Commitment and Extortion

Paul Harrenstein, Felix Brandt, Felix Fischer

Computer Science Department,
Ludwig-Maximilians-Universität, München
{harrenst,brandtf,fischerf}@tcs.ifi.lmu.de

PREFERENCE AGGREGATION IN MULTIAGENT SYSTEMS

1. Commitments, Promises and Threats

- Game Theory: efficient application of actual force *vs.* exploitation of potential force.
- Promises and threats are examples of the latter.
- Commitment is a limitation of one's freedom of action to improve one's strategic position. One can even do so by committing to a strongly dominated action. (Schelling 1960)

$$\begin{bmatrix} (1,3) & (3,2) \\ (0,0) & (2,1) \end{bmatrix} \qquad \begin{bmatrix} (2,2) & (0,0) \\ (1,3) & (3,1) \end{bmatrix}$$

- For a commitment to be effective it should be made credible. The credibility of a commitment may depend on its binding power.
- Illustration: Stanley Kubrick's *Dr Strangelove*. A rogue American general has started a nuclear attack on the USSR. The Soviets, however, have built a *Doomsday Machine*, a device that will destroy all human and animal life on Earth if the USSR is attacked. The following discussion between the US President and the Soviet ambassador ensues:
  - President Merkin Muffley: *"I am afraid I do not understand something, Alexi. Is the premier threatening to explode this device, if our planes carry out their attack?"*
  - Ambassador Alexi de Sadesky: *"No, it is not a thing a sane man would do. The Doomsday Machine is designed to trigger itself automatically"*
  - President Muffley: *"But surely you can disarm it somehow."*
  - Ambassador de Sadesky: *"It is designed to explode if any attempt is ever made to un-trigger it"*
  - President Muffley: *"Automatically?" [...] But this is absolute madness, Ambassador. Why should you build such a thing?..."*

    [Enter Dr. Strangelove]
  - President Muffley to Dr Strangelove: *"But how is it possible for this thing to be triggered automatically and at the same time impossible to un-trigger?"*
  - Dr Strangelove: *"Mr President, it is not only possible, it is essential. That is the whole idea of this machine, you know. Deterrence is the art of producing in the mind of the enemy the fear to attack. And so because of the automated and irrevocable decision-making process, which rules out human meddling, the Doomsday machine is terrifying and simple to understand and completely credible and convincing."*
- Different Types of Commitment:
  - Commitments to actions
  - Conditional commitments
  - Commitments on commitments
  - Randomized types of commitment

2. Threatening Inconsistency and Commitment Order

Conditional commitments may give rise to a type of inconsistency:

$$\begin{bmatrix} (3,1) & (0,2) \\ (1,3) & (2,0) \end{bmatrix}$$

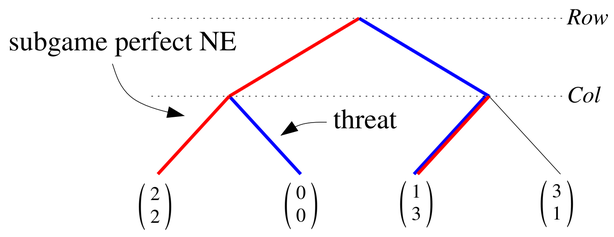
Row: "I play *top* if *Col* plays *left*, *bot* otherwise."
Col: "I play *right* if *Row* plays *top*, *left* otherwise."

Which actions are to be played without anyone's bluff being called?

A solution: agents commit in a particular order.

3. Backward Induction and Credible Threats

In Selten's argument equilibria that are not subgame perfect were excluded because they involved *incredible threats*. If commitments are possible these threats are thought of as credible.



$$\begin{bmatrix} \mathbf{(2,2)} & (2,2) & (0,0) & (0,0) \\ (1,3) & (3,1) & \mathbf{(1,3)} & (3,1) \end{bmatrix}$$

4. Formal Definitions of Commitment Types and Extortions

Definition (*Commitment types*) Be $\pi = (\pi_1, \ldots, \pi_n)$ a commitment order of agents and S a set of strategy profiles. A *commitment type* is a tuple $(X_{\pi_1}, \ldots, X_{\pi_n}, \phi)$, where:

- X_{π_i} an abstract set of commitments for π_i
- $\phi\colon X_{\pi_1} \times \cdots \times X_{\pi_n} \to S$

Definition (*Extortions*) For $G = (N, (A_i)_{i \in N}, (U_i)_{i \in N})$, $\pi = (\pi_1 \ldots \pi_n)$ and τ:

- Any commitment profile f is an extortion of order 0.
- A commitment profile $(f_{\pi_1}, \ldots, f_{\pi_m} h_{\pi_{m+1}}, \ldots, h_{\pi_n})$ is an extortion of order m, if for all extortions $(g_{\pi_1}, \ldots, g_{\pi_m}, h_{\pi_1}, \ldots, h_{\pi_{m-1}})$ of order $m-1$.:
  - $(f_{\pi_1}, \ldots, f_{\pi_m}, h_{\pi_{m+1}}, \ldots, h_{\pi_n})$ is an extortion of order $m-1$,
  - $(g_{\pi_1}, \ldots, g_{\pi_m}, h_{\pi_{m+1}}, \ldots, h_{\pi_n})' \precsim_{\pi_m} (f_{\pi_1}, \ldots, f_{\pi_m}, g_{\pi_{m+1}}, \ldots, g_{\pi_n})'$
- A commitment profile f is an extortion if it is an extortion of all orders.

Theorem 1 *Let G a game, π an order of commitment and τ an commitment type. If G^* is the extensive game in which the players have their commitments of type τ as their strategies and move in the order that is the reverse of π, then a commitment profile f is an extortion in G if and only if f is a backward induction solution in G^*.*

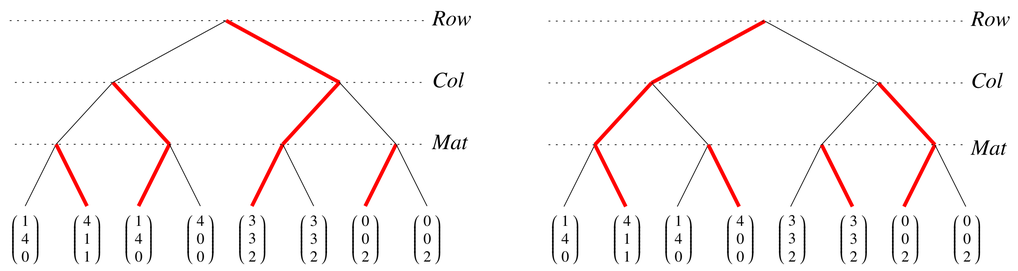
5. Conditional Commitments: Commitments on Actions of other Players

Conditional commitments allow players to commit earlier to commit on the possible actions of the players to commit later.

Definition (Conditional Commitments) For $G = (N, (A_i)_{i \in N}, (U_i)_{i \in N})$ and $\pi = (\pi_1 \ldots \pi_n)$ let $(F_{\pi_1}, \ldots, F_{\pi_n}, ')$ be such that:

- $F_{\pi_1} = A_{\pi_1}$
- F_{π_i} is the set of functions mapping $A_{\pi_1} \times \cdots \times A_{\pi_{1-1}}$ on $A_{\pi_i} \quad (i > 1)$
- For $f = (f_{\pi_1}, \ldots, f_{\pi_n})$, define $f' = (f'_{\pi_1}, \ldots, f'_{\pi_n}) \in A$ such that:

$$f'_{\pi_1} = f_{\pi_1} \quad \text{and} \quad f'_{\pi_{i+1}} = f_{\pi_{i+1}}(f'_{\pi_1}, \ldots, f'_{\pi_i})$$



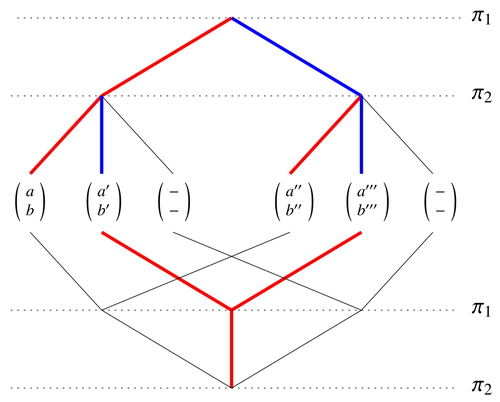
An extortion of order 1 (left) and an extortion of order 3 (right) in conditional commitments.

Theorem 2 *In each game, Pareto efficient conditional extortions exist. Moreover, any strategy profile that Pareto dominates an extortion is also extortionable through a conditional commitment.*

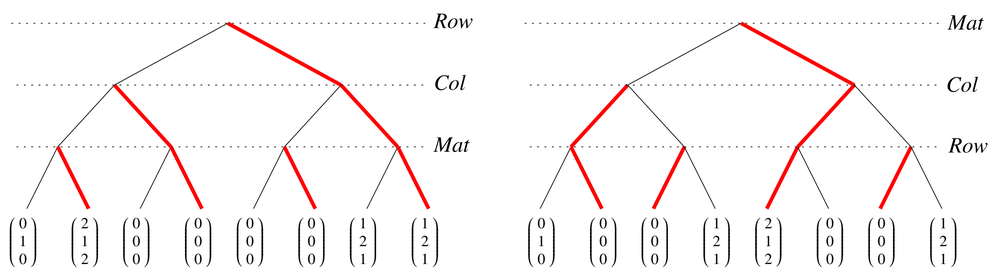
6. When To Commit?

It is always better to commit earlier than later in the two-player case with conditional commitments.

Theorem 3 *Let G be a two-player strategic game with player i and let f be an extortion in which i commits first, and g an extortion in which i commits second. Then, $g' \precsim_i f'$.*

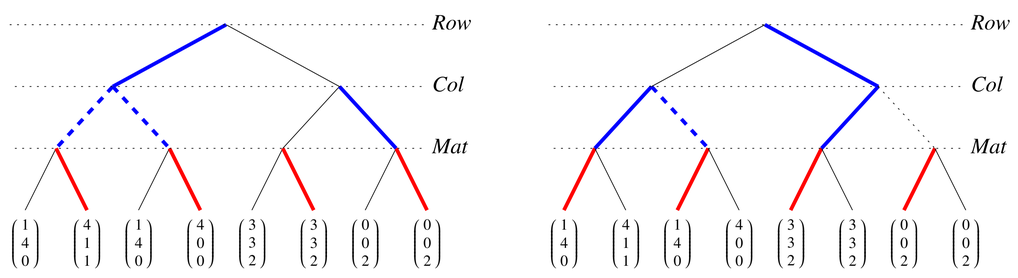


It is *not*, however, always better to commit earlier than later in the general n-player case.



7. Inductive Commitments: Commitments on Commitments

Idea: Given a commitment order $(pi_1, \ldots, \pi_n)$, player π_1 only commits to an action, player π_1 commits to the *commitments* of π_1, player π_2 commits to the *commitments* of the players π_1 and π_2, and so on.



An outcome in which *Mat* becomes a payoff of 2 is extortionable in inductive commitments.

8. Future Research

- A comparative study of different commitment types.
- Partial commitments, commitment violation.
- Computational complexity of extortions.

References

[1] V. Conitzer and T. Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce (ACM-EC)*, pages 82–90. ACM Press, 2006.

[2] D. Samet. How to commit to cooperation, 2005. Invited talk at the 4th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS).

[3] T. C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

[4] R. Selten. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft*, 121:301–324, 1965.

[5] M. Tennenholtz. Program equilibrium. *Games and Economic Behavior*, 49:363–373, 2004.

[6] H. von Stackelberg. *Marktform und Gleichgewicht*. Julius Springer Verlag, 1934.