Trust Model Architecture: Defining Prejudice by Learning

Marika Wojcik[1]
Jan Eloff[2], Hendrik Venter[2]

Information and Computer Security Architectures Research Group (ICSA)
Department of Computer Science, University of Pretoria
[1]{hibiki}@tuks.co.za
[2]{eloff, hventer}@cs.up.ac.za

Abstract. Due to technological change, businesses have become information driven, wanting to use information in order to improve business function. This perspective change has flooded the economy with information and left businesses with the problem of finding information that is accurate, relevant and trustworthy. Further risk exists when a business is required to share information in order to gain new information. Trust models allow technology to assist by allowing agents to make trust decisions about other agents without direct human intervention. Information is only shared and trusted if the other agent is trusted. To prevent a trust model from having to analyse every interaction it comes across – thereby potentially flooding the network with communications and taking up processing power – prejudice filters filter out unwanted communications before such analysis is required. This paper, through literary study, explores how this is achieved and how various prejudice filters can be implemented in conjunction with one another.

1 Introduction

Technological development has influenced the principles required to run a successful economy [1]. However, the advent of new technologies and the subsequent implementations thereof have resulted in exposure to new risks. Two risk factors exist that continually drive research towards lessening the risks encountered: effective communication and security.

In order to accomplish an organisation's desired task, effective and timely communication is required. An organisation makes use of technology to communicate and share information. This information is an asset to the organisation and is used to assist decision-making processes. It is important that this information be reliable and accurate so that it can be trusted [2].

Organisation-owned information is usually sensitive in nature and requires protection against threats such as tampering. This creates the dilemma where the risk of gathering new information for organisational growth is weighed against the value of protecting information currently in an organisation's possession. Existing information often needs to be shared in order to acquire new information. Compromise of sensitive information can lead to serious negative consequences.

Trust models have been proposed in order to minimise the risk of sharing and successfully analysing information [3], [4]. Trust models rely on the abstract principle of trust in order to control what information is shared and with whom. Trust models evaluate the participants of a transaction and assign a numerical value, known as a trust value, to the interaction. This numerical value is used to determine the restrictions placed on the transaction and the nature of information shared. Information is classified by sensitivity and highly sensitive information requires a transaction to have a high trust value before such information is to be shared. This process occurs with all interactions a trust model encounters. In order to control the number of interactions a trust model encounters, prejudice filters have been proposed.

This paper introduces and defines the concepts of prejudice, trust and trust models in Section 2 by introducing a basic trust management architecture and expanding on work already done in these areas. The concept of prejudice filters and their interdependencies is explored in Section 3, with special focus on one relationship involving the learning filter. This is followed by a discussion of concepts in Section 4 and a conclusion in Section 5.

2 Background

Since trust model architecture is based on the concept of trust, a basic understanding of trust is required. This section introduces the concept of trust in the context of human relationships and then explores how this concept is put into practice by trust model architecture. The concept of prejudice is also explored, with special attention to how this concept can lighten communication load required to make trust-based decisions.

2.1 Trust Models and Trust

Trust models rely on the concept of agents [4]. Within the context of trust models, an agent refers to a non-human-coded entity used to form and participate in machine-based trust relationships. This agent would usually be situated on a computer and implement some form of logical rules to analyse the interactions with which it comes into contact in order to determine whether another agent is to be trusted or not. These logical rules may be static or adjustable by the agent in a dynamic manner, based on results of transactions the agent has participated in.

Trust is a subjective concept – the perception of which is unique to each individual. Trust is based on experience and cognitive templates. Cognitive templates are templates formed by experiences that are later used to analyse future experiences of a similar nature. Trust is dynamic in nature and influenced by environment, state and situation. According to Nooteboom [5], "*[s]omeone* has trust in *something*, in some *respect* and under some *conditions*".

Each of the four key concepts highlighted by Nooteboom exists within trust model architecture. *Someone* and *something* define two agents participating in an interaction. The former refers to the instigator of the interaction whiles the latter refers to the agent accepting the request. The *respect* is defined by the reason for instigating an

interaction. Finally, the *conditions* refer to the situational factors that influence the success of an interaction.

2.2 Trust Model Architecture

Trust models assist agents that have not previously encountered one another by forming and participating in trust-based interactions. Various experts have already proposed numerous trust models [6], [7], [8]. A survey of the literature conducted by the author has identified four components that have been used in trust model implementation: trust representation, initial trust, trust dynamics and trust evaluation.

Catholijn M. Jonker and Jan Treur [9] focus on how trust is represented by agents in order to simulate intelligence and make trust-based decisions. They propose a simple qualitative method of representing trust that defines four basic trust values. These values include unconditional distrust, conditional distrust, conditional trust and unconditional trust. Other issues of trust representation include whether the data used is qualitative or quantitative and even whether distrust parameters should be incorporated as separate values as proposed by Guha *et al.* [10].

Jonker and Treur in further research state that trust models incorporate trust characteristics that can be divided into two states. These states refer to initial trust – the initial trust state of an agent – or trust dynamics – the mechanisms that allow for the change in and updating of trust [9]. The initial trust state of an agent determines the agent's predisposition wherein the agent can be predisposed towards trust, distrust or neutrality. Taking the dynamic nature of trust in consideration, Marx and Treur [8] concentrate on a continuous process of updating trust over time. Experiences are evaluated and used by a trust evolution function.

Changing trust values requires that some form of trust evaluation should take place. The reputation-based model of Li Xiong and Ling Liu [11], known as PeerTrust, emphasises the importance of this evaluation process by evaluating various parameters, such as nature of information shared and purpose of interaction, in order to update the trust value an agent retains.

Trust models are able to obtain trust values in several manners. Trust information and state can be pre-programmed into the agent as a list of parameters. These parameters can also be dynamically formulated, based on pre-defined and logically formed trust rules that an agent uses to evaluate trust.

2.3 Example of a Typical Trust Architecture

According to Ramchurn *et al*. [12] basic interactions among agents go through three main phases. These phases are negotiation, execution and outcome evaluation. Trust plays an essential part in all three of these phases. This is illustrated by Figure 1.

Two agents attempting to communicate with one another are first required to establish a communication link, usually initiated by one agent and accepted by another. This process initiates a negotiation process whereby two agents negotiate various parameters, such as the security level of information that is to be shared or the services for which permission will be granted, that will define boundaries of the interaction. A trust value for the interaction is defined through comprehensive analysis

of logical rules. The simplest way of storing and implementing these rules is to have them present in a list that the agent accesses and processes. In Figure 1, storage of these rules occurs in the trust definition list.

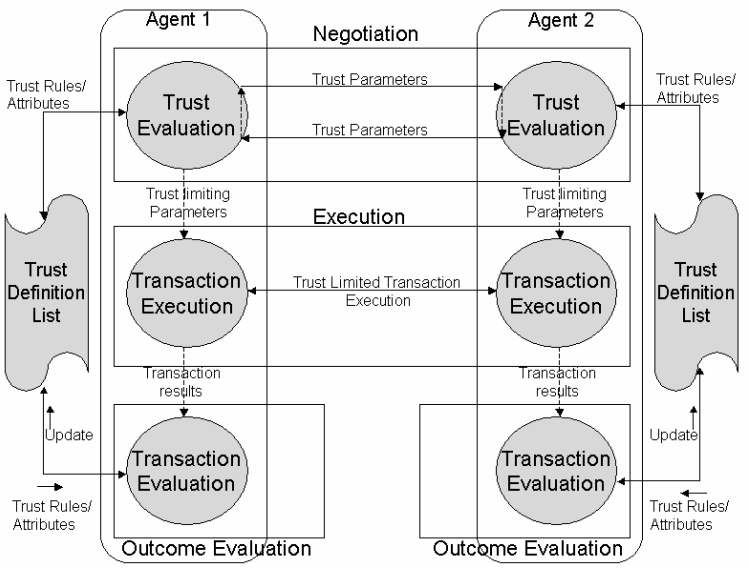

Figure 1: Operation of an agent using a trust model

The successful negotiation and establishment of a trust value triggers an analysis of the trust value. Provided the trust value is above a certain acceptable threshold, the transaction execution process is started. Trust models control the context of the interaction during the execution phase, limiting trust given and hence controlling which information or services are accessible and which are not.

Once transaction execution has terminated, the results of the interaction are sent to the transaction evaluation process. This process evaluates the results and updates the trust definition list in either a positive or negative manner. Negative updating of the logical rules occurs due to business transaction failure, while business transaction success will trigger a positive update.

The evaluation of trust among agents is a time-consuming process that requires comprehensive evaluation of the defined logical rules in order to attain an accurate trust value to be used during an interaction. Only once the trust value has been obtained, the agents will decide whether to participate in a transaction or not.

In a networking environment, the amount of possible agents that will request participation in such an interaction can be vast. To successfully assess another agent, agents pass several messages to obtain the required information that is to be analysed against the defined trust parameters. For instance, the formal model for trust in dynamic networks proposed by Carbone, Nielson and Sassone [7] passes delegation information between agents in order to create a global trust scheme. Delegation allows a particular agent to trust another agent, based on the fact that the other agent is trusted by agents that the agent in question trusts. This reliance on the passing of

messages exposes the network to the possibility of network overload. Another potential problem arising during the process of establishing trust is the level of comprehensiveness required by the analysis process. Having a large number of strict rules define a trust relationship limits the communications an agent will be able to participate in, while at the same time adding to the analysis load. Rules that are too generic open the system up to a higher level of risk by allowing an agent to participate in interactions with other agents that have not been fully analysed for trustworthiness.

Prejudice filters have been proposed to lessen the number of interactions that require comprehensive trust evaluation [13] so as to solve the problems mentioned above. Stereotyped grouping of interactions allows for characteristics to be assumed instead of evaluated in detail. It also allows trust evaluation to focus on characteristics that are not assumed, instead of evaluating the interaction against the entire list of logical rules that represent expectations.

3 Prejudice Filters

In order to understand the concept of prejudice filters, an understanding of prejudice is required. Prejudice is an extension of the concept of trust-building processes and is defined as a negative attitude towards an entity, based on stereotype. All entities of a certain stereotyped group are placed in the same category, allowing assumptions to be made and simplifying the processing required before trust can be established [14].

Agents see prejudice filters as simplified trust rules that rely on the concept of prejudice in order to limit the number of interactions an agent needs to analyse in detail. Prejudice filters rely on broad definitions of attributes that lead to distrusted interactions, thus denying interactions that can be defined by these attributes. For example, if an agent has interacted with another agent from a specific organisation and the interaction failed in terms of expectations, future requests from agents belonging to the same organisation will be discriminated against. Figure 2 illustrates where prejudice filters extend the trust architecture as originally depicted in Figure 1.

Prejudice filters affect two phases of the three-phase interaction cycle: the negotiation and outcome evaluation phases. In the negotiation phase, the prejudice filters are consulted first to provide a quick, simplistic evaluation of trust in order to filter unwanted communications before they are required to go through detailed trust evaluation and definition. Once an interaction has passed the prejudice evaluation, it moves onto the trust evaluation in order to acquire a trust value. When the execution phase concludes, the outcome evaluation phase includes the prejudice parameters when it evaluates the interaction. Failed transactions update the prejudice filters in order to filter out other transactions of a similar nature at an earlier stage.

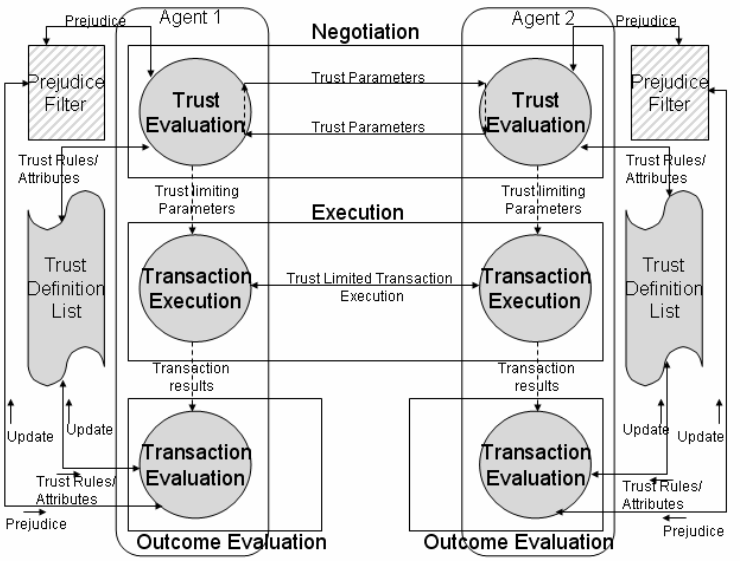

Figure 2: Operation of an agent using a trust model with prejudice filters

3.1 Extending Existing Models to Include Prejudice Filters

Existing trust models rely on various means of establishing trust, which include recommendation, reputation, third party reference, observation, propagation, collaboration, negotiation and experience [1], [2], [4], [6], [7], [8], [9], [10], [11], [12]. Based on these, five means of implementing prejudice filters have been identified by the author in order to simplify the extension of existing models to include prejudice. These five are as follows [13]:

Learning: When using the learning filter, prejudice is not defined explicitly. An agent relies on 'first impressions' to learn prejudice. If an interaction fails, the agent analyses the interaction's attributes and looks for unique attributes of other interactions that were previously encountered and found to be successful. These unique attributes are used to create a category to be used as a prejudice filter.

Categorisation: An agent creates various categories that are trusted. If an interaction request does not fall into a trusted category, the agent filters out that interaction in a prejudiced manner. This can also be implemented in a reverse manner where an agent creates categories that are distrusted and filters out communications that fall into those categories. Categories can also be created to represent various levels of trust. Any interactions falling into such categories are assigned the default trust value associated with that particular category.

Policy: Policies define the operational environment in which an agent exists and affect parameters of interactions that are regarded acceptable. Policy-based prejudice filters out interactions with agents whose policies differ from the agent doing the filtering. One way of doing this is to request data on the country an agent resides in. Such data defines the laws and culture that bind business interactions for that agent, as well as controls the means in which data and confidentiality are is handled.

Path: Path-related prejudice allows an agent to refuse an interaction, simply because of the fact that the path of communication between two agents passes through a distrusted agent.

Recommendation: Agents that are trusted to make recommendations are known as recommender agents. Implementing prejudice by using recommendation allows a particular agent to only trust other agents that are trusted by the particular agent's recommender agents.

The above five filters can be incorporated into current trust models to extend their capability, while at the same time allowing for these filters to merge with a particular trust model's main philosophy. Just as some models use a combination of concepts to implement the concept of trust, interrelated filters can be implemented in different combinations in order to optimise their effectiveness.

3.2 Defining Interrelationships between Filters

The five prejudice filters discussed above can be organised into a structure of relationships as shown in Figure 3. This structure depicts relationships that exist between these filters. The root node of a relationship between two prejudice filters indicates the dominant filter. The second filter can be incorporated into the workings of the dominant filter when the two are implemented together. The directional arrows in Figure 3 illustrate this. The dominant filter is situated at the tail of the directional arrows. Two prejudice filters emerge as more dominant than the others: learning and policy. These prejudice filters are always situated at the tail end of the arrows in Figure 3 and can be implemented in conjunction with all the other lesser filters.

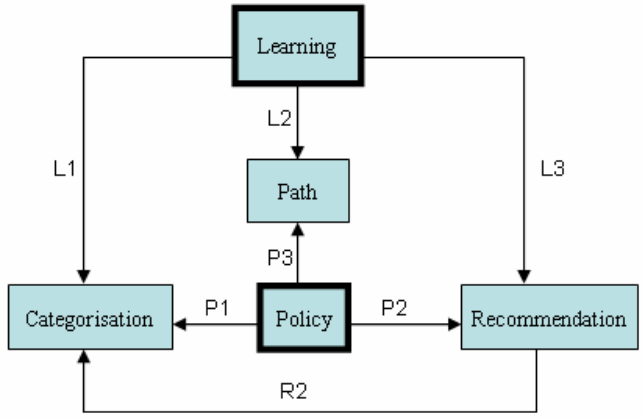

Figure 3: Overview of the inter-relationships between prejudice filters

Due to space constraints, only one of the illustrated relationships is explored, leaving the rest for further discussion in future work. The relationship discussed has the learning filter as its root node and is labelled L1 – linking the learning and categorisation prejudice filters.

Learning-Dominated Relationships. The nature and success of learning is governed by the nature and variety of information and experience that an agent is exposed to [15]. Experiences and information are filtered to form templates unique to each agent. Templates are default rules that have been formed by experiences and that are subsequently used to evaluate other similar experiences.

When using learning, prejudice is not defined explicitly, and an agent relies on 'first impressions' to learn prejudice. Possible implementation of this concept allows an agent to begin with a basic set of rules that it uses to evaluate the success of an interaction. Initially, the agent will interact with any agent with which it comes into contact, under restricted conditions of trust. Each interaction instigates an analysis process by means of which the agent will identify parameters such as location of an agent, security of information required, and even factors such as an agent's reputation. These parameters become the characteristics of the particular interaction and should the interaction fail, they will be analysed in order to identify a means of filtering out future interactions of a similar nature.

Due to the fact that learning creates various forms of templates [16], learning various forms of prejudice can be accomplished. One of these is discussed below.

Learning by Categorisation (L1). Categorisation is an umbrella term that allows for objects or concepts with similar attributes to be grouped together. This allows for certain assumptions to be made in order to simplify analysis of such objects. The attributes that can be assumed are those that define a certain object or concept as belonging to a specific category. For instance, agents that belong to the same policy category are assumed to hold similar policy values, such as information privacy constraints. Only agents from acceptable categories will be sent for trust evaluation by an agent wishing to interact with another. Agents that are defined as unacceptable at the onset of the interaction are discarded before entering the comprehensive trust evaluation phase. This eases the processing load by filtering out undesirable categories before sending the interaction to the trust evaluation process which determines a trust value.

The process of learning prejudice relies heavily on categorisation. Learning analyses a transaction to determine its unique features. If the transaction fails, the agent uses this analysis process to create a category of failure to be used in future category-based prejudice decisions. Implementation of this concept relies on allowing an agent to form categories defined by the trust rules in place. For instance, if the trust rules in place require transactions to be analysed in order to determine the policies used by the agents in question, these agents can be categorised by their policies and characteristics. Agents can be categorised by their core services, products and policies [17].

An agent is required to either keep a list of categories that are trusted or categories that are not trusted. Whenever a new interaction is encountered, the interaction is analysed against the characteristics of the various categories in order to define the category the interaction belongs to. Once the category has been defined, the agent checks its list of trusted or distrusted categories in order to determine whether interactions of that nature are trusted. If the interaction type exists in the distrusted categories list or alternately does not exist in the trusted list, the interaction is seen as

distrusted and is then discarded. Unknown or undefined categories are by default considered to be distrusted.

Categorisation can also be used to define different levels of trust. This is accomplished by assigning a default trust value associated with a category to agents that fall into that category. The rights delegated to an interaction are consequently limited by the category to which it belongs [6]. An example of such a category is role. Various roles are given differing rights. An administrative role is given more access rights than a client role.

4 Discussion

The concept of implementing prejudice as discussed in this paper is a very new concept that still requires further experimentation and analysis. One of the shortcomings of these filters is related to the fact that they allow machines to deny access due to the values of prejudice that were obtained.

This can lead to a situation in which agents that are in actual fact trustworthy are seen as untrustworthy, simply because of the prejudice filter in place. A situation like this, however, can be controlled by allowing agents to interact with several agents with similar defined characteristics before deciding prejudice against them. Increasing the number of interactions in which an agent participates increases the risk an agent is exposed to. Thus, there is a trade-off between accuracy of prejudice prediction, and the risk an agent is willing to take.

5 Conclusion

This paper has introduced the concept of trust models and prejudice. Different means of incorporating prejudice include categories, policies, path, recommendation and learning. Several of these filters are related in such a manner that they may be implemented in conjunction with one another. One of these relationships, namely that between learning and categorisation, has been explored and defined by this paper.

The authors have explored this topic from a conceptual standing that requires implementation and testing. Since only one relationship was scrutinised in this paper, further work requires more detailed investigation of the other defined existing relationships. More in-depth work needs to be done on means to standardise the representation of trust-related data, thus allowing agents from various platforms and using various models to efficiently interact with one another.

References

1. Hultkrantz, O., Lumsden, K., E-commerce and consequences for the logistics industry. In: Proceedings for Seminar on “The Impact of E-Commerce on Transport.” Paris (2001)
2. Patton, M.A., Josang, A., Technologies for trust in electronic commerce. In Electronic Commerce Research, Vol 4. (2004) 9-21

3. Abdul-Rahman, A., Hailes, S., A distributed trust model: new security paradigms workshop. In Proceedings of the 1997 workshop on new security paradigms, Langdale, Cumbria, United Kingdom, (1998) 48-60
4. Ramchurn S.R., Sierra, C., Jennings, N.R., Godo, L., A Computational Trust Model for Multi-Agent Interactions based on Confidence and Reputation. In: Proceedings of 6th International Workshop of Deception, Fraud and Trust in Agent Societies, Melbourne, Australia, (2003) 69-75
5. Nooteboom, B., Trust: forms, foundations, functions, failures, and figures. Edward Elgar Publishing, Ltd., Cheltenham UK. Edward Elgar Publishing, Inc. Massachusettes, USA. ISBN: 1 84064 545 8 (2002)
6. Papadopou, P., Andreou, A., Kanellis, P., Martakos, D., Trust and relationship building in electronic commerce. In: Internet Research: Electronic Networking Applications and Policy, Vol 11. No. 4 (2001) 322-332
7. Carbone, M., Nielsen, M., & Sassone, V., A formal model for trust in dynamic networks. In: Software Engineering and Formal Methods. In: Proceedings of the First International Conference on 25-26 Sept. (2003) 54-61
8. Marx, M., Treur, J., Trust dynamics formalised in temporal logic. In: Proceedings of the Third International Conference on Cognitive Science, ICCS (2001) 359-362
9. Jonker, C.M., Treur, J., Formal Analysis of Models for the Dynamics of Trust based on Experiences. In: Proceedings of MAAMAW'99. LNAI (1999).
10. Guha, R., Kumar, R., Raghaven, P., Tomkins, A., Propagation of Trust and Distrust. International World Wide Web Conference. In: Proceedings of the 13th international conference on World Wide Web . New York, NY, USA. (2004) 403-412
11. Xiong L., Lui L., A Reputation-Based Trust Model for Peer-to-Peer eCommerce Communities. E-Commerce, IEEE International Conference on 24-27 June (2003) 275-284
12. S.R., Sierra, C., Jennings, N.R., Godo, L., Devising a trust model for multi-agent interactions using confidence and reputation. In: Applied Artificial Intelligence. , Vol. 18., (2004) 833-852
13. Wojcik, M., Venter, H.S., Eloff, J.H.P., Olivier, M.S., Incorporating prejudice into trust models to reduce network overload. In: Proceedings of South African Telecommunications and Networking Application Conference (SATNAC 2005). SATNAC, Telkom, CD ROM Publication. (2005)
14. Bagley, C., Verma, G., Mallick, K., Young, L., Personality, self-esteem and prejudice. Saxon House. , Teakfield Ltd, Westmead. Farnborough, Hants. England. ISBN: 0 566 00265 5 (1979)
15. Bowling, M., Manuela, V., Multiagent learning using variable rate. In: Artificial Intelligence. Vol. 136 (2002) 215-250
16. Dasgupta, D., Artificial neural networks and artificial immune systems: similarities and differences. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '97), Orlando, October 12-15 (1997)
17. Siyal, M.Y., Barkat, B., A novel Trust Service Provider for the Internet based commerce applications. In Internet research: electronic networking applications and policy, Vol. 12(1) (2002) 55-65