DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 50 (63) 2017

DOI: 10.1515/slgr-2017-0021

Katarzyna Abramczuk
University of Warsaw

Jan Obłój
Oxford University

THIRD PARTY SANCTIONS IN GAMES WITH COMMUNICATION*

Abstract. This paper discusses the relation between communication and preservation of social norms guarded by third-party sanctions. In 2001 Jonathan Bendor and Piotr Swistak derived deductively the existence of such norms from a simple boundedly rational choice model. Their analysis was based on a perfect public information case. We take into account communication and analyse at the micro level the process of production and interpretation of information on which decisions are based. We show that when information is fully private and we allow for communication a state of anomie can result. If some social control mechanisms are available, social stability can be maintained. The less efficient the social control mechanisms however, the more restrictive rules will be needed to sustain the social norms. Furthermore not all cognitive strategies for interpreting received messages are equally effective. Strategies based on reputation are better than strategies based on profit analysis.

Keywords: game theory, third-party sanctions, social norms, social control, private information, communication, lies detection.

1. Introduction

In recent years we have observed a growing interest in formal modelling of the sociological and psychological aspects of human behaviour. Since the traditional *homo oeconomicus* approach fails to explain a vast majority of both experimental and field observations a considerable effort has been made to account for such phenomena as altruism, social norms, or costly sanctions. All of them are easiest to analyse under an assumption of perfect

* This paper is based on our MA dissertations (Abramczuk, 2003; Obłój, 2004). We want to thank Grzegorz Lissowski for his enthusiastic support and supervision back then and his continued encouragement to publish our results. We are also grateful to Piotr Świstak for his valuable comments and advice during his stay at the University of Warsaw in 2003.

public information. This often allows the obtaining of important first results but at the same time is greatly simplifying and eliminates an important layer of social interactions. In what follows we turn our attention to the consequences of introducing the component of communication into a setting where third party sanctions are being employed.

In reality, the information available to us is crucial for our decision making. Most of the information we use is acquired in our interactions and communication, direct or indirect, with other people. Information may be biased, incomplete or deliberately misleading. In interpreting the information, we take this into account and we try to correct for potential errors. We ourselves also provide information for other social actors to signal our reliability, affluence, status etc. Our knowing communication is also often strategic, the signal being optimised to make us look as favourable as possible. The result is a complex system in which actions, declarations, interpretations and statement verification overlap and interconnect, and which constitutes the environment for our interactions. Intuitively the consequences of this construct for efficiency of third party sanctions should be significant. Third party sanctions are inflicted upon norm breakers by actors who were not directly affected by deviation and do not have immediate data on what exactly happened. As the threat of sanctions for deviating is common knowledge the actors involved often try to obscure the reality even more. Hence the sanctioning parties face a problem of reconciling conflicting versions of events.

In this paper we are trying to gain an insight into the workings of the process described above. We analyse a simple evolutionary game theoretic model to show the consequences of imperfect or strategic information communication and interpretation for stability of socially enforced norms. Our basic model is the one proposed by Jonathan Bendor and Piotr Swistak (Bendor & Swistak, 2001) who showed that simple social categorizations into friends and foes can serve as stabilizing devices for many different rules of behaviour in a wide range of situations. We analyse the consequences of introducing an explicit information flow to their model. We consider a case in which there are no control mechanisms that would allow for establishing and communicating the actual course of interactions taking place and detecting the norm breaches. Instead we assume that actors produce (not necessarily truthful) information that thereafter has to be interpreted and reconciled by others. In this scenario the actors have to have at their disposal strategies for coping with information inconsistencies. We show how in this condition a state of anomie may result. Next we move on to a scheme in which a limited random control is possible. We show that in this case stability of norms may be sustained but the weaker the social control mechanism

is, the more restrictive rules are needed. We also analyse which strategies for interpreting information coming from other actors are most effective in supporting the stability. We consider strategies based on profit analysis and on reputation.

The remainder of the paper is organized as follows. First we provide a short overview of the existing literature. Next we describe the basic assumptions of our model in which information production and interpretation take place. In the consecutive sections we reiterate the most relevant findings of Bendor and Swistak pertaining to a simple model with perfect information. We move on to outlining the consequences of introducing an explicit communication component in a full model. First we present a model without additional control mechanisms. Second we enhance it by adding such mechanisms. The paper closes with conclusions and a brief discussion of possible extensions of the model. The whole presentation is kept simple and with as few formulae as possible. Definitions and proofs are postponed to the Appendix.

2. Related work

Communication might seem to be a secondary issue when the primary problem of rationality of cooperation and costly sanctioning is still discussed. There are a number of theories that try to solve these mysteries including such prominent examples as the theory of kin-selection (Hamilton, 1964), direct (Trivers, 1971; Axelrod & Hamilton, 1981) and indirect (Alexander, 1987; Nowak & Sigmund, 1998) reciprocity, strong reciprocity (Gintis, 2000; Fehr et al., 2002), conformist transmission (Henrich & Boyd, 2001), and multi-level selection (Sober & Wilson, 1998). Generally rather than being a part of the game the social actors play with each other, information in these approaches is an externally shaped part of the environment they are operating in. Some of the proposed solutions require it to be rather extensive and accurate (Leimar & Hammerstein, 2001) while others are suitable when its coverage is limited (Nowak & Sigmund, 1998) or virtually non-existent (Riolo et al., 2001). In the simplest of models it is typically assumed that information is always correct and easily accessible. More complex approaches often include the possibility of production or transmission errors. Yet, these are still just random disturbances while we view information production as potentially strategic.

An analysis of the rich literature on the influence of information source, form, and coherence, on decision making in other contexts (e.g. (Hertwig

& Hoffrage, 2013; Kahneman, 2003)), shows that the origin and strategically shaped information distribution might be crucial for the types of strategies that win the social game. Dealing with random disturbances and untrustworthy information generated on purpose by other social actors can require different cognitive abilities and favour different strategies. For example in the case of randomly revealed lying, one may assume that the truth will transpire at some later point in time. The actor-shaped misrepresentations on the other hand will probably remain consistently untruthful throughout the interaction. To analyse this problem it is necessary to model the creation and interpretation of information on the micro level. This should allow for observation of how individual decisions regarding communication shape the decision environment for all the social actors in a vein portrayed famously by James Coleman in his diagram (Coleman, 1990, p. 8).

Notably there is one strain of literature that looks at a similar problem in a slightly different context. We refer to papers on discounted repeated games with private monitoring. Their authors explore the possibility of cooperation in long-lasting relations where each agent receives a private imperfect signal about the opponents' actions (Ben-Porath & Kahneman, 1996; Kandori & Matsushima, 1998; Compte, 1998; Kandori, 2002; Sugaya & Wolitzky, in press). The model refers to such economic phenomena as collusion under secret price-cutting or exchange of goods with uncertain quality. In analogous models with (perfect and imperfect) public monitoring there are a number of papers deriving Folk Theorem (Abreu et al., 1990; Fudenberg & Maskin, 1986)[1]. The solutions applied there do not work in the private information scenario primarily because the lack of publicly observed history deprives the game of a recursive structure (Abreu et al., 1990) used to construct these equilibria. Several solutions to this problem have been put forward. The most relevant for us is the introduction of communication (Ben-Porath & Kahneman, 1996; Compte, 1998; Kandori & Matsushima, 1998; Awaya, 2014). Usually the actors are assumed to observe some signal regarding others' choices and they are free to report this signal to the whole community. Alternatively (Ben-Porath & Kahneman, 1996) the interactions are monitored by some subset of other actors that can publicly reveal their course. In any case the messages can be either truthful or not. Hence the strategies, similarly as in our model, apart from the action component contain also a mechanism for generating announcements. Furthermore the construed equilibria involve some proposals for handling the generated information. These include majority rule (McLean et al., 2014) and checking for the compatibility of reports of different interaction monitors (Ben-Porath & Kahneman, 1996). In our case similar solutions will

be used to formulate a third component of strategy, namely an interpretation mechanism.

Furthermore, it is worth commenting on a paper by Kandori (Kandori, 1992) which precedes the above-mentioned literature. This work, similarly as ours, concentrated specifically on community enforcement of social norms under a lack of public information. Two solutions for equilibria were discussed: contagious defection and local information processing. The latter refers to attaching publicly observable tags to the actors. A similar construct has been proposed by Okuno-Fujiwara and Postlewaite (Okuno-Fujiwara & Postlewaite, 1995). It is analogous to assigning labels of friends and foes in the Bendor and Swistak (Bendor & Swistak, 2001) model described in Section 4 that is a starting point for our analysis. In Bendor & Swistak (2001), however, unlike in Kandori's paper, the labels are processed separately by various actors and in principle every population member can categorize others in a different way. Consequently, the information requirements are altered in comparison to Kandori (1992). What was phrased as local information processing in our model becomes a special case of the perfect public information scenario. This formulation gives rise to a whole set of new problems. It allows us to go one step further and analyse the reconciliation mechanisms for the actors' strategies.

Is it important to stress that, unlike many of the papers mentioned above, we are not seeking to characterize a set of game equilibria. Instead we follow Bendor and Swistak (2001) and are analysing an evolutionary game to find strategies that are uniformly stable. The details of the model are given below.

3. The model

Our model is based on the model proposed by Bendor and Swistak (2001). It is an evolutionary game with uniformly stable strategies that utilize simple social categorizations. Actors divide others into friends and foes depending on their adherence to a prevailing rule of behaviour. Those who follow the rule are considered friends and treated well. Those who are deviants are considered foes and punished. A friend of a foe and a foe of a friend are also deemed foes. This simple mechanism assures the existence of uniformly stable strategies in a broad class of games. We enhance this model by adding rules for information dissemination and interpretation. This allows us to consider the process of information flow on a micro level tied directly to decisions made by individual actors. It is the individual

actors who decide what to communicate and what to ultimately believe. We will now provide some basic definitions and clarify the workings of this model.

First, only games of enforceable cooperation are being considered. A **game of enforceable cooperation** is a symmetric two-person game with a feasible punishment action. In a one-shot game each actor can choose an action from a set $\mathcal{A} = \{A_1, \ldots, A_\kappa\}$. An actor using action A_k against action A_l ($1 \leq k, l \leq \kappa$) earns payoff $\mathcal{V}(k, l)$. Without loss of generality we assume that the actions are numbered so that $\mathcal{V}(1,1) \geq \mathcal{V}(2,2) \geq \ldots \geq \mathcal{V}(\kappa, \kappa)$. When both actors choose action A_1 we say they are (maximally) cooperating[2]. For this cooperation to be enforceable there has to exist a punishment action A_p such that: $\max_{1 \leq k \leq \kappa} \mathcal{V}(k, p) < \mathcal{V}(1,1)$. The consequence of this assumption is that when somebody does not cooperate they can be effectively disciplined using action A_p and their payoff will be smaller than it would have been had they cooperated. One example of a game of enforceable cooperation is the famous Prisoner's Dilemma, whose exemplary payoff matrix is given in Table 1 below. In this case cooperation is the action A_1, and defection is the punishing action. The best reply to punishment is also defection which gives to the punished a payoff that by definition is smaller than the payoff for mutual cooperation.

The game is repeated. A standard simplifying assumption is that the payoff matrix of the stage-game (a one-shot game) stays the same across all iterations. Each interaction is continued independently of all the other interactions with probability δ ($0 < \delta < 1$). Throughout, when referring to *future being sufficiently important* we mean δ being sufficiently close to one. This is an important criterion necessary for the validity of some of our statements. It ensures the repeated character of the game is predominantly important. The condition $\delta < 1$ ensures that the game ends a.s. after a finite number of rounds. Yet, after each iteration it is generally expected to continue (Aaumann, 1959). The total **payoff** of an actor i from interaction with an actor j equals an expected normalized sum of his/her payoffs throughout the whole such interaction. This is mathematically equivalent to computing limits in the following form: $\lim_{\delta \to 1}(1-\delta)\sum_{t=1}^{\infty} \delta^t v_{ij}^t$, where v_{ij}^t stands for the payoff gained by actor i in interaction with actor j in round t.

The game is played within an **unstructured group of n actors**, where everybody interacts with everybody else[3]. It is the context of this group in which strategies of the actors can be defined. Traditionally strategy in a repeated interaction is thought of as a complete plan of the game given the heretofore course of the interaction. The most famous example of a strat-

egy in this meaning is Tit For Tat (Axelrod, 1984) meant to solve the problem of cooperation in the repeated Prisoner's Dilemma. TFT is very austere because the only piece of information it uses is the information about the most recent action of the partner. It simply says that an actor entering a new interaction should cooperate, and next they should repeat the last action of their partner. In general, however, strategies of this type can use information about any of the previous doings of the partner as long as these are executed within the given pair. In other words an actor can punish or reward his/her partner for actions taken towards him/her but cannot interfere with what the partner chooses to do when interacting with other actors in the group. Strategies with this property will be called **dyadic**.

Reliance on dyadic strategies only, though useful when modelling certain types of interactions, is intuitively inadequate when considering groups of people living in a community. In the latter case it is sociologically evident that they will be interested in how their friends, neighbours and acquaintances behave towards other group members. Hence, Bendor and Swistak (Bendor & Swistak, 2001) revoke the constraint imposed on the domain of a strategy. We follow them in assuming that actors' **behavioural rules** can be based on the whole history of all the interactions in the group and the actors can punish and reward their partners for what they do both to them and to the others. Strategies with this property will be called **social**. It is important to note that using social strategies does not necessarily imply that actors store away all the events that take place for future reference. As a matter of fact it suffices if they use the information to run and update a classification of all the members of the group. The precise nature of this mechanism will be described in the next section.

Since we are mostly interested in the process of communication we introduce two important modifications. They both pertain to what is known about the history of all the interactions in the group. Since the information an actor uses transcends his/her own interactions, we need to consider (1) how it comes about (2) what the recipients do with it. In other words we model explicitly the process of dissemination and interpretation of social information[4].

First we assume that actors can have different beliefs about what happened. A **belief** of an actor concerning the history is necessarily true in the case of interactions in which the actor took part him/herself. In the remaining cases the beliefs have to be formed on the basis of either observation or the declarations of other actors. The observation is assumed to be a form of public social control mechanism that allows the whole community to learn

about the true course of an interaction. Its prevalence varies in different models we describe from a perfect information condition to a complete lack of observation. In the latter case the actors produce all the information themselves and they can lie.

To model the strategic creation and dissemination of information by the actors, we assume that actors have **information rules** that describe what actors will report about interactions they directly participated in. The information rule defines what the actor will be claiming about his interactions both with regard to his/her own action and the action of his/her partners for each possible belief about the history so far and each possible actual course of interactions of a given actor. Please note that what an actor decides to communicate can depend on the actual course of the interaction reported, as well as on the course of his/her other interactions and the beliefs of the actor concerning the history of all the interactions in the group so far. For example actor i who did not cooperate towards actor j may try to hide this fact when he/she thinks that j has always cooperated so far and is highly regarded by the group, but may also admit the lack of cooperation when he/she thinks that j is a notorious norm-breaker.

Obviously when information is provided by the actors themselves it may lead to conflicting versions of the same events. Therefore some strategies for reconciling the conflicting versions are needed. These strategies will be called **interpretation rules**. They are decisive for which version of the past events an actor adapts as his/her belief regarding the course of a given interaction. An interpretation rule defines a belief regarding what happened in the most recent round for each possible belief about history so far, each possible set of reports on the interactions in this round, and each possible course of one's own interactions. For example an actor i who is faced with conflicting reports about interaction between j and k may decide to believe j because he/she is convinced that j has always cooperated so far while k committed several transgressions. In another case, discussed later in more detail, i may believe j because j is involved in a smaller number of conflicting reports i.e. other reports on j's action generally agree with what j claims while for k there are many other actors who deny k's reports.

Important additional comments on three related simplifying assumptions we make are required here. First, we assume that the communication is public i.e. an actor cannot vary his/her messages depending on whom they are directed to. This is an important postulate especially when one considers the possibility of untruthful reports. It is justified provided we are analysing communities small enough to track all the events in the group.

In the case of such communities varying one's claims depending on their receiver would likely result in a quick discovery of the inconsistencies produced and discredit their author. Second, while we do not assume that all the actors necessarily communicate directly, we do assume that information is passed around without any distortion. In other words, actor i does not have to ask j him/herself about what happened to j. Actor i might learn what j reported from somebody else. However, this information is necessarily consistent with the actual claim of j. This assumption does not allow e.g. for modelling the process of rumour formation. Third, it is important to note that the communication, as well as the actions, in a given round are simultaneous. When the round starts all actors simultaneously decide how to behave towards all their partners and next all of them produce reports about their interactions in the present round at the same time. Hence an actor cannot use information about what his/her partner claimed about their interaction to adjust his/her own report.

Having defined all the necessary components we are ready to define a strategy in our model. Summarizing, a **strategy** is a triple consisting of an information rule, an interpretation rule, and a behavioural rule[5]. All of them are based on beliefs regarding the whole history of interactions in the group that are formed by the actors as a result of interpretations of information produced via the information rules. Furthermore all strategies are assumed to be **non-discriminatory**; i.e. they do not condition their choices (actions, messages or interpretations) on the identity of the other actors.

Our approach to evolution is consistent with (Bendor & Swistak, 2001). An evolutionary process is, similarly to communication, defined at the micro level. We assume that the actors can change their strategies in a way that leads to increase their utility. No specific form of the utility function and no specific evolutionary dynamics are proposed to model these changes. It is only assumed that **utility functions** of all the actors are normal (Bendor & Swistak, 2000) i.e they are increasing functions of both (and only) the payoffs generated by the strategies and their relative frequencies. In other words actors value more those strategies that lead to better outcomes and are more common in the group. Notably different actors may have different utility functions. We want to establish which strategies will survive under these conditions. This reduces to searching for strategies that are weakly uniformly stable. A strategy is said to be weakly **uniformly stable** if and only if, once sufficiently common in the population, it does not decrease in frequency regardless of the particular form of the (monotonous) evolutionary dynamics. This happens only if the strategy is unbeatable

i.e. when used by most of the actors, it assures them the highest payoff in the population[6].

At this point we can start talking about a population of strategies rather than actors. This population can be thought of as a vector of strategies and their frequencies in the population. Since payoffs achieved by actors in our model do not depend on their identities we will be referring to a total payoff gained by an actor using strategy S as a payoff of this strategy and we will denote it by $\mathcal{V}_S$. The freqency of S will be denoted as p_S. Using this notation we can define uniform stability in a more precise way. We say that S is uniformly stable iff there is a $p_S^* < 1$ called the **minimal stabilizing frequency** of S such that $\mathcal{V}_S \geq \mathcal{V}_T$ for any strategy T.

A final comment on other possible strategies is necessary. Similarly as in most game-theoretic models we assume that **deviations/mutations** are possible. We do not investigate the process that might generate them. We simply accept the fact that any new strategy might appear in the population of strategies in some (originally) small frequency. Importantly we consider it unrealistic to assume only homogeneous mutations. Therefore we allow for several different new strategies to enter the population at the same time.

4. Perfect information

In this section we explain the basics of the original Bendor and Swistak model (Bendor & Swistak, 2001). It is a model with perfect observation. Information about the whole history of interactions is always available and credible. Hence the information and interpretation rules are superfluous and will not be considered.

Throughout this and the following sections we will be using an example of Prisoner's Dilemma whose payoff matrix is presented in Table 1. It is important to remember however, that our conclusions apply to all games of enforceable cooperation.

Table 1

An exemplary payoff matrix of a game of enforceable cooperation – Prisoner's Dilemma

	C	D
C	3 3	0 5
D	5 0	1 1

C stands for Cooperation and D stands for Defection

The first important observation is given by Theorem 1 (Bendor & Swistak, 2001, p. 1512)[7]:

Theorem 1

In a symmetric nontrivial iterated two-person game, any dyadic pure strategy is not uniformly stable if the future is sufficiently important.

The Theorem holds because for every conceivable dyadic strategy it is possible to design a pair of mutants one of which would be a dydically neutral mutant of the original strategy (i.e. would behave the same as the native strategy in its interaction with it) while the other would boost the payoff of the first mutant and lower the payoff of the original strategy. To see how this can happen let us take an example of the game in Table 1. Say the population is dominated by TFT and is invaded by TF2T and STFT[8]. TF2T (Tit For 2 Tats) is a strategy that cooperates unless the partner defected in the two most recent interactions. STFT (Suspicious Tit For Tat) plays the same way as TFT but it starts a new interaction with a defection. The patterns of interactions for each possible pair of strategies in this example are given in Table 2.

Table 2

Patterns of interactions for every possible pair of strategies in an exemplary population consisting of TFT, TF2T and STFT

round	TFT, TFT	TFT, TF2T	TFT, STFT	TF2T, TF2T	TF2T, STFT	STFT, STFT
1	C,C	C,C	C,D	C,C	C,D	D,D
2	C,C	C,C	D,C	C,C	C,C	D,D
3	C,C	C,C	C,D	C,C	C,C	D,D
4	C,C	C,C	D,C	C,C	C,C	D,D
5	C,C	C,C	C,D	C,C	C,C	D,D
6	C,C	C,C	D,C	C,C	C,C	D,D
...	...	...	...	...	...	...

In this population TFT and TF2T attain perfect cooperation with themselves and each other. However, things get more involved when we analyse their interactions with STFT. TF2T fares better in this case because it forgives the first defection of STFT and cooperates with it from the second iteration on. TFT on the other hand gets locked in an infinite sequence of exploiting and being exploited. Since the interactions with STFT are the only ones that differentiate between TFT's and TF2T's payoffs they are also decisive for which strategy wins. More precisely the expected payoffs for the two equal:

$$\mathcal{V}_{TFT} = (p_{TFT} + p_{TF2T}) \times 3 + p_{STFT} \times 2,5$$
$$\mathcal{V}_{TF2T} = (p_{TFT} + p_{TF2T}) \times 3 + p_{STFT} \times 3$$

Clearly it is TF2T that gets the highest payoff in this case and outperforms TFT. TFT is not uniformly stable.

Bendor and Swistak show that to overcome this problem one needs a meta-construction that would allow for punishing any norm departure including the deviations occurring outside of the given dyad. This enables punishing not only defections, but also e.g. lack of punishing of the defectors. More precisely it was proven that (Bendor & Swistak, 2001, p. 1517)[9]:

Theorem 2

Uniformly stable pure strategies exist in a non-trivial symmetric two-person game with sufficiently important future iff the game is a game of enforceable cooperation.

An exemplary uniformly stable strategy is CNF (Conformity). It consists of the following rules[10]:

1. Categorize all actors as either friends or foes.
2. In the beginning consider everybody friends.
3. Cooperate with all friends and punish all the foes.
4. After each interaction add to the set of foes all actors to whom at least one of the following applies:
   (a) They did not cooperate with some friend
   (b) They cooperated with some foe

This strategy is an application of a fairly intuitive logic entailing three simple rules: (1) the friend of my friend is my friend (2) the foe of my friend is my foe (3) the friend of my foe is my foe. These rules make CNF a social strategy. CNF not only pays attention to what happens in its own interactions, but interferes with interactions of others. Hence it can be thought of as a social norm. It has two basic features of all norms: it says what actors should be doing and it is guarded by social sanctions (Homans, 1950).

No mutant can ever fare better than CNF in a population dominated by it. To see this, note that there are only two possible scenarios. First, the mutant can be completely neutral i.e. actors using it will always make exactly the same choices as actors using CNF and they will be treated in the same way earning the exact same payoff. Second, it can at some point make a different choice than CNF i.e. cooperate with some foe or defect

towards some friend. The latter implies however that actors using it will be categorized as foes from this moment on and in most interactions they will receive at most $\max_{1\leq k\leq \kappa}\mathcal{V}(k,p)$, where A_p is the punishment action. This, by definition of a game of enforceable cooperation, is less than the payoff for cooperation achieved by CNF[11].

We can track this process using an example of CNF dominated population playing an iterated Prisoner's Dilemma from Table 1. Let us assume that this population is invaded by TF2T and STFT. The patterns of interactions for each possible pair of strategies in this example are given in Table 3.

Table 3

Patterns of interactions for every possible pair of strategies in an exemplary population consisting of CNF, TF2T and STFT

round	CNF, CNF	CNF, TF2T	CNF, STFT	TF2T, TF2T	TF2T, STFT	STFT, STFT
1	C,C	C,C	C,D	C,C	C,D	D,D
2	C,C	C,C	D,C	C,C	C,C	D,D
3	C,C	D,C	D,D	C,C	C,C	D,D
4	C,C	D,C	D,D	C,C	C,C	D,D
5	C,C	D,D	D,D	C,C	C,C	D,D
6	C,C	D,D	D,D	C,C	C,C	D,D
...	...	...	...	...	...	...

In the beginning it seems everything will play out according to the previous scenario. CNF and TF2T cooperate while STFT defects. The defection causes CNF to classify STFT as a foe, while TF2T is more forgiving and continues to cooperate with STFT. STFT seeing this situation unfold cooperates towards TF2T but defects towards CNF. Afterwards however CNF, unlike TFT, does not consider TF2T worthy of further cooperation. It observes TF2T's failure to punish STFT for the first detection (cooperation with a foe in the second iteration) and considers actors using TF2T foes from the iteration 3 on. Ultimately in iteration 5 it causes CNF and TF2T to become mutual foes. Using the exemplary payoff matrix we can write:

$$\mathcal{V}_{CNF} = p_{CNF} \times 3 + (p_{TF2T} + p_{STFT}) \times 1$$

$$\mathcal{V}_{TF2T} = p_{CNF} \times 1 + (p_{TF2T} + p_{STFT}) \times 3$$

Since the population consists mostly of CNF actors, it is the first component of these expressions that is decisive. Obviously CNF can sustain its domination.

5. Private information

Now we will move on to discussing what happens in the model described above when we introduce communication and leave no space for independent social control of the truthfulness of actors' declarations. To discuss this problem we need to introduce strategies that aside from the behavioural rules have also information and interpretation rules. We will start by examining an example akin to those discussed in the previous section. The native strategy dominating the population will be the truthful CNF (TCNF). TCNF always gives truthful reports. We will also assume that it uses the following simple interpretation rule:

1. When reports agree, believe them true.
2. When reports differ, for each actor involved in the conflict count in how many other information conflicts he/she is involved in and believe the report of the actor with fewer conflicts.
3. When the number of conflicts for both actors is the same, assume they both broke the rules and deem them foes.

The population will be invaded by the two familiar mutants: TF2T and STFT. More precisely we define TF2TP – TF2T pretending CNF, and GSTFT – gloomy STFT. We can think of actors playing TF2TP as of people who have decided it is ineffective to resign from cooperating with GSTFT. They do realize however that the operating norm will expect them to punish the other mutant. Hence, even though they cooperate with everybody, they decide to hide this fact. Their information mechanism requires them to always claim about all their interactions what they know, or believe, to be the social expectation. In the words of Robert Sudgen (Sudgen, 1985) they are applying a modified version of a famous maxim: "When in Rome, APPEAR as the Romans do". To simplify keeping up the pretence of following CNF, TF2TP uses the same interpretation rule, to update the history as CNF does. However it is not a social strategy and it bases its behavioural decisions only on the history of its own interactions. It does of course always know what their real course was.

GSTFT is a fully dyadic strategy that does not use information about history other than its own, neither to decide which action to choose nor to determine the content of its reports. Due to this feature it does not need any interpretation rule. It is a suspicious strategy that starts with defection. It is also a gloomy type that irrespective of the actual development of an interaction claims that defection is all that anybody ever did.

The patterns of interactions for each possible pair of strategies alongside with all the reports are given in Table 4.

Table 4
Patterns of interactions and reports for every possible pair of strategies in an exemplary population consisting of CNFT, TF2TP and GSTFT

round	TCNF, TCNF	TCNF, TF2TP	TCNF, GSTFT	TF2TP, TF2TP	TF2TP, GSTFT	GSTFT, GSTFT
1	C,C	C,C	C,D	C,C	C,D	D,D
reports 1	*(C,C) (C,C)*	*(C,C) (C,C)*	*(C,D) (D,D)*	*(C,C) (C,C)*	*(C,D) (D,D)*	*(D,D) (D,D)*
2	C,C	C,C	D,C	C,C	C,C	D,D
reports 2	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,C) (D,D)*	*(C,C) (C,C)*	*(D,C) (D,D)*	*(D,D) (D,D)*
3	C,C	C,C	D,D	C,C	C,C	D,D
reports 3	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(D,D) (D,D)*
4	C,C	C,C	D,D	C,C	C,C	D,D
reports 4	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(D,D) (D,D)*
5	C,C	C,C	D,D	C,C	C,C	D,D
reports 5	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(D,D) (D,D)*
6	C,C	C,C	D,D	C,C	C,C	D,D
reports 6	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(D,D) (D,D)*
...	...	...	...	...	...	...

Recall that in the model with perfect information the two mutants did not pose any danger to CNF. Here however the problem of lack of stability returns. After the first iteration it becomes clear that GSTFT is different. TCNF immediately classifies actors playing it as foes. TF2TP realizes that doing this is required by the operating norm but wants to go on cooperating with GSTFT. Hence it reports the factual events from the first round but decides to refrain from retaliation. Subsequently it cooperates with everybody while claiming that it punishes GSTFT. TCNF needs to resolve a reporting conflict pertaining to interactions between TF2TP and GSTFT. It notes that GSTFT is in reporting conflicts with all TCNF actors that form the majority of the population, while TF2TP is only conflicted with GSTFT that are rare mutants. Hence it decides to trust TF2TP. As a result it becomes blind to TF2TP's treacherousness. The latter can get away with cooperating with everyone while maintaining the status of a friend. Using the exemplary payoff matrix from Table 1 we can write:

$$\mathcal{V}_{TCNF} = (p_{TCNF} + p_{TF2TP}) \times 3 + p_{GSTFT} \times 1$$
$$\mathcal{V}_{TF2TP} = (p_{TCNF} + p_{TF2TP}) \times 3 + p_{GSTFT} \times 3$$

Clearly TF2TP gets a higher payoff, which suffices to prove that TCNF is not uniformly stable. One will of course wonder whether there is some other strategy in this condition, perhaps with a better interpretation rule that is still stable. Our first main result asserts that this is not the case: in the absence of some element of control which allows one to observe, even if on rare occasions, the actual behaviours of actors, no strategy can be stable.

Theorem 3

In a symmetric nontrivial iterated two-person game with no independent control of the information disseminated by the actors, no pure strategy can be uniformly stable.

A detailed proof of this theorem is presented in the Appendix. Its outline coincides with the story of TCNF. Basically for each pure strategy in this context it is possible to construct two mutants. The primary mutant is an "almost" neutral mutant (an analogue for TF2TP) which gains in interactions with enemies more than the native strategy does but hides this fact by always reporting what is expected by the dominating rule. The additional mutant is a strategy that cooperates with the first mutant while lowering the payoff of the native strategy. The additional mutant also has to have an information rule that makes it impossible for the native strategy to differentiate between the first mutant and its own kin.

The situation only seemingly will become more complicated if we impose a restriction on the type of lies, by requiring them to be only strategic. The reports of GSTFT in the example above do not meet this requirement because there is no rational calculation backing up its gloomy nature. However, we can easily make it rational by modifying TF2TP so that it stops cooperating with agents that ever defected towards it, if they report anything else than mutual defection about any interaction. Admittedly, this modification might seem a little strained and it points to some important observations related to the model with lies.

First, lying is a complex and tricky task. It requires knowledge of the partner that often exceeds what can be thought of as a reasonable set of assumptions. While it is naturally expected that most of the actors know the contents of the most common norm, it is more disputable to presume they know that the other norm breakers will smoothly join a conspiracy. If the other mutants will not do so, the whole plot might collapse. For example the behaviour of TF2TP can only be profitable if GSTFT makes it possible. If we replaced GSTFT with its truthful version TSTFT that never lies, TCNF would realize quickly that there is something wrong with TF2TP's reports. The situation after the first two rounds is presented in Table 5. Because both TCNF and TSTFT are honest their reports are always the same. There is however a conflict between what TF2TP and TSTFT claim about their interactions. Recall that in such situations TCNF's interpretation rule requires it to compare the number of reporting conflicts in which a given actor is involved to decide whom

to trust. In this case the turn of events will depend on the frequencies of the two mutants. If the proportion of TF2TP is higher than the proportion of STFT it might still escape with its lie. In any other case it will be punished.

Table 5
Patterns of interactions and reports in the first two rounds for every possible pair of strategies in an exemplary population consisting of CNFT, TF2TP and TSTFT

round	TCNF, TCNF	TCNF, TF2TP	TCNF, TSTFT	TF2TP, TF2TP	TF2TP, TSTFT	TSTFT, TSTFT
1	C,C	C,C	C,D	C,C	C,D	D,D
reports 1	*(C,C) (C,C)*	*(C,C) (C,C)*	*(C,D) (C,D)*	*(C,C) (C,C)*	*(C,D) (C,D)*	*(D,D) (D,D)*
2	C,C	C,C	D,C	C,C	C,C	D,D
reports 2	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,C) (D,C)*	*(C,C) (C,C)*	*(D,C) (C,C)*	*(D,D) (D,D)*

We can make things even harder for TF2TP by revising TCNF's interpretation rule. Note that the difference in TSTFT's reports concerning its interactions with TCNF and TF2TP is a clear indicator that at some point TF2TP must have made some choices that differed from the choices of TCNF. Consequently it must have broken the norm and should be deemed a foe irrespective of the number of conflicts it is involved in. A new interpretation rule would therefore be one of a complete uniformity unwilling to accept any (even seemingly harmless) difference. It can be described as follows:

1. Compare the received reports with the reports on your own interactions with the given actor.
2. Categorize as a foe anybody:
   (a) whose reports concerning interaction with the given actor differ from your own reports about interactions with the same actor
   (b) for whose interactions the given actor reports something else than the same actor reports for your interactions

In our example we are dealing with the case 2b above – TSTFT claims something else about its interactions with TF2TP than about its interactions with TCNF. Therefore TF2TP is automatically classified as a foe. Unfortunately even with this improved interpretation rule TCNF fails to be uniformly stable. It can still be invaded if TSTFT is replaced with‘ GSTFT. Furthermore, using the uniform interpretation rule can prove difficult when, as possible here, some interactions end while others still last or when we introduce elements of stochasticity. The latter remark points to possible enhancements of the model. Some of them are discussed in the last section of this paper.

Above we described a situation in which one of the mutants betrays the other mutant by producing revealing reports. Yet, the additional mutant may also make conspiracy difficult by varying its behaviour towards the native strategy and the primary norm breaker. In that case the first mutant might lose track of what it should report about its interactions, to convincingly pretend to be following the norm. Imagine for example that GSTFT cooperates every tenth interaction with all the actors that punished it in the second round (i.e. TCNF). In round 12, 22, and so on the truthful CNF will report that it experienced cooperation from GSTFT. Unless TF2TP knows the pattern in good time, it will have no way of learning what it should report in these rounds. It experiences the continuous cooperation of the other mutant and only learns about GSTFT choices in interactions with TCNF after they are announced. The reports however are produced simultaneously. TF2TP has to produce a false report aimed at pretending to be TCNF before it knows what the actual TCNF report for the given round is.

Apart from deception being a difficult task, one might wonder about the nature of the evolutionary process in the model with lies. As noted above, it is generally assumed that actors learn which strategies fare best in the ecology and imitate them or innovate to improve their payoff. A successful imitation or innovation requires the actors to know some details of the successful approach. Yet when the actors lie, these details become obscure. The successful mutant cannot be imitated for exactly the same reasons for which it cannot be punished. What consequences it might have for the evolution remains an open question.

The final note is partially related to the problem above. Let us assume for the time being that an invasion was successful e.g. TF2TP won over the population of TCNF. What we get is an interesting discrepancy between what actors do and what they claim is right to do. The winner is an effectively dyadic strategy that avoids meddling with other actors' business, but it does keep up the pretence of a strong social moralist. It achieves its goal of high payoff in ways that are not legitimate. Such inconsistency is known in social sciences as anomie (Merton, 1968), a state of social and cultural chaos in which individuals have to choose between following the verbalized norms and achieving socially approved goals because the two do not align with each other. TCNF would be an example of a ritualist that sticks to the norm, while TF2TP is an innovator who puts greater emphasis on the goals. The model shows that, without an effective social control mechanism, anomie might easily result. We will now move on to show how social control can stop this process.

6. Partial social control

In this section we analyse a model that lies in between the first two. We assume that in general the information is produced by the actors themselves but there is some *partial social control*. More specifically, this is understood in the following way: we assume that each interaction independently is observed by the community with probability γ ($1 > \gamma > 0$). With probability $1 - \gamma$ the community has to rely solely on the actors' reports. In this case the following theorem can be proved, which is the second main result of our work:

Theorem 4

In a symmetric nontrivial iterated two-person game with partial social control and a stage game of enforceable cooperation there are pure strategies that are uniformly stable, provided the future is sufficiently important.

Proof.

The proof of this theorem is straightforward. We will show that TCNF is uniformly stable in this case. From the model with lies we have already learned that a successful mutant has to be neutral in the sense that it always appears to be making the same choices as TCNF i.e. cooperating with friends and punishing foes. Let us assume that there are two mutants in the population: S^1 and S^2 with corresponding frequencies p_1 and p_2. We will also assume that S^1 gets a higher payoff from its interactions with S^2 than TCNF does. For this to happen S^1 must, at some point, make a different choice than TCNF in its interaction with S^2. This choice will involve breaking the TCNF rules and will become known to the community with probability γ. However large the benefit of S^1 from interactions with S^2, and however small γ, its total expected payoff of S^1 will be lower than the payoff of TCNF. Because the frequency of the latter can be arbitrarily close to one, nothing can make up for the non-zero probability of getting punished by it indefinitely. □

Even though TCNF is uniformly stable in the model with partial social control, it does not represent an optimal solution for this setting. We can improve it further by changing slightly its behavioural rule and make it harder for the mutants to invade the population.

First let us compute a minimal stabilizing frequency of TCNF in its current form. Remember that A_p is a punishment action in the game of enforceable cooperation. Let P denote the maximal payoff of a punished actor i.e. $\max_{1 \leq k \leq \kappa} \mathcal{V}(k, p)$. Denote also payoff to maximal

cooperation as $R = \mathcal{V}(1,1)$, the maximal possible payoff in the game as $T = \max_{1\leq k,i\leq\kappa}\mathcal{V}(k,i)$, and the minimal possible payoff in the game as $S = \min_{1\leq k,i\leq\kappa}\mathcal{V}(k,i)$.

We can conservatively assume that a minimal total normalized payoff of TCNF when δ goes to one equals $p_{TCNF}R+(1-p_{TCNF})S$. Furthermore the maximal total normalized payoff of any mutant depends on whether his/her detour from the operating norm will be detected. In the most favourable case there will be only one such detour that will be observed by the community with probability γ. As a consequence the mutant can never earn more than $p_{TCNF}(1-\gamma)R + p_{TCNF}\gamma P + (1-p_{TCNF})T$. This leads to the conclusion that the frequency that guarantees TCNF uniform stability depends on the probability of observation γ. For TCNF payoff to be larger than the payoff of any mutant we need:

$$p_{TCNF} > \frac{T-S}{T-S+\gamma(R-P)} \tag{1}$$

Obviously when γ goes to zero, i.e. it is very difficult to control others' interactions: the minimal stabilizing frequency of TCNF goes to one.

We can illustrate the source of the problem using once again an exemplary payoff matrix from Table 1. We will construe a population consisting of TCNF, LCNF and GR2. LCNF (Late CNF) is a strategy that behaves like CNF but it assigns actors to enemies with a one round delay. Similarly as TCNF, LCNF pretends to follow the dominating norm. It always claims about its interactions that they were exactly like the interactions of TCNF with the same partners. GR2 (Gloomy Repeat Two) is a dyadic strategy that starts with 2 defections and next repeats the second choice of its partner indefinitely. Similar to GSTFT it is a gloomy type that, irrespective of the actual development of an interaction, claims that defection is all that anybody ever did. Table 6 illustrates the course of all the interactions and reports in this population under an assumption that no interactions were observed.

Please note that LCNF breaks the operating norm only in the second round. If it goes unnoticed it generates only a small loss. It is however also the source of a never-ending benefit because GR2 cooperates with LCNF indefinitely and assures it gets the highest possible payoff. Admittedly TCNF might observe this fact in some later interaction. However in its current form the CNF behavioural rule for ascribing status as friend or foe does not punish agents that are treated by enemies better than the TCNF itself. As a result the expected payoff of LCNF equals $p_{TCNF}(1-\gamma)3 + p_{TCNF}\gamma 1 + (1-p_{TCNF})5$. For it to be smaller than

Table 6
Patterns of interactions and reports for every possible pair of strategies in an exemplary population consisting of TCNF, LCNF and GR2

round	TCNF, TCNF	TCNF, LCNF	TCNF, GR2	LCNF, LCNF	LCNF, GR2	GR2, GR2
1	C,C	C,C	C,D	C,C	C,D	D,D
reports 1	*(C,C) (C,C)*	*(C,C) (C,C)*	*(C,D) (D,D)*	*(C,C) (C,C)*	*(C,D) (D,D)*	*(D,D) (D,D)*
2	C,C	C,C	D,D	C,C	C,D	D,D
reports 2	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(D,D) (D,D)*
3	C,C	C,C	D,D	C,C	D,C	D,D
reports 3	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(D,D) (D,D)*
4	C,C	C,C	D,D	C,C	D,C	D,D
reports 4	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(D,D) (D,D)*
5	C,C	C,C	D,D	C,C	D,C	D,D
reports 5	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(D,D) (D,D)*
6	C,C	C,C	D,D	C,C	D,C	D,D
reports 6	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(C,C) (C,C)*	*(D,D) (D,D)*	*(D,D) (D,D)*
...	...	...	...	...	...	...

the payoff to TCNF we need $p_{TCNF} > 5/(5 + 2\gamma)$. However for fixed γ we can lower the minimal stabilizing frequency of TCNF by adding one additional point to the TCNF categorization rule (marked in bold below):

1. Categorize all actors as either friends or foes.
2. In the beginning consider everybody friends.
3. Cooperate with all friends and punish all the foes.
4. After each interaction add to the set of foes all actors to whom at least one of the following applies:
   (a) They did not cooperate with some friend
   (b) They cooperated with some foe
   (c) **They received higher payoff in some interaction with some foe than the native strategy with the same foe.**

The addition reflects yet another notion of what it means to be the friend of a foe. It was unnecessary in the model with full observation. It becomes vital when information is not perfect – especially if we endeavour to investigate forgiving behavioural rules and various trembles (see the Summary section below). Its major benefit is that any infinite sequence of profiting from breaking the rule at some point will be detected a.s. after a finished number of rounds. Hence the expected payoff of any mutant cannot exceed $p_{TCNF}P + (1 - p_{TCNF})T$. In that case the minimal stabilizing frequency does not depend on the probability of observation presuming that δ goes to one. For the example analysed above it equals 5/7. On a practical

level the problem described above shows that when social control is weak, more restrictive rules are needed to sustain the norms.

There is one more interesting observation concerning the model with partial social control. It refers to how the interpretation rule of a uniformly stable strategy can be constructed. Recall that the interpretation rule of TCNF requires it to rely on the number of information conflicts in which the given actor is involved. It presumes that the actor less conflicted with the whole population is a trustworthy one. This mechanism is based on reputation. One might think that rules based on profit analysis would be more effective. This however is not the case, primarily because it is often unclear what is in the interest of particular actors. As has been shown in the section above, an originally gloomy lie can be made rational if the rule of behaviour for other actors is changed. There is however an even more fundamental reason for which profit analysis is not a good solution to solving reporting conflicts – its use may lead to unjustified punishing of one's kin.

To see why let us take a look at a population consisting of PACNF, TF2TP and CSTFT. PACNF is the CNF strategy that uses Profit Analysis instead of relying on the interpretation rule defined above. CSTFT (Cheerful STFT) is a sister to GSTFT. The only difference is that it always reports mutual cooperation. In the second round TF2TP cooperates with CSTFT, hence breaking the norm. It reports however that it punished its partner. CSTFT reports mutual cooperation. The pattern of interactions and reports under an assumption that nothing is observed is given in Table 7.

Table 7
Patterns of interactions and reports for every possible pair of strategies in an exemplary population consisting of PACNF, TF2TP and CSTFT

round	PACNF, PACNF	PACNF, TF2TP	PACNF, CSTFT	TF2TP, TF2TP	TF2TP, CSTFT	CSTFT, CSTFT
1	C,C	C,C	C,D	C,C	C,D	D,D
reports 1	*(C,C) (C,C)*	*(C,C) (C,C)*	*(C,D) (C,C)*	*(C,C) (C,C)*	*(C,D) (C,C)*	*(C,C) (C,C)*
2	C,C	C,C	D,C	C,C	C,C	D,D
reports 2	*(C,C) (C,C)*	*(C,C) (C,C)*	*(D,C) (C, C)*	*(C,C) (C,C)*	*(D,C) (C,C)*	*(C,C) (C,C)*
3	D, D	D,C	D,D	C,C	C,C	D,D
reports 3	*(D,D) (D,D)*	*(D,C) (D,D)*	*(D,D) (C, C)*	*(D,D) (D,D)*	*(D,D) (C,C)*	*(C,C) (C,C)*
4	D, D	D,C	D,D	C,C	C,C	D,D
reports 4	*(D,D) (D,D)*	*(D,C) (D,D)*	*(D,D) (C, C)*	*(D,D) (D,D)*	*(D,D) (C,C)*	*(C,C) (C,C)*
5	D, D	D, D	D,D	C,C	C,C	D,D
reports 5	*(D,D) (D,D)*	*(D,D) (D,D)*	*(D,D) (C, C)*	*(D,D) (D,D)*	*(D,D) (C,C)*	*(C,C) (C,C)*
6	D, D	D, D	D,D	C,C	C,C	D,D
reports 6	*(D,D) (D,D)*	*(D,D) (D,D)*	*(D,D) (C, C)*	*(D,D) (D,D)*	*(D,D) (C,C)*	*(C,C) (C,C)*
...	...	...	...	...	...	...

If PACNF was to analyze in whose interest it is to lie it would have to refer to the expected consequences of the lie for the total payoff of the actors involved in the information conflict. This analysis should rely on the most commonly applied norm because it is decisive for the actors' profit. Obviously in the case above it is in the interest of TF2TP to claim it punished CSTFT in the second round. This guarantees further cooperation with the majority of the actors, while admitting cooperation would lead to infinite punishment. At the same time from the perspective of CSTFT the course of interaction is irrelevant as it has already been classified as a foe. Hence PACNF believes the latter and punishes T2T2P. This seems to be a good thing. Unfortunately exactly the same reasoning will lead PACNF to think that other PACNF actors are its foes. This is because CSTFT and TF2TP make it impossible to differentiate between the native strategy and the TF2TP mutant. The error of wrongly assigning friends the label of foes is in this case particularly difficult to correct. Even if PACNF was a forgiving strategy there would be little for it to learn from random observation of interactions of other PACNF actors. This is because all of them would suddenly have completely different sets of foes and friends and would not cooperate with each other.

Luckily, as shown before, reliance on reputation assures that all[12] the native norm actors will have the same sets of foes and friends and will be able to use the random control mechanism to detect the foes hidden among the friends. Because the native strategy is also the most common one, the frequency of conflicts is a very good cue indicative of whom to trust. In other words reputation, unlike profit analysis, guarantees uniform stability in the model with partial social control.

7. Summary and Discussion

This paper has discussed the relation between communication and preservation of social norms guarded by third-party sanctions. We used a model proposed by Jonathan Bendor and Piotr Swistak (Bendor & Swistak, 2001) and augmented it with a component describing how information is strategically created and disseminated, and then interpreted, on the micro level. We show that in the private information case with communication, stability can be destroyed by a mutant that keeps up a pretense of following the norm in spite of breaking it in interactions with another mutant strategy. The resulting state is kin to the state of anomie described in classic sociological literature, where individuals have to choose between following verbalized norms and achieving socially approved goals because the two

do not align with each other. Introducing public social control restores the stability of the third-party sanctions. The efficiency of social control can influence the minimal stabilizing frequency of the operating norm unless this norm is very restrictive in terms of whom to categorize as a friend and as a foe. While highly efficient control allows for turning a blind eye on some signs of misconduct, the weak control increases the need for being prohibitive. Finally it has been shown that when deciding whom to believe in case of inconsistent information, one should rely on the social reputation of the sources rather than on an analysis of their interests.

There are several ways in which the approach presented in this paper could be advanced. First, we can introduce stochasticity to the model. It may for instance be a consequence of various types of implementation and communication errors. We may also consider mixed strategies. Second, it is worth examining the possible consequences of reformulating some of the basic assumptions of the model. For example we could investigate what happens if the matching rules change or when δ for some interactions gets smaller and the threat of punishment weakens.

The above mentioned modifications will likely lead to further developments. In the current paper we restrict our attention to unforgiving strategies. This is justified given we wanted to offer a clear introduction to the communication component we put forward. However a natural next step is to compare their properties with the characteristics of more lenient rules of behaviour that will become necessary once stochasticity is introduced. The latter may also call for more fundamental changes to the communication component. Please note that in its current form the model assumes that actors form their beliefs concerning other people's deeds directly after the communication phase and they do not change them afterwards. Consequently they are unable e.g. to track the number of reporting conflicts of a given actor throughout the whole game and apply some threshold rules to the number of those that can be tolerated. To enable these types of strategies we have to equip the agents with a capacity for some sort of double accounting that would allow them to form beliefs that can be revised when more evidence is gathered.

Finally, while we did model the communication process at the micro level, the dissemination of the information and the social control are still framed in terms of public observability. It would be very interesting to see what happens if they are also attached to specific decisions of specific actors. If we allow for information to be spread unreliably or we model observation as a local activity we might for instance be able to understand better the process of rumour formation.

Appendix A. Modelling Setup

Recall that in a one-shot game each actor can choose an action from a set $\mathcal{A} = \{A_1, \ldots, A_\kappa\}$. An actor using action A_k against action A_l ($1 \leq k, l \leq \kappa$) earns payoff $\mathcal{V}(k, l)$. Without loss of generality we assume that the actions are numbered so that $\mathcal{V}(1,1) \geq \mathcal{V}(2,2) \geq \ldots \geq \mathcal{V}(\kappa, \kappa)$. We suppose there are n agents who play a repeated game against each other. The set of agents is denoted by $\mathcal{N}$.

Let us recall that the game is expected to finish in finite time. Let ξ_{ij}^t, $1 \leq i, j \leq n$, $t \in \mathbb{N}$ be a family of independent random variables with the same distribution $\mathbb{P}(\xi_{ij}^t = 1) = \delta = 1 - \mathbb{P}(\xi_{ij}^t = 0)$ in some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\eta_{ij}^t = \xi_{ij}^1 \cdot \ldots \cdot \xi_{ij}^t$. Recall that $v_{ij}^t = \mathcal{V}(a_{ij}^t, a_{ji}^t)$ stands for payoff gained by actor i in interaction with actor j in round t. A random payoff of i from the whole interaction with j is a sum: $\sum_{t=0}^{\infty} \eta_{ij}^t v_{ij}^t$ and the total payoff of i from the whole interaction with j is the expected value of this sum i.e.: $\mathbb{E}(\sum_{t=0}^{\infty} \eta_{ij}^t v_{ij}^t) = \sum_{t=0}^{\infty} \delta^t v_{ij}^t$. δ is a discount factor. Any reference to *sufficiently important future* should be interpreted in terms of $\delta > \delta_0$, where $\delta_0 < 1$ is some threshold value that exists. This is equivalent to using *normalised payoffs* given in the following form: $\lim_{\delta \to 1}(1-\delta)\sum_{i=1}^{\infty} \delta^t v_{ij}^t$, where $(1-\delta)$ is a normalization factor assuring that the limit is finite.

We fix an abstract probability space $(\Omega, \mathcal{G}, \mathbb{P})$ on which we model the game. The most important ingredient in the description of our formal setup are flows of information available to different actors at different times. We denote $\mathcal{F}_{\text{pub}}^t$ the σ-algebra representing public information available to all the actors at time t and $\mathcal{F}_i^t$ the one representing the information available at time t to actor i. Clearly both sequences should be increasing in t and $\mathcal{F}_{\text{pub}}^t \subset \mathcal{F}_i^t$. We consider strategies which do not have any initial informational advantage so that $\mathcal{F}_{\text{pub}}^0 = \mathcal{F}_i^0$, $i \in \mathcal{N}$.

An action chosen by actor i in an interaction with actor j in round t, $t \geq 1$, will be denoted by $a_{ij}^t \in \mathcal{A}$. This action is decided based on the information available after the previous round, i.e. a_{ij}^t is $\mathcal{F}_i^{t-1}$ measurable for all $i, j \in \mathcal{N}$, $i \neq j$ and $t \geq 1$. After their interactions the agents make announcements, truthful or not, about the strategies used in the interactions they were involved in. We let $m_{ij}^t \in \mathcal{A}^2$ denote the information that agent i broadcasts about the value of (a_{ij}^t, a_{ji}^t), $i \neq j$. Note that when choosing what to reveal an agent can also use her knowledge of her own interactions at time t, i.e. m_{ij}^t is measurable with respect to $\mathcal{F}_i^{t-1} \vee \sigma(a_{ji}^t : i \neq j \in \mathcal{N})$.

As the last part of an iteration of the game, each interaction is verified, independently of everything else, with some probability γ. Formally, we

suppose that there is a large collection $\epsilon^t_{ij} = \epsilon^t_{ji}$, $i, j \in \mathcal{N}$, $i \neq j$, $t \geq 1$, of i.i.d. Bernoulli random variables, independent of $\mathcal{F}^0_{\text{pub}}$, with $\mathbb{P}(\epsilon^t_{ij} = 1) = \gamma$. The control process can be represented as

$$c^t_{ij} = \begin{cases} (a^t_{ij}, a^t_{ji}) & \text{if } \epsilon^t_{ij} = 1 \text{ and either } (a^t_{ij}, a^t_{ji}) \neq m^t_{ij} \\ & \text{or } (a^t_{ji}, a^t_{ij}) \neq m^t_{ji}, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

for $i, j \in \mathcal{N}$, $i \neq j$ and $t \geq 1$. In words, c^t_{ij} returns the true course of an interaction if this interaction was controlled and at least one of the announcements was false. Note that we think of controlling interactions so that c^t_{ij} and c^t_{ji} are the same (up to re-ordering of the two actions). Note also that the way the particular control process is set up in (2), the actors only learn that a particular interaction was controlled when one of the announcements is false. We could modify the definition to return 0 if there was no control, 1 if the interaction was controlled but the announcements were truthful and to return the true course of actions if someone was lying. This would allow for slightly more sophisticated strategies but otherwise would not change anything in our setup.

Finally, in order to progress to time $t + 1$ we need to specify how the information is updated. The public knowledge is enlarged with the public announcements and outcomes of the random control:

$$\mathcal{F}^t_{\text{pub}} = \mathcal{F}^{t-1}_{\text{pub}} \vee \sigma(m^t_{ij}, c^t_{ij} : i, j \in \mathcal{N}, i \neq j), \quad t \geq 1.$$

The private information further contains the true course of interactions that the particular agent was involved in:

$$\mathcal{F}^t_i = \mathcal{F}^{t-1}_i \vee \mathcal{F}^t_{\text{pub}} \vee \sigma(a^t_{ij}, a^t_{ji} : j \in \mathcal{N}, j \neq i).$$

It may be of interest to consider smaller information sets to define subclasses of strategic behaviour. For example, we say that an agent is using a *dyadic strategy* if for all $t \geq 1$, $i \neq j$, a^t_{ij} is measurable with respect to

$$\mathcal{F}^{t-1}_{\text{priv},ij} = \sigma(a^u_{ij}, a^u_{ji} : u < t).$$

The announcements are irrelevant for such a strategy so we will also take it to be truthful: $m^t_{ij} = (a^t_{ij}, a^t_{ji})$.

More formally, we think of *a strategy* as a behavioural rule, a prescription which in any such setting generates the actions and announcements,

i.e. generates an $(\mathcal{F}_i^t)$-predictable process $(a_{i\cdot}^t)_{t\geq 1}$ valued in $\mathcal{A}^{n-1}$ and a process $(m_{ij}^t)_{t\geq 1}$ valued in $\mathcal{A}^{2(n-1)}$ adapted to the filtration $\mathcal{F}_i^{t-1} \vee \sigma(a_{ji}^t : i \neq j \in \mathcal{N})$. The σ-algebras being finitely generated, this corresponds to saying that the action component is a mapping which, for a fixed t, takes the historic inputs

$$\{m_{kj}^u, c_{kj}^u, a_{ij}^u . a_{ji}^u : u < t, i \neq j, k \in \mathcal{N}\}$$

and returns a vector of actions $a_{i\cdot}^t$, with analogous description for the announcement component. We can then distinguish strategies based on their use of information, e.g. the dyadic strategies introduced above, but equally on their functional form. For example, it is natural to restrict to non-discriminating strategies which do not distinguish agents based on their labels but only on the history of collective past behaviour. That is, if we apply a permutation to the labels of agents in $\mathcal{N}$, the output vector will contain actions and announcements with the same permutation.

Appendix B. Proof of Theorem 3

Proof.

First, based on Theorem 2, we can assume that the game is a game of enforceable cooperation. Let us further assume that there is some strategy S that is uniformly stable in this game. From Theorem 1 we know that S has to be social. We will introduce two mutants to a population dominated by S: S^1 and S^2. S^1 will be assumed to fulfill the following two conditions[13]:

- It always makes the same choices as S in its interactions with S
- It always produces the same reports about its interactions with other actors as S

S^2 will be assumed to always produce the same reports about its interactions with S^1 as about its interactions with S. These conditions ensure that S and S^1 are indistinguishable from each other.

Let S^2 start its interactions with some action A_i that is different from the action chosen in the beginning of the interaction by both S and S^1. Let us further assume that S and S^1 react to A_i differently and choose different actions in their interactions with S^2 in round 2. Hence, they become distinguishable from the point of view of S^2. From interaction 3 on we will assume that S^2 in its interactions with S^1 always chooses action that enables S^1 to earn the highest possible payoff in the game (T defined on page 126), while S^1 chooses the best reply to it. Furthermore we will assume that in their interactions with S, actors using S^2 always choose A_p.

We can now inspect the expected payoffs of S and S^1. First note that interactions between two actors playing S, two actors playing S^1, and two actors one of whom plays S and the other S^1 necessarily all have the same course. What follows the difference in payoffs of the two strategies can only be a result of what happens in their respective interactions with S^2. The maximum S can earn in its interactions with S^2 equals $\max_{1\leq k\leq\kappa}\mathcal{V}(k,p)$. S^1 on the other hand earns in its interactions with S^2 the maximal payoff in the game. Because the game is a game of enforceable cooperation, the latter is necessarily larger. Hence S cannot be uniformly stable. □

NOTES

[1] Folk Theorem is actually a whole class of theorems referring to feasible Nash equilibria in repeated games. They generally state that any payoff vector in which each actor receives at least their minimax payoff are feasible under some Nash equilibrium.

[2] We concentrate on this case, but it is worth remembering that lower levels of cooperation, where the assured payoff is not maximal, also can be uniformly stable.

[3] For convenience it is also assumed that actors play with themselves.

[4] As a result of this approach and all the additions that we make a formal definition of a strategy becomes more involved. It is postponed to the Appendix.

[5] In this paper we restrict our analysis to pure strategies. In principle however it is possible to include mixed strategies in the model.

[6] Please note that some other strategy may assure equally high payoff to some actors (hence weak stability), however, it can never out beat the uniformly stable strategy.

[7] See (Bendor & Swistak, 1998) for a full proof.

[8] This example is taken from (Boyd & Loberbaum, 1987).

[9] See the reference for a full proof.

[10] Similar "moral" strategies were considered earlier e.g. by Boyd and Richerson (Boyd & Richerson, 1992, p. 182).

[11] CNF, unlike TFT, is an unforgiving strategy. This does not imply that all uniformly stable strategies have to be equally unforgiving. This assumption simply makes the problem easier to explain. For a discussion see (Bendor & Swistak, 2001).

[12] Most, if we consider mutants using mixed strategies.

[13] The formulation has to remain general as S does not have to be a truthful strategy.

REFERENCES

Abramczuk, K. (2003). *Mechanizmy kontroli społecznej z perspektywy teorii gier* (Unpublished master's thesis). Instytut Socjologii, Uniwersytet Warszawski.

Abreu, D., Pearce, D., & Stacchetti, E. (1990). Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica: Journal of the Econometric Society*, *58* (5), 1041–1063.

Alexander, R. D. (1987). *The biology of moral systems*. New York: Aldine de Gruyter.

Aumann, R. J. (1959). Acceptable points in general cooperative n-person games. *Contributions to the Theory of Games (AM-40)*, *4*, 287.

Awaya, Y. (2014). *Private monitoring and communication in repeated prisoners' dilemma* (Tech. Rep.). Working Paper, Penn State University, http://www.personal.psu.edu/yxa120/research.html.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*, 1390–1396.

Axelrod, R. M. (1984). *The evolution of cooperation*. New York: Basic Books.

Bendor, J., & Swistak, P. (1998). Evolutionary equilibria: Characterization theorems and their implications. *Theory and decision*, *45*(2), 99–159.

Bendor, J., & Swistak, P. (2000). The impossibility of pure homo economicus. In *Annual meeting of the american political science association*. Marriott Wardman Park.

Bendor, J., & Swistak, P. (2001). The evolution of norms. *American Journal of Sociology*, *106*(6), 1493–1545.

Ben-Porath, E., & Kahneman, M. (1996). Communication in repeated games with private monitoring. *Journal of Economic Theory*, *70*(2), 281–297.

Boyd, R., & Lorberbaum, J. P. (1987). No pure strategy is evolutionarily stable in the repeated prisoner's dilemma game. *Nature*, *327*, 58–59.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology*, *13*(3), 171–195.

Coleman, J. S. (1990). *Foundations of social theory*. Cambridge Mass.: The Belknap Press of Harvard University Press.

Compte, O. (1998). Communication in repeated games with imperfect private monitoring. *Econometrica*, *66*(3), 597–626.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity. *Human Nature*, *13*, 1–25.

Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica: Journal of the Econometric Society*, *54*(3), 533–554.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*(2), 169–179.

Hamilton, W. D. (1964). Genetical evolution of social behavior i, ii. *Journal of Theoretical Biology*, *7*(1), 1–52.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*(1), 79–89.

Hertwig, R., Hoffrage, U., & the ABC Research Group (Eds.). (2013). *Simple heuristics in a social world*. New York: Oxford University Press.

Homans, G. C. (1950). *The human group*. Harcourt: Brace and Company.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioural economics. *The American economic review*, *93*(5), 1449–1475.

Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies*, *59*(1), 63–80.

Kandori, M. (2002). Introduction to repeated games with private monitoring. *Journal of Economic Theory*, *102*(1), 1–15.

Kandori, M., & Matsushima, H. (1998). Private observation, communication and collusion. *Econometrica*, *66*(3), 627–652.

Leimar, O., & Hammerstein, P. (2001). volution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London: Biological Sciences*, *268*, 745–753.

McLean, R., Obara, I., & Postlewaite, A. (2014). Robustness of public equilibria in repeated games with private monitoring. *Journal of Economic Theory*, *153*, 191–212.

Merton, R. K. (1968). *Social theory and social structure*. New York: The Free Press.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*, 573–577.

Obłój, J. (2004). *Normy społeczne i kontrola zachowań w ujęciu dedukcyjnym* (Unpublished master's thesis). Instytut Socjologii, Uniwersytet Warszawski.

Okuno-Fujiwara, M., & Postlewaite, A. (1995). Social norms and random matching games. *Games and Economic behavior*, *9*(1), 79–109.

Riolo, R. L., Cohen, M. D., & Axelrod, R. (2001). Evolution of cooperation without reciprocity. *Nature*, *414*(22), 441–443.

Sober, E., & Wilson, D. S. (1998). *Unto others – the evolution and psychology of unselfish behavior*. Cambridge: Harvard University Press.

Sudgen, R. (1985). Review of the economics of conformism. *Economic Journal*, *95*, 502–504.

Sugaya, T., & Wolitzky, A. (in press). Bounding equilibrium payoffs in repeated games with private monitoring. *Theoretical Economics.*

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.