PUNISHMENT WITHOUT CRIME: A TALE OF COOPERATION AND COMPETITION IN PUBLIC GOODS GAMES

Alexis Belianin*

First draft: July 2011 This draft: March 2013

Abstract

Punishment is known to be one of the major factor of cooperation in the public goods (PG) games. However, the exact nature and reasons why people punish each other to a large extent remains unexplored. In this work we study the punishment strategies in a systematic way, disentangling several possible explanations for punishing behaviour, including competitive, emotional and preemptive motives, alongside with availability and tolerance towards punishment. We set and ran a series of experiments in different regions of Russia, which establishes that actual disapproval of others' contirbutions is the determinant of punishment in a minority of cases. Using a structural statistical model, we offer a classification of behavioural strategies of the punishers for our sample, as well as in cross-regional perspective. This analysis establishes that, besides ethical considerations, willingness to outperform other players in the group, and precautionary punishment in anticipation of the punishment from the other player, play a major role in determination of the direction and size of spiteful punishments.

JEL codes: C92, C72, C31

Keywords:public goods games, punishment, spite, competition, preemption, insurance, structural experimetrics

*International College of Economics and Finance, Higher School of Economics, Moscow, Russia. e-mail: icef-research@hse.ru. I am thankful to Yulia Safarbakova and Diana Kolesnikova for excellent research assistance, and to Fuad Aleskerov, Daniel Houser, Benedikt Herrmann and Marco Novarese for helpful comments. All remaining errors are my own.

1 Introduction

Since the seminal paper by Ernst Fehr and Simon Gächter (2000), punishment in public goods (PG) games is one of the key research area in experimental economics. The voluntary costly punishment (VCP) mechanism, introduced into the PG games in the late 1980s–early 1990s (Yamagishi, 1986; Ostrom, Walker and Gardner, 1992), gave rise to a broad literature, primarily because of its cooperation-sustaining properties, but soon became a topic on its own. Recent behavioural literature (Fehr and Schmidt, 1999; 2003; Bolton and Ockenfels, 2000; Falk and Fischbacher, 2004; Dufwenberg and Kirchsteiger, 2006, to name a few) suggests natural rationalization for punishment in terms of preferences towards fairness, reciprocity and equity in social interactions. Following this track, punishment in the PG context is most often viewed as a mean to express disapproval of low contributions of players who contribute relatively less from the part of those players who contribute relatively more. This *altruistic punishment* (Fehr and Gächter, 2002) is costly to the latter players, and is thus interpreted as altruistic expression of social preferences (Fischbacher and Gaechter, 2010).

Our paper belongs to the strand in literature which challenges this view. We argue that behaviour which is interpreted as 'punishing' may correspond to quite different motives, some of which have nothing to deal with social attitudes, but are induced by the experimental design (Zizzo, 2010, 2011) or competitive intentions (Houser and Xiao, 2010). To disentangle these, we set up and run one of the most elaborated experiment which is explicitly focused on punishment motives in a single-stage punishment game. To facilitate comparison, we start by closely replicating the design of Herrmann and Gächter (2006, henceforth abbreviated as GH), including the country where the data has been collected (Russia), but extending it in a number of directions: including elicitation of prior beliefs, different punishment costs, possibilities to insure against such deductions, reassign these amounts to different players, and ex post questionnaires asking for motives of their usage. This broader perspective not only allows to separate various motives for deductions, but also explain some bits of experimental evidence which are inconsistent with the theoretical picture of prosocially-motivated punishment. Specifically, it explains co-existence of altruistic and *spiteful punishment* (Herrmann e.a., 2008; 2011; Saijo, 2008), or costly deductions of payoffs of those players who contributed *more* than the punisher[1]. In the conventional behavioural framework, this last phenomenon is interpreted as disapproval of prosocial behavour clustered in several regions of the world — specifically, Middle East and Eastern Europe (Russia and Ukraine), which in the literature usually receives cultural explanation (Cason e.a., 2002; Herrmann e.a., 2008; Herrmann and Thöni, 2009).

As a result of our experimental treatment we are able to segregate individual behaviour into different motives, and figure out that for prosocial punishers, the main motive is ethical (disapproval of antisocial/noncooperative behaviour), while the second important motive among them is fear of being punished without an opportunity to respond — a motive we call 'preemptive' punishment. For spiteful people, this second motive is also present, and is actually the main one, while the second is not disapproval of cooperative behavior, but willingness

[1]This notion is not to be mixed with 'spiteful contributions' introduced by Cason e.a. (2004) in a different context of sequential contibutions in PG games with interior equilibria

to outperform the other players in a group in terms of overall gain. We call this motive 'competitive', and in our sample, it accounts for about 1/3 of all spiteful punishments. A contribution of our paper is that we calibrate and assess the validity of these terms by three complementary methods: preferences revealed through experimental institution, solicitation of intentions through survey questionnaire, and structural experimetric model estimating the probabilistic predominance of each motive by means of a latent class model.

The rest of the paper is organized as follows. Section 2 recaps the background for the public goods game with and without punishment, together with its main theoreitcal properties. In section 3 we review the related literature, and describe the connection of our work with the previous works. Section 4 describes our experimental setup, and section 5 contains major results. Section 6 is devoted to the development and estimation of a behavioural model, which analysis leads to the posterior taxonomy of the factors of punishment, quite different from the prior. Finally, section 7 concludes by indicating the implications of the present and directions for further research.

2 The Public Goods Game with punishment

The classical linear PG voluntary contributions mechanism (VCM) is implemented in groups of $n \geq 2$ players endowed with w experimental currency units (points) per period each. The game consists of a number of periods, in which every player i independently of the others bid any integer amount $c_i, 0 \geq c_i \geq w$ she wishes to the public account, and retains the rest $(w - c_i)$. Each retained point contributes one to the final utility of that *individual*, while each unit deposited on public account is an increasing linear function of the number of points deposited by the *entire group*, $k \cdot \sum_i c_i = \alpha \bar{c}$, where $\bar{c} = \frac{\sum_i c_i}{n}$ is average contribution of the group and $\alpha = kn, k < 1 < kn$. Possible revenues from the public account are available to the participants as a table which shows the worth of public good for any amount contributed by the group. Expected value of individual contributing c_i given the average contribution $\bar{c}$ is thus given by

$$v_i(c_i, \bar{c}) = w - c_i + \alpha \bar{c} = w - c_i + k \cdot \sum_i c_i \tag{1}$$

written as a function of observables to the player; and the participants are communicated the amount contributed to the public good after all contribution decisions were made. Since $1 < \alpha$, the socially efficient outcome is to contribute everything to the public account. However, $k < 1$ implies that the game has a prisoners' dilemma structure: any individual is better-off depositing nothing on that account in a single-period version of this game. A unique Nash equilibrium in this game is free-riding, which is also a unique subgame perfect equilibrium in any finitely repeated game, where backward induction stipulates non-cooperative behaviour at every stage game[2].

Since Fehr and Gächter (1999), a typical punishment mechanism works as follows. After the contribution stage, all players are communicated not only the sum of contributions, $\sum_i c_i$, but the entire contributions vector, $\mathbf{c} = [c_1, \ldots c_n]$,

[2]Infinitely repeated version of the game, of course, admits other solutions via the Folk theorem. Cason e.a. (2004) is an example of an experimental design with interior equilibrium contributions.

and thus everyone's contribution (without, of course, identity of the contributing player) is known to all participants. Each player i then can punish each other player j (not herself!) by p_{ij} units, which expenditure decreases gain v_i of the punished player sp_{ij} units, where $s < 1$ i.e. punishment is less costly to the punisher than to the punished player. Given the punishment matrix $\mathbf{P} = [p_{ij}]$, each player knows the row vector P_i (her own punishment) and learns the total punishment imposed on that player by the others, $\sum_{j \neq i} p_{ji}$. Total payoff to player i is then

$$V_i(\mathbf{c}, \mathbf{P}) = v_i - s \sum_{j \neq i} p_{ij} - \sum_{j \neq i} p_{ji} \tag{2}$$

where again, the arguments of the V_i function contain only the observables — in particular, the players do not know the identity of those who punished them. This typical payoff structure obviously makes punishment socially suboptimal in the short run: it results in decrease of of punishment an efficient tool to maximize the difference between one's post-contribution payoff and the payoff of the punished player. This strategy as such offers incentives to a strategically-minded individual motivated by competitiveness.

2.1 Reasons for punishment

Starting from Fehr and Gächter, traditional literature attributes punishment to prosocial attitudes, which are hypothesizes to pre-exist in the mind of the experimental subjects. In reality, however, subjects who enter the experimental lab act in an environment whose significance and meaning need not necessarily coincide with that of the experimenter. In particular, it is farily possible that the experimental institution is not sufficiently salient (Smith, 1982) to trigger payoff-maximizing behaviour in the public goods game. In such cases, which are likely to be unobservable, any limited experimental remuneration in the range of 10-100 euro is not sufficient to seek maximization of experimental gain, if the amount at stake is of order of units, not hundreds of euro. This feature might be common to quite a few of experiments, and are largerly beyond experimenter's control. A related issue refers to real incentives to contribute and punish. Game-theoretic solutions, bearing on neoclassical precepts, tacitly assume that subjects should maximize their own payoffs notwithstanding the payoffs of the others — hence costly punishment, which is a pure deduction of the punisher's wealth, should not have been observed at first instance. But, once it exists, it reveals failure of this payoff-maximizing model — in particular, it offers a nice way to improve one's standing in the game relative to the other players. And indeed, in literally every experimental session, at least one subject asked the experimenter: who gained the most in our group? For methodological reasons, this information was not revealed to them; however, this interest is telling in itself. To some extent, this feature may be mitigated by the size of experimental reward — however, marginal rate of substitution between material gain and such preferences is unobservable as well, and can be inferred only indirectly, and using a broader model of decision-making in experimental game, which we develop later on.

To sum up, real experimental subjects might have a large variety of punishment motives, which can be classified and summarized as follows:

Availability — presense of punishment option may be suggestive by itself: once this option is available, some people might try it just out of interest, inasmuch as the cost of punishment is small in real terms. We call it the *Chekhov motive*, remnicieint of the famous sentence by Anton Chekhov, a Russian novelist and playwriter, who reportedly said that, 'if in the first scene of the play, there is a gun on the wall, by the third scene it mush shoot'. By the same logic, players may reason that if the punishment option is there, it must be used somehow. To test whether this explanation is viable, one should *compare punishments when this option is always available to punishments when the player has to purposedly switch it on before making use of it.* For thak sake, in our design we let our subjects who want to punish somebody to explicitly make this option available to themselves, to check if this results in substantial decline in punishment relatively to the other experiments. Moreover, availability of punishment as real alternative may also depend on the size of punishment cost: the lower it is, the more people may consider bearing it. In view of that, half of our sessions were conducted under high-cost conditions, with $p_{ij} = 0.5$, and the other half, under low-cost, with $p_{ij} = 0.1$ per unit punishment.

Preemption — another explanation sometimes raised in the literature suggests that people may punish because of expected punishments from the others, or the *Brodsky motive* who once, somewhat metaphysically, said 'A man is more frightening than its skeleton'[3]. This motive draws on the unwillingness of some people to tolerate someone's penalties imposing on you for social reasons, as seems to be the case in some countries of the Arab world, or simply by force of the personal temper. This motive, albeit mentioned by Herrmann and Gächter (2009) in their explanation for spiteful punishment in Saudi Arabia and Oman, has not been systematically explored so far. If this explanation is correct, *we would not observe punishment if the subjects could defend themselves against punishments in other ways.* To separate these motives,we introduce alternative method of preemptive behaviour — namely, insurance against punishments. After knowing contributions and making punishment decisions, but before knowing whether punishments have been applied to them or not, subjects in our low-cost treatments were introduced to the possibility to buy insurance policies against possible punishments from the other players[4]. Punishments were individual, unexpensive, could be bought up to to the maximum possible punishment, and work like a conventional insurance policy: if player i insures against punishment of player j of size r_{ij}, then any punishment $p_{ji} \leq r_{ij}$ is not applied to player i, while punishment $p_{ji} > r_{ij}$ does, with player j still bears the full cost of punishment. Insurance in our design could be paid for in two forms: 1) transfer of the money spent on punishment to insurance — that is, instead of punishing other players, our subjects could relocate the money they have previously

[3]Joseph Brodsky, one of the best Russian poets of the XXth century, was convicted in the 1960s in USSR for allegations of 'social parasitism', notwithstanding his poetic work has already earned him a reputation of one of the best poets of his generation. In the 1970s, he was sent out of the USSR, and died in the US in 1986 as Nobel Laureate and Poet Laureate of America.

[4]Subjects were not aware of the availability of that option until after all punishment decisions have been made — see the instructions in Appendix 1.

spent on punishment for insurance, parting with the right to punish; or 2) procurement of insurance with extra money, while maintaining the punishment option; any linear combination of these two options was also allowed, as described later. Under these settings, a player who made punishment decision for preemptive reasons, but was not motivated otherwise, should be willing to swop the resources he or she spent on punishment to buy protection, and possibly adding some more.

By contrast, player who were intrinsically motivated to apply punishment per se, could or could not purchase insurance using additional money, but would in either case maintain punishment as an active option. One can think of several motives of that kind:

Contention — for some people, clash with others per se may be source of joy and utility, just like gambling or extreme sports. Punishment is a method of clash premitted within the experimental institution, and people in some countries (like Russia) may arguably be culturally accustomed to them[5]. Accordingly, in some cultures at least, costly clashes may be viewed not as something extraordinary, but as 'customary' and 'acceptable' thing, promoting tolerant attitude towards punishment itself. We term this the *Tjutchev motive*, after a famous Russian poet and diplomat of the XIXth century, who somewhat aphoristically mentioned that 'The entire Russian history before Peter the Great is an entire commemoration service, and after Peter the Great — an entire criminal case'. Contentious players should enjoy clashes, hence, unlike preemptive ones, they should maintain the punishment option even in the presence of opportunity to insure.

Competitiveness — similarly to previous motive, this one stipulates that people derive utility from punishment, but this time not from the fact of fight, but from the desire to win it. Given per unit cost, $p_{ij} < 1$, punishment as an efficient method to improve one's relative standing in the group, or 'outperform' the other players. Some experimental subjects seem to be overly interested not in their own material gains, but in their performance relative to the other players. In literally every session after this, and many other experiments,one or several participants are asking the same question: who won most in our game?[6] Such questions seem to reveal more than intellectual curiosity: some competitively-minded subjects may be really driven by this strive to take over the others, even if this comes at their expense. This may be called *Dostoyevsky motive*, following the quest of his hero, Rodion Raskol'nikov: 'Am I a trembling beast, or I daresay?'. A person with this motivation *would never be willing to improve relative standing of the other players, such as relocate money in favour of them.*

[5]For instance, collective boxing exercises, or *fist figthings* were customary and widespread among the Russians for ages (http://en.wikipedia.org/wiki/Russian_fist_fighting)

[6]Thus, some of our subjects in Moscow study at the most prestigious faculty of economics in the city (ICEF, http://icef.hse.ru), and come from fairly wealthy background, which entitles them for personal disposable income (net of rental costs) of several thousand euro per month. In such cases, even experimental rewards of order of 50 euro is next to negligible, so such participants may well have reasons to maximize not their own material payoff, but subjective utility of participation in an experimental session, e.g. by gaining more than the other people. This logic seems to be rather general, for individual disposable incomes are most often, and marginal rates of substitution between material payoff and subjective utility — always unobservable to the experimenter.

To test for this motive, in our sessions with low punishment cost, following insurance decisions, and again without knowing that ex ante, punishers had to decide how to dispose of the amount of their punishment, if any: at their discretion, it can be either 1) burned off the punished subject's account (if not insured against, which is unobservable for the punisher anyway), or 2) redistributed among the players other than the punisher and the punished. A person who is competitive should never relocate money to other players, as this will improve their relative standing, while contentious player might; and both of them should have had punishment option in use in the presence of insurance.

Retaliation — negative feeling at what the others have contributed, whatever is the specific cause or reason. The most typical reason is of coruse dissatisfaction due to low level of contribution of the punished player, which causes justified desire to revenge. This feeling may be called also the *Pushkin motive*[7]. This classical motive can be manifested in several ways, and be either prosocial (punishment of the player who contributed less than you did), or spiteful (punishment of the more generous partner), with the canonical interpretations of punishment due to disapproval of behaviour that is not 'sufficiently cooperative' or 'overly cooperative', respectively. Yet even these intentional motivations may be driven by several comparisons, in either prosocial or spiteful directions:

Congruence in contributions $c_i - c_j$ — comparison of contribution fo the other player (j) to that of the punisher (i). In terms of Fehr and Gaechter (1999), this is the most natural justification for punishment, based on the *absolute* deviations of the punisher's contribution from those of the punished player.

Conformity of individual behaviour to the group $c_j - \bar{c}$ — difference between contributions of the other player and factual average contribution in the group (Carpenter, 2004). In this case, the punisher has no strong predispositions as to what should the average contributions look like, but believes it is bad to look too different from the group standard.

Conformability to the norm $c_j - \hat{c}$ — difference between factual contribution of the other player and due average contribution, or what the punisher believes one ought to contribute. If this is the motive for anger, one might expect prosocial punishments if this difference is negative, and spiteful when it is positive.

[7] After a masterpiece of the greatest Russian poet based on an old legend about death of the second variag prince of the Xth century Russia:
'Like now is going Prophetic Oleg
To revenge irrational khazars
He made their towns, for bloody attack
Subjected to different hazards
(transl. A.Artemov)'.
Prince Oleg, son of Rurik of Kiev Russia, was reported to be a brave warlord, who not only successfully defended his lands against nomad tribes, but also undertook a victorious assault on the Bysantine empire, imposed contrbution on its emperor, and put his shield on the gate of Constantinoples. Legend says he died of a serpent's bite, hiding in the skull of his past horse — a story that Pushkin laid on poetry.

Many of these possibilities have been noticed in the literature from the very beginning (see, e.g. Fehr and Gächter (2000), p.990, footnote 10) — yet they have not been systematically explored hitherto. In our experiment, we explicitly control for the previous motives, and attribute punishment to retaliation in the residual sense, further corroborating it with other supporting evidence, such as subjects' ex post questionnaires and statistical analysis using a structural model with individual heterogeneity through observable and latent controls.

3 Related literature

Since Fehr and Gächter (2000), punishment in PG context has been studied by many authors, including Page e.a. (2005), Nikiforakis (2008; 2010), Bochet et.al. (2006), Carpenter (2007a, 2007b), Chaudhuri (2011), Herrmann e.a. (2008), Sefton e.a. (2007); Houser and Kurzban (2002), Xiao and Houser (2010), including some works in medical and biological studies (Fowler, 2005; de Quervain e.a., 2008; Marlowe e.a., 2011). A survey by Herrmann and Gäachter (2009) is available.

As theoretical background, contemporary economists use existing behavioural literature is (largely implicitly) drawn on the model of maximization of the extended utility function including psychological phenomena (Croson, 2007; Masclet e.a., 2003), such as inequity aversion (Fehr and Schmidt, 1999), reciprocity (Dufwenberg and Kirschteiger, 2004) or conditional cooperation (Fischbacher and Gächer, 2011), up to the point that even critique of the existing approaches is based on them (Casari, 2005). Further critique came from a few authors who study the conventional PG games under the assumption that players might misinterpret the model (Houser and Kurzban, 2002; Ferraro and Vossler, 2008). In a pioneering work of that kind, Andreoni (1995) hypothecized that players might simply misunderstand the rules of the game and/or be motivated by their relative standing. To capture these effects, Andreoni compares contributions to PG to ordinal game in which players are rewarded lump-sum depending on their final revenue standing; Houser and Kurzban, (2002) and Ferraro and Vossler (2008) extend this idea to game against computers (see Bardsley (2000) for a similar empirical implementation).

Neoclassical approach has been implemented by Carpenter (2007a, b) who studied demand for punishment as normal good. Cason e.a. (2004) study two-players contribution games with payoff nonlinear in each contributions x, y given by a Cobb-Douglas production function of a form $u(x, y) = f(x^{\alpha} \cdot y^{\beta})$, which yields interior solution. Saijo and Yamato (1999) study a variant of this game with precommitment at first stage, which defines a game in strategic form with two asymmetric and one mixed strategy equilibrium (of the Hawk-Dove type). A combination of both features in an experiment yielded interesting punishment result: players who face an opponent who committed to contribute nothing tended to undercontribute relatively to the first-best interior solution (which behaviour the authors call 'spiteful'), giving lower revenue to the deviating opponent.

Punishment, of course, is not the only disciplining device. In a series of interesting works, Talbot Page and Louis Putterman with coauthors investigate the role of distributed power as disciplining device. Cinyabuguma e.a. (2005) show that cooperation increases if instead of punishment players could decide

themselves if there are two groups (big one, of 16 players and more lucrative, and less lucrative), and the big group may expel a particular player who moves to the other group. Kamei e.a. (2011) and Putterman e.a. (2011) in recent papers show that moral pressure can be no more powerful than material one. In related works Cinyabuguma e.a. (2006) introduce the secondary effect of punishment by allowing players to punish those who have punished, and find no significant effect on contributions, but some variety over punishment choices. Availability was expored in an endogenous context: Ertan e.a. (2009) have used 3- and 5-times voting on whether to allow or disallow altruistic or spiteful punishment, and the sanctioning mechanism was allowed only if the majority of 4 players approves it. The authors found that low-but-not-high contributors punishment rule prevails. Similarly, Casari and Luini (2005) in an unpublished paper using a sample of 240 Italian students have obtained higher contribution rates under majority voting and sequential voting procedures. These experiments, however, cannot serve as pure tests of individual intentions to punish, as punishment decisions depend on social approval, which is known to the subjects from the outset, and hence might affect their choice of punishment rationale. Further, the voting mechanism may not be robust against different trading rules.

Our design is related to this and other literature in several respects. Solicitation of expected values is common (Gächter and Renner (2010), whereas redistribution of punishment proceedings has been realized by Page e.a. (2008), who found that this mechanism generates larger contributions, presumably because players believe they are going to be rewarded with the punishment money. The introduction of insurance to VCP games appears to be new.

Finally, our empirical model is closest in techniques, if not in spirit, to Brandts and Schram (2001) and Bardsley and Moffatt (2007). Both papers present a structural model of strategy choice, but they are directed towards contributions' motives, whereas our main interest is in punishment.

4 Experimental setup

As a starting point, we have used the design of Gächter and Herrmann (2006), and conducted an experiment consisting of only two games: PG without punishment, followed by PG with punishment and some further actions which were announced, but whose contents was not known to the subjects before they actually were about to be done. Inasmuch as we are interested in factors of punishment per se, we used only this (direct) sequence of the games, wherein the first PG essentially worked as training for just one punishment game, to avoid any reputational/learning effects, and treat punishments as independent events. The same experimental currency was used to denote gains and punishments, worded in neutral terms as 'deductions' (see Appendix 1 for full set of instructions). For reasons to be clarified in a while, we have adopted the experimental setup with groups of 4 players, and efficiency factor $k = 1.6$ (i.e. value of $\alpha = 0.4$ for all participants). As treatment variable, we varied the cost of punishment, which was be either low (0.1) or high (0.5) per each unit of punishment, which was itself limited in size to closed interval from 0 to 10 units.

Upon arrival and signing the consent form, the subjects were introduced into the PG framework without punishment, explained in details with worked

examples and followed by calculation exercises. Experimenters checked all solutions to make sure the subjects do not proceed with the experiment until these solutions are absolutely clear to them. Prior to the actual game, subject were to complete a pre-session questionnaire in which we ask for their *planned* contirubtions to the public good, the *due* average and *expected* average contributions in their group, as well as their desired contribution level if the group average turns out to take discrete values of 0, 3, 6, 10, 14 and 17 units. This evaluation replicates the strategy method, except that it it is payoff immaterial, and bears no consequences for actual decisions. From the beginning, subjects were aware that there will be two games in the experiment, and that the matching protocol will be partner.

A single stage of the PG game without punishment followed this, after which the subjects were reported their own contribution and profit, as well as contributions and profits of the other participants, and average contributions (not profits) in their group. Following this, participants were introduced to the punishment treatment (called 'deduction') of the post-contribution profits of each other at the second game. It was especially stressed that reasons to assign deduction points to anyone can be whatever, and are entirely left at the participants' discretion. This instruction was again followed by worked examples and calculation exercise whose correctness was checked by the experimenters prior the game. Participants were also alerted that following these decisions, there will be some more actions at the end of that stage, but they were not told what it will consist of. At the end of this game, they were again informed of the contributions and profits of all individual players and average contributions in their group.

Time flow of the second game is displayed in Figure 6.3 in the Appendix. The punishment stage following this display began by the switching screen with two buttons. If the participant was unwilling to assign deduction points to anyone in his or her group, he should choose 'No' (the top button), and was shown the waiting screen at the next stage. By contrast, if he wanted to assign deduction points to at least one of his or her group fellows, he was to say 'Yes', and was given move at the next screen, where he could assign any legitimate number of deduction points to his group members.[8] On the right side of the screen, subjects could see the reminder of individual and group average contributions and profits in this period. Subjects had a chance to skip punishing anyone by putting 0 in the respective placeholder.

Folloiwng this stage, insurance option was introduced: subjects in the low-cost punishment treatment could purchase insurance against punishment from each particular player in her group. It is important to stress that insurance as an option was not mentioned before, so no subject could get an idea of this opportunity at the time of contribution or punishment decisions. Insurance can be in size from 0 (no insurance) to 10 (full insurance at a maximum level of possible punishment), so that purchase of r_{ij} units of insurance of player i from possible punishment of player j reduces this punishment from p_{ji} to $\max(p_{ji} - r_{ij}, 0)$. Insurance can be bought by all participants, including those who have chosen did not punish at all, at a cost of 0.2 per unit of insurance. Those who did

[8] In that screen, all four group members were shown, and the number of this particular player was displayed at the top. This was done as an additional control for rationality: subjects who have mistakenly assigned deduction points to themselves were excluded from the data.

punish could procure insurance in two ways: using additional money, at a cost of 0.2 (i.e. in the same way as non-punishers), or by reallocating the money they have used to punish other players to the cause of insurance, at the cost of punishment of 0.1[9]. Punishing subjects could have relocated to insurance less than the full amount they have spent on punishment (in which case the rest of their punishment expenditure went on for punishment), or insure by more than the amount they have punished (in which case the cost of insurance up to the value of punishment was 0.1, and the cost of additional units of insurance — 0.2). All this was explained to the subjects prior the beginning of that part of the experiment, and illustrated in examples. Subjects have received an opportunity to ask any number of clarifying questions before proceeding to that stage. Technically, all insurance tasks were collected in one screen, wherein subjects saw total contributions, and have had to indicate their individual insurance decisions (amount of insurance purchased), and tick the funding source (relocation of resources or usage of additional sources). For example, suppose a preemptive a player originally wanted to punish an opponent by 5 but, having learned of the insurance option, preferred instead to buy insurance of 10. This person had to put 10 in the insurance box; by ticking 'relocate funds', he would part with the punishment of 5, and would convert his cost of 0.5 to buy 5 units of punishment to purchase of insurance, and buy additional 5 units of insurance at a cost of 0.2 each, i.e. his total insurance cost would be 1.5.

Finally, after these decisions, the final treatment was announced to all participants. Any amount they have deduced from each other, and which was not insured against, may be disposed of in two ways: 'burned out', i.e. simply subtracted from the income of the punished player, or redistributed among the members of the group other than the punishing and the punished player. Proportion of this distribution was set as the fraction of the contributions of those players whom this addition may be assigned to: thus, if players 3 and 4 were to receive the proceeding of 6 from punishment of player 1 to player 2 (and if this punishment was not insured by this latter player), and if player 3 and 4 have contributed 10 and 5 units, respectively, then reallocation would mean that 4 units of these 6 would go to player 3, and 2 units — to player 4. This is an option which a competitive punisher would have never selected, while contentious punicher might. Again, this option was explained through examples, questions and answers, made available to all players (those who have not punished anyone had to place zeros in the placeholder form), and was not mentioned until the very moment the players had to proceed, so it did not interfere with the previous decisions. This was the last effective stage of the game; upon its completion, players observed the outcomes of the game, and had to proceed to the ex post questionnaire and payments.

The experiment was programmed in z-tree (Fischbacher, 2007), using Russian interface; instructions and exercises were handled individually in paper form. Altogether, 320 subjects took part in 21 experimental session hosting various number of players (8 to 24) in three cities: Moscow (capital and the largest city, 148 subjects), Perm (big city in the Urals, some 1500 km east of Moscow, 76 subjects) and Tomsk (regional capital in Eastern Siberia, about 4000 km east of Moscow, 96 subjects). The only explicit treatment variable

[9]This is the reason why we did not employ insurance in high-cost treatment: given the cost of punishment of 0.5, twice this cost would have meant the cost of insurance of 1, which makes no sense to purchase it.

in our experiment was punishment cost: low (0.1) in Moscow and Tomsk and high (0.5) in Moscow and Perm — in sum, 164 players were subjected to low-cost, and 156 to high-cost conditions. This setup might create some interaction between punishment behaviour and the locus of the session. Indeed, we noticed some differences between contribution patterns across cities, but as we will see shortly, this is not the case of punishment behaviour, which is quite uniform across cities, but strongly depends on punishment costs. Subjects were recruited through open calls in all locations; they were students of different departments (primarily economics) of Higher School of Economics (Moscow and Perm) and of different universities in Tomsk, who had no exposure to experimental games before. Mean age of participants was 20.5 years, gender balance was almost exactly equal. Experiments took part in Autumn 2010–Spring 2011. Subjects were paid on performance, the average payment being 215 RuR, or about 8.5 US$ according to the average market exchange rate at the times of the experiment, which were paid in cash at the end of each session[10]

5 Results

5.1 Prior projections and expectations

First of all, we look at the difference between expectations across the three cities as elicited in the ex ante questionnaire (see Figure 1).

The Figure shows, left to right: mean projected own contributions, mean normative (average due contributions, according to subjects' prior judgments), and mean expected (average expected contributions, which are generally, lower than normative) and factual mean contributions in game 1 (without punishment). Several facts are worth noting. First, projected own contributions (first column, with overall mean 9.10 and median 10), are always larger than factual contributions of the same subjects (mean 6.54, median 5). This difference is significant overall, and in all three cities separately, in line with the previous results. Indeed, Gächter and Renner (2010) in a repeated PG experiment with the UK and Swiss students of 4 participants and payoff parameters identical to ours, also elicit non-incentivized expectation of contribution, and report mean underestimation of contribution in the first round of around 5 of 20 units — pretty similar to our results. Second, and yet more strikingly, people expect other fellow players to undercontribute relatively to the normative standard, which difference is significant overall (for Perm, the difference is marginally significant, with Wilcoxon matched pairs test $z = -1.70, p < 0.088$). This suggests that people are ex ante sceptical of cooperation from the others, which might be source of potential disappointment. This feeling, however, is likely to be mitigated by the behaviour of the players themselves: planned contribution in all three cities (first column) was smaller than the normative contribution (second column; for the city of Perm only the difference is not significant). Furthermore, in factual behaviour subjects are even less generous than they expect of the average player overall measure (comparison of columns 3 and 4, Wilcoxon

[10]This relatively low material reward was due to the fact that, after completing this experiment, the same subjects took part in another, unrelated experiment. Subjects were paid at the end of the session consisting of two experiments, getting an average payoff of about 20 US$ (roughly, about 15 euro).

$z = 2.25, p < 0.023$), if not by cities. Altogether, these observations reveals systematic undercontribution relatively to normative standards of behaviour, which feature may be termed 'reflective scepticism': subjects not only anticipate other people to be less cooperative than they ought to, but intend to fall short of the standard themselves, and in fact behave even less cooperatively than they expect other people to be.

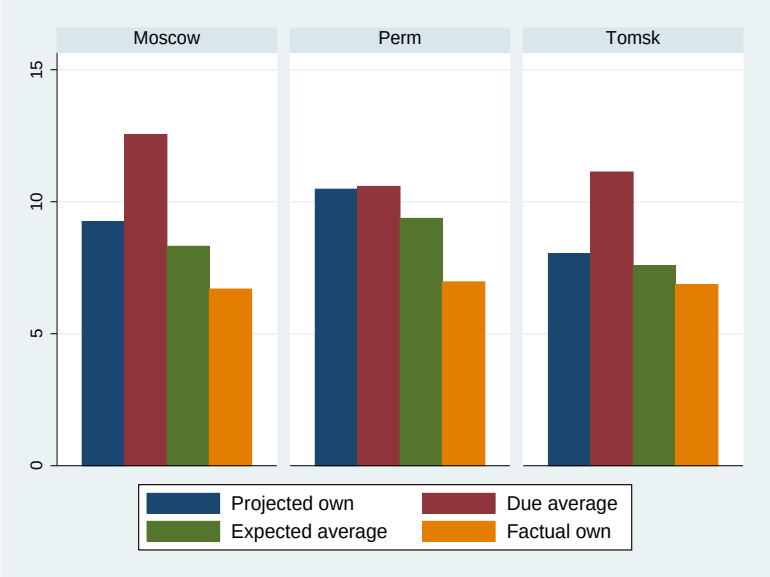

Figure 1: Projected, expected and factual contributions, game 1

Differences across cities are also of some interest. Factual contributions are almost the same across cities, while projected own and average contributions are overly optimistic in Moscow and Perm, but not in Tomsk, where expectations are most realistic. Normative contributions are highest in Moscow, followed by Perm and Tomsk, which differences from Moscow to both provincial cities being significant at 5% level (for Perm vs.Moscow, Wilcoxon $z = -2.20, p < 0.027$ and $z = 2.47, p < 0.013$; for Tomsk, $z = 2.12, p < 0.033$), while expected factual contributions are not significantly different in either pairwise comparison. Difference between expected ought and expected factual contributions of the group is lowest in Perm (means of 1.21 vs.3.54 in Tomsk and 4.22 in Moscow; medians 0, 3.5 and 3, respectively), meaning that people in the former city believe other people would contribute almost what they ought to, while participants from Moscow and Tomsk were more skeptical about the morale of their partners. Further, people in Perm, on average, project themselves to contribute what they ought to (mean difference between these variables is -0.10, median difference is 0), whereas in the other cities this deviation is statistically different from zero. This means that our Perm sample is somewhat more coherent in its behaviour and view of the fellow players from their home city — something which cannot be said of other two cities. Nevertheless, in general, both factual contributions and expectations of own contributions across the range (the strategy method type of elicitation, as shown on Figure 2) are similar across cities, as are changes in contribution in the second game vs. the first. In view of that,

we aggregate data across treatments, and henceforth concentrate on the main treatment effects.

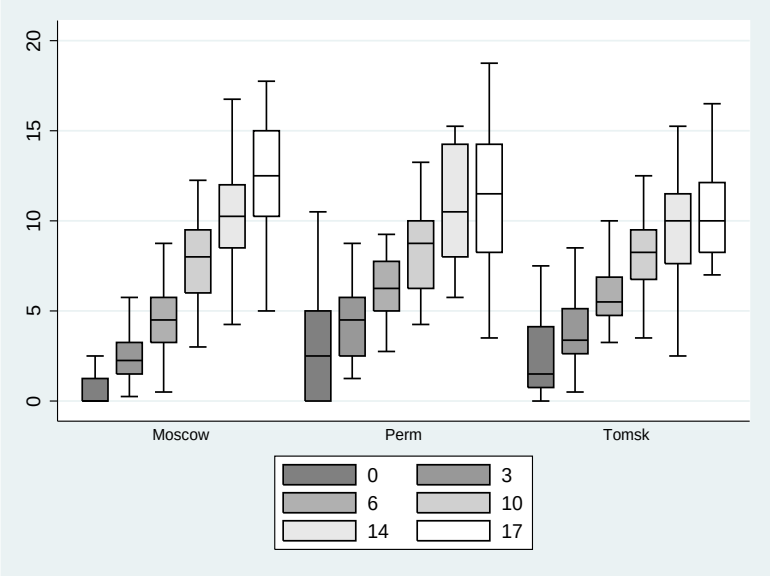

Figure 2: Projected contributions across the range of average contributions of the others. Displayed are total and interquartile ranges and medians of means of intergroup projections.

5.2 Contributions

Distribution of contributions by cities in the first and second games, together with normal probability plots, are shown on Figure 3. Mean contributions in the second game slightly increased relatively to the first (from 6.81 to 6.99), with median 5 in both cases. This equality result from two opposite factors: threat of punishment supports cooperation, but accummulated experience after stage 1 pushes it in the opposite direction. Overall, contributions are not significantly different across games (Kruskall-Wallis $\chi^2 = 0.012, p < 0.911$), and are significant only in one city (Tomsk — Kruskall-Wallis $\chi^2 = 6.152, p < 0.013$).

At the same time, contributions in the second game systematically vary with punishment cost, as shown in table 5.2: median contributions are 9 for the high, and 5 for the low-cost treatments, again indistinguishable across cities. This suggests that when punishments are more efficient, subjects become less cooperative. These differences are clearly significant at any degree of confidence, both overall and across cities. The most likely explanation to this observation is that people under low cost treatment were more afraid of punishment for misbehaviour, and as a result, contributed more. This explanation is supported by the fact that differences between contributions in stage 1 were not significant, with overall mean of 6.79 for low and 6.96 for high cost treatments.

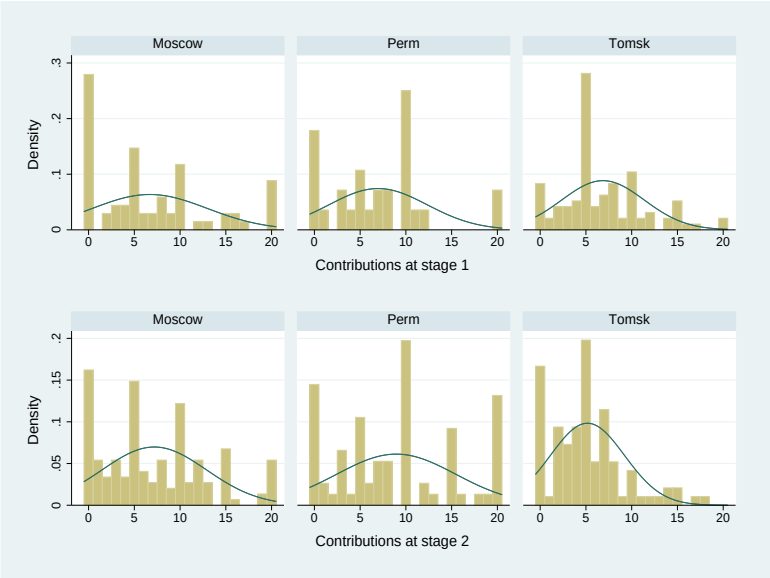

Figure 3: Distributions of contributions by cities

Sample	N	mean	median	st.dev.
Overall	320	6.99	5	5.64
Moscow, high cost	80	8.56	9	5.83
Perm, high cost	76	8.94	9	6.47
Moscow, low cost	68	5.55	5	5.08
Tomsk, low cost	96	5.16	5	4.04

Table 1: Contributions in game 2

5.3 Punishments

We now concentrate on the game 2, and consider the treatment effects of punishment. The first of these is willingness to punish at all, which has been introduced at stage 2 in our design (see Figure 6.3) to make sure this decision is entirely conscious. And indeed, on average, 54% of the respondents wanted to punish at least once, without significant difference between willingness to punish across the the cost of punishment treatment (WMW test statistic $0.30, prob < 0.76$). This statistics is generally in line with the previous findings: thus, Espin e.a. (2013) in a recent experiment, report 37.5% of punishments of at least one person in similar groups of 4 players, which is even less than in our case.

These findings, together with some to follow, suggest that our data are generally in line with the previous findings on cooperation and punishments in Russia. However there is one aspect in which this is not the case: of 173 individuals who have spelled out willingness to punish at least once, over 10% (18 subjects in all three cities, of which all but one under high punishment cost, of which 9 in Moscow and 8 in Perm) have never used this right in practice. This

fraction is significantly different from zero at any degree of confidence (Wilcoxon test $z = -4.24, p < 0.000$). This suggests that at least some subjects may have been incentivized to think twice about their true willingness to 'put the gun at work', and at a second thought, refuse from punishment action. The fact that this feature is almost completely limited to high-cost treatments suggests that these decisions are grounded on cost considerations.

Result 1 *About 10% of subjects who expressed willingness to punish did not complete that threat at the punishment stage. This feature is limited to high-cost treatment, where these rejections amount to 20% of instances (17 out of 83).*

Further statistics of punishments is presented in Table 2. From this Table, as well as from subsequent analysis, we exclude data on those 18 subjects who have wished to punish but did not punish anybody, as well as those 13 subjects who exhibited inconsistent behaviour has been allocated positive punishments to themselves. This leaves us with 289 subjects, who committed 278 instances of punishment, i.e. 32% of punishments out of all possible, and 0.96 punishments per person on average. This is compatible with previous results: Nikiforakis (2004, Figures 9 and 10) reports about 1 punishment per player on average in the first round of the repeated PG game with 4 players. Carpenter (2007b), in a small-scale experiment in the US reports 22% of punishments in repeated PG game.

First three columns of the table report statistics for overall contributions and for contributions of those who have never punished and have done so at least once. As reported earlier, high-cost treatments result in systematically larger contributions for the restricted subsample. Further, punishers have somewhat larger contributions, although this difference is not systematic: the most significant difference between contributions of punishers and non-punishers is for low-cost treatments, where it is only marginally significant (Kruskall-Wallis $\chi^2 = 2.63, p < 0.105$).

Last three columns report, respectively, statistics for punishment size, number of punished partners (out of 3 possible) and total punishment sizes averaged per person, again overall and by punishment costs. It strikes that punishers in the low-cost sessions not only have contributed less than in high-cost sessions, but also applied larger average punishment (means by treatments of 5.27 vs. 4.00, medians 5 and 3, WMW rank-sum test $z = 3.14, p < 0.017$) and imposed larger total puishments (means 12.90 vs. 9.03, medians 10 and 6, WMW $z = 3.45, p < 0.000$), even though average number of punished partners is not significantly different. Total number of punishment instances (165 vs. 113) is also larger for low-cost punishments. This is consistent with intuition, as well as previous evidence: Carpenter (2007b) evaluates demand for punishment substitution effects, and finds them to vary with punishment cost scales in the same direction: the higher the cost, the lower are punichments. Falk e.a. (2001) report the same tendency, even though their cost structure is more elaborate.

Result 2 *Contributions under low costs are systematically lower, and punishments are systematically larger than under high cost treatment.*

As a first look at the causes of this result, consider Figure 4 which plots mean punishments categorized by deviations of the punished subject's contribution from contribution of the punisher. Positive values on the horizontal axis

	contributions			*punishments*		
	overall	*non-punishers*	*punishers*	*size*	*number pp*	*sum*
Overall ($N = 289$)						
N	289	144	145	278 instances		
mean	7.06	6.88	7.23	4.76	2.28	11.33
median	6	5	6	4	3	10
std.dev.	5.65	6.21	5.04	3.22	0.79	9.02
Low cost= 0.1 ($N = 155$)						
N	155	71	84	165 instances		
mean	5.46	5.15	5.73	5.27	2.31	12.90
median	5	5	5	5	2	10
std.dev.	4.56	5.22	3.93	3.34	0.76	9.50
High cost= 0.5 ($N = 134$)						
N	134	73	61	113 instances		
mean	8.90	8.56	9.31	4.01	2.26	9.04
median	10	9	10	3	3	6
std.dev.	6.21	6.65	5.67	2.89	0.83	7.76

Table 2: Contributions and punishments by treatments

correspond to contribution of the punished person that are larger than those of the punisher, hence punishments in this range are spiteful, while punishments in the negative range are prosocial. As shown in this figure, overall punishment sizes are generally in the range from 3.7 to 6.2, with one peak at zero deviations[11]. A notable thing is that mean punishments are not monotonic in deviations for either prosocial or spiteful punishments, specifically when the cost of punishments are high. Prosocial (retaliation) interpretations of punishment suggest that its size should increase in under-contribution of the punished player relatively to the punisher — and indeed, this trend is observable in the overall data. Similar tendency, however, takes place for spiteful punishments: the more cooperative is the punished player, relatively to punisher, the larger is punishment size. This bilateral tendency has been observed in many countries (Herrmann e.a., 2008). Decomposition by treatments, however, shows that in our data this tendency is due to low-cost treatment, but breaks down for high-cost treatments. Difference between punishment patterns is significant across treatments (ANOVA $F = 10.68, p < 0.001$), and remains robust to alternative measures, including medians and deviations of contributions from group averages. This irregularity has also been reported in other studies — e.g. Gächter and Herrmann (2007, Figure 4)[12]. Also similar to previous studies were the fractions of punishments of deviants in each category, presented in Figure .

Of 278 instances of punishment, 78 (28%) were spiteful (punished player contributed more than the punisher), with 187 instances (67%) of prosocial

[11] Although players were allowed to use decimals, all contributions and punishments in our experiment were in integer numbers.

[12] Another striking observation is that punishment of fellow players who contributed exactly the same amount as the punisher were overall the largest. This suggests that ethical considerations were not the main driving force behind such punishments; however, this category is the smallest in size (only 13 observations, of which 4 are under high costs).

punishment (in 13 more cases, punished players contributed exactly as much as the punisher). Gächter and Herrmann (2007) also construct the measure of spitefulness of punishment, μ, defined by these authors as ratio of mean sizes of spiteful punishments to prosocial punishments. For the whole sample, this value is 1.18 (= 5.25/4.44), for low cost treatment it is higher at 1.47 (= 6.67/4.54), and for the high cost treatment, lower at 0.74 (= 3.21/4.31). Only this last value is compatible with their value (in range of 0.35-0.78), which is understandable, as unit punishment in their experiment costs 0.33 units. This again suggests that punishment size is sensitive to costs.

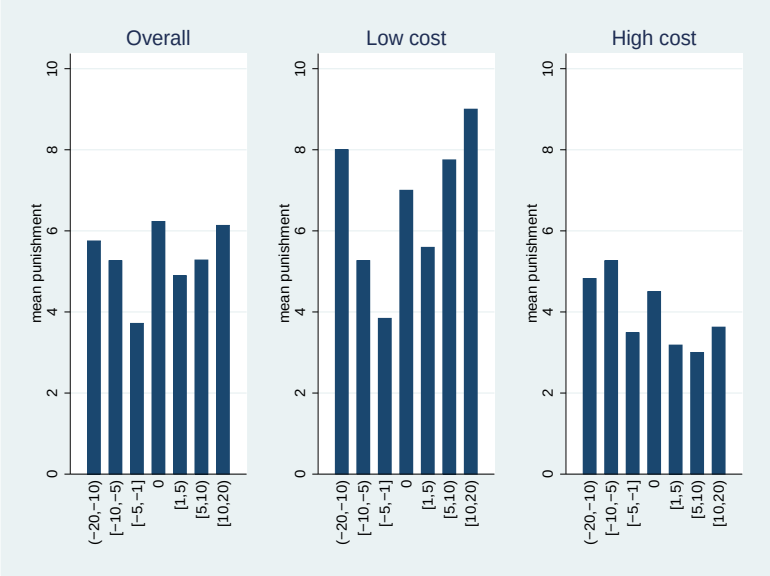

Figure 4: Mean punishments by categories of deviations
This figure is based on 278 valid cases, of which 165 are under low, and 113 under high cost treatments.

In general, spiteful and prosocial punishments do not systematically differ across punishment cost treatments (chi-square test value $\chi^2(1) = 0.001, p < 0.98$). In a more detailed way, of the 78 instances of spiteful punishments, 46 took place when the cost of punishment was low, and over a half of these (24 instances) occur when punishments themselves reach the possible maximum of 10 units. Such extreme punishments were much less common for the high-cost treatment, with only 4 instances out of 32. This tendency does not extend to extreme prosocial punishments, when the cost of punishment seems to be immaterial: 17 out of 110 punishments are maximal under the low, and 10 out of 77 — under the high cost, amounting to 15 and 13% of all punishments, respectively. These conjectures are confirmed statistically: Wilcoxon-Mann-Whitney (WMW) rank sum test confirms significant difference between spiteful punishments under low and high costs ($z = 4.16, p < 0.000$) and no difference between prosocial punishments across these treatments ($z = 0.48, p < 0.625$). The is our first observation concerning the nature of 'spiteful' behaviour in the light of our experiment:

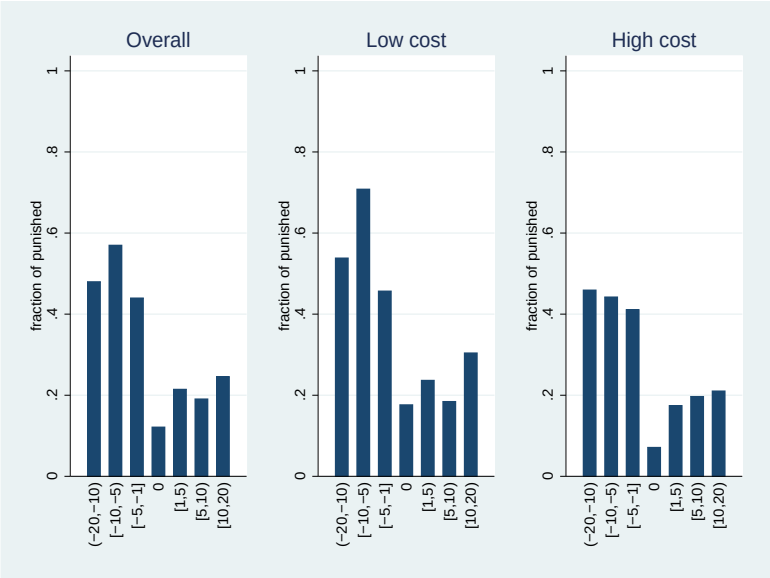

Figure 5: Fraction of punishments by categories of deviations
This figure is based on 867 cases, of which 465 are under low, and 402 under high cost treatments.

Result 3 *Spiteful punishments are systematically larger under low cost than under high cost, which tendency is not true for prosocial punishments.*

5.4 Spite

Our data allows to look further at factors and conditions related to spiteful and prosocial punishments. First, in addition to summary statistics, our 4-players experimental setup justifies a look at the numbers of punished partneers, as well as total and average punishments imposed on them. These are summarised in the left part of Table 3 (ignore for the time being the right part. This Table confirms the above tendencies: spiteful punishments are larger in size and in total, as well as more frequent. Looking more closely, consider the distribution of punishments. Of 289 valid subjects in our study, 145 (50%) punished at least once. Of these 145 punishers, 58 subjects punished only once, 41 twice and 46 three times. Splitting by treatments, 84, or 58% of punishments (30+27+27 by numbers) took place under low, and 61, or 32% (28+14+19 by numbers) under high costs conditions. These figures reveal an expected decrease of overall number, and dropout in seriality of punishments in case of high costs — a difference which is not significant though (ANOVA $F = 2.49, p < 0.115$). What is significant is significant change in punshment size, which is illustrated by Figure 6.

Clearly there is an increasing trend of mean punishment with the seriality of punishment, which trend is significant overall (ANOVA $F = 50.91, p < 0.000$, Kruskall-Wallis with ties $\chi^2 = 104.98, p < 0.000$), and for both treatments. Figure 7 reveals why: punishment size increases with seriality for spiteful punishments (ANOVA $F = 4.10, prob < 0.021$, Kruskall-Wallis $\chi^2 = 4.62, p <$

Spiteful punishments								
	Full sample (N=78)				*Cleaned sample(N=57)*			
stats	size	number	sum	average	size	number	sum	average
mean	5.26	2.62	15.04	5.48	4.43	2.06	9.33	4.43
median	4	3	12	4	4	2	7	4
st.dev.	3.75	0.63	11.02	3.52	3	.819	7.53	2.79
min	1	1	1	1	1	1	1	1
max	10	3	30	10	10	3	30	10
Prosocial punishments								
	Full sample (N=187)				*Cleaned sample(N=164)*			
mean	4.45	2.15	9.51	4.36	5.93	2.58	16.3	5.93
median	4	2	7	4	5	3	12	5
st.dev.	2.92	0.82	7.44	2.73	3.8	.68	11.6	3.65
min	2	1	2	2	1	1	1	1
max	10	3	30	10	10	3	30	10

Table 3: Punishment statistics by punishment characteristics

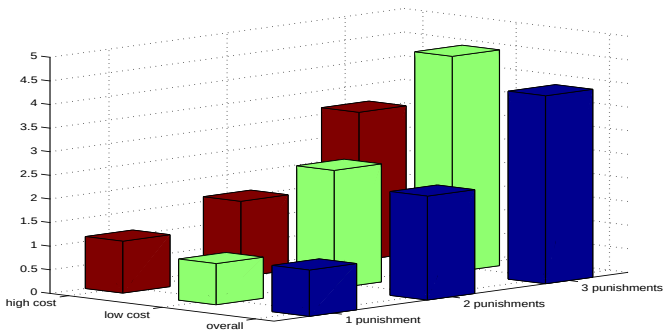

Figure 6: Mean punishments sizes by number of punishments, treatment conditions

0.099), but not for prosocial ones ($F = 0.74, prob < 0.477$, Kruskall-Wallis $\chi^2 = 1.04, p < 0.593$). Further look corroborated by result 3 shows that is effect is primarily due to low-cost treatment (ANOVA $F = 7.49, prob < 0.002$, Kruskall-Wallis $\chi^2 = 12.20, p < 0.0.003$), while other contrasts are not significant. Distribution of frequencies of punishment instances is illustrated in Figure 8, which again confirms the same tendency: spiteful punishments under low cost are unusually large.

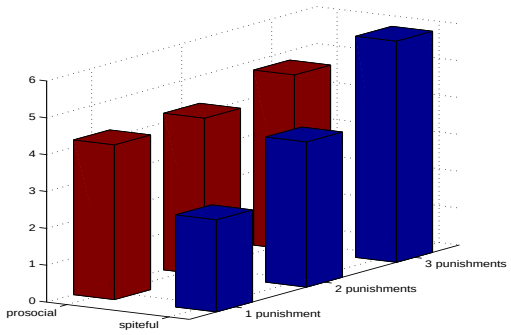

Figure 7: Mean punishments sizes by number of punishments, character of punishment

One might wonder if this increase in seriality in case of spiteful punishment (and its absence in case of prosocial punishments) is not an artefact of data analysis, in the sense that some of the 2– and 3–players punishments might come from subjects some of whom punish prosocially, and some spitefully (punishments of one of the subjects are classified unambiguously). To check for this, we restrict the classification by punishment character to those subjects who punish only prosocially or only spitefully, all over the range. Relevant statistics are provided in the right part of Table 3, where 50 subjects punished prosocially once, 27 prosocially and only prosocially twice, and 20 — prosocially and only prosocially three times. Corresponding figures for spiteful only punishments are 6, 6 and 13, which confirms presence of a small but very persistent cluster of spiteful punishers. Altogether, 122 (84%) out of 145 punishing subjects were either keen prosocial or keen spiteful punishers. Differences between them in this comparison become even more striking than before, and significant for all indicators in Table 3 at any reasonable degree of confidence. This finding warrants our next result:

Result 4 *Spiteful punishments are systematically larger and more serial than prosocial ones.*

Splitting data by punishment character allows us to shed first light into the main question: what are the motives for prosocial and spiteful punishments. This can be usefully done with the help of our hypothetical ('strategic form') questions which lead to testable implications for the range of retaliation motives. Consider first the *congruence* interpretation, according to which punishment us based on the differences between individual contributions of the two involved players, $c_i - c_j$, and driven by ethical standard embedded into punisher's mind.

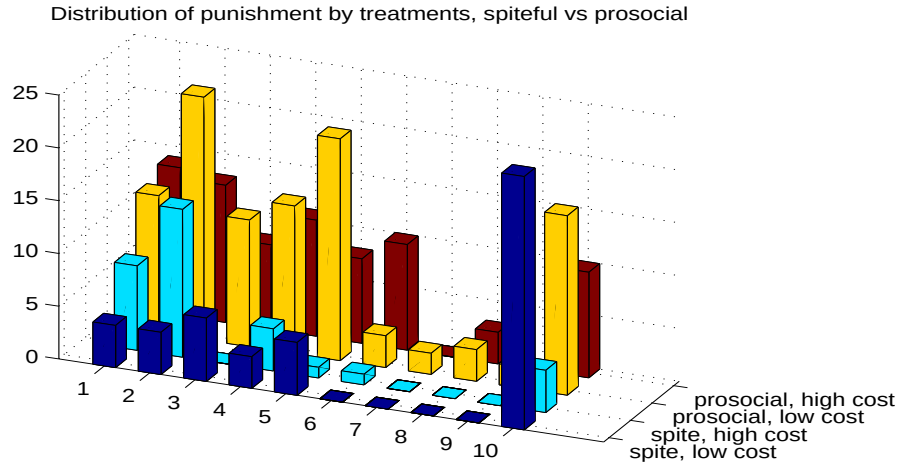

Figure 8: Distribution of punishments by treatment and character of punishment

One way to interpret prosocial punishment along this way is that players i punish players j who are more greedy than themselves, i.e. whenever this difference is positive. Spiteful punishers i can also be driven by this comparison: in a sort of 'generosity pride', they may believe that it is their right and duty to be the most generous in the group. In such case, they will be be unhappy if players j daresay to contribute more than they did, i.e. punish if this difference is negative. We shall refer to the interpretation according to this logic as to *pride*.

Unfortunately, no test of this interpretation can be based directly on $c_i - c_j$ difference, because this same value classifies punishments as prosocial or spiteful. Fortunately, our design allows to construct a proxy variable for this source of unhappiness, based on the reported contributions measured by 'strategy method'. Relevant variables are reported in the first three columns of Table 4 (the last three columns are discussed later). To check for congruence, we take deviations of the reported to-be-contribution of each individual player from the group category (0,3,6 etc.), and then average out these deviations. Resulting variable *dexpcav* may be taken as measure of what the player deems *a priori* 'ethical' behaviour all over the range; it statistics is reported in Table 4, and its histogram for prosocial and spiteful punishers is plotted on Figure 9. Taking the group category as an expected contribution of a representative 'other' player, c_j, intentional prosocial punishers should be unhappy if *dexpcav* is positive, i.e. if their projected contribution will be larger than that of the representative other player. By the same token, spiteful punishers driven by pride shall be unhappy if this variable is negative. Figure 9 reveals systematic *ex ante* differences of the projected deviations for spiteful and prosocial punishers: for the latter, the distribution of *dexpcav* is bell-shaped with mean not significantly different from zero (see also Table 4), while for the former, is clearly biased towards the left, and differences between the two are significant (Kruskall-Wallis $\chi^2 = 34.34, p < 0.000$).

stats	dexpcap	cavg	difavg	insurance size	new funding share	redistribution share
Spiteful punishment ($N = 78$)						
mean	-3.56	-0.38	1.86	4.65	0.26	0.41
median	-3.16	-1	1	5	0	0
st.dev.	3.28	6.60	3.03	3.20	0.44	0.49
Prosocial punishment ($N = 187$)						
mean	-0.75	-8.63	-3.15	2.01	0.46	0.52
median	-0.83	-8	-3	1	0	1
st.dev.	3.51	5.61	2.66	2.41	0.50	0.50
No punishment ($N = 1102$)						
mean	-1.60	-4.23	0.49	1.21	0.62	0.23
median	-1.16	-4	0.25	0	1	0
st.dev.	3.60	7.73	4.04	2.24	0.49	0.42

Table 4: Behavioural indicators split by punishment characteristics

For many prosocial punishers (namely, 58 out of 187 instances) *dexpcav* is positive, hence their punishments might be driven by pride. However, in a significant fraction of cases — to be specific, 104 out of 187 instances — this value is negative. These people from the very beginning have not intended to contribute more than the 'representative' players whom they have punished, which is incompatible with the hypothesis of ethically-driven prosocial pride. For spiteful punishers, the case against this hypothesis is even stronger: they should have planned to contribute more than the mean player in the group (for otherwise they cannot be the most generous players!), which is the case of an overwhelming minority — to be specific, of only 7 of 78 instances. It follows that pride could be the reason for punishment only for those prosocial punishers who have intended to conribute more than the average player, and cannot be the reason for spiteful punishments.

One might also conjecture that punishers' dissatisfaction with the size of $c_i - c_j$ is to be interpreted as tribute to rationality reasoning, either in the behavioural sense (prosocial punishers should endorse cooperation, and punish those individuals who do not contribute), or in traditional game-theoretic sense (people cannot be that stupid to contribute, and my punishment would work as a natural selection mechanism which gets rid of irrationalities at the level of the society). In these cases, punishments should be applied primarily to those players who have under/overcontributed most relatively to the punisher. As universal explanation, this *sanitation* reason does not work either: only about one half of either subsamples (47 and 56 for spiteful and prosocial punishments, respectively) are applied to those players whose contributions constituted maximum deviation from the punisher's contribution in their subgroups. These observations again do not preclude the possibility that some punishers were motivated by such factors; however, they require more detailed analysis of individual strategies, which we present later.

Turning to the other interpretations of retaliation, *conformability* derives punishment from dissatisfaction of contribution of the punished player, c_j relatively to the normative standard, which is directly available in our data elicited

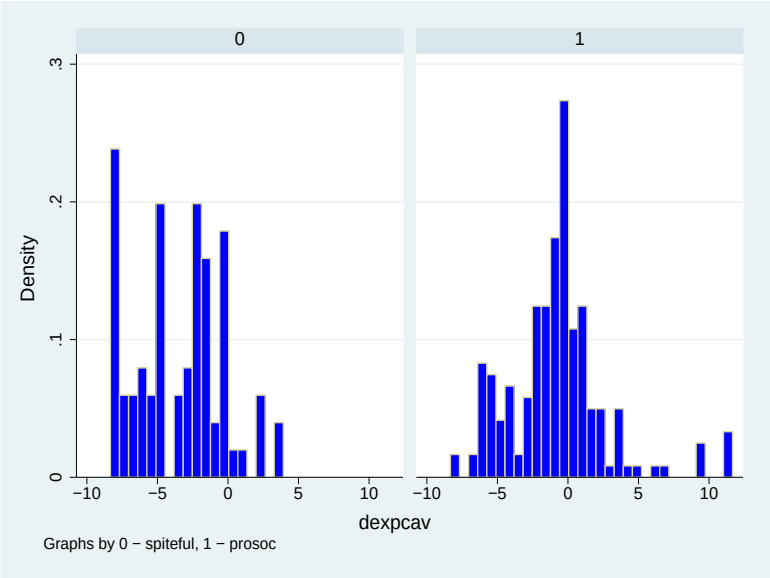

Figure 9: Distribution of mean deviations of projected from group category contributions

by 'strategy' method. Difference between contribution of the punished player and this standard, *cavg*, should be significantly negative for prosocial punishers (punished player contributed less than (s)he should have to), and positive for spiteful ones. The first tendency is confirmed at any degree of confidence (see table 4), while the second, with mean value of -0.38 and median of -1, is not significantly different from zero. Hence, this explanation would work for prosocial but not for spiteful punishers.

Finally *conformity* interprets punishment as manifestation of dissatisfaction of deviation of contribution of the punished player, c_j, from factual group norm (mean group contribution), which is again available in our data. Difference, between them, *difavg*, should again be negative for prosocial (individual contribution is lower than expected) and positive for spiteful punishment. This interpretation is confirmed: mean values of this variable for prosocial (-3.15) and spiteful (1.86) punishments are significantly different from zero and have the predicted signs.

Further reinforcing these conclusions is Table 5, which is analogous to Table 4, but uses punishment instances cleaned in the same way as in the right part of Table 3. Tendencies in this table are the same, confirming the validity of our conclusions: prosocial punishments on average work as predicted by the congruence (both pride and sanitation), conformity and conformability, all statistics of the respective variables being significantly different from zero at any reasonable degree of confidence. By contrast, for spiteful punishments the hypotheses are confirmed only for conformability, where *difavg* is significantly different from zero (Wilcoxon matched pairs test $z = 3.02, p < 0.002$). Altogether,

Result 5 *Prosocial punishments are compatible with retaliation punishment motives, spiteful punishments are not except for conformability to group standard.*

stats	dexpcap	cavg	difavg	insurance size	new funding share	redistribution share
Spiteful, 1 punishment (6 players)						
mean	-3.83	-3	0.75	2	.33	.16
median	-3.83	-3.5	1	3	0	0
std.dev.	2.62	5.72	0.63	1.73	.57	.41
Spiteful, 2 punishments (6 players)						
mean	-4.77	2.16	1.20	3.17	.167	.5
median	-4.91	-0.5	0.63	2	0	.5
std.dev.	2.68	7.16	3.50	3.66	.41	.52
Spiteful, 3 punishments (13 players)						
mean	-4.03	-0.72	1.53	5.74	.15	.39
median	-3.33	0	1	5	0	0
std.dev.	3.25	6.66	3.25	3.19	.36	.49
Prosocial, 1 punishment (50 players)						
mean	-1.16	-8.86	-4.69	1.92	.4	.58
median	-1.16	-8.5	-4.5	1	0	1
std.dev.	3.29	6.30	2.54	2.36	.5	.49
Prosocial, 2 punishments (27 players)						
mean	0.31	-9.98	-3.28	2.17	.583	.5
median	0	-9.5	-3.25	2	1	.5
std.dev.	3.85	5.30	2.11	2.42	.5	.51
Prosocial, 3 punishments (20 players)						
mean	-0.85	-8.31	-1.67	1.28	.41	.57
median	-0.66	-8	-0.87	1	0	1
std.dev.	3.14	4.87	2.70	1.62	.5	.5

Table 5: Behavioural indicators of cleaned sample of punishers

5.5 Insurance

Consider now the effects of insurance, which were allowed in low-cost treatment. There were 238 instances of 468 admissible cases, which is slightly above 50%, meaning that insurance was more popular than punishments. Of these insurance instances, 146 (61%) came from those who did not punish anyone, and 92 were purchased by the punishers (55% of the 165 punishers in low-cost treatment). Of the 146 punishing players, 40 were spiteful and 68 prosocial, with the remaining 38 unclassified.

The fraction of spiteful punishers who bought insurance is larger than overall spite: Only 6 of the 46 instances of spiteful punishments were not accompanied by insurance — a sharp contrast with 42 of 110 prosocial punishments that were not. This difference is clearly significant statistically (chi-square test statistic $9.62, p < 0.002$, and Fisher exact test $p < 0.002$). Further, as shown in the fourth column of by Table 4, insurance sizes for spiteful punishers are systematically larger (median 5) than that of prosocial ones and those who did not punish (median 1 and 0, respectively); figures for the cleaned sample in Table 5 are similar. Finally, spiteful punishers have funded their insurances using new funds (column 'new funding share' in Tables 4 and 5) much less often than prosocial punishers: for Table 4, the respective fractions and 26% vs 46%.

These statistics imply that insurance motives seem to be very different for those categories of punishers. This is explicit from Figures 10 and 11, which show the relative frequencies of insurance purchases of punishing players in four categories (no insurance policy, insurance against small punishment of 1 to 3 units, of moderate punishment of 4 to 7 units, of large punishment of 8 to 10 units), by the categories of deviations of contribution of the punished player from those of the punisher (and the insured player). Differences are striking: most of insurance of prosocial punishers is small in size, and directed towards possible punishments of people who are slightly less cooperative than they are. This predominant pattern means that prosocial punishers purchase mostly small insurances against occasional punishments of slightly more antisocial players. By contrast, spiteful punishers purchase large insurances, and are mostly directed towards slightly more prosocial players; further, unlike prosocial players, they relatively unfrequently insure against possible punishment of prosocial players.

These results allow us to note a striking difference between insurances of prosocial and spiteful punishers:

Result 6 *Insurances of spiteful punishers is more systematic, larger in size and is much more frequently purchased against spiteful punishers than against prosocial ones. Insurance of prosocial punishers is smaller in size, and mostly purchased against occasional punishments of similarly prosocial partners.*

We also control for sex, profession and differences across cities (not reported here) — none of these comparisons reveal systematically significant effects on punishment strategies.

An important role of insurance in our experiment is to control for contentious (Tjutchev) motive: punishment takes place because people view it as 'natural' or 'normal'. Data contradict this hypothesis: of those who punish at least one opponent, 51% (43 individuals) purchase insurance, while the corresponding purchases from among those who did not punish is only 31% (23 individuals).

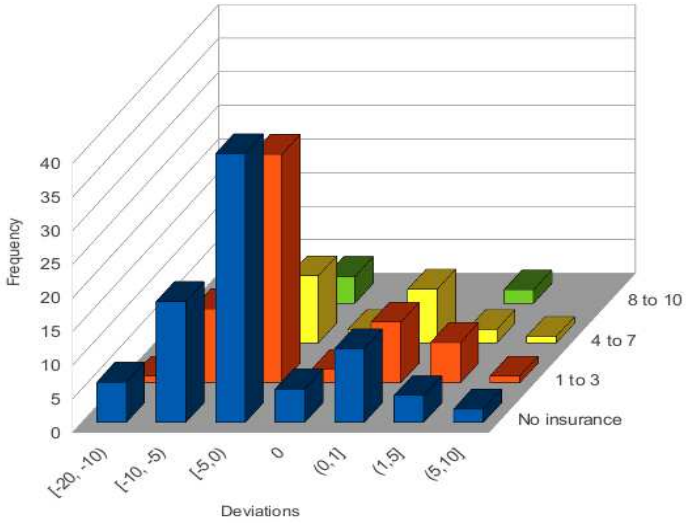

Figure 10: Insurance of prosocial punishers

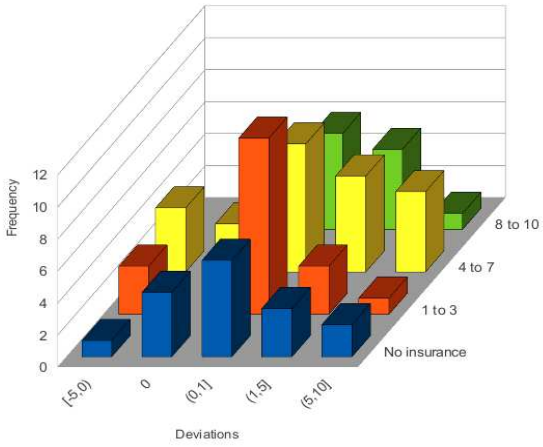

Figure 11: Insurance of spiteful punishers

Further, as shown in Tables 4 and 5, spiteful punishers more often than prosocials tend to relocate money from punishment to insurance, which suggests that preemptive motive rather than contention is more common for the former, but not for the later players. This is confirmed by Figure 12, which shows the distribution of insurance by funding sources and categories of punishment. A detailed analysis should of course be done on individual grounds, but as a general tendency, this Figure implies that prosocial punishers buy smaller insurance, which is somewhat more often funded by redistribution (56% of all cases, especially for prosocials), while spiteful insure more, and predominantly redistribute (75%). Both findings imply that preemption rather than contention seems to be the main driving force behind punishments.

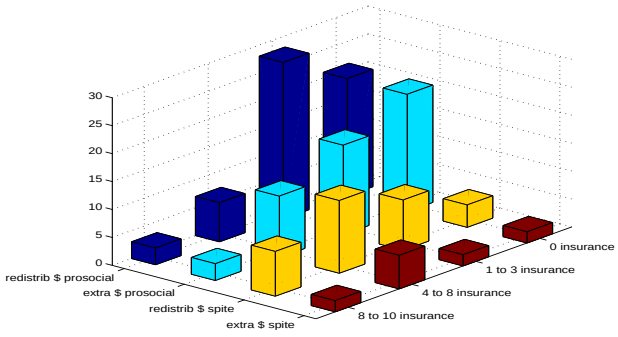

Figure 12: Insurance by sources

This conclusion is further corroborated by Figures 13 and 14, which display sizes of residual punishments applied by spiteful and prosocial punishers (left after all redistributions of expenditure from punishment to insurahce) have been made). These graphs categorize frequencies of residual punishments (on the left axis) against purchased insurance on the right. There are two clearly different patterns: most prosocial punishers leave heavy punishments with either very low or no insurance. By contrast, spiteful punishers seem to have bimodal distrbituion, with one cluster of small insurances and small punishment, and another, well expressed in the category of insurance of 4 to 8 units, for large insurances. Behaviour of the former subgroup is compatible with preemption, while that of the latter — with competitive motive, as we discuss next.

5.6 Assignments

Finally, we look at assignments of punishments to other players vs. burning it down: proportion of the former strategy is shown in the last column, 'redistribution share' in Tables 4 and 5. Prosocial punishers burn and redistribute punishment proceedings in almost equal proportions (48% and 52%, respectively), while spiteful punishers mostly burn (59% vs.41%). Further statistics of assignments are provided in Table 6, which splits data on prosocial and spiteful punishers by directions of money usage (burning vs. redistribution). It shows that spiteful punishers who burn are also the harshest punishers (mean 6.56, median 8) and insurers (mean 5.12, median 5), which figures are significantly

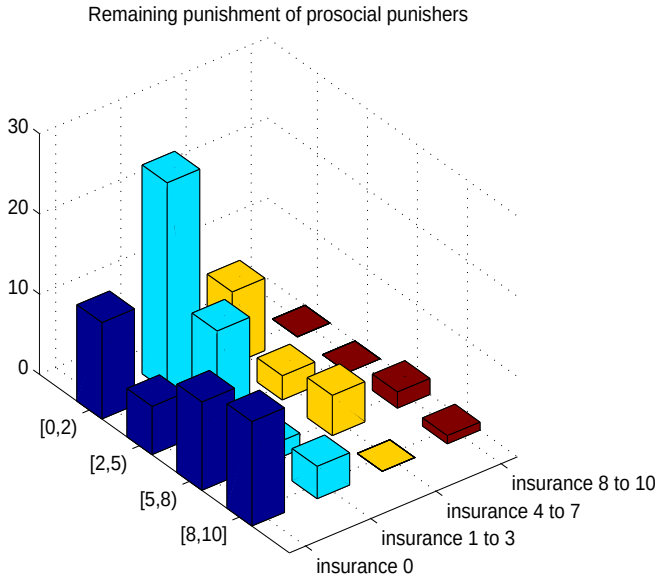

Figure 13: Residual prosocial punishments and insurance

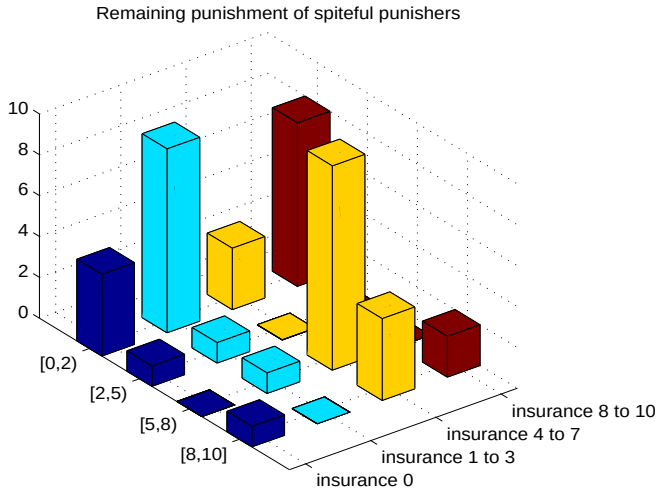

Figure 14: Residual spiteful punishments and insurance

different from the rest of the sample. This clearly suggests that competitiveness hypothesis tends to be valid for the category of spiteful punishers, while other motives might be appropriate for the rest of the sample.

	contribution			insurance	
stats	punisher	punished	punishment	size	new fund.
	Spiteful, burn (N=46)				
mean	2.58	7.65	6.56	5.12	0.24
median	2.00	7.50	8.00	5.00	0.00
std.dev.	2.74	3.98	3.61	3.18	0.43
	Spiteful, redistribute (N=32)				
mean	4.71	10.81	3.37	3.46	0.30
median	4.00	10.00	2.00	3.00	0.00
std.dev.	4.20	5.18	3.11	3.04	0.48
	Prosocial, burn (N=89)				
mean	8.16	3.16	4.24	1.85	0.48
median	7.00	2.00	4.00	1.00	0.00
std.dev.	4.86	3.26	2.68	2.31	0.50
	Prosocial, redistribute (N=98)				
mean	9.88	3.28	4.63	2.16	0.44
median	10.00	3.00	4.00	2.00	0.00
std.dev.	4.69	3.53	3.12	2.53	0.50

Table 6: Statistics of assignments by contributions, punishments and insurances

5.7 Survey data

Last, consider supporting evidence of the survey questionnaires, which we have distributed to participants at the end of the experiment. Table 7 shows percentages of the most popular answers to two questions (participants were free to choose up as many motives as they like):

1. On the ground of what considerations did you made the decision about deducing something off other player's revenues, or not deducing anything at all?, and

2. If you have made deductions, on what grounds did you decide about their size?

Ordering answers by priority, we see that the modal motive for punishment is lower contribution/level of cooperation of the punished, while modal among spiteful punishers is the competitive motive (to gain more than the punished player, which makes full sense given cost efficiency of punishment as revenue deduction mechanism). The main determinant of punishment size of prosocial punishers is again the relative contributions, while spiteful punishers usually apply the indiscriminately largest possible punishment.

This last finding is further calibrated by the evidence of Table 8, which shows mean and median answers at 0–10 scale (from immaterial to crucial) to the following determinants of punishment: difference in contributions of the punisher

Reasons for punishments		
Variable	Prosocial (N=121)	Spiteful (N=53)
Lower (than average) contribution	47.1	20.8
To stop them lowering our revenues	13.2	7.5
To gain more than they will	12.4	43.4
Afraid of them reducing my revenue	11.8	9.4
To equalize revenue within group	9.1	15.1
Intuitively/to experiment	7.5	1.9
Size determinants		
Variable	Prosocial (N=121)	Spiteful (N=50)
Inverse to their contribution	29.0	6.0
Maximal to the smallest contributor	18.5	8.0
To average out revenue	15.5	16.0
To put all revenues down to mine	11.5	–
Intuitively	8.7	14.0
Depending on my costs	6.8	–
Maximal to all	2.9	38.0
Minimal to all	1.9	8.0

Table 7: Main punishment factors, % of subjects choosing that answer
Note: – less than 2%

and the punished (*diffcontrib*), difference of the group average contribution and that of the punished player (*diffgroup*), fear of being punished by the punished player (*retaliate*), desire to gain more than the opponents (*competit*) and the cost of punishment (*costs*). As can be clearly seen, considerations of relative contribution and retaliation loom larger for prosocial punishers, while the most significant difference is for competitive motives, which is ways more important for spiteful motives, which is mentioned by over 80% of spiteful punishers. Taken together with the previous findings, this verbal evidence provides futher support to the following result:

Result 7 *Prosocial punishments are called upon by low cooperation level of the punished player, whereas spiteful punishments are associated with the competitiveness motive, or attempt to gain more than other people in one's group. Preemptive motive presumably takes place for both subgroups, revealing heterogeneity in their compositions.*

This evidence seem to qualify the behaviour which appears to be 'spiteful' as unrelated to disapproval of prosocial behaviour, but akin to competitiveness hypothesis.

5.8 Econometric evidence

These observations can be further calibrated econometrically. Table 9 shows results of tobit model estimates for the factors affecting spiteful and prosocial punishment (best models only are presented), as well as for the pooled sample (for all samples, clustered standard errors were used). Not surprisingly, cost is the only variable that matters for punishment sizes in both directions. Prosocial

statistics	*diffcontrib*	*diffgroup*	*retaliate*	*competit*	*costs*
	Spiteful punishment (N=45)				
mean	3	2.64	4.64	5.6	4.26
p50	2	1	3	6	4
sd	2.98	3.08	3.90	4.20	3.46
	Prosocial punishment (N=101)				
mean	5.23	3.69	2.67	2.98	3.32
p50	6	3	2	2	2
sd	3.82	3.37	3.15	3.53	3.20

Table 8: Importance for punishment decision

behaviour is conditional upon contribution of the punishing person, *contr*, with the minus sign (the more I contributed, the less I punish as cost consideration do matter), difference between contributions of the punisher and the punished player, *difcontr* (with positive sign: the larger the difference, the larger is anger), and marginally — on the difference between own planned and expected factual contributions of the other players, *homxavg*, which implies that failure to match prior expectations is instrumental in causing prosocial punishments.

The story of spiteful punishment is somewhat different: what matters is difference between one's contributions, and difference between contribution of the punisher and average contribution of the group, *relcontr*. This sign is negative again, implying that the lower is contribution of the punishing player relatively to the group average contribution, the higher is spiteful punishment. This observation futher supports the competitiveness hypothesis: the lower is one's position in the group, the larger are the chances for spiteful punishments, as further confirmed by our earlier discussion of insurance.

Table 9: Estimation results for punishment

	Spiteful		Prosocial		Total	
Variable	**Coef.**	**Std.Err.**	**Coef.**	**Std.Err.**	**Coef.**	**Std.Err.**
contr			-0.409^{***}	(0.103)	-0.658	(0.061)
difcontr	-0.865^{***}	(0.224)	1.312^{***}	(0.122)	0.695^{***}	(0.098)
relcontr	-1.583^{*}	(0.947)			-0.451^{**}	(0.182)
homxavg			0.175^{*}	(0.112)	0.029	(0.079)
cost	-22.17^{***}	(6.263)	-6.290^{***}	(1.575)	-8.753^{***}	(1.635)
Intercept	-20.025^{**}	(4.859)	-5.216^{***}	(0.606)	-4.259^{***}	(0.564)
Log pseudolik.	-368.55		-739.23		-1167.29	
N	958		1060		1148	

Tobit model estimates. *** — significant at 1% level, ** significant at 5% level, * significant at 10% level

6 Factors of punishment

Let us now briefly recap our conclusions concerning the importance of the various explanations for spiteful and prosocial punishment behaviour.

Due to sample size restrictions, we could not directly test for the **availability** motive. However, between–experimental comparisons (Gächter and Herrmann, 2008; Carpenter, 2007a,b; Nikiforakis, 2008, 2010) suggest that our subjects did not punish significantly less because of our introduction of that punishment stage. A striking and interesting feature of our design is, however, a highly visible dropout of punishments from among those subjects who have expressed their willingness to punish at first instance. A possible explanation to this fact might be that our availability option gave them additional impetus to think of the reasons why they might punish anyone in their group.

Preemption as motive seems to be valid as motive, for both prosocial and spiteful punishers. Insurance was more popular than punishments; furthermore, we saw that that spiteful punishers were even more willing to make use of money transfers than prosocial punishers to cover their insurance. Difference between insurance patterns across punishment characteritics (Figure 12) is also very strking: prosocial punishers tend to insure against small punishments, while spiteful ones insure against large ones. Constellations of these conclusions implies that, while preemptive motive seems to be valid for all players, there is substantial intra-type heterogeneity between motives of prosocial and spiteful players. Careful analysis of this heterogeneity requires individual-level scrutiny.

By contrast, *contentious* motive does not seem to gain support. If subjects view punishment as something normal, they should value fight per se, preferences towards punishments should not be affected if other reasons for applying it (such as insurance) disappear. This is not the case: when in the last part of the game, subjects are asked whether they want relocate their funds from punishments to insurance, an overwhelming majority (75%) of spiteful and half of prosocial punishers have done so, although here again we cannot exclude heterogeneity in punishment strategies.

Larger share of maintained punishments (see Figure 14), unwillingness to reassign punishments to other players, and questionnaire survey evidence all suggest that *competitiveness* as motive is valid for at least some of the spiteful punishers. By contrast, most of retaliation motives do not work for this subgroup as a whole, but is very valid and important for the prosocials. This last conclusion is further supported by the questionnaire and econometric evidence, as well as by the large punishment sizes left over by those prosocials who did not insure at all (the leftmost row in Figure 13).

To sum up, we are left with the following major picture for punishment reasons: retaliation, clustered mostly among prosocial punishers; competitiveness, typical for spiteful ones; and preemption, common to anyone. To disentangle these, we will make use of the following behavioural model.

6.1 Behavioural modeling framework

The last two hypotheses suggest the following enlargement of the total payoff (utility) of an individual with behavioural components:

$$u_i = V_i - \eta_{1i}\frac{\sum_j \sum_k \varphi_{kij}}{p_{ij}} - \eta_{2i}\sum_j \frac{\mathsf{E}p_{ji}}{p_{ij}} - \pi\left[\eta_{1i}\sum_j\left(p_{ji}|\sum_k(\varphi_{kij})\right) + \eta_{2i}\sum_j(p_{ij}|\mathsf{E}p_{ji})\right] \tag{3}$$

where V_i is the material payoff of player i as given by (2), φ is the dissatisfaction function showing the anger of player i at the contribution of player j associated with the k factors, which are assumed additive, and correspond to either retaliation or competitiveness, $\mathsf{E}p_{ji}$ is expectation of player i of punishment from player j (corresponding to preemption), and π is the cost of punishment. To make sure the funciton is well-defined, we let the behavioural component to be defined whenever punishment is positive, i.e. $p_{ij} > 0 \Leftrightarrow \eta_{ij} > 0$, and zero otherwise. Variants of the k arguments for the φ_{kij} function include:

1. $c_i - c_j$, difference between contribution of player i and j
2. $\bar{c} - c_j$, difference between average contribution in the group and that of player j
3. $\hat{c}_i - c_j$, difference between normative (believed to be appropriate by the player i) contribution and factual contribution of player j
4. $\bar{c} - \hat{c}$, difference between factual and normative average contributions.

other specifications are also possible, including nonlinear functions of these arguments. Latent variables η_{1i} and η_{2i} are individual-specific weights to two factors, retaliation and preemption for prosocial, and separately, competitiveness and preemption for spiteful punishers[13]. Division of utility by p_{ij} is a formalization of the fact that the more player i punishes player j, the lower is player i's dissatisfaction from j's action, while conditioning $|$ in the square brackets mean that punishments p_{ij} are caused by the respective factors. Assuming punishment takes place, and maximizing (3) wrt p_{ij} for the punishment of each j yields

$$\begin{aligned}
\frac{d\pi}{dp_{ij}} &= \eta_{1i} \sum_k \frac{\varphi_{ki}}{p_{ij}^2} + \eta_{2i} \frac{\mathsf{E}p_{ji}}{p_{ij}^2} - \pi \left[\eta_{1i} + \eta_{2i}\right] \Rightarrow \\
0 &= \eta_{1i} \sum_k \frac{\varphi_{ki}}{p_{ij}^2} + \eta_{2i} \frac{\mathsf{E}p_{ji}}{p_{ij}^2} - \pi \Rightarrow \pi p_{ij}^2 = \eta_{1i} \sum_k \varphi_{ki} + \eta_{2i} \sum_k \mathsf{E}p_{ji}
\end{aligned}$$

Further analysis of the model could be simplified by some assumptions about individual perceptions and beliefs concerning each others' behaviour. A key simplification could be the assumption of symmetric punishment strategies, i.e. $p_{ij} = p_{ji}$, which is quite common in many related economic models, such as oligopoly or bargaining games. In our case, this assumption can be justified using data on factual contributions are amazingly similar on average to the projected contributions evaluated in strategic form. To evaluate these, we calculate the factual average contributions by groups, take the ex ante projected contribution of each subject when the average is factual (variables $ex\#$), and take the difference between factual contribution and this projected contribution at factual average. Figure 6.1 shows the distribution of this variable. As can be seen, its modal value is zero for all cities, especially prominent for Moscow and Perm — in other words, on average, even though own projected contributions

[13] Richer models involving more factors are also possible, and even to some extent warrranted by our data. However, for the time being, and for simplicity, we keep attention to the two-factor latent model.

fall short of the factual, participants did contribute as much as they would like, given the factual average contribution in their group. This can be seen as a demonstration of rationality, akin to other rationality results, such as Rapoport e.a. (1998) on entry games or Camerer e.a. (2003) on cognitive hierarchies — here it takes the form of proper forecasting of group cooperation level. Taken together, these findings imply that on average, players understand pretty well the average bidding behaviour of each other; and it is precisely this property which allows us to simplify the derivative of (3) by symmetry of the players. Putting it otherwise, unbiasedness of each player's perception of average contributions implies that, on average, players view others as a copy of themselves within the institution of the game.

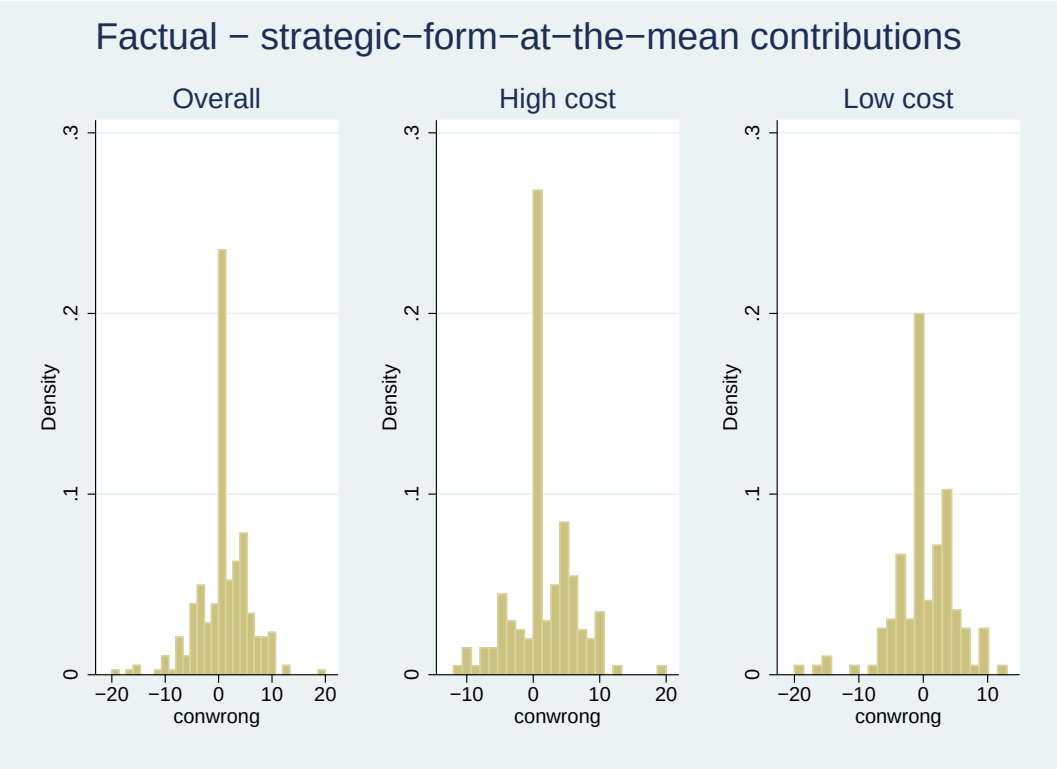

Figure 15: Deviations of factual from projected-at-the-mean contribtuions

This observation allows us to simplify the model, assuming symmetry among all expected and factual punishments. Setting $p_{ij} = \mathsf{E}p_{ji} \equiv p, \forall i, j$, the FOC may be rewritten as

$$\pi p^2 = \eta_{1i} \sum_k \varphi_{ki} + \eta_{2i}(n-1)p \Rightarrow p^* = \eta_{1i} \frac{\sum_k \varphi_{ki}}{\pi p} + \eta_{2i} \frac{n-1}{\pi} \qquad (4)$$

whose parameters can be estmated econometrically as structural model.

In particular, we use the finite mixture model to evaluate individual-specific weights η which correspond to two punishment motives — upset and preemption, which motives can be further disaggregated into competitive motive and

punishment per se, as presented in what follows. In principle, we can think of more categories in the finite mixture regression model, evaluated as non-parametric maximum likelihood (Lindsay, 1995), but for simplicity and at first approximation, we limit attention to these two.

6.2 Model estimates

The finite mixture regression model takes the form

$$pun_i = \beta_0 + \beta_i \boldsymbol{\varphi_i} + \eta_{1i} \boldsymbol{\varphi_i} + \eta_{2i} pcons + \varepsilon_i \tag{5}$$

i.e. punishment size for every individual is modeled as a function of $\boldsymbol{\varphi}_i$, vector of observation-specific explanatory variables corresponding to the upset motive, β_0, constant capturing the effect of remaining (preemptive) motives, and two latent variables η_{1i} and η_{2i}, which measure the contribution of each of these two factors to the desider punishment size, and are functions of the same parameters $\boldsymbol{\varphi_i}$ and $pcons$.

This model is estimated using the two-level GLLAMM model for Stata 10 (Skrondal e.a., 2004), the first level being individual-specific random effect, which allows us to use all obervations (up to 3) for every individual while controlling for interdependence between observations belonging to the same individual. We limit our estimation sample to positive punishments alone, as this allows us to isolate the effect of punishment behaviour per se. Based on the results of Table 9, we take $\varphi_{\mathbf{i}} = expcontrx_i$ (difference between factual and projected own contribution) for model of spiteful punishment, and $\varphi_{\mathbf{i}} = [diffcontr_i, expcontravg_i]$ (differences between contributions of punisher and punished subject, and between punisher's factual and normatively projected contributions) for prosocial one. These variables are transformed as specified in (4), and interpreted as 'competitive' and 'retaliation' motives, respectively; complementary, 'preemptive' component in both models is defined as a constant $pcons = (n-1)/\pi$.[14]

Estimates of the model are in Table 10, separately for prosocial and spiteful samples. The main purpose of the two models is classification of the punishment motives through latent random effects, separately for prosocial and spiteful punishers. The estimated probabilities show that most of the prosocial punishments (72%) are related to the second random effect, associated with preemption, while a minority (27%) is due to retaliation. On the contrary, most of the spiteful punishments (82%) are related to competitive motive, and only a minority is due to preemption.

Further qualifications of this model can be obtained using predicted probabilities of the lambdas for each individual. Table 11 presents a split of the various variables wrt two variables: whether punishment is spiteful or prosocial, and whether the predicted probabilities of η_1 is less than 0.5 (punishments driven primarily by preemption motive) or $\eta_1 \geq 0.5$ (punishments caused by upset; probabilities η_2 are complementary, and convey the same information). All differences in this table are significant at 5% level at least, and the differences

[14] Besides this, we have also tried a number of alternative specifications, including pure finite mixed model and multilevel model with group effects (level 3 in terms of GLLAMM). However, for the present analysis we concentrate on individual behaviour, and so limit attention to 2-level models.

Table 10: Estimation results : gllamm

	Prosocial		Spiteful	
Variable	Coeff.	Std.Err.	Coeff.	Std.Err.
Equation for pun				
Intercept β_0	4.653	(0.312)	4.143	(0.052)
pdiffcontr	0.125	(0.033)		
pexpcontravg	-0.060	(0.0265)		
pexpcontrx			0.059	(0.054)
Residual variance				
Intercept	1.924	(0.228)	1.806	(0.322)
Loadings for locus 1 (regret)				
cons	1	(fixed)	1	(fixed)
pdiffcontr	0.043	(0.009)		
pexpcontravg	-0.019	(0.006)		
expcontrx			-0.017	(0.013)
Locations of random effects				
	Intercept	Slope	Intercept	Slope
locus1	4.372	-0.087	-1.339	0.003
locus2	-1.671	0.033	6.151	-0.014
Covariance matrix of random effects				
	locus1	locus2	locus1	locus2
locus1	7.306	-	8.237	-
locus2	-0.146	0.002	-0.019	0.0001
Probabilities of random effects				
η_1	0.276		0.821	
η_2	0.724		0.179	
Log-likelihood				
	-318.49		-124.04	

are telling. First, we see that high probabilities of competition vs. preemption are associated with much higher (closer to maximal) sizes of punishment, be it prosocial or spiteful. Second, by far the maximal insurance against punishment is typical of just one category — spiteful competitive punishers. Third, the largest underpayment of the punished player vs. the normative (variable *diffexpavg*) occurs for prosocial retaliative punishers. This is pretty much in line with the conventional wisdom; but perhaps less obvious is the fact that the lowest underpayment of the punished player vs. the normative (zero in the median, or no deviations) corresponds to spiteful punishments unrelated to ethical motives, but driven by competition. Finally, subjects who expect average contributions are the lowest (variable *contrib*) also tend to punish spitefully and preemptively, which means they are rather afraid of being punished themselves than willing to punish those who are too cooperative.

Table 11: **Breakdown statistics by estimated punishment motives**

stats	*pun*	*insp*	*difexpavg*	*contr*
Type 1: Prosocial, $\eta_1 < 0.5$, N=131				
mean	3.45	2.5	7.58	3.84
p50	3	2	8	3
sd	1.74	2.18	5.36	3.88
Type 2: Prosocial, $\eta_1 \geq 0.5$, N=26				
mean	9.73	1.28	11.04	2.92
p50	10	0	11	2.5
sd	.66	2.70	6.16	2.99
Type 3: Spiteful, $\eta_1 < 0.5$, N=47				
mean	2.57	2.5	-.85	10.04
p50	2	3	0	4
sd	1.66	1.92	6.27	10.83
Type 4: Spiteful, $\eta_1 \geq 0.5$, N=17				
mean	10	7.38	2.94	6.37
p50	10	8	3	5
sd	0	3.15	7.43	4.19

These observations allow us to construct the following fourfold classification, which in Table 11 is denoted types 1 through 4. Type 1 consists of those instances which are prosocial: on average the punished player contributed by 8 less than punishing one. However, the size of punishment is quite modest, and our model tells that this is because of expected preemption: members of that class are unwilling to bear risks of being retaliated, and do not want to spend on insure either. All in all, this group (in our sample, the most numerous, accounting for almost 60% of all punishments) consists of those instances where the punisher feels unhappy about what the punished subject contributed, but he is afraid or unwilling to punish — in short, this is the strategy of typical 'small people', or **philistine** mass who are not deprived of some values, but are not ready to fight for them.

By contrast, type 2 consists of **fair** prosocial people who are upset by the low contribution of those who they punish. High mean punishment corresponds for

Table 12:

stats	contrib	cexpx	homexp	pun	insp	from
Retaliating prosocial — 12% (M 13%, P 4%, T 13%)						
mean	10.16	-10.51	4.61	9.54	1.9	.65
p50	10	-9	4	10	0	1
sd	4.04	5.868	5.37	1.09	3.59	.48
Preemptive prosocial — 59% (M 56%, P 67%, T 58%)						
mean	8.85	-7.02	3.00	3.43	2.03	.42
p50	8	-6	2	3	2	0
sd	4.96	4.85	4.72	1.94	2.09	.49
Competitive spite — 11% (M 15%, P 0%, T 12%)						
mean	1.37	1.82	2.06	9.65	6.5	.23
p50	1	1	2	10	5	0
sd	2.029	6.25	5.16	1.284	2.74	.42
Preemptive spite — 18% (M, T 16%, P 30%)						
mean	4.69	2.85	1.85	2.65	2.25	.3
p50	5	2	0	2	2.5	0
sd	3.70	5.49	5.50	1.71	1.88	.47

them with low insurances, that is, they are not only upset at poor cooperativeness of their counterparts, but also are not afraid of retaliation from the others: they punish, and believe they are on their rights. This group is not numerous — about 12%, but one may safely say that they perhaps would constitute the core of active civil society which is ready to fight for their prosocial values and to stop others from breaking them (their behaviour may be called 'preemptive' in this particular sense).

Type 3 is preemptive too, but unlike the first type, their punishment is spiteful; further, they punish primarily those who have contributed a lot more than they did (cf. difference between variables *diffcontrib* and *contr* in this Table), being also yet more angry at the need to spend some of their own resources for that sake. This punishment is hardly norm-driven — remember their punishment for those who comply with the norm. This type may be properly called **spiteful** or 'envious' at those who have been more in line with the social norm than they were themselves — note, however, theat their spite is not really active, and their share in the population is only about 20%.

Finally, type 4, being the smallest (under 8%), is also the most interesting. First, it consists exclusively of extremal spiteful punishers by 10, with no variations, and up to the point that all but one instance of maximum spiteful punishments fall into this category. These players are also clearly competitiveness-driven, but they are also extremely unwilling to bear risk, buying more insurance than anyone else. This type may be termed **aggressive** spiters, perhaps closest in its spirit to the Russian 'bratki' — groupds of gangsters who are willing to be at the top at the expense of the others, but only to the extent this is not dangerous for their own well-being.

This classification can be further illustrated graphically. Using the estimated individual probabilities, we can caluclate utilities of each particular participant as given by (3) whose plot against punishments is provided in Figure 16. Top

panel shows utilities of philistine prosocial punishers, which has inverse U-shape in the punishment-utility plane. Most people in this category are of type 1, who are somewhat upset by low contributions of the others, hence their utility is increasing up to the maximum in a somewhat 'wave-shaped' way. Later on, risks and cost of punishment becomes too large for them relatively to their upset (in which they are not very determined anyway), hence utility is decreasing, and almost none of these players applies punishment. However, at the right edge comes a cluster of players of type 2 (fair prosocial), whose utility is lower for another reason: they are very much unhappy about what the others have contributed and, unlike the first type, they are happy to punish to the extreme (and in fact, would be happy to punish by more had they been able to pass the upper threshold of 10).

The bottom panel shows the pattern for spiteful punishers, which is generally U-shaped, and also consists of two groups. At first, a decreasing scatter of points descends from left to right, corresponding to lowering dissatisfaction of players of type 3 (spiteful in proper sense of the word), who would generally benefit from boosting their pride by lowering utilities of the others, but not at a cost to themselves — these are competitive punishers. Finally, at the right edge, again comes a cohort of sensored punishments of type 4 — the same players who report desire to hurt other people, and are characterized by serial punishments, who are preemptive: over 85% of them are willing to trade their punishments for insurance, revealing that this aggression may itself be a manifestation of fear of punishment from the others.

6.3 Summary and conclusions

To summarize our results, it follows that the apparently intuitive explanation of 'punishment' behaviour in PG games is not as straightforward as it appears. Our experiment and its analysis suggest the fourfold classification of punishment motives in the Russian society. Strategic choices of the subjects and their non-parametric clusterization warrant fourfold classification of our sample. Most of the subjects (almost 3/4) share prosocial values of cooperation, however, an overwhelming majority of them (60% of the entire population) behave like philisines: sharing them in principle, they are unwilling to fight for them; and only a large minority of these are fair in their actions (and have lower utility as a result of this!). A minority of under 1/4 of population generally do not share prosocial values, but a minority of these does punish mostly for the sake of not being deemed 'losers' — as a matter of fear or preemption; and only the rest (20% of the total population) can be deemed 'spiteful' in proper sense of the word. Albeit this fourfold classifcation has been obtained on a nonrepresentative sample and warrants further confirmation and support, it is remarkably similar to the composition of the contemporary Russian society, and may be viewed as its momentary portrait obtained by experimental methods.

The main implication of this work is that punishment in PG context at least, should not always be interpreted as a revelation of dissatisfaction with contributions of the other players. In experimental games, players may have a much larger variety of motives, and researchers should be cautious in attibuting it to the most apparent of these, no matter how robust is this in the light of the existing literature.

One more implication is that the multiplicity of the principles on which 'pun-

ishment' behaviour may rest. In Russia, these were quite heterogeneous, while in Western Europe, for instance, interpretations of deductions as 'punishments for antisocial actions' appears to be more straightforward, and closer to THE explanation. Comparison of these factors across countries may be interesting and important for the diagnosis of the state of the respective societies.

The last conclusion we draw is more general. To economists, it is customary to think of human behaviour primarily in comparison to some specific benchmark, be it substantive (neoclassical) of behavioural. While illuminating theoretically, and even necessary to structure our thoughts, this approach may be misleading when it's a matter of understanding the reasons and motives of real behaviour in experiments, as well as in broader ecological contexts. Yet however mad the behaviour might appear from the viewpoint of the prevailing theories, if there's a method in it, our goal as of positive scientists is to do our best to understand and explain it.

References

[1] Anderson, Christopher M. and Louis Putterman (2006) Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Games and Economic Behavior, v.54(1), p. 1–24.

[2] Andreoni, James (1995). Cooperation in Public-Goods Experiments: Kindness or Confusion? American Economic Review, v.85(4), p.891-904.

[3] Bardsley, Nicholas (2000) Cournot without deception: individual behaviour in free-riding experiments revisited. Experimental Economics, v.3, p.215-240.

[4] Bardsley, Nicholas, and Peter G.Moffatt (2007). The experimetrics of public goods: inferring motivation from contribution. Theory and Decision, v.62, p.161-193.

[5] Bochet, Olivier, Talbot Page and Louis Putterman (2006) Communication and punishment in voluntary contribution experiments. Journal of Economic Behavior and Organization, v.60(1), p.11-26.

[6] Bolton G. and Ockenfels A. (2000) A theory of equity, reciprocity and competition. American Economic Review, v.90(1), p.

[7] Bochet, Olivier, Talbott Page and Louis Putterman (2006). Communication and Punishment in Voluntary Contribution Experiments, Journal of Economic Behavior and Organization, v.60(1), p.11-26.

[8] Brandts, Jordi, and Arthur Schram (2001). Cooperation and noise in public goods experiments: applying the contribution function approach. Journal of Public Economics, v.79, p.399-427.

[9] Brosig, J., J. Weimann, C.-L.Yang (2003). The hot versus cold effect in a simple bargaining experiment. Experimental Economics, 6, 75-90.

[10] Camerer, Colin, T. Ho, J. Chong. (2003) A Cognitive Hierarchy Model of Behavior in Games. Quarterly Journal of Economics, v.119(3), p.861-98.

[11] Carpenter, Jeffrey (2004). When in Rome: conformity and the provision of public goods. Journal of Socio-Economics, Elsevier, vol. 33(4), pages 395–408.

[12] Carpenter, Jeffrey (2007a). Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods. Games and Economic Behavior, v.60(1), p.31-52.

[13] Carpenter, Jeffrey (2007b). The Demand for Punishment, Journal of Economic Behavior and Organization, v.62(4), p.522-542.

[14] Casari, Marco (2005). On the Design of Peer Punishment Experiments. Experimental Economics, v.8(2), p.107-115.

[15] Casari, Marco, and Luigi Luini (2005) Group cooperation under alternative peer punishment technologies: an experiment. Working paper 1176, Purdue University.

[16] Cason, Timothy, Tatsuyoshi Saijo and Takehiko Yamato (2002). Voluntary Participation and Spite in Public Good Provision Experiments: An International Comparison. Experimental Economics, v.5(2), p.133-153.

[17] Cason, Timothy N., Tatsuyoshi Saijo, Takehiko Yamato and Konomu Yokotani (2004) Non-excludable public good experiments. Games and Economic Behavior, v.49(1), p.81-102.

[18] Chaudhuri, Ananish (2011) "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature," Experimental Economics, Springer, vol. 14(1), p.47-83.

[19] Cinyabuguma, Matthias, Page, Talbot and Louis Putterman (2005). Cooperation under the threat of expulsion in a public goods experiment. Journal of Public Economics, v.89(8), p.1421-1435.

[20] Cinyabuguma, Matthias, Page, Talbot and Louis Putterman (2006). Can second-order punishment deter perverse punishment? Experimental Economics, v.9(3), p.265-279.

[21] Croson, Rachel (2000). Thinking like a game theorist: Factors affecting the frequency of equilibrium play. Journal of Economic Behavior and Organization, 41, 299–314.

[22] Croson, Rachel (2007). Theories of commitment, altruism and reciprocity: Evidence from linear public goods games. Economic Inquiry, 45, 199–216.

[23] Denant-Boemont, Lauren, David Masclet and Charels Noussair (2007). Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment. Economic Theory, v.33(1), p.145-167.

[24] Dufwenberg Martin and Georg Kirchsteiger (2004) A Theory of Sequential Reciprocity. Games and Economic Behavior, v.47(2), p.268-298.

[25] Ertan, Arhan, Talbot Page and Louis Putterman (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. European Economic Review, v.53(5), p.495–511.

[26] Falk, Armin, and Urs Fischbacher (2006). A theory of reciprocity. Games and Economic Behavior, v.54, p.293-315.

[27] Falk, Armin, and Ernst Fehr and Urs Fischbacher (2001). Driving Forces of Informal Sanctions," IEW - Working Papers 059, Institute for Empirical Research in Economics - University of Zurich.

[28] Fehr, Ernst and Fischbacher, Urs. (2004). Third Party Punishment and Social Norms. Evolution and Human Behavior, 2004, 25, pp. 63-87.

[29] Fehr, Ernst, and Simon Gächter (2000). Cooperation and Punishment in Public Goods Experiments, American Economic Review, v.90, p.980-994.

[30] Fehr, Ernst, and Simon Gächter (2002). Altruistic Punishment in Humans, Nature 415, 137-140.

[31] Fehr, Ernst, and Klaus Schmidt (1999). A Theory of Fairness, Competition and Co-operation. Quarterly Journal of Economics v.114, p.817-868.

[32] Fehr, Ernst, and Klaus Schmidt (2003). Theories of Fairness and Reciprocity: Evidence and Economic Applications, In: Dewatripont, M. et al. (eds), Advances in Economic Theory, Eighth World Congress of the Econometric Society, Vol. I, 208-257, Cambridge: Cambridge University Press.

[33] Ferraro, Paul J. and Christian Vossler (2010). The Source and Significance of Confusion in Public Goods Experiments. B.E. Journal of Economic Analysis and Policy (Contributions) 10(1), article 53.

[34] Fischbacher, Urs (2007) .z-Tree: Zurich Toolbox for Ready-made Economic Experiments, Experimental Economics 10 (2), 171-178.

[35] Fischbacher, Urs, and Simon Gächter (2010). Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments. American Economic Review, 100(1), 541-556.

[36] Fischbacher, Urs, Simon Gächter and Ernst Fehr (2001). Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters, 71, 397–404.

[37] Fowler, James H. (2005) Altruistic punishment and the origin of cooperation. Proc Natl Acad Sci U S A., v.102(19), p.7047–7049.

[38] Gächter, Simon and Herrmann, Benedikt (2011). "The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. European Economic Review, vol. 55(2), p.193-210.

[39] Gächter, Simon, Elke Renner, and Martin Sefton (2008). The Long-Run Benefits of Punishment. Science, 322(5907), 1510.

[40] Gächter, Simon, and Elke Renner (2010). The effects of (incentivized) belief elicitation in public goods experiments. Experimental Economics, v.13, p.364–377.

[41] Gürerk, Özgür, Irlenbusch, Brend and Bettina Rockenbach (2006). The Competitive Advantage of Sanctioning Institutions. Science 312: 108-110.

[42] Herrmann, Benedikt, and Simon Gächter (2006). The Limits of Self-Governance in the Presence of Spite: Experimental Evidence from Urban and Rural Russia. Working Paper, University of Nottingham.

[43] Herrmann, Benedikt, and Simon Gächter. Reciprocity, Culture and Human Cooperation: Previous Insights and Cross-Cultural Experiment. Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 364, March 2009, pp. 791-806.

[44] Herrmann, Benedikt, Christian Thöni, Simon Gächter (2008). Antisocial Punishment Across Societies. Science, v.319, pp.1362-1367

[45] Herrmann, Benedikt, and Christian Thöni (2009).Measuring conditional cooperation: a replication study in Russia. Experimental Economics, v.12, p.87–92.

[46] Hoff, Karla, Mayuresh Kshetramade and Ernst Fehr (2011). Caste and Punishment: the Legacy of Caste Culture in Norm Enforcement. Economic Journal, v.121(556), p.F449-F475.

[47] Houser, Daniel and Robert Kurzban. (2002). Revisiting kindness and confusion in public good experiments. American Economic Review, 92, 1062-1069.

[48] Kamei, Kenju, Louis Putterman and Jean-Robert Tyran (2011) State or Nature? Formal vs. Informal Sanctioning in the Voluntary Provision of Public Goods. University of Coppenhangen Discussion Paper 11-05.

[49] Lindsay, B. G. (1995) Mixture models: theory, geometry and applications. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5., Institute of Mathematical Statistics, Hayward, CA.

[50] Masclet, David, Charles Noussair, Steven Tucker and Marie-Claire Villeval. Monetary and nonmonetary punishment in the voluntary contributions mechanism. American Economic Review, v.93, p.366-380.

[51] Marlowe FW, Berbesque JC, Barrett C, Bolyanatz A, Gurven M, Tracer D. (2011). The 'spiteful' origins of human cooperation. Proc Biol Sci. v.278(1715), p.2159-64.

[52] Nikiforakis, Nikos (2008). Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" Journal of Public Economics, v.92, p.91-112.

[53] Nikiforakis, Nikos (2010). Feedback, punishment and cooperation in public good experiments. Games and Economic Behavior, v.68, p.689-702.

[54] Noussair, Charles and Stephen Tucker (2005). Combining Monetary and Social Sanctions to Promote Cooperation, Economic Inquiry 43(3), 649-660.

[55] Ones, Umut and Louis Putterman (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation. Journal of Economic Behavior and Organization, v. 62(4), p.495-521.

[56] Ostrom, Elinor, James M. Walker, and Roy Gardner (1992). Covenants with and without a Sword–Self–Governance Is Possible. American Political Science Review, v.86, p.404-417.

[57] Page, Talbot, Louis Putterman, and B.Unel (2005). Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency. Economic Journal, v.115 (506), p.1032-1053.

[58] Page, Talbot, Louis Putterman and Bruno Garcia (2008) Getting Punishment Right: Do Costly Monitoring or Redistributive Punishment Help? Working paper, Brown University.

[59] Putterman, Louis, Jean-Robert Tyran, and Kenju Kamei (2010) Public Goods and Voting on Formal Sanction Schemes: An Experiment. University of Coppenhagen Discussion Paper 10-02.

[60] de Quervain, D. J. F., Urs Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, A., Ernst Fehr (2004). The Neural Basis of Altruistic Punishment, Science 305 (5688), 1254-1258.

[61] Rapoport, Amnon, Darryl A.Seale, Ido Erev and James Sundali (1998). Equilibrium play in large group market entry games. Management Science, v.44(1), p.129-141.

[62] Saijo, Tatsuyoshi (2008). Spiteful Behavior in Voluntary Contribution Mechanism Experiments. In: C.Plott and V.Smith, eds. Handbook of Experimental Economics Results, Elsevier.

[63] Saijo, Tatsuyoshi and Takehiko Yamato (1999). A Voluntary Participation Game with a Non-Excludable Public Good. J. Econ. Theory,, v.84, p.227-242.

[64] Sefton, Martin, R. Shupp, and James Walker (2007). The Effect of Rewards and Sanctions in Provision of Public Goods, Economic Inquiry, v.45, p.671-690.

[65] Skrondal, Anders, Sophia Rabe-Hesketh and Andrew Pickles (2004). GLLAMM textbook.

[66] Houser, Daniel and Erte Xiao (2010). Inequality-seeking punishment. Economics Letters, v.109(1), p.20–23.

[67] Xiao, Erte, and Daniel Houser (2010) Punish in public. Journal of public economics, 95(7-8), pages 1006-1017

[68] Yamagishi, Toshio (1986). The Provision of a Sanctioning System as a Public Good. Journal of Personality and Social Psychology, v.51, p.110-116.

[69] Zizzo, Daniel J. (2010). Experimenter demand effects in economic experiments. Experimental Economics, v.13(1), p.75-98.

[70] Zizzo, Daniel J. (2011). Do Dictator Games Measure Altruism? Manuscript for the Handbook on the Economics of Philanthropy, Reciprocity and Social Enterprise, ed. by Luigino Bruni and Stefano Zamagni.

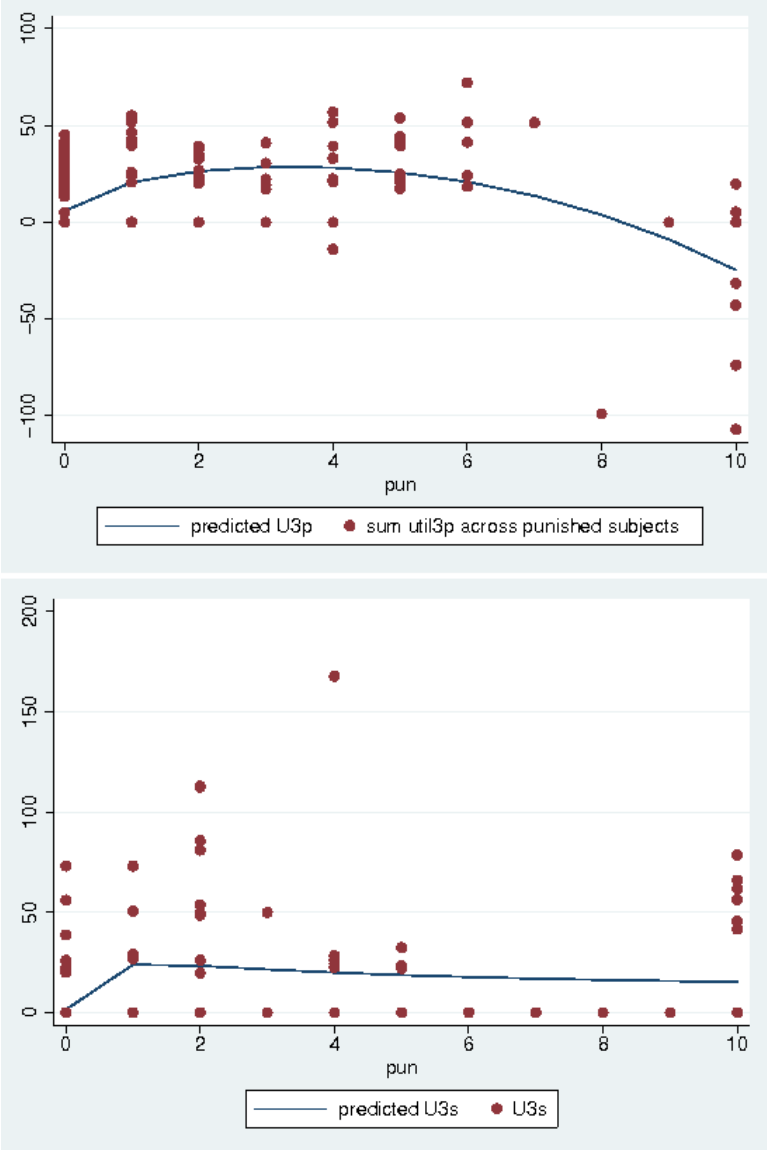

Figure 16: Utility profiles for prosocial (top) and spiteful (bottom) punishments

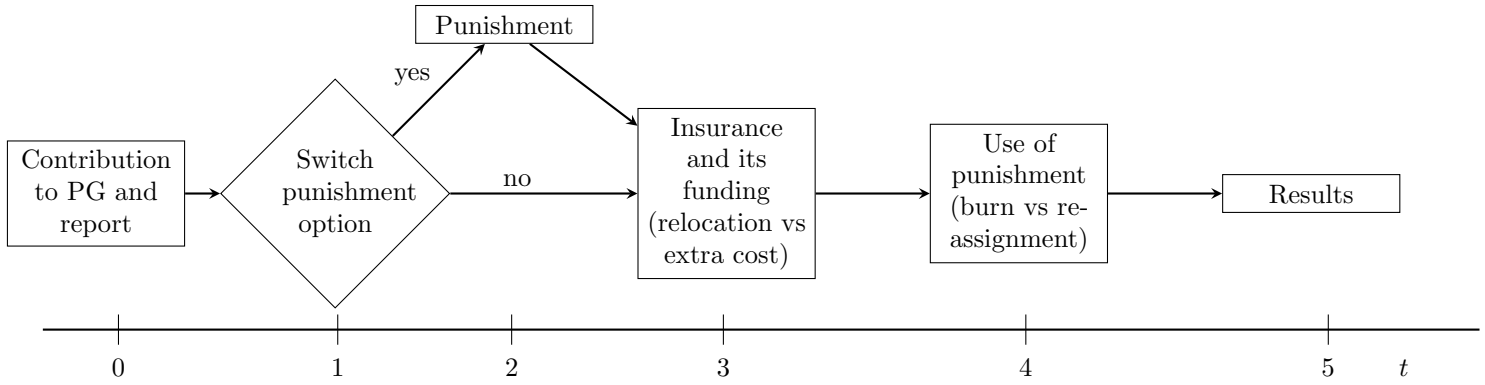

Figure 17: Timeline of the experiment under low cost, stage 2