Agreements with Reciprocity: Co-financing and MOUs*

Dooseok Jang† Amrish Patel‡and Martin Dufwenberg§

March 20, 2018

Abstract

Institutions for co-financing agreements often exist to encourage public good investment. Can such frameworks deliver maximal investment when agents are motivated by reciprocity? We demonstrate that indeed they can, but not how one might expect. If maximal investment is impossible in the absence of the institution and public good returns are high, then an agreement signed by all parties cannot lead to full investment. However, if all parties reject the agreement, then full investment is attainable via a gentlemen's agreement or memorandum of understanding (MOU). Agreement institutions may thus do more than just facilitate the signing of binding agreements; they may play a critical role in igniting informal cooperation underpinned by reciprocity.

Keywords: Co-financing agreements, gentlemen's agreements, MOUs, public goods, reciprocity

JEL codes: C72, D03, F53, H41

*This paper extends and replaces Jang (2015). We thank Scott Barrett, Oana Borcan, Gautam Gowrisankaran, Bård Harstad, Doruk Iris, Kiryl Khalmetski, Georg Kirchsteiger, Ashley Langer, Derek Lemoine, David Reinstein, Stanley Reynolds, Alexander Sebald, Mark Stegeman, two anonymous reviewers and numerous seminar audiences for useful comments or discussions. Funding is gratefully acknowledged from the Jan Wallander and Tom Hedelius Foundation (P2012-0097-1).

†Korea Advanced Institute of Science and Technology; jangds@kaist.ac.kr

‡University of East Anglia; amrish.patel@uea.ac.uk

§University of Arizona, University of Gothenburg, CESifo; martind@eller.arizona.edu

1 Introduction

Institutions play an important role in creating the conditions for investment in public goods. Among other things, they facilitate the negotiation and enforcement of binding agreements. One common type of agreement is a co-financing, or cost-sharing, agreement; signatories make a binding commitment to co-finance each other's future public good investments. The agreement does not commit a signatory to invest in public goods per se. However, should any signatory initiate a public good investment, its co-signatories are committed to share the cost.[1] Such agreements have been used to finance critical investment in public goods, ranging from disease eradication to climate change mitigation.[2]

Theoretically, co-financing agreements can increase public good investment (cf. Varian 1994).[3] This is because a signatory can be pivotal in inducing other signatories to invest in public goods, as only with its participation would the private cost of a public good be less than the private benefit. However, full investment remains impossible. This is because when there are many signatories, an individual signatory is no longer pivotal, thus it deviates to not signing and not investing.

These insights rely on the assumption that agents care only about their material payoffs. Yet behaviour in public good contexts often exhibits conditional cooperation (e.g. Fischbacher et al. 2001), cooperating only if others do. Such behaviour can be rationalised using reciprocity theory (Rabin 1993, Dufwenberg and Kirchsteiger 2004 (D&K), Falk and Fischbacher 2006). It describes agents as having a desire to be kind to those who are kind to them, and unkind to those who are unkind to them. For example, if agent A invests, agent B may view A as kind and invest himself.

[1] These agreements are often politically more feasible than binding commitments to actually invest in public goods.

[2] In April 2016 The World Bank and The Asian Infrastructure Investment Bank signed a co-financing agreement focusing on water, transport and energy. Each party contributed $216 million to the first project, upgrading slums in Indonesia: www.brettonwoodsproject.org/2016/06/world-bank-and-aiib-signs-joint-co-financing-agreement. Such agreements are also signed by private companies. The Asian Development Bank, for instance, has an agreement with Chevron to invest in IT, construction and engineering education: www.adb.org/site/cofinancing/partners. One area where cost-sharing agreements are extensively used is in R&D investments (Katz 1986).

[3] Indeed higher investment is observed in related experimental games (Andreoni and Varian 1999, Falkinger et al. 2000 and Charness et al. 2007.)

The implications of reciprocity for public good provision are both well established (e.g. Sugden 1984) and straightforward. If agents care enough about reciprocity, maximal investment is possible, otherwise it is not. Such investment is not due to a formal agreement to invest, but rather an informal one (referred to as a gentlemen's agreement, a tacit agreement or a memorandum of understanding (MOU), for example).[4] By contrast, little is known about the implications of reciprocity for formal agreements over public goods.

An obvious question follows: How does an opportunity to strike a formal co-financing agreement perform under reciprocity? More specifically, one may wonder: Can a co-financing mechanism deliver full investment? Does such investment follow if all players sign the formal agreement? Is it impossible if all players reject the formal agreement? To answer these questions, we apply D&K's model of reciprocity to an agreements game where players choose whether or not to sign a formal cost-sharing agreement, then play a public good game. We find that if in the absence of the mechanism, full investment via an informal MOU is impossible and the public good return is high, then such investment remains impossible if all players sign. However, if all players reject the formal agreement, then full investment becomes attainable via an informal MOU. Despite not being the unique equilibrium, this outcome is both stark and surprising.

For some intuition, consider the interaction of kindness and co-financing agreements. Roughly, D&K say that agent i is kind to agent j if i could have given j a much lower payoff by changing his behaviour. Agent i deviating from a situation where all players sign and invest does reduce j's payoff, but not by much, as the cost-sharing agreement still has many signatories thus provides large investment incentives. By contrast, if i deviates from a situation where no-one signs and all invest, j's payoff is reduced considerably as there is no such cost-sharing agreement. Kindness and hence reciprocity incentives to invest in public goods are thus larger when there are no signatories than when there are many.

Our results provide several important insights. First, the existence of an institution for making binding agreements is potentially critical for triggering informal cooperation via MOUs and other informal agreements. Second, since our main result exemplifies a more general point that "high investment is possible with few signatories and impossible with many", formal agree-

[4]We shall refer to informal agreements to invest as MOUs throughout.

ments with few signatories may achieve better outcomes than those with many. Third, and pointing to the more general feature underlying the previous insights, prior stages in games (here, an agreement stage) can increase a player's influence over others' payoffs, since others may condition their actions on his early choice. This increase in payoff influence, can "amplify" psychological payoffs (in our case kindness) and make otherwise impossible outcomes attainable.

We add to the literature on agreements (e.g. Barrett 2003, Battaglini and Harstad 2016, Martimort and Sand-Zantman 2016) and an emerging literature on mechanism design where players have reciprocity preferences (Netzer and Volk 2014, Bierbrauer and Netzer 2016, Bierbrauer et al. 2017, Dufwenberg and Patel 2017).

Our particular mechanism, cost-sharing agreements, falls into a class of mechanisms where commitments on strategy-conditional side-payments are made before a game is played (Jackson and Wilkie 2005, Ellingsen and Paltseva 2016). Cost-sharing is an important case of models where agents make commitments to match others' public good investments (Guttman 1978, 1987, Boadway et al. 2007) or to compensate others for their investment (Varian 1994). Our game may also be relevant for agreements on R&D investment (Katz 1986) and International Environmental Agreements (IEAs) (Barrett 1994), if they involve binding co-financing.

Understanding the role of reciprocity in IEAs is important for environmental economists. Nyborg (2018) concurrently developed a model that extends D&K to cooperative games in order to apply it to Barrett's IEAs model. She finds that reciprocity can create weakly larger stable coalitions that exhibit higher abatement. Less closely related are Hadjiyiannis et al. (2012) and Kolstad (2014). The former studies the effect of a different notion of reciprocity on abatement in a two-player game with no possibility to sign an agreement. The latter examines the effect of equity- and efficiency-concerns (Charness and Rabin 2000) in an IEAs game.

We structure the paper as follows. Section 2 presents a set of preliminaries needed for our main result. Section 3 states and explains our main result on how full investment is impossible if all players sign, but is possible if no-one signs. Section 4 argues that our result illustrates a more general principle that high investment is possible with few signatories but not with many, examines comparative statics and identifies a zero-sign-zero-invest equilibrium. Section 5 offers reflections on alternative definitions of reciprocity and other game forms. Section 6 concludes.

2 Preliminaries

We introduce a public good game (2.1), an agreements game (2.2), and D&K's reciprocity model (2.3) which we apply to the public good game (2.4).

2.1 The public good game (Γ_P)

Let $N = \{1, \ldots, n\}$ be the set of players where $n \geq 4$.[5] Each $i \in N$ simultaneously chooses $a_i \in \{0, 1\}$, where 1 corresponds to investing in a public good and 0 to not doing so. Let $a = (a_i)_{i \in N}$. Player i's material payoff is

$$\pi_i(a) = \beta \sum_{j \in N} a_j - \gamma a_i,$$

where β is the public good benefit and γ the cost. Assume that $n\beta > \gamma > \beta > 0$ so that the individual cost exceeds the benefit and all players investing maximises total payoffs. Γ_P has a unique NE where for all $i \in N$, $a_i = 0$.

While there are no formal agreements in this game form, a legitimate interpretation of the strategy profile where all players invest is that players have a MOU (or gentleman's agreement) to invest. Clearly the MOU profile is not a NE of Γ_P.

2.2 The agreements game (Γ_A)

The agreements game appends a prior stage to the public good game where each player decides whether or not to sign a co-financing agreement: a binding commitment to share public good investment costs. In addition to capturing this formal agreement, the game form still nests the previous informal agreement, a MOU. If all players refuse to sign the co-financing agreement, but still invest in the public good, they can be said to have an informal agreement to invest, a MOU. Rejecting a formal agreement does not necessarily imply a MOU as players may not plan to invest. We now describe the game more precisely.

As before, $N = \{1, \ldots, n\}$, $n \geq 4$. In stage 1, the "sign-up stage", each $i \in N$ simultaneously chooses $a_i^1 \in \{0, 1\}$; 1 means signing the agreement, 0

[5] Our main result requires $n \geq 4$, we thus focus on this in Sections 2 and 3. Two and three player games are discussed in Section 4.

not doing so. In stage 2, the "investment stage", each $i \in N$ simultaneously chooses $a_i^2 \in \{0,1\}$; 1 means investing in the public good, 0 not doing so. Let $a^1 = (a_i^1)_{i \in N}$, $a^2 = (a_i^2)_{i \in N}$ and $a = (a^1, a^2)$.

Let $m(a^1) = \sum_{i \in N} a_i^1$ be the number of signatories, $x(a^2) = \sum_{i \in N} a_i^2$ the number of players who invest, $x^m(a) = \sum_{i \in N} a_i^1 a_i^2$ the number of signatories who invest. Player i's material payoff is

$$\pi_i(a) = \begin{cases} \beta x(a^2) - \gamma a_i^2 & \text{if } a_i^1 = 0 \\ \beta x(a^2) - \gamma \frac{x^m(a)}{m(a^1)} & \text{if } a_i^1 = 1 \end{cases},$$

where β is the public good benefit and γ the cost. The investment stage involves the same decision as in Γ_P, but with different payoff consequences for signatories. Signatories have made a binding commitment to share their investment costs. As before, $n\beta > \gamma > \beta > 0$ and to avoid knife-edge cases assume $\frac{\gamma}{\beta}$ is non-integer and $n > \left\lceil \frac{\gamma}{\beta} \right\rceil$.[6] A strategy for i, $s_i \in S_i$, specifies an initial choice and a choice for each stage 1 history. We focus on pure strategies throughout. Finally, let $S = \times_{i \in N} S_i$ and $s \in S$.

As previously explained, we interpret the strategy profile where zero players choose to sign but all invest, as players having a MOU to invest.

The following observation highlights important properties of SPE of Γ_A.

Observation 1 (Agreements game SPE)

(a) *There does not exist a full investment SPE in* Γ_A.

(b) *There exist zero investment SPE in* Γ_A *with* $[0, \left\lceil \frac{\gamma}{\beta} \right\rceil - 1)$ *signatories.*

(c) *There exist positive investment SPE in* Γ_A *with* $\left\lceil \frac{\gamma}{\beta} \right\rceil$ *signatories.*

Our proofs are rather tedious. We provide them in the appendix. In the main text we highlight key intuitions. Two aspects of Observation 1 are particularly noteworthy: a MOU is not a SPE and SPE exhibit less than full investment. To understand why, solve backwards. Non-signatories never invest as they bear their full investment cost (thus a MOU to invest is not a SPE). A signatory only invests if his costs are shared with sufficiently many co-signatories, namely if $m(a^1) \geq \left\lceil \frac{\gamma}{\beta} \right\rceil$. Given this, consider the sign-up stage. There exist SPE with very few signatories, (b), as stage 1 deviation does not

[6] $\left\lceil \frac{\gamma}{\beta} \right\rceil$ is the lowest integer strictly greater than $\frac{\gamma}{\beta}$.

induce investment. There also exist SPE with $\lceil \frac{\gamma}{\beta} \rceil$ signatories, (c), as each signatory is pivotal in inducing all signatories to invest. If there are greater than $\lceil \frac{\gamma}{\beta} \rceil$ signatories, a signatory can increase his payoff by deviating to not signing as other signatories continue to invest, hence full investment is not SPE, (a).

Observation 1 echos the results of Barrett (1994) and Varian (1994).[7]

2.3 Modelling reciprocity

We now incorporate reciprocity following D&K.[8] Their approach uses "kindness functions". To determine whether i is kind to j one needs a reference point, the equitable payoff. In defining this reference point, D&K argue that only the set of efficient strategies, E_i, those not involving "wasteful" play, are relevant.[9] For Γ_P, $E_i = S_i$; for Γ_A, $s_i \notin E_i$ iff s_i prescribes not investing after a history of i signing and $m(a^1) \geq \lceil \frac{\gamma}{\beta} \rceil$.

Let $b_{ik} \in S_k$ denote i's (point) belief of k's strategy, then given $(b_{ik})_{k \neq i}$ the *equitable payoff* for j is the average of the most i believes he can "give" to j given his strategy set and the least he believes he can "give" given his efficient strategies:

$$\pi_j^{e_i}\left((b_{ik})_{k\neq i}\right) = \tfrac{1}{2}\left[\max\{\pi_j\left(s_i,(b_{ik})_{k\neq i}\right)|s_i \in S_i\} + \min\{\pi_j\left(s_i,(b_{ik})_{k\neq i}\right)|s_i \in E_i\}\right].$$

To define i's kindness to j, let $s_i(h)$ be i's (updated) strategy which is identical to s_i except that a_i^1 must be consistent with reaching h. Let $b_{ik}(h)$ be i's (updated) belief of k's strategy. Given $s_i(h)$ and $(b_{ik}(h))_{k\neq i}$, i's *kindness* to j at h is

$$\kappa_{ij}\left(s_i(h),(b_{ik}(h))_{k\neq i}\right) = \pi_j\left(s_i(h),(b_{ik}(h))_{k\neq i}\right) - \pi_j^{e_i}\left((b_{ik}(h))_{k\neq i}\right).$$

The material payoff i believes j receives (first term on RHS) is compared to the equitable payoff (second term on RHS). If $\kappa_{ij}(.) > 0$, i is kind to j. If $\kappa_{ij}(.) < 0$, i is unkind to j. If $\kappa_{ij}(.) = 0$, i has zero kindness toward j.

[7] Both authors suggest mechanisms that give some SPE with partial investment and some with zero investment.

[8] Our presentation is tailored to Γ_P and Γ_A; see D&K for general games.

[9] A strategy is efficient if there does not exist another strategy which for all histories and others' strategies gives no player a lower payoff, and for some history and others' strategies gives at least one player a strictly higher payoff. See D&K pp. 275-6 for more on motivation, and the precise definition for general games. See Section 5.1 of the current paper for the implications of an alternative definition of efficient strategies.

To capture reciprocity incentives D&K need a function reflecting how kind i perceives j as being. Let $c_{ijk}(h)$ denote i's updated (point) belief about j's (point) belief about k's strategy. Given $b_{ij}(h)$ and $(c_{ijk}(h))_{k\neq j}$, i's *perceived kindness* of j towards i at history h is

$$\lambda_{iji}\left(b_{ij}(h),(c_{ijk}(h))_{k\neq j}\right)=\pi_i\left(b_{ij}(h),(c_{ijk}(h))_{k\neq j}\right)-\pi_i^{e_j}\left((c_{ijk}(h))_{k\neq j}\right).$$

The material payoff that i believes j believes i receives (first term on RHS) is compared to the equitable payoff (second term on RHS). If $\lambda_{iji}(.)>0$ then i perceives j as kind to him, etc.

Player i 's utility at history h sums material and reciprocity payoffs

$$\begin{aligned}&U_i\left(s_i(h),\left(b_{ik}(h),(c_{ikl}(h))_{l\neq k}\right)_{k\neq i}\right)\\&=\underbrace{\pi_i\left(s_i(h),(b_{ik}(h))_{k\neq i}\right)}_{\text{material payoff}}\\&+\underbrace{Y\cdot\sum_{j\in N\setminus\{i\}}\left(\kappa_{ij}\left(s_i(h),(b_{ik}(h))_{k\neq i}\right)\cdot\lambda_{iji}\left(b_{ij}(h),(c_{ijl}(h))_{l\neq j}\right)\right)}_{\text{reciprocity payoff}},\end{aligned}\tag{1}$$

where $Y\geq 0$ is the sensitivity to reciprocity payoffs. If $Y>0$ then i's preference for reciprocation toward j is captured by i's utility increasing when $\kappa_{ij}(.)$ and $\lambda_{iji}(.)$ are non-zero with matching signs, reflecting mutual kindness or unkindness.

Let $\widehat{\Gamma}_P$ and $\widehat{\Gamma}_A$ denote the public good game and agreements game where players' utilities are (1).[10] To define an appropriate solution concept, let $s_i'(h)$ be the strategy identical to $s_i(h)$ at all histories except at h, where it differs (given binary action choices there is only one such strategy).

[10] $\widehat{\Gamma}_P$ and $\widehat{\Gamma}_A$ are psychological games (Geanakoplos et al. 1989, Battigalli and Dufwenberg 2009).

Definition 1 *$s \in S$ is a sequential reciprocity equilibrium (SRE) if for all i and at each history h it holds that*

(i) $U_i\left(s_i(h), \left(b_{ik}(h), (c_{ikl}(h))_{l\neq k}\right)_{k\neq i}\right) \geq U_i\left(s_i'(h), \left(b_{ik}(h), (c_{ikl}(h))_{l\neq k}\right)_{k\neq i}\right)$,

(ii) $c_{ijk} = b_{jk} = s_k$ *for all* $j \neq i$ *and* $k \neq j$.

Condition (i) implies i best-responds at each history given his beliefs. Condition (ii) requires beliefs to be correct. If $Y = 0$ then Definition 1 describes a SPE (+ correct beliefs) in a game where utility equals material payoffs.

2.4 Reciprocity in the public good game

Full investment via a MOU was impossible in Γ_P (2.1). Can it be a SRE in $\widehat{\Gamma}_P$? If so, given equilibrium beliefs, for all $i, j \in N$, $\kappa_{ij}(1,.) = \lambda_{iji}(1,.) = \frac{\beta}{2}$ and $\kappa_{ij}(0,.) = -\frac{\beta}{2}$. Therefore i would not deviate to not investing if $n\beta - \gamma + (n-1)Y(\frac{\beta}{2})^2 \geq (n-1)\beta + (n-1)Y(-\frac{\beta}{2})(\frac{\beta}{2})$. Put differently, a MOU for full investment is a SRE of $\widehat{\Gamma}_P$ if $Y \geq Y^*$, where

$$Y^* = \frac{2(\gamma - \beta)}{\beta^2(n-1)}.$$

Intuitively, a sufficiently high reciprocity sensitivity ensures that the reciprocity cost of deviating to not investing (being unkind to co-players who are kind to you) outweighs the material gain.

When a MOU for full investment is possible in the public good game ($Y \geq Y^*$), unsurprisingly, it is also possible in the agreements game.[11] Our main result, in the next section, concerns the less obvious question of whether the agreements mechanism ($\widehat{\Gamma}_A$) can deliver full investment when a MOU for full investment is impossible in $\widehat{\Gamma}_P$, i.e. $Y < Y^*$.

3 Main result

Can the agreements game deliver full investment if this is impossible without the agreements mechanism ($Y < Y^*$) and the stakes are high (large β)?

[11] See the Online Appendix for full details. The Online Appendix contains additional formal results and proofs not included in the paper.

Indeed it can, but not how one might expect. If all players sign the co-financing agreement, full investment remains out of reach. However, full investment is possible if all players refuse to sign and have a MOU to invest instead. We first state the result formally (3.1), then provide intuition via examples (3.2-3).

3.1 Formal statement

To state our result and help with later analysis, we introduce some notation. Define

$$Y' := \frac{2(\gamma - \beta)}{(n-1)\beta(n\beta - \gamma)}, \tag{2}$$

and note that if $\beta > \frac{\gamma}{n-1}$ then $Y' < Y^*$.

Theorem 1 *For all $\gamma > 0$ and $n \geq 4$, there exists $\beta' \in (\frac{\gamma}{n-1}, \gamma)$ such that,*

- ***(a)*** *if $\beta \geq \beta'$ and $Y < Y^*$, there does not exist a full investment SRE in $\widehat{\Gamma}_A$ with n signatories,*
- ***(b)*** *if $\beta \geq \beta'$ and $Y \in [Y', Y^*)$, there exists a full investment SRE in $\widehat{\Gamma}_A$ with 0 signatories. The SRE is described by 0 signing, then non-signatories investing iff there are 0 signatories and signatories investing iff there are at least $\lceil \frac{\gamma}{\beta} \rceil$ signatories.*

Note that Theorem 1(b) does not imply equilibrium uniqueness. For example, we shall provide results on a zero-sign-zero-invest SRE in Section 4. Nonetheless, we consider the equilibrium highlighted in Theorem 1(b) to be our main contribution.

The theorem suggests that for high return public goods where full investment is impossible, the introduction of an agreements mechanism makes full investment possible. Encouraging all players to actually sign however is counterproductive. Instead, it is potentially a good thing if all parties reject the co-financing agreement as a MOU to invest is then incentive compatible.

The strategies that support the full investment result – part (b) of the theorem – may seem to require a deep understanding of the game by the players. If a player deviates in stage 1 by signing, then no players invest in stage 2. So a deviation in stage 1 would result in a large decrease in material payoff for all players, and (as we show below) this is central for

incentive-compatibility. This requires that all players understand the full consequences of varying degrees of take-up of the agreement, which may seem a tall order. However, under our MOU interpretation where coordination is achieved through discussions, such consequences would soon become common knowledge as negotiators reveal the implications of deviation.

3.2 Highlighting the intuition: part (b)

We begin with part (b) of Theorem 1, the SRE involving a MOU to invest, as it is more straightforward. Example 1 takes the stated profile, where no-one signs and all invest on path, and shows that it is a SRE for an interval of Y less than Y^*.

Example 1 (No-one signs): Take $\widehat{\Gamma}_A$ and let $n = 4$, $\gamma = 10$, $\beta = 9$ and $Y \in [\frac{1}{351}, Y^*)$, where $Y^* = \frac{2(10-9)}{9^2(4-1)} = \frac{2}{243}$. Consider the profile where,

- no-one signs,
- non-signatories invest iff there are 0 signatories,
- signatories invest iff there are at least 2 signatories.

Reason as follows to see that the profile is a SRE. Following a history of 0 signatories, i does not deviate to not investing if

$$4 \cdot 9 - 10 + 3Y \cdot \left(26 - \tfrac{1}{2}(26+0)\right) \cdot \underbrace{\left(26 - \tfrac{1}{2}(26+0)\right)}_{\lambda_{iji}(.)}$$

$$\geq 3 \cdot 9 + 3Y \cdot \left(17 - \tfrac{1}{2}(26+0)\right) \cdot \underbrace{\left(26 - \frac{1}{2}(26+0)\right)}_{\lambda_{iji}(.)}. \tag{3}$$

Solving gives $Y \geq \frac{1}{351}$, note that $\frac{1}{351} < Y^* = \frac{2}{243}$.

Following a history of 1 signatory, i does not deviate to investing for any Y. Following a history of 2 signatories, signatory i does not deviate to not investing if $Y \leq \frac{4}{157}$; note that $Y^* = \frac{2}{243} < \frac{4}{157}$. Following histories of 2 or 3 signatories, non-signatory i does not deviate to investing if $Y \leq Y^*$. Following a history of 3 or 4 signatories, signatory i does not deviate to not investing for any Y. Player i does not deviate to signing in the sign-up stage for any Y. The profile is thus a SRE for $Y \in [\frac{1}{351}, Y^*)$. ▲

Kindness amplification
Why is a MOU to invest incentive compatible in Example 1, when it would not be if there had been no option to sign a co-financing agreement ($\widehat{\Gamma}_P$ with $Y < Y^*$)? The explanation centers on what we call *kindness amplification*.

To see this, contrast the 0 signatory subgame in Example 1 with the full investment profile in $\widehat{\Gamma}_P$. In both cases, i has the same material incentive to deviate to not investing: a material payoff increase of $3 \cdot 9 - (4 \cdot 9 - 10) = 1$. His reciprocity incentive however differs as kindness is higher (or amplified) in Example 1 (e.g. $\lambda_{iji}(.) = \frac{9}{2}$ in $\widehat{\Gamma}_P$, see section 2.4, but $\lambda_{iji}(.) = 13$ in Example 1, see (3)). The amplified kindness implies that for a given Y, the decrease in i's reciprocity payoff from deviating to not investing is larger in Example 1 ($3Y \cdot 13 \cdot 13 - 3Y \cdot 4 \cdot 13 = 351Y$) than in $\widehat{\Gamma}_P$ ($3Y(\frac{9}{2})(\frac{9}{2}) - 3Y(-\frac{9}{2})(\frac{9}{2}) = \frac{243}{2}Y$). A lower Y is thus sufficient to prevent i deviating in Example 1 than in $\widehat{\Gamma}_P$, hence our result.

The fundamental reason as to why the agreements mechanism can amplify kindness is that it can increase a player's ability to influence others' behaviour. For instance, if j signs in Example 1, *no-one* invests, implying i's material payoff falls by $4 \cdot 9 - 10 = 26$. If j deviates to not investing in $\widehat{\Gamma}_P$, others' actions are unchanged so i 's material payoff falls by only $9 < 26$. Thus i perceives j as kinder in Example 1 than in $\widehat{\Gamma}_P$.

3.3 Highlighting the intuition: part (a)

We now illustrate part (a) of Theorem 1. Example 2 has three parts, it takes the parameters of Example 1 and demonstrates that for $Y < Y^*$, no profile where all players sign and invest on path is a SRE. The key intuition is that kindness cannot be amplified as effectively as when there are zero signatories on path (3.2). Part (i) demonstrates that a particular off-path investment-stage behaviour implies sign-up stage deviation. Parts (ii) and (iii) show that while alternative off-path investment-stage behaviours may avoid sign-up stage deviation, they necessarily involve investment-stage deviation.

Example 2(i) (Sign-up stage deviation): Take $\widehat{\Gamma}_A$ and let $n = 4$, $\gamma = 10$, $\beta = 9$ and $Y < Y^* = \frac{2}{243}$. Consider any profile where

- all players sign & invest on path,
- only signatories invest if there are 3 signatories.

Player i's sign-up stage incentives are identical to those for the full investment profile in $\widehat{\Gamma}_P$, i thus deviates to not signing as $Y < Y^*$. ▲

It follows from Example 2(i) that all remaining candidate SRE (where all players sign and invest on path) must involve: (a) a signatory who does not invest when there are 3 signatories, or, (b) a non-signatory who invests when there are 3 signatories. Examples 2(ii) and 2(iii) rule out candidate SRE with properties (a) and (b) respectively by demonstrating that players deviate in the investment stage.

Example 2(ii) (Signatory not investing with 3 signatories): Take $\widehat{\Gamma}_A$ and let $n = 4$, $\gamma = 10$, $\beta = 9$ and $Y < Y^* = \frac{2}{243}$. Consider a profile where

- all players sign & invest on path,
- at history h where there are 3 signatories, some signatory i does not invest.

By deviating to investing at h, signatory i can increase his material payoff by $9 - \frac{10}{3} = \frac{17}{3}$. Can i's reciprocity incentives prevent this deviation? Since i not investing is less kind than investing, to maximise the reciprocity cost of i's deviation examine the profile where

- all players sign & invest on path,
- no-one invests if there are 3 signatories,
- all players invest if there are 2 signatories.

Player i's reduction in reciprocity payoff from deviation at h is then $\frac{1144}{3}Y < \frac{2288}{729} < \frac{17}{3}$ (first inequality from $Y < Y^* = \frac{2}{243}$), thus i deviates. ▲

Weak kindness amplification

Why can kindness be sufficiently amplified to prevent deviation in Example 1 but not in Example 2(ii)? The answer concerns the size of the material loss that the reciprocity payoff need compensate in the relevant subgames. In Example 1, following a history of zero signatories, the material loss of investing was 1. In Example 2(ii), following a history of 3 signatories, the material loss to a signatory of not investing was $\frac{17}{3} > 1$. More generally, for high β, the size of the material loss from not deviating in the relevant

subgame is lower for the case when there are zero signatories than when there are many. Lower kindness is thus sufficient for non-deviation when there are few signatories than when there are many.

To complete Example 2, part (iii) examines the remaining candidate SRE (where all sign and invest on path) which involve all players investing when there are 3 signatories (by Examples 2(i)-(ii)). It shows that for a non-signatory to invest when there are 3 signatories, a signatory must not invest when there are 2 signatories. However, since this signatory has an incentive to deviate to investing, these profiles are not SRE either.

Example 2(iii) (Non-signatory investing with 3 signatories): Take $\widehat{\Gamma}_A$ and let $n = 4$, $\gamma = 10$, $\beta = 9$ and $Y < Y^* = \frac{2}{243}$. Consider a profile where

- all players sign & invest on path,
- all players invest at h where all but i have signed

By deviating to not investing at h, non-signatory i increases his material payoff by $10 - 9 = 1$. Whether i's reciprocity incentives prevent deviation depends on behaviour following histories with 2 signatories. If all invest when there are 2 signatories, i's incentives are identical to the full investment profile in $\widehat{\Gamma}_P$, thus i deviates at h ($Y < Y^*$). If only the signatories invest when there are 2 signatories, then i deviates at h if $Y < \frac{1}{108}$, which is true ($Y < Y^* = \frac{2}{243} < \frac{1}{108}$). For i not to deviate at h, it must be that at least 3 players do not invest when there are 2 signatories.

Suppose then that some signatory j does not invest when there are 2 signatories. Signatory j has a material incentive to deviate to investing. To maximise the reciprocity cost of this deviation, suppose no-one invests if there are 2 signatories and all invest if there is 1 signatory. Signatory j deviates to investing when there are 2 signatories if $Y < \frac{2}{143}$, which is true ($Y < Y^* = \frac{2}{243} < \frac{2}{143}$). ▲

For the non-signatory to invest when there are 3 signatories his perceptions of others' kindness must be sufficiently amplified, this requires signatories to not invest when there are 2 signatories. However, reciprocity incentives are not sufficiently large to prevent a signatory deviating to investing for reasons analogous to those discussed following Example 2(ii). Thus there does not exist a full investment SRE where all sign on path.

To summarise, the fundamental intuition behind part (a) of Theorem 1 is the difficulty of amplifying kindness when many players sign. In order to prevent sign-up stage deviation, kindness must be amplified, this requires low investment in subgames where there are $n-1$ or $n-2$ (3 or 2 in Example 2) signatories. However, signatories have relatively large material incentives to invest in such subgames (high β) and reciprocity payoffs are relatively low ($Y < Y^*$), hence the impossibility.

4 Robustness and relevance

In this section we examine various aspects of the robustness and relevance of our main result.

4.1 Few signing versus many signing

Theorem 1 examined whether particular profiles (those with either zero or n signatories, and full investment) are SRE. While these are clearly interesting profiles, one may nonetheless wonder whether Theorem 1 exemplifies the more general finding that "high investment is possible with few signatories and impossible with many signatories" or whether the equilibria in focus are just a special case of what may happen? We now provide a result that suggests the equilibria studied in Theorem 1 are more than just a special case.

Thus far we have defined a MOU to invest as the profile where zero players sign a co-financing agreement, then all players invest. It seems equally reasonable to interpret a profile where very few players sign the agreement, then all players invest, as a MOU to invest (at least for the many non-signatories). Proposition 1(b) demonstrates that such a MOU is also a SRE (cf. Theorem 1(b)). Theorem 1(a) identified the impossibility of full investment if there are n signatories. Proposition 1(a) shows that this impossibility remains if there are $n-1$ signatories.

Proposition 1 *For all $\gamma > 0$ and $n \geq 4$, there exists $\beta'' \in \left(\frac{\gamma}{n-1}, \gamma\right)$ and $Y'' \in (0, Y^*)$ such that,*

- ***(a)*** *if $\beta \geq \beta''$ and $Y < Y^*$, there does not exist a full investment SRE in $\widehat{\Gamma}_A$ with $n-1$ signatories,*
- ***(b)*** *if $\beta \geq \beta''$and $Y \in [Y'', Y^*)$, there exists a full investment SRE in $\widehat{\Gamma}_A$ with one signatory. The SRE is described by one player signing, then non-signatories investing iff there is exactly one signatory, and signatories investing always.*

The intuition behind the result is analogous to that driving Theorem 1 (see section 3.2-3), kindness can be sufficiently amplified when there is only one signatory but not when there are $n-1$. Proposition 1 is a reassuring robustness check and suggests an even more general insight that for high value public goods, if the social goal is full investment, it is feasible with few signatories but not with many.

4.2 When should no-one sign?

Theorem 1 states that for high value public goods, zero-signatories can lead to full investment. However the result provides little idea of its importance in the sense of answering questions like: How large is the interval of reciprocity sensitivities for which the zero-sign-all-invest equilibrium exists? And what are the determinants of the size of this interval? We now consider these questions.

In order to explore the relevance of our SRE in the parameter space we first amend our model. Under the current setup, as n tends to infinity, reciprocity incentives outweigh material incentives by construction.[12] In order to avoid this feature, replace Y with $\frac{y}{n-1}$ in utility function (1), where $y \geq 0$ is the reciprocity sensitivity parameter. Normalising i's reciprocity incentives by the number of co-players implies that such incentives do not increase simply because there are more players in the game. Repeating the analysis given this amendment, one finds $y^* = \frac{2(\gamma-\beta)}{\beta^2}$, $y' = \frac{2(\gamma-\beta)}{\beta(n\beta-\gamma)}$ and the following analog

[12] Also since $\lim_{n\to\infty} Y^* = 0$, Theorem 1 applies to a negligible set of parameters for very large populations.

of Theorem 1(b): if $\beta \geq \beta'$ and $y \in [y', y^*)$, then there exists a full investment SRE in $\widehat{\Gamma}_A$ with zero signatories.[13]

Given this amendment, we study the following ratio

$$\rho = \frac{y^* - y'}{y^* - 0} = 1 - \frac{\beta}{n\beta - \gamma}. \tag{4}$$

Note that 0 to y^* are all the values of reciprocity sensitivity that imply full investment is not a SRE in the public good game. ρ is the fraction of these that allow zero-sign-all-invest to be a SRE in the agreements game. We are only interested in ρ for parameters where the zero-sign-all-invest equilibrium exists, that is $\beta \geq \beta'$. For such β, $y^* > y'$, thus ρ is strictly positive and its comparative statics are as follows.

Proposition 2 *For all $\gamma > 0$, $n \geq 4$ and $\beta \geq \beta'$,*

- ***(a)*** *ρ is strictly increasing in n and as n tends to infinity, ρ tends to unity;*
- ***(b)*** *ρ is strictly increasing in β and as β tends to γ, ρ tends to $1 - \frac{1}{n-1}$;*
- ***(c)*** *ρ is strictly decreasing in γ.*

The result states that for high return public goods, a large share of reciprocity sensitivities can lead to full investment via no-one signing if there are many players, the return of the public good is high or the cost is low. This is very intuitive. For example, consider the effect of n. Having more players does not affect the y that can support a full investment SRE in the public good game since i's reciprocity incentive is unaffected by more players who he perceives as just as kind as the existing co-players; thus y^* does not change. However, having more players does decrease y'. This is because with more players, kindness is amplified more as the difference in material payoff between everyone investing and no-one investing is larger, thus a lower y is sufficient to prevent deviation. This increase in the difference between y^* and y' implies a larger share of sensitivities supporting the SRE, hence, a higher ρ.[14] Intuitions for the effect of changes in the cost and benefit of the public good are equally intuitive.

[13] Note also that with normalised reciprocity incentives, our result is relevant even for large populations since y^* does not depend on n.

[14] On the issue of the number of players, recall that Theorem 1 only applies to $n \geq 4$. When $n = 3$, an analog of Theorem 1(a) cannot be established because players have

Proposition 2 is not only informative on how different parameters affect ρ, but also on the magnitude of ρ. Part (b) of the proposition provides an upper bound to the share of sensitivity parameters supporting our SRE for a given n and γ, $1 - \frac{1}{n-1}$. This upper bound is relatively high: since $n \geq 4$, it is at least $\frac{2}{3}$. Thus when the return of the public good is high, a large share of sensitivities implying no investment in the public good game imply zero-sign-all-invest is a SRE.

A lower bound to ρ would be informative, however this is difficult to come by.[15] Nonetheless, examples illustrate that even for slightly lower values of β, the value of ρ is still relatively close to the upper bound. For instance, recall Example 1 where $n = 4$, $\gamma = 10$ and $\beta = 9$. For this example, $\rho = 0.654$, that is, zero-sign-all-invest is a SRE for 65.4% of reciprocity sensitivity parameters such that full investment is impossible in the public good game (note that given n and γ, the upper bound of ρ is 0.667).

4.3 Equilibrium multiplicity: Zero-sign-zero-invest

Even when the set of parameters supporting our zero-sign-zero-invest SRE is large, it is not the case that no-one signing an agreement necessarily implies full investment. The reason for this is simply that the agreements game has many SRE, including those where no-one signs followed by low investment. More specifically, we have the following negative result.

Proposition 3 *For all $\gamma > 0$ and $n \geq 4$, there exists $\beta' \in \left(\frac{\gamma}{n-1}, \gamma\right)$ such that if $\beta \geq \beta'$ and $Y \in [Y', Y^*)$, there exists a 0 investment SRE with 0 signatories. The SRE is described by 0 signing, then non-signatories not investing and signatories investing iff there are at least $\left\lceil \frac{\gamma}{\beta} \right\rceil$ signatories.*

That is, whenever a zero-sign-all-invest SRE exists, a zero-sign-zero-invest SRE also exists. To see the intuition as to why, contrast the strategy profiles in Theorem 1(b) and Proposition 3. Note that they are identical other

no material incentive to invest when very few sign, thus kindness can be sufficiently amplified on path to support the all-sign-all-invest SRE. When $n = 2$, an analog of Theorem 1(b) cannot be established since i can no longer amplify his kindness towards k by influencing the behaviour of j, thus a zero-sign-zero-invest SRE does not exist. See the Online Appendix for full details.

[15] It is difficult to evaluate the limit of ρ as β tends to β' from above since we do not have an explicit expression for β'. Taking the limit of ρ as γ tends to $n\beta$ from below is also not possible since the relevant β are such that $\beta \geq \beta'$ and β' is in general a function of γ.

than at the history following zero signatories. In both cases, players have no material incentive to invest at this history. If zero invest at this history, players do not view others as kind and thus have no incentive to deviate. If all invest at this history, then players view each other as kind and reciprocating kindness prevents non-deviation only if Y is sufficiently high. Thus the conditions supporting the zero-sign-all-invest SRE are more restrictive than those supporting the zero-sign-zero-invest SRE.

Proposition 3 implies a clear need for caution when no-one signs an agreement. It may be that full investment is more likely than zero investment following zero signatories as the outcome is more unusual and also Pareto dominant, so more salient and easier to coordinate on. However, such equilibrium selection arguments are very context dependent so would need to be considered on a case by case basis. The existence of the zero-investment SRE does not affect our main insight: full investment is possible if very few sign, but not if many do. However, it does underscore the need to tackle a difficult coordination problem; simply not signing an agreement is not sufficient to guarantee full investment.

One might expect a zero-sign-zero-invest SRE to always exist. This turns out not to be the case as our next result states (as in 4.2, we normalise i's reciprocity incentives by the number of co-players, however a qualitatively similar result can be established without the normalisation).

Proposition 4 *For all* $\gamma > 0$, $n > 13$ *and* $\beta < \frac{3\gamma}{2(n+2)}$, *there does not exist a SRE where* 0 *sign and* 0 *invest if* $y \in \left(\frac{4(\gamma-\beta)}{\beta(\gamma-4\beta)}, \min\left\{\frac{2(\gamma-\beta)}{3\beta^2}, \frac{2(\gamma-\beta)}{\beta(n\beta-\gamma)}\right\}\right)$.

A zero-sign-zero-invest SRE does not exist when there are many players, the return to the public good is sufficiently low and the reciprocity sensitivity is at an intermediate value. Deviations from candidate equilibria are typically in the one signatory subgame. In such subgames, non-signatories have no material incentive to invest and given the low return to the public good and relatively low reciprocity concern, any reciprocity incentive to invest is not sufficiently large motivative investment. The signatory not investing is not part of an equilibrium as he perceives non-signatories as kind (as they could have signed and invested, forcing a lower material payoff on the signatory), thus the signatory's reciprocity incentives motivate deviation to investing. Equally however, the signatory investing is not part of an equilibrium as then the signatories views the signatories as unkind and thus has both material and reciprocity incentives to deviate to not investing.

While Proposition 4 implies that a zero-invest-zero-sign SRE does not exist for all parameters, taken together with previous results we can begin to infer something about how the existence of such equilibria are affected by n. It is often found that larger groups provide fewer public goods (e.g. Bergstrom et al. 1986). Reason as follows to see how this is reflected in our results. Note that for $y < y^*$, Propositions 2 and 3 imply that as n increases a larger share of y support a zero-sign-zero-invest SRE. In a similar vein, notice that for the sufficient conditions presented in Proposition 4, an increase in n reduces the size of the interval of y that implies non-existence of such equilibria. Since Proposition 4 only identifies sufficient conditions for non-existence of zero-sign-zero-invest equilibria, one must be careful not to over-infer from this comparative static. Nonetheless, both findings go in the expected direction, larger groups may struggle to provide public goods.

5 Two reflections

5.1 The role of efficient strategies a la D&K

Our main result depends critically on using D&K's notion of "efficient strategies," rather than the corresponding notion in Rabin (1993). In this section we present and discuss the issue.

The key point concerns how to calculate the zero-kindness threshold (i.e. the equitable payoff, $\pi_j^{e_i}$). In D&K's theory this is done with respect to the average of the lowest and highest payoff that i can achieve for j, if i uses an efficient strategy, a notion D&K define independently of player i's beliefs (see p. 276 in D&K; footnote 9 in this paper). Rabin, by contrast (see his p. 1286) calculates the zero kindness threshold with respect to those strategies that produce payoffs on the Pareto frontier, *given* i's beliefs. With such a definition the equilibrium highlighted in our main theorem would collapse: Under Rabin's definition a player who deviated at the sign-up stage would do something inefficient, hence irrelevant for calculating the zero-kindness threshold. In effect, the dramatic kindness amplification discussed in section 3.2. would be lost.

While this is not the place for a detailed discussion of the relative merits of the two assumptions, we briefly describe two reasons why the D&K assumption seems intuitive.[16]

[16]See section 5 of D&K for more discussion of the two assumptions.

First, consider the kindness associated with a deviation from a purported equilibrium. What beliefs could this player have had? Probably not equilibrium beliefs since he is deviating. When assessing the kindness of a deviating co-player, one might ask oneself what payoff consequences the deviator had in mind, and in particular whether he is inefficiently attempting to lower the payoffs of all players. D&K's key definition identifies the set of strategies (the inefficient ones) for which no story can be told that explains the behaviour unless that story involves "waste" at some history. This is a set with a meaningful interpretation, which helps one think about and interpret the zero kindness threshold.

Second, D&K's definition implies kindness changes continuously with beliefs. This seems psychologically intuitive. It also implies that equilibrium existence is guaranteed, a convenient implication.

5.2 The phenomenon of kindness amplification

The main motivation of our work is to explore the impact of reciprocity in our particular game forms, given our particular co-financing agreements and MOUs interpretation. Our results uncovered a phenomenon that we labelled kindness amplification. Analogous effects can be found in related game forms with alternative economic interpretations. While it is beyond the scope of this paper to explore the topic at length we now present the simplest example which demonstrates that the phenomenon can imply similar but not identical implications for public good provision in other game forms.

Example 4 (Cheap-talk first stage): Define Γ_C as identical to Γ_A except that i's material payoff is $\pi_i(a) = \beta x(a^2) - \gamma a_i^2$. First-stage actions in Γ_C thus have no material payoff consequences and can be interpreted as cheap-talk messages sent prior to a public good game. For all γ and $n \geq 4$, there exists $\beta_C \in (\gamma/n, \gamma)$ and $Y_C \in (0, Y^*)$ such that

(i) if $\beta \geq \beta_C$ and $Y \in [Y_C, Y^*)$, there exists a full investment SRE where all players choose 0 in stage 1. The SRE is described by all choosing 0, then players investing iff an even number of players chose 0 in the first stage.

(ii) if $\beta \geq \beta_C$ and $Y \in [Y_C, Y^*)$, there exists a full investment SRE where all players choose 1 in stage 1. The SRE is described by all choosing 1, then players investing iff an odd number of players chose 0 in the first stage.[17]

[17] Proof available on request.

The example illustrates that the existence of any material-payoff-irrelevant first-stage action choice can amplify kindness and thus lead to full investment (the profile may be interpreted as a MOU to invest). The intuition is identical to that behind the Theorem 1(b) where no-one signed (i.e. took a material-payoff-irrelevant-action) and then all players invested. In this cheap-talk game however, since there are no material payoff consequences of sending a particular message (unlike where at least two players choose to sign a co-financing agreement), there is no analog to Theorem 1(a) i.e. full investment is not impossible if all players take a particular first-stage action. The asymmetric effects of all signing versus no-one signing in our main result cannot be generated by a cheap-talk first stage.

6 Conclusion

In this paper we studied the ability of a co-financing agreements mechanism to raise public good investment given reciprocity motivations. Our main result is that when full investment is impossible in the absence of an agreements mechanism and the return to the public good is high, then full investment remains impossible if all players sign the agreement, but is possible (via a MOU) if all players reject the agreement. Critically, the informal agreement to invest (the MOU) would not be possible if players had not met, tried to sign a binding agreement, but then explicitly rejected it in favour of an informal agreement.

Given our result, it is natural to wonder whether in reality we ever see parties meet, try to sign a binding agreement, then walk away with a non-binding agreement? As a possible example, take the 2015 UN Climate Change Conference in Paris (aka COP21/CMP11). Many commentators have pointed out that the agreement is not binding, in the sense that there are no penalties for non-compliance (e.g. Clémençon (2016), Jacquet and Jamieson (2016)). Arguably, negotiators could have made a binding, or formal, commitment to co-finance each other's efforts to reduce greenhouse gas emissions, but instead they struck a non-binding MOU to exert such effort anyway, without co-financing mandates (at least between industrial countries). This may be an outcome along the lines that our Theorem 1 points to. Agents do not sign a formal co-financing agreement, but rather reject the agreement and coordinate on full investment nonetheless. The "soft-touch" Paris agreement may be the kind of situation pointed to by Theorem 1: informal deals gain

traction as other formal deals are shunned. And institutions that allow formal agreements for co-financing to be signed, even when they are not, are critical for spurring informal agreements to invest in public goods

Despite the usefulness of non-binding agreements, binding ones are often also signed in reality. Where such an agreement is signed, only a subset of potential signatories would typically actually sign. Strictly, our Theorem 1 is silent about binding agreements with few signatories. However, as argued for in Section 4.1, it is reasonable to view the result as illustrating a more general point that higher investment is feasible when few people sign a co-financing agreement than when many do. This interpretation suggests that reciprocity motivations may be able to explain why agreements with few signatories nonetheless manage to achieve desirable outcomes

If MOUs are indeed the route to desirable outcomes, why shouldn't institution designers just allow parties to discuss MOUs directly rather than offering them the possibility to sign binding co-financing agreements? In Section 5.2 we illustrated that such a meeting with cheap-talk messaging can lead to full investment via many different action profiles, this may make coordination difficult. By contrast, the number of routes to full investment is smaller with a co-financing agreements institution, as if there are many signatories, full investment is impossible. This difference would presumably make coordination easier with a co-financing mechanism, a potentially important advantage of using co-financing agreements. Even with the option of co-financing agreements, coordination problems are not solved as multiple equilibria remain (e.g. see Section 4.3).

The overall lesson is clear: Institutions that offer opportunities for binding agreements may be important even if such agreements are not struck or there is low uptake. In our case, creating the possibility of a formal co-financing agreement may promote actual investment in public goods, even if a co-financing agreement is not signed.

Appendix

Proof of Observation 1

Apply backward induction to identify the SPE of Γ_A. At $h = a^1$, non-signatories do not invest and signatory i invests iff $\beta \geq \frac{\gamma}{m(a^1)}$. Since $\frac{\gamma}{\beta}$ non-

integer, write the condition as iff $m(a^1) \geq \lceil \frac{\gamma}{\beta} \rceil$. Given optimal behaviour at all a^1, consider the first-stage. First suppose there are less than $\lceil \frac{\gamma}{\beta} \rceil - 1$ signatories. Player i does not deviate in the first-stage as $\pi_i(.) = 0$ regardless. Thus this is a SPE. Second suppose there are $\lceil \frac{\gamma}{\beta} \rceil - 1$ signatories. Non-signatory i deviates to signing if $0 < \beta \lceil \frac{\gamma}{\beta} \rceil - \gamma$, which is true. Thus this is not a SPE. Third suppose there are $\lceil \frac{\gamma}{\beta} \rceil$ signatories. Signatory i does not deviate to not signing if $\lceil \frac{\gamma}{\beta} \rceil \beta - \gamma \geq 0$, which is true. Non-signatory i does not deviate to signing if $\lceil \frac{\gamma}{\beta} \rceil \beta \geq (\lceil \frac{\gamma}{\beta} \rceil + 1)\beta - \gamma$, which is also true. Thus this is a SPE. Fourth suppose there are more than $\lceil \frac{\gamma}{\beta} \rceil$ signatories. Signatory i deviates to not signing as other players invest regardless of his choice. Thus this is not a SPE. The four cases are exhaustive. ■

Proof of Theorem 1

(a) Take the set of strategy profiles that involve n players signing and investing on path. Partition this set into 3 subsets of profiles distinguished by behaviour following a history of a^1 such that $m(a^1) = n - 1$: *subset 1*, all invest; *subset 2*, all signatories invest only; and *subset 3*, all remaining profiles. We take each subset in turn and demonstrate that no profile in the subset is a SRE if β is sufficiently high and $Y < Y^*$.

Subset 1: Consider any candidate SRE profile s^* such that each $i \in N$ signs, then invests if a^1 is such that $m(a^1) \geq n - 1$. Reason as follows to show that there is no behaviour at histories such that $m(a^1) < n - 1$ that would imply s^* is a SRE.

Consider $h = a^1$ such that $m(a^1) = n - 1$, so all players invest. Non-signatory i has the same material incentive to deviate to not investing as in Γ_P. Given $Y < Y^*$, a necessary condition for i not to deviate is that $\lambda_{iji}(s_j^*, .) > \frac{\beta}{2}$ (recall $\lambda_{iji}(1, .) = \frac{\beta}{2}$ in $\widehat{\Gamma}_P$). The value of $\lambda_{iji}(s_j^*, .)$ at h, depends on the action choices s^* prescribes at history h' where all except i and j sign. If s^* were such that n invest or all except j invest at h', then $\lambda_{iji}(s_j^*, .) = n\beta - \gamma - \frac{1}{2}(n\beta - \gamma + (n-1)\beta - \gamma) = \frac{\beta}{2}$ at h, thus i would deviate at h. If s^* were such that all except i invest or all except i and j invest at h', then $\lambda_{iji}(s_j^*, .) = n\beta - \gamma - \frac{1}{2}((n-1)\beta + (n-2)\beta) = \frac{3}{2}\beta - \gamma < \frac{\beta}{2}$, thus i would deviate at h. Therefore a necessary condition for i to not deviate at h, is that s^* must be such that some signatory l does not invest at h'.

Consider h' and suppose s^* prescribes signatory l does not invest. Signatory l has a material incentive to deviate to invest at h'. We now show

that for β sufficiently high and all $Y < Y^*$, l's reciprocity incentives are insufficient to prevent deviation at h'. The change in signatory l's reciprocity payoff from playing s_l^* rather than $s_l'(h', s_l^*)$ equals $Y\Psi$, where

$$\Psi := \sum_{k \in N \setminus \{l\}} \left(\kappa_{lk}(s_l^*, .) - \kappa_{lk}\left(s_l'(h', s_l^*), .\right) \right) \lambda_{lkl}(s_k^*, .).$$

If $\Psi \leq 0$, then l deviates at h'. Suppose $\Psi > 0$. Signatory l does not deviate at h' if

$$Y \geq \frac{(n-2)\beta - \gamma}{(n-2)\Psi}. \tag{5}$$

Let $\widehat{Y}(\beta)$ denote the RHS of (5) as a function of β. For $Y < Y^*$ we require

$$\widehat{Y}(\beta) = \frac{(n-2)\beta - \gamma}{(n-2)\Psi} < \frac{2(\gamma - \beta)}{\beta^2 (n-1)} = Y^*.$$

Now argue that for sufficiently high β either $Y < Y^*$ does not hold or l has an incentive to deviate at h'. Note that $\lim_{\beta \to \gamma} Y^* = 0$. Therefore for β in the neighbourhood of γ, $Y < Y^*$ requires $\lim_{\beta \to \gamma} \widehat{Y}(\beta) \leq 0$. Evaluating $\widehat{Y}(\gamma)$, note that the numerator of $\widehat{Y}(\gamma)$ is positive thus it must be that $\Psi \leq 0$. However if $\Psi < 0$ then l would deviate at h' for β slightly lower than γ as already argued, thus $\Psi = 0$ when $\beta = \gamma$. That we have supposed that $\Psi > 0$ for sufficiently high β and deduced $\Psi = 0$ at $\beta = \gamma$, implies that the denominator of $\widehat{Y}(\beta)$ approaches zero from above and hence the one-sided limit $\lim_{\beta \to \gamma^-} \widehat{Y}(\beta) = +\infty$ which is greater than $\lim_{\beta \to \gamma} Y^* = 0$, violating $Y < Y^*$. Thus for all $Y < Y^*$ and β sufficiently high, l would deviate to investing at h'.

Subset 2: Consider a candidate SRE profile s^* such that each $i \in N$ signs, then invests if a^1 is such that $m(a^1) = n$, and if a^1 is such that $m(a^1) = n - 1$, all except the non-signatory invest. At the initial node, i's incentives are identical to those in the full investment profile in $\widehat{\Gamma}_P$. Thus i deviates to not signing at the initial node for all $Y < Y^*$. Hence s^* is not a SRE.

Subset 3: Consider any remaining candidate SRE profile s^* such that each $i \in N$ signs, then invests if a^1 is such that $m(a^1) = n$, it must be that for history $h'' = a^1$ such that $m(a^1) = n - 1$, there exists some signatory r who does not invest. Reasoning analogous to that used to show signatory l deviates to investing at h' (end of subset 1) establishes that signatory r deviates to investing at h''. Hence s^* is not a SRE.

(b) Consider s^* such that each $i \in N$ does not sign, then invests if a^1 is such that $m(a^1) = 0$ or $a_i^1 = 1$ and $m(a^1) \geq 2$, and does not invest otherwise. We demonstrate that there exists $Y' < Y^*$ such that no player deviates at any h if $Y \in [Y', Y^*)$ and β is sufficiently large.

First consider $h = a^1$ such that $m(a^1) = 0$. Player i has a material incentive to deviate to not investing and a reciprocity incentive to not do so. His increase in reciprocity payoff from playing s_i^* instead of $s_i'(h, s_i^*)$ is

$$\begin{aligned}&\left(n - m(a^1) - 1\right) \cdot Y \cdot \left(\kappa_{ij}\left(s_i^*, .\right) - \kappa_{ij}\left(s_i'\left(h, s_i^*\right), .\right)\right) \cdot \lambda_{iji}\left(s_j^*, .\right) \\ &+ m(a^1) \cdot Y \cdot \left(\kappa_{il}\left(s_i^*, .\right) - \kappa_{il}\left(s_i'\left(h, s_i^*\right), .\right)\right) \cdot \lambda_{ili}\left(s_l^*, .\right), \end{aligned} \tag{6}$$

where j is a non-signatory and l a signatory. Since $\kappa_{ij}\left(s_i^*, .\right) = \lambda_{iji}\left(s_j^*, .\right) = \frac{1}{2}(n\beta - \gamma)$ and $\kappa_{ij}\left(s_i'(h, s_i^*)\right) = \frac{1}{2}(n\beta - \gamma) - \beta$, this increase in reciprocity payoff is larger than the reduction in material payoff if $Y \geq \Phi$, where

$$\Phi := \frac{2(\gamma - \beta)}{(n-1)\beta(n\beta - \gamma)}.$$

Note that if $\beta > \frac{\gamma}{n-1}$, then $\Phi < Y^*$. Thus i does not deviate for $Y \in [\Phi, Y^*)$.

Now consider $h = a^1$ such that $m(a^1) > 0$. No player has a material incentive to deviate at a^1 since for $\beta > \frac{\gamma}{2}$, $\lceil \frac{\gamma}{\beta} \rceil = 2$, thus action choices following all a^1 are identical to SPE profiles. We will demonstrate for sufficiently high β and Y (but less than Y^*), any reciprocity incentive to deviate is less than the material incentive to not do so.

At $h = a^1$ such that $m(a^1) = 1$, player i has no reciprocity incentive to deviate to investing, thus does not deviate. At $h = a^1$ such that $m(a^1) = 2$, the change in signatory i's reciprocity payoff from playing s_i^* rather than $s_i'(h, s_i^*)$ is

$$\begin{aligned}&\left(n - m(a^1)\right) \cdot Y \cdot \left(\kappa_{ij}\left(s_i^*, .\right) - \kappa_{ij}\left(s_i'\left(h, s_i^*\right), .\right)\right) \cdot \lambda_{iji}\left(s_j^*, .\right) \\ &+ \left(m(a^1) - 1\right) \cdot Y \cdot \left(\kappa_{il}\left(s_i^*, .\right) - \kappa_{il}\left(s_i'\left(h, s_i^*\right), .\right)\right) \cdot \lambda_{ili}\left(s_l^*, .\right), \end{aligned} \tag{7}$$

where j is a non-signatory and l a signatory. For non-signatory j, $\kappa_{ij}\left(s_i^*, .\right) = \beta$, $\kappa_{ij}\left(s_i'(h, s_i^*), .\right) = 0$ and $\lambda_{iji}\left(s_j^*, .\right) = -\frac{\beta}{2}$. For signatory l, $\kappa_{il}\left(s_i^*, .\right) = \frac{3}{2}\beta - \gamma$, $\kappa_{il}\left(s_i'(h, s_i^*), .\right) = \frac{1}{2}(\beta - \gamma)$ and $\lambda_{ili}\left(s_l^*, .\right) = \frac{3}{2}\beta - \gamma$. Substituting into (7) gives $Y\left(\left(\beta - \frac{\gamma}{2}\right)\left(\frac{3}{2}\beta - \gamma\right) - (n-2)\frac{\beta^2}{2}\right)$, which is negative for sufficiently large β. Signatory i's reciprocity incentive to deviate to not investing at h is no larger than the material incentive to not do so if

$$Y \leq \left[\frac{\frac{\gamma}{2} - \beta}{\left(\beta - \frac{\gamma}{2}\right)\left(\frac{3}{2}\beta - \gamma\right) - (n-2)\frac{\beta^2}{2}}\right].$$

For β sufficiently large, there exists Y satisfying the inequality and that $Y \geq \Phi$ (as $\lim_{\beta \to \gamma} [.] > 0$ and $\lim_{\beta \to \gamma} \Phi = 0$).

Using (6), non-signatory i's change in reciprocity payoff from playing s_i^* rather than $s_i'(h, s_i^*)$ is $\frac{1}{2} Y \beta^2 (n-7)$. This is non-negative for $n \geq 7$, thus i does not deviate at h. For $n \in [4, 6]$, i does not deviate at h if $Y \leq \frac{2(\gamma - \beta)}{\beta^2 (7-n)}$. There exists Y satisfying the inequality and that $Y \geq \Phi$, since the RHS of this inequality is greater than Φ if $(n^2 - 7)\beta + (1-n)\gamma > 0$, which holds given $n \geq 4$.

Consider $h = a^1$ such that $m(a^1) \in [3, n-1]$. Using (7), signatory i's change in reciprocity payoff from playing s_i^* rather than $s_i'(h, s_i^*)$ is $Y \Xi$ where

$$\Xi := \left(m\left(a^1\right) - 1\right) \left(\beta - \frac{\gamma}{m(a^1)}\right) \frac{\beta}{2} - \left(n - m\left(a^1\right)\right) \frac{\beta^2}{2}.$$

If $\Xi \geq 0$, then i has no reciprocity incentive to deviate at h. Suppose $\Xi < 0$, then i does not deviate if $Y \leq \left[\frac{\gamma - \beta m(a^1)}{\Xi m(a^1)}\right]$. For β sufficiently large, there exists Y satisfying the inequality and that $Y \geq \Phi$ (as $\lim_{\beta \to \gamma} [.] > 0 > \lim_{\beta \to \gamma} \Phi = 0$).

Non-signatory i's change in reciprocity payoff from playing s_i^* rather than $s_i'(h, s_i^*)$ is $\frac{1}{2} Y \beta^2 (n - 2m(a^1) - 1)$ (use (6)). If $n - 2m(a^1) - 1 \geq 0$, then $\frac{1}{2} Y \beta^2 (n - 2m(a^1) - 1) \geq 0$, thus i has no reciprocity incentive to deviate at h. If $n - 2m(a^1) - 1 < 0$, then i does not deviate at h if

$$Y \leq \frac{2(\gamma - \beta)}{\beta^2 (2m(a^1) + 1 - n)}.$$

The RHS is strictly greater than Φ if $(n^2 - 2m(a^1) - 1)\beta + (1-n)\gamma > 0$ which is true as β tends to γ.

Finally, at $h = a^1$ such that $m(a^1) = n$ and the initial node, player i has neither material nor reciprocity incentives to deviate. Thus for β sufficiently large, the profile is a SRE for $Y \in [\Phi, Y^*)$, that is, $Y \in [Y', Y^*)$. ■

Proof of Proposition 1

(a) Take the set of strategy profiles that involve $n-1$ signing and all investing on path. First consider $h = a^1$ such that $m(a^1) = n$ and suppose that some i does not invest. Since i has a material incentive to invest, in order to maximise his reciprocity incentive not to do so, suppose no-one invests at a^1

(to maximise how unkind he perceives his co-players as). Player i then does not deviate to investing if $Y \geq \frac{2}{(n\beta-\gamma)(n-1)}$, however note that as β tends to γ, then RHS tends to $\frac{2}{\gamma(n-1)^2}$ which is greater than $\lim_{\beta\to\gamma} Y^* = 0$. Thus for sufficiently high β, i deviates to investing. Given this restrict attention to strategy profiles where all invest if all sign.

Note that if all players invest when there are $n-2$ signatories then i would deviate to not investing when there are $n-1$ signatories given $Y < Y^*$. It must thus be that some players do not invest when there are $n-2$ signatories. Consider $h = a^1$ such that $m(a^1) = n-2$ and suppose that signatory i does not invest. Since i has a material incnetive to invest, in order to maximise his reciprocity incentive to not do so, suppose no-one invests at a^1. Signatory i then does not deviate to investing if $Y \geq \frac{2(\beta(n-2)-\gamma)}{(n\beta-\gamma)(2\beta(n-2)-\gamma)}$. However note that as β tends to γ, the RHS converges to a strictly positive value while Y^* converges to zero. Thus for β sufficiently high, signatory i deviates to investing when $Y < Y^*$. Given this restrict attention to strategy profiles where all signatories invest when there are $n-2$ signatories.

Finally, given that all players invest if there are at least $n-1$ signatories and all signatories invest if there are $n-2$ signatories, for $h = a^1$ such that $m(a^1) = n-1$, non-signatory i invests if $Y \geq \frac{2(\gamma-\beta)}{\beta(3\beta-2\gamma)(n-1)}$. However for β sufficiently high, the RHS is strictly greater than Y^*, thus i deviates to not investing for $Y < Y^*$. Hence there exists no SRE where all invest when there are $n-1$ signatories.

(b) Consider s^* such that only one player signs, then i invests if only this player signs or if $a_i^1 = 1$ and $m(a^1) \geq 2$, and does not invest otherwise. We demonstrate that there exists $Y'' < Y^*$ such that no player deviates at any h if $Y \in [Y'', Y^*)$ and β is sufficiently large.

First consider $h = a^1$ such that $m(a^1) = 0$. Player i has no material incentive to deviate to investing and given that he views all co-players as unkind, he also has no reciprocity incentive to do so.

Next consider $h = a^1$ such that $m(a^1) = 1$. The signatory does not deviate to not investing if $Y \geq \frac{2(\gamma-\beta)}{\beta^2(n-1)(n-2)}$; clearly the RHS is strictly less than Y^*. Non-signatory i does not deviate to not investing if $Y \geq \frac{2(\gamma-\beta)}{\beta(\beta n-\gamma+(\beta(n-2)-\gamma)(n-2)}$. For a sufficiently high β, the RHS is strictly less than Y^*. Thus if $Y \geq \max\{\frac{2(\gamma-\beta)}{\beta^2(n-1)(n-2)}, \frac{2(\gamma-\beta)}{\beta(\beta n-\gamma+(\beta(n-2)-\gamma)(n-2)}\}$, i does not deviate at h.

Now consider $h = a^1$ such that $m(a^1) = 2$. Signatory i does not deviate to not investing if $Y \leq \frac{2\beta-\gamma}{\beta(4\beta-\gamma)}$. Note that as β tends to γ, the RHS of the

inequality tends to $\frac{1}{3\gamma}$, while Y^* tends to zero. Thus for a sufficiently high β, signatory i does not deviate. Non-signatory i does not deviate to investing since he has no material incentive to do so and perceives all his co-players as unkind, so no reciprocity incentive to invest either.

For all $h = a^1$ such that $m(a^1) > 2$, see the proof of Theorem 1. Finally, in the sign-up stage, player i has neither material nor reciprocity incentive to deviate, thus the profile is a SRE. ■

Proof of Proposition 2

(a) Write $\rho(n)$ to denote the dependence of ρ on n. Note that for all $n \geq 4$ and $k > 0$, $\rho(n+k) - \rho(n) = 1 - \frac{\beta}{(n+k)\beta-\gamma} - 1 + \frac{\beta}{n\beta-\gamma} = \frac{\beta^2 k}{(n\beta-\gamma)((n+k)\beta-\gamma)} > 0$. Also note that $\lim_{n\to\infty} \rho = 1$. **(b)** Note that $\frac{\partial\rho}{\partial\beta} = \frac{\gamma}{(n\beta-\gamma)^2} > 0$ and that $\lim_{\beta\to\gamma} \rho = 1 - \frac{1}{n-1}$. **(c)** Note that $\frac{\partial\rho}{\partial\gamma} = -\frac{\beta\gamma}{(n\beta-\gamma)^2} < 0$. ■

Proof of Proposition 3

Note that the profile is identical to that examined in Theorem 1, other than that following a history of zero signing, zero invest. Given Theorem 1, we need only verify non-deviation following three histories: the initial history, that following zero signatories and that following one signatory.

At the initial history deviation has no effect on any player's material payoff, thus i has no incentive to deviate. Following a history of zero or one signatories, i has no material incentive to invest and since he perceives others as unkind, he has no reciprocity incentive to invest either. ■

Proof of Proposition 4

Restrict attention to strategy profiles where zero sign and zero invest on path throughout this proof. We demonstrate that given the conditions in Proposition 4, within the permitted set of strategy profiles: **(a)** any profile where there is a non-signatory who invests at some $m(a^1) = 1$ is not a SRE; **(b)** given non-signatories do not invest at $m(a^1) = 1$, any profile where there is a non-signatory who invests at some $m(a^1) = 2$ is not a SRE; **(c)** given that non-signatories do not invest if $m(a^1) \in \{1, 2\}$, any profile where a signatory does not invest at some $m(a^1) = 1$ is not a SRE; **(d)** given that non-signatories do not invest if $m(a^1) \in \{1, 2\}$, any profile where a signatory

invests at some $m(a^1) = 1$ is not a SRE; and hence there exist no SRE where zero sign and zero invest on path.

(a) Consider a strategy profile where zero sign and zero invest on path where there exists a non-signatory, i, who invests at some $m(a^1) = 1$. At this decision node, h, non-signatory i deviates to investing if

$$\beta - \gamma + \frac{y}{n-1}\beta[m_1\lambda_{SI} + m_2\lambda_{SN} + m_3\lambda_{NI} + m_4\lambda_{NN}] < 0, \tag{8}$$

where $m_1 + m_2 = m(a^1)$, $m_1 + m_2 + m_3 + m_4 = n - 1$, λ_{SI} is non-signatory i's perception of the kindness of a signatory who invests at h; λ_{SN} is non-signatory i's perception of the kindness of a signatory who does not invest at h; and λ_{NI} and λ_{NN} are defined analogously for i's perceptions of fellow non-signatories' kindness to i.

To ensure i deviates at h, we let [.] in (8) take its maximum value, then identify a condition on y such that (8) holds. To identify an upper bound to $m_1\lambda_{SI} + m_2\lambda_{SN}$, first let w denote the number of players other than non-signatory i and signatory j who invest at h. Note that $\lambda_{SI} = \lambda_{SN} = \frac{1}{2}(\beta(w+2) - \gamma)$, which is maximised if $w = n - 2$. Using $m_1 + m_2 = 1$, the maximum value of $m_1\lambda_{SI} + m_2\lambda_{SN}$ is thus $\frac{1}{2}(\beta n - \gamma)$.

We next identify an upper bound to $m_3\lambda_{NI} + m_4\lambda_{NN}$ for two different cases. First consider profiles where non-signatory i does not invest if non-signatory j were to sign. Let x and z be the number of players other than non-signatory i and non-signatory j who invest at h and the node where j has also signed respectively. Then $\lambda_{NI} = \frac{1}{2}(\beta(x-z+2)-\gamma) \leq \frac{1}{2}(n\beta-\gamma)$ and $\lambda_{NN} = \frac{1}{2}(\beta(x-z)-\gamma) < \frac{1}{2}(n\beta-\gamma)$. Thus $m_3\lambda_{NI}+m_4\lambda_{NN} \leq \frac{1}{2}(n\beta-\gamma)(n-1)$. Substituting the bounds into (8), implies that for profiles where i does not invest if any non-signatory j signs, i deviates at h if $y < \frac{2(\gamma-\beta)}{\beta(n\beta-\gamma)}$.

Second consider profiles where non-signatory i invests if non-signatory j were to sign. Then $\lambda_{NI} = \frac{\beta}{2}(x-z+1)$ and $\lambda_{NN} = \frac{\beta}{2}(x-z-1)$. Consider the profile where all non-signatories who invest at h also invest in the subgame where j also signs and $z = x - 1$. Then $\lambda_{NI} = \beta$ and $\lambda_{NN} = 0$, thus for β sufficiently small $\lambda_{NI} > \frac{1}{2}(n\beta - \gamma)$. For such β, the upper bound of $m_3\lambda_{NI}+m_4\lambda_{NN} = (n-1)\beta$. Substituting this into (8), implies that i deviates at h if $y < \frac{\gamma-\beta}{\beta^2}$.

One can show that for any other profile where zero sign and zero invest on path, at least one of the two conditions on y that we have derived is sufficient to ensure non-signatory i deviates at h. Thus if $y < \min\left\{\frac{2(\gamma-\beta)}{\beta(n\beta-\gamma)}, \frac{\gamma-\beta}{\beta^2}\right\}$, non-signatories cannot invest if there is only 1 signatory.

(b) Consider strategy profiles where non-signatories do not invest at $m(a^1) = 1$ and some non-signatory i invests at some h where $m(a^1) = 2$. Reasoning analogous to that in part (a) establishes that a sufficient condition for deviation by i at h is $y < \frac{2(\gamma-\beta)}{3\beta^2}$. Note that this implies the second condition on y derived in part (a).

(c) Consider strategy profiles where non-signatories do not invest when $m(a^1) \in \{1, 2\}$ and some signatory i does not invest at h where i is the only signatory. Let x equal 1 if i would invest if non-signatory j signs and 0 otherwise. If $\beta < \frac{\gamma}{4}$, signatory i deviates at h if $y > \frac{4(\gamma-\beta)}{\beta(\gamma-2\beta)(x+1)-2\beta)}$. Note that the RHS of this condition is highest if $x = 0$, thus if $y > \frac{4(\gamma-\beta)}{\beta(\gamma-4\beta)}$ then i deviates at h.

(d) Consider strategy profiles where non-signatories do not invest when $m(a^1) \in \{1, 2\}$ and some signatory i invests at h where i is the only signatory. Let x equal 1 if i would invest if non-signatory j signs and 0 otherwise. For $\beta < \gamma/2$, i's perception of j's kindness to i is $\frac{1}{2}(\frac{\gamma}{2}(d-2) + \beta(1-d))$ which is negative for all d. Thus i has neither reciprocity nor material incentives to invest at h and thus deviates.

To conclude the proof simply note that: $\frac{2(\gamma-\beta)}{3\beta^2} > \frac{4(\gamma-\beta)}{\beta(\gamma-4\beta)}$ if $\beta < \frac{\gamma}{10}$; $\frac{2(\gamma-\beta)}{\beta(n\beta-\gamma)} > \frac{4(\gamma-\beta)}{\beta(\gamma-4\beta)}$ if $\beta < \frac{3\gamma}{2(n+2)}$; and $\frac{3\gamma}{2(n+2)} < \frac{\gamma}{10}$ if $n > 13$. ■

References

[1] Andreoni, J. and H. Varian (1999) "Preplay contracting in the prisoners' dilemma" *Proceedings of the National Academy of Science* 96: 10933-10938.

[2] Barrett, S. (1994) "Self-enforcing international environmental agreements" *Oxford Economic Papers* 46: 878-894.

[3] Barrett, S. (2003) *Environment and Statecraft: The Strategy of Environmental Treaty-Making*. Oxford. Oxford University Press.

[4] Battaglini, M. and B. Harstad (2016) "The political economy of weak treaties" *NBER working paper* no. 22968.

[5] Battigalli, P. and M. Dufwenberg (2009) "Dynamic psychological games" *Journal of Economic Theory* 144: 1-35.

[6] Bergstrom, T., Blume, L. and H. Varian (1986) "On the private provision of public goods" *Journal of Public Economics* 29: 25-49.

[7] Bierbrauer, F. and N. Netzer (2016) "Mechanism design and intentions" *Journal of Economic Theory* 163: 557-603.

[8] Bierbrauer, F., Ockenfels, A., Pollak, A. and D. Rückert (2017) "Robust mechanism design and social preferences" *Journal of Public Economics* 149: 59-80.

[9] Boadway, R., Song, Z. and J.-F. Tremblay (2007) "Commitment and matching contributions to public goods" *Journal of Public Economics* 91: 1664-1683.

[10] Charness, G., Fréchette, G. R. and C.-Z. Qin (2007) "Endogenous transfers in the prisoner's dilemma game: An experimental test of cooperation and coordination" *Games and Economic Behavior* 60: 287-306.

[11] Charness, G. and M. Rabin (2002) "Understanding social preferences with simple tests" *Quarterly Journal of Economics* 117: 817-869.

[12] Clémençon, R. (2016) "The two sides of the Paris climate agreement" *Journal of Environment & Development* 25(1): 3-24.

[13] Dufwenberg, M. and G. Kirchsteiger (2004) "A theory of sequential reciprocity" *Games and Economic Behavior* 47: 268-298.

[14] Dufwenberg, M. and A. Patel (2017) "Reciprocity networks and the participation problem" *Games and Economic Behavior* 101: 260-272.

[15] Ellingsen, T. and E. Paltseva (2016) "Confining the Coase theorem: contracting, ownership and free-riding" *Review of Economic Studies* 83: 547-586.

[16] Falk, A. and U. Fischbacher (2006) "A theory of reciprocity" *Games and Economic Behavior* 54: 293-315.

[17] Falkinger, J., Fehr, E., Gächter, S. and R. Winter-Ebmer (2000) "A simple mechanism for the efficient provision of public goods: Experimental evidence" *American Economic Review* 90: 247-264.

[18] Fischbacher, U., Fehr, E. and S. Gächter (2001) "Are people conditionally cooperative? Evidence from a public goods experiment" *Economic Letters* 71: 397-404.

[19] Geanakoplos, J., Pearce, D. and E. Stacchetti (1989) "Psychological games and sequential rationality" *Games and Economic Behavior* 1: 60-79.

[20] Guttman, J. (1978) "Understanding collective action: Matching behavior" *American Economic Review* 68: 251-255.

[21] Guttman, J. (1987) "A non-Cournot model of voluntary collective action" *Economica* 54: 1-19.

[22] Hadjiyiannis, C., Iris, D. and C. Tabakis (2012) "International environmental cooperation under fairness and reciprocity" *B.E. Journal of Economic Analysis & Policy (Topics)* 12(1): Article 33.

[23] Jackson, M. O. and S. Wilkie (2005) "Endogenous games and mechanisms: Side payments among players" *Review of Economic Studies* 72: 543-566.

[24] Jang, D. (2015) "Reciprocity and International Environmental Agreements" in *Two Essays of Other-Regarding Preferences on Social Decision Making*, Chapter 2, PhD thesis. University of Arizona.

[25] Jaquet, J. and D. Jamieson (2016) "Soft but significant powers in Paris agreement" *Nature Climate Change* 6: 643-646.

[26] Katz, M. L. (1986) "An analysis of cooperative R&D" *RAND Journal of Economics* 17(4): 527-543.

[27] Kolstad, C. D. (2014) "International Environmental Agreements among heterogenous countries with social preferences" *NBER Working Paper* No. 20204.

[28] Martimort, D. and W. Sand-Zantman (2016) "A mechanism design approach to climate-change agreements" *Journal of the European Economic Association* 14(3): 669-718.

[29] Netzer, N. and A. Volk (2014) "Intentions and ex-post implementation" mimeo.

[30] Nyborg, K. (2018) "Reciprocal climate negotiators" *Journal of Environmental Economics and Management*, forthcoming.

[31] Rabin, M. (1993) "Incorporating fairness into game theory and economics" *American Economic Review* 83: 1281-1302.

[32] Sugden, R. (1984) "Reciprocity: the supply of public goods through voluntary contributions" *Economic Journal* 94: 772-787.

[33] Varian, H. (1994) "A solution to the problem of externalities when agents are well-informed" *American Economic Review* 84: 1278-1293.