Facilitating trust and trustworthiness: The impact of mediation*

Saori Chiba[†] Michiko Ogaku[‡]

30 June 2024

Abstract

This paper examines how mediation can foster trust and trustworthiness between socially distant parties, focusing on investor-receiver interactions. We analyze three mediation scenarios: an omniscient mediation, a communication device mediation, and an influential mediation affecting game dynamics. Using communication game models, we find that while omniscient mediation enhances transaction likelihood, a communication device mediation provides no advantage over unmediated interactions. An influential mediation, however, can motivate trustworthy behavior when receivers are incentivized to act contrary to recommendations. Unexpectedly, we demonstrate a case where the communication revelation principle (Sugaya and Wolitzky 2021) doesn't hold under Nash equilibrium when the mediator influences the game. This study contributes to the understanding of trust building mechanisms in socially distant transactions and has implications for game theory involving communication and practical mediation strategies.

[*] Financial support from the Trust Forum Foundation and the Kyoto University Foundation is gratefully acknowledged.

[†] Faculty of Economics, Kyoto Sangyo University, Motoyama Kamigano, Kita-ku, Kyoto, 603-8555, Japan

[‡] Faculty of Economics, Nagasaki University, 4-2-1 Katafuchi, Nagasaki, 850-8506 Japan

Key words: Trust, Trustworthiness, Bayes correlated equilibrium, Communication revelation principle, Sequential communication equilibrium

JEL: D82, D83.

1 Introduction

Large social distance[1] is, in itself, a significant obstacle to building and maintaining trust and trustworthiness.[2] Conversely, spatial and non-spatial proximity have been shown to enhance the reciprocal establishment of such trust and trustworthiness.[3] Trust and trustworthiness are key to initiating transactions and subsequently building relationships. How can we facilitate the initiation of successful transactions when the social distance between people cannot be reduced?

This paper examines how mediation can foster trust and trustworthiness between an investor and a receiver who are not socially close, thereby facilitating the initiation of transactions. We consider three scenarios with differing mediation characteristics. First, we consider an ideal case where the mediator possesses perfect knowledge of each party's type. We test this using a mechanism whose solution concept is a Bayes correlated equilibrium (Bergemann and Morris 2016). As intuition suggests and our results confirm, such omniscient mediation could significantly increase the likelihood of successful transactions.

However, mediators often lack complete information about the parties involved. Therefore, we analyze a case where the mediator does not know the parties' types

[1] This social distance could be interpreted as any socio-economic difference.

[2] Alesina and La Ferrara (2002) found that interpersonal trust was lower in more racially heterogeneous communities using data from the General Social Survey. Several authors supported the hypothesis put forward by Alesina and La Ferrara (2002). For example, Hoffman et al. (1996) provided the experimental evidence that smaller social distance lead to more other-regarding actions in dictator games. Binzel and Fehr (2013) conducted a lab-in-the-field experiment with residents of an informal settlement of Cairo. They showed that trust is higher among friends (members of a given social network) than among strangers, and that higher trust among friends is related to the trustor's belief in the other's trustworthiness.

[3] Fisman et al. (2017) provided empirical evidence that cultural proximity (shared codes, belief and ethnicity) between lenders and borrowers raises the volume of and reduces default rate using data from a state-owned bank in India.

and only acts as a communication device, as defined by Forges (1986). In this scenario, the mediator receives input (self-reported types) from the players and passes output (recommendations) to them, without direct involvement in their original game. We test this using a mechanism whose solution concept is a sequential communication equilibrium (Myerson 1986). Unfortunately, we find that the mediator's role in this case becomes insignificant, with the use of a communication device providing no advantage over the parties transacting without a mediator. This result stems from the adverse selection incentive of the bad type receiver, suggesting that if a mediator is just a communication device, the mediator should at least know the types of receivers.

Given these contrasting outcomes, we finally examine an alternative form of mediation where the mediator can directly influence the original game by justifying or discounting parties' actions. This influential mediator can be modeled by adapting a mediator in Sugaya and Wolitzky (2021). We modify it so that the receiver is motivated to disobey recommendations. In the presence of a large social distance where the decision to betray is rational for the receiver, behaving trustworthily after being recommended to betray provides a positive surprise for the investor and psychological benefits for the receiver. Under this assumption, we show that the mediator could motivate trustworthy behavior, albeit because good actions become more meaningful when not simply following recommendations.

Our analysis of the last case yields an unexpected result. We find that our model could produce an example where the communication revelation principle, defined by Sugaya and Wolitzky (2021), does not hold under the solution concept of Nash equilibrium. The communication revelation principle is an extension of the revelation principle for multistage games with communications. This violation occurs due to the mediator's recommendation inducing the receiver to act contrary to expectations. Indeed, the principle requires that players canonically follow the recommendations. If the mediator is just a communication device, this violation does not occur in our example.

Using a mediator who could potentially influence the game dynamics, Sugaya and Wolitzky (2021) showed that the communication revelation principle does not hold under the solution concept of sequential communication equilibrium, although it does hold under the solution concept of Nash equilibrium. According to our finding, if the mediator acts solely as a communication device, "does the communication revelation principle still hold under the solution concept of sequential communication equilibrium?" is still an open question.

This paper is structured as follows. Section 2 describes our basic game. Section 3 presents results of the basic game, where the players choose their actions without any communication. Section 4 provides results of the mediated game, examining three cases: (1) where the mediator knows the types perfectly, (2) where the mediator is only a communication device, and (3) where the mediator could potentially influence the game directly. Section 5 concludes. All the proofs are in the appendix.

2 Model

We use a game with spacial matching adapted from Tabellini (2008) and Okada (2020). The game involves two roles of players: an investor (or player 1) and a receiver (or player 2). Players of each type are continuously distributed along a line segment of size $\overline{y}$. Each player is randomly matched with a player of the opposite role. Their distance, denoted by the random variable Y on a probability space (Ω, Σ, P) with the values in $(0, \overline{y}]$, follows a distribution with probability density function $g(y)$.

The matched players may initiate a transaction. Each player has two possible actions: player 1 chooses to invest (TR) or not invest (N), while player 2 chooses to return some rewards (TW) or not (BE). Let A_i be the set of actions of player i and $A := A_1 \times A_2$.

Players could obtain both material payoffs and psychological benefits from their transaction. Table 1 presents the material payoffs, where $\lambda > 0$ is player 1's initial endowment, $\Lambda\lambda$ is player 2's gross earning upon receiving λ, and $\alpha \in (0, 1)$ is the allocation rule for the gross earnings.

	TW	BE
TR	$\alpha\Lambda\lambda$, $(1-\alpha)\Lambda\lambda$	0, $\Lambda\lambda$
N	λ, 0	λ, 0

Table 1: Material payoffs

If players were to consider only material payoffs, player 2 would have an incentive to betray (choose BE), which would lead player 1 to choose N. Consequently, (N, BE) would be the equilibrium.

Psychological benefits represent concepts of trust and trustworthiness. We make the following assumption:

Assumption 1: $d_1 > \lambda,\ d_2 > \alpha\lambda\Lambda > \lambda$.

This assumption gives the possibility of successful transactions depending on the social distance between players, where player 1 chooses TR and player 2 chooses TW. The value of psychological benefits for each player i is $d_i e^{-\theta_i y}$, where d_i is the benefit value in a match with zero distance, and $\theta_i > 0$ is the decaying rate. These benefits are earned regardless of the other player's action and make (TR, TW) the socially optimal outcome.

Let $u_i : A_i \times A_{-i} \times \Theta \to \mathbb{R}$ be the utility function of player i, where $\Theta := \{g, b\} \times \{g, b\} \times (0, \bar{y}]$ is the set of all states. Each element $\theta \in \Theta$ comprises of player 1's type, player 2's type (defined later), and their distance y, respectively.

For $(a_i, a_{-i}, \theta) \in A_i \times A_{-i} \times \Theta$, we define:

$$u_i(a_i, a_{-i}, \theta) = \zeta_i(a_i, a_{-i}) + d_i e^{-\theta_i y} \delta_i,$$

where $\zeta_i(a_i, a_{-i})$ is the material payoff for player i when $(a_i, a_{-i}) \in A_i \times A_{-i}$ is chosen, and δ_i is 1 if a_i is TR or TW, and 0 if otherwise. Under Assumption 1, if the social distance y is sufficiently small, TR and TW become dominant strategies for the players.

We further assume that each player has a type: "good" (g) or "bad" (b). Hereafter we use "good (bad) type" or "type (j = good, bad)" to describe player type. The proportion of good players among players i is n_i. Good players are more tolerant of social distance than bad players. Let θ_i^j be the decay rate for player i of type j, with $0 < \theta_i^g < \theta_i^b$. Consequently, good players choose TR and TW over a wider interval of y.

The basic game G consists of: (i) the set of actions A_i and utility function u_i for each player i and (ii) a common prior $\psi \in \Delta_{++}(\Theta)$.

We also define signals on types. Let $T := T_1 \times T_2$ be a set of random vectors on the probability space (Ω, Σ, P). Each element $t \in T$ and Y are independent. Let π be the distribution of $t \in T$. For each i, $T_i := T_i^1 \times T_i^2$ denotes the set of signals that player i receives at the game's start. A typical element is $t_i = (t_i^1, t_i^2) \in T_i$, where t_i^1 is a perfect signal about player i's type, and t_i^2 is a signal about the other player $-i$'s type. Although t_1^1 and t_2^1 are independent of each other, t_1^2 and t_2^2 are potentially correlated.

Let S^1 be an information structure composed of (T^1, π^1), given at the game's start. This is the information structure studied in Okada (2020). We also consider an information structure composed of $T = (T, \pi)$, which we'll use when discussing a mediator-provided information structure. The two tuple (G, S^1) is a standard incomplete information game played by the players in our model.

Let $\beta_i : T_i \times (0, \overline{y}] \to \Delta(A_i)$ be a strategy for player i in a match with distance $y \in (0, \overline{y}]$ when given signals $t_i \in T_i$, where $\Delta(A_i)$ is the set of all Borel probability measures on A_i. Let $B_i(C|t_{-i}, y)$ be the average probability that an individual in the population of player i chooses either TR or TW in a match with distance y.

Player 1 of type j is indifferent between choosing TR and N if and only if:

$$\alpha\Lambda\lambda B_2(C|t_1, y) + d_1 e^{-\theta_1^j y} = \lambda. \tag{1}$$

The distance y satisfying (1) is given by:

$$y = I_1^j(B_2(C|t_1, y)) = \frac{1}{\theta_1^j} \log \frac{d_1}{\lambda(1 - \alpha\Lambda B_2(C|t_1, y))}.$$

Similarly, for player 2, given $B_1(C|t_2, y)$, the distance y that makes player 2 indifferent between choosing TW and BE is:

$$y = I_2^j(B_1(C|t_2, y)) = \frac{1}{\theta_2^j} \log \frac{d_2}{\lambda(\alpha\Lambda B_1(C|t_2, y))}.$$

We consider that these $I_i^j(\cdot)$ as mappings from $[0, 1]$ to $\mathbb{R}_+$. Let $I_i^j(p) = +\infty$ if $I_i^j(p)$ cannot be defined for a given $p \in [0, 1]$. We call $I_i^j(\cdot)$ a trust threshold function. Let $\nu_i^j(y)$ be the inverse function of I_i^j, $\nu_i^j(y) = (I_i^j)^{-1}(y)$ for $y \in (0, \overline{y}]$. We define $\nu_1^j(y) = 0$ for $y < I_1^j(0)$ and $\nu_2^j(y) = 0$ for $y < I_2^j(1)$.

Note that I_1^j is increasing in the probability $B_2(C|t_1^1, t_1^2, y)$ and I_2^k is decreasing in the probability $B_1(C|t_2^1, t_2^2, y)$. Furthermore, since good players have lower decay rates, the inequality $I_i^g > I_i^b$ holds for $i = 1, 2$.

3 Unmediated game

To understand the results without mediation, we can refer to Okada (2020). Okada (2020) demonstrates the Bayes Nash equilibrium of the incomplete information game (G, S^1). A partial excerpt of his findings is as follows. Let

$(a,b)_+ := \max(a,b)$ and $(a,b)_- := \min(a,b)$.

Proposition 1. *(Okada 2020) Suppose Assumption 1 holds. For* $(I_1^b(n_2), I_2^b(1))_+ < y < I_2^b(n_1)$*:*

- *Good players 1 and 2 take* TR *and* TW *respectively with probability* 1
- *Bad players employ mixed strategies:*
  - *Bad player 1 chooses* TR *with probability* $(\nu_2^b(y) - n_1)/(1 - n_1)$
  - *Bad player 2 chooses* TW *with probability* $(\nu_1^b(y) - n_2)/(1 - n_2)$

As we will demonstrate in the following section, introducing a mediator with perfect knowledge of each player's type improves upon this outcome.

4 Mediated game

We now consider the case where (G, S^1) is played with mediation. Before the game, the mediator privately recommends actions to each player. Each player can either follow or dismiss this recommendation. The optimal condition occurs when the mediator perfectly knows each player's type and aims to maximize social welfare, which in this model achieved when the action pair (TR, TW) is played. We first consider this ideal case. The equilibrium concept used is Bayes correlated equilibrium, as defined by Bergemann and Morris (2016).

4.1 Case: Mediator with perfect knowledge of types

The mediator's recommendations are defined as follows:

Let $\sigma : \Theta \to \Delta(A_1 \times A_2)$ be the mediator's decision rule in the game (G, S^1). Recommendations follow this the decision rule σ. The mediator's goal is to increase the probability of the action pair (TR, TW) being played. We assume the mediator only recommends actions that players will follow, aligns with our equilibrium concept.

Definition 1. (Bayes correlated equilibrium) A decision rule σ is a Bayes correlated equilibrium of (G, S^1) if it is obedient for (G, S^1).

In Definition 1, obedience satisfies the following constraints.

Definition 2. (Obedience) A decision rule σ is obedient for (G, S^1) if, for $i \in \{1, 2\}$, $t_i^1 \in \{g, b\}$ and $a_i \in A_i$, we have:

$$\sum_{a_{-i}, t_{-i}^1} \pi^1(t_i^1, t_{-i}^1)\sigma(a_i, a_{-i}|\theta)u_i(a_i, a_{-i}, \theta)$$
$$\geq \sum_{a_{-i}, t_{-i}^1} \pi^1(t_i^1, t_{-i}^1)\sigma(a_i, a_{-i}|\theta)u_i(a_i', a_{-i}, \theta)$$

for all $a_i' \in A_i$.

The optimal recommendation is given as a solution to the following mediator's problem:

Problem A: For each y for $t_1^1, t_2^1 \in \{g, b\}$,

$$\max_{\sigma(TR, TW|t_1^1, t_2^1, y)} \sum_{t_1^1, t_2^1 \in \{g, b\}} \pi^1(t_1^1, t_2^1)\sigma(TR, TW|t_1^1, t_2^1, y)$$

subject to the obedience constraints.

We focus on the case where the mediator's recommendation is potentially effective, in particular when the players' social distance is in the interval $(I_2^b(1), I_2^b(n_1)]$, with the assumption of $I_1^g(0) < I_2^b(1)$.

Assumption 2: $I_1^g(0) < I_2^b(1)$.

Since social distance y is publicly observable, it often determines action choices. For very small y affecting both players ($y \leq I_2^b(1)$), (TR, TW) is achieved with probability one without mediation. For y that is very small only from the perspective of player 1 ($y \in (I_2^g(1), I_1^b(0)]$), (TR, BE) is inevitable even with mediation. We focus on $y > \max(I_1^g(0), I_2^b(1))$, where mediation can significantly impact outcomes. In this range, good players can potentially choose (TR, TW) with probability one, while bad players could choose (N, BE).

Proposition 2. *Suppose Assumptions 1 and 2 hold. For players with social distance $y \in (I_2^b(1), I_2^b(n_1)]$ playing (G, S^1), the mediator's optimal recommendations are as follows:*

If the relative weights of the indifference probabilities satisfy the inequality

$$\nu_1^b(y) \geq \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}, \tag{2}$$

then

$$\begin{aligned}
\sigma(TR, TW|\theta_1) &= 1,\ \sigma(TR, BE|\theta_1) = 0, \\
\sigma(TR, TW|\theta_2) &= \nu_1^b(y) - \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}, \\
\sigma(TR, BE|\theta_2) &= 1 - \nu_1^b(y) + \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}, \\
\sigma(TR, TW|\theta_3) &= \frac{\nu_2^g(y) - n_1}{1-n_1},\ \sigma(TR, BE|\theta_3) = 0, \\
\sigma(N, BE|\theta_3) &= 0,\ \sigma(N, TW|\theta_3) = \frac{1-\nu_2^g(y)}{1-n_1}, \\
\sigma(TR, TW|\theta_4) &= \frac{\nu_2^b(y) - n_1}{1-n_1}\left[\nu_1^b(y) - \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}\right], \\
\sigma(N, TW|\theta_4) &= \frac{1-\nu_2^b(y)}{1-n_1} \cdot \left[\nu_1^b(y) - \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}\right], \\
\sigma(TR, BE|\theta_4) &= \frac{\nu_2^b(y) - n_1}{1-n_1}\left[1 - \nu_1^b(y) + \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}\right], \\
\sigma(N, BE|\theta_4) &= \frac{1-\nu_2^b(y)}{1-n_1}\left[1 - \nu_1^b(y) + \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}\right],
\end{aligned}$$

where $\theta_1 = (g, g, y)$, $\theta_2 = (g, b, y)$, $\theta_3 = (b, g, y)$, $\theta_4 = (b, b, y)$, respectively.

The optimal recommendation in Proposition 2 advises the bad player 2 to choose TW more often than in his mixed strategy. Additionally, since $\nu_1^b(y) \geq n_2$ implies (2), good players 1 and 2 are recommended to choose TR and TW

respectively with probability one for a larger interval of social distance than in the unmediated game. This leads to the following result:

Corollary 1. *The optimal recommendation in the mediated game given in Proposition 2 provides a higher probability of a successful transaction than the unmediated game shown in Proposition 1.*

However, as we will see in the following subsections, if the mediator does not know each player's type, the mediation effect will be limited.

4.2 Case: Mediator as a communication device

We now consider the case where the mediator does not know the players' types and is not directly involved in their transactions. This scenario aligns with the concept of a communication device as described by Forges (1986), where the mediator receives inputs from players and passes output to players without directly participating in the original game. We demonstrate that in this case, a canonical communication equilibrium (as defined in Forges (1986)) is achieved, with the mediator only recommending the actions that players would have taken without mediation.

To address potential ambiguity in player identification when the mediator functions solely as a communication device, we assume a fixed communication order: Player 1 privately communicates with the mediator first, followed by player 2. As in the previous case, we focus on the scenarios where the players' social distance falls within the interval $(I_2^b(1), I_2^b(n_1)]$, maintaining the assumption that $I_1^g(0) < I_2^b(1)$.

The mediator's problem, requiring inputs (type reporting) to provide outputs (recommendations), is formulated as follows.

Problem B: For each y for $t_1^1, t_2^1 \in \{g, b\}$,

$$\max_{\sigma(TR,TW|t_1^1,t_2^1,y)} \sum_{t_1^1,t_2^1 \in \{g,b\}} \pi^1(t_1^1, t_2^1)\sigma(TR, TW|t_1^1, t_2^1, y)$$

subject to truth-telling and obedience constraints.

The truth-telling constraints are defined as:

Definition 3. (Truth-telling constrants) A decision rule σ for (G, S^1) induces players' truth telling if, for $i \in \{1, 2\}$, $t_i^1 \in \{g, b\}$, we have:

$$\sum_{a_i,a_{-i},t_{-i}^1} \pi^1(t_i^1, t_{-i}^1)\sigma(a_i, a_{-i}|t_i^1, t_{-i}^1, y)u_i(a_i, a_{-i}, t_i^1, t_{-i}^1, y)$$
$$\geq \sum_{a_i,a_{-i},t_{-i}^1} \pi^1(t_i^1, t_{-i}^1)\sigma(a_i, a_{-i}|t_i^{1\prime}, t_{-i}^1, y)u_i(a_i^\prime, a_{-i}, t_i^1, t_{-i}^1, y)$$

for all $a_i^\prime \in A_i$ and $t_i^{1\prime} \in \{g, b\}$.

The solution to Problem B can be viewed as a one-period version of the sequential communication equilibrium described in Myerson (1986). While not explicitly stated, it is understood that the mediator's recommendations are restricted to $\{TR\}$ for good player 1 and $\{TW\}$ for good player 2, aligning with the rule of recommendations in Myerson (1986) that excludes codominated actions.

Problem B does not explicitly describe the conditional probability systems introduced in Myerson (1986). However, it is known that any sequential equilibrium in the sense of Kreps and Wilson (1982) is a sequential communication equilibrium as defined by Myerson (1986). Therefore, it suffices to show that the solution to Problem B can be viewed as a one-period version of sequential equilibrium. This can be demonstrated as follows: Proposition 4 shows that the equilibrium of Problem B is equivalent to the Bayes Nash equilibrium of a one-period game without a mediator, where two players can have two possible types independently. Furthermore, the assessment at the terminal nodes can be considered reasonable in the sense of Fudenberg and Tirole (1991), and in this case, the concept of reasonable assessments is equivalent to the concept of consistent assessment as defined by Kreps and Wilson (1982). As shown by Fudenberg and Tirole (1991), this implies the equivalence of perfect Bayesian equilibrium and sequential equilibrium, where the assessments are considered only at the terminal nodes.

If the mediator ignores the truth-telling constraints and maintains the optimal recommendation from Proposition 2, the bad player 2 has an incentive to misreport their type as good and then choose N with probablity one. The truth-telling and the obedience constraints ensure that it is always rational for player i of type j to be honest and obedient at both the reporting and action stages, provided they have not previously lied.

However, the addition of truth-telling constraints renders the mediator ineffective, as shown in the following propositions.

Proposition 3. *The optimal decision rule σ reduces to the strategy profile in the unmediated game shown in Proposition 1.*

This result is intuitive: the rationality of honesty for bad player 2 requires that player 1 chooses TR with the same probability against both good and bad player 2. This can be implemented when player 1 chooses the strategy without knowing player 2's type. For a successful transactions, it is better not to inform bad player 2 of player 1's type. If uninformed, bad player 2 chooses the mixed stragegy for both types of player 1, which is sufficient for good players 1 and 2 to choose TR and TW respectively with probability one.

More generally, when a mediator functions solely as a communication device, the optimal mediation rule is equivalent to a Bayes Nash equilibrium of (G, S^*), where S^* is either an expansion of S^1 or S^1 itself. This can be easily proved by extending the theorem by Bergemann and Morris (2016), which states that if a mediator has knowledge of the players' types, a decision rule is a Bayes correlated equilibrium if and only if there exists a Bayes Nash equilibrium of (G, S^*) that induces σ. The "if" part of this theorem still holds true even when the mediator serves as a communication device, provided that the decision rule satisfies truth-telling constraints and is a Bayes correlated equilibrium.

Proposition 4. *Suppose a mediator is simply a communication device. If a decision rule σ of the mediator for (G, S^1) satisfies truth-telling constraints and is a Bayes correlated equilibrium, then there exists a Bayes Nash equilibrium of (G, S^*) that induces σ, where S^* is either an expansion of S^1 or S^1 itself.*

Truth-telling constraints typically limit the informativeness of the expansion S^* of S^1 as seen in Proposition 4. If there are no conflicts of interest between players, these constraints may not be binding, allowing the mediator to function as if it had perfect knowledge. Otherwise, in extreme cases, players may dismiss their communication device entirely. This case corresponds to Proposition 3.

4.3 Case: Influential mediator

Given the contrasting outcomes discussed above, we now examine a case where the mediator can directly influence the original game. The mediator's recommendation justifies or discounts players' actions, potentially altering their payoffs based on whether they follow the recommended action. Under certain assumptions, we demonstrate that even if it lacks type knowledge, if it is influential a mediator could motivate trustworthy behavior, thereby facilitate successful transactions. Furthermore, we unexpectedly find that this case exemplifies a situation where the communication revelation principle, as defined by Sugaya and Wolitzky (2021), does not hold under either sequential communication equilibrium or Nash equilibrium solution concepts.

The case is as follows: The mediator's recommendation discounts the players' actions. Player 1 may greatly appreciate if player 2 chooses TW, even when the mediator must recommend BE. In this scenario, player 2 might gain extra psychological benefits by giving player 1 such a pleasant surprise. This situation does not arise with a mediator serving only as a communication device, but can be considered by adapting a mediator model proposed by Sugaya and Wolitzky (2021). In their model, the mediator is allowed to take an action in each period that directly affects the players' utilities. While it may not be natural to consider a case where the mediator invests or returns a part of profits, this action could be interpreted as a record of the mediator's recommendation. If the mediator records a recommended action for a player, this could influence whether the players follow the recommendation.

We consider a scenario where the social distance falls within the range $I_2^g(1) < y \leq I_1^b(0)$. In this range Okada (2020) demonstrated that in the absence of a mediator, players adopt pure strategies: player 1 takes TR and player 2 takes BE. Since the players adopt pure strategies, a communication device alone cannot alter the equilibrium presented by Okada (2020).

We assume a mediator who records the action recommended for player 1. While maintaining both the monetary and psychological benefits for player 1 as before, we modify the psychological benefits of player 2 as follows: Suppose that if the mediator recommends BE, if player 2 disobeys, and if this provides player 1 with a pleasant surprise, then player 2 gains φ as an extra psychological benefit such that

$$(1-\alpha)\Lambda\lambda + d_2 e^{-\theta_2^b y} + \varphi > \Lambda\lambda.$$

The monetary benefits of player 2 remain unchanged. Under these conditions, both types of player 2 will be motivated to disobey the BE recommendation.

First, let's consider the outcome of this change under the solution concept of sequential communication equilibrium. We restrict the range of the mediator's recommendation for both types of player 1 to $\{TR\}$ because N is the codominated action for them. Furthermore, we must restrict the ranges of the mediator's recommendations for both types of player 2 to empty sets. This is because both TW and BE are codominated actions for both types of player 2. If TW is recommended, they would disobey because both types of player 1 always choose TR with probability 1, given that the social distance in $I_2^g(1) < y \leq I_1^b(0)$ is too large for them to take TW. They also disobey BE because they are better off by disobeying the BE recommendation. If the range of the mediator's recommended actions is the empty set, both types of player 2 play BE. Consequently, the equilibrium is equivalent to the Bayes Nash equilibrium in the absence of the mediator. In this situation, the communication revelation principle does not hold. Truth-telling constraints are satisfied because the mediator's recommendations are the same for each type of player. Obedience constraints hold in the sense that players do not disobey the mediator's recommendation. However, the communication revelation principle requires players to canonically follow the recommended actions.

Next, let's examine the outcome of this change under the Nash equilibrium solution concept. In this analysis, we assume that all players are good types, i.e., $n_1 = n_2 = 1$. We do not restrict the range of the mediator's recommendations. If the mediator were only a communication device, it would be classified as autonomous, as defined in Forges (1986), where the mediator does not receive any inputs.

Under these assumptions, we observe that player 2 deviates from both TW and BE recommendations. Although the mediator may recommend BE because it guarantees a transaction success with probability one, this recommendation fails to meet the criteria of the communication revelation principle, which limits the actions to those that canonically follow the mediator's recommendations.

Proposition 5. *The communication revelation principle does not hold under the Nash equilibrium solution concept if the mediator is allowed to take an action in each period.*

If we interpret the revelation principle as a principle stating that a social choice function defined on the set of types is implementable with an induced direct mechanism, such a principle holds for this example. Knowing that player 2 is contrarian, the mediator can induce the most preferable outcome by recommending the opposite action to player 2 than the one the mediator wishes them to take, when recommending to player 1 the action that the mediator wishes them to take.

5 Conclusion

This paper applies various communication games with a mediator to an investment game where trust and trustworthiness are crucial. Our findings provide multiple important implications. First, when the mediator knows the types of the investor and receiver, the mediation increases the probability of successful transactions, even without mandatory compliance. Second, however, if the mediator lacks knowledge of the parties' types, the effect is limited. When acting solely as a communication device without influencing the original game, the mediation effect disappears and the recommendations fail to convey useful information for improving transactions. This suggests that, in order to facilitate transactions where trust is paramount, a communication device mediator should be selected from among those who, at minimum, know the receivers' types. Third, when the mediator can influence the parties' original game, a positive mediation effect can be expected. We specifically examine cases where a mediator's recommendation discounts the players' actions, and where the trustworthy behavior of disobeying a betrayal recommendation provides psychological benefits to the receiver. Successful mediation occurs when the mediator recommends to the receiver the opposite action than desired, while recommending the desired action to the investor. Finally, and unexpectedly, our model demonstrates that the communication revelation principle may not hold under the Nash equilibrium solution concept when the mediator is influential, which could change the payoffs of the parties' original game. This study contributes to the understanding of mechanisms that foster trust and trustworthiness in socially distant transactions, with implications for both communication game theory and practical mediation strategies.

References

Alberto Alesina and Eliana La Ferrara. Who trusts others? *Journal of Public Economics*, 85(2):207–234, 2002.

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016.

Christine Binzel and Dietmar Fehr. Social distance and trust: Experimental evidence from a slum in cairo. *Journal of Development Economics*, 103:99–106, 2013.

Raymond Fisman, Daniel Paravisini, and Vikrant Vig. Cultural proximity and loan outcomes. *American Economic Review*, 107(2):457–492, 2017.

Francoise Forges. An approach to communication equilibria. *Econometrica: Journal of the Econometric Society*, pages 1375–1385, 1986.

Drew Fudenberg and Jean Tirole. Perfect bayesian equilibrium and sequential equilibrium. *journal of Economic Theory*, 53(2):236–260, 1991.

Elizabeth Hoffman, Kevin McCabe, and Vernon L. Smith. Social distance and other-regarding behavior in dictator games. *American Economic Review*, 86(3): 653–660, 1996.

David M Kreps and Robert Wilson. Sequential equilibria. *Econometrica: Journal of the Econometric Society*, pages 863–894, 1982.

Roger B Myerson. Multistage games with communication. *Econometrica: Journal of the Econometric Society*, pages 323–358, 1986.

Akira Okada. The cultural transmission of trust and trustworthiness. *Journal of Economic Behavior & Organization*, 169:53–69, 2020.

Takuo Sugaya and Alexander Wolitzky. The revelation principle in multistage games. *The Review of Economic Studies*, 88(3):1503–1540, 2021.

Guido Tabellini. The scope of cooperation: Values and incentives. *Quarterly Journal of Economics*, 123(3):905–950, 2008.

Appendix A Proofs

Proof of Proposition 2

To reduce the notation for fixed y, let $\theta_1 = (g, g, y)$, $\theta_2 = (g, b, y)$, $\theta_3 = (b, g, y)$, $\theta_4 = (b, b, y)$, respectively. Write

$$q_i = \sigma(TR, TW|\theta_i),\ r_i = \sigma(TR, BE|\theta_i),\ s_i = \sigma(N, BE|\theta_i),\ t_i = \sigma(N, TW|\theta_i).$$

Write the average probabilities of action pairs (TR, TW), (TR, BE), (N, BE), (N, TW) from the perspective of player 2 as follows: For the pairs $(i, j) = (1, 3), (2, 4)$

$$\overline{q}_{ij} := n_1 q_i + (1 - n_1) q_j,\ \overline{r}_{ij} := n_1 r_i + (1 - n_1) r_j$$
$$\overline{s}_{ij} := n_1 s_i + (1 - n_1) s_j,\ \overline{t}_{ij} := n_1 t_i + (1 - n_1) t_j.$$

Similarly, write the average probabilities of (TR, TW), (TR, BE), (N, BE), (N, TW) from the perspective of player 1 as follows: For the pairs $(i, j) =$

$(1,2),(3,4)$

$$\overline{q}^1_{ij} := n_2 q_i + (1-n_2)q_j,\ \overline{r}^1_{ij} := n_2 r_i + (1-n_2)r_j$$
$$\overline{s}^1_{ij} := n_2 s_i + (1-n_2)s_j,\ \overline{t}^1_{ij} := n_2 t_i + (1-n_2)t_j.$$

The mediator's problem is rewritten as for any fixed y

$$\max_{\sigma(TR,TW|t^1,y)} n_2\overline{q}_{13} + (1-n_2)\overline{q}_{24}$$

subject to

$$\overline{s}^1_{12} \geq (1-\nu^g_1(y))(\overline{s}^1_{12} + \overline{t}^1_{12}) \tag{A.3}$$
$$\overline{q}^1_{12} \geq \nu^g_1(y)(\overline{q}^1_{12} + \overline{r}^1_{12}) \tag{A.4}$$
$$\overline{s}^1_{34} \geq (1-\nu^b_1(y))(\overline{s}^1_{34} + \overline{t}^1_{34}) \tag{A.5}$$
$$\overline{q}^1_{34} \geq \nu^b_1(y)(\overline{q}^1_{34} + \overline{r}^1_{34}) \tag{A.6}$$
$$\overline{q}_{13} \leq \nu^g_2(y)(\overline{q}_{13} + \overline{t}_{13}) \tag{A.7}$$
$$\overline{s}_{13} \leq (1-\nu^g_2(y))(\overline{r}_{13} + \overline{s}_{13}) \tag{A.8}$$
$$\overline{q}_{24} \leq \nu^b_2(y)(\overline{q}_{24} + \overline{t}_{24}) \tag{A.9}$$
$$\overline{s}_{24} \leq (1-\nu^b_2(y))(\overline{r}_{24} + \overline{s}_{24}) \tag{A.10}$$
$$q_i, r_i, s_i, t_i \in [0,1]$$
$$1 = q_i + r_i + s_i + t_i \quad \text{for } i = 1,2,3,4.$$

The constraints from (A.3) to (A.10) are obedience constraints.

Proof of Proposition 2. Consider the case of $I^b_2(1) < y \leq I^b_2(n_1)$. Since larger values of $\overline{q}_{13}$ and $\overline{q}_{24}$ increase the value of objective function, the constraints (A.7) and (A.9) are binding. Write

$$h = \overline{q}_{13} + \overline{t}_{13} \text{ and } h' = \overline{q}_{24} + \overline{t}_{24}.$$

Then, from (A.7) and (A.9)

$$\overline{q}_{13} = \nu_2^g(y)h, \quad \overline{t}_{13} = (1 - \nu_2^g(y))h,$$
$$\overline{q}_{24} = \nu_2^b(y)h', \quad \overline{t}_{24} = (1 - \nu_2^b(y))h'.$$

Since $\overline{r}_{13} + \overline{s}_{13} = 1 - h$ and $\overline{r}_{24} + \overline{s}_{24} = 1 - h'$, using parameters $Y \geq \nu_2^g(y)$ and $Y' \geq \nu_2^b(y)$, $\overline{s}_{13}$, $\overline{r}_{13}$, $\overline{s}_{24}$, $\overline{r}_{24}$ could be written as

$$\overline{s}_{13} = (1 - Y)(1 - h), \quad \overline{r}_{13} = Y(1 - h),$$
$$\overline{s}_{24} = (1 - Y')(1 - h'), \ \overline{r}_{24} = Y'(1 - h').$$

Similarly, let $l = \overline{s}_{34}^1 + \overline{t}_{34}^1$, and using parameter $X \geq \nu_1^b(y)$ and $X' \leq \nu_1^b(y)$, $\overline{s}_{34}^1$, $\overline{t}_{34}^1$, $\overline{q}_{34}^1$, $\overline{r}_{34}^1$ are written as

$$\overline{s}_{34}^1 = (1 - X')l, \ \overline{t}_{34}^1 = X'l$$
$$\overline{q}_{34}^1 = X(1 - l), \quad \overline{r}_{34}^1 = (1 - X)(1 - l).$$

Show (A.3) is binding: Using the relations $n_1\overline{s}_{12}^1 = n_2\overline{s}_{13} + (1-n_2)\overline{s}_{24} - (1-n_1)\overline{s}_{34}^1$ and $n_1\overline{t}_{12}^1 = n_2\overline{t}_{13} + (1 - n_2)\overline{t}_{24} - (1 - n_1)\overline{t}_{34}^1$, (A.3) is rewritten as

$$\begin{aligned}(1 - n_1)l(X' - \nu_1^g(y)) &+ \Big[n_2(1 - Y) + (1 - n_2)(1 - Y')\Big]\nu_1^g(y) \\ &\geq \Big[\nu_1^g(y)n_2(1 - Y) + (1 - \nu_1^g(y))n_2(1 - \nu_2^g(y))\Big]h \\ &+ \Big[\nu_1^g(y)(1 - n_2)(1 - Y') + (1 - \nu_1^g(y))(1 - n_2)(1 - \nu_2^b(y))\Big]h' \quad (*)\end{aligned}$$

To increase the value of the objective function, which is increasing in h and h', for given l, X', Y and Y', the pairs of h and h' rise until the above equation holds. Thus, (A.3) is binding.

Show $X' = \nu_1^b(y)$, $Y = \nu_2^g(y)$, $Y' = \nu_2^b(y)$ and $l = n_2(1-\nu_2^g(y)) + (1-n_2)(1-\nu_2^b(y))/(1-n_1)$: From binding $(*)$, both h and h' are increasing in X'. So $X' = \nu_1^b(y)$.

From binding $(*)$, the partial derivative of h with respect to $(1-Y)$ is positive for $h' \geq \nu_1^g(y)$, $Y' \geq \nu_2^b(y)$, and that with respective to $(1-Y')$ is also positive. So the smaller the values of Y and Y', the larger the value of h. Similar arguments hold for h' as well. Thus, $Y = \nu_2^g(y)$ and $Y' = \nu_2^b(y)$.

Substituting $X' = \nu_1^b(y)$, $Y = \nu_2^g(y)$ and $Y' = \nu_2^b(y)$ into $(*)$, we obtain

$$(1-n_1)(\nu_1^b(y)-\nu_1^g(y))l = n_2(1-\nu_2^g(y))(h-\nu_1^g(y))+(1-n_2)(1-\nu_2^b(y))(h'-\nu_1^g(y)). \quad (**)$$

On the other hand, from $n_1\bar{s}_{12}^1 + n_1\bar{t}_{12}^1 \geq 0$,

$$l \leq \frac{n_2(1-\nu_2^g(y)) + (1-n_2)(1-\nu_2^b(y))}{1-n_1}. \quad (***)$$

Since from $(**)$, l should be large as much as possible, the above inequality is binding at the maximum. This would be the case that $n_1\bar{s}_{12}^1 + n_1\bar{t}_{12}^1 = 0$.

Show binding $(***)$ implies $q_1 + r_1 = q_2 + r_2 = 1$, $s_1 = s_2 = t_1 = t = 2 = 0$: From $n_1\bar{s}_{12}^1 + n_1\bar{t}_{12}^1 = 0$, we have $q_1 + r_1 = q_2 + r_2 = 1$ and $s_1 = s_2 = t_1 = t_2 = 0$.

Show the objective function $n_2\nu_2^g(y)h + (1-n_2)\nu_2^b h'$ satisfying $(**)$ and binding $(***)$ is increasing in h: From $(**)$ and binding $(***)$, we have the equation

$$n_2(1-\nu_2^g(y))h+(1-n_2)(1-\nu_2^b(y))h' = \nu_1^b(y)\Big[n_2(1-\nu_2^g(y))+(1-n_2)(1-\nu_2^b(y))\Big]. \quad (\dagger)$$

Using $(\dagger)$, the objective function could be expressed as a linear function of h such that

$$n_2\nu_2^g(y)h + (1-n_2)\nu_2^b(y)h' = \\ \frac{n_2(\nu_2^g(y)-\nu_2^b(y))}{1-\nu_2^b(y)}h + (1-n_2)\nu_1^b(y)\nu_2^b(y)\left[1+\frac{n_2(1-\nu_1^g(y))}{(1-n_2)(1-\nu_2^b(y))}\right].$$

Since $\nu_2^g(y) > \nu_2^b(y)$, this function is increasing in h.

Show the optimal h, h' and corresponding X: From the equation $n_2\overline{q}_{13} + (1-n_2)\overline{q}_{24} = n_1\overline{q}_{12}^1 + (1-n_1)\overline{q}_{34}^1$, we have

$$n_1\overline{q}_{12}^1 = n_2\nu_2^g(y)h + (1-n_2)\nu_2^b(y)h' - X\Big[n_2\nu_2^g(y) + (1-n_2)\nu_2^b(y) - n_1\Big].$$

Set $q_1 = h$ and $q_2 = h'$. It does not restrict the value of objective function. (Since $s_1 = t_1 = 0$ and $h = 1$ make $r_1 = 0$, $h = 1$ is only allowed if $q_1 = h = 1$. $q_2 = h'$ can be set because $q_2 = h'$ implies $q_4 = (\nu_2^b(y) - n_1)h'/(1-n_1)$ and $y \leq I_2^b(n_1)$.) Then, using (†) we obtain

$$X[n_2\nu_2^g(y) + (1-n_2)\nu_2^b(y) - n_1] = \frac{n_2(\nu_2^g(y)-\nu_2^b(y))(1-n_1)}{1-\nu_2^b(y)}h \\ + [n_2(1-\nu_2^g(y)) + (1-n_2)(1-\nu_2^b(y))]\frac{\nu_1^b(y)(\nu_2^b(y)-n_1)}{1-\nu_2^b(y)} \quad (\dagger\dagger)$$

(a) For $\nu_1^b(y) \geq \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}$: Since the value objective function is increasing in h, set $h = 1$. Then, from (††) X is determined such that

$$X = \frac{n_2(\nu_2^g(y)-\nu_2^b(y))(1-n_1) + [n_2(1-\nu_2^g(y)) + (1-n_2)(1-\nu_2^b(y))]\nu_1^b(y)(\nu_2^b(y)-n_1)}{[n_2\nu_2^g(y) + (1-n_2)\nu_2^b(y) - n_1](1-\nu_2^b(y))} < 1. \tag{A.11}$$

The last inequality follows from the fact: On the (h, X) plane, the line (††) and the line $X = h$ intersects at $(h, X) = (\nu_1^b(y), \nu_1^b(y))$. Since the slope of (††) is less than 1, when $h = 1$, $X < h = 1$.

From (†)

$$h' = \nu_1^b(y) - \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}. \tag{A.12}$$

Show remained probabilities: Since l, h, h', X' and X are determined, q_1, r_1, q_2, r_2, q_i, r_i, s_i and t_i, $i = 3, 4$ are given by

$$\begin{aligned}
&q_1 = 1,\ r_1 = 0,\\
&q_2 = \nu_1^b(y) - \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))},\ r_2 = 1 - \nu_1^b(y) + \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}\\
&q_3 = \frac{\nu_2^g(y) - n_1}{1-n_1},\ r_3 = 0,\\
&s_3 = 0,\ t_3 = \frac{1-\nu_2^g(y)}{1-n_1},\\
&q_4 = \frac{\nu_2^b(y) - n_1}{1-n_1}\left[\nu_1^b(y) - \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}\right],\\
&t_4 = \frac{1-\nu_2^b(y)}{1-n_1}\cdot\left[\nu_1^b(y) - \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}\right],\\
&r_4 = \frac{\nu_2^b(y) - n_1}{1-n_1}\left[1 - \nu_1^b(y) + \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}\right],\\
&s_4 = \frac{1-\nu_2^b(y)}{1-n_1}\left[1 - \nu_1^b(y) + \frac{n_2(1-\nu_2^g(y))(1-\nu_1^b(y))}{(1-n_2)(1-\nu_2^b(y))}\right].
\end{aligned}$$

□

Proof of Proposition 3

The mediator's problem is rewritten as for any fixed y

$$\max_{\sigma(TR,TW|t^1,y)} n_2\overline{q}_{13} + (1-n_2)\overline{q}_{24}$$

subject to the truth telling constraints for each type of player1 and 2, obedience constraints from (A.3) to (A.10), $q_i, r_i, s_i, t_i \in [0,1]$, $1 = q_i + r_i + s_i + t_i$ for $i = 1, 2, 3, 4$.

It is easy to see that only the truth-telling constraints for the bad player 2 need to change the optimal recommendation in proposition 2. The truth-telling constraints for bad player 2 that change the optimal recommendation are written as follows.

$$\begin{aligned}&\overline{q}_{24}(1-\alpha+\alpha\nu_2^b(y))+\overline{r}_{24}+\overline{t}_{24}\alpha\nu_2^b(y)\\&\qquad\geq \overline{q}_{13}(1-\alpha+\alpha\nu_2^b(y))+\overline{r}_{13}+\overline{t}_{13}\alpha\nu_2^b(y). \qquad\text{(A.13)}\end{aligned}$$

$$\begin{aligned}&(1-\alpha)\overline{q}_{24}+\overline{r}_{24}+\alpha\nu_2^b(y)(\overline{q}_{24}+\overline{t}_{24})\\&\qquad\geq \overline{q}_{13}+\overline{r}_{13} \qquad\text{(A.14)}\end{aligned}$$

$$\begin{aligned}&(1-\alpha)\overline{q}_{24}+\overline{r}_{24}+\alpha\nu_2^b(y)(\overline{q}_{24}+\overline{t}_{24})\\&\qquad\geq (1-\alpha+\alpha\nu_2^b(y))(\overline{q}_{13}+\overline{r}_{13})+\alpha\nu_2^b(y)(\overline{t}_{13}+\overline{s}_{13}) \qquad\text{(A.15)}\end{aligned}$$

$$\begin{aligned}&(1-\alpha)\overline{q}_{24}+\overline{r}_{24}+\alpha\nu_2^b(y)(\overline{q}_{24}+\overline{t}_{24})\\&\qquad\geq \overline{q}_{13}+(1-\alpha)\overline{r}_{13})+\alpha\nu_2^b(y)(\overline{r}_{13}+\overline{s}_{13}) \qquad\text{(A.16)}\end{aligned}$$

These constraints induce bad player2 to report his type honestly. Each constraint differs in the bad player's strategy after his misreporting. (A.13) requires honest reporting and obedience is preferable to the case where he misreports and then has the strategy to obey. (A.13) requires honest reporting and obedience is preferred to the case where he misreports and has the strategy of taking N with probability 1 afterwards. (A.15) requires that this is preferred to the case where he misreports and has the strategy of taking C with probability 1 afterwards. (A.15) requires that this is preferable to the case where he misreports and has the strategy of disobeying afterwards.

Proof of Proposition 3. The fact that only the truth-telling constraints for the bad player 2 need to change the optimal recommendation in proposition 2 implies that

(A.9) is binding. Write

$$h = \overline{q}_{13} + \overline{t}_{13} \text{ and } h' = \overline{q}_{24} + \overline{t}_{24}.$$

Then, from (A.9)

$$\overline{q}_{24} = \nu_2^b(y)h',\ \overline{t}_{24} = (1 - \nu_2^b(y))h'.$$

From (A.7), using the parameter $X \leq \nu_2^g(y)$, $\overline{q}_{13}$ and $\overline{t}_{13}$ are written as

$$\overline{q}_{13} = Xh,\ \overline{t}_{13} = (1 - X)h.$$

Since $\overline{r}_{13} + \overline{s}_{13} = 1 - h$ and $\overline{r}_{24} + \overline{s}_{24} = 1 - h'$, using the parameters $Y \geq \nu_2^g(y)$ and $Y' \geq \nu_2^b(y)$, $\overline{s}_{13}$, $\overline{r}_{13}$, $\overline{s}_{24}$ and $\overline{r}_{24}$ are written as

$$\begin{aligned} \overline{s}_{13} &= (1 - Y)(1 - h),\ \overline{r}_{13} = Y(1 - h) \\ \overline{s}_{24} &= (1 - Y')(1 - h'),\ \overline{r}_{24} = Y'(1 - h'). \end{aligned}$$

Similarly, let $l = \overline{s}_{34}^1 + \overline{t}_{34}^1$, and using the parameter $X' \leq \nu_b^1(y)$, $\overline{s}_{34}^1$ and $\overline{t}_{34}^1$ are written as

$$\overline{s}_{34}^1 = (1 - X')l,\ \overline{t}_{34}^1 = X'l.$$

If $X \geq \nu_2^b(y)$, then (A.16) is binding and we have

$$\nu_2^b(y)h' + Y'(1 - h') = Xh + Y(1 - h) + (1 - h)\alpha\nu_2^b(y).$$

If $X \leq \nu_2^b(y)$, then (A.15) or (A.16) binds and we have

$$\nu_2^b(y)h' + Y'(1 - h') = (1 - \alpha)[Xh + Y(1 - h)] + \alpha\nu_2^b(y) \text{ or}$$

$$\nu_2^b(y)h' + Y'(1 - h') = Xh + Y(1 - h) + (1 - h)\alpha\nu_2^b(y).$$

In both cases, if $h' > \nu_2^b(y)$, then h is increasing in h', which makes the mediator's objective function $n_2 X h + (1 - n_2)\nu_2^b(y)h'$ increasing in h. If $h' = \nu_2^b(y)$, then the objective function is again increasing in h. So $h = 1$ could be a necessary condition for the optimal recommendation.

Suppose $h = 1$. If $X \geq \nu_2^b(y)$, then since (A.16) and (A.14) bind, we have

$$X = \nu_2^b(y)h' + Y'(1 - h'). \tag{A.17}$$

If $X < \nu_2^b(y)$, then (A.15), (A.14) or (A.16) bind. If (A.16) is binding, then we have

$$(1 - \alpha)X + \alpha\nu_2^b(y) = \nu_2^b(y)h' + Y'(1 - h').$$

However, this implies that $Y' < \nu_2^b(y)$. A contradiction. Thus, we need to have the equation (A.17).

Show that (A.3) is binding. Using the relation $n_1\bar{s}_{12}^1 = n_2\bar{s}_{13} + (1 - n_2)\bar{s}_{24} - (1 - n_1)\bar{s}_{34}^1$ and $n_1\bar{t}_{12}^1 = n_2\bar{t}_{13} + (1 - n_2)\bar{t}_{24} - (1 - n_1)\bar{t}_{34}^1$, and substituting $\bar{s}_{13} = 0$, $h = 1$ and (A.17) into (A.3), (A.3) is rewritten as

$$\begin{aligned}(X' - \nu_1^g(y))(1 - n_1)l + (\nu_1^g(y) - n_2)(1 - Y')& \\ \geq [(\nu_1^g(y) - n_2)(1 - Y') + (1 - \nu_1^g(y))(1 - \nu_2^b(y))]h'&\end{aligned} \tag{A.18}$$

To increase the value of the objective function that is increasing in h', increase h' until the equality in (A.18) holds. So (A.3) is binding.

Show $X' = \nu_1^b(y)$, $Y' = \nu_2^b(y)$ and $X = \nu_2^b(y)$. Since (A.18) is binding, h' is increasing in X'. So $X' = \nu_1^b(y)$. Also, since the partial derivative of h' with respect to $(1 - Y')$ is positive, Y' should be the smallest, and we have $Y' = \nu_2^b(y)$. From (A.17), we have $X = \nu_2^b(y)$.

Show $\overline{s}^1_{12} = \overline{t}^1_{12} = 0$ and $l = (1 - \nu^b_2(y))/(1 - n_1)$. From $n_1\overline{s}^1_{12} + n_1\overline{t}^1_{12} \geq 0$, l is bounded from above as follows.

$$l \leq \frac{n_2(1 - X) + (1 - n_2)(1 - \nu^b_2(y))}{1 - n_1} = \frac{1 - \nu^b_2(y)}{1 - n_1}. \tag{A.19}$$

Since the binding (A.18) implies that l should be as large as possible for larger h', (A.19) is binding and this is when $n_1\overline{s}^1_{12} + n_1\overline{t}^1_{12} = 0$.

Show $h' = (\nu_1(y) - n_2)/(1 - n_2)$ for $y \geq I^b_1(n_2)$. Since (A.18) and (A.19) are binding, we have

$$h' = \frac{\nu^b_1(y) - n_2}{1 - n_2}.$$

This h' is zero or positive for $y \geq I^b_1(n_2)$.

Show q_i, r_i, s_i, t_i for $i = 1, \ldots, 4$. The optimal values of parameters $h = 1$, $h' = (\nu^b_1(y) - n_2)/(1 - n_2)$ and $X = \nu^b_2(y)$ are valid for $(I^b_2(1), I^b_1(n_1))_+ < yleqI^b_2(n_1)$ if

$$\begin{aligned}
q_1 &= 1,\ r_1 = s_1 = t_1 = 0, \\
q_2 &= \frac{\nu^b_1(y) - n_2}{1 - n_2},\ r_2 = \frac{1 - \nu^b_1(y)}{1 - n_2},\ s_2 = 0,\ t_2 = 0 \\
q_3 &= \frac{\nu^b_2(y) - n_1}{1 - n_1},\ r_3 = 0,\ s_3 = 0,\ t_3 = \frac{1 - \nu^b_2(y)}{1 - n_1} \\
q_4 &= \frac{(\nu^b_2(y) - n_1)(\nu^b_1(y) - n_2)}{(1 - n_1)(1 - n_2)} \\
r_4 &= \frac{(\nu^b_2(y) - n_1)(1 - \nu^b_1(y))}{(1 - n_1)(1 - n_2)} \\
s_4 &= \frac{(1 - \nu^b_2(y))(1 - \nu^b_1(y))}{(1 - n_1)(1 - n_2)} \\
t_4 &= \frac{(1 - \nu^b_2(y))(\nu^b_1(y) - n_2)}{(1 - n_1)(1 - n_2)}.
\end{aligned}$$

This result is equivalent to the result in the silent game shown in proposition 1. □

Proof of Corollary 1

Proof. The optimal recommendation in proposition 2 is obtained under the condition that the obedience constraint inducing to take TR is not binding for bad player 1. This is true for good player 1 for $y \in (I_2^b(1), I_2^g(1))$. It implies that bad player 2 is recommended to take TW more than his mixed strategy in the silent game. In addition, since $\nu_1^b(y) \geq n_2$ implies (2), good players 1 and 2 are recommended to take TR and TW respectively with probability 1 for a larger interval of social distance. As a result, the recommendation provides a higher probability of successful transactions than in the silent game shown in proposition 1. □

Proof of Proposition 4

It is proved by a minor change of the proof of Theorem 1 of Bergemann and Morris (2016).

Proof of Proposition 3. Suppose that σ is a Bayes correlated equilibrium satisfying truth-telling constraints. Then we have for each y, i, $t_i^1 \in T^1$, $a_i \in A$ and $a_i' \in A_i$

$$\sum_{a_{-i}, t_{-i}^1} \pi^1(t_i^1, t_{-i}^1)\sigma((a_i, a_{-i})|t_i^1, t_{-i}^1, y) u_i((a_i, a_{-i}), t_i^1, t_{-i}^1, y)$$
$$\geq \sum_{a_{-i}, t_{-i}^1} \pi^1(t_i^1, t_{-i}^1)\sigma((a_i, a_{-i})|t_i^1, t_{-i}^1, y) u_i((a_i', a_{-i}), t_i^1, t_{-i}^1, y).$$

Note that $\pi(t_i^1, t_{-i}^1|y) = \pi(t_i^1, t_{-i}^1)$ from the independence of types t^1 and social distances y, and social distance y is observable for all players and the mediator. Let $S^* = ((T_i^*)_{i=1}^2, \pi^*)$ be an expansion of S^1, where $T_i^* = T_i^1 \times T_i^2$, $T_i^2 = A_i$ for $i = 1, 2$, and π^* is defined such that

$$\pi^*((t_i^1, a_i)_{i=1}^2|y) = \pi(t^1|y)\sigma(a|t^1, y) = \pi(t^1)\sigma(a|t^1, y)$$

for $t^1 \in T^1$, $a \in A$ and $y \in (0, \overline{y}]$.

Consider the strategy β_j^* for player j in the game (G, S^*) such that

$$\beta_j^*(a_j'|(t_j^1, a_j)) = \begin{cases} 1 & \text{if } a_j' = a_j, \\ 0 & \text{if } a_j' \neq a_j, \end{cases}$$

for all $t_j^1 \in T_j^1$ and $a_j \in A_j$. Then for a social distance $y \in (0, \overline{y}]$, the expected utility of player i who is observing signal (t_i^1, a_i) and chosing action $a_i' \in A_i$ in the game (G, S^*) is propotional to

$$\begin{aligned} &\sum_{a_{-i}', a_{-i}, t_{-i}^1} \pi^*((t_i^1, t_{-i}^1), (a_i, a_{-i})|y) \left(\prod_{j \neq i} \beta_j^*(a_j'|t_j^1, a_j) \right) u_i((a_i', a_{-i}'), t_i^1, t_{-i}^1, y) \\ &\quad = \sum_{a_{-i}, t_{-i}^1} \pi^*((t_i^1, t_{-i}^1), (a_i, a_{-i})|y) u_i((a_i', a_{-i}), t_i^1, t_{-i}^1, y) \\ &\quad = \sum_{a_{-i}, t_{-i}^1} \pi(t^1) \sigma(a|t^1, y) u_i((a_i', a_{-i}), t_i^1, t_{-i}^1, y). \end{aligned}$$

So the Bayes correlated equilibrium σ satisfying truth-telling constraints implies Bayes Nash equilibrium conditions under the strategy profile β^*. By its constraction, the stragegy profile β^* induces σ.

□