RESEARCH ARTICLE

Can Centralized Sanctioning Promote Trust in Social Dilemmas? A Two-Level Trust Game with Incomplete Information

Raymond Yu Wang[1]*, Cho Nam Ng[2]

[1] Faculty of Social Sciences, The University of Hong Kong, Pokfulam, Hong Kong, Hong Kong,
[2] Department of Geography, The University of Hong Kong, Pokfulam, Hong Kong, Hong Kong

* wangyuray@connect.hku.hk

Abstract

The problem of trust is a paradigmatic social dilemma. Previous literature has paid much academic attention on effects of peer punishment and altruistic third-party punishment on trust and human cooperation in dyadic interactions. However, the effects of centralized sanctioning institutions on decentralized reciprocity in hierarchical interactions remain to be further explored. This paper presents a formal two-level trust game with incomplete information which adds an authority as a strategic purposive actor into the traditional trust game. This model allows scholars to examine the problem of trust in more complex game theoretic configurations. The analysis demonstrates how the centralized institutions might change the dynamics of reciprocity between the trustor and the trustee. Findings suggest that the sequential equilibria of the newly proposed two-level model simultaneously include the risk of placing trust for the trustor and the temptation of short-term defection for the trustee. Moreover, they have shown that even a slight uncertainty about the type of the newly introduced authority might facilitate the establishment of trust and reciprocity in social dilemmas.

Citation: Wang RY, Ng CN (2015) Can Centralized Sanctioning Promote Trust in Social Dilemmas? A Two-Level Trust Game with Incomplete Information. PLoS ONE 10(4): e0124513. doi:10.1371/journal.pone.0124513

Academic Editor: Daniel Houser, George Mason University, UNITED STATES

Received: November 12, 2014

Accepted: March 15, 2015

Published: April 16, 2015

Data Availability Statement: All relevant data are within the paper.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Trust is a critical social factor which is considered to be highly conducive to preventing opportunistic behaviour, decreasing transaction costs and maintaining cooperation in human activities [1–5]. Traditional game theoretic analysis of trust is usually built upon a standard trust game as shown in Fig 1 [6–11]. In the trust game, "not placing trust" is the rational individual action in a one-shot game [12]. This is a suboptimal social outcome because "placing trust, honouring trust" is a strict improvement for both the trustor and the trustee. From this point of view, Fig 1 concisely illustrates a paradigmatic social dilemma in human society, where R_1, S_1, P_1 denote the utility of the trustor and R_2, T_2, P_2 denote the utility of the trustee.

However, the traditional trust game has limitations of analysing hierarchical interactions in complex systems. In particular, what is missing in the traditional model is a dynamic perspective which distinguishes different types of trust. In many conventional studies, trust is a concept at the individual or interpersonal level [13]. Some scholars focus on interpretations of one

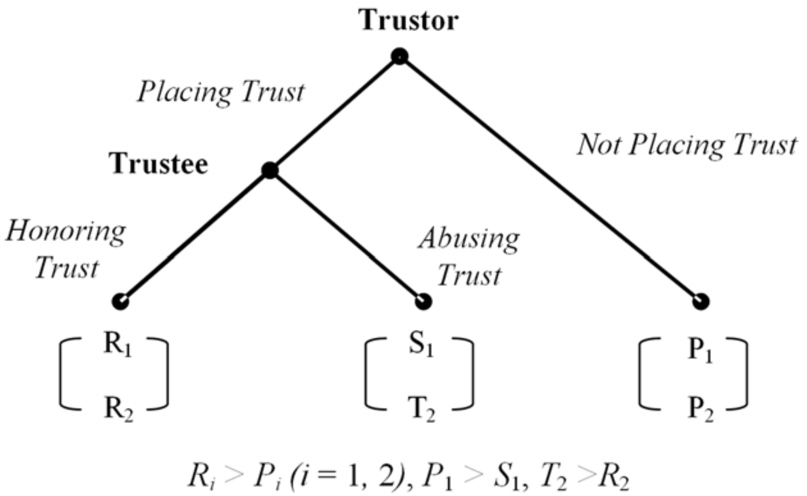

Fig 1. Extensive form of the Trust Game.

doi:10.1371/journal.pone.0124513.g001

single type of trust such as inter-personal relationships in which trust is a stationary result under certain socio-economic conditions [14–23]. Others highlight the sources of trust and trace them back to the process of interactions or characteristics and competence of individuals [24–26]. Yet if one analyses the problem of trust in more complex contexts, then it is important to recognize that it involves interactions between more than just individuals [27]. Consider common-pool resource (CPR) governance in a large-scale social-ecological system (SES), it is evident that trust should be attributed to a variety of relationships between multiple groups, organizations and institutions. Under diverse social, economic, political and spatial-temporal conditions, noticeably, an actor has to measure his belief not only in the trustworthiness of his peer actors, but also in institutions and governance agencies which influence collective and individual behaviour through external intervention. These trust relationships are beyond social-psychological understandings of inter-personal activities. Against the backdrop of increasing complexity, the problem of trust thus should be examined through an alternative structure which integrates interactions between actors at different levels.

From the perspective of rational choice theory, formal models and experiments have been developed to investigate dynamics and mechanisms of trust relationships [10,28,29]. Built on dyadic interactions between different actors, these theoretical and experimental studies have shown that reciprocity and peer punishment can facilitate trust in social dilemmas [12,30,31]. However, the major limitation of these models is that they are established on a horizontal actor-network in which only individual trust is examined. One should notice that peer punishment on defectors through reciprocity is difficult in more complicated systems [32,33]. Recently, scholars have started raising questions in regard to "the ability of spontaneous, uncoordinated and decentralized peer punishment to sustain cooperation in complex societies". Based on experiments of public good games (PGG), some scholars argue that peer punishment might be only effective under conditions such that the group size is sufficiently small or exit option is provided to participants [34–37]. In empirical settings of CPR governance, we frequently observe centralized institutions which impose top-down sanctions to prevent opportunistic behaviour. Some experimental evidence derived from PGG also has shown that centralized sanctioning system can promote human cooperation and be more efficient than peer punishment due to its ability to overcome coordination failure and free-riding problems [38–40].

This paper engages in discussions about peer punishment and centralized sanctioning by developing a formal iterative two-level trust game with incomplete information. More specifically, it proposes a hierarchical structure which not only simultaneously includes individual trust and institutional trust, but also examines how centralized sanctioning might affect reciprocity between the trustor and the trustee. In each period of the newly proposed game, an authority who moves after the trustee is added. If the trustee honoured trust, then the focal period of the game ends; if the trustee abused trust, then a choice is granted to the authority, who can either impose a costly punishment on the defective trustee or not punish the defection. Meanwhile, information is incomplete in the sense that actors are not fully informed on other actors' utility functions and preferences. Note the hierarchical structure and the role of the authority in this paper differ from those in previous literature on altruistic third-party punishment and human cooperation [41–49]. In our model, the authority engages in repeated interactions rather than a one-shot game. In addition, the authority's total utilities are dependent on his own and other actors' behaviour. Thus he is considered a strategic purposive actor rather than an altruist who gains no economic benefits from costly punishment. Therefore, the configurations of this model do not intend to examine human altruism; instead, we focus on how centralized sanctioning institutions and incomplete information may affect reciprocity at the individual level. This extension could improve the applicability of the two-level trust game in more complex settings. By linking equilibrium strategies of the traditional trust game with those of the newly proposed model, one could compare the effects of centralized sanctioning and peer punishment on trust and reciprocity.

This paper is organized as follows. First, a two-level trust game which is built upon the traditional baseline trust game is introduced. Then, the sequential equilibrium is formally derived for the two-level trust game under a scenario in which information is incomplete about both the trustee and the authority. Finally, the paper concludes with theoretical and empirical implications of the two-level trust game.

Method

The baseline trust game with incomplete information

The formal game theoretic analysis begins with a review of a baseline model which is built upon the trust game presented in Fig 1 [7,10,50,51]. The baseline model includes two important features. The first one is a move by nature, before the game starts, deciding which type of trustee will participate in the game (see Fig 2). This entails that it is assumed two types of trustees exist in nature—the G-type (good) and the B-type (bad). Both types of trustees are utility maximisers. Yet they have different preferences. The G-type trustees have stronger altruistic tendencies and therefore always feel more satisfied by honouring trust than abusing trust. On the contrary, the B-type trustees have stronger selfish tendencies and therefore prefer abusing trust than honouring trust in a one-shot game. This is a plausible assumption as it reflects the coexistence of opportunists and altruists in empirical settings [52–54]. The trustee knows his type, yet information is incomplete in the sense that, at the beginning of the game, the trustor does not know which type of trustee will be his counterpart. Let $\pi_1{}^E$ be the probability that the trustor assigns at the beginning of the game to the event that the trustee is a G-type.

The second feature is a continuation of the game. The game is finitely repeated without the assumption of discounting utilities. Hence the total utility any actor receives during the game is the total undiscounted sum of utility that he obtains in each game period. Moreover, anyone in this game knows exactly how many periods will last in the game.

Under circumstances of complete information, backward induction informs us that no trust should be placed in finitely repeated trust games if the trustor knew he would encounter a B-

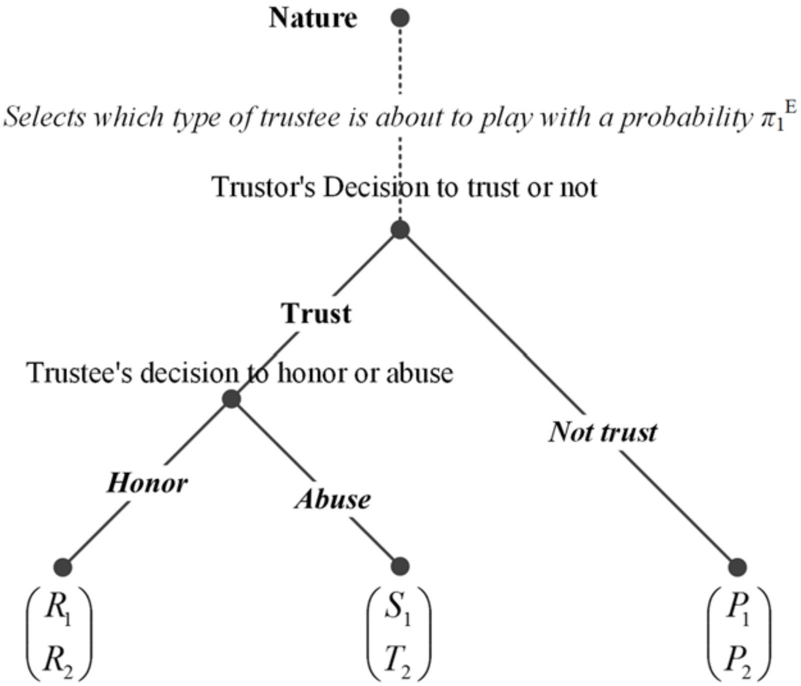
Nature
Selects which type of trustee is about to play with a probability π_1^E
Trustor's Decision to trust or not
Trust
Not trust
Trustee's decision to honor or abuse
Honor
Abuse
$\begin{pmatrix} R_1 \\ R_2 \end{pmatrix}$
$\begin{pmatrix} S_1 \\ T_2 \end{pmatrix}$
$\begin{pmatrix} P_1 \\ P_2 \end{pmatrix}$

Fig 2. Extensive form of a baseline trust game with incomplete information.

doi:10.1371/journal.pone.0124513.g002

type trustee; however, with slight uncertainties about the type of the trustee, the outcome of the game would have been substantially changed in a way that even B-type trustees might honour trust in early periods of the game. The sequential equilibrium of the baseline trust game with incomplete information consists of three phases; namely, stable trust, randomization and no trust [51,55]. Bower, Garber, and Watson (1996) provided a comprehensive proof of this sequential equilibrium when the baseline trust game is played twice. To avoid complexity and ensure consistency in the following analysis, the results for the two-period baseline game are summarised in Table 1, where the notations are defined as follows:

w_1 & w_2 = probabilities that the trustor A_1 places trust at period I & II.

r_1 & r_2 = probabilities that the B-type trustee A_2 honours trust at period I & II.

Table 1. Sequential equilibria for the baseline trust game.

Case	Equilibrium strategies
1. $\pi_1^E > \frac{S_1}{R_1 - S_1}$	$w_1 = r_1 = w_2 = r_2 = 1$
2. $\pi_1^E = \frac{S_1}{R_1 - S_1}$	$w_1 = r_1 = 1$; $w_2 \geq (T_2—R_2)/T_2$
3. $\left(\frac{S_1}{R_1 - S_1}\right)^2 < \pi_1^E < \frac{S_1}{R_1 - S_1}$	$w_1 = 1; r_1 = -\pi_1^E/S_1(1 - \pi_1^E); w_2 = (T_2 - R_2)/T_2$
4. $\pi_1^E = \left(\frac{S_1}{R_1 - S_1}\right)^2$	$w_1 = 1; r_1 = -\pi_1^E/S_1(1 - \pi_1^E); w_2 \leq (T_2 - R_2)/T_2$
5. $\pi_1^E < \left(\frac{S_1}{R_1 - S_1}\right)^2$	$w_1 = w_2 = 0$

Source: adapted from (Bower, et al., 1996). For analytical simplicity, it is assumed that $P_1 = P_2 = 0$ when no trust is placed, and hence $S_1 < 0$. π_1^E is common knowledge.

doi:10.1371/journal.pone.0124513.t001

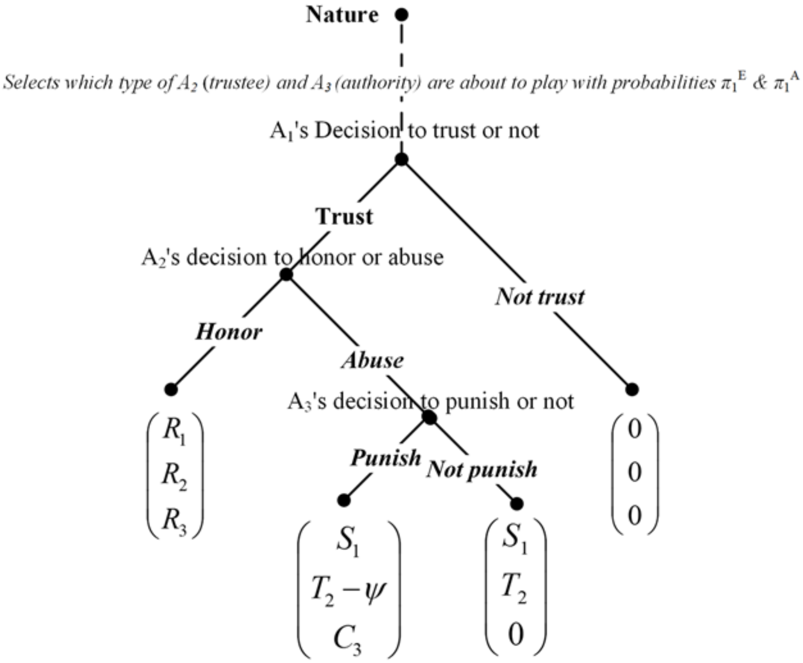
Nature
Selects which type of A_2 (trustee) and A_3 (authority) are about to play with probabilities π_1^E & π_1^A
A_1's Decision to trust or not
Trust
A_2's decision to honor or abuse
Not trust
Honor
Abuse
A_3's decision to punish or not
Punish
Not punish
(R_1, R_2, R_3)
$(S_1, T_2-\psi, C_3)$
$(S_1, T_2, 0)$
$(0, 0, 0)$

Fig 3. Extensive form of a two-level trust game with incomplete information.

doi:10.1371/journal.pone.0124513.g003

This result exhibits that whether the trustor places trust is mainly dependent on his *ex ante* belief about the probability that the trustee is a G-type, the number of periods to be played and the RISK for him to place trust, where RISK = $S_1/(S_1\text{-}R_1)$ [55]. Therefore, it is easy to conclude that the first (trust) phase of the model will be longer under three conditions; namely, a higher *ex ante* probability that the trustee is a G-type, a larger number of periods to be played and a smaller risk of placing trust for the trustor.

The two-level Trust game with incomplete information

The two-level trust game is an extension of the baseline model (see Fig 3). The term "two-level" emphasises a hierarchical structure and a newly introduced actor—the authority.

Let the trustor, the trustee and the authority be named by the order of their actions as A_1 and A_2 and A_3. In a focal period of the two-level trust game, the authority would receive a reward R_3 if trust was placed and honoured; nevertheless, if trust was placed and abused, then the authority would not receive the reward R_3 and he must make a binary choice between punishing (C) and not punishing (D) the defective trustee. By punishing defective trustees, the authority A_3 can impose a punishment ψ on them. This punishment is costly for the authority as it associates with a negative cost C_3. The punishment will not change the utility of the trustor; however, it reduces the utility of the defective trustee to $T_2\text{-}\psi$. It is assumed that $(T_2\text{-}\psi)<R_2$ and hence a rational trustee would not have abused trust if he knew that he was going to be punished. In the baseline model, the type of the trustee is unknown to the trustor at the beginning of the game. Likewise, it is sensible to make a similar assumption about the type of the authority. Namely, two types of authorities are assumed to co-exist in nature—an altruistic G-type who always prefer to punish defective trustees and an opportunistic B-type who are reluctant to punish if it was only a one-shot game. It is assumed that $R_3+C_3>0$ and thus the B-type A_3 has an incentive to bear a short-term cost for a long-term reward in repeated interactions. It is continued to assume $P_1 = P_2 = P_3 = 0$ when no trust is placed for analytical simplicity.

Results

In this section, the sequential equilibrium strategies for all actors in the two-level trust game are derived. Let $\pi_n{}^E$ and $\pi_n{}^A$ respectively denote the probabilities assigned to the event that the trustee A_2 and the authority A_3 are a G-type at period n, where the subscript n denotes the period of the game. Note that $\pi_n{}^E$ and $\pi_n{}^A$ are common knowledge. For analytical simplicity, the game is assumed to be played twice ($n = 1, 2$) and there will be no discounted utilities in the game. To derive the sequential equilibrium of the game, this paper follows the approach developed by Bower, et al. (1996). Each decision node on the game tree is indexed and worked backwards starting with the second period.

Node 6: Period II, A_3's decision when A_2 abused trust

There are two types of A_3. The G-type A_3 surely punishes. The B-type A_3 surely defects because the game will immediately end after A_3's action and there will be no future utilities to offset the B-type A_3's costs for punishing. It is a unique equilibrium continuation.

Node 5: Period II, A_2's decision to honour or abuse when trusted by A_1

There are two types of A_2. The G-type A_2 always honours trust. The B-type A_2 honours trust only if his expected utilities for choosing C are larger than those of choosing D. In other words, the B-type A_2 honours trust when the authority is sufficiently trustworthy and the punishment on defectors is sufficiently high.

Mathematically that is,$R_2 > \pi_2^A(T_2 - \psi) + (1 - \pi_2^A)T_2$, and algebra yields,

$$\pi_2^A > \frac{T_2 - R_2}{\psi} \tag{1}$$

Otherwise, the B-type A_2 abuses trust. Note that the B-type A_2 may randomize with probability $r_2 \in [0,1]$ when $\pi_2^A = (T_2 - R_2)/\psi$.

Node 4: Period II, A_1's decision to place trust or not

A_1 only places trust when his expected utilities for placing trust are larger than zero, the latter of which is his expected utilities for not placing trust in period II. Different from the baseline model, now A_1's expected utilities not only depend on the probability that A_2 is a G-type who will honour trust for sure, they also depend on the probability that A_3 is a G-type who will punish the B-type A_2; in the latter case, even the B-type A_2 might honour trust in the last round of the game given that there is a high belief that A_3 would punish A_2 should he abused trust. In other words, A_1 will place trust if either the probability that A_2 is a G-type or the probability that A_3 is a G-type is sufficiently high.

It is already established in the analysis of Node 4 that the B-type A_2 honours trust on condition of Inequality (1). Thus, it is clear that A_1 will place trust when $\pi_2{}^A > (T_2 - R_2)/\psi$. In addition, A_1 will also place trust if he believes the trustee is sufficiently trustworthy even when Inequality (1) is not satisfied. That is to say although the B-type A_2 surely defects, A_1 will place trust when $\pi_2^E R_1 + (1 - \pi_2^E)S_1 > 0$, and algebra yields

$$\pi_2^E > \frac{S_1}{S_1 - R_1} \tag{2}$$

Note that A_1 may randomize with probability $w_2 \in [0,1]$ when $\pi_2{}^E = S_1/(S_1\text{-}R_1)$. Here the first two-level outcome is reached; namely, the trustor A_1 will place trust at period II if either Inequality (1) or (2) is satisfied.

Node 3: Period I, A_3's decision to punish or not when A_2 defected

Node 3 is only reached when the trustee A_2 abused trust in period I, which suggests A_2 is B-type and $\pi_2{}^E = 0$. At node 3, the G-type A_3 always punishes. Note that if the B-type A_3 did not punish the defective A_2 in period I, then it would be revealed that A_3 is B-type too, which leads to $\pi_2{}^A = 0$ as well as zero utility for all actors in period II. The B-type A_3 would, however, punish only if his expected utilities in period II are no smaller than his costs for punishing the defective A_2 in period I.

Suppose in equilibrium a B-type A_3 chooses to punish with a probability q_r. Note that by Bayes' rule,

$$\pi_2^A = \frac{\pi_1^A}{\pi_1^A + (1 - \pi_1^A)q_r} \tag{3}$$

Let $q_r{}^*$ be defined by

$$\frac{T_2 - R_2}{\psi} = \frac{\pi_1^A}{\pi_1^A + (1 - \pi_1^A)q_r^*}$$

and thus

$$q_r^* \equiv \frac{\pi_1^A(\psi + R_2 - T_2)}{(T_2 - R_2)(1 - \pi_1^A)} \tag{4}$$

In the following it is shown by contradiction that $q_r{}^*$ is the equilibrium value of q_r.

If $q_r > q_r{}^*$, then $\pi_2{}^A < (T_2\text{-}R_2)/\psi$ and the B-type A_3's continuation utility in period II would be zero; because, A_1 knew that A_2 would have abused trust if A_1 places trust in period II, therefore A_1 would not place trust in the first place and everyone receives zero in the second round. In this case, the B-type A_3 would strictly prefer choosing not to punish in period I, which implies $q_r = 0$ (a contradiction).

If $q_r < q_r{}^*$, then $\pi_2{}^A > (T_2\text{-}R_2)/\psi$ and the B-type A_3 expects a continuation utility of R_3 in period II; because A_1 knew that A_2 would have honoured trust if A_1 places trust in the first place. Thus the trust would be placed and honoured in period II. Given it was stipulated that R_3 is strictly larger than the cost for a B-type A_3 to punish defective A_2, the B-type A_3 then strictly prefers choosing to punish, which implies $q_r = 1$ (a contradiction).

Furthermore, the randomization probabilities for A_1 and A_2 in equilibrium, when $\pi_2{}^E = S_1/(S_1\text{-}R_1)$ and $\pi_2{}^A = (T_2\text{-}R_2)/\psi$, are selected such that the actors who respectively move prior to A_1 and A_2 are indifferent in the periods before [55]. For A_2, the trustor A_1 must be indifferent between placing and not placing trust, that is, $r_2R_1+(1\text{-}r_1)S_1 = 0$ and thus $r_2 = S_1/(S_1\text{-}R_1)$; for A_1, the authority must be indifferent between imposing and not imposing punishment, that is $w_2r_2R_3+C_3 = 0$. Replacing r_2 with $S_1/(S_1\text{-}R_1)$ and simple algebra yields $w_2 = C_3(R_1\text{-}S_1)/(R_3S_1)$. In a nutshell, the randomization probability cannot be either too high or too low to alter the choice of the actor who moves in the period before.

Node 2: Period I, A_2's decision to honour or abuse trust when trust is placed

The G-type A_2 always cooperates. With regard to the strategy for the B-type A_2, one needs to consider a two-level deduction. More specifically, the B-type A_2 now simultaneously faces an internal and an external factor that influence his choice of action. The internal factor is peer punishment which might be imposed on the B-type A_2 and reduces his potential long-term benefits. The peer punishment might drive the B-type A_2 pretend to be a G-type. This is a well-established notion according to the baseline model and it has been tested by many laboratory experiments [12,56]. The external factor, on the other hand, is introduced by the imposition of the authority. In this two-level trust game, it is a centralized sanctioning institution and an extra constrain on possible defective behaviours of the B-type A_2. Therefore, the B-type A_2 has another motivation to cooperate in addition to his concern for peer punishment.

The centralized sanctioning is reflected by A_2's expected utilities. He will honour trust if his expected utilities for choosing C are no smaller than choosing D at period I. That is,

$$R_2 > (T_2 - \psi)\pi_1^A + (1 - \pi_1^A)q_r^*(T_2 - \psi) + (1 - \pi_1^A)(1 - q_r^*)T_2$$

replace $q_r{}^*$ with Eq (4) and algebra yields,

$$\pi_1^A > \left(\frac{T_2 - R_2}{\psi}\right)^2 \tag{5}$$

When Inequality (5) cannot be satisfied, the scenario can be simply viewed as a baseline model in which the authority does not exist. It basically suggests that the authority is so untrustworthy to such extent that $\pi_1{}^A < [(T_2 - R_2)/\psi]^2$. In this case, the B-type A_2 may randomize with probability $r_1 \in [0,1]$in equilibrium as proved by Buskens [55] and Bower et al. [51]. The process of randomization probability selection is similar to what has been shown in the analysis of Node 3. Without further duplication, the randomization probability is,

$$r_1 = \frac{\pi_1^E R_1}{(\pi_1^E - 1)S_1} \tag{6}$$

Node 1: Period I, A_1's decision to place trust or not

Similar to the analysis of Node 4, A_1 chooses to place trust under two circumstances. One is that the B-type trustee honours trust given their anticipation that there is a sufficiently high probability the authority will punish. This condition is illustrated by the analysis of Node 2 and Inequality (5). When Inequality (5) cannot be satisfied, or to put it another way, when the authority's trustworthiness does not reach an adequate level, an alternative condition for the trust placement is that the probability that the trustee is a G-type is sufficiently high. This is clearly true when $\pi_1{}^E > S_1/(S_1 - R_1)$. If $\pi_1{}^E \leq S_1/(S_1 - R_1)$, then it implies

$$\pi_1^E R_1 + (1 - \pi_1^E)r_1 R_1 + (1 - \pi_1^E)(1 - r_1)S_1 > 0$$

Substitution of r_1 according to Eq (6) and algebra yields

$$\pi_1^E < \left(\frac{S_1}{R_1 - S_1}\right)^2 \tag{7}$$

Table 2. Categories for the two-level trust game with incomplete information about both the authority and the trustee.

Category I: Belief that A_3 is a G-type	Category II: Belief that A_2 is a G-type	Description
$\pi_1^A > \frac{T_2 - R_2}{\psi}$	$\pi_1^E > \frac{S_1}{R_1 - S_1}$	Full optimism (FO)
$\pi_1^A = \frac{T_2 - R_2}{\psi}$	$\pi_1^E = \frac{S_1}{R_1 - S_1}$	High optimism (HO)
$\left(\frac{T_2 - R_2}{\psi}\right)^2 < \pi_1^A < \frac{T_2 - R_2}{\psi}$	$\left(\frac{S_1}{R_1 - S_1}\right)^2 < \pi_1^E < \frac{S_1}{R_1 - S_1}$	Intermediate optimism (IO)
$\pi_1^A = \left(\frac{T_2 - R_2}{\psi}\right)^2$	$\pi_1^E = \left(\frac{S_1}{R_1 - S_1}\right)^2$	High pessimism (HP)
$\pi_1^A < \left(\frac{T_2 - R_2}{\psi}\right)^2$	$\pi_1^E < \left(\frac{S_1}{R_1 - S_1}\right)^2$	Full pessimism (FP)

doi:10.1371/journal.pone.0124513.t002

Therefore, A_1 places trust with certainty at period I when either Inequality (5) or Inequality (7) is satisfied. There will be no trust placement if neither inequality can be satisfied.

To sum up, the sequential equilibrium of the two-level trust game inherits characteristics of the baseline model. In particular, π_1^E (Actors' *ex ante* belief in the probability that the trustee A_2 is a G-type) and RISK still constitute important factors for trust. In addition, new features have been developed. π_1^A (Actors' *ex ante* belief in the probability that the authority A_3 is a G-type) and an alternative version of temptation, TEMPP = $(T_2\text{-}R_2)/\psi$ (recall that TEMP = $(T_2\text{-}R_2)/T_2$), are incorporated into the equilibrium. It should be noted that the *ex ante* beliefs in the probability that A_2 and A_3 is a G-type respectively fall into five categories as illustrated in Table 2. These five categories are labelled as Full optimism (FO), High optimism (HO), Intermediate optimism (IO), High pessimism (HP) and Full pessimism (FP) with a descending degree of confidence in A_2 and A_3 being a G-type.

The incomplete information about both the trustee and the authority reconstructs the baseline trust game. The result of the newly proposed two-level trust game suggests that the equilibrium strategies for both the trustor and B-type actors are altered. There are 25 cases which involve different combinations of *ex ante* beliefs in the types of the trustee and the authority. The equilibrium strategies for the trustor, the B-type trustee and the B-type authority in each scenario are summarized in Table 3.

Case I-V. When $\pi_1^A = FO$, the actors are initially very optimistic about the authority's type. In this case, maximum trust and reciprocity can be achieved regardless of actors' *ex ante* beliefs in the trustee. The sequential equilibrium only includes pure strategies in the sense that the trustor places trust and the trust is honoured in both periods. Even the B-type trustee honours trust in the last period of the game. The authority enjoys the reward for the establishment of trust without entering the game. These cases yield the highest group utilities for all actors.

Case VI-X. When $\pi_1^A = HO$, the actors are still optimistic about the authority's type, but their equilibrium strategies might vary. The trustor will place trust in both periods if $\pi_1^E = FO$ given it indicates his full confidence in the trustee. He will otherwise place trust in period I and randomize with a high probability in period II regardless of his *ex ante* beliefs in the trustee. On the other hand, the B-type trustee will always honour trust in Period I and always randomize in period II. His randomization probability is determined by a polynomial which represents the risk for the trustor to place trust.

Case VI-X. When $\pi_1^A = IO$, it implies an intermediate degree of optimism about the authority's type. The trustor will place trust in both periods if he is fully confident in the trustee ($\pi_1^E = FO$) and he will randomize in period II otherwise. His randomization probability is determined by a polynomial consists of various parameters such as his risk for placing trust, the

Table 3. Sequential equilibrium strategies for the two-level Trust game.

Case	Belief in the authority	Belief in the trustee	Equilibrium Strategies
I-V	$\pi_1^A = FO$	$\pi_1^E = FO, HO, IO, HP, FP$	$w_1 = r_1 = w_2 = r_2 = 1;$
VI-X	$\pi_1^A = HO$	$\pi_1^E = FO$	$w_1 = r_1 = w_2 = 1; r_2 = S_1/(S_1\text{-}R_1)$
		$\pi_1^E = HO, IO, HP, FP$	$w_1 = r_1 = 1; w_2 \geq C_3(R_1\text{-}S_1)/(R_3S_1); r_2 = S_1/(S_1\text{-}R_1)$
XI-XV	$\pi_1^A = IO$	$\pi_1^E = FO$	$w_1 = r_1 = w_2 = 1; r_2 = 0$
		$\pi_1^E = HO$	$w_1 = r_1 = 1; w_2 \geq C_3(R_1\text{-}S_1)/(R_3S_1); r_2 = 0$
		$\pi_1^E = IO, HP, FP$	$w_1 = r_1 = 1; w_2 = C_3(R_1\text{-}S_1)/(R_3S_1); r_2 = 0$
XVI-XX	$\pi_1^A = HP$	$\pi_1^E = FO$	$w_1 = r_1 = w_2 = 1; r_2 = 0$
		$\pi_1^E = HO$	$w_1 = r_1 = 1; w_2 \geq C_3(R_1\text{-}S_1)/(R_3S_1); r_2 = 0$
		$\pi_1^E = IO, HP, FP$	$w_1 = 1; r_1 = \pi_1^E R_1/(S_1\text{-}\pi_1^E S_1); q_r = \pi_1^A(\psi+R_2\text{-}T_2)/[(T_2\text{-} R_2)(1\text{-}\pi_1^A)]; w_2 = C_3(R_1\text{-}S_1)/(R_3S_1); r_2 = 0$
XXI-XXV	$\pi_1^A = FP$	$\pi_1^E = FO$	$w_1 = r_1 = w_2 = 1; r_2 = 0$
		$\pi_1^E = HO$	$w_1 = r_1 = 1; w_2 \geq C_3(R_1\text{-}S_1)/(R_3S_1); r_2 = 0$
		$\pi_1^E = IO, HP$	$w_1 = 1; r_1 = \pi_1^E R_1/(S_1\text{-}\pi_1^E S_1); q_r = \pi_1^A(\psi+R_2\text{-}T_2)/[(T_2\text{-} R_2)(1\text{-}\pi_1^A)]; w_2 = C_3(R_1\text{-}S_1)/(R_3S_1); r_2 = 0$
		$\pi_1^E = FP$	$w_1 = w_2 = 0$

doi:10.1371/journal.pone.0124513.t003

costs C_3 for the authority to impose punishment and the rewards R_3 for the authority if trust is placed and honoured. The B-type trustee will surely cooperate in period I and always defect in period II considering the *ex ante* belief in the authority's type is not sufficiently high.

Case XVI-XX. When $\pi_1^A = HP$, the actors are generally pessimistic about the authority's type. However, trust can be still placed and honoured in period I if π_1^E is relatively high; because, the B-type trustee has incentives to pretend to be a G-type in order to receive higher long-term benefits. If the degree of optimism about the trustee's type is intermediate or relatively low, the B-type trustee will randomize in Period I and always abuse trust in Period II. The B-type authority will also randomize if the trustee abused in Period I. The trustor may also place trust with a probability in Period II. There will be no trust until the trustor stops placing trust or trust is abused and defection is unpunished.

Case XXI-XXV. When $\pi_1^A = FP$, it implies that the actors are fully pessimistic about the authority's type. The game then turns into a scenario similar to the baseline game in the sense that the authority barely has any impact over the actors' choices of action. Basically, the sequential equilibrium strategies for the trustor and the B-type trustee are similar to what is presented in Table 1.

Discussion and Conclusion

This paper has presented a two-level trust game with incomplete information. This two-level configuration introduces several new features and it depicts a hierarchical structure of interactions in which the effects of centralized sanctioning on trust and reciprocity can be examined. Many new insights have been developed from the game theoretic analysis of the two-level trust game.

Firstly, the sequential equilibrium of our model simultaneously includes both the risk for placing trust (RISK) and the temptation for abusing trust (TEMPP) as key factors for cooperation at the individual level. This is an important advancement because traditional formal models only include one of them as a key factor for cooperation in social dilemmas. For instance, the Perfect Folk Theorem explains cooperation with indefinitely repeated games. It suggests that rational actors might cooperate in social dilemmas as long as the discounting factor β is sufficiently large when compared to TEMP. The baseline trust game explains cooperation with incomplete information. As shown in Table 1, RISK plays a much more important role than TEMP in determining the length of cooperation period in the baseline trust game. Although these two mechanisms (indefinitely repetition and incomplete information) are well-known for bringing about cooperation in social dilemmas, no previous formal models have simultaneously incorporated TEMP and RISK into critical conditions for cooperation. The proposed two-level model, however, reaches such an integrated result as indicated by Inequalities 5 and 7.

Secondly, the two-level trust game provides an opportunity to compare its equilibrium strategies with those of the baseline model in which no overarching authority participates. This comparison leads to an interesting postulation suggesting that even a slight uncertainty about the authority's type might significantly increase the level of trust and reciprocity at the individual level. In previous laboratory experiments, scholars often observe a decrease of trustworthiness in the last few rounds of the baseline trust game. This phenomenon is referred as the end-game effect [29]. Yet in some cases of the two-level trust game, full trustfulness and full trustworthiness can be obtained such that the end-game effect can be alleviated. This is mainly due to two reasons. One is that the conditions for trust placement is relaxed—a high trustworthiness of either the trustee or the authority is sufficient for the trustor to place trust. The other reason is that the condition for the B-type trustee to honour trust is relaxed. Specifically, the imposition of the authority creates an additional incentive for the B-type trustee to pretend to be a G-type since he is afraid to be punished when the *ex ante* belief ${\pi_1}^A$ is sufficiently high. Therefore, the incomplete information with regard to the type of the authority reshapes each actor's equilibrium strategies. It produces favourably impacts on conditions for trust and reciprocity.

Lastly, the two-level set up of the trust game brings new factors, including ψ, R_3 and C_3, into the baseline model. They all have different effects on the sequential equilibrium. The amount of centralized sanction imposed on defective trustees plays a key role in determining equilibrium strategies for both the trustor and the trustee. When the *ex ante* beliefs are common knowledge and fixed, the harsher the sanction is the more likely trust is placed and honoured. Yet the utilities associated with the authority's reward R_3 and costs C_3 only play a peripheral role of determining the randomization period for the trustor.

Despite the above theoretical interest in the effects of centralized sanctioning and incomplete information on human cooperation, empirical evidence testing these theories is limited. Pluralistic methods, including experimental approaches, are needed to develop more comprehensive, accurate and well-specified explanations of these game theoretic postulations [57,58]. Testable hypotheses could be generated from the formal two-level trust game and examined in experiments in future research. For instance, Buskens et al. [29,55,59] and Anderhub et al. [30] have analysed, in experimental settings, the sequential equilibrium of the baseline trust game with two or three actors [6,51,60]. An important finding from these studies is the end-game effect. The analyses of our two-level trust game, however, suggests that the end-game effect may be weakened by the imposition of a centralized sanctioning institution and incomplete information. This leads to the following hypotheses: 1) In the condition of the two-level trust game, compared to the condition of the baseline trust game, the likelihood of trustworthiness of the trustor and trustworthiness of the trustee is higher; 2) The likelihood of trustfulness and trustworthiness decrease slower in the last few rounds of the two-level trust game than that in the baseline trust game. By moderate modifications to previous experimental configurations by Buskens et al. and Anderhub et al., one could test these hypotheses and compare the effects of centralized sanctioning on trust and reciprocity in alternative lab experiments.

To put the two-level model in a broader context, it could shed some light on two interrelated types of trust; namely, individual trust between the trustor and trustee as well as institutional trust between the individuals and the authority. The result of the two-level game demonstrates that the imposition of an authority produces a synergetic effect on trust and reciprocity. This result, to some extent, echoes with extensive arguments concerning the relationship between "trust in the state" and "social trust" [14,61–63]. Future research could further explore such conjectures with more empirical evidence collected in the field.

Author Contributions

Conceived and designed the experiments: RYW. Performed the experiments: RYW CNN. Analyzed the data: RYW CNN. Contributed reagents/materials/analysis tools: RYW. Wrote the paper: RYW CNN.

References

1. Bachmann R (2001) Trust, Power and Control in Trans-Organizational Relations. Organization Studies 22: 337–365.
2. Ostrom E (1998) A behavioral approach to the rational choice theory of collective action. American Political Science Review 92: 1–22.
3. Arrow K (1974) The limits of organization. New York: Norton
4. Arrow K (1970) Political and Economic Evaluation of Social Effects and Externalities. In: Margolis J, editor. The Analysis of Public Output. New York: National Bureau of Economic Research, Columbia University Press
5. Ouchi WG (1980) Markets, Bureaucracies, and Clans. Administrative Science Quarterly 25: 129–141.
6. Camerer C, Weigelt K (1988) Experimental Tests of a Sequential Equilibrium Reputation Model. Econometrica 56: 1–36.
7. Kreps DM (1990) Corporate Culture and Economic Theory. In: Alt JE, Shepsle KA, editors. Perspectives on Positive Political Economy. Cambridge: Cambridge University Press. pp. 90–143.
8. Snijders C (1996) Trust and commitments: Purdue University Press.
9. Buskens V (2002) Social networks and trust. Boston: Kluwer Academic Publishers.
10. Buskens V (1998) The social structure of trust. Social Networks 20: 265–289.
11. Coleman JS (1990) Foundations of Social Theory. Cambridge, Massachusetts: Harvard University Press.
12. Buskens V, Raub W (2013) Rational choice research on social dilemmas: embeddedness effects on trust. In: Wittek R, Snijders TAB, Nee V, editors. The handbook of rational choice social research. Stanford, California: Stanford University Press.
13. Berg L (2004) Trust in food in the age of mad cow disease: a comparative study of consumers' evaluation of food safety in Belgium, Britain and Norway. Appetite 42: 21–32. PMID: 15036780
14. Rothstein B (2005) Social Traps and the Problem of Trust. Cambridge: Cambridge University Press.
15. Uslaner EM (2002) The moral foundations of trust. New York: Cambridge University Press.
16. Putnam RD (1993) Making democracy work: civic traditions in modern Italy. Princeton, NJ: Princeton University Press.
17. Norris P (2002) Democratic Phoenix: Reinventing Political Activism. Cambridge, UK: Cambridge University Press.
18. Lane C, Bachmann R (1998) Trust within and between organizations: conceptual issues and empirical applications. New York: Oxford University Press.
19. Lorenz EH (1988) Neither friends nor strangers: Informal networks of subcontracting in French industry. In: Gambetta D, editor. Trust: making and breaking cooperative relations. Oxford, UK: Oxford University Press. pp. 194–210.
20. Nooteboom B, Berger H, Noorderhaven NG (1997) Effects of trust and governance on relational risk. Academy of Management Journal 40: 308–338.
21. Ratnasingam P (2003) Inter-Organizational Trust for Business to Business E-Commerce. Hershey, PA: IRM Press.

22. Ring PS, van de Ven AH (1992) Structuring cooperative relationships between organizations. Strategic Management Journal 13: 483–498.

23. Woolthuis RK (1999) Sleeping with the Enemy: Trust, Dependence and Contracts in Interorganisational Relationships. Enschede, the Netherlands: Universiteit Twente.

24. Edelenbos J, Eshuis J (2012) The Interplay Between Trust and Control in Governance Processes: A Conceptual and Empirical Investigation. Administration & Society 44: 647–674.

25. Zucker LG (1986) Production of trust: Institutional sources of economic structure, 1840–1920. Research in Organizational Behavior 8: 53–111.

26. Lindgreen A (2003) Trust as a valuable strategic variable in the food industry: Different types of trust and their implementation. British Food Journal 105: 310–327.

27. Ng CN, Wang RY, Zhao TJ (2013) Joint Effects of Asymmetric Payoff and Reciprocity Mechanisms on Collective Cooperation in Water Sharing Interactions: A Game Theoretic Perspective. Plos One 8.

28. Buskens V, Raub W (2002) Embedded trust: Control and learning. Advances in Group Processes 19: 167–202.

29. Buskens V, Raub W, van der Veer J (2010) Trust in triads: An experimental study. Social Networks 32: 301–312.

30. Anderhub V, Engelmann D, Guth W (2002) An experimental study of the repeated trust game with incomplete information. Journal of Economic Behavior & Organization 48: 197–216.

31. Camerer C (2003) Behavioral game theory: Experiments in strategic interaction: Princeton University Press. PMID: 12757825

32. Gachter S, Renner E, Sefton M (2008) The Long-Run Benefits of Punishment. Science 322: 1510–1510. doi: 10.1126/science.1164744 PMID: 19056978

33. Rockenbach B, Milinski M (2006) The efficient interaction of indirect reciprocity and costly punishment. Nature 444: 718–723. PMID: 17151660

34. Boyd R, Gintis H, Bowles S (2010) Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare. Science 328: 617–620. doi: 10.1126/science.1183665 PMID: 20431013

35. O'Gorman R, Henrich J, Van Vugt M (2009) Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. Proceedings of the Royal Society B-Biological Sciences 276: 323–329. doi: 10.1098/rspb.2008.1082 PMID: 18812292

36. Fowler JH (2005) Altruistic punishment and the origin of cooperation. Proceedings of the National Academy of Sciences of the United States of America 102: 7047–7049. PMID: 15857950

37. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K (2007) Via freedom to coercion: The emergence of costly punishment. Science 316: 1905–1907. PMID: 17600218

38. Traulsen A, Rohl T, Milinski M (2012) An economic experiment reveals that humans prefer pool punishment to maintain the commons. Proceedings of the Royal Society B-Biological Sciences 279: 3716–3721. doi: 10.1098/rspb.2012.0937 PMID: 22764167

39. Baldassarri D, Grossman G (2011) Centralized sanctioning and legitimate authority promote cooperation in humans. Proceedings of the National Academy of Sciences of the United States of America 108: 11023–11027. doi: 10.1073/pnas.1105456108 PMID: 21690401

40. Sigmund K, De Silva H, Traulsen A, Hauert C (2010) Social learning promotes institutions for governing the commons. Nature 466: 861–863. doi: 10.1038/nature09203 PMID: 20631710

41. Fehr E, Fischbacher U (2004) Social norms and human cooperation. Trends in Cognitive Sciences 8: 185–190. PMID: 15050515

42. Fehr E, Fischbacher U (2003) The nature of human altruism. Nature 425: 785–791. PMID: 14574401

43. Fehr E, Gachter S (2002) Altruistic punishment in humans. Nature 415: 137–140. PMID: 11805825

44. Fehr E, Fischbacher U (2004) Third-party punishment and social norms. Evolution and Human Behavior 25: 63–87.

45. Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. Proceedings of the National Academy of Sciences of the United States of America 100: 3531–3535. PMID: 12631700

46. Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, Bolyanatz A., et al. (2006) Costly punishment across human societies. Science 312: 1767–1770. PMID: 16794075

47. Kurzban R, DeScioli P, O'Brien E (2007) Audience effects on moralistic punishment. Evolution and Human Behavior 28: 75–84.

48. Pedersen EJ, Kurzban R, McCullough ME (2013) Do humans really punish altruistically? A closer look. Proceedings of the Royal Society B-Biological Sciences 280.

49. Bernhard H, Fischbacher U, Fehr E (2006) Parochial altruism in humans. Nature 442: 912–915. PMID: 16929297

50. Dasgupta P (1988) Trust as a Commodity. In: Gambetta D, editor. Trust: Making and Breaking Cooperative Relations. Blackwell, Oxford. pp. 49–72.

51. Bower AG, Garber S, Watson JC (1996) Learning about a population of agents and the evolution of trust and cooperation. International Journal of Industrial Organization 15: 165–190.

52. Sethi R, Somanathan E (2003) Understanding reciprocity. Journal of Economic Behavior & Organization 50: 1–27.

53. Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114: 817–868.

54. Bolton GE, Ockenfels A (2000) ERC: A theory of equity, reciprocity, and competition. American Economic Review 90: 166–193.

55. Buskens V (2003) Trust in triads: effects of exit, control, and learning. Games and Economic Behavior 42: 235–252.

56. Cook KS, Cooper RM (2003) Experimental Studies of Cooperation, Trust and Social Exchange. In: Ostrom E, Walker J, editors. Trust and Reciprocity: Interdisciplinary Lessons for Experimental Research. New York: Russell Sage. pp. 209–244.

57. Poteete AR, Janssen MA, Ostrom E (2010) Working Together: Collective Action, the Commons, and Multiple Methods in Practice. New Jersey: Princeton University Press.

58. Janssen MA, Lee A, Waring TM (2014) Experimental platforms for behavioral experiments on social-ecological systems. Ecology and Society 19.

59. Buskens V, Weesie J (2000) An experiment on the effects of embeddedness in trust situations—Buying a used car. Rationality and Society 12: 227–253.

60. Kreps DM, Wilson R (1982) Sequential Equilibria. Econometrica 50: 863–894.

61. Mishler W, Rose R (2001) What are the origins of political trust? Testing institutional and cultural theories in post-communist societies. Comparative Political Studies 34: 30–62.

62. Levi M, Stoker L (2000) Political trust and trustworthiness. Annual Review of Political Science 3: 475–507.

63. Li LJ (2013) The Magnitude and Resilience of Trust in the Center: Evidence from Interviews with Petitioners in Beijing and a Local Survey in Rural China. Modern China 39: 3–36.