Workshop on New Forms of Collaborative Production and Innovation: Economic, Social, Legal and Technical Characteristics and Conditions. SOFI, University of Göttingen, 5-7 May 2010.

Trust Management in Online Communities

***Audun Jøsang**, University of Oslo*

Abstract

Our perception of trust works as a compass for safe navigation through a world of uncertainty. On the one hand it helps us find peers with whom interaction and collaboration is most likely to be fruitful. On the other hand it helps us steer away from unacceptable risks and potential deceptions. While human intuition about trust often fails, it seems to provide us with guidance very quickly in most situations, which has the effect of significantly saving cognitive effort. Online communities represent a new environment for human interaction, and we often find that our capability to reason about trust is not well adapted to online environments. The reason for this can be the limitation of current digital interfaces which thereby reduces the richness of information we receive about others, and also the fact that people actually encounter misrepresentation or deceptive online services and behaviour relatively often. Trust management in online communities aims at making trust reasoning more powerful and reliable by collecting, analysing and disseminating information that is relevant for trust and trust based decision making. This article describes semantic aspects of trust as well as principles and methods for building online trust and reputation systems. The problems and challenges for designing and implementing reliable trust and reputation systems are invoked and some potential solutions are mentioned. Finally, the article articulates our vision for trust management in online communities.

Introduction

Trust is a fundamental consideration for the growth and stability of markets and communities because trust guides decisions about interactions between humans and organizations. Large-scale online environments with participants from diverse geographical and cultural groups are primary arenas for human interaction, but the very nature of online environments makes trust management challenging. For example, it is common to request services from a website we have never heard of before, and from which we might never request a service again in the future.

The relative difficulty of assessing trust in online environments leads to security problems on many levels. On the technology level, the exploitation of global network mechanisms can enable attackers to disrupt services on a massive scale. On the psychological level, cleverly designed deceptions can dupe a significant percentage of online users into divulging sensitive information. On the commercial level, automated agents can exploit market platforms to commit fraud and gain unfair advantages. On the social and political levels, online media and communities can be manipulated to create unnatural opinion biases and to hijack democratic processes. There are currently very few practical methods for assessing the reliability or good faith of entities and the quality of resources in the online environment. It is challenging to enforce policies or to sanction non-compliance, and in many cases it is even difficult to know which policies apply in specific online environments. This uncertainty makes it difficult to know which resources can be relied upon and which entities it is safe to interact with, which thereby represents a serious obstacle to the creation and cultivation of quality online markets and communities. However, it is in this environment of risk and uncertainty that online communities and markets must grow.

Innovation in traditional security technologies is an important and a necessary factor for creating reliable online environments, but it is certainly not enough. The traditional definition of information security is the preservation of confidentiality, integrity and availability. Traditional information security assumes that the information resources have an owner who wants to protect their confidentiality, integrity and availability. The owner then defines policies and implements security controls to enforce them and to prevent misuse of the resources or keep misuse to a minimum. Unfortunately this model does not fit well with reality on the open Internet. We can be harmed simply by accessing low-quality, misrepresented or deceptive resources. Even if deceptive resources do not affect our information systems directly, they can have a negative effect on our knowledge and our business processes. This type of harm is not addressed by the traditional interpretation of information security. In fact, traditional information security mechanisms are not designed to protect against this type of harm because the classic security paradigm is reversed. Security is not only about controlling who can access information assets that we own or control. We also need methods to identify which agents and third party information assets and services can be safely accessed and which should be avoided. Trust management, sometimes called soft security, can provide

the type of security required for this purpose and is a crucial complement to traditional information security. Trust management makes our approach to solving security problems more general.

Trust management is the activity of assisting participants in online markets and communities to assess the quality, reliability and good faith of online services and of each other in order to make better decisions about which parties it is safe to transact with, and which services are correctly represented. Trust management also allows providers of quality services to market themselves as such; so that it serves parties on both sides of a trust relationship. The combination of providing an incentive for quality services and good-faith behaviour, and of providing a mechanism for sanctioning low-quality services and deceptive behaviour is the primary effect that trust management brings to online communities. The secondary effect is that this stimulates the emergence of quality markets and communities. The challenge is not only to design effective models, but also to design robust methods and mechanisms for trust management.

Trust concepts

Trust allows people to interact spontaneously and helps the economic system to run smoothly. Lack of trust, on the other hand, is like sand in our social and economic systems, it makes us spend inordinate amounts of time and resources on protecting ourselves against possible harm and thus slows transactions considerably. Fukuyama (Fukuyama 1995) describes the role mutual trust plays in the formation of social structures, and it is natural to assume that this also applies to the creation of quality online communities and markets. However, distrust can also serve as a useful state of mind, as it helps us to avoid harm when confronted with unreliable systems or dishonest people and organizations. The question of whom to trust online is, according to Craig Newmark, the biggest challenge for the Internet in the next decade (Ingram 2010). To face this challenge, he believes that the Web needs a "distributed trust network" that allows us to manage our online relationships and reputations.

Trust is a directional relationship between a relying party and a trusted party. One must assume the relying party to be a "thinking entity" in some form, meaning that it has the ability to make assessments and decisions based on received information and past experience. The trusted party can be anything from a person, organisation or physical entity to an abstract notion such as information or a cryptographic key. A trust relationship has a scope, meaning that it applies to a specific purpose or domain of action, such as "being authentic" in the case of an agent's trust in a cryptographic key, or "providing reliable information" as in the case of a person's trust in the correctness of an entry in Wikipedia. Mutual trust is when both parties trust each other within the same scope, but this is obviously only possible when both parties are thinking entities. Trust can be seen as a state of mind of the relying party, but can also have effects on the trusted party and other

elements in the environment; for example, by stimulating reciprocal trust. The term "trust" is used in the literature with a variety of meanings; we will focus on just two types of trust. On the one hand, we shall look at trust as a subjective evaluation of the reliability or quality of something or somebody (i.e. the trusted party), which we will call "evaluation trust."[1] On the other hand, we have the view of trust as a decision to enter into a situation of dependence on the trusted party, which we call "decision trust."

As the name suggests, evaluation trust can be interpreted as the evaluation of something or somebody independently of any actual commitment. Decision trust, on the other hand, indicates that the relying party has actually made a commitment to depend on the trusted party. To illustrate the difference between evaluation trust and decision trust with a practical example, consider a fire drill where participants are asked to escape from the third floor window of a house using a rope that looks old and appears to be in a state of deterioration. In this situation, the participants would assess the probability that the rope will hold their weight. A person who thinks that the rope could rupture would distrust the rope and refuse to use it. This is illustrated on the left-hand side of Fig. 1.

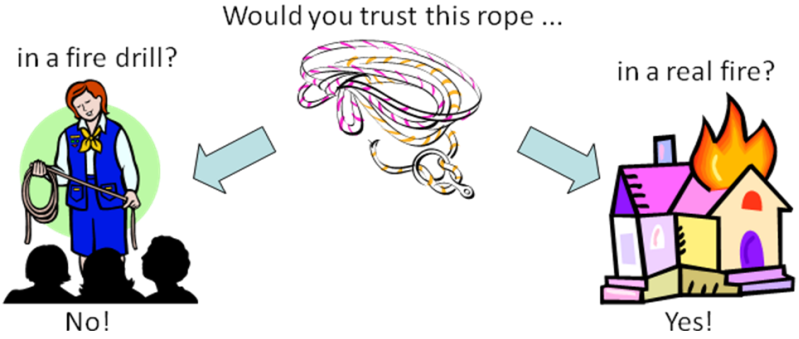

Figure 1. Same evaluation trust, but different decision trust

Imagine now that the same person is trapped in a real fire, and that the only escape is to descend from the third floor window using the same ragged-looking rope. In this situation, illustrated on the right-hand side of Fig. 1, it is likely that the person would trust the rope, even if he thinks it might break. This change in trust decision is perfectly rational because the likelihood of injury or death while descending is weighed against the hazards of smoke suffocation and death by fire. Although the evaluation trust in the rope is the same in both situations, the decision trust changes as a function of the different utility values associated with the different courses of action in the two situations.

[1] Also called "reliability trust."

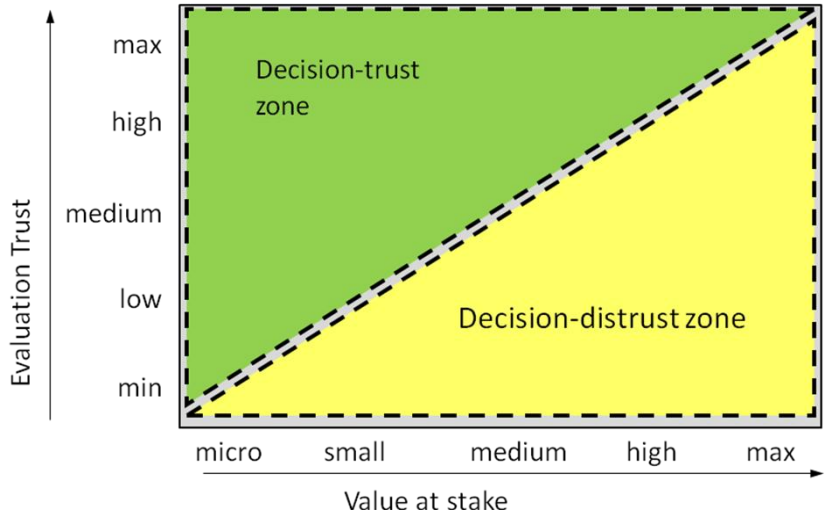

Figure 2. Relationship between evaluation trust and decision trust

This difference shows that decision trust depends on many factors, as illustrated in Fig. 2. If the value at stake is very high, then a relying party normally requires higher evaluation trust before making a trust decision. On the other hand, buying a 1 Euro lottery ticket puts little value at stake and does not require much evaluation trust. In addition to these factors, one must also consider subjective risk attitudes among many other factors. This simple analysis shows that decision trust can be a complex measure whereas evaluation trust is simply an evaluation of the trusted entity in isolation.

Trust and reputation systems, abbreviated as TRS hereafter, are mechanisms for the computation of trust/reputation measures. Different types of TRSs have different properties, so it is interesting to identify typical categories. One way to classify them is according to whether they utilize aspects of trust transitivity and whether the computed trust/reputation scores are private or public. This classification results in 4 different categories, as illustrated in Table 1.

1) Trust Systems	**2) Public Trust Systems**
Private Scores and Transitivity	Public Scores and Transitivity
Examples:	Examples:
Rummble.com, LinkedIn	PageRank, Slashdot moderation
3) Private Reputation Systems	**4) Reputation Systems**
Private Scores, No Transitivity	Public Scores, No Transitivity
Example:	Examples:
Customer feedback analysis	eBay Feedback Forum, epinions.com

Table 1. Trust and reputation system (TRS) categories

Category 1 contains pure trust systems with transitive trust and private scores, and category 4 contains pure reputation systems with public scores and without transitivity. It can be argued that pure reputation systems do use transitivity in the sense that the computed scores are derived from ratings in a transitive way. However, the transitivity goes no further than that, and these systems do not explicitly take the relying party's trust in the reputation system into account. There are systems that are neither pure trust systems nor pure reputation systems. For example, Category 2 systems, where scores are public and where transitivity is a significant factor, can be called public trust systems; one example is the Google PageRank algorithm and model. Another example is Category 3 systems, in which community participants provide ratings but the computed scores are private. This can be called a private reputation system; e.g., a customer feedback analysis performed by an organisation. The abbreviation TRS is used below to indicate any type of trust and reputation system.

Trust transitivity merits a closer look, as it relies on specific semantic constraints in order to be operable. This is illustrated in Fig. 3 below.

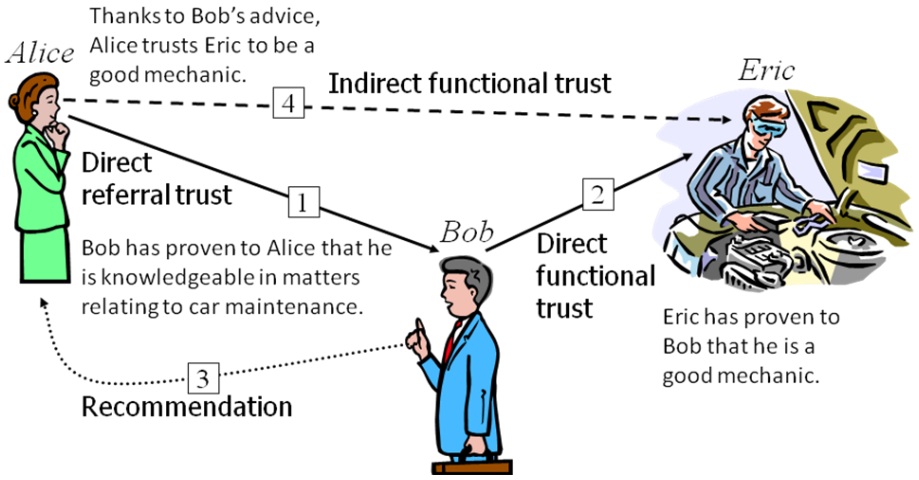

Figure 3. Trust transitivity

Assume that Alice needs to get her car fixed, but she has just arrived in town and does not know any mechanics. She asks her colleague Bob for his recommendation, because she has seen that Bob's car is always well maintained. She has direct trust in Bob on matters of car maintenance, but she would not trust Bob to actually service her car, so this is only referral trust. Assume further that Bob has had his car serviced by Eric many times and is very satisfied with Eric's work. As a result Bob's trust is both direct and functional, because Eric actually does the job. Assume now that Bob provides a recommendation to Alice about Eric. Alice can then derive functional trust in Eric because he is going to fix the car, but this trust is indirect because Alice has not had any direct experience with this mechanic. However, once Eric has serviced Alice's car she will have direct experience, so her functional trust in Eric will be based on both recommendation and direct experience. Studies show that direct experience carries more weight than indirect recommendations; as the amount of direct experience increases, the influence of indirect recommendation decreases.

One important observation from this example is that edges of referral trust make transitivity and recommendations operable, whereas the final edge of functional trust enables derivation of functional trust. A transitive trust path must thus consist of one or several consecutive referral trust edges followed by a final functional trust edge.

Another important observation is that all the trust edges have the same scope – in this example, that of car repair. Although Alice does not trust Bob to fix her car, she trusts him to recommend somebody who can do so. Trust transitivity thus requires that each edge have the same trust scope. If that were not the case, e.g. if Alice trusts Bob to look after her children, and Bob trusts Eric to fix cars, this

would not enable Alice to derive trust in Eric, neither for fixing cars nor for looking after her children.

The level of detail in the representation and analysis of trust described above is not normally considered in practical TRSs, i.e. such systems do not distinguish between functional and referral trust, nor between direct and indirect trust.

An aspect of trust not mentioned so far is the trust or reputation values which can be binary, discrete or continuous. Humans prefer discrete verbal categories such as *"low trust"*, *"medium trust"* or *"high trust"*, but such measures must often be mapped to numerical values to facilitate computational analysis. Expressing trust directly as numerical values can simplify the analysis, but the derived values must often be mapped to discrete categories to facilitate human cognition.

Trust and reputation models

There are a large number of proposed and implemented TRSs; we will only describe some general principles here. It is worth comparing the physical and the online world in terms of their potential for trust management. Table 2 illustrates some general aspects.

	Availability and richness of trust evidence	*Efficiency of communication and processing*
Brick & mortar world	Good	Poor
Online world	Poor	Good

Table 2. Potential for trust management in the physical world and online world

In general, the physical world provides rich and varied input evidence, but does not support highly efficient communication and analysis of this evidence. The online world, on the other hand, offers a rather limited variety of evidence, but consists of powerful networks and computers that enable extremely efficient communication and analysis of evidence.

In the case of reputation systems, for example, members of a community typically provide ratings to a reputation center as illustrated in Fig. 4 below.

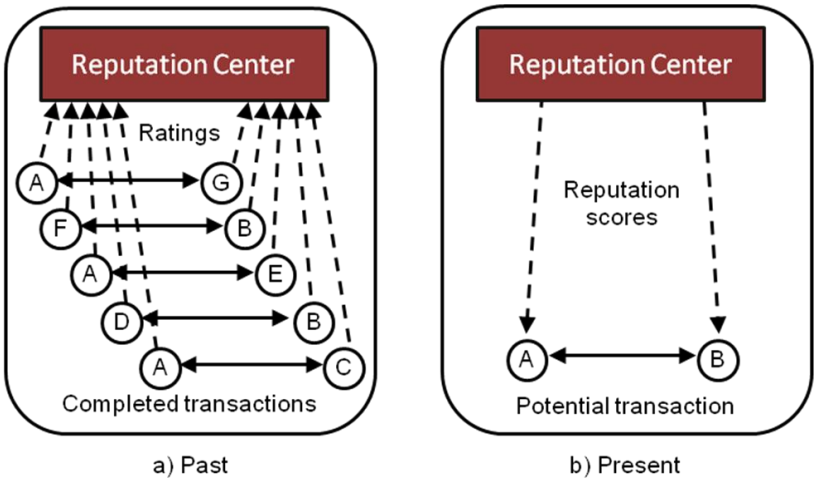

Figure 4. Centralised Reputation System

As depicted in Fig. 4.a, the reputation system receives ratings that reflect the direct experience of community members. From these ratings, the reputation center computes reputation scores that are published online. When participants contemplate transacting with one another, they can use the reputation scores as a basis for making their decisions, as illustrated in Fig. 4.b. A reputation score thus represents a degree of evaluation trust, whereas a decision to transact represents binary decision trust, where the former supports the latter.

A simple trust network is illustrated in Fig. 5 where levels of trust are expressed as a subjective opinion visually represented by a dot within an opinion triangle. The closer the dot is to the right hand side of the triangle, the greater the trust. Conversely, proximity to the left-hand side indicates distrust. The height of the dot within the triangle indicates the level of uncertainty in the trust value. Subjective logic defines operators and methods for modeling and analyzing this type of trust networks where trust edges are represented as subjective opinions.

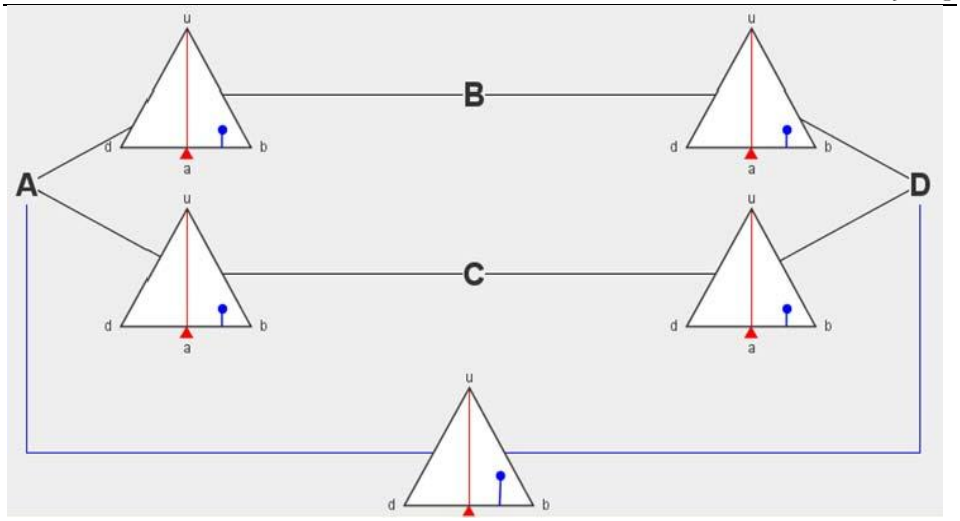

Figure 5. Simple trust network with trust values represented as subjective opinions

The two parallel trust paths $A \rightarrow B \rightarrow D$ and $A \rightarrow C \rightarrow D$ in the upper part of Fig. 2 represent input arguments for deriving agent A's trust in agent D, which is illustrated at the bottom of the diagram. More specifically, the input arguments are the opinions for the trust edges $[A,B]$, $[B,D]$, $[A,C]$ and $[C,D]$. It must be assumed that agent A has already formed opinions for the trust edges $[A,B]$ and $[A,C]$. Agent B must then inform agent A of its opinion concerning $[B,D]$, and agent C must inform agent A of its opinion concerning $[C,D]$. Agent A can then analyse the entire trust network, expressed as $[A,D]=([A,B]:[B,D])\Diamond([A,C]:[C,D])$.

The theoretical models illustrated in Figs. 4 and 5 give little detail about how a TRS should be implemented in practice. It seems that there are no general architectures that fit in all situations, so that each community or market requires a specially designed architecture in order for the TRS to function well. One of the more advanced architectures for a trust system is the moderation system used on Slashdot, the general architecture of which is illustrated in Fig. 6.

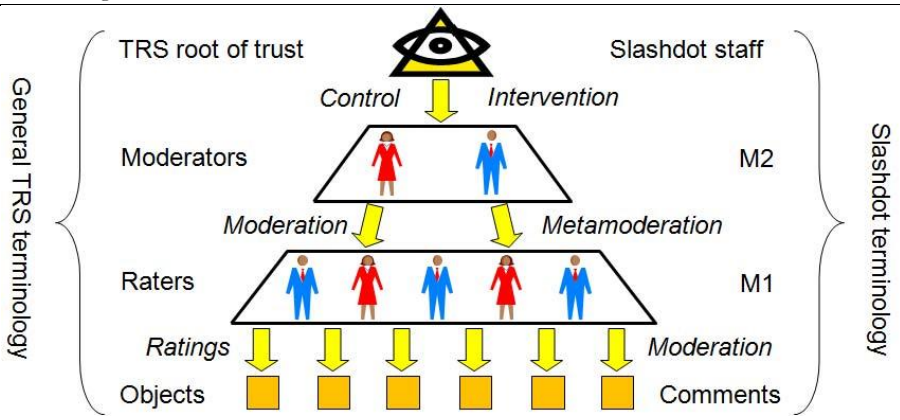

Figure 6. Slashdot moderation system architecture

Articles are posted on the Slashdot website by Slashdot staff. Once an article has been posted, anyone can comment on that article. The purpose of the Slashdot moderation system is to allow readers to filter the comments as a function of their quality. The moderation scheme actually consists of two moderation layers, where M1 is for moderating comments to articles, and M2 is for moderating M1 moderators. Users can rate comments; thus each comment gets a score. A user who only wants to read the best comments can set the threshold to only read high-scoring comments. To reduce the likelihood of unfair moderations, Slashdot implements the metamoderation layer M2 to moderate the M1 moderators. A user who wants to metamoderate will be asked to moderate the M1 ratings on 10 randomly selected comments. The metamoderator decides if a moderator's rating was fair, unfair, or neither. This moderation affects the Karma of the M1 moderators which in turn influences their eligibility for being M1 moderators in the future. The Slashdot TRS directs and stimulates the mass-collaborative effort of moderating thousands of postings every day. The system is constantly being tuned and modified and can be described as an ongoing experiment in the pursuit of the best practical way to promote quality postings, discourage noise and to make Slashdot as readable and useful as possible for a large community.

Challenges for trust and reputation systems

The primary purpose of TRSs is to provide decision support for users. The value of this decision support function depends on the reliability and accuracy of the trust and reputation scores produced. Unfortunately, it seems that TRSs have many types of vulnerabilities which make them relatively easy targets for attacks and manipulation. Vulnerabilities and attacks mentioned in the literature are e.g.:

- Ad hoc computation
- Playbooks
- Unfair ratings
- Discrimination
- Collusion
- Proliferation
- Reputation lag
- Re-entry/Change of identity
- Value imbalance
- The Sybil Attack
- No incentive to provide ratings
- Hard to elicit negative feedback
- Notorious attackers

Ad hoc computation means that the algorithm or model for deriving trust and reputation scores is unsound; i.e., that they simply produce erratic scores. A playbook consists of a sequence of actions that maximises profit or fitness of a participant according to certain criteria. There is an infinite set of possible playbook sequences, and the actual profit resulting from any particular sequence will be influenced by the actions (and playbooks) of other participants in the community. Unfair rating attacks consist in providing ratings that do not reflect the genuine opinion of the rater. Discrimination means that a service entity provides high-quality services to one group of relying parties, and low-quality services to another group of relying parties. Collusion means that a group of agents coordinate their behaviour, which e.g. can consist in running playbooks, of providing unfair recommendations, or practicing discrimination. Proliferation means that an agent offers the same service through many different channels, thereby increasing the probability of being chosen by a relying party. The reputation lag attack means that the attacker uses the time lag between an instance of a service provision and corresponding rating's effect on the service entity's score, e.g. to offer and provide a large number of low-quality services over a short period before the rating suffers any significant degradation. Re-entry means that an agent with a low score leaves a community and subsequently reenters the community under a different identity. The effect is that the agent can start from fresh, and thereby avoiding the consequences of the low score associated with the previous identity. The value imbalance attack is possible when the weight of a rating is not related to the value of the transaction. The effect of providing a large number of high-quality, low-value services and a small number of deceptive high-value services would then result in a high profit resulting from high value deception without any significant loss in scores. The Sybil attack is when a single entity establishes multiple pseudonym identities within a TRS domain to provide multiple ratings on the same service object. The name Sybil attack comes from a book of the same name by Flora Rheta Schreibe (1973) about a woman suffering from multiple personality disorder.

A TRS needs ratings to function properly. However, participants have little incentive to provide ratings after direct experiences because the ratings are only beneficial to others, not to themselves. It therefore seems that altruism plays a certain role in providing ratings, but reliance on altruism could potentially be con-

sidered as a weakness of TRSs. In some situations it can be especially challenging to obtain feedback about negative experiences because of people's reluctance to offend and because some might fear some form of retaliation from the rated party.

Finally, there will always be participants whose sole purpose it is to disrupt the order in a community or market, and for whom incentives for good behavior or the sanctioning of bad behavior will have no effect. While there is little a TRS can do to moderate the behavior of such participants, the community or the TRS itself should at least not risk breaking down when confronted with such participants.

Evaluation is an obvious approach to determine the strength and improve the robustness of TRSs, but TRS evaluation seems particularly challenging. A few approaches have been proposed:

- TRS evaluations can be conducted from a theoretical perspective, e.g. through simulation. However, this would only provide a partial exposure to potential threats.
- A comprehensive set of robustness evaluation methods and criteria can be defined. This would make it possible for TRS designers to produce comparable evaluations. However, the great variety in types of TRS makes it difficult to apply the same criteria to different TRSs.
- TRS robustness can be evaluated by implementing the TRS in a real environment where a certain proportion of participants have an interest in manipulating the TRS. However, establishing a real online community with a representative population of participants can be difficult.

When we see that TRSs often cannot be considered robust, it seems surprising that they still can provide significant value and that they have become so widespread. One might therefore say that TRSs follow the paradoxical *"Yhprum's Law,"* which is the inverse of Murphy's Law, expressed by: *"Something that shouldn't work sometimes does work."*

One possible explanation of why TRSs are useful despite their weaknesses is that in many situations, a TRS does not necessarily need to be robust. Resnick & Zeckhauser (Resnick and Zeckhauser 2002) consider two explanations in relation to eBay's reputation system: (a) Even though a reputation system is not robust it might serve its purpose of providing an incentive for good behaviour if the participants think it works, and (b) even though the system might not work well in the statistical normative sense, it may function successfully if it reacts swiftly to bad behavior and imposes costs for a participant to get established.

Finally, it could be argued that the TRS in an online community serves as a kind of social glue. A TRS provides an interface through which participants can communicate and relate to each other, which in itself is valuable. Any TRS with user participation will depend on how people can use it to better connect to other participants and to the community as a whole, and must be designed with that perspective in mind.

Conclusion and Vision for online trust management

We are witnessing the emergence of new forms of cultures in which human and automated agents interact, and where it is often impossible to distinguish between the two. Cultural and biological evolution has resulted in our current set of civilized communities, despite continuous failures and setbacks along the way. When considering our options for cultivating the best possible online communities that are beneficial for local or global communities, we must remember that we have the power to make certain design choices and to implement constraints – at both the technological and behavioural levels of human and automated agents and platforms – in the way they interact in online environments. Trust management is an important element of such a culture-by-design. Trust management can enable service consumers to reliably assess the quality of services and the reliability of entities before they decide to use a particular service, or to interact with or depend on a given entity. Trust management will also enable serious service providers and online players to correctly represent their own reliability and the quality of their services, so that in effect it becomes a marketing tool as well as a compass for safe navigation of online environments.

References

Fukuyama, F.(1995). "*Trust: The Social Virtues and the Creation of Prosperity*". New York: The Free Press.

Ingram, Matthew.(2010). *"Craig Newmark on the Web's Next Big Problem"*. Online Article featuring an interview with Craig Newmark, founder of www.craigslist.org. http://gigaom.com/2010/03/18/craig-newmark-on-the-webs-next-big-problem/, accessed on 31.03.2010.

Resnick, P. and Zeckhauser, R.. (2002). *"Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System"*. In: M.R. Baye (Ed.). *The Economics of the Internet and E-Commerce 11 of Advances in Applied Microeconomics*. Elsevier Science.