Potsdam Institute for
Climate Impact Research

Game Theory for Climate Coalitions: Strategies for Compliance & Hierarchical Coalition Formation

Jobst Heitzig (PIK RD IV)
with contributions by Kai Lessmann and Yong Zou

Statistics Norway, Oslo, 10 January 2012

Leibniz
Gemeinschaft

Overview

- Problem:
  GHG emissions and free-riding

- Game theoretic framework
- Existing literature
- General model of the emissions game

- Making agreements self-enforcing: The LinC strategy

- Outlook & Conclusion

Problem:

GHG emissions and free-riding

Basic Facts

Emission of greenhouse gases (GHG: CO2, methane, ...)
▶ **Global warming** (increase in global mean temperature)
▶ **Climate change** (diverse regional effects, extreme events)
▶ **Damages** (economic, loss of life & biodiversity, ...)

- conservative estimates: IPCC's 4th assessment report 2007

GHG distribute fast & climate is a globally connected system
▶ Damages at place X **independent from place of origin** of GHG

- hence abatement (emissions reduction) is a **public good**

▶ Country X can hope that damages in X will be avoided because GHG emissions in *other* places will be reduced!

- **Free-riding** = "The others will solve the problem for me"

“Non-cooperative” game theoretic framework

Two Approaches to Studying Games

Basic distinction: How can agreements be enforced?

- "Cooperative" game theory assumes that **players can reach *binding agreements*** which are enforced by measures that are not themselves analysed (e.g. powerful courts)
- "Non-cooperative" game theory assumes that **agreements might at best be *self-enforcing strategies*** studied inside the game model (e.g. using threats of reciprocation)
- "Nash's program" tries to base the former on the latter

Non-Cooperative Formulation of the Emissions Conflict

- Countries can choose their own emissions levels
- Large **externalities**
  - *Globally,* a social planner would choose low emissions
  - *Individually,* marginal costs of emissions reductions soon exceed the individual benefits of avoided damages
- If a player treats the emissions levels of the others as *given* (at whatever level), it is best to emit a lot
  - ▶ Nash equilibrium payoffs are **inefficient** (similar to Prisoners' Dilemma)

- International agreements are not easily enforceable
- **Free-rider incentive:** Even if I *agree* with others to emit less, I can profit even more by *not complying*

My Basic Approach at a Solution in the Non-Cooperative context

- To make the others cooperate and reduce emissions, I have to reach a self-enforcing agreement with them that
  - encourages to emit less (by sharing the reduction burden)
  - discourages free-riding

- The latter can only be done via *threats*, so it requires a game model that allows for **reacting** on others' actions
  - e.g., using issue linkage (trade, ...)
  - or a game with a small number of different *stages*
  - or a **repeated game** with infinitely many similar *periods* allowing for **strategies** that react suitably to non-compliance

Examples of Strategies in the Repeated Prisoners' Dilemma

	defect	coop.
defect	1 1	0 5
coop.	5 0	3 3

- **Trigger strategies**
  - **Grim**: Cooperate as long as *the other* never defected before
  - **SymT**: Cooperate as long as *no player* ever defected before
- **Tit For Tat** (TFT)
  - Start to cooperate, then do what the other did the last time
- **Getting Even** (GE) avoids the “echoing” problem of TFT
  - Start to coop., then defect if the other has defected more often in the past
- **Contrite Tit For Tat** (CTFT)
  - Start to coop., then defect whenever the other is in “bad standing”
    - A player is in “bad standing” iff, in the previous period, he defected although CTFT told him to cooperate
  - **We will use a similar recursive idea in the emissions game!**

Some Formal Stability Concepts in Games with Stages or Periods

- **Equilibrium Concepts**

pure strat. eq., Nash, correl.
no *individual player* wants to switch strategy right away

TFT

strong Nash, coal.-proof, …
no *group of players* wants to switch strategies right away

subgame-perfect
no *individual player* wants to switch strategy *after any history*

~~TFT~~ ~~Grim~~ SymT

groupwise subg.-perfect
no *group of players* wants to switch strategy *after any history*

new!?

- **Renegotiation-Proofness** (Farrell & Maskin '89, Bergin & MacLeod '93)

weakly reneg.-proof (WRP)
after no history it profits all players to *pretend history was different*

~~TFT~~ ~~Grim~~ ~~SymT~~ GE

strongly reneg.-proof
after no history it profits all players to *switch to a different WRP agreement*

"strong perfect": future payoffs are Pareto-efficient after each history

Discounting and Folk Theorems

- **Discounting** future payoffs $P_i(t)$

  Exponentially (with a constant discount factor δ)
    - Utilities (= discounted long-term payoffs) $U_i(t) = \sum_{t' \geqslant t} P_i(t)\,\delta^{t'-t}$
  - Hyperbolically (with a declining discount rate)
  - ...? (inter-generational discounting seems a hard philosophical question)

- **Folk Theorems** are of this form:
  - *For a repeated game and a given payoff vector: If both fulfil some conditions and **if δ is close enough to 1,** there is a (usually Grim-like) strategy vector that realizes these payoffs and has some stability property X*
  - No known folk theorem seems to suffice in our case...

Existing literature

in the non-cooperative framework

The Emissions Game as a Multi-Player Repeated Prisoners' Dilemma

- Cooperate = emit little
  Defect = emit much
- **Froyn & Hovi 2008** present a CTFT-like strategy which...
  - punishes a *unilateral* deviation
    with defection by a carefully chosen subset of other players
  - is **subgame-perfect** (but not groupwise)
  - is **weakly renegotiation-proof** (but not strongly)
- **Asheim & Holtsmark 2009** show that this still works if...
  - emissions levels can be chosen more freely
  - the game has a certain *symmetric* payoff structure

Scott Barrett's Work

- Many eloquent papers on the problem since 1989
- Overall rather pessimistic findings
- But **CAUTION!**
  - Mostly uses quite specific and symmetric payoff structures (results don't always carry over to other payoff structures)
  - Formal arguments sometimes incomplete or even flawed
  - Game-theoretic terminology and definitions sometimes non-standard
- E.g., the pessimistic claim in his chapter in the Handbook of Environmental Economics (2005), p. 1491–93, is implicitly disproved by Asheim & Holtsmark 2009

A General Model of the Emissions Game with Emissions Trading

A General Model of the Emissions Game with Emissions Trading (1)

- Repeated game in **periods** (e.g. 4-year), between n countries or regions
  - Critical simplification: Same payoff structure in all periods (in reality, GHG gases are stock pollutants & technology lowers costs)

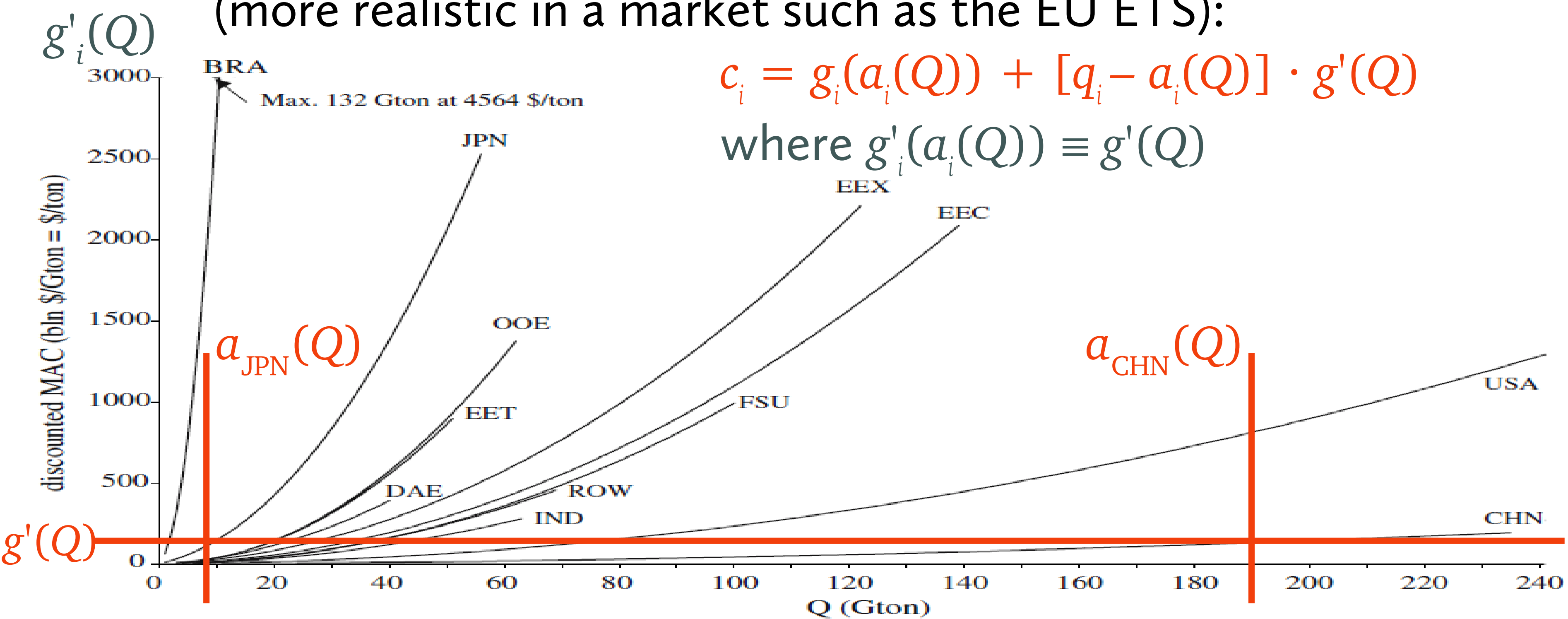
- **Individual contribution** of player i in period t is $\boldsymbol{q_i(t) = \textit{reference emissions} - \textit{net emissions}}$
  - may be negative, since large amounts of permits might be traded!
- Total contributions $Q(t)$ lead to
  - **total period costs** $\boldsymbol{C(t) = g(Q(t))}$
    - for some convex function g with $g(Q \leq 0) = 0$
  - **individual period benefits** $\boldsymbol{B_i(t) = f_i(Q(t))}$
    - for increasing functions f with $f_i(Q=0) = 0$ and $\lim_{Q \to -\infty} f_i(Q) = -\infty$
    - e.g. discounted consumption losses for i avoided after t

A General Model of the Emissions Game with Emissions Trading (2)

- **Total period costs $g(Q)$ are shared in some way, leading to individual period costs c_i**
  - e.g. **proportionally**: $c_i = q_i \cdot g(Q)/Q$
  - or with **marginal cost pricing** based on indiv. cost fcts. g_i (more realistic in a market such as the EU ETS):

$$c_i = g_i(a_i(Q)) + [q_i - a_i(Q)] \cdot g'(Q)$$

where $g'_i(a_i(Q)) \equiv g'(Q)$



Example: Individual Costs If Cost Functions Are Equal

- **Typical in the literature (without emissions trade):**
  - quadratic individual costs $\boldsymbol{c_i = q_i^2/2}$

- **Similar structure with emissions trading:**
  - quadratic individual cost functions: $g_i(x) = x^2/2$
  - marginal cost pricing requires $g'_i(a_i(Q)) = g'_j(a_j(Q))$
    hence $a_i(Q) = a_j(Q) = Q/N$, $g(Q) = Q^2/2N$, $g'(Q) = Q/N$
  - individual costs: $\boldsymbol{c_i} = g_i(a_i(Q)) + [q_i - a_i(Q)]\, g'(Q)$
    $= (Q/N)^2/2 + [q_i - Q/N]\, Q/N$
    $\boldsymbol{= q_i Q/N - Q^2/2N^2}$

A General Model of the Emissions Game with Emissions Trading (3)

- Individual period payoffs $P_i(t) = f_i(Q(t)) - c_i(t)$
    - or a concave increasing function of this, e.g. $\log[f_i(Q(t)) - c_i(t)]$
- Usual assumptions of classical non-coop. game theory
    - Common knowledge of rationality
        - All know that all know that … that all are rational
    - Complete information
        - For all i,j and $t' < t$, $q_j(t')$ is known to i before she chooses $q_i(t)$

- **Goal: find a strategy vector that**
    - **realizes the optimal emissions level**
    - **has as good stability properties as possible**

A Crucial Consequence of Convexity

- If g, g_i are convex, both sharing rules are also convex in a sense: there is a "**cost sensitivity**" $\gamma(Q)$ so that
  - **reducing** contribution q_i by some amount $x > 0$ lowers the costs c_i by at most $x \cdot \gamma(Q)$
  - **redistributing** some amount $x > 0$ from q_{-i} to q_i raises the costs c_i by at least $x \cdot \gamma(Q)$
    - with proportional sharing, $\gamma(Q)$ equals average costs: $c_i = q_i \cdot g(Q)/Q$, $\gamma(Q) = g(Q)/Q$
    - with marginal cost pricing, $\gamma(Q)$ equals marginal costs: $c_i = g_i(a_i(Q)) + [q_i - a_i(Q)] \cdot g'(Q)$, $\gamma(Q) = g'(Q) \equiv g'_i(Q)$
- This relationship between the effects of reducing and redistributing contributions motivates the strategy LinC...

Making agreements self-enforcing:

The LinC strategy

(Heitzig, Lessmann, Zou 2011)

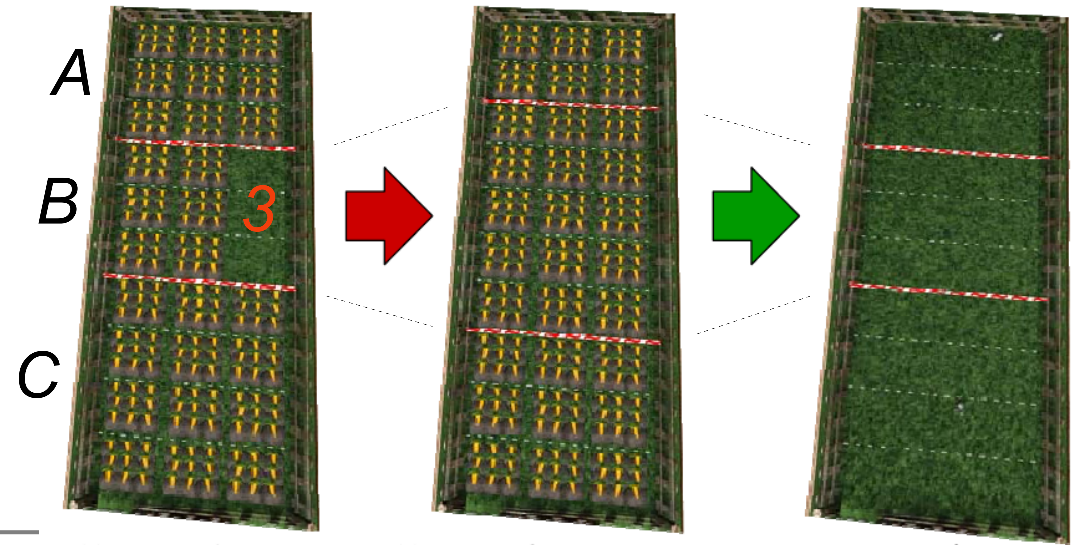
Solution: the Strategy "LinC" (Linear Compensation of Shortfalls)

- Q^* = global optimum contributions, maximizing the total payoff
- Let q^*_j be *any* allocation of Q^* into individual **targets**
  (emissions trading makes the total payoff independent of this allocation!)
- Define dynamic **liabilities** $l_j(t)$
  - initially equal to the targets: $l_j(1) = q^*_j$
  - always comply with your liability: put $q_i(t) = l_i(t)$
- After each t, compute the **shortfalls** $d_j(t)$
  - $d_j(t) = l_j(t) - q_j(t)$ if $q_j(t) < l_j(t)$, otherwise $d_j(t) = 0$
  - $\bar{d}(t) = (\text{average shortfalls in } t) = \sum_j d_j(t)/n$
- **Redistribute the liabilities** linearly for compensation:
  - $l_j(t+1) = q^*_j + [d_j(t) - \bar{d}(t)] \cdot \alpha$ with a sufficiently large α

$$l_j(t+1) = q^*_j + [d_j(t) - \bar{d}(t)] \cdot \alpha$$

Small Example: Growing Carrots in a Community Garden

- Assume $n = 3$, optimal contributions $Q^* = 30$, and individual targets $q^*_A = q^*_B = 9$, $q^*_C = 12$
- Initial liabilities equal the targets: $l^*_{(A,B,C)}(1) = (9,\mathbf{9},12)$
- B falls short by $d_B(1) = \mathbf{3}$ units, so next period's liabilities are redistributed, say using $\alpha=2$: $l^*_{(A,B,C)}(2) = (6,\mathbf{15},9)$
- In that period, all fulfil their liabilities, so in period 3, they are back to normal: $l^*_{(A,B,C)}(3) = (9,9,12)$



$$l_j(t+1) = q^*_j + [d_j(t) - \bar{d}(t)] \cdot \alpha$$

Results: If All Players Apply LinC, this is...

- **Pareto-efficient** in every subgame ("strongly perfect")
  - **because of emissions trading**, it only matters that $Q(t) = Q^*$
- hence **strongly renegotiation-proof**
  - no deviating group can hope to afterwards convince the others to overlook their deviation or to switch to a new strategy
- a strong Nash equilibrium in every subgame (proof later) ("**groupwise subgame-perfect**")
  - no group of players can increase their joint discounted future payoffs by deviating from LinC, even when some deviations have already happened, assuming that the other players will apply LinC
- **timely, proportionate & robust** against small errors
  - If $d_i(t) \sim \mathrm{N}(0,\sigma^2)$, then $l_i(t+1) - q^*_i \sim \mathrm{N}(0,\sigma^2\alpha^2(n-1)/n)$
  - errors do not accumulate (similar to "trembling hands perfectness")

Proof of Groupwise Subgame-Perfectness (1)

- Contributing *too much* does never pay
  (otherwise it would raise the total payoff which is impossible since Q^* is optimal)
- **Proof of *one-shot* groupwise subgame-perfectness:**
  If some proper subgroup G of players deviates in one period t only, together contributing an amount x too little, then...
  - Joint shortfalls are $d_G(t) = l_G(t) - d_G(t) = x$, avg. shortfalls $\bar{d}(t) = x/n$
  - By convexity, G's joint **gains** in t are less than $\gamma(Q^*) \cdot x$
  - In $t+1$, the amount of liability that is redistributed towards G is
    $$(x - |G|x/n) \cdot \alpha$$
  - By convexity, G's **losses** in $t+1$, discounted because of the delay, are at least $\gamma(Q^*) \cdot x \cdot (1 - |G|/n) \cdot \alpha \cdot \delta$
  - These losses are larger than the above gains if α is sufficiently large
    (see paper for details)

Proof of Groupwise Subgame-Perfectness (2)

- **Proof of *finite-shots* groupwise subgame-perfectness**, using a standard argument
  - Assume the shortest length of deviations that can increase some group G's utility is m, with a return to LinC afterwards
  - After the first $m-1$ deviations, the group will not want to deviate another time (because of one-shot subgame-perfectness)
  - Hence alread the first $m-1$ deviations alone must have been profitable, so there is a shorter profitable sequence of deviations – a contradiction to the choice of m

Proof of Groupwise Subgame-Perfectness (3)

- **Sketch of remaining proof:** (see paper for details)
  Assume G plays an *infinite* sequence of shortfalls that pays.
  - If the discounted long-term shortfalls are *finite*, one can find a length m so that it would still pay to play only the first m shortfalls and then returning to LinC
    - But we proved already that such a finite sequence cannot exist
  - If the discounted long-term shortfalls are *infinite*, one can show that the cut down long-term costs are finite while the long-term benefits decrease infinitely
    - Hence such a sequence of deviations is infinitely bad
    - This is because of a period-by-period **escalation** in which the other players emit more each period as a punishment

Remarks (1)

- The proof requires that individual emissions could *in principle* be raised **unboundedly** (at least step-by-step)
  - If this is not so, a variant with bounded liabilities can be used
    - Then the condition for groupwise subgame-perfectness is more complicated
      - First simulations with estimated cost/benefit models from the literature show that this might still work
- It is essential that both...
  - the deviators are required to **make up** for their shortfalls
    - similar to the current Kyoto/Marrakach rules
  - the others are allowed to *emit more* as a **punishment**
    - similar to defection as punishment in the Prisoners' Dilemma

Remarks (2)

- LinC needs few information to be implemented
  - global emissions target Q^* and some regional allocation q^*_i
  - estimate of global marginal costs and benefits at this target
  - monitoring of regional emissions $q_i(t)$
- **LinC can stabilize *any* target allocation q^*_i**

  ➔ **Problem of equilibrium selection:**
  Which allocation will be realized?

  ➔ Negotiations & agreement about the allocation are necessary

  ➔ LinC will mainly be useful to ensure **compliance**,
  not to ensure initial **participation** in a climate coalition

  ➔ "Cooperative" analysis needed to study coalition formation!

Outlook & Conclusion

Possible Political Roadmap using LinC

- **One or more "coalitions of the willing"** each agree...
  - on an **internal Cap & Trade** regime with some initial individual caps
    - maybe sub-optimal/pragmatic ("hot air", "grandfathering") to ensure participation
  - internal usage of **LinC to ensure compliance**
    - requires sufficient monitoring capabilities (e.g. satellite-based)
  - usage of e.g. **border taxes against non-members**
- **Caps get adjusted** each time when...
  - **non-members join** a coalition to avoid the border taxes
  - several **coalitions merge**
    - to be more efficient with a merged emissions market
  - major **changes in cost/benefit estimates**
  - ...keeping track of shortfalls, not "letting bygones by bygones"
- Hope: eventually, **a grand coalition forms**
  - and the global cap approaches the optimum

Cooperative Formulation of the Emissions Conflict

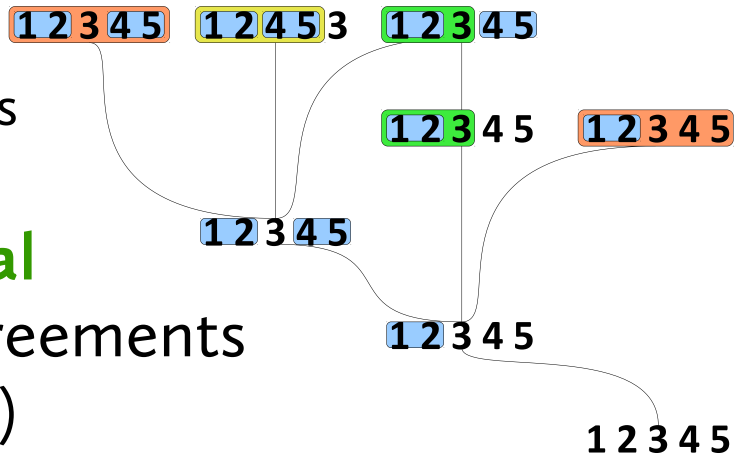
- Players can choose to form **coalitions** in some way
  - each coalition tries to maximize its joint long-term utility
    - based on some assumptions on the other players' behaviour

- **Free-rider incentive:**
  I might gain by leaving/not joining a coalition
  - depending on how coalition(s) will then change
    - **models of coal. formation, farsightedness**

- If large coalitions are **unstable**, only small ones form
  - *resulting global emissions are then inefficiently high*

My Basic Approach at a Solution in the Cooperative context

- Assume that already formed coalitions can enter further agreements to form larger coalitions
  - **hierarchical agreements, coalitions of coalitions**
  - corresponds to some proposals from political science
    - negotiations between groups of players
    - regional climate agreements
    - merging of existing carbon markets



- in a suitable **model of hierarchical coalition formation,** efficient agreements might be stable (in a suitable sense)

To Do

- Better models of (hierarchical) coalition formation when agreements are **reversible** (as in reality)
  - Some first approaches: Konishi&Ray 2003, my SSRN paper
- Numerical **simulations** of LinC with recent cost/benefit estimates
- Model **non-identical periods**
  - declining costs due to **technology** (exo- or endogenous)
  - **stock pollutant** nature of GHG
  - long-term **investment** decisions
- Issue linkage, network structure, ...

Take Home Messages

**With emissions trading,
redistribution of liabilities can be
a credible threat against non-compliance**

- e.g. simply using linear compensation

**If coalitions can build hierarchically,
a global coalition might emerge
even when externalities are large**

*Thank you for your attention
– I'm curious for your comments!*