined by calculating the minimum and maximum possible values obtained by applying the above calculation to the paths that contain no cycles. Complexity of this is clearly polynomial, as it is trivially equivalent to the shortest path problem.

More realistically, in a multi-agent and therefore distributed setting, where the graph's edge values are not centrally known, the problem of calculation of the trust interval becomes equivalent to the problem of routing in a communication network. Distributed algorithms such as in RIP [18] can apply. The merit of such algorithms is also that they allow the regular and incremental update of trust values.

Another problem with propagation is that the notion of trust might vary for each agent-agent relationship. Agents might build trust for different aspects of their acquaintances, for example assign trust for a particular task. Therefore we need to have colored edges, with a color per task or type of trust. And we would have a "multi-colored" edge for "general" trust. Trust would only propagate through edges of the same color.

4.1 Use of the Trust Interval for Decision-Making

In an optimistic setting the agent can use the max value as a decision threshold, whereas in a pessimistic setting the agent will use the min value. The optimistic setting models effectively the learning process used by children. The pessimistic one can be used in the case of delegation decisions in security related problems.

Long paths tend to exaggeratedly decrease the trust value. Does then low trust mean distrust, or rather a lack of a direct enough knowledge of that trust? A possible improvement in our propagation mechanism would be to restrict our calculation to paths shorter than an arbitrary n value. For instance, we could only accept up to third-hand information. Or we could accept the "freshest" value (shortest path, least number of intermediate agents involved), with an interval in the case of multiple paths. This is similar to the HISTOS model suggested by [28]. Calculating the average value can also be a good indication.

5 Conclusion and Future Work

This paper proposed, and referred to, a list of simple mechanisms for trust acquisition and propogation. Our trust propogation model is closely related to [28], which uses a directed graph method to arrive at trust values for human agents in e-commerce settings. The model uses trust evaluations provided by humans. We believe that as software agents integrate into virtual communities, the human-agent distinction will be blurred, and the issue of trust acquisition will become more and more important. We are currently planning on applying some of the proposed techniques in the RoboCup competition, as well as continue the application to network management.

One big limitation of our definition is that it does not make a distinction between distrust and lack of knowledge about trust. It would be interesting to use Ginsberg's bilattice [15] as an extension to our definition in order to capture both the trust value and the agent's certainty about that value.

We are also exploring how the notion of identity is related to trust, and how agents can be provided with identities. This is related to [22], which lists a set of mechanisms to enforce and verify trust in an e-commerce setting, such as relying on a transaction history, establishing contracts and guarantees with indemnities, and using Trust Authorities. It also proposes trust models to allow decision-making.

References

1. Abdul-Rahman, A. and Hailes, S.: Supporting Trust in Virtual Communities, Proceedings of the Hawaii International Conference on System Sciences, 2000

2. Banerjee B et al: Using Bayesian Networks to Model Agent Relationships, Applied Artificial Intelligence, 14:867-879, 2000

3. Brainov, S. and Sandholm, T.: Contracting with Uncertain Levels of Trust, Proceedings of the Autonomous Agents Workshop on Deception, Fraud and Trust in Agent Societies, 1999

4 Brooks, R. (1991).: Intelligence Without Representation, reprinted in Mind Design II, Ed. Haugeland, J. (1997), MIT Press, Cambridge, Mass. (1991)

5. Castelfranchi, C. and Falcone, R.: The Dynamics of Trust: From Beliefs to Action, Proceedings of the Autonomous Agents Workshop on Deception, Fraud and Trust in Agent Societies, 1999

6. Chandrasekharan, S. and Esfandiari, B.: Software Agents and Situatedness: Being Where?, in the Proceedings of the Eleventh Mid-west Conference on Artificial Intelligence and Cognitive Science, 2000, AAAI Press, Menlo Park, CA (2000)

7. Chandrasekharan, S. and Esfandiari, B.: Building Trustable Agents, Carleton University Research Report SCE-00-04, Ottawa, Canada (2000)

8. Dasgupta, P.: Trust as a Commodity. In Trust: Making and Breaking Cooperative Relations, Gambetta, D (ed.)., Basil Blackwell, Oxford (1990).

9. Dunn, J.: Trust and Political Agency. In Trust: Making and Breaking Cooperative Relations, Gambetta, D (ed.)., Basil Blackwell, Oxford (1990)

10. Esfandiari B. et al : Agent-oriented Techniques for Network Supervision, In Annals of Telecommunications, 51, no 9-10, 1996, p521-529.

11. Esfandiari, B.: Application de Techniques Orientees Agent en Gestion et Supervision de Reseaux, Ph.D. Thesis, Universite Montpellier II, 1997.

12. Esfandiari, B and Weiss, M.: System for Discounting in a Bidding Process Based on Quality of Service, Mitel Patent.

13. Ganzorali et al: The Social and Institutional Context of Trust in Electronic Commerce, Proceedings of the Workshop on Deception, Fraud and Trust in Agent Societies, Seattle, 1999
14. Gambetta, D. : Can We Trust Trust? In Trust: Making and Breaking Cooperative Relations, Gambetta, D (ed.)., Basil Blackwell, Oxford (1990).
15. Ginsberg M.: " Multivalued logics". Computational Intelligence 4(3), 1988.
16. Griffiths, N. and M. Luck.: Cooperative Plan Selection Through Trust, In Multi-Agent System Engineering - Proceedings of the Ninth European Workshop on Modelling Autonomous Agents in a Multi-Agent World, F. J. Garijo and M. Boman (eds.), Lecture Notes in Artificial Intelligence, 1647, 162-174, Springer-Verlag, Berlin. (1999)
17. Jones, A.J and Firozabadi, B.S.: On the Characterisation of a Trusting Agent – Aspects of a Formal Approach. Proceedings of the Formal models for Electronic Commerce (FMEC) workshop, 1999, available at (http://www.ecec.fbk.eur.nl/services/FMEC/proceedings.html)
18. Hedrick, C.: Routing Information Protocol, rfc1058.
19. Khare,R and Rifkin, A.: Weaving a Web of Trust. World Wide Web Journal, Volume 2, Number 3, Pages 77-112, Summer 1997.
20. Luhmann, N.: Familiarity, Confidence, Trust: Problems and Alternatives. In Trust: Making and Breaking Cooperative Relations, Gambetta, D (ed.). Basil Blackwell, Oxford, (1990).
21. Lashkari, Y., Metral, M. and Maes, P.: Collaborative Interface Agents. In Proceedings of the Twelfth National Conference on Artificial Intelligence, AAAI-Press, 1994.
22. Manchala, D.: E-Commerce Trust Metrics and Models, IEEE Internet Computing, March-April 2000, p36-44.
23. Marsh, S.: Formalising Trust as a Computational Concept, Ph.D. Thesis, Department of Computing Science and Mathematics, University of Stirling, Stirling, Scotland, (1994).
24. Misztal, B.: Trust in Modern Societies, Polity Press, Cambridge, Mass., (1996).
25. Riely, J. and Hennessy, M.: Trust and Partial Typing in Open Systems of Mobile Agents, Internal Report, School of Cognitive and Computing Sciences, University of Sussex, Sussex, England, (1998).
26. Russel S. and Norvig P.: Artificial Intelligence: A Modern Approach, Prentice Hall, New Jersey, (1995).
27. Wasserman, S. and Faust, K.: Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, England, (1994).
28. Zacharia, G., Maes, P.: Trust Management Through Reputation Mechanisms, Applied Artificial Intelligence, 14:881-907, 2000