Computational Trust Model for Repeated Trust Games

Quang-Vinh Dang
Université de Lorraine, LORIA, F-54506
Inria, F-54600
CNRS, LORIA, F-54506
quang-vinh.dang@inria.fr

Claudia-Lavinia Ignat
Inria, F-54600
Université de Lorraine, LORIA, F-54506
CNRS, LORIA, F-54506
claudia.ignat@inria.fr

***Abstract*—Trust game is a money exchange game that has been widely used in behavioral economics for studying trust and collaboration between humans. In this game, exchange of money is entirely attributable to the existence of trust between users. The trust game could be one-shot, i.e. the game ends after one round of money exchange, or repeated, i.e. it lasts several rounds. Predicting user behavior in the repeated trust game is of critical importance for the next movement of the partners. However, existing behavior prediction approaches uniquely rely on players personal information such as their age, gender and income and do not consider their past behavior in the game.**

In this paper, we propose a computational trust metric that is uniquely based on users past behavior and can predict the future behavior in repeated trust game. Our trust metric can distinguish between users having different behavioral profiles and is resistant to fluctuating user behavior. We validate our model by using an empirical approach against data sets collected from several trust game experiments. We show that our model is consistent with rating opinions of users, and our model can provide higher accuracy on predicting users' behavior compared with other naive models.

***Index Terms*—trust management, computational trust, non-cooperative games, behavioral game theory**

I. Introduction

Studying cooperation between users is a main topic of research in economy, psychology and computer science. Investment game or trust game proposed by Berg [1] is an important tool to study trust and collaboration [2]. The experiment has been repeated thousands of times all over the world[1]. In this game two users exchange money in anonymous mode. The sender sends an amount to the receiver. This amount is tripled when it arrives to the receiver. The receiver then selects an amount that should not exceed the sum received to send back to the sender. This time, the amount is not tripled when it arrives to the sender. The money exchange between users is entirely attributable to the existence of trust between them. The model of trust game can be applied for the cooperation between users that are located in different places that do not know each other. For instance, in e-commerce, buyers usually need to pay before they can receive the goods, so they need to trust the sellers.

[1]According to Google Scholar, the paper has been cited 3,810 times, as on 27-Jun-2016. An incomplete study counted that there are more than 23,000 participants involved in trust game experiments [3].

The trust game could be one-shot, i.e. the game ends after one round of money exchange, or repeated, i.e. it lasts several rounds [4], [5], [6]. The pairs of users could be fixed [7] or re-assigned before each round [8]. The total number of rounds in the experiment can be reported to users [9], [10] or not [6]. These games provide different kinds of partner information to players, such as their gender, age and income [11], or their past interaction history [12], [8]. Some of these games allow communication between players [9], [13] or specification of user contracts [14], [15].

Predicting partner behavior in the repeated trust game is of critical importance. In e-commerce, for instance, it is very important to predict the behavior of the partners before transaction completion. Game theory predicts that, in trust game, sender will send 0 and receiver will send back 0 [16]. However, in experimental game theory we usually do not observe this user behavior. In fact, the sending behavior of users in large-scale settings follows the normal distribution [3].

Existing approaches on predicting behavior of participants in one-shot trust game were proposed using additional information related to players, such as the personal information (age, gender, income, etc.) or the data collected through some tests / questionnaires. However, the only reliable available information for predicting users behavior is their behavior during previous transactions. No other work proposed prediction of user behavior based on analysis of previous interactions.

In this paper, we present a computational trust metric, which serves as a model for user behavior, to reflect and predict user's behavior in repeated trust games. We claim that, based on users past behavior in the repeated trust game, we can model and predict their next behavior. We only focus on repeated version of original anonymous trust game, i.e. the players have no information about their partners. We show that our model is robust to several types of attacks. We validate our model by using an empirical approach against data sets collected from several trust game experiments. We prove that our trust metric is consistent with users' opinion about partners trustworthiness.

The paper is structured as follows. We present related works in Section II. In Section III we present our trust metric and in section IV we describe its evaluation. Section V presents some

limitations of our metric. Concluding remarks are presented in Section VI.

II. Related work

Several approaches on predicting behavior of participants in one-shot trust game were proposed using additional information related to players, such as their personal information (e.g. age, gender, income) or evaluation collected through some tests or questionnaires users needed to fill in before the experiment.

Gunnthorsdottir et al. [17] used the Mach test on users personality to predict their behavior in the game. Evans et al. [18] predicted users behavior based on the results of *Propensity to Trust Survey*. Using a similar idea, Yamagishi et al. [19] defined a new model called *attitudinal trust* to predict users' behavior in one-shot trust game. The *attitudinal trust* is calculated based on data collected through a questionnaire.

Yen [20] claimed that users with higher income send more to their partners than users with lower income. Falk et al. [21] confirmed this suggestion by showing that, students tend to send less than other social groups.

Ashraf et al. [22] used the data collected in the dictator game to predict the behavior of the same users in the trust game.

In real world applications, sociometric information of users is usually not available [23]. Even if this information is available, such as a user login to a website by using Facebook account, it is not always reliable as users can declare false personal information. Our method to model and predict users' behavior does not require any additional personal data.

We are not aware of any prior work that predicted users' behavior in repeated trust game. The theoretical prediction is that in the repeated trust game participants keep exchanging an amount of 0 [16].

Glaeser et al. [24] used average value of previous sending amount as the trust measurement of users in trust game. This trust metric as average of previous sending amounts was used later on by other research works [8], [3], [25]. However, this average trust metric can not deal with user fluctuating behavior, as we discuss in Section IV.

III. Trust calculation

We define our trust metric as a measure of how well a user behaved in the past. We claim that we can predict users behavior based on the computed trust metric. For instance, a user with high trust score tends to behave well in the future. As we discuss in Section IV-C2, some users try to behave well at the beginning and then suddenly deviate. Our trust calculation will take into account this strategy that we call *fluctuate strategy*. In general, for trust games, trustworthiness of a user depends on the amount sent to her partners [8], [24]. A higher sending amount should lead to higher trustworthiness.

The trust score formula needs to satisfy the following requirements:

1) The trust value is higher if the sending amount is higher.
2) The trust value can distinguish between different types of users.
3) The trust value considers user behavior over time.
4) The trust value encourages a stable behavior rather than a fluctuating one.
5) The trust value is robust against attacks.

A. Parameters initial values

The values of the parameters used for the trust metric computation are displayed in Table I. The left side of the table contains the initial values of the corresponding parameters, while the right side of the table contains the constant values of the corresponding parameters.

TABLE I
Parameter Initial values.

α_0	0.	ϵ	0.3
β_0	0.	ϕ	0.1
atf_0	0.	MAX_ATF	2.
$expect\_trust_0$	0.	$threshold$	0.25
$trend\_factor_0$	0.	c	0.9
$current\_trust_0$	0.		
$aggregate\_trust_0$	0.		
$change\_rate_0$	0.		
$trust\_value_0$	0.5		

B. Current trust

As we described in Section I, in each round, two users interact by sending a non-negative amount. For senders, the maximum amount they can send is set to 10, and for receivers, the maximum amount they can send is the amount they received from the sender (i.e. three times of what the sender sent). For both roles, we normalize the $send\_proportion_t$ as the sending proportion of a user at round t, with $t \geq 1$:

$$send\_proportion_t = \frac{sending\_amount_t}{maximum\_sending\_amount_t} \quad (1)$$

It is obvious that $\forall t, 0 \leq send\_proportion_t \leq 1$.

We define next the trust metric for a single interaction between users that we call $current\_trust$. $current\_trust_t$ is a function of $send\_proportion_t$, meaning that the trustworthiness of a user in a single interaction depends on how much she sends to her partner in round t. We define $current\_trust_t$ as a value between 0 and 1 inclusive. This function should satisfy the following properties (for convenience, we use the notation $f(x), f : [0,1] \rightarrow [0,1]$ for the function of $current\_trust_t$, with x being $send\_proportion_t$):

- $f(x)$ is continuous in $[0,1]$.
- $f(0) = 0$, meaning that $current\_trust$ is 0 if the user sends nothing.
- $f(1) = 1$, meaning that $current\_trust$ is 1 if the user sends the maximum possible amount.
- $f'(x) > 0$ with $x \in [0,1]$, meaning that $current\_trust$ is strictly increasing when $send\_proportion$ increases from 0 to 1. $f'(x)$ denotes the derivative of function $f(x)$.

- $f''(x) \leq 0$ with $x \in [0,1]$ meaning that the function is concave, i.e. the closer to 1 the value of $current\_trust$ is, the harder is to increment it.
- $f'(x^-) = f'(x^+), \forall x \in [0,1]$, meaning that the function is smooth, i.e. there is no reason that at some point the current trust increases roughly less than previously.

We proposed the following function that satisfies the above mentioned conditions:

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

where $current\_trust_t$ is the $current\_trust$ function at round t and $send\_proportion_t$ is the value of $send\_proportion$ at round t.

Explanation about the selection of the formula 2 will be provided in Section IV-A2.

C. Aggregate Trust

$current\_trust_t$ uniquely computes the value of trust based on the current interaction t. However, the previous interactions between two users have to be taken into account. The calculation of trust for multiple interactions is inspired by the trust model SecuredTrust [26]. The main drawback of SecuredTrust is that the metric assumes the existence of $current\_trust_t$ value. However, as shown in the previous subsection, computing $current\_trust$ is not an easy task as it has to satisfy certain requirements. Moreover, SecuredTrust was mainly designed for peer-to-peer network systems, where computation of the trust in a peer node relies on information provided by the neighbours in the network. In this way, the trust value in one peer is in fact the reputation of that peer computed as an aggregation of the neighbor trust values on that peer. Nevertheless, in collaborative environments different users have different experiences with a certain user and therefore their trust values on that user are different.

Furthermore, SecuredTrust uses a constant value of α as forgetting factor. If this property can be valid in the peer-to-peer network field, it does not hold for human users. Based on psychological peak-end rule [27] we applied a dynamic α. The peak-end rule claims that, in a series of experiences, humans remember the extreme and the last experience, and forget the other ones.

We calculate $aggregate\_trust$ as follows:

$$\delta_t = |current\_trust_t - current\_trust_{t-1}| \quad (3)$$

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

$$\alpha_t = threshold + \frac{c \times \delta_t}{1+\beta_t} \quad (5)$$

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

As we described in the section III-A, $current\_trust_0 = 0$.

The δ_t is the change of current trust value by two sequential interactions t and $t-1$ between two users. 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. It means that, 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.

D. Dealing with fluctuating behavior

Some users may try to collaborate in the beginning and then suddenly betray. We added a $change\_rate_t$ variable into our model to punish this kind of activity.

First, we calculate the $trend\_factor_t$ at round t representing the recent trend of user behavior, with higher value meaning that users improved lately their behavior:

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

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

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

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

In formula 9, we present the accumulated trust fluctuation (atf) function. Both kinds of *fluctuate behaviors* are punished: the latest sending amount is suddenly higher or lower than usual behavior. However, it is obviously that the latter case is more critical than the former one. Therefore, the punishment in the latter case should be stronger.

The accumulated trust fluctuation is a non-decreasing function. The increase depends on the change over time of 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, it means that accumulated change in user behavior over time reaches the level of betrayal and therefore $change\_rate_t$ drops to 0. Otherwise, as shown by Equation 10, $change\_rate_t$ decreases if atf_t increases.

The cosine function is used in formula 10 because the cos function has a low degradation rate in the initial stage, and a high degradation rate in the case of repeated fluctuating behavior[26]. It means that, if a user starts adopting a fluctuating behavior the punishment is low, but it increases fast while fluctuating behavior persists.

Finally, the trust value after round t is calculated by:

$$trust\_value_t = expect\_trust_t \times change\_rate_t \quad (11)$$

where,

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

The trust value is updated on every round.

IV. Evaluation of trust metric

In this section, we evaluate the performance of our trust metric according to the following three aspects:

1) *Evaluation with simulated data.* We evaluate our trust metric with simulated data. More specifically, we analyse whether our trust metric can distinguish between user types and cope with a fluctuating strategy.
2) *Consistency with human opinions.* We study how our trust metric can be validated with real user data. More specifically, given the same data set, we analyse whether our trust metric provides the same ratings of user behavior as the ones manually assigned by humans.
3) *Evaluation of prediction with real data.* We study whether our trust metric can predict users future behavior. In other words, we analyse whether the trust score assigned by the trust metric to a user reflects her future behavior.

A. Evaluation with simulated data

Our trust metric should punish fluctuating user behaviors. Moreover, it should detect user behavior patterns, i.e. it should be able to distinguish different types of user profiles: low, medium and high. In this subsection, we verify that our trust metric satisfies these criteria.

1) Fluctuating user behaviors: We define three types of user profiles according to the values of $send\_proportion$: low, medium and high. Similar to [28], we define that a user with a low profile sends in average 20% of the maximum possible amount, while for a medium profile user the $send\_proportion$ is 50% and for a high profile user it is 80%. We also define a *fluctuate profile* user who first tries to behave well and then deviates.

By means of simulations for the above user profiles, we compare the behavior of our trust metric with the average trust metric where the trust score is calculated by an average of previous sending amounts [4], [14], [7], [8], [6], [24], [3]. The behavior of our trust metric in first 10 rounds is displayed in Fig. 1 and the behavior of the simple average trust metric is displayed in Fig. 2.

We can easily see that, our trust metric can cope and punish the *fluctuating* behavior very well, as it reduces the trust score of *fluctuating* user to the same as of a *low profile* user. On the other side, the simple average metric cannot distinguish between *fluctuating* and *high profile* users.

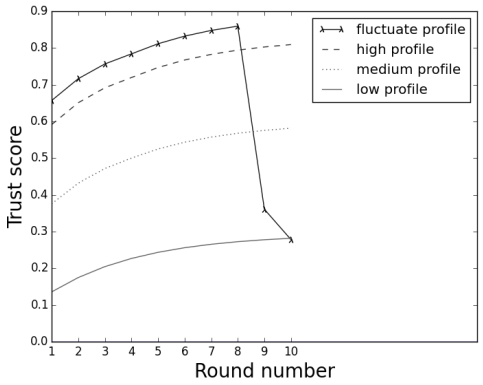

Fig. 1. Our trust metric on different user types in the first 10 interactions.

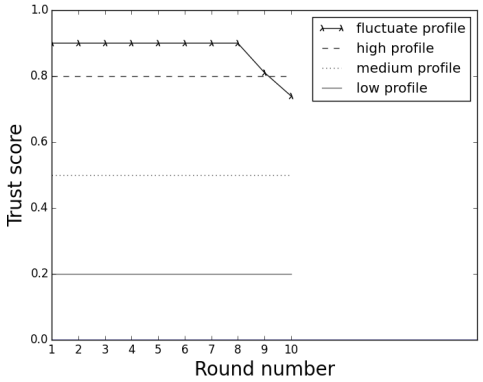

Fig. 2. The average trust metric on different user types.

2) Distinction between user types: In subsection IV-A1, we analyzed the behavior of our trust metric on constant sending behavior versus fluctuating behavior. However, the constant sending behavior is not realistic, and in this section, we relax our user profiles by allowing them to vary their behavior around the average value. In particular, we defined the behavior of *low profile*, *medium profile* and *high profile* as normal distributions with means of 0.2, 0.5 and 0.8 respectively, with standard deviation of 0.15 (this standard deviation value has been approximated from [3]). In what follows we analyze whether our trust metric can distinguish between different user types. Hence, after a large number of rounds, trust scores of different users will follow a distribution. In order to distinguish between different profiles, these distributions must satisfy the following properties:

- The trust values assigned to *fluctuating* users should be similar with the trust values assigned to *low profile* users, and should not overlap with the trust values assigned to *medium profile* users.
- The difference between two mean values should be at least the sum of two standard deviations. If we denote by $mean_{low}$, $mean_{medium}$ and $mean_{high}$ the mean values of trust scores of *bad profile*, *medium profile* and *high*

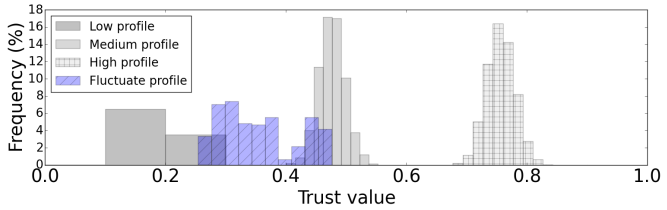

Fig. 3. Distribution of the trust metric $current\_trust = send\_proportion$ after ten rounds. The trust scores assigned to *fluctuating* users overlap with trust scores assigned to *medium profile* users.

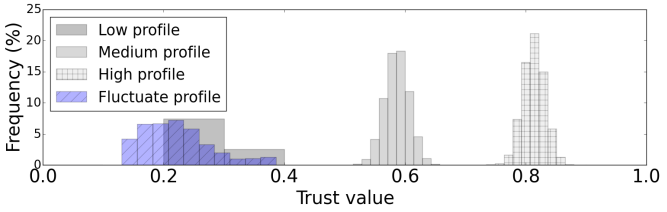

Fig. 4. Distribution of our trust metric after ten rounds. The trust scores assigned to *fluctuating* users do not overlap with trust scores assigned to *medium profile* users.

profile respectively, and by std_{low}, std_{medium} and std_{high} the corresponding standard deviations, then:

$$mean_{low} + std_{low} \leq mean_{medium} + std_{medium} \quad (12)$$

$$mean_{medium} + std_{medium} \leq mean_{high} + std_{high} \quad (13)$$

- The ratio of any two variances of these distributions should not be larger than 3, as suggested by Keppel [29].

It is not easy to find a $current\_trust$ function which can satisfy these above requirements. After an empirical process, the formula presented in Equation 2 is the only function we found so far that can satisfy these requirements.

For instance, if we replace our $current\_trust$ formula by a new formula such as $current\_trust = send\_proportion$, this trust metric will not be able to distinguish between *medium profile* and *fluctuating* users. As we show in Fig. 3, after ten rounds, the new trust metric will assign overlapping trust scores to *medium profile* and *fluctuating* users, but our metric still can distinguish between these two user profiles as displayed in Fig. 4.

B. Consistency with human opinions

In this section, we evaluate our trust metric according to user ratings obtained by an existing experimental study of the repeated trust game [30].

Keser [30] organized a repeated trust game experiment where users could rate in each round their partners' sending behavior. The three levels proposed were: negative, neutral or positive. Based on the data published in this study, we created three virtual users called *positive user*, *neutral user* and *negative user* respectively corresponding to the levels of possible ratings. These virtual users follow the average behavior of real users who have the corresponding rating.

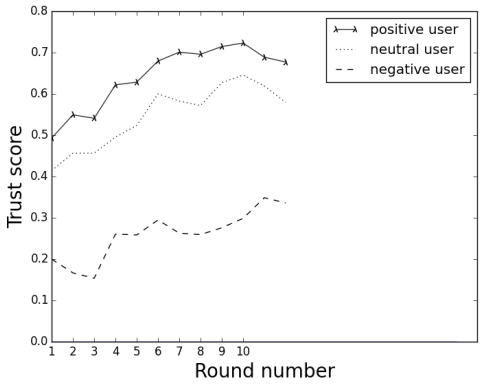

Fig. 5. Validating trust metric with real users' ratings.

In what follows we analyse the results obtained by our trust metric applied for the behavior of these virtual users. Since we are using a continuous rating score and Keser was using a discrete rating score, the two rating scores do not match completely. However, we should expect that our trust metric does not conflict with Keser's results, i.e. for any two behaviors A and B, if A was rated higher than B (for instance, positive versus neutral or positive versus negative), our trust metric should assign a higher trust score to A than B.

The analysis is displayed in Fig. 5. As expected, our trust metric assigns in all cases higher trust values to *positive user* than *neutral user*, and higher trust values to *neutral user* than *negative user*.

The conclusion is that our trust metric and people's opinion about trustworthiness of behavior in repeated trust games do not contradict each other.

C. Evaluation with real data

We showed that our trust metric matches real people's opinions about partner's behavior in the past. In this section we address the issue whether it can predict the future behavior of users. For instance, if our trust metric assigns a high trust value for a user, we are interested whether this particular user behaves well or badly in the future.

We note that, a low R^2 value is usual in predicting human behavior, but in many cases, it does not mean that the prediction is useless [17]. For instance, Ashraf et al. [22] used a list of ten factors to predict users' behavior in one-shot trust game, and achieved the average R^2 of 0.25.

1) Data sets: We analyze the performance of our trust metric on three data sets.

The first data set was collected by an experiment we conducted in our laboratory. We recruited 30 participants through a public announcement and we ran five sessions, each with six people. Before a session started, participants were asked to read user instructions and sign a consent form. During the experiment, users were asked to play a repeated anonymous trust game that we implemented by using zTree [31]. However, participants were not told the total number of rounds in order to avoid them behaving differently at the end of the session.

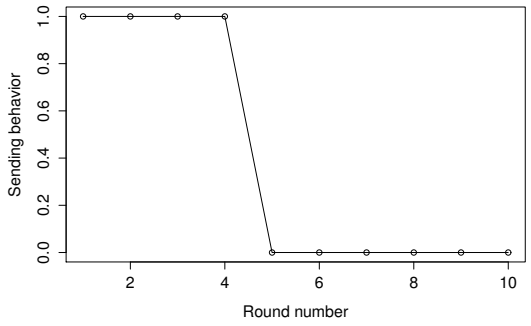

Fig. 6. An observation of fluctuating behavior from our data set.

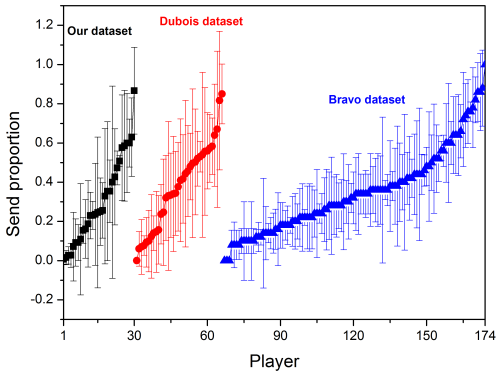

Fig. 7. Average and standard deviation of sending proportions in datasets.

In each round, the six participants were paired randomly, and for each pair the roles were assigned randomly. We ensured that during the session each user interacted exactly five times with each other user.

The two other data sets were obtained by the experiments described in [8] involving 36 participants and in [10] involving 108 participants. Both of these experiments are done anonymously, and the data is provided under the form of a behavior log of participants.

The total data set comprises behavior of 174 participants in repeated trust games.

2) Observation on data: First, we show that our models on user profiles (*low, medium, high and fluctuate profiles*) are consistent with data collected throughout experiments. Next, we show that real data proves the existence of different user types such as participants who send in average a high amount and those who send in average a low amount. We also show that real data proves the existence of users with a fluctuating behavior.

We notice that changes in user behavior in repeated trust games are very usual. Fig. 7 illustrates the average and standard deviation of sending amount proportions of each user in the three datasets previously mentioned. The standard deviations of user sending proportions are large compared with their average sending proportions, meaning that users often change their sending behavior during the experiments. For instance, Fig. 6 illustrates a selected user behavior from our dataset: this player cooperates very well at beginning then deviates and never cooperates again. We observed that in all data sets, only few players send a constant amount throughout a session.

Fig. 7 shows that for all three datasets, the average sending proportions of participants vary from 0 to 1, matching with our defined profiles: *low*, *medium* and *high* corresponding to a sending proportion of 0.2, 0.5 and 0.8 respectively.

We can conclude that, fluctuating behavior is a fact in all three data sets, and for this reason, it is important to design a trust metric that copes with this behavior.

3) Predicting users' behavior: Based on the behavior log we applied our trust metric on users' behavior at a certain

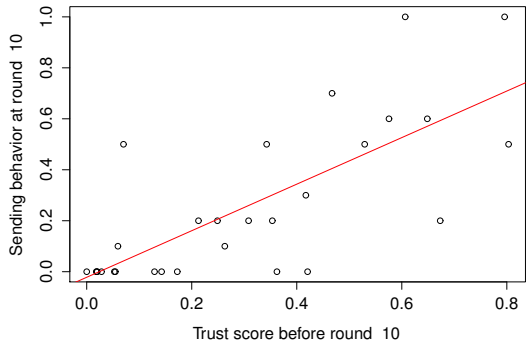

Fig. 8. Relationship between trust metric and user behavior at round ten in our own experiment.

round, then used the output trust score as the independent variable to predict the user's behavior in the next round. For all rounds starting with round five, we found a high correlation between the output trust scores and user behavior in the next round. However, due to space limitation, we present the results of our analysis only for rounds five and ten.

In our analysis the independent variables are the trust values for each user after fourth and ninth interaction and the dependent variables are the sending proportions of users in the fifth and tenth round. For the data in [8], we tested the relationship between our trust metric and the user behavior at round five and ten. However, because of the design of the experiment in [10], we could only test the relationship between our trust metric and user behavior at round five. Fig. 8 displays the prediction of user sending behavior at round ten by using the data set from our experiment. Fig. 9 displays the prediction of user sending behavior at round five by using the Bravo dataset.

The summary of all linear regressions previously mentioned is displayed in Table II, where the independent variable (x-axis in Fig. 8 and Fig. 9) is the trust value our metric assigned to each user before a particular round, and the dependent variable

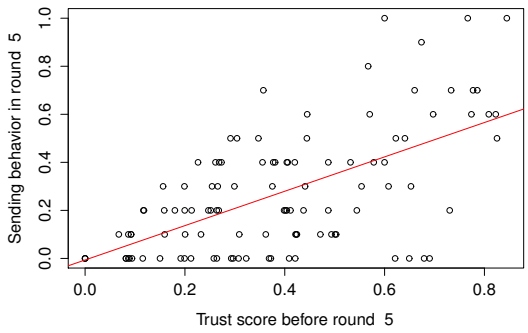

Fig. 9. Relationship between trust metric and user behavior at round five in the Bravo dataset.

TABLE II
REGRESSION BETWEEN TRUST METRIC AND FUTURE USERS' BEHAVIOR.

	Intercept	Slope	Adj.R^2
Our dataset (round 5)	0.071	0.701***	0.319
Our dataset (round 10)	-0.022	0.913***	0.542
Bravo dataset (round 5)	-0.006	0.715***	0.362
Dubois dataset (round 5)	0.072	0.848***	0.356
Dubois dataset (round 10)	0.027	0.855***	0.357

We denote '***' as significant level of 99.9%.

is the behavior of this user in this round (y-axis in Fig. 8 and Fig. 9). We can notice that the slopes of all regressions are significant, meaning that our trust metric predicts well user's behavior. Similar results were obtained for the same analysis in other rounds (i.e. a significant slope value and a positive r-value), but, as previously mentioned, due to lack of space, we do not present them in detail.

4) Performance comparison: As previously mentioned, there is no prior work in predicting users behavior in repeated trust game. For this reason, in this section, we compare our model with two other baseline models: average model and null model.

Average model predicts that, the next sending amount of a user is equal to the average of her previous sending amounts. On the other hand, the null model predicts that, the next sending amount of a user is equal to her previous sending amount.

In order to compare the performance of these three models, we calculate the predicting values of each of these models. We compute the adjusted R^2 value for each model from round five to round ten and then calculate the average of adjusted R^2. The higher average R^2 a model achieves, the better this model is in predicting users behavior.

The comparison of performance of different models is displayed in Table III. For our data and data of Dubois, we calculated an average adjusted R^2 values in predicting users behavior from round five to ten. As Bravo's dataset contains only five rounds, we computed the average adjusted R^2 values in predicting users behavior at round five.

TABLE III
COMPARISON OF R^2 VALUES OF DIFFERENT PREDICTING MODELS.

	Average model	Null model	Our model
Our data	0.42	0.43	0.55
Dubois's data	0.28	0.34	0.40
Bravo's data	0.3	0.32	0.36

We can see that, our model outperforms the other two baseline models in predicting users behavior in repeated trust games.

D. Robustness of the trust model

In this section, we consider the robustness of our model against some common attacks to the reputation systems as addressed in [32] with an adaptation to trust games context. We show that our trust metric is quite robust against some common kinds of attacks to reputation systems.

- Self-promoting attack implies that "attackers manipulate their own reputation by falsely increasing it". In the context of trust games, we can define *self-promoting strategy* as *fluctuating strategy*. Our trust metric can punish the fluctuating strategy by assigning people who use this strategy same score with low-profile users.
- Whitewashing attack implies that "attackers try to repair and restore their reputation". In trust game context, the *whitewashing* strategy corresponds to users that try to repair their reputation by behaving better than in the past. Users adopting this strategy will be classified as having a *fluctuating strategy* and therefore they need a long period to repair their reputation.
- Slandering attack implies that "attackers try to report false data". There are two types of *slandering* attacks: the attackers can report falsely bad information about honest users, or can report falsely good information about malicious users. The first type of *slandering* attack is not possible to appear in our method as data is collected by the system and not by user report. Our metric is robust against the second type of *slandering* attack. In order to falsely increase the reputation level, malicious users can use *fake accounts* to increase the trust score of their main account by providing positive reviews about it. However, our trust metric is based on personal experience of a user with another particular partner, and does not rely on the relationship of this partner with other users.

V. DISCUSSION

Nowadays reputation-based systems are widely used and trust scores assigned to users could be observed in many popular websites, such as Wikipedia, eBay, Amazon and on-line discussion forums (e.g. Stack Overflow). A good trust metric can suggest honest users to select the right partners and reduce the probability that they are cheated in on-line environments.

Our model has a limitation: the cold-start problem. The model requires the data for several interactions between users in order to calculate and predict their next behavior. This drawback generally holds for behavior modelling approaches: without data on previous behavior no prediction can be done [23]. Several potential solutions for this problem exist, such as querying the information from common friends or computing trust path through network [33].

On the other hand, our model has several advantages compared with other approaches presented in Section II. Our trust metric only requires the information about the interaction between the user and the partner in the past. This information is available on users side, without querying any central server or other nodes in the network. The trust metric also does not require any personal information such as gender, age, income or Mach score, which is not always available in collaborative systems today. These advantages facilitate the application of our trust metric to real systems.

VI. Conclusion

In this paper, we presented a trust metric to measure the trustworthiness of users in repeated trust games. We prove that our trust metric can deal with fluctuating user behavior and can distinguish different types of participants. Moreover, we show that our trust metric is consistent with human opinions. Last but not least, we show that our trust metric achieves a better performance in predicting user behavior in repeated trust games in comparison with other baseline approaches.

As repeated trust game is a general model for studying human cooperation, we expect that our trust metric could be applied to collaborative systems [34] where users need to trust and collaborate with other users such as Wikipedia or open source software development projects.

VII. Acknowledgements

This research was partially supported by USCoast Inria associate team and PSPC OpenPaaS::NG project funded by the *Investissements d'Avenir* French government program managed by the *Commissariat général à l'investissement* (CGI). We are thankful to Prof. Valerie L. Shalin, Wright State University, Department of Psychology, for assistance with distribution analysis.

References

[1] J. Berg, J. Dickhaut, and K. McCabe, "Trust, social history, and reciprocity," *Games and Econ. Behav.*, vol. 10, no. 1, pp. 122–142, 1995.

[2] S. Chakravarty, D. Friedman, G. Gupta, N. Hatekar, S. Mitra, and S. Sunder, "Experimental economics: a survey," *Econ. and Political Weekly*, vol. 46, no. 35, pp. 39–78, 2011.

[3] N. D. Johnson and A. A. Mislin, "Trust games: A meta-analysis," *Jour. of Econ. Psychology*, vol. 32, no. 5, pp. 865–889, 2011.

[4] V. Anderhub, D. Engelmann, and W. Güth, "An experimental study of the repeated trust game with incomplete information," *Jour. of Econ. Behav. & Organization*, vol. 48, no. 2, pp. 197–216, 2002.

[5] S. V. Burks, J. P. Carpenter, and E. Verhoogen, "Playing both roles in the trust game," *Jour. of Econ. Behav. & Organization*, vol. 51, no. 2, pp. 195–216, 2003.

[6] J. Engle-Warnick and R. L. Slonim, "Learning to trust in indefinitely repeated games," *Games and Econ. Behav.*, vol. 54, no. 1, pp. 95–114, 2006.

[7] F. Cochard, P. N. Van, and M. Willinger, "Trusting behavior in a repeated investment game," *Jour. of Econ. Behav. & Organization*, vol. 55, no. 1, pp. 31–44, 2004.

[8] D. Dubois, M. Willinger, and T. Blayac, "Does players' identification affect trust and reciprocity in the lab?" *Jour. of Econ. Psychology*, vol. 33, no. 1, pp. 303 – 317, 2012.

[9] J. Bracht and N. Feltovich, "Whatever you say, your reputation precedes you: Observation and cheap talk in the trust game," *Jour. of Public Econ.*, vol. 93, no. 9, pp. 1036–1044, 2009.

[10] G. Bravo, F. Squazzoni, and R. Boero, "Trust and partner selection in social networks: An experimentally grounded model," *Social Networks*, vol. 34, no. 4, pp. 481–492, 2012.

[11] R. Slonim and E. Garbarino, "Increases in trust and altruism from partner selection: Experimental evidence," *Exp. Econ.*, vol. 11, no. 2, pp. 134–153, 2008.

[12] G. E. Bolton, E. Katok, and A. Ockenfels, "Cooperation among strangers with limited information about reputation," *Jour. of Public Econ.*, vol. 89, no. 8, pp. 1457–1468, 2005.

[13] C. Vanberg, "Why do people keep their promises? an experimental test of two explanations," *Econometrica*, vol. 76, no. 6, pp. 1467–1480, 2008.

[14] A. Ben-Ner and L. Putterman, "Trust, communication and contracts: An experiment," *Jour. of Econ. Behav. & Organization*, vol. 70, no. 1, pp. 106–121, 2009.

[15] N. Feltovich and J. Swierzbinski, "The role of strategic uncertainty in games: An experimental study of cheap talk and contracts in the nash demand game," *European Econ. Review*, vol. 55, no. 4, pp. 554–574, 2011.

[16] C. Camerer, *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press, 2003.

[17] A. Gunnthorsdottir, K. McCabe, and V. Smith, "Using the machiavellianism instrument to predict trustworthiness in a bargaining game," *Jour. of Econ. Psychology*, vol. 23, no. 1, pp. 49–66, 2002.

[18] A. M. Evans and W. Revelle, "Survey and behavioral measurements of interpersonal trust," *Jour. of Research in Personality*, vol. 42, no. 6, pp. 1585–1593, 2008.

[19] T. Yamagishi, S. Akutsu, K. Cho, Y. Inoue, Y. Li, and Y. Matsumoto, "Two-component model of general trust: Predicting behavioral trust from attitudinal trust," *Social Cognition*, vol. 33, no. 5, p. 436, 2015.

[20] S. T. Yen, "An econometric analysis of household donations in the usa," *Applied Econ. Letters*, vol. 9, no. 13, pp. 837–841, 2002.

[21] A. Falk, S. Meier, and C. Zehnder, "Did we overestimate the role of social preferences? the case of self-selected student samples," *CESifo Working Paper Series*, 2010.

[22] N. Ashraf, I. Bohnet, and N. Piankov, "Decomposing trust and trustworthiness," *Exp. Econ.*, vol. 9, no. 3, pp. 193–208, 2006.

[23] J. Tang, H. Gao, H. Liu, and A. D. Sarma, "eTrust: understanding trust evolution in an online world," in *Proceedings of KDD*. ACM, 2012, pp. 253–261.

[24] E. L. Glaeser, D. I. Laibson, J. A. Scheinkman, and C. L. Soutter, "Measuring trust," *Quarterly Jour. of Econ.*, pp. 811–846, 2000.

[25] S. Altmann, T. Dohmen, and M. Wibral, "Do the reciprocal trust less?" *Econ. Letters*, vol. 99, no. 3, pp. 454–457, 2008.

[26] A. Das and M. M. Islam, "SecuredTrust: A dynamic trust computation model for secured communication in multiagent systems," *IEEE Trans. Dependable Sec. Comput.*, vol. 9, no. 2, pp. 261–274, 2012.

[27] B. L. Fredrickson and D. Kahneman, "Duration neglect in retrospective evaluations of affective episodes," *Jour. of personality and social psychology*, vol. 65, no. 1, p. 45, 1993.

[28] C. Buntain and J. Golbeck, "Trust transfer between contexts," *Jour. of Trust Management*, vol. 2, no. 1, 2015.

[29] G. Keppel, *Design and analysis: A researcher's handbook* . Prentice-Hall, Inc, 1991.

[30] C. Keser, "Trust and reputation building in e-commerce. IBM Watson Research Center," CIRANO working paper, Tech. Rep., 2002.

[31] U. Fischbacher, "z-Tree: Zurich toolbox for ready-made economic experiments," *Exp. Econ.*, vol. 10, no. 2, pp. 171–178, 2007.

[32] K. J. Hoffman, D. Zage, and C. Nita-Rotaru, "A survey of attack and defense techniques for reputation systems," *ACM Comput. Surv.*, vol. 42, no. 1, 2009.

[33] X. Liu, A. Datta, and E.-P. Lim, *Computational Trust Models and Machine Learning*. CRC Press, 2014.

[34] H. T. T. Truong, C. Ignat, and P. Molli, "A contract-extended push-pull-clone model for multi-synchronous collaboration," *Jour. of Cooperative Inf. Syst.*, vol. 21, no. 3, pp. 221–262, 2012.