The Origins of In-Group Bias and the Cost of Signaling Sociality*

Moti Michaeli†

Abstract

All around us we see that people form groups, that these groups are often indifferent to other groups in the best case, or hostile to other groups in the worst case, and that many cohesive groups push their members to signal their belonging to the group by performing actions that involve some self-sacrifice. In this paper we show that the tendency of people to form groups of limited size and to show in-group favoritism can be traced back to a fundamental characteristic of our mentality – the psychological cost we, as social individuals, pay for not reciprocating the kind actions of others. Moreover, a difficultly in spotting asocial individuals, who are not subject to this cost, may lead to the emergence of costly signaling of sociality. Groups that adopt such practices are characterized by a high level of cooperation among group members, and can coexist alongside groups with no signaling and a lower level of cooperation. When the proportion of asocial individuals is not too high, the welfare of all individuals in society that contains signaling groups is strictly lower than would have been if signaling was impossible. Thus,

*I would like to thank Moshe Shayo and Eyal Winter for their guidance and precious advices. I wish also to thank Stefan Behringer, Elchanan Ben-Porat, Bård Harstad, Sergiu Hart, Andrea Ichino, Chloe Le Coq, Yosef Rinott, Assaf Romm, Tomer Siedner, Daniel Spiro, Elyashiv Wiedman, and participants at the Hebrew University, the University of Oslo, and the 8th Nordic Conference on Behavioral and Experimental Economics in Stockholm, for their valuable comments.

†Department of Economics and the Center for the Study of Rationality, the Hebrew University, motimich@gmail.com.

the cost of signaling is twofold: the individual cost of producing it, and the social cost of ending up in an inferior equilibrium.

Keywords: In-Group Bias, Costly Signaling, Group Formation, Evolution of Cooperation, Prisoner's Dilemma Game.

JEL Classification: D82, Z13, D03, D7, C72.

1 Introduction

In the literature on human group size and on the development of sociality in Homo Sapiens, one prominent hypothesis suggests that the size of a "natural" human group is bounded by our cognitive skills - the need to memorize all the interactions and relationships between all members of the group consumes a lot of memory space and thus limits the group size. This hypothesis is based on research of various animals, which showed a positive correlation between animal group size and the relative size of the neocortex in the animal's brain, a correlation that led the researches to hypothesize that the bound is actually on the number of relationships that an individual animal can successfully monitor (Sawaguchi and Kudo 1990, Dunbar 1992).[1] These results were extended to anatomically modern humans, for whom a maximal group size of 148, commonly known as "Dunbar's number", was predicted (Dunbar 1993).[2]

Although there is some evidence in support of "Dunbar's number", larger groups are also known to exist, even in hunter-gatherer societies (Stewart 1955, Service 1962, and Birdsell 1970). Moreover, this theory can explain the limit on the cooperative group size, but cannot explain cooperation itself. We suggest here an alternative theory. We believe that at least when it comes to modern human beings, the bound on group size is not due to cognitive limitations, but rather due to the nature of human social conscientiousness. In particular, most human beings are endowed with a "psychological cost of cheating", i.e., they have disutility from cheating or betraying another person by not reciprocating

[1] For example, an increase in group size from 40 to 50, entails an increase from 780 to 1225 in the number of pairwise interactions to memorize, suggesting that brain complexity crucially limits group size (Aiello & Dunbar 1992).

[2] According to this theory, when groups significantly exceed this size, they can no longer be egalitarian in their organization but must increasingly develop stratification involving specialized roles relating to social control (Naroll 1956, Forge 1972).

the other person's kind actions.[3] But one should be careful not to automatically assume that this cost rises linearly with the number of betrayed individuals. In fact, though it is quite reasonable that this cost of cheating increases in the number of cheated individuals, the most salient difference is probably between cheating no one and cheating someone. Moreover, the marginal cost is bound to decrease in the number of cheated individuals. Thus, a plausible assumption would be that the "psychological cost of cheating" is concave.[4] Since the gain from cheating increases more or less linearly in the number of cheated individuals, one is inclined to be tempted to cheat if the number of cooperators exceeds a certain threshold.[5] Thus, belonging to a group of limited size ensures that the temptation is resistible, and that others can be trusted to cooperate because their temptation is resistible too.

Note that as opposed to the hypothesis about cognitive limitations as the source of restriction on group size, our hypothesis does not imply that the human brain imposes a hard-wired constraint on group size. Therefore, we do not predict a fixed limit on group size, but rather a flexible bound that is sensitive to the material returns to cheating. In particular, groups of larger size can be sustained if they find reliable ways to reduce the material gains from unilateral defection of a group member.

The assumption that the cost of cheating is increasing in the number of cheated individuals, yet it does so in a concave manner, generates two distinct refutable predictions. In natural situations, where the material benefit from

[3]Note that cheating here is not lying: one's actions are what counts, and not the consistency between one's statements and one's actions. Lopez-Perez (2012) demonstrates that indeed lying aversion is not enough to induce cooperation in PD.

[4]One may think of this concavity as depicting a state where the more social connections one has, the weaker is one's empathy to one's weakest connection, and as a consequence the psychological cost of breaking the weakest connection decreases with the total number of connections. However, we will not assume the existence of groups, so there is no reason to presuppose that some people are inherently closer than others. Moreover, this depiction seems to suggest that groups are formed because people end up cheating only those who are detached enough from them, and that the cheated people will be considered the out-group members. However, we do not assume such discrimination exists, and therefore we show a different mechanism that leads to group formation, where in fact there is no cheating at all in equilibrium.

[5]Unless one assumes that the utility from monetary gains is even more concave than the cost of cheating.

unilateral defection is (more or less linearly) *increasing* in group size (i.e., more "suckers" to exploit if one defects from cooperation), we expect the tendency to cheat on the group to increase with group size. However, if for some reason the material benefit from unilateral defection is *constant* in group size, the individual should be less prone to cheat when the group is larger, because cheating more people would inflict a higher cost on him, with no increase in benefit. These two distinct predictions were neatly demonstrated in experiments of the public good game conducted by Isaac et al (1994) and surveyed in Ledyard (1995) and Holt and Laury (2008). In these experiments, subjects divide their allocation of tokens between a private account and a group account. In order to create an incentive to free-ride, the experimenters set the *marginal per capita return* from the group account (MPCR, defined as the ratio of benefits to costs for moving a single token from the individual to the group account) to be in the range of $(0, 1)$. Moreover, the design of the experiments was such that the monetary return to unilateral defection was inversely related to the MPCR. Isaac et al. showed that in treatments in which MPCR was independent of group size (i.e., the material benefit from unilateral defection was *constant* across group sizes), the rates of defection in groups of size 40 and 100 were lower than in groups of size 4 and 10, in line with our theory (and contrary to most economists' expectations to find more free riding in larger groups). On the other hand, when they compared the rates of defection in groups of different size in situations where the MPCR was decreasing in group size (i.e., monetary return to unilateral defection *increasing* in group size), they found higher defection rates in the larger groups, again, in line with our prediction.[6] Although this is not a validation of our hypothesis, these experiments demonstrate its potential to explain some prominent group behaviors.

The limit on group size has another important implication. As we show in the paper, cooperation *within* groups (of limited size) emerges side by side

[6]The most striking evidence was probably the comparison of groups of sizes 4 and 10, where in both kinds of groups a token contributed to the group account was multiplied by the same multiplier, 3 (corresponding to MPCR's of 0.75 and 0.3 respectively). The experimenters found a significantly higher rate of defection in groups of size 10. This result was not replicated in a different experiment that compared groups of sizes 40 and 100, but this experiment with larger size groups was not conducted with monetary incentives, which may possibly affect the results.

with defection *between* groups. That is, the cost of cheating leads at the same time to the formation of groups and to the development of *in-group bias* - an inclination to cooperate only with members of one's own group. Otherwise, a person would have "too many" cooperative partners, and the temptation to defect would destroy cooperation both within and between groups. This is true in particular to *social types*, i.e., people whose social conscientiousness makes them subject to the aforementioned psychological cost of cheating. In our model we do not presuppose any initial difference in their empathy or commitment towards different individuals, yet we show that in equilibrium they are all non-cooperative toward out-group members. This result is in line with the experimental findings of Tajfel (1970), Tajfel et al. (1971), and more recently Chen and Li (2009), Efferson et al (2008) and de Cremer et al (2008), who show that the effect of in-group bias can be easily triggered by even the most trivial and arbitrary group categorization.

We distinguish the social types from *asocial types,* i.e., people who are *not* subject to the cost of cheating. When these people are easily spotted, they cannot form any social connections at all. However, when it is hard to spot the asocial types, the *social types* cannot form cooperative groups without having to lose something. In particular, if one's type is one's private information, then any partition of society into mutually exclusive groups contains at most two distinct kinds of groups. The first kind, which we call a *mixed group*, contains individuals of both types, where a minority of asocial types free ride at the expense of the social types. In a sense, in any modern state where most people pay taxes but some do not, yet everyone enjoys the social benefits provided by the state, a similar situation prevails. We show that this kind of group can always be sustained in equilibrium, but the limit on the group size is stricter than before. So in the absence of enforceable contracts and central authority, higher proportion of selfish or asocial individuals will be correlated with smaller social structures (e.g., families instead of tribes).[7]

Groups of the second kind, which we call *signaling groups*, consist only of

[7]Note that the explanation that goes in the other direction, saying that people are asocial *because* they live in small families and not in big tribes, takes the social structure as exogenous, while we believe it should be treated as endogenous.

social types, and their members fully cooperate with one another. Yet they need to screen out potential free riders. They do so by enforcing a practice of *costly signaling*, which means that members of the group obtain the trust and cooperation of the other group members only by exhibiting some self-sacrifice. We further show that the two kinds of groups can coexist in equilibrium. However, the existence of signaling groups strictly decreases the expected utility of *all* members of mixed groups, regardless of their type. Thus, beyond the private cost for the individual who signals, signaling as a phenomenon imposes a public cost on society. This public cost represents society's loss of "good guys", who form their own exclusive clubs instead of mixing with the other parts of society and lifting the average willing to cooperate. As for the members of these signaling groups, unless the proportion of asocial type is high enough, the possibility to signal decreases their welfare too, and signaling is shown to be unstable.[8]

The structure of the paper is derived mostly from the stylized facts that we wish to explain. After presenting few illustrative examples of in-group bias and covering some of the related literature, we present our benchmark model with complete information in Section 3. This model captures the tendency of people to form groups that exhibit in-group bias, and the tendency of social individuals to be "kind" (cooperative) only to in-group members.[9] Our model with incomplete information (Section 4) captures the connection between the cohesiveness of a group and the use of costly signaling of sociality, and analyzes the prospects for having a society in which groups with different levels of cooperation coexist. In Section 5 we show that signaling has a negative effect on the welfare of members of groups who do not use signaling, and that there is a tipping point for society in the form of a critical proportion of asocial types, such that below it signaling is not Pareto optimal even for the signalers themselves (and is unstable too), while above it signaling maximizes the welfare of social types. Section 6 demonstrates these properties of costly signaling using two real-life examples. Section 7 concludes.

[8] In the sense of core-stability of a partition of society into groups such that some of them use signaling.

[9] For experimental findings that support this assertion about social individuals, see de Dreu (2010) and the discussion of these findings in Section 3.

2 Examples of in-group bias and related literature

Although in-group bias is a pervasive phenomenon that everyone encounters, it is worthwhile to present some illustrative examples. These specific examples were chosen because they contain an individual incentive to free ride, and yet cooperation within the confined limits of the group is sustainable. These examples also demonstrate that in-group bias does not necessarily imply hostility or resentment towards out-group members, and can take the form of merely a higher level of cooperation within the group than with out-group members.

The first example is that of the Israeli Kibbutz. The Kibbutz is a special form of a collective community unique to the state of Israel, which can be thought of as a large scale experiment in Collectivism. Members of the Kibbutz contribute all their productivity to the Kibbutz (as a public good), and then equally share the aggregate product. The incentive to free ride is therefore crystal clear. Nevertheless, many of these communes survived for dozens of years (70 of them still function nowadays as collective communes who share their property between the members). The prediction of the model that such cooperation can be sustained only in limited size groups is plausible when comparing the relative success of many such Kibbutz communes (note that each Kibbutz is a separate group of "social types") to the almost total failure of Communism at the state level. As for showing in-group bias towards outsiders, this was neatly demonstrated in an experiment done by Ruffle and Sosis (2006), who let Kibbutz members play a game of sharing money with either an anonymous other member of their Kibbutz or a person "from another place". When paired with another Kibbutz member, subjects shared significantly more (on average) than when paired with a person "from another place".

The second example takes us one step further, into the model with incomplete information. Here we focus on virtual communities who share files, knowledge, or even tastes using the Internet. A representative example is that of Peer-to-Peer networks for sharing files. While the ability to download music and films from the folders of other users of the network displays a clear benefit to the individual user, having other users uploading files from the individual's own folders comes usually at a cost, such as a slowdown of his Internet connec-

tion. Together with the anonymity that the Internet provides (so that one's type is one's private information), it is clear why free-riding (called "Leeching" in the context of these Internet communities) is bound to thrive. Nevertheless, our model predicts that some level of resource sharing is still possible, even when providing resources imposes a cost on users. Using data collected from the *OpenNap* Peer-to-Peer network, Asvanund et al (2004) show that free riding limits the network size, yet small size networks are sustainable. Moreover, both sharing and free-riding are observed in these networks (see also Adar and Huberman 2000), which is in line with our model's prediction that free-riding is sustainable in equilibrium under incomplete information (Section 4). Nevertheless, some groups do set borders by utilizing authentication and access control technologies. By doing so they indirectly create in-group bias – those who do not have the password that is required for joining the network are left out, even if they would have shared their own files with existing members had they have the chance to do so. We refer the reader also to Demange (2010), who develops a theoretical model to explain the setting of group borders in the context of information sharing communities, where restricting access serves to achieve homogeneity of tastes.

A third example is *CouchSurfing*. A CouchSurfer is welcomed to sleep on the living room couch of a peer when traveling far from home, but is expected to host other peers on his own living room couch when asked to do so. While, for most people, hosting others imposes a cost, being hosted by others while traveling saves costs. This again clearly creates an incentive to free ride. And although the network of CouchSurfers does not maintain anonymity, the inherent globality of this network approximates anonymity (as most users use it in order to find solutions for sleeping abroad), and the use of the Internet as the infrastructure opens up opportunities for faking identities. Since membership is free, it seems that nothing can prevent free riders from getting free meals (even literally). So how are free riders screened out in practice? As a complementary initiative, the CouchSurfing community organizes all kinds of social activities, from camping trips and bar hoppings to meetings and sporting events. Though not formally presented as "qualification tests", it is quite clear that participation in such activities raises the chances of being invited to sleep over. In the

terminology of our model, these activities serve as a form of *costly signaling of sociality*, which can be successful in screening out the asocial types, who do not enjoy the social interaction with unknown people, and therefore will also not want to host strangers in their houses. As for the social types, the cost of attending these social activities may even be negative, as they might enjoy the presence of other social types, and so will also be delighted to host in their homes.

In Section 6, after presenting all the layers of the model, we demonstrate the costly signaling of sociality using two other examples of behaviors that are not often presented as such. The choice of the examples is motivated by the ability to demonstrate through them how the two kinds of groups, the mixed and the cohesive, can coexist, and how signaling inflicts a cost on society as a whole. The first example is related to the term "acting white". This term is mostly used to describe the pressure that is imposed on Black people who invest in particular behaviors (especially acquiring higher education) by their social peer group (Fordham and Ogbu 1986; Austen-Smith and Fryer 2005). We suggest to interpret the personal sacrifice of a Black individual who concedes to the pressure and *refrains* from these behaviors as a form of costly signaling of sociality. That is, by not acquiring higher education, the individual signals that he can be trusted not to forsake the Black brotherhood in pursuit of selfish goals at the expense of others. The second example we discuss there is that of religious rituals. There is extensive literature on the relationship between religious rituals and intra-group cohesiveness and cooperation (e.g., Sosis and Ruffle 2003, Ruffle and Sosis 2007, Hayden 1987, Turner 1969, Wilson 2002). However, with regard to out-group members, at least some of the religious groups "condemn deviance, shun dissenters, and *repudiate the outside world*" (lannaccone 1994). We show that religious rituals (such as Sunday prayers) enable the practitioners to signal their social value to the community and to screen-out potential free-riders, but the segregation of the practitioners from the other parts of society comes at the expense of the non-practitioners.[10]

[10]Levy and Razin (2012) develop a model where religious organizations play a significant role in enhancing cooperation through establishing belief in reward and punishment and through the possibility to signal membership in these organizations. However, a belief in punishment for bad deeds is conceptually different than our "cost of cheating", as the

The paper relates mostly to four literatures. The first is the literature on cheating and deception, the second is the literature on the link between cooperation and group size, the third is the literature on in-group bias, and the fourth is the literature on costly signaling.

Cheating, deception, lying and dishonesty, have all been recently in the spotlight of experimental study in behavioral economics (e.g., Gino, Norton and Ariely 2010, Hurkens and Kartik 2009, Gneezy et al. 2013, and Lundquist et al. 2009). The assumption of the current paper that the psychological cost of cheating is concave in nature is related to a concept called the "what the hell effect", which is generally used to describe behaviors that, once triggered, burst into full-fledge expression instead of developing gradually. Gino et al (2010) documented the "what the hell effect" of cheating in the dimension of time. They found that a person may be unwilling to cheat for a long period of time, but once he cheats for the first time, he often succumbs to full-blown cheating afterwards. Another dimension of the "what the hell effect" of cheating, the dimension of the size of lie, was reported by Gneezy et al (2013), who showed that when monetary payoffs were positively correlated to the size of lying, most subjects who decided to cheat a fellow participant chose the maximum size of lie. Moreover, Hurkens and Kartik (2009) found that their subjects could be divided into two distinctive types - those who lie whenever they can monetarily gain from lying (our asocial types), and those who never lie (our social types, assuming that the monetary temptation used in the experiment was not big enough).

The second related literature is the one about the link between cooperation and limited group size. The problem of sustaining cooperation in sizable groups was raised already by Olsen (1965). Bonacich et al. (1976), Bendor and Mookherjee (1987), Boyd and Richerson (1988), and Suzuki and Akiyama (2005), have all used the N-person Prisoner's Dilemma game in order to analyze this problem under various assumptions. However, these works do not try to explain the tendency to cooperate only with in-group members. Choi and

punishment for bad deeds is conditioned only upon one's own actions, thus ignoring the reciprocal element. In Section 6.2 we highlight the differences between the mechanism of Levy and Razin (2012) and ours.

Bowles (2007) and Fu et al (2012) do provide evolutionary models that explain at the same time altruism within the group and parochialism between groups, but do not account for group size. These works are closely linked to the third related literature, which is the one that documents in-group bias.

We already mentioned some lab experiments that demonstrated the minimal group effect, i.e., that in-group bias can be triggered by arbitrary group categorization. Goette et al. (2006) showed a similar effect in a field experiment, where the arbitrary group categorization was the division of soldiers into platoons in the Swiss army. When it comes to naturally formed groups, such as ethnic or racial groups, Bernhard et al (2006) showed in-group bias among ethnic groups in Papua New Guinea, and Fong and Luttmer (2009) showed racial in-group bias among contributors to Hurricane Katrina victims.[11] All these works, whether in the lab or in the field, whether with randomly assigned groups or with natural ones, involved subjects playing canonical experimental games, such as the dictator game and the Prisoner's Dilemma. But recently, in-group bias was verified also using naturally occurring data. Shayo and Zussman (2011) were able to expose in-group bias in real-life decisions by professionals, where the decisions had significant implications to the parties involved. They analyzed judicial decisions in Israeli courts, where strong nondiscriminatory norm applies, and demonstrated empirically the existence of in-group bias in the decisions of judges.

The fourth related literature is the one on costly signaling. The canonical works in this literature are Spence's (1974) model of education as a signal in the labor market, and the models of reputation signaling in firm competition by Kreps and Wilson (1982) and Milgrom and Roberts (1982). When it comes to signaling as a means to acquire cooperation and social connections, Gintis et al (2001) develop an evolutionary model of costly signaling as a promoter

[11] It is interesting to note that experiments that use the Trust Game instead of allocation games tend to show much more variation in behavior towards out-group members. Hennig-Schmidt et al. (2009) find no in-group bias when letting Germans, Israelis and Palestinians play the Trust Game with in-group and with out-group members. Similarly, Bornhorst et al. (2010) find no regional discrimination in an experiment involving students of different European nationalities who are matched to play this game in mix-nationality groups. Even more strikingly, Fershtman and Gneezy (2001) reveal out-group favoritism among Israeli Jews of eastern decent, who show more trust towards Israeli Jews of western decent.

of cooperation in the group level, and Camerer (1988) analyzes gift exchange as signaling intentions for future investments in pairwise relations. Even more closely related to our paper are the work of Akerlof and Kranton (2000), who discuss costly signaling of one's identity, lannaccone's (1992) work on social clubs, where signaling is used by individuals as a means to be accepted to desired groups, and the work of Levy and Razin (2012), where participation in religious rituals signals a greater inclination to cooperate. Finally, Benabou and Jean Tirole (2006) suggest a model where pro-social behavior (charity in their case) is used as a means for signaling quality, and not as an indicator of it's independent existence.

3 A basic model of in-group bias

We model society as a continuum of individuals of size $N = 1$, who simultaneously interact with each other to play one-shot Prisoner's Dilemma (PD) games. We follow the notations of Kandori (1992) and Ellison (1994) and use the following payoff matrix for the game:

	C	D
C	$1, 1$	$-l, 1+g$
D	$1+g, -l$	$0, 0$

The zero payoff for mutual defection suggests that there is no difference between mutual defection and no interaction at all, thus relaxing the somewhat unrealistic assumption that each individual is practically engaged in simultaneous plays against all members of society (an assumption that aims to keep the model as parsimonious as possible). Furthermore, it implies that the payoff for mutual cooperation is strictly positive, hence the total return to cooperation increases in group size (nevertheless, groups will be of limited size in equilibrium). g stands for the *gain* from unilateral defection, and l for the *loss* from being the victim of the opponent's unilateral defection. We assume strategic complementarity, i.e., $l > g$, which implies that if one's opponent is more prone to defect, one is more prone to defect too. Our analysis considers only pure strategies at the pairwise level, but individuals can discriminate between

opponents, i.e., cooperate with some while defecting against others.[12]

Society is composed of two types of individuals, $\tau \in \{s, as\}$, where s stands for *social type* and as stands for *asocial type*. Asocial types are affected only by the material payoffs of the game, and so for them defection is a dominant strategy against any opponent. Unlike them, social types may lose utility by *cheating*, where cheating means playing D against an opponent who plays C.[13]

Let $t(k)$ denote the cost of cheating against a measure k of individuals. This can be thought of as a psychological cost caused by the arousal of uncomfortable feelings such as shame or guilt on the side of the defector.[14] We naturally assume that $t(0) = 0$, and that $t(k)$ is weakly increasing in k - the more people are cheated by the individual, the (weakly) more it costs him. Additionally, we put some restriction on the form of this increase. In particular, we assume that the "what the hell effect" of cheating, as discussed in the introduction, applies. With regard to modeling, this effect can be modeled as a cost function $t(k)$ that is concave in k. We do not require smooth concavity or even continuity, so that any cost function with a discrete jump at 0 and a weakly increasing and concave continuation afterwards satisfies our concavity condition, and in particular this includes one with a fixed cost of cheating for any $k > 0$.[15] The other requirements are similar in spirit to the INADA conditions – an infinite slope at 0 (or a discrete "jump"), and a "small enough" cost as k

[12] Mixed strategies pose here a modeling ambiguity. Since part of the payoff is going to be related to disutility from defecting against a cooperative opponent, it is not clear how one should feel when defecting against an opponent who uses a mixed strategy - is it the realization that counts, or maybe the (impure) intention to cooperate? We prefer to leave these potential controversies aside.

[13] Note the difference between defecting, i.e., playing D, and cheating, i.e., playing D against an opponent who plays C. The idea here is basically that the people whom we call "social types" do not like to exploit others (e.g., with respect to the examples in Section 2, these are the people who would not like to let other Kibbutz members work for them, or who would not like to refrain from sharing their own files or couch in exchange for others' sharing). Miettinen and Suetens (2008) indeed show that (most) people feel guilty when defecting in the PD game only if the partner has not defected as well.

[14] This interpretation is in line with that of Lopez-Perez (2008), with the exception that he would treat the k cooperators as those who respect the norm, and the defector as the norm breaker. In Lopez-Perez (2008) $t(k)$ is linear in k and the groups are of fixed size.

[15] In terms of social identity theory, a discrete jump captures the change in one's perceived self image from a self image of someone who never cheats, to a self image of someone who potentially cheats (the border between these two distinct characters is nicely captured in the recent experiments on lying aversion of Hurkens & Kartik 2009 and Gneezy et al. 2013).

goes to 1 (this condition could have been replaced with a flat slope as N goes to infinity if N was unbounded). Formally, the assumptions on $t(k)$ beyond positive monotonicity and concavity are:

$$t(0) = 0,\ t(1) < g, \text{ and } \lim_{k \to 0} t'(k) = \infty$$

$$(\text{or } \lim_{k \to 0^+} t(k) > 0 \text{ if } \lim_{k \to 0} t'(k) \text{ is not defined})$$

In the benchmark model with complete information that we analyze in this section, we assume that the type of each individual is common knowledge. The strategy of player i subscribes the action played in the PD encounter with any other player j. We denote this element by s_{ij}. We say that society is in (Nash) equilibrium if, given the strategies of all other individuals, no individual has a profitable deviation from his strategy. The following result implies that cooperation can be sustained within groups of social types, but *in-group bias*, i.e., defection against out-group members, is bound to emerge too.[16]

Lemma 1 *The equation $t(K) = Kg$ has a unique strictly positive solution $\bar{K}$ in $]0,1[$. Moreover, $t(K) > Kg$ for every $K \in \left]0, \bar{K}\right[$ while $t(K) < Kg$ for every $K \in \left]\bar{K}, 1\right[$.*

Proof. See the appendix. ■

Proposition 2 *Let $\bar{K} \in (0,1)$ be the unique strictly positive solution to the equation $t(K) = Kg$. Then in equilibrium:*

1. Every asocial type plays D against everyone else, and everyone else plays D against him.

2. Every social type plays C against a mass of individuals of size $\bar{K}$ or less, who play C against him too, and plays D against everyone else.

Proof. Since, for both types, defection is a best response against an opponent playing D himself, we get that in equilibrium, if $s_{ij} = D$ then $s_{ji} = D$. Hence,

[16] The degenerate state where every individual plays D against everyone else is naturally sustainable in equilibrium too.

since D is a dominant strategy for asocial types, and types are common knowledge, we get (1). Next, it follows from Lemma 1 that $t(K) > Kg$ for every $K < \bar{K}$ while $t(K) < Kg$ for every $K > \bar{K}$. If a mass of K players play C against a social type, and $K \leq \bar{K}$, then his best response to all of them is C, since deviating to defection against any subset of them (of size $k \leq K$) would impose on him a net cost of $t(k) - kg \geq 0$. Otherwise, if $K > \bar{K}$, then playing C against all of them cannot be his best response, because deviating to playing D against all of them would increase his total payoff by $Kg - t(K) > 0$. This completes the proof of (2). ■

Our main focus in this model will be on partitions of society into mutually exclusive groups, so we present now the following definition and an important corollary of the proposition:

Definition 3 *Let a cohesive group be a collection of individuals who cooperate with each other, and defect against all out-group members.*

Corollary 4 *Any partition of the social types into cohesive groups whose sizes are bounded by $\bar{K}$ can be sustained in equilibrium.*

This result implies that it is easier to sustain cooperation in smaller groups.[17] It sounds plausible when considering the limited size of tribes and clans, especially in societies with no central authority, where groups are presumed to form spontaneously. The driving force behind this result is the "what the hell effect" of cheating - as the size of the group increases, it becomes harder to avoid the temptation to defect and achieve the ever growing material benefits of unilateral defection. At some point this effect is going to burst out, leading to cheating across the board. The limit on group size in equilibrium is the threshold above which such across the board defection is bound to occur.

Another aspect of the result is its built-in in-group bias. It turns out that social types would show the same level of asociality towards out-group members as would asocial types, while exhibiting sociality only towards in-group members. This result is in line with the behavior of Kibbutz members in the

[17] If $\bar{K}$ is larger than the mass of social types then the social types can be united in one group, showing in-group bias only towards the asocial types.

experiment of Ruffle and Sosis (2006) mentioned earlier, who exhibited the same level of generosity as that of city residents towards anonymous out-group peers, while showing higher levels of generosity towards anonymous in-group peers. It is also in line with experimental studies of the Prisoner's Dilemma. For example, Wilson and Kayatani (1968) and Dion (1973) find that the competitiveness which characterizes inter-group behavior resembles that of individual players, whereas it is the increased proportion of cooperative choices exhibited in intra-group decisions that deviates from typical inter-personal play (see further analysis in Brewer 1979). More recently, de Dreu (2010) used the Intergroup Prisoner's Dilemma to show that compared to individuals with a "chronic pro-self orientation", those with a "chronic prosocial orientation" (these would be the social types in the jargon of the current paper) display stronger ingroup trust and ingroup love — they self-sacrifice to benefit their ingroup — but not more or less outgroup distrust and outgroup hate. As we show in the next section, the self-sacrifice practiced by social types is not always intended to benefit the ingroup, but can rather be a means of signaling membership in the group.

4 In-group bias under incomplete information

4.1 Cooperation and sustainable free riding

The basic model with complete information implicitly assumed that a social type can consider the cooperation of other group members as guaranteed. This assumption is a bit unrealistic when considering pairwise PD game. Moreover, the assumption that asocial types can be easily distinguished from social types is quite strong. We therefore turn now to consider the case where the individual's type is his private information. We assume that the mass of asocial types in society is p, and that this is common knowledge. Can there still be an equilibrium with some cooperation in it? The following proposition, preceded by a definition, shows that the answer is affirmative.

Definition 5 *A mixed group is a collection of individuals of both types, such that:*

- *All social types in the group play C against all in-group members, and D against all out-group members.*
- *All asocial types in the group play D against both in-group and out-group members*

Proposition 6 *Given $p \in (0,1)$, $\exists K_p \in (0, \bar{K})$ such that a mixed group of size K is sustainable in equilibrium if and only if $K \leq K_p$. Furthermore, K_p is decreasing in p.*

Proof. First, it is obvious that asocial types have no profitable deviation, since as members of mixed groups they already play their dominant strategy D against everyone else in society. As for the social types, consider an individual of type s who is a member of a mixed group of size K. By the definition of a mixed group, all the social types in the group play C against all other group members, including him. However, due to the incomplete information, the individual cannot choose to defect only against the asocial types in the group.[18] Defecting against any subset of the group, of mass $k \leq K$, of which only a fraction of $(1-p)$ are of type s,[19] would result in an increase in expected material payoff of $k[(1-p)g + pl]$, but the expected total payoff would also decrease by $t((1-p)\,k)$ due to the cost of cheating. The individual would have no profitable deviation if and only if $t((1-p)\,k) \geq k[(1-p)g + pl]$ for every $k \leq K$. Let $\Delta\,(k,p) \equiv t((1-p)\,k) - k[(1-p)g + pl]$. The conditions on $t(k)$ and on the payoffs of the game imply that for any given $p \in (0,1)$, we have $\Delta\,(0,p) = 0$ and $\Delta\,(k,p) < 0$ for every $k > \bar{K}$, because $t((1-p)\,k) \leq t(k)$, $[(1-p)g + pl] > g$,

[18] One can think of this setup as having an initial stage where the groups are formed, followed by a stage in which each individual chooses a type-contingent strategy, and a final stage in which each individual randomly draws his type. If no one in the group wishes to deviate from his type-contingent strategy, then these strategies are sustainable in equilibrium.

[19] Strictly speaking, the distribution of realizations (in terms of the exact proportion of social types) over an interval of size k is not defined. However, it is common to assume that the measure p applies to any subinterval of the original range $[0,1]$. One way to ensure it is to have society represented by $[0,1]^2$, such that the choice of partners is applied only in one dimension (represented by choosing a subinterval of $[0,1]$), while the other dimension guarantees that the proportion of asocial types is p for every chosen set of partners. Anyway, the proposition holds also for a model with a discrete number of individuals instead of a continuum, where the realizations are clearly defined – see the appendix for this more elaborate proof.

and $t(k) < kg$ for every $k > \bar{K}$. Moreover, $\lim_{k\to 0} \Delta'(k,p) = +\infty$ (or, if $\lim_{k\to 0} t'(k)$ is not defined, $\lim_{k\to 0^+} \Delta(k,p) = \lim_{k\to 0^+} t(k) > 0$), and $\Delta(k,p)$ is weakly concave in k. Thus, $\exists K_p \in (0,\bar{K})$ such that $\Delta(K_p,p) = 0$, $\Delta(k,p) > 0$ for every $k < K_p$, and $\Delta(k,p) < 0$ for every $k > K_p$,[20] which proves that mixed groups of size K are sustainable if and only if $K \leq K_p$. Finally, $\Delta(k,p)$ is strictly decreasing in p, which means that for any $\{p,q|p<q\}$ we have $\Delta(k,q) < 0$ for every $k \geq K_p$, and so $K_q < K_p$, i.e., K_p is decreasing in p. ■

Corollary 7 *Any partition of society into mixed groups whose sizes are bounded by K_p forms a Bayesian equilibrium.*

In these equilibria, social types would show in-group bias, by playing C against all group members and D against all outsiders, while asocial types would play D against everyone, thus "free-riding" on the social types in their group.[21] Such groups are bound to be smaller than the groups of purely social types in the complete information case (i.e., $K_p \leq \bar{K}$), because here the temptation to defect is larger (avoiding the sucker payoff l is assumed to increase expected payoff at least as much as gaining g by defecting against a cooperative opponent) and the cost of cheating is lower (because defecting against an asocial type who defects himself is not psychologically costly). Moreover, that the maximal group size is decreasing in p implies that the greater is the proportion of asocial types in society, the more it is tempting for social types to defect, thus the smaller are the groups that can sustain cooperation.[22] The behavior of social types in this kind of equilibrium can be illustrated by reconsidering some of the examples of Section 2. In the context of these examples, the social types who endure the free riding are the Peer-to-Peer network users

[20] If $\Delta(k,p)$ is discontinuous due to discontinuity of $t(k)$, then the same logic of the proof to Lemma 1 applies here too.

[21] Strictly speaking, we can also get Bayesian equilibria in which not all social types in mixed groups cooperate, accompanied by an appropriate system of beliefs, but in such equilibria all mixed groups must be of the same size, which is the unique size that would make social types indifferent between cooperation and defection. We believe that such restrictions make these equilibria less interesting.

[22] However, groups that are small enough can still sustain cooperation, because in such groups the material payoffs are low so there is not much to gain by defection, yet the psychological cost of cheating kicks in already with the first potential cheating.

who keep sharing their music folders knowing there are free riders ("leeches") using the network and enjoying their shared folders without reciprocating, or the CouchSurfers who host on their living room couch someone who has not hosted anyone yet, and can potentially be a free rider.[23]

An interesting scenario is revealed when considering the case of $p(1+l) > 1$. In this case, the proportion of asocial types in society is high enough to make the expected payoff of a social type in a cohesive group of size K *negative*, regardless of the exact group size ($K[1-p(1+l)] < 0$, $\forall K$). This means that the social types would have been better off in a society with full-blown defection (where the payoff of everyone is zero), yet, if $K \leq K_p$, they end up playing C against their group members for a negative expected payoff. One can think of this situation as resembling the frustrating state of someone who pays taxes in order not to free ride other people like him, in a country that is so corrupt that he would be better off with no tax system and no public service at all. This state of affairs invites the use of costly signaling of sociality. Can such signaling be efficiently used by social types to distinguish themselves from asocial types? We consider this option in the next subsection.[24]

[23] The behavior of social types in this equilibrium bears some similarities to the behavior of "conditional altruists" in Palfrey and Rosental (1988). However, Palfrey and Rosental assume that the payoffs of a contributor (= cooperator) are unaffected by the opponent's strategy, and so even "conditional altruists", who condition their strategy on their expectations from the opponent, contribute only because they fear from mutual defection and not because they feel obliged to contribute when others do so. The main differences between our results and theirs are that in ours group size plays a significant role in determining the players' strategies, and the threshold for cooperation of social types is affected by beliefs that are derived from the actual proportion of social types in society.

[24] In a similar model, Camerer (1988) models gift exchange as a system of costly signalling intentions for a long term investment in relationship. For some values of the parameters in his model, he gets that the "Willing" types, who prefer investing if and only if their partner invests too (similar to our social types), would choose to invest when playing against an unknown type when there is incomplete information. This is equivalent to the case of social types playing C against an unknown group member in our model. However, Camerer jumps then to the conclusion that in this case there is no potential for signaling, because the purpose of signaling is to elicit investment by a "Willing" opponent, who anyway invests. We claim that not only is it possible to prove that separating equilibria with costly signaling *can* exist in such a case, but also that the emergence of signaling is plausible and almost even self evident in some cases, e.g., when $p(1+l) > 1$ in our model.

4.2 Suckers and signalers

Costly signaling is a well-known means to achieve a separating equilibrium (e.g., Spence 1974). In our model, the signal (if it indeed achieves separation) can be interpreted as indicating "sociality", because anyone, regardless of one's type, would like to be regarded as a social type and achieve the cooperation of his opponents.[25] So a necessary condition for a separating equilibrium is that the cost of signaling sociality will be lower to social types, compared to asocial types. However, this condition is not sufficient, since the gain of an asocial type from being considered social exceeds that of a truly social type, so an asocial type will be willing to pay a higher cost in order to be perceived as social.

Let x_s and x_{as} be the cost of signaling sociality for types s and as respectively. The signal is not directed at any specific opponent, but is rather a singular payoff-irrelevant sacrifice that is observable by everyone else.[26] A person would signal sociality if this signal made others treat him as social, and if his increase in total expected payoff from being treated as social exceeded the cost of signaling. Being treated as social means getting the cooperation of potential group members. Recall now that the return to both cooperation and defection is increasing in the number of cooperative partners. So what is the lower bound on the number of cooperative partners that makes signaling sociality profitable to each type?

Assume that a fully separating equilibrium exists, i.e., types are believed to

[25] This interpretation of the signal as revealing sociality holds even in the case of $p(1+l) > 1$ discussed above, where social types get negative payoffs as members of cooperative groups, thus do worse than they could do by being perceived as asocial types. In this case, it seems that they have incentive to signal *asociality*. If believed, their opponents would defect, and they would then be able to defect too without any pangs of conscience, while improving their expected payoff. But should they be believed indeed? Since truly asocial types would not like to be revealed as such (they have a strictly positive expected payoff as members of a mixed group), they themselves would not want to signal asociality. Clearly, in such a case, signaling asociality would in fact reveal one as *social*.

[26] It is not unreasonable to have signaling in the level of the group or the society as a means to promote pairwise relationships. As Gintis et al (2001) write in their paper titled 'costly signaling and cooperation', "it is often the case that biological signals in other domains such as mate choice, resource competition, and even predator-prey interactions are not private to an intended receiver, but are emitted without the signaler knowing exactly with which among a population of possible observers it might influence". Evidence from Meriam turtle hunters is consistent with this claim (Smith et al 2003).

be social if and only if they signal, and these beliefs are consistent with reality. Then asocial types would not be trusted by anyone and thus have a zero payoff, while social types would be able to form cohesive groups of size K and get a payoff of $K - x_s$. Of course, the upper limit on K from the benchmark model with complete information still applies here, i.e., $K \leq \bar{K}$. Furthermore, to be indeed consistent with reality, no one should be able to profit by changing his signaling decision. If a social type deviates to not signaling, he can expect to be treated as asocial and thus meet only defecting opponents and get a zero payoff, but to save the cost x_s. This deviation would not be profitable if $K > x_s$, so $K_s \equiv x_s$ is a lower bound on the number of cooperative partners that makes signaling sociality profitable to a social type. If an asocial type deviates to signaling, he can make $1+g$ against every social type that plays C against him, but signaling would cost him x_{as}. So this deviation would be profitable if he can expect to meet at least $K_{as} \equiv \frac{x_{as}}{1+g}$ such cooperative opponents. Therefore, a social type can be sure that his K signalling group mates are truly social only if $K < K_{as}$. This means that $K_s < K_{as}$ is a necessary condition for a separating equilibrium, or, written as a condition on the ratio of the costs of signaling, $\frac{x_{as}}{x_s} > 1+g$. That is, to get separation it is not sufficient that signaling sociality would be cheaper to social types (a reasonable assumption in itself, reflecting the notion that it should cost more to fake sociality than to signal it when it indeed exists, as in Frank 1987), but the ratio of costs must also exceed the ratio of marginal gains from a cooperative opponent. If this condition of separability holds, and if the condition of individual rationality, $x_s \leq \bar{K}$, holds too, then in a separating equilibrium we can get signaling groups of purely social types, where the size of each such group will be $K \in \left[K_s, \min\left\{K_{as}, \bar{K}\right\}\right]$. Otherwise, if $\hat{K} \equiv \min\left\{K_{as}, \bar{K}\right\} < K_s$, then social types cannot distinct themselves from the asocial types by signaling, either because the cost of signaling that is needed to get the cooperation of the other social types exceeds the maximal benefit from this cooperation (if $\bar{K} < K_s$), or because they can be imitated by asocial types (if $K_{as} < K_s$).

The result that signaling groups are bounded in size both from below and from above is in line with evidence reported in Iannaccone (1994). With respect to the lower bound set by the cost of signaling, Iannaccone reports that stricter

churches tend to be larger, and relates it to the capacity of the high cost of adherence to their rules of conduct (which we interpret as signaling – see Section 6) to screen out potential free riders. In our model too, the higher cost of signaling used by stricter churches implies that they must attract many members in order to survive, thus the surviving strict churches are bound to be large. However, we believe that in terms of analyzing the ability of strict churches to attract followers and not just screen out free riders, it is equally important that imitation would be even more costly, so that by paying the high cost of signaling the followers can distinguish themselves from all those for whom the cost is just not worth it. As for the upper bound on the size of the signaling groups, Iannaccone further writes that the data "imply 'optimal' levels of strictness, beyond which strictness discourages most people from joining or remaining within the group," i.e., signaling should be individually rational in order to be pursued.

It is important to note that $K_s < \hat{K}$ is only a necessary condition for a separating equilibrium, i.e., it does not *guarantee* that a fully separating equilibrium will indeed emerge. There is always an equilibrium where everyone plays D, and there are always pooling equilibria in which no one signals yet cooperation is maintained in mixed groups whose sizes are bounded by K_p. In these cases, a social type cannot hope to gain from a unilateral deviation to signaling his type, even if the signal is known to be truthful.

The fact that separation is not guaranteed even when the signal is reliable supports a stylized fact we want to explain here - the coexistence of cohesive groups with costly signaling (which will be referred to as *signaling groups*) side by side with groups that are less cooperative, and whose members do not engage in costly signaling. This situation may emerge if a fraction $1-\lambda$ of the social types form signaling groups of purely social types (of sizes at the range $[K_s, \hat{K}]$), while the other social types are members of mixed groups in which asocial types "free-ride" on the social ones. The splitting of the social types into two kinds of groups increases the proportion of asocial types in the mixed groups, thus further constrains the size of this kind of groups. That is, let $q \equiv \frac{p}{p+\lambda(1-p)}$ be the proportion of asocial types in the mixed groups. Following the same analysis as before, K_q $(< K_p)$ will be the new upper bound on the

size of mixed groups. We call such an equilibrium with both kinds of groups a *hybrid equilibrium*. Figure 1 maps the possible equilibria as a function of the cost of signaling for each type.

5 Welfare and stability of equilibria with signaling

The multiplicity of equilibria invites a comparison of them in terms of stability and welfare. As will be shown here, these concepts are tightly related. We start by analyzing the effect of signaling on the welfare of individuals who do not signal and on the signalers themselves, and then introduce a stability concept that helps to formalize the welfare result. We will explicitly focus now on partitions of society into mutually exclusive groups, where each group can be either a mixed group or a signaling group.

Definition 8 *Let a coalition formation be a partition of society into mutually exclusive groups such that the individuals' strategies under this partition form an equilibrium.*

The following result is about the negative externality of signaling on society. It essentially states that the existence of signaling groups strictly decreases the expected utility of *all* members of mixed groups, regardless of their type.

Proposition 9 *Assume that $K_s < \hat{K}$, and let p be given. Then the expected payoff of all the non-signalers in any coalition formation that contains a non-zero mass of signalers can be strictly increased by prohibiting signaling.*

Proof. Take any coalition formation Π that contains a non-zero mass of signalers. Then q, the proportion of asocial types among the non-signalers, is strictly greater than p, their proportion in society. Take now a different partition Π' with no signaling, such that all the members of mixed groups under partition Π are still members of mixed groups of the same size under Π'. This partition can be sustained in equilibrium since $K_q < K_p$ (see Proposition 6). Moreover, under partition Π' each of these individuals gains a higher expected payoff than under partition Π, because, for any given group size, the expected payoff of members of mixed groups (whether they are social or asocial) is strictly decreasing in the proportion of asocial types among the non-signalers. ■

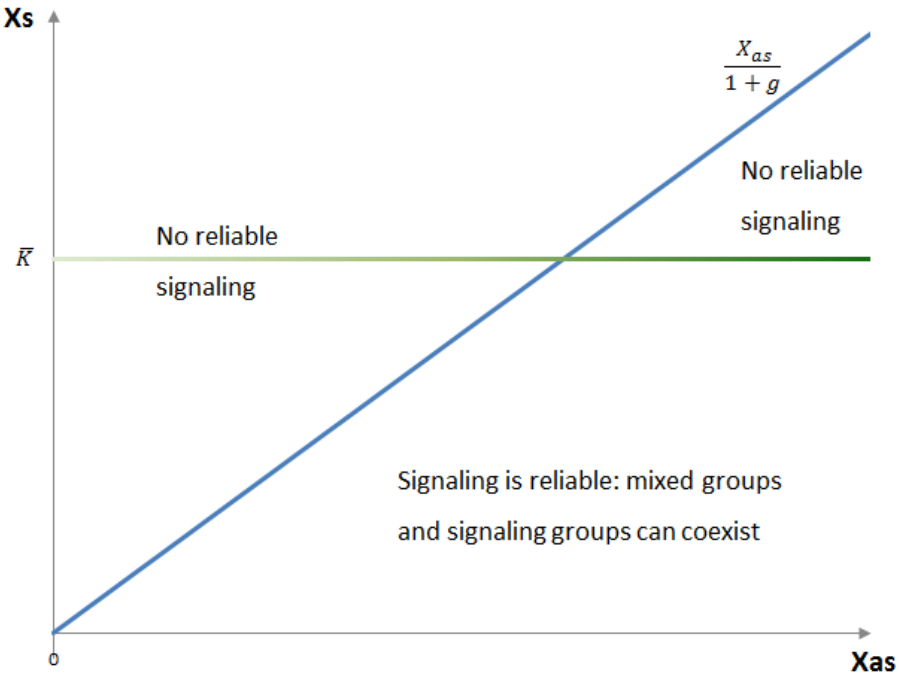

Figure 1: Displaying the necessary conditions for reliable signaling and the potential for separating and hybrid equilibria as a function of the cost of signaling for asocial types (x_{as}) and for social types (x_s). The blue diagonal line is where the ratio of these costs is equal to the ratio of the benefits from cooperative partners, i.e., $\frac{x_{as}}{x_s} = 1 + g$. It marks the border between the region where social types can distinguish themselves from the asocial types by signaling (below it to the right) and the region where they cannot (above it to the left). Moreover, if the cost of signaling for the social types is above the green line (marking a cost of $\bar{K}$), then signaling is not individually rational for a social type, as the gain from cooperation cannot exceed $\bar{K}$ in equilibrium.

Thus, beyond the individual cost for the signaler, signaling as a social phenomenon imposes a public cost on society. This public cost represents society's loss of "good guys", who form their own exclusive clubs instead of mixing with the other parts of society and lifting the average willingness to cooperate. One may think that at least for the signalers themselves signaling is an optimal strategy, otherwise there wouldn't have been an equilibrium with signaling. However, the following lemma states that this is the case only for large enough values of p.

Lemma 10 *Assume that $K_s < \hat{K}$, and let p_c be the unique implicit solution to the equation*

$$\hat{K} - x_s = K_p[1 - p(1 + l)]. \tag{1}$$

Then there is a tipping point for social types – in the equilibria that maximize their expected payoff, they are signaling if and only if $p \geq p_c$.

Proof. First recall that the expected payoff of a social type in a mixed group of size K is $K[1 - p(1 + l)]$, which is negative if $p > \frac{1}{1+l}$, but positive and increasing in the group size if $p \leq \frac{1}{1+l}$, where, given p, it reaches its maximal value $f(p) \equiv K_p[1 - p(1 + l)]$ when the group is of maximal size, K_p. Since, for $p \leq \frac{1}{1+l}$, both K_p and $[1 - p(1 + l)]$ are positive and decreasing in p (see Proposition 6), we get that $f(p)$ is also positive and (strictly) decreasing in p at $p \in \left[0, \frac{1}{1+l}\right]$. Moreover, $f(0) = \bar{K} > \hat{K} - x_s$, and $f\left(\frac{1}{1+l}\right) = 0$. Hence, given that $K_s < \hat{K}$, there is a unique solution to equation (1), denoted by p_c, and $p_c \in \left(0, \frac{1}{1+l}\right)$. Next, since $\hat{K} - x_s$ is the maximal equilibrium payoff in signaling groups, we get that if $p < p_c$ this payoff is strictly smaller than the payoff achievable by social types in mixed groups. If on the other hand $p \geq p_c$, then the converse is true – the maximal payoff achievable by social types in mixed groups is smaller than $\hat{K} - x_s$, which is achievable in signaling groups. ■

Lemma 10 roughly says that social types can be better-off by signaling if and only if the proportion of asocial types is high enough, with p_c being the tipping point. Figure 2 illustrates this result. The following proposition links this result to the concept of *core stable* coalition formations, as introduced by

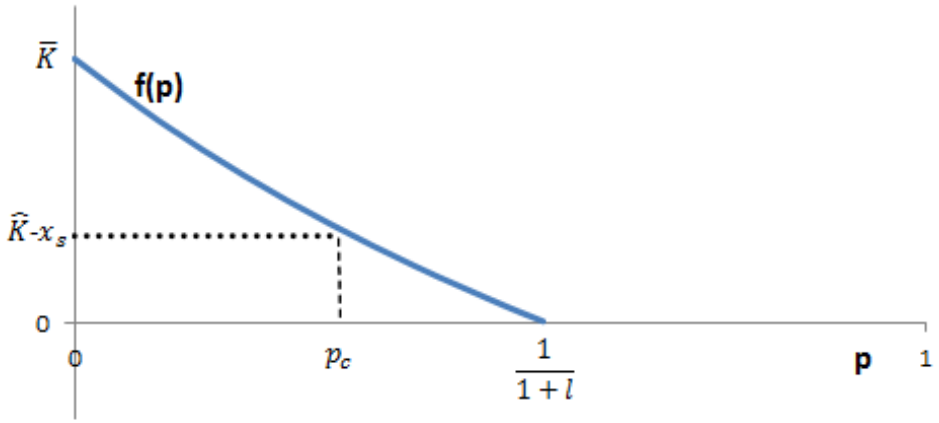

Figure 2: When groups are of maximal size, $f(p)$ is the expected payoff of social types in mixed groups in a pooling equilibrium (i.e., when there is no signaling in society) and $\hat{K} - x_s$ is their expected payoff in signaling groups. If p, the proportion of asocial types in society, is smaller than p_c, then a pooling equilibrium where all groups are of maximal size Pareto dominates all other equilibria, and so signaling is wasteful. If $p_c < p$, then social types can get a payoff of $\hat{K} - x_s$ in signaling groups, which is strictly greater than the expected payoff they can achieve in mixed groups.

Bogomolnaia and Jackson (2002) and Banerjee et al (2001).[27]

Definition 11 *A coalition formation* Π *is said to be unstable if there exists another coalition formation* Π' *and in it a coalition* $T \notin \Pi$*, such that all members of* T *have strictly higher payoff under* Π' *than under* Π*.*

In our context T would be a mixed group or a signaling group that does not exist under the considered coalition formation Π, yet is feasible in equilibrium.

Proposition 12 *If* $p < p_c$*, then any coalition formation with a non-zero mass of signalers is unstable.*

Proof. Let Π be a coalition formation with a non-zero mass of signalers, and let q denote the proportion of asocial types among the non-signalers under this

[27] In Bogomolnaia & Jackson (2002) and Banerjee et al (2001) the number of players N is finite, but this does not prevent the adoption of their core stability concept. The only adjustment needed here is to restrict the proposition to the case of a non-zero mass of signalers.

coalition formation. Consider now a different coalition formation Π' in which there is no signaling, and which contains a mixed group T of size K_p. If $p < p_c$, if follows that $p < \frac{1}{1+l}$, in which case the expected payoff of every member of T is strictly higher than the maximal expected payoff he can obtain in a mixed group under partition Π (because $p < q \Rightarrow f(p) > \max\{f(q), 0\}$ – see the proof to Lemma 10 and Figure 2). Furthermore, since T is of maximal size, the fact that $p < p_c$ implies (by Lemma 10) that the expected payoff of social types in T is higher than the expected payoff of any member of a signaling group under partition Π. Thus, the expected payoffs of all the members of T are strictly higher than their expected payoffs under coalition formation Π, and so this coalition formation is unstable. ■

As already implied earlier, the signaling groups can be thought of as social clubs, cults or communes, as in the work of Iannaccone (1992). It is interesting to note that Proposition 2 in Iannaccone's paper says that if society consists of two types of people, type 1 and type 2, such that type 1 people participate in group activities and value group quality less than type 2 people, "then, *as long as people of type 1 constitute a sufficiently large fraction of the population*, there will exist a signaling equilibrium in which type 2 people end up in groups that require their members to sacrifice a valued resource or opportunity". The equivalents to type 1 and type 2 people in our model are asocial types and social types respectively. So Iannaccone's result as stated in his Proposition 2 is similar in spirit to our Proposition 12, in the sense that if one restricts attention only to stable equilibria, then these equilibria can contain signaling only as long as the asocial types constitute a sufficiently large fraction of the population.

6 Two Examples of Costly Signaling of Sociality

6.1 On Acting White

One salient case of costly (and probably wasteful) signaling in cohesive communities is the one related to the "acting White" accusation in the Black American society. When thinking about "acting White", many tend to focus on those who do try to acquire education, and the social cost they have to bear by doing

so, but we believe that the focus should be instead on those who *do not* try to acquire education. That is, the cost is in fact for "remaining Black", not for "acting White".

In order to see why "acting White" can be explained with our model, let action D in the PD game be interpreted as pursuing individual goals, and let action C be interpreted as contributing to the Black community one comes from. Using the PD game to model this situation implies that from a selfish perspective, pursuing individual goals is always better, but everyone in the Black community would be better off if all contributed to it than if all pursued individual goals.[28] The social types are the Black individuals who are willing to sacrifice some self profits for the benefit of their community if others do it too (unless the number of contributors is big enough to make them free-ride). In the case of incomplete information, people in the community cannot know who will eventually comeback to the community to contribute and who will shirk from contribution. Then, the costly signal is naturally the self-sacrifice of a Black person who refrains from the pursuit of individual goals such as education or career opportunities in order to avoid being perceived as 'acting like White people do'.

Consider now the case of a hybrid equilibrium, where the costly signal is the personal cost of giving up education and staying in the Black neighborhood. In such equilibrium, some people will give up education and form signaling groups in their communities, and some will acquire education. Those giving up education will enjoy the cooperation and support of their group mates, at the cost of staying uneducated. Those acquiring education will consist of social types who go back to the community to contribute, and asocial types who leave their communities in pursuit of their individual goals. Note that in our model, these are only the social types in the mixed groups, i.e., those who acquire education and comeback to contribute, that suffer from the defection of the asocial types (the members of the signaling groups are only affected indirectly through the need to costly signal in order to distinguish themselves). A plausible explanation for that would be that the departure of the asocial

[28] This should not necessarily apply to the White community too for various reasons, such as differences in socioeconomic status or in community structure.

educated Blacks imposes a higher burden on the educated Blacks who return to the community (because they share this burden with less people), while from the point of view of those who stay in the community, the total contribution acquired is the same.

The payoff structure captures correctly the fact that the asocial types are clearly better-off by acquiring education, and are much better-off if others (the educated social types) pay back to the community on their behalf too. As for the social types, the lesson from the previous section is that unless the proportion of asocial types is so large that even if all social types acquire education still the burden of coming back to serve the community afterwards is high ($p > p_c$), social types would have been better-off if all of them acquired education and absorbed the absence of the asocial types together as a group. By splitting into cohesive groups of non-educated people on the one hand, and a fraction who become educated on the other hand, the social types are all worse-off: the non-educated could have had higher utility by acquiring education, and the educated could have gained from sharing their burden with all the other social types. In this sense, "acting white" is a shameful waste of human capital. When it comes to high education it is not reasonable to apply policies that eliminate signaling by making this education mandatory. However, if the gains from education will continue to increase, the cost of signaling is bound to increase too, and the model predicts that eventually signaling would stop being individually rational, and consequentially would cease to exist.

6.2 Religious practices

That stronger social ties (i.e., higher levels of cooperation) and religious practice are positively correlated is not new and was empirically demonstrated by Ellison and George (1994). Indeed, generally speaking, people who go to church every Sunday are usually considered to be normative people who possess good features such as sympathy, compassion, concern for others, etc., what we call in this paper "social types". But is it going to church, dressed in their nice Sunday suits, that makes people behave nicely (maybe because they are affected by the reverend's sermon), or is there a different mechanism here? Levy and Razin (2012) develop a model where the mechanism is mainly the former, i.e.,

a religious person goes to church and is then endowed with a belief in reward and punishment, making him a "social type". However, a quite different explanation is possible, in which Sunday prayers are only a costly signal.[29] That is, by attending Sunday prayers and enlisting to the locally active religious congregation, people signal they are trustworthy, and gain the cooperation of other people like themselves. As a result, the social types confine all their sympathy and concern to their congregation members, believing they are social too because they also come to church. This is in line with various allusions in modern culture that people observe the devoutness of their neighbors in order to evaluate their trustworthiness.

This explanation can account not only for the mere existence of religions, but also for the large variety of religious congregations. Even in the US alone one may find Baptists, Anabaptists, Methodists, Evangelists, Presbyterians, Lutheran, Mormons, and many others, including various independent "congregational churches". This observation is in line with our theory, in which groups have to be clearly segregated in order to maintain their cooperative level. For this story to hold, signaling should be a reliable screening device, i.e., it should be costly enough for the "asocial types". One can clearly see why it cannot be a *perfect* screening device, but it is certainly plausible that, on average, and compared to sociopaths, normative people find it easier to conform to religious practices. Note that if religious practice is only a signal, then one can be a believer and not go to church at all. This separation between belief and actual religious practice is well demonstrated by Huber (2005), who finds a large variability in the degree to which religious beliefs are associated with decisions to participate in religious services.

Our model highlights the negative impact of the religious communities' segregation: by joining the other parts of society, the members of these groups could have lifted the proportion of the contributing members of society, thus

[29] To be accurate, Levy and Razin (2012) do incorporate the signaling aspect of religion into their model, but the driving mechanism in their model is that religious people believe in future reward (or punishment), and get utility from their expected reward. This is particularly apparent by the existence of equilibria in their model, in which individuals cooperate in the PD game if and only if they are religious, regardless of their opponent. In these equilibria clearly there is no signaling aspect of religion.

raising the average level of cooperation.[30]

7 Conclusion

The main conclusion of the paper is that a simple and quite intuitive assumption on our social conscientiousness, and more specifically – on the psychological cost of defecting from cooperation with others who wish to cooperate with us, can explain a plethora of prevailing group behaviors. These range from the mere existence of groups, through in-group bias, to costly signaling of sociality and the positive relation between the use of such signaling in a particular group and the cohesiveness of that group, a relation that was demonstrated recently in the lab by Ahn et al (2009).[31] Moreover, quite intuitively, inability to distinguish between social types, who are characterized by such social conscientiousness, and asocial types, who are not, gives rise either to costly signaling or to free riding. The trade-off between the cost of signaling on the one hand, and the cost of having free riders in the group on the other hand, explains why cohesive groups who engage in costly signaling can coexist side by side with mixed groups where no signaling is practiced, but free riding is likely to happen. Of these two costs, it is the signaling cost that is bound to be more harmful from the point of view of society, unless the proportion of asocial types is so large that the mere existence of mixed groups in equilibrium is questionable. Finally, it would be interesting to directly investigate the exact shape of the psychological cost of cheating as a function of the number of cheated partners (e.g., is it fixed, smoothly concave, or is characterized by a "jump"?), possibly in experiments.

[30] For a related but more economically oriented analysis of signaling in the Ultraorthodox Jewish communities see Berman (2000).

[31] The key result in Ahn et al (2009) was that subjects in a restricted entry treatment (where one was incentivized to signal "sociality" or "cooperativeness" in order to be accepted as a group member) achieved substantially higher earnings, due to higher levels of cooperation, than subjects in the other treatments. Note that the total earning of a signaling member could have been lower, depending on the amount he spent on signaling before entering the group. Charness and Yang (2008) report similar evidence using a different mechanism.

8 References

References

[1] Adar, E., and Huberman, B. A. (2000), "Free riding on gnutella," *First Monday*, 5(10).

[2] Ahn, T. K., Isaac, R. M., and Salmon, T. C. (2009), "Coming and going: Experiments on endogenous group sizes for excludable public goods," *Journal of Public Economics*, 93, 336–351.

[3] Aiello, L. C. and Dunbar, R. I. M. (1992), "Neocortex Size, Group Size, and the Evolution of Language," *Current Anthropology*, 34(2), 184-193.

[4] Akerlof, G. A. and Kranton, R. E. (2000), "Economics and Identity", *The Quarterly Journal of Economics*, 115(3), 715-753.

[5] Asvanund, A., Clay, K., Krishnan, R., and Smith, M. D. (2004), "An empirical analysis of network externalities in peer-to-peer music-sharing networks," *Information Systems Research*, 15(2), 155-174.

[6] Austen-Smith, D. and Fryer, R. G. (2005), "An Economic Analysis of "Acting White"," *The Quarterly Journal of Economics*, 120(2), 551-583.

[7] Banerjee, S., Konishi, H., and Sönmez, T. (2001), "Core in a simple coalition formation game," *Social Choice and Welfare*, 18(1), 135-153.

[8] Benabou, R. J. M., and Tirole, J. (2006), "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

[9] Bendor, J., and Mookherjee, D. (1987) "Institutional Structure and the Logic of Ongoing Collective Action," *The American Political Science Review*, 81(1), 129-154.

[10] Bernhard, H., Fischbacher, U., and Fehr, E. (2006), "Parochial Altruism in Humans," *Nature*, 442, 912-915.

[11] Berman, E. (2000), "Sect, Subsidy, and Sacrifice: An Economist's View of Ultra-Orthodox Jews," *The Quarterly Journal of Economics*, 115(3), 905-953.

[12] Birdsell, J. B. (1970), "Local group composition among the Australian Aborigines: A critique of the evidence from fieldwork conducted since 1930," *Current Anthropology*, 11, 115-142.

[13] Bogomolnaia, A., and Jackson, M. O. (2002), "The stability of hedonic

coalition structures," *Games and Economic Behavior*, 38(2), 201-230.

[14] Bonacich, P., Shure, G. H., Kahan, J. P., and Meeker, R. J. (1976), "Cooperation and Group Size in the N-Person Prisoners' Dilemma," *The Journal of Conflict Resolution*, 20(4), 687-706.

[15] Bornhorst, F., Ichino, A., Kirchkamp, O., Schlag, K. H. and Winter, E. (2010), "Similarities and differences when building trust: the role of cultures," *Experimental Economics*, 13(3), 260-283.

[16] Boyd, R. and Richardson, P. J. (1988), "The evolution of reciprocity in sizable groups," *Journal of Theoretical Biology*, 132, 337–356.

[17] Boyer, P. (2001), Religion Explained. Basic Books, New York, NY.

[18] Brewer, M. B. (1979), "In-Group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis," *Psychological Bulletin*, 86(2), 307-324.

[19] Camerer, C. (1988), "Gifts as Economic Signals and Social Symbols," *American Journal of Sociology*, 94, Supplement: S180-S214.

[20] Charness, G. B. and C. Yang (2010), "Endogenous Group Formation and Public Goods Provision: Exclusion, Exit, Mergers, and Redemption," Department of Economics, UC Santa Barbara, Working Paper.

[21] Chen, Y., and Li, Sh. X. (2009), "Group Identity and Social Preferences," *American Economic Review*, 99, 431-457.

[22] Choi, J., and Bowles, S. (2007), "The Coevolution of Parochial Altruism and War," *Science*, 318, 636-640.

[23] de Cremer, D., van Knippenberg, D. L., van Dijk, E., and van Leeuwen, E. (2008), "Cooperating if one's goals are collective-based: Social identification effects in social dilemmas as a function of goal-transformation," *Journal of Applied Social Psychology*, 38(6), 1562–1579.

[24] Demange, G. (2010), "Sharing information in web communities," *Games and Economic Behavior*, 68(2), 580-601.

[25] Dion, K. L. (1973), "Cohesiveness as a determinant of in-group-outgroup bias," *Journal of Personality and Social Psychology*, 28, 163-171.

[26] de Dreu, C. K. W. (2010), "Social value orientation moderates ingroup love but not outgroup hate in competitive intergroup conflict," *Group Processes Intergroup Relations*, 13(6), 701-713.

[27] Dunbar, R. I. M. (1992), "Neocortex size as a constraint on group size in primates," *Journal of Human Evolution*, 22, 469-93.

[28] ——— (1993), "Coevolution of neocortical size, group size and language in humans," *Behavior and Brain Sciences*, 16, 681–735.

[29] Efferson, C., Lalive, R., and Fehr, E. (2008), "The coevolution of cultural groups and ingroup favoritism," *Science*, 321, 1844–1849.

[30] Ellison, C. G., and George, L. K. (1994), "Religious Involvement, Social Ties, and Social Support in a Southeastern Community," *Journal for the Scientific Study of Religion*, 33, 46-61.

[31] Ellison, G. (1994), "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching," *The Review of Economic Studies*, 61(3), 567-588.

[32] Fershtman, C., and Gneezy, U. (2001), "Discrimination in a Segmented Society: An Experimental Approach," *The Quarterly Journal of Economics*, 116(1), 351-377.

[33] Fong, C. M., and Luttmer E. F. P.(2009) "What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty," *American Economic Journal: Applied Economics*, 1(2), 64-87.

[34] Fordham, S., and Ogbu, J. (1986), "Black Students' School Successes: Coping with the Burden of 'Acting White'," *The Urban Review*, 18 (3), 176-206.

[35] Forge, A. (1972), "Normative factors in the settlement size of Neolithic cultivators (New Guinea)," in Man, settlement, and urbanism. Edited by P. Ucko, R. Tringham, and G. Dimbleby, 363-376. London: Duckworth.

[36] Frank, R. H. (1987), "If Homo Economicus Could Choose His Own Utility Function Would He Want One with a Conscience?," *The American Economic Review*, 77(4), 593-604.

[37] Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., and Nowak, M. A. (2012), "Evolution of in-group favoritism," *Scientific reports*, 2.

[38] Gino, F., Norton, M. I., and Ariely, D. (2010), "The Counterfeit Self : The Deceptive Costs of Faking It," *Psychological Science*, 21(5), 712-720.

[39] Gintis, H., Smith, E. A., and Bowles, S. (2001), "Costly signaling and

cooperation," *Journal of Theoretical Biology*, 213, 103-119.

[40] Gneezy, U., Rockenbach, R., and Serra-Garcia, M. (2013), "Measuring lying aversion," *Journal of Economic Behavior & Organization*, 93, 293–300.

[41] Goette L., Huffman, D. and Meier, S. (2006), "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups," *The American Economic Review*, 96(2), 212-216.

[42] Hayden, B. (1987), "Alliances and ritual ecstasy: Human responses to resource stress," *Journal for the Scientific Study of Religion*, 26, 81–91.

[43] Hennig-Schmidt, H., Selten, R., Walkowitz G., and Winter, E. (2009), "Cultural Biases in Bilateral Trust Among Israelis, Palestinians, and Germans," *mimeo*.

[44] Holt, C. A. and Laury, S. K. (2008), "Theoretical Explanations of Treatment Effects in Voluntary Contributions Experiments," Handbook of Experimental Economics Results, Volume 1, Ch. 90, Elsevier B.V.

[45] Huber, J.D. (2005), "Religious belief, religious participation, and social policy attitudes across countries," working paper, Columbia University.

[46] Hurkens, S. and Kartik, N. (2009), "Would I lie to you? On social preferences and lying aversion," *Experimental Economics*, 12(2), 180-192.

[47] lannaccone, L. R. (1992), "Sacrifice and Stigma: Reducing Free-riding in Cults, Communes, and Other Collectives," *Journal of Political Economy*, 100(2), 271-291.

[48] Iannaccone, L. R. (1994), "Why strict churches are strong," *American Journal of Sociology*, 99(5), 1180-1211.

[49] Isaac, R. M., Walker, J. M., and Williams, A. W. (1994) "Group size and the voluntary provision of public goods," *Journal of Public Economics*, 54, 1-36.

[50] Kandori, M. (1992), "Social Norms and Community Enforcement," *The Review of Economic Studies*, 59(1), 63-80.

[51] Kreps, D. M., and Wilson, R. (1982), "Reputation and Imperfect Information," *Journal of Economic Theory*, 27, 253-279.

[52] Ledyard, J. O. (1995), "Public Goods: A Survey of Experimental Re-

search". In: Kagel, J., Roth, A. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, NJ.

[53] Levy, G., and Razin, R. (2012), "Religious beliefs, religious participation, and cooperation," *American economic journal: microeconomics*, 4(3), 121-151.

[54] Lopez-Perez, R., (2008), "Aversion to norm-breaking: A model," *Games and Economic Behavior*, 64, 237–267.

[55] Lopez-Perez, R., (2012), "The power of words: A model of honesty and fairness," *Journal of Economic Psychology*, 33, 642–658.

[56] Lundquist, T., Ellingsen, T., Gribbe, E. and Johannesson, M. (2009), "The Aversion to Lying," *Journal of Economic Behavior and Organization*, 70(1), 81-92.

[57] Miettinen, T., and Suetens, S., (2008), "Communication and guilt in a Prisoner's dilemma," *Journal of Conflict Resolution*, 52, 945–960.

[58] Milgrom, P., and Roberts, J. (1982), "Predation, Reputation and Entry Deterrence," *Journal of Economic Theory*, 27, 280-312.

[59] Naroll, R. (1956), "A preliminary index of social development," *American Anthropologist*, 58, 687-715.

[60] Olson, M.(1965), "The Logic of Collective Action". Cambridge: Harvard University Press.

[61] Palfrey, T. R., and Rosenthal, H. (1988), "Private incentives in social dilemmas: The effects of incomplete information and altruism," *Journal of Public Economics*, 35(3), 309-332.

[62] Ruffle, B. J., and Sosis, R. (2006), "Cooperation and the In-Group-Out-Group Bias: A Field Test on Israeli Kibbutz Members and City Residents," *Journal of Economic Behavior and Organization*, 60(2), 147-163.

[63] Ruffle, B. J., and Sosis, R. (2007), "Does it Pay to Pray? Costly Ritual and Cooperation," *The B.E. Journal of Economic Analysis & Policy*, 7(1), Article 18.

[64] Sawaguchi, T., and Kudo, H. (1990), "Neocortical development and social structure in primates," *Primates*, 31, 283-290.

[65] Service, E. R. (1962), "Primitive social organization: An evolutionary perspective." New York: Random House.

[66] Smith, E. A., Bliege Bird, R. L., and Bird, D. W. (2003), "The benefits of costly signalling: Meriam turtle hunters," *Behavioral Ecology*, 14, 116-126.

[67] Shayo, M, and Zussman, A. (2011), "Judicial Ingroup Bias in the Shadow of Terrorism," *The Quarterly Journal of Economics*, Vol.126(3), 1447-1484.

[68] Sosis, R. H., and Ruffle, B. J. (2003), "Religious Ritual and Cooperation: Testing for a Relationship on Israeli Religious and Secular Kibbutzim," *Current Anthropology*, 44(5), 713-722.

[69] Spence, A. M. (1974), "Market Signalling". Harvard University Press.

[70] Steward, J. H. (1955), "Theory of culture change: The methodology of multilinear evolution". Urbana: University of Illinois Press.

[71] Suzuki, S., and Akiyama, E.(2005), "Reputation and the evolution of cooperation in sizable groups," *Proceeding of the Royal Society B*, 272, 1373–1377.

[72] Tajfel, H. (1970), "Experiments in intergroup discrimination," *Scientific American*, 223, pp.96-102.

[73] Tajfel, H., Billig, M. G., Bundy, R. P., and Flament, C. (1971), "Social categorization and intergroup behavior," *European Journal of Social Psychology*, 1, 149-178.

[74] Turner, V. (1969), The ritual process. Chicago: Aldine.

[75] Wilson, D. (2002), Darwin's cathedral: Evolution, religion, and the nature of society. Chicago: University of Chicago Press.

[76] Wilson, W., and Kayatani, M. (1968), "Intergroup attitudes and strategies in games between opponents of the same or of a different race," Journal of Personality and Social Psychology, 9, 24-30.

9 Appendix

9.1 Proof of Lemma 1

First assume by negation that there are at least two solutions to equation $t(K) = Kg$ in the interval $]0, 1[$, denoted by K_1 and K_2. Since the conditions

on $\lim_{k\to 0} t'(k)$ and $\lim_{k\to 0^+} t(k)$ imply that for $\varepsilon \to 0^+$ we have $t(\varepsilon) > \varepsilon g$, we get that

$$\left[1 - \frac{K_1 - \varepsilon}{K_2 - \varepsilon}\right] t(\varepsilon) + \frac{K_1 - \varepsilon}{K_2 - \varepsilon} t(K_2) > \left[1 - \frac{K_1 - \varepsilon}{K_2 - \varepsilon}\right] \varepsilon g + \frac{K_1 - \varepsilon}{K_2 - \varepsilon} K_2 g = K_1 g,$$

while the concavity of $t(\cdot)$ implies that

$$\left[1 - \frac{K_1 - \varepsilon}{K_2 - \varepsilon}\right] t(\varepsilon) + \frac{K_1 - \varepsilon}{K_2 - \varepsilon} t(K_2) \leq t(K_1),$$

which contradicts the assumption that K_1 solves the equation $t(K) = Kg$. Next, note that $t(K) - Kg$ is strictly positive at $K = \varepsilon$, strictly negative at $K = 1$ (by assumption), and any possible discontinuity in between is an increase. Thus $t(K) - Kg = 0$ at least once in the range $[\varepsilon, 1]$. Finally, since we have shown that $t(K) - Kg = 0$ *exactly* once in the range $[\varepsilon, 1]$, it follows that $t(K) - Kg$ changes signs from positive to negative only once in $[\varepsilon, 1]$, and so $t(K) > Kg$ for every $K \in \left]0, \bar{K}\right[$ while $t(K) < Kg$ for every $K \in \left]\bar{K}, 1\right[$.

9.2 Proofs for discrete N

Let the number of individuals be an integer (instead of a continuum), with N individuals in society. We show here that the results of Section 4 hold in this case too, where the main difference is that the realization of K partners does not necessarily consist of exactly $(1-p)K$ social types. In particular, it is natural to assume now that each individual is randomly assigned a type, with probability p to be assigned the asocial type. The move from a continuum to a discrete number of individuals requires replacing the condition $t(1) < g$, which ensured that the material payoff from cheating the whole society exceeded the psychological cost, with the parallel condition $\lim_{k\to\infty} t(k) - kg < 0$.

Proposition 13 *If $p \leq \frac{t(1)-g}{t(1)+l-g}$, then there exists a unique integer $K_p \in [1, \bar{K} - 1]$ such that a mixed group of size $K+1$ is sustainable in equilibrium if and only if $K \leq K_p$. Furthermore, K_p is decreasing in p.*

The proof of the proposition follows the next lemma.

Lemma 14 *Let $h(x)$ be an increasing and concave function defined for $x \geq 0$ with $h(0) = 0$. If $x \sim Bin(n, p)$, then:*

1. *Given a fixed $p \in [0, 1]$, $E_n h(x)$ is increasing and concave in n.*
2. *Given a fixed $n > 0$, $E_n h(x)$ is increasing in p.*

Proof. *(1) That $E_n h(x)$ is increasing in n is clear from the fact that*

$$E_{n+1}h(x) = pE_n h(x+1) + (1-p)E_n h(x),$$

and $h(x+1) \geq h(x)$. For proving concavity, we can write

$$E_{n+2}h(x) = (1-p)^2 E_n h(x) + 2p(1-p)E_n h(x+1) + p^2 E_n h(x+2).$$

Then we need to show that $E_{n+2}h(x) + E_n h(x) \leq 2E_{n+1}h(x)$. Substituting the above expressions in this inequality, it boils down to showing that $p^2 E_n h(x) + p^2 E_n h(x+2) \leq 2p^2 E_n h(x+1)$, which indeed holds by the concavity of $h(x)$ and the linearity of the expectation operator. (2) We will prove by induction. For $n = 1$ the inequality holds: if $x \sim Bin(n, p)$ and $y \sim Bin(n, q)$ with $q > p$, then $E_1 h(y) = qh(1) \geq ph(1) = E_1 h(x)$. Assume now that the inequality holds also for some n, so that $E_n h(y) \geq E_n h(x)$. Then

$$\begin{aligned} E_{n+1}h(y) &= qE_n h(y+1) + (1-q)E_n h(y) \\ &\geq pE_n h(x+1) + (1-p)E_n h(x) = E_{n+1}h(x), \end{aligned}$$

which completes the proof by induction. ■

Proof of Proposition 13

The proof for asocial types is the same as in the proof of Proposition 6. As for the social types, consider an individual of type s who is a member of a mixed group of size K. Defecting against any $k \leq K$ of them, of which $X \in [0, k]$ are of type s, would result in an increase in expected material payoff of $Xg + (k - X)\, l$, but the expected total payoff would also decrease by $t(X)$ due to the cost of cheating. Since $X \sim Bin(k, 1-p)$, the individual would

have no profitable deviation if and only if $E_k[t(X)] \geq E[Xg + (k-X)l] = k[(1-p)g + pl]$ for every $k \leq K$. Let $\Delta(k,p) \equiv E_k[t(X)] - k[(1-p)g + pl]$. The conditions on $t(k)$ and on the payoffs of the game imply that for any given $p \in (0,1)$, we have $\Delta(0,p) = 0$ and $\lim_{k\to\infty} \Delta(k,p) < \lim_{k\to\infty} t(k) - kg < 0$. From Lemma 14 part (1) we know that $E_k[t(X)]$ is concave in k, and therefore so is $\Delta(k,p)$. It can be verified that if $p \leq \frac{t(1)-g}{t(1)+l-g}$ then $\Delta(1,p) \geq 0$, in which case $K_p \geq 1$ is the floor of K_p^*, the unique strictly positive solution to the equation $\Delta(k,p) = 0$. Moreover, $\Delta(k,p) > 0$ for every $k < K_p^*$, and $\Delta(k,p) < 0$ for every $k > K_p^*$, which proves that mixed groups of size K are sustainable if and only if $K \leq K_p$. Furthermore, $K_p < \bar{K}$ because $l > g$ and $E_k[t(X)] < t(k)$, and so $\Delta\left(\bar{K},p\right) < t(\bar{K}) - \bar{K}g = 0$. Next, from Lemma 14 part (2), we get that $E_k[t(X)]$ is decreasing in p for a fixed value of k, and so $\Delta(k,p)$ is also decreasing in p for any $k > 0$ (remembering that $l > g$ and so $[(1-p)g + pl]$ is increasing in p). That is, for any $\{p, q | p < q\}$ we have $\Delta(k,q) < 0$ for every $k \geq K_p^*$, which in turn implies that $K_q^* < K_p^*$ and so $K_q \leq K_p$, i.e., K_p is (weakly) decreasing in p.