

Prospectives for Online Trust Management

Audun Jøsang

Abstract—Innovation in technologies for online service provision has been explosive during the relatively short history of the Internet, and this trend shows no sign of slowing down. As a result there is an enormous amount and variety of resources available online. Currently there is little control over the quality of online resources by users and peer service providers, with the result that many Web sites provide low quality and even deliberately malicious or misleading services with little fear of being sanctioned. Without any practical means of assessing the quality of online resources, users are left in the dark when making decisions about which resources they can safely access, and that best fit their needs. This situation can be described in terms of trust management, meaning that serious service providers should be able to correctly represent the quality of their services, and service consumers should be able to reliably assess the quality of online services before they decide to use and depend on a particular service. This paper discusses the perspectives for reliable trust management for provision and consumption of online resources. As it turns out, this is a hard problem that no single method alone can solve. Instead, we need to start experimenting with various approaches to identify methods or combinations of such methods that give participants the best trust management capability.

Index Terms—Reputation systems, trust, deception, Web services

I. INTRODUCTION

Physical encounter and traditional forms of communication allow people to assess a much wider range of cues related to trustworthiness than is currently possible through computer mediated communication such as the Web. The time and investment it takes to establish a traditional brick-and-mortar street presence provides some assurance that those who do it are serious players. This stands in sharp contrast to the relative simplicity and low cost of establishing a good looking Internet and Web presence which in reality gives little evidence about the solidity or seriousness of the organisation behind it. This situation can be described in terms of trust management, meaning that serious service providers should be able to correctly represent the quality of their services, and service consumers should be able to reliably assess the quality of online services before they decide to use and depend on a particular service.

The problem of determining whether something or somebody can be trusted was not thought of as a problem when the Internet and the Web were conceived, because the community consisted of a group of users motivated by the same goals, and with strong trust in each other. Each time new and groundbreaking Internet technologies are being developed and deployed, the early adopters typically have *good intentions* because they are motivated by the desire to make the new

technology successful. Deceptive and fraudulent behaviour only occur after important services start to be supported by the new technology. The legacy technical architecture and the governance structure of the Internet are clearly inspired by the assumption of well intentioned participants. However, people and organisations currently engaging in Internet activities are not uniformly well intentioned in the same sense, because they are increasingly motivated by financial profit and personal gain. The result is that we are poorly prepared for controlling and sanctioning users and service providers with this type of selfish, or even worse, malicious and criminal intentions.

Allowing malicious entities to participate in online communities where no control or sanctioning mechanisms are in place, is analogous to introducing a new aggressive animal species into an environment where it has no predators. In such cases the new species will multiply at the expense of the existing species and the environment, and a new equilibrium will be reached after some time. It is worrisome to think about what the new equilibrium will look like in the case of the Internet if misrepresentation and deceptive activities are allowed to continue relatively unsanctioned.

This article describes how online trust management can play an important role for fostering quality online communities and markets. This includes methods for assessing the quality of online resources, and methods for sanctioning low quality service providers. The purpose of such technologies is to provide decision support to consumers of online resources, and to discourage providers of online resources from providing low quality and misleading resources.

The article is organised as follows. Sec.II describes the soft security paradigm as a complement to traditional information security. Sec.III describes the concepts of trust and reputation in a way that is relevant for online activities and proposes a definition for trust management. Sec.IV describes important trust management challenges, and Sec.V presents general solutions in terms of reputation systems and trust engines. Sec.VI proposes a research agenda for online trust management.

II. HARD AND SOFT SECURITY

Information security technologies might seem as a possible way to control online activities. Many people and organisations assume that information security technologies, when properly designed, can provide protection against viruses, worms, Trojans, spam email and any other undesirable content the users can be exposed to through the Internet. However, it is questionable whether traditional information security technologies alone are sufficient for solving these problems. To better understand why, it is useful to look at the definitions of security and of information security separately.

Security can generally be defined as “*the quality or state of being secure - to be free from danger*” [1]. This definition

A. Jøsang is Associate Professor at the Faculty of Information Technology at Queensland University of Technology, Brisbane, Australia, Phone: (+61) 7-3864-2960, email: a.josang@qut.edu.au

is very broad and covers the protection of life and assets from intentional and unintentional human actions, as well as from natural threats such as storms and earthquakes. In case of protection of information assets, the term *information security* is being used. Information security is commonly defined as “the preservation of confidentiality, integrity and availability of information” [2]. It is here assumed that it is the owner of information assets who wants to keep those assets free from danger, and to preserve their confidentiality, integrity and availability. However, in many situations we have to protect ourselves from harmful information assets and from those who offer resources, so that the problem in fact is reversed. Information providers can for example act deceitfully by providing false or misleading information, and we have to admit that traditional information security mechanisms provide little protection against this type of threats. False, misleading, and malicious online resources can inflict harm on people and organisations without directly affecting their information assets, and this is a security problem not covered by the traditional definition of information security. In other words, online threats are not necessarily information security threats in the traditional sense.

We are thus in a situation where we are faced with serious threats, against which there are no established and effective protection. There is not even a well defined vocabulary for describing these threats and the security risks they are posing. In the case of traditional information security, vocabulary and technologies are well defined and standardised. ISO’s IS7498-2 Security Architecture [3] is an example of an early security standard that was instrumental in defining the information security vocabulary. In the present situation of confusion and lack of understanding of the current online threats, the spreading of malicious software and disinformation takes place almost unchallenged.

The difference between traditional information security and the extended view of online security described above, was first described by Rasmussen & Jansson (1996) [4] who used the term *hard security* for traditional mechanisms like authentication and access control, and *soft security* for what they called social control mechanisms. Designing and implementing such mechanisms represents a formidable challenge, in that the long term acceptance of the Internet as a platform for supporting open markets and communities depends on the success of this endeavour.

In case of hard security, the existence of a security policy is always assumed, whereby the owner of information resources authorises specific actions by certain parties. *White Hats* (i.e. the good guys) and *Black Hats* (i.e. the bad guys) are easily identified depending on whether they act according to, or against the security policy. In the case of soft security however, this distinction can become blurred, because there is generally no formally defined or generally accepted policy that defines what constitutes acceptable behaviour. Misrepresentation of the quality of an online service might not even be illegal in the jurisdiction of the service provider, yet a consumer who feels deceived by an online service would most likely define the service provider as a Black Hat. Social control methods that can provide protection against this type of online threats

can be called soft security mechanisms.

We propose to define soft security as follows:

Definition 1 (Soft Security): Soft security is the adherence to common ethical norms by interacting parties.

While the goal of traditional (hard) information security is to preserve the CIA properties (Confidentiality, Integrity and Availability) of assets within a specific domain, the goal of soft security mechanisms is to stimulate the quality of a specific community in terms of the ethical behaviour and integrity of its members.

It is interesting to try to establish a vocabulary around the notion of soft security similarly to the well established vocabulary around hard security. For example, what constitutes a soft security attack, or a soft security breach? Hard security mechanisms defend against attacks inside the victim domain, whereas soft security mechanisms defend against attacks outside the victim domain, before the attacks reach it. For example, virus scanners are hard security mechanisms that will try to detect viruses that have entered a domain through some communication channel. If the virus originates from a Website that offers free software, a reputation system can prevent people from downloading the software so that the virus never reaches the domain.

Soft security mechanisms are collaborative methods for assessing the behaviour of members in the community against the ethical norms, making it possible to identify and sanction members who breach the norms, and to recognise and reward members who adhere to the norms.

What constitutes ethical norms within a community will in general not be precisely defined. Misrepresentation of an item for sale on eBay would constitute unethical behaviour, but this still leaves open what misrepresentation means. For example representing a wrist-watch with a high quality photo and a description saying that it works perfectly might *a priori* be considered as correct representation. But failing to mention that the same wrist-watch has a name engraved on it will by most people be seen as misrepresentation. It would be quite impossible to define the requirements for correct representation in every case. As a consequence it is impossible to clearly define what is ethical and what is not in every situation. Still, most people will have a common understanding of this question, and it is this common understanding of ethical behaviour that forms the basis for soft security.

It can be noted that traditional law enforcement has limited capability for maintaining online social control, for example because of multiple jurisdictions, the difficulty of obtaining forensic evidence, and the prohibitive cost associated with investigation and prosecution.

III. DEFINING TRUST MANAGEMENT

A. What is Trust?

Trust is a directional relationship between two parties that can be called *trustor* and *trustee*. One must assume the trustor to be a “thinking entity” in some form, whereas the trustee can be anything from a person, organisation or physical entity, to abstract notions such as information or a cryptographic key [5].

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” in case of a person’s trust in an article from Wikipedia¹. Mutual trust is when both parties trust each other with the same scope, but this is obviously only possible when both parties are thinking entities. Trust influences the trustor’s attitudes and actions, but can also have effects on the trustee and other elements in the environment, for example, by stimulating reciprocal trust [6].

The literature uses the term trust with a variety of meanings [7]. A distinction between *situation independent trust*, (called “reliability trust”), and *situation dependent trust* (called “decision trust”), can often be recognised in the literature, although usually not explicitly expressed in those terms.

As the name suggests, reliability trust can be interpreted as the reliability of something or somebody independently of the situation, and the definition by Gambetta (1988) [8] provides an example of how this can be formulated:

Definition 2 (Reliability Trust): *Reliability trust is the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which its welfare depends.*

In Def.2, trust is primarily defined as the trustor’s estimate of the trustee’s reliability (in the sense of probability) in the context of *dependence* on the trustee. It can be noted that this definition does not take the situation into account.

However, trust can be more complex than Gambetta’s definition indicates. For example, Falcone & Castelfranchi (2001) [9] recognise that having high (reliability) trust in a person in general is not necessarily enough to decide to enter into a situation of dependence on that person. In [9] they write: “For example it is possible that the value of the damage *per se* (in case of failure) is too high to choose a given decision branch, and this independently either from the probability of the failure (even if it is very low) or from the possible payoff (even if it is very high). In other words, that danger might seem to the agent an intolerable risk.” For example, consider a person who during a fire drill is asked to abseil from the third floor window of a house using a rope that looks very old. In this situation, the participant would try to assess the probability that the rope will hold him while abseiling. If the participant thinks it is possible that the rope would rupture, it is likely that he would distrust the rope and refuse to use it. This is illustrated in Fig.1.

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 old rope. In this situation it is likely that the person would trust the rope, even if he thinks it is possible that it could rupture. This change in trust decision is perfectly rational because the likelihood of injury or death while abseiling is assessed to be lesser than in case of any other practical alternative. Although the *reliability 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.

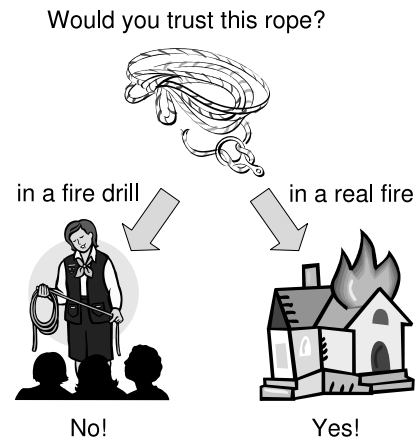


Fig. 1. Same reliability trust, but different decision trust

The following definition captures the concept of trust seen within a given situation.

Definition 3 (Decision Trust): *Decision trust is the extent to which a given party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible.*

In Def.3, trust is primarily defined as the willingness to rely on a given object, and specifically includes the notions of *dependence* on the trustee, and its *reliability*. In addition, Def.3 implicitly also covers situational elements such as *utility* (of possible outcomes), *environmental factors* (law enforcement, contracts, security mechanisms etc.) and *risk attitude* (risk taking, risk averse, etc.).

Both reliability trust and decision trust reflect a positive belief about something that the trustor depends on for his welfare. Reliability trust is most naturally measured as a discrete or continuous degree of reliability, whereas decision trust is most naturally measured in terms of a binary decision. Both types of trust are relevant for online trust management.

B. What is Reputation?

The concept of reputation is closely linked to that of trustworthiness, but it is evident that there is a clear and important difference. For the purpose of this study, we will define reputation according to the Concise Oxford dictionary.

Definition 4 (Reputation): *Reputation is what is generally said or believed about a person’s or thing’s character or standing.*

This definition corresponds well with the view of social network researchers [10], [11] that reputation is a quantity derived from the underlying social network which is globally visible to all members of the network. The difference between trust and reputation can be illustrated by the following perfectly normal and plausible statements:

- 1) “I trust you because of your good reputation.”
- 2) “I trust you despite your bad reputation.”

Assuming that the two sentences relate to identical situations, statement (1) reflects that the relying party is aware of the trustee’s reputation, and bases his trust on that. Statement (2) reflects that the relying party has some private knowledge

¹<http://www.wikipedia.org/>

about the trustee, e.g. through direct experience or intimate relationship, and that these factors overrule any reputation that a person might have. This observation reflects that trust ultimately is a personal and subjective phenomenon that is based on various factors or evidence, and that some of those carry more weight than others. Personal experience typically carries more weight than second hand recommendation or reputation, but in the absence of personal experience, trust often has to be based on recommendations from others.

Reputation can be considered as a collective measure of trustworthiness (in the sense of reliability) based on the recommendations or ratings from members in a community. An individual's subjective trust can be derived from a combination of received recommendations and personal experience.

Reputation can relate to a group or to an individual. A group's reputation can for example be modelled as the average of all its members' individual reputations, or as the average of how the group is perceived as a whole by external parties. Tadelis' (2001) [12] study shows that an individual belonging to a given group will inherit an *a priori* reputation based on that group's reputation. If the group is reputable all its individual members will *a priori* be perceived as reputable and vice versa.

C. What is Online Trust Management?

Because trust is an asymmetric relationship between a trustor and a trustee, trust management can be seen from two sides. The ability to gain the trust of others is an important criterion for the success and survival of an entity in a community, because it makes others willing to collaborate. Humans and animals therefore have a set of genetically determined and culturally acquired strategies for appearing trustworthy. The safest and most used strategy is probably to simply act in a responsible and trustworthy manner in any interaction with other members of the community. However, it is not uncommon that people attempt to give a false impression of trustworthiness for various reasons, such as for financial profit and personal gain. The ability to reliably assess the trustworthiness of an object is therefore an equally important criterion for the success and survival of an entity in a community. Humans and animals have a strong ability to assess trustworthiness of other entities, based on a multitude of cognitive and affective factors. Based on this duality of trust, we define trust management for online environments as follows:

Definition 5 (Online Trust Management): The activity of implementing and using systems and methods

- that allow a relying party to make assessments and decisions regarding the dependability of objects and agents outside its control,
- that allow agents to correctly represent to a relying party the dependability of themselves and of objects that they control.

There is currently a strong need for better methods that can enable relying parties to determine the trustworthiness of remote parties and services through Internet mediated communication and collaboration. At the same time, trustworthy entities need methods that enable them to be recognised as such

in a way that gives them compensation for their investment in quality of service. These methods should also, directly or indirectly, sanction untrustworthy players and thereby contribute to increasing the quality of online markets and communities. In a nutshell, developing and applying methods and technologies that aim at achieving this can be called trust management, and as we become increasingly dependent on Internet services, this activity becomes more and more important.

It is worth mentioning that term trust management is often being used within the scope of distributed access authorisation and control with a much narrower meaning, see e.g. Blaze *et al.* (1996) [13]. The main idea behind this approach is that a system or its owner does not need to know the identities of those who are accessing it, only that they are trusted to do so. This type of trust management is thus about verifying access credentials without necessarily authenticating entities. This is a much narrower definition of the term than that of Def.5, but it is of course covered by Def.5. Translated into the terms of Def.5, the system or system owner can be considered as the relying party that needs to assess the dependability of agents that request access to the system based on access credentials. A more descriptive name for that purpose could e.g. be *access credentials management*.

IV. TRUST MANAGEMENT CHALLENGES

A. Semantic Misrepresentation

The Semantic Web is the concept of encoding semantics in Web pages in a way that is readable and understandable by computers. The W3C² has established various standardisation activities³ in order to develop the technology for turning this concept into practical reality.

The main components of the semantic web are the standards and tools of XML (Extensible Markup Language) and of WOL (Web Ontology Language). XML provides a syntax for structured documents, but imposes no semantic constraints on the meaning of these documents. WOL provides a standardised vocabulary for describing properties, characteristics, classes and relationships of XML encoded data, i.e. a rudimentary (from a human perspective) form of semantics.

The idea is that WOL metadata can be added to a Web resource in order to describe in a machine readable form certain characteristics of the resource, and how it relates to other Web resources.

The semantic web is still mostly a theoretic concept without any large scale implementations, however the idea itself is fascinating. It would for example make it possible for automated software agents to search and combine Web resources in an independent manner on behalf of their human masters.

A potential problem that the W3C community is becoming increasingly aware of, is that having good technical standards is not enough to make the semantic Web practical. It is also important to consider the social context into which it will be implemented. For example, no provision has been made to verify and enforce that WOL metadata associated with a Web resource correctly represents that resource. In

²World Wide Web Consortium

³<http://www.w3.org/2001/sw/>

case the semantic Web is widely adopted, it is obvious that individuals and organisations will be tempted to misrepresent OWL metadata for their personal or financial advantage.

The semantic Web is thus another example of a new technology that is being developed assuming that all users will be well intentioned. Unfortunately, the reality will most likely be quite different should the semantic Web become a success, and unless the W3C is able to take this potential problem adequately into account, deliberate misrepresentation of OWL metadata will make the semantic Web much less usefully than anticipated.

B. The Wiki Problem

A Wiki is a web site where users can create and edit content collaboratively. Some wikis restrict edit access to authenticated and authorised users only, whereas other wikis allow unrestricted edit access. Intermediate implementations can for example be to moderate changes before they go online, or to freeze certain pages so that they can no longer be edited by the public.

The Wikipedia⁴ is currently the most prominent wiki on the Internet. It provides an online encyclopedia, where anybody who considers themselves an expert on a topic can add or edit an article about that topic. The Wikipedia has become one of Internet's great success stories, and is being used by many as an authoritative source of knowledge.

However, the inherent problem with all wikis is that open access allows ill intended parties to publish false and misleading information, or to use it as an advertisement channel to promote commercial products or political and ideological views. This is a serious problem because there is no simple way for wikis to stop and sanction this behaviour.

Because of Wikipedia's popularity, it provides a unique opportunity for malicious parties to spread false and misleading information in breach with Wikipedia's policy. The term *wiki vandalism* is used to describe this practice, but vandalism does not cover the case of conflicting opinions (e.g. political, religious or scientific) about a controversial topic. This can cause an article to be repeatedly overwritten by people with opposing views, and the particular time that the article is accessed will determine the version of the article that an unsuspecting user will see. To solve this problem, Wikipedia has introduced the concept of *stable version* which prevents further editing of an article by the public. The set of stable version articles is thus no longer a wiki. In order to separate between the two types of articles, those that still can be edited are called *working versions*. In order for a working version to become a stable version it must be checked by editors to meet a standard of quality. In case of a controversial topic, the editor will have the difficult task of extracting the essence from the various conflicting versions of an article, and present it in a balanced and neutral way. Controversy can emerge, and obvious errors can be discovered after an article has been made stable, so there must be a way of changing stable versions, which would require even more work. The question of who is accountable for an article must also be addressed.

This approach to collecting knowledge leads to an hybrid medium which is half wiki and half editorial, but where the content is collected through the wiki in some form or another. While the creation of content on Wikipedia is a collaborative unpaid effort, it could be difficult to find suitable editors, and costly to get them to do the work. Wikipedia's challenge is to give the articles credibility as authoritative sources of information, and the editorial process will be crucial to achieve this goal. This could for example be based on establishing a collaborative network of editors, accompanied with an efficient sanctioning system against editorial misbehaviour. The system must of course provide sufficiently strong incentives to make the right people willing to participate. These incentives could for example be financial as a proportion of Wikipedia's profit, or immaterial in the form of social status. Despite the extra cost needed to set up and run such a system in order to give the assurance of reliability and quality, this approach is significantly less expensive than the traditional way of creating an encyclopedia. The Wikipedia model is also far superior in providing up-to-date information, provided that a proper model for managing the quality of the collected information can be designed.

C. The Search Engine Problem

The early web search engines such as Altavista simply presented every web page that matched the key words entered by the user, which often resulted in too many and irrelevant pages being listed in the search results. Altavista's proposal for handling this problem was to offer advanced ways to combine keywords based on binary logic. This was too complex for users, and therefore did not provide a good solution.

PageRank proposed by Page *et al.* (1998) [14] represents a way of ranking the best search results based on a page's score according to a specific metric. Roughly speaking, PageRank computes the score for any Web page as the sum of the normalised weights of hyperlinks pointing to it, where a normalised hyperlink weight is determined by the score of the page containing the hyperlink, divided by the total number of hyperlinks from that page. This can be described as a reputation system, because the set of hyperlinks linking Web pages together can be seen as public information that can be used as a basis for deriving reputation scores. A single hyperlink to a given web page can be seen as a positive rating of that web page. Google's search engine⁵ is based on the PageRank algorithm, and the rapidly rising popularity of Google at the cost of Altavista was obviously caused by the superior search results that the PageRank algorithm delivered. The definition of the PageRank algorithm from Page *et al.* (1998) [14] is given below:

Definition 6 (PageRank): Let P be a set of hyperlinked web pages and let u and v denote web pages in P . Let $N^-(u)$ denote the set of web pages pointing to u and let $N^+(v)$ denote the set of web pages that v points to. Let s be some vector over P corresponding to a distribution of initial score

⁴<http://en.wikipedia.org/>

⁵<http://www.google.com/>

such that $\sum_{u \in P} s(u) = 1$. Then, the rank of a web page u is:

$$r(u) = d s(u) + (1 - d) \sum_{v \in N^-(u)} \frac{r(v)}{|N^+(v)|}, \quad (1)$$

In [14] it is recommended that d be chosen such that $d = 0.15$. The first term $d s(u)$ in Eq.(1) gives rank value based on initial score. The second term $(1 - d) \sum_{v \in N^-(u)} \frac{r(v)}{|N^+(v)|}$ gives rank value as a function of normalised weights of hyperlinks pointing at u . The algorithm of Def.6 must be iterated over the whole Web until the scores for all Web pages stabilise.

The PageRank algorithm provides an algorithmic representation of the “random surfer model”, i.e. the value $r(u)$ represents the probability of arriving at Web page u by randomly surfing the Web. Intuitively, because of the very large total number of hyperlinked Web pages in the Internet, this probability value is very close to zero for any random web page.

According to Def.6, $r(u) \in [0, 1]$, but the PageRank values that Google provides to the public are scaled to the range $[0, 10]$. We will denote the public PageRank of a page u as $R(u)$. This public PageRank measure can be viewed for any web page using Google’s toolbar which is a plug-in to the MS Internet Explorer. Although Google do not specify exactly how the public PageRank is computed, it is widely conjectured that it measured on a logarithmic scale with base close to 10. An approximate expression for computing the public PageRank could for example be:

$$R(u) = l + \log_{10} r(u) \quad (2)$$

where l is a constant that defines the cut-off value, so that only pages with $r(u) > 10^{-l}$ will be listed by Google. A typical value is $l = 11$.

It is not publicly known how the source rank vector s is defined, but it would be natural to distribute it over the root web pages of all domains weighted by the cost of buying each domain name. Assuming that the only way to improve a page’s PageRank is to buy domain names, Clausen (2004) [15] shows that there is a lower bound to the cost of obtaining an arbitrarily good $R(u)$ for a Web page u .

Without specifying many details, Google state that the PageRank algorithm they are using also takes other elements into account, with the purpose of making it difficult or expensive to deliberately influence PageRank.

In order to provide a semantic interpretation of a PageRank value, a hyperlink can be seen as a positive rating of the page it points to. Negative ratings do not exist in PageRank so that it is impossible to blacklist web pages with the PageRank algorithm of Eq.(1) alone. Before Google with its PageRank algorithm entered the search engine arena, some webmasters would promote web sites by filling web pages with large amounts of commonly used search key words as invisible text or metadata in order for the page to have a high probability of being listed by a search engine no matter what the user searched for. The PageRank algorithm compensates for problem because a high R is also needed in addition to matching key words in order for a page to be presented to the user.

The growing importance of having a high score in search

engines has made many owners of Web sites very restrictive with placing hyperlinks to other websites, because outgoing hyperlinks normally result in decreased scores for Web pages on the own Web site. The very existence of search engines thus had the inevitable effect of interfering with the structure of the Web.

The increasing popularity and economic importance of search engines has also lead to more damaging methods for artificially boosting the score of Web pages. One such example is the phenomenon called *link spam* which consists of placing many hyperlinks to the same Web page on open Web fora such as online discussion boards, guest books, weblogs and wikis. The motivation behind this attack is that search engines will give an increased score for the Web page that these hyperlinks point to.

In order to counter the link spam attacks Google announced in early 2005 that hyperlinks marked with `rel="nofollow"` would not influence the hyperlink target’s score in the search engine’s index. This is implemented as follows:

```
<a href="http://some-spammer-website.com"
rel="nofollow" >Link</a>
```

Most open Web fora now mark user-submitted hyperlinks this way by default, with no option to disable it by the users, and most search engines take it into account when computing scores. This is an example where a simple technical solution was able to solve a growing problem. However, it might lead to other problems.

The increasing usage of `rel="nofollow"` in Web pages will have the effect that scores computed by Google and search engines no longer reflect the real structure of the Web, and removes the model more and more from the random surfer model. The random surfer follows any link, whereas search engines only follow those that are not marked by `rel="nofollow"`. A likely development is that most outgoing hyperlinks will be marked in this way in a selfish manner in order not to suffer decreased scores. The search engines will then face the problem of scarcity of cross links between Web sites, making the computed scores increasingly unreliable.

As a substitute for the hyperlinks, search engines need to use other types of evidence. An obvious source of information is the links that users actually select after a specific search. algorithms can for example be designed that increase the rank of a specific Web page when many people select the link to that page after a Web search. However, the value of this information is limited, because it only becomes available during searches, and does not reflect which Web pages people go to when not using search engines.

It would be more valuable for search engines to know the link to every page that people visit. By encouraging people to use toolbars, search engines can get precisely that information. A toolbar provides some value-added functionality to users, such as displaying the PageRank of every page the user visits. In return for this functionality, the engine is informed about every single Web page that the user visits. This architecture is

illustrated in Fig.2 below.

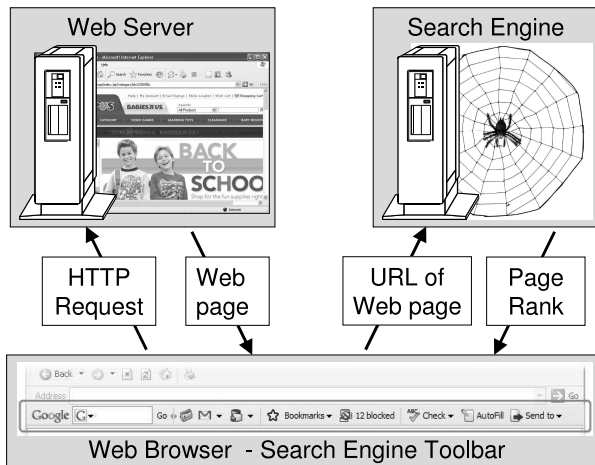


Fig. 2. Network architecture for search engine toolbars

Users often ignored that the toolbar provides this information. Constantly providing the search engine with information about Web pages visited by the user can be considered quite intrusive, and this functionality is usually also found in so-called spyware.

On the basis of information provided by toolbars, search engines are able to compute the probability that an intentional surfer will go to any particular Web page. This can be called the *intentional surfer model*, which represents an improvement over the random surfer model of the original PageRank algorithm.

However, it is likely that the current model is already under attack with the purpose of artificially increasing the ranking of certain Web pages. An obvious attack method is e.g. to install search engine toolbars on a large number of computers, and let programs automatically browse specific Websites. Google and other toolbar providers are aware of this potential problem, and usually registers each individual toolbar installation in order to identify possible “click spamming”. It is still unclear to what degree click spamming already is or will be a problem in the near future. The “pay per click” business model is being abused through click spamming, and it is therefore to be expected that the intentional surfer model that bases rank on the number of clicks to Web pages already is under attack as well.

In general it can be observed that any new method for improving rank computation becomes the subject of new attacks as soon as it is implemented. The robustness and reliability of searching and Web navigation has become a cat-and-mouse game, similarly to that of traditional information security.

D. Assessing Online Quality of Service

The basic problem with regard to online trust management is that it is difficult to obtain reliable information about the quality of online resources.

In theory it would require somebody to do online QoS assessments, and provide the conclusions available on the

Internet. In practice, that can happen in many different ways. Service assessments can be done as a global collaborative effort, or by specialised organisations. Rating quality of service can be a labour intensive and costly activity, and also creates new trust management problems. Can we trust the assessors, and who will assess the assessors? Many different approaches can be used for assessing the reliability of online services and entities, for example to pay people for doing the job as a full time activity, to base it on collaborative voluntary effort, or a combination of both. In any case, the activity must be based on a business model in a general sense, meaning that some form of compensation or incentive must be provided in order for people and organisations to be willing to contribute to the trust management effort within an online community.

Selecting service providers based on trust and perceived QoS is closely related to that of searching for services and service providers. The idea of assessing QoS could therefore be linked with the search activity, and could be integrated with search engines.

V. TRUST MANAGEMENT SOLUTIONS

A. Reputation Systems

Reputation systems represent an important type of technology for encouraging good online behaviour and for sanctioning poor performers. The basic idea is to let parties rate each other, for example after the completion of a transaction, and use the aggregated ratings about a given party to derive a trust or reputation score, which can assist other parties in deciding whether or not to transact with that party in the future.

A natural side effect of implementing reputation systems in applications is that it also provides an incentive for good behaviour, and therefore tends to have a positive effect on market quality. Reputation systems can be called collaborative sanctioning systems to reflect their collaborative nature, and are related to collaborative filtering systems. Reputation systems are already being used in successful commercial online applications. There is also a rapidly growing literature around trust and reputation systems, but unfortunately this activity is not very coherent.

For a centralised reputation system, the communication aspects are relatively straightforward in the sense that a central authority collects ratings and computes scores. This is illustrated in Fig.3.

In a centralised reputation system, ratings from each transaction are sent to the server which computes and publishes a reputation score for every entity. When two entities consider transacting, they can fetch each others reputation scores in order to have a better basis for making the decision.

In certain applications, such as P2P systems, a distributed model is needed, as illustrated in Fig.4.

Communication protocols for distributed reputation systems can be complex. In Fig.4, the problem faced by the two potential transaction partners *A* and *B* (on the right hand side of the figure), is to discover who the other party has interacted with in the past, and to obtain ratings from them.

The two fundamental aspects of reputation systems are:

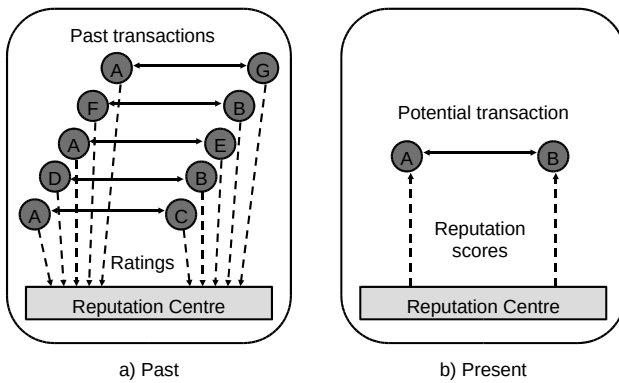


Fig. 3. General architecture for a centralised reputation system

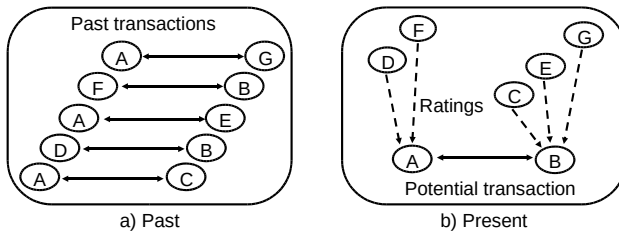


Fig. 4. General architecture for a distributed reputation system

- 1) *Communication protocols* that allow reputation centres or relying parties to obtain ratings from participants about each other, as well as allowing relying parties to obtain reputation scores of potential transaction partners.
- 2) A *reputation computation engine* used by a reputation centre or by relying parties to derive reputation scores for each participant, based on received ratings, and possibly also on other information.

Communities are important to the issue of reputation because the reputation of any given entity may vary widely from community to community, and even within a community. E.g. What the community of Microsoft users may think of Microsoft is likely to be very different to what the community of Linux users think of Microsoft. When users want to find the reputation score of an entity, community membership is a very important but largely implicit context. Therefore, a good reputation system should present the user with the reputation as calculated for the community that the system calculates is the 'best fit' for the user's search criteria and user profile. It should also allow the user to explicitly change their community context to view the reputation of the entity from the viewpoint of other communities of which the user is a member.

As a method for using QoS assessments during the Web search and service selection process, reputation systems can be integrated into search engines. A possible implementation of this could be by having a browser toolbar that allows users to rate Web pages and Web sites. This functionality could be implemented as part of the the search engines toolbars according to Fig.2. The reputation aspect of such a toolbar is illustrated in Fig.5

While the browser is fetching a Web page, the reputation toolbar will query the reputation server about the reputation score of that Web page or Web site. This is provided as a

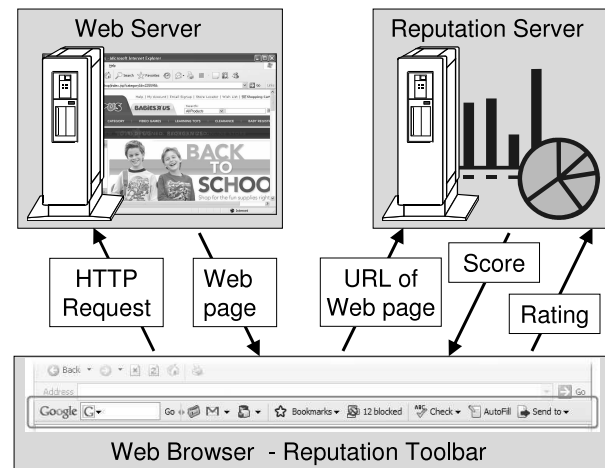


Fig. 5. Network architecture for reputation toolbars

reputation score. The user is also invited to rate the same Web site through the toolbar. This rating is sent to the reputation server, and taken into account when computing the reputation score in the future.

The functionality of the search engine toolbar of Fig.2 and the reputation toolbar of Fig.5 can very well be integrated into one single toolbar. The reputation scores can be taken into account for computing ranking when searching Web resources, or can be presented as a separate score for each search query result. In the latter case, the reputation server and the search engine do not need to be co-located. The reputation score can simply be fetched as part of a search query, either by the search engine itself, or by a shell on the client machine.

The addition of a reputation system to the traditional search engine will result in the *critical surfer model*, which represents an improvement over the random surfer model and the intentional surfer model. The critical surfer model provides search results according to the critical judgements of Web surfers. This is not possible with the existing models.

B. Trust Engines

There is a wide variety of reputation engines ranging from practical systems implemented in commercial applications, to theoretical proposals from the scientific literature. An overview of reputation systems and engines with their strengths and weaknesses can be found in [16].

A part from varying robustness against strategic attacks, there are other fundamental properties of reputation systems that need to be considered. Reputation engines implement a specific representation of ratings and reputation scores, and each representation has specific limitations regarding expressiveness, and the ease of rating collection and reputation score computation. In general, there is a trade-off between those two aspects. In many applications and contexts, information elements that are available as part of the communication protocols and infrastructure can also be used as a basis for computing reputation scores, so that no additional information needs to be generated and collected. However, these information elements usually have limited expressiveness, and therefore are not optimal for computing reputation scores.

For example, it would be interesting to rank web pages as a function of the number of people and Web services user agents (not Web crawlers) who access each page. However, there is no reliable information available for this purpose, but the existing hyperlinks linking Web pages together can be used as an approximation, and this is the principle on which Google's PageRank algorithm and other search engine algorithms are based. One obvious limitation with this approach is that it is impossible to express a negative rating through a hyperlink, for example associated with the text: "*This web page is very bad, don't go there*", because by encoding a hyperlink in this way will be seen as a positive rating by the search engines and contribute to increasing that page's rank. This is similar to the inability of scientometric references to distinguish between positive and negative references. Citing a scientific article for providing wrong or useless information should intuitively contribute to decreasing the ranking of the cited article and the journal where it appears, but in reality it will contribute to increasing the ranking of the article and the journal. By providing a citation regarding the usage of the term trust management in the sense of distributed access authorisation [13] from Sec.III-C, we unintentionally, but inevitably, contribute to increasing the scientometric ranking of that paper. By this we do not argue for the establishment of negative citations, because that could introduce more problems than it solves, e.g. in the form of citation wars. We simply want to illustrate that when metrics have limited expressiveness, their interpretation can be problematic.

Ratings and reputation scores can apply to different aspects of the service or goods offered. For example, an online newspaper could be rated for accuracy of information, readability of articles, and level of advertisement. Ratings from users are typically given as discrete fuzzy values such as Poor, Average, Good. Instead of fuzzy ratings, numerical values can also be used, such 1..5 stars, values in the range 1..10, or any range of numbers (discrete or continuous). The reputation score of a service can be visualised in various ways to provide optimal usability. New ratings should carry more weight than old ratings. Businesses do not remain static, they are bought and sold, change operations, change staff, move to new premises, and generally change in response to their environment. As a result, older ratings should be considered less reliable than newer ratings and the overall reputation should change accordingly. A forgetting factor defines how slowly old ratings are forgotten as a function of time and/or frequency.

C. Stimulating Social Control

Trust management relates directly to people's behaviour and perceptions, and these aspects can not always be automatically captured by reputation systems. Trust management therefore requires mental effort and manual input from people, which inevitably will carry a cost in terms of time spent. The question boils down to finding incentives for people to provide the necessary input, and support mechanisms for directing this input in a controlled fashion. In other words, there needs to be a business model in order for reputation systems to work in a practical setting. There are cases where ratings can be

collected automatically, such as when computing Web page scores based on hyperlinks, but in general manual input is needed.

Specific design choices need to be made to facilitate manual input. The ability to provide ratings can for example be restricted to designated experts, be restricted to certain user groups, or be open to anybody. Incentives to rate can be based on monetary payment, on other financial advantages, or simply on gaining prestige in the community.

A fundamental problem of reputation systems is that the relying party can not verify the sincerity of the ratings when they are provided on a subjective basis. Various technical solutions for detecting and excluding unfair ratings have been described in the literature [16], but none seem to provide a reliable solution.

Since unfair ratings is a sociological problem, it seems natural that a more sociological approach is needed to find a good solution. Using a hierarchic approach to sanctioning in trust management systems has been successful in online communities such as Slashdot.

1) *The Slashdot Model*: An approach that seems to work relatively well is that of meta-moderation used on Slashdot⁶ which is a "*news for nerds*" message board started in 1997. More precisely it is a forum for posting articles and comments to articles. In the early days when the community was small, the signal to noise ratio was very high. As is the case with all mailing lists and discussion fora where the number of members grow rapidly, spam and low quality postings emerged to become a major problem, and this forced Slashdot to introduce moderation. To start with there was a team of 25 moderators which after a while grew to 400 moderators to keep pace with the growing number of users and the amount of spam that followed. In order to create a more democratic and healthy moderation scheme, automated moderator selection was introduced.

The moderation scheme actually consists of two moderation layers where M1 is for moderating comments to articles, and M2 is for moderating M1 moderators. The purpose of M1 is to be able to filter the good comments from the bad. The purpose of M2 is to address the issue of unfair moderations, or more precisely to sanction M1 moderators. Above M2 in the hierarchy are the staff of Slashdot with omnipotent powers to sanction any M1 or M2 moderator who is detected in abusing the system. Details of the Slashdot reputation systems are described in [16].

The Slashdot reputation system directs and stimulates the massive 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 search for the best practical way to promote quality postings, discourage noise and to make Slashdot as readable and useful as possible for a large community.

2) *Hierarchic Trust Management*: In a hierarchical reputation system, ratings occur at different levels, and scores can be computed for elements on each level. Here we describe a general approach to designing hierarchical reputation systems.

⁶<http://slashdot.org/>

Service objects can have a reputation score based on ratings from service users. Users who provide ratings have a credit score based on ratings from moderators. Moderators have a credit score based on ratings from Controllers who represent the top of the hierarchy. Users rate service objects positively or negatively based on direct experience with those services. A reputation score can be computed for each object as a function of those ratings. Moderators can rate users depending on whether they provide fair or unfair service ratings. A credit score can be computed for each user based on the user's fairness in rating services. The idea is that service ratings provided by discredited users will carry relatively less weight than service ratings provided by credited users, when the reputation scores for service objects are derived. Controllers, who for example can be representatives from the reputation centre, can rate moderators, and depending on the design, can also rate users. This model is illustrated in Fig.6.

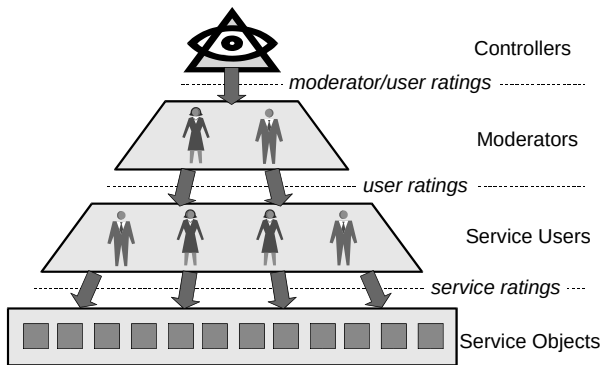


Fig. 6. Hierarchic model for reputation systems

The idea being this design is to spread the workload of providing ratings over all the service users, and provide a mechanisms for stabilising the system and sanctioning unfair raters. Design questions are for example to determine the optimal Moderator/User and Controller/Moderator ratios, and to determine adequate incentives for each party to do the job. From a purely rational viewpoint, a user has little incentive to rate a service after the fact, because providing ratings benefits others, not oneself. A study from eBay [17] shows that 60.7% of the buyers and 51.7% of the sellers on eBay provided ratings about each other. Possible explanations for these relatively high values can for example be that providing reciprocal ratings simply is an expression of politeness. However lack of incentives for providing ratings is a general problem that needs special attention when designing reputation systems.

VI. A RESEARCH AGENDA FOR ONLINE TRUST MANAGEMENT

There are two fundamental differences between traditional and online environments regarding how trust and reputation are, and can be, used.

Firstly, the traditional cues of trust and reputation that we are used to observe and depend on in the physical world are missing in online environments, so that electronic substitutes are needed. Secondly, communicating and sharing information related to trust and reputation is relatively difficult, and

normally constrained to local communities in the physical world, whereas IT systems combined with the Internet can be leveraged to design extremely efficient systems for exchanging and collecting such information on a global scale.

Motivated by this basic observation, the purposes of research in trust and reputation systems should be to:

- i. Find adequate online substitutes for the traditional cues to trust and reputation that we are used to in the physical world, and identify new information elements (specific to a particular online application) which are suitable for deriving measures of trust and reputation.
- ii. Take advantage of IT and the Internet to create efficient systems for collecting that information, and for deriving measures of trust and reputation, in order to support decision making and to improve the quality of online markets.

These simple principles invite rigorous research in order to answer some fundamental questions: What information elements are most suitable for deriving measures of trust and reputation in a given application? How can these information elements be captured and collected? What are the best principles for designing such systems from a theoretic and from a usability point of view? Can they be made resistant to attacks of manipulation by strategic agents? How should users include the information provided by such systems into their decision process? What role can these systems play in the business model of commercial companies? Do these systems truly improve the quality of online trade and interactions? These are important questions that need good answers in order to determine the potential for trust and reputation systems in online environments.

In particular, the robustness of trust management mechanisms is one of the critical factor for the success of this technology, and which currently is not being sufficiently addressed. Traditional security mechanisms can be used to achieve goals such as anonymity and integrity of ratings [18]. Identity and credentials management can be used to control when and by whom ratings can be provided, e.g. to prevent ballot stuffing [19]. The robustness of soft security mechanisms will thus depend on hard security mechanisms.

Social acceptance of trust and reputation systems is another critical factor, which many commercial systems have addressed quite well. However, for the more widespread and general usage of these systems, social acceptance by all parties is an issue that needs to be considered.

VII. CONCLUDING REMARKS

When accessing online services and resources, service consumers often have insufficient information about the service provider, and about the goods and services offered. Even if consumers are determined to act prudently when accessing online resources, in practice, they have little opportunity to verify the quality of those resources, or said metaphorically, to “squeeze the oranges before they buy”.

There is a need to find adequate online substitutes for the traditional cues to trust and reputation that we are used to in the physical world, and to take advantage of IT and the Internet

to create efficient systems for analysing this information in order to support decision making and to improve the quality of online markets.

When new Internet applications are being developed, people often make the simplifying, but too often erroneous assumption that all participants will behave with good intentions. The reason why this gets ignored is that Inventors and developers of new Internet applications only focus on successful uptake of the new application, and fail to see the consequences that a success might have. The problem of misbehaviour only start to occur when a new application has become successful and important business models depend on it.

Designing and implementing mechanisms that can provide adequate social control of online markets and communities represents a formidable trust management challenge. The long term acceptance of the Internet as a platform for supporting open markets and communities depends on the success of this endeavour.

REFERENCES

- [1] Merriam-Webster, *Merriam-Webster Online*. accessed January 2006: Available from <http://www.m-w.com/>.
- [2] ISO, *ISO/IEC IS17799 - Information technology – Code of practice for information security management*. ISO/IEC, 2005.
- [3] —, *IS 7498-2. Basic Reference Model For Open Systems Interconnection - Part 2: Security Architecture*. International Organisation for Standardization, 1989.
- [4] L. Rasmussen and S. Janssen, "Simulated Social Control for Secure Internet Commerce," in *Proceedings of the 1996 New Security Paradigms Workshop*, C. Meadows, Ed. ACM, 1996.
- [5] A. Jøsang, "The right type of trust for distributed systems," in *Proc. of the 1996 New Security Paradigms Workshop*, C. Meadows, Ed. ACM, 1996.
- [6] R. Falcone and C. Castelfranchi, "How trust enhances and spread trust," in *Proceedings of the 4th Int. Workshop on Deception Fraud and Trust in Agent Societies, in the 5th International Conference on Autonomous Agents (AGENTS'01)*, May 2001.
- [7] D. McKnight and N. Chervany, "The Meanings of Trust," <http://www.misrc.umn.edu/wpaper/wp96-04.htm>, University of Minnesota, Management Information Systems Research Center, Tech. Rep. MISRC Working Paper Series 96-04, 1996.
- [8] D. Gambetta, "Can We Trust Trust?" in *Trust: Making and Breaking Cooperative Relations*, D. Gambetta, Ed. Basil Blackwell. Oxford, 1990, pp. 213–238.
- [9] R. Falcone and C. Castelfranchi, *Social Trust: A Cognitive Approach*. Kluwer, 2001, pp. 55–99.
- [10] L. Freeman, "Centrality on Social Networks," *Social Networks*, vol. 1, pp. 215–239, 1979.
- [11] P. Marsden and N. Lin, Eds., *Social Structure and Network Analysis*. Beverly Hills: Sage Publications, 1982.
- [12] S. Tadelis, "Firm Reputation with Hidden Information," *Economic Theory*, vol. 21, no. 2, pp. 635–651, 2003.
- [13] M. Blaze, J. Feigenbaum, and J. Lacy, "Decentralized trust management," in *Proceedings of the 1996 IEEE Conference on Security and Privacy*, Oakland, CA, 1996.
- [14] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford Digital Library Technologies Project, Tech. Rep., 1998.
- [15] A. Clausen, "The Cost of Attack of PageRank," in *Proceedings of The International Conference on Agents, Web Technologies and Internet Commerce (IAWTIC'2004)*, Gold Coast, July 2004.
- [16] A. Jøsang, R. Ismail, and C. Boyd, "A Survey of Trust and Reputation Systems for Online Service Provision (in press, accepted June 2005)," *Decision Support Systems*, vol. 00, no. 00, pp. 00–00, 2006.
- [17] P. Resnick and R. Zeckhauser, "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System," in *The Economics of the Internet and E-Commerce*, ser. Advances in Applied Microeconomics., M. Baye, Ed. Elsevier Science, 2002, vol. 11.
- [18] R. Ismail, C. Boyd, A. Jøsang, and S. Russel, "Strong Privacy in Reputation Systems," in *Proceedings of the 4th International Workshop on Information Security Applications (WISA)*, Jeju Island, Korea, August 2003.
- [19] —, "An Efficient Off-Line Reputation Scheme Using Articulated Certificates," in *Proceedings of the Second International Workshop on Security in Information Systems (WOSIS-2004)*, 2004.