Potsdam Institute for
Climate Impact Research

Bottom-Up Strategic Linking of Carbon Markets: Which Climate Coalitions Would Farsighted Players Form?

Jobst Heitzig

Potsdam Institute for Climate Impact Research, Transdisciplinary Concepts and Methods,
P. O. Box 60 12 03, 14412 Potsdam, Germany, heitzig@pik-potsdam.de

Presented to EAERE2012 and GAMES2012 conferences
This version July 28, 2012

Abstract

We present typical scenarios and general insights from a novel dynamic model of farsighted climate coalition formation involving market linkage and cap coordination, using a simple analytical model of the underlying cost-benefit structure. In our model, the six major emitters of CO_2 can link domestic cap-and-trade systems to form one or several international carbon markets, and can either choose their emissions caps non-cooperatively or form a hierarchy of cap-coordinating coalitions inside each market. Based on individual and collective rationality and an assumed distribution of bargaining power, we derive scenarios of such a climate coalition formation process which show that a first-best state with a coordinated global carbon market might well emerge bottom-up, and underline the importance of coordinating caps immediately when linking carbon markets. Surprisingly, the process tends to involve less uncertainty when agreements can be terminated unanimously or unilaterally, depending on the level of farsightedness.

keywords: climate policy, international environmental agreements, cap and trade, coalition formation, farsightedness

Contents

1 Introduction **3**

1.1 Background 3

1.2 Modelling strategic behaviour in a process of market linkage and cap coordination 4

2 Static cost-benefit analysis **10**

2.1 Game structure 10

2.2 A simple analytical cost-benefit model 12

3 Results from the dynamic model **15**

3.1 Influence of the level of rationality 16

3.2 Importance of the possibility of immediate cap coordination 17

3.3 Influence of the level of farsightedness 21

3.4 Influence of the level of reversibility of agreements 22

3.5 Influence of the distribution of bargaining power 24

4 Conclusion **25**

Appendix: A general dynamic model of farsighted coalition formation **32**

Equilibrium processes 32

Algorithms 32

Market structures, coalition hierarchies, states, codes, and moves 33

The Random Proposer Amended Move Bargaining Game 37

1 Introduction

1.1 Background

Failed efforts of international GHG emissions mitigation. Since the 2009 Copenhagen Accord there seems to be almost unanimous agreement in the international community about the fact that one should not allow global mean temperatures to exceed two degrees Celsius above pre-industrial levels, and that this can only be ensured by reducing greenhouse gas (GHG) emissions vastly. Still, the 2011 Durban result shows that it seems quite unlikely that the current process of global negotiations will soon lead to a global climate treaty that can be expected to achieve that goal.

To a large extent, this problem of forming a “grand coalition” is related to the special payoff structure of the conflict, involving large externalities both due to the global effect of GHG emissions, making emissions reductions a global public good, and due to expected spillovers from investment into mitigation technology R&D. These give many countries incentives to “free-ride” in a number of ways, e.g., by not joining an emissions market, not agreeing to introduce an emissions cap or to lower their existing cap, or waiting with their own mitigation until others have reduced the costs by investing into R&D.

Such free-riding incentives typically exist both for the signatories of an agreement after it has come into effect, making various forms and levels of non-compliance likely if the agreement cannot be enforced in some way, and they exist even more at the pre-agreement negotiation stage where players can hope that if they do not participate, other players might still form a substantial coalition since most countries cannot convincingly commit themselves to not cooperate with others unless there is a global treaty.

While the problem of *non-compliance* might be solved by trying to make the agreement sufficiently binding via some sanctioning mechanism, e.g., by setting liabilities via a self-enforcing dynamic redistribution strategy (Heitzig et al. 2011), the *non-participation* problem can be shown to remain even when some form of emissions trading is possible (Helm 2003), seemingly in contrast to the intuition given by the Coasian argument that in such a situation, rational players will eventually implement the collectively rational outcome by signing a suitable agreement. This can be seen as one of the main reasons why international efforts to reduce GHG emissions have achieved so little yet beyond the meager level of emissions abatement that seem profitable for each individual player in a non-cooperative setting, other reasons being the long time scales and high uncertainties involved and the supposedly bounded rationality of the players.

Reaching global cooperation by linking carbon markets and then coordinating caps. In view of this situation, the recent literature on international climate policy has developed a number of proposals for bilateral or regional cooperation (see, e.g., the excessive collection in Aldy and Stavins, eds 2009), involving some but not all of main emitters of GHG, and some of these ideas can be combined into a story of the following kind:

1. First, individual countries or regions establish domestic or regional emissions trading schemes with individually chosen emissions caps to achieve their individually rational mitigation goals in a cost-efficient way.

2. **Market linkage.** Later on, several such emissions markets might be linked in order to further increase efficiency and reduce mitigation costs by equalizing the different abatement cost curves of the individual markets, which might in turn lead to a lowering of the individually rational emissions cap of these players (e.g., Flachsland et al. 2009b; Tuerk et al. 2009).

3. **Cap coordination.** If the member countries of such a linked emissions market agree to coordinate the amounts of permits each member issues, they might further reduce the market-wide cap to the collectively rational level by internalizing the effect of their respective emissions on each other (e.g., Jaffe and Stavins 2008; Flachsland et al. 2009a).

4. Eventually, a (near-)global emissions trading scheme might emerge in this way in which all major emitters coordinate their caps to achieve the globally optimal level of emissions reductions in a cost-efficient way (e.g., as in Fig. 1).

Although the very first steps along this line are taken already in the sense that several domestic carbon markets are in the process of being linked and the respective caps are coordinated to some extent, the stepwise nature of the whole envisioned bottom-up process and the complex effects of linking markets and coordinating caps on the payoffs of both the members and non-members of the respective agreements entail several possibilities for strategic behaviour by the relevant players in such a process, the more so the more large emitters there are. It is therefore far from obvious which markets can be expected to be linked and which caps can be expected to be coordinated in which order, how likely this will lead to a global market with a globally efficient total cap, and what final distribution of the surplus generated by this process can be expected.

Such long-term expectations would however be relevant already for the decisions taken at the present beginning of the process. Fig. 2, for instance, shows the difference in static and long-term evaluation of the initial moves of the process for Europe (for a certain choice of parameters and if our simplified model of the underlying payoff were correct). E.g., it might even be long-term profitable to form a small, initially uncordinated market with the Former Soviet Union or with Japan, although this would reduce Europe's payoff if it were to remain the final state.

1.2 Modelling strategic behaviour in a process of market linkage and cap coordination

To give tentative answers to this key question of international climate policy, the present paper describes several scenarios of how such a bottom-up process of market linkage and cap coordination might evolve, but not considering the strategic effects of R&D or carbon leakage at this time.[1]

After deriving a simplified analytical model of the relevant inter-player cost-benefit structure of GHG emissions abatement in Sect. 2, we present several of the resulting scenarios in Sect. 3 and discuss by their means the various strategic effects that the model suggests might occur also in the real-world process. In the Appendix, we give a concise formal description of the game-theoretical model of farsighted coalition formation and the numerical algorithm used to find a set of transition probabilities between the possible intermediate states of the process that are consistent with individual and collective rationality and the assumed bargaining power.

Existing literature. Considering market linkage without the possibility of cap coordination, Helm 2003 and Carbone et al. 2009 showed that a global market is unlikely to emerge at once and that market linkage might actually lead to higher global emissions. He analyses a three-stage game in which countries can first agree to establish a global carbon market, then choose their individual emissions caps non-cooperatively, and then trade these permits if in the beginning all have agreed to establish the market. He observes that "this reflects a rather simplified view of international negotiations. Considering coalition formation would, however, substantially complicate the analysis."

Even without the complication of modelling emissions trading, the literature on coalition formation in the climate context has so far usually assumed a symmetric cost-benefit structure that does not reflect the asymmetries in both the benefit and cost functions of real-world players which Helm 2003 shows to be quite influential in his model. Also, the literature has mainly considered only quite restricted possibilities of coalition formation, typically assuming an "Open Membership Single Coalition Game" in which at most one instead of several disjoint coalitions might form, and in which players can join or leave the coalition without consent of the other members, leaving the remaining coalition intact (e.g., Carraro and Siniscalco 1993; Barrett 1994). The effects that a decision to not participate may have on the participation decisions of the other players are usually not considered. Only few authors consider closed-membership coalitions or the simultaneous formation of several coalitions, and without the possibility of later merging these into larger coalitions (see Finus 2003 for an excellent overview). Instead of using different models of coalition formation, the newer literature tends to focus on varying the underlying payoff structure, for instance by discussing transfer schemes (e.g., Nagashima et al. 2009 and refs. therein) or by including policy instruments such as tariffs (e.g., Lessmann et al. 2009).

[1] R&D cooperation could be included in several ways, either by assuming that cap coordinating coalitions also coordinate R&D investment (e.g., Buchner et al. 2005), or by assuming agreements on R&D are independent from cap coordination, but possibly restricted to the members of a market or a cap-coordinating coalition.

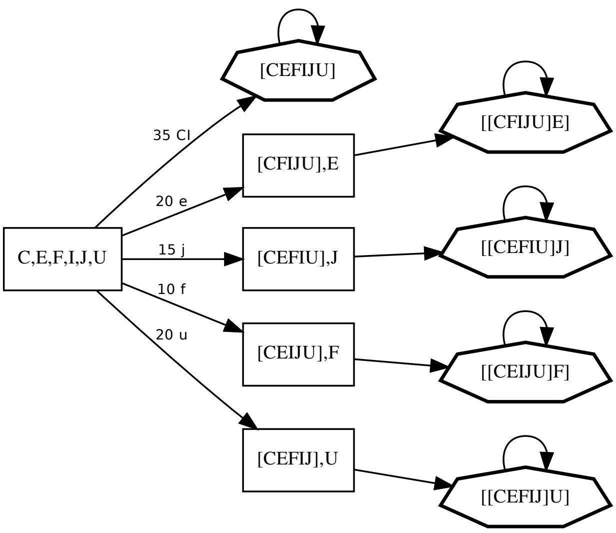


Figure 1: Process of climate coalition formation, typical model result for a wide range of parameter values, in which a fully coordinated global carbon market results after one period (with 35% probability, if the permit sellers C(hina) and I(ndia) get their way) or two periods (with 65% probability, if E(urope), F(ormer Soviet Union), J(apan), or U(SA) manage to stay out of the market at first). Boxes and diamonds are states, e.g., the left box C,E,F,I,J,U is a world with six domestic carbon markets, [CEFIJU] is a global carbon market in which caps are coordinated marketwide, [CEFIJ],U is a world with a large coordinated market excluding U, and [[CEFIJ]U] denotes that U has now joined that market. Arrows are moves between states. In case of conflicting interests, arrow labels indicate transition probabilities [%] and those players who prefer the move over the other moves shown (lowercase letters for players who are not among the initiators of the move). Diamond-shaped nodes are stable states with an optimal global cap, differing only in the burden- or surplus-sharing result. See Table 1 for payoffs.

State	**Static payoffs**						**Discounted average long-term payoffs**					
	C	E	F	I	J	U	C	E	F	I	J	U
C,E,F,I,J,U	94	394	115	86	304	347	254	609	200	206	458	555
[CEFIJU]	484	785	310	379	597	738	*484	785	310	*379	597	738
[CFIJU],E	330	1182	233	263	481	584	352	*1204	244	280	498	606
[CEIJU],F	421	721	378	331	550	675	443	743	*389	348	566	696
[CEFIU],J	375	676	255	297	960	629	385	686	260	305	*968	639
[CEFIJ],U	326	626	231	260	478	1055	357	658	246	284	502	*1087
C-E-F-I-J-U	180	>360	231	217	321	>338	380	≫510	306	329	≫434	≫488
C-E-F-I-J,U	147	>381	189	169	317	439	≫237	≫466	232	230	≫378	659
C,E,F,I,J-U	91	385	112	84	306	>338	175	480	160	147	512	607
C,[EFIJU]	215	565	200	214	433	519	329	680	258	300	519	633
[CFJU],E,I	298	1020	217	219	457	551	344	1078	240	255	492	597
[CEJU],F,I	381	681	328	245	520	635	430	730	354	282	556	684
[CEFJU],I	443	743	289	280	566	696	470	771	303	301	587	724
[CU],E,F,I,J	164	677	195	146	512	418	254	808	245	214	606	≫523
[[CFIJU]E]	374	1226	255	296	514	628	374	1226	255	296	514	628
[[CEIJU]F]	464	765	400	364	582	718	464	765	400	364	582	718
[[CEFIU]J]	395	696	265	312	975	649	395	696	265	312	975	649
[[CEFIJ]U]	389	689	262	307	526	1119	389	689	262	307	526	1119

Table 1: Static and long-term payoffs [bln. US$ per 100 years] in the process shown in Fig. 1 for a typical choice of parameters (medium farsightedness $\delta = 0.5$, subjective bargaining power distribution, agreements unilaterally terminable), for states reached with positive probability (codes in boldface) and some alternative states. * Favourite undominated move of this player in state C,E,F,I,J,U. > Move is not statically profitable for this initiating player. ≫ Move is not long-term profitable for this initiating player.

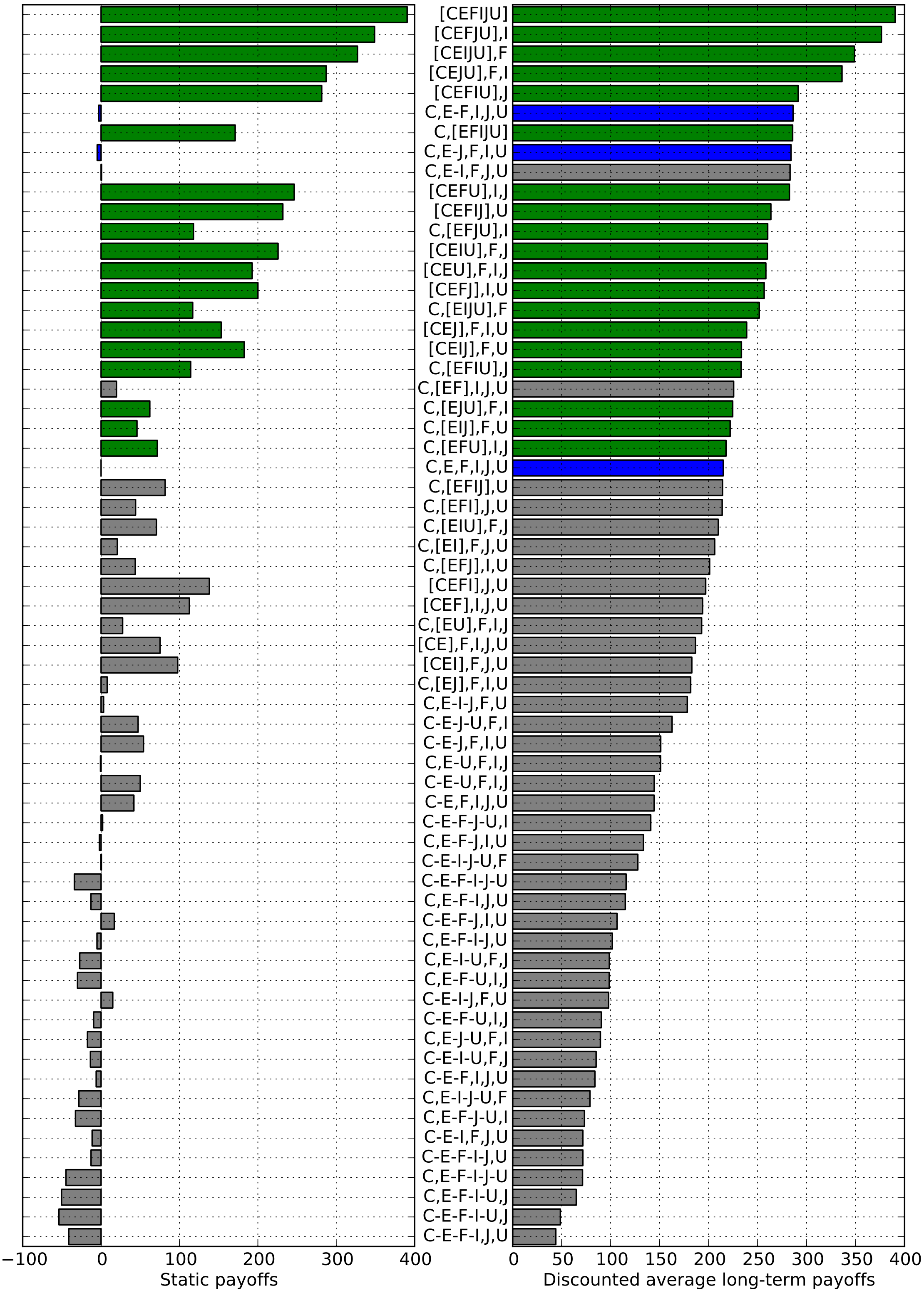


Figure 2: Relative payoffs [bln. US$ per 100 years] for E(urope) in the scenario of Fig. 1 and Table 1, in states reachable by one move involving E (blue=uncoordinated, green=coordinated market linkage; grey: move is dominated or unprofitable for another required player). If a coordinated global market cannot be formed, E would prefer to rather leave out only I(ndia) rather than leave out another player. Note that although forming an uncoordinated market with F or J (top two blue bars) is not statically profitable, it is long-term profitable because of what would happen afterwards.

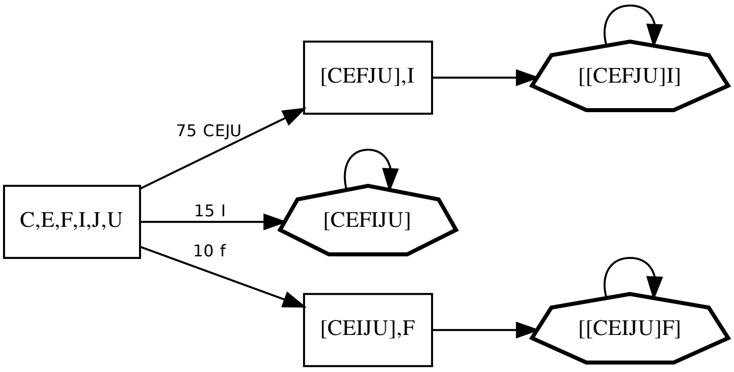


Figure 3: Alternative to Figs. 1 and 4 for the case when players are highly farsighted ($\delta = 0.9$) and any agreement can be terminated unanimously (or even unilaterally) by its signatories. The most likely path is now that all but I establish a coordinated market which is later joined by I.

Outside the environmental economics community, however, the general game-theoretic literature on coalition formation has produced a number of much more sophisticated models of coalition formation, allowing not only for multiple disjoint coalitions to form, but also assuming *farsighted* players that anticipate the possible effects of participation decisions on other players' participation. Such models, e.g., assume some kind of Rubinstein-type protocol-based bargaining game (Ray and Vohra 1999), or allow for groups of players to "block" coalitions (Ray and Vohra 1997), or model the coalition formation process as a dynamical process that moves between possible states with certain transition probabilities (Konishi and Ray 2003). One such concept (of Chwe 1994) has been applied to the climate context (see Osmani 2011; Osmani and Tol 2010 and refs. therein), focussing on the stability of a certain partition of the players into coalitions under a set of possible changes to that partition, but not modelling the process of coalition formation itself.

Approach of the present paper. In view of the observable complications of international climate negotiations, it seems plausible that real-world players are indeed farsighted to a certain extent, and we believe that the above envisioned process of successive market linkage and cap coordination is best modelled in a way similar to the *dynamic* model of Konishi and Ray 2003.

To this end, we derive in Sect. 2 an analytical cost-benefit structure that estimates the payoffs to the six major emitters of GHG in each imaginable *state* of the process, i.e., given any possible configuration of which markets are already linked and which members have already coordinated their caps, extending the derivation from Helm 2003 to the multiple-markets and cap coordination case. In addition, we extend the dynamic coalition formation model of Konishi and Ray 2003 to allow for two types of "coalitions" (linked markets, and groups coordinating their caps) and different types of transitions between the states of the process, as described in the Appendix.[2]

Both ingredients are then used to derive several instructive scenarios of how the dynamic formation of linked carbon markets with coordinated caps might evolve depending on a small number of parameters representing the amount of farsightedness, the types of transitions considered possible, and the distribution of bargaining power.

These scenarios take the form of coalition formation process diagrams like those in Figs. 1 and 4, and they indicate that despite the pessimistic results of earlier game-theoretic studies, a global carbon market with a first-best cap might well emerge eventually if players manage to combine the linking of carbon markets with an immediate coordination of the respective emissions caps.

[2] Although at this point, this cost-benefit structure is assumed to stay constant through time, the dynamic nature of our model would easily allow us to replace it by a path-dependent or explicitly time-dependent payoff structure derived from a sophisticated dynamic integrated assessment model in a later study.

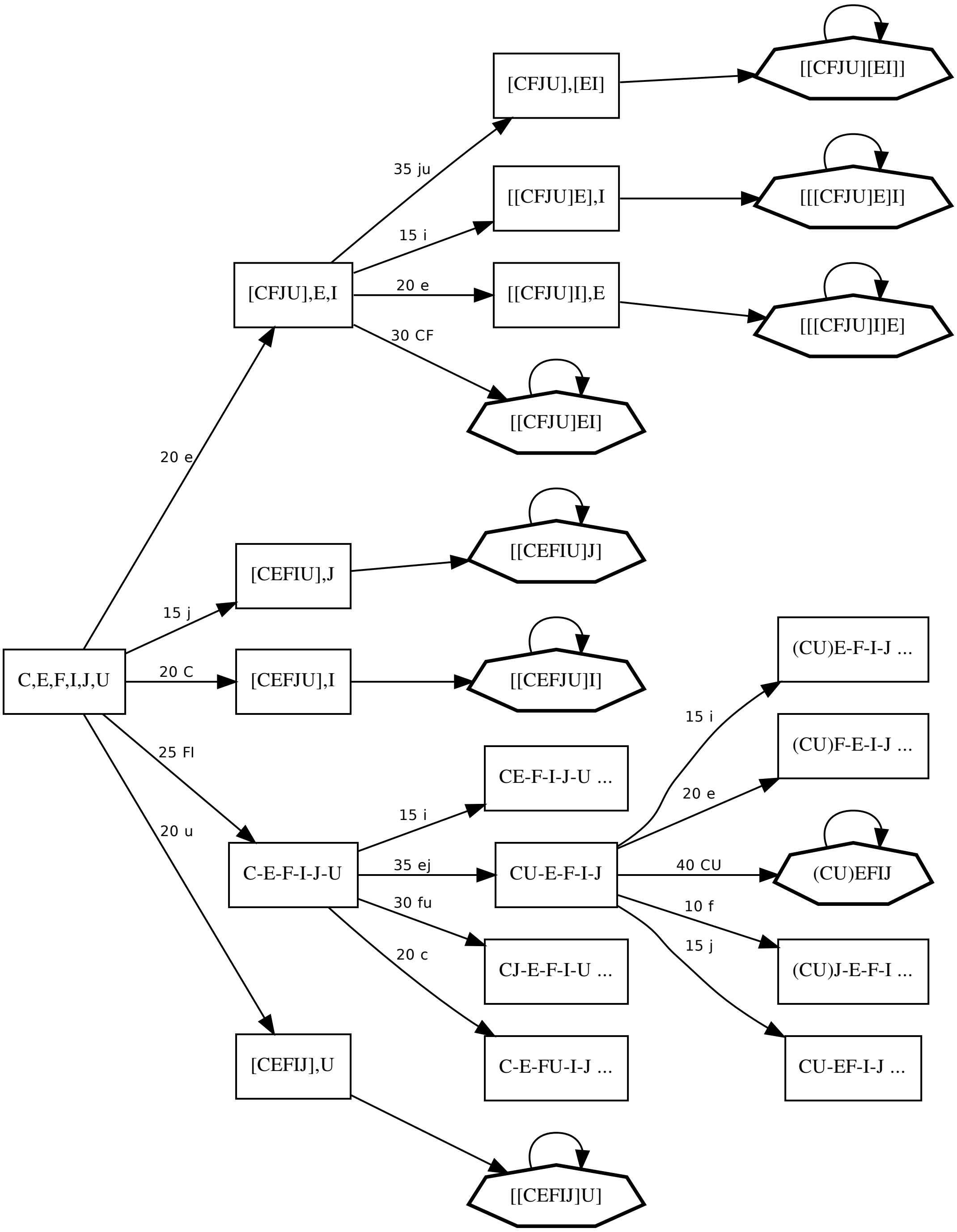


Figure 4: Alternative to Fig. 1 with typical complications occurring if players are highly farsighted and agreements are irreversible (see Fig. 3 for the reversible case). In view of the expected later moves, F and I now prefer to establish a global market C-E-F-I-J-U that only later coordinates its caps and in which all members prefer to join cap coordination late. Only the continuation path with highest probability is shown here, ending in a fully coordinated market (CU)EFIJ in which C and U have formed a cap coordinating coalition first before agreeing with the others to coordinate further. Compare also the difference in the permit seller C's favourite moves in states C,E,F,I,J,U and [CFJU],E,I (see main text for a detailed discussion).

State	**Static payoffs**						**Discounted average long-term payoffs**					
	C	E	F	I	J	U	C	E	F	I	J	U
C,E,F,I,J,U	94	394	115	86	304	347	360	744	289	316	579	695
moves from C,E,F,I,J,U to …												
[CEFJU],I	443	743	289	280	566	696	*492	793	314	317	603	746
[CEFIU],J	375	676	255	297	960	629	393	694	264	311	*974	647
[CEFIJ],U	326	626	231	260	478	1055	383	683	259	303	521	*1112
[CFJU],E,I	298	1020	217	219	457	551	378	*1124	260	283	529	647
C-E-F-I-J-U	180	>360	231	217	321	>338	318	638	*408	*459	531	542
[CEFIJU]	484	785	310	379	597	738	484	785	310	379	597	738
C,E,[FI],J,U	104	432	117	89	332	384	385	756	266	295	610	759
[CFIJU],E	330	1182	233	263	481	584	370	+1222	252	293	511	623
[CEJU],F,I	381	681	328	245	520	635	460	774	368	302	589	727
[CEIJU],F	421	721	378	331	550	675	460	760	398	361	579	714
C,[EFIJU]	215	565	200	214	433	519	421	772	303	369	587	725
C-E-F-I-J,U	147	>381	189	169	317	439	335	596	325	343	482	843
move from [CEFJU],I to …												
[[CEFJU]I]	498	798	317	322	607	751	498	798	317	322	607	751
moves from [CFJU],E,I to …												
[CFJU],[EI]	316	1041	237	234	509	619	377	1101	267	280	*555	*680
[[CFJU]I],E	335	1182	235	247	485	589	375	*1222	255	276	515	628
[[CFJU]E],I	343	1066	239	280	492	597	393	1115	264	*317	529	647
[[CFJU]EI]	404	1126	270	298	537	658	*404	1126	*270	298	537	658
[CFJU],E-I	301	1021	220	236	467	565	369	1085	262	284	546	668
moves from C-E-F-I-J-U to …												
CE-F-I-J-U	>122	>302	298	281	405	421	265	475	459	*498	636	687
CJ-E-F-I-U	>110	448	298	281	>269	421	246	699	*487	496	394	*701
CU-E-F-I-J	>119	448	298	281	405	>277	283	*729	427	497	*648	460
C-E-FU-I-J	231	448	>196	281	405	>268	*603	664	318	443	569	451
CI-E-F-J-U	>100	448	298	>156	405	421	307	696	408	418	579	641
CFIJU-E	216	1400	249	244	348	374	299	+1483	290	306	410	458
CEFIU-J	220	400	251	247	1335	378	303	483	292	309	+1398	461
C-EJ-F-I-U	231	>312	298	281	>285	421	582	582	375	382	522	564
CEIJU-F	250	430	1101	269	373	408	333	513	+1142	331	436	491
CEFIJ-U	234	414	258	257	361	1309	317	497	299	319	424	+1392
moves from CU-E-F-I-J to …												
(CU)EFIJ	412	741	445	501	624	570	*412	741	445	501	624	*570
(CU)E-F-I-J	>80	>410	409	388	538	>239	228	667	527	*578	703	417
(CU)F-E-I-J	>61	588	>270	388	538	>219	187	*925	355	533	754	353
(CU)J-E-F-I	>75	588	409	388	>372	>233	213	922	*550	532	509	381
CU-EF-I-J	156	>374	>261	388	538	346	292	577	390	501	*787	539
CU-E-FJ-I	156	588	252	388	334	346	258	892	408	510	539	472
(CU)FIJ-E	132	1400	305	291	414	290	215	+1483	346	353	477	373
(CU)EFJ-I	172	501	325	1059	444	330	255	584	366	+1121	507	413
move from C-EJ-F-I-U to …												
CFI-EJ-U	>111	548	>239	>192	508	785	≻205	739	314	281	669	952

Table 2: Payoffs in the process shown in Fig. 4 (high farsightedness $\delta = 0.9$, subjective bargaining power distribution, irreversible agreements), for selected states.
* Favourite move of this player in this conflicting state.
> Move is not statically profitable for this initiating player.
≻ Move is not mixed-profitable for this initiating player (long-term payoff of target vs. static payoff of origin).
+ This move would be preferred by this player but is dominated by the move above (with a non-indented code), i.e., all initiators prefer the other move and can initiate it.

2 Static cost-benefit analysis

2.1 Game structure

Players and notation for carbon market structure. We assume a set P consisting of $N > 0$ *players* representing disjoint countries or world regions. In our notation, each specific player is represented by an uppercase letter, and we consider as players the $N = 6$ currently largest GHG emitters, C(hina), E(urope), F(ormer Soviet Union), I(ndia), J(apan), and U(SA). General players are denoted $i, j, \dots$.

In each period, each player i (i.e., a central authority in the respective country or world region, e.g., the government) first chooses their *domestic emissions cap* c_i individually, issuing that many permits to their domestic industry or population. These can then be traded freely in a domestic or international emissions market such as the EU ETS. In the terminology of Flachsland et al. 2009a, this means that we consider a "bottom-up" cap-and-trade architecture in which companies or households are trading permits in a sufficiently "integrated" international market at a market-wide equalized price, while governments only issue permits but do not trade them directly, instead of a "top-down" architecture in which governments trade permits directly (as in the Kyoto protocol). For simplicity, if several carbon markets have been linked, we treat them as one large market and do not analyse the trade in its parts individually while they are linked.[3] Notationally, we represent the *market structure* by a code in which the markets are separated by commas and the members by dashes. E.g., the code C-U,E-F-J,I represents three markets, a domestic one consisting of player I, one international with members E, F, and J, and one international with members C and U. After trading, player i's *actual emissions* $e_i(t)$ equal its post-trade amount of permits, so that in particular

$$\sum_{i \in P} e_i = \sum_{i \in P} c_i =: E, \tag{1}$$

and she gets a payoff of $\pi_i = f(e_j : j \in P)$ depending on everyone's actual emissions via some function f to be specified later.

Notation for cap coordinating coalitions. Within each market, players might be organized in a tree-like *hierarchy of coalitions* as in Heitzig 2011. A *coalition* in our sense is a subset K of the members of a market M that agree to coordinate their cap choices in some way. Such an agreement might have been signed by individual players or by *sub-coalitions* that have already formed earlier. We assume that the cap choices are coordinated in such a way that the *surplus* (the difference between the post- and pre-agreement coalitional payoffs) is distributed among the signatories in some fixed proportions given by the *bargaining power* of the individual players (see below). For convenience, we treat individual players as one-member (singleton) "coalitions". There is no explicit cap coordination *between* the top-level coalitions in a market.

Notationally, we represent the coalition hierarchy in a market by a code in which the top-level coalitions are separated by dashes and the lower-level coalitions are identified by parentheses. E.g., the code EF-J represents a market with members E, F, and J, in which E and F have formed a coalition by agreeing to coordinate their cap choices, while J chooses its caps individually. If the coalition EF in a later period signs a further agreement with J, the code becomes (EF)J. If all three had agreed immediately without a preceding bilateral agreement, we would write EFJ instead. Note that because of the assumed proportional surplus-sharing rule, the actual cap choices of E, F, and J will in general not be the same in these two situations since the pre-agreement payoffs are those in EF-J in the first case but those in E-F-J in the second case. Hence the payoffs will depend not only on the current top-level coalitions but also on what lower-level coalitions exist, and it is thus important to distinguish the cases (EF)J and EFJ.[4]

Market linkage and notation for states and moves. Markets can be linked in two ways: Either several markets such as C-U and EF-J are linked *without* immediate coordination of caps, thus becoming a new larger market C-EF-J-U, or several markets that have already reached full internal cap coordination, such as CU and

[3] This is justified, e.g., for "two-way direct links" in the terminology of Jaffe and Stavins 2008 aka "formally linked" markets in the terminology of Flachsland et al. 2009a.

[4] Alternatively, one might enlarge the set of possible states from a finite set to a continuum by representing the state of a market as a pair consisting of a partition of the market's members into coalitions and a set of payoff allocations for these coalitions.

(EF)J, are linked *with* immediate overarching cap coordination, which is then indicated by square brackets: [(CU)((EF)J)]. Once a market is formed by the second type of agreement, i.e., with immediate cap coordination, it is assumed that it can no longer be linked with further markets by the first kind of agreement, i.e., without immediate further cap coordination. In other words, the markets [(CU)((EF)J)] and I can only be linked to form [[(CU)((EF)J)]I], while "[(CU)((EF)J)]-I" is an impossible structure. Of course, the markets CU and (EF)J could also develop into CU-(EF)J, then into (CU)((EF)J) in a second step, and then into (CU)((EF)J)-I. But although (CU)((EF)J) and [(CU)((EF)J)] will get the same joint payoff, their cap distributions will differ, again because of the surplus-sharing rule which compares the payoff in (CU)((EF)J) with that in the one market CU-(EF)J but compares the payoff in [(CU)((EF)J)] with that in the two markets CU,(EF)J instead to determine surplusses.

Combining the market structure and coalition hierarchy codes to *state codes* and indicating *moves* between states with arrows labelled by the subset of players who are required for *initiating* that move, the above fictitious example process would then be denoted

$$\begin{array}{cc} & \text{C-U,E-F-J,I} \\ \xrightarrow{\text{EF}} & \text{C-U,EF-J,I} \\ \xrightarrow{\text{EFJ}} & \text{C-U,(EF)J,I} \\ \xrightarrow{\text{CU}} & \text{CU,(EF)J,I} \\ \xrightarrow{\text{CEFJU}} & \text{[(CU)((EF)J)],I} \\ \xrightarrow{\text{CEFIJU}} & \text{[[(CU)((EF)J)]I]} \end{array}$$

where we use an alphabetical order of the subcomponents of each code in order to get a unique coding scheme. Note that the number of theoretically possible states grows faster than exponentially in the number of players. For five or six players, the model has already 2729 or 41 106 states, respectively, which is why we restrict our analysis to the chosen set of players at this time. Fortunately, our results will verify the intuition that only a very small number of these possible states will occur with positive probability. The actual process might then, e.g., look as depicted in Fig. 1 or Fig. 4, where the arrows are labelled with transition probabilities and those players that favour the move.

Individual and collective rationality, farsightedness. In order to decide which moves to consider, we assume that players apply certain principles of individual and collective rationality, trying to influence the market structure and coalition hierarchy to optimize their (average, properly discounted) long-term payoffs ℓ_i. We assume that they do so in a farsighted way, anticipating the further development of the structure. We model the level of this farsightedness via a number $\delta \in (0,1)$ used in the discounting of prospective future states' static payoffs π_i. This *farsightedness* δ can be interpreted as a combined measure of time discounting, period length, and trust in the process (see below for details).

In contrast to some other game-theoretic models of coalition formation, we do not assume that the changes to the market structure and coalition hierarchy follow a specific bargaining protocol precisely prescribing who can propose which move at what time to whom, since in the climate context negotiations are probably not following such restrictive rules. Instead, we assume that in each period, the set of initiators of any feasible move can consider its realization if they all agree to do so. If several different moves are considered in a period, however, it will depend on other factors than only rationality principles which move will actually get realized. In the model this is represented by assigning probabilities to moves on the basis of all players' preferences and on assumptions about their bargaining power.

We consider different *levels of rationality.* In the weakest case, any move might be considered that is individually *profitable* for each of its initiators, using one of several concepts of profitability to be discussed below. On the medium level of rationality, only those profitable moves might be considered which are *undominated* in the sense that its initiators cannot initiate a different move which they all prefer (this corresponds to the approach in Konishi and Ray 2003). Even stronger, we will eventually assume that an undominated move will only be considered if it is the *favourite* undominated move of at least one player, be it an initiator of the move or not, based on the assumption that no international agreement will come about without at least one country pressing for its realization.[5]

[5] One might also consider an even higher level of collective rationality in which players can find a *consensus* move which no player favours but which all players prefer to the otherwise resulting lottery of favourite moves, as in Heitzig and Simmons 2012. With the long-term profitability concept of our model, however, such consensus

The remaining uncertainty about which move will actually be realized is then expressed as a probability distribution over the thus determined set of considered moves, assuming that only one of them will be realized in each period even when there are several moves considered by disjoint sets of initiators which could in principle be realized at the same time. The latter assumption is justified by the fact that usually a move by one set of players also affects the payoffs of other players, so that when a certain move is about to be made by some of the major emitters, it seems plausible to assume that the other players will wait with their attempt of an additional move until it becomes clear whether the first move will actually be realized.

2.2 A simple analytical cost-benefit model

Assumptions. We will use an analytically derived form of the payoff function f that results from the following assumptions:

- Abatement costs are cubic functions of actual domestic abatement.
- Abatement benefits are linear functions of global abatement.
- Emissions trading equalizes the price with all marginal abatement costs.
- Before the trading, all top-level coalitions simultaneously choose their coalitional caps to maximize their respective joint payoffs, anticipating its effect on trading (i.e., on traded amounts and price), leading to a global Nash equilibrium between all top-level coalitions of all markets.
- Each coalition allocates their coalitional cap to its members so that the surplus is shared in some exogenously given fixed proportions.

The functional form and coefficients of the abatement cost and benefit functions are taken at this point from the STACO model (Finus et al. 2006 version) which calibrates its benefit estimates to the vastly used DICE model of Nordhaus 1994 and takes its cost estimates from Ellerman and Decaux 1998, because that model presents a good trade-off between tractability and qualitative real-world relevance.[6] To keep our numbers comparable to those in Finus et al. 2006, we report e_i in Gton CO_2 emissions per 100 years and π_i in bln. US$ per 100 years.

Derivation of coalitional payoffs. Given the actual emissions e_i, the STACO model expresses individual payoff in terms of *individual abatement contributions* $q_i = e_i^0 - e_i > 0$ with respect to some fixed reference (“business as usual”) emissions e_i^0 since this formulation makes it easier to compare the abatement game with other public good games. In the linearized static version of STACO that we will use here, benefits from global abatement (avoided damages from climate change) are a linear function $\sigma_i Q$ of global contributions $Q = \sum_{i\in P} q_i = E^0 - E$, and costs of abatement are a cubic function

$$g_i(q_i) = a_i q_i^3/3 + b_i q_i^2/2 \qquad (2)$$

of individual contributions, where the coefficients σ_i, a_i, b_i are given in Table 3 using calibration I from Finus et al. 2006. Together with the emissions trade balance, individual payoffs of a member i of a market M in terms of caps and emissions are then

$$\pi_i = \sigma_i(E^0 - E) - g_i(e_i^0 - e_i) + p_M(c_i - e_i), \quad (3)$$

where p_M is the market price in M.

The remaining derivation is a straightforward application of the one in Helm 2003 to the case of several markets. We assume that each emissions market M has perfect competition, so that the marginal abatement costs at the post-trade abatement levels are equal to the market price for all market members,

$$g_i'(e_i^0 - e_i) = p_M \qquad (4)$$

for all $i \in M$ (see Fig. 5 for the corresponding marginal abatement cost curves). Since the market's cap equals the market's emissions,

$$\begin{aligned} c_M &= \sum_{i\in M} c_i \\ = \ e_M &= \sum_{i\in M} e_i = \sum_{i\in M} [e_i^0 - (g_i')^{-1}(p_M)], \end{aligned} \qquad (5)$$

moves are automatically identified as the only profitable moves in an equilibrium process.

[6] For future versions of our model we plan to use newer estimates, e.g., derived from Nordhaus 2010 or from more sophisticated models such as the one in Carbone et al. 2009.

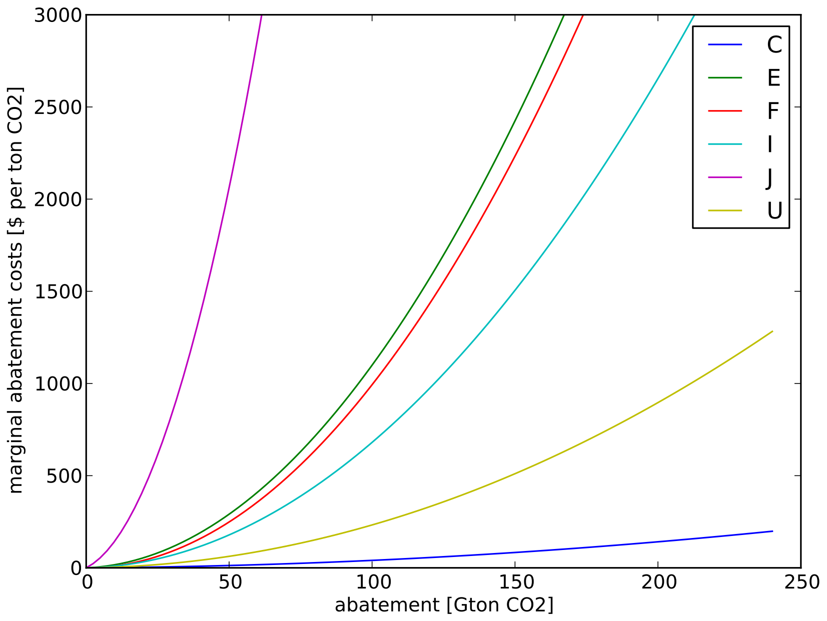


Figure 5: Marginal abatement cost curves of the individual players (quadratic functions as estimated by Ellerman and Decaux 1998). Given a market price, actual abatements are determined by intersecting the corresponding horizontal line with the curves, so that in a market including C, most actual abatement takes place there. Still, also F and I are very likely to be permit sellers in a competitive market that includes E, J, or U since because of their low vulnerabilities σ_i (see Table 3), they will issue many permits. Likewise, even if C is not in the market, U it is very unlikely to be a permit seller since because of its high vulnerability it will issue only few permits.

player i	**population**	**GDP**	σ_i	**egalitarian**	**subj.** w_i	a_i	b_i
C(hina)	1400	610	62	298	20	0.0030	0.1030
E(urope)	374	7720	236	437	20	0.1034	0.6478
F(ormer Sov. Un.)	321	777	67	82	10	0.0991	0.0181
I(ndia)	1196	365	50	197	15	0.0647	0.3392
J(apan)	127	3298	173	224	15	0.6681	7.8270
U(SA)	300	6738	226	267	20	0.0216	0.1715

Table 3: Relative distributions of bargaining power considered (arbitrary units), and coefficients for cubic abatement cost curves derived from Ellerman and Decaux 1998. See main text for definitions.

the price p_M can be seen as a function of c_M whose derivative is related to individual emissions via the theorem on implicit functions as

$$\frac{d}{dc_M} p_M = -1 \Big/ \sum_{i \in M} \frac{1}{g_i''(e_i^0 - e_i)} < 0. \qquad (6)$$

Now we assume that each top-level coalition K in M acts as an output cartel that chooses its cap $c_K = \sum_{i \in K} c_i$ to maximize its joint payoffs,

$$\begin{aligned} \pi_K &= \sigma_K Q - \sum_{i \in K} g_i(e_i^0 - e_i) + p_M(c_K - e_K) \qquad (7) \\ &= \sigma_K Q - \sum_{i \in K} [g_i(e_i^0 - e_i) + p_M e_i] + p_M c_K, \end{aligned}$$

taking the caps $c_{K'}$ of all other top-level coalitions $K' \neq K$ as given, where σ_K, e_K are the coalitional aggregates of σ_i, e_i. The corresponding first-order condition is

$$\begin{aligned} 0 &= \frac{d}{dc_K} \pi_K \\ &= \sigma_K \frac{d}{dc_K} Q + \sum_{i \in K} [g_i'(e_i^0 - e_i) - p_M] \frac{d}{dc_K} e_i \\ &\quad + p_M + (c_K - e_K) \frac{d}{dc_K} p_M \\ &= p_M - \sigma_K + (c_K - e_K) \frac{d}{dc_M} p_M \qquad (8) \end{aligned}$$

by Eq. 4, where the last term reflects the fact that the coalition is not a “price-taker” but is aware of its choice’s effect on the price. If there are n_M top-level coalitions in M, their simultaneous optimization leads to a unique Nash equilibrium which can easily be found analytically by summing the above condition over all n_M coalitions, giving

$$\begin{aligned} 0 &= n_M p_M - \sigma_M + (c_M - e_M) \frac{d}{dc_M} p_M \\ &= n_M p_M - \sigma_M \qquad (9) \end{aligned}$$

by Eq. 5. Hence the market price is simply

$$p_M = \sigma_M / n_M, \qquad (10)$$

actual individual emissions are

$$\begin{aligned} e_i &= e_i^0 - (g_i')^{-1}(p_M) \\ &= e_i^0 - \frac{\sqrt{b_i^2 + 4a_i p_M} - b_i}{2a_i} \qquad (11) \end{aligned}$$

by Eq. 4, the coalition’s cap choice is

$$c_K = e_K + (p_M - \sigma_K) \sum_{i \in M} \frac{1}{2a_i(e_i^0 - e_i)}, \qquad (12)$$

by Eqs. 6, 8, and 11, and all coalitions’ payoffs are given by Eq. 7.

From this general payoff structure, Helm 2003 derives several effects of establishing a global carbon market without cap coordination that translate into our setting as follows:

- A coalition K in a market M is a permit seller iff $\sigma_K < p_M$ (follows from Eq. 8).
- When markets are linked without coordinating caps further than before, permit sellers might increase their caps and global emissions might actually increase instead of decrease.
- Independently of whether such a linkage decreases or increases the market’s cap, it might or might not be profitable for all members.

At first glance, all this might indicate that the immediate coordination of caps when linking markets is the preferable option since it surely gives a positive surplus that can be distributed via cap redistribution to make sure that all members profit from it. Such myopic reasoning however neglects the possibility that also after a linkage *without* cap coordination, caps might later on be coordinated, and some coalitions might prefer such a two-step process since its first step puts them in a more comfortable bargaining situation for the second step. It is precisely such effects and the resulting conflicts that our dynamic model will uncover, and our results show that indeed such complications will increasingly occur when players get more farsighted (see, e.g., Fig. 4).

Surplus-sharing and bargaining power. Finally, each top-level coalition K determines their surplus payoff $\Delta\pi_K = \pi_K - \pi_K^0$ by comparing their joint payoff π_K with the joint payoff $\pi_K^0 = \sum_{i \in K} \pi_i^0$ their members i would get in the following reference state: remove coalition K from the coalition hierarchy, and if K is of the immediate-coordination form [...], also split the corresponding market into one market for each of the resulting top-level coalitions. E.g., for K = (EJ)U in state C-(EJ)U,FI the reference state is C-EJ-U,FI, while for K = [C(EJ)U] in state [C(EJ)U],FI the reference state is C,EJ,FI,U. Then coalition K allocates their joint cap c_K in such a way that each player $i \in K$ gets a share of this surplus that is proportional to their bargaining power w_i,[7] so that

$$\pi_i = \pi_i^0 + \Delta\pi_K w_i / \sum_{j\in K} w_j. \qquad (13)$$

We consider the following five distributions of bargaining power, taking the respective numbers from a reference year 1995 (see Table 3):

- $w_i =$ population of i.
- $w_i =$ GDP of i in US$.
- $w_i = \sigma_i$ (climate "vulnerability").
- $w_i = 1$ (equal bargaining power).
- An "egalitarian" approach that leads to equal per-capita surplus in purchasing power parities (PPP):

  $$\begin{aligned} w_i = &\text{(population of } i)\\ &\times \text{(PPP in currency of } i)\\ &\times \text{(exchange rate from } i \text{ to US\$)}. \end{aligned}$$

- A subjectively chosen distribution that represents a simple compromise between the other four.

3 Results from the dynamic model

The observations that can be made from our model results for various parameter settings can be summarized as follows:

- If it is always possible to *immediately* include a binding agreement on caps into an agreement to link existing carbon markets, then it seems likely that a global carbon market with a first-best cap will emerge eventually, but probably not in one move (e.g., Fig. 1).
- Such a process is likely to involve high *uncertainty* about which moves will happen in the beginning, and will get more deterministic towards the end. Somewhat counterintuitively, there might remain less uncertainty about later moves when agreements are reversible, i.e., can not only be signed but also be terminated unanimously or even unilaterally (e.g., compare Figs. 4 and 3).
- When agreements are *reversible,* it seems still very unlikely that in the emerging process agreements will actually be terminated. Rather, agreements that would likely be terminated later on would not be signed in the first place.
- When agreements are *irreversible,* the process might contain branches in which a large market is established at first with non-cooperatively chosen caps which then eventually get fully coordinated in several further moves (e.g., the C-E-F-I-J-U branch in Fig. 4). This seems to get more likely the more farsighted the players are (e.g., compare Figs. 4 and 1).
- Independently of the history of the process, once a set of markets has been linked, it is very likely that they *stay linked* and get fully coordinated caps eventually. In rare cases it might however be profitable for a market member to unilaterally terminate coordination or linkage agreements, hoping to get a larger share of the eventual surplus from global cap coordination.
- If it is *not* possible to immediately agree on caps at the time of market linkage but only later via separate agreements, it is not clear that a global carbon market with a first-best cap will emerge eventually. If players are not sufficiently farsighted, the process might get stuck in a state without global cap coordination, with several disjoint carbon markets in each of which there is full internal cap coordination, but which are unlikely to be linked further (e.g., Figs. 7 and 8).
- Even when joining a market eventually, prospective permit *buyers* tend to have an incentive to free-ride on the abatements of that market by joining it rather later than earlier. Prospective permit *sellers,* on the other hand, tend to profit even more from

[7] A possible (though somewhat trivial) interpretation of this surplus-sharing rule that relates it to traditional solution concepts of cooperative game theory is that each player gets its weighted Shapley value in the unanimity game v with $v(K') = \Delta\pi_K$ if $K' \supseteq K$ and $v(K') = 0$ otherwise, using the weights w_i (compare Kalai and Samet 1987 who also discuss using population as weight). The underlying rationale is that the reference state is the only alternative state that could realistically be reached on short notice, by terminating only one top-level agreement, so that the value of each player's outside option is simply its payoffs in that reference state.

being included early into the market (e.g., compare favourite moves and payoffs of C and U in Fig. 1 and Table 1). Depending on the amount of farsightedness and the other members of the market, a permit seller might or might not prefer if its main competitor joins the market only later (e.g., compare the long-term payoffs for C in state [CEFJU],I in Tables 1 and 2, and C's favourite moves in states C,E,F,I,J,U and [CFJU],E,I of Fig. 4).

- In a *not yet fully coordinated* market, both permit buyers and sellers usually have an incentive to free-ride on the abatements of others by entering a cap-coordinating coalition rather later than earlier (e.g., in the C-E-F-I-J-U branch in Fig. 4).

In the following detailed explanation of the occurring effects, we will first assume that all agreements are irreversible before discussing reversibility in Sect. 3.4, and will assume the subjective bargaining power distribution before discussing bargaining power in Sect. 3.5.

3.1 Influence of the level of rationality

Let us first assume myopic players, i.e., with very small $\delta \ll 1$, so that a "profitable" move is basically one to a state with higher static payoff than the going state (let's call this a *statically profitable* move).

When all undominated moves are probable. Assume that the set of moves that might be realized in each state with positive probability is only insofar restricted by individual and collective rationality as all moves which are *unprofitable* for at least one of its initiators have zero probability, and all others have equal probability. Then there would be vast uncertainty about the process throughout its evolution, involving thousands of states reached with positive probability and already 72 possible moves from the "root" state C,E,F,I,J,U to linked markets of all sizes, with and without immediate cap coordination. Still, a fully coordinated global market would evolve eventually already at this low level of rationality because of the option of combining market linkage with immediate cap coordination, but this "first-best" state would be reached on a highly uncertain path with correspondingly high uncertainty about the speed and final payoff distribution.

Most of these moves are however dominated in the sense that their initiators could initiate a different move that they all prefer. If we assume such *dominated* moves have zero probability as well, there remain 58 moves from C,E,F,I,J,U since each uncoordinated linkage move is dominated by the corresponding immediate-coordination linkage move. The only exception is the move to C,E-I,F,J,U which I slightly prefers to C,[EI],F,J,U because it can profit more from selling permits to E in the uncoordinated market E-I than from getting its share of $15/(15+20)$ of the surplus of the coordinated market [EI] over the non-market state E,I. More generally, the less bargaining power a permit seller has, the more likely it prefers the uncoordinated to the immediately coordinated market. The whole process would still lead to a first-best solution with high uncertainty about speed and payoff distribution.

When only favourite moves are probable. For several reasons, it seems plausible that individual rationality restricts the set of moves further than described above. First, a necessary initiator of a move might block that move by not agreeing to it if it favours a different undominated move. Second, it might seem realistic that a move is only realized if at least one player presses for its realization because it is the move which that player prefers over all other undominated moves.

Consider, e.g., the situation in the root state C,E,F,I,J,U in the above process. There, players C and E have an undominated move to [CE],F,I,J,U, but they know that also players CEFIJU have an undominated move to the global market [CEFIJU] and that CFIJU have an undominated move to the E-excluding market [CFIJU],E, among other undominated moves. Now C prefers [CEFIJU] to all other target states of undominated moves (mainly because C is the cheapest and largest seller of permits) and E prefers to be a free-rider in [CFIJU],E to all other target states of undominated moves. This lets it seem plausible to assume that C will not agree to form [CE] or to other moves involving C but will try to press for [CEFIJU]. Player E, on the other hand, might try to block the formation of the coordinated markets [CE] or [CEFIJU] and other moves involving E, hoping that then a large market without itself will come about, using its power to urge the relevant players to make that undominated move instead of any other (with farsightedness, however, this move might no longer be attractive for E, as shown in Fig. 2).

If it is assumed that the probability of each undominated move is proportional to the aggregate bargaining power (the subjectively chosen one in this case) of those players who *favour* this move, no matter whether they belong to its initiators or not, then the set of nodes reached with positive probability reduces to only ten nodes, and the process looks as in Fig. 1, now reaching a first-best state much faster.

Although it might seem a little ad-hoc to assume that the bargaining power influences the move probability in this simple proportional way, a certain game-theoretic model of the actual bargaining process will result in exactly these probabilities if it is assumed that the bargaining power reflects the frequency with which a party might propose a motion which might then however be amended by the initiators of the proposed move (see the Appendix).

3.2 Importance of the possibility of immediate cap coordination

Let us still assume myopic players and study why it is important that players agreeing to link their carbon markets can agree to coordinate their caps in the same move. If the latter was not possible and caps could be coordinated only after market linkage in later moves, the simple process described above (Fig. 1) would change to the more complicated one shown in Fig. 6, where no branch reaches a first-best state but all get stuck with several disjoint markets.

Although (similar to what is shown in Fig. 1) C would still favour a coordinated global market CEFIJU *eventually,* the latter could only be reached in two moves via the uncoordinated global market C-E-F-I-J-U. Moving to the latter from the no-market state C,E,F,I,J,U is however (statically) unprofitable for both E and U, and even if it were profitable, C would still (myopically) prefer a move to C-E-J-U,F,I instead of C-E-F-I-J-U, as shown in Fig. 6, since that excludes its competitor permit sellers F and I. Similarly, I now favours C-E-I-J-U,F.

Likewise similar to what is shown in Fig. 1, U would still favour a coordinated global market (CEFIJ)U in which CEFIJ have formed a cap-coordinating coalition first, improving U's bargaining position for the final formation of the grand coalition. But this could now only be reached via C-E-F-I-J,U and CEFIJ,U, and the move from C,E,F,I,J,U to C-E-F-I-J,U is not (statically) profitable for E, and even if if were profitable it would have zero probability since each of its initiators prefers a different undominated move (see Fig. 6). The largest market without itself that U can hope to form is C-E-J, so U presses for this to happen.

For F, the incentives are even more different from the situation in Fig. 1 where it wanted to stay out of the market. Although it is still possible to reach CEIJU,F via C-E-I-J-U,F, the immediate gains from moving to C-E-F-J-U,I are larger since that market excludes I, making F the second permit seller after C.

More generally, a recurring motif in most of our results is that if a market has formed without cap coordination, the next move will very likely be that either all or all but one of its members form a coalition which will afterwards agree with the remaining market member to form a 2nd-level coalition.

If joining the market is no longer an option, outsiders usually prefer a faster over a slower coalition formation inside the market, since that will decrease that market's emissions faster. E.g., in state C-E-J-U,F,I, one can expect that both F and I try to convince CEJU to form the coalition CEJU instead of CEJ or EJU. If there are several outsiders, as in state C-J-U,E,F,I, they might profit from building a second market (here E-I), but often prefer to wait with this move until the existing market has reached full cooperation (here CJU). This might also be the preferable order for the existing market's members, as in this example, but it might also be the case that a member of the existing market prefers if the second market forms and coordinates before the first market coordinates its caps (not shown here).

After a large but non-global market has formed and coordinated in such a way, as in CEFJU,I, it might still be the case that all players would prefer to this state the outcome of joining and then coordinating with the remaining player, i.e., the state (CEFJU)I. But it is quite likely that the intermediate step CEFJU-I is not (statically) profitable for all players, so that the process gets stuck. In this example, the move from CEFJU,I to CEFJU-I is (statically) profitable only for player

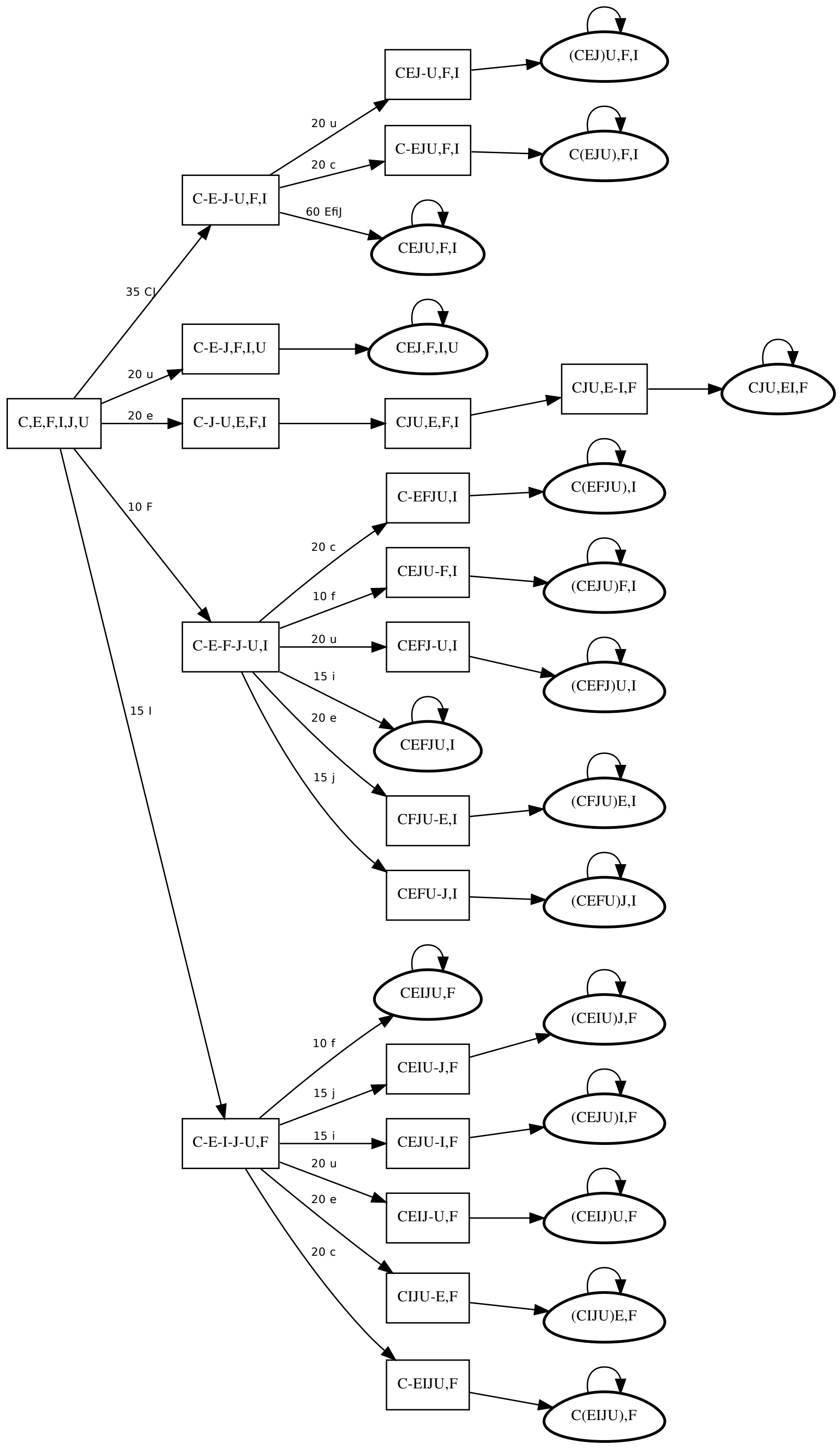


Figure 6: Alternative scenario to Fig. 1 in a world where caps cannot immediately be coordinated when markets are linked but only later in separate moves (myopic players with $\delta = 0.00001$ and irreversible agreements; see Figs. 7 and 8 for farsighted players or reversible agreements). No first-best state can be reached and the process gets stuck in a state with one or more non-global markets (egg-shaped nodes). See main text in Sect. 3.2 for a detailed discussion.

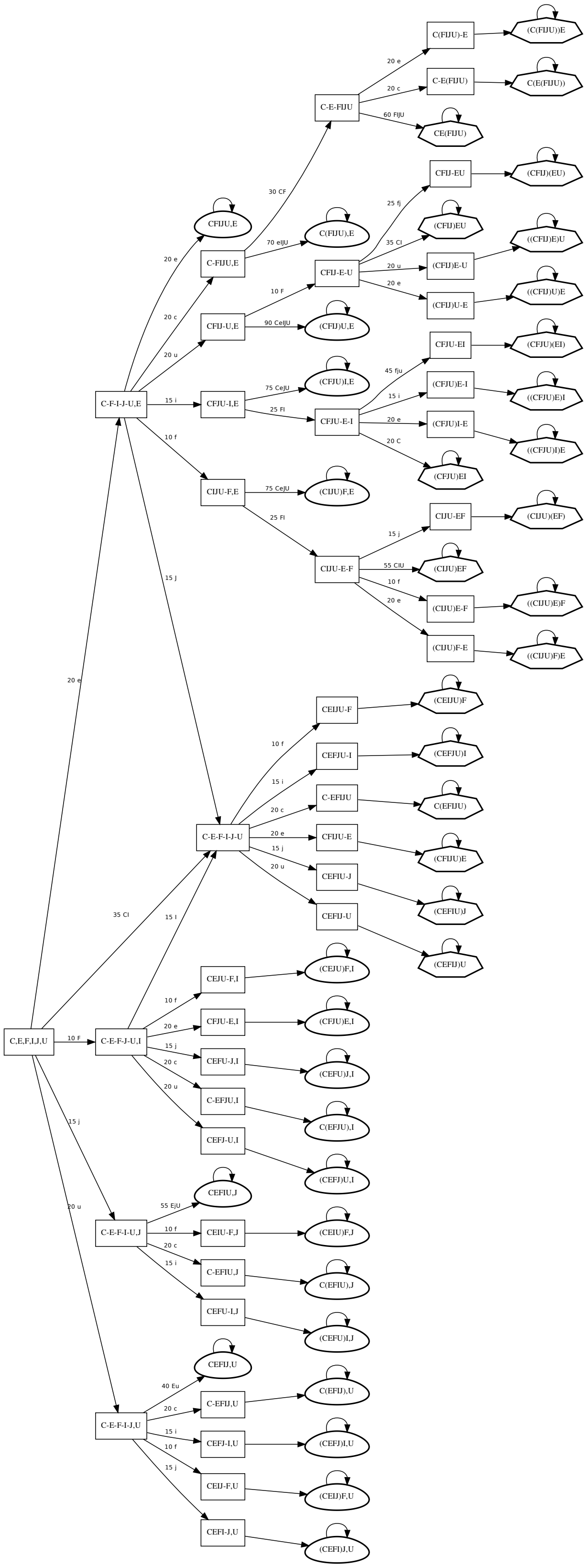


Figure 7: Same conditions as in Fig. 6 but with medium farsightedness ($\delta = 0.5$), showing quite similar effects but reaching first-best states (diamond-shaped) at least with some probability. Note that C-E-F-I-J-U can be reached via three paths here.

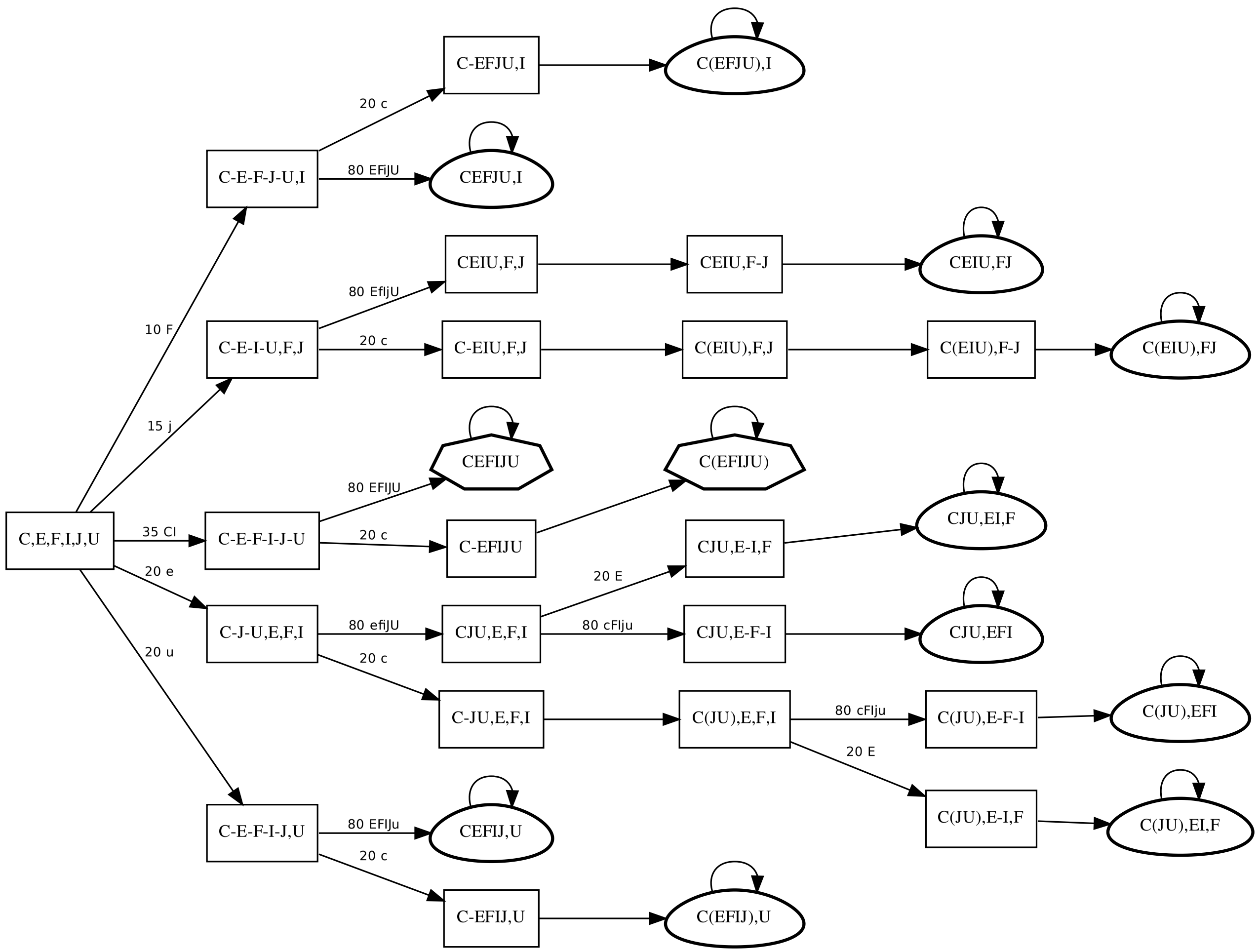


Figure 8: Same conditions as in Figs. 6 and 7 but with $\delta = 0.5$ and unilaterally terminable agreements, still showing similar effects and reaching first-best states only with some probability.

I, so that CEFJU would not agree to it.

Although it might seem that many of the effects described above might be due to a lack of farsightedness, similar processes can be observed with higher values of δ, as shown for $\delta = 0.5$ in Figs. 7 (for the assumption that agreements are irreversible) and 8 (for reversible agreements), where a first-best state is reached eventually with only 42% or 35% probability, respectively.

All this shows that the possibility of market linkage with *immediate* cap coordination is essential for an efficient process towards a first-best state.

3.3 Influence of the level of farsightedness

In the examples discussed above, players were assumed to be *myopic* in the sense that they base their agreements more or less only on a comparison of their payoff in the immediate target state of a possible move with that in the going state. Real-world governments, however, must be considered to be at least to some extent more farsighted and to also consider the possible further development of the state after a certain move. In our model, we assume that they take into account all possibly resulting states and their payoffs, discounting them with a factor $\delta \in (0,1)$ per period according to the number of moves it takes to get to them, and weighing them with the respective probabilities of reaching that state.

Although we treat the measure of *farsightesness* δ as a primary parameter that is not analysed further here, it will depend on the governments' normal discount rate, on their trust in the process not breaking down, and on the period length, so it could be modelled as $\delta = \delta_0^{\Delta t}\tau$, where

- δ_0 is the normal yearly discounting factor of the governments,
- Δt is the period length in years,
- and τ is the subjectively believed probability that the process will not break down from a period to the next.

Let us still assume that agreements are irreversible (before discussing reversibility in the next section), so that the process cannot contain cycles. Then the probability of all feasible moves can be determined via backward induction, starting from the states in which a fully coordinated global market is reached, and using the already determined probability of later moves to determine the profitability, dominatedness and probabilities of earlier moves. A simple way to do this is to assume that the players assess the profitability of a move in the following farsighted way: A move $x \xrightarrow{S} y$ is *mixed-profitable* iff for each player i in the set of initiators S, the static payoff $\pi_i(x)$ which i would get if the going state were to remain forever is less than the discounted average long-term payoff $\ell_i(y)$ which i would get in the already determined part of the process after the considered move, which can be calculated recursively as

$$\ell_i(y) = (1-\delta)\pi_i(y) + \delta \sum_z p_{y\to z}\ell_i(z), \quad (14)$$

where $p_{y\to z}$ is the probability that the state following y is z.

As δ grows from 0.00001 in Fig. 1, the final process diagram doesn't change until well beyond $\delta = 0.7$. At $\delta = 0.9$ it reaches the one depicted in Fig. 4, containing 269 possible states. There, J and U still try to press for a market that excludes them at first, C, E, F, and I now change their behaviour for certain reasons.

E would still prefer a move to [CFIJU],E, as can be seen in Table 2 (marked by a '+' sign), but although that move is (mixed-) profitable, it is *dominated* by the move to C,E,[FI],J,U in the sense that for each player i in the initiating set $S =$ CFIJU, the long-term payoff $\ell_i(y)$ is larger at $y =$ C,E,[FI],J,U than at $y =$ [CFIJU],E. The best E can hope for in view of which moves are undominated is that the market [[CFJU]I] is formed in two steps, but at the intermediate state [CFJU],E,I, all other players favour different continuations.

C now no longer prefers to move to [CEFIJU] directly but is farsighted enough to prefer the two-step path to [[CEFJU]I] via [CEFJU],I since the latter intermediate step improves C's bargaining position vs. I (see Table 2).

F and I can also hope to get a larger share from the final surplus, but in contrast to C they are patient enough to prefer a global but initially uncoordinated market in which others then start to coordinate their caps before eventually including F and I into the emerging hierarchical coalition, despite the highly uncertain outcome of this widely branching process (Fig. 4 depicts only an initial part of it).

At $\delta = 0.99999$, corresponding to a very high farsightedness, e.g., if there is almost no normal discounting, periods are very short, and trust in the process is very high, the process finally contains hundreds of possible states, starting with a move to either [CEFJU],I (preferred by C), C-F-J-U,E,I (preferred by E), C-E-F-J-U,I (preferred by F), C-E-I-J-U,F (preferred by I), C-E-F-I-U,J (preferred by J), or C,[EF],I,J,U (preferred by U). The paths following the highest probability from C,[EF],I,J,U on are depicted in Fig. 9.

We have thus seen that the level of farsightedness can have a large influence on the process and that, somewhat counter-intuitively, more farsightedness can lead to more uncertainty. The latter might however be in part due to our specific assumption of what a farsighted player considers to be a profitable move, and we will see next that the processes become much simpler again when agreements are assumed to be reversible and a more consistent concept of long-term profitability is used.

3.4 Influence of the level of reversibility of agreements

When market and coalition formation can be reverted unanimously. In contrast to the preceding discussion, let us now assume that for each market linkage or cap coordination move, there is also a reverse move which splits a market or removes a top-level coalition, which can be initiated by the very same set of players, i.e., all members of the market that gets split or of the top-level coalition that gets removed.

It might seem at first glance that such "unanimous" reversals should never occur if the original move was profitable, so that their possibility should not change the result. Our current definition of (mixed-) profitable, however, only required that the discounted average long-term payoff after the move be larger than the *static* payoff in the present state. Depending on what other moves are probable in the present state, it might happen that the discounted average long-term payoff in the present state exceeds both the static payoff in the present state and the discounted average long-term payoff after a certain move. In that case, the initiators of that move might realize after the move that they could have expected higher payoffs before the move, and in that case we say that the reverse move is *long-term profitable*. Formally, $x \xrightarrow{S} y$ is long-term profitable iff $\ell_i(x) < \ell_i(y)$ for all $i \in S$.

Note that with this concept of profitability, it would no longer be possible to calculate the move probabilities at a state x via backward induction since $\ell_i(x)$ already depends on these probabilities. But with reversible moves, this is not possible anyway since the graph of feasible moves now contains cycles. Still, an assignment $p_{x\to y}$ of transition probabilities for all possible pairs of states can at least be considered to represent a plausible process if we get the same probabilities back when we use them to calculate the long-term payoffs $\ell_i(x)$ for all i and x, then use those to find the favourite undominated long-term profitable moves, and finally use the latter to derive a new assignment of probabilities $p'_{x\to y}$ for all x, y. If this cyclic plausibility check reproduces the probabilities it started with, we call the assignment $p_\cdot$ an *equilibrium process* here. This novel equilibrium concept is a refinement of the EPCF concept suggested in Konishi and Ray 2003 where we added the restriction that move probabilities are determined from bargaining power (see the Appendix for formal definitions).

The following iterative procedure might be considered an intuitive approach at finding such equilibrium processes: Determine an initial estimation of p as in the irreversible case, using mixed profitability and backward induction. Then enable reverse moves, switch to the long-term profitability concept, and iterate the above plausibility check, always using the resulting new assignment p' as the next estimate of p, until the algorithms stops with $p' = p$ (see the Appendix for a formal definition of this algorithm). The final assignment of transition probabilities is then an equilibrium process which can be considered as representing a *consistent set of common beliefs* of what transitions will happen in what states, where "consistent" means that together with individual and collective rationality and with the given distribution of bargaining power, these beliefs about the transition probabilities lead to no contradictions.[8]

When we compare the equilibrium processes derived in this way for the case of reversible agreements with the processes derived by backward induction for the case of irreversible agree-

[8] Although in our examples the algorithm almost always converged to an equilibrium process, the latter is not guaranteed to be unique and there might be parameters for which the algorithm does not converge whether or not an equilibrium process exists in the first place. We plan to analyse the existence of equilibrium processes and the formal properties of the algorithm in a separate, more theoretical paper.

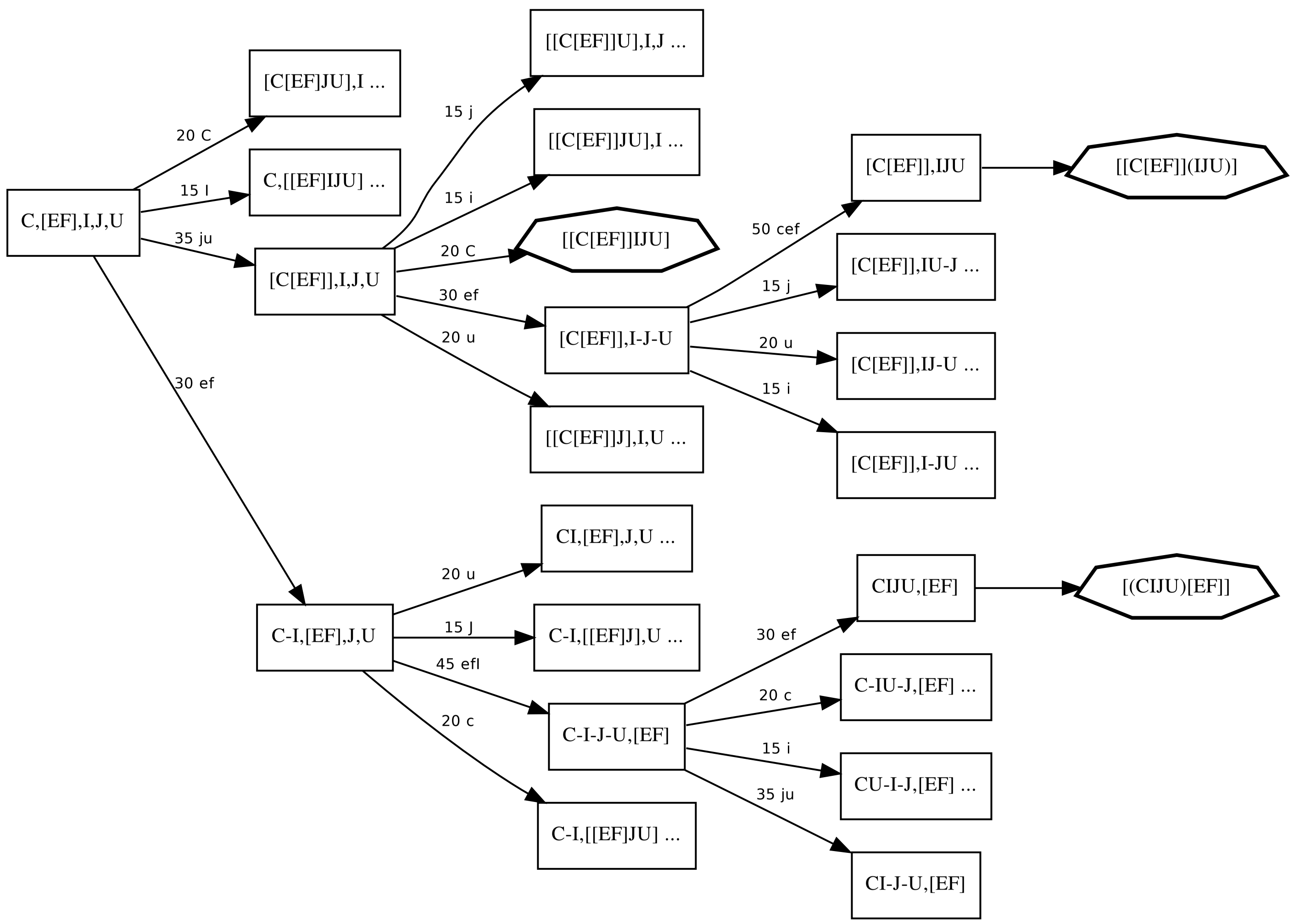


Figure 9: Detail of a complicated alternative process to Fig. 1 for the case of very high farsightedness ($\delta = 0.99999$) and irreversible agreements (see Fig. 3 for a reversible case with $\delta = 0.9$). Branches following less likely moves have been truncated.

ments, we see that unanimous reversibility of moves can result in a considerable simplification of the process, leading to a faster emergence of a first-best state and less uncertainty, involving fewer moves with positive probability although the set of feasible moves has been doubled by the inclusion of "reverse" moves. For $\delta \in \{0.1, 0.5, 0.7\}$, the result is the simple process in Fig. 1.

The branch of an equilibrium process that starts in the no-linkage root state C,E,F,I,J,U (which is the one we have focussed on until here) typically does not involve the termination of any established agreements with positive probability. When starting in a different state, however, it might easily be preferable to terminate an agreement in the hope of getting a better one eventually. E.g., when $\delta = 0.5$ and starting in state C,E,F,I,J-U, both members of the uncoordinated market J-U can profit from splitting that market since the further process from the resulting state C,E,F,I,J,U promises higher long-term payoffs than the further process from the going state C,E,F,I,J-U, and for at least one of these members this market splitting is the favourite move. The same is true in those states in which the only existing market is C-E-I, C-F-I, C-F-J, C-F-I-J, C-F-I-U, C-I, C-I-J, or E-I-J.

With unanimously terminable agreements, the overall behaviour with respect to the level of farsightedness is now that a larger δ tends to lead to simpler processes, e.g., for $\delta = 0.9$ the process in Fig. 1 simplifies to the one in Fig. 3.

When agreements can be terminated unilaterally. In most international agreements, its signatories might not only unanimously agree to terminate it, but there is usually also an "exit" option by which individual signatories can leave. After such a change in the agreement's membership, the agreement usually needs some time of adjustment and/or the incentives for membership of the remaining members often change as well. We therefore assume here that once a player exits an agreement, the agreement is automatically terminated and the remaining members can sign a new agreement in the next period.[9]

When any agreement can be terminated unilaterally by any signatory in this way, the set of possible moves becomes much larger and there is no longer a unique initiating set for most of the possible transitions between states. Still, the resulting processes are only rarely different from those resulting when only unanimous termination of agreements is possible. For $\delta \in \{0.5, 0.9\}$, e.g., the processes from the starting state C,E,F,I,J,U on are still those of Figs. 1 and 3, where Table 1 shows payoffs for the case of $\delta = 0.5$, highlighting several typical cases of relevant payoff comparisons.

From a different starting state, however, the process might be different when unilateral termination is possible. Fig. 10 on page 28 shows the result for the starting state CU,EI,FJ in which there are three coordinated markets, each consisting of a low-vulnerability country as a permit seller and a high-vulnerability country as a permit buyer. While in the unanimous case, J favours the move to [(CU)(EI)],FJ, in the unilateral case it now favours to terminate first the cap-coordination agreement with F and then the market linking agreement with F, since that moves J into a more comfortable bargaining position for the final linking of all markets. Due to this changed behaviour by J and the resulting changes in the long-term payoffs at CU,EI,FJ, the set of profitable moves at CU,EI,FJ changes so that the moves to [(CU)(FJ)],EI and CU,[(EI)(FJ)] now become profitable and are actually favoured by E and U above the move to [(CU)(EI)(FJ)]. Likewise, the move to [(CU)(EI)],FJ is now no longer profitable for J, so F has to settle for the next-best move to [(CU)(EI)(FJ)]. Examples of other states in which a player wants to unilaterally terminate an agreement are C-E-F-I-J,U where E wants to exit the market and continue as in Fig. 1, and [((CF)EI)J],U where C and E want to exit the market.

3.5 Influence of the distribution of bargaining power

Finally, let us study shortly how the processes derived by backward induction or by searching for an equilibrium process are affected by the choice of assumed bargaining power distribution, concentrating on the model version with unilateral terminability of agreements for this.

Comparing the processes in Figs. 11 and 12 with the one in Fig. 1, we see that with my-

[9] This is a deliberate contrast to what is assumed in the concept of "internal stability" that is used in many other studies, in particular in the analysis of the single-coalition open-membership game.

opic or moderately farsighted players, a low-vulnerability player such as C tends to press either for a large market including itself and excluding its major competitor (if its power is large), or for a global market (if it has medium power), or for a larger market *ex*cluding itself (if it has small power, since its free-rider payoffs then exceed the small share of the global market's surplus it would get). A high-vulnerability player such as E apparently always wants to remain a free-rider.

Under high farsightedness, shown in Fig. 13, the picture gets more complex when the power distribution is highly unequal, since many of the moves establishing a large market are then no longer long-term profitable and the linkage proceeds in much smaller steps. For some highly unequal power distributions, our algorithm did not even converge on an equilibrium process in the case of high farsightedness, and for others we conjecture that there are likely several other equilibrium processes in addition to the ones we show here.

4 Conclusion

In this paper, we presented several scenarios of how the dynamics of the suggested bottom-up process of climate coalition formation through carbon market linkage might look like with farsighted players that consider only undominated long-term profitable moves and use their bargaining power to press for the realization of their favourite such move. When compared to the often quite pessimistic picture painted by much of the existing literature on coalition formation in the climate context, our results seem to justify more hope that a first-best state with a global cap-and-trade system and a global agreement on the caps will evolve eventually, at least if the major emitters manage to include immediate agreements on caps into agreements to link carbon markets.

Still, the transition probabilities and resulting long-term payoffs in these scenarios are not meant to be quantitatively accurate representations of what will happen in reality, since they are based on a quite simple static cost-benefit model, although the latter was calibrated to a more sophisticated integrated assessment model. Also, we deliberately introduced and varied two subjectively chosen parameters, the measure of farsightedness δ and the bargaining power distribution, in order to explore a range of possible geo-political environments.

Future simulations with our model should thus use more accurate payoff structures, also involving path-dependent payoffs instead of static ones, welfare effects of leakage and the feedback between trade in emissions and other goods, and policy instruments such as tariffs vis-a-vis free-riders. Since such payoff estimates can probably only be found by running a numerical integrated assessment model for each of the numerous possible states of the process, it will be very important to find a trade-off between accuracy and computation time, hence the MICA model used in Lessmann et al. 2009 might be a good choice. We also plan to explore model refinements such as different farsightedness of individual players, the possibility of reversible cap-coordination agreements in irreversibly linked markets, and the inclusion of other relevant forms of agreements, e.g., on R&D.

Acknowledgements. The author thanks Ulrike Kornek, Kai Lessmann, Robert Marschinski, Ottmar Edenhofer, and the Policy Instruments Group at the Potsdam Institute for Climate Impact Research for many intense discussions; Bjart Holtsmark, Mads Greaker, Cathrine Hagem, Robert Schmidt, and the CREW project for inspiring work; and Rajiv Vohra, Peter Menck, Norbert Marwan, Jonathan Donges, and Carl-Friedrich Schleussner for helpful comments. This work was supported by the Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability (PROGRESS).

References

Aldy, Joseph E and Robert Norman Stavins, eds, *Post-Kyoto international climate policy: implementing architectures for agreement*, Cambridge University Press, 2009.

Barrett, Scott, "Self-enforcing international environmental agreements," *Oxford Economic Papers*, 1994.

Buchner, Barbara, Carlo Carraro, Igor Cersosimo, and Carmen Marchiori, "Back to Kyoto? US participation and the linkage between R&D and climate cooperation," *The Coupling of Climate and Economic Dynamics*, 2005, pp. 173–204.

Carbone, Jared C, Carsten Helm, and Thomas F Rutherford, "The case for international emission trade in the absence of cooperative climate policy," *Journal of Environmental Economics and Management*, 2009, *58* (3), 266–280.

Carraro, Carlo and Domenico Siniscalco, "Strategies for the international protection of the environment," *Journal of Public Economics*, 1993, *52*, 309–328.

Chwe, Michael Suk-Young, "Farsighted coalitional stability," *Journal of Economic Theory*, 1994, *63* (2), 299–325.

Ellerman, A Denny and Annelene Decaux, "Analysis of Post-Kyoto CO2 Emissions Trading Using Marginal Abatement Curves," *MIT Joint Program on the Science and Policy of Global Change*, 1998, *Report 40*, 1–33.

Finus, Michael, "New Developments in Coalition Theory," in Laura Marsiliani, Michael Rauscher, and Cees Withagen, eds., *Environmental policy in an international perspective*, Kluwer, 2003, pp. 19–49.

__ , Ekko Van Ierland, and Rob Dellink, "Stability of Climate Coalitions in a Cartel Formation Game," *Economics of Governance*, March 2006, *7* (3), 271–291.

Flachsland, C., R. Marschinski, and O. Edenhofer, "Global trading versus linking: Architectures for international emissions trading," *Energy Policy*, 2009, *37* (5), 1637–1647.

__ , __ , and __ , "To link or not to link: benefits and disadvantages of linking cap-and-trade systems," *Climate Policy*, 2009, *9* (4), 358–372.

Heitzig, Jobst, "Efficiency in face of externalities when binding hierarchical agreements are possible," *Game Theory & Bargaining Theory eJournal*, 2011, *3* (40), 1–16.

__ and Forest W Simmons, "Some chance for consensus: voting methods for which consensus is an equilibrium," *Social Choice and Welfare*, November 2012, *38* (1), 43–57.

__ , Kai Lessmann, and Yong Zou, "Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games," *Proceedings of the National Academy of Sciences of the United States of America*, 2011, *108* (38), 15739–15744.

Helm, Carsten, "International emissions trading with endogenous allowance choices," *Journal of Public Economics*, December 2003, *87* (12), 2737–2747.

Jaffe, Judson and Robert Norman Stavins, "Linkage of Tradable Permit Systems in International Climate Policy Architecture," *NBER Working Paper 14432*, 2008.

Kalai, Ehud and Dov Samet, "On weighted Shapley values," *International Journal of Game Theory*, 1987, *16* (3), 205–222.

Konishi, H and Debraj Ray, "Coalition formation as a dynamic process," *Journal of Economic Theory*, 2003, *110* (1), 1–41.

Lessmann, Kai, Robert Marschinski, and Ottmar Edenhofer, "The effects of tariffs on coalition formation in a dynamic global warming game," *Economic Modelling*, May 2009, *26* (3), 641–649.

Nagashima, Miyuki, Rob Dellink, Ekko van Ierland, and Hans-Peter Weikard, "Stability of international climate coalitions – A comparison of transfer schemes," *Ecological Economics*, March 2009, *68* (5), 1476–1487.

Nordhaus, William D, *Managing the global commons: the economics of climate change*, Cambridge, MA: MIT Press, 1994.

__ , "Economic aspects of global warming in a post-Copenhagen environment.," *Proceedings of the National Academy of Sciences of the United States of America*, June 2010, *107* (26), 11721–11726.

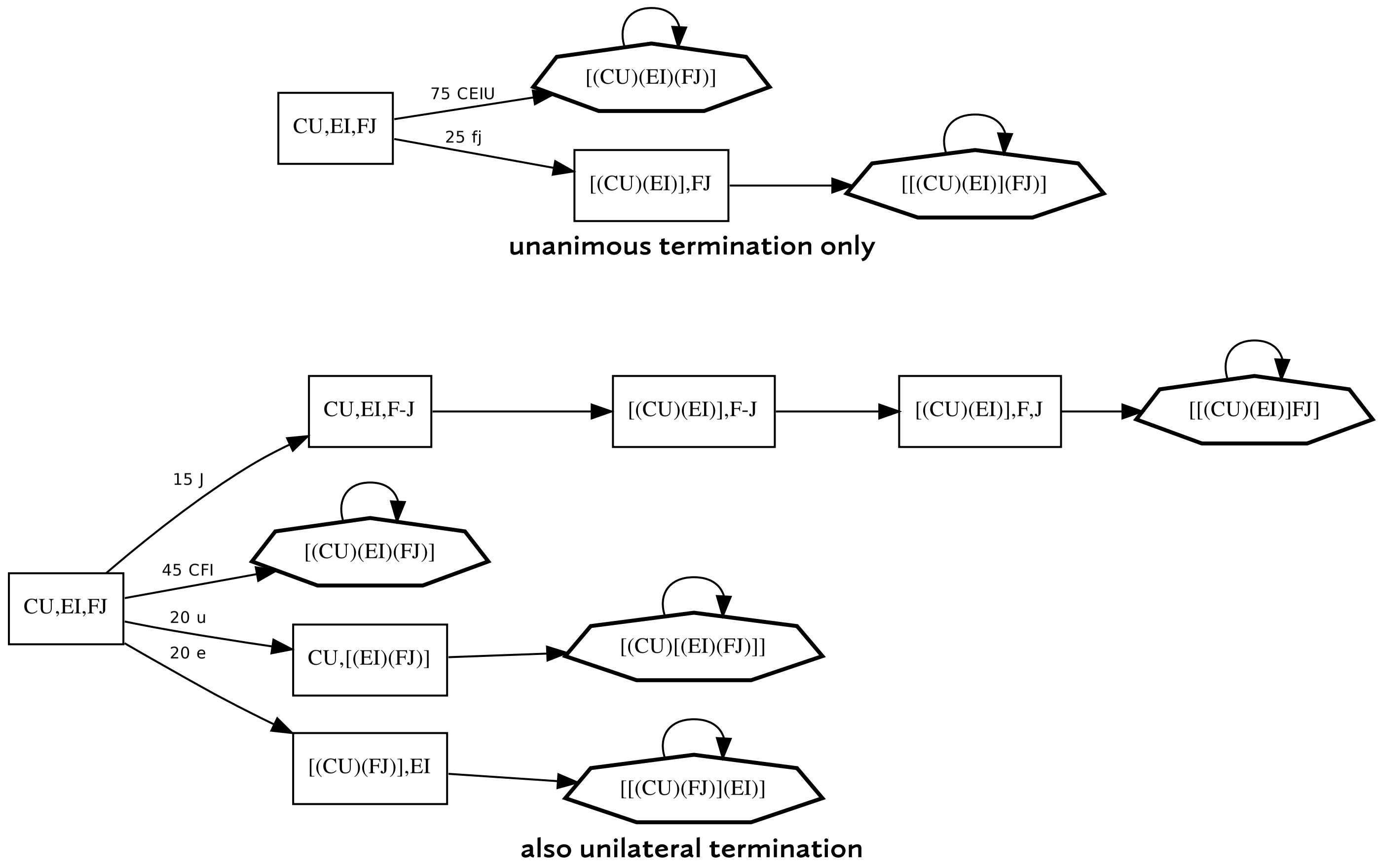
Osmani, Dritan, "A Note on Computational Aspects of Farsighted Coalitional Stability," 2011.

__ and Richard S J Tol, "On the efficiency gains of emissions trading when climate deals are non-cooperative," 2010.

Ray, Debraj and Rajiv Vohra, "Equilibrium Binding Agreements," *Journal of Economic Theory*, 1997, *73*, 30–78.

__ and __ , "A theory of endogenous coalition structures," *Games and Economic Behavior*, 1999, *26* (2), 286–336.

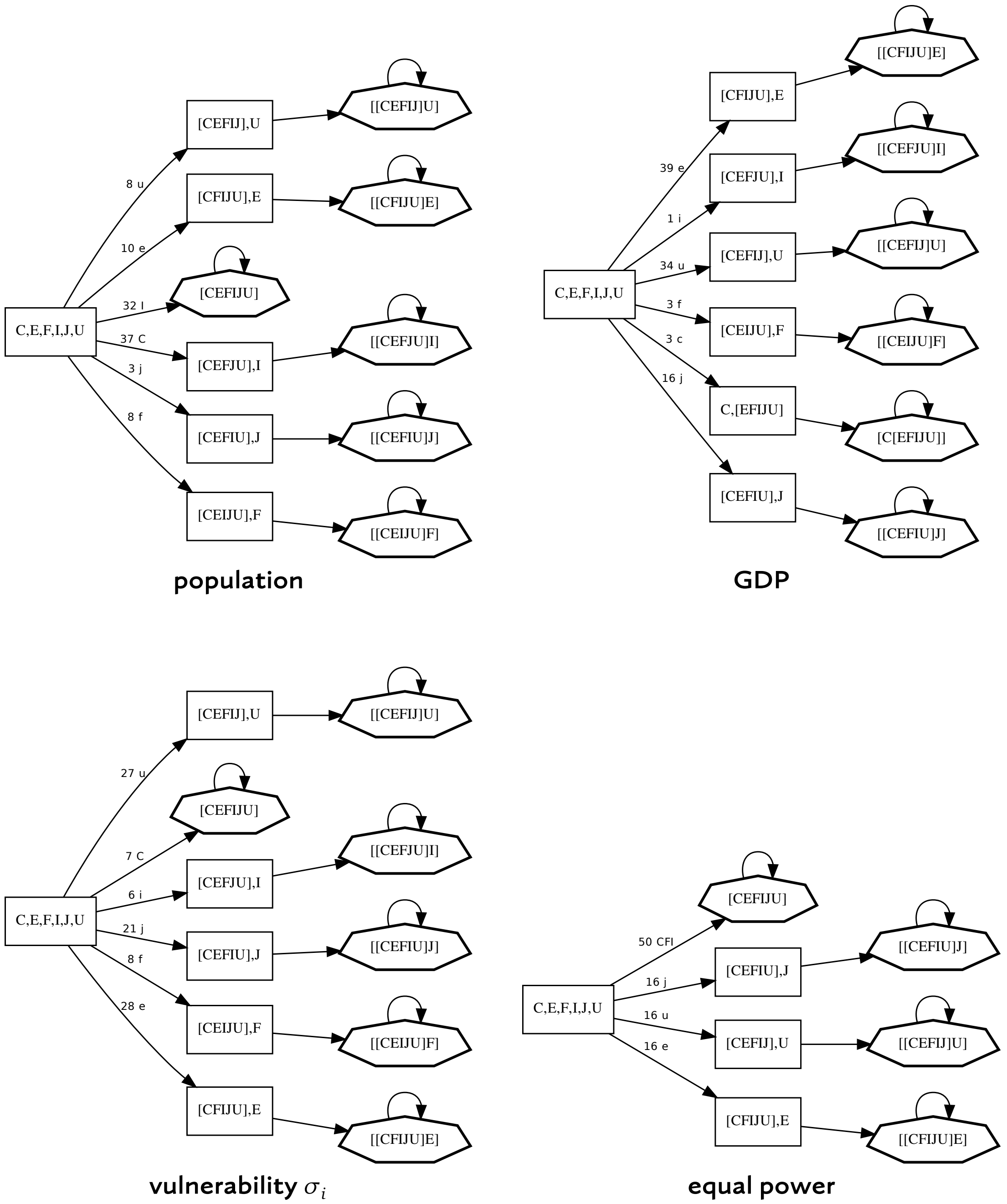
Tuerk, A., M. Mehling, C. Flachsland, and W. Sterk, "Linking carbon markets: concepts, case studies and pathways," *Climate Policy*, 2009, *9* (4), 341–357.



unanimous termination only

also unilateral termination

Figure 10: Effect of unilateral terminability on a process starting at state CU,EI,FJ with $\delta = 0.5$. J can now move into a better bargaining position by terminating its agreements with F, and this in turn changes the behaviour of C,F, and U in the beginning.



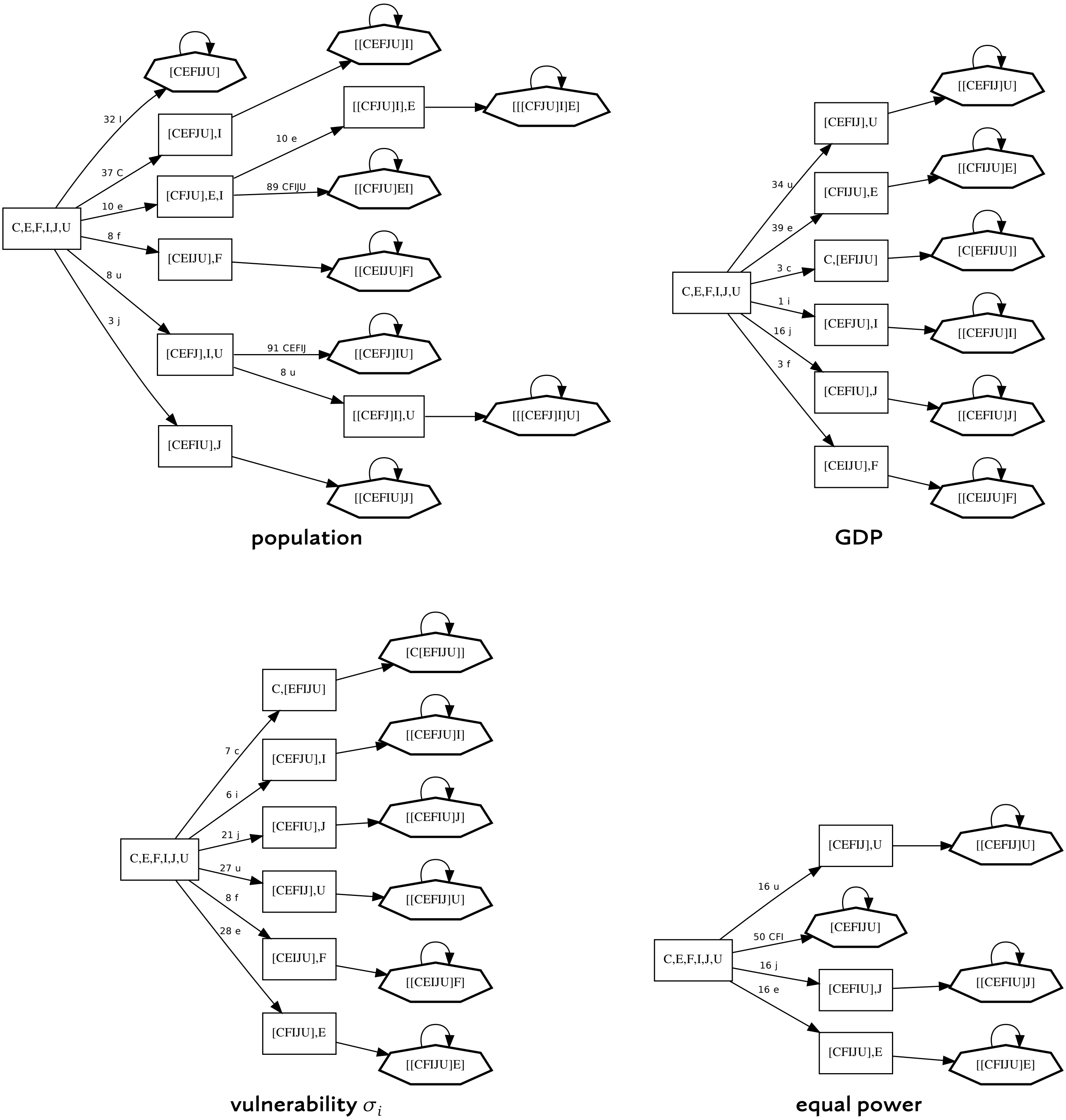
population

GDP

vulnerability σ_i

equal power

Figure 11: Alternatives to Fig. 1 for different bargaining power distributions (see Table 3) with low farsightedness ($\delta = 0.1$).



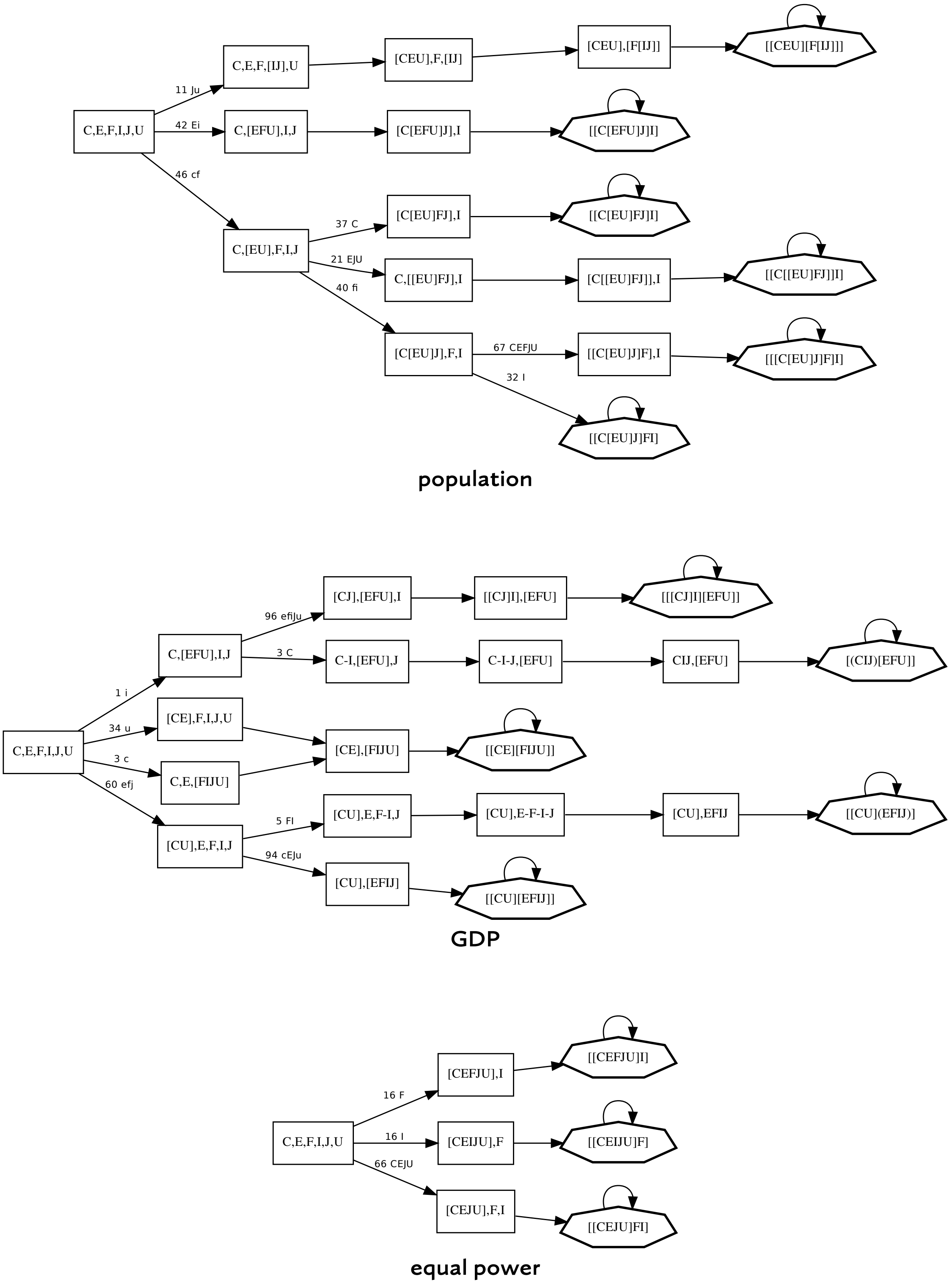
population

GDP

vulnerability σ_i

equal power

Figure 12: Alternatives to Fig. 1 for different bargaining power distributions (see Table 3) with medium farsightedness ($\delta = 0.5$).



population

GDP

equal power

Figure 13: Alternatives to Fig. 1 for different bargaining power distributions (see Table 3) with high farsightedness ($\delta = 0.9$). For power proportional to vulnerability σ_i, the search algorithm did not converge on an equilibrium process.

Appendix: A general dynamic model of farsighted coalition formation

Here we give a concise formal description of our model.

Equilibrium processes

A *process* $\mathscr{P} = (P, \mathscr{X}, \pi, \mathscr{F}, p)$ consists of...

- a finite set P of players,
- a finite set $\mathscr{X}$ of states,
- a function π assigning to each player $i \in P$ and each state $x \in \mathscr{X}$ a *static payoff* $\pi_i(x)$,
- a finite set $\mathscr{F}$ of *feasible moves* for P and $\mathscr{X}$, and
- a functon p assigning to each pair of states $x, y \in \mathscr{X}$ a *transition probability* $p_{x\to y} \in [0,1]$ with $\sum_{y\in\mathscr{X}} p_{x\to y} = 1$ for all $x \in \mathscr{X}$.

A *move* $x \xrightarrow{S} y$ for a set of players P and a set of states $\mathscr{X}$ consists of...

- an *origin state* $x \in \mathscr{X}$,
- a *target state* $y \in \mathscr{X}$,
- and a set of *initiators* $S \subseteq P$.

Let us fix a process $\mathscr{P}$ and a level of *farsightedness* $\delta \in (0,1)$. Then we define a function ℓ assigning to each player $i \in P$ and each state $x \in \mathscr{X}$ a *long-term payoff*

$$\ell_i(x) = (1-\delta)\pi_i(x) + \delta \sum_{y\in\mathscr{X}} p_{x\to y}\ell_i(y). \quad (15)$$

Note that this recursive definition leads to a system of linear equations which have a unique solution in general. A feasible move $(x \xrightarrow{S} y) \in \mathscr{F}$ is then called...

- *weakly mixed-profitable* iff $\ell_i(y) \geqslant \pi_i(x)$ for all $i \in S$ and $\ell_i(y) > \pi_i(x)$ for at least one $i \in S$,
- *weakly long-term profitable* iff $\ell_i(y) \geqslant \ell_i(x)$ for all $i \in S$ and $\ell_i(y) > \ell_i(x)$ for at least one $i \in S$,
- *mixed- [long-term] undominated* iff it is weakly mixed- [long-term] profitable and there is no weakly mixed- [long-term] profitable move $x \xrightarrow{T} z$ such that $T \subseteq S$ and $\ell_i(z) > \ell_i(y)$ for all $i \in S$.

Let $\mathscr{U}_\pi$ [$\mathscr{U}_\ell$] denote the set of all mixed- [long-term] undominated moves.

Given δ and a function w assigning to each player $i \in P$ some *bargaining power* $w_i > 0$, the process $\mathscr{P}$ (and by way of abbreviation the function p) is called a *mixed- [long-term] equilibrium process* iff the following holds for each $x \in \mathscr{X}$:

- If there exists no $(x \xrightarrow{S} y) \in \mathscr{U}_\pi$ [$\mathscr{U}_\ell$], then $p_{x\to x} = 1$.
- Otherwise,
  - for each player $i \in P$, there is a *favourite move* $(x \xrightarrow{S_i} y_i) \in \mathscr{U}_\pi$ [$\mathscr{U}_\ell$] such that

    $$\ell_i(y_i) \geqslant \ell_i(z) \text{ for all moves } (x \xrightarrow{T} z) \in \mathscr{U}_\pi\ [\mathscr{U}_\ell], \quad (16)$$

  - and for all $y \in \mathscr{X}$,

    $$p_{x\to y} = \sum_{i\in P,\, y_i = y} w_i \Big/ \sum_{j\in P} w_j. \quad (17)$$

Note that given P, $\mathscr{X}$, π, $\mathscr{F}$, δ, and w, there might in general be no, one, or several equilibrium processes p.

Algorithms

A set $\mathscr{F}$ of moves for P and $\mathscr{X}$ is *acyclic* iff there exist no $k > 1$, $x_1, \dots, x_k \in \mathscr{X}$, and $S_1, \dots, S_k \subseteq P$ such that $(x_k \xrightarrow{S_1} x_1) \in \mathscr{F}$ and $(x_{j-1} \xrightarrow{S_j} x_j) \in \mathscr{F}$ for all $j \in \{2, \dots, k\}$. Given a configuration P, $\mathscr{X}$, π, $\mathscr{F}$, δ, and w with acyclic $\mathscr{F}$, a corresponding mixed-equilibrium process p can be found using the following *backward induction* algorithm.

Iterate the following until $p_{x\to y}$ and $\ell_i(x)$ is known for all $x, y \in \mathscr{X}$ and all $i \in P$:

- Find an $x \in \mathscr{X}$ such that $\ell_i(x)$ is not yet known for all $i \in P$, but for all $(x \xrightarrow{S} y) \in \mathscr{F}$, $\ell_i(y)$ is already known for all $i \in P$.
- Determine the moves $(x \xrightarrow{S_i} y_i) \in \mathscr{U}_\pi$ from these known values $\ell_i(y)$.
- For each $i \in P$, pick a favourite move $(x \xrightarrow{S_i} y_i) \in \mathscr{U}_\pi$ that fulfills Eq. 16.
- Define $p_{x\to y}$ for all $y \in \mathscr{X}$ using Eq. 17.
- Calculate $\ell_i(x)$ for all $i \in P$ using Eq. 15.

Note that since $\mathscr{X}$ is finite and $\mathscr{F}$ is acyclic, there is at least one $x \in \mathscr{X}$ for which there is no $(x \xrightarrow{S} y) \in \mathscr{F}$, and the algorithm will start with these states. Also note that if π is generic, favourite moves are unique and hence p is the unique mixed-equilibrium process.

To search for a long-term equilibrium process, backward induction is not possible and we use the following iterative algorithm instead. Pick some initial assignment p^0 of transition probabilities with $\sum_{y \in \mathscr{X}} p_{x \to y} = 1$ for all $x \in \mathscr{X}$, then iterate the following for $k = 1, 2, \dots$ until $p^k = p^{k-1}$ or until $k = k_{\max}$:

- Calculate the function ℓ using Eq. 15 using p^{k-1} as p.
- Determine $\mathscr{U}_\pi$ from this ℓ.
- For each $x \in \mathscr{X}$ and $i \in P$, pick a favourite move $(x \xrightarrow{S_i} y_i) \in \mathscr{U}_\pi$ that fulfills Eq. 16.
- Determine the function p using Eq. 17 and put $p^k =$ this p.

Market structures, coalition hierarchies, states, codes, and moves

States. In the current version of our model, a *state* $x = (\mathscr{M}, \mathscr{K}, \mathscr{K}^+)$ is given by three families of subsets of P, namely...

- a partition $\mathscr{M}$ of P, called the *market structure* in x;
- a family $\mathscr{K}$ of non-empty subsets of P, called the *coalition hierarchy* in x, such that $\mathscr{K}$ is a clustering tree, i.e.,
  - there is a "singleton" coalition $\{i\} \in \mathscr{K}$ for each $i \in P$,
  - for each coalition $K \in \mathscr{K}$ there is (a unique) $M \in \mathscr{M}$ with $M \supseteq K$,
  - for all $K, K' \in \mathscr{K}$, either $K \subseteq K'$ or $K \supseteq K'$ or $K \cap K' = \emptyset$;
- and a subset $\mathscr{K}^+ \subseteq \mathscr{K}$, called the set of *immediately coordinated* coalitions in x, such that for all $K \in \mathscr{K}^+$,
  - $|K| > 1$,
  - $M \in \mathscr{K}^+$ for that $M \in \mathscr{M}$ with $M \supseteq K$, and
  - $K' \in \mathscr{K}^+$ for all $K' \in \mathscr{K}$ with $K' \supseteq K$.

The set of *signatories* of a non-singleton coalition $K \in \mathscr{K}$ with $|K| > 1$ is

$$\mathscr{S}(K, \mathscr{K}) = \{K' \in \mathscr{K} : K' \subset K, \\ \neg\exists K'' \in \mathscr{K}\,(K' \subset K'' \subset K)\}, \quad (18)$$

where $\subset$ denotes *proper* set inclusion. The set of *top-level* coalitions in a market $M \in \mathscr{M}$ is

$$\mathscr{T}(M, \mathscr{K}) = \{K \in \mathscr{K} : K \subseteq M, \\ \forall K' \in \mathscr{K}\,(K' \not\supset K)\}. \quad (19)$$

When dealing with several states $x, y, \dots$, we designate their respective components $\mathscr{M}, \mathscr{K}, \mathscr{K}^+$ by adding and index $x, y, \dots$, so that $x = (\mathscr{M}_x, \mathscr{K}_x, \mathscr{K}_x^+)$, $y = (\mathscr{M}_y, \mathscr{K}_y, \mathscr{K}_y^+), \dots$.

Codes. We use the following notation for individual players and states:

- Each player i and each singleton coalition $\{i\}$ is represented by an upright uppercase letter, e.g. $i = \mathrm{C}$.
- Each non-singleton coalition $K \in \mathscr{K} \setminus \mathscr{K}^+$ is represented by a *coalition code* which is a concatenation of the coalition codes of all $K' \in \mathscr{S}(K, \mathscr{K})$ in lexicographic order, placed between round brackets, e.g., $K =$ (C(EJ)(FI)). The outermost round brackets can be dropped if no confusion can arise, i.e., $K =$ C(EJ)(FI).
- Each coalition $K \in \mathscr{K}^+$ is represented by a *coalition code* which is a concatenation of the coalition codes of all $K' \in \mathscr{S}(K, \mathscr{K})$ in lexicographic order, placed between square brackets, e.g., $K =$ [C[EJ](FI)]. The outermost square brackets are not dropped.
- A market $M \in \mathscr{M}$ is represented by a *market code* which is a concatenation of the coalition codes of all $K \in \mathscr{T}(M, \mathscr{K})$ in lexicographic order, separated by dashes, e.g., $M =$ C-(EJ)U-FI or $M =$ [C[EJ](FI)]. Note that because of the requirements on $\mathscr{K}^+$, no market code can contain both dashes and square brackets.[10]
- A state $x = (\mathscr{M}, \mathscr{K}, \mathscr{K}^+)$ is represented by a *state code* which is a concatenation of the

[10] This formalized the assumption that once a market is formed with immediate cap coordination, it can no longer be linked further without immediate cap coordination.

market codes of all $M \in \mathscr{M}$ in lexicographic order, separated by commas, e.g., $x =$ C-(EJ)U,FI.

Irreversible agreements. In the *irreversible agreements model,* $\mathscr{F}$ is acyclic and consists of three types of feasible moves, uncoordinated market linkage, cap coordination, and coordinated market linkage.

An *uncoordinated market linkage* move is a move $x \xrightarrow{S} y$ in which a subset $\mathscr{M}'$ of at least two of the markets in x that do not contain any immediately coordinated coalitions is linked to establish one larger market, while the remaining markets and the coalition hierarchy remain unchanged. The initiators of an uncoordinated market linkage move are the members of the new market. Formally,

$$\begin{aligned} \mathscr{M}' &\subseteq \mathscr{M}_x \setminus \mathscr{K}_x^+, \quad |\mathscr{M}'| \geqslant 2, \\ S &= \bigcup \mathscr{M}', \\ \mathscr{M}_y &= (\mathscr{M}_x \setminus \mathscr{M}') \cup \{S\}, \\ \mathscr{K}_y &= \mathscr{K}_x, \\ \mathscr{K}_y^+ &= \mathscr{K}_x^+. \end{aligned} \tag{20}$$

For example, in $x =$ A,B,C-F,D,E, the players $S =$ BDE might initiate the linkage of the existing (domestic) markets B, D, and E, establishing an uncoordinated (international) market B-D-E, which is a move to $y =$ A,B-D-E,C-F, denoted

$$\text{A,B,C-F,D,E} \xrightarrow{\text{BDE}} \text{A,B-D-E,C-F.}$$

A *cap coordination* move is a move $x \xrightarrow{S} y$ in which a subset $\mathscr{T}'$ of at least two of the top-level coalitions of a market in x forms an additional (overarching) coalition, while all existing coalitions, including those in $\mathscr{T}'$, remain intact and the market structure remains unchanged. The initiators of a cap coordination move are the members of the new coalition. Formally,

$$\begin{aligned} M &\in \mathscr{M}_x, \quad \mathscr{T}' \subseteq \mathscr{T}(M, \mathscr{K}_x), \quad |\mathscr{T}'| \geqslant 2, \\ S &= \bigcup \mathscr{T}', \\ \mathscr{M}_y &= \mathscr{M}_x, \\ \mathscr{K}_y &= \mathscr{K}_x \cup \{S\}, \\ \mathscr{K}_y^+ &= \mathscr{K}_x^+. \end{aligned} \tag{21}$$

For example, in $x =$ A,BE-D,C-F, the players $S =$ BDE might initiate that the existing coalition BE and the existing one-member "coalition" D form an overarching new coalition (BE)D, which is a move to $y =$ A,(BE)D,C-F, denoted

$$\text{A,BE-D,C-F} \xrightarrow{\text{BDE}} \text{A,(BE)D,C-F.}$$

A *coordinated market linkage* move is a move $x \xrightarrow{S} y$ in which a subset $\mathscr{M}'$ of at least two of those markets in x which have only one top-level coalition is linked to establish one larger market, which is also added to the coalition hierarchy as a new immediately coordinated top-level coalition, while the remaining markets and their coalition hierarchy remain unchanged. The initiators of a coordinated market linkage move are the members of the new market. Formally,

$$\begin{aligned} \mathscr{M}' &\subseteq \mathscr{M}_x, \quad |\mathscr{M}'| \geqslant 2, \\ |\mathscr{T}(M, \mathscr{K})| &= 1 \text{ for all } M \in \mathscr{M}, \\ S &= \bigcup \mathscr{M}', \\ \mathscr{M}_y &= (\mathscr{M}_x \setminus \mathscr{M}') \cup \{S\}, \\ \mathscr{K}_y &= \mathscr{K}_x \cup \{S\}, \\ \mathscr{K}_y^+ &= \mathscr{K}_x^+ \cup \{S\}. \end{aligned} \tag{22}$$

Unanimously terminable agreements. In the *unanimously terminable agreements model,* $\mathscr{F}$ is no longer acyclic but contains three additional types of moves, consisting of the exact inverses of the above types:

If $y \xrightarrow{S} x$ is an uncoordinated market linkage move, then $x \xrightarrow{S} y$ is a *unanimous uncoordinated market splitting* move. In other words, a unanimous uncoordinated market splitting move is a move $x \xrightarrow{S} y$ in which a market M in x is split into at least two smaller markets, where each top-level coalition stays in one of the new markets, while the remaining markets and the coalition hierarchy remain unchanged. The initiators are the members of the old market. For example, in $x =$ A,BE-D,C-F, the players $S =$ BDE might initiate the splitting of the uncoordinated (international) market BE-D into an (international) market BE and a (domestic) market D, which is a move to $y =$ A,BE,C-F,D, denoted

$$\text{A,BE-D,C-F} \xrightarrow{\text{BDE}} \text{A,BE,C-F,D.}$$

If $y \xrightarrow{S} x$ is a cap coordination move, then $x \xrightarrow{S} y$ is a *unanimous cap de-coordination* move. In other words, a unanimous cap de-coordination move is a move $x \xrightarrow{S} y$ in which an existing top-level coalition K in x is removed from the coalition hierarchy, while all other existing coalitions, including those contained in K, remain intact and the market structure remains unchanged. The initiators are the members of the terminated coalition. For example, in $x =$

A,(BE)D,C-F, the players $S = \text{BDE}$ might initiate that the existing coalition (BE)D is terminated, leaving the coalition BE intact, which is a move to $y = $ A,BE-D,C-F, denoted

$$\text{A,(BE)D,C-F} \xrightarrow{\text{BDE}} \text{A,BE-D,C-F.}$$

If $y \xrightarrow{S} x$ is a coordinated market linkage move, then $x \xrightarrow{S} y$ is a *unanimous coordinated market splitting* move. In other words, a unanimous coordinated market splitting move is a move $x \xrightarrow{S} y$ in which a market M in x that was formed by coordinated market linkage is split into at least two smaller markets, where each signatory of the lone top-level coalition of M stays in one of the new markets, while the remaining markets and coalition hierarchy remain unchanged. The initiators are the members of the old market. For example, in $x = $ A,[(BE)D],C-F, the players $S = \text{BDE}$ might initiate the splitting of the immediately coordinated (international) market [(BE)D] into an (international) market BE and a (domestic) market D, which is a move to $y = $ A,BE,C-F,D, denoted

$$\text{A,[(BE)D],C-F} \xrightarrow{\text{BDE}} \text{A,BE,C-F,D.}$$

Unilaterally terminable agreements. In the *unilaterally terminable agreements model,* $\mathscr{F}$ contains two more types of moves:

A *unilateral uncoordinated market splitting* move is a move $x \xrightarrow{S} y$ in which a top-level coalition K of a market M of x that has at least two top-level coalitons leaves M, causing a complete split of M into as many markets as M has top-level coalitions, so that each top-level coalition of M stays in its own market, while the remaining markets and the coalition hierarchy remain unchanged. The initiators are the members of the leaving coalition. Formally,

$$\begin{aligned}
&M \in \mathscr{M}_x, \quad K \in \mathscr{T}(M, \mathscr{K}_x), \quad |\mathscr{T}(M, \mathscr{K}_x)| \geqslant 2,\\
&\quad S = K,\\
&\mathscr{M}_y = (\mathscr{M}_x \setminus \{M\}) \cup \mathscr{T}(M, \mathscr{K}_x), \qquad (23)\\
&\mathscr{K}_y = \mathscr{K}_x,\\
&\mathscr{K}_y^+ = \mathscr{K}_x^+.
\end{aligned}$$

For example, in $x = $ A-BE-D,C-F, the players $S = \text{BE}$ might leave the market A-BE-D and thereby unilaterally initiate its the splitting into three markets A, BE, and D, which is a move to $y = $ A,BE,C-F,D, denoted

$$\text{A-BE-D,C-F} \xrightarrow{\text{BE}} \text{A,BE,C-F,D.}$$

A *unilateral cap de-coordination* move is a move $x \xrightarrow{S} y$ in which an existing *non*-top-level coalition K in x in a market that was formed by uncoordinated market linkage leaves all higher-level coalitions in which it is contained, causing a removal of all those coalitions from the coalition hierarchy, while all other existing coalitions, including K and those contained in K, remain intact and the market structure remains unchanged. The initiators are the members of the leaving coalition. Formally,

$$\begin{aligned}
&M \in \mathscr{M}_x \setminus \mathscr{K}_x^+, \quad K \subseteq M, \quad K \in \mathscr{K}_x \setminus \mathscr{T}(M, \mathscr{K}_x)\\
&\quad S = K,\\
&\mathscr{M}_y = \mathscr{M}_x, \qquad (24)\\
&\mathscr{K}_y = \mathscr{K}_x \setminus \{K' \in \mathscr{K}_x : K' \supset K\},\\
&\mathscr{K}_y^+ = \mathscr{K}_x^+.
\end{aligned}$$

For example, in $x = $ A((BE)D),C-F, the players $S = \text{BE}$ might leave all higher-up coalitions and thereby unilaterally initiate the removal of the coalitions (BE)D and A((BE)D), leaving the coalition BE intact, which is a move to $y = $ A-BE-D,C-F, denoted

$$\text{A((BE)D),C-F} \xrightarrow{\text{BE}} \text{A-BE-D,C-F.}$$

A *unilateral coordinated market splitting* move is a move $x \xrightarrow{S} y$ in which an existing *non*-top-level coalition K in x in a market that was formed by coordinated market linkage leaves all higher-level coalitions in which it is contained, causing a removal of all those coalitions from the coalition hierarchy and a splitting of the market into the resulting top-level coalitions, while all other existing coalitions, including K and those contained in K, remain intact and the other markets remains unchanged. The initiators are the members of the leaving coalition. Formally,

$$\begin{aligned}
&M \in \mathscr{M}_x \cap \mathscr{K}_x^+, \quad K \subseteq M, \quad K \in \mathscr{K}_x \setminus \mathscr{T}(M, \mathscr{K}_x)\\
&\quad S = K,\\
&\mathscr{K}_y = \mathscr{K}_x \setminus \{K' \in \mathscr{K}_x : K' \supset K\},\\
&\mathscr{K}_y^+ = \mathscr{K}_x^+ \setminus \{K' \in \mathscr{K}_x^+ : K' \supset K\} \qquad (25)\\
&\mathscr{M}_y = (\mathscr{M}_x \setminus \{M\}) \cup\\
&\qquad \{K' \in \mathscr{K}_y : K' \subset M, \forall K'' \in \mathscr{K}_y \, (K'' \not\supset K')\}.
\end{aligned}$$

For example, in $x = $ [A[(BE)D]],C-F, the players $S = \text{BE}$ might leave all higher-up coalitions and thereby unilaterally initiate the removal of the coalitions [(BE)D] and [A((BE)D)] and the splitting of the market into three markets A,BE,D, leaving the coalition BE intact, which is a move to $y = $ A,BE,C-F,D, denoted

$$[A[(BE)D]],C\text{-}F \xrightarrow{BE} A,BE,C\text{-}F,D.$$

The Random Proposer Amended Move Bargaining Game

In this section, we give a game-theoretic foundation to the assumption that move probabilities are proportional to bargaining power, by deriving these probabilities from a certain Rubinstein-like game based on a specific bargaining protocol.

Let us assume in each period of the process of coalition formation, the next move from the *going state* x is determined by the *Random Proposer Amended Move Bargaining Game at* x, consisting of a potentially infinite number of *rounds* all occurring within the same period of the coalition formation process. Each round consists of the following stages:

1. Nature chooses a *proposer* $i \in P$ at random according to an exogenously given distribution $w_i \geqslant 0$ of *bargaining power,*

2. The proposer chooses a feasible move $x \xrightarrow{S} y$ as the proposed *focal motion.*

3. Nature sorts the $k = |S| \geqslant 1$ many initiators in some arbitrary way into an order $S = \{j_1, \dots, j_k\}$.

4.1. . . 4.k
If $i \in S$, each of the initiators $j_1, \dots, j_k$ in turn chooses to *reject* or *accept* the motion. If $i \notin S$, j_1 chooses to *reject* or *accept* the focal motion or proposes a different move $x \xrightarrow{T} z$ with $z \neq y$ and $T \subseteq S$ as the *amendment,* and then each of the remaining initiators $j_2, \dots, j_k$ in turn chooses to *reject,* to *accept the focal motion,* or to *accept the amendment,* where the latter option only exists if j_1 did propose an amendment.

As soon as a player rejects, the remaining stages of the going round are skipped and the next round starts after an infinitesimal delay. If no player rejects and either no amendment was proposed or some player did not accept the amendment, the game ends and the focal motion is realized. If all players accept an amendment $x \xrightarrow{T} z$, $x \xrightarrow{T} z$ becomes the focal motion, i.e., S is replaced by T and $x \xrightarrow{S} y$ by $x \xrightarrow{T} z$, and stages 3 and 4 are repeated. If the game does not end, the move $x \xrightarrow{\emptyset} x$ is realized.

Now let L be the lottery which selects each of the long-term undominated moves $x \xrightarrow{S} y$ with a probability proportional to the aggregate bargaining power of those players whose favourite long-term undominated move at x is $x \xrightarrow{S} y$, assuming each player has a strict preference among each pair of moves at x. Then it is easy to see that the following is a subgame-perfect Nash equilibrium of the Random Proposer Amended Move Bargaining Game at x that consists of pure Markov strategies only and implements the lottery L:

- When you are the proposer i, propose your favourite long-term undominated move $x \xrightarrow{S} y$.

- When $i \in S$ and you are an initiator j_ℓ, accept the focal motion if you weakly prefer it to the lottery L, else reject it.

- When $i \notin S$ and you are j_1, check whether there is a different move $x \xrightarrow{T} z$ with $z \neq y$ and $T \subseteq S$ that all $j \in S$ strictly prefer to both $x \xrightarrow{S} y$ and to the lottery L. If so, propose your favourite such move $x \xrightarrow{T} z$ as the amendment. Otherwise, accept the focal motion $x \xrightarrow{S} y$ if you weakly prefer it to the lottery L, else reject it.

- When $i \notin S$ and you are one of $j_2, \dots, j_k$, accept the amendment $x \xrightarrow{T} z$ if one was suggested and you strictly prefer it to $x \xrightarrow{S} y$, otherwise accept the focal motion $x \xrightarrow{S} y$ if you weakly prefer it to the lottery L, else reject it.

Without the amendment option, there is a similar implementation of those lotteries that select each of the weakly long-term profitable moves, whether dominated or not, with a probability proportional to the aggregate bargaining power of those players favouring this weakly long-term profitable move.