

CAN TRUST BE ENGINEERED?

Ben Humphreys (bh704@soton.ac.uk)

1 Introduction

Trust is a social phenomenon that has been present in human society for thousands of years and is a key aspect of local and global interactions, whether personal or business-related. It is beneficial for individuals to be trusted, as this trust can often be used to gain something from the trustee. This benefit leads individuals to try to manipulate others to trust them more, and to reduce the trust in competitors. Trust can also lead to betrayal, in which a trusted individual can exploit the trust in them for their own purposes. In all these situations, there is a reward and a cost involved in each action. Trusting an individual has risks, and being betrayed is costly.

Trust can be broken down into four elements: the *truster* who is placing their trust in a *trustee* concerning a particular *topic*, and trusting them to a certain *degree*. To elaborate, trusters and trustees can range in scale from individual human beings to international entities. The subjective nature of trust is also clear, individuals trust others concerning a specific set of topics. For example one might trust a chef to make a meal, but not to give financial advice. It is also important to point out that trust is not binary, there is infinite variation in the amount of trust individuals can place in others, which has led to some issues of how to represent this trust (see Section 3.3).

1.1 Benefits and Betrayals

Being trusted is beneficial to an individual and is manifested in the individual's ability to interact with others. A trusted businessman can conduct more business transactions than someone one who is untrustworthy, and therefore is more successful. However by placing trust in someone, there is the risk of being betrayed.

The Theory of Games and Economic Behaviour [1] created the field of game theory which postulated that social interactions between humans could be quantified and simplified to resemble games. From this ideal strategies could be created and more could be learnt about the way humans interact. One of the most famous games to come from game theory is the Prisoner's Dilemma, in which two individuals A and B can choose to stay silent or betray each other (see Figure 1).

The idea of game theory is important when considering how to engineer trust, as it can give some indications of the way in which humans interact. While the technical considerations (covered in section 2) are extremely important for protecting honest users, by understanding the reasons *why* some users try to cheat the system, we can hope to design the benefits for co-operation to far outweigh the rewards for betrayal.

It is important to remember that while simplistic models can be used to give an indication of how greater society operates, if the values associated with each action do not accurately represent the real world, the usefulness of the model is greatly reduced. To illustrate the imperfection of some models, the safest strategy for the prisoner's dilemma is to betray your opponent every time. However when a similar game was given to the secretaries of a famous game theory researcher,

		B	
		Stays Silent	Betrays
A	Stays Silent	Both: 10	A: -20 B: 20
	Betrays	A: 20 B: -20	Both: -10

Table 1: *Prisoner's Dilemma*: A version of the prisoner's dilemma. Players do not know what their opponent will choose each round. The safest option to avoid betrayal and incurring a huge loss is to always betray your opponent. However the greatest stable solution is through co-operation.

they often co-operated and chose not to betray the other [2]. This demonstrates the limitations of game theory models, in that they cannot represent the good-willed and sometimes irrational nature of human beings. People gamble and take risks if the perceived benefit is large enough.

1.2 Digital Age

Trustworthiness is a property that is ingrained in human society. An individual or institution can appear to be more trustworthy by projecting certain characteristics. Humans are able to pick up on these signals even subconsciously and constantly make judgements about the trustworthiness of people. For example men that are clean-shaven and people that have more symmetrical features are deemed to be more trustworthy than others [3, 4]. In the case of shops, those that appear to be doing a large amount of business, with many items in stock and many customers are perceived to be more trustworthy than smaller shops.

However these indicators are not available when deciding whether to trust an individual remotely in the digital age. Technology has been developed – with varying degrees of success – to try to replace these social indicators by more reliable guarantees and engineer trust. From the technical aspect of secure connections and centralised authentication institutions, to the visual aspect of display ‘secure’ icons on web browsers and digital certificates on web pages. Even with these visual markings, the vast majority (around 90%) of both new users and expert users can be fooled into trusting a well-made fake site [5].

On the subject of how to engineer trust in a digital environment, the problem is two-fold: the social and emotional aspects of human trust must be catered for, along with the supporting increasing number of automated systems that are used for data gathering, shopping and competing in auctions. This essay is concerned with trust in the modern age, and whether it is possible to engineer trust.

2 Authentication

One important aspect of trust that is often not explicitly mentioned is that of authentication. Before any amount of trust can be placed in another individual, the identity of that individual must be confirmed. This idea of trust through authentication is shown throughout human society: driver's licenses, passports, birth certificates can all used to confirm the identity of someone.

In the digital world there are a number of similar methods that can be used to verify a user or entity's identity. Indeed most practical work to date concerning trust on the internet has focussed more on security, authentication and identity verification than that of the social aspect of what makes someone ‘trustworthy.’

2.1 Public-Key Encryption

Public-key cryptography is a technique that is at the core of authentication. It can be used to encrypt messages using a pair of keys, one public and one private [6, 7]. The encryption method used is such that a message encrypted by a public key can only be decrypted through its private counterpart. As illustrated by Diagram 1, the sender encrypts the message using the receivers public key and sends it to the receiver. The receiver then decrypts the message using their private key.

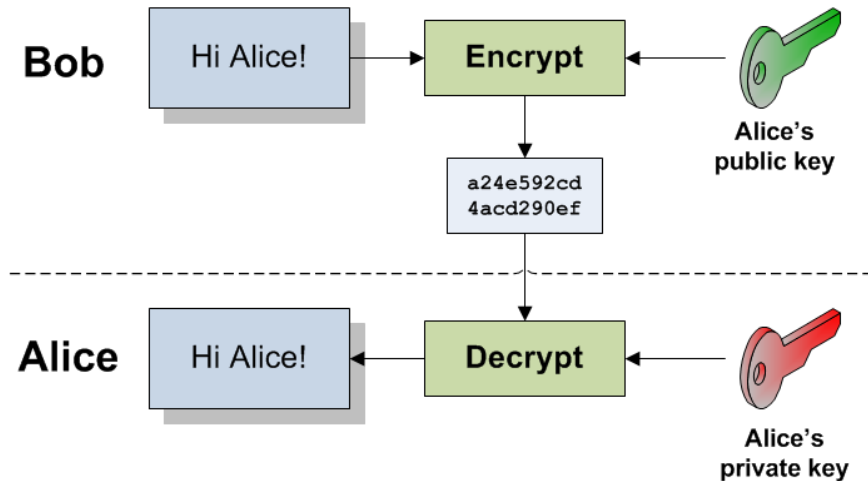


Figure 1: *Public-private Demonstration*: Overview of the system architecture.

Public-key encryption can also be used to add ‘digital signatures’ to documents. A hash of the document to be signed is created, and encrypted using the author’s private key. This encrypted hash is sent with the original document and decrypted using the recipient using the author’s public key. The decrypted hash is checked with the hash of the received document and if found to match, the signature is deemed valid.

3 Trust

Once the identity of both parties has been confirmed, the next step is to create a method of evaluating the trustworthiness of the other party. At present, the majority of widely-deployed trust systems are simple feedback forms in which users can rate each other. Despite this crudeness, they give users some indication of the trustworthiness of others and have proved extremely popular. However as section 3.2 goes on to show, there are a number of more complex strategies that can be deployed.

3.1 Pseudo-trust Systems

Through necessity, some applications with a large number of users have implemented pseudo-trust systems. While not as theoretically rigorous as some of the trust strategies being researched, it is useful to examine their effectiveness.

User Reputation Users of the system have a public rating which indicates their reputability or trustworthiness within that community. Individuals can rate other users they interact with, based on how they felt about the interaction. This leads to users with high ratings from a large number of people are deemed more trustworthy. To take the example of eBay, both buyers and sellers can

rate the other party after an auction, from the options negative, neutral or positive. This kind of system is relatively simple to implement and quickly pays off for honest customers that use the site often. However it has a number of drawbacks [8]:

- New users have no rating either way, and so are not trusted.
- Users must trust the system itself for it to work – it will not work if there are perceived flaws in the system.
- Users may be less likely to give bad feedback for fear of ‘retaliatory’ bad feedback from the other party. If users are restraining themselves the system becomes less useful.
- The evaluation system must be compulsory or the incentive for rating the other party must be high for the system to work. A study of eBay showed that less than half of all transactions received a rating, which would lead to reduced effectiveness [9].

3.2 Trust Strategies

While the current state of the web focusses on human users, current research is aiming towards creating more automated systems that can interact with each other independently [10, 11]. A simple example would be a shopping agent, a small program that automatically checks shop prices and quality reviews to find the best value for money. However the issue of trust in sources of information arises. This trust in other individuals and sources is not static, but a process of constantly revising trust levels based on events. A *trust strategy* is the name given to the function that defines an agent’s initial trust unknown individuals and how this trust is modified by trustworthy acts and betrayals.

Human beings have an extremely flexible and complex system of trust strategies that depend on a number of factors: the subject in question, the individual’s relationship with the other party and other situational. While humans possess an infinite number of variations on trust strategies, most research on intelligent agents has used strategies that fall into a small number of distinct categories [12]. Each strategy has it’s own strengths and weaknesses that determine which environments it is best used in.

3.2.1 Pessimistic

Pessimistic agents do not trust any sources at first, until they have evidence to do so. As this technique takes longer than others to build up a list of trustworthy sources, it is often combined with a technique known as ‘bootstrapping’ which manually adds an initial set of trustworthy sources. This strategy is very conservative and is in theory less likely to mistakenly trust untrustworthy sources. This approach is suited to applications and tasks that have extremely high costs for betrayal that outweigh the need for trusting more individuals, such as those involving financial transactions.

3.2.2 Optimistic

The logical opposite to the pessimistic approach, optimistic agents will naively trust all sources until they have reason to do otherwise. This strategy is best used when interacting with large amounts of other agents and sources. For example as the agent trusts all sources, it would allow for data-gathering from a huge number of sources. However optimistic strategies are prone to exploitation by untrustworthy agents.

3.2.3 Centralised

A centralised trust strategy uses a number of trustworthy institutions to provide lists of other trustworthy agents and sources. This reduces the decision-making needed by individual agents, but creates a single point of failure in the trust system. If trust in the centralised system is eliminated, whether through security compromise or social changes, the whole system fails as no agent can determine which other agents are trustworthy. There are also issues of scalability, central servers being faked and the possibility of bias.

3.2.4 Investigation

In contrast to the centralised approach, an investigative trust strategy is one in which agents communicate their opinions on each other in a peer-to-peer manner. Through these messages, the agents create a distributed network of agreed trust levels. However this strategy can be exploited by groups of malicious agents that recommend each other while bad-mouthing competitors.

3.2.5 Transitivity

While not a trust strategy in itself, the issue of transitivity is key to most collaborative trust strategies. When agents have a direct connection to their source, the trust they place in that source is determined by them alone. However when trust information is gathered indirectly through one or more third-parties, it becomes necessary to modify this third-party information in a certain way. For example one strategy might be to place more credence in information coming via trusted sources, but also reduce the amount of trust placed by half for every step the information is removed from the truster.

At first the issue of transitivity is simple to grasp, but there are a number of more complex issues: How should the trust information be modified depending on the task? How is it best to prevent malicious users from creating large numbers of agents that falsely recommend certain services? How should multiple routes via different agents be dealt with?

It is clear from the large amount of research that there is no single perfect trust strategy [13]. It seems that rather than a single best strategy for the semantic web being developed, the myriad of strategies created by research will be used in different situations. More conservative strategies lend themselves to agents that look for reliable scientific data, whereas those for more casual use are able to be more nave.

3.3 Representation

In order to engineer and communicate trust information, it must be represented in some way. As stated in the introduction, trust is composed of four elements: truster, trustee, topic and degree. At first the issue of representation seems trivial, but on closer inspection it is extremely difficult. Unique references can be used to represent the two parties involved. The topic area is slightly more difficult, if one trusts an individual for one topic, is it prudent to trust them on all sub-topics?

The degree to which you trust the fact is also more complex, most commonly it is thought to be best represented as a continuous range from 0 to 1.0, but what do these numbers actually represent? Should facts with a degree of 0.0 be totally disregarded? As mentioned in section 3.2.5, in transitive systems trust can come from third-parties, and so a piece of trust information cannot be reduced to a single fact. There is a chain of trust information that stretches back along the line from which the fact has come. The trustworthiness of every agent in the chain will affect the final calculation of how much credence to put in the fact.

4 Conclusion

The issue of trust in the digital age is still a new topic area – the idea of conducting business online with a total stranger was unthinkable ten years ago. Since then a number of simple solutions have been developed to give an indication of the reliability of other parties. While these systems have proved extremely effective in providing customers with the level of reliability that they require, as automation of services increase, a more sophisticated system that more closely emulated the complexities of trust in social interaction will need to be developed.

This paper has outlined some of the problems associated with engineering trust on a more advanced level – problems that must be solved before trust can be represented with sufficient detail to enable the use of automated agents.

The social issues of *how* people develop trust and *why* some people try to exploit systems must not be forgotten. Good solutions should be designed such that the benefit of co-operation far outweighs possible gains from ‘cheating’ the system or betrayal.

References

- [1] *Theory of Games and Economic Behavior*. Princeton University Press, 1944.
- [2] Lszl Mr. *Moral Calculations: Game Theory, Logic, and Human Frailty*. Springer, 1998.
- [3] Shiela Brownlow. Seeing is believing: Facial appearance, credibility, and attitude change. *Journal of Nonverbal Behavior*, 16:101–115, 06 1992.
- [4] Sheila Brownlow. Facial appearance, gender, and credibility in television commercials. *Journal of Nonverbal Behavior*, 14(1):51–60, March 1990.
- [5] R Dhamija, J Tygar, and M Hearst. Why phishing works. *Proceedings of the SIGCHI conference on Human Factors in Computing*, Jan 2006.
- [6] W Diffie and M Hellman. New directions in cryptography. *Information Theory*, Jan 1976.
- [7] R Rivest, A Shamir, and L Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications*, Jan 1978.
- [8] P Resnick, K Kuwabara, R Zeckhauser, and E Friedman. Reputation systems. *Communications of the ACM*, Jan 2000.
- [9] P Resnick and R Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay’s reputation system. *The Economics of the Internet and E-Commerce*, Jan 2002.
- [10] T Berners-Lee, J Hendler, and O Lassila. The semantic web. *Scientific American*, Jan 2001.
- [11] D Fensel. Spinning the semantic web: Bringing the world wide web to its full potential. *books.google.com*, Jan 2005.
- [12] Trust strategies for the semantic web. Sep 2004.
- [13] J Golbeck, B Parsia, and J Hendler. Trust networks on the semantic web. *Proceedings of Cooperative Intelligent Agents*, Jan 2003.