The Influence of Trust Score on Cooperative Behavior

Claudia-Lavinia Ignat, Quang-Vinh Dang, Valerie Shalin

The influence of trust score on cooperative behavior

CLAUDIA-LAVINIA IGNAT, Université de Lorraine, CNRS, Inria, LORIA, France
QUANG-VINH DANG*, Université de Lorraine, CNRS, Inria, LORIA, France
VALERIE L. SHALIN, Department of Psychology and Kno.e.sis, Wright State University, USA

The assessment of trust between users is essential for collaboration. General reputation and ID mechanisms may support users' trust assessment. However, these mechanisms lack sensitivity to pairwise interactions and specific experience such as betrayal over time. Moreover, they place an interpretation burden that does not scale to dynamic, large-scale systems. While several pairwise trust mechanisms have been proposed, no empirical research examines trust score influence on participant behavior. We study the influence of showing a partner trust score and/or ID on participants behavior in a small-group collaborative laboratory experiment based on the trust game. We show that trust score availability has the same effect as an ID to improve cooperation as measured by sending behavior and receiver response. Excellent models based on the trust score predict sender behavior, and document participant sensitivity to the provision of partner information. Models based on the trust score for recipient behavior have some predictive ability regarding trustworthiness, but suggest the need for more complex functions relating experience to participant response. We conclude that the parameters of a trust score, including pairwise interactions and betrayal influence the different roles of participants in the trust game differently, but complement traditional ID and have the advantage of scalability.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; **Empirical studies in collaborative and social computing**; • **Security and privacy** → *Economics of security and privacy*; *Usability in security and privacy*;

Additional Key Words and Phrases: trust, reputation, cooperation, trust game

ACM Reference Format:
Claudia-Lavinia Ignat, Quang-Vinh Dang, and Valerie L. Shalin. 2019. The influence of trust score on cooperative behavior. *ACM Trans. Internet Technol.* 19, 4, Article 46 (September 2019), 22 pages. https://doi.org/https://doi.org/10.1145/3329250

1 INTRODUCTION

Flattened organizational hierarchies promote reliance on direct peer-to-peer interactions [McChrystal et al. 2015]. However, this increases both the number of interactions and critically, the number of peers interacting with each other. At the same time, *ad-hoc* work groups increasingly respond to transient need, as in Wikipedia modifications, crisis response, political activism and software development. Technology facilitates these ad-hoc work groups, allowing users from different locations to collaborate without the need for face-to-face interaction.

However, psychological considerations exist concerning globally distributed work, particularly the cognitive demand associated with developing, maintaining and accessing a large number of

*now at Data Innovation Lab, Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam

Authors' addresses: Claudia-Lavinia Ignat, Université de Lorraine, CNRS, Inria, LORIA, Inria Nancy – Grand-Est, 615 rue du Jardin Botanique, 54600, Villers-lès-Nancy, France, claudia.ignat@inria.fr; Quang-Vinh Dang, Université de Lorraine, CNRS, Inria, LORIA, Inria Nancy – Grand-Est, 615 rue du Jardin Botanique, 54600, Villers-lès-Nancy, France, quang-vinh.dang@inria.fr; Valerie L. Shalin, Department of Psychology and Kno.e.sis, Wright State University, Fawcett Hall 447, 3640 Colonel Glenn Hwy, Dayton, OH 45435-0001, USA, valerie.shalin@wright.edu.

interactions with a large network of partners. Rentsch and Klimoski [2001] identify shared knowledge, including schemas, goals and values as fundamental to effective collaboration. Accordingly, overall performance benefits from participants who share expertise and world views. Critically, acquiring knowledge about a collaborator's expertise and worldview is a *process*. Some researchers specifically identify *transactive memory* as essential to successful performance in distributed work [Chang 2004], allowing participants to solicit assistance and information from the best resource for the task at hand. Exploiting transactive memory entails cognitive demand for encoding, retrieving and updating representations of each participant's capabilities, stored in declarative memory.

In a conventional work environment, participant names serve as a retrieval cue for recalling a participant's expertise and personally observed, previous behavior maintained in declarative memory. Consistent with this practice, virtual identities or IDs, either assigned by a system or chosen by users, distinguish between participants. Similar to the notion of branding, when participant A sees the ID of B, participant A can recall the experiences she had with B, and engage accordingly. Of course, nonsense ID strings such as "p67718an22187bOz" [Dix 2009] are not remembered well. More concerning is that psychological research has established the persisting response time penalties of increasing the size and interconnectedness of declarative content such as ID [Anderson and Reder 1999]. As a result, increasingly large, dense networks with rarely accessed nodes, such as those made possible by internet collaboration, pose retrieval problems, and hence access to the knowledge that supports effective collaboration.

In this paper we seek a computationally derived, behavior-based substitute for the above-mentioned demands that scales better than ID with increasing network size. We suggest the presentation of trust scores as a substitute for maintaining detailed, qualitative accounts of prior experience between partners. We use the definition of trust as "a cognitive learning process obtained from social experiences based on the consequences of trusting behaviors" [Cho et al. 2015], where trust is built based on observations in the past.

Trust and *reputation* are sometimes used interchangeably [Pecori 2016; Vu et al. 2010]. Though related, they are not the same constructs [Fetchenhauer and Dunning 2009]. Consistent with [Breitmoser 2015], we consider reputation as the *collective opinion of a community* regarding a particular participant, while trust is the *specific* relationship between a pair of participants. Participant reputation is a *global* value, while trust in a participant is a *personal* value and differs by partners [Hoelz and Ralha 2015]. This distinction allows us to accommodate the different concerns and corresponding weights that one participant has relative to another. Psychological research supports the claim that different participants view the trustworthiness of the same target differently [Bergman et al. 2010]. Personality and perceptual bias may also influence an observer's assessment of a target's trustworthiness. For these reasons we seek a metric that characterizes pairwise trust.

Using these definitions, the widely used Internet scoring systems are reputation–not trust–systems [Resnick et al. 2000]. Examples include the Amazon reputation score, calculated by averaging all rating scores from *all buyers*, such that every buyer will see the same score when they examine the seller's profile. The heart of a reputation system is therefore indirectness: one benefits from interacting with participants who have been shown to be trustworthy with other people. Studies such as [Tadelis 1999] established a strong connection between reputation and name. Bolton [Bolton et al. 2004] demonstrated the effectiveness of reputation scores in e-commerce. General reputation is particularly relevant in the absence of repeated interaction with a particular individual. In [Resnick et al. 2006] the authors conducted a controlled field experiment of an Internet reputation mechanism where they ruled out several potential confounds appearing in previous observational studies such as seller skill, product quality and seller responsiveness to customer inquiries. The study found that sellers with high reputation fared better; buyers were willing to pay a well established reputation seller 8.1% more on average than a new seller for the same item.

However, the study did not analyze how repeated interaction between a customer and a seller is influenced by reputation score.

The main limitation of reputation systems is their vulnerability to third party manipulation [Hoffman et al. 2009]. Suppose Alice wants to know the reputation score of Bob, that is, the community opinion of Bob. Alice may query other users about Bob, say Carol or Dave. Or she may acquire a reputation score from a central server, which has collected the opinions about Bob from all users. In any case, Alice relies on information from third parties. The information might not be available, i.e. a central server might go down or is unavailable in a peer-to-peer system, or Carol never answers Alice. The information might not be reliable, i.e. Carol might give an unfair opinion of Bob [Jøsang et al. 2007]. Additionally Bob can create multiple virtual identities to provide deceptively high rating scores for himself. In order to address reputation attacks some researchers [Sänger et al. 2016] have proposed an enhanced presentation of reputation data. Their interactive visualization increased a participant's ability to detect and understand malicious seller behavior in e-commerce. While the approach was efficient for participants with less experience, the additional information also distracted some participants.

Reputation systems are also vulnerable to the *playbook* attack [Jøsang and Golbeck 2009] where a service provider provides bad service only to a subset of participants, gaining both revenue and reputation score at the expense of a few unhappy participants. In this case, the assumption that a participant behaves identically with all participants is clearly wrong. Nevertheless, in the popular averaging reputation system [Jøsang et al. 2007], all behaviors have the same weight regardless of the surrounding context. There is not yet an effective technique to deal with the playbook strategy [Sun and Ku 2014]. More generally, the psychometric foundations of reputation scores are unclear, such as the treatment of variability or the weighting of recent information, especially betrayal.

Furthermore, reputation scores lack personalization. According to Wang and Vassileva [2007] a personalized score is required in the presence of subjective factors, i.e. user needs or interests. A *personal trust* scoring system can accommodate subjectivity. Such a trust score is ideally calculated and attached to a participant. Because the trust score reflects personal experience between pairs of users, the playbook attack is not possible. A user can compute the trust score of a partner locally without querying information from third parties. Crucially, with effective trust scoring, participants do not need to recall anything. Continued, context sensitive interaction therefore proceeds with limited cognitive demand.

We employed a dynamic trust function that calculates participant trust values based on behavioral history [Dang and Ignat 2016]. We examined the effect of presenting this trust metric and ID on trust behavior, with a post-hoc analysis of reputation as a predictor. For this purpose, we adapted the *trust game* [Berg et al. 1995], a money exchange game that is widely used in economics to study human trust behavior [Johnson and Mislin 2011; Lewicki and Brinsfield 2015]. We particularly investigated whether the availability of either trust scores or ID improves user cooperation. The set of analyses and results of our study are available at https://github.com/coast-team/trust_influence_analysis.

2 TRUST GAME

We employed the trust game as an analogue for the exchanges between pairs of interdependent participants in distributed work, where partner identity is relevant to the assessment of partner behavior. Numerous studies documented behavior in the trust game context. This provides experimental design and performance standards and allows us to associate observed behavior with our specific manipulations and/or dismiss idiosyncrasies as non-influential. In this section we describe the trust game and give an overview of several studies of cooperation that employed the trust game.

Berg et al. [1995] developed the trust game, or the "investment game", to study economic reciprocity. Participants are organized in pairs. For each pair, one participant is assigned the role

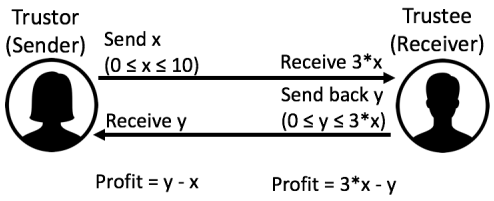

Fig. 1. One trial trust game.

of “sender" while the other is assigned the role of “receiver". The two roles are sometimes called “trustor" and "trustee" respectively. As shown in Figure 1, initially the sender sends an integer amount between 0 and 10 units to the receiver. The receiver gains three times the amount sent. For instance, if the sender sent 7 money units, the receiver will gain $3 * 7 = 21$ units. Subsequently, the receiver can select an amount between 0 and the gained amount (in this case, 21) to return to the sender. However, the returned amount is not further multiplied. Suppose the receiver returned 11. The final payoff to the sender is 11 units, and the payoff to the receiver is $21 - 11 = 10$ units.

The game pits joint payoff against individual payoff. Joint payoff is maximized if the sender sends 10 to the receiver, so the total profit is 20, calculated as the total payoff of 30 minus the initial endowment of 10 to the sender. However, assuming that participants only seek to maximize their own profit, normative game theory predicts that the sender will send 0 and upon receiving any sum the receiver will send back 0. Any amount other than 0 will reduce the receiver's profit. According to normative theory for the one-round trust game, the sender knows this fact, so at her turn, she should send 0 to the receiver. If she sends any greater amount, she must assume that she will not receive anything back [Camerer 2003].

In fact, participants do not behave according to normative theory, but choose to maximize their joint profit. The trust game is therefore considered to be *cooperative* [Balliet and Van Lange 2013; Cesarini et al. 2008], or said differently, increases in payoff reflect cooperation. Researchers interpret the initial sending amount as an expression of sender trust in the receiver [Glaeser et al. 2000]. Receiver response is an expression of trustworthiness [Fehr et al. 2002]. Brülhart and Usunier [2012] equate trust game behavior as a measure of trust and trustworthiness.

Repetition of the exchange enhances cooperation [Cochard et al. 2004]. In the repeated game, trust potentially accrues as both players react to their partner's past behavior in subsequent rounds. To our knowledge, the theoretical analysis for participant behavior in the infinitely repeated trust game is still an open question [Breitmoser 2015; Bruttel and Kamecke 2012]. However, we note that the two players experience consequence at different times. The sender receives immediate feedback regarding his trust in the form of the receiver's response. The receiver on the other hand will not incur consequence to his trustworthiness until a subsequent exchange with this sender. Studies of the repeated trust game [Engle-Warnick and Slonim 2004] document a decline in cooperation towards the end of a session with known length.

Some research work examines the influence of reputation in the trust game or public good games. Huck et al. [2012]; Semmann et al. [2004] associate reputation with user identity, showing that cooperation declines when individual identities switch from being recognizable to being unrecognizable. Bente et al. [2014] tested the influence of avatar and reputation levels on buyer decisions, i.e. senders in the trust game. Both reputation scores and avatars can encourage the investment decision of buyers. However, the authors did not study the behavior of sellers, i.e. receivers in the trust game. Moreover, reputation scores were artificial rather than computed from real behavior. Keser [2003] represents user reputation as a set of trinary ratings ("positive",

"negative" or "neutral") manually assigned by senders to rate the behavior of receivers in the trust game. These ratings were presented to subsequent senders before they made their decisions. The study found that reputation information significantly increases the overall cooperation levels of the game. A follow-up experiment [Boero et al. 2008] employed three games: in the first one, similar to Keser [2003], senders could rate receivers; in the second game receivers could rate senders and in the third one both senders and receivers could rate their partners. The study confirmed the findings of Keser [2003] regarding senders and reported on similar results regarding receivers. The study found that a bidirectional reputation scheme does not perform necessarily better than a single-way reputation scheme. However, in both studies Keser [2003] and Boero et al. [2008] reputation levels were manually assigned by participants, who had to examine all previous partner reputation levels before making a decision.

Our view is that general reputation is not always available, and that a participant is only aware of his own experience with specific partners but is not aware of, or does not necessarily want to rely on, other users' experience with those partners. In particular, we examined the influence of an available *trust score* on participant behavior, controlling for the availability of user ID. The trust score is automatically computed by the system based on participants behavior during the game, taking variability and misbehavior into account. There is no burden on participants to assign or calculate partner trust scores manually. We study the influence of trust score for both sender and receiver roles. We did depart from the standard repeated trust game paradigm in one notable way: to maximize power, our manipulations concerning the availability of trust score or participant name are within-subjects.

3 PRELIMINARY STUDY

We present a preliminary analysis of the predictive power of participants' future behavior in the trust game comparing trust and reputation scores.

According to Malaga [2001], reputation score (and by inference) trust score comprises a prediction about future behavior. For instance, if Alice has a high score, we could expect that she will behave well in the future. If she fails to do so, the score assigned to her is inaccurate. Below we compare the relative predictive power of trust and reputation scores.

We employed two external datasets from two repeated simple trust game experiments independently conducted by Dubois et al. [2012] and Bravo et al. [2012]. The experiment in Bravo et al. [2012] involved 108 participants and contained five rounds. The experiment in Dubois et al. [2012] involved 36 participants and contained ten rounds. Both experiments employ groups of 18 participants. For computing trust scores we employed the trust function proposed in Dang and Ignat [2016], shown to reflect and predict participants' behavior in the repeated trust game, with resistance to fluctuating participant behavior. As a reputation measure we used participants' average sending proportion up to the moment reputation is computed, which is similar to many real-world reputation scoring methods [Jøsang et al. 2007; Tavakolifard and Almeroth 2012].

We conducted one regression analysis using the trust score computed by our trust function as a predictor and with observed sending proportion as the criterion. Starting with round 4 when the trust metric has stabilized, we predicted the send proportion of participants using the trust score calculated after the previous round. We employed a similar regression analysis with reputation score as a predictor and sending proportion as the criterion. The results for the sender role appear in Table 1 and for the receiver, in Table 2. The corresponding t-values for the trust value assigned to each participant in predicting their future behavior are all significant for both senders and receivers, i.e. the trust score calculated by our trust function is predictive for external datasets. Moreover, adjusted R^2 values are higher for predictive models using trust values than for reputation values in

Dataset	df	t-value for trust	Adj. R^2 for trust	t-value for reputation	Adj. R^2 for reputation
Bravo dataset (round 4)	106	7.85^{***}	0.36	3.05^{**}	0.19
Bravo dataset (round 5)	106	10.0^{***}	0.48	8.86^{***}	0.42
Dubois dataset (round 4)	34	4.41^{***}	0.35	3.24^{**}	0.21
Dubois dataset (round 5)	34	4.51^{***}	0.36	2.84^{**}	0.17
Dubois dataset (round 6)	34	4.68^{***}	0.37	4.26^{***}	0.32
Dubois dataset (round 7)	*34*	*4.05^{***}*	*0.31*	*4.29^{***}*	*0.33*
Dubois dataset (round 8)	*34*	*4.15^{***}*	*0.32*	*4.83^{***}*	*0.39*
Dubois dataset (round 9)	34	4.25^{***}	0.33	3.17^{**}	0.21
Dubois dataset (round 10)	34	4.52^{***}	0.36	2.36^{*}	0.11

Table 1. Regression analysis of our trust function and reputation applied on external datasets for sender role. Italicized entries have a higher or equal Adj. R^2 for reputation than for trust. '*' $p < 0.05$, '**' $p < 0.01$, '***' $p < 0.001$.

all cases except for round 7 and 8 for senders in the Dubois dataset and equal in round 4 of the Bravo dataset for receivers.

This preliminary analysis provides compelling evidence for the predictive power of trust scores. The trust model requires *less raw information* (only information observed by the user) than the reputation model, which requires complete information from all users. These are somewhat surprising findings given that none of these participants were aware of their partners. In fact, the trust function uses less raw data but has more contextual parameters than reputation, accounting for partner, cumulative behavior over time, and punishment of misbehavior. In the next sections we present our research questions and experimental design for demonstrating the influence of trust scores on user cooperative behavior and how our trust metric predicts behavior.

4 RESEARCH QUESTIONS

We study how the availability of partner trust score and ID impacts participant behavior and the appropriateness of the used trust metric for computing trust scores in the repeated trust game. We grouped our research questions as follows:

Dataset	df	t-value for trust	Adj. R^2 for trust	t-value for reputation	Adj. R^2 for reputation
Bravo dataset (round 4)	*93*	*4.72^{***}*	*0.18*	*4.71^{***}*	*0.18*
Bravo dataset (round 5)	64	5.04^{***}	0.27	4.61^{***}	0.24
Dubois dataset (round 4)	30	3.84^{***}	0.31	3.15^{**}	0.22
Dubois dataset (round 5)	31	4.58^{***}	0.35	2.95^{**}	0.19
Dubois dataset (round 6)	31	6.06^{***}	0.53	2.20^{*}	0.11
Dubois dataset (round 7)	29	6.52^{***}	0.58	2.93^{**}	0.20
Dubois dataset (round 8)	30	6.69^{***}	0.64	4.88^{***}	0.42
Dubois dataset (round 9)	26	3.86^{***}	0.34	1.59	0.05
Dubois dataset (round 10)	27	4.88^{***}	0.45	4.38^{***}	0.39

Table 2. Regression analysis of our trust function and reputation applied on external datasets for receiver role. Italicized entries have a higher or equal Adj. R^2 for reputation than for trust. '*' $p < 0.05$, '**' $p < 0.01$, '***' $p < 0.001$.

RQ1 *Does showing partner trust score or ID change user cooperative behavior?* If so, is there a significant difference in cooperative user behavior with only trust scores relative to ID only? Is there a significant difference in user cooperative behavior resulting from the availability of both trust score and ID compared to the availability of only one of these two features? Does cooperative behavior change over time?

RQ2 *Does the trust calculation predict participant's future behavior? Do participants follow the guidance of the trust calculation?*

As senders and receivers have two different roles and may behave differently, we analyze these research questions separately from both the senders' and receivers' points of view.

5 METHODS

5.1 Participants

Participants were recruited through a public announcement. Five independent groups of six participants resulted in a total 30 of participants. Four of the five groups included one female participant, while the fifth group included two female participants. The ages of participants ranged from 19 to 45 with an average age of 28.5.

Typically researchers compensate participants using an exchange rate between virtual money in the experiment and real money, then pay the participants an amount based on how much they earned during the experiment. To assure continuing incentive throughout the session, each person who participated received a coupon of ten euros, but the person who earned most, i.e. who had the highest payoff among other people in the group, received an additional coupon of ten euros.

5.2 Task

In each game a participant played at least 25 rounds with the other five partners in the group in a random order, namely five rounds with each of these partners where she served as sender and receiver equally often. At the beginning of the first game each participant received 10 money units. In each round, the sender moved first. She knew how much money she had, and had to decide the amount she wanted to send to the receiver. After that, the receiver received a message indicating how much she had at the beginning of this round, how much she received from the sender, and how much she will have after having received. Then, the receiver decided how much she wanted to return.

5.3 Independent Variables

We crossed the availability of ID and partner trust scores to create four different games as shown in Table 3. IDs, such as "Mr. Black" or "Mrs. Green", were assigned to participants, fixed during a game and varied between games. Trust scores were calculated as in Dang and Ignat [2016] (see

		ID presented	
		False	**True**
Trust presented	**False**	**Simple Game**: The trust game when participants are given no information about partners	**Identity Game**: The trust game when participants are given only partner ID
	True	**Score Game**: The trust game when participants are given only trust scores of partners	**Combined Game**: The trust game when participants are given both trust the scores and ID of their partners

Table 3. Game descriptions

Appendix A). Trust scores were always calculated for each participant in a pair, but only displayed according to experimental condition and only partner scores were available. The theoretical trust score value ranges from 0.0 to 1.0 inclusive, presented when available with two significant digits. Participants started with the neutral value of 0.5 [Abbass et al. 2016].

We calculated user reputation score as distinct from trust score by averaging all previous sending proportion amounts of that user in both roles sender or receiver.

5.4 Design

The experimental conditions were organized as a split-plot factorial with group as a between subjects factor and Show-ID and Show-Trust as within subjects, such that each group of six participants participated in the set of four randomly ordered games. In each round, participants were paired randomly within their group and assigned randomly the sender or receiver role. We ensured that within each game, a participant was paired with a particular other participant at least five times.

5.5 Dependent Measures

The four dependent measures used in our study are: **sending proportion by senders**, **sending proportion by receivers**, **average sending proportion by senders** and **average sending proportion by receivers**.

Sending proportion by senders is the net amount the sender sends to the receiver over 10, which is the maximum amount the sender could send.

Sending proportion by receivers is the net amount the receiver sends back over the amount she received after being tripled.

Other studies [Burks et al. 2003; Dubois et al. 2012] also used sending proportion measures in order to normalize the sending behavior of receivers for comparison. For example, sender A sent 6 to receiver B, and B sent back 9 to A. In this round, the net sending amount of A and B are 6 and 9 respectively, the sending proportion of A is $6/10 = 0.6$ and the sending proportion of B is $9/18 = 0.5$.

Consistent with Burks et al. [2003]; Dubois et al. [2012] for all analyses of receiver behavior, we eliminated the zero transaction between the sender and the receivers, (i.e. the sender sends 0 and the receiver is obliged to send 0), for two reasons. First, receiver behavior is completely determined by the sender, so that the receiver's behavior is not informative. Moreover, in this case, the sending proportion for the receiver (0 divided by 0) is not calculable. We note that the zero-sending amount is retained in the analysis of sender behavior.

For the sender, there are exactly 375 sending proportion data points in each game (25/2 senders $\times$6 players in a group $\times$5 groups). For receiver, the number of sending proportion data points varies between 250 and 340 due to the elimination of the zero transaction.

Average sending proportion by senders is the average of sending proportions by each sender over all trials in the game. Taking an average distributes the effect of the zero transaction and also eliminates trial as a repeated factor in analysis.

Average sending proportion by receivers is the average sending proportion the receiver sends back to the sender over all trials in the game, without the zero transaction case.

There are 30 average sending proportion data points corresponding to 30 participants, for both sender and receiver. For the receiver, the zero-transaction data is removed before calculating the means.

5.6 Procedure

All groups participated independently using z-Tree [Fischbacher 2007] hosted on our laboratory computers. At the beginning of each session, all participants read the instructions presenting the

purpose of the experiment, a short description of the four games, the payment procedure and some example screenshots illustrating the interaction of users with the z-Tree tool. Instructions informed participants that they would play the games in an arbitrary order. For each of the games participants were told what partner information would be displayed during each interaction: for the Simple Game no information, for the Identity Game the partner identity in the form of an ID, for the Score Game a partner trust score computed according to her behaviour in previous interactions (without any details about the metric) and for the Combined Game, the partner identity and trust score. Participants did not know the number of rounds they would play in each game. After confirming that they had read and understood the instructions, participants reviewed and signed an informed consent form prior to commencing the experiment. Participants sat in different rooms to avoid any communication during the experiment. Each participant used a computer running our z-Tree application. All senders in the group finished their decision making process before proceeding to the next trial. Play then waited for every receiver to respond before starting a new round. This eliminated response time cues as an indication of player identity. No other means of communication or identification were available. Participants were informed of their cumulative earnings at each round. It was possible to play with a negative balance but this never occurred.

The repeated measures design resulted in 100 rounds across the four games. A session usually lasted two hours. At the end of the experiment participants filled out a questionnaire regarding general information such as university major and game preference.

6 RESULTS

We organize our results into two main subsections: sender behavior and receiver behavior.

6.1 Sender Behavior

The following analyses address how the sender (trustor) responded to our manipulations. We demonstrate that both trust score and ID increase sending generosity with equivalent improvement and no combined effect. To examine cooperation, we study the 0 exchange condition and rule-out round effects as influential for all games except the Simple Game with no partner information. Finally, we illustrate the dependence of performance on trust score metrics.

6.1.1 Omnibus ANOVA. A basic ANOVA with Subject, Show-Trust and Show-ID as predictors reveals an interaction, $F(1,29) = 19.36$, $p < 0.001$ as measured by average proportion sent for each game. The interaction between the availability of trust score and ID on average sending proportion for senders appears in Figure 2. We note that showing either trust score or ID improves sending proportion but showing both partner information sources does not change the sent proportion relative to one source, which suggests the need for paired comparisons between games. Johnson and Mislin [2011] claimed that in large-scale the send proportion of users in trust game follows the normal distribution. Therefore we present paired t-test based confidence intervals (yoking

Game in Comparison with Simple Game	95% confidence interval	Df
Identity Game	(-0.32, -0.14)	29
Score Game	(-0.35, -0.13)	29
Combined Game	(-0.35, -0.14)	29

Table 4. Paired t-based confidence intervals for senders' average sending proportion in the Simple Game compared to other games. The negative signs indicate that the sending amount of participants in Simple Game is less than the sending amount of these participants in other games.

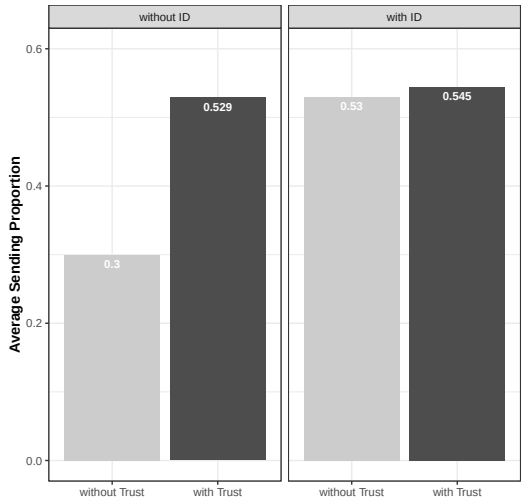

Fig. 2. Interaction between trust score and ID availability for sender.

by sender ID) in Table 4 to examine differences between the Simple Game and any other tested game, demonstrating that either trust score or ID increases sending amounts with no additive effect. The differences between the other three games (Identity, Score and Combined Games) are not significant, i.e. $p > 0.10$. To rule out any possible difference between sender performance with Show-ID and Show-Trust, we followed up with a paired t-test, yoking the results from the Identity Game and the Score Game for each sender-receiver pair for each trial $t(266) = -0.175$, $p > 0.10$. We conclude $IdentityGame \approx ScoreGame \approx CombinedGame > SimpleGame$ for average sending proportion.

6.1.2 Cooperative Behavior. Below we address the claim that providing identification or trust score controls cooperative behavior, explaining the above results. We consider the cases of non cooperation where senders send 0, the change in trust scores over time and the dependence of sending behavior on trust score values.

The percentage of times that a sender sends 0 in Simple Game, Identity Game, Score Game and Combined Game are 33.3%, 9.3%, 13.6% and 12.7% respectively. A logistic regression on the frequency of 0 transactions for all rounds with sending participant, Show-Trust and Show-ID as predictors indicates an interaction between Show-Trust and Show-ID $z = 5.607$, $p < 0.001$. Senders are more likely to send 0 in the Simple Game.

To examine the potential change in sending behavior over round, we regressed sending behavior on participant ID to remove general participant effects that would contaminate a regression analysis. We then used the resulting residuals as the criterion in a regression with round number as the predictor, reducing the df in the error term due to the prior regression. The only game with a significant round effect was the game with no information (Simple Game), revealing decreasing cooperation over time $F(1,116) = 7.3$, $p < 0.01$. No other game indicated a round effect: Identity Game, $F(1,114) = 0.05$, $p > 0.10$, Score Game $F(1,115) = 0.42$, $p > 0.10$ and Combined Game $F(1,116) = 0.008$, $p > 0.10$. Partner information eliminates decreasing cooperation over time and end game effects for senders.

	without Trust		with Trust	
	without ID (Simple)	**with ID** (Identity)	**without ID** (Score)	**with ID** (Combine)
Own trust	12.80***	9.31***	7.36***	8.33***
Partner trust	1.65	1.73	5.69***	4.69***
Adjusted R^2	0.85	0.75	0.88	0.89
F(2,27)	86.03	43.57	106.9	117.1

Table 5. Trust regression analysis for average sending behavior of senders. The table reports on t(27) values. '*' $p < 0.05$, '**' $p < 0.01$, '***' $p < 0.001$.

Finally, in Table 5 we present regression analyses between average sending behavior as the criterion with sender trust values and participant trust values as predictors. Sender behavior is positively correlated with his own trust value for all games. The trust function predicts sender behavior well. Moreover, when partner trust is available, it controls sending behavior. Notably, this is the only analysis suggesting any difference between the availability of partner identity and the trust score, as partner trust score does not predict sending behavior in games without a trust score. We conclude that partner trust score availability controls cooperation. We also note the relatively high adjusted R^2 for the Simple Game. We attribute this to range restriction on trust score values that eliminates non-linear influences at higher levels of trust.

6.1.3 Summary of Sender Behavior. Senders are less cooperative in the Simple Game than all other games. Decreasing cooperation in the form of round effects only appears in the Simple Game. Good models for sending behavior show predictive effects of own trust in all conditions, and partner trust when trust scores are available. The availability of partner trust score therefore controls sending behavior.

6.2 Receiver Behavior

The following analyses address how the receiver (trustee) responded to our manipulations. We demonstrate that both trust score and ID increase generosity with equivalent improvement and no combined effect. To examine cooperation, we study the 0 exchange condition when the receiver received a positive amount from the sender but decided to send back 0. We rule-out round effects and examine the dependence of performance on trust score metrics.

6.2.1 Omnibus ANOVA. A basic ANOVA with Subject, Show-Trust and Show-ID as predictors reveals an interaction, $F(1,29) = 14.36$, $p < 0.001$ as measured by average sending proportion. The interaction between the availability of trust score and ID on average sending proportion appears in Figure 3. We note that showing either trust score or ID improves receiver return proportions, but showing both partner information sources does not change the sent amount

Game in Comparison with Simple Game	95% confidence interval	Df
Identity Game	(-0.23, -0.10)	29
Score Game	(-0.25, -0.08)	29
Combine Game	(-0.26, -0.11)	29

Table 6. Paired t-test confidence intervals for receivers' average sending proportion in Simple Game compared to other games. The negative signs indicate that the sending amount of participants in Simple Game is less than the sending amount of these participants in other games.

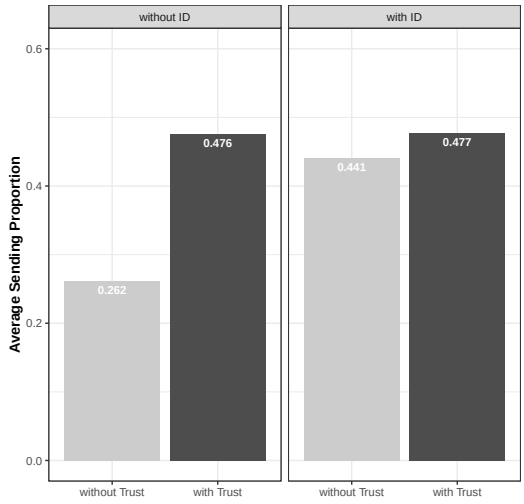

Fig. 3. Interaction between trust score and ID availability for receiver.

relative to one source which suggests the need for paired comparisons between games. As above, and consistent with Johnson and Mislin [2011] we assume that the sending proportion of receivers follows the normal distribution in large-scale. We used paired-t based confidence intervals (yoking by receiver ID) in Table 6 to examine differences between the Simple Game and any other tested game. Showing either trust score or ID increases the amount sent back with no additive effect. To rule out any possible difference between receiver performance with Show-ID and Show-Trust, we followed up with a paired t-test yoking the results from the Identity Game and the Score Game for each receiver-sender pair for each trial. The results of the paired t-test, i.e. $t(219) = -0.458$, $p > 0.10$ confirmed the absence of difference between Identity Game and Score Game. We conclude $CombinedGame \approx ScoreGame \approx IdentityGame > SimpleGame$ for receiving behavior.

6.2.2 Cooperative Behavior. Below we address the claim that providing identification or trust score increases cooperative behavior, explaining the above results. We consider the cases of sending 0, the change in trust scores over time and the dependence of receiver behavior on trust score values.

The percentage of times that a receiver sends 0 in Simple Game, Identity Game, Score Game and Combine Game are 36.8%, 8.5%, 8.3% and 4.5% respectively. A logistic regression on the frequency of 0 transactions for all trials with sending participant, Show-Trust and Show-ID as predictors indicates an interaction between Show-Trust and Show-ID $z = 3.68$, $p < 0.01$. Receivers are more likely to return 0 in the Simple Game.

To examine the potential change in receiver behavior over round, we regressed receiver behavior on participant ID to remove general participant effects that would contaminate a regression analysis. We then used the resulting residuals as the criterion in a regression with round number as the predictor, reducing the df in the error term due to the prior regression. Round is not significant for any game: Simple Game $F(1,100) = 0.052$, $p > 0.10$, Identity Game, $F(1,114) = 1.44$, $p > 0.10$, Score Game $F(1,108) = 0.019$, $p > 0.10$ and Combined Game $F(1,110) = 0.027$, $p > 0.10$. Participant information therefore has no effect on the prevention of end-game effects, which do not exist.

Finally, in Table 7 we present regression analyses between average sending behavior as the criterion with sender trust values, participant trust values and amount received from the sender as

	without Trust		with Trust	
	no ID (Simple)	**with ID** (Identity)	**no ID** (Score)	**with ID** (Combined)
Own trust	6.003***	8.936***	4.617***	3.927***
Partner trust	0.687	0.978	0.237	-2.158*
Partner sending amount	-2.214*	-1.849	-1.469	0.587
Adjusted R^2	0.565	0.746	0.415	0.494
F(3,26)	13.53	29.36	7.854	10.44

Table 7. Trust regression analysis for average sending behavior of receivers. The table reports on t(26) values.'*' $p < 0.05$, '**' $p < 0.01$, '***' $p < 0.001$.

predictors. Receiver behavior is positively correlated with his own trust value for all games. This confirms our ability to predict receiver cooperation (i.e., receiver trustworthiness) from past trust values. However, receiver behavior is only related to partner trust in the Combined Game. Moreover, model fits are not as good for receivers as they are for senders. We have explored models that include interactions between amount received and trust values. These often improve the relatively smaller adjusted R^2 we obtain for receiver behavior. Such models suggest the need for different trust functions for sender and receiver, to accommodate the asymmetry in their relationship.

6.2.3 Summary of Receiver Behavior. Receivers are less cooperative in the Simple Game than all other games. There is no evidence of round effects in any game. Fair models for returning behavior show predictive effects of own trust in all conditions confirming our trustworthiness predictions. However, partner trust is only predictive in the combined game.

7 EXPERIMENTAL DESIGN ISSUES

In this section we investigate the properties of our experiment, comparing our results with other trust game experiments, evaluating the accuracy of our trust function, and addressing repeated measures concerns such as the nesting of participants in groups.

7.1 Comparison with other trust game data sets

We compared the average sending proportions of participants in our Simple Game (30 data points) with two external datasets from Dubois et al. [2012] with 36 data points and Bravo et al. [2012] with 108 data points. Table 8 shows Welch two-sample t-test values comparing our results in the simple game to their results, assuming unequal variances. None of the comparisons are statistically significant. The observed behavior in the simple game in our experimental design is consistent with other experiments. The findings are illustrated in Figure 4.

7.2 Trust function analysis

In the previous sections, we demonstrated that showing the trust score improves cooperation, but how good is the trust function? We provide two forms of support for the quality of the trust

	Dubois et al. [2012]	**Bravo et al. [2012]**
Sender	t(61.6) = -1.33	t(45.3) = -0.991
Receiver	t(55.9) = 1.69	t(45.6) = -0.598

Table 8. Welch two-sample t values between our Simple Game average send proportion data with two external datasets.

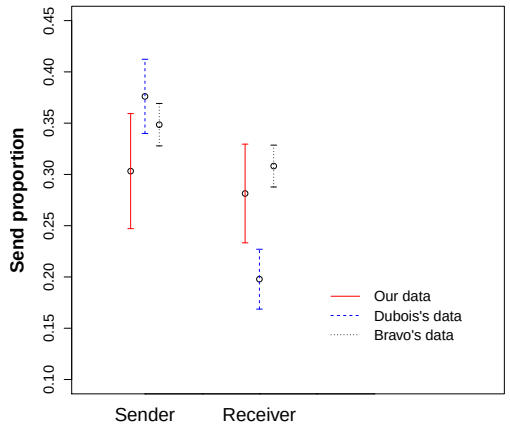

Fig. 4. Average values and standard errors of users' sending proportions in three datasets.

function: detailed prediction of participant behavior in our experiment and prediction of participant behavior using calculated reputation.

7.2.1 Predicting behavior in our experiment. The trust score models participant behavior, even when, as in Simple and Identity Games, the trust score is not made available to participants. Thus participant behavior should correlate with their own trust scores [Dang and Ignat 2016]. In the games with presented trust scores (Score and Combine Games), participant behavior should appear to react to partner trust values. The R^2 values in Tables 5 and 7 provide some evidence of prediction accuracy, although we noted less satisfactory models for receivers, and less evidence for the relevance of partner trust values in receiver behavior. Here we rule out interactions between trust values themselves as better behavior predictors. We also examine correlations between behavior and trust scores separately for rounds 4 and 5 when trust scores have sufficient data to stabilize.

Regressions of sender behavior, i.e. average sending proportion, on the interaction of sender and receiver trust values in the presence of both predictors as main effects provide no evidence of interaction effects in any game: Score Game $t(26) = 1.079$, $p > 0.1$, Combined Game $t(26) = 0.022$, $p > 0.1$, Simple Game $t(26) = -0.352$, $p > 0.1$ nor Identity Game $t(26) = 0.725$, $p > 0.1$.

Regressions of receiver behavior, i.e., average return proportion, on the interaction of sender and receiver trust values in the presence of both predictors as main effects provide no evidence of interaction effects in any game: Score Game $t(26) = -0.122$, $p > 0.1$, Combined Game $t(26) = -0.776$, $p > 0.1$, Simple Game $t(26) = 0.706$, $p > 0.1$ nor Combined Game $t(26) = 0.080$, $p > 0.1$. Adding interactions between trust predictors does not improve our models.

To further examine the predictive capability of the trust function, we performed separate multiple regression analyses for each game, for rounds 4 and 5 when trust scores have accrued sufficient data. The criterion variable is the sending proportion of the participants to their partners. Table 9 provides the results of a regression of the senders sending proportion on a model with her trust value and the trust value of her partner for both rounds. In all cases, the sender's trust value predicts sending behavior. Moreover, the partner's trust value also predicts sending behavior in the presence of ID or trust score information, confirming sender attention to these sources. Adjusted R^2 values range from 0.26 to 0.70, with lower values resulting from the game with no information.

Table 10 provides comparable information for receiver behavior, answering the question of how well we can predict whether a participant is trustworthy. These regression models included own trust value, partner trust value and the amount just received (i.e., three times the amount sent).

	without Trust		with Trust	
	no ID (Simple)	**with ID** (Identity)	**no ID** (Score)	**with ID** (Combine)
Round 4	df = 72	df = 72	df = 72	df = 72
Own trust value	6.46***	5.80***	3.89***	7.28***
Partner's trust value	0.67	3.24**	6.98***	4.41***
Adj. R^2	0.36***	0.40***	0.66***	0.70***
Round 5	df = 72	df = 72	df = 72	df = 72
Own trust value	4.87***	7.13***	3.19**	7.11***
Partner's trust value	1.16	4.54***	7.38***	3.52***
Adj. R^2	0.26***	0.55***	0.67***	0.70***

Table 9. Trust regression analysis on senders' sending proportion with t-values for individual slope tests. '*' $p < 0.05$, '**' $p < 0.01$, '***' $p < 0.001$.

While receivers were never aware of their own trust values, our trust function is a good predictor of receiver behavior *when trust score is not provided.* This does support our claim that the trust function is a good predictor of trustworthiness. However, the mere presence of trust scores in the trust score conditions dampens its predictive capability. Partner trust value is rarely predictive. Receivers did not rely on this systematically. Adjusted R^2values range from 0.08 to 0.45 with higher values in the conditions where trust score is *not* provided.

7.2.2 Post-hoc Reputation Analysis. We present a post-hoc analysis to compare the predictive power of participants future behavior between trust and reputation scores.

In our analyses presented in Tables 11 and 12 we substituted calculated reputation predictors for trust predictors, using average sending proportion as the criterion. These models differ from those in Tables 5 and 7 by the absence of own-score predictors. These reduced models were necessary because of the close relationship between average reputation and average sending amount. However, the absence of own-values does inflate the error term. As in Table 5, in Table 11 partner reputation values predict sender behavior when trust values are shown. As measured by Adjusted R^2, the resulting models of sender behavior with trust predictors are better than models with reputation

	without Trust		with Trust	
	no ID (Simple)	**with ID** (Identity)	**no ID** (Score)	**with ID** (Combine)
Round 4	df = 42	df = 62	df = 60	df = 60
Own trust value	3.41**	7.21***	1.98	1.76
Partner's trust value	0.02	1.40	1.63	0.50
Amount received	-0.53	-1.62	-2.37*	0.33
Adj. R^2	0.18*	0.45***	0.08	0.10*
Round 5	df = 39	df = 61	df = 61	df = 60
Own trust value	4.21***	3.56***	3.06**	1.09
Partner's trust value	0.14	2.10*	0.74	1.53
Amount received	-2.19*	0.06	-1.75	-0.16
Adj. R^2	0.30***	0.29***	0.13*	0.09*

Table 10. Trust regression analysis on receivers' sending proportion with t-values for individual slope tests. '*' $p < 0.05$, '**' $p < 0.01$, '***' $p < 0.001$.

	without Trust		with Trust	
	no ID (Simple)	**with ID** (Identity)	**no ID** (Score)	**with ID** (Combine)
Trust predictors				
Partner trust	1.09	0.33	7.42***	6.92***
Adjusted R^2	0.007	-0.03	0.65	0.62
F(1,28)	1.202	0.11	55.07***	47.86***
Reputation predictors				
Partner reputation	0.69	-1.14	4.55***	3.78***
Adjusted R^2	-0.01	0.01	0.40	0.31
F(1,28)	0.48	1.3	20.72***	14.31***

Table 11. Trust and reputation analysis for average sending proportion of senders. The table reports on t(28) values. '*' $p < .05$, '**' $p < .01$, '***' $p < .001$.

predictors. This is not surprising given that we provided trust values and not reputation values in these conditions. However, the superior fit provides further evidence that participants were attending to the trust values. Regarding receiver behavior, in Table 7 partner trust is only significant in the Combined game. In Table 12 partner reputation predicts receiver behavior for the ID game, no doubt assisted by the significant effect of partner sending amount. We note that in those cases with significant partner effects, the direction is negative regarding to the amount received. Model fits are poor. Adjusted R^2 are, however, better for trust predictors than reputation predictors for the games where trust information was present.

7.3 Group Effects

While data on the trust game are typically collected in groups, concern for group effects has received little attention in trust game analyses. Moreover, in our experiment, group is confounded with treatment order. In order to consider group effects, we conducted a three factor split-plot ANOVA with group as a between subjects effect and Show-ID and Show-Trust as within subjects effects [Keppel 1991]. If group is regarded as a random (sampled) factor, then the independent variables are properly tested against the interaction of group with the independent variables.

	without Trust		with Trust	
	no ID (Simple)	**with ID** (Identity)	**no ID** (Score)	**with ID** (Combine)
Trust predictors				
Partner trust	-0.71	-0.41	-0.26	-2.73*
Partner sending amount	-0.22	1.35	0.85	3.20*
Adjusted R^2	0.00	0.00	-0.02	0.22
F(2,27)	0.99	1.05	0.64	5.18*
Reputation predictors				
Partner reputation	-1.70	-2.72*	-0.15	-1.40
Partner sending amount	0.23	2.33*	0.45	2.07*
Adjusted R^2	0.08	0.21	-0.02	0.08
F(2,27)	2.26	4.93	0.62	2.21

Table 12. Trust and reputation analysis for average sending proportion of receivers. The table reports on t(27) values. '*' $p < .05$, '**' $p < .01$, '***' $p < .001$.

Our sole concern here therefore is the robustness of manipulation effects in a very conservative, low power test owing to the reduced df in the error term. We tested our effects considering group as a random factor, and interactions with group as an error term. Our analysis of sending behavior, as measured by relative sending proportion, withstands even this less powerful test. The omnibus test for the interaction of ID and Trust is $F(1,4) = 8.86$, $p < 0.05$. Moreover, none of the Group by Treatment interactions are significant: with Show-Trust $F(4,25) = 2.610$, $p > 0.05$, with Show-ID $F(4,25) = 1.253$, $p > 0.05$, or the interaction $F(4,25) = 2.698$, $p > 0.05$. Regarding receiver behavior, as measured by relative returned proportion, the omnibus interaction contrast just misses significance $F(1,4) = 6.966$, $p < 0.1$. These findings are best captured as two main effects: for Show-Trust $F(1,4) = 74.44$, $p < 0.001$ and for Show-ID $F(1,4) = 35.862$, $p < 0.01$). As above, none of the Group by treatment interactions are significant: with Show-Trust $F(4,25) = 0.153$, $p > 0.75$, with Show-ID $F(4,25) = 0.553$, $p > 0.75$, or the interaction $F(4,25) = 2.484$, $p > 0.05$. These analyses limit concern for group effects in general, and the game order differences confounded with group in particular.

.

8 DISCUSSION

We analyzed our research questions distinguishing between sender's and receiver's points of view.

RQ1 *Does showing partner trust score or ID change user cooperative behavior?*

We provided several forms of evidence regarding the influence of these interventions on cooperation. These include overall increases in the proportion returned and reductions in the frequency of 0 unit returns for both senders and receivers. Only the Simple Game differs from the alternatives, in paired-t tests of sending behavior and in the persistence of end-game effects for senders. Otherwise, we eliminated end game effects. Large-n, yoked dependent t-tests by round failed to reveal any difference in behavior between the availability of names and the availability of trust scores.

RQ2 *Does the trust calculation predict participants' future behavior ?*

With respect to senders, we provide excellent predictive models for average behavior. These average models always depend positively on own trust values, and on partner trust values when trust values are available. Sender behavior is also well modeled at the round level, always depending upon own trust values and on partner trust values for all games except the Simple Game. Senders are attending to the specific values shown for partners, as predictions based on reputation are not as good as predictions based on the trust values displayed. We note that the effect is not to encourage blind cooperation, but rather cooperation in response to the available information. Low partner trust scores elicit low sending amounts.

With respect to receivers, models of average return proportions behavior do depend on own-trust. This supports a claim for some ability to predict trustworthiness. Models at the round level are best when the trust score is *not* available. This unexpected result is possibly due to strategic differences in receiver behavior. Models are quite poor when own-values are removed in order to compare with reputation predictions. While receiver models did include an additional factor (partner sending amount), our general impression is that the models of receiver behavior are more complex than models of sender behavior and not yet accommodated by the trust function used. Moreover, unlike the sender, duplicitous receiver behavior is not punished until the subsequent round. These considerations suggest that the trust function should differ for sender and receiver.

We have not identified the source of leverage on the success of the trust function for senders. Relative to an average reputation calculation, we have noted three different influences: the specification of partners, the management of change over time and the treatment of variability, particularly punishment in response to non-cooperative behavior. Limitations in the receiver model highlight this claim, where the role of amount received may interact with the partner trust values in ways

that we have not yet captured. These influences cast the trust function as a psychometric issue, concerning the psychological factors that influence the response to experience.

Our preliminary study suggested the predictive power of trust scores compared to that of reputation. However, our experimental study did not include a condition with computed reputation score for the partner in order to be able to compare the influence of trust score and reputation.

The trust function used considers only the sending proportion as a parameter, but not for instance the amount sent by the partner. This trust model fits well for a sender that initiates the interaction by sending an initial amount. But the trustworthiness value associated with a receiver should depend not only on the return proportion but also on the amount received. We might consider associating a higher trustworthiness with a receiver that received 6 and returned 1 than to someone that received 30, but returned the same proportion. The receiver that received 30 obtained the maximum possible amount but did not reciprocate the granted trust. These suggestions further reinforce the need to consider the measurement of trust from a psychometric perspective, capturing the relationship between physical quantities and behavioral response.

We have demonstrated that the presence of partner information benefits cooperative behavior. The burden of recalling past experience with participants is just one justification for the use of trust values as a source of this information [Tang et al. 2013]. Compared with reputation scores, trust scores have several advantages. Reputation scores are globally computed values that are stored on a central server that is vulnerable to attack [Hoffman et al. 2009]. Trust scores are suitable for *distributed* architectures and do not require a central server. Trust scores are computed in a distributed way for each user: each member of the network locally computes trust levels of her partners. Moreover, trust scores emphasize *personal* experience and value. For instance, in reputation systems, if ten thousand participants rated a seller, the next participant does not have a high motivation to provide a rating because it will not change the average rating score of this seller. However, in trust-based systems, her impression has a great influence because the trust value is calculated for her only based on her experience.

On the other hand, as our experiment suggested, the trust score has a similar effect on cooperative behavior relative to ID. Therefore, trust scores may complement current systems that employ ID to identify users, helping users define the trustworthiness of their connections. While it is possible for participants to change their ID in on line systems, they cannot change the trust level other participants assigned to them. If a trust score is available, participants do not need to remember individuals by name, nor do they need to assess previous experience with imprecise mental calculations. Instead, they can make decisions based on their partner's current trust score.

Such a system greatly facilitates engagement with large scale collaborative networks. Our proposed solution for computing partner trust scores scales well with the number of partners. For each user u_i, where $1 \leq i \leq n$ and n is the total number of partners, the system stores m_i trust values t_{ij}, with $1 \leq j \leq m_i$, associated with the m_i partners with whom he is interacting. Each time a participant u_i interacts with another partner u_j, the trust score corresponding to that interaction is aggregated to the old trust value t_{ij}. The new aggregated value becomes the new value of t_{ij}. The time complexity of the computation of the trust score from an interaction is $O(1)$, i.e. constant. The space complexity for a participant to keep track of the trust scores of the other participants is linear with the number of participants with whom he interacts.

Regarding generalizability, significant effort remains in developing trust functions for other domains. Our claim is not that the specific function we used [Dang and Ignat 2016] is suitable for every domain, but rather that the dimensions we have identified (partner specificity, the representation of cumulative experience, and the treatment of variability) are candidates for inclusion.

9 CONCLUSIONS

We showed that trust score or ID availability could significantly improve the level of cooperation between users. We also demonstrated that the availability of a trust score has a similar impact on boosting cooperation as the availability of identities. Finally we showed that the availability of both features has no additional benefit to cooperation as the availability of only one of these features. Our study suggests that trust score could function as an enhancement or even replacement of traditional ID systems. We plan to study a closer comparison between the influence of trust score and reputation score on the collaborative behavior by designing a trust game experiment where we analyse the effect of showing partner trust score and reputation score.

10 ACKNOWLEDGMENTS

This work was partially supported by Inria Associated Team USCoast and NSF Grant 1520870.

REFERENCES

Hussein A. Abbass, Garrison W. Greenwood, and Eleni Petraki. 2016. The N-Player Trust Game and its Replicator Dynamics. *IEEE Trans. Evolutionary Computation* 20, 3 (2016), 470–474.

John R Anderson and Lynne M Reder. 1999. The fan effect: New results and new theories. *Journal of Experimental Psychology-General* 128, 2 (1999), 186–197.

Daniel Balliet and Paul AM Van Lange. 2013. Trust, conflict, and cooperation: a meta-analysis. *Psychological Bulletin* 139, 5 (2013), 1090.

Gary Bente, Thomas Dratsch, Simon Rehbach, Matthias Reyl, and Blerta Lushaj. 2014. Do You Trust My Avatar? Effects of Photo-Realistic Seller Avatars and Reputation Scores on Trust in Online Transactions. In *HCI in Business*. Springer International Publishing, Cham, 461–470.

Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 1 (1995), 122–142.

Jacqueline Z Bergman, Erika E Small, Shawn M Bergman, and Joan R Rentsch. 2010. Asymmetry in perceptions of trustworthiness: It's not you; it's me. *Negotiation and Conflict Management Research* 3, 4 (2010), 379–399.

Riccardo Boero, Giangiacomo Bravo, Marco Castellani, and Flaminio Squazzoni. 2008. *Reputation and judgment effects in repeated trust games.* Technical Report. Department of Social Sciences, University of Brescia, Working Paper SOC 01-08.

Gary E Bolton, Elena Katok, and Axel Ockenfels. 2004. How effective are online reputation mechanisms? An experimental investigation. In *Management Science*. Informs, 1587 – 1602.

Giangiacomo Bravo, Flaminio Squazzoni, and Riccardo Boero. 2012. Trust and partner selection in social networks: An experimentally grounded model. *Social Networks* 34, 4 (2012), 481–492.

Yves Breitmoser. 2015. Cooperation, but no reciprocity: Individual strategies in the repeated Prisoner's Dilemma. *The American Economic Review* 105, 9 (2015), 2882–2910.

Marius Brülhart and Jean-Claude Usunier. 2012. Does the trust game measure trust? *Econ. Letters* 115, 1 (2012), 20–23.

Lisa Bruttel and Ulrich Kamecke. 2012. Infinity in the lab. How do people play repeated games? *Theory and Decision* 72, 2 (2012), 205–219.

Stephen V Burks, Jeffrey P Carpenter, and Eric Verhoogen. 2003. Playing both roles in the trust game. *Journal of Economic Behavior & Organization* 51, 2 (2003), 195–216.

Colin Camerer. 2003. *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press.

David Cesarini, Christopher T Dawes, James H Fowler, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace. 2008. Heritability of cooperative behavior in the trust game. *Proceedings of the National Academy of sciences* 105, 10 (2008), 3721–3726.

Klarissa Chang. 2004. Transactive Memory and Trust Networks in Computer-Supported Teams. In *PACIS*. AISeL, 121.

Jin-Hee Cho, Kevin S. Chan, and Sibel Adali. 2015. A Survey on Trust Modeling. *ACM Comput. Surv.* 48, 2 (2015), 28.

Francois Cochard, Phu Nguyen Van, and Marc Willinger. 2004. Trusting behavior in a repeated investment game. *Jour. of Econ. Behavior & Organization* 55, 1 (2004), 31–44.

Quang Vinh Dang and Claudia-Lavinia Ignat. 2016. Computational Trust Model for Repeated Trust Games. In *Trustcom/BigDataSE/ISPA*. IEEE, 34–41.

Alan Dix. 2009. *Human-computer interaction.* Springer.

Dimitri Dubois, Marc Willinger, and Thierry Blayac. 2012. Does players' identification affect trust and reciprocity in the lab? *Jour. of Econ. Psychology* 33, 1 (2012), 303–317.

Jim Engle-Warnick and Robert L. Slonim. 2004. The evolution of strategies in a repeated trust game. *Jour. of Econ. Behavior & Organization* 55, 4 (2004), 553–573.

Ernst Fehr, Urs Fischbacher, Bernhard von Rosenbladt, Jürgen Schupp, and Gert G. Wagner. 2002. A nation-wide laboratory: examining trust and trustworthiness by integrating behavioral experiments into representative survey. *Schmollers Jahrbuch: Zeitschrift für Wirtschafts-und Sozialwissenschaften/ Journal of Applied Social Science Studies* 122, 4 (2002), 519–542.

Detlef Fetchenhauer and David Dunning. 2009. Do people trust too much or too little? *Journal of Economic Psychology* 30, 3 (2009), 263–276.

Urs Fischbacher. 2007. z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. economics* 10, 2 (2007), 171–178.

Edward L. Glaeser, David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter. 2000. Measuring trust. *The Quarterly Journal of Economics* 115, 3 (2000), 811–846.

Bruno W. P. Hoelz and Célia Ghedini Ralha. 2015. Towards a cognitive meta-model for adaptive trust and reputation in open multi-agent systems. *Auto. Agents and Mul.-Agent Systems* 29, 6 (2015), 1125–1156.

Kevin J. Hoffman, David Zage, and Cristina Nita-Rotaru. 2009. A survey of attack and defense techniques for reputation systems. *ACM Comput. Surv.* 42, 1 (2009), 1–31.

Steffen Huck, Gabriele K. Lünser, and Jean-Robert Tyran. 2012. Competition fosters trust. *Games and Economic Behavior* 76, 1 (2012), 195–209.

Noel D Johnson and Alexandra A Mislin. 2011. Trust games: A meta-analysis. *Journal of Economic Psychology* 32, 5 (2011), 865–889.

Audun Jøsang and Jennifer Golbeck. 2009. Challenges for robust trust and reputation systems. In *5th International Workshop on Security and Trust Management (STM 2009), Saint.*

Audun Jøsang, Roslan Ismail, and Colin Boyd. 2007. A Survey of Trust and Reputation Systems for Online Service Provision. *Decis. Support Syst.* 43, 2 (March 2007), 618–644. DOI: http://dx.doi.org/10.1016/j.dss.2005.05.019

Geoffrey Keppel. 1991. *Design and analysis: A researcher's handbook* . Prentice-Hall, Inc.

Claudia Keser. 2003. Experimental games for the design of reputation management systems. *IBM Systems Journal* 42, 3 (2003), 498–506.

Roy J. Lewicki and Chad Brinsfield. 2015. Trust research: measuring trust beliefs and behaviours. In *Handbook of research methods on trust.* Edward Elgar Publishing, 46–64.

Ross A. Malaga. 2001. Web-Based Reputation Management Systems: Problems and Suggested Solutions. *Electronic Commerce Research* 1, 4 (2001), 403–417.

General Stanley McChrystal, Tantum Collins, David Silverman, and Chris Fussell. 2015. *Team of Teams: New Rules of Engagement for a Complex World.* Penguin Publishing Group.

Riccardo Pecori. 2016. S-Kademlia: A trust and reputation method to mitigate a Sybil attack in Kademlia. *Computer Networks* 94 (2016), 205–218.

Joan R Rentsch and Richard J Klimoski. 2001. Why do 'great minds' think alike?: Antecedents of team member schema agreement. *Journal of Organizational Behavior* 22, 2 (2001), 107–120.

Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.

Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Experimental Economics* 9, 2 (01 Jun 2006), 79–101. DOI: http://dx.doi.org/10.1007/s10683-006-4309-2

Johannes Sänger, Norman Hänsch, Brian Glass, Zinaida Benenson, Robert Landwirth, and M. Angela Sasse. 2016. Look Before You Leap: Improving the Users' Ability to Detect Fraud in Electronic Marketplaces. In *CHI*. ACM, 3870–3882.

Dirk Semmann, Hans-Jürgen Krambeck, and Manfred Milinski. 2004. Strategic investment in reputation. *Behavioral Ecology and Sociobiology* 56, 3 (2004), 248–252.

Po-Ling Sun and Cheng-Yuan Ku. 2014. Review of threats on trust and reputation models. *Industrial Management and Data Systems* 114, 3 (2014), 472–483.

Steven Tadelis. 1999. What's in a Name? Reputation as a Tradeable Asset. *The American Economic Review* 89, 3 (1999), 548–563.

Jiliang Tang, Xia Hu, and Huan Liu. 2013. Social recommendation: a review. *Social Network Analysis and Mining* 3, 4 (2013), 1113–1133.

Mozhgan Tavakolifard and Kevin C. Almeroth. 2012. A Taxonomy to Express Open Challenges in Trust and Reputation Systems. *JCM* 7, 7 (2012), 538–551.

Quang Hieu Vu, Mihai Lupu, and Beng Chin Ooi. 2010. Trust and reputation. In *Peer-to-Peer Computing*. Springer, 183–214.

Yao Wang and Julita Vassileva. 2007. A Review on Trust and Reputation for Web Service Selection. In *ICDCS Workshops*. IEEE Computer Society, 25.

A TRUST SCORE CALCULATION

Separate trust scores are calculated for each player for each round, i.e. for each interaction between two players. The round number is denoted as t.

In our experiment, a user might be assigned different role in different round, i.e. in a round she can be a sender and in another round she can be a receiver. The maximum amounts she can send are different by role, which is 10 in case of sender and $3 * received\_amount$ in case of receiver. Therefore, we firstly normalize the sending amount of both roles to $send\_proportion_t$ for round t.

$$send\_proportion_t = \frac{sending\_amount_t}{maximum\_sending\_amount_t} \tag{1}$$

The zero-transaction is eliminated on the receiver's side, i.e. if the sender sends 0 to the receiver the trust score of the receiver is kept the same for the next interaction because her send proportion is 0/0, which is undefined. In this case, for this round, the trust score of the sender is updated to 0, send proportion being 0/10=0.0.

Then we calculate the trust score for a single current round t:

$$current\_trust_t = \log(send\_proportion_t \times (e-1) + 1) \tag{2}$$

Then we calculate the aggregate trust score, which is the cumulative trust score over multiple interactions.

$$\delta_t = |current\_trust_t - current\_trust_{t-1}| \tag{3}$$

$$\beta_t = c \times \delta_t + (1-c) \times \beta_{t-1} \tag{4}$$

$$\alpha_t = threshold + \frac{c \times \delta_t}{1+\beta_t} \tag{5}$$

$$aggregate\_trust_t = \alpha_t \times current\_trust_t + (1-\alpha_t) \times aggregate\_trust_{t-1} \tag{6}$$

The δ_t is the change in current trust value by two sequential interactions t and $t-1$ between two users with $current\_trust_0 = 0$. We calculated δ_t to see how much a person changes her behavior since her last activity. It is easy to prove that, α_t is bigger if δ_t is bigger, and vice versa. Therefore, if the trust of the current interaction is much different from accumulated trust of all previous interactions, the current interaction will play a more important role in the final trust value.

Now we can calculate the $trend\_factor_t$ at round t representing the recent trend of user behavior, with higher values meaning that users improved lately their behavior. $trend\_factor_t$ helps us to deal with *fluctuating* behavior, i.e. a user firstly cooperates to gain trust from partners then suddenly deviates: this kind of behavior will be punished immediately by our trust metric.

atf represents accumulated trust fluctuation. Both kinds of *fluctuating behaviors* are punished: whether the latest sending amount is suddenly higher or lower than usual behavior. However, it is obvious that the latter case is more critical than the former one. Therefore, punishment in the latter case should be greater. The accumulated trust fluctuation is a non-decreasing function. The increase depends on the change over time of the user's behavior. If the behavior is stable or it changes within the allowed range (defined by the constant ϕ), atf_t will not change. When atf_t reaches the threshold value MAX_ATF, accumulated change in user behavior over time reaches the level of betrayal and therefore $change\_rate_t$ drops to 0. Otherwise, $change\_rate_t$ decreases if atf_t increases. The cosine function is used in the formula as it has a low degradation rate in the initial stage, and a high degradation rate in the case of repeated fluctuating behavior. Therefore, if

a user begins to adopt fluctuating behavior the punishment is small, but it increases quickly when fluctuating behavior persists.

$$trend\_factor_t = \begin{cases} trend\_factor_{t-1} + \phi & \text{if } current\_trust_t - aggregate\_trust_t > \epsilon \\ trend\_factor_{t-1} - \phi & \text{if } aggregate\_trust_t - current\_trust_t > \epsilon \\ trend\_factor_{t-1} & \text{otherwise} \end{cases} \tag{7}$$

$$adj\_atf_t = \begin{cases} \frac{atf_t}{2} & \text{if } atf_t > \text{MAX\_ATF} \\ atf_t & \text{otherwise} \end{cases} \tag{8}$$

$$atf_t = \begin{cases} adj\_atf_{t-1} + \frac{(current\_trust_t - aggregate\_trust_t)}{2} & \text{if } current\_trust_t - aggregate\_trust_t > \phi \\ adj\_atf_{t-1} + (aggregate\_trust_t - current\_trust_t) & \text{if } aggregate\_trust_t - current\_trust_t > \phi \\ adj\_atf_{t-1} & \text{otherwise} \end{cases} \tag{9}$$

$$change\_rate_t = \begin{cases} 0 & \text{if } atf_t > \text{MAX\_ATF} \\ \cos\left(\frac{\pi}{2} \times \frac{atf_t}{\text{MAX\_ATF}}\right) & \text{otherwise} \end{cases} \tag{10}$$

Finally, the trust score is calculated:

$trust\_value_t = expect\_trust_t \times change\_rate_t$ where,

$expect\_trust_t = trend\_factor_t \times current\_trust_t + (1 - trend\_factor_t) \times aggregate\_trust_t$

More details about our trust function and its evaluation can be found in Dang and Ignat [2016].