Endogenous Institutions and Economic Policy[*]

PRELIMINARY DRAFT

James A. Robinson[†] **Alexander Vostroknutov**[‡]

Ekaterina Vostroknutova[§]

November 18, 2022

Abstract

Although the importance of institutions for economic growth is well-understood, no model with endogenously emerging institutions exists that would lend itself to policy analysis. We propose a new modelling framework, designed to fill this gap, where agents not only maximize their personal consumption, but also care about morality: they receive additional utility from cooperating with others by following social norms. Under these assumptions, we model the emergence of formal/informal and inclusive/extractive institutions for facilitating cooperation. Institutional change happens when moral agents choose between existing institutions on the basis of their profitability and "fairness" that is determined endogenously from the context of the game. We show with examples how the framework can accommodate well-known structural distortions that go along with poor quality of public institutions, such as informality and clientelism. We also demonstrate the usefulness of this approach for designing economic policy that can directly take institutions and their functioning into account. The framework allows for case-specific calibration that can help evaluate policy effectiveness, such as for example tax policy in the presence of a large informal sector.

Keywords: *endogenous institutions, social norms, inclusive/extractive institutions, growth, institutional policy.*

[*] All mistakes are our own. The findings expressed in this paper are those of the authors and do not necessarily represent the views of the World Bank or its member countries.

[†] Harris School of Public Policy and Department of Political Science, University of Chicago. e-mail: jamesrobinson@uchicago.edu.

[‡] Department of Microeconomics and Public Economics (MPE), Maastricht University, The Netherlands. e-mail: a.vostroknutov@maastrichtuniversity.nl.

[§] The World Bank. e-mail: evostroknutova@worldbank.org. (Corresponding author)

1 Introduction

In the past decades, neoclassical economics has been rather successful in accounting for the process of economic growth in industrialized countries using the Solow-Swan model and its extensions [Solow, 1956, Swan, 1956, Aghion and Howitt, 2008]. However, it has failed to predict the slow catch up in income per capita by developing countries [North et al., 1993, Keefer and Knack, 1997, Caselli, 2005, Rodrik, 2013]. While recent methodological breakthroughs allow to conclude that significant welfare gains from eliminating structural distortions can be achieved [Baqaee and Farhi, 2020], the literature does not explain the staggering proportions of these "distortions" or suggest policies to reduce them. For example, the informal sector employs the majority of workers and produces more than half of GDP in many parts of the world [Loayza, 2018, Perry et al., 2007], while political institutions in transplanted democracies are often overwhelmed with clientelism and corruption to the detriment of growth [Stokes, 2011]. The attempts to eliminate these distortions directly have failed [Floridi et al., 2020], suggesting that they are structural in the sense of being "the way of life" in many developing countries [De Soto et al., 1989], and need to be taken into account in policy making.

Chief among possible explanations of these failures has been the importance of economic institutions [North, 1990, Ostrom, 1990, Hall and Jones, 1999]. Historically, the formal institutions of the Western world (such as democratic political systems) have been shown to create long-term growth in income per capita [Fukuyama, 1995]. However, multiple studies have shown that transplanting Western institutions to other countries does not necessarily lead to a commensurate welfare improvement [Putnam et al., 1992, Raiser, 1997]. One potential reason is that institutional change is a long-term process, where institutions need to grow organically based on the existing economic system including the informal institutions and social norms already in place [Acemoglu and Robinson, 2020].

Indeed, Acemoglu and Robinson [2020] have argued that building and maintaining institutions conducive to economic growth requires enormous effort from governments and societies. Such *inclusive* institutions are located in the "narrow corridor" between a fully disorganised "state of nature," or Warre, and an all-controlling Leviathan of a powerful government. Acemoglu and Robinson painstakingly collect evidence of institutional design and historical path-dependence that have led countries to or away from this narrow corridor. But while this analysis clarifies what features the desired inclusive institutions possess that makes them so effective in creating growth, in practice only a handful of developing countries have been able to properly engineer such institutions to create some convergence to the narrow corridor. The majority of developing countries suffer through what Acemoglu and Robinson call *extractive* institutions, where a small proportion of population enjoys the benefits that all workers of the economy create, and are not therefore conducive to long-term economic growth per capita.

The general consensus across various literatures is that institutions arise to support *cooperation* of economic agents [e.g., Nee, 1998, Ostrom, 2000, Keefer and Knack, 2008, Henrich, 2015a]. Cooperation is also an important feature of inclusive institutions. As Acemoglu and Robinson [2020] note, "in the corridor [of inclusive institutions] the state and society do not just compete, they also cooperate. This cooperation engenders greater capacity for the state to deliver the things that society wants and foments greater societal mobilization to monitor this capacity." Yet, in neoclassical economics transactions are costless, agents are only interested in maximizing their own consumption utility, and firms—or collectives of agents who create surplus—are treated as exogenously given production functions. The fact that standard models do not account for cooperation is mainly the reason they are unable to model emergence of firms and other institutions [Coase, 1937].

Alternative approaches that tried to modify neoclassical paradigm to incorporate institutions also did not produce a general, viable framework for policy applications. For example, according to views expressed in institutional economics, models of institutions have lagged the intuitive and historical understanding of their importance [Williamson, 2000, Furubotn and Richter, 2010]. With majority of models being at country-level and including exogenous variables to account for quality of institutions, institutional economics has not yet arrived at the practicality of how to induce positive institutional change or orchestrate a shift from extractive to inclusive institutions. To obtain a workable model of institutional change, there is a need for a more fundamental bottom-up approach to understanding the formation and evolution of institutions that takes into account the motivation of individual agents to self-organize into cooperative "units" or firms [Sala-i Martin, 2002]. Such approach can lead to a framework where institutions arise *endogenously* and where policies can be devised to change them. To our best knowledge, no such framework currently exists.

Our study aims at filling this long-standing gap in the understanding of economic institutions and policies directed at changing them. We propose a parsimonious framework of how institutions arise and how they function within an economic environment defined by other institutions. In the version of the model described here, many of the "distortions" mentioned above (such as informality, clientelism, or corruption) appear organically and can be changed through appropriate policies suggested by the model. More importantly, the model can be adapted to various forms of institutions and economic environments, which can make it indispensable for applied and policy-oriented work related to institutional change.

Conceptually, our framework is different from neoclassical approach in only one small but fundamental aspect. While in neoclassical economic agents are assumed to have a *single motivation*, namely the desire to have more consumption utility, we assume that agents trade-off *two motivations:* the standard preference to have more consumption and *the desire to cooperate with others*. While the introduction of the latter motivation in the utility function is novel, large swaths of literatures in various social sciences have been converging on the view that humans are a social species that has evolved special mental capacity for cooperation.[1] According to this view, the innate preference for cooperation is operationalized through the desire to adhere to *social norms*, or common beliefs about how things ought to be done. In economics, this second motivation is modeled as an additional term in *norm-dependent utility function* that was first introduced by Kessler and Leider [2012]: norm-following agents maximize the sum of their consumption utility and a measure of social appropriateness defined on outcomes (the agents

[1] The reason this capacity has evolved is that cooperation is more profitable than no cooperation [e.g., Bicchieri, 2006, Henrich, 2015a, Fehr and Schurtenberger, 2018, Laland, 2018].

prefer more socially appropriate outcomes to less socially appropriate ones).

The introduction of the cooperative motivation changes dramatically the way agents behave and helps understand how institutions emerge. To be more specific, in neoclassics the concept of a cooperative enterprise (a firm) and the concept of an agent are necessarily separated because selfish agents by definition are unwilling to cooperate in social dilemmas and thus have no motivation to self-organize into firms. Firms however do exist in reality and therefore are simply brought in as production functions.[2] In our framework, the norm-following agents care about social appropriateness of their actions, which tends to be the highest when agents successfully cooperate with each other. This is the motivational force that pushes them to find ways to create new cooperative enterprises, and leads to the formation of institutions.

We assume that the desire to cooperate is not negligible and has a tangible effect on preferences. Therefore, we should expect that norm-following agents are comparatively as willing to enter cooperative relationships and form institutions as they are willing to maximize their consumption utility. This implies that institutions for cooperation should emerge in all economic environments, even those in the "state of nature," similarly to how we expect humans to strive for more consumption in all economic environments. In this view then, informality is not a distortion, but rather an attempt of norm-following agents to self-organize and cooperate in the absence of other means to do so [Perry et al., 2007]. This view also suggests that government plays a crucial role of a facilitator who helps agents to fulfill their desire to cooperate. With strong government that provides good and efficient services for doing business (e.g., in the form of enforceable laws), we should then expect that agents will prefer to cooperate legally through government rather than through informal institutions that are far less efficient given their internal structure. This argument provides a flavor of institutional policy that comes out of our framework: agents are attracted to institutions that give them the best available options to cooperate; while this is happening, the less attractive institutions die out thus bringing about institutional change.

The paper is structured as follows. In Section 2 we describe our general methodology based on the new framework to study the emergence of social norms by Kimbrough and Vostroknutov [2022a], further KV, and embed it in various literatures. This framework allows us to talk about social norms that emerge in individual interactions as building blocks of more complex institutions also termed "packages of social norms" [Henrich, 2015b].

In Section 3, we analyze the behavior of norm-following (or moral) agents in the Public Goods game. This game serves as an abstract representation of a cooperative enterprise (a "firm") that can emerge or not depending on whether agents decide to cooperate. We relate the parameters coming from the framework of KV to the intuitive notion of *trust* that is considered by many authors to be the necessary ingredient of economic development [Keefer and Scartascini, 2022]. We demonstrate that trust, operationalized by KV as having two dimensions, can account for a wide variety of intuitive social constraints

[2] The neoclassical approach remains silent about how firms and institutions emerge. A vigorous attempt to fix this came in the form of New Institutional Economics that proposed a set of models of firm formation based on various principles [e.g., Hart, 1989, Holmström and Roberts, 1998].

that prevent agents from successfully doing business with each other.

In Section 4, we define *institution for facilitation:* the simplest form of institutional arrangement among three agents that allows two of them, who do not trust each other, to nevertheless cooperate with the facilitation from a common "acquaintance" whom both agents trust (the third agent called *facilitator*). We show how the norm with respect to the amount of *rent* paid to facilitator emerges endogenously from the parameters of the Public Goods game and trust weights. This in its turn paves the way to natural definitions of *inclusive* (or norm-abiding) and *extractive* (or norm-breaking) institutions.

In Section 5, we consider how agents choose between institutions. We give definitions of *formal* and *informal* institutions, and model the choice of a norm-following agent between them. Given that the agent intrinsically cares about the normative arrangement in the two types of institutions, her choice will depend not only on the consumption utility expected from the two institutions, but also on how "fair" (inclusive or extractive) these institutions are. We finish this section with a stylized model of the dynamics of trust in government, the size of informal sector, institutional policy, and growth. This model suggests that increasing the quality of public services leads to higher trust in government, which then leads to the increase in the relative size of the formal sector.

In Section 6 we consider arbitrary trust networks and discuss how and why some agents might prefer to work (participate in the Public Goods game) or facilitate (collect rent from facilitating cooperation of others). We show with an example that depending on the structure of the trust arrangements in the population, different forms of *clientelistic networks* may emerge. Sometimes, one or two large informal institutions like that take hold over the whole network (similar to a monopoly or a duopoly), which leads to high barriers to enter and elitism. In other cases, many facilitators compete for clients (similar to perfect competition), which creates the opposite situation where clients are paid for their attention to facilitators.

Finally, Sections 7 and 8 conclude with a discussion of further directions of research and the implications of our framework for economic and institutional policy.

2 Literature and Methodology

It has been noted in the literature that institutions—both economic and political—can be a major obstacle to development [North, 2008]. The policy dilemma of the century that can be succinctly formulated as "How to change institutions?" has attracted a lot of attention in the literature, but still remains unsolved. According to the long tradition of institutional economics, institutions can be good or bad for economic development, and Western-style institutions have been shown to correlate with economic growth. However, literature also documents many failed attempts at imposing them in developing countries [Putnam et al., 1992].[3]

[3] Acemoglu and Robinson [2020] provide many examples of "transplanted" formal institutions that have never taken root because they had been overwhelmed by pre-existing social norms and informal institutions. For example, the Peruvian pre-colonial *mita* system is among the origins of extractive institutions in Latin America. The political economy literature documents many cases of the failure of transplanted democratic institutions, leading to corruption and clientelism [Acemoglu

Empirical analysis of the importance of institutions for development has traditionally treated institutions as exogenous variables with evidence based on cross-country regressions that include proxies for "institutions" such as democracy index, etc. The literature analysing specific institutions, such as for example the legal regulations guiding financial market functioning, is not typically generalizable to other situations and takes the form of case studies. Other models do not take into account the context of economic interactions, rendering policy analysis impossible. It is no surprise that, similarly to Shirley [2005] and Knack and Keefer [1995], Rodrik et al. [2002] note that empirical studies show significant regularities in how institutional variables tend to dominate others in explaining growth and social progress, "but these studies lack a theory that would transform regularities into causal explanations."

Not many formal models of institutions—with the purpose to shed some light on their role in economic activity—have been proposed in the literature. For example, Ostrom [2000] uses social dilemmas to reason about the nature of collective action and social norms in Common Pool Resource problems [see also Kimbrough and Vostroknutov, 2015]. Some other approaches appear in the literature on the emergence of firms [e.g., Hart, 1989, Holmström and Roberts, 1998]. Even though these models are capable of capturing certain elements of institutional influence on economic behavior, they cannot be used as the general models of institutional change or made more specific to arrive to more granular policy guidance.

What has been lacking is a workhorse model of institutional change. Such a model needs to be general enough to be able to account for a wide range of institutions, but at the same time it should allow for specific adaptations to accommodate the impacts of specific policies. More importantly, in such a model institutions would need to enter *endogenously* [see also Engerman and Sokoloff, 2005], so that it is possible to consider the emergence and dynamic development of institutions depending on a set of parameters. This is needed for testing the model in practice as well as for policy implications and can only be achieved if institutions are introduced at the micro level, as social norms that aggregate into institutional packages [Henrich, 2015b]. This bottom-up approach would allow to organically incorporate both formal and informal institutions [Keefer and Knack, 2008]. Moreover, apart from straightforward variables describing "institutional quality" or specific institutions, the model should directly account for the structural distortions that accompany them. For example, inclusive institutions would correlate with lower informality as they reduce exclusion [Perry et al., 2007]. Clientelism can also emerge organically and coexist with the existing, extractive political system [Muñoz, 2019].

In this paper, we offer one such model. As was mentioned above, the main difference from neoclassical approach is that agents are assumed to be moral (norm-following), or having an additional term in their norm-dependent utility that tracks social appropriateness of alternative outcomes they choose. This addition to selfish motivation makes agents want to cooperate and enter "business relationships" with each other.

It is worth mentioning that this idea is not new and that the connection between norm-abiding behav-

and Robinson, 2006]. Among many cases when formal institutions have been unsuccessfully over-imposed on existing informal institutions, the Latin American lettered city literature highlights the origins of modern widespread informality in the inconsistency of design of cities and reality of life for the poor, that has resulted in their continuous exclusion from the formal economy [Angel and Charles, 1996, Lazar, 2008].

ior and business activity has been pointed out in economics for a long time [e.g., Arrow, 1972]. However, the traditional definitions that simply suggest this connection are not enough for policy and understanding institutional change. To have a workable model of institutions, the moral agents' utility-maximizing behavior should be pinned down exactly, which means that the normative values (social appropriateness) of all outcomes should be specified.

It may seem that in a given institution the values of social appropriateness of various actions and outcomes (what is considered appropriate to do within this institution) may heavily depend on the past history of the institution, local beliefs, and other specific cultural features. This is indeed true for many existing institutions. However, it is also true that people who consider entering a new business relationship face a new context (the set of feasible allocations) that depends on the specific business opportunities available to them at that time. So, it is plausible that there are no pre-specified norms from the past, customs, or traditions that prescribe what is moral in this new context. KV propose that for such cases people have evolved *moral psychology*, or specific machinery in the brain that computes social appropriateness of outcomes solely from the information available in any current context (feasible allocations). Such computations, if done in the same way by all moral agents, should lead to formation of similar normative beliefs about social appropriateness and consequently to coordinated, cooperative behavior [see also Kimbrough and Vostroknutov, 2022b].

KV describe a specific moral psychology of norm formation in any context that is based on the idea of aggregating *dissatisfaction*. Specifically, it is assumed that the social appropriateness of each given allocation is inversely proportional to the summed dissatisfactions of all agents in that allocation. The dissatisfaction of each agent in a given allocation is high when there are many other counterfactual allocations in the context that give the agent more consumption utility and low otherwise. Thus, dissatisfactions express personal grievances of agents in a given allocation, and the normative value of the allocation is formed by aggregating all these dissatisfactions. In other words, the allocation is considered more socially appropriate, the less aggregated dissatisfaction it evokes.

This construction may seem overly complex and unnecessary. However, tests of existing experimental data have shown that the model of dissatisfaction-based norms can account for many puzzles in social behavior known to behavioral economists.[4] Specifically, it resolves many issues pertaining to the existing models of social preferences and reciprocal kindness [see Vostroknutov, 2020, for discussion] that are not context-dependent enough to accommodate certain stylized facts (e.g., radical context-dependence of social behavior [List, 2007, Bardsley, 2008] or extreme switches of moral principles with small changes in payoffs [Galeotti et al., 2018]). Moreover, the model of KV allows to introduce social weights (aka trust weights) that measure the importance of dissatisfaction of each agent in the aggregation.[5] This is an im-

[4] Experiments by Merguei et al. [2022] and Panizza et al. [2021] perform direct and rather stringent tests of KV's model and find a very good fit with the data.

[5] In each allocation, the dissatisfaction is aggregated by summing up individual dissatisfactions of all agents. The weights multiplying these individual dissatisfactions act as measures of trust to individual agents. When an agent is not trusted (low weight), her dissatisfaction will count little in the aggregation, so that it may become socially appropriate to give her less than others (and vice versa for the agents with high trust weight).

portant mechanism that accounts for various observations about social behavior within and without an in-group [Chen and Li, 2009], and that translates intuitively to one of the dimensions of trust.

We base our model on the framework of KV exactly because it seems to fit well the moral intuitions used by people in *new* contexts and because it provides a way to determine what norms may emerge in these contexts without assuming some ad hoc social motives like, for example, inequality aversion [Fehr and Schmidt, 1999] or payoff efficiency maximization [Engelmann and Strobel, 2004] that may or may not play a role in a given institution. Most importantly though, the model of KV shows how norms—or rules of social conduct that define how to behave in certain situation—can emerge *endogenously* and later become the accepted rules of an emerging institution. The fact that norms are context-dependent and change with the environment provides the way to conceptualize institutional change that we follow in this paper.

3 Model

3.1 Public Goods Game Played by Moral Agents

We start with describing the simplest possible incentive structure and the type of interaction among moral agents that will be used as a building block in applications presented later in the paper. Specifically, we assume that two moral (norm-following) agents desire to profit from mutual cooperation in the Public Goods game (PG), which is similar to game forms used by Ostrom [1990] to study common pool resource problems.[6] This interaction can be seen as representing various forms of "doing business," which can produce surplus and consequently growth of wealth through cooperation (or not). This assumption can be seen as an embodiment of the various views in institutional economics, where institutions facilitate interactions of economic agents and impact transaction costs [North, 2016]. Similarly to Ostrom [1990], we focus on a specific form of such costs related to trust and problems with cooperation as any form of economic activity requires a certain degree of trust and cooperation/coordination among the people involved. Notice also that we consider PG as a versatile example of a social dilemma that embodies a potentially variable levels of productivity (through the PG multiplier). In applications, any other game with social dilemma flavor can be utilized.

To describe how PG is played, suppose that before the game each player $i \in \{1, 2\}$ has some endowment $w \geq 0$.[7] In the game, player i chooses an amount to contribute to the public good with some PG multiplier $p \in [\frac{1}{2}, 1)$. Suppose that player i contributes $x_i \in [0, w]$, then her wealth after the game is

$$w_i = w - x_i + p(x_i + x_{-i}). \tag{1}$$

In other words, player i keeps the part of the endowment that was not contributed $(w - x_i)$ and gains the

[6] The argument easily generalizes to more than two players.

[7] A generalization to PG with unequal endowments is very interesting and easy to implement theoretically. However, for the sake of expositional simplicity we choose to not introduce it in this paper and leave it for future research.

return from the public good (the sum of contributions $x_i + x_{-i}$, where x_{-i} stands for the contribution of the other player, times p). Notice that the multiplier p can be thought of as a measure of *productivity* and defines how profitable the cooperation is. It can be influenced by various factors (e.g., it is high when good-quality public services are provided and low otherwise).

When agents are standard consumption utility maximizers, the unique Nash equilibrium of this game is to contribute nothing ($x_i^* = 0$ for all i). This represents the classic case of underprovision of public goods. When player i is assumed to be a moral agent, she chooses x_i to maximize her *norm-dependent utility* [Kessler and Leider, 2012, Krupka and Weber, 2013]

$$U_i(x_i; x_{-i}) = w_i + \phi_i \eta_i(w_i, w_{-i}). \tag{2}$$

The utility function U_i consists of two terms: the standard linear consumption utility (w_i) and the normative term ($\phi_i \eta_i(w_i, w_{-i})$) representing the desire to follow norms. Here we implicitly see w_i as the function of x_i and x_{-i} as in (1); the coefficient $\phi_i \geq 0$ defines i's propensity to follow norms; and $\eta_i : C \to \mathbb{R}$ is the *norm function of player* i that defines the *social appropriateness* of each possible outcome (w_1, w_2) in the game and arises endogenously from the set of all achievable wealth allocations defined by $C = \{(w_1, w_2) \mid (x_1, x_2) \in [0, w]^2\}$. Specifically, following KV we define η_i as

$$\begin{aligned}\eta_i(w_i, w_{-i}) = \eta_i(w_i, w_{-i}; \tau) &= -[D_i(w_i) + \tau D_{-i}(w_{-i})] \\ &= -\left[\int_{c \in C} \max\{c_i - w_i, 0\} dc + \tau \int_{c \in C} \max\{c_{-i} - w_{-i}, 0\} dc\right].\end{aligned} \tag{3}$$

Here, the notation $\eta_i(w_i, w_{-i}; \tau)$ emphasizes that the social appropriateness expressed by the norm function depends on $\tau \geq 0$, the *social weight* that player i attaches to the other player ($-i$). $D_i(w_i)$ stands for the *personal dissatisfaction* that player i feels should she receive wealth w_i after the game. This personal dissatisfaction, in its turn, is the sum of dissatisfactions due to all possible other allocations in C where player i receives more consumption utility than w_i (so, i is dissatisfied because she could have had more consumption). This is expressed with the max operator in the second line of (3), that measures dissatisfaction at w_i due to some other possible consumption utility c_i (we notate $c = (c_i, c_{-i})$ since C is a subset of $\mathbb{R}^2$).

Thus, the social appropriateness $\eta_i(w_i, w_{-i})$ of outcome (w_i, w_{-i}) from the perspective of player i is the negative of the weighted sum of the personal dissatisfactions of the two players ($D_i(w_i)$ and $D_{-i}(w_{-i})$) with the latter weighted by the exogenously given social weight τ. The idea here is that player i feels that outcomes in C are more socially appropriate when they evoke less dissatisfaction from both players, though the dissatisfaction of the other player ($-i$) can be inflated or discounted in comparison to i's own dissatisfaction (which has weight 1). For example, $\tau = 0$ would imply that player i does not normatively care about player $-i$ as a human being at all (e.g., slavery). From her perspective social appropriateness only increases when her own dissatisfaction goes down (regardless of what happens to the dissatisfaction of $-i$). The left panel on Figure 1 shows such norm function. When $\tau = 1$ we get

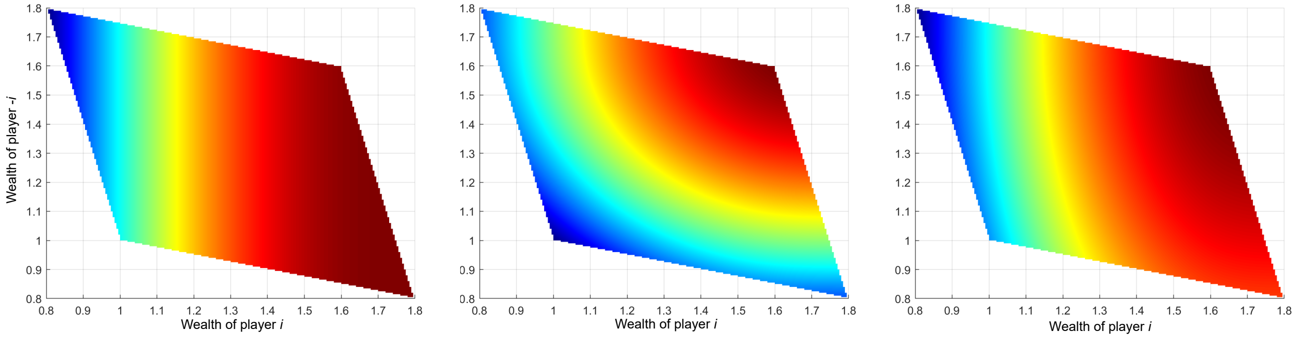

Figure 1: **Left.** The set of allocations C with $w = 1$, $p = 0.8$, and $\tau = 0$ (dark blue - the lowest social appropriateness; dark red - the highest social appropriateness). **Middle.** Same as the left panel only with $\tau = 1$. **Right.** Same as the left panel only with $\tau = 0.2$.

the case when player i treats the other player equally to herself. She cares about the dissatisfaction of the other player as much as she cares about her own (e.g., family). The norm function η_i for this case is shown in the middle panel of Figure 1. When $\tau \in (0, 1)$ we get the situation where i cares somewhat about the dissatisfaction of $-i$, but not as much as about her own. This is the case, for example, when $-i$ belongs to another social group than i. Equal partners, who are not related through family ties, might have a τ above a certain threshold, but less than 1. The right panel of Figure 1 illustrates.[8]

Given these definitions, we can talk about the *norm from the perspective of player* i as an allocation $c_i^* = \arg\max_{c \in C} \eta_i(c)$ that has the highest social appropriateness. It is clear that in general, when $\tau \neq 1$, the two players will not "see eye to eye" with respect to what they consider as the norm in this game. However, when both players treat each other as equals ($\tau = 1$ for both of them), they will agree on the norm, which should make cooperation easier to achieve (in this case $c_i^* = c_{-i}^*$ correspond to the outcome where both players contribute fully to the public good, see the middle panel of Figure 1). As we show formally in Appendix A.1, PG with moral agents has two possible Nash equilibria depending on the parameters. Given high enough ϕ_i and ϕ_{-i}, so that players want to follow norms, and when $\tau \in (0, 1]$ is high enough for both players (they care a lot about each other), the unique equilibrium is to contribute full amounts ($x_i^* = x_{-i}^* = w$). When ϕ's are high and τ's are low, the unique equilibrium is to contribute nothing ($x_i^* = x_{-i}^* = 0$). For the case of low ϕ's (when players do not care much about following norms in general), the unique Nash equilibrium is also to contribute nothing as in the case of selfish players who only maximize their own consumption utility. These possibilities are exhaustive.

3.2 Norms and Trust

As described in the previous section, agents cooperate or do business more willingly with people whom they trust, so the concept of trust is important to understanding the behavior and beliefs of moral agents. A widely used definition of trust states that it is the belief that someone will not behave opportunistically,

[8] It is important to emphasize at this point that this model has been tested against many experimental datasets (see KV) and shows a very good fit to actual human behavior.

as in Keefer and Scartascini [2022]. A more general definition is that trust is a situation-specific expectation about other agent's behavior [Bauer and Freitag, 2018]. In other words, trust is context-dependent. In particular, trust can be linked to propensity to follow specific social norms. For example, agents exhibit higher trust towards people who belong to their social group [Chen and Li, 2009]. To reflect this, we add a second dimension to the commonly used definition of trust.

The first dimension of trust corresponds directly to the definition of Keefer and Scartascini [2022]. This kind of trust is defined by the individual propensities to follow norms ϕ_i and ϕ_{-i} (and players' beliefs about them). Indeed, if player i believes, for example, that $\phi_{-i} = 0$, then she is sure that the other player is selfish and only maximizes his consumption utility. She *does not trust him,* because she believes that $-i$ will disregard the norms (whatever they may be) and contribute nothing in PG, which is $-i$'s strictly dominant strategy.

For the second dimension of trust, notice that in our model i's believing that $-i$ has high ϕ_{-i} is not enough for her to trust him. If i believes that ϕ_{-i} is high but that $-i$ does not care about her (low τ in η_{-i}), then she also will not trust $-i$. This happens not because i thinks that $-i$ is selfish, but rather because $-i$ has *different norms* that discount her, i's, importance in the eyes of $-i$. For example, the members of two warring tribes might be absolutely sure that the opponent is a highly moral norm-following individual. However, they will still not cooperate with each other (contribute in the PG) because they know that both of them do not care about the dissatisfaction of the other. They *do not trust each other due to different normative views*, [see e.g., Akerlof and Kranton, 2000, Chen and Li, 2009]. This is an important case of trust that we believe is the key to understanding the formation of institutions given that most human beings do follow norms of their—sometimes highly specific or imaginary—social groups most of the time.

In the rest of the paper, we will assume therefore that all players are moral agents who, by definition, have some propensity to follow norms (ϕ's are above zero), but that their normative views are different due to low social weights τ put on other players. Thus, we describe the world where all agents want to cooperate by following *some* norms (e.g., those of their social group), but might fail to cooperate with strangers because they think that they come from the out-group.[9]

4 Institutions for Facilitation

Institutions are "humanly devised constraints that structure human interaction," including formal constraints such as constitutions and laws and informal constraints, such as norms, conventions and self-imposed codes of conduct [North, 1990]. Describing the incentives for "doing business," Coase argued that "if the costs of making an exchange are greater than the gains which that exchange would bring, that exchange would not take place." Institutions thus emerge to reduce transaction costs associated with production [Coase, 1937].

To understand the functioning of endogenous institutions in our model, we must understand their

[9] The case with low ϕ's is not interesting anyway, as it is approximated well by standard players with selfish preferences. In this case, we know already that no cooperation can happen at all in any social dilemma.

emergence and trace their development. We start from the Hobbes' primordial "state of nature," devoid of any institutions or social norms and observe the emergence of institutions to support cooperation. We eventually arrive at a model with endogenous institutions, where institutional change is possible between inclusive and extractive institutions, depending on personal characteristics of agents involved (ϕ) and their trust to each other (τ).

4.1 The State of Nature

To understand how and for what purpose institutions may emerge, we need to consider the motivation and incentives of agents in what moral philosophers call the *state of nature* or *Warre* [Hobbes, 1651, Acemoglu and Robinson, 2020]. This is a hypothetical world of unorganized individuals who do not hold any common views on how things should be done (social norms) and thus act mostly in their own self-interest in the absence of any shared norms, regulations, or laws. Translated to the language of our model in Section 3.1, this would correspond to the situation where moral agents, although willing to follow norms, are unable to cooperate because they do not trust anyone (high ϕ, low τ). In this case, in equilibrium no two agents will contribute anything to the public good and thus there will be no growth of wealth. Such conditions are not inconceivable and existed for brief periods of time even in recent history [Leeson, 2007].[10]

Even though the state of nature precludes wide-spread cooperation with strangers due to generally low trust, some episodic cooperation and growth of wealth might exist among moral agents who know each other well and have high social weights τ attached to each other (e.g., friends). A pair of friends with high enough τ will contribute fully to the public good, thus benefiting themselves. However, it is also reasonable to think that after some time friends will exhaust all possibilities for profitable cooperation due to the fact that they most likely live in the same area and have access to the same resources. This will make the PG multiplier p, applicable to their interaction, low and the resulting growth will be negligible.

From this we can argue that moral agents would strive to cooperate with others who have access to different resources and have potentially different skills and know-how (which makes the multiplier p in the potential PG high). However, such others will typically be strangers that cannot be trusted in the state of nature (low τ). Thus, we have a situation where moral agents *want* to cooperate because it is profitable, but are *unable to do so* due to low trust. In this case, it is reasonable to believe that they will try to find *some arrangements* that will make cooperation possible.

4.2 Facilitation

The simplest possible arrangement that can expand cooperation beyond what is possible in the state of nature is *facilitation* of cooperative relationship between two moral agents by a common friend or

[10] Hobbes describes the state of nature as "war of all against all," which can be likened to negative growth. In our model, such situation is also possible when social weights attached to other agents are negative ($\tau < 0$). In this case, moral agents will find it socially appropriate to *increase* dissatisfaction of others (e.g., by appropriating or destroying their wealth).

facilitator.

Consider two moral agents in $\{1,2\}$ as in Section 3.1 with equal and low social weights $\tau_2 < \tau^*$ attached to each other, which is below the threshold τ^* above which the Nash equilibrium in the PG is to fully contribute (see Appendix A.1 for the values of τ^* as dependent on p).[11] Thus, in the PG the agents will optimally choose to contribute nothing when their mutual social weights are τ_2 (the meaning of subindex 2 will become clear in Section 6). Now suppose that there is a facilitator agent f with whom both agents maintain friendly relationships, so that the mutual social weights between any agent $i \in \{1,2\}$ and f are $\tau_1 > \tau^*$. This means that agent i would cooperate with f if the opportunity arose. The left panel of Figure 2 illustrates: the dotted line emphasizes the impossibility of cooperation.

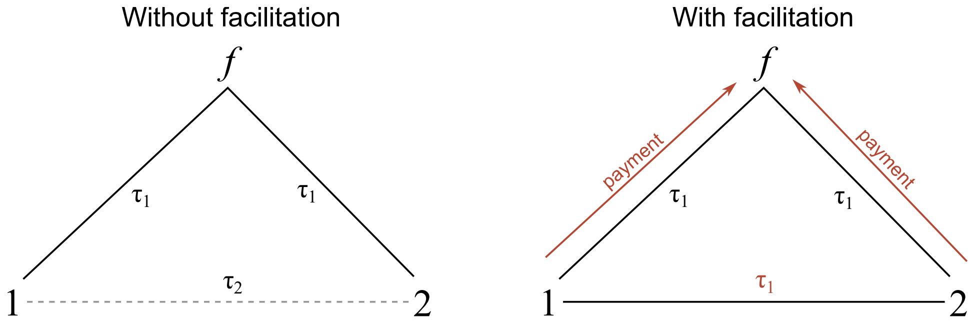

Figure 2: **Left.** Without facilitation players 1 and 2 cannot cooperate due to low social weights τ_2. **Right.** Players 1 and 2 can pay f for facilitation and cooperate with the same social weight τ_1 that they have with the facilitator.

Despite the possibility of cooperation with f, the agents want to cooperate with each other as this brings more profit than cooperation with f (see the argument above). Therefore, it is possible that the three agents agree on the following deal. Agent f facilitates the relationship between agents 1 and 2 by "vouching" for each agent as a good cooperative partner (since both are his friends). In the presence of such facilitation, agents 1 and 2 can cooperate with each other *as if* they share a friendly social weight τ_1 that allows for an equilibrium with full contributions. Such deal is of course not free and agents also agree to pay the facilitator f some amount z for his services. Suppose that each agent pays f the amount $z/2$. The right panel of Figure 2 illustrates: now players 1 and 2 can cooperate if they pay to the facilitator.

It is clear that such deals can be profitable for all three agents as long as the payment z to facilitator is "reasonable." In this case, the profit from cooperation for agents 1 and 2 minus the payment $z/2$ can exceed the profit from all other available options, of which there are not so many, and the deal will be made. This form of interaction can be thought of as the simplest form of *institution for facilitation* that allows for cooperation among strangers and that can emerge spontaneously due to its profitability to all parties involved. Of course, such institution is not perfect as it cannot connect two agents who do not share a friend. However, with time more complex institutions can evolve from this simple interaction gradually involving more and more unrelated agents.

[11] When $p = 0.8$, we have $\tau^* = 0.25$.

4.3 Norms in Institutions for Facilitation

In Section 3.1 we analyzed how normativity and the notion of social appropriateness of outcomes emerges in a simple PG. The behavior of moral agents was consequently guided by these notions and allowed us to talk about morally "right" and "wrong" outcomes, or appropriate and inappropriate behavior (which leads to right or wrong outcomes). Using the same technique from KV, we can also analyze social appropriateness of outcomes in the institution for facilitation that includes an additional player, the facilitator. This analysis will allow us to understand the behavior of moral agents as well as define norms that become the inalienable part of this institutional arrangement. Thus, we can calculate, for example, what would be considered reasonable amount of payment z to facilitator that agents 1 and 2 would find "fair." This amount is the *norm* associated with the institution for facilitation. This is an important notion that leads to the extended welfare analysis for moral agents (with norm-dependent utility) that includes standard consumption considerations as well as their utility losses or gains due to fairness or unfairness of the outcomes resulting from the institution for facilitation (see Section 5).

To do this, we consider the set of all allocations to three players (agents 1, 2, and f) that can be achieved in some (yet unspecified) non-cooperative game that represents how exactly agents interact in this institution.[12] As in Section 3.1, suppose that players 1 and 2 have endowments w before the game (facilitator does not have or need an endowment since he is not playing the PG). Suppose as well that if each player i pays the amount $z/2$ to facilitator, then both of them have $w - z/2$ left to play the PG. Thus, we can say that the wealth of player i after the game where z was paid to the facilitator is given by

$$w_i^z = w - z/2 - x_i^z + p(x_i^z + x_{-i}^z) \tag{4}$$

where $x_i^z \in [0, w - z/2]$ for both $i \in \{1, 2\}$ is the amount contributed. We define the set of achievable allocations as $C_F = \{(w_1^z, w_2^z, z) \mid z \in [0, 2w] \text{ and } (x_1^z, x_2^z) \in [0, w - z/2]^2\}$. Here we assume that all payments z are possible: from 0 (facilitator works for free) to full endowments of both players $2w$ (facilitator gets all the money and players 1 and 2 have nothing left). The set of allocations C_F is a pyramid in $\mathbb{R}^3$ shown on the left panel of Figure 3.

The definition of C_F is all we need to specify the norm functions of the three players. For player $i \in \{1, 2\}$, we have

$$\begin{aligned}&\eta_i(w_i^z, w_{-i}^z, z) = -[D_i(w_i^z) + \tau_1 D_{-i}(w_{-i}^z) + \tau_1 D_f(z)] \\ &= -\left[\int_{c \in C_F} \max\{c_i - w_i^z, 0\}dc + \tau_1 \int_{c \in C_F} \max\{c_{-i} - w_{-i}^z, 0\}dc + \tau_1 \int_{c \in C_F} \max\{c_f - z, 0\}dc\right].\end{aligned} \tag{5}$$

Here we notate $c = (c_i, c_{-i}, c_f) \in C_F \subseteq \mathbb{R}^3$. In words, the social appropriateness of an outcome

[12] An important feature of the framework of KV is that norm functions or measures of social appropriateness are defined on the sets of allocations that ignore the non-cooperative game structure (who moves when, which actions are available, etc.). It allows us to analyze games in the style of cooperative game theory without specifying ex ante how the game unfolds. This property can be very useful for applications where the exact game structure is often unknown or changes depending on circumstances. See Kimbrough and Vostroknutov [2022b] for the additional discussion.

$(w_i^z, w_{-i}^z, z) \in C_F$ from the perspective of player i is the weighted sum of personal dissatisfactions of the three players, with social weights τ_1 for the other two players (as discussed in Section 4.2).

The norm function for the facilitator f is

$$\begin{aligned}\eta_f(w_1^z, w_2^z, z) &= -[D_f(z) + \tau_1 D_1(w_1^z) + \tau_1 D_2(w_2^z)] \\ &= -\left[\int_{c \in C_F} \max\{c_f - z, 0\} dc + \tau_1 \int_{c \in C_F} \max\{c_1 - w_1^z, 0\} dc + \tau_1 \int_{c \in C_F} \max\{c_2 - w_2^z, 0\} dc\right]. \quad (6)\end{aligned}$$

This is just the reshuffling of the same personal dissatisfactions as above, only with social weights τ_1 applied to the players 1 and 2.

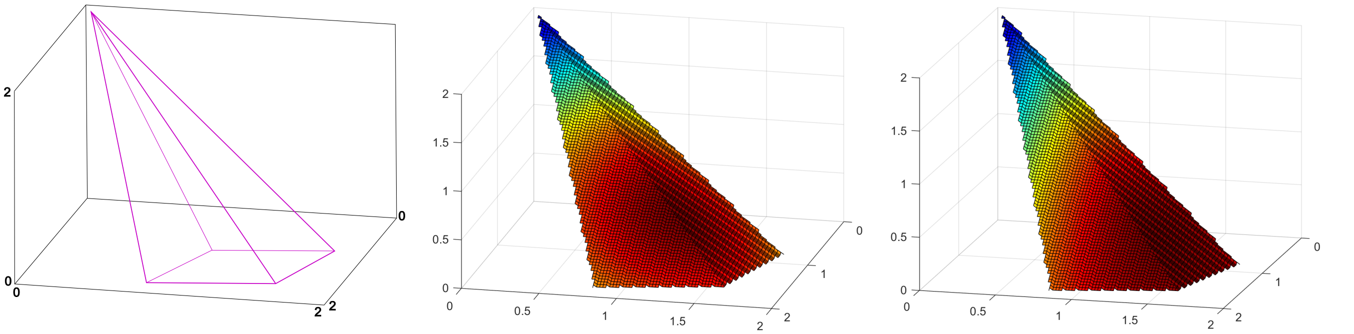

Figure 3: **Left.** The set of allocations C_F (with $w = 1$ and $p = 0.8$). The wealths of players 1 and 2 are on the x- and y-axes; the wealth of the facilitator on the z-axis. **Middle.** The Pareto frontier of the set of allocations C_F in the PG with facilitation. The colors denote the norm function η_i with $\tau_1 = 1$ (dark blue - the lowest social appropriateness; dark red - the highest social appropriateness). **Right.** Same as the middle panel only with $\tau_1 = 0.2$.

We illustrate these functions in Figure 3. Specifically, the middle panel shows the unique norm function for all three players that arises when $\tau_1 = 1$, or when both players 1 and 2 trust the facilitator as they trust themselves (e.g., family). This case represents the most cooperative situation achievable within the institution where all players treat each other as they treat themselves (of course, players 1 and 2 do not trust each other much outside the institution for facilitation, they have weight τ_2, see Section 4.2). Notice that in this case the norm (or the most appropriate allocation) obtained from maximization of (6) and marked on the graph in dark red gives 26% of $2w$ to the facilitator ($z = 0.26 \cdot 2w$) and players 1 and 2 should choose to contribute full amounts of what is left after this payment. The right panel of Figure 3 shows the norm function of player 1 for the case when $\tau_1 = 0.2$. In this case, unsurprisingly, player 1 considers it socially appropriate to have more than both other players, which can be seen from the dark red spot being shifted to one side of the pyramid.

The important implication of these calculations, especially that the payment to the facilitator in the most cooperative case is 26% of the players' endowments, is that moral agents can use this as a benchmark to judge how *fair* the payment to the facilitator is. For example, when the facilitator is an actual family member of players 1 and 2 (e.g., brother to one and son-in-law to the other) and demands the payment of, say, 30%, then both players will find it inappropriate and will think that the facilitator is breaking the norm. They might not agree to this arrangement on the basis of their moral convictions and might seek facilitation elsewhere. Similar arguments can be made about taxes and facilitation by the government or some informal institution.

4.4 Inclusive and Extractive Institutions

Now that we know which norms emerge in institutions for facilitation, we can add the non-cooperative game structure to the institution (the explicit specification of the sequence of moves and available actions of the three players) and analyze the various types of equilibrium behavior that result depending on the parameters. In this section, we will assume that there is only one institution for facilitation available (this is relaxed in the next section) and that players 1 and 2 are "forced" to participate in it, given that without this institution they cannot make any profit at all (they do not trust each other enough to cooperate themselves).[13]

This analysis will allow us to discern two broad classes of equilibria (for different norm-dependent-utility parameters) that can be likened to what literature calls *inclusive* and *extractive* institutions [Robinson and Acemoglu, 2012]. Note that the definitions used in the model are more narrow and model-dependent than the broad concepts of inclusive and extractive institutions used in institutional economics. The key properties of these institutions remain the same however: the pluralistic nature of inclusive institutions, where all can participate (at an affordable cost), that leads to more cooperation and production; and exclusive nature of extractive institutions that allow for a group of agents to extract rents above the "fair" cost, from the rest of the population.

Given that the state of nature does not provide players 1 and 2 with much outside options, it is not so hard to imagine that the facilitator can exercise power over them since without him players 1 and 2 cannot make any profits. Thus, we define the non-cooperative game as follows. First, the facilitator makes players 1 and 2 a take-it-or-leave-it offer of the payment $z \in [0, 2w]$ that he wants for his services. Then, players 1 and 2 pay the requested amount z and play the PG with endowments $w - z/2$. As before, we assume that players 1 and 2 attach social weight $\tau_1 > \tau^*$ to the facilitator (and each other within the institution), so that they optimally cooperate in each subgame happening after any offer z. Notice that in the non-cooperative game all that matters is that players 1 and 2 *believe* that the social weight that they attach to the facilitator and guaranteed by him for their own relationship is τ_1 (high enough for cooperation to be the Nash equilibrium in the subgames). This is important since, in principle, the facilitator does not have to respect or genuinely care about players 1 and 2 at all. All he needs to do for the institution to work (which leads to cooperation and payment of z to himself) is to *convince* the players that he can be trusted.[14]

The last ingredient of the non-cooperative game are the utilities that the players eventually receive.

[13] The nature of this assumption is not that far-fetched as it may seem. It is often the case in developing countries that people are in such dire economic circumstances that they cling to any opportunity to make money.

[14] In Section 6 we will show how this can lead to the emergence of clientelistic networks.

This is straightforward: the norm-dependent utility of player $i \in \{1, 2\}$ is given by

$$U_i(x_i; x_{-i}, z) = w_i^z + \phi_i \eta_i(w_i^z, w_{-i}^z, z), \tag{7}$$

and the norm-dependent utility of the facilitator is

$$U_f(z; x_1, x_2) = z + \phi_f \eta_f(w_1^z, w_2^z, z). \tag{8}$$

The analysis of the equilibrium behavior proceeds by backward induction. Given that players 1 and 2 attach high social weight τ_1 to each other within the institution and assuming that their propensities to follow norms ϕ_1 and ϕ_2 are high enough, they will choose full contributions $x_i^*(z) = w - z/2$ given any offer z (see Appendix A.2 for details).[15] In the next step of backward induction, the facilitator takes the equilibrium actions $x_i^*(z)$ as given and solves the following maximization problem:

$$\max_{z \in [0,2w]} z + \phi_f \eta_f(p(2w - z), p(2w - z), z).$$

Notice that here we use the wealth of player i given equilibrium play $x_i^*(z)$ and $x_{-i}^*(z)$ determined as $w_i^z = w - z/2 - x_i^*(z) + p(x_i^*(z) + x_{-i}^*(z)) = p(2w - z)$ for $i \in \{1, 2\}$.

What will determine optimal z^* that solves this maximization problem? As follows from the proof of Midpoint Theorem [see Proposition 9 in Kimbrough and Vostroknutov, 2022b], in the vicinity of the optimum the maximand function is a downward sloping parabola such that $\lim_{\phi_f \to 0} z^* = 2w$ and $\lim_{\phi_f \to \infty} z^* = \eta_f^*(w, p, \tau_1)$, where $\eta_f^*(w, p, \tau_1)$ is the maximum of η_f given parameters w, p and τ_1.[16] This means that selfish facilitator with $\phi_f = 0$ will demand the highest possible payment of $2w$ thus leaving players 1 and 2 with nothing, and the most rule-following facilitator with $\phi_f \to \infty$ will ask for the payment that is the most socially appropriate from his perspective (the maximum of η_f) given the optimal actions $x_i^*(z)$ of players 1 and 2 in the subgames. This completely describes the Subgame-Perfect Nash equilibrium in this game.

From the analysis above, it is clear that the "kind" of institution we get in equilibrium depends solely on the payment z^* demanded by the facilitator (since players 1 and 2 fully cooperate in all subgames). Thus, we can classify institutions depending on the value of z^* that also determines the resulting welfare of players 1 and 2. When the facilitator is selfish ($\phi_f = 0$), he extracts everything from the two players ($z^* = 2w$), which leads to the wealth allocation $(0, 0, 2w)$. This is the worst allocation from the perspective of the welfare of the two players not only because they are left with nothing, but also because this allocation is considered highly inappropriate by both of them. Players 1 and 2 get the lowest possible level of normative part of their norm-dependent utility at $(0, 0, 2w)$, see Figure 3. We formulate this as a

[15] The case when ϕ_1 and/or ϕ_2 are low is uninteresting. Here players 1 and 2 will not contribute anything to the public good and cooperation will fail. Given that in our framework everyone strives to cooperate to make money, we assume that players 1 and 2 are norm-followers to a sufficient degree.

[16] By the symmetric nature of η_f (with respect to players 1 and 2), its global maximum coincides with the maximum of $\eta_f(p(2w - z), p(2w - z), z)$ as a function of z.

definition.

Definition 1. *We call an institution for facilitation* **fully extractive** *when players* 1 *and* 2 *pay their full endowments to the facilitator, which gives them the lowest possible level of both consumption and normative utility, and the facilitator gets the highest possible payment.*

Fully extractive institutions can emerge for two different reasons that are related to the two dimensions of trust. The first possibility is when the facilitator is completely selfish ($\phi_f = 0$) and thus disregards the norms: he will act as a standard consumption utility maximizer and extract all surplus from players 1 and 2. The second possibility, mentioned above, is that the facilitator is not selfish ($\phi_f > 0$) but cares little for players 1 and 2 (e.g., his real trust weight towards them is very low). However, facilitator manages to convince players 1 and 2 that he can be trusted ($\tau_1 > \tau^*$). Then, players 1 and 2 may agree to participate in the institution, only to learn later that they lose all their endowments.[17]

Notice that fully extractive institutions can emerge for arbitrary levels of trust τ_1 among agents in the institution. This is so because the other dimension of trust, namely the propensity to follow norms ϕ_f, is responsible for creating such conditions, which eliminates the dependency of the equilibrium on trust all together. In order to define the other extreme where the propensity to follow norms is high, we need therefore to take into account both dimensions of trust (ϕ's and τ's). We formulate it as follows.

Definition 2. *We call an institution for facilitation* **fully inclusive** *when* $\tau_1 = 1$*, or when all players in the institution care about each other as they care for themselves, and when the payment to the facilitator is consistent with the norm and is equal to* $\eta_f^*(w, p, 1)$*, or the payment that all three players find most socially appropriate.*

These conditions are reached in equilibrium when $\phi_f \to \infty$, or when the facilitator only cares about social appropriateness and does not care about his own consumption (and when the trust in the institution is the highest, $\tau_1 = 1$).[18]

In between fully extractive and fully inclusive institutions lies a continuum of possibilities that are characterized by non-extreme values of the two dimensions of trust: ϕ_f and τ_1. The problem with classifying this continuum into some flavors of "extractiveness" and "inclusiveness" lies in the fact that for $\tau_1 \in (0, 1)$ the norm functions η_1, η_2, and η_f will in general be different and have three different maxima (all players 1, 2, and f individually believe that they deserve more than the others). Thus, we need to pick a moral perspective from which to judge the welfare qualities of the institution. We propose to take the perspective of the players 1 and 2 since it is their welfare (and not that of the facilitator) that is important for economic policy. Moreover, the maxima of the norm functions η_1 and η_2 are reached at the same value $\bar{z}(w, p, \tau_1)$ of payment to the facilitator (the functions are symmetric). So, players 1 and 2, although disagreeing on how much they themselves should get in the most socially appropriate allocation, nonetheless agree on the most socially appropriate payment to the facilitator. We use this

[17] An example of fully extractive institution is slavery.

[18] For example, family or a well-functioning democracy are inclusive institutions.

fact in the definition.

Definition 3. *Suppose that $\tau_1 \in (0,1)$. Then call an institution for facilitation* **inclusive** *if the payment to the facilitator is consistent with the norms of players 1 and 2 and is equal to $\bar{z}(w,p,\tau_1)$. Such payment is considered fair by players 1 and 2. Else, call the institution for facilitation* **extractive**.

Notice that in the last definition we do not specify what parameters of the norm-dependent utility of the facilitator lead in equilibrium to inclusive or extractive institutions. The reason is that, in general, when $\tau_1 \in (0,1)$ the facilitator who maximizes his norm-dependent utility will always ask for strictly more payment than $\bar{z}(w,p,\tau_1)$. This is so simply because the facilitator cares more about his own dissatisfaction in the norm function η_f than he cares about the dissatisfactions of players 1 and 2. Thus, any informal relationship with $\tau_1 \in (0,1)$ will *always* lead to extractive institutions. Only government that, in principle, does not have to follow its own norm-dependent utility since it is not a human being can set the payment to itself to be equal to $\bar{z}(w,p,\tau_1)$, thus creating an inclusive institution.

5 Choice of an Institution

Today, most people live in developing countries, many of which have failed to create or sustain strong inclusive institutions [Acemoglu and Robinson, 2020]. As formal institutions are ineffective or not accessible, ”individuals enforce most bargains using informal institutions...and they have little trust in or trade with people not subject to these mechanisms” [Shirley, 2005]. We therefore assume that both formal and informal institutions exist, and that agents can choose between them.

In the previous section, we have provided the analysis of the behavior of moral agents in a single institution for facilitation in which players 1 and 2 were forced to participate. We now assume that agents can try to cooperate with others through an informal institution (and pay some amount to the informal facilitator) or they can do business through government (and pay taxes). In other words, people have outside options, and in this section we consider the implications of the availability of such options; their influence on the optimal demand of payment by the informal facilitator; and the resulting choice of institution by players 1 and 2.

5.1 Formal and Informal Institutions

In our model, both *informal institutions* and *government* (formal institution) work using the same facilitation mechanism described above. However, there are two tangible differences that we would like to emphasize. In the economy with players 1 and 2, an informal institution is an institution for facilitation where the facilitator is another player f with his specific norm-dependent utility as in (8). This means that such *informal facilitator* is a human being with some specific perspective on morality (as expressed by η_f) who maximizes his norm-dependent utility. Moreover, informal facilitator collects the payment from the two players for himself and does not invest it in anything useful for them (e.g., public services). Thus, the informal institution is characterized by the norm-dependent utility of the informal facilitator

U_f (as in (8)) and the trust in informal facilitator τ_1 that determines the fairness of different levels of payments to him from the perspective of players 1 and 2.

Government is a formal institution and another facilitator that also collects payment for its services, or taxes. But unlike the informal facilitator, government is not a human being, thus it does not have a specific morality described by the norm-dependent utility and consequently does not maximize it. Government's job is to facilitate business transactions, collect taxes, and invest them into public services that enhance productivity p in the future (see Section 5.3). Another important parameter that defines government's facilitation is the amount of trust in it determined by the social weight τ_g that enters players' norm-dependent utilities (with τ_g instead of τ_1 in (5)) and determines the fairness of tax payments from the perspective of players 1 and 2.

5.2 Model of Choice between Institutions

To model the choice of institution by players 1 and 2, we consider the following parsimonious set-up that captures the main intuition of such choice. Suppose that the payment that government asks in taxes, in case players 1 and 2 use its facilitation services, is fixed before the game at z_g and known to all players. We assume that z_g is given and fixed since, realistically, government cannot adjust taxes quickly at any desired moment in time. Though, we can still work out the consequences of different choices of z_g by comparing the optimal responses of other players in the game (comparative statics). Suppose as well that other parameters, namely, w, p, τ_1, and τ_g are known and fixed. The game proceeds as follows. First, the informal facilitator makes an offer for payment z to him from players 1 and 2. Then players 1 and 2 compute the norm-dependent utilities they would get in both possible institutions (informal and formal) and choose an institution that brings them the highest utility. To avoid coordination problems (when the two players want to choose different institutions), we assume for the sake of the argument that $\phi_1 = \phi_2$, so that both players 1 and 2 will prefer the same institution. We also assume as above that ϕ_1 and ϕ_2 are high enough for players to be able to cooperate. These assumptions will be relaxed in the network economies with more than two players that we consider in later sections.

To analyze the game resulting from the above description, we again use backward induction and consider the choices of players 1 and 2 between institutions (the choices within institutions are already computed in Section 4). Player $i \in \{1, 2\}$ chooses formal institution when in the corresponding subgame her equilibrium norm-dependent utility in it is higher than that in the informal institution. This can be expressed using (7) as

$$w_i^{z_g} + \phi_i \eta_i(w_i^{z_g}, w_{-i}^{z_g}, z_g; \tau_g) \geq w_i^z + \phi_i \eta_i(w_i^z, w_{-i}^z, z; \tau_1).$$

Plugging equilibrium choices corresponding to full cooperation $x_i^z = w - z/2$ (and the same for z_g), we obtain

$$p(2w - z_g) + \phi_i \eta_i(p(2w - z_g), p(2w - z_g), z_g; \tau_g) \geq p(2w - z) + \phi_i \eta_i(p(2w - z), p(2w - z), z; \tau_1)$$

or

$$\phi_i[\eta_i(p(2w - z_g), p(2w - z_g), z_g; \tau_g) - \eta_i(p(2w - z), p(2w - z), z; \tau_1)] \geq p(z_g - z).$$

Notice that now η_i is a function of z_g and τ_g or z and τ_1, so we simplify the notation and rewrite the above as

$$\phi_i[\eta_i(z_g; \tau_g) - \eta_i(z; \tau_1)] \geq p(z_g - z). \tag{9}$$

Inequality (9) is the condition under which player i will choose the formal institution (government) for facilitation over the informal one (remember, player $-i$ has the exactly same condition, since by assumption $\phi_1 = \phi_2$). For player i with high ϕ_i, who cares about norms enough to be able to cooperate (an assumption made above that is also necessary for the full contributions in the subgames), this inequality represents a trade-off between the social appropriateness of participating in the two institutions (which is important for player i) and the payments that should be made in them.

Going up the game tree, we now consider the choice of offer z by the informal facilitator f, who understands that he can only make profit z when (9) is *not* satisfied (otherwise f gets nothing).[19] In general, this choice of z depends on z_g, τ_g, τ_1 in potentially non-trivial way. However, it is not hard to see what the choice of z will be for the realistic case when 1) player i considers the tax z_g too high and inappropriate given the level of trust τ_g that player i has for the government; and 2) when $\tau_g = \tau_1 = \tau$. The latter condition is not too restrictive as the most socially appropriate level of taxes (or payments to informal facilitator) are not very sensitive to changes in τ when it is higher than the minimum level of trust τ^* necessary for cooperation.

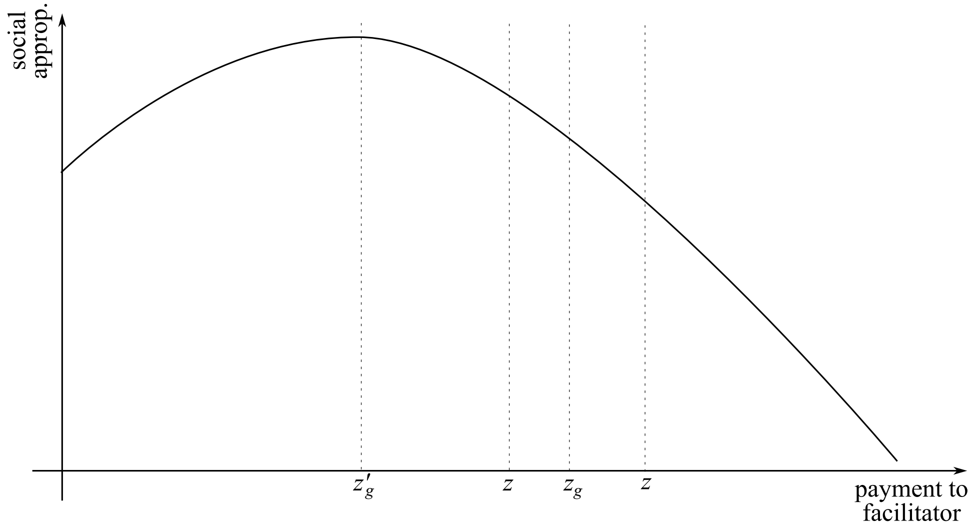

Figure 4: The graph of $\eta_i(z; \tau)$. Amount z_g is the realistic, inappropriate level of taxation by the government and z'_g is the appropriate level of taxation. The choices of z by informal facilitator either make player i choose the government (z to the right of z_g) or the informal institution (z to the left of z_g).

For this case, as Figure 4 illustrates, the choice of z is straightforward. If the informal facilitator

[19] Technically, the informal facilitator compares his norm-dependent utility U_f as in (8) in case z is accepted to his U_f when the players choose the formal institution, in which case we assume the informal facilitator's normative part of the utility is the lowest: the deal that could have been made was not made, which creates dissatisfaction.

chooses $z > z_g$, the inequality (9) will be trivially satisfied as $z_g - z < 0$ and $\eta_i(z_g;\tau) - \eta_i(z;\tau) > 0$. In this case players 1 and 2 will prefer the government. Thus, the informal facilitator will choose some $z < z_g$ to guarantee the opposite. The optimal value of z will be just below z_g if the informal facilitator is selfish ($\phi_f = 0$) and potentially something lower if he cares about norms ($\phi_f > 0$). Either way, by choosing z smaller than z_g the informal facilitator can always lure players 1 and 2 to use his institution instead of the government.

This situation is often observed in reality. For example, the informal sector in some developing countries employs more than 80% of the labor force and produces well over half of GDP. Following the seminal review in Perry et al. [2007], we can recognize two mechanisms that are at play. First, informal workers and firms are excluded from the formal economy in the sense that while they may not be paying taxes they also cannot enjoy the benefits provided by formality, such as contract enforcement or pensions. Second, economic agents and organizations (firms) in the informal sector have likely opted out of the formal sector based on a cost-benefit analysis weighing taxes and regulations against state's enforcement effort and capability [following Hirschman, 1970].

How can the government avoid such unappealing situation? The model suggests that the only way to do that is to make sure that z_g is as socially appropriate as possible from the perspective of player i with trust τ_g (e.g., z'_g on Figure 4). In this case, $\eta_i(z_g;\tau_g)$ will be the highest and the undercutting by informal facilitator will be problematic, because now he will have to compromise the social appropriateness of his institution.

This suggests that in order to make players prefer the government to informal institution, the government should try to *align the level of taxes with their perceived social appropriateness*. In the model, this can be achieved in two ways. First, the government can lower the taxes. This option is however very limited. Leaving aside the very inefficiency that prompts the use of informal institutions, the government (unlike the informal facilitator) is a complex institution that has costs and obligations to provide public services, so the taxes cannot be too low. The second way is to increase trust in government τ_g that will then make the existing level of taxes seem more socially appropriate. This latter strategy would be consistent with reducing the main driver of informality: "a massive opting out of formal institutions by firms and individuals [...] implies a blunt societal indictment of the quality of the state's service provision and its enforcement capability" [Perry et al., 2007, p. 2].

5.3 Dynamics of Trust in Government and Policy

How the trust in government τ_g can be increased? The classic view is based on its interconnection to the quality of institutions, where high quality of institutions can compensate for mistrust; for example Keefer and Scartascini [2022] emphasize prosecutors and audit agencies, dispute resolution systems, and electoral processes. Other approaches focus on the role of the social contract, citizenship, and collective action [Keefer et al., 2021]. In the latter vein, a policy that can increase collective action also increases trust in government. For example, Falconi and Robinson [2021] suggest promoting collective identities, ideas that have a unifying effect on the population (e.g., Peruvian cuisine). In theory, such a unifying national idea can spill over to the government and make people believe that it represents their national interests thus increasing τ_g. In this section, we take a more traditional view and propose one intuition using a reduced-form dynamic model of public service provision. We link trust to the relative appropriateness of formal and informal institutions, and to quality of public services. In doing so, we also shed some light on the relationship between the size of formal and informal sectors and growth that is stimulated through the connection between the amount of public services and the quality of cooperation in PG or productivity (represented by the multiplier p).

For this exercise, we make several simplifying assumptions that do not compromise the main intuition, but make things more tractable. First, we assume that the informal facilitator is selfish ($\phi_f = 0$). This assumption has one implication that will simplify the analysis. Selfish informal facilitator in the settings described above will always optimally choose $z = z_g - \varepsilon$, where ε is some small number, to barely undercut the government and attract players to his institution (the logic of this is similar to the standard Bertrand competition). This makes the right-hand side of inequality (9) equal to 0. Thus, the condition for the players to use the formal institution becomes

$$\phi_i[\eta_i(z_g; \tau_g) - \eta_i(z; \tau_1)] \geq 0. \tag{10}$$

Given this, we assume that there is a continuum of players on $[0, 1]$, who in each discrete time period $t = 1, 2, ...$ choose whether to join the formal or informal institution. We make a reduced-form assumption that, given the quantity $d_t = \eta_i(z_g; \tau_{gt}) - \eta_i(z_g; \tau_1)$, the measure $R(d_t)$ of players joins the formal institution. We assume that: $R(0) = 0$ (when the appropriateness of formal and informal institutions is the same everyone joins the informal one) and $R()$ is an increasing function (the more appropriate the formal institution is, relatively to the informal one, the more players join it). The latter assumption is not that unrealistic, as heterogeneity in ϕ_i in the population will generate optimal solutions of this sort anyway.

Next, we assume that the more players join the formal institution, the more taxes they pay, so the amount of total tax collected in period t is equal to $Z_t = R(d_t)z_g$. The quantity Z_t is spent by the government on public services that influence productivity, or multiplier p in the PG. Suppose that in period t we have productivity $p_t = p_{t-1} + \pi(Z_{t-1})$, where $\pi()$ is an increasing function with $\pi(0) = -\pi < 0$. The idea here is that without constant provision of public services, the productivity can decrease due to depreciation of the existing ones.

Finally, assume that trust in government τ_{gt} in period t is related to the amount of public services that it provided in the previous period. When people see that the government is doing what it is supposed to do (provide public services), they start trusting it more. So, $\tau_{gt} = \tau_{gt-1} + g(Z_{t-1})$, where $g(Z_{t-1}) = \bar{g} > 0$ when Z_{t-1} is higher than some fixed threshold $\bar{Z}$ and $g(Z_{t-1}) = -\bar{g} < 0$ otherwise (we proxy the amount of public services by the amount of taxes collected). The idea here is that people's trust in government increases by a constant when they see that enough public services is provided and decreases otherwise.

Thus, we have the system of equations

$$\begin{aligned} d_t &= \eta_i(z_g; \tau_{gt}) - \eta_i(z_g; \tau_1) \\ \tau_{gt} &= \tau_{gt-1} + g(R(d_{t-1})z_g) \\ p_t &= p_{t-1} + \pi(R(d_{t-1})z_g) \end{aligned}$$

The first two equations determine the joint dynamics of trust τ_{gt}, the difference in the appropriateness of the two institutions d_t, and other related variables. Growth in represented by the third equation and expressed in terms of productivity of the PG. To understand how this system evolves, notice first that if the government does not pay specific attention to d_t, then the situation in the economy will deteriorate with time. Suppose that $d_1 = 0$, which happens when $\tau_{g1} = \tau_1$, or the trust in the formal and informal institutions is the same. In this case, informal sector constitutes 100% of the economy, trust in government decreases with time (because no new public services are provided), and productivity deteriorates due to depreciation for the same reason, which leads to negative growth. As time unfolds, the situation will get worse and worse and the economy will get to the state where no public services are provided at all.

If however, we start from a good initial conditions where d_1 is positive enough so that enough people join the formal sector, pay taxes, and $R(d_t)z_g > \bar{Z}$, then the positive feedback loop in the first two equations will start. High tax revenues will generate a lot of public services that will increase trust. That in its turn will increase d_t even more. This will increase the proportion of the formal sector $R(d_t)$, which will consequently lead to more taxes and more public services. This in its turn will increase trust even further, etc. On this path, the informal sector will shrink with time and disappear, with all agents choosing the formal sector for their transactions. All this will be accompanied by increasing productivity p_t, which will lead to economic growth.

In this model, the opting out of the formal sector happens because taxes are considered too high for the public services that are provided, and trust in government is low. As a result, a vicious cycle of high informality and low trust ensues. According to our dynamic system, a way to start a virtuous cycle would be to create a positive shock to public services provision. If the government could improve public services consistently (for a number of periods), trust τ_{gt} will gradually increase. As trust increases, a threshold will be reached at some point after which d_t will become positive. At this moment, some players will join the formal sector, which will then be able to increase taxes and provide more public services. This will feed back into growing trust and more people joining the formal sector until at some point we will have $R(d_t)z_g > \bar{Z}$. Once in this positive feedback loop, the economy will be growing, the formal sector will expand, and the informal sector will shrink.

6 General Trust Network

Up to this point, our arguments were built on a micro-level interactions between four agents: players 1 and 2, the informal facilitator f, and the government. The relationship between the two players and the informal facilitator was based on a small trust network (see Figure 2) where nodes represent people and the links are the social weights between them. Also, the roles of the players were fixed: players 1 and 2 were those who tried to cooperate and the informal facilitator always facilitated and did not participate in any cooperative activities. In this section, we expand this idea to a general trust network covering arbitrary number of agents that can represent the whole economy. In such general network, moral agents have a choice between cooperation (doing business with each other) or facilitation (collecting rent). We will provide some ideas about what determines the choice of becoming cooperator or facilitator, which will consequently shed some light on the emergence and evolution of informal institutions under various conditions. Then we apply these ideas to understanding economics of clientelistic networks [Stokes, 2011, Muñoz, 2019].

Suppose that there is a set of players N with connections defined by $c_{ij} \in \{0, 1\}$ for $i, j \in N$. If i and j "know each other" then $c_{ij} = c_{ji} = 1$, else $c_{ij} = c_{ji} = 0$. The weights c_{ij} describe the "network of friends." Let us define $n_{ij} \in \mathbb{N}$ as the *minimal* number of links that connect players i and j. If they are direct friends with $c_{ij} = 1$, or *1-friends*, then $n_{ij} = 1$. If i and j share a friend but themselves are not friends, then they are *2-friends* with $n_{ij} = 2$, etc.[20]

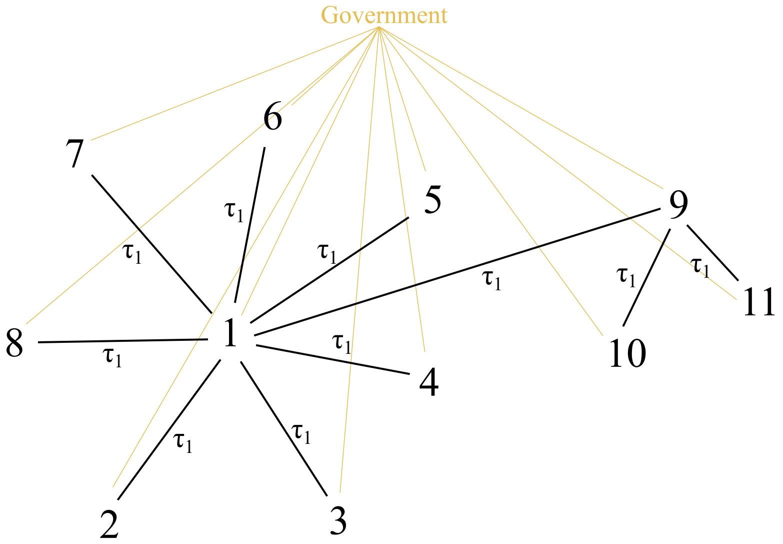

Figure 5: An example of a trust network (informal, in black) and the government network (formal, in orange).

The network described above not only defines the structure of friendship connections but also the levels of trust among them. If two players are 1-friends ($n_{ij} = 1$) then when they play some game with each other they use equal social weights τ_1 in their norm-dependent utilities (as in (3)) and trust each other to the degree $\tau_{n_{ij}} = \tau_1 \in [0, 1]$. In general, we assume that the level of trust is weakly decreasing

[20] In the previous sections, players 1 and 2 are 2-friends.

in n_{ij}, so the further away the players are on the network, the less they trust each other. In other words $\tau_1 \geq \tau_2 \geq \tau_3 \geq \ldots$ This definition takes into account possible "parochial tendencies" of the players [Kimbrough and Vostroknutov, 2022c] when the weights quickly decrease to 0 (e.g., if $\tau_2 = 0$, then players only trust their friends and no one else), but at the same time leaves room for some trust towards strangers $\tau > 0$ when $\lim_{k\to\infty} \tau_k = \tau$. For convenience, let us define the level of trust between players i and j as $\tau_{ij} = \tau_{n_{ij}}$ and assume that $\tau_{ii} = \tau_0 \equiv 1$ for all $i \in N$ (any player trusts herself at least as much as her 1-friends, $\tau_0 \geq \tau_1$). Figure 5 shows an example of a trust network with links (in black) representing 1-friendships.

The trust network is the foundation on which informal institutions are built: as time unfolds, informal facilitation will generate some cooperation between 2-friends (and in more complex settings possibly between 3-friends, etc.). However, formal institutions, or government, are based on a slightly different idea. Government is modeled here as an agent to whom *all* agents in the economy are connected by a direct link (in orange in Figure 5). This means that—through laws, regulations, and public services—government can act as a facilitator for any two agents, thus overcoming the parochial nature of informal facilitation. This makes formal institutions a potentially much more powerful mechanism for enhancing cooperation and eventually growth.

Given these definitions, we can now think about what makes agents want to cooperate or facilitate on the trust network. Simply from the structure of the network it is clear that some agents are better suited for cooperation and some for facilitation. Consider agent 1 on Figure 5. He is connected to many other agents, who in their turn are only connected to him. Thus, agent 1 is in perfect position for facilitation. He can make a lot of money by simply connecting his friends to each other, who might not be able to cooperate without him due to low trust. At the same time, agents with only one connection cannot be facilitators, because they have only one friend, so they are best suited for cooperation. Thus, it is reasonable to expect that for agents with many friends facilitation will be more profitable than cooperation and vice versa. Well-connected agents will choose facilitation as their primary source of income, whereas agents with few friends will choose cooperation.

Another way to think about who should facilitate and who should cooperate is by imagining that new friendships can be forged at a cost. New friendships create opportunities for facilitation and thus profit, but constitute an investment. Thus, the question is Which agents would be willing to invest in finding new friends? To answer this, notice that in the previous sections we always assumed that players 1 and 2 have high propensities to follow norms ϕ_1 and ϕ_2, which made them able to cooperate with each other *in principle*. Even if the agents do not originally trust each other, they can use facilitation to enter a successful cooperative relationship and make profit. However, experimental estimates of propensities to follow norms [e.g., Kimbrough and Vostroknutov, 2016, 2018] suggest that the distribution of ϕ's is bimodal: around 20% of people have very high ϕ's, around 20% very low ϕ's, and the rest is spread somewhere in between, which is more or less the same across countries (see Figure 10 in Appendix B). This evidence suggests that around 20% of the population with high ϕ's are able to cooperate, thus they might prefer that to investing into finding new friends. The other 20% with low ϕ's are not able to cooperate at all, because in PG or any other social dilemma, they would always free ride. Given this, it is reasonable to think that selfish agents like that would invest in collecting new friends with the purpose of making money through facilitation. This argument suggests that the number of friendship links on a network might not be random, but rather modulated by agents' propensity to follow norms: selfish agents in trust networks are motivated to invest into new connections with the purpose of facilitation, whereas norm-following agents might prefer to benefit from cooperation.

6.1 Clientelistic Networks

In this final section, we use the idea about the motivation of selfish and norm-following agents explained above to understand the nature and purpose of clientelistic networks. The literature on this topic [Stokes, 2011, Acemoglu, 2005] has catalogued many instances of this phenomenon that is reflected in the attempts of businessmen and politicians to "collect" clients through various forms of gifts, bribes, lobbying, public political events, etc. with the purpose of increasing their own power (through votes, increased political influence, rent, etc.). The literature has described the typologies and even network nature of clientelism [Muñoz, 2019]. A new understanding of clientelism proposes that it arises organically to fill in the vacuum created by the weak state [Acemoglu, 2005, Lazar, 2008]. More specifically, Falconi and Robinson [2021] suggest that clientelism and state weakness in Latin America inhibit provision of basic public goods. Raiser [1997] highlights the role of informal institutions in fostering trust in third-party enforcement through the state during transition from a planned to a market economy in Eastern Europe; he describes how, if trust is eroded by persistent fiscal crises, the control over institutional change can be taken over by competing informal agents (”roving bandits”). But the literature stops short of an explanation as to why this is happening, why this phenomenon takes so many different forms, and how to counteract it. The model of trust network and facilitation provide some new ideas in that regard.

One example of clientelistic network is reported in Falconi and Robinson [2021], where the story of Jhon "Calzones," a Columbian businessman turned politician, is analyzed. Calzones used his connections to provide public services to (illegal) plots of land, $10{,}000$ of which he then sold to the poorest people in the area with a subsidy for constructing a house, before he ran for mayor. Falconi and Robinson note that through his "philanthropic work [Calzones] formed a clientele, a political base, by using his wealth [and connections] to provide services that the Colombian government does not provide, and this allowed him to fill a niche intermediating between the people and the state." To this day, the people of Yopal, or the 'Ciudadela la Bendición' exist as a society independent of the state.

Dramatically, this is not a stand-alone case in Colombia (or several other countries in Latin America). In fact, these arrangements are viewed as a legitimate political process. When state underprovides public goods, it often legalizes the local elites having emerged in similar way, and distributes resources and jobs through their networks. Other examples of structures parallel to the state exist and are sometimes described by the concept of the ”lettered city” [Lazar, 2008]. In this set-up, clientelism is no longer dismissed as undemocratic or inefficient. On the contrary, it provides a solid alternative to the weak formal political and economic institutions, and can even include competition between patrons for clients, who decide which network to support [Lazar, 2008, Falconi and Robinson, 2021].

The connection to the political system and parallel provision of public goods are important for understanding the nature of clientelistic networks and their impact on institutions. Fergusson et al. [2022] show that tax evasion, as a measure of state weakness, and vote buying, as a measure of clientelism, are highly correlated at individual level. Other studies have drawn parallels between the strength and degree of fragmentation of the political system and the level of clientelism [e.g., Levitsky and Zavaleta, 2016]. In another set of studies, Muñoz [2019] analyzes the political system in Peru, where candidates employ clientelistic strategies in the absence of political machines or other means of bargain enforcement. She stresses that approaches focused on monitoring, reciprocity, and conditional loyalty cannot account for the widespread electoral clientelism. The absence of political machinery, Munõz claims, gives rise to clientelism as a means of determining candidates viability and election. Similar ideas have been expressed in the context of other countries with better organized political systems, which nevertheless are plagued with clientelism.

One puzzling feature of Peruvian clientelism is that many "facilitators" distribute gifts to populations in various ways seemingly without any means to guarantee that they get something in return (in form of votes or political influence), which puzzled many a political economist [Muñoz, 2019]. Our network model provides an interesting account of this phenomenon. For clients living in remote areas or working in the informal sector, joining a patron network might be the only means to cooperate or "do business," as they do not have access to formal networks. Figure 6 illustrates.

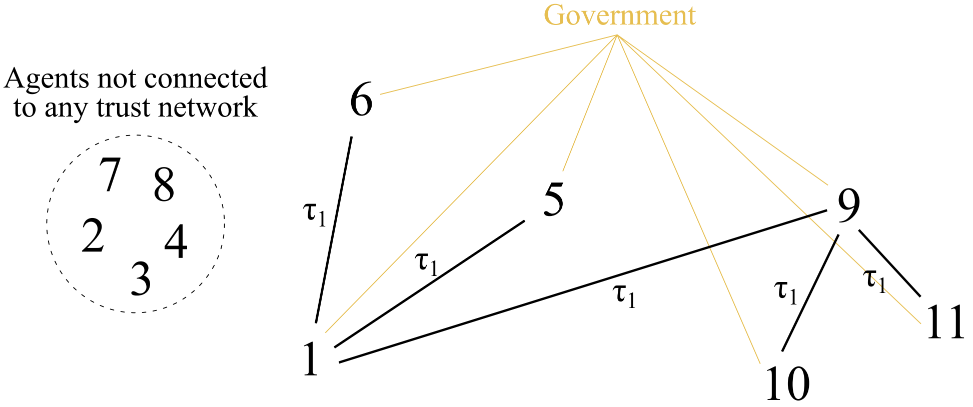

Figure 6: An example of a network that has agents who are not connected to anything and do not have means to cooperate with others.

In such circumstances, selfish agents who are not constrained by the political system try to collect rent by enticing the unconnected individuals with promises of becoming connected to a trust network. At the same time, such individuals desperately want to be connected to any trust network, since they realize that such connections can potentially bring them profitable opportunities for cooperation through facilitation. Thus, both sides of this "market" want something from the other. On the one side we have selfish agents who compete for the connections with poor people. On the other side of the market, we have poor people who want to be connected to some trust networks. As a result, we obtain a situation similar to *perfect competition* among selfish facilitators, who are ready to do anything to obtain new connections. This competition decreases the price for facilitation services and also makes facilitators seek any new connections even if they cannot guarantee any tangible returns from the clients.

Clientelistic networks express themselves differently in countries where political system is more established than in Peru (e.g., Brazil or the US). To continue the analogy with market organization, imagine that one clientelistic network manages to create a *monopoly* in the form of an elite that facilitates cooperation of the whole country. Such a network will create the conditions where people who need to join the trust network have to pay high rents to enter, since there is only one network that they can use (the government is weak and does not provide good facilitation services). At the same time, the network will actively fight any attempts by outside facilitators to lure clients on their own, thus keeping the monopoly power and not allowing more "wild" forms of clientelism as in Peru. In this case, the network can spread to all strata of the society and become entrenched in the political system.

Yet other "market structures" for clientelistic networks are possible. The constant switch of governments from left to right in many Latin American countries can be accounted for as the system with two large clientelistic networks that constantly battle for political power and influence. Such *duopoly* can have negative consequences for the economy and growth as the two networks have to reorient government organizations towards themselves after each election.

7 Discussion

In this paper, we show how introducing an additional motivation to the economic agent—namely, the motivation to cooperate with others that works through the innate desire to adhere to social norms—can give rise to a theory of institution formation and institutional change. Moral agents strive to follow norms and have good cooperative relationships with others, which incentivizes them to seek business opportunities in any economic environment. This implies that when the quality of the government is low and agents are not willing to use its services they will create informal institutions that allow them to cooperate outside the official channels. Various forms of informality ranging from elite capture to "wild" clientelistic networks can emerge depending on the political landscape and current economic situation.

It is important to note that the models presented above are just an illustration of this general approach and do not constitute some final word on how specifically institutions should be modeled. For example, we use the Public Goods game to represent a small business arrangement in which various agents may choose to cooperate with each other or not. However, given the specific context of some country or industry that needs to be modeled, other more tailored games might be used. Also, the way agents change their trust weights towards others in an institution for facilitation can be adapted to specific cases. What we try to present in this paper is the general style of modeling institutions that, similarly to game-theoretic Industrial Organization, can be flexible in treating specific details. Given all this, we believe that our approach contributes to the understanding of institutional change and can help develop new institutional policy.

Although, the models of institutions with moral agents can get arbitrarily specific, we think that the value of our framework lies in its ability to model institutional change and to make general predictions about a variety of institutional arrangements. For example, the idea that moral agents choose whether to work in formal or informal sectors, and do so depending on the two institutions' comparative economic and normative characteristics, shows how easily standard economic policy can go awry. Imagine that the government decides to increase taxes expecting more tax revenue. However, in reality the revenues might fall because after the tax increase agents may leave the formal institution and join the informal one. This process will take place regardless of the specifics of the informal institutions. This example shows that taking institutions into account in a structural way within the model can have tangible and important policy implications.

Another important lesson to be learned from our framework is that moral agents, who have two motivations (selfish and moral), can be heterogeneous in how they mix them.[21] Some agents may be very selfish and disregard the norms, whereas others may be extremely norm-abiding. This fact implies that a policy may have differential effect on different agents and can fail because not all of them react to it in the desired way. For example, introducing a lower speed limit on a highway might make norm-abiding agents drive slower, which will decrease their utility. However, it may not decrease the number of accidents, because selfish agents who do not follow rules will keep driving very fast regardless of the change in the speed limit. This shows how a seemingly benign policy can have adverse effects in the world with moral agents. Our framework can be used to take situations like this into account.

The value of the model with endogenous institutions for applications, as well as empirical and policy work lies in the possibility to calibrate it to specific country or circumstances, existing interest groups, etc. For example, the version of the model as it is presented in this paper can be calibrated to a case of specific country where the estimates for trust weights (τ), propensities to follow norms (ϕ); choice context; and actual rents to facilitators can be obtained from surveys. Then, conclusions can be drawn regarding the "fairness" of the institutions, or how extractive they are, as well as on the effectiveness of specific policies. For example, having estimated the parameters pertaining to moral agents, it would be possible to say how effective a policy to increase tax rate would be in increasing revenue: it would be possible to estimate how many agents will leave for the informal sector, thus providing estimates for the revenue reduction through this channel.

On a final note, we would like to point out that our framework can be used not only for studying "current" institutional change in a given country, but also for understanding the broader historical processes of economic development that start with the emergence of informal institutions in the original state of nature, then proceed to more complex institutions like tribes, and finally lead to the formation of larger arrangements like kingdoms or states. Indeed, in the current version of the model we assume, for example, that government simply exists. However, we do not explicitly describe how it came to be. We believe that this is possible. The versatility of our framework that can be essentially seen as an upgrade of game theory with norms should in principle allow to model institutions of any complexity.

[21] Notice that this problem does not arise in the neoclassical world where agents have only one motivation.

8 Conclusion

In this paper, we propose a new framework to model institutions and institutional change based on the theory of norms by Kimbrough and Vostroknutov [2022a]. We show how moral agents, who care about following social norms and thus strive to cooperate with others, can form institutions that facilitate cooperation. Similarly to game-theoretic Industrial Organization, the framework allows to model informal as well as formal institutions as games played by moral agents. It is possible to conceptualize institutions as inclusive or extractive and model institutional change as a consequence of choice of moral agents among available institutions as time unfolds. We demonstrate how the framework can be used to understand the origins of informal institutions. With a series of examples of clientelistic networks, we show that our framework can be useful for understanding how and why such networks form and persist.

While in the paper we show how our framework is able to account for some interesting phenomena related to institutions in developing countries, these models are just mere illustrations of the full power of the framework that can be used to model any interactions among moral agents, thus giving rise to a wide variety of possible institutional settings. This opens new possibilities to model institutional change; devise new types of policies that take into account normative considerations of the agents; and understand informal and formal institutions in greater detail.

We use the example of clientelistic networks to show how the framework can be applied to specific types of institutions. In principle, one direction of future research could be to create more specific and applied models of clientelistic networks or other types of institutions. However, we believe that the power of the framework lies in its ability to produce general conclusions about broader classes of institutions. For that, more theoretical research is needed that could explore general game-theoretic properties of institutions with moral agents.

Another important direction of future research is to make the framework more practical; to test it with real or experimental/survey data; and to see how to calibrate the model's parameters from surveys or specifically designed tasks. In the next step, policy implications can be considered and estimated, and const-benefit analysis of impact could be performed. We believe that with a developed applied methodology, our framework can become an indispensable tool for studying institutions, conducting economic policy, and suggesting paths to economic prosperity.

References

Robert M. Solow. A contribution to the theory of economic growth. *The quarterly journal of economics*, 70(1):65–94, 1956.

Trevor W. Swan. Economic growth and capital accumulation. *Economic record*, 32(2):334–361, 1956.

Philippe Aghion and Peter W. Howitt. *The economics of growth.* MIT press, 2008.

Douglass C. North et al. The new institutional economics and development. *Economic History*, 9309002: 1–8, 1993.

Philip Keefer and Stephen Knack. Why don't poor countries catch up? A cross-national test of an institutional explanation. *Economic inquiry*, 35(3):590–602, 1997.

Francesco Caselli. Accounting for cross-country income differences. In *Handbook of economic growth*, volume 1, pages 679–741. Elsevier, 2005.

Dani Rodrik. Unconditional convergence in manufacturing. *The quarterly journal of economics*, 128(1): 165–204, 2013.

David Rezza Baqaee and Emmanuel Farhi. Productivity and misallocation in general equilibrium. *The Quarterly Journal of Economics*, 135(1):105–163, 2020.

Norman Loayza. Informality: why is it so widespread and how can it be reduced? *World Bank Research and Policy Briefs*, (133110), 2018.

Guillermo Perry, William Maloney, Omar Arias, Pablo Fajnzylber, Andrew Mason, and Jaime Saavedra-Chanduvi. *Informality: Exit and exclusion.* World Bank Publications, 2007.

Susan C. Stokes. Political clientelism. In Robert Goodin, editor, *The Oxford Handbook of Political Science.* Oxford Academic, 2011.

Andrea Floridi, Binyam Afewerk Demena, and Natascha Wagner. Shedding light on the shadows of informality: A meta-analysis of formalization interventions targeted at informal firms. *Labour Economics*, 67:101925, 2020.

Hernando De Soto et al. The other path. 1989.

Douglass C. North. *Institutions, Institutional Change and Economic Performance.* Cambridge University Press, 1990.

Elinor Ostrom. *Governing the Commons: the Evolution of Institutions for Collective Action.* Political economy of institutions and decisions. Cambridge University Press, Cambridge, 1990. ISBN 9780521405997.

Robert E Hall and Charles I Jones. Why do some countries produce so much more output per worker than others? *The quarterly journal of economics*, 114(1):83–116, 1999.

F. Fukuyama. *Trust: The Social Virtues and the Creation of Prosperity.* The Free Press, New York, 1995.

Robert D. Putnam, Robert Leonardi, and Rafaella Y. Nanetti. *Making democracy work: Civic traditions in modern Italy.* Princeton University Press, 1992.

Martin Raiser. *Informal institutions, social capital and economic transition: reflections on a neglected dimension*, volume 25. EBRD London, 1997.

Daron Acemoglu and James A. Robinson. *The narrow corridor: States, societies, and the fate of liberty*. Penguin, 2020.

Victor Nee. Norms and networks in economic and organizational performance. *The American Economic Review*, 88(2):85–89, 1998.

Elinor Ostrom. Collective action and the evolution of social norms. *Journal of economic perspectives*, 14 (3):137–158, 2000.

Philip Keefer and Stephen Knack. Social capital, social norms and the new institutional economics. In *Handbook of new institutional economics*, pages 701–725. Springer, 2008.

Joseph Henrich. *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press, 2015a.

Ronald Harry Coase. The nature of the firm. *Economica*, 4(16):386–405, 1937.

Oliver E. Williamson. The new institutional economics: Taking stock, looking ahead. *Journal of economic literature*, 38(3):595–613, 2000.

Eirik G. Furubotn and Rudolf Richter. *Institutions and economic theory: The contribution of the new institutional economics*. University of Michigan Press, 2010.

Xavier Sala-i Martin. 15 years of new growth economics: What have we learnt? 2002.

Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2006.

Ernst Fehr and Ivo Schurtenberger. Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7):458–468, 7 2018. doi: 10.1038/s41562-018-0385-5. URL `https://rdcu.be/2Jjo`.

Kevin N. Laland. *Darwin's unfinished symphony: how culture made the human mind*. Princeton University Press, 2018.

Judd B. Kessler and Stephen Leider. Norms and contracting. *Management Science*, 58(1):62–77, 2012.

Oliver Hart. An economist's perspective on the theory of the firm. *Colum. L. Rev.*, 89:1757, 1989.

Bengt Holmström and John Roberts. The boundaries of the firm revisited. *Journal of Economic perspectives*, 12(4):73–94, 1998.

E. O. Kimbrough and Alexander Vostroknutov. A theory of injunctive norms. SSRN Working Paper, Chapman University and Maastricht University, 2022a.

Joseph Henrich. Culture and social behavior. *Current opinion in behavioral sciences*, 3:84–89, 2015b.

Phillip Keefer and Carlos Scartascini, editors. *Trust, social cohesion, and growth in Latin America and the Caribbean*. IDB Publications, 2022.

Douglass C. North. Institutions and the performance of economies over time. In Claude Ménard and Mary M. Shirley, editors, *Handbook of new institutional economics*, pages 21–30. Springer, 2008.

Daron Acemoglu and James A. Robinson. *Economic origins of dictatorship and democracy.* Cambridge University Press, 2006.

Rama Angel and Chasteen John Charles. *The Lettered City.* Duke University Press, Durham, NC, 1996.

Sian Lazar. *El Alto, rebel city: Self and citizenship in Andean Bolivia.* Duke University Press, 2008.

Mary M. Shirley. Institutions and development. In Claude Ménard and Mary M. Shirley, editors, *Handbook of new institutional economics*, pages 611–638. Springer, 2005.

S. Knack and P. Keefer. Institutions and economic performance: Cross-country tests using alternative-institutional measures. *Economics and Politics*, 7(3):207–227, 1995.

Dani Rodrik, Arvind Subramanian, and Francesco Trebbi. Institutions rule: the primacy of institutions over integration and geography in economic development. IMF Working Paper No. 02/189, 2002.

Erik O. Kimbrough and Alexander Vostroknutov. The social and ecological determinants of common pool resource sustainability. *Journal of Environmental Economics and Management*, 72:38–53, 2015.

Stanley L. Engerman and Kenneth L. Sokoloff. Institutional and non-institutional explanations of economic differences. In Claude Ménard and Mary M. Shirley, editors, *Handbook of new institutional economics*, pages 639–665. Springer, 2005.

Paula Muñoz. *Buying audiences.* Cambridge University Press, 2019.

Kenneth J. Arrow. Gifts and exchanges. *Philosophy & Public Affairs*, pages 343–362, 1972.

E. O. Kimbrough and Alexander Vostroknutov. A theory of moral reasoning. SSRN Working Paper, Chapman University and Maastricht University, 2022b.

Nitzan Merguei, Martin Strobel, and Alexander Vostroknutov. Moral opportunism as a consequence of decision making under uncertainty. *Journal of Economic Behavior and Organization*, 197:624–642, 2022.

Folco Panizza, Alexander Vostroknutov, and Giorgio Coricelli. The role of meta-context in moral decisions. mimeo, Maastricht University, University of Trento, and University of Southern California, 2021.

Alexander Vostroknutov. Social norms in experimental economics: Towards a unified theory of normative decision making. *Analyse & Kritik*, 42(1):3–39, 2020.

John A. List. On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3): 482–493, 2007.

Nicholas Bardsley. Dictator game giving: altruism or artefact? *Experimental Economics*, 11:122–133, 2008.

Fabio Galeotti, Maria Montero, and Anders Poulsen. Efficiency versus equality in bargaining. *Journal of European Economic Association*, forthcoming, 2018.

Yan Chen and Sherry Xin Li. Group identity and social preferences. *American Economic Review*, 99(1): 431–57, 2009.

Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868, August 1999.

Dirk Engelmann and Martin Strobel. Inequality aversion, efficiency, and maximin preferences in simple distribution. *American Economic Review*, 94(4):857–869, 2004.

Douglass C. North. Institutions and economic theory. *The american economist*, 61(1):72–76, 2016.

Erin L. Krupka and Roberto A. Weber. Identifying social norms using coordination games: why does dictator game sharing vary? *Journal of European Economic Association*, 11(3):495–524, 2013.

Paul C. Bauer and Markus Freitag. Measuring trust. In Uslaner, editor, *The Oxford handbook of social and political trust*, volume 15. Oxford University Press, Oxford, 2018.

George A. Akerlof and Rachel E. Kranton. Economics and identity. *Quarterly Journal of Economics*, 115 (3):715–753, 2000.

T. Hobbes. *Leviathan.* Menston, Scolar P., 1651.

Peter T. Leeson. Better off stateless: Somalia before and after government collapse. *Journal of comparative economics*, 35(4):689–710, 2007.

James A. Robinson and Daron Acemoglu. *Why nations fail: The origins of power, prosperity and poverty.* Profile London, 2012.

Albert O. Hirschman. *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*, volume 25. Harvard university press, 1970.

Philip Keefer, Carlos Scartascini, and Razvan Vlaicu. Trust, populism, and the quality of government. *The Oxford Handbook of the Quality of Government*, page 249, 2021.

José Luis Falconi and James A. Robinson. The political economy of latin america: New visions. working paper, 2021.

Erik O. Kimbrough and Alexander Vostroknutov. Affective decision-making and moral sentiments. mimeo, Chapman University and Maastricht University, 2022c.

Erik O. Kimbrough and Alexander Vostroknutov. Norms make preferences social. *Journal of European Economic Association*, 14(3):608–638, 2016.

Erik O. Kimbrough and Alexander Vostroknutov. A portable method of eliciting respect for social norms. *Economics Letters*, 168:147–150, 2018.

Daron Acemoglu. Politics and economics in weak and strong states. *Journal of monetary Economics*, 52 (7):1199–1226, 2005.

Leopoldo Fergusson, Carlos A. Molina, and James A. Robinson. The weak state trap. *Economica*, 89(354): 293–331, 2022.

Steven Levitsky and Mauricio Zavaleta. Why no party-building in Peru? *Challenges of party-building in Latin America*, pages 412–439, 2016.

Robinson, Vostroknutov, Vostroknutova
"Endogenous Institutions and Economic Policy"

Appendix

A Additional Analyses

A.1 Best Responses in Public Goods Game

In this appendix we analyze the Nash Equilibria (NE) of the Public Goods game considered in Section 3.1. We assume that the propensities to follow norms ϕ_1 and ϕ_2 are high enough so that the consumption part of the norm-dependent utility (2) is negligible. Thus, we analyze the Public Goods game where the payoffs are equal to the norm function $\eta_i(w_i, w_j; \tau) = -D_i(w_i) - \tau D_j(w_j)$ for player i (we will call j the player who is not i, or $j = -i$).

Our main goal is to prove that there is a threshold value τ^* of trust to the other player such that the Public Goods game with utilities defined by $\eta_i(w_i, w_{-i}; \tau)$ has either 1) unique NE where both players contribute fully (for all $\tau > \tau^*$); 2) unique NE where both players contribute nothing (for all $\tau < \tau^*$); and 3) all pairs of choices of equal contributions are NE (when $\tau = \tau^*$).

We prove this by construction. First, notice that from the proof of the Midpoint Theorem [Proposition 9 in Kimbrough and Vostroknutov, 2022b] we know that on convex polytopes of allocations—to which class the set of allocations C in the Public Goods game belongs—we can write

$$D_i(w_i) = \sum_{k \in \Omega_{w_i}} a^i_k (b^i_k - w_i)^2 + c^i_k,$$

where Ω_{w_i} is the set of vertices of the polytope that have player i's consumption utility higher than w_i and $a^i_k, b^i_k, c^i_k \in \mathbb{R}$ are some coefficients (they cover all possible quadratic equations). Notice as well that $D_i(w_i)$ is a piece-wise parabola with fixed coefficients for three ranges of w_i. When w_i is high there is only one vertex with higher wealth than that (one set of coefficients). When w_i is lower, there are two vertices with wealth of player i higher than w_i, so the coefficients change. For even lower w_i there are three vertices with higher wealth, so the coefficients change yet again. What is important for us though is the fact that for two players i and j the personal dissatisfaction functions are the same, or that $D_i(w) = D_j(w)$. This comes from the symmetry of the game. Also, since D_i are quadratic, so is η_i, which is a collection of piece-wise-stitched concave quadratic forms.

Now, we want to focus on the allocations that are obtained when both players i and j choose the same contributions $x = x_i = x_j$ that result in some symmetric allocation (w, w). In this case, using Lemma 1 in Appendix A.3, we have

$$D_i(w) = D_j(w) = a(b - w)^2 + c,$$

where $a, b, c \in \mathbb{R}$ are some coefficients common for both players. These coefficients are also common for all allocations (w, w) because all such allocations have two vertices with wealths larger or same as w (for either player).

Now that this fact is established, we can use it to show that there is a specific value of the trust weight $\tau = \tau^*$ such that the best response of both players with this τ^* is to choose the same contribution as the other player, *no matter what that contribution is.* We show that such τ^* indeed exists and that it also satisfies the first order condition: the derivative at (w, w) along the choices of one player having the other player's action fixed is zero (this guarantees that it is a best response).

Let us write down the derivative. Notice that by definition $\eta_i(w_i, w_j; \tau) = -D_i(w_i) - \tau D_j(w_j)$, which can be rewritten in terms of contribution choices as

$$\eta_i(x_i, x_j) = -D_i(w - x_i + p(x_i + x_j)) - \tau D_j(w - x_j + p(x_i + x_j)).$$

Suppose that x_j is considered fixed and we look at the maximization problem of player i, who chooses x_i to maximize the above (best response). Using the fact that $D_i(w) = D_j(w) = a(b-w)^2 + c$, we can write the first order condition (the derivative of the above with respect to x_i equal to zero) as

$$-2a(1-p)(b-(w-x_i+p(x_i+x_j))) + 2ap\tau(b-(w-x_j+p(x_i+x_j))) = 0.$$

Notice that when $x_i = x_j$, or when the players choose equal allocations we are interested in, the big parentheses cancel out and we get the condition

$$-(1-p) + p\tau = 0 \tag{11}$$

or

$$\tau = \frac{1-p}{p} = \tau^*.$$

This means that the first order condition above is satisfied for all $x_i = x_j$ only when $\tau = \tau^* = (1-p)/p$. The left panel on Figure 7 illustrates. Here we show the representation of the set of allocations in the Public Goods game (the set of points within the polytope $ABCD$) together with allocations that give both players equal wealths (all in magenta lines). Suppose that x_j is fixed and that player i chooses x_i. Then her choices fall along the lines parallel to AB and DC depending on the value of x_j (AB when $x_j = w$ and DC when $x_j = 0$). The result above means that when $\tau = \tau^*$ the derivatives of η_i are zero for all points (w, w) on the diagonal. This is shown graphically by the dashed black lines (along the edges AB and DC) and little grey dashed lines in between.

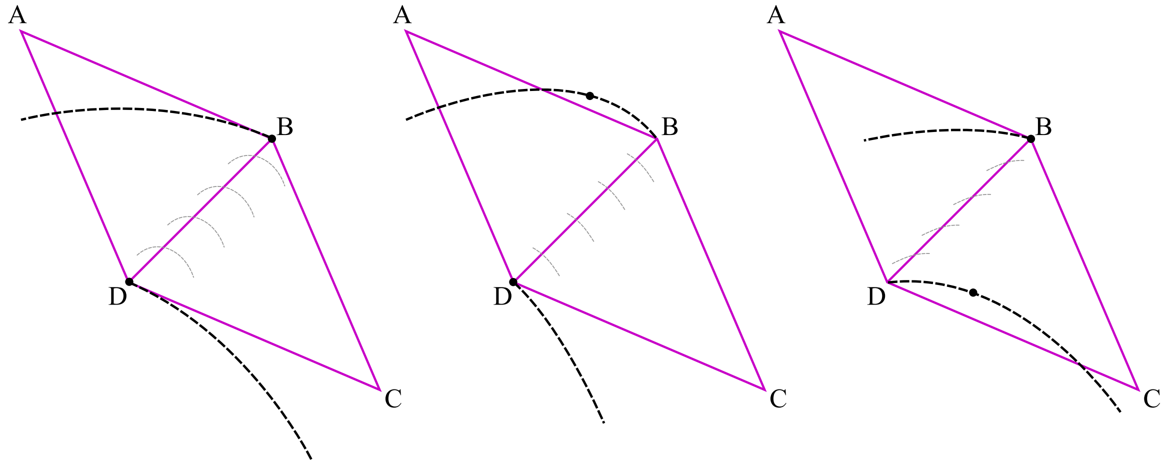

Figure 7: Illustration of best responses in the Public Goods game.

This is also true for the other player. Thus, all this together implies that when $\tau = \tau^*$ the derivatives on the diagonal are zero and are best responses of both players (symmetry). Given that these are mutual best responses, we have a continuum of NE on the diagonal when $\tau = \tau^*$.

This argument demonstrates that there is $\tau^* = (1-p)/p$ such that when the two players with trust τ^* to each other play the Public Goods game, they have a continuum of NE on the diagonal (when they choose the same contributions). This is also a unique τ^* because only it can satisfy the first order condition (11) on the diagonal. The next thing we need to show is that there is a unique NE for all other values of τ, below and above τ^*.

To do that, notice that the condition (11), the left-hand side of which defines the derivative at all diagonal allocations (with equal wealths), tells us that these derivatives are always the same for all such allocations: they are all either positive, negative, or zero depending on the value of τ. For example, the middle panel of Figure 7 shows the case when the derivatives are negative. This means the following. Given that η_i is concave, the negative derivative at point D on the figure implies that it is the maximum on the edge DC (the range of choices of player i given fixed $x_j = 0$) and thus the best response (marked by a black circle). To the contrary, negative derivative at point B means that B cannot be the best response on the edge AB, because the function then grows in the direction of point A. So, the best response is somewhere on the edge AB, but not at B. This last observation also holds in the same way for all points on the diagonal in between B and D. Thus, except for the point D, all best responses of player i lie to the left of the diagonal. Similarly, we can work out that for positive derivatives (the right panel of Figure 7), we have best response at point B and not in D (all best responses in this case are on the right side of the diagonal).

We can use these findings when looking for mutual best responses for the two players. Indeed, when $\tau < \tau^*$ we have the case of negative derivatives. This means that the best responses of the two players are symmetrically situated on the opposite sides of the diagonal except for the point D, where the best responses coincide. They do not coincide anywhere else, since they are divided by the diagonal in all other places. Thus, we can conclude that when $\tau < \tau^*$, we have a unique NE of the game with zero contributions (point D, $x_i = x_j = 0$).

Similarly, when $\tau > \tau^*$ the best responses of the two players coincide at point B, but not anywhere else, since for all other actions the best responses again lie on the opposite sides of the diagonal. Thus, for $\tau > \tau^*$ the unique NE is to contribute fully (point B, $x_i = x_j = w$).

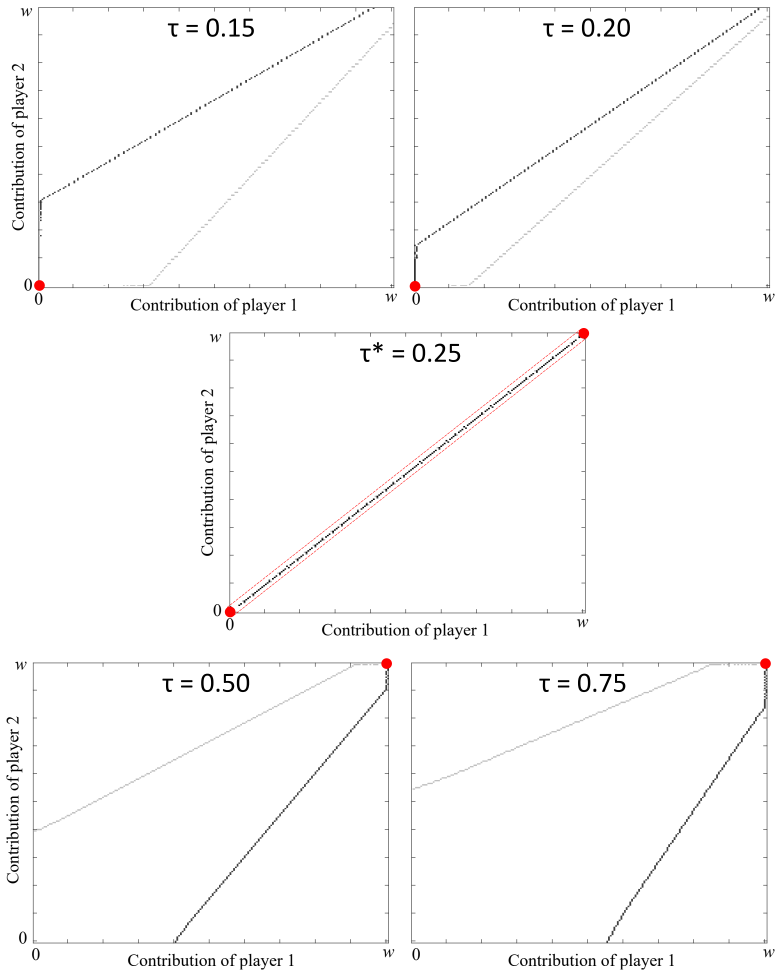

Figure 8: Best response correspondences in Public Goods game with $p = 0.8$ and different levels of τ.

Figure 8 shows numerically computed best response correspondences and Nash equilibria in the Public Goods with $p = 0.8$ for different levels of trust τ of the players to each other. The black lines correspond to the best response of player 1 and grey lines to those of player 2. Equilibria are marked with red circles and dashed lines denote the continuum of equilibria. One can see that for low τ, the unique NE is to contribute nothing as is demonstrated on the top two graphs of Figure 8. For high τ, the NE is to contribute full amounts (the bottom graphs). Finally, when $\tau = \tau^* = 0.25$ we have an intermediate case where any choice of equal contributions constitutes a NE (the middle graph). This same structure of best responses and NE is present for any value of $p \in [0.5, 1)$ when $\tau^* = (1 - p)/p$.

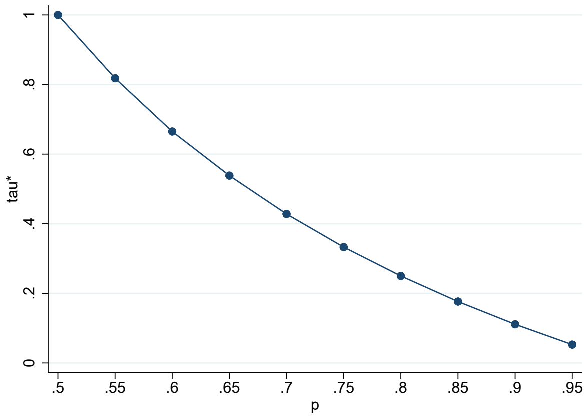

Figure 9: The relationship between productivity p and the cooperation threshold τ^*.

To illustrate, we plot the values of τ^* as dependent on p in Figure 9. Notice that the values on the graph were obtained from the numerical computations of best responses and the resulting values of τ^* are in perfect alignment with the function $(1-p)/p$, which validates our computations.

A.2 Best Responses in Public Goods Game with Facilitation

In this appendix, we show that the threshold property of NE holds also in the Public Goods with facilitation (three players). The logic of this is exactly the same as in the previous section where we considered a simple Public Goods game with two players. To see this notice first that after the facilitator chooses the payment z, the two players find themselves in a subgame, which is essentially a Public Goods game with endowments $w - z/2$ instead of w (z is a given constant). However, we cannot treat this subgame as a simple instance of the Public Goods because the norm function is computed differently using all the allocations in the set C_F (the 3D pyramid on Figure 3).

Despite the complication with the norm function, the logic of the proof of the threshold property stays the same. Indeed, by the results in the proof of Midpoint Theorem in Kimbrough and Vostroknutov [2022b], we know that for any game with any number of players as long as the allocations are represented by a convex polytope, we have personal dissatisfactions D_i represented by piece-wise connected quadratic functions as in Appendix A.1. Moreover, by symmetry these functions are the same for players i and j on the diagonal allocations because again they are defined only by the vertices with higher consumption utility and on the diagonal such vertices are always the same for both players (they may differ for different points (w, w), but are the same for a given point (w, w)). This observation—together with the fact that the dissatisfaction of the facilitator in any subgame is constant and can be ignored—allows us to do the same reasoning as in Appendix A.1 and conclude that there is a unique threshold $\tau^* = (1-p)/p$ such that for any $\tau > \tau^*$ the unique NE in all subgames is to contribute fully and for $\tau < \tau^*$ the unique NE in all subgames is to contribute nothing. This result suggests the optimal behavior of the facilitator as described in the main text.

A.3 Lemmata

Lemma 1. *Any non-constant function* $f(x) = \sum_k a_k(b_k - x)^2 + c_k$ *with some coefficients* $a_k, b_k, c_k \in \mathbb{R}$ *can be represented as* $f(x) = a(b - x)^2 + c$ *where* $a, b, c \in \mathbb{R}$ *are also some coefficients.*

Proof. When we open up the squared terms in f, we get

$$f(x) = (a_1 + ... + a_k)x^2 - 2(a_1 b_1 + ... + a_k b_k)x + d,$$

where d is some constant. Then

$$f(x) = (a_1 + ... + a_k)\left[x^2 - 2\frac{a_1 b_1 + ... + a_k b_k}{a_1 + ... + a_k}x + \left(\frac{a_1 b_1 + ... + a_k b_k}{a_1 + ... + a_k}\right)^2\right] + e,$$

where e is some constant. This can be rewritten as

$$f(x) = a\left(\frac{a_1 b_1 + ... + a_k b_k}{a_1 + ... + a_k} - x\right)^2 + c,$$

where $a = a_1 + ... + a_k$ and c is some constant. From the above we have $b = \frac{a_1 b_1 + ... + a_k b_k}{a_1 + ... + a_k}$. □

B Additional Graphs

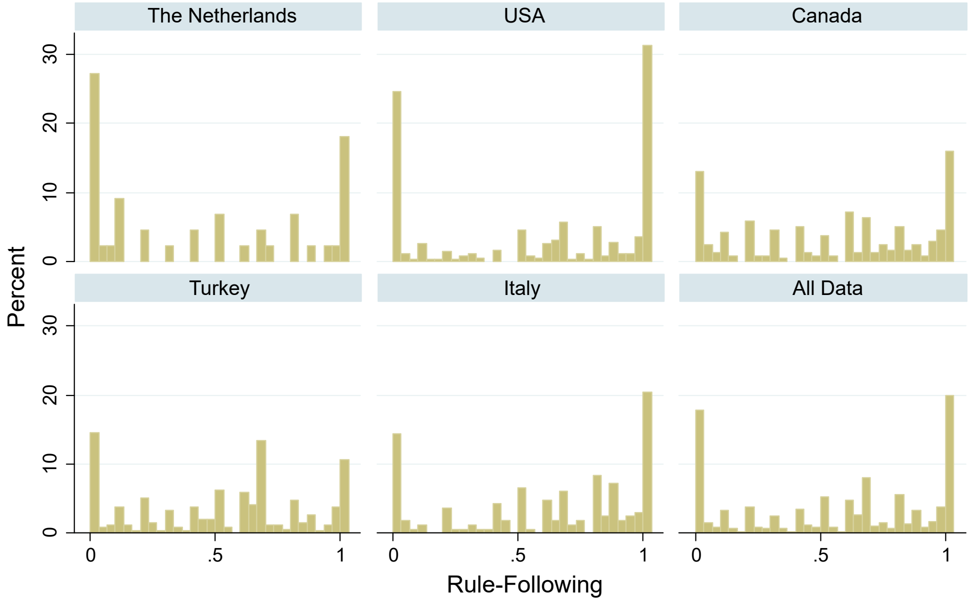

Figure 10: The graph from Kimbrough and Vostroknutov [2018] showing the distributions of ϕ's in five countries and overall. All distributions (except Turkey) are bimodal with a large measure of experimental subjects complying fully with an artificial but costly rule (rule-following = 1), and a large measure of subjects not complying with the rule at all (rule-following = 0).