Towards a Theory of Trust in Networks of Humans and Computers

Virgil Gligor and Jeannette M. Wing

September 8, 2011

CMU-CyLab-11-016

CyLab
Carnegie Mellon University
Pittsburgh, PA 15213

Presented at the 19th International Workshop on Security Protocols, Cambridge, UK, March 28-30, 2011 (to appear in LNCS, Springer Verlag)

Towards a Theory of Trust in Networks of Humans and Computers

Virgil Gligor and Jeannette M. Wing

Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
gligor@cmu.edu,wing@cs.cmu.edu

Abstract. We argue that a general theory of trust in networks of humans and computers must be build on both a theory of *behavioral trust* and a theory of *computational trust*. This argument is motivated by increased participation of people in social networking, crowdsourcing, human computation, and socio-economic protocols, e.g., protocols modeled by trust and gift-exchange games [3, 10, 11], norms-establishing contracts [1], and scams [6, 35, 33]. User participation in these protocols relies primarily on trust, since on-line verification of protocol compliance is often impractical; e.g., verification can lead to undecidable problems, co-NP complete test procedures, and user inconvenience. Trust is captured by participant preferences (i.e., risk and betrayal aversion) and beliefs in the trustworthiness of other protocol participants [11, 10]. Both preferences and beliefs can be enhanced whenever protocol non-compliance leads to punishment of untrustworthy participants [11, 23]; i.e., it seems natural that betrayal aversion can be decreased and belief in trustworthiness increased by properly defined punishment [1]. We argue that a general theory of trust should focus on the establishment of new trust relations where none were possible before. This focus would help create new economic opportunities by increasing the pool of usable services, removing cooperation barriers among users, and at the very least, taking advantage of "network effects." Hence a new theory of trust would also help focus security research in areas that promote trust-enhancement infrastructures in human and computer networks. Finally, we argue that a general theory of trust should mirror, to the largest possible extent, human expectations and mental models of trust without relying on false methaphors and analogies with the physical world.

1 Introduction

Consider this fundamental question: How can I, a human, trust the information I receive through the Internet? This question's relevance has grown with the advent of socially intelligent computing, which includes social networking, crowd sourcing, and human computation. Socially intelligent computing recognizes the increasing opportunities for humans to work with each other relying on input from both humans and computers in order to solve problems and make decisions. When we read a Wikipedia page, how can we trust its contents? We need to

trust the person who wrote the page, the computer that hosts the page, the channel over which the message that contains the page contents are sent to the reader, and finally the computer that receives the message that contains the page contents.

We seek a general theory of trust for networks of humans and computers. Our main idea is that for a general theory we need to build on both a theory of *behavioral trust* to complement and reinforce a theory of *computational trust*. Behavioral trust defines trust relations among people and organizations; computational trust, among devices, computers, and networks. Towards building a general theory of trust through combining ideas from behavioral trust and computational trust, we moreover argue that there is *new economic value* to be gained, raising new opportunities for technological innovation.

The True State of Affairs. Toward a general theory of trust, let's review from computer science and the social and economic sciences, the state of the art, since it is not as rosy as we would like. Over the past three decades, research on trust in computer networks focused on specific properties, e.g., authentication and access-control trust, in traditional distributed systems and networks [4, 22, 13]), mobile ad-hoc networks [12, 31, 25, 36, 5, 32, 38, 27, 26, 18], and applications [19, 21]. Lack of a formal theory of trust has had visible consequences: definitions of trust are often ad-hoc, and trust relations among different network components and applications are hidden or unknown. Often trust definitions underlying the design of secure protocols are misunderstood by both network designers and users, and lead to unforeseen attacks [30]. Similarly, despite a vast body of work on trust in the social sciences [16, 24], we do not have a formal theory of trust among groups of humans, social and business organizations. Instead, trust is defined by example in different areas of economics, sociology and psychology, and no generally accepted theory of trust exists to date. Hence, we neither have a formal theory of trust for computers nor one for humans; and we certainly do not have a formal theory of trust for networks of humans *and* computers to date. Yet it is quite clear that such a theory is needed in the light of complex interactions in networks of humans and computers in the Internet.

This paper's main contributions to the computer security community are: (1) asking our opening question of trust where humans are as much a part of the system as computers; (2) introducing behavioral trust as a seed toward answering the question; (3) arguing the new economic value introduced by a general theory of trust based on the combination of behavioral trust and computational trust.

2 Impact of a Theory of Trust

We anticipate that a new theory of trust will have significant impact on several important areas of network economics, security, and usability.

New Economic Value. A new theory of trust should explain the establishment of new trust relations where none existed before. The expansion in the kinds of and numbers of trust relations in human and computer networks clearly helps create new economic opportunities and value. New trust relations increase the

pool of usable services, remove cooperation barriers among users, and at the very least, take advantage of "network effects." Cloud computing is the most obvious example of new kinds and numbers of trust relations: people trust companies, e.g., Google, Facebook, and Amazon, with all sorts of personal data, and moreover people trust these companies' computing infrastructure to store and manage their data. New trust relations also help increase competition among network service providers, which spurs innovation, productivity, expanded markets, and ultimately economic development.

New Focus for Security Research. Much of the past research on trust establishment focused on (formal) derivation of new trust relations from old ones; e.g., trusted third-party services, transitive trust relations, delegation. In contrast with prior research, we seek a theory of trust which explains how to create new trust relations that are *not* derived from old ones, and create new opportunities for cooperation among users and among services. For example, it should be possible to establish private, pair-wise trust relations between two untrusting parties that do not necessarily share a common trusted service, such as eBay, which enables reputation systems to work; or a common social network, which might enable recommendations systems to work. While helpful in many cases, trusted third parties create additional complexity and uncertainty, and sometimes become an attractive attack target (e.g., Google). Instead, we seek network infrastructures that enable *unmediated* trust relations, which take us beyond the realm of various (re)interpretations of the end-to-end argument at the application level [7]. In this paper, we argue that network infrastructures that support the establishment of *behavioral trust*, which lower risk and betrayal aversion between untrusting parties and increase beliefs in trustworthiness between these parties, will spur establishment of unmediated trust relations, and as a consequence create new economic value.

Usable Models of Trust. A new theory of trust should be useful for casual users, not just for network and application-service designers. To be useful, such a theory must be easily understandable by designers and users alike. And to be understandable, *a theory of trust has to mirror, to a large extent, human expectations and mental models of trust.* For example, users understand how to separate and protect physical and financial assets in everyday activity. Similarly, they would understand and expect computer systems networks to enable them to separate information assets, be they local system services or financial data held by bank servers. Furthermore, a new theory of trust must not create false metaphors and analogies with the physical world. The email trust model is an example of false expectations: the widespread user expectation that electronic mail would mirror the trust model of physical mail (e.g., authenticity, confidentiality, non-repudiation of receipt, delivery in bounded time) has misled many unsuspecting users into accepting spam, misleading ads, and malware. In contrast, the trust model of eBay follows a well-established, traditional human trust example: it establishes trust relations based on reputation, and to counter inevitable protoctol non-compliance and trust failures, it uses insurance-based recovery mechanisms.

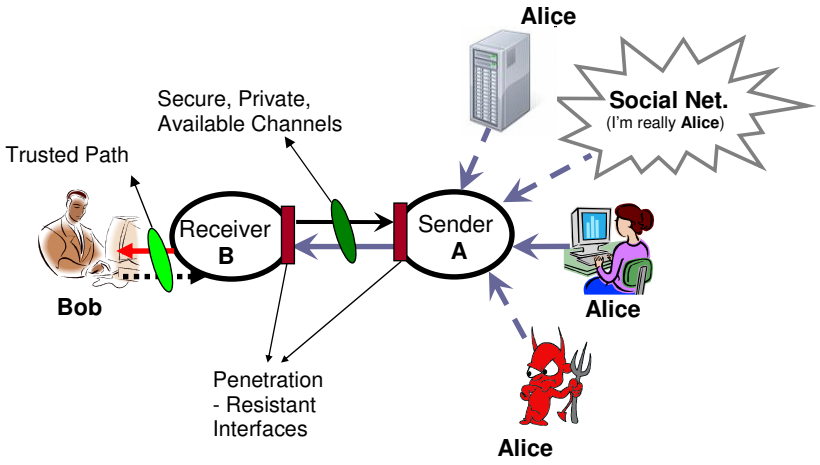


Fig. 1. Simple Communication Model: Entities and Channels

We begin with a simple communication model (Section 3), primarily to state our assumptions. In addressing our opening motivating question, we focus on notions from computational trust (Section 4) that leads us naturally to introduce notions from behavioral trust (Section 5). We explore in Section 6 the implications of behavioral trust for the creation of novel computing and institutional infrastructures. We close in Section 7 with directions for future research.

3 A Simple Communication Model

We assume a network of parties who exchange information with each other over communication channels. When party A (Alice) sends information to party B (Bob), we call A the *sender* and B the *receiver*. Senders and receivers could be human users, network hosts, network applications, and even purveyors of malware (see Figure 1). To focus on trust in networks of humans and computers, we need a simple communication model that allows us to separate questions of sender/receiver *behavior* from more traditional questions of trust specific to communication-channel security, privacy and availability.

Communication-Model Assumptions. Figure 1 summarizes our main assumptions: communication channels are secure and private, and trusted paths can be used by users to communicate with their local computers. We also assume that all communication channels are available for sender-receiver communication whenever needed. Furthermore, we assume that the receiver's and sender's interfaces to communication channels are penetration resistant. Properties of penetration resistance are relatively well understood and have already been formally defined [15]. Recent research [17, 28] illustrate how attack surfaces exposed by these, and other, interfaces can be measured and reduced.

4 Computational Trust Overview

Let's tease apart the opening motivating question "How can I trust the information I receive?" in a purely computational setting. This information is produced by another, possibly anonymous party, or multiple parties, and arrives via messages on communication channels. Thus, first we equate "information" to be literally "contents of messages" (or for brevity, "messages"). We then decompose this question into two separate sub-questions: "Is the communication channel via which I receive messages from a sender secure?" and "How can I trust the sender of the messages I receive?" This decomposition allows us to equate trust in the information I receive with the *act of trusting the sender*. Let's consider the first sub-question. Since we assume secure communication channels, we rule out undetectable corruption, insertion, and replay of the messages received by a third party. Similarly, receipt of secret information over a secure channel rules out non-negligible secret leakage to a third party and reduces any confidentiality concern to trusting the sender with keeping that secret.

In short, assuming the security of communication channels, the question of "How can I trust the information I receive?" (possibly via an anonymous channel) reduces to: "How can I trust the sender of the message I receive?" For example, if I click on a link to a website, can I trust the contents of the message received from that website? Of course, similar questions can be asked in the context of multiple message exchanges and multiple protocol participants. For much of the balance of this paper, we focus on possible answers to this question and assume that communication takes place over secure channels; i.e., we separate concerns of communication security from the act of trusting the sender by the receiver.

Value Underlying the Act of Trusting the Sender. In sender-receiver communication, both the sender and the receiver must derive some *value*, i.e., the benefit of communication must exceed the cost. If there is no value for either sender or receiver, the question of trusting the message sender does not arise, since communication does not take place. For the receiver, value materializes whenever the receiver depends on that information for further action, e.g., the receiver uses the information to invest resources. For example, reading the contents of a message may be very beneficial to the receiver, as the message may contain a program which solves a problem the receiver could not solve himself. However, executing that program on the receiver's system may also have an unexpectedly high cost, e.g., the program may contain "malware" that damages his system, or provides incorrect results, which may cause the receiver to take costly future actions. Hence, the receiver has to decide whether the value derived from reading a message received exceeds the cost of the potential damage caused by that message.

Similarly, a sender derives value in responding to a request from a receiver. The sender benefit materializes whenever there is a payment transfer from the receiver, or when a request represented by a click on a link to the sender's website is monetized. However, when processing the response, a sender uses its resources at some cost, which must be lower than its benefit.

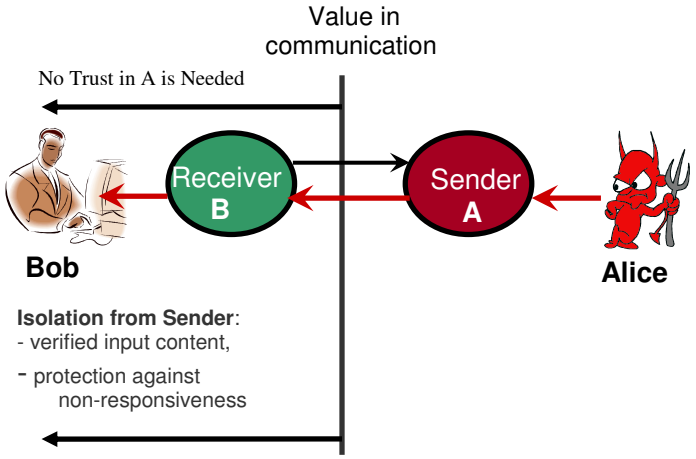


Fig. 2. Receiver's Isolation from Sender's Behavior

4.1 Isolation

Isolation from Sender's Behavior Implies No Trust is Needed. Let's return to our simple question of "How can I trust the sender of the message I receive?" Suppose that the receiver can verify the validity of the message at a reasonably low cost, certainly lower than the benefit derived from it. That is, the receiver can derive value from the message received by verifying all its properties that could affect his resource investment and future state. For example, suppose that the receiver can verify whether the message contains malware, incorrect solutions to a problem of interest, incorrect proofs, or other disinformation. If verification indicates that the received information is valid, then the receiver can act on it and be better off than before. Otherwise, the receiver can reject it, and avoid negative consequences. In short, verification of message properties isolates the receiver from a malicious or incompetent sender. That is, regardless of what message the sender sends the receiver, her input to the receiver cannot damage the receiver. Hence, there is no need to trust the sender. Isolation establishes a well-defined boundary or interface between the sender's and receiver's interests. Figure 2 illustrates the role of a receiver's isolation from the sender's behavior.

In practice, support for isolation can be implemented by means other than direct receiver verification. One way is to get a *second opinion* from a different independent sender, thereby protecting the receiver from the interests and/or competence of the first sender. For example, two independent *authorities* that provide acceptable answers on a subject matter may offer independent advice in response to a query. A positive comparison, which indicates that the two independent advice messages are similar, would cause the receiver to trust the (original) message more than before, and perhaps simply to accept the first sender's message as valid without further checking. However, a negative com-

parison would mean that the advice messages are contradictory, and thus the second opinion would offer no help in isolating the receiver from a malicious or incompetent authority. Furthermore, for a second opinion to be useful, it also has to be available whenever needed. Receiving messages from an authority on a particular subject of interest to the receiver can be expensive, and the cost of soliciting advice from multiple authorities may exceed the receiver's benefit. The notion of "second opinion" generalizes to "n opinions," but at a cost.

Another way to achieve isolation from a sender is by *outsourcing validation*, e.g., via a recommendation system whereby multiple parties independently evaluate (e.g., score) the validity of the message received from that sender. Unlike the case where the receiver gets a second opinion from an independent *authority*, a recommendation system allows inconsistent evaluation of a sender's message by independent parties. In effect, the receiver outsources the decision to trust the message (and thus implicitly trust the sender) to a crowd of recommenders. The receiver's decision whether to accept a message from a sender would be based on the dominant recommendation made by a crowd of independent evaluators. To be useful, however, recommender systems must make available a sufficient number of independent recommendations when needed by a receiver. Availability of a sufficiently large number of independent evaluations helps assure that various biases about the particular message content (and thus sender) are diminished. A more challenging situation is when there is lack of independence of the recommenders.

While receiver verification of a sender's input message represents a major part of receiver isolation from sender behavior, it is not the only isolation property that protects against a malicious or incompetent sender. Isolation from a sender's behavior also requires the receiver to protect against the event that the sender does not reply to the receiver's query in due time. In short, because a sender may not comply with communication protocol requirements (e.g., protocol norms, use agreements) *receiver's isolation from sender's behavior*, comprises: (1) the verification of a received message and (2) protection against deliberate sender non-responsiveness (independent of any network availability concerns, which we assume away by our communication model).

4.2 Trustworthiness and Correctness

Trustworthiness is Stronger than Correctness. Traditionally, the notion of trustworthiness in computer security has been identified with that of correctness. For example, this interpretation of trustworthy computer systems is embodied in all evaluation criteria, from the Orange Book (1985), to EU's ITSEC (1987), and to Common Criteria (1996). In fact, these evaluation criteria have tiered levels of correctness (assurance). However, the notion of trustworthiness is *strictly stronger* than the notion of correctness. In our example, the correctness of the sender's code is a necessary condition for sender's trustworthiness, but it is not a sufficient condition. Why? A sender (computational entity) that satisfies its specifications may still use input of uncertain validity from a human in preparing its response to the receiver. Or the sender may not start in a correct initial

state (e.g., after a crash). Or it may not be configured correctly. As a result, we must think of a sender's trustworthiness as a combination of computational correctness properties *and* human trustworthy behavior. In other words, whenever we have a *human in the trust loop*, the notion of trustworthiness must incorporate both a computational primitive, namely correctness, and a behavior trust primitive, namely beliefs of human trustworthiness.

To establish a sender's trustworthiness a receiver needs to verify the validity of *trustworthiness evidence* both in a computational (e.g., by checking proofs of response validity or using proof-carrying code, to protect herself from incorrect, stale, inconsistent evidence) and a behavioral setting (e.g., beliefs established as a results of trust and gift-exchange games—see discussion below). Also, the delivery mechanism for obtaining evidence of sender's trustworthiness, whether on-line via secure and available channels or off-line, ultimately raises different, but separate, questions of trust, as already discussed above.

We note in passing that sender trustworthiness itself may require sender's isolation from receivers. For example, in traditional security settings, a sender responds to multiple receivers' requests via channels with fairly complex interfaces (e.g., large, complex sender code typically has a wide attack surface). Competing receivers may, in fact, penetrate the sender and invalidate all sender's evidence of trustworthiness. This explains the key underlying requirement that security kernels (and trustworthy computing bases, in general) need isolation to be trustworthy, and illustrates further the need for our penetration-resistance assumption of channel interfaces.

Separation of Sender's Trustworthiness from Receiver Isolation. Suppose that the receiver can obtain incontrovertible evidence that the sender's trustworthiness; i.e., both computationally and behaviorally. Then the receiver can accept all messages from the sender *without additional* verification. The sender is *trustworthy* by definition, and hence the receiver need *not* isolate himself from the sender.

4.3 The Act of Trusting the Sender

Now suppose that the receiver cannot isolate himself from a sender. For example, isolation is not possible in cases when the receiver cannot verify the output behavior of a program contained in a sender's message, since such verification may lead to an undecidable problem. Or, the verification cost may exceed the value of the response to the receiver; e.g., the receiver may have to verify the solution to a co-NP complete problem, which is very unlikely to be possible in polynomial time. Or, a second opinion regarding the validity of a message received and outsourced message-validity checking may be unavailable. Furthermore, suppose that the cost of obtaining and checking evidence of sender's trustworthiness may be too high.

In such cases, the receiver can extract the value of communicating with the sender only if the receiver agrees: to wait for the response to his query *without* any guarantee of a response, and to accept the sender's message received *without* verifying its validity. That is, the receiver benefits only if the receiver *trusts the*

sender. There must be value to the receiver in waiting for the sender's message and accepting it without verification, even in the absence of evidence of sender trustworthiness or protection against invalid sender messages. The essence of trusting the sender is that the receiver places himself and his resources at the disposal of the sender *without* any protection against being cheated, misinformed, or otherwise led down a garden path. Why would the receiver do this? Because there is tangible value to be had.

Value Outcomes of the Act of Trust: The act of trusting transfers value from the receiver of information to the sender and back, at a cost.

Here now is the link between computational trust and behavioral trust. Consider henceforth the receiver to be the "trustor" or "investor" and the sender to be the "trustee." The trustor incurs a cost by clicking on a link to the trustee. This cost, for example, could be the cost of executing the action or opportunity cost (of not clicking on a different link). Yet the trustor willingly transfers value to (e.g., pays) the trustee. In addition, the market may amplify the trustor's transfer, as it may help the trustee monetize the trustor's click. The trustee can choose to transfer value back to the trustor, as it expands resources to respond correctly to the trustor's request, instead of cheating, i.e., instead of providing an arbitrary response, if not a malicious one, a partial response, or no response at all. However, if the trustor anticipates that the trustee will cheat, the trustor would avoid contacting the trustee (i.e., by not clicking on the link for a website), and no value would be exchanged. In summary, the *act of trusting* has the following three possible value outcomes:

1) if the trustor trusts the trustee *and* the trustee is *trustworthy*, then both the trustor and trustee are better off than before executing the protocol, i.e., cooperation pays off.

2) if the trustor trusts the trustee *and* the trustee is *untrustworthy*, the trustee is better off and the trustor is worse off than before, i.e., trustee has strong incentive to cheat in the absence of a mechanism that protects the trustor.

3) if the trustor suspects that the trustee will cheat and hence, does not initiate the protocol, no value is exchanged.

By building on computational techniques (e.g., cryptography, verification, fault-tolerance) that give us trust among computational entities (e.g., computers and networks) we are left with non-computational entities. Thus now when we talk about trustors/receivers and trustees/senders we are talking about humans: the person (sender) who wrote the code to be run remotely or who wrote the Wikipedia page and the person (received) who requested the message contents, e.g., by clicking on the webpage. This means we need to look at theories that explain trust relations among humans.

5 Behavioral Trust Overview

One-shot Trust Games. Recent research in behavioral economics defines a game that captures the notion of trust described above [3, 11, 10], which we present as a motivation for the foundations of *behavioral trust*. In a typical one-shot trust

game (illustrated in Figure 3), two players, the trustee A and trustor B, both of whom remain anonymous, are endowed with $10 each by a dealer. Player B must decide whether to hold on to his endowment or send it to A. If he sends it to A, the dealer quadruples the transferred value received by A. Player A is supposed to return half of her accumulated value (i.e., ($40 +$10)/2) to B. So both players end up with $25 each if they cooperate, and only with their initial $10 endowment, if they do not. If B transfers his endowment to A and A does not comply with the protocol (i.e., A cheats and transfers less than $25), B loses his endowment and receives less than the $25 expected, if anything. Depending upon her behavior, A is left with as much as $50, and certainly with more than the $25 guaranteed by cooperation, in worst-case cheating. That is, A retains her initial endowment of $10 plus $40 representing the quadrupling of the value transferred by B ($10) by the dealer. Hence, player A has an incentive to cheat.

The trust game mimics a sequential economic exchange in *the absence of any mechanism* that would enforce a contract between players A and B, and has the same flavor as the sender-receiver protocol outlined above in the absence of the receiver B's isolation and sender A's trustworthiness; i.e., it illustrates receiver B's *act of trusting* sender A. Experiments in neuro-economics [11] show that the vast majority of players in the role of B, namely 14 of 15 players, trust the players in the role of A contrary to rational economic theory, which would have each player pocket the $10 endowment and not play.

The trust game also illustrates the definition of cooperation and fairness *norms in sender-receiver protocols*. Specifically, player A is supposed to send half of her accumulated value to player B; i.e., ($40+$10)/2. If A decides not send $25 to player B, then B interprets this as a norms violation, or a protocol non-compliance act.

The same experiments [11] also show that protocol non-compliance evokes a sense of betrayal in player B and a desire to punish violator A, even if *punishment comes at a cost* to B. In the continuation of the trust game, player B is informed of A's protocol violation, is given one minute to decide whether, and how much, to punish A, and his neural reaction is measured. Specifically, in a first experiment, where punishment costs, player B is given 20 punishment units, U, each costing B $1 and A $2; i.e., for every punishment unit U, B pays $1 so that A loses $2. In another experiment, punishment is free; i.e., it costs B nothing to reduce A's ill-gotten gain by $2. In the third experiment punishment is symbolic; i.e., neither B nor A incur any cost. Finally, in the last experiment A's decision to cheat is made by a device, at random; i.e., the device replaces cheating player A. To measure player B's neural response to punishment, a PET scan of B's brain was taken during the one-minute interval B's decision was made. Figure 4 summarizes three outcomes of these experiments. First, player B is willing to *incur a cost to punish*, and the amount of punishment inflicted was higher when punishment was free. Second, player B derived satisfaction (i.e., felt rewarded) proportional to the amount of punishment inflicted on cheating player A; i.e., the stronger the satisfaction player B derived, the higher the cost he was willing to incur. This indicates the strength of B's *aversion to being betrayed* by A. It also

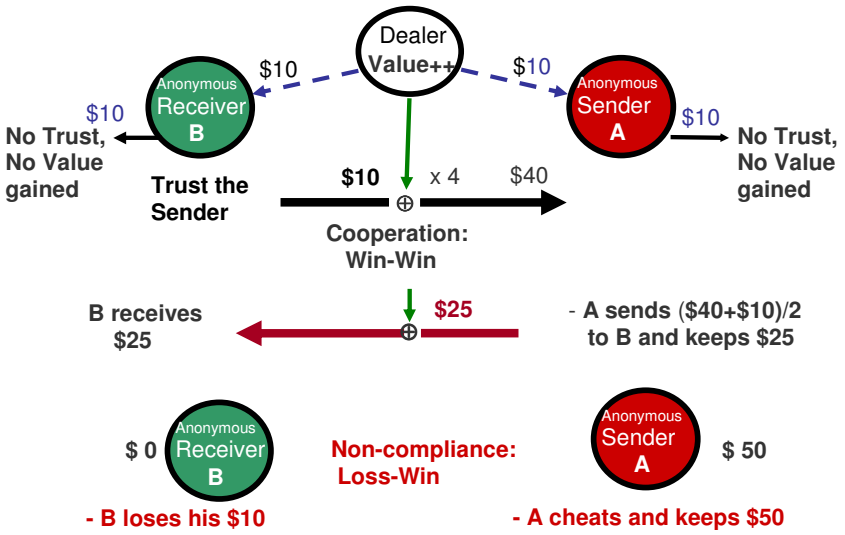


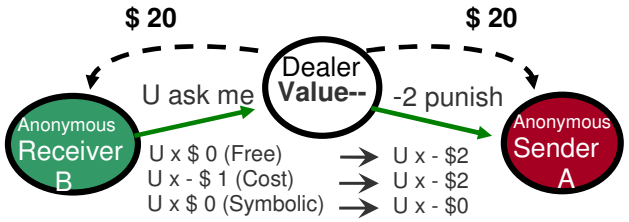
Fig. 3. Outcomes of an One-shot Trust Game

illustrates the fact that B's *punishment is altruistic*, since he is willing to pay to punish even though he is not deriving any material gain. Third, when cheating player A was replaced by a random device that is protocol non-compliant, B's desire to punish is negligible. This indicates that B's *aversion to the risk of losing* money when faced with an ambiguous outcome was different (i.e., lower) from his aversion to being betrayed.

The analogy between our sender-receiver protocol and a one-shot trust game in economics can be valuable because behavioral economics has already discovered the fundamental primitives comprising the act of trusting. Some of these primitives have analogues in computer networks. Others complement these primitives as they refer to *the act of trusting* among humans and apply directly to human networks. In both cases, they address the need for network infrastructures that promote value creation by trustor and trustee cooperation (as suggested by Figure 5).

Behavioral-Trust Primitives. Behavioral trust is characterized by three *independent* primitives studied extensively in economics. As the trust game suggests, these are:

1) A trustor's *risk preferences*. This captures the trustor's (player B above) degree of risk aversion when faced with ambiguous outcomes; i.e., the trustee's (player A above) trustworthiness is uncertain. Much of the research on decision making under uncertainty relies on these *asocial* preferences [37, 20].
2) A trustor's *social preferences*. This captures the trustor's degree of *betrayal, or non-reciprocation, aversion* when faced with a trustee who may cheat or perpetrate a scam [10, 8, 9]. Recent research also indicates that betrayal



1) "Punishment": Receiver B asks Dealer to punish cheating Sender A in units U
Free ~ 18 U (1 U -> $0 cost) punishment: ~ - $ 36
Cost ~ 11 U (1 U -> - $1 cost) ~ - $ 22

2) PET scan of Receiver B's brain striatum indicates reward
- betrayal aversion measure: higher reward <=> higher the punishment
- punishment is altruistic

3) Betrayal Aversion ≠ Risk Aversion: Sender A is a random device
=> Receiver B: very small desire to punish and very little (reward) satisfaction
cost ~ $ 2 punishment: < $ 4

Fig. 4. Experiments Illustrating Betrayal and Risk Aversion

aversion—a social preference—is very different from and independent of asocial risk preferences. For example, in the one-shot trust game above, player B's neurological reaction is very different when the cheating player A is a human as opposed to a device, e.g., a computer. Recent research also suggests that trust is partially heritable [10].

3) A trustor's *belief in trustworthiness* of trustees. The trustor forms probability beliefs about a trustee's actions and uses these beliefs to guide the trusting act; e.g., it measures the difference between the expected utilities of trust and distrust, both of which are computed using probability beliefs. Recent research also indicates that *trustworthiness beliefs* are independent of a trustor's *risk and social preferences*. Trustworthiness beliefs can be enhanced by reputation building opportunities, including reputation protocols, with or without recommender features, etc. Gift-exchange games capture reputation building in behavioral economics [10].

Standard models of trust, both in the social sciences (i.e., economics, sociology and psychology) and computer science, rely on two of the three primitives above: risk preferences and formation of trustworthiness beliefs. Hundreds of articles have been published on these two primitives separately in the social sciences [16, 24] and computer science. However, there is little practical experience applying these notions in networks of computers and humans; e.g., only eBay's model of trust attempts to diminish risk aversion by offering insurance

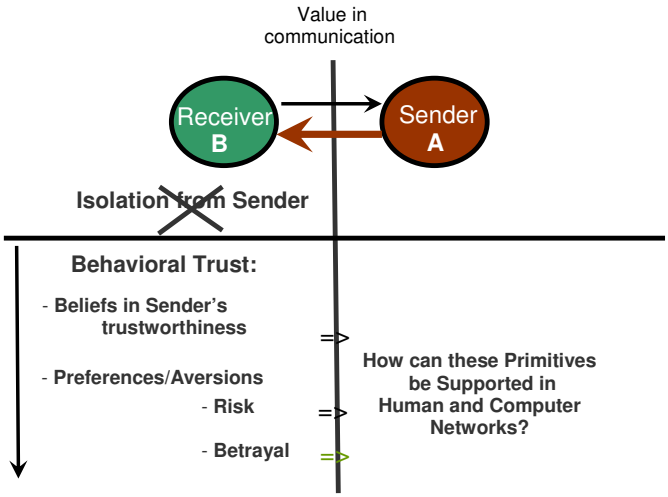


Fig. 5. Summary of Behavioral-Trust Primitives

to trustors and enhance trustors' beliefs in trustees' trustworthiness by reputation protocols. With the exception of recent research in behavioral trust, both theory and practice missed the notion of betrayal aversion and its application in networks of humans and computers.

6 Usefulness of Behavioral Trust in Networks of Humans and Computers

Is there Anything New in "Behavioral Trust?" Yes, betrayal aversion (as characterized by social preferences). Standard economic models of trust account only for trustors' risk aversion and the role of reputation-formation institutions. These models emphasize reputation and guided-belief-formation games, typically using Bayesian updating procedures. Similarly, in computing, there have been several reputation-building exchanges [19] and recommender systems with collaborative filtering [29, 14]. The experiments in neuro-economics mentioned above show that betrayal aversion has a larger inhibiting role in the act of trusting than risk aversion. Intuitively, people are more willing to take risks when faced with a certain probability of bad luck than when faced with an identical probability of being cheated by others.

Behavioral trust shows that standard economic models miss the betrayal-aversion component of trust altogether, and hence they miss opportunities to deal with it in a constructive way, i.e., build system infrastructures that deter untrustworthy trustees from cheating. Intuition suggests that *deterrence* requires punishment, and punishment requires accountability [23]. (Accountability is also necessary in reputation-building exchanges, not just for deterrence.) However,

intuition does not explain what punishment and accountability infrastructures are sufficient for deterrence; e.g., accountability does not prescribe any punishment and it is unclear what type of punishment deters cheating and encourages trustworthy behavior. Research is needed in this area.

Why Do We Expect Behavioral Trust to Create Value in Networks of Humans and Computers? Trustworthiness beliefs are positively correlated with economic prosperity (as measured by GDP), though no causal relationship has been established, yet [10]. More acts of trusting (e.g., based on lower risk- and betrayal-aversion and higher belief in trustworthiness), however, should lower the cost of economic transactions, promote trade, and hence create wealth [10]. Also, restrictive security mechanisms designed to protect parties involved in economic transactions should diminish the need for acts of trusting but increase transaction cost. By analogy, the spread of the Internet and the Web, which was largely built on (often unwarranted) acts of trusting, could be curbed by overly restrictive security mechanisms.

If causality between the act of trusting and wealth creation has not been proven, why should we really look at economic primitives to encourage acts of trusting in networks of computers and humans? Because we need *usable security* by humans. Current approaches have not been particularly successful and are not scaling up to meet the increases in both functional demands and numbers of participating entities. Most communication infrastructures built in computer systems and networks to date rely either on receiver-isolation mechanisms, namely on verification of input content, or on (often unwarranted) beliefs of sender's trustworthiness. While isolation mechanisms are often necessary, they can be: (1) cumbersome to use, (2) costly to obtain, set up and administer [23], and (3) often psychologically less than acceptable; e.g., enforcing them promotes mistrust and may be unappealing to use. In contrast, unwarranted acts of sender trust have proven to be very dangerous as they can lead to breaches of security, ranging from social scams, to identity theft, and to failures to protect the national interests. We propose this alternative: the *act of trusting* based on *evidence-driven* beliefs of trustworthiness and *robust infrastructures* for diminished risk- and betrayal-aversion. In short, behavioral trust appears to be a good place to start.

Are the Three Primitives of Behavioral Trust Sufficient? Or Are There Other Trust "Determinants" That Are Unaccounted For (e.g., Religion, Ethnicity, Geography, History)? Current economic theory emphasizes that *all* individual behaviors, such as the act trusting others, are captured by (1) preferences and (2) beliefs. To date, economic thought indicates that all other determinants are already captured by preferences and beliefs primitives. The three independent primitives discussed above, namely *risk preferences*, *social preferences*, and *beliefs in trustworthiness* are sufficient to characterize behavioral trust completely [10].

Behavioral Trust Answers the Question "How can I believe the input I receive from a trustee?" What About Trusting the Trustee with the Secrecy of the Trustor's Sensitive Output? Behavioral trust captures both cases. In the lat-

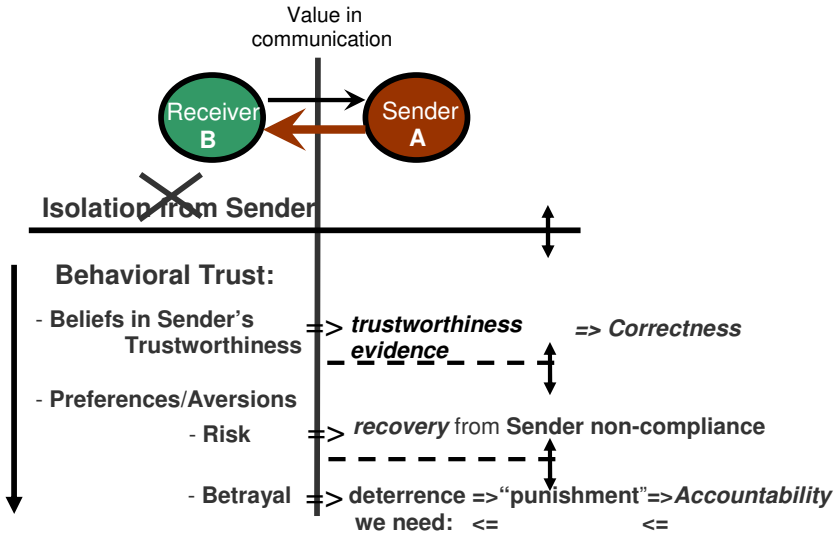


Fig. 6. Behavioral-Trust Primitives and their Infrastructure Requirements

ter, the trust game can also be gauged in terms of trusting the trustee with maintaining the secrecy of output, e.g., digital rights management, information leakage via covert channels, traitor tracing, and personal data. Output trust is not explicitly addressed herein because we suspect it has similar behavioral characteristics of those for input trust, though the details need to be worked out.

How Can Social "Institutions" (aka Infrastructures) Enhance The Act Of Trusting? Economic theory shows that risk and social preferences are less malleable than beliefs in trustworthiness. Then does it make sense to try to affect them in an attempt to induce more trusting behaviors? First, since *betrayal aversion* is an inhibitor for the act of trusting, any institution that promotes *deterrence* against cheating trustees can lower betrayal aversion and increase trust. Second, beliefs in trustworthiness of others, the most malleable of all three trust primitives, can be institutionally increased; viz., different gift-exchange games presented by Fehr et al [10] illustrate the formation of trustworthiness beliefs. In computer networks, the formation of trustworthiness beliefs can illustrated by reputation protocols, such as the ones used by eBay. Hence, creating network and social infrastructures that lower betrayal and risk aversions and increase beliefs in trustworthiness makes intuitive sense. This is summarized in Figure 6.

Like betrayal aversion, risk aversion can also be reduced institutionally to encourage acts of trusting. However, institutional reduction of risk aversion may create "moral hazard," and hence trustors may become imprudent, e.g., may take unwarranted risks in transactions where risk reductions are absent. Moral hazard can be countered by infrastructure mechanisms that reduce risk at some cost to trustors; for example, cost is incurred in invoking recovery mechanisms,

in purchasing and exercising insurance, and with credit misuse. Ebay's insurance protocol illustrates this idea in practice.

All proposed infrastructures must be carefully designed and analyzed since adversaries (e.g., cheaters, scammers) may find ways to exploit weak mechanisms and perpetrate more fraud. As always, the design of robust infrastructures must account for the presence of (rigorously defined) adversaries. Note that the mix of infrastructure protocols that would respond to adversary attacks can vary. The vertical arrows of Figure 6 represent a graphical illustration of the idea that the boundaries between different behavioral trust primitives and isolation mechanisms are flexible, and can be moved as applications demand. For example, in some applications, one can rely on isolation alone, and then one need not develop protocols that enhance beliefs of trustworthiness, reduce risk and deter. In other applications where isolation is impractical, one might rely more on recovery and less on deterrence. Yet in others, accountability may lead to deterrence, in which case one would rely on recovery to a smaller extent.

Are Beliefs in Trustworthiness of Others Long Lasting? Are They Self Reinforcing and Can They Reach Stable Equilibrium? Available evidence in this area is mixed; e.g, see research work cited in [10]. For example, it is unclear whether trustworthiness beliefs are self-reinforcing, despite some empirical evidence that this may be the case. Furthermore, there is some evidence which indicates that, although beliefs in trustworthiness can increase, initial high levels of trustworthiness beliefs are not always sustainable. A consequence of this possibility would be to sustain acts of trusting by external factors; e.g., periodic administrative action.

Can We Explain Social Scams as Games Based on Unwarranted Acts of Trusting? Some scams illustrate how a trustee, often aided by social collaborators, can induce an unsuspecting trustor to engage in social protocols and games where eventually the trustor is cheated. The social scams described by Stajano and Wilson [35] and Ryan [33] include, but are not necessarily limited to, some trust-and gift-exchange games [10]. Christin et al. [6] illustrates how a trustee extracts value from a trustor in one-click requests by delivering only blackmail. The challenge seems to be in representing scams as sequences of games and then studying their composition.

If Behavioral Trust Is So Useful, Why Hasn't It Been Already Adopted By Security Economists? We see three reasons for this: (1) most security economists [2] have used the standard economic models mentioned above, which include only risk preferences and beliefs in trustworthiness of others, but have not incorporated betrayal aversion in their models, and hence missed the related notions of *institutionalized deterrence*; (2) recent work on social scams and what we can learn from them in security [35], though very interesting, has focused only on the relevance of risk preferences of unsuspecting victims, decision making under uncertainty [20, 37], and bounded rationality [34]; (3) the value of institutionalized deterrence via norms-based punishment has been investigated only very recently [1] and is largely unknown in security economics to date. Provably effective norms definitions must be provided to deter. Unreasonable norms seem to

fail; e.g., folk tales have it that during the Paris Commune (May–March, 1871), public hangings for those who pick-pocket did not deter pick-pocketing at public hangings.

7 Directions for Further Research

7.1 Foundations

In our introduction we argued for the need of a general theory of trust in networks of humans and computers. We pointed out the lack a theory of computational trust, though we believe elements of this theory need to build on concepts such as isolation, correctness, and trustworthiness as introduced in Section 4. So towards a general theory of trust, our first task is to identify primitives of computational trust akin to the three primitives of behavioral trust as introduced in Section 5.

A second task is to understand the precise relationship between concepts from computational trust and those from behavioral trust. Where are they the same? Where are they different? How do they combine and interact? For example, behavioral trust shares with computational trust the notion of belief in others trustworthiness, e.g., gift-exchange games in economics and reputation-building protocols in computer science. Both areas target robust game design, which cannot be manipulated by unscrupulous trustees. The notion of risk aversion and risk management are common to both but are treated differently in practice to account for specific domain differences; e.g., one domain addresses social sciences concerns whereas the other addresses concerns of computer network design.

7.2 Computer Systems, Network Architecture, Computer Security

Another premise of our work is that a general theory of trust will elucidate potential for new economic value. In networks of humans and computers, there are new trust relations that neither computational trust nor behavioral trust address. A general theory will make these trust relations explicit and these new trust relations can be monetized. What implications does this potential economic value have on the design and implementation of computer systems, e.g., cloud computing, and network architectures of the future?

A new theory should provide necessary and sufficient conditions for practical network infrastructures that can enhance the creation of trust relations where none were possible before. Focus on network infrastructures that support behavioral trust poses an interesting challenge for security research: What security mechanisms and policies spur creation of new economic opportunities? This question is motivated, in part, by a direct analogy between security mechanisms in computing networks and those in transportation systems. For example, air breaks in railcars (a safety invention of George Westinghouses in 1869) and automated railway signals and train stops (developed at Westinghouses Union Switch and Signal during 1882-1901) played major roles in substantial safe increases of train speeds, railroad commerce, and economic opportunities. Much of the past

security research has been dedicated to mechanisms and policies that prevent loss. In contrast, we expect that a similar causality to that between increased railroad commerce in the 19th century and increased economic value, which has been firmly established by economics, will be proven to hold between network infrastructures that promote trust and thus economic development. What evidence do we have for this seemingly unusual expectation? The correlation between increased trust and economic development, while not proven to be an economic causality yet, is unambiguous: countries where survey-measured trust among people is high have a higher GDP than countries where trust is lower [10].

Further research is needed to show the practical usefulness of a new theory of trust. In particular, the new theory should address new protocol areas. For example, it should model properties of protocols where trustees are required to maintain the secrecy of data provided by the trustor; it should model social scams and deception; and it should incorporate users' mental models of trust relations.

7.3 Promoting Cooperation in Networks of Humans and Computers

A new theory of trust should address the stability properties of trust relations. We hypothesize that we can get better security, namely more usable security at lower cost, with less emphasis on traditional mechanisms (e.g., isolation, restrictive security policies) and more reliance on system primitives that enhance trust (e.g., deterrence against non-compliance in trust games to decrease betrayal aversion, recovery and resiliency mechanisms to decrease risk aversion, and reputation-enhancement protocols to increase beliefs in others trustworthiness). Furthermore, if trust is eventually shown to be self-reinforcing (e.g., high levels of trust beget high levels of trust), then system primitives that support cooperation/trust become very desirable. If not, acts of trusting could be sustained by external factors; e.g., periodic administrative action that bring networks to known levels of trust.

To test this hypothesis, we should analyze and build network primitives and institutions (e.g., network infrastructures, protocols) that promote cooperation between trustors and trustees. To do so, we can draw inspiration from behavioral economics regarding the malleability of the three separate primitives that characterize behavioral trust: risk aversion, betrayal aversion, and beliefs in trustworthiness.

Addressing all these research tasks and more would bring a tangible contribution to understanding trust in today's networks of humans and computers.

Acknowledgments

We thank Butler Lampson and Manuel Blum who independently raised our opening question. Lampson first raised this question in the context of accepting input in green from red [23]. Blum raised it specifically in the context of reading Web

pages but more generally in the context of networks of humans and computers. Both inspired us to explore this new research direction. We thank Jun Zhao for helping with the preparation of this manuscript.

This research was supported in part by CyLab at Carnegie Mellon under grant DAAD19-02-1-0389 from the US Army Research Office. The first author was also partially supported by the MURI grant W 911 NF 0710287 from the Army Research Office, and by the National Science Foundation (NSF) under grants CNS083142 and CNS105224. The second author was also partially supported by by the National Science Foundation (NSF) under grant number CNS-1064688. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

References

1. Akerlof, R.: Punishment, Compliance, and Anger in Equilibrium. Job Market Paper, MIT Sloan School, November 18 (2010) http://mit.academia.edu/RobertAkerlof/Papers/163148/Punishment_Compliance_and_Anger_in_Equilibrium_JOB_MARKET_PAPER_
2. Anderson, R. J.: Security Engineering, (second edition), Wiley (2008)
3. Berg, J., Dickhaut, J., McCabe, K., Trust, Reciprocity, and Social History. Games and Economic Behavior, no. 100 (1995).
4. Birrell, A., Lampson, B., Needham, R., Schroeder, M.: Global Authentication without Global Trust. In: IEEE Symp. on Security and Privacy, Oakland, CA (1987).
5. Chen, C-H., Chen, C-W., Kuo, C., Lai, Y-H., McCune, J. M., Studer, A., Perrig, A., Yang, B-Y., Wu, T-C.: GAnGS: Gather Authenticate 'n Group Securely. In: ACM Annual International Conference on Mobile Computing and Networking (MobiCom), San Francisco, CA, September 13 - 19, (2008)
6. Christin, N., Yanagihara, S.S., Kamataki, K.: Dissecting One click Frauds. In: ACM CCS, Chicago, Illinois (2010)
7. Clark, D.D., Blumenthal, M.S.: The End-to-End Argument and Application Design: the Role of Trust. Federal Communications Law Journal, vol. 63, No. 2, pp. 357-390 (2011)
8. Falk A., Fischbacher, U.: A Theory of Reciprocity. Games and Economic Behavior. 54, 293-315 (2006)
9. Fehr, E., Fischbacher, U., Kosfeld, M.: Neuroeconomic Foundations of Trust and Social Preferences. Forschungsinstitut zur Zukunft der Arbeit, IZA (Institute for the Study of Labor), Bonn, Germany (2005)
10. Fehr, E.: The Economics and Biology of Trust. Journal of the European Economics Association (2009)
11. de Quervain, D., Fishbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E.: The Neural Basis for Altruistic Punishment. Science, Aug. 27 (2004)
12. Eschenauer, L., Gligor, V., Baras, J.: On Trust Establishment in Mobile Ad-Hoc Networks. In: Security Protocols Workshop, Cambridge, U.K. (2002) (also in Springer LNCS Volume 2845, 2003).
13. Gligor, V., Luan, S.-W., Pato, J.: Inter-Realm Authentication in Large Distributed Systems. In: IEEE Symp. on Security and Privacy, Oakland, CA (1992) (also in the Journal of Computer Security, 1993)

14. Gligor, V., Perrig, A., Zhao, J.: Brief Encounters with a Random Key Graph. In: International Workshop on Security Protocols, Cambridge, UK, April (2009) (to appear in LNCS, Springer Verlag)
15. Gupta, S., Gligor, V.: Towards a Theory of Penetration-Resistant Systems and its Applications. In: IEEE Computer Security Foundations Workshop, Franconia, NH, June 1991 (also in the Journal of Computer Security, Vol. 1, No. 2, pp. 133-158 (1992)
16. Bachmann, R., Zaheer, A.: Handbook of Trust Research. Edward Elgar Publishing (2006)
17. Howard, M., Pincus, J., Wing, J.M.: Measuring Relative Attack Surfaces. Computer Security in the 21st Century, D.T. Lee, S.P. Shieh, and J.D. Tygar, editors, Springer, pp. 109-137, March (2005)
18. Hsiao, H.-C., Studer, A., Chen, C., Perrig, A., Bai, F., Bellur, B., Iyer, A.: Flooding-Resilient Broadcast Authentication for VANETs. In: ACM Annual International Conference on Mobile Computing and Networking (MobiCom), Las Vegas, Nevada, September (2011)
19. Josang, A., Goldbeck, J.: Challenges for Robust Trust and Reputation Systems. In: 5th International Workshop on Security and Trust Management (STM), Saint Malo, France, September (2009)
20. Kahneman, D., Tversky, A.: Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2):263-291, Mar (1979)
21. Kim, H. J., Owusu, E., Gligor, V., Perrig, A., Hong, J., Yamada, A.: Robust Online Trust Establishment through Visualization of Tie Strength, CMU CyLab Technical Report, Feb. (2011)
22. Lampson, B., Abadi, M., Burrows, M., Wobber, T.: Authentication: Theory and Practice. ACM TOCS (1992)
23. Lampson, B.W., "Usable Security: How to Get It," in Comm. ACM, Nov (2009)
24. Bachmann, R., Zaheer, A.: Landmark Papers on Trust (The International Library of Critical Writings on Business and Management). Vol I and II, Edward Elgar Publishing (2008)
25. Luk, M., Whillock, B., Perrig, A.: Seven Cardinal Properties of Sensor Network Broadcast Authentication. In: ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN), October (2006)
26. Lin, Y-H., Studer, A., Chen, Y.-H., Hsiao, H.-C., Kuo, L.-H., Lee, J., McCune, J. M., Wang, K-H., Krohn, M., Lin, P-L., Perrig, A., Sun, H-M., Yang, B-Y.: SPATE: Small-Group PKI-Less Authenticated Trust Establishment. IEEE Transactions on Mobile Computing, vol. 9, no. 12, December (2010)
27. Lin, Y-H., Studer, A., Hsiao, H-C., McCune, J. M., Wang, K-H., Krohn, M., Lin, P-L., Perrig, A., Sun, H-M, Yang, B-Y.: SPATE: Small-group PKI-less Authenticated Trust Establishment. In: International Conference on Mobile Systems, Applications and Services (MobiSys), 22-25 June (2009)
28. Manadhata, P.K., Wing, J.M.: An Attack Surface Metric. IEEE Transactions on Software Engineering, June (2010)
29. Marbach, P.: A Lower Bound on the umber of Rankings Required in Recommender Systems Using Collaborative Filtering. In: IEEE Conference on Information Sciences and Systems. pp 292-297, Princeton University, NJ. March (2008)
30. Marlinspike, M.: SSL Attack demonstrated at Blackhat (2009) `https://www.blackhat.com/presentations/bh-usa-09/MARLINSPIKE/BHUSA09-Marlinspike-DefeatSSL-SLIDES.pdf`

31. McCune, J. M., Perrig, A., Reiter, M. K.: Seeing-Is-Believing: Using Camera Phones for Human-Verifiable Authentication. In: IEEE Symposium on Security and Privacy (2005)
32. Raya, M., Papadimitratos, P., Gligor, V., Hubaux, J.P.: On Data-Centric Trust Establishment in Ephemeral Ad Hoc Networks. In: INFOCOM (2008)
33. Ryan, T.: Getting in Bed with Robin Sage. Black Hat Conference, (2010).
34. Simon, H.A.: Rational choice and the structure of the environment. Psychological Review, vol. 63, pp. 129-138, Mar (1956)
35. Stajano, F., Wilson, P.: Understanding Scam Victims: Seven Principles for Systems Security. Comm. ACM, vol. 54, no. 3, pp. 70-75, (2011)
36. Surie, A., Perrig, A., Satyanarayanan, M., Farber, D.: Rapid Trust Establishment for Pervasive Personal Computing. IEEE Pervasive Computing, 6(4):24-30, October-December (2007)
37. Tversky, A., Kahneman, D.: Judgment under Uncertainty: Heuristics and Biases. Science, 185(4157):1124-1131, Sep (1974)
38. Wendlandt, D., Andersen, D., Perrig, A.: Perspectives: Improving SSH-style Host Authentication with Multi-Path Probing. In: USENIX Annual Technical Conference, June (2008)