Individual incentives and collective goods: experimental results at HSE

by Alexis Belianin[1],
based on a series of joint works with Fuad Aleskerov, Kirill Pogorelskiy, Leon Kosals and other colleagues at the Laboratory of Experimental and Behavioural Economics
http://epee.hse.ru

[1]ICEF at HSE, and
Institute for World Economy and International Relations, Moscow, Russia

19 October 2011
Prepared for the Moscow Mathematical Seminar

Measures of bargaining power

- Voting in bodies matters in many contexts (boards of directors, shareholders' meetings, parliaments, United Nations...)

Measures of bargaining power

- Voting in bodies matters in many contexts (boards of directors, shareholders' meetings, parliaments, United Nations...)
- Formally, voting can be viewed as an instance of *simple* splitting-the-pie cooperative game

Measures of bargaining power

- Voting in bodies matters in many contexts (boards of directors, shareholders' meetings, parliaments, United Nations...)
- Formally, voting can be viewed as an instance of *simple* splitting-the-pie cooperative game
- Theoretical models of voting power — both classical (Shapley and Shubik, 1954; Banzhaf, 1965) and more recent (Aleskerov, 2006) are based on the calculations of values in such games

Measures of bargaining power

- Voting in bodies matters in many contexts (boards of directors, shareholders' meetings, parliaments, United Nations...)
- Formally, voting can be viewed as an instance of *simple* splitting-the-pie cooperative game
- Theoretical models of voting power — both classical (Shapley and Shubik, 1954; Banzhaf, 1965) and more recent (Aleskerov, 2006) are based on the calculations of values in such games
- Experimental evidence (Montero, Sefton, Zhang, 2008; Aleskerov, Belianin, Pogorelskiy, 2009) gives only partial support to this approach: specifically, real subjects tend to concentrate on minimal coalitions and choose some specific allocaions among many possible ones.

Measures of bargaining power

- Voting in bodies matters in many contexts (boards of directors, shareholders' meetings, parliaments, United Nations...)
- Formally, voting can be viewed as an instance of *simple* splitting-the-pie cooperative game
- Theoretical models of voting power — both classical (Shapley and Shubik, 1954; Banzhaf, 1965) and more recent (Aleskerov, 2006) are based on the calculations of values in such games
- Experimental evidence (Montero, Sefton, Zhang, 2008; Aleskerov, Belianin, Pogorelskiy, 2009) gives only partial support to this approach: specifically, real subjects tend to concentrate on minimal coalitions and choose some specific allocaions among many possible ones.
- Existing solution concepts (Shapley, 1953; Aumann and Maschler, 1965; Schmeilder, 1969; Myerson, 1977) tend to neglect this property of the bargaining process.

Measures of bargaining power

- Voting in bodies matters in many contexts (boards of directors, shareholders' meetings, parliaments, United Nations...)
- Formally, voting can be viewed as an instance of *simple* splitting-the-pie cooperative game
- Theoretical models of voting power — both classical (Shapley and Shubik, 1954; Banzhaf, 1965) and more recent (Aleskerov, 2006) are based on the calculations of values in such games
- Experimental evidence (Montero, Sefton, Zhang, 2008; Aleskerov, Belianin, Pogorelskiy, 2009) gives only partial support to this approach: specifically, real subjects tend to concentrate on minimal coalitions and choose some specific allocaions among many possible ones.
- Existing solution concepts (Shapley, 1953; Aumann and Maschler, 1965; Schmeilder, 1969; Myerson, 1977) tend to neglect this property of the bargaining process.
- How can the process of bargaining be described theoretically, and how can its outcome be predicted?

Main issues

- How to measure voting power in multilateral bargaining

Main issues

- How to measure voting power in multilateral bargaining
- What changes if parties in bagraining have non-uniform preferences

Main issues

- How to measure voting power in multilateral bargaining
- What changes if parties in bagraining have non-uniform preferences
- Can the outcome of the bargaining process be predicted?

Main issues

- How to measure voting power in multilateral bargaining
- What changes if parties in bagraining have non-uniform preferences
- Can the outcome of the bargaining process be predicted?
- What are the strategic incentives of players in this process?

Main issues

- How to measure voting power in multilateral bargaining
- What changes if parties in bagraining have non-uniform preferences
- Can the outcome of the bargaining process be predicted?
- What are the strategic incentives of players in this process?
- What kind of theoretical concepts can be used to explain the bargaining outcome, given the evidence about bargaining process?

Voting power: main notions

- N — set of agents (players), $|N| = n$, with generic player i
- $w_i > 0$ — number of votes i possesses
- q — quota (minimum number of votes for a bill to pass)

Voting power: main notions

- N — set of agents (players), $|N| = n$, with generic player i
- $w_i > 0$ — number of votes i possesses
- q — quota (minimum number of votes for a bill to pass)
- Coalition $S \subseteq 2^N$ is *winning* iff $\sum_{i \in S} w_i \geq q$ (denote $|S| = s$ and let W be the set of all winning coalitions)
- $v(S)$ — payoff to the coalition S. Let $v(S) = 1$ iff $S \in W$; $v(S) = 0$ iff $S \notin W$

Voting power: main notions

- N — set of agents (players), $|N| = n$, with generic player i
- $w_i > 0$ — number of votes i possesses
- q — quota (minimum number of votes for a bill to pass)
- Coalition $S \subseteq 2^N$ is *winning* iff $\sum_{i \in S} w_i \geq q$ (denote $|S| = s$ and let W be the set of all winning coalitions)
- $v(S)$ — payoff to the coalition S. Let $v(S) = 1$ iff $S \in W$; $v(S) = 0$ iff $S \notin W$
- Player $i \notin S$ is *pivotal* for the coalition S iff S is losing, while $S \cup \{i\}$ is not (thus, i is *decisive*)

Classical power indices

- Banzhaf (1965): $\beta_i = \frac{\sum_{S \subseteq N \setminus \{i\}} (v(S) - v(S \cup \{i\}))}{\sum_{j=1}^{N} \sum_{S \subseteq N \setminus \{j\}} (v(S) - v(S \cup \{j\}))}$
  This is a share of player i's decisiveness in the total decisiveness.

Classical power indices

- Banzhaf (1965): $\beta_i = \frac{\sum_{S \subseteq N \setminus \{i\}} (v(S) - v(S \cup \{i\}))}{\sum_{j=1}^{N} \sum_{S \subseteq N \setminus \{j\}} (v(S) - v(S \cup \{j\}))}$
  This is a share of player i's decisiveness in the total decisiveness.

- Shapley-Shubik (1954):
  $\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} (v(S) - v(S \cup \{i\}))$.
  This is the share of permutations of all coalitions S in which player i is pivotal in the total number of permutations, i.e. the Shapley value for the cooperative voting game.

Preference-based power indices

(Aleskerov, 2006). Assume we know the preference profile of each player i about coalescing with any other player: $P_i = (p_{i1}, ..., p_{in})$.

Let p_{ij} be (ordinal or cardinal) *measure* of, or *explicit modifiers* of player i's preferences towards coalescing player j.

Preference-based power indices

(Aleskerov, 2006). Assume we know the preference profile of each player i about coalescing with any other player: $P_i = (p_{i1}, ..., p_{in})$.

Let p_{ij} be (ordinal or cardinal) *measure* of, or *explicit modifiers* of player i's preferences towards coalescing player j.

Let $f_i(S) : \times_{j \in S} P_j \to \mathbb{R}$ be the *intensity of connections* of player i with other members of the winning coalition S she is part of.

Preference-based power indices

(Aleskerov, 2006). Assume we know the preference profile of each player i about coalescing with any other player: $P_i = (p_{i1}, ..., p_{in})$.

Let p_{ij} be (ordinal or cardinal) *measure* of, or *explicit modifiers* of player i's preferences towards coalescing player j.

Let $f_i(S) : \times_{j \in S} P_j \to \mathbb{R}$ be the *intensity of connections* of player i with other members of the winning coalition S she is part of.

Let $\chi_i = \sum_{S \subseteq N \setminus \{i\}} f_i(S) \left(v(S) - v(S \cup \{i\})\right)$ be the sum of intensities of connection of player i over all the winning coalitions in which she is pivotal.

Preference-based power indices

(Aleskerov, 2006). Assume we know the preference profile of each player i about coalescing with any other player: $P_i = (p_{i1}, ..., p_{in})$.

Let p_{ij} be (ordinal or cardinal) *measure* of, or *explicit modifiers* of player i's preferences towards coalescing player j.

Let $f_i(S) : \times_{j \in S} P_j \to \mathbb{R}$ be the *intensity of connections* of player i with other members of the winning coalition S she is part of.

Let $\chi_i = \sum_{S \subseteq N \setminus \{i\}} f_i(S) \left(v(S) - v(S \cup \{i\})\right)$ be the sum of intensities of connection of player i over all the winning coalitions in which she is pivotal.

Similarly to the Banzhaf index, let

$$\alpha_i = \frac{\sum_{S \subseteq N \setminus \{i\}} f_i(S)(v(S) - v(S \cup \{i\}))}{\sum_{j=1}^{N} \sum_{S \subseteq N \setminus \{j\}} f_j(S)(v(S) - v(S \cup \{i\}))} = \frac{\chi_i}{\sum_{j=1}^{N} \chi_j}$$

Power indices with preferences

Particular forms of the intensity of connections functions include

- $f_i^{\times}(S) = \prod_{j \in S \setminus \{i\}} p_{ij}$ — multiplicative intensity of i's preferences.
- $f_i^{\div}(S) = \prod_{j \in S \setminus \{i\}} p_{ji}$ — dual multiplicative intensity.
- ... and many others.

Experimental design (Montero, Sefton, Zhang, Soc Choice Welfare, 2008)

- Unstructured bargaining game in groups of 3 or 4 players (12 or 16 participants per session).
- In each round of each game the players of a group decide on how to divide 120 points among them. Each player can post at most one offer at a time, and can vote for any offer on the board.
- The first offer to meet the quota is accepted, and the players receive the corresponding number of points unless they fail to come to an agreement within 300 seconds, in which case all receive 0 points.
- All players are randomly rematched from round to round.

Features of our experiment (Aleskerov, Belianin, Pogorelskiy, 2009)

- 2 games are played in each experimental session in randomized block order.
- With or without preferences (explicit modifiers).
- All games were played at HSE campus during October 2008 - May 2009, using specially developed experimental software.
- Participants - 136 students at various department, gender composition 50-50, average age 19.1 years.
- Gains of participants in 10-round games: average 7.62 EUR, minimum — 3.81 EUR, maximum — 13.68 EUR; gains in 20-round games: average 10.65 EUR, minimum 5.38 EUR, maximum 16.81 EUR per 1- to 1.5-hour session.

Screenshot of a typical game Standard (S)

Player number	1	2 (You)	3
Votes	3	2	2
Proposed shares	60	60	

Submit your proposal

- Instruction
- Shares should sum up to 120
- You can replace your proposal by a newly submitted one
- 4 votes are required to pass a proposal
- You are marked in red where applicable
- Please note that your login name shown at the bottom of this page is NOT your in-game player number! Also beware that your in-game player number may change between rounds

121 seconds left

Player #2's proposal (Total votes accumulated: 2)

Player number	1	2 (You)	3	
Votes	3	2	2	You have voted for this proposal
Proposed shares	45	75	0	
Acceptance		Y		

Player #3's proposal (Total votes accumulated: 2)

Player number	1	2 (You)	3	
Votes	3	2	2	Vote for this proposal
Proposed shares	20	30	70	
Acceptance			Y	

Games S–1 (Standard) [4; 3, 2, 2]

Game S: quota is 4 votes

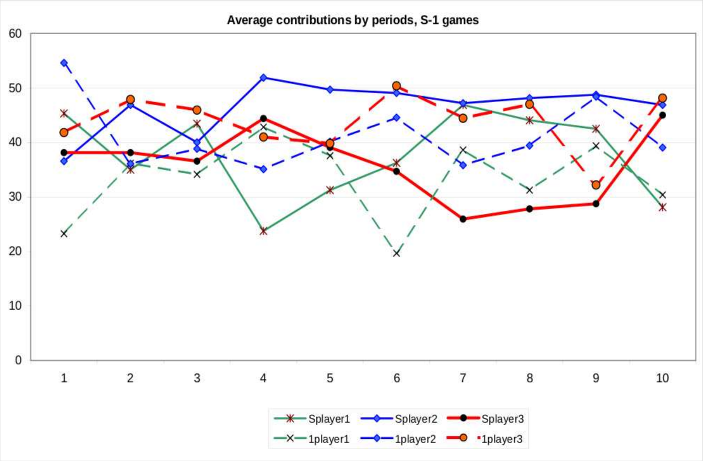
player#	1	2	3
votes	3	2	2

Winning coalitions: $W = \{\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$.
Banzhaf (and Shapley-Shubik) index: $\beta_1 = \beta_2 = \beta_3 = 1/3$, predicting that all players get around **40 pts** each. **Game 1** uses the following *explicit modifiers*:

	1	2	3
1	-	1	1
2	1	-	1.01
3	1	1	-

α indices based on the $f^{\div}$ intensity function:
$\alpha_1 = \alpha_2 = 0.3328, \alpha_3 = 0.3344$

The S-1 games


Average contributions by periods, S-1 games
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Splayer1
Splayer2
Splayer3
1player1
1player2
1player3

The S-1 games

- Player 3 on average receives systematically more in the 1–treatment (43.85) than in S–treatment (35.84), which difference is significant. Hence **explicit modifiers work** for player 3: 'being loved is better than love'.
- There are no treatment effects for players 1 and 2, but ...

The S-1 games

- Player 3 on average receives systematically more in the 1–treatment (43.85) than in S–treatment (35.84), which difference is significant. Hence **explicit modifiers work** for player 3: 'being loved is better than love'.
- There are no treatment effects for players 1 and 2, but ...
- Player 2 receives systematically more than either of the other players in the S-treatment (46.5 vs. 35.84 or 37.66), the difference being significant.
  - Same effect as in MSZ, who attribute it to 'framing effect'
  - We attribute it to the position of player 2 in the middle of the table on the screen: player 2 has two neighbours (1 and 3), whereas the other two players — just one (player 2).

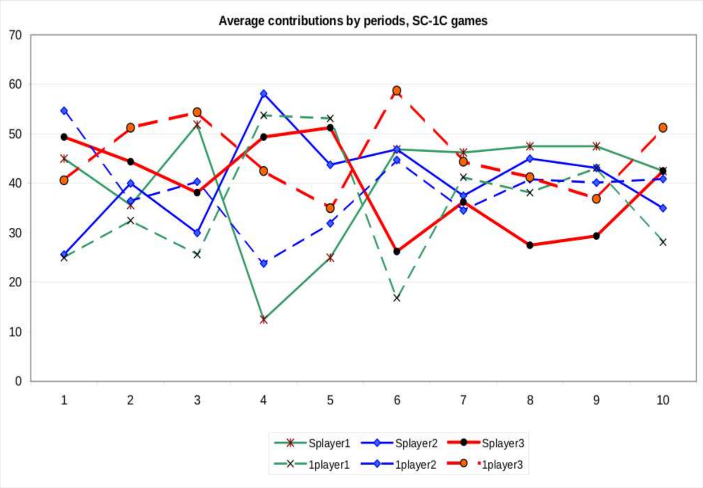
player#	1	2	3
votes	3	2	2
proposed shares	x	y	z

  - Effects in an **implicit modifier** to player 2's payoff.

Way out: symmetric positioning

- In SC-1C games, each player is shown in the middle of the table in a systematic (clockwise) rotation.
- The difference between player 2 and the others in S-games is mitigated to (40.50 vs. 39.40 or 40.06), and becomes insignificant
  - We conjecture that the effect of *implicit modifier* is to completely disappear in a fully symmetric treatment.
- *Explicit modifiers'* effect persists for player 3, although to a somewhat smaller extent and over the last rounds.
- Average number of offers in games S (1) — 2.13 (resp., 2.42).
- Average time per round in games S (1) — 30 (resp., 37) seconds.

Results: SC–games



Summary of the S-1 games

All ($N = 320$)	mean	s.d.	min	max
player 1	35.36	29.04	0	80
player 2	44.53	24.42	0	100
player 3	40.1	27.56	0	111
Game S				
player 1	37.40	29.44	0	80
player 2	46.25	23.89	0	100
player 3	**36.34**	28.05	0	110
Game 1				
player 1	33.32	28.57	0	80
player 2	42.81	24.91	0	99
player 3	**43.85**	26.62	0	111

- No significant difference in payoffs for players 1 and 2.
- Significant difference for player 3 at 1-2% confidence level.
- Centered treatment suppresses implicit modifiers.

Games V–2 (Veto)

Game V: quota is 5 votes

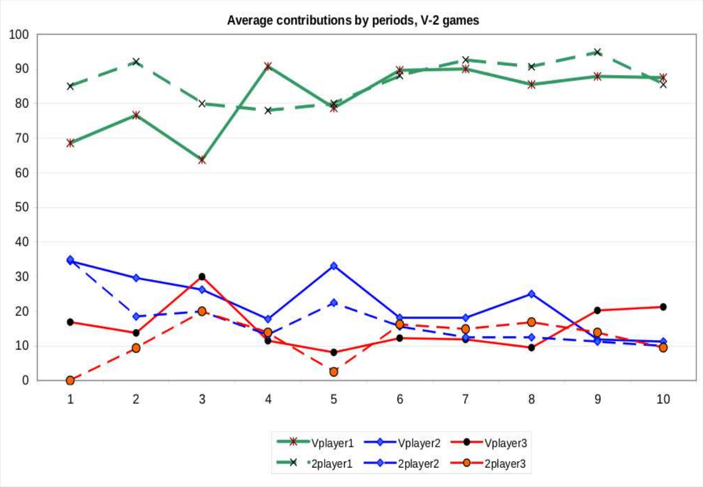
player#	1	2	3
votes	3	2	2

Winning coalitions $W = \{\{1,2\},\{1,3\},\{1,2,3\}\}$.
Banzhaf: $\beta_1 = 3/5, \beta_2 = \beta_3 = 1/5$, shares [72, 24, 24].
Shapley-Shubik: $\sigma_1 = 2/3, \sigma_2 = \sigma_3 = 1/6$, shares [80, 20, 20].
Game 2 uses the following *explicit modifiers*:

	1	2	3
1	-	1	1
2	0.99	-	1
3	0.99	1	-

α indices based on the $f^{\times}$ intensity function:
$\alpha_1 = 0.6016, \alpha_2 = \alpha_3 = 0.1992$

Results:V–2 games



Summary of the V-2 games

All ($N = 160$)	mean	s.d.	min	max
player 1	84.29	24.99	0	120
player 2	19.76	20.97	0	70
player 3	13.56	18.42	0	60
Game V				
player 1	81.90	24.76	0	119
player 2	22.56	23.40	0	70
player 3	15.53	19.99	0	60
Game 2				
player 1	86.68	25.14	0	120
player 2	16.96	17.94	0	60
player 3	11.60	16.59	0	60

Results: V–2 games

- Player 1 (the veto player) gets even more than the Banzhaf index predicts.
- No significant difference across treatments.
- Effects of greater negative modifiers might be larger.
- Average number of offers in games V (2) — 5.94 (resp., 5.63).
- Average decision time in games V (2) — 147 (resp., 141) seconds. Timing of decisions requires further attention.

Maya's question: does smuggling and waiting in the V-game reflect the fact that there are three players rather than 2 as in the ultimatum game, or is it about changing notion of generosity?

Games E–3 (Enlarged)

Game E: Again, **5** votes are required to reach an agreement

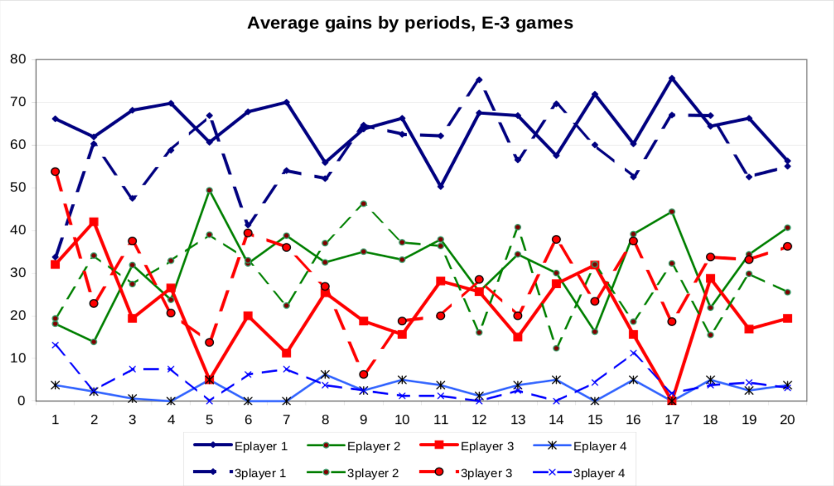
player#	1	2	3	4
votes	3	2	2	1

Winning coalitions $W = \{\{1,2\},\{1,3\},\{1,2,3\},\{1,2,4\},$ $\{1,3,4\},\{2,3,4\},\{1,2,3,4\}\}$. Banzhaf (and Shapley-Shubik): $\beta_1 = 5/12, \beta_2 = \beta_3 = 3/12, \beta_4 = 1/12$, shares [50, 30, 30, 10].
Game 3 employs the following modifiers:

	1	2	3	4
1	-	1	1	1
2	0.99	-	1	1
3	1	1	-	1
4	1	1	1	-

α indices based on the $f^{\times}$ intensity function:
$\alpha_1 = 0.5005, \alpha_2 = 0.1992, \alpha_3 = 0.2002, \alpha_4 = 0.1001$

The E–3 games



Average gains by periods, E-3 games
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Eplayer 1
Eplayer 2
Eplayer 3
Eplayer 4
3player 1
3player 2
3player 3
3player 4

Summary of the E-3 games

All ($N = 160$)	mean	s.d.	min	max
player 1	61.15	25.76	0	100
player 2	30.63	23.82	0	70
player 3	24.73	24.54	0	70
player 4	3.49	9.00	0	70
Game E				
player 1	**64.34**	22.36	0	95
player 2	31.65	23.17	0	70
player 3	**21.23**	23.72	0	70
player 4	2.76	7.40	0	40
Game 3				
player 1	**57.95**	28.47	0	100
player 2	29.59	24.48	0	70
player 3	**28.23**	24.90	0	65
player 4	4.21	10.33	0	70

Summary of the E-3 games

- In **E-game** player 1 gets systematically more than the Banzhaf index prediction at the expense of player 4, while gains of players 2 and 3 are in line with the index, and are *greater than in the V–2 treatment*.

Summary of the E-3 games

- In **E-game** player 1 gets systematically more than the Banzhaf index prediction at the expense of player 4, while gains of players 2 and 3 are in line with the index, and are *greater than in the V–2 treatment*.
- Player 3 gains a statistically significant increase in the average payoff.
  - Thus, a small negative modifier towards player 1 indirectly benefits player 3, (gain per session increases by 25%).

Summary of the E-3 games

- In **E-game** player 1 gets systematically more than the Banzhaf index prediction at the expense of player 4, while gains of players 2 and 3 are in line with the index, and are *greater than in the V–2 treatment*.
- Player 3 gains a statistically significant increase in the average payoff.
    - Thus, a small negative modifier towards player 1 indirectly benefits player 3, (gain per session increases by 25%).
- Frequency of coalitions $\{2, 3, 4\}$ is $\times 2$ higher in the 3–game than in the E–game.
    - Means that players 2, realizing they do not like player 1, tend to switch to a larger coalition, even though it is clearly more difficult and may involve lowering one's share of the pie (has to be divided among 3 players instead of 2).

Games F–4

Game F: 6 votes required to reach an agreement

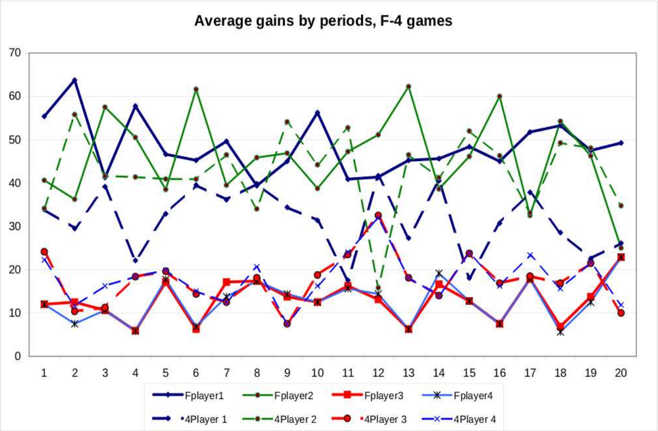
player#	1	2	3	4
votes	3	3	2	2

Winning coalitions
$W = \{\{1,2\},\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\},\{1,2,3,4\}\}$.
Banzhaf index is $\beta_1 = \beta_2 = 1/3, \beta_3 = \beta_4 = 1/6$, 1 and 2 get 40, 3 and 4 get 20 each.
Game 4 employs the following modifiers:

	1	2	3	4
1	-	0.8	1	1.01
2	0.8	-	1	1.1
3	1	1	-	1
4	1	1	1	-

α indices based on $f^{\times}$ intensity function:
$\alpha_1 = 0.3107, \alpha_2 = 0.3583, \alpha_3 = \alpha_4 = 0.2002$.

The F–4 games


Average gains by periods, F-4 games
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Fplayer1
Fplayer2
Fplayer3
Fplayer4
4Player 1
4Player 2
4Player 3
4Player 4

Summary of the F–4 games

All ($N = 160$)	mean	s.d.	min	max
player 1	39.95	10.97	17.55	63.75
player 2	44.32	9.68	15.81	62.25
player 3	15.24	5.75	5.88	32.50
player 4	15.35	5.87	5.63	31.88
Game F				
player 1	**48.43**	6.34	39.38	63.75
player 2	45.97	10.02	25.00	62.25
player 3	**12.95**	4.68	5.88	22.88
player 4	**12.66**	4.95	5.63	22.88
Game 4				
player 1	**31.48**	7.46	17.55	41.66
player 2	42.67	9.30	15.81	55.80
player 3	**17.53**	5.90	7.50	32.50
player 4	**18.04**	5.58	7.50	31.88

Results: F-4 games

- 'Large' negative modifier of player 2 for player 1 significantly lowers her earnings (48.43 vs 31.48).
- On the contrast, player 2's payoff does not change much.

Results: F-4 games

- 'Large' negative modifier of player 2 for player 1 significantly lowers her earnings (48.43 vs 31.48).
- On the contrast, player 2's payoff does not change much.
- Complex interaction of modifiers: high 'dislike' modifiers of 0.8 tend to hurt player 1 more than player 2 because player 2 more strongly prefers larger coalitions.
- We also investigated another explanation – that the *psychological features* of the subjects' characters essentially influence their behaviour.

Coalitional outcomes across treatments

coalitions ╲ games	S–1 games			V–2 games	
S-1 coalitions	S	1	V-2 coalitions	V	2
1&2	54	33	**1&2**	41	40
1&3	29	33	**2&3**	27	26
2&3	56	59	1&2&3	12	10
1&2&3	21	35	1 alone	0	1
other	0	0	none	0	3
total	160	160	total	80	80

coalitions ╲ games	E–3 games			F–4 games	
E-3 coalitions	E	3	F-4 coalitions	F	4
1&2	73	74	**1&2**	82	64
1&3	57	51	**1&3&4**	38	31
2&3&4	1 3	26	**2&3&4**	33	56
1&2&3	5	1	1&2&3	1	1
1&2&4	1	3	1&2&4	1	0
1&3&4	1	1	1&3	0	1
1&2&3&4	9	3	1&4	0	0
none	1	0	1&2&3&4	4	6
total	160	160	total	160	160

Offers by roles, S-1 games

player	observations	mean	std.dev
all coalitions			
1	1155	35.46	28.86
2	1155	43.02	24.97
3	1155	41.51	27.12
{1,2} coalitions			
1	196	61.74	7.48
2	196	58.25	7.48
3	196	0	0
{1,3} coalitions			
1	166	59.88	10.47
2	166	0	0
3	166	60.12	10.47
{2,3} coalitions			
1	274	0	0
2	274	59.26	4.51
3	274	60.73	4.51

Offers by roles, F-4 games

player	observations	mean	std.dev
all coalitions			
1	1280	42.59	27.66
2	1280	46.01	25.82
3	1280	15.49	15.00
4	1280	15.53	15.25
{1,2} coalitions			
1	580	60.26	3.72
2	580	59.75	3.72
3	580	0	0
4	580	0	0
{1,3,4} coalitions			
1	276	60.65	10.30
2	276	0	0
3	276	29.83	5.17
4	276	29.52	5.41
{2,3,4} coalitions			
1	356	0	0
2	356	61.73	8.08
3	356	29.01	4.42
4	356	29.26	4.61

Composition of the winning coalition explains over 90% of shares' variations!

Interpretation

- An overwhelming majority of the outcomes result in *minimal* winning coalitions.
- Explicit modifiers are of secondary importance; by contrast, people use simple heuristic strategies that are not captured by either classical or generalized power indices (in their present formulation).
- The best predictors for the model are players' roles and the composition of the winning coalition.
  - S-1 60-60 for all three winning coalitions
  - V-2 85-25 for the coalitions $\{1, 2\}$ and $\{1, 3\}$, the rest being 'noise'
  - E-3 70-50 for the $\{1, 2\}$ and $\{1, 3\}$ coalitions, and 50-50-20 for the $\{2, 3, 4\}$ coalition.
  - F-4 60–60 for the $\{1, 2\}$ coalition, and the 60–30–30 for the $\{1, 3, 4\}$ and $\{2, 3, 4\}$ coalitions.
- How can we describe this evidence theoretically?

Punishment in public goods game (PG VCM)

- Factors of cooperative behaviour are of interest, especially when this behaviour is disequilibrium (e.g. investment game, trust game, ultimatum game, public goods game)
- Recent behavioural explanations (e.g. McKelvey and Palfrey, 1998; Fehr and Schmidt, 1999; Falk and Fischbacher, 2003) are important, but sometimes lack empirical background
- Empirical attempts (e.g. Camerer e.a., 2003; Stahl, 2008) are useful, albeit restrictive.
- One more of these: estimation of factors of punishment in public goods games using experimental technique and structural model.

This paper

Claims that conventional attribution of punishment to ‘dissatisfaction with low contribution’ (and by the same token, to disapproval of antisocial behaviour) is too quick/impudent: In the PG game context, people may punish each other for different (strategic) reasons driven by the experimental institution.

In particular, this may explain the divergence between the fractions of spiteful behaviour (punishing those who contributed *more* than you did) obsereved in some (developing) countries to a much more substantial extent that in other (developed).

Contributions:

- New experimental design (insurance against punishment)
- Behavioural model of strategic incentives for punishment
- Empirical estimates of latent classes of motives in a convenience sample of Russian subjects.

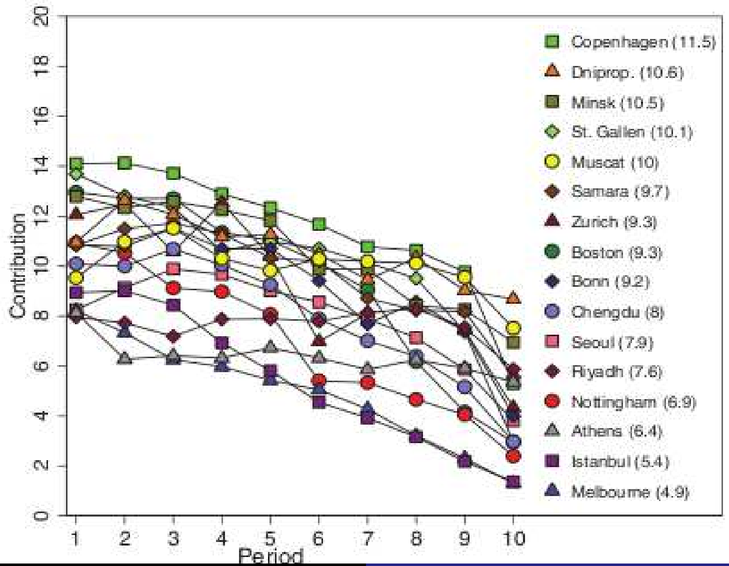
Public goods (PG) game with voluntary contribution mechanism (VCM)

- $n \geq 2$ players endowed with w units per period each (normalized to 1)
- Each player i independently decides what fraction $c_i, 0 \geq c_i \geq 1$ she will contribute to the public good, retaining $1 - c_i$.
- Return from public good is $k \cdot \sum_i c_i = \alpha \bar{c}$, where $\bar{c} = \frac{\sum_i c_i}{n}$ and $\alpha = kn$, $k < 1 < kn$ is efficiency factor.

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

The only Nash equilibrium is zero contribution, while Pareto-optimal is 100% contribution

PG with VCM: typical results (Herrmann, Gächter, Thoni, 2009)



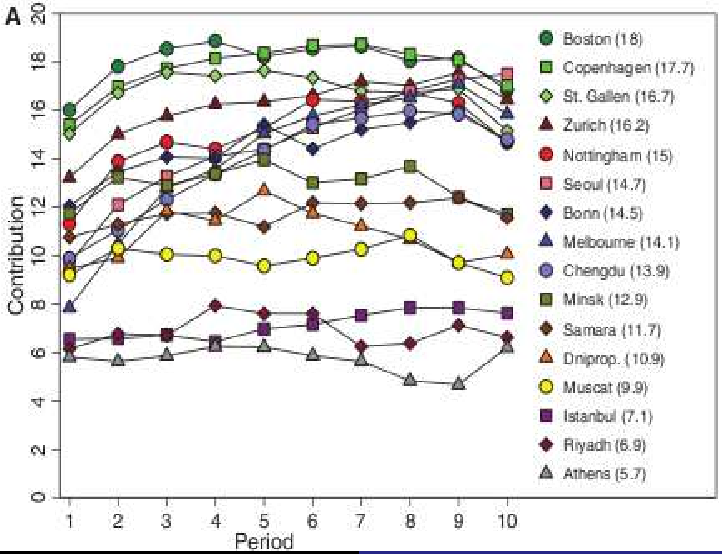
Public goods game with VCM and punishment

After the contribution stage, all players are informed about individual contributions, and can punish each other player j (not herself!) by p_{ij} units at a cost sp_{ij} units to themselves, where $s < 1$. Total payoff to player i is then

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

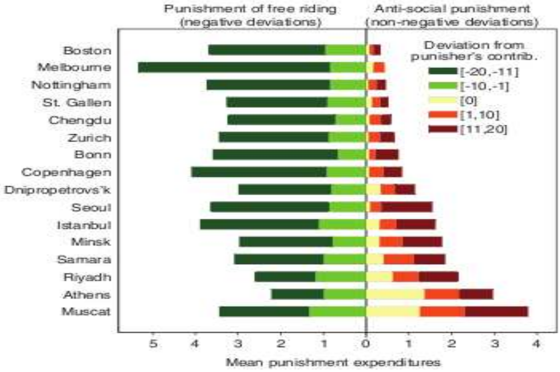
Punishments are known to increase the degree of cooperativeness, especially in with time and in partner treatments.
Mechanism: punishment (threaten, expression of disapproval) of those who free-ride boosts up cooperativeness.

PG with VCM: typical results (Herrmann, Gächter, Thoni, 2009)



Spiteful (antisocial) punishment (Herrmann, Gächter, Thoni, 2009)

Sometimes players punish not only those who contributed less, (free-riders — *prosocial* punishment), but also those who conributed more than they did (*spiteful*, or antisocial punishment)



Middle East, Russia and Eastern Europe are world leaders in spite

Spiteful (antisocial) punishment

...or are they?

- What are the origins for spiteful punishment?

Spiteful (antisocial) punishment

...or are they?

- What are the origins for spiteful punishment?
- More generally: Is punishment necessarily an expression of ethical disapproval (retaliation for low contributions?)

Spiteful (antisocial) punishment

...or are they?

- What are the origins for spiteful punishment?
- More generally: Is punishment necessarily an expression of ethical disapproval (retaliation for low contributions?)
- Yet more generally: what are the motives for punishment behaviour?

Classification of possible motives for punishment

Availability — presense of punishment option is suggestive in itself — the Chekhov motive.
'If in the first scene of the play, there is a gun on the wall, by the third scene it mush shut'

Classification of possible motives for punishment

Availability — presense of punishment option is suggestive in itself — the Chekhov motive.
'If in the first scene of the play, there is a gun on the wall, by the third scene it mush shut'

Tolerance — culturally-defined punishment is something 'customary' and 'acceptable' — the Tjutchev motive.
'The entire Russian history before Peter the Great is an entire commemoration cervice, and after Peter the Great — an entire criminal case'

Classification of possible motives for punishment

Availability — presense of punishment option is suggestive in itself — the Chekhov motive.
'If in the first scene of the play, there is a gun on the wall, by the third scene it mush shut'

Tolerance — culturally-defined punishment is something 'customary' and 'acceptable' — the Tjutchev motive.
'The entire Russian history before Peter the Great is an entire commemoration cervice, and after Peter the Great — an entire criminal case'

Competitiveness — punishment as an efficient way to improve own relative standing in the group — the Dostoyevsky motive.
'Am I a trembling biest, or I daresay?'

Classification of possible motives for punishment

Availability — presense of punishment option is suggestive in itself — the Chekhov motive.
'If in the first scene of the play, there is a gun on the wall, by the third scene it mush shut'

Tolerance — culturally-defined punishment is something 'customary' and 'acceptable' — the Tjutchev motive.
'The entire Russian history before Peter the Great is an entire commemoration cervice, and after Peter the Great — an entire criminal case'

Competitiveness — punishment as an efficient way to improve own relative standing in the group — the Dostoyevsky motive.
'Am I a trembling biest, or I daresay?'

Preemption — penalizing because one expects penalties from the others — the Brodsky motive
'A man is more frightening than its skeleton'.

Classification of possible motives for punishment

Availability — presense of punishment option is suggestive in itself — the Chekhov motive.
'If in the first scene of the play, there is a gun on the wall, by the third scene it mush shut'

Tolerance — culturally-defined punishment is something 'customary' and 'acceptable' — the Tjutchev motive.
'The entire Russian history before Peter the Great is an entire commemoration cervice, and after Peter the Great — an entire criminal case'

Competitiveness — punishment as an efficient way to improve own relative standing in the group — the Dostoyevsky motive.
'Am I a trembling biest, or I daresay?'

Preemption — penalizing because one expects penalties from the others — the Brodsky motive
'A man is more frightening than its skeleton'.

Retaliation — negative feeling at what the others have contributed, leading to the desire for retaliation.

$c_i - c_j$, difference between contributions.
$\hat{c}_i - c_j$, difference between believed norm and factual contribution.
$\bar{c} - c_j$, , difference between group norm (mean) and factual contribution.

Classification of possible motives for punishment

Availability — presense of punishment option is suggestive in itself — the Chekhov motive.
'If in the first scene of the play, there is a gun on the wall, by the third scene it mush shut'

Tolerance — culturally-defined punishment is something 'customary' and 'acceptable' — the Tjutchev motive.
'The entire Russian history before Peter the Great is an entire commemoration cervice, and after Peter the Great — an entire criminal case'

Competitiveness — punishment as an efficient way to improve own relative standing in the group — the Dostoyevsky motive.
'Am I a trembling biest, or I daresay?'

Preemption — penalizing because one expects penalties from the others — the Brodsky motive
'A man is more frightening than its skeleton'.

Retaliation — negative feeling at what the others have contributed, leading to the desire for retaliation.

$c_i - c_j$, difference between contributions.

$\hat{c}_i - c_j$, difference between believed norm and factual contribution.

$\bar{c} - c_j$, , difference between group norm (mean) and factual contribution.

Spite per se — genuine disapproval of those who behave pro-socially.

Design: baseline after Gächter and Herrmann (2008)

- 2 single-shot games: VCM without punishment, followed by VCM with punishment (2 games altogether).
- Groups of $n = 4$ players, endowment 20, efficiency factor $k = 1.6$ ($\alpha = 0.4$) for all subjects.
- After each contributions stage, participants observe contributions and payoffs of all groupmates.
- Cost of punishment from 0 to 10 either low (0.1) or high (0.5).
- Preceding instructions with worked examples and exercises to check understanding.
- Ex ante intentions questionnaire other than oneself and the punished one, in proportion to their contributions.
- Post-punishment treatments introduced at the end.

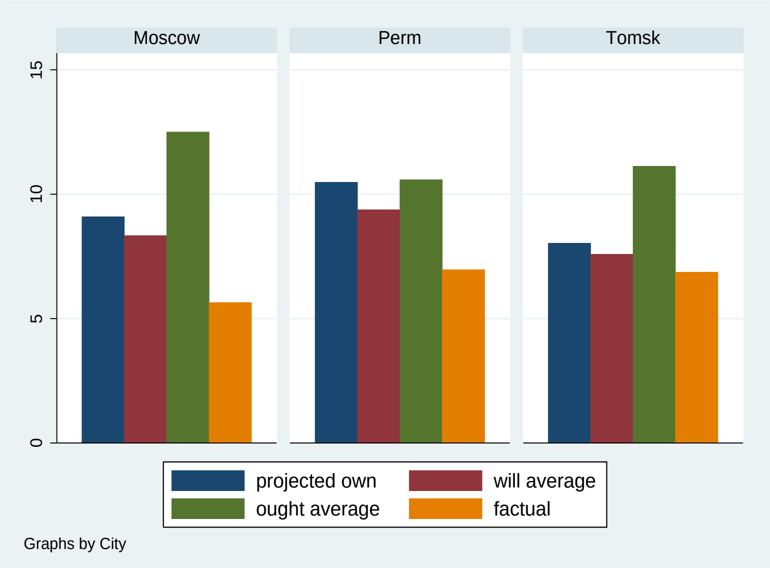
Participants: 300 full-time and part-time students from Moscow (128), Perm (76) and Tomsk (96). Gender composition — 50/50, average payoff — 208 RuR.

Experiment on the map

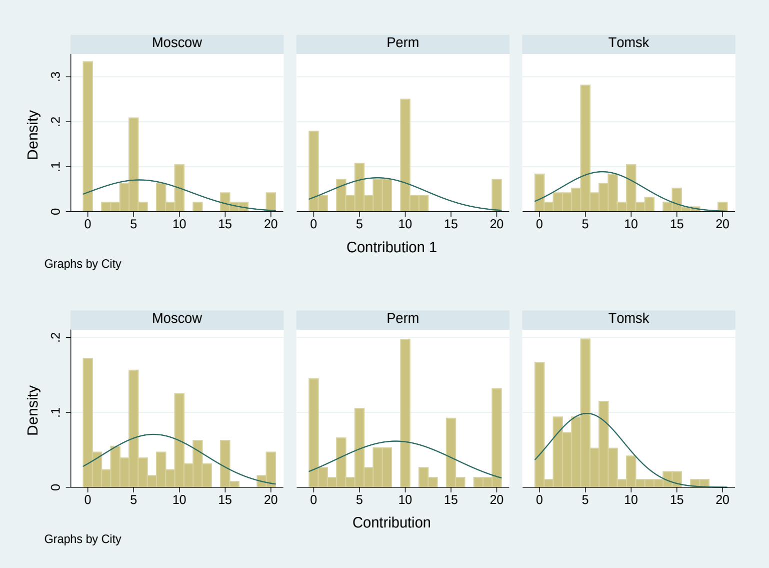
Design: additions

- Intentions questionnaire asks for *planned* own contirubtions, the *due* average and *expected* average contributions in their group, and desired contribution level if the group average turns out to take discrete values of 0, 3, 6, 10, 14 and 17 units, evaluated by strategy method.
- In a separate screen with *yes-no* button shown after the contributions stage, the subject has to choose 'yes' iff (s)he wants to assign deduction points to at least one of his or her group fellows (test for availability).
- After punishment stage, subjects in the low cost of punishment sessions could purchase *insurance against punishment* of up to 10 units from each individual player in her group, at a cost of 0.1 if redistributed from punishment, and 0.2 per unit of insurance.

Contributions



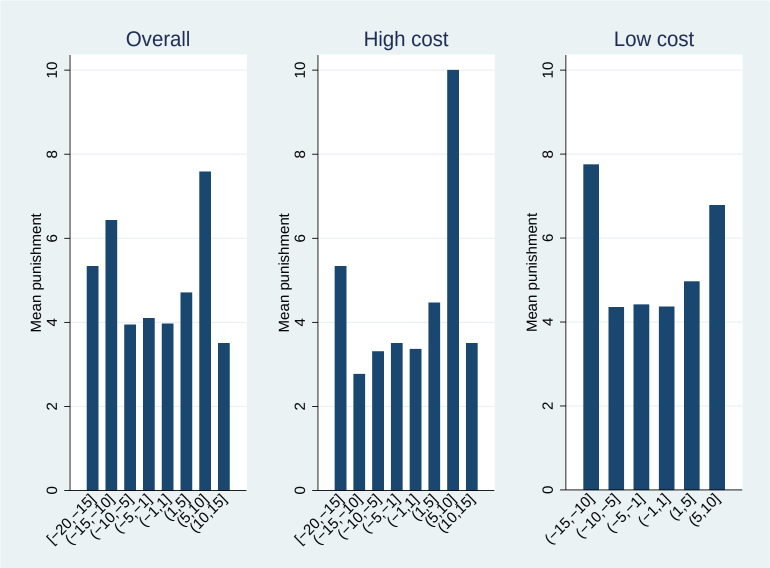
Contributions: first (upper) and second (lower) stage



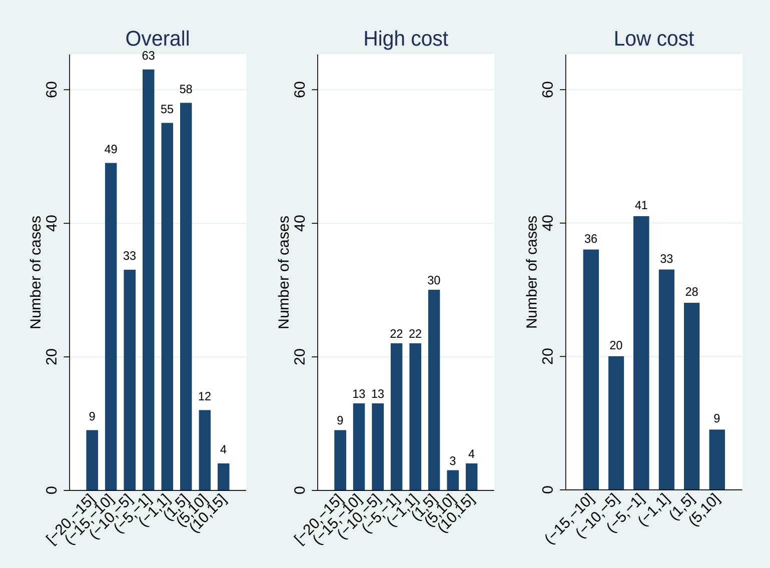
Contributions

- Contributions in line with the previous experience.
- Factual own contributions always lower than projected and (especially) normative.
- Expected undercontribution.
- In one-round span, disciplining role of punishment is limited at best.
- Second-stage contributions are stable across cities at low (median 5) and high (median 9) costs.

Mean punishments by treatments



Number of punishments by treatments



Punishments: statistics

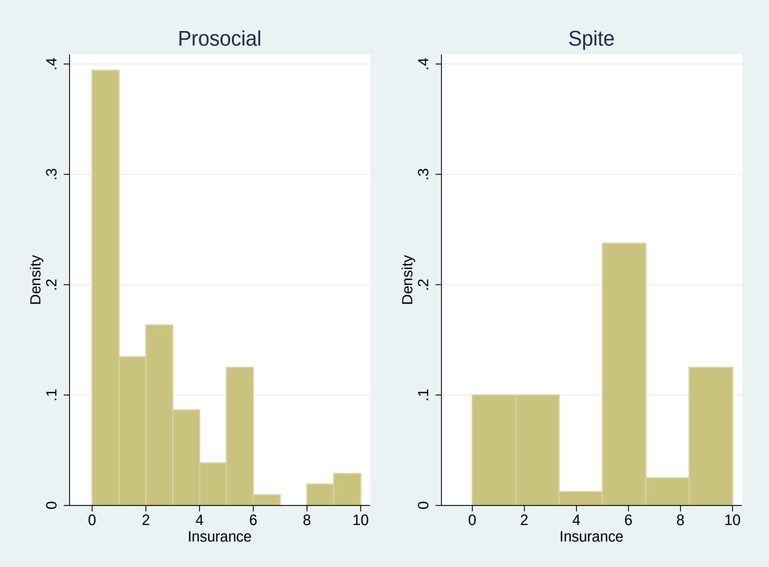
statistics	*contrib*	*wish*	*qpun*	*pun*
Overall (N=295)				
mean	6.99	.55	.95	4.75
median	5	1	0	4
sd	5.60	.49	1.14	3.23
Low cost = 0.1 (N=143)				
mean	5.09	.59	1.16	5.33
median	5	1	1	5
sd	4.17	.49	1.17	3.34
High cost = 0.5 (N=152)				
mean	8.78	.51	.76	3.93
median	9	1	0	3
sd	6.16	.50	1.07	2.89

Number of punishments by treatments

stat's	*c1*	*c2*	*mpun*	*c1*	*c2*	*mpun*	*c1*	*c2*	*mpun*
Spiteful punishment									
	1 punish't (N=6)			2 punish'ts (N=10)			3 punish'ts (N=39)		
mean	3.16	6	2.5	1.4	8	4.7	2.69	8.82	6.82
median	2	4	2.5	1	7.5	2.5	2	8	10
sd	2.31	4.42	.04	1.57	4.08	3.71	2.61	4.95	3.83
Prosocial punishment									
	1 punish't (N=49)			2 punish'ts (N=46)			3 punish'ts (N=60)		
mean	7.89	2.48	4.26	9.86	2.97	4.19	10.8	4.08	4.76
median	8	0	4	9	3	3	10.5	5	4
sd	4.36	3.24	2.75	5.41	2.87	2.99	4.20	3.56	3.24

c1 — contribution of the punisher, c2 — contribution of the punished,
mpun — mean punishment size per person punished

Insurance decisions



Motives, % in ex-post questionnaire

Reasons		
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

Note: — less than 2%

Preliminary conclusions

confirmed: Mean frequency and size of spiteful punishments are compatible with those of the previous experiments

Preliminary conclusions

confirmed: Mean frequency and size of spiteful punishments are compatible with those of the previous experiments

confirmed: Mean punishment size decreases with cost, and is on average the same for prosocial and spiteful punishments (similar rationality)

Preliminary conclusions

confirmed: Mean frequency and size of spiteful punishments are compatible with those of the previous experiments

confirmed: Mean punishment size decreases with cost, and is on average the same for prosocial and spiteful punishments (similar rationality)

new! Spite increases in low-cost conditions

Preliminary conclusions

confirmed: Mean frequency and size of spiteful punishments are compatible with those of the previous experiments

confirmed: Mean punishment size decreases with cost, and is on average the same for prosocial and spiteful punishments (similar rationality)

new! Spite increases in low-cost conditions

new! Spiteful punishments are more serial and larger on average than prosocial punishments

Preliminary conclusions

- confirmed: Mean frequency and size of spiteful punishments are compatible with those of the previous experiments
- confirmed: Mean punishment size decreases with cost, and is on average the same for prosocial and spiteful punishments (similar rationality)
- new! Spite increases in low-cost conditions
- new! Spiteful punishments are more serial and larger on average than prosocial punishments
- new! Spiteful punishers insure significantly more often and use more extra money than prosocial punishers

Preliminary conclusions

- confirmed: Mean frequency and size of spiteful punishments are compatible with those of the previous experiments
- confirmed: Mean punishment size decreases with cost, and is on average the same for prosocial and spiteful punishments (similar rationality)
- new! Spite increases in low-cost conditions
- new! Spiteful punishments are more serial and larger on average than prosocial punishments
- new! Spiteful punishers insure significantly more often and use more extra money than prosocial punishers
- new! In the ex post questionnaire, over 3/4 of spiteful punishers report desire to increase their relative standing as the main motive for punishment

Punishments factors: Tobit model estimates

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

Cluster-robust standard errors in parentheses. *** –1%, ** –5%, * – 10% sign.level

contr – c_j, contribution of punisher, difcontr – $c_i - c_j$, relcontr – $c_i - Ec_j$, homxavg – $Ec_i - E\bar{c}_j$, cost – cost treatment dummy

Punishment factors revisited

- *Availability* appears to be immaterial: average willingness to punish insignificantly smaller than elsewhere.
- *Tolerance* is immaterial: 51% of prosocial and 75% of spiteful punishers have relocated their funds from punishment to insurance.
- Prosocial punishments driven by **retaliation/upset**: differences in contributions are the major explanatory factor.
- Spiteful punishments driven **competition**: willingness to beat the others prevails.
- Separate factor of **preemption** may apply to both.

How can we disentangle competitive/retaliation and preemption motives for prosocial and spiteful punishments, respectively?

Behavioural model of punishment motives

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

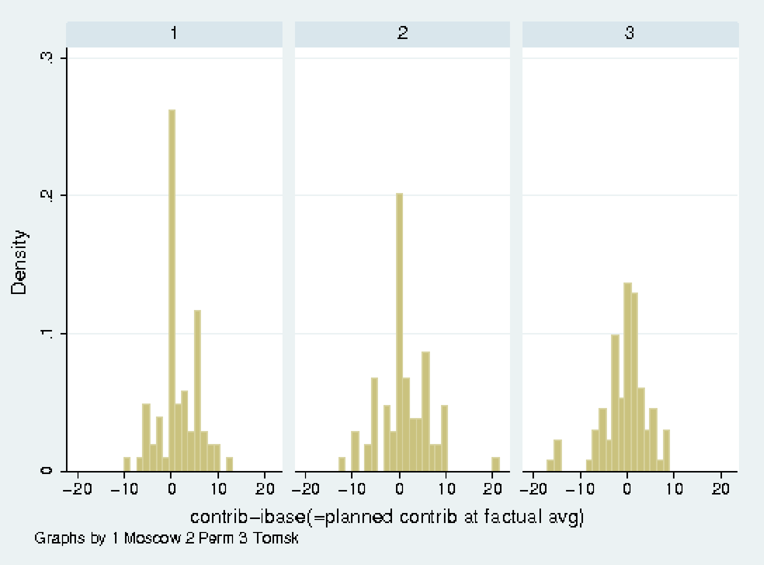
- V_i —material payoff,
- φ — dissatisfaction function of player i at player j,
- $\mathrm{E}p_{ji}$ — expectation of player i of punishment from player j,
- π — cost of punishment,
- η_{1i} and η_{2i} — individual-specific weights to retaliation and preemption for expected punishment (η's are zero in case of no punishment)

Maximizing (3) wrt punishment p_{ij} and rearranging,

$$p^*_{ij} = \eta_{1i} \frac{\sum_k \gamma_k \varphi_{kij}}{p_{ij}\pi} + \eta_{2i} \frac{n-1}{\pi} \tag{4}$$

wherein linear weights η attached to normal densities of the latent factors are estimable using GLLAMM

Factual vs strategic form planned contributions



Model estimates

For prosocial punishment:

$$pun = \alpha + \eta_1\phi(prcontr + pcontr) + \eta_2\phi(pcons) + \varepsilon \quad (5)$$

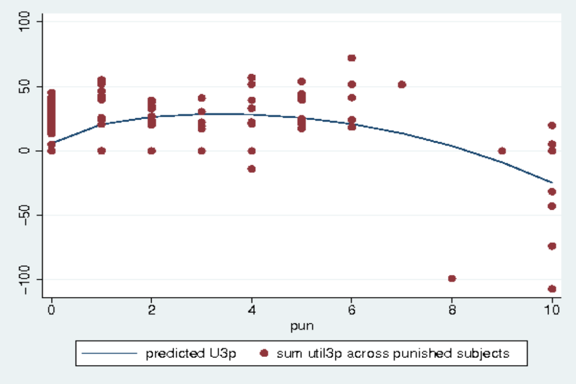
Weights are $\eta_1^p = 0.207, \eta_2^p = 0.793$, implying larger weight on preemption

For spiteful punishment:

$$pun = \alpha + \eta_1\phi(pcondev) + \eta_2\phi(pcons) + \varepsilon \quad (6)$$

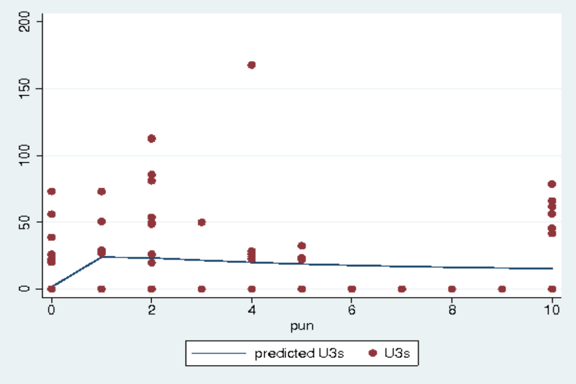
Weights $\eta_1^p = 0.826, \eta_2^p = 0.176$, imply larger weight on retaliation

Estimated utility for prosocial punishers



Inverse U-shape of utility vs. punishment size: at lower levels, larger punishments correspond to low utility of the punisher as they reflect their unhappiness with the social behaviour.

Estimated utility for spiteful punishers



U-shape graph with high dispersion at low punishment levels and large utility for those with extreme punishments.

Cluster statistics

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

Classification: the four punishment caterogires

Fair prosocial (12%) Punishments motivated by low contributions of the punished relative to the group standard (retaliation). Believe they are on their right, punish by a lot (mean 9.78), and almost do not insure (mean 1.28).

Philistine prosocial (60%) Fairness motivated, but afraid of expression for fear of preemption and/or cost. Punishment is low (3.51), insurance yet lower (2.5)

Jealous (spite per se) (20%) Afraid of being exploited by the society, try to decrease payoffs of more successful players, but not at own cost. Both punishments (2.66) and insurance (2.5) are low.

Aggressive spite (8%) Motivated by competitiveness, but also very afraid of preemption: use maximal punishments (10 in 100% cases) and insurance (7.38%).

Conclusions and extensions

- Punishment in PG context at least, should not always be interpreted as a revelation of dissatisfaction with contributions of the other players: there is a variety of competing explanations.
- These results suggest a multiplicity of principles on which 'punishment' behaviour may rest. In Russia, these were quite heterogeneous, while in Western Europe, for instance, 'spiteful' punishments are minor. However, if we exclude strategic punishments from apparently spiteful ones in Russia, its 'spitefulness' would shrink/become non significant.
- Decomposition of punishment motives may be interesting and important for the diagnosis of the state of the respective societies.

Thank you!
PS: Full version of the paper available at http://epee.hse.ru

Why police?

- Important institution with high public prominence.
- Police reform is badly needed and becomes part of the agenda with the new Law of Police.
- Thanks to the project of the Center for the studes of institions and development of the HSE, we are involved in policy analysis and can perform the first in Russia (if not in the world) laboratory experiment with real police officers.
- Long-run objective — field experiment with new systems of police assessment.

Background I

- Russian police currently employs 1,280,000 people — twice as much as in the US, per more than two times smaller population.
- Russian police registers about 3 mln crimes per year (US police — about 12 mln).
- Crime rates per 100,000 inhabitants (2006)

Crime	Russia	USA
Murder	19.2	5.7
Rape	6.2	92.4
Robbery	41,9	149.4

- At least in part, this discrepancy is due to misreported statistics: Russian police officers (POs) are being assesses by the system of internal indicators collected by the police itself (палочная система).

Background II

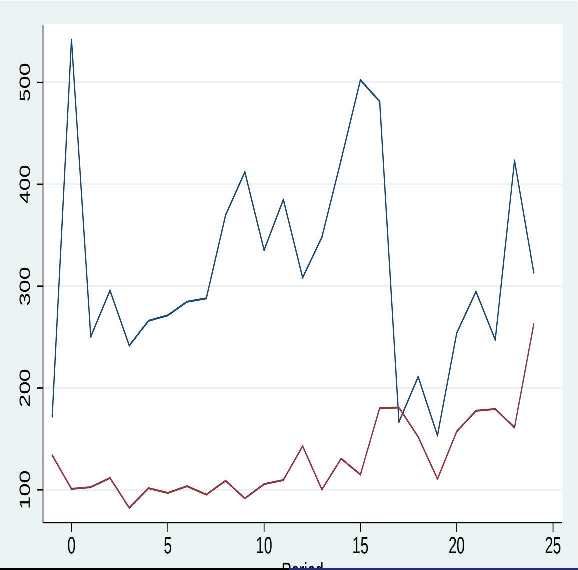
- Underpayment: as of 2005, average baseline salary of freshman conscript — 7,000 RuR (about 250 US$), captain with 10-year tenure was 12,000 RuR per month (400 US$), major of criminal police with 10-year tenure —20,000 RuR (700 US$). Real wage can be larger due to benefits, but legally not by much. Social security system is poor: pension for an injured officer would be 1,500 RuR.
- Police officers are overloaded with paperworks: a beat cop (участковый милиционер) has about 20 reports to fill continuously.
- Information is very poor: MVD is one of the most closed structures of the Russian government, and people's view about what is police are most often very naive.
- Quality of personnel is dubious, and is being reported to decrease over time.
- Mistrust: Over 50% of population believes POs are corrupt, and about 60% of the Russians do not trust the police (worst of all institutions, along with customs officers), trusting share — about 25%.

In total, the system of incentives among the police is such that officers have little incentives to serve the citizens, even if they want and can.

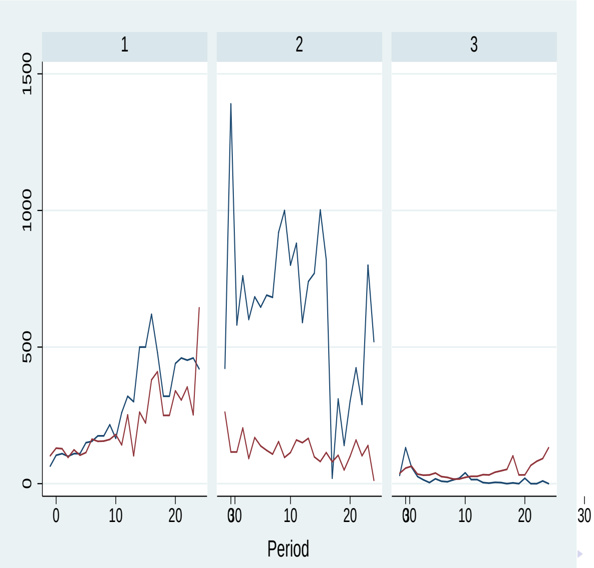
Experiment on factors of corruption

- Game of three stage, in groups of 5 participants.
- **Stage 1 — 8 rounds**: Police officers receive *basic income* 100 and can take *additional income* in any amount, but if the sum of group income exceeds 5,000 per period, everybody receives 0.
- External monitoring of additional income: if caught by taking additional income, this income is confiscated and the guilty player is fined by 50.
- Probability of monitoring: 0.8, but if the group is able to raise 500 or more as an *insurance fund* by means of independent contributions, and lose what is contributed, then the probability decreases to 0.1.
- **Stage 2 — 8 rounds**: Insurance fund may or may not work (new boss appointed), which is not known to the players. If it does not, probability of monitoring is 0.8 and contributions are returned. In fact, the fund never works from now on.
- **Stage 3 — 8 rounds**: Same as stage 2, but basic income increases to 300, fine in case of discovery increases to 300, and if caught, basic income is 100.

Overall contribution and bribery



Contribution and bribery by groups



Conclusions and extensions

- Tacit coordination works in both threshold public good games
- End-of-punishment effect: when corruption is being fought, people start to take more bribes.
- (Unique to police) role assignment effect: players stick to group norms to a much higher extent than ordinary players (students of HSE).

For cooperation, support and feedback, big thanks are due to

- Fuad Aleskerov
- Elena Tougareva
- Kirill Pogorelskiy
- Alexander Poddiakov
- Leon Kosals
- Benedikt Herrmann
- Daniel Houser
- Diana Kolesnikova
- Alexey Charkov
- Irina Karpenko
- Ivan Filatov
- Yulia Safarbakova
- Ekaterina Shilkina
- Stas Moskovtsev
- ...

For further details, please visit http://epee.hse.ru

Thank you!